



Investigation of protein structure sustainability and bio-inspiration for urban systems

Lorenza Pacini

► To cite this version:

Lorenza Pacini. Investigation of protein structure sustainability and bio-inspiration for urban systems. Bioengineering. Université de Lyon, 2021. English. NNT : 2021LYSE1195 . tel-03640284

HAL Id: tel-03640284

<https://theses.hal.science/tel-03640284>

Submitted on 13 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



N°d'ordre NNT : 2021LYSE1195

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale N° 160
(Electronique, Electrotechnique, Automatique de Lyon)

Spécialité de doctorat : Ingénierie pour le vivant

Soutenue publiquement le 24/09/2021, par :
Lorenza Pacini

Investigation of Protein Structure Sustainability and Bio-Inspiration for Urban Systems

*Investigation de la durabilité des structures protéiques et bio-
inspiration pour les systèmes urbains*

Devant le jury composé de :

Keskin, Ozlem	Professeure	Université Koc	Rapporteure
Leitner, David	Professeur	Université du Nevada, Reno	Rapporteur
Di Paola, Luisa	Professeure Associée	Université de Rome	Examinatrice
Hologne, Maggy	Maître de Conférences	Université Lyon 1	Examinatrice
Menezo, Christophe	Professeur des Universités	Université Savoie Mont Blanc	Examineur
Schabanel, Nicolas	Directeur de Recherche	CNRS	Examineur
Lesieur, Claire	Chargée de Recherche	CNRS	Directrice de thèse
Vuillon, Laurent	Professeur des Universités	Université Savoie Mont Blanc	Co-directeur de thèse

Université Claude Bernard – LYON 1

Président de l'Université	M. Frédéric FLEURY
Président du Conseil Académique	M. Hamda BEN HADID
Vice-Président du Conseil d'Administration	M. Didier REVEL
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. Philippe CHEVALLIER
Vice-Président de la Commission de Recherche	M. Jean-François MORNEX
Directeur Général des Services	M. Pierre ROLLAND

COMPOSANTES SANTE

Département de Formation et Centre de Recherche en Biologie Humaine	Directrice : Mme Anne-Marie SCHOTT
Faculté d'Odontologie	Doyenne : Mme Dominique SEUX
Faculté de Médecine et Maïeutique Lyon Sud - Charles Mérieux	Doyenne : Mme Carole BURILLON
Faculté de Médecine Lyon-Est	Doyen : M. Gilles RODE
Institut des Sciences et Techniques de la Réadaptation (ISTR)	Directeur : M. Xavier PERROT
Institut des Sciences Pharmaceutiques et Biologiques (ISBP)	Directrice : Mme Christine VINCIGUERRA

COMPOSANTES & DEPARTEMENTS DE SCIENCES & TECHNOLOGIE

Département Génie Electrique et des Procédés (GEP)	Directrice : Mme Rosaria FERRIGNO
Département Informatique	Directeur : M. Behzad SHARIAT
Département Mécanique	Directeur M. Marc BUFFAT
Ecole Supérieure de Chimie, Physique, Electronique (CPE Lyon)	Directeur : Gérard PIGNAULT
Institut de Science Financière et d'Assurances (ISFA)	Directeur : M. Nicolas LEBOISNE
Institut National du Professorat et de l'Education	Administrateur Provisoire : M. Pierre CHAREYRON
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe VITON
Observatoire de Lyon	Directrice : Mme Isabelle DANIEL
Polytechnique Lyon	Directeur : Emmanuel PERRIN
UFR Biosciences	Administratrice provisoire : Mme Kathrin GIESELER
UFR des Sciences et Techniques des Activités Physiques et Sportives (STAPS)	Directeur : M. Yannick VANPOULLE
UFR Faculté des Sciences	Directeur : M. Bruno ANDRIOLETTI

Thèse réalisée aux laboratoires:

AMPERE (UMR 5005)

Université Claude Bernard Lyon 1
43 Boulevard du 11 Novembre 1918
69622 Villeurbanne cedex



IXXI

École Normale Supérieure de Lyon
46 Allée d'Italie
69007 Lyon



Abstract

Proteins are sustainable as they keep their function in time and in different conditions. As a first approximation of functional behavior, we investigate how the structure design yields sustainability, concentrating on the problem of diagnosing of structural and dynamical perturbations caused by mutations of amino-acid, basic components of proteins. This is relevant to decipher how diseases arise from mutations. We also open the possibility of designing sustainable urban systems bio-inspired from proteins. Proteins and cities are modeled using weighted spatial networks that measure space occupancy by amino acids and buildings, respectively. This allows inferring the empty space available for the dynamics. Proteins have empty space that we assume to be exploited for the accommodation of mutations. In contrast, in the city high packing is present, suggesting spatial unsustainability (impossibility to change). We verify the relation between local packing and the impact of mutations on the protein structure and dynamics. We find that mutations keeping structural integrity require links rearrangement at diverse scales that will perturb the protein dynamics depending on the impacted structural level (2D, 3D, 4D). We propose a tool to describe the direction of the perturbations and extract the impact of multiple mutations acting collectively. This work contributes to the decoding of the protein's design for sustainability, opening perspectives in the fields of biomedicine and the design of sustainable bio-inspired systems.

Keywords: sustainability, protein structure, urban structure, biomimicry, network analysis, complex systems, amino-acid mutation, protein dynamics.

Résumé

Les protéines sont durables car elles conservent la fonction dans le temps et dans différentes conditions. Comme première approximation du comportement fonctionnel, nous étudions comment la conception de la structure assure la durabilité, en nous concentrant sur le problème du diagnostic des perturbations structurales et dynamiques causées par les mutations des acides aminés, composants élémentaires des protéines. Cette question est pertinente pour déchiffrer comment les maladies résultent des mutations. Nous ouvrons également la possibilité de concevoir des systèmes urbains durables bio-inspirés des protéines. Protéines et villes sont modélisées par des réseaux spatiaux pondérés qui mesurent l'occupation de l'espace par acides aminés et bâtiments, respectivement. Cela permet d'inférer l'espace vide disponible pour la dynamique. Les protéines ont de l'espace vide que nous supposons être exploité pour accueillir des mutations. En revanche, dans la ville, un encombrement élevé est présent, suggérant une non-durabilité spatiale (impossibilité de changer). Nous vérifions la relation entre encombrement local et impact des mutations sur la structure et la dynamique des protéines. Nous constatons que les mutations conservant l'intégrité structurale nécessitent un réarrangement des liens à diverses échelles qui va perturber la dynamique de la protéine en fonction du niveau structural impacté (2D, 3D, 4D). Nous proposons un outil pour décrire la direction des perturbations et extraire l'impact de mutations multiples agissant collectivement. Ce travail contribue au décodage du design pour la durabilité des protéines, ouvrant des perspectives dans les domaines de la biomédecine et la conception de systèmes bio-inspirés durables.

Mots clés : durabilité, structure protéique, structure urbaine, biomimétisme, analyse de réseaux, systèmes complexes, mutation d'acides aminés, dynamique protéique.

Contents

List of Figures	11
List of Tables	17
List of Abbreviations	18
1 Introduction	19
2 State of the art	26
2.1 Sustainability	26
2.1.1 Defining sustainability	26
2.1.2 Designing a sustainable system	28
2.2 Proteins as functional spatial systems	29
2.2.1 The amino acid sequence	29
2.2.2 The protein structure	31
2.2.3 The protein dynamics	34
2.2.4 The protein function	35
2.2.5 Then things get complicated	35
2.3 Sustainability of proteins	37
2.4 Proteins as Complex Systems	39
2.4.1 Network models of protein structures	39
2.4.2 Complex Networks	40
2.4.3 Network Analysis methods	40
2.5 Applications of network models of protein structures	44
2.6 Cities as functional spatial Complex Systems	49
3 Methods	53
3.1 Protein structure analysis	53
3.1.1 Protein structures data	53
3.1.2 Amino Acid Network	53
3.1.3 Perturbation Network and Induced Perturbation Network	55

3.1.4	Electrostatic Network	57
3.1.5	Local structural-level allocation of the atomic interactions	58
3.2	In-silico mutagenesis	59
3.2.1	In-silico mutations	59
3.2.2	Classification of in-silico amino acid mutations	59
3.2.3	GCAT Network	63
3.3	Urban Structures analysis	64
3.3.1	Building Network	64
3.3.2	Agent-based modeling of random mobility in urban systems	69
4	Database analysis: Can the protein- and urban structures accommodate substitutions of their components without structural rearrangement?	70
4.1	Introduction	71
4.2	Measuring space occupancy in protein- and urban structures.	72
4.2.1	Spatial network models	72
4.2.2	Node and link measures in the spatial network models	73
4.2.3	Case studies of urban structures	74
4.2.4	Case studies of protein structures	76
4.3	Statistics of space occupancy in proteins and cities	78
4.3.1	Proteins database	78
4.3.2	Buildings database: the city of Lyon	83
4.4	Conclusion	87
5	Database analysis: If necessary, can protein- and urban structures accommodate substitutions of their components with structural rearrangement?	90
5.1	Introduction	91
5.2	Jamming of granular materials	92
5.3	The analogy between granular materials, proteins and cities	93
5.4	Detecting “force chains” in spatial networks	94
5.5	Are proteins and urban structures jammed?	95
5.5.1	Protein structures	96
5.5.2	Urban structures	97
5.5.3	Agent-based modeling of random mobility in urban systems	99
5.6	Conclusion	101

6 Case study: Third PDZ domain. Means to accommodate substitutions in the protein structure.	103
6.1 Introduction	104
6.2 Classification of in-silico amino-acid mutations based on the scale of neighborhoods rearrangement they provoke	104
6.3 In-silico mutagenesis of PSD95 ^{pdz3}	107
6.4 Conclusion	112
7 Database analysis: How to measure protein dynamics from a static structure?	114
7.1 Introduction	115
7.2 Classification of atomic interactions into structural levels	116
7.3 Statistics of link weights in the four structural levels in the proteins database	118
7.4 Measure and visualization of the local structural-level allocation of the atomic interactions	123
7.5 Statistics of the local structural-level allocation of the atomic interactions in the proteins database	124
7.6 Conclusion	128
8 Case study: Transthyretin, B-subunits toxin pentamers, Main Proteases of SARS coronaviruses. Measure of the impact of amino-acid mutations on the protein dynamics.	130
8.1 Introduction	131
8.2 Comparison of local structural-level allocation of atomic interactions in protein variants	132
8.2.1 B-subunit pentamers of AB ₅ toxins	132
8.2.2 Transthyretin variants	134
8.2.3 Main proteases of SARS coronaviruses	136
8.3 Conclusion	138
9 Case study: Transthyretin. Large scale changes in dynamics upon mutation leading to amyloid fiber formation.	139
9.1 Introduction	140
9.2 Induced Perturbation Network	140
9.3 Tiling model of fiber formation	142
9.4 L55P Transthyretin amyloid fiber model	143
9.4.1 Structural comparison of Transthyretin variants from node properties in the AAN	143

9.4.2	Dynamical comparison of Transthyretin variants from link weights in the AAN	144
9.4.3	Tiling model of the L55P TTR amyloid fiber	147
9.5	Conclusion	149
10	Case study: B-subunits toxin pentamers. Integrative approaches to protein dynamics perturbations related to mutations.	150
10.1	Introduction	151
10.2	Experimental measure of protein multi-scale dynamics	151
10.2.1	Nanoconfinement of proteins	151
10.2.2	Broadband Dielectric Spectroscopy	152
10.2.3	Experimental protocol	154
10.3	Electrostatic Network models of dipole interactions in the protein structure	155
10.4	Integrative approach to probe multi-scale dynamics of the CtxB ₅ toxin . .	156
10.4.1	Experimental measure: Broadband Dielectric Spectroscopy	156
10.4.2	Electrostatic Network model to validate the peak-assignment of BDS signal	158
10.5	Measure of multi-scale dynamical perturbation upon mutation: CtxB ₅ ver- sus LTB ₅ toxins	163
10.5.1	Differences in experimental molecular dynamics	163
10.5.2	Electrostatic Network models	166
10.6	Conclusion	169
11	Proof of concept: Third PDZ domain. Directions of perturbations caused by amino-acid mutations in the protein structure measured by a directed network.	171
11.1	Introduction	172
11.2	The GCAT network tool	173
11.3	Exploratory analysis of the GCAT network of PSD95 ^{pdz3}	176
11.4	Conclusion	194
12	Conclusion and perspectives	196
A	Proteins database	205
B	Supplementary Figures	211
C	Résumé étendu	214
C.1	Introduction	214

C.2	État de l'art	217
C.3	Méthodes	222
C.4	Analyse de la base des données : Les structures protéiques et urbaines peuvent-elles s'adapter aux substitutions de leurs composants sans réarrangement structurel ?	227
C.5	Analyse de la base de données : Si nécessaire, les structures protéiques et urbaines peuvent-elles s'adapter aux substitutions de leurs composants par le biais d'un réarrangement structurel ?	231
C.6	Étude de cas : Troisième domaine PDZ. Moyens pour s'adapter aux sub- stitutions dans la structure protéique.	235
C.7	Analyse de base de données : Comment mesurer la dynamique des protéines à partir d'une structure statique ?	238
C.8	Étude de cas : Transthyréline, pentamères des sous-unités B de toxines, protéases principales des coronavirus du SRAS. Mesure de l'impact des mutations des acides aminés sur la dynamique des protéines.	241
C.9	Étude de cas : Transthyréline. Changements à grande échelle de la dy- namique lors de la mutation conduisant à la formation de fibres amyloïdes.	245
C.10	Étude de cas : pentamères des sous-unités B de toxines. Approches intégratives des perturbations de la dynamique des protéines liées aux mutations.	247
C.11	Preuve de concept : Troisième domaine PDZ. Directions des perturbations causées par les mutations des acides aminés dans la structure des protéines mesurées par un réseau dirigé.	252
C.12	Conclusion et perspectives	256
Bibliography		259
Acknowledgements		274

List of Figures

2.1	Schematics of the proposed classification of Sustainability approaches. . . .	27
2.2	Schematics of the system design procedures.	29
2.3	The twenty amino acids.	30
2.4	Protein structure visualization.	31
2.5	The four structural levels of the protein structure.	32
2.6	Domains in protein structures.	33
2.7	From the protein sequence to the protein function.	36
2.8	Examples of Network Measures on Network models and on the Amino Acid Network of a protein.	41
3.1	Amino Acids Network (AAN) of the human wild-type Transthyretin dimer (PDB: 1F41).	55
3.2	Link weight assignment in the AAN	55
3.3	Space occupancy around amino acids in the Amino Acid Network.	56
3.4	Perturbation Network and Induced Perturbation Network of a toy example.	57
3.5	Relations between the AAN, the Electrostatic Network, the Induced Elec- trostatic Network and the equivalent intermolecular networks	58
3.6	Flow chart of the classification procedure of the in-silico amino acid muta- tions.	62
3.7	Schematics of the construction of the GCAT network.	64
3.8	Definition of the Building Network.	67
3.9	Space occupancy around buildings in the Buildings Network.	68
4.1	Schematics of the substitution of a component in a spatial system.	71
4.2	Spatial network models of the protein and urban structure.	72
4.3	Schematics of the space-occupancy measures in the spatial networks. . . .	73
4.4	Neighborhood watch in the BN of the area of Monplaisir in Lyon (France).	75
4.5	Building Networks (BNs) of three urban structures.	75
4.6	Amino Acid Network (AAN) of three protein structures.	77

4.7	Statistics of the node and link properties in the Amino Acid Networks (AANs) the database of globular proteins.	78
4.8	Mean values of the Neighborhood watch measure in the AANs of the proteins in the database.	80
4.9	Statistics of the node properties in the database of globular proteins based on the amino acid type.	81
4.10	Examples of the spatial orientation of the links in the AAN when all links are considered (left) or only 2D links are considered (right) in β -strands (top) and α -helices (bottom)	82
4.11	Distribution of the Neighborhood watch measure in two-dimensional Amino Acid Networks of the protein database.	83
4.12	Building Network (BN) of the city of Lyon.	84
4.13	Form factor of the buildings of Lyon.	84
4.14	Statistics of the node and link properties in the Building Network (BN) of the city of Lyon (France).	85
4.15	Neighborhood watch versus the geometry of the buildings in the buildings database.	86
4.16	Examples of building neighborhoods involving the same number of buildings ($k = 6$) but made by buildings of different size and resulting in different values of node weight w	87
5.1	Schematics of the substitution of a component in a spatial system through system rearrangement.	91
5.2	Contact between grains in a granular material.	94
5.3	Connected and disconnected networks.	95
5.4	Schematics of the methodology to extract “force chains” of high-weight links in the networks.	96
5.5	Sub-networks of high-weight links in the Amino Acid Network of three protein structures.	96
5.6	Sub-networks of high-weight links in the Building Network of three urban structures.	98
5.7	Sub-networks of high-weight links in the Building Network of the city of Lyon (France).	98
5.8	Schematics of the random mobility simulation algorithm.	100
5.9	Simulation of random mobility in three urban structures.	101

6.1	Examples of amino-acid mutations leading to no neighborhoods rearrangement (Z mutation), local rearrangement (L mutation) and large-scale rearrangement (F mutation).	105
6.2	Examples of mutations leading to different neighborhoods rearrangement. .	106
6.3	Classification of all the in-silico mutations of PSD95 ^{pdz3} based on the neighborhoods rearrangement they provoke.	108
6.4	Classification of the in-silico mutations of the surface-exposed amino acids of PSD95 ^{pdz3} based on the neighborhoods rearrangement they provoke. . .	111
6.5	Classification of the in-silico mutations of the buried amino acids of PSD95 ^{pdz3} based on the neighborhoods rearrangement they provoke.	111
7.1	Classification of atomic interactions into four structural levels.	117
7.2	Allocation of atomic interactions into structural levels.	118
7.3	Distribution of the link weights $w_{i,j}$ in the proteins database after links classification.	119
7.4	Distribution of the 2D link weights $w_{i,j}$ in the proteins database, classified based on the secondary structure of the nodes they connect and the distance between the two amino acids in the amino-acid sequence ($ i - j $).	120
7.5	2D atomic contacts between amino acids at distance $ i - j = 2$ in the sequence in an α -helix (left) and in a β -strand (right).	120
7.6	Distribution of the node weight (w) and Neighborhood watch (Nw) in the proteins database, classified based on the secondary structure.	121
7.7	Distribution of the node weight (w) allocated to each structural levels in the proteins database, classified based on the secondary structure.	121
7.8	Ternary plot of the structural-level allocation of the atomic interactions to the 1D, 2D and 3-4D structural levels.	124
7.9	Allocation of the atomic packing into structural levels for the nodes of the proteins database.	125
7.10	Allocation of the atomic packing into structural levels for the nodes of the proteins database, classified based on the amino-acid type.	126
7.11	Allocation of the atomic packing into structural levels for the nodes of the proteins database, classified based on the secondary structure.	127
7.12	Allocation of the atomic packing into structural levels for the nodes of the proteins database, classified based on the relative Accessible Surface Area (rASA).	128
8.1	B subunits of AB ₅ toxins.	133
8.2	Transthyretin (TTR) variants.	134

8.3	Main proteases (Mpro) of coronaviruses SARS-CoV and SARS-CoV-2. . .	137
9.1	Comparison between the Induced Perturbation Network and local structural-level allocation of atomic interactions measure.	141
9.2	Schematics of two possible cases of position of the candidate interacting interface in the protein oligomer.	143
9.3	Node properties (degree k , weight w and Neighborhood watch $Nw = w/k$) of the amino acids of the TTR variants in their Amino Acid Networks, along the protein amino acid sequence.	144
9.4	Induced Perturbation Networks (threshold $\bar{w} = 4$) for four TTR variants. .	145
9.5	TTR L55P variant: position of the candidate new interacting interface. . .	146
9.6	Tetrameric structure of TTR.	147
9.7	Model for fibers resulting from the creation of an active interface (arrows) on each chain of a tetrameric protein with dihedral symmetry.	148
9.8	Fiber models for the TTR L55P variant.	148
10.1	Relation between the BDS signal collected at fixed temperature and the BDS signal collected at fixed frequency.	154
10.2	Dielectric loss as a function of temperature after 3 hours of thermal treatment at 80°C for the cholera toxin B pentamer at frequencies from 1 to 106 Hz.	157
10.3	Temperature dependencies of the relaxation times for the three relaxation processes P_{1C} , P_{2C} and P_{3C} identified for the cholera toxin B pentamer after thermal treatments at 80°C.	157
10.4	The toxin interface.	160
10.5	Toxin interface state and dipole detected in P_{1C}	161
10.6	Toxin interface state and dipoles detected in P_{2C} and P_{3C}	162
10.7	LTB ₅ and CtxB ₅ thermal dynamics changes measured by BDS.	164
10.8	Schematics of the unfolding mechanisms based on the different relaxation peaks detected by BDS along the thermal treatments.	165
10.9	Electrostatic Networks of CtxB ₅ and LTB ₅	168
10.10	Structural analyses of the dynamic differences between LTB ₅ and CtxB ₅ . .	169
11.1	Perturbation of amino acids H372 and I336 due to the G330T mutation in PSD95 ^{pdz3} . The position of the ligand is shown in red. <i>Left</i> : the wild-type structure (PDB: 1BE9). <i>Right</i> : the G330T in-silico mutant produced with FoldX [161]. The images are produced with VMD [38].	174
11.2	Classification of nodes as G (Generate), C (Connect), A (Absorb) and T (Transmit) in the GCAT network, based on their in- and out-degree. . . .	175

11.3	The GCAT network of PSD95 ^{pdz3}	176
11.4	Distribution of $k_{out}^{d=1}/k_{AAN}$, that measures the fraction of chemical neighbors of an amino acid i that are perturbed upon mutations of i	178
11.5	The directed shortest paths from position 330 to position 372 and vice-versa in the GCAT network.	179
11.6	In- and out-degree for all the nodes of the GCAT network of the PSD95 ^{pdz3} protein.	180
11.7	Relation between the in-degree k_{in} , out-degree k_{out} and total degree $k_{tot} =$ $k_{in} + k_{out}$ in the GCAT network with the degree in the Amino Acid Network k_{AAN}	181
11.8	Core-periphery structure of the GCAT network of PSD95 ^{pdz3}	183
11.9	The inner core (13-core) of the GCAT network of PSD95 ^{pdz3}	184
11.10	The sub-network of functionally-sensitive positions (functionally-sensitive positions) of the GCAT network of PSD95 ^{pdz3}	185
11.11	The paths from the functionally-sensitive position 341 to functionally- sensitive hotspots in the GCAT network.	186
11.12	Sub-network of the GCAT network of the amino acids of the α -helix of the binding pocket of PSD95 ^{pdz3}	187
11.13	Candidate positions whose mutation may provide a rescue mechanism for the mutation of position i	191
11.14	Jaccard similarity between out-neighborhoods in the GCAT network of PSD95 ^{pdz3}	193
11.15	Two possible error-correction mechanisms of structural perturbations of PSD95 ^{pdz3}	194
12.1	Example of Building Networks that model the urban structure at different granularity.	198
12.2	Building Networks of Manhattan (New York City, USA) with different cutoff distance δ	198
12.3	Amino-acid neighborhood of the I359 amino acid of PSD95 ^{pdz3} for different cutoff distances of the Amino Acid Network model.	199
12.4	Examples of dependencies of ILE amino-acid packings (k) with the cutoff of the Amino Acid Network of PSD95 ^{pdz3}	200
12.5	Scheme of the evolution of degree and weight as a function of the cutoff distance for different neighbors sizes.	201
12.6	Changes in the local structural-level allocation of atomic interactions in Tt-cpn10 compared to hm-cpn10 (chain T)	202

B.1	Comparison of the local structural-level allocation of atomic interactions in Transthyretin T119Y and L55P variants.	211
B.2	Comparison of the local structural-level allocation of atomic interactions in Transthyretin V30M and T119M variants.	212
B.3	Comparison of local structural-level allocation of atomic interactions in LTB ₅ and CtxB ₅	213

List of Tables

3.1	Classification of in-silico mutations based on the distance of the perturbed amino acid neighborhoods from the mutation site.	61
4.1	Urban morphologies depicted by geometrical and network measures.	76
4.2	Protein structures depicted by geometrical and network measures.	77
5.1	Connected components in the sub-networks of high-weight links in the AANs of three protein structures and the proteins database.	97
5.2	Connected components in the sub-networks of high-weight links in the BNs of urban structures.	98
6.1	Side-chain length of all amino-acid types.	107
6.2	Number of amino acids in each class of structural perturbations (Z, L, F or M) for each type of secondary-structure element.	109
10.1	Interface interaction and dipole features inferred from amino acid network based models.	159
10.2	LTB_5 and $CtxB_5$ (underlined if different) interface interactions and inter-molecular electrostatic dipoles inferred from network-based models.	167
11.1	Partition of the surface-exposed and buried amino acids into the GCAT classes.	182

List of Abbreviations

AAN	Amino Acid Network
BA	Barabási-Albert
BDS	Broadband Dielectric Spectroscopy
BN	Building Network
BSA	Building Surface Area
CtxB ₅	B-pentamer of the Cholera toxin
EN	Electrostatic Network
ER	Erdős-Rényi
IPN	Induced Perturbation Network
LCC	Largest Connected Component
LTB ₅	B-pentamer of the human Heat Labile toxin
MD	Molecular Dynamics
Mpro	Main protease
PDB	Protein Data Bank
PN	Perturbation Network
rASA	relative Surface Accessible Area
RMSD	Root Mean Square Deviation
SARS	Severe Acute Respiratory Syndrome
TTR	Transthyretin
WT	wild-type

Chapter 1

Introduction

Proteins fulfil the biological functions in organisms on Earth since more than three-billion years. This means that proteins keep functioning in time and in different environmental conditions, i.e. proteins are sustainable systems.

This work investigates the protein structural design to elucidate how it leads to the functional sustainability and it uses the results on proteins to explore the sustainability of completely unrelated systems: urban structures. This second task is made possible by the use of similar modeling strategies, so that the two system designs can be compared despite their functional differences. Comparing proteins and cities is unconventional because of the obvious differences between the two systems. However, the analogy between the two systems, their common challenges and the benefice of considering both systems together to broaden our understanding of each one will be illustrated throughout this manuscript.

The term “sustainability” takes various meanings across disciplines [1]. For example, a sustainable material is defined as a product that uses a low amount of energy and produces little environmental pollution during its life cycle [2] and similarly sustainable urban forms have been defined as contributing to low energy consumption and low pollution levels [3], while a sustainable software has a cost-efficient longevity [4]. The issue of defining sustainability is developed more in details in Chapter 2, Section 2.1. Because the industrial-design point of view has the closest definition of sustainability to our investigation, we use its definition of a sustainable system, namely a sustainable system is a system that remains functional across time and in a changing environment [5, 6].

In time, a system is subject to perturbations, that are internal and external. In the city, an internal perturbation is the demolition and re-construction of a building and an external perturbation is a natural disaster. In the protein, an internal perturbation is the mutation of an amino acid (amino acids are the protein components) with an amino acid of a different type, and an external perturbation is a change in the environment.

The fact that systems are subject to internal and external perturbations implies two requirements for a system to keep its function regardless perturbations, i.e. to be sustainable.

The first requirement is that, in absence of external perturbations, internal perturbations do not inhibit the system’s functionality. Thus, either the system’s functionality is unaffected by the perturbation (the system is robust to the perturbation because the modified system is an alternative solution for the fulfillment of the function), or it must implement some error-correction mechanism that allows restoring the functionality upon

perturbation.

The second requirement is that the system must be able to evolve together with its environment, i.e. the system must adapt its function to novel environments. It should be noted that internal perturbations are a mean to modulate the system function, providing adaptability to future environmental changes. Thus, internal perturbations are at the same time challenges that the system faces during time but also means to adapt.

One can thus list some key ingredients for the design of sustainable systems: robustness to perturbations through the diversity of solutions, robustness to perturbations through error-correction mechanisms and adaptability through internal perturbations. This manuscript investigates sustainability according to the proposed ingredients focusing on internal perturbations and their possible outcomes: robustness or adaptability. External perturbations are not taken directly into consideration in this work. Instead, in the context of adaptability to environmental changes, internal perturbations are assessed as a mean for functional change. Internal perturbations are here considered to be substitutions of the system components (amino acids in proteins and buildings in cities).

Considering the impact of the substitutions of the system components illustrates how sustainability requires balancing two contradicting criteria: on the one hand, robustness requires the system to be un-impacted by the substitution; and on the other hand, adaptability requires the system to be responsive to the substitution, in order to evolve. Following the definition proposed by Brian H. Walker that says “Resilience is [...] the ability to adapt and change, to reorganize, while coping with disturbance.” [7], sustainable systems are resilient and robust to the substitution of their components.

Moreover, the possibility for correcting errors implies that the impact of substitutions depends on the environment of the substituted component. This means that sustainable systems implementing error-correction mechanisms are by definition Complex Systems, i.e. systems whose global behavior cannot be inferred from the knowledge of the behavior of its components taken individually [8].

In this work, sustainability is assessed from the point of view of the structural design of the system, i.e. the spatial organization of the system components in the Cartesian space. Accordingly, the structure of proteins and cities is modeled using spatial networks that describe the proximity between the system components. Spatial proximity between the system components determines how the empty space available for the system dynamics is distributed in the system’s structure. As a result, the system’s structure defines the geometrical constraints that govern the system dynamics, in turn providing control on the system function. Thus, an internal perturbation of the system that modifies the proximity between the components in the spatial system, as the substitution of one component with a component of different shape or size, has the potential to impact the system structure, dynamics and/or function.

The components of proteins are amino acids. The set of amino acids composing

a protein is determined by the protein sequence, encoded in the DNA or RNA of the organism, and the way the amino acids of the protein are arranged in the three-dimensional space constitutes the protein structure. Amino acids are made of atoms and the biological function of proteins is fulfilled through the controlled dynamics of its atoms [9]. Since the position of the amino acids in the protein structure determines which atomic interactions will be present, then the spatial arrangement of amino acids in space controls the atomic dynamics in the protein structure, in turn controlling the biological function. This makes proteins spatial systems.

Proteins are sustainable systems because:

1. The substitution of their components (amino-acid mutations) only rarely impacts their function [10], i.e. proteins are robust to internal perturbations. This property means that several functional solutions exist for a protein to be functional. A direct consequence is that, despite the fact that the genetic code contained in the DNA is unique for each individual, i.e. several amino acid mutations exist among the sequences of a protein across different individuals [11], all human bodies are functional.
2. Amino-acid mutations do not act independently of one another, allowing for error-correction mechanisms. For example, rescue mutations restore the protein activity after it has been perturbed by a first amino-acid mutation [12].
3. The rare function-impacting amino-acid mutations have been exploited for evolution, i.e. proteins are adaptable to new environments [13].

Studying how sustainability is designed in the protein structures has a double interest. The first interest is biomedical. Despite the general functional sustainability of proteins, certain amino acid mutations can lead to the development of diseases [14]. Understanding how robustness versus functional-change upon amino acid mutations is determined by the protein structure will help deciphering the molecular mechanisms of diseases caused by amino acid mutations [15–19]. Moreover, understanding the impact of amino-acid mutations will help develop personalized medicine, that aims to assess the risk of disease development and select adapted drugs depending on the genetic backgrounds of the patients, i.e. considering the amino-acid mutations existing in the proteins of one patient compared to another [20]. Finally, if error-correction mechanisms happening in proteins upon single- and multiple mutations are deciphered, then they may be mimicked for the development of therapies against pathogenic mutations.

The second interest concerns the design of man-made sustainable systems. Since proteins are spatial sustainable systems, they are a good model to be followed to design sustainable spatial systems in general. In this study, the possibility of applying this bio-mimetic approach to the design of sustainable urban structures is explored.

As stated before, the goal of this study is to extract the design rules for sustainability of proteins through the analysis of the impact of amino acid mutations (substitutions of the system component) on the protein functionality (resilience versus change of function versus loss of function). In the context of spatial sustainability, the geometry of the amino-acids is considered and not their chemistry. To this goal, we assess the following questions:

1. Can the new component simply substitute the old one or is there a need for a structural rearrangement? If yes, is it local or global?
2. Does such structural rearrangement impact the system dynamics?
3. If many substitutions are performed, what is their collective impact?

Three main difficulties exist in assessing these questions. The first difficulty rises from missing knowledge. Decoding the dynamics and function of proteins from the sole knowledge of their structure is an open problem in structural biology. Because of this reason, in this manuscript the problem of predicting the impact of mutations on the protein function is not assessed, and the inter-dependence of mutations (question 3) is considered from the structural and not functional point of view. That is, we do not try and predict the impact of amino-acid mutations on the protein function, we only focus on measuring the impact on the protein structure and dynamics. This is a pre-requisite for the establishing the relation between the protein's structure, dynamics and function and to predict the functional impact of mutations: predicting whether the protein function is impacted by the mutation first requires assessing whether the protein structure is modified and at what scale, then whether the dynamics are impacted, and finally whether the structural and dynamical changes lead to functional change or not.

The second difficulty is also caused by missing knowledge and is intrinsic to the study of biological systems: we only have access to protein sequences that correspond to functional solutions, otherwise they would not be produced by an organism. Likewise, we do not have access to protein sequences that result in structures that lead to the death of the organism. As a consequence, it is not possible to extract "sustainability features" by comparing sustainable and unsustainable cases, because unsustainable cases are inaccessible.

The third difficulty is technical: protein structures have a complex three-dimensional structure, made of multiple structural levels (Chapter 2, Section 2.2), making it difficult to anticipate whether amino acid mutations can be accommodated with or without structural rearrangement, and if yes, at what scale (question 1).

The second and third difficulties are alleviated by studying, in parallel to protein structures, simpler spatial systems: urban structures. Following the same modeling approach used for proteins, urban systems are investigated from the point of view of structural design: the urban structure of a city is defined as the way buildings are arranged on

the urban area, regardless of traffic management (e.g. the road network) and any social factors. As the protein is composed of amino acids of different size and shape, the city is filled by buildings of different size and shape. In both systems, the size, shape and relative position of the system components (amino acids and buildings) carves the empty space that will be available for the dynamics (atomic motions in the protein and traffic in the city) and for the accommodation of the mutation of components (question 1).

Since urban structures are modeled as two-dimensional systems, they are easily visualized making it easier to assess whether the substitution of the system components (buildings) can be accommodated by the system geometry (question 1). As the impact of amino-acid mutations on the protein functional dynamics are not studied, similarly we do not investigate how the substitution of buildings will impact the traffic management in the city. Nevertheless, we can safely assume that any building substitution that modifies the empty space in the urban structure will have an impact on the possible dynamics in the system: less empty space between buildings implies fewer space available for mobility. Thanks to their easier visualization, the study of urban structures can guide the analysis of protein structures, of more complex geometry. Moreover, urban structures are not guaranteed to be sustainable. If unsustainable urban structures are identified, then they may be used as models of unsustainable spatial systems, that are inaccessible when studying protein structures alone. Then, it will be easier to extract “sustainable features” of spatial designs by comparing proteins (necessarily sustainable) and sustainable urban solutions with unsustainable urban solutions.

On one side the challenge of assessing protein sustainability is eased by the comparison with urban structures, and on the other side the comparison between sustainable protein structures and urban structures allows diagnosing sustainable and unsustainable urban structures. Then, when the rules for the design of sustainable systems will be completely understood from protein structures, they may be employed to guide the design of bio-inspired sustainable urban structures. However, this remains a perspective at the present time.

Obviously, comparing urban structures and proteins introduces an additional difficulty, that is building models that measure how space is shared among the components in both systems, despite the differences in size, number of dimensions and function between proteins and cities. The problem of abstracting the relevant features from technological and natural systems to allow their comparison is typical of biomimicry approaches [21]. As presented in Chapter 3, this abstraction is obtained here by modeling protein structures and cities as weighted complex spatial networks.

The rest of the manuscript is organized as follows.

- Chapter 2 provides a review on the state of the art on the sustainability of proteins, cities and complex systems in general and on the use of network models to study proteins and cities.

- Chapter 3 describes the details the methods used in the following. In the first page of all following chapters, the reference to the relevant sections of Chapter 3 is provided.
- Chapter 4 establishes network measures to determine whether proteins and cities can accommodate substitutions of their components (amino acids and buildings, respectively) with no need of structural rearrangement (question 1). The study presented in Chapter 4 was developed starting from the doctoral work of Rodrigo Dorantes-Gilardi on the protein structural robustness to mutations, supervised by Claire Lesieur and Laurent Vuillon [22, 23]. A scientific paper presenting the part of the results on urban systems in Chapter 4 is in preparation.
- Chapter 5 establishes network measures to determine whether substitutions can be accommodated in proteins and cities with structural rearrangement. From Chapters 4 and 5, it is concluded that protein structures have the potential to accommodate substitutions with or without structural rearrangement (question 1). This is not necessarily the case in urban structures.
- Chapter 6 verifies that upon mutations both no structural rearrangement and structural rearrangement at different scales take place in protein structures (question 1). The scale of structural rearrangement is measured by studying in-silico mutations performed on the third PDZ domain of the synaptic protein PDS-95 as a case study.
- Chapter 7 proposes a network measure to dig how the dynamics are encoded in the static protein structure (question 2).
- Chapter 8 investigates the validity of the measure proposed in Chapter 7 by studying amino-acid mutations that lead to dynamical changes in three cases of study (B-subunits of AB₅ toxins, Transthyretin variants, Main proteases of SARS coronavirus) (question 2). The results presented in Chapter 8 have been submitted to the Bioinformatics journal and are currently under review.
- In Chapter 9, Transthyretin variants are studied more in details, and a mechanism of amyloid fiber formation caused by the dynamical impact of a single amino-acid mutation is proposed. The results presented in Chapter 9 have been published in [24].
- In Chapter 10, the differences in dynamics between the B-subunits of AB₅ toxins variants are studied more in details using an integrative approach, combining experimental dynamics measures with network models. This chapter is the result of a collaboration with Laëtitia Bourgeat, that has performed the experimental measure of protein dynamics during her doctoral work supervised by Claire Lesieur and

Anatoli Sergei [25, 26]. The results presented in Chapter 10 have been published in [27] and [28].

- Chapter 11 proposes a perspective on how to measure the direction of perturbations caused by amino-acid mutations, a requirement for assessing the impact of multiple mutations (question 3). As a proof of concept, the methodology is applied to the third PDZ domain of the synaptic protein PDS-95.
- Finally, some conclusions and perspectives are drawn in Chapter 12.

Chapter 2

State of the art

2.1 Sustainability

2.1.1 Defining sustainability

The term “sustainability” is employed in various contexts, and despite attempts of providing a unified definition of sustainability [29, 30], its meaning is highly dependent on the discipline and on the boundaries of the system under consideration [1]. As an example, Ecological and Environmental sustainability have been defined in relation to the impact of human activity on the health of ecosystems [31, 32]. The so-called systemic approaches also includes economic, social and cultural criteria into sustainability, where the term “systemic” emphasizes the necessity to take into consideration all facets of a sustainable society [33–35]. Other approaches instead refer to the literal definition of sustainability as “the quality of being able to continue over a period of time” (Oxford Dictionary, <https://dictionary.cambridge.org/dictionary/english/sustainability>, accessed 23 February 2021), as in the framework of sustainable software architectures [4] and sustainable product design [5, 6].

I propose that different sustainability approaches reflect different scales of the system taken into consideration, while the goal is always the maintenance of the system functionality over time, implying a system capable of coping with perturbations. Specifically, one could classify sustainability approaches roughly into three categories:

- “Global Sustainability”, that focuses on Nature and Society [33–35]. The system boundaries extend to the entire Earth, including Nature and the human society. The goal is to preserve the functionality in time of the whole planet, including societies that populate it, where the functionality is to host life.
- “Ecological Sustainability”, that focuses on the use of natural resources by one or more species [31, 32]. The system boundaries are the ones of the ecosystem(s) taken into consideration. The goal is to preserve or restore the functionality of the ecosystem(s) in time, where the functionality is the survival of all species of the ecosystem(s).
- “Functional Sustainability”, that focuses on the durability of the function of an object of some sort (material [5, 6] or immaterial [4]), as provided by the design

and architecture of its components. The boundaries are the ones of the object to be designed and goal is to preserve the functionality of the object in time.

As summarized in Figure 2.1, I propose a hierarchical structure of these three classes, where “Functional Sustainability” is included into “Ecological Sustainability”, in turn included into “Global Sustainability”. Indeed, the components used for the “Functional Sustainability” of an object exploit some resources that are part of an ecosystem (object of “Ecological Sustainability”) and the function of the object is exploited by the human species, also part of an ecosystem. Then, the resources and species of the ecosystems studied in “Ecological Sustainability” are part of Nature and society, that are object of the “Global Sustainability”.

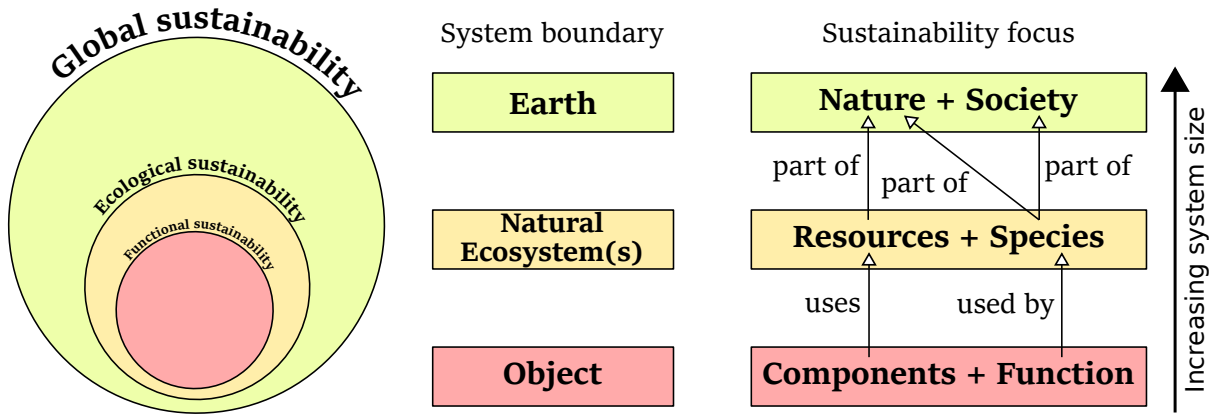


Figure 2.1: Schematics of the classification of Sustainability approaches and their hierarchical structure (left), as well as their interactions (right).

In this work, the “Functional Sustainability” approach is used, with a focus on the sustainability of spatial systems. In particular, the sustainability of natural systems (proteins) is investigated, with a long-term perspective to implement bio-inspired sustainable design to man-made systems. Obtaining functional sustainability of man-made products will imply longer life-times for the products and thus a scarcer use of natural resources and a reduced production of waste, to the benefit of the Ecological and Global Sustainability goals [36]. However, the quantification of the impact of Functional Sustainability on the Earth’s ecosystems and on society is out of the scope of this study and will not be discussed.

The goal is to unravel the criteria to be applied to the design of a spatial system in order to obtain the durability of its function upon perturbations. For a system to be durable, two conditions are necessary:

1. Functional resilience to perturbations, i.e. the capacity to cope with perturbations that may be internal (e.g. substitution of components, failure of the components, ageing) and external (e.g. environmental changes).

2. Evolvability of the function upon changes in the functional requirements (i.e. the system specifications).

We adopt the definition of “resilience” proposed by Brian H. Walker as “the ability to cope with shocks and to keep functioning in much the same kind of way” [7]. In the same work, B. H. Walker also underlines the difference between robustness and resilient systems: while robust systems generally resist perturbations by not changing, resilient systems change and adapt as a response to a perturbation.

2.1.2 Designing a sustainable system

The classical approach for the design of a system consists in determining the system characteristics necessary to fulfil a specific function (specifications) under a certain number of constraints. Designing a sustainable system is a more challenging task, as it requires the determination of the system characteristics necessary to obtain an adaptable function (the specifications may evolve in time), under a set of evolving constraints and that is resilient to perturbations.

As an example, a telephone designed in the 1960s needed to fulfil the function of allowing long-range audio communication under the constraints of a limited cost, ergonomics and the compatibility with the land-line communication technology. Nowadays, smartphones are designed to fulfil numerous functions (long-range audio and video communication, access to the web, recording videos and photos, etc.) under the constraints of reduced size to fit in a pocket, long-lasting battery, compatibility with cellular technology, etc. Obviously, the telephone produced in 1960 could never be functional as a smartphone because it was designed under a different set of specifications and constraints. The challenge of sustainable design is that it is generally impossible to anticipate the specifications and constraints that the system will have to fulfil in the future.

However, a different approach is possible: instead of trying to predict the future specifications and constraints the system will have to fulfil, the adaptability to new specifications and constraints can be added to the initial design specifications of the system. For example, modular smartphones allow the replacement of individual components providing adaptability to novel specifications (e.g. a higher-resolution camera, see for example the ©Fairphone <https://www.fairphone.com/en/2020/07/31/how-sustainable-is-the-fairphone-3/>). Obviously, the additional requirement (adaptability to new specifications) comes with additional constraints. For example, the substitution of one module should not cause the failure of the others.

Figure 2.2 summarizes the traditional system design procedure (left) and the design procedure for an ideally sustainable system (right): while the rise of novel specifications or constraints requires the re-design of traditional systems, an ideally-sustainable system would keep its functionality despite the changes in specifications or constraints, because

sustainability has been implemented in the original system design.

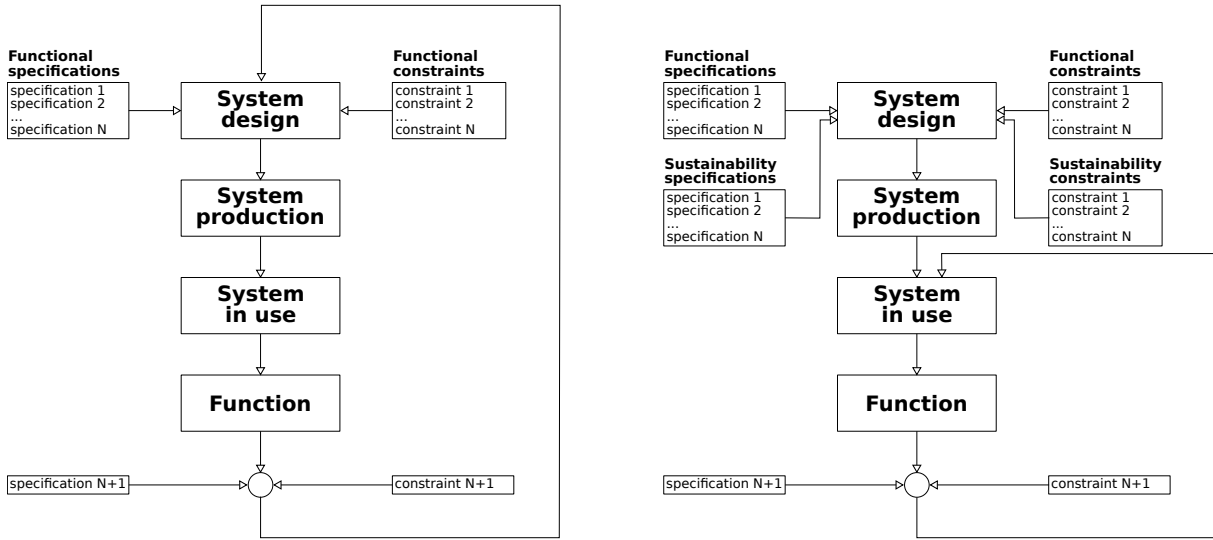


Figure 2.2: Schematics of the system design procedures. Left: the traditional design procedure. The system is designed based on a set of specifications and constraints, so that it can be exploited to provide a specific function. If additional specifications or constraints arise, the system needs to be re-designed and produced to be exploited. Right: the design procedure for an ideally sustainable systems. At the design step, the specifications and constraints for sustainability are added to the functional specifications and constraints. If additional specifications or constraints arise, the system continues to be functional.

The challenge is to determine what are the specifications and constraints for sustainability, and how to implement them in the system design.

In this work, we focus on spatial systems where the system design is reduced to the arrangement of the system components in space, i.e. the determination of the system's structure, to provide the system's dynamics and function. A sustainable design of such systems requires that the system's structure may be modified by most internal perturbations (e.g. changes in the system components) without impacting the system's functionality (robustness), while some internal perturbations may be exploited to modulate the system's function to novel specifications (evolvability). In particular, two very different spatial systems are studied: protein structures and urban structures.

2.2 Proteins as functional spatial systems

2.2.1 The amino acid sequence

The components of proteins are amino acids, which are organic molecules composed of a common part (the so-called backbone atoms: an amino group, a carboxyl group attached to a carbon called C_α) and a side chain that is specific of each particular amino acid [37]. Twenty-two types of amino acids exist, among which twenty are abundant in nature. The latter are represented in Figure 2.3, where they are classified based on their chemical properties. Figure 2.3 shows that amino acids adopt a high variety of sizes and shapes. The amino acids are denoted either with a 1-letter or 3-letter code (Figure 2.3).

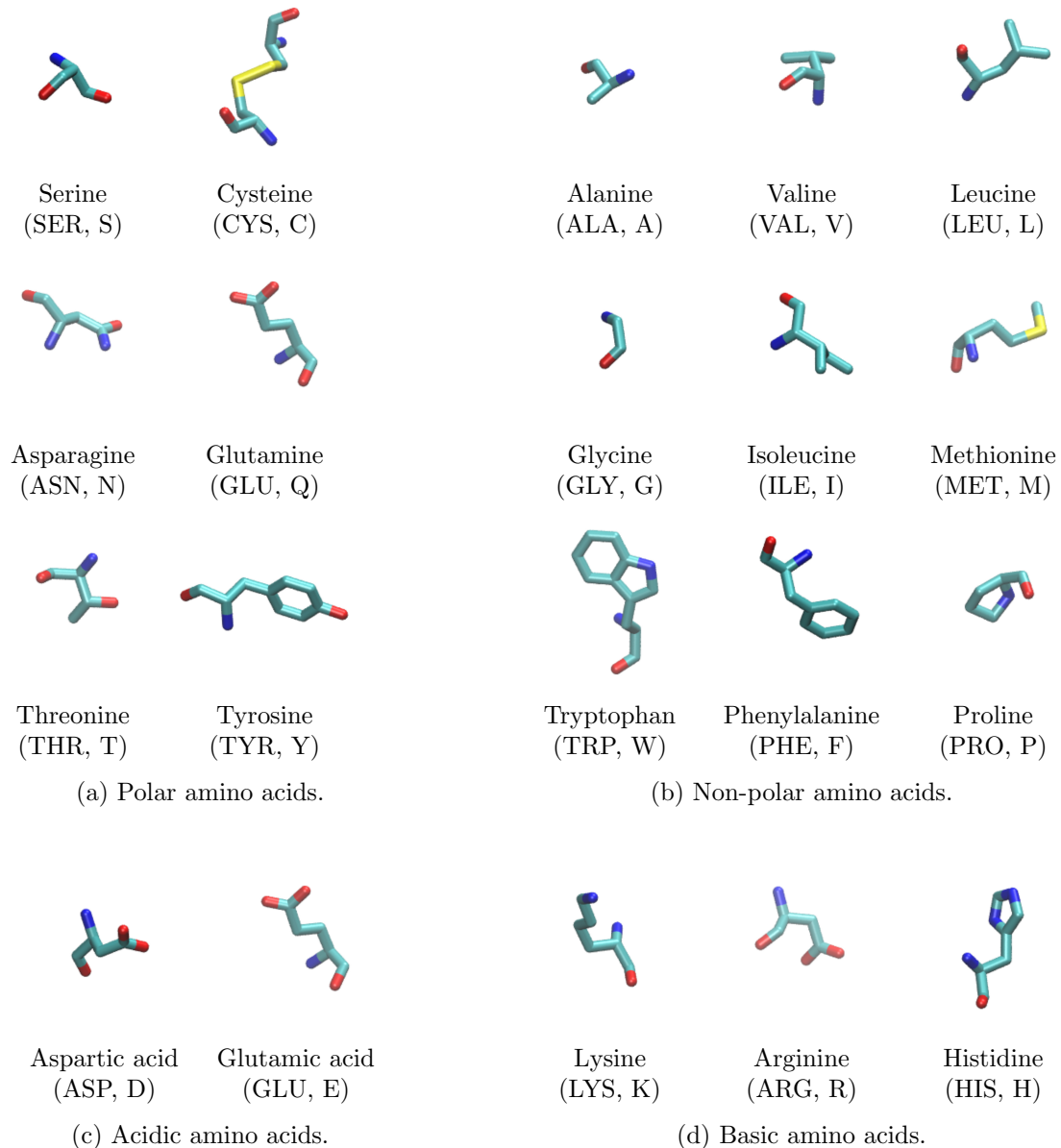


Figure 2.3: The twenty amino acids. In brackets, their 3-letter and 1-letter notations are provided. Different atoms have different colors (carbon is cyan, oxygen is red, nitrogen is blue and sulphur is yellow.) Hydrogen atoms are not represented.

The amino-acids in a protein are covalently bound to form a chain called amino-acid sequence (Fig. 2.5A). The amino-acid sequences of all proteins produced in the cells of an organism is encoded in the DNA. Importantly, the amino-acid sequence determines not only which amino acids are present in the proteins, but also in which order they are chained (see next paragraph, primary structure): if the amino-acid sequence is reshuffled, the resulting chain is not the same. This is important, because the amino acid sequence encodes for the protein structure, dynamics and function.

2.2.2 The protein structure

In the cell, the open-reading frame of the DNA, that encodes for the amino-acid sequence of a protein, is read by messenger RNA molecules which is then translated into the amino-acid sequence by the ribosomes. When the corresponding amino-acid chain is synthesized, it folds into a three-dimensional structure (Fig. 2.4) according to a process called protein folding.

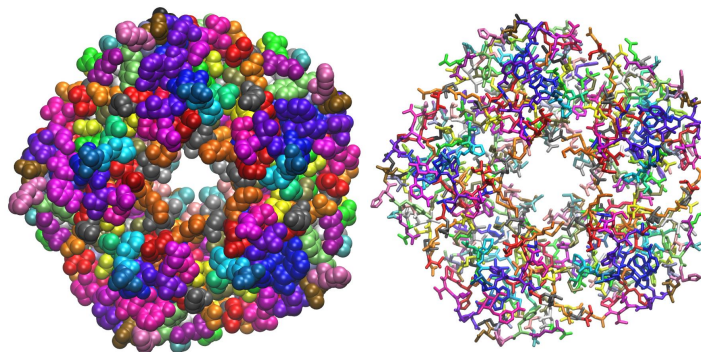


Figure 2.4: Protein structure visualization using Visual Molecular Dynamics [38]. Left: each atom is represented by a sphere. Atoms belonging to the same amino acid have the same color. Right: each covalent bond within each amino acid is represented by a stick. The color-code is the same as in the representation on the left.

The protein structure is characterized according to four structural levels, representing four types of spatial constraints between amino acids (Figure 2.5) [37].

The spatial constraints are fulfilled by chemical interactions between the atoms of the amino acids and yield to different geometrical shapes. The four structural levels are the following:

- The primary structure is the amino acids sequence: it defines the two first neighbors (left- and right-side) of all amino acids (Figure 2.5A). The amino acids are covalently bonded to their sequence neighbors via the backbone atoms through peptide bonds. Apart from the end-points, all amino acids have two neighbors in the protein sequence. The two ends of the chain are called N-terminus and C-terminus, having an amino- and a carboxyl- free groups that have not performed any peptide bond, respectively. The amino-acids along the sequence are enumerated starting from the N-terminus.
- The secondary structure represents the conformations of portions of the amino-acid sequence resulting from the backbone-atoms interactions pattern between amino acids that are close in the sequence. Based on the type of constraints fulfilled by the amino acids that belong to a portion of the sequence, secondary-structure elements as α -helices (with helical shape, Figure 2.5B, left), β -strands (with a flat

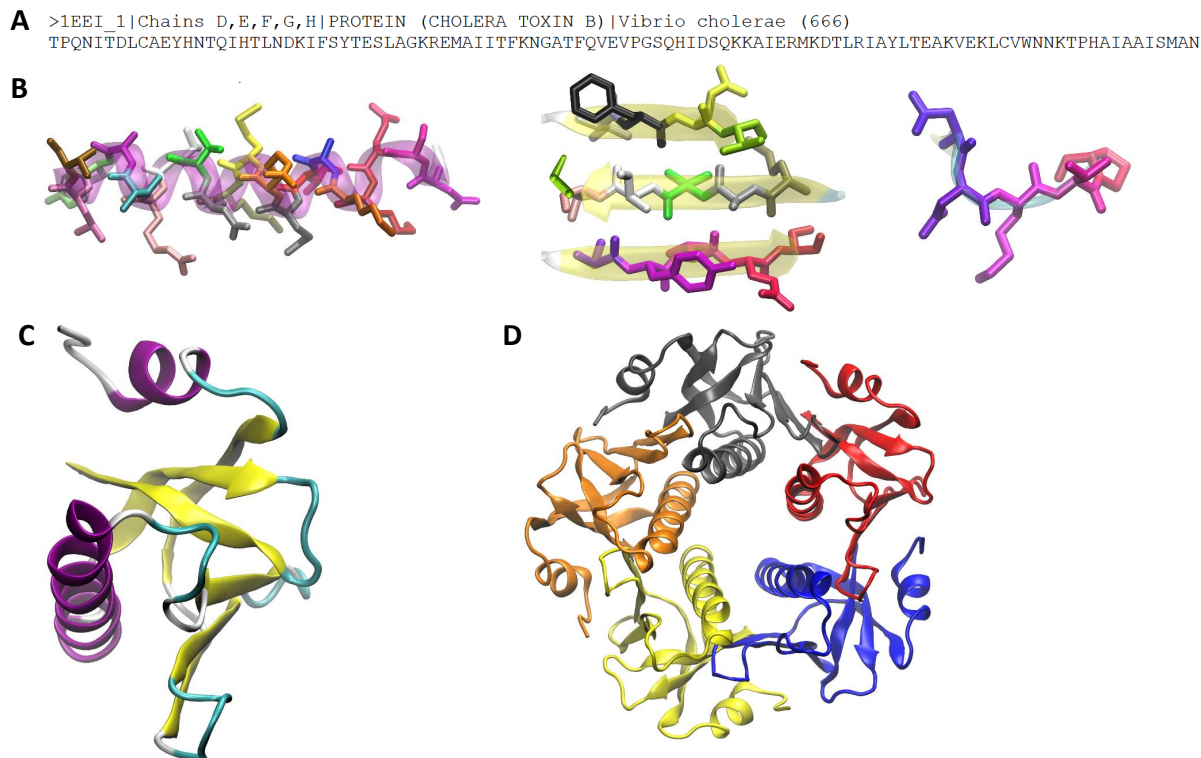


Figure 2.5: The four structural level of the protein structure. A: The primary structure. It is defined by the amino acid sequence of the protein, here reported in its FASTA format, where each amino acid is reported using its one-letter code. B: The secondary structure. Depending on the geometrical constraints fulfilled by the backbone and side-chains of the amino acids, the amino acids form three main types of secondary-structure elements: α -helices (left), β -strands (center), or disordered elements (coils, right). The alignment of β -strands forms a β -sheet (center) C: The tertiary structure. It corresponds to the arrangement of the secondary-structure elements into the three-dimensional space. α -helices are colored in purple, β -sheets are colored in yellow and coils are colored in blue. D: The quaternary structure. It corresponds to the three-dimensional arrangement of the chains in space. Each chain is represented with a different color.

geometry, Figure 2.5B, center) and coils (with no fixed geometry, 2.5B, right) are recognized. The amino acids that belong to the same secondary structure element form hydrogen bonds through their backbone atoms.

- The tertiary structure corresponds to the arrangement of the secondary-structure elements into the three-dimensional space (Figure 2.5C). A particular form of tertiary structure is the β -sheet, made by the planar alignment of β -strands via backbone or side-chain atomic interactions between amino acids that are far apart in the sequence (Figure 2.5B, center).
- The quaternary structure is defined only for proteins made of more than one chain (oligomers) and corresponds to the three-dimensional arrangement of the chains in space (Fig. 2.5D). Each chain is represented with a different color. It should be noted that the chains can be identical (homo-oligomer) or different (hetero-oligomer). In this work, only homo-oligomers are studied.

While the primary structure is the first one to be formed, the order by which the other structural levels are folded is protein-dependent [39].

A complementary description of protein structures is based on the recognition of domains, structural building blocks that are found in evolutionary-related proteins. For example, Fig. 2.6 shows two distinct proteins that share domains (each domain is depicted with a different color), but arranged differently.

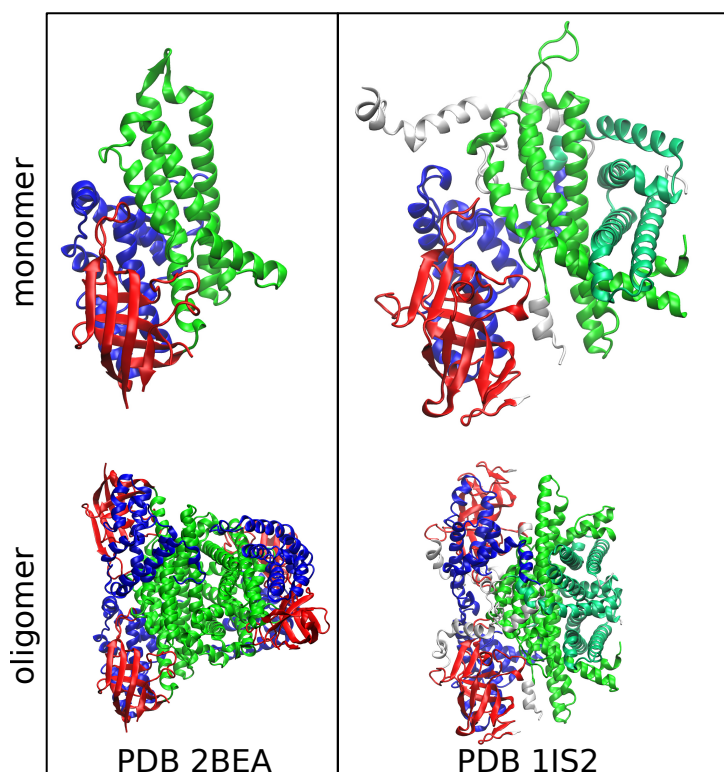


Figure 2.6: Domains in protein structures. Two protein structures (left: PDB 2BEA, right: PDB 1IS2) that share similar domains, but arranged differently in space. Each domain is depicted with a different color.

The most common experimental techniques for protein structure determination at the atomic level are X-Ray Crystallography and Nuclear Magnetic Resonance Spectroscopy. Also Electron Microscopy is now able to provide protein structures at atomic [40]. When the structures of new proteins are evaluated and validated, they are added to international archives, the most important one being the Protein Data Bank (PDB, <https://www.rcsb.org/>). To date, more than 175k entries are present in the PDB and the number of structures released per year is in constant growth (e.g. 14047 structures were released in 2020, 11508 in 2019 and 11176 in 2018, <https://www.rcsb.org/stats/growth/growth-released-structures>). The vast majority (88%) of the entries in the PDB come from X-Ray Crystallography experiments (<https://www.rcsb.org/stats/summary>).

2.2.3 The protein dynamics

Proteins are dynamical objects and the protein dynamics consist in atomic motions. In the previous section, it was mentioned that proteins are synthesized as a chain of amino acids (primary structure), that subsequently folds into the three-dimensional protein structure, by forming the secondary, tertiary and eventually quaternary structural levels. Conversely, folded proteins can unfold, transitioning from their three-dimensional structure to their disordered form by losing the quaternary, tertiary and secondary structures. Protein unfolding is exploited in the cell to regulate the protein activity and degrade misfolded proteins [41]. Protein folding and unfolding are dynamical processes that involve multiple spatial scales (from the primary to the quaternary structure) and multiple time scales [39, 42]. Moreover, proteins can transition from a folded- to a mis-folded conformational state, as typical of proteins involved in Alzheimer-like diseases [19]. Similarly to folding and unfolding, also protein mis-folding involves multi-scale atomic motions in the protein.

Moreover, the biological function of proteins also relies on controlled multi-scale atomic motions (the protein's functional dynamics), as is illustrated in the next section.

The protein dynamics consist in the controlled motions of the atoms of the amino acids that constitute the protein structure. The relation between the protein structure, dynamics and function makes proteins spatial systems: the protein's functionality is determined by how its components (the amino acids) are arranged in space (the protein structure).

On the one hand, the protein structure determines the interactions between the atoms of the protein, and thus the forces to which each atom is subject. As a result, the atomic dynamics are expected to be driven by the interaction forces between the atoms. Based on this relation between the structure and the dynamics of proteins, Molecular Dynamics (MD) simulations are employed to predict protein dynamics starting from the protein structure and even starting from the initial un-folded amino acid chain, coupled or not with experiments [43–45]. MD simulations consist in solving Newton's equation of motions for all or some of the atoms of the protein at each time-step of the simulation. Because of this reason, MD simulations are limited in terms of protein size and simulated time [46]. Moreover, the multi-scale dynamics of proteins underlying large-scale conformational changes remain inaccessible at high resolution [47, 48].

On the other hand, the protein structure also determines the empty space that is available for the atomic motions, as was formulated in 1977 by Frederic M. Richards [49]. From this point of view, the atomic interactions represent the full space that constrains the atomic motions. Importantly, despite overall compactness, protein structures present numerous pockets and voids, where pockets are concavities in the protein surface and voids are holes in the interior of the protein structure [50]. As an example, the shape of the pockets has been studied in relation to protein-protein interactions, i.e. the formation of

quaternary structures, and ligand binding flexibility [51–53]. This work adopts this second approach, that focuses on the geometrical problem of inferring possible dynamics from the empty space carved by the full space in the structure. This choice is made because the focus is on the structural design of sustainable systems, in such a way that very different objects as proteins and cities may be compared in terms of spatial characteristics.

2.2.4 The protein function

The biological functions of proteins are very diverse: for example, proteins can have a structural roles, they can control the transport of substances, they can help repair the cell and they can transmit signals.

The biological function of a protein is fulfilled through controlled atomic motions [9]. We refer to these atomic motions as the functional dynamics of the protein. The spatial scale of the functional dynamics is highly varied depending on the protein function. For example, perfringolysin O (PFO) is a protein toxin that, when close to the cell membrane of its host, performs major and well-controlled structural rearrangements, involving the secondary, tertiary and quaternary structure, that initiate the insertion into the cell membrane, impossible when PFO is in its initial shape [54]. An example of functional dynamics involving much smaller scales is the one of the catabolite gene activator protein (CAP), an allosteric protein that activates the transcription of DNA after CAP binds to a ligand (cAMP) [55]. Binding to cAMP causes a change in the amplitude of the atomic fluctuations in CAP, but no global structural changes [56].

These two examples show how the functional dynamics of proteins cover multiple scales. Local-scale motions are vibrations and side chain motions, at the of Angströms (10^{-10} m), temporally ranging from femtoseconds (10^{-15} s, or 10^{15} Hz) to nanoseconds (10^{-9} s, or 10^9 Hz). Larger collective motions involve changes in the secondary, tertiary and quaternary structures and span up to the nanometers scale (10^{-9} m), temporally ranging from microseconds (10^{-6} s, or 10^6 Hz) to seconds (or 1 Hz) [57, 58].

2.2.5 Then things get complicated

In the previous sections, it was illustrated that the amino acid sequence determines the protein structure, the protein structure determines the protein dynamics and the protein dynamics control the protein function. Following this reasoning, one could conclude that - at least theoretically - it should be possible to predict the protein function from the sole knowledge of the protein sequence after the determination of the protein structure and dynamics. Then, assessing whether an amino acid mutation impacts the protein function (robustness versus evolvability of the protein) would be straightforward: one would just need to calculate and compare the protein structures corresponding to the original sequence (wild-type WT sequence) and the sequence modified by the mutation

(variant sequence); if the structures are similar, then they should encode for the same function (the protein is robust to the mutation) and otherwise the mutation has caused a change in the protein function (evolvability). Obviously, a minimal requirement for such strategy would be the possibility to predict the protein structure from the amino acid sequence, that is still an open issue in Structural Biology.

Relying on the increasing amount of data on protein sequences and structures released every year, predicting protein structures may in the future reduce to obtaining enough computational power to train state-of-the machine learning algorithms or perform accurate *ab initio* MD simulations that simulate the folding process. However, the problem of assessing the protein function from the sequence is more complex and cannot be reduced to the prediction of the protein structure, as schematized in Figure 2.7 and developed in the following.

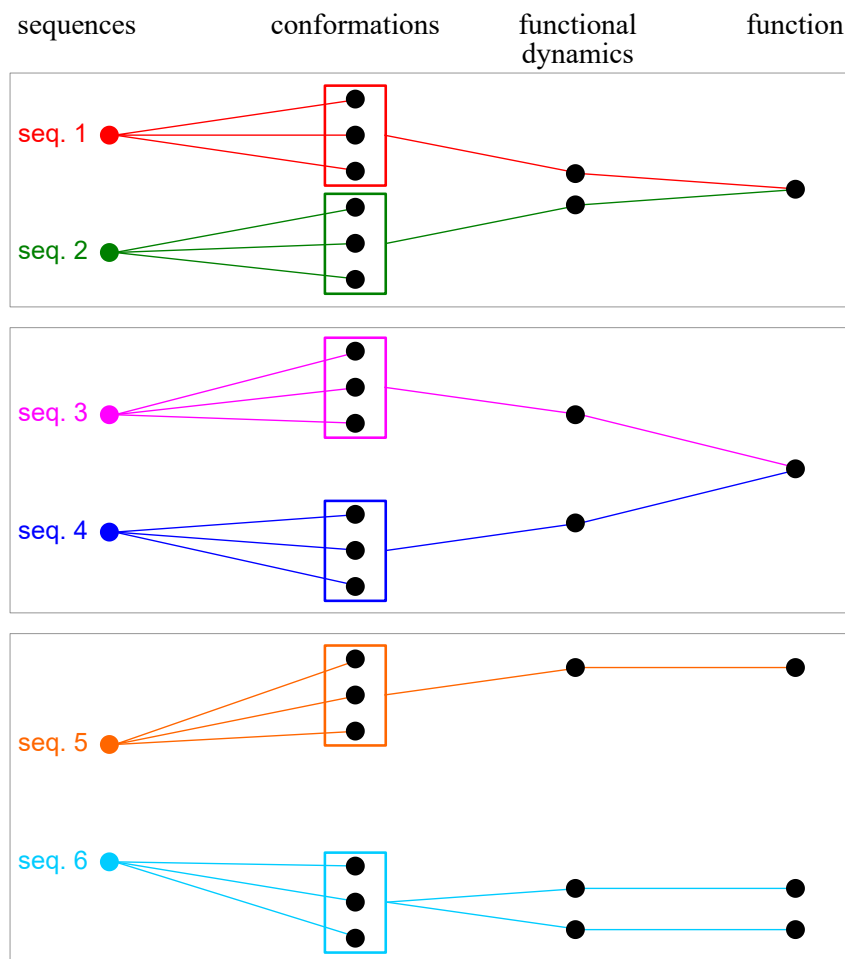


Figure 2.7: From the protein sequence to the protein function. Each sequence is associated to a set of conformations, that encodes for some functional dynamics, in turn determining the protein function (or functions). Depending on the difference between the structural conformations corresponding to sequence variants, their dynamics may be similar, resulting in a similar function (top); the dynamics may be different but leading to the same function (center), or the function of the sequence variants may be different (bottom).

First, one sequence does not correspond to one structure, but to a set of conformations

(the protein conformational ensemble) that are accessible to the protein sequence (Figure 2.7, from sequences to conformations). This is well-described by the concept of the energy landscape of proteins, where stable protein conformations corresponds to the free-energy minima of a high-dimensional space of conformational coordinates [59–61]. The protein conformations can be highly diverse, and it has been proposed that such structural diversity favors adaptability by promoting functional diversity [62]. The existence of multiple conformations for a protein structure is also evident from the fact that protein functions may involve substantial conformational changes, as for the case of PFO presented in Section 2.2.4: if the protein can convert from a conformation A to a conformation B , then both conformation A and conformation B are valid structures associated to the protein amino-acid sequence. Because of the existence of a conformational ensemble for a given protein sequence, the protein structure and dynamics cannot be assessed separately: the protein structure *is* dynamic. Protein dynamics include not only folding, mis-folding and functional dynamics, but also the dynamical changes of the protein structure among its accessible conformations.

Second, the same protein sequence may fulfill multiple functions, associated to different structural conformations [63] (Figure 2.7, seq. 6).

Third, different sequences and structures can fulfill the same function. This is evident from the fact that most amino-acid mutations do not have an impact on the protein function [10], while mutating an amino acid means changing the protein sequence and thus changing the atoms in the protein structure and so, by definition, changing the protein structure (Figure 2.7, seq. 1 and seq. 2). Moreover, structurally-different proteins have been shown to perform the same activity through different mechanisms [64] (Figure 2.7, seq. 3 and seq. 4, assuming that large changes in the dynamics imply different functional mechanisms).

In summary, each amino acid sequence is associated to a set of (dynamical) structural conformations and different protein sequences will provide the same biological function(s) or not depending on the functional dynamics encoded by their conformational ensemble (Figure 2.7). The challenge is to determine how different should conformational ensembles be to encode for different dynamics, and how different the dynamics should be to provide different functions. Finally, for biomedical application, an additional challenge would be to determine what changes in structure, dynamics and function lead to functional failure and development of diseases.

2.3 Sustainability of proteins

The previous section illustrates the complexity of studying proteins because of the redundancy observed at each level of the system: for a given composition (amino-acid sequence), several outcomes are possible in terms of system structure, dynamics and function, and

inversely the same structure, dynamics and function can be obtained by different compositions. However, while this complexity is problematic when one wants to decode the protein function from its sequence, redundancy of solutions at all system levels is also at the basis of protein sustainability.

From the fact that one amino-acid sequence is associated to a set of structural conformations and not to a single rigid structure, it follows that amino-acid mutations, that consist in a modification of some atoms and thus a modification of the protein structure, do not necessarily impact the protein function (functional robustness). Error-correction mechanisms [12] allow the restoration of the protein functionality upon mutations and may be regarded as the control system of the protein structure.

Moreover, it has been proposed that the conformational and functional diversity of proteins, i.e. the fact that one protein sequence can encode for multiple structures and multiple functions (functional promiscuity), allow proteins to evolve [63]. Importantly, functional evolution through functional promiscuity can be attained by exploiting robustness to mutations: the protein may evolve towards a novel function through mutations that maintain the original protein function (robustness) but also promote the emergence of a novel secondary function (evolvability) [65]. These examples show how robustness and evolvability are not necessarily in conflict, and can concurrently contribute to the system's resilience to perturbations and changes in the environment [7].

How can the system be robust and evolvable at the same time? One possibility lies in the use of a modular architecture, where modules evolve independently while the system robustness is obtained through system control (feedback mechanisms to provide a robust response to perturbations) and alternative solutions, as has been proposed in the field of systems biology [66, 67]. For what concerns protein structures, the modular-architecture view is compatible with the existence of multiple structural levels (primary, secondary, tertiary and quaternary structures) and domains (Section 2.2.2, Figures 2.5 and 2.6).

In summary, we are able to recognize alternative solutions, error-correction mechanisms and modularity as ingredients for protein structure sustainability. However, the observation of these properties is not enough to explain the mechanisms by which sustainability is obtained in proteins. With the ultimate goals of predicting the impact of amino-acid mutations, exploit error-correction mechanisms to cure pathologies, and mimic protein sustainability in the design of man-made sustainable systems, it is necessary to decipher how the possibility for alternative solutions at all levels (sequence, structure, dynamics and function) and error-correction mechanisms are designed in the protein structure.

2.4 Proteins as Complex Systems

Protein robustness and evolvability, as well as self-organization into a folded structure and the fulfillment of a biological function are emergent properties of the protein structure, meaning that they are system properties that arise from the system's structure itself.

Self-organization and emergent properties are characteristic of complex systems, i.e. systems whose global behavior cannot be inferred from the knowledge of the behavior of its components taken individually [8]. Complex systems are networked systems, meaning that the set of interactions between the system components determines the global behavior of the system.

An example of complex system are social networks, where the components are individuals. The study of epidemic spreading in a population is an example of how the global output of the system depends on the network of interactions between the system components: depending on the set of interactions between the individuals (how many links are present in the network and how they are distributed), the dynamics of the epidemic spreading will differ [68].

In the protein structure, the components are the atoms forming the amino acids, that interact through atomic interactions. One illustration of the fact that proteins are complex systems comes from the example of intrinsically-disordered proteins that transition from a disordered (unfolded) to an ordered (folded) state before fulfilling their function [69]: the system components (the amino-acid sequence) is identical in both states, what changes is the set of atomic interactions between them. When the protein is in its ordered state, i.e. the native atomic interactions corresponding to the folded structure are made by the amino acids, it acquires a specific biological function. This example shows how a system property (here, a biological function) emerges from the network of interactions (atomic interactions) between the system components. Another evidence of the complexity of protein structures is the non-additivity of the effects caused by multiple amino-acid mutations in terms of function [70–73] and thermodynamic stability [74]: the result of multiple perturbations is not the sum of the perturbations taken individually.

2.4.1 Network models of protein structures

In the last twenty years, network models have been used to model protein structures, as common for the modeling of Complex Systems [75, 76]. In contact-based models, the nodes are amino-acids or atoms (usually the C_α , each amino acid has one C_α atom), that are linked based on spatial proximity. The cutoff-distance varies from 3.8 to 9 Å depending on the study [77]. The Amino Acid Network (AAN) model used in the following chapters uses a cutoff equal to 5 Å, that was proposed as an optimal cutoff [78]. Alternative names found in the literature are Protein Contact Network, Residue Network, Protein Structure Network, Residue Interactions Graph. We will refer to all these models as Amino Acid

Network, because the elementary components (nodes of the network) are amino acids.

In Section 2.5, some examples of applications of network models to study protein dynamics and robustness versus fragility to amino acid mutations will be presented. Before doing so, the remaining of this section introduces network modeling of Complex Systems in general, as well as some measures widely-used in the context of structural biology.

2.4.2 Complex Networks

Complex systems are modeled as networks, that are mathematically represented by graphs. Networks and graphs are composed by nodes (or vertices) representing the system components and links (or edges) representing the interactions between the components. The notation for a graph is $G(V, E)$, where V is the set of nodes (vertices) and E the set of links (edges):

$$V = \{i | i \text{ is a component of } G\} \quad (2.1)$$

and

$$E = \{(i, j) | i, j \in V \text{ and } i \text{ and } j \text{ interact}\} \quad (2.2)$$

Moreover, the links in the network may be weighted according to some definition of strength or importance of interaction, that depends on the study. In this case, a link-weight $w_{i,j}$ is assigned to each link and the network is said to be weighted.

It should be noted that the term “interaction” should be here intended with a very broad meaning and not necessarily as a physical interaction. For example, links in social networks may represent friendship relations, the fact that two people are communicating at a given moment or else, depending on the study.

A sub-class of networks is the one of spatial networks, where nodes have coordinates in the Cartesian space and the links represent proximity between the nodes.

In this study, protein structures are studied using the established model called the Amino Acid Network (AAN), that is a weighted spatial network where the nodes are the amino acids, the links connect nodes whose amino acids are at a maximal cutoff distance and the link weights are given by the number of atomic interactions between the two amino acids. The details of the construction of the AAN are provided in Section 3.1.2.

2.4.3 Network Analysis methods

In the field of Network Science, networks modeling complex systems are studied using measures that describe the system properties from the local level (node) to the global level (whole network). The goal obviously depends on the study, but is generally related to classify networks and to extract the roles of single nodes or links. In the following, some of the basic network-analysis measures are defined. The definitions of the presented

methods can be found in Network Analysis textbooks, for example [79] and Barabási's Network Science Book online (<http://www.networksciencebook.com/>).

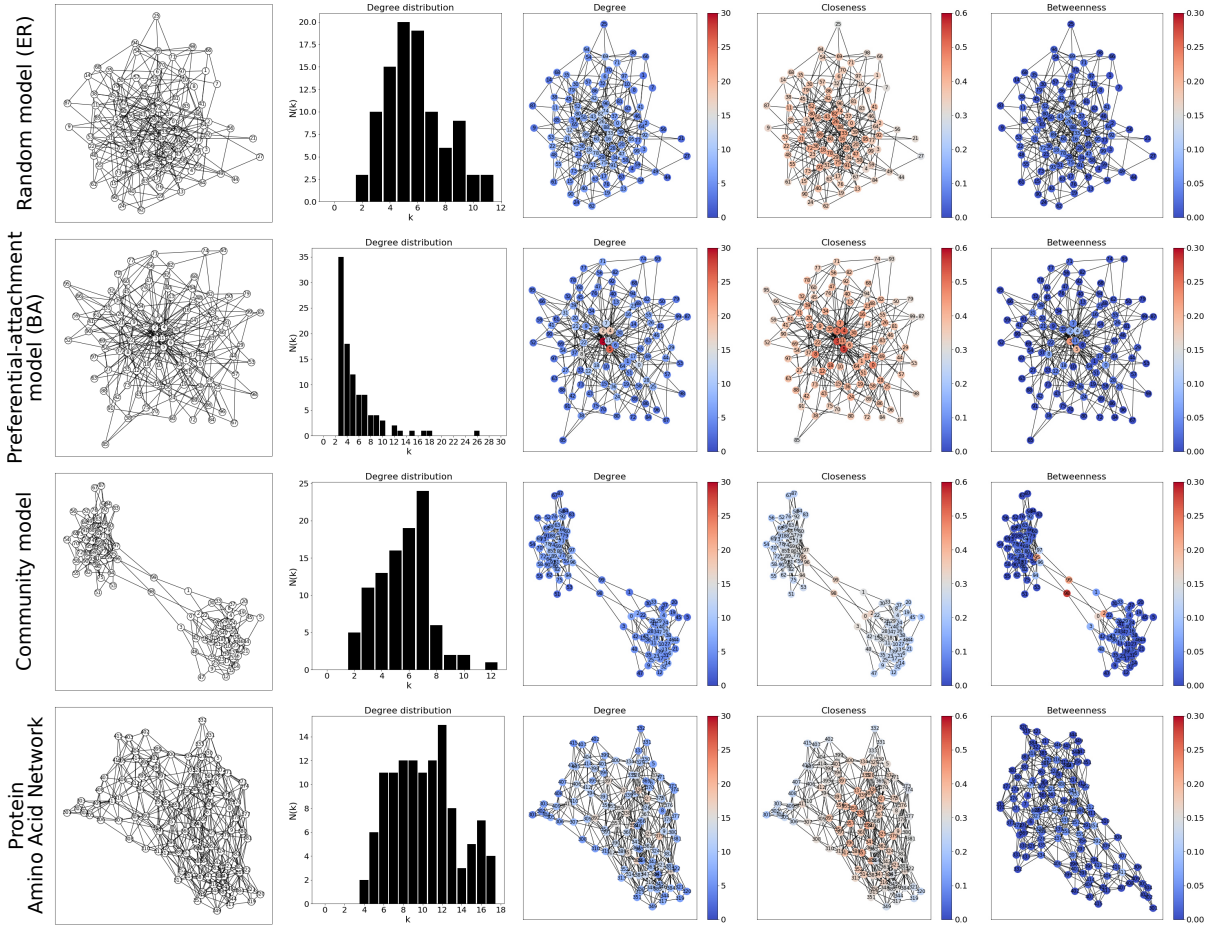


Figure 2.8: Examples of Network Measures on Network models and on the Amino Acid Network of a protein.

Network models Networks may be classified according to their topology by comparing their properties to some models. The easiest example is the the Erdős-Rényi (ER) random model, in which the number of nodes is fixed and links between pairs of nodes are added with some probability p [80]. An example of ER network with 100 nodes and $p = 0.06$, resulting in 298 edges, is represented in Fig. 2.8 (first row). Another widely-used model is the Barabási-Albert (BA) model, that reproduces preferential attachment, typical of some real networks, i.e. the fact the nodes that have many neighbors tend to attract even more neighbors, “rich gets richer” [81]. An example of BA network with 100 nodes and 291 edges is represented in Fig. 2.8 (second row).

Degree distribution The degree k_i of a node i in a network is the number of links the node i makes in the network, i.e. the number of neighbors of i . A simple way to compare networks is via their degree distribution. The ER networks have a Poissonian degree

distribution, while BA networks have a scale-free degree distribution, i.e. they contain many nodes of low degree and few nodes of very large degree (Fig. 2.8). The nodes of high degree present in scale-free networks are called hubs. The differences in degree distributions between the ER and BA network models are expected, because the two models follow different strategies for creating links: a random process for the ER model and preferential attachment for the BA model. Thus, by comparing the degree distribution of real networks with the degree distribution of model networks of the same size, one may try and infer the connectivity rules underlying the observed network topology.

However, comparing degree distributions alone may not be sufficient to describe how nodes are connected in a network, as illustrated by two examples. The first example is a network with a community structure (Fig. 2.8, third row). The network in the figure has 100 nodes and 284 links and has been produced by creating two ER networks of 49 nodes, each representing one community, and by adding two nodes, each of them having two links towards each of the two communities. Unsurprisingly, the degree distribution of the network has the same shape as the one of the ER network (first row); however, the topologies of the two networks are substantially different. Several techniques exist to identify communities in networks, but their description is out of the scope of this manuscript.

The second example is the Amino Acid Network (AAN) of a protein where nodes correspond to amino-acids and links exist between nodes that are at distance lower or equal than 5\AA (Fig. 2.8, last row). The AAN in the figure models the structure of the protein with PDB 1BE9 and has 120 nodes and 613 edges. The link weights are not taken into account here. The AAN is a spatial network and its nodes are connected if the corresponding amino acids are close in space. Because the AAN is a spatial, the connectivity rule underlying the AAN is far from random: the maximal number of neighbors of each node is bounded by how many amino acids can occupy the space surrounding any other amino acid. Moreover, the position of the amino acids has to fulfill the spatial constraints imposed by the primary, secondary, tertiary and quaternary structural levels (Section 2.2.2) Yet, the degree distribution of the AAN (third column) is again similar to the one of the ER random model.

Apart from comparing degree distributions, other network measures can be used to compare the connectivities in different networks, as distances between nodes and node centralities, defined in the following.

Network distances In a network, the distance between two nodes is the length of the shortest path connecting them, expressed as the number of “hops” necessary to travel from one node to the other using links on the network. Two measures that characterize a network are its diameter, i.e. the largest distance among all the possible couples of nodes, and the average shortest path length, i.e. the average distance between nodes. The ER

and BA models of Fig. 2.8 have both diameter = 5 and similar average shortest path length (2.8 and 2.6, respectively). On the contrary, the network model with communities has diameter = 8 and average shortest path length = 3.8, reflecting the fact that the distances between nodes of different communities are large. The AAN has diameter = 7 and average shortest path length = 3.2, intermediate between the previous cases.

Node centralities While the degree distribution, the diameter and the average shortest path length of a network are a global measures of the network characteristics, it is sometimes relevant to extract information on single nodes of the network. More in detail, network centralities measures are used to extract what nodes are the most “important” (the most central) in the network.

The easiest centrality measures is simply the degree of the node: one node is said to be central if it is connected to many other nodes in the network. In the third column of Fig. 2.8, the nodes of the networks are colored based on their degree. As expected from the degree distribution, all nodes have similar degree in the networks, apart from the hubs in the BA model that have a very high degree (red in the figure).

A second measure of node centrality is the closeness centrality, that is calculated by measuring the distance from a node to all other nodes of the network. A node is said to be central if is at low distance from the other nodes of the network (i.e. the shortest path from the node to all the other nodes has a small length). In the fourth column of Fig. 2.8, the nodes of the networks are colored based on their closeness centrality. In the ER network, all nodes have a similar closeness centrality, as expected by the fact that nodes are connected randomly as opposed to preferential connection. In the BA network, the hubs have higher closeness centrality compared to the other nodes. This is also expected from the way the BA network is built: hubs have a high number of neighbors because of preferential attachment and thus a high number of distances equal to one (with their neighbors), two (with the neighbors of their neighbors), etc. In the network with communities, almost all nodes have low closeness centrality apart from the two nodes that connect the two communities, as expected from the fact the nodes in one community are far from the nodes of the other community. Finally, in the AAN, the closeness centrality ranges from relatively-low to relatively-high values. However, the closeness centrality in spatial network is by definition higher for the nodes that occupy the central part of the Cartesian space in which the nodes are embedded: since the links in the network are based on spatial proximity, links that are at the center of the Cartesian space will also be central in the network [82].

A third measure of node centrality is the betweenness centrality, that measures the fraction of shortest paths that pass through a node among the shortest paths between all the possible node pairs of the network. A node is said to be central if it is “in-between” many node pairs. Similarly to the closeness centrality, the betweenness centrality is similar

for all nodes in the ER network and higher for the hubs of the BA network. Again, this result is expected from the fact that the links are assigned randomly in the ER network and that hubs have a high degree in the BA network, and thus a higher probability to connect pairs of nodes. The betweenness centrality is also high in the network with communities for the nodes that connect the two communities, because all shortest paths between nodes of different communities necessarily pass through one of these nodes. The betweenness centrality is similar for all nodes of the AAN.

As illustrated by the examples, centrality measures may be useful to extract roles of nodes in the network, specially when extreme values are present, e.g. the hubs in the BA network or the nodes with low-degree and high-betweenness that highlight the presence of a community-structure in the network. Nevertheless, attention should be paid when interpreting these results, as shown by the closeness centrality of spatial networks (here, the AAN), where high values are associated to a purely geometrical property (distance from the center of the Cartesian space) and not by a higher “importance” of the nodes in the network [82]. Moreover, centrality anomalies may arise from an over-simplification of the network model: using the example of transportation networks, that are also spatial networks, Alves et al. have shown that anomalies in betweenness centrality found in the unweighted network models disappear when link weights are added to the network models [83]. In proteins where the link weight varies significantly between amino-acid pairs, the use of unweighted network models may be misleading.

The following sections reports some applications of the network measures presented here for the study of the impact of amino-acid mutations in proteins and the investigation of protein dynamics.

2.5 Applications of network models of protein structures

Exploratory studies of the network topology: importance of the model parameters

One of the first studies applying Network Science to Structural Biology was published in 2003 by Greene and Higman [84]. Similarly to the model used in this study, they modeled protein structures as networks of amino acids, connected with multiple links when any of their heavy atoms (carbon, oxygen, nitrogen and sulfur) are at a maximal distance of 5Å and compared the network characteristics when including all links or only the ones involving residues far apart in the amino acid sequence (long-range contacts, corresponding to tertiary-structure contacts). They found that the networks including all links have a Poisson-like degree distribution, consistent with the example in Fig. 2.8 (last row), while the networks of long-range contacts have a scale-free distribution, similar to the BA network model in Fig. 2.8 (second row), implying the existence of a few hubs. Susequent studies have confirmed that the degree-distribution of the AAN is not scale-free

when all amino-acid contacts are considered [85, 86].

Study [84] puts in evidence the fact that the choice of the model parameters can deeply change the model interpretation: a given node can be a hub in the network of long-range contacts, but loose its “centrality” when all atomic contacts are taken into consideration. This phenomenon is similar to the disappearance of betweenness centrality anomalies when link weights are added to transportation networks [83]. Similarly, Yu et al. have shown that the properties of AANs are highly dependent on the cutoff distance employed and that the hierarchical structure typical of scale-free networks having hubs [87] is observed at large cutoff (12Å) and not at lower cutoffs (3 and 5Å) [88].

The dependence of the existence hubs on the model parameters raises an important issue because determining whether the AAN is scale-free (i.e. with hubs) or not is not only a problem of vocabulary. Indeed, the presence of hubs in scale-free networks has been related to a network’s vulnerability to targeted attacks: removing hubs in the network can create a cascade failure in the system [89]. However, attention should be paid to the definition of hubs, that is highly community-dependent. In AANs, hubs have been defined as nodes having at degree $k \geq 3$ [78], $k \geq 4$ [90], or having degree higher than average [91]. In contrast, hubs of power grids are defined as nodes having $k \geq 10$ and they have degree up to 25 (the minimal degree is equal to 1, as in the AAN) [92]. As another example, the 25 largest hubs of the airline network, defined as 25 cities that are connected by a direct flight to the highest number of main cities worldwide, have degree between 70 and 168 [93]. In conclusions, the role of hubs in protein mutations fragility is debatable [18].

Node centralities to predict the functional impact of mutations

Following the recognition of central nodes in networks, amino-acid positions having high closeness-centrality or similar measures (e.g. the node is central if its removal significantly increases the average shortest path length of the network) have been hypothesized to be functionally important for the protein [91, 94, 95]. These functionally-important amino acids have been proposed to represent positions whose mutation would be deleterious for the protein function [91]. However, predicting functionally-important amino acids from closeness-centrality measure alone provides only 46.4% sensitivity (fraction of identified active site amino-acids positions) and 9.4% of corrected predictions [94]. The predictive performance is significantly increased (80% sensitivity and 63% of corrected predictions) when considering evolutionary-conserved high closeness centrality as an indicator of functional importance [96]. In fact, there exists a positive correlation between closeness centrality, evolutionary conservation (the amino acid is rarely mutated) and proximity to the protein’s center of mass of amino acids (the amino acid is buried in the protein structure) [91, 97]. This makes it difficult to draw conclusions on the causality between centrality and functional importance.

An additional difficulty in assessing the functional impact of amino-acid mutations from protein structures comes from the fact that amino-acid mutations may disrupt the protein functionality because of different reasons. For example, the mutation of two residues of the allosteric enzyme phenylalanine hydroxylase, A330 and P314, both buried in the protein structure and located at the active center of the protein, have very different consequences: A330S leads to conformational destabilization while the enzymatic activity is not affected, while P314S stabilizes the protein conformation but leads to lower enzymatic activity [98]. Thus, the amount of atomic contacts made by an amino-acid is not a good indicator of functional sensitivity upon mutations *per se*, what matters is rather which atomic contacts are gained or lost upon mutation, and what is the impact of this change in atomic contacts on the protein dynamics, for example leading to misfolding (A330S) or loss of functional dynamics (P314S). Moreover, the consequences of mutations may depend on the type of substituting amino acid, while AANs are built based on the wild-type structure of the protein. As a consequence, to assess the mechanisms of functional robustness and sensitivity upon mutations it is necessary to first describe the impact of mutations on the protein dynamics. To do so from the sole knowledge of the protein static structure, it is in turn needed to decode how the protein dynamics are encoded by the protein structure.

Methods to study protein dynamics

AANs have been used alone or coupled to simulations or experiments to study protein dynamics and the impact of amino-acid mutations on the protein dynamics.

Differences in amino acid networks inferred from Nuclear Magnetic Resonance experiments between the *resting* and *working* states of an allosteric enzyme in its wild-type form and after mutation confirm the role of atomic contacts in the control of the functional dynamics and that mutations impacting the functional dynamics reflect in differences in the AAN [99, 100]. Employing only static structures of the apo- and bound forms of allosteric proteins, differences in atomic contacts in the networks corresponding to the apo- (i.e. without the ligand) and holo- (i.e. bound to the ligand) form have been employed to investigate allosteric paths [101] and changes in the connectivity between clusters of nodes have been related to changes in protein dynamics upon binding [102].

A study from 2002 put in evidence that the dynamical process of protein folding can be modeled as a rewiring of the AAN of the protein structure, by comparing the contact networks of protein conformations adopted just before and just after the transition state of the protein folding procedure [103]. More in detail, it was shown that despite the fact that the number of contacts is similar in the two conformations, the average shortest path lengths between the nodes after the transition states are shorter. It was proposed that shorter path lengths correspond to higher cooperativity between amino acids [103]. Subsequent studies have proposed node-centrality measures to identify the folding nucleus

of a protein, i.e. the set of amino acids that first acquire the folded structure during folding, from the AAN of the folded structure [104] or the one of the transition state [105]. Moreover, the number of nodes performing more than 4 long-range contacts in the AAN has been shown to be inversely related to the folding rate of proteins [106].

AANs models have also been employed coupled with Molecular Dynamics (MD) simulations [107]. In MD simulations, the equation of motions for the atoms of the protein are solved. Employing MD simulations on top of static-structure analysis has the advantage of directly incorporating the dynamical information. For example, calculating contact networks at different time-stamps of the MD simulation of an enzyme has unraveled functional dynamics and the functional impact of amino-acid mutations distant from the active site from patterns of atomic interactions that are not observed in the crystalline structure of the protein [108]. These approaches have been employed to study functional dynamics of proteins using networks where the links between the nodes represent either atomic contacts (averaged over the MD simulations frames or consistent across multiple frames) [108–111] or correlated motions obtained from the MD results [112, 113]. Methods exploiting information on both atomic contacts and correlated motions are reported [90, 114]. Similarly, Energy Exchange Networks (EENs) model energy transport in protein contact networks based on both the protein structure and dynamics simulated using MD [115]. The comparison of the EENs of the apo and holo forms of an allosteric protein has identified residues involved in the functional dynamics [116].

Simulated folding of protein single amino-acid variants having similar structure showed that the impact of mutations on the folding dynamics is dependent on the neighborhood of the mutated positions [117], indicating once again the role of atomic contacts in governing the protein dynamics and the dynamical impact of mutations. The impact of the mutations is explained in terms of “local frustration” of atomic contacts induced by the mutation, i.e. the introduction of non-native interactions that impact the folding process of the protein. The concept of “frustration” has been used to explain changes in structural conformation of proteins across their conformational landscape not only upon mutation, but also upon ligand binding (functional dynamics) [118]. Again, the changes in structural conformations can be described as rewirings of the AAN.

The main limitation of MD-based methods is the high computational cost of the simulations. Apart from its computational cost, additional limitations are the strong dependence of the simulation results on the choice of the force field to which the atoms are subject [119] and on the length of the simulated period [120].

A computationally-lighter alternative to MD simulations to incorporate dynamical information on top of AANs is represented by the Elastic Network model, or Gaussian Network. In the Elastic Network, the nodes are usually the C_α atoms of the amino acids and, similarly to the contact-based networks, the nodes are connected based on a distance cutoff [121]. The difference from AANs is that the links in the Gaussian Network are not

rigid entities but are springs with some force constant. Because the nodes are masses connected by springs, the nodes of the network are subject to fluctuations, simulated using Normal Mode Analysis and characterized in terms of amplitude (mean square fluctuations of a node), temporal scale (frequency of the fluctuations) and correlations (sets of nodes having correlated motions). Slow motions (low frequency) have been associated with collective motions (global motions, involving many residues) and inversely fast motions (high frequency) have been associated with local motions [122].

As from AANs analysis ([102]), the roles of links connecting different amino-acid communities in AAN in controlling the protein dynamics has been highlighted also by Elastic Network models, where local perturbations at these links caused changes in the predicted low-frequency vibrational modes of amino acids far apart from the perturbed site [123]. The authors of [123] propose that amino-acid mutations at the positions involved in these inter-community links may cause significant structural rearrangement of the protein structure and similarly the same mechanism could model the conformational changes of transmembrane proteins upon binding. Elastic Network models were also able to identify dynamical differences in the low-to-intermediate-frequency regime among proteins with similar monomeric structure but different function and/or oligomerization states [124, 125]. However, the Elastic Network Model is not suitable to study large-scale domain movements with high-energy barriers [125].

A study of 2011 compared the protein dynamics obtained using Elastic Network models and MD simulations and raised the issue of the choice of the spring-force parameter in Elastic Network models, that may have a significant impact on the results [120]. The same study measured that Elastic Networks tend to under-estimate slow motions and over-estimate fast motions, but that the same effect is obtained when MD simulations are performed on short simulated time lengths [120]. On the other side, extensive MD simulations have revealed the existence of multiple trajectories underlying allosteric functional dynamics, something that is sometimes overlooked in methods based on static pictures of AANs or Elastic Network models [126]. However, another comparative study of 2018 claims that Elastic Network models are slightly better at reproducing experimental conformational ensemble of a benchmark of proteins compared to MD simulations [127]. In fact, the large number of parameters involved in the definition of the AAN, its corresponding Elastic Network and in the definition of the MD simulation conditions makes it hard to effectively compare the two methodologies.

The positioning of this study

The literature presented above proves the relevancy of network-based models, coupled or not with experiments and dynamics simulations, to predict the structural, dynamical and functional impact of amino-acid mutations and to investigate protein dynamics. AANs are complementary to experimental measures and simulations and are compatible with

the view of protein structures as dynamical objects that transition across states in a conformational landscape. Using network-based models to study protein structures benefits the availability of a wide range of well-known measures coming from Network Science, together with the possibility to define ad-hoc measures.

The goal of this study is to investigate the structural design of protein structures and their amino-acid networks to unravel the design criteria underlying different impacts to mutations (structural and dynamical robustness versus evolvability) and different dynamics. The focus is on the role of atomic contacts in carving the empty space available for the protein dynamics, and how changes in atomic contacts upon amino-acid mutations impact the protein structure and dynamics.

Unraveling how the protein dynamics are encoded by the protein structure and assessing the impact of mutations on the protein structure and dynamics is a prerequisite for the assessment of the functional impact of mutations: functional robustness or functional changes, where functional changes may lead to the development of diseases and/or adaptation to novel environments.

Similarly to proteins, other Complex Systems are spatial, meaning that the system's functionality is encoded by the spatial arrangement of its components in space. The following section illustrates that cities are spatial Complex Systems, and outlines network approaches used to model them. In the framework of this study, modeling cities as spatial networks, similarly to the AAN of protein structures, allows highlighting the structural-design characteristics that are common or different in these two very different systems. Thanks to this analogy, the identification of the spatial-design parameters that provide sustainability to proteins is eased by the comparison with urban structures, that have the advantage of being modeled as two-dimensional systems. Moreover, the design criteria for sustainability extracted from the analysis of protein structures may in the future be employed to design sustainable urban structures.

2.6 Cities as functional spatial Complex Systems

The 11th goal of the United Nations Sustainable Development Goals is to “Make cities inclusive, safe, resilient and sustainable” in view of an expected raise in urban population (60% of the global population is expected to live in cities in 2030) and based on the observation that cities account for 70% of global carbon emissions and over 60% of resource use (<https://www.un.org/sustainabledevelopment/cities/>, accessed 6th April 2021). This goal calls for methods to identify urban sustainability, resilience and transformation capacity [128].

Cities have been recognized as Complex Systems because of the complex networks of interactions existing in the urban system at multiple levels (e.g. transportation, mobility, economic activities etc.), that generate emergent behaviors [129–132]. Examples of

emergent behaviors are the transportation patterns that depend on the location of the amenities and on the population density and the dramatic racial or social segregation that may arise from mild preferential bias of people for being surrounded by others of the same ethnicity or social status, as reviewed by Batty in 2009 [130].

From the spatial point of view, cities are structures made of buildings (“filled space”) that occupy the urban land and shape the “empty space” (roads, parks, etc.) available for the mobility of pedestrians, cars, bikes, public transport, etc.

The topology of the “empty space” has been largely investigated to study urban mobility, mostly within the frameworks of urban road networks [133–140], the Space Syntax theory [141, 142] and named-streets networks [143], that focus on the spatial arrangement of roads in the city.

In parallel, Urban Morphology studies the geometrical features of the so-called urban forms, including buildings, blocks of buildings and streets. Thus, Urban Morphology studies incorporate both the “filled” and “empty” space of the urban structure. Urban Morphology has been applied to study of urban growth during time [144] and the comparison of different cities [145, 146]. These methods rely on multi-variate statistics involving a large number of indicators (e.g. built compactness, fractal dimension of the built area, spatial autocorrelation of the built area, area of blocks of buildings, lengths of streets, etc.). The results of Urban Morphology studies provide either a global description of the urban systems [144, 146] or a classification of the single urban forms [147].

Additionally, Urban Morphology coupled with physical simulations has shown the role of the build geometry (the “filled” space) in modulating the urban microclimate [148, 149], including air quality [150–152] and the amount of urban-heat-island effect [153], as well as the buildings energy consumption [149], all factors influencing the environmental sustainability of the urban system.

The positioning of this study

Among the variety of levels of description of urban systems, this study focuses on the spatial structure created by building footprints on the urban landscape, i.e. the filled space of the urban system, and how this filled space carves the empty space. The goal is to investigate the relation between the local arrangement of geometrical entities (the buildings) over the spatial sustainability of the system, defined as the possibility for accommodating perturbations (robustness) and adapt to changes (evolvability).

Cities need to cope with perturbations, that can be of different nature. For example, urban growth leads to the addition of buildings and to the enlargement of the existing ones, the amount of cars in the system can increase in a gradual way due to an increase in population or in a sudden way due to exceptional events and streets can be temporarily closed. All these phenomena have the potential to affect the amount of traffic in the city. This has many consequences on the urban system, including influencing the quality of

air, the quality of life of commuters and the city's economics [154]. Moreover, cities have to face climate change, that has direct and indirect impacts on the urban system e.g. increase in temperature, higher exposure to natural disasters, drought, etc. [131].

Due to the complexity of the problem of studying a city's resistance to perturbations, in this study only one type of perturbation will be taken into consideration, that is the modification of buildings geometry (change in shape of buildings or introduction of novel buildings). Within this framework, it is appropriate to reduce the complexity of urban systems to a (still complex) spatial system made of buildings.

The objective of this study is to diagnose spatial buildings layouts that are able to accommodate changes (spatial sustainability) versus layouts that are too tightly packed and represent unsustainable solutions. By comparing with the way amino acids are organized to sustain perturbations, we also propose to look for mechanisms to correct building layouts in cities. Notice that increasing a building surface reduces the space available between buildings, as does an accumulation of cars upon traffic jam. Hence, monitoring and comparing the space left available between buildings in cities, might also be relevant to assess tolerance to traffic jam. Importantly, interstitial spaces, including empty space or unused built space, have been recognized as possible locations to implement strategies for climate-change adaptation, e.g. the installation of photo-voltaic panels or the creation of urban agriculture sites [155] or to be converted into green areas to mitigate the urban-heat-island effect [156]. Thus, the empty space made available for changes of buildings geometries in the spatially-sustainable layouts, may be exploited to the benefit of other sustainability issues.

The problem of identifying a diagnostic measure to discriminate buildings packings that are "too tight", "too loose" or "optimal" is far from trivial: indeed, the questioning of whether dense cities are to be favoured or not is under debate. On the one hand, highly-compact cities reduce energy consumption, improve transport and encourage social interactions, but on the other hand they create problems of health, quality of life and ecology [3, 144, 156–158]. The difficulty comes from the complexity of urban systems, where different objectives (e.g. traffic efficiency versus inclusion of green spaces) are affected by the same parameter (urban density) and thus become inter-dependent.

The representation of cities as spatial systems of buildings, emptied of their population, their roads, their parks etc., is a crude approximation of the urban system but allows focusing on a single aspect of the urban sustainability, that is the ability to face modifications in the geometry of the buildings.

Similarly to protein structures, that are complex spatial systems constituted by atoms arranged in space, urban structures are complex spatial systems constituted by buildings arranged in space, and as such, they can be modeled using spatial networks. The Building Network model is developed to this purpose (see Section 3.3.1), where the geometrical features of the single buildings and of their local arrangements are taken into

consideration.

Chapter 3

Methods

This chapter describes the methods used in the following of the manuscript, divided according to the type of data: protein structures, in-silico protein mutants and urban structures.

For all methods, networks have been manipulated using the Python library NetworkX [159] and the software Gephi [160].

3.1 Protein structure analysis

3.1.1 Protein structures data

All protein structures data have been downloaded from the Protein Data Bank (PDB, <https://www.rcsb.org/>). In the PDB, protein structure data are encoded in the form of a text file in the .pdb format, where the spatial coordinates of the atoms of the protein are reported, as well as the amino acid sequence and the secondary-structure assignment. Each protein structure in the PDB is identified by a unique 4-characters code. In this study, only X-ray structures of proteins are used. Thus, hydrogen atoms are not taken into consideration because they are not detected by X-ray Crystallography. For consistency, hydrogen atoms are discarded even when their position has been reconstructed and included in the PDB structure file of the protein. The number of atoms in all amino acids is checked to make sure that differences in AANs do not arise from different X-ray resolutions.

3.1.2 Amino Acid Network

Starting from the PDB data, protein structures are modeled using the Amino Acid Network (AAN) [23], an established model in Computational Biology. An example of AAN is reported in Fig. 3.1. The AAN is a graph $G = (V, E)$, with V is the set of the N nodes of the network (vertices of the graph) and E the set of links of the network (edges of the graph).

Nodes of the AAN Each node in the AAN corresponds to one amino acid of the protein's sequence (Fig. 3.1, center):

$$V = \{i \mid i \text{ is an amino acid}\} . \quad (3.1)$$

Links of the AAN A link is an atomic interaction defined by atomic proximity: two amino acids i and j are connected if there exists at least one couple of atoms, one belonging to i and one belonging to j , whose distance is lower or equal than a given threshold (Fig. 3.1, right). If not otherwise specified, the threshold is fixed to 5\AA :

$$E = \{(i, j) \mid i, j \in V \text{ with } i \neq j \text{ and } \exists (\text{atom}_i \in i, \text{atom}_j \in j) \text{ with } \text{dist}(\text{atom}_i, \text{atom}_j) \leq 5\text{\AA}\} . \quad (3.2)$$

Link weights in the AAN Each link is weighted according to the number of atomic couples that respect the cutoff condition (Fig. 3.2):

$$w_{ij} = | \{ (\text{atom}_i \in i, \text{atom}_j \in j) \text{ with } \text{dist}(\text{atom}_i, \text{atom}_j) \leq 5\text{\AA} \text{ and } i \neq j \} | . \quad (3.3)$$

As a result, the link weights measure the number of atomic interactions between two amino acids, i.e. the number of atomic couples present in the volume of intersection of the 5\AA -surrounding of each amino acid (Fig. 3.2, red volume).

Packing around amino acids In the AAN, the node degree k_i , defined as the number of neighbors of a node i , measures the amino-acid packing around the amino acid i . The node weight w_i is defined as the sum over all the weights of the links that connect the node i to its neighbors ($w_i = \sum_{j \in N(i)} w_{ij}$, with $N(i)$ the set of neighbors of node i)^a. The node weight measures the atomic packing around the amino acid i . Finally, the node's Neighborhood watch Nw_i is defined as the ratio between the node's weight and the node degree ($Nw_i = w_i/k_i$) [23]. Nw_i represents the average link weight that node i performs with its neighbors, i.e. the average number of atomic interaction amino acid i performs with its neighbors. As a consequence Nw is a measure of the average local packing.

Because of the larger surface and larger number of atoms, bigger amino acids have a higher potential for linking to many other amino acids (high k_i) and performing many atomic interactions (high w_i) compared to smaller amino acids, even though such theoretical limits are never attained by amino acids in protein structures [23]. The Nw measure allows comparing the atomic packing around amino acids of different size, as the total number of atomic interactions (w_i) is normalized by the number of neighbors of the amino acid k_i . An example of the Nw measure is shown in Fig. 3.3: the GLY106 amino acid has a loose neighborhood resulting in a low value of Nw , while the ASN90 amino acid has a tighter packing, resulting in a higher value of Nw . The Nw of the two amino acids has to be employed to compare packings; indeed, the w value is highly dependent on the amino-acid size.

^aThis quantity is often called node strength. However, we employ "node weight" to prevent confusion with the strength of chemical interactions in terms of bonding energies.

Code The AAN is built using Rodrigo Dorantes-Gilardi’s implementation in the Biographs module available at <https://github.com/rodogi/biographs>.

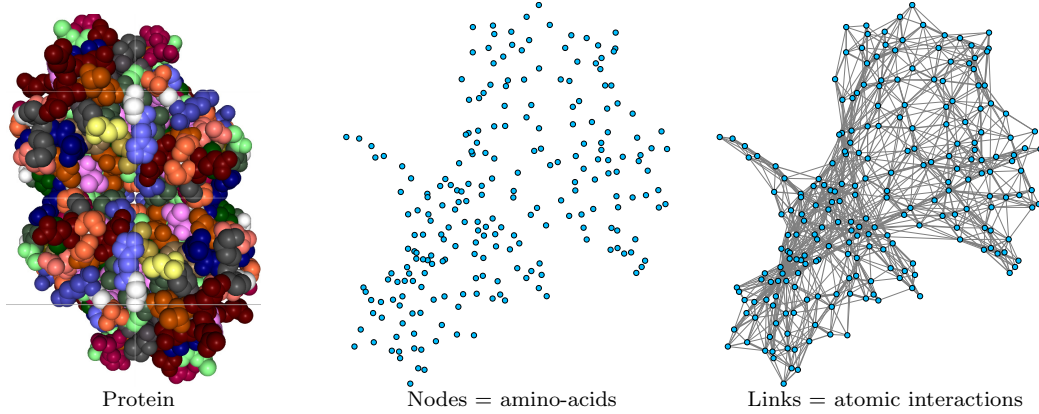


Figure 3.1: Amino Acids Network (AAN) of the human wild-type Transthyretin dimer (PDB: 1F41). *Left*: the protein structure, where the atoms are represented as spheres and colored according to the amino acid they belong to. *Center*: nodes of the AAN, corresponding to the protein amino acids. *Right*: links of the AAN, connecting amino acids that have at least two atoms at distance $\leq 5\text{\AA}$ and weighted according to the number of atomic couples at distance $\leq 5\text{\AA}$. *This figure is published in [24].*

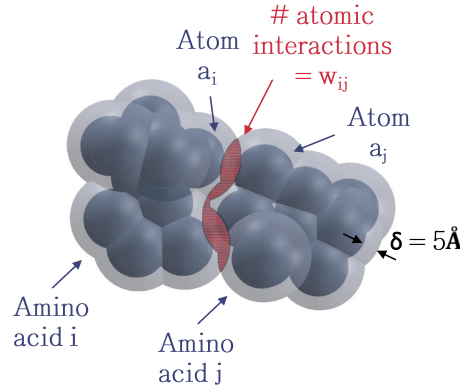


Figure 3.2: Link weight assignment in the AAN. The weight of the link between two nodes i and j is given by the number of atomic interactions between amino acids i and j , i.e. the number of atomic couples in the area of intersection of the 5\AA -surrounding of each amino acid (red volume).

3.1.3 Perturbation Network and Induced Perturbation Network

Perturbation Network The Perturbation Network (PN) is used to compare the AANs of protein structures with the same number of atoms (the sequences can be aligned). For example, to compare the AAN of protein variants [23] or the apo- and bound- states of enzymatic proteins [111]. In Chapter 9, it is used for the analysis of Transthyretin (TTR) single-amino-acid variants [24].

The amino acids interactions of the variant (*var*) and the wild-type (*WT*) are compared through the analysis of the Perturbation Network (PN) of threshold \bar{w} , represented by a graph $G_p = (V_p, E_p)$ with links $E_p = \{(i, j) \in E_{WT} \cup E_{var} \text{ s.t. } |w_{var}(i, j) -$

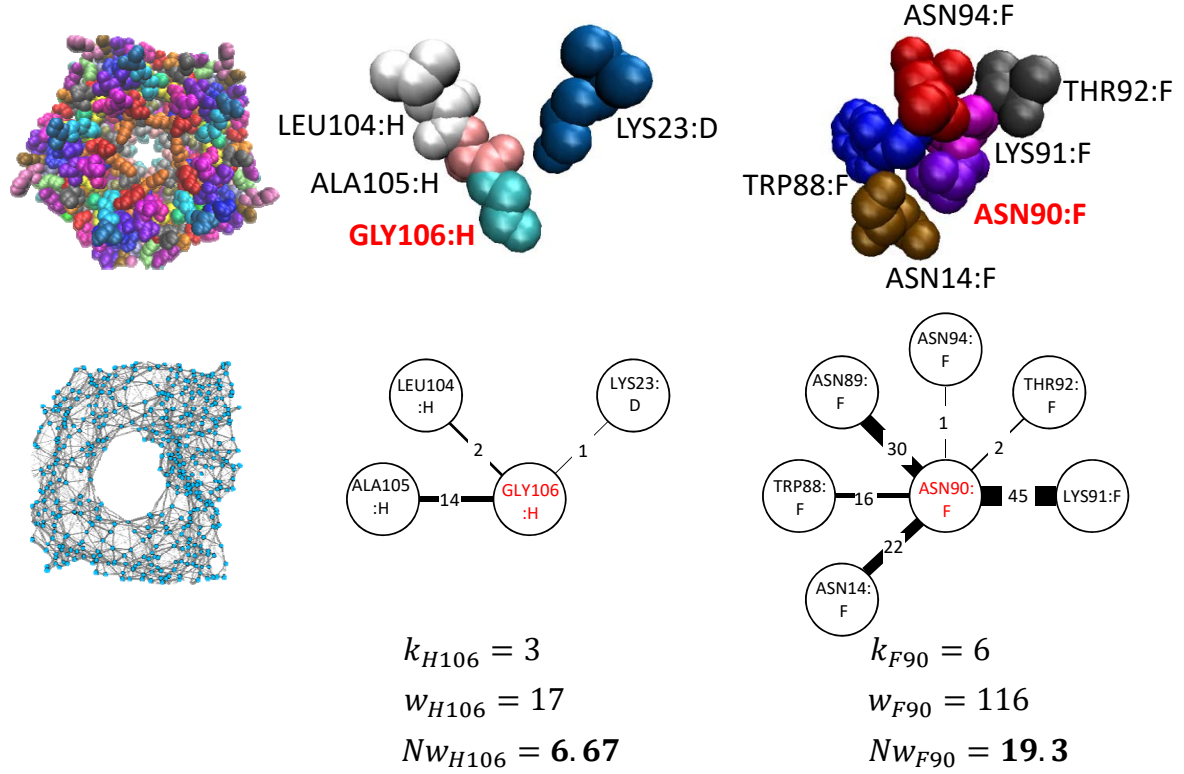


Figure 3.3: Space occupancy around amino acids in the Amino Acid Network. The examples are taken from the structure (top left) and AAN (bottom left) of the PDB structure with code 1B44. Amino acids are labeled by amino acid type, amino acid position in the protein sequence, and the protein chain (one-letter) to which they belong. Using the Nw measure (see main text), the loose packing around the GLY106 amino acid of chain H is distinguished from tight packing around the ASN90 amino acid of chain F.

$w_{WT}(i, j) > \bar{w}\}$, link weights $w_p(i, j) = |\Delta w(i, j)| = |w_{var}(i, j) - w_{WT}(i, j)|$ and link color

$$color(i, j) = \begin{cases} \text{red} & \text{if } w_{var}(i, j) - w_{WT}(i, j) < -\bar{w} \\ \text{green} & \text{if } w_{var}(i, j) - w_{WT}(i, j) > \bar{w} \end{cases}.$$

Thus, amino-acid pairs that make fewer atomic interactions in the variant compared to the WT are corrected by a red link in the PN and inversely amino-acid pairs that make more atomic interactions in the variant compared to the WT are corrected by a green link in the PN. Amino-acid pairs that make the same number of atomic interactions plus or minus \bar{w} in the variant and in the WT are not connected in the PN. The set of nodes V_p is a subset of V , including all nodes for which at least one link has different weight in the variant's AAN compared to the WT AAN according to the \bar{w} threshold (i.e. nodes with degree zero are removed from the PN). Figure 3.4 (third panel from the left) shows a toy example of PN.

Induced Perturbation Network The Induced Perturbation Network (IPN) of a single-amino-acid variant is the connected component of the PN that contains the mutation site.

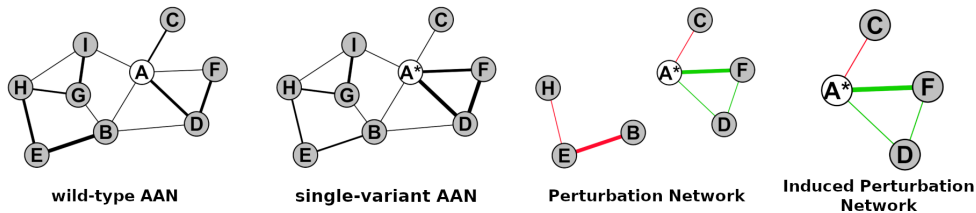


Figure 3.4: Perturbation Network and Induced Perturbation Network of a toy example. From left to right: Amino Acid Network of the wild-type protein; Amino Acid Network of the variant (with single mutation from A to A*); Perturbation Network of the variant; Induced Perturbation Network of the variant. The graphs have been reproduced with the Gephi software [160]. *This figure is published in [24].*

With respect to the PN, the construction of the IPN of a mutation allows to maintain only the information on the perturbations that have propagated from the mutation point to other areas of the protein structure. As a first approximation, the other perturbations, not connected to the mutation point, are here considered to be uncorrelated with the mutation, reflecting alternative atomic arrangements that the protein adopts when crystallized. As a result, the IPN represents the network of changes in atomic contacts in the protein structure that have been caused in the protein structure by the single mutation. Figure 3.4 (right-most panel) shows a toy example of IPN.

The Python code to produce the PN and IPN of a protein structure compared to another variant of the same protein can be found at https://github.com/lorpac/amino_acid_network.

Choice of the perturbation threshold \bar{w} The choice of the threshold \bar{w} impacts the number of nodes and links in the PN and in the IPN. In the present study, $\bar{w} = 4$ is employed, providing a compromise that allows showing perturbations while maintaining the size of the IPN network small enough. A threshold $\bar{w} = 4$ (at least four atomic interactions gained or lost for a link to be in the PN) is reasonably conservative, considering that the average link weight in the WT TTR AAN is $\langle w_{i,j} \rangle = 12.4$ with $std(w_{i,j}) = 10.3$, over a total of 1175 links.

3.1.4 Electrostatic Network

The Electrostatic network (EN) is employed to interpret the experimental protein dynamics signal obtained with Broadband Dielectric Spectroscopy [27]. The EN is a sub-graph of the AAN where only the links connecting charged amino acids of opposite charge are kept. The Intermolecular EN (4D-EN) is a sub-graph of the EN where only the links connecting two amino acids belonging to different chains are kept. Inversely, the Intramolecular EN is a sub-graph of the EN where only the links connecting two amino acids belonging to the same chain are kept. The Induced EN is given by the EN plus all the first neighbors of the charged nodes in the AAN, with the respective links.

The Python code to produce the EN of a protein structure can be found at <https://>

github.com/lorpac/amino_acid_network/blob/master/electrostatic_network.ipynb.

The relations between the AAN, the EN, the Induced EN and the equivalent intermolecular networks are summarized in Fig. 3.5.

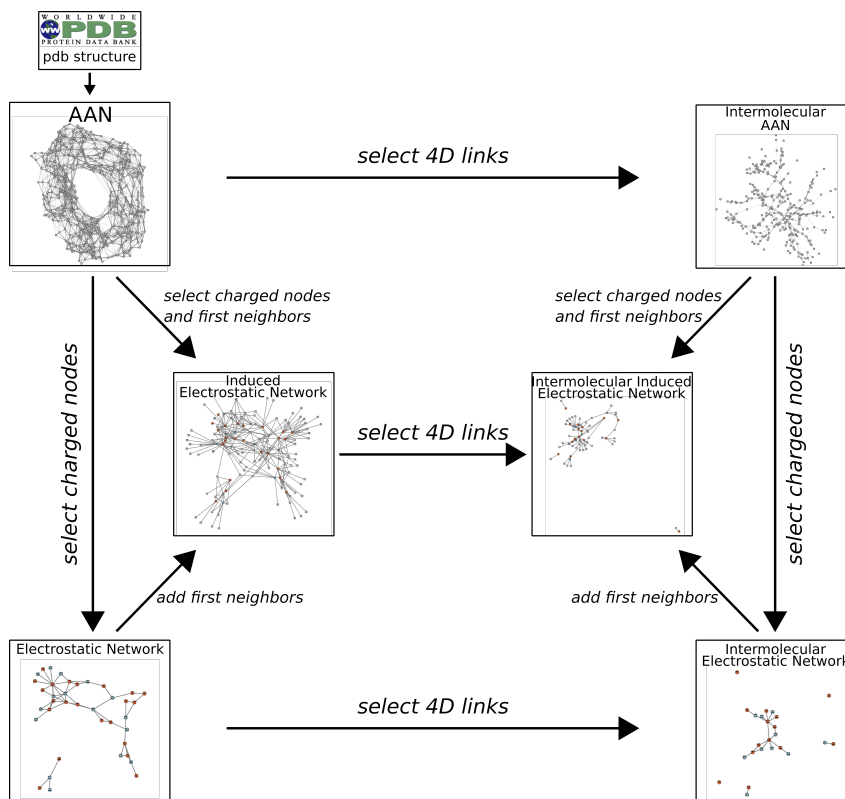


Figure 3.5: Relations between the AAN, the Electrostatic Network, the Induced Electrostatic Network and the equivalent intermolecular networks

3.1.5 Local structural-level allocation of the atomic interactions

The local structural-level allocation of the atomic interactions quantifies the fraction of atomic interactions that an amino acid i allocates to the different structural levels (primary, secondary, tertiary and quaternary structure), as extracted from the AAN of the protein.

Links classification A (i, j) link between amino acid i and amino acid j in the AAN of a protein is classified as: 1D if i and j belong to the same chain and are first neighbors in the protein amino-acids sequence; 2D if i and j belong to the same secondary structure element (α -helix, β -strand or coil as retrieved from the PDB file of the protein structure) and i and j are distant by more than one and less than five positions in the protein amino-acids sequence; 3D if i and j belong to different secondary structure elements or are distant by at least five positions in the amino-acids sequence; 4D if i and j belong to two different chains. With this rule, the contacts between β -strands in β -sheets are classified as 3D.

Local structural-level allocation of the atomic interactions The amount of atomic contacts allocated to each structural level x by an amino acid i is given by the sum of the weights of the links belonging to the structural level (Eq. 3.4):

$$w_{x,i} = \sum_{j=1}^{n_x} w_{i,j} \quad (3.4)$$

with $x \in (1D, 2D, 3D, 4D)$ and n_x the number of neighbors of category x of amino acid i . Then, the fraction of atomic contacts allocated to each structural level x by the amino acid i is given by Eq. 3.5:

$$f_{x,i} = \frac{w_{x,i}}{w_i} . \quad (3.5)$$

We define local structural-level allocation of the atomic interactions of an amino acid i the set of $(f_{1D,i}, f_{2D,i}, f_{3D,i}, f_{4D,i})$ values. By construction, $f_{1D,i} + f_{2D,i} + f_{3D,i} + f_{4D,i} = 1$.

Protein variants comparison The local structural-level allocation of atomic interactions in two protein variants is compared through the differences $(\Delta f_{w_{1D,i}}, \Delta f_{w_{2D,i}}, \Delta f_{w_{3D,i}}, \Delta f_{w_{4D,i}})$ of $(f_{w_{1D,i}}, f_{w_{2D,i}}, f_{w_{3D,i}}, f_{w_{4D,i}})$ values for each amino acid i of the protein structure.

3.2 In-silico mutagenesis

3.2.1 In-silico mutations

All the 19 possible mutations of all the amino acids of a protein are produced in-silico using FoldX [161] version 5, producing 19 single-amino acid mutants per amino acid position. First, the PDB structure of the protein is repaired using the FoldX command RepairPDB. Then, the in-silico mutations are performed on the repaired structure using the BuildModel command. All parameters were set at their default values. The RepairPDB procedure is advised in FoldX before making the in-silico mutation because it makes sure that the WT protein structure on which the mutations are performed does not contain steric clashes, bad torsion angles and side-chains orientations that don't correspond to energy minima. The details can be found in the FoldX documentation (<http://foldxsuite.crg.eu/command/RepairPDB>).

3.2.2 Classification of in-silico amino acid mutations

We classify in-silico single-amino acid mutations of a protein based on the impact they have on the amino-acid neighborhoods in the protein structure, as measured by changes in amino-acid neighborhoods in the Amino Acid Network (AAN).

We say that an amino acid is perturbed by some amino-acid mutation if its set of neighbors in the AAN of the mutant is different from its set of neighbors in the AAN of

the wild-type (WT). We define as source the mutated amino-acid position.

It should be noted that the number of atomic interactions (i.e. the link weights in the AAN) are not taken into consideration: if an amino-acid mutation causes a change in the link-weights made by an amino acid with its neighbors but the list of neighbors of the amino acid is not changed, then the amino acid is said to be unperturbed by the amino-acid mutation. This allows distinguishing two mechanisms of perturbations, named degree- and weight-mechanism, defined below.

The amino-acid mutation class is:

- **Zero (Z)** if no nodes are perturbed;
- **Local (L)** if only the mutation point's first neighbors (distance $\leq 5\text{\AA}$) are perturbed;
- **Far (F)** if amino acids at distance $> 5\text{\AA}$ from the mutation point are perturbed.

A further classification of **L** perturbations is made based on whether the perturbation passes from the source to the neighbors via a rearrangement of atoms not involving a change in the source's neighbors (**Lw**, weight-mechanism) or the source loses (**Ll**) or gains (**Lg**) neighbors (degree-mechanism). If the source loses some neighbors but gains others, the perturbation is classified as **Lg**. This is because it means that the mutant amino acid is able to reach neighbors that were not reachable before.

Similarly, a **F** perturbation is defined as **Fk** (degree-mechanism) if the perturbations spreads through changes in neighborhoods from the source to the other perturbed nodes, **Fw** (weight-mechanism) if the source does not gain or loses neighbors, or **Fkw** if the source gains or loses neighbors but a weight mechanism is involved afterwards.

To practically classify mutations into the Z, Ll, Lw, Lg, Fk, Fw and Fkw classes, it is useful to define two graphs:

The “type 1 graph” that has:

- Nodes: perturbed nodes and the source (even if not perturbed);
- Links: between nodes that are first neighbors in the WT AAN.

The “type 1 graph” provides information on the connectivity of perturbed amino acids before the mutation.

The “type 2 graph” that has:

- Nodes: perturbed nodes;
- Links: lost or gained interactions.

The “type 2 graph” provides information on the changes in connectivity caused by the mutation.

The classification flow chart based on the use of these graphs is reported in Figure 3.6. Table 3.1 summarizes the possible impacts of mutations and the corresponding classification.





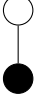
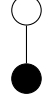




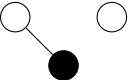
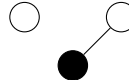
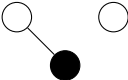
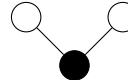
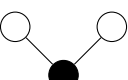
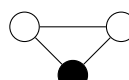
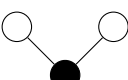
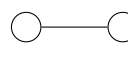
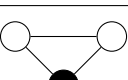
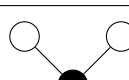
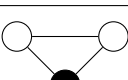
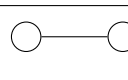
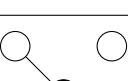


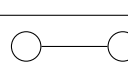
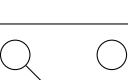

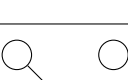



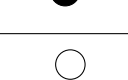
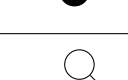




AAN WT	AAN MUT	Type 1	Type 2	Class
				Z
				Ll
				Lg
				Lg
				Lw
				Lw
				Fw
				Fk
				Fw
				Fkw

Table 3.1: Classification of in-silico mutations based on the distance of the perturbed amino acid neighborhoods from the mutation site (see main text). In black, the node corresponding to the mutation point (source).

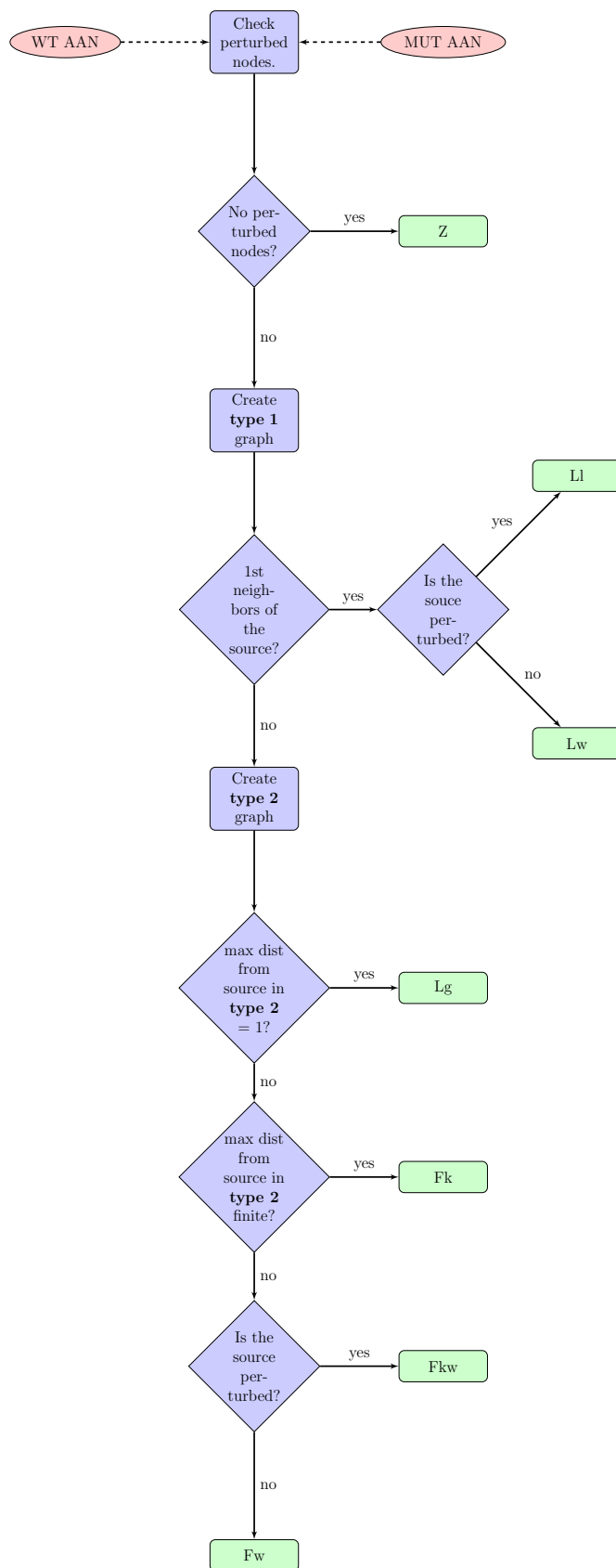


Figure 3.6: Flow chart of the classification procedure of the in-silico amino acid mutations. The classification is based on the distance of the perturbed amino acid neighborhoods from the mutation site (see main text). *Note: the question "first neighbors of the source?" is equivalent to "Is the maximal geodesic distance from the source in "type 1 graph" equal to 0 or 1?"*.

3.2.3 GCAT Network

The directed GCAT network models the directions of perturbations caused by in-silico amino-acid mutations in a protein structure.

In general, a mutation XiY , where the amino acid in position i is mutated from type X to type Y , can provoke a rearrangement of the atoms in the protein structure, that can eventually result in differences in the AAN of the mutant with respect to the WT structure. As in Section 3.2.2, we say that an amino acid i perturbs an amino acid j , with $j \neq i$, if there exists a mutation of the amino acid in position i that causes a change in the neighborhood of the amino acid in position j in the mutant's AAN, compared to the AAN of the wild-type, in terms of list of neighbors.

The GCAT network of a protein is defined as a directed graph $G(V, E)$ with V its set of nodes and E its set of arcs^b, where:

- V is the set of amino acids of the protein;
- An arc $(i, j) \in E$ if i perturbs j .

Once all the AANs of the WT protein and of all its possible single-amino acid mutants (19 times the length of the amino acid sequence) are created, we compare the neighborhood of each node in each mutant AAN respect to the WT AAN and we use this information to build the GCAT network: if a node j has changed neighborhood in the AAN of a $i \rightarrow i'$ mutant with respect to the WT AAN, then we say that the mutation of i has perturbed the amino acid j , and we add a $i \rightarrow j$ arc in the GCAT network. It should be noted that because **F** mutations exists, an arc in the GCAT network may connect nodes that are not chemical neighbors (they are neighbors neither in the WT AAN nor in the mutant's AAN). In the toy-case of Figure 3.7, the $1 \rightarrow 1'$ mutation has caused a change in the neighborhood of nodes 2, 3 and 4 (panel B) compared to the WT AAN (panel A), and thus the arcs from node 1 to nodes 2, 3, and 4 are added in the GCAT network (Panel C). Please note that we do not add self-edges (e.g. the perturbation of the neighborhood of node 1 caused by the mutation of itself) in the GCAT network. Similarly, the arcs from node 2 to nodes 1 and 4 and from node 3 to nodes 1 and 2 are added in the GCAT network because the $2 \rightarrow 2'$ mutation causes a change in the neighborhoods of nodes 1 and 4 and the $3 \rightarrow 3'$ mutation causes a change in the neighborhoods of nodes 1 and 2. The $4 \rightarrow 4'$ mutation does not cause any change in neighborhoods compared to the WT AAN, and thus no arcs leave node 4 in the GCAT network (panel C).

In the GCAT network, an arc leaving a node i (corresponding to the amino acid i) represents the potential perturbation of another amino acid j caused by the mutation of i , while an arc entering node i represents a potential perturbation of i caused by the mutation of another amino acid. Thus, the out-degree $k_{out,i}$ of a node i (i.e. the number

^bLinks in directed networks are called arc.

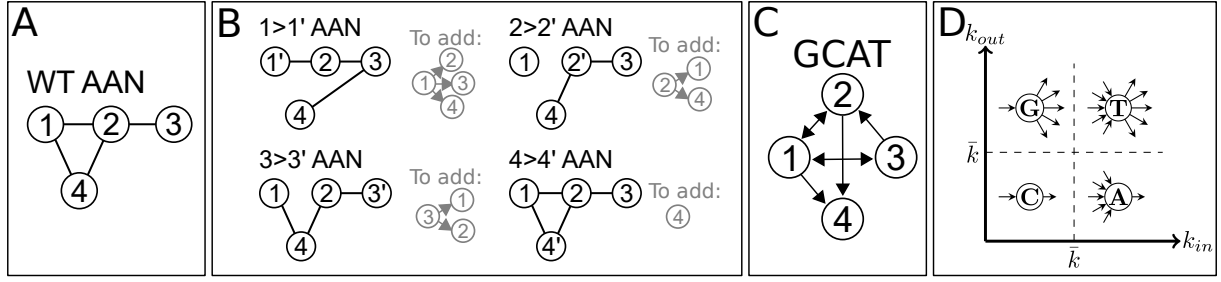


Figure 3.7: Schematics of the construction of the GCAT network. *A*: the Amino Acid Network (AAN) of the wild-type protein in created. *B*: for each amino acid (i.e. each node of the AAN), the AAN of all its 19 mutants is created. For each mutation $i \rightarrow i'$, if a residue j has changed amino-acid neighborhood in the mutant's AAN compared to the WT AAN, then an arc from the mutation site to j is added in the GCAT network. Here, only one mutation per amino acid is represented, for simplicity. *C*: the resulting GCAT network. *D*: Classification of nodes as G (Generate), C (Connect), A (Absorb) and T (Transmit) in the GCAT network, based on their in- and out-degree.

of arcs that leave node i) represents its potential to perturb, while the in-degree $k_{in,i}$ (i.e. the number of arcs that enter node i) represents its potential to be perturbed. We classify the nodes (amino acid positions) in the GCAT network into four classes based on their in- and out-degree:

- **Generate (G)** class: nodes with $k_{in} \leq \bar{k}_{in}$ and $k_{out} > \bar{k}_{out}$;
- **Connect (C)** class: nodes with $k_{in} \leq \bar{k}_{in}$ and $k_{out} \leq \bar{k}_{out}$;
- **Absorb (A)** class: nodes with $k_{in} > \bar{k}_{in}$ and $k_{out} \leq \bar{k}_{out}$;
- **Transmit (T)** class: nodes with $k_{in} > \bar{k}_{in}$ and $k_{out} > \bar{k}_{out}$.

With \bar{k}_{in} the average of k_{in} among all the nodes of the GCAT network and \bar{k}_{out} the average of k_{out} among all the nodes of the GCAT network. It must be noted that by definition $\bar{k}_{in} = \bar{k}_{out} = \bar{k} = \frac{|E|}{N}$, with $|E|$ the number of edges and N the number of nodes, for any directed network. A scheme of the GCAT classification is reported in Figure 3.7D.

3.3 Urban Structures analysis

3.3.1 Building Network

The Building Network (BN) models urban structures. The input is the buildings footprint (Fig. 3.8A, left) available in OpenStreetMaps under the Open Database Licence, see <https://www.openstreetmap.org/copyright/en>. The building footprints are downloaded from OpenStreet Maps using the OSMnx Python package [162] and manipulated through the GeoPandas [163] and Shapely [164] packages.

Nodes of the BN The BN aims at quantifying the space occupied in the urban systems to extract - by difference - the space available for mobility and for the modification of

its components. The nodes of the network represent the components of the system, i.e. the geometrical entities that occupy space, similarly to amino acids that occupy space in the protein structure. Thus, adjacent buildings are merged together, as they represent a unique geometric element, even if they are administratively distinct. Then, the merged buildings are substituted with their convex hull. This is done because the cavities in the building's shape (e.g. courtyards) cannot be exploited for mobility and reconstruction. Then again, overlapping convex hulls are considered as a unique geometrical entity and merged together and replaced by the convex hull of their union. This procedure is repeated iteratively until all convex hulls are disjoint (Fig. 3.8B). The resulting convex hulls of merged buildings represent the nodes of the BN (Fig. 3.8A, middle).

Links of the BN After the merging procedure, the convex hulls of merged buildings define the nodes of the building network. For simplicity, in the following, the convex hulls of merged buildings will be referred to as buildings. Two nodes are linked when two buildings i and j are at distances lower or equal to a distance threshold δ equal to 30m and have no other building in between them (no other building crosses the segment joining the centres of buildings i and j). More precisely, two nodes are connected when at least two points, one of the building i and of the building j are at a distance ≤ 30 m (3.8A, right). The number of linked buildings, namely the number of building neighbors of a node is k , the degree of the node. The threshold distance δ equal to 30 m is chosen because a maximum of around 25m of width for streets appears in the recommendations for urban planning of the municipality of Lyon (France)[165]. Thus, two buildings are linked when the empty space between them mainly represents a road and not a square, a park, a river etc. Buildings separated by a larger distance are not considered linked and therefore are not neighbors. What is important here is to link buildings that are one road apart to assess to our question on urban mobility and growth.

Link weights in the BN Counting the number of neighbors of a building in the BN (i.e. the node degree k_i) is not sufficient to measure the space occupied in the urban space. Indeed, the space occupancy depends on the relative distance (Fig. 3.8C, case 1 versus case 2) and orientation (Fig. 3.8C, case 2 versus case 3) between a building and its neighbors, as well as the size of the building (Fig. 3.8C, case 2 versus case 5) and the size of its neighbors (Fig. 3.8C, case 4 versus case 5 versus case 6). Thus, links have to be weighted with some measure that takes all these parameters into account. This is accomplished using the following procedure (Fig. 3.8D). The surface B_i of a building i (3.8D, blue area) is increased by a length $\delta = 30$ m (3.8D, grey area) to give a new diluted surface BD_i . The same is performed on the building neighbor j to give a surface BD_j diluted of the original surface B_j (3.8D). The intersection area A_i between the area BD_i and the area B_j and the intersection area A_j between the area BD_j and B_i are calculated

and summed to give the link weight w_{ij} (3.8C, red areas).

As a result of the link-weight procedure, lower weights are assigned to links involving more spared surface between two buildings (Fig. 3.8E). The link weight takes into consideration the relative distance (Fig. 3.8E, case 1 versus case 2) and orientation (Fig. 3.8E, case 2 versus case 3) between a building and its neighbors, as well as the size of the building (Fig. 3.8D, case 2 versus case 5) and the size of its neighbors (Fig. 3.8D, case 4 versus case 5 versus case 6).

Packing around buildings The node degree k_i is the number of neighbors of a node in the BN and the node weight is defined as the sum over all the weights of the links that connect the node i to its neighbors ($w_i = \sum_{j \in N(i)} w_{ij}$, with $N(i)$ the set of neighbors of node i). In the BN, the node weight w_i is a measure of the space occupied by a building and its neighbors. However, similarly to the atomic packing of amino acids measured with the AAN, the w_i measure does not allow comparing the space occupancy around buildings of difference sizes, because bigger buildings will have a higher potential for large space occupancy (high w_i because they have a large area) but also a higher potential of sparing this space occupancy with more neighbors (higher k_i because they have a large perimeter) compared to smaller buildings. As for the comparison of the atomic packing around amino acids with the AAN, the comparison of the space occupancy around buildings in the BN is done using the Neighborhood watch measure $Nw_i = w_i/k_i$. The Nw_i values represents the average weight of the links building i performs with its neighbors. As a result, loosely-packed neighborhoods will have a low values of Nw (Fig. 3.9, node A) while tightly-packed neighborhoods will have a large value of Nw (Fig. 3.9, node B): $Nw_B > Nw_A$ despite the fact that building A is larger than building B because building A has many small neighbors and at large distance, while on the contrary building B has few neighbors, tightly packed around it. Finally, building C is the largest node of the BN of Fig. 3.9 but has a moderate value of Nw because again it has many small neighbors, not tightly packed.

Code The Python code used to produce the buildings networks is available at <https://github.com/lorpac/building-network>.

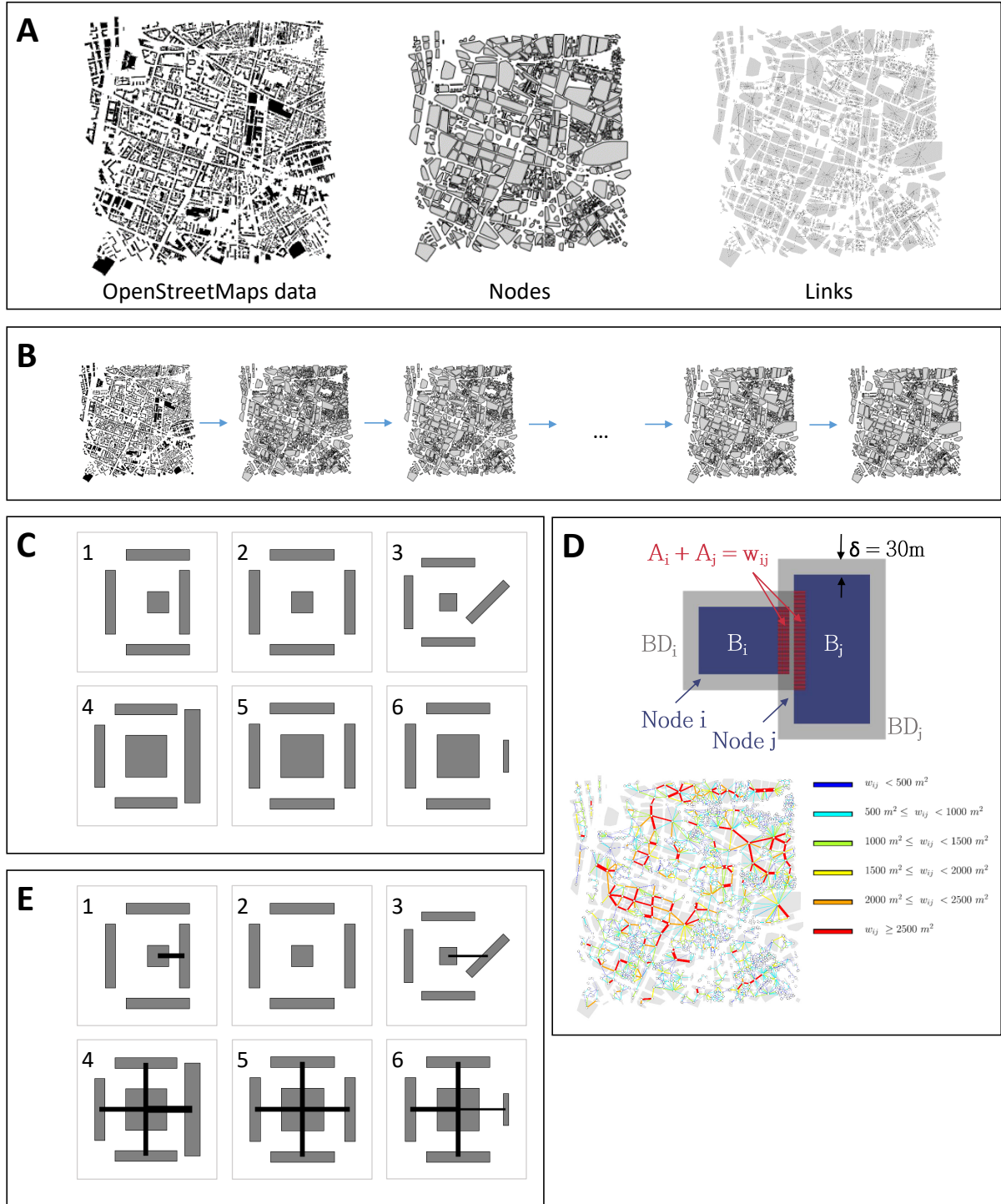


Figure 3.8: Definition of the Building Network (BN). A: The buildings footprint data from OpenStreetMaps (left), the BN nodes (center) and links (right). B: the merging procedure for assigning the BN nodes (see main text). C: examples of different space occupancies depending on the buildings distance and orientation (top) and buildings size (bottom). D: Definition of the link weights (see main text). E: same as C, with the link-weight assignment.

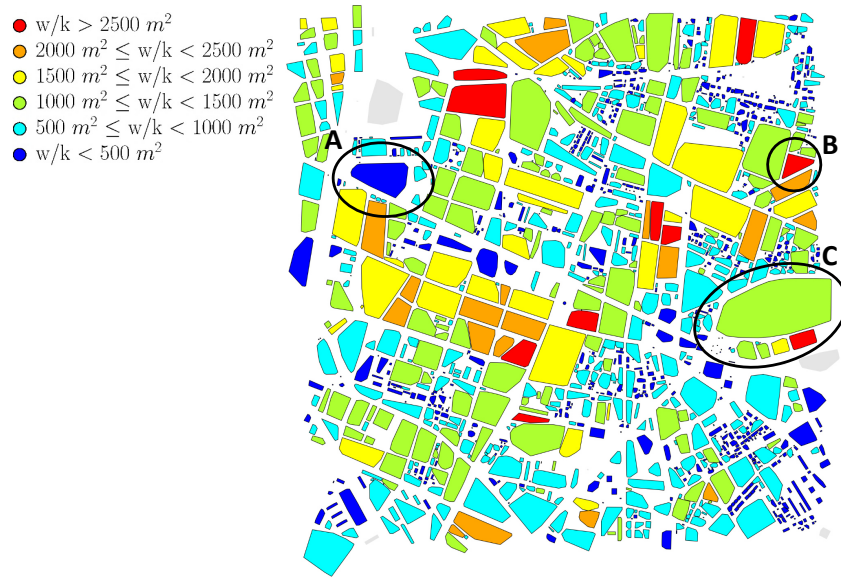


Figure 3.9: Space occupancy around buildings in the Buildings Network. Using the Nw measure (see main text), the loose packings (A) are distinguished from tight packings (B) around buildings, irrespectively of the building size. Building C is the largest building, but it has a moderate value of Nw because it is surrounded by many small buildings, not tightly packed.

3.3.2 Agent-based modeling of random mobility in urban systems

The w_{ij} measure allows quantifying the space occupancy around buildings in the urban space. It is reasonable to assume that larger space occupancy results in a reduced freedom for mobility around the buildings.

Agent-based simulation To quantify the relation between the link weights w_{ij} in the BN and the potential for mobility in the urban system, agent-based simulations of random mobility in the urban system were performed with GAMA [166] using the following procedure.

1. Creation of the BN. The BN of an area of fixed size (2km X 2km) is produced.
2. Set-up of the simulation space. The simulation space consists of a 2km X 2km square filled by the BN nodes. Thus, the simulation space consists of full space (occupied by buildings) and empty space. The simulation space is subdivided into cells (resolution = $n_{cell} \times n_{cell}$), cells can be completely filled by buildings, completely empty, or partially filled.
3. Initialization of the agents. N_{agents} agents are initialized. Each agent is initialized with a random initial position in the empty space and an initial random direction.
4. At each time step, each agent:
 - Tries to move along its direction;
 - If it cannot move (it is stuck by a building or by the space borders), it chooses a new random direction;
 - Is re-initialized with probability $p_{restart}$.

The simulation is halted after N_t time steps. The restart probability $p_{restart}$ is included to assure the exploration of the whole simulation space.

At each time step of the simulation, the number of agents occupying each cell of the simulation space is counted. At the end of the simulation, the mean exit time per cell is calculated, defined as the average time agents spent inside the cell before leaving it.

The following values for the parameters were used: $N_t = 5000$, $N_{agents} = 50$, $p_{restart} = 0.05$, $n_{cell} = 100$.

Code The GAMA model file for the simulation is available at <https://github.com/lorpac/building-network-agents>.

Chapter 4

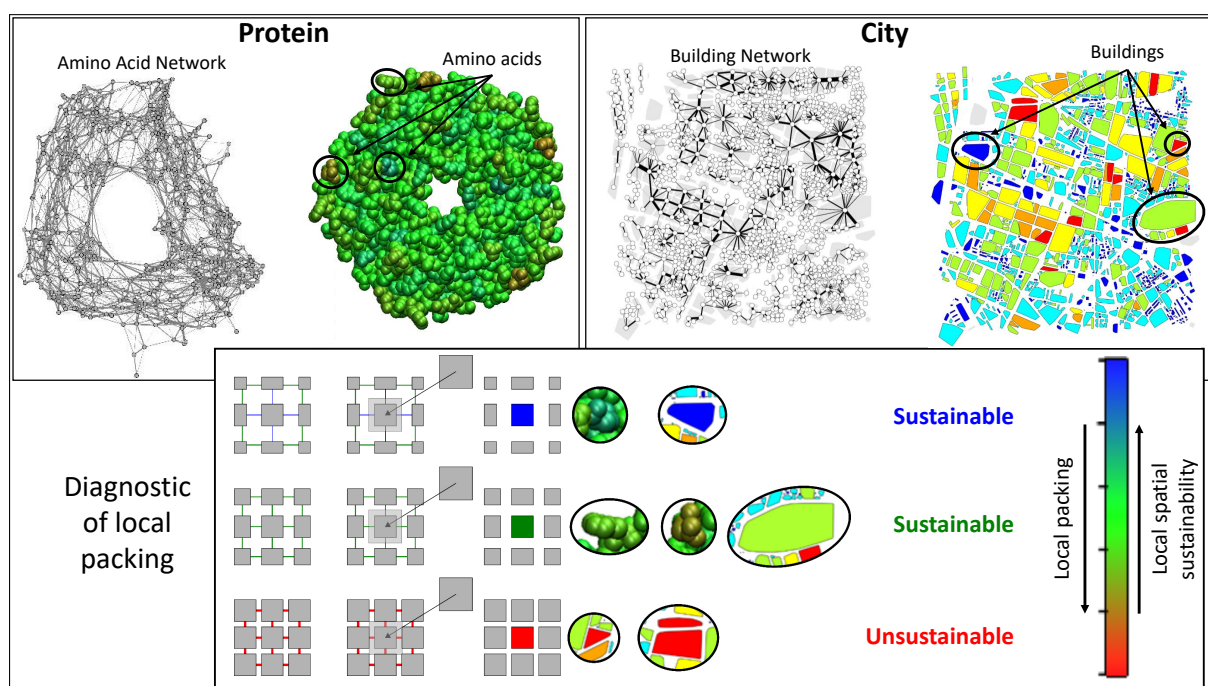
Database analysis: Can the protein- and urban structures accommodate substitutions of their components without structural rearrangement?

Highlights:

- Amino-acids- and buildings packing in protein- and urban structures respectively are measured using spatial network models.
- Based on packing, local structures (one component and its first neighbors) are diagnosed for their ability to tolerate substitutions of amino acids or buildings.
- In proteins, the average local packing around amino acids is constant and moderate so it can accommodate substitutions, making proteins spatially sustainable.
- In cities, packed local structures preclude accommodating substitutions, resulting in spatial unsustainability, an issue for urban growth.

Abstract: Space occupancy in protein- and urban structures is measured using network models. Local structures in proteins leave empty space to be exploited to accommodate substitutions (amino acid mutations). In the city, packing of local structures precludes accommodating substitutions with bigger buildings (spatial unsustainability).

Graphical abstract:



Methods: Amino Acid Network (Section 3.1.2), Building Network (Section 3.3.1).

Publication: Dorantes-Gilardi R., Bourgeat L., Pacini L., Vuillon L. and Lesieur C. In proteins, the structural responses of a position to mutation rely on the Goldilocks principle: Not too many links, not too few. *Physical Chemistry Chemical Physics*, 2018, <https://doi.org/10.1039/c8cp04530e>

4.1 Introduction

Doing a mutation in a spatial system means changing the size, shape and orientation of one component. For a spatial system to keep its functionality upon the mutation of one of its components, it is necessary (yet not sufficient) that the system structure can geometrically accommodate the new component. We define spatially-sustainable a system whose structure leaves enough space to spatially accommodate mutations of its components, regardless of the impact this will have on the system function.

The mutation of a component may be accommodated with or without a rearrangement of the system structure. The mutation of a component is possible with no need of structural rearrangement of the system if there is enough space around the component to accommodate a change in size, shape or orientation. The goal of this chapter is the definition of a spatial measure of the local packing in spatial systems that allows distinguishing cases in which the mutation may be accommodated without the need for structural rearrangement (local spatial sustainability) from the ones where structural rearrangement is necessary to accommodate the mutation (local spatial unsustainability). The measure is applied to the diagnosis of protein- and urban structures. The possibility for the accommodation of mutations with structural rearrangement (multi-scale spatial sustainability) will be assessed in the next chapter.

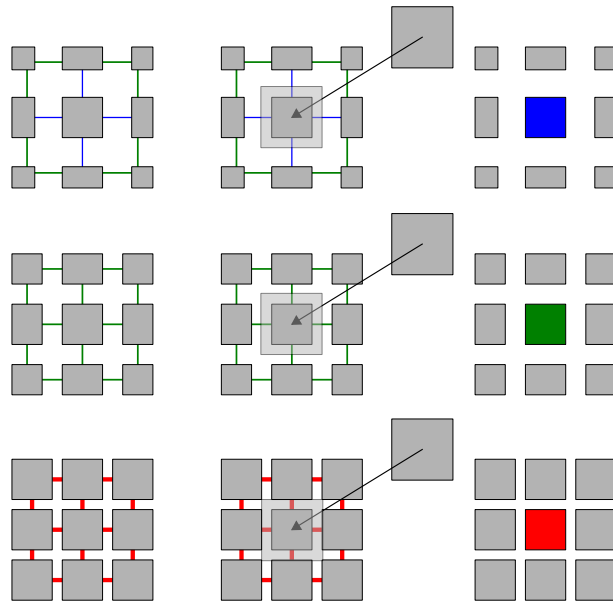


Figure 4.1: Schematics of the substitution of a component in a spatial system. From top to bottom, the neighborhood of the component to be substituted is more and more packed. The blue color corresponds to low packing, the green color corresponds to medium packing, and the red color corresponds to high packing. The component substitution can be accommodated in the first two cases, but not in the last one.

As schematized in Fig. 4.1, substituting one component with a component of bigger size will be possible (first two rows) or not (last row) depending on the local packing of the neighboring components around the component to be mutated. In the last row, the

neighborhood of the component to be mutated is too packed to accommodate the novel component. This simple case shows how the possibility for mutating one component of the system depends on the geometry (size and shape) and the spatial arrangement of the component's neighbors. Thus, local spatial sustainability is a property of neighborhoods (a component and its neighbors) and not a property of the component alone.

Because of the necessity of quantifying the spatial occupancy at the scale of the components' neighborhoods, using spatial networks models to assess the spatial sustainability of the systems comes as a natural choice. The challenge is to define a proper metric to diagnose neighborhoods that can accommodate mutations (locally spatially sustainable) versus neighborhoods that cannot (locally spatially unsustainable).

In the following, weighted spatial network models of protein- and urban structures are proposed as a mean to measure space occupancy and a node measure called the Neighborhood watch is proposed as a diagnostic metric of local packing of neighborhoods in terms of local spatial sustainability.

4.2 Measuring space occupancy in protein- and urban structures.

4.2.1 Spatial network models

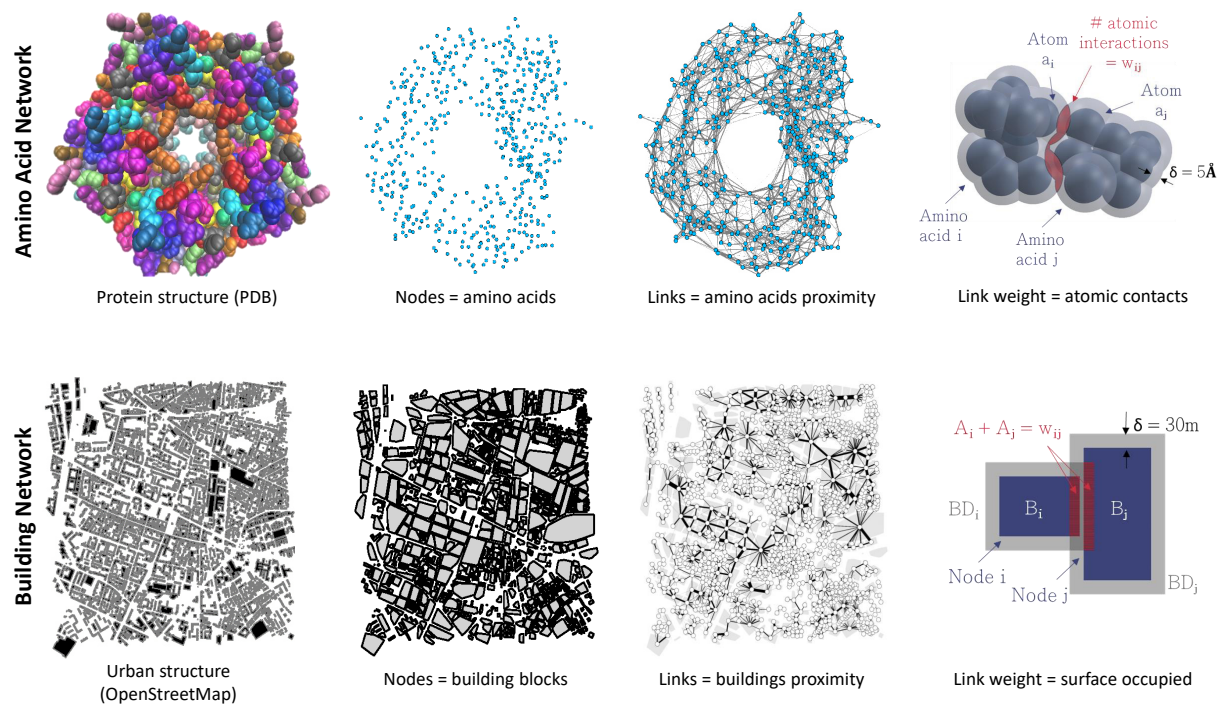


Figure 4.2: Spatial network models of the protein and urban structure. Top: the Amino Acid Network (AAN, Section 3.1.2). Bottom: the Building Network (BN, Section 3.3.1).

Spatial occupancy in protein- and urban structures is measured using the Amino Acid Network (AAN) and Building Network (BN) models, respectively. The AAN is an

established model in Computational Biology, while the BN model is defined ad-hoc for the analysis. Both the AAN and the BN are weighted spatial network models. The details of the construction of the AAN and BN are reported in the Methods (Chapter 3, Sections 3.1.2 and 3.3.1, respectively) and their main features are described in the following. Examples of AAN and BN are reported in Fig. 4.2.

In the AAN of a protein structure, the nodes i are amino acids and the links (i, j) connect amino acids that are at distance $\leq 5\text{\AA}$. The link weights w_{ij} are defined as the number of atomic contacts between the amino acids involved in the links, and thus are a proxy of the volume occupied by atoms in-between the two amino acids (Fig. 4.2, top).

In the BN of a urban structure, the nodes i are building blocks and links (i, j) connect building blocks that are at distance $\leq 30\text{m}$. For simplicity, the nodes of the BN are called simply buildings in the following. The link weights w_{ij} are defined by the area of intersection between the buildings and their area dilated by 30m, and thus are a proxy of the surface occupied in-between the two buildings (Fig. 4.2, bottom).

4.2.2 Node and link measures in the spatial network models

To quantify the space occupied locally by amino acids in proteins and buildings in cities, node and links measures are performed on the AAN and the BN, respectively. Fig. 4.3 shows a schematic of these measures applied to toy examples.

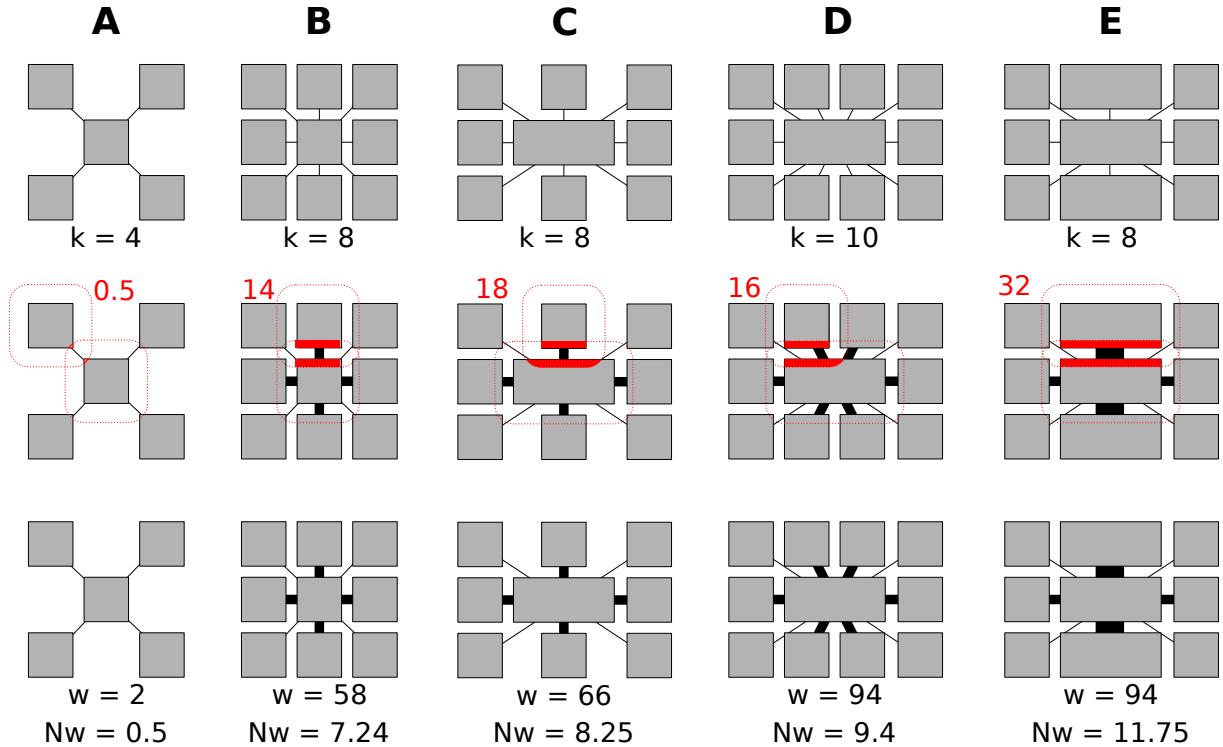


Figure 4.3: Schematics of the space-occupancy measures in the spatial networks. Top: the node degree k_i . Middle: the link weight w_{ij} . Bottom: the node weight w_i and the node Neighborhood watch $Nw_i = w_i/k_i$.

The degree k_i of a node i counts the number of spatial neighbors of the node i (Fig.

4.3, top). Because the node degree does not take into account the geometry of the nodes, it may be used only to compare packing around nodes of exactly the same size and shapes, surrounded by neighbors of exactly the same size and shape (case A versus case B and case C versus case D), while it is not adapted to compare the packing around nodes of different size (case B versus case C) or whose neighbors have different size (case C versus case E).

The link weights (Fig. 4.3, middle) depend on the size and shape of the two nodes forming a link as well as their relative distance and orientation, both in the AAN and the BN (Chapter 3, Sections 3.1.2 and 3.3.1). Thus, the node weight w_i , defined as the sum of the weights of the links made by the node i with its neighbors, is a more accurate measure of the packing around the node compared to the node degree (Fig. 4.3, bottom). As an example, in both cases C and E the node degree is $k = 8$, but $w_C < w_E$, consistent with the fact that more space is occupied in case E with respect to case C. However, using the node weight to compare neighborhoods poses two problems. The first problem is illustrated by the comparison between cases B and D: because node D is bigger than node B, filling their neighborhood with nodes of identical sizes makes node D performing more links ($k_D > k_B$) and higher link weights ($w_D > w_B$), while the packing of space is similar in the two cases. The second problem is illustrated by the comparison of cases D and E: the node weight is identical in the two cases, but the packing is higher in case E.

The two problems are solved by the use of a third measure, called the node Neighborhood watch (Nw), introduced for the study of protein structures [23] and defined as $Nw_i = w_i/k_i$. The Nw represents the average weight of the link made by the node with its neighbors. Since link weights measure space occupancy in the AAN and the BN, the Nw is a measure of the average space occupancy around the nodes of the AAN and BN. Because the node weight is normalized by the node degree in the Nw measure, the difference in Nw between cases B and D is lower compared to the difference in node weights: $w_D/w_B = 1.6$ and $Nw_D/Nw_B = 1.3$. Moreover, $Nw_E > Nw_D$, consistent with the fact that the neighborhood of node E is more tightly packed compared to the neighborhood of D.

4.2.3 Case studies of urban structures

Fig. 4.3 shows the node measures applied to simple toy cases. To provide a more meaningful taste of the relevancy of the Nw metric to measure space occupancy in spatial systems, the example of the BN of a real urban structure is provided in Fig. 4.4. The figure represents the nodes of the BN of the area of Monplaisir in Lyon (France), defined as a 2km x 2km square around the point of geographical coordinates (45.745591, 4.871167), where the nodes are colored based on their Nw value. Building A in Fig. 4.4 is relatively large, but has a low value of Nw because it has several small neighbors, at relatively large distance. On the contrary, building B is relatively small but has a large value of

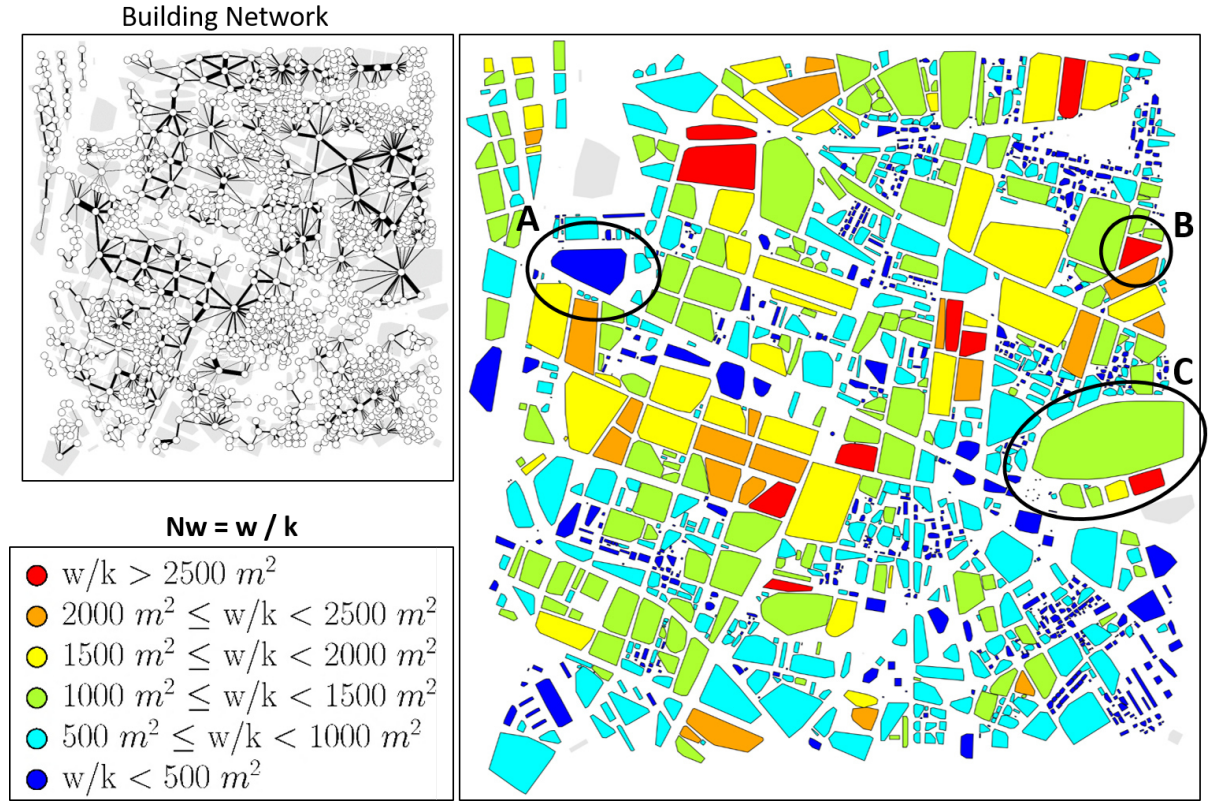


Figure 4.4: Neighborhood watch in the BN of the area of Monplaisir in Lyon (France).

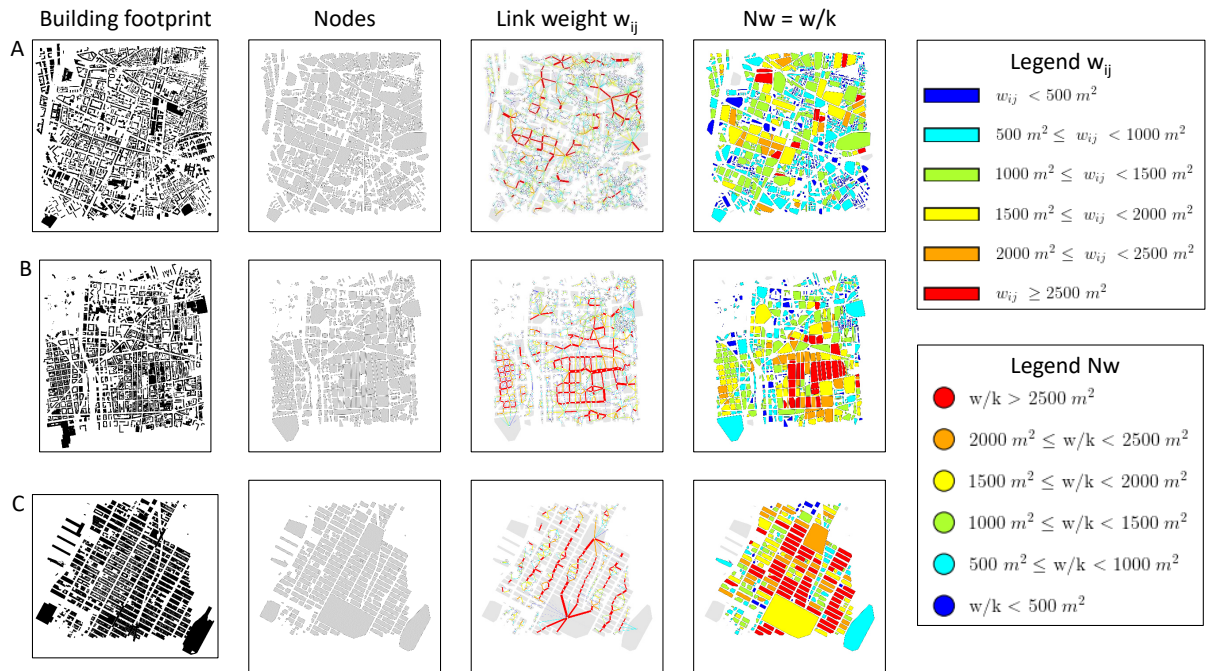


Figure 4.5: Building Networks (BNs) of three urban structures. A: area of Monplaisir in Lyon (France). B: area of Charpenne in Villeurbanne (metropolitan area of Lyon, France). C: area of Manhattan in New York City (USA). From left to right: raw data from OpenStreetMaps, BN nodes, BN links colored based on the link weight w_{ij} , BN nodes colored based on the Nw measure.

Nw because it has few neighbors, that are large and tightly packed around building B.

Finally, building C is the largest of the system but has a moderate value of Nw because again its neighbors are small and not too tightly packed around it.

Fig. 4.5 compares the BN of Monplaisir (Figures 4.4 and 4.5A) with the BNs of two other urban structures: the area of Charpennes in Villeurbanne (metropolitan area of Lyon, France), defined as a 2km x 2km square around the point of geographical coordinates (45.7711641, 4.8658947) (Fig. 4.5B) and the area of Manhattan in New York City (USA), defined as a 2km x 2km square around the point of geographical coordinates (40.763582, -73.988943) (Fig. 4.5C). The building footprints of the three areas were downloaded from OpenStreetMap on November 12th, 2020.

The three cases have similar built surface density but different urban morphology as illustrated by the number of buildings, building surface areas, degree k , weight w and Neighborhood watch Nw (Table 4.1). Manhattan is a more modern urban area with buildings of higher surface areas and less buildings in the 4 km² area, lower node degrees and larger node weights compared to Charpennes and Monplaisir. The values of Neighborhood watch averaged over all the buildings of the area follow the order Monplaisir < Charpennes < Manhattan. In Fig. 4.5, the nodes with maximal values of Nw are colored in red. While red buildings are rare in the case of Monplaisir, a cluster of red buildings is observed in the area of Charpennes and red buildings across all the area of Manhattan are observed.

The comparison of the three case studies shows that urban structures exploit diverse solutions in terms of building sizes and shapes, buildings proximities (k) and buildings packings (Nw). In particular, the Nw values are highly heterogeneous, showing that the Nw measure is sensitive to differences in building layouts.

Table 4.1: Urban morphologies depicted by geometrical and network measures. The building surface area (BSA), degree (k), weight (w) and neighborhood watch Nw are calculated for every individual building block and averaged over all buildings in the city area. The density is calculated as the sum of the building surface areas divided by 4 km².

City area	Building number	$\mu_{BSA} \pm \sigma_{BSA}$ [m ²]	$\mu_k \pm \sigma_k$	$\mu_w \pm \sigma_w$ [km ²]	$\mu_{Nw} \pm \sigma_{Nw}$ [km ²]	Density
Monplaisir	1298	1646 \pm 4350	5.5 \pm 2.6	3.4 \pm 3.4	0.6 \pm 0.5	0.53
Charpennes	776	2801 \pm 5856	4.6 \pm 2.4	4.3 \pm 4.4	0.8 \pm 0.7	0.54
Manhattan	267	8830 \pm 25737	3.0 \pm 1.4	4.9 \pm 3.7	1.7 \pm 1.1	0.58

4.2.4 Case studies of protein structures

Fig. 4.6 shows the AANs of three proteins (PDB 1B44, 1CUL and 1B9L) and their structures with the amino acids colored based on the Nw value of their corresponding nodes in the AAN. Table 4.2 compares the node properties of the three proteins. Compared to the urban cases (Fig. 4.5), it is noticeable how the Nw values are homogeneous among all the amino acids in the three proteins structures and assume moderate values. Few extreme values of Nw exist in the 1CUL and 1B9L proteins (red amino acids in Fig. 4.6), but

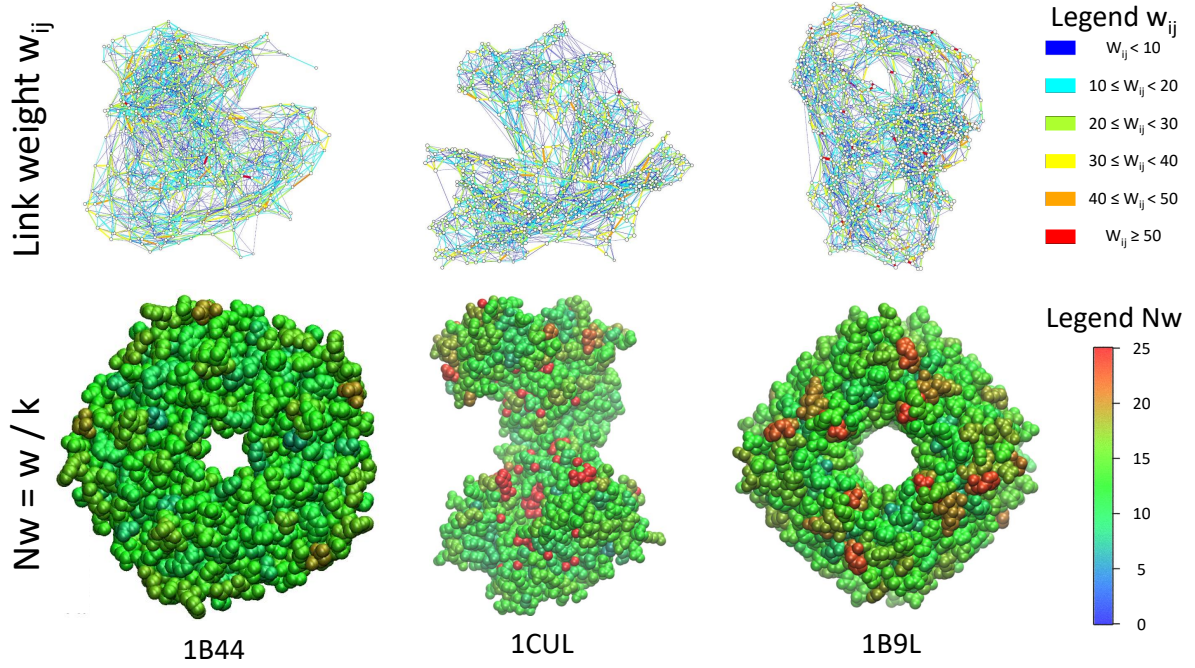


Figure 4.6: Amino Acid Network (AAN) of three protein structures. Top: the AAN, with the links colored based on the link weigh w_{ij} . Bottom: the protein structures, with the amino acids based on the Nw values of the corresponding nodes in the AAN.

they are rare compared to the number of amino acids with moderate Nw , as confirmed by moderate values of the Nw average and standard deviation across all the amino acids of the protein structures (Table 4.2). The values of Table 4.2 also put in evidence how the node characteristics are homogeneous across the three protein structures.

Table 4.2: Protein structures depicted by geometrical and network measures. The amino acid size (number of atoms), degree (k), weight (w) and neighborhood watch Nw are calculated for every individual amino acids and averaged over all amino acids in the protein structure.

Protein	Amino acid number	$\mu_{size} \pm \sigma_{size}$	$\mu_k \pm \sigma_k$	$\mu_w \pm \sigma_w$ [km ²]	$\mu_{Nw} \pm \sigma_{Nw}$
1B44	531	7.9 ± 1.9	11.2 ± 3.6	135 ± 38	12.4 ± 2.3
1CUL	707	7.9 ± 2.1	9.3 ± 3.1	114 ± 36	12.9 ± 3.1
1B9L	952	8.3 ± 1.9	10.0 ± 2.8	126 ± 32	13.1 ± 3.5

In summary, heterogeneous local packing solutions, measured by the Nw values, are observed in the urban case studies, while homogeneous local packing solutions are observed in the protein case studies. The following section is dedicated to the statistical analysis of the Nw metric in a database of protein structures and a database of buildings, with the goal of investigating whether the differences in Nw solutions, heterogeneous in the urban structure and homogeneous in the protein structure, are conserved in large datasets. For seek of clarity, the proteins database is analyzed first and the buildings database is analyzed next.

4.3 Statistics of space occupancy in proteins and cities

4.3.1 Proteins database

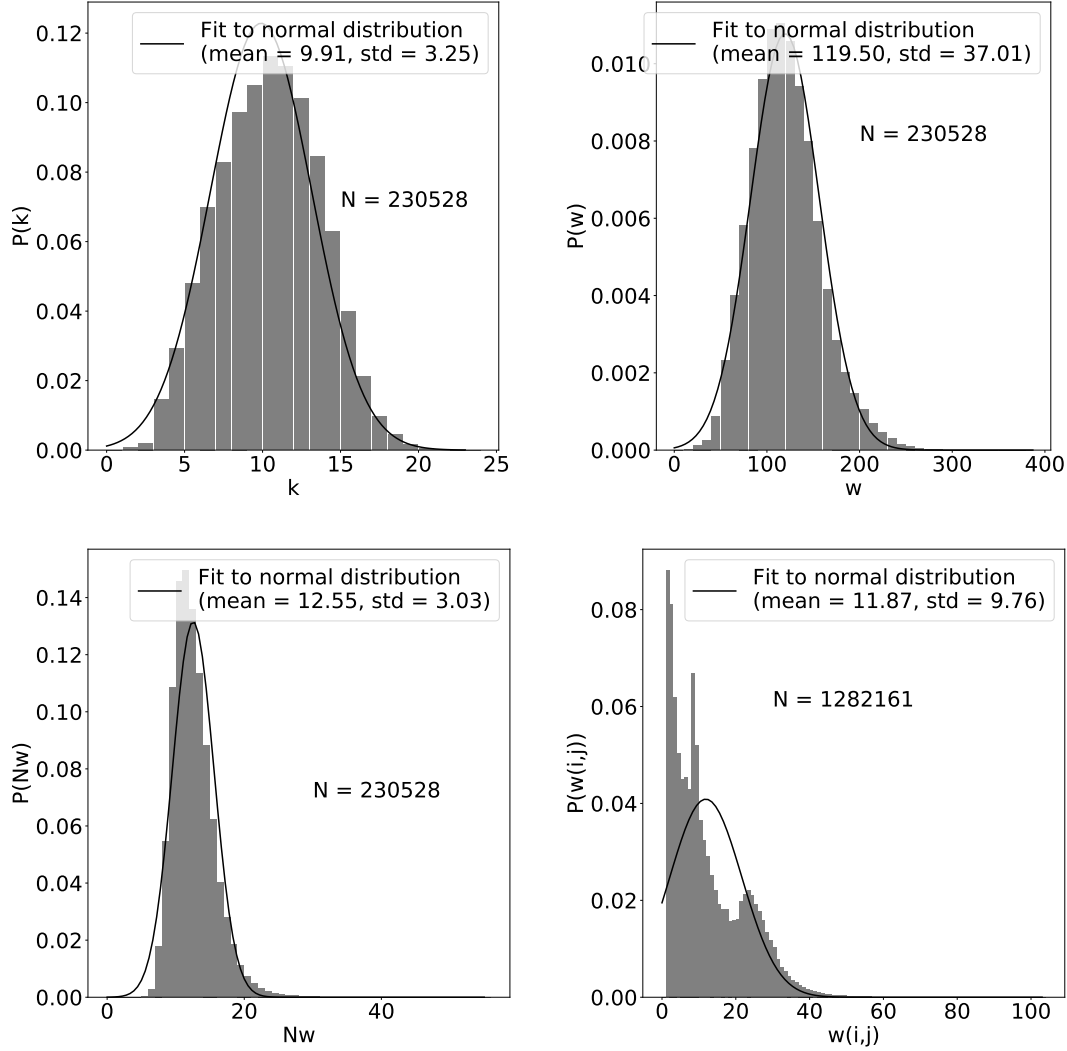


Figure 4.7: Statistics of the node and link properties in the Amino Acid Networks (AANs) the database of globular proteins. Top left: distribution of the node degree k . Top right: distribution of the node weight w . Bottom left: distribution of the node's Neighborhood watch $Nw = w/k$. Bottom right: distribution of the link's weight $w_{i,j}$. All distributions apart from the $P(w_{i,j})$ distribution are well-fitted by a normal distribution.

The AANs of a database of 250 globular protein structures has been analyzed by Rodrigo Dorantes Gilardi in his doctoral work and published in ref. [23]. In the database, all proteins are homo-oligomers made of 2 to 28 chains. The protein sequence lengths range from 28 to 681 amino acids (average: 189, median: 147), resulting in 114 to 7336 nodes per AAN (average: 922, median: 633). The list of the PDB id, chains, and chain lengths for all the proteins of the database is provided in Appendix A. Considering all AANs together, the database comprises 230528 nodes.

Fig. 4.7 shows the statistics of the node measures k , w and Nw and of the link weights

$w(i, j)$ over the whole set of nodes of the database. The distributions $P(k)$, $P(w)$ and $P(Nw)$ show a good fit with normal distributions, meaning that exceptionally-high and exceptionally-low values of k , w and Nw are rare. On the contrary, the $P(w_{ij})$ distribution is not normal. The fact that the node degree k is bounded is expected from the fact that the AAN is spatial (Chapter 2, Section 2.4.3). Similarly, the fact that the w is bounded is expected from the fact that the number of atomic interactions made by an amino acid, measured by the w value, is constrained by the number of atoms of the amino acid and by steric hindrance. However, the fact that the Nw distribution follows a normal distribution is not easily explained, because the Nw represents the average of the link weights made by a node, and the link weights $w(i, j)$ do not follow a normal distribution.

The fact that the Nw measure is normally distributed around a mean value $\mu_{Nw} = 12.6$ in the database does not necessarily imply that the Nw is homogeneous across all amino-acid types and across all protein structures, because the statistics have been performed over all the nodes in the database, regardless the protein structure to which the amino acid belong and regardless the amino-acid type. In the following, the Nw statistics are calculated by considering first the protein structure to which the amino acid belongs and then the amino-acid type of the nodes.

To verify whether the observed statistics for the Nw remain valid when considering single proteins, the average and standard deviation of the Nw values among the nodes of the AAN of each of the proteins in the database have been calculated. Fig. 4.8 shows that the average and standard deviation of the Nw in the AANs taken individually are similar to the average and standard deviation of the Nw across the whole database, confirming that the Nw measure is uniform across the nodes of the AANs also at the single-protein level. It remains to be verified whether the Nw measure is also uniform across different types of amino acids.

Because different amino acids types have different sizes, it may be reasonable to assume that large amino acids would have more neighbors (high k) and/or make a higher number of atomic contacts with their neighbors (high w), compared to small amino acids. Fig. 4.9 reports the distribution of k , w and Nw values based on the amino acid type, where amino acids are ordered from the smallest (GLY) to the largest (TRP) in terms of number of atoms. It is found that the number of amino acid neighbors (k) does not strongly depend on the amino acid type, while on average larger amino acids make a higher number of atomic contacts (w) compared to smaller amino acids. However, the average number of atomic contacts made by the amino acid ($Nw = w/k$) is independent on the amino acid type. Thus, the observation of a normal distribution of the Nw measure around an average value remains valid when the type of amino acid is taken into account.

As discussed in [23], the independence of the Nw on the amino acid type, while w depends on the amino acid type, means that the neighborhood of an amino acid is customized in terms of amino-acid type of the neighbors, so that the average number of

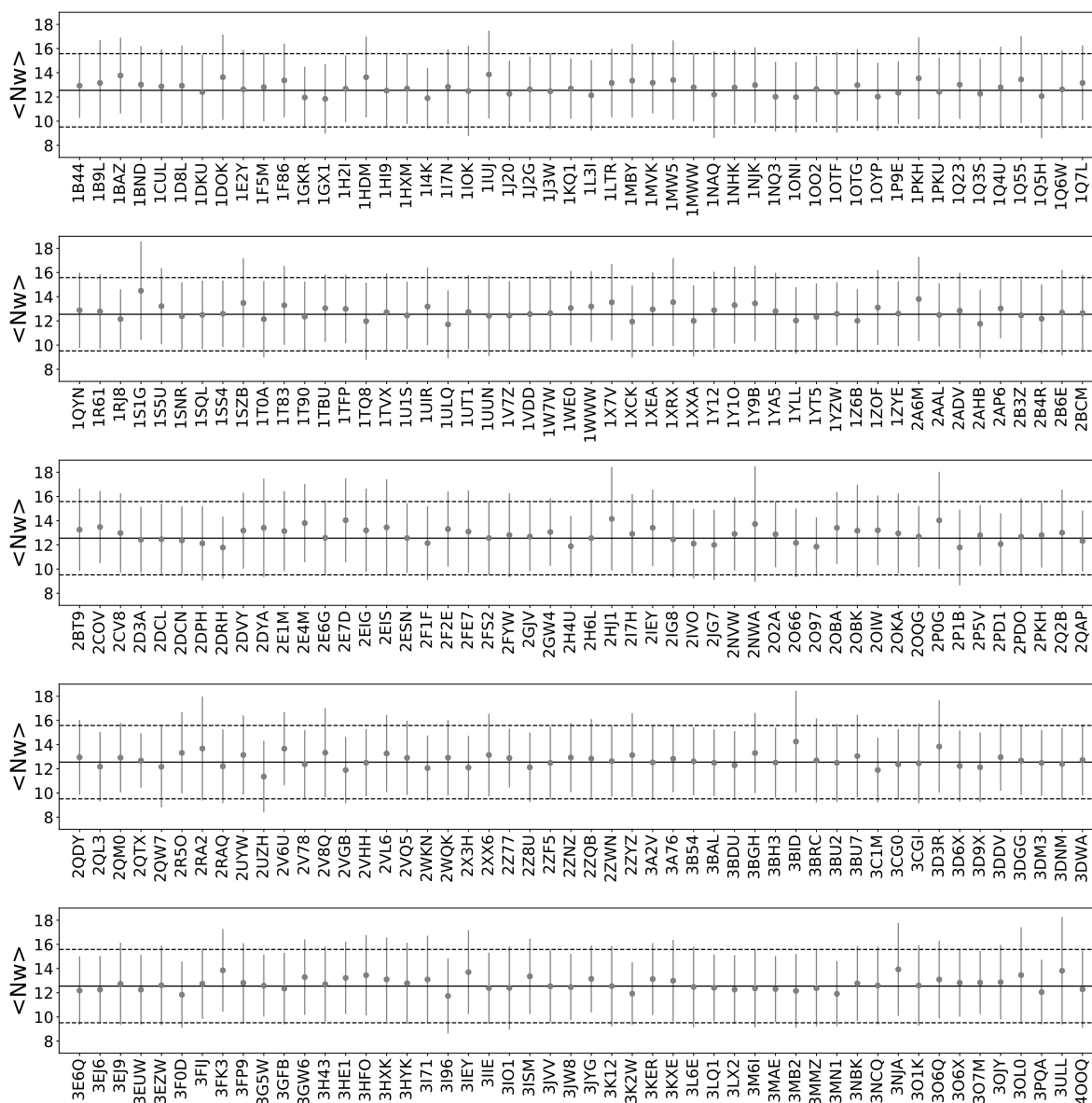


Figure 4.8: Mean values of the Neighborhood watch measure in the AANs of the proteins in the database. Each point represents the average of the Nw value among the nodes of the AAN (one per protein), labeled by the PDB of the protein. The error bars represent the standard deviation. The black solid line marks the average Nw value among all the nodes in the database ($\langle Nw \rangle = 12.55$) and the dashed black lines mark the values of $\langle Nw \rangle \pm \sigma$ with $\sigma = 3.25$ the standard deviation of the Nw among all the nodes of the database (Fig. 4.7, bottom left).

atomic contacts made by all amino acids is constant. Moreover, in the same study it was shown that the observed Nw values are significantly lower than the theoretical upper bound of Nw that would be observed if amino acids would use the totality of their atoms to perform atomic interactions. This is formulated as a Goldilock principle: amino acids make not too many and not too few atomic contacts. Finally, it was proposed that the use of a moderate and uniform value of Nw is the key to the structural robustness of proteins to amino acid mutations: as all amino acids perform a similar number of atomic contacts, all positions in the protein structure may be occupied by virtually all twenty

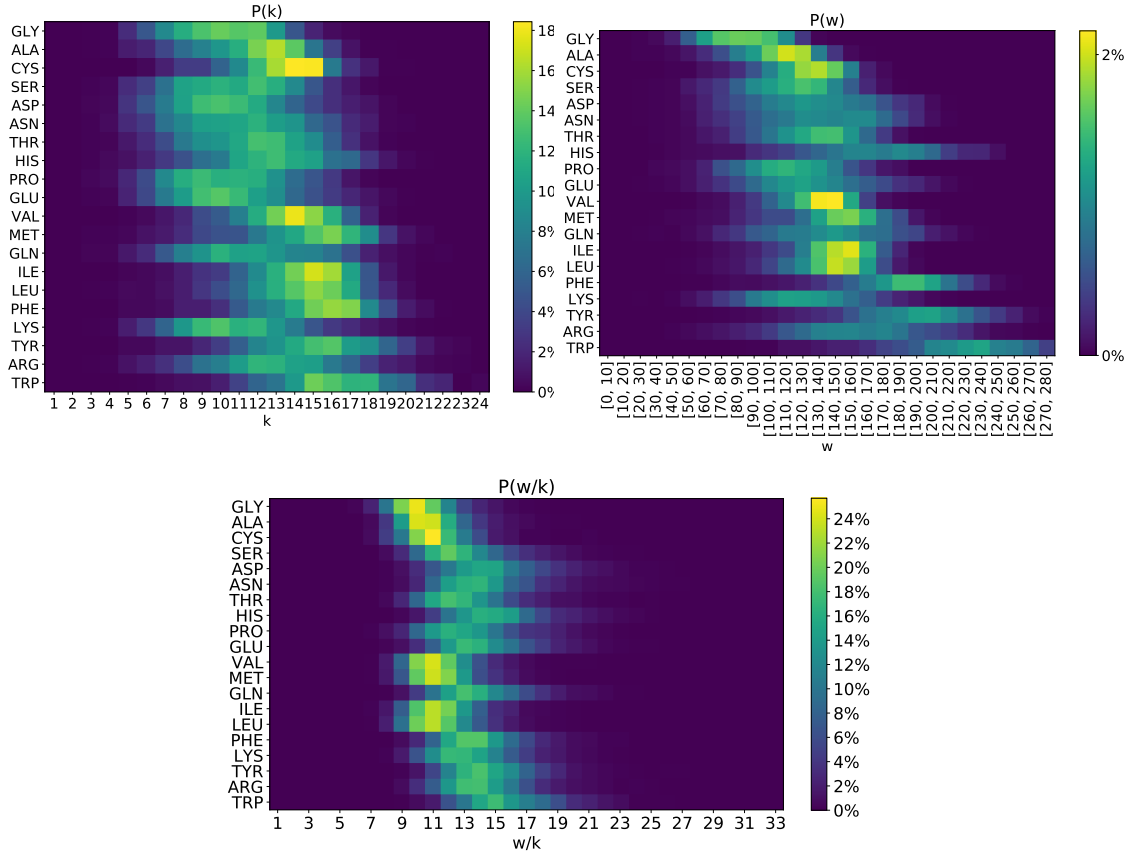


Figure 4.9: Statistics of the node properties in the database of globular proteins based on the amino acid type. From top left to bottom: frequency distribution of node degree k , node weight w and node's average link weight w/k in the database of robust proteins, based on the amino acid type. The amino acids are ordered based on their number of atoms.

amino acid types, and this is made possible by the fact that the neighborhoods of amino acids are never too packed to allow for mutation of the amino acids geometry [23]. In other words, empty space is available in protein structures to accommodate substitutions of the system components, i.e. proteins are locally spatially sustainable.

The Nw measure was found to be moderate in proteins thanks to customized neighborhoods around amino acids. The next question is whether customized neighborhood solutions are possible also in urban structures to provide sustainability (moderate and similar values of Nw across the buildings), regardless on whether the neighborhoods are two-dimensional or three-dimensional.

An issue may rise when transferring the sustainability measure Nw from the three-dimensional protein structures to the two-dimensional urban structures. Indeed, the moderate packing, measured by moderate values of Nw , may be obtained in proteins thanks to the exploitation of the third dimension. Adding buildings heights would provide three-dimensionality to the BN model; however, it would not help solving the question of how to arrange buildings in the urban space to maintain moderate and average packing. To verify the applicability of three-dimensional protein structures as models for two-dimensional

spatial structures, we have calculated the Nw_{2D} measure as follows. 2D-links of the AANs have been defined as links connecting amino acids that belong to the same secondary-structure element (an α -helix, a β -strand, a coil or loop) and are between 2 and 5-positions distant in the protein sequence. 2D-links represent almost two-dimensional spatial constraints in the protein structures. Then, 2D-AANs were built by considering only 2D-links and the Nw_{2D} values for the nodes in the 2D-AANs were calculated. As illustrated by the examples in Fig. 4.10, the 2D-links made by an amino acid (right column) do not lie exactly on a plane; nevertheless, they are much less scattered in the third dimension compared to 3D and 4D links (compare with left column).

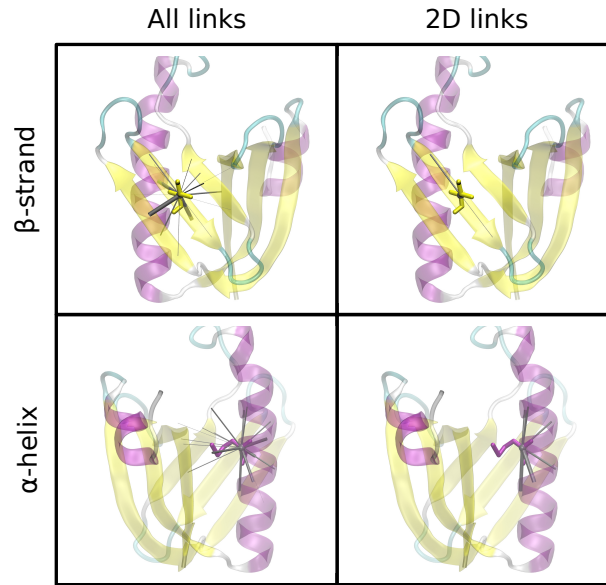


Figure 4.10: Examples of the spatial orientation of the links in the AAN when all links are considered (left) or only 2D links are considered (right) in β -strands (top) and α -helices (bottom)

Across the proteins database, the average of the Nw_{2D} is $\mu_{Nw_{2D}} = 9.1$ and the standard deviation is $\sigma_{Nw_{2D}} = 5.4$. Figure 4.11 shows the $P(Nw_{2D})$ distribution.

Compared to the distribution of the Nw measure, the Nw_{2D} distribution is less well fit to a normal distribution, because of the exceptionally-high number of very low Nw_{2D} values and the presence of some very-high values of Nw_{2D} . Nevertheless, the average and standard deviation remain moderate, confirming that the protein structures are spatially-sustainable also at the quasi-two-dimensional scale. As a consequence, local spatial sustainability of urban structures can be assessed using the Nw as it is done for the diagnostic of protein structures. Moreover, it should be underlined that the Nw values are compared across urban cases and across protein cases. The comparison between proteins and cities is made from the distributions of the Nw values within each type of system.

The urban case studies analysis has revealed that unsustainable solutions (red buildings) are adopted in certain cases (Manhattan and partially Charpennes) but that spatially-sustainable solutions are possible (Monplaisir). The fact that sustainable solutions are possible in the urban system is a proof that low values of Nw can be obtained also in

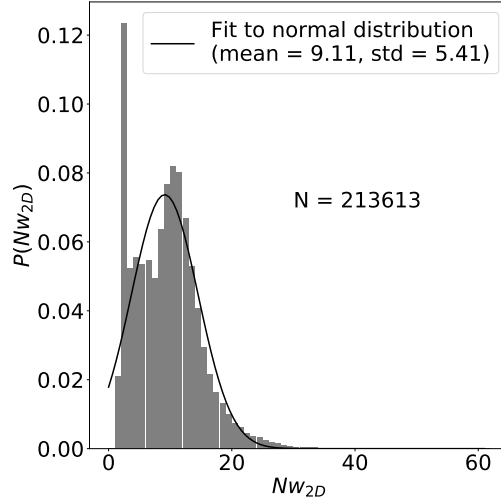


Figure 4.11: Distribution of the Neighborhood watch measure in two-dimensional Amino Acid Networks of the protein database.

two-dimensional systems, and thus the sustainability of proteins does not come from the use of the third dimension. In the following, a database of buildings is studied to determine whether the same variety of solutions of urban neighborhoods is observed in a large dataset.

4.3.2 Buildings database: the city of Lyon

A database of buildings was obtained by creating the BN of the whole city of Lyon (France). The system boundary was set to the administrative boundary reported in OpenStreetMap and the building footprint was downloaded on October 6th, 2020. Fig. 4.12 shows the BN of Lyon on the left and the nodes of the BN on the right, colored according to the Nw value.

The database comprises 12353 buildings, with building surface area ranging from 0.8 m² to 0.8 km² (average: 1339 m², standard deviation: 8178 m², median: 218 m²). The diversity of the buildings shapes was checked from the measure of form factor ϕ of the buildings. The form factor ϕ of a polygon is defined as the ratio between the area of the polygon A and the area of the smallest enclosing circle A_C to the polygon:

$$\phi = \frac{A}{A_C}, \quad \phi \in (0, 1] . \quad (4.1)$$

A perfect circle has $\phi = 1$ and ϕ decreases as the anisotropy (i.e. the elongation along a direction) of the polygon increases (Fig. 4.13, left).

The $N(\phi)$ distribution in the buildings database (Fig. 4.13, right) shows that most of the buildings have shape factor $0.1 < \phi < 0.7$, with near circular ($\phi > 0.8$) and very thin ($\phi < 0.1$) shapes being rare. The broadness of the $N(\phi)$ distribution in the $0.1 < \phi < 0.7$

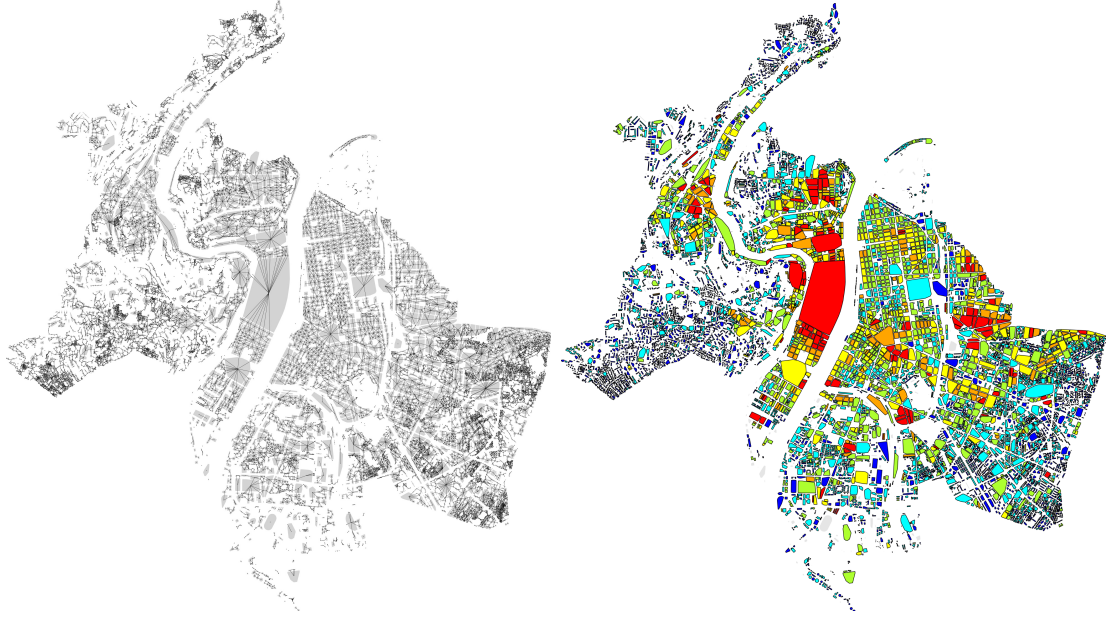


Figure 4.12: Building Network (BN) of the city of Lyon. Left: the BN links. Right: the BN nodes colored based on the Nw value. The color-code is the same as in Fig. 4.5.

range shows that a high diversity in buildings shapes is present in the database.

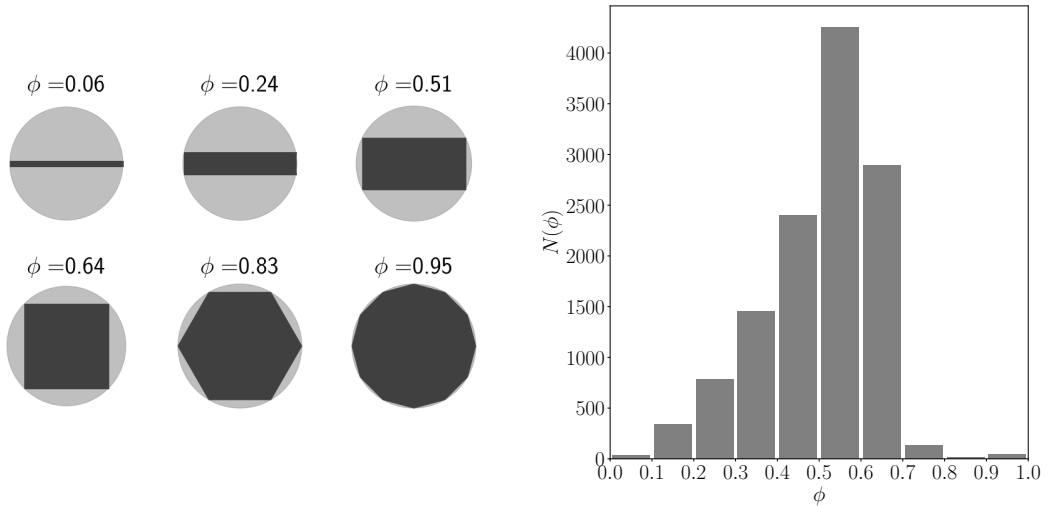


Figure 4.13: Form factor of the buildings of Lyon. Left: Examples of the form factors of polygons. The form factor ϕ is defined as the ratio between the area of the polygon (black) and the area of the smallest enclosing circle (gray). Right: Distribution of the form factor of the buildings in the city of Lyon.

Fig. 4.14 shows the statistics of the node measures k , w and Nw and the link weights w_{ij} in the buildings database. Similarly to the protein database, the degree distribution $P(k)$ is well-fitted to a normal distribution, consistent with the fact that the BN is a spatial network and thus its degree is constrained. However, none of the other quantities (w , Nw and w_{ij}) follow a normal distribution. In particular, the $P(Nw)$ distribution shows the existence of few nodes of very high value of Nw , and a higher heterogeneity (σ/μ) compared to protein structures. It is notable that the Nw values are not homogeneously

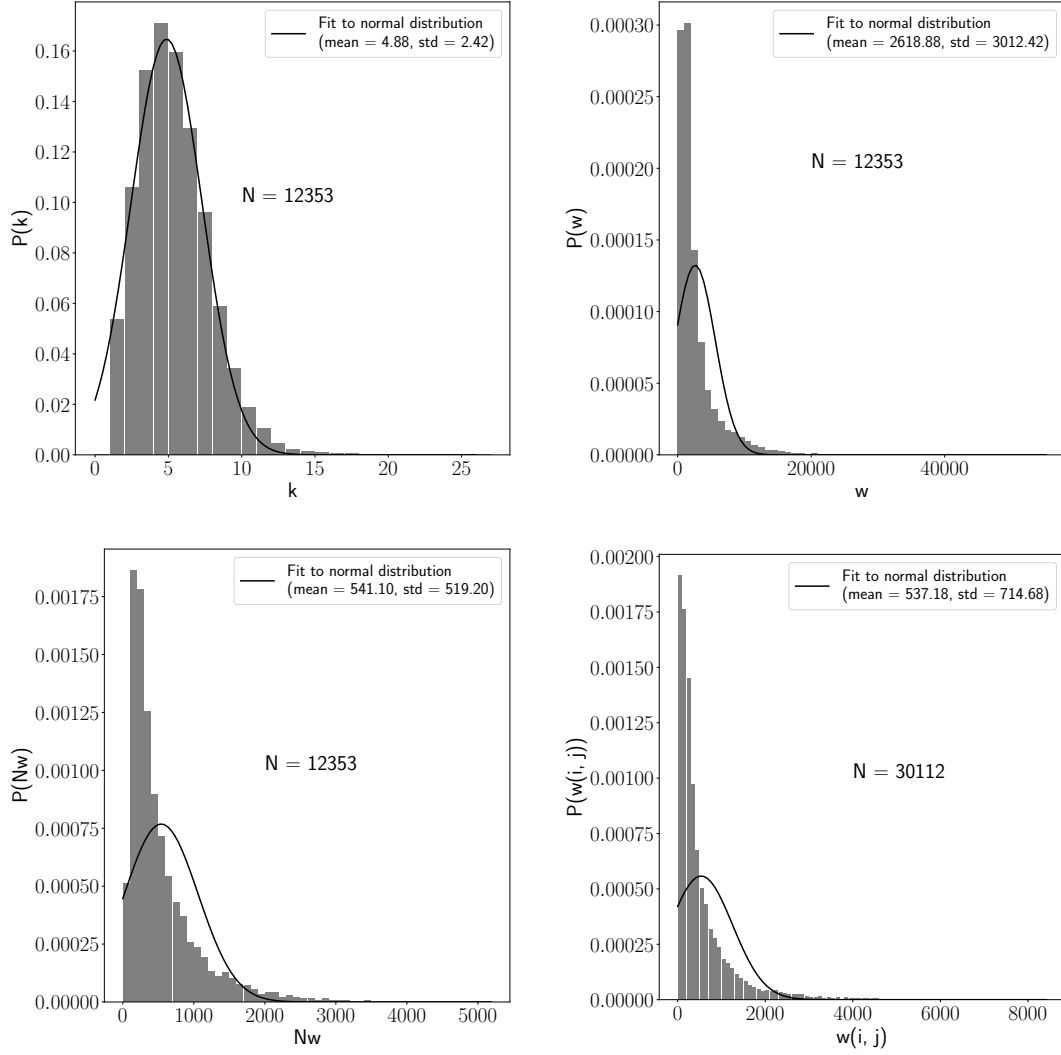


Figure 4.14: Statistics of the node and link properties in the Building Network (BN) of the city of Lyon (France). The black curve represents a normal distribution with the same mean and standard deviation as the observed distribution. Top left: distribution of the node degree k . Top right: distribution of the node weight w . Bottom left: distribution of the node's Neighborhood watch $Nw = w/k$. Bottom right: distribution of the link's weight $w_{i,j}$. Only the $P(k)$ distribution is well-fitted by a normal distribution.

distributed on the urban surface of the city of Lyon (Fig. 4.12, right): high values (red nodes) are found mostly at the city center, and low values (blue nodes) mostly in the suburbs.

To quantify the number of outliers in the Nw distribution, the Herfindahl index of the $P(Nw)$ distribution was calculated. The Herfindahl index (Equation 4.2) is used for example in economy to assess the uniform distribution of holding within banks or industries and monitor anti-trust or risk of failure in the market [167]. For a distribution $P(x)$,

$$H_x = \frac{\sum_1^N (x^2)}{\left(\sum_1^N (x)\right)^2} \quad (4.2)$$

where N is the number of samples of the distribution. H_x varies from $1/N$ to 1. When

$1/H_x$ is close to N , the samples have a uniform distribution, whereas when it is not, $(N - 1/H_x)$ indicates the number of samples that have x deviating from the others. As a result, the fraction of outliers can be estimated as $\frac{N - (1/H_x)}{N}$.

Applying Herfindahl index measure to the $P(Nw)$ distribution of the Neighborhood watch in the city of Lyon, 48% of the buildings are defined as outliers, consistent with the high non-uniformity of the $P(Nw)$ distribution. For comparison, only 3% of the amino acids of the protein database are outliers.

In the protein database, it was found that the Nw measure is independent of the amino-acid type, meaning that amino acids of different size and shape have neighborhoods of similar packing, implying customized spatial arrangements of the amino-acid neighborhoods. To investigate whether this would be possible also in the urban system, we have measured the correlation between three buildings geometrical features - building surface area, perimeter and form factor - and their Nw value (Fig. 4.15).

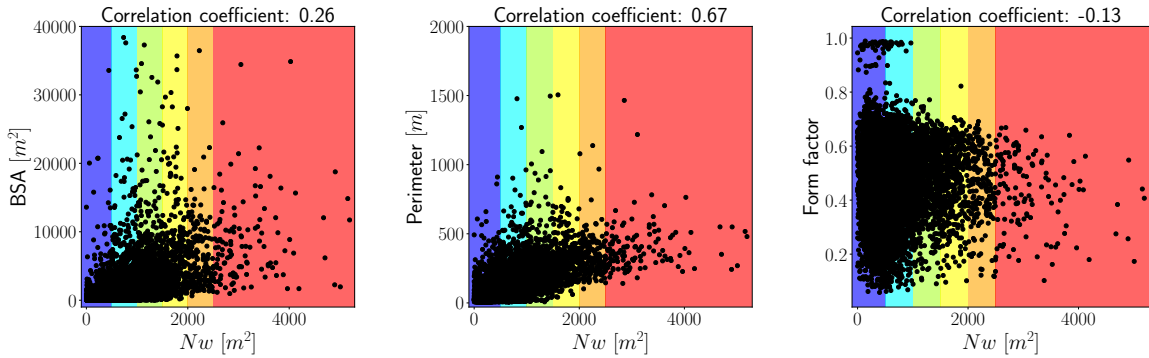


Figure 4.15: Neighborhood watch versus the geometry of the buildings in the buildings database. From left to right: scatter plot of the Nw versus the Building Surface Area (BSA), the perimeter and the form factor of the buildings. For clarity of illustration, the largest building of the system, with $BSA = 785283 \text{ m}^2$, $perimeter = 3610 \text{ m}$ and $Nw = 2906^2$ has been removed from the first two plots.

The correlation coefficients between the Nw and the building surface area, perimeter and form factor are 0.26, 0.67 and -0.13, respectively (Fig. 4.15). Thus, the Nw is not correlated with the building surface area and form factor of the building, while it is partially positively correlated with the perimeter. Nevertheless, there exist cases of buildings of high perimeter ($2p > 1000\text{m}^2$) and moderate Nw (blue, light-blue and green areas in Fig. 4.15). This means that the neighborhood of a building may be chosen to result in a moderate value of Nw , regardless the geometry of the building itself.

To get a better insight on how to modulate the Nw around a building through the choice of the building neighborhood, we have selected examples of buildings with different perimeters (class 1: $2p \sim 100\text{m}$, class 2: $2p \sim 200\text{m}$, class 3: $2p \sim 300\text{m}$) and Nw values. To ease the comparison, we have selected buildings with the same degree $k = 6$ but different weight w (low: $w \sim 2000\text{m}^2$, moderate: $w \sim 5000\text{m}^2$, high: $w \sim 7000\text{m}^2$), resulting in three classes of Nw (low: $Nw \sim 300\text{m}^2$, moderate: $Nw \sim 800\text{m}^2$, high: $Nw \sim 1000\text{m}^2$).

The examples are reported in Fig. 4.16. The result pinpoints that sustainable surface management (low or moderate Nw , first and second column in Fig. 4.16) is achieved by having sizes of building neighbors compensating the size of the central building. Small buildings can accommodate big to medium buildings in their neighborhoods. Medium-sized buildings can accommodate big buildings in their neighborhoods as well but only if associated with small neighbors or medium-sized neighbors at larger distance. Big buildings need to have small or medium-sized building neighbors. If space is seen as a resource to be allocated among the buildings in a neighborhood, moderate values of Nw are obtained by frugal use of the space resource: if the central building is large (i.e. it consumes much resource), then its neighbors need be small (i.e. they consume little resource).

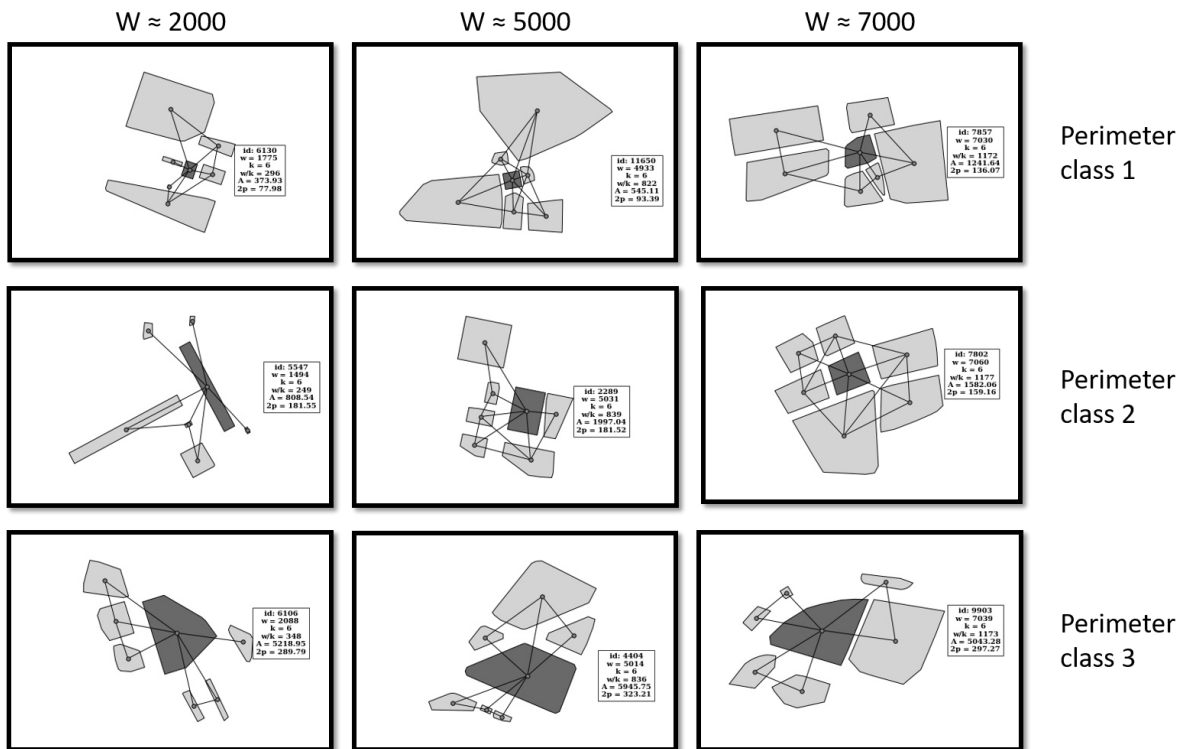


Figure 4.16: Examples of building neighborhoods involving the same number of buildings ($k = 6$) but made by buildings of different size and resulting in different values of node weight w .

4.4 Conclusion

The Nw metric, defined for the analysis of protein structures and here applied to urban structures, is an appropriate diagnostic measure for local spatial sustainability in spatial systems modeled using spatial networks. Moderate values of Nw correspond to neighborhoods that can tolerate a mutation of the central component with no need of structural rearrangement, because enough space is spared between the components (local spatial sustainability). Inversely, extreme values of Nw correspond to neighborhoods where the mutation of the central component would require structural rearrangement, because the

packing of the neighborhood is too high (local spatial unsustainability).

Local packing is uniform moderate in protein structures, providing local spatial sustainability. This means that in general amino acids may be substituted by amino acids of different types. Instead, spatially unsustainable packing solutions are observed in urban structures. Nevertheless, low packing around buildings is possible in the urban system regardless of the geometrical feature of the building itself, provided that the neighboring buildings are well-chosen. This is what is observed in the neighborhoods of amino-acids in protein structures: the amino-acid neighbors are customized to the central amino acid, so that not too many atomic contacts are performed, and not too few.

The BN model proposed here has been shown to be a viable tool to measure local packing in urban systems. The Nw measure may be employed at the early stages of urban design to check for the durability of the proposed solution: urban design solutions resulting in low values of Nw leave space for urban growth, that requires creating new buildings and replacing small buildings with buildings of bigger size. The BN parameter δ , that defines the threshold distance for two buildings to be linked, may be adjusted depending on the spatial scale to be analyzed: smaller scales can be selected using low δ values and larger scales using higher δ values. Moreover, the BN may be constructed using raw buildings footprints data instead of merged buildings to study the space available for pedestrian mobility.

Interestingly, a recent study on Urban Morphology showed that increasing the number of unattached buildings while keeping the Floor Area Ratio (i.e. the ratio between the built area and the total area of a neighborhood, equivalent to the density measure in Table 4.1) improves the air quality of urban neighborhoods [150]. As shown by the comparison of cases D and E in Fig. 4.3, increasing the number of unattached buildings reduces the Nw value. Thus, optimizing the Nw values in the urban design procedure may result in better air circulation and not only better spatial sustainability, even though experimental data or simulations would be needed to verify this hypothesis.

Studying protein- and urban structures in parallel raised questions that would have been hardly formulated from the analysis of one system alone. The study of protein structures allowed to identify a measure, the Nw , that accounts for the local spatial sustainability. Then, the analysis of urban structure has proven that having low and universal values of Nw is not a property of any spatial systems, but results from an optimization. The statistics of the Nw measure among amino acids of different types has evidenced that optimized values of Nw are obtained by choosing neighborhoods adapted to the central amino acid. This evidence has guided the search for optimal (low Nw) and non-optimal (high Nw) neighborhoods of buildings of different size.

Choosing the adapted neighborhood of a building or amino-acid to obtain low values of Nw means choosing the size, shape, orientation and distance of the building neighbors, i.e. determining the link weights $\{w_{ij}\}$ between the building i and its neighbors $\{j\}$.

Because the Nw measure is an average of w_{ij} values, low values of Nw may be obtained even in presence of high values of w_{ij} , if i has many neighbors and the fraction of high-weight links it makes with its neighbors is low. This provides a versatility in the design of sustainable spatial systems: it is not forbidden to place two large buildings (or amino acids) close in space (high w_{ij}), provided that the rest of the neighborhood leaves enough space (low values of w_{ij}) to compensate for the local high-space occupancy. Now, what are the consequences on the system's spatial sustainability if such high-weight links are present in the system, and what happens if high-weight links are scattered in space or instead adjacent one to another? The next chapter assesses this question in relation to the multi-scale spatial sustainability of proteins and cities, defined as the possibility of accommodating mutations of the components via a rearrangement of the spatial structure of the system at multiple scales.

Chapter 5

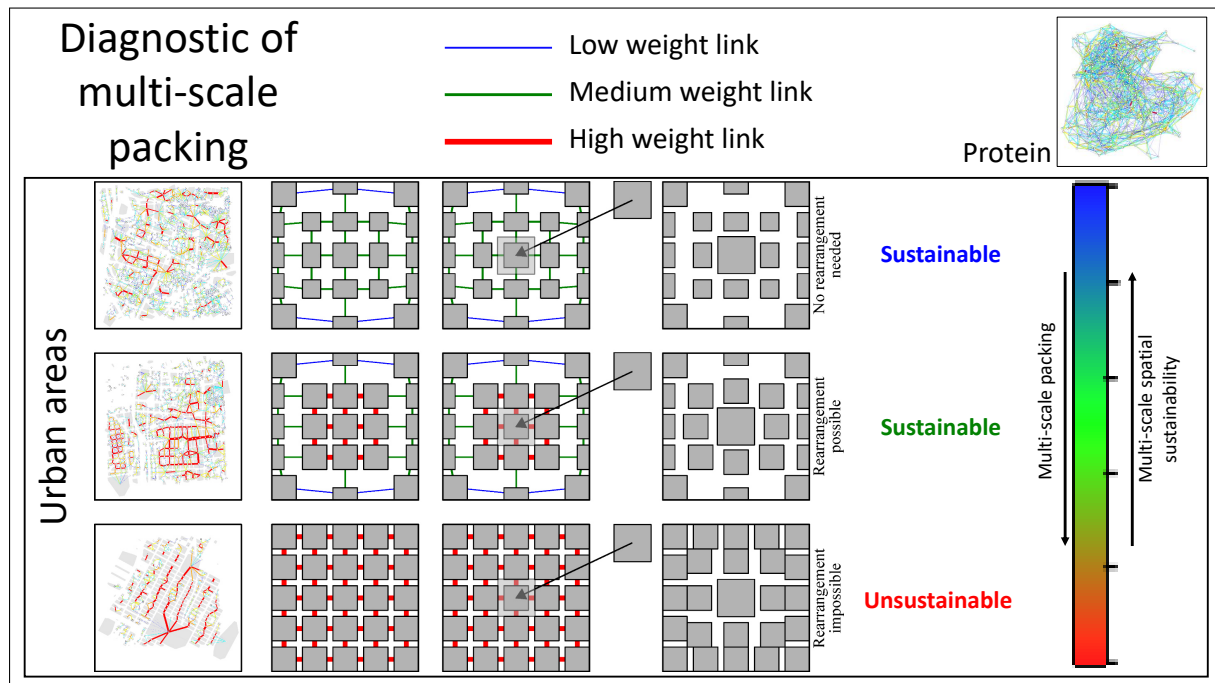
Database analysis: If necessary, can protein- and urban structures accommodate substitutions of their components though structural rearrangement?

Highlights:

- Substitutions in system components can be accommodated using multi-scale structural rearrangement.
- Multi-scale space occupancy in spatial systems is measured by weighted spatial networks.
- Areas of high packing, measured by adjacent high-weight links, preclude rearrangements in cities.
- High-weight links are scattered in protein structures, allowing structural rearrangement at any scale.

Abstract: Where local packing is high, accommodating substitutions of system components in spatial systems may require a rearrangement at large scale. Rearrangement requires empty space around system components, so the potential for rearrangement depends on the multi-scale space occupancy, measured using weighted spatial networks. Areas of high packing are measured by adjacent high-weight links. Contrary to urban systems, high-weight links are never adjacent in proteins, allowing for rearrangement.

Graphical abstract:



Methods: Amino Acid Network (Section 3.1.2), Building Network (Section 3.3.1), Agent-based modeling of random mobility in urban systems (Section 3.3.2).

5.1 Introduction

In the previous chapter, local spatial sustainability of spatial systems was defined as the possibility of substituting components with components of different size and shape with no need of structural rearrangement of the system. In this chapter, the possibility of accommodating substitutions of the system components through a structural rearrangement of the system structure is assessed. We call this property multi-scale spatial sustainability, because the rearrangement may involve multiple spatial scales, from the vicinity of the mutated components (local scale) to a global rearrangement of the system.

Fig. 5.1 shows a schematic of three toy examples of spatial systems where the central component is to be substituted with a component of larger size. In the first row, the component to be substituted has a low value of Nw because its neighborhood is not tightly packed around it, and the substitution can be accommodated with no need for structural rearrangement of the system (Chapter 4). In the second row, the neighborhood of the component to be substituted is tightly packed (high Nw), so structural rearrangement is needed to accommodate the substitution. Structural rearrangement is made possible by the low packing of the second-shell neighborhood of the component to be substituted: empty space is available for the dynamics of the system components. In the last row, the structural packing is tight at all scales, resulting in the impossibility of rearranging the system structure to accommodate the substitution. We say that the first example is spatially sustainable at the local level; the second example is spatially sustainable at the multi-scale level; the third example is spatially unsustainable.

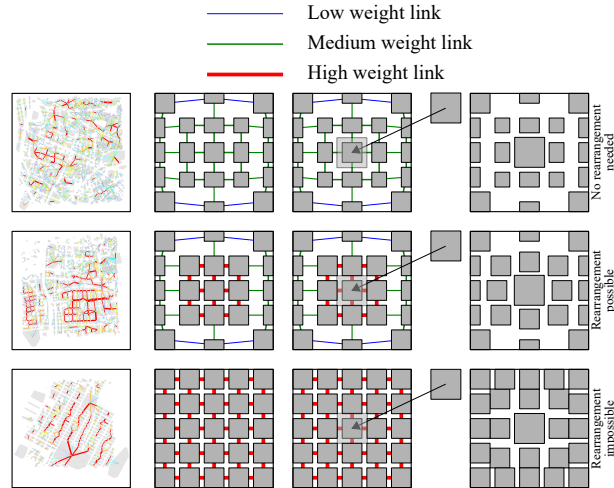


Figure 5.1: Schematics of the substitution of a component in a spatial system through structural rearrangement. From top to bottom, the multi-scale neighborhood of the component to be substituted is more and more packed. Top: the component substitution can be accommodated with no need for rearrangement. Center: the component substitution can be accommodated with structural rearrangement. Bottom: the component substitution cannot be accommodated.

The simple examples of Fig. 5.1 show that multi-scale spatial sustainability depends on the multi-scale space occupancy in the spatial system. The scale up to which rear-

rearrangement of the components is possible depends on the scale up to which packing is not too tight to allow for the mobility of the components. Because link weights in the spatial networks of proteins and cities (AAN and BN models, respectively) measure space occupancy, adjacent high-weight links designate areas that are tightly-packed. Sequences of adjacent high-weight links correspond to areas that are tightly packed at multiple scales: in the second row of Fig. 5.1 packing is tight at the local scale but not at largest scale, while in the third row packing is tight at all scales.

How do high-weight links generate in the spatial system? Fig. 5.1 shows that the link weights depend on the size of the system components: in the three cases, the size and number of the components increases from top to bottom, resulting in an increase in the number of high-weight links (higher packing) from top to bottom. Similarly, for fixed components size, higher-weight links correspond to shorter distances (Chapter 3, Fig. 3.8 E).

The relation between components size, link weights and the possibility for rearrangement (i.e. the mobility of the system components) is better clarified using an analogy with granular materials. This analogy comes from the fact that granular materials may be in a jammed state (no rearrangement of grains is possible) or in an un-jammed state (grains are mobile).

5.2 Jamming of granular materials

Granular materials are made of grains (or particles) in contact and surrounding voids [168]. These materials have been shown to behave as complex fluids: their mechanics can be liquid-like or solid-like depending on the friction coefficient between the particles, the packing density and the applied stress and the size distribution of the grains [169, 170]. The transition from the liquid-like behavior (un-jammed system) to the solid-like behavior (jammed system) is called jamming transition [171–173]. The jammed system’s ability to resist shear or isotropic stress has been explained through the building up of force chains within the contact network of the particles [174]. The contact network of the particles can be represented as a graph where each particle corresponds to a node and a weighted link exists between particles that are in physical contact, with the link weight the magnitude of the contact force between the two particles. Experimentally, the contact network of a granular system can be obtained using photoelastic granular materials [175–177], in which different areas rotate polarized light differently based on their state of stress [178]. Alternatively, discrete element method (DEM) simulations of granular materials can be performed and the synthetic contact network can be obtained [171–173, 179]. Such simulations present the advantage of a complete control over the particles’ size distribution and the loading.

The force chains are connected sub-networks of the contact network where the link

weights are higher than the average [180]. They can be identified in the contact network by filtering links based on their weight and analyzing the topology of the remaining graph [181] or by applying a community detection analysis [177, 182].

Minh and Cheng have studied the influence of the particle-size distribution of granular materials on one-dimensional compression, by simulating the behavior of systems with more or less uniform size distributions and by classifying the links in the contact network as small-small, big-big or small-big based on the size of the two involved particles [171]. They found that big particles are always involved in the transmission of strong forces and that the big particles are all involved in the force chains, while the “weak” regions, that flow under stress, are composed by small particles. In another study of the same authors, it was found that small particles may be involved in the transmission of force chains when present in high percentage ($> 60\%$) [173]. Nevertheless, both studies conclude that higher heterogeneity in particles size results in higher particles’ mobility [171, 173].

In summary,

- Granular materials can be modeled by spatial networks where the nodes corresponds to the grains, the links connect grains that are in contact, and the link weights are the contact forces between the grains;
- Jamming of granular materials occurs with the building-up of force chains, that are connected sub-networks of grain contacts with link weights higher than average;
- Higher heterogeneity in grains sizes results in higher mobility (the system is less jammed).

5.3 The analogy between granular materials, proteins and cities

Protein structures and cities can be described as granular systems where the grains are amino acids and buildings, respectively. The analogy is justified by the fact that in both cases the system components (amino acids and buildings) arrange in space to constitute the system structure, in the same way as grains arrange in space to constitute the granular material’s structure.

In the modeling of granular materials using spatial networks, the links connect grains that are in physical contact. In a similar way, the links in the AAN and the BN connect “grains” that are in proximity. Links in the AAN connect amino acids that would be in physical contact if their volume were enlarged by 5\AA and the links in the BN connect buildings that would be in contact if their surface were enlarged by 30 m (Chapter 4, Fig. 4.2, last column).

As a first approximation, ignoring friction, the contact force between two grains in a granular material is proportional to the contact surface between the two grains (Eq. 5.1). If σ is the normal stress acting on the grains and A_c is the contact surface between the

two grains (Fig. 5.2), then the contact force is

$$F_c = \sigma A_c . \quad (5.1)$$

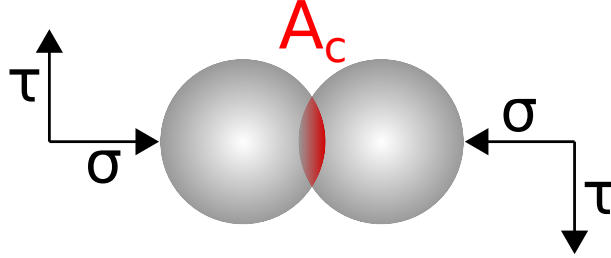


Figure 5.2: Contact between grains in a granular material. A_c is the contact area and σ is the normal stress.

In the spatial network of granular materials, the link weights are given by the contact forces, thus the link weights are proportional to the contact surface between the two grains.

In the AAN, the link weights are proportional to the “contact surface” between the amino acid enlarged by 5\AA and similarly in the BN the link weights are proportional to the “contact surface” between the buildings enlarged by 30 m (Chapter 4, Fig. 4.2, last column, red areas).

Because unjammed systems allow for mobility of their components, as needed for structural rearrangement upon substitution of the system components, we classify them as spatially sustainable at the multi-scale level. Inversely, jammed systems do not allow for mobility of the components and we classify them as spatially unsustainable. In analogy with granular materials, proteins and cities will be considered as jammed (spatially unsustainable) or unjammed (spatially sustainable at the multi-scale level) depending on whether “force chains” of adjacent high-weight links are present in the AAN and BN, respectively. The issue is then how to measure whether the AANs and BNs contain “force chains” or not.

5.4 Detecting “force chains” in spatial networks

A network is said to be connected if all the distances between the nodes of the network are finite, i.e. it is possible to “travel” from any node of the network to any other node of the network using a finite number of steps (Fig. 5.3, left). If a network is disconnected, then it is made of a finite number of connected components (Fig. 5.3, right). The largest connected component (LCC) of the network is the connected component that contains the highest number of nodes. The size of the largest connected component is given by the number of nodes it contains.

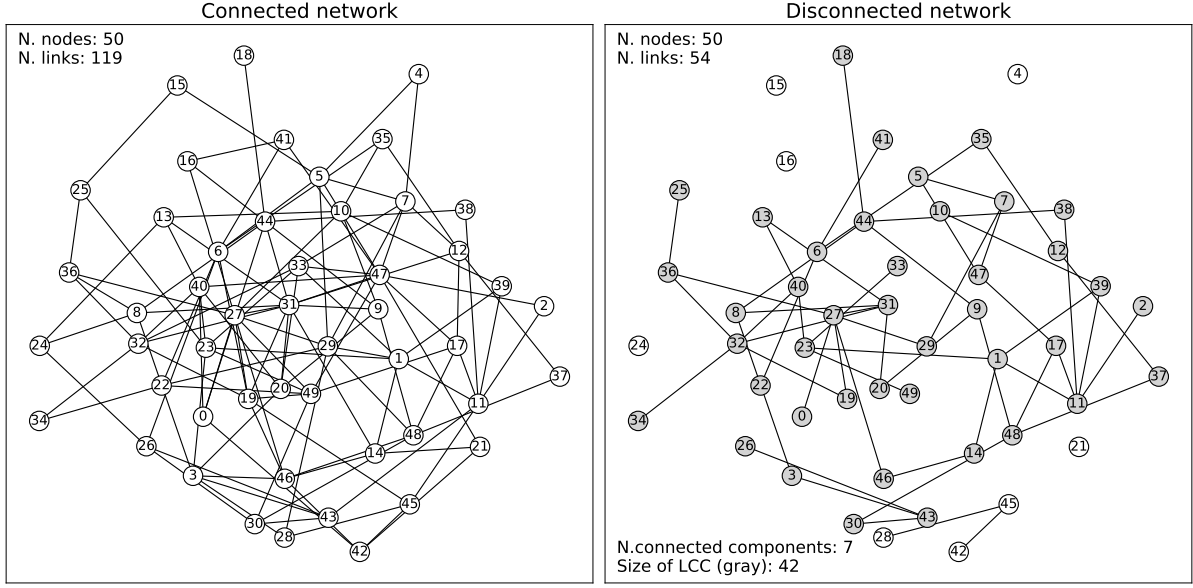


Figure 5.3: Connected and disconnected networks. Left: a connected network. Right: a disconnected network. The largest connected component (LCC) is colored in gray.

To detect “force chains” in the spatial networks the following approach is adopted. The links are classified as low-weight or high-weight depending on whether the link weight is respectively lower or higher than $\mu_{w_{ij}} + 2\sigma_{w_{ij}}$, with $\mu_{w_{ij}}$ the average link weight and $\sigma_{w_{ij}}$ the standard deviation, based on the statics of the proteins and buildings databases (Chapter 4). Then, the low-weight links are removed from the spatial network and the number of connected components and size of the LCC of the remaining sub-network are calculated.

As schematized in Fig. 5.4, a network with a force chain (top) will have fewer connected components and larger LCC size in the sub-network of high-weight links compared to a network with no force chains (bottom) with the same number of nodes, links and high-weight links.

In the following, jamming in proteins and urban structures is assessed from the analysis of the sub-networks of high-weight links in the AANs and BNs, respectively.

5.5 Are proteins and urban structures jammed?

In Chapter 4 it was found that the distributions of the link weights in both the AANs and the BNs are right-skewed, meaning that most of the link weights are low, but there exists a significant amount of links of exceptionally-high weight (Fig. 4.7 and Fig. 4.14).

In the proteins database, the average link weight is $\mu_{w_{ij}} = 10.24$ and the standard deviation is $\sigma_{w_{ij}} = 9.78$. We define high-weight links in AANs the links having $w_{ij} > \mu_{w_{ij}} + 2\sigma_{w_{ij}} \sim 30$.

In the buildings database, the average link weight is $\mu_{w_{ij}} = 537.18 \text{ m}^2$ and the standard

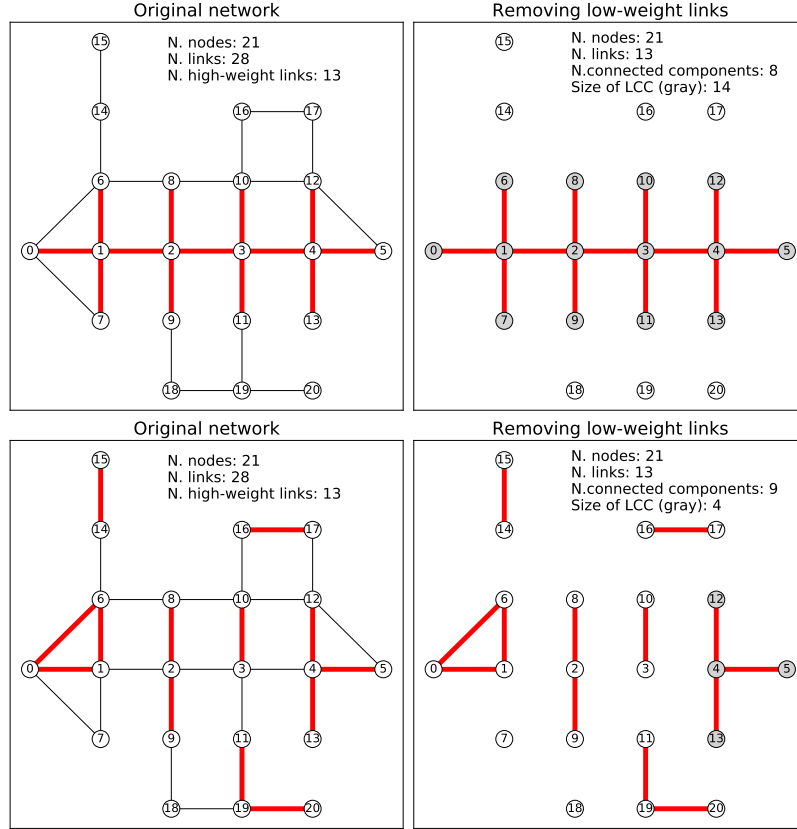


Figure 5.4: Schematics of the methodology to extract “force chains” of high-weight links in the networks. High-weight links are red and thick and low-weight links are black and thin. Top: network with a force chain of high-weight links. After removing low-weight links, the largest connected component (LCC) of the remaining disconnected network corresponds to the force chain. Bottom: network with no force chains of high-weight links. After removing low-weight links, the largest connected component (LCC) is smaller compared to Top and the number of connected components is higher.

deviation is $\sigma_{w_{ij}} = 714.68 \text{ m}^2$. We define high-weight links in BNs the links having $w_{ij} > \mu_{w_{ij}} + 2\sigma_{w_{ij}} \sim 2000 \text{ m}^2$.

5.5.1 Protein structures

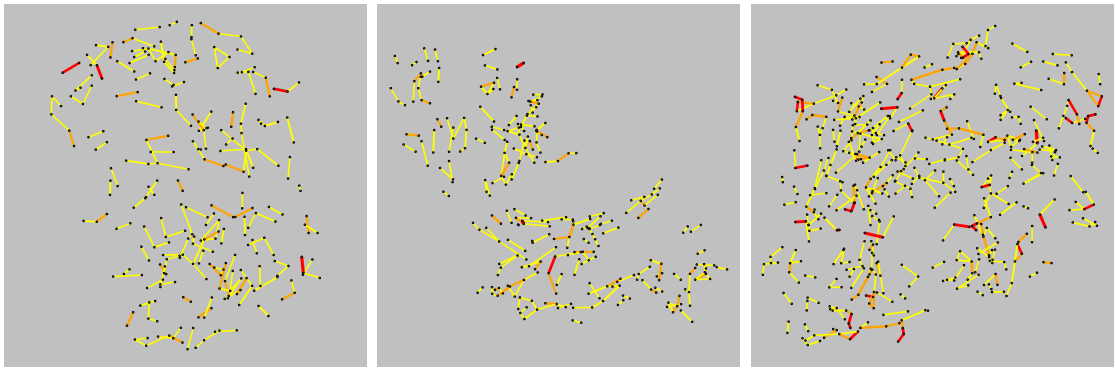


Figure 5.5: Sub-networks of high-weight links in the Amino Acid Network of three protein structures. From left to right: 1B44, 1CUL and 1B9L. The color code is the same as in Fig. 4.6.

Table 5.1: Connected components in the sub-networks of high-weight links in the AAN of three protein structures and the proteins database. For the protein database, the results are reported as average \pm standard deviation.

Protein	N	m	m_{high}	$\% m_{\text{high}}$	N. CCs	LCC	LCC/ N
1B44	531	2974	180	6.0 %	109	6	1.1 %
1CUL	707	3836	202	5.2 %	106	12	1.7 %
1B9L	952	5305	385	7.3 %	191	8	0.1 %
Database	922 ± 906	5150 ± 5266	$6.0 \% \pm 1.3 \%$	288 ± 256	156 ± 142	10 ± 4	$1.8 \% \pm 1.3 \%$

Fig. 5.5 shows the sub-networks of high-weight links of the AANs of the same protein structures studied in Chapter 4, Section 4.2.4 (PDBs 1B44, 1CUL and 1B9L). For the three protein case studies and for the proteins in the proteins database (Chapter 4, Section 4.3.1), Table 5.1 reports the number of nodes (N), number of links (m), number and percentage of high-weight links (m_{high} and $\% m_{\text{high}}$) in the AANs and the number of connected components (N. CCs), size of the LCC, and percentage of nodes belonging to the LCC (LCC/ N) in the sub-networks of high-weight links of the AANs.

The results show that links with maximal values of weight (red in Fig. 5.5) are scattered in the protein structures and that the sub-networks of high-weight links of the AANs of protein structures have small connected components compared to the size of the network (LCC/ $N \sim 2 \%$). Thus, protein structures are unjammed spatial systems.

Protein structures being unjammed is consistent with the observation that proteins are dynamical objects, and that their biological function relies on atomic motions (Chapter 2, Section 2.2). Moreover, if amino acids are unjammed in the protein structure, then amino-acid mutations may be accommodated through multi-scale structural rearrangement, if the local structure around the amino acid to be mutated is too packed: proteins are spatially sustainable at the multi-scale level.

In the next section, the same measures are performed on the BNs of urban structures to investigate whether urban structures are also spatially-sustainable at the multi-scale level or not.

5.5.2 Urban structures

Fig. 5.6 shows the sub-networks of high-weight links of the BNs of the same urban structures studied in Section 4.2.3 (Monplaisir, Charpennes and Manhattan) and Fig. 5.7 shows the sub-network of high-weight links of the BN of the city of Lyon. Table 5.2 reports the number of nodes (N), number of links (m), number and percentage of high-weight links (m_{high} and $\%m_{\text{high}}$) in the BNs and the number of connected components (N. CCs), size of the LCC, and percentage of nodes belonging to the LCC (LCC/ N) in the sub-networks of high-weight links of the BNs.

Compared to the protein structures, the urban case studies have a higher diversity in the percentage of nodes that belong to the LCC of the sub-network of high-weight links



Figure 5.6: Sub-networks of high-weight links in the Building Network of three urban structures. From left to right: Monplaisir in Lyon (France), Charpennes in Villeurbanne (metropolitan area of Lyon, France), and Manhattan in New York City (USA). The color code is the same as in Fig. 4.5.

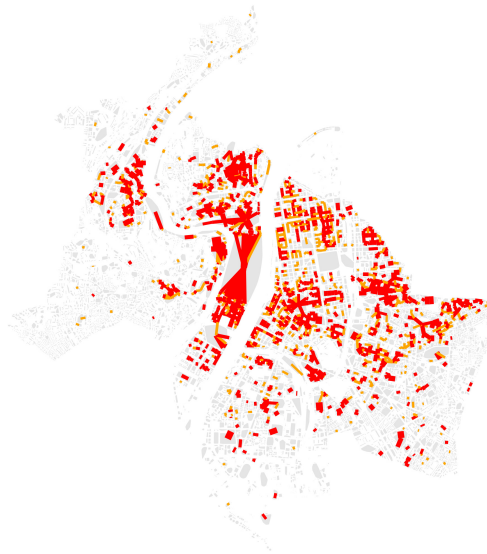


Figure 5.7: Sub-networks of high-weight links in the Building Network of the city of Lyon (France). The color code is the same as in Fig. 4.5.

Table 5.2: Connected components in the sub-networks of high-weight links in the BN of urban structures.

Protein	N	m	m_{high}	$\%m_{\text{high}}$	N. CCs	LCC	LCC/ N
Monplaisir	1298	3553	173	4.9 %	1133	35	2.7 %
Charpennes	776	1798	233	13.0 %	576	96	12.4 %
Manhattan	267	400	125	31.3 %	145	41	15.3 %
Lyon	12344	30097	1425	4.7 %	11064	124	1.0 %

(LCC/ N). Monplaisir is the least-jammed area and has values of LCC/ N and $\%m_{\text{high}}$ similar to the proteins'. On the contrary, Charpennes and Manhattan have higher values of LCC/ N and $\%m_{\text{high}}$, with Manhattan being the most jammed area. Indeed, “force chains” of high-weight links are observed in the BN of Manhattan, and in a part of the BN of Charpennes, while high-weight links are scattered in the BN of Monplaisir (Fig. 5.6). Interestingly, Monplaisir is the most heterogeneous area in terms of buildings geometries

and Manhattan is the most homogeneous (Fig. 5.6), consistent with the analogy with granular materials, where higher heterogeneity is associated with higher mobility (less jamming).

In the city of Lyon, the percentage of nodes belonging to the LCC is low ($LCC/N = 1\%$), leading to the classification of the city as unjammed. However, attention should be paid to the fact that the distribution of the high-weight links is not uniform in the BN of Lyon: high-weight links are concentrated and adjacent in the city center, while are rare in the suburbs (Fig. 5.7). Consistently, even though LCC/N is low, 124 buildings belong to the LCC of the sub-network of high-weight links of Lyon (Table 5.2). It can be concluded that the city of Lyon is partially jammed, with the city center jammed and the suburbs unjammed.

The results show that the locally spatially unsustainable urban areas (Manhattan, part of Charpennes and city center of Lyon, Chapter 4) are also spatially unsustainable at the multi-scale level. Following the parallel with amino-acid mutations, jammed urban areas cannot sustain the substitution of buildings, because the local neighborhoods are highly packed (high Nw) and the multi-scale packing is also high (“force chains” are present).

Obviously, even in unjammed urban systems (e.g. Monplaisir), multi-scale structural rearrangement is practically infeasible, because it requires destroying and reconstructing several buildings. However, high multi-scale packing, corresponding to jammed systems, also means that little empty space is available over large distances. Urban mobility exploits this empty space, thus high multi-scale packing (i.e. the presence of adjacent high-weight links in the BN) suggests hindered mobility and a higher potential to accumulate traffic. The next session investigates the relation between mobility in the urban system and link weights in the corresponding BN model.

5.5.3 Agent-based modeling of random mobility in urban systems

To investigate the relation between mobility and the multi-scale packing in the urban system, measured by the link-weights in the BN, a simulation of the random mobility in the three urban system case studies is performed. The details of the simulation algorithm are provided in the Methods (Chapter 3, Section 3.3.2). In brief, the urban system is subdivided into small cells, agents are randomly placed in the cells and allowed to move in the urban space following a random direction. When an agent encounters an obstacle (i.e. a building), it chooses a new random direction to follow. If no obstacles are present, the agent will move along the chosen direction; otherwise, it will remain still and will choose a new random direction to be followed at the next time step. Figure 5.8 shows a schematic of the simulation algorithm.

At each time step, the cell occupied by each agent is memorized. At the end of the simulation, the following quantities are reported for each cell:

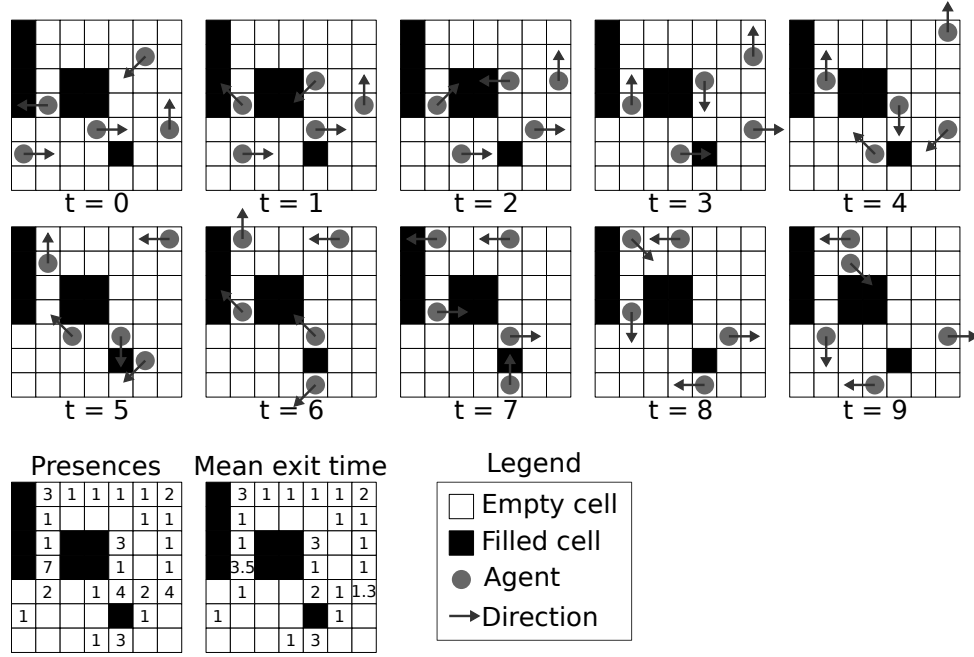


Figure 5.8: Schematics of the random mobility simulation algorithm (see main text).

- Presences: the cumulative number of agents that have been in the cell (one presence per agent per time step);
- Mean exit time: average number of time steps required by an agent to leave the cell after entering it. High mean exit time means that the mobility is hindered at the cell.

The random mobility simulated here is obviously not representative of the real traffic in the city, that depends on the origin and destination of the trips, the width of the streets, the traffic management, etc. However, modeling random mobility allows focusing on only one aspect of the urban mobility, that is the influence of the surface occupied by buildings or available between buildings, as measured by the BN.

The results of the simulations of the random mobility in the three case studies of urban structures are shown in Fig. 5.9.

The Presences measure (top) assures that all space has been explored. More importantly, the mean exit time measure (bottom) is not uniform in the urban spaces: higher mean exit time (dark red cells in Fig. 5.9, bottom) is measured where high-weight links are present (Fig. 5.6). Thus, the dynamics of the random mobility in the urban system, favoured by low packing and hindered by high packing at multiple scales, may be inferred from the BN model. More in detail, high-weight links in the BN correspond to locations where mobility is hindered, because the buildings packing is high. Adjacent high-weight link, i.e. “force chains” of the urban system, cause hindered mobility at multiple scales. A perturbation in these “jammed” areas, for example the closure of a street, will be more hardly sustained compared to a perturbation in “unjammed” areas, because buildings are tightly packed at multiple scales, i.e. they leave few empty space available

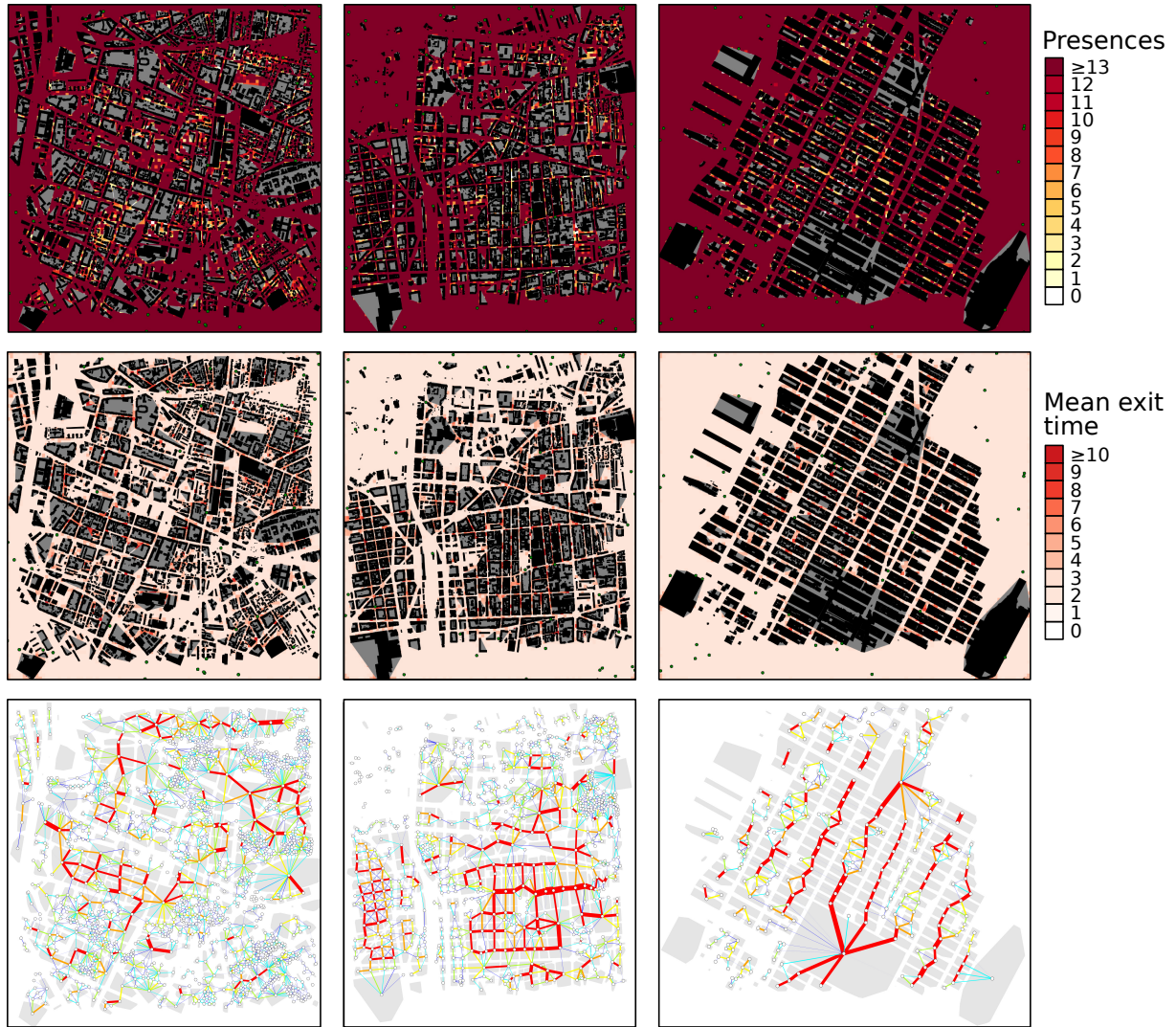


Figure 5.9: Simulation of random mobility in three urban structures. From left to right: Monplaisir in Lyon (France), Charpenne in Villeurbanne (metropolitan area of Lyon, France), and Manhattan in New York City (USA). Top: Cells are colored based on the presences count. Middle: Cells are colored based on the mean exit time. Bottom: The BN of the three areas as in Fig. 5.6, for comparison.

to accommodate mobility.

5.6 Conclusion

Using an analogy with granular materials, multi-scale spatial sustainability in proteins and urban structures has been assessed from the classification of the systems as “jammed” or “unjammed”. “Jammed” urban areas have been identified as areas where adjacent high-weight links are present. In these areas, mobility is hindered by the high packing of buildings at multiple scales. Contrary to urban systems, protein structures are always “unjammed”: the high-weight links are scattered in the protein structures and do not cluster into “force chains” of adjacent high-weight links. The absence of “force chains” in protein structures means that empty space is available for atomic mobility at all scales.

This property makes protein spatially sustainable at the multi-scale level: amino acid mutations can be accommodated even when the local packing is high (high Nw) by a structural rearrangement of the system at multiple scales.

The next chapter investigates whether such structural rearrangement is indeed observed upon amino-acid mutations in protein structures and if it is, what scales it involves.

Chapter 6

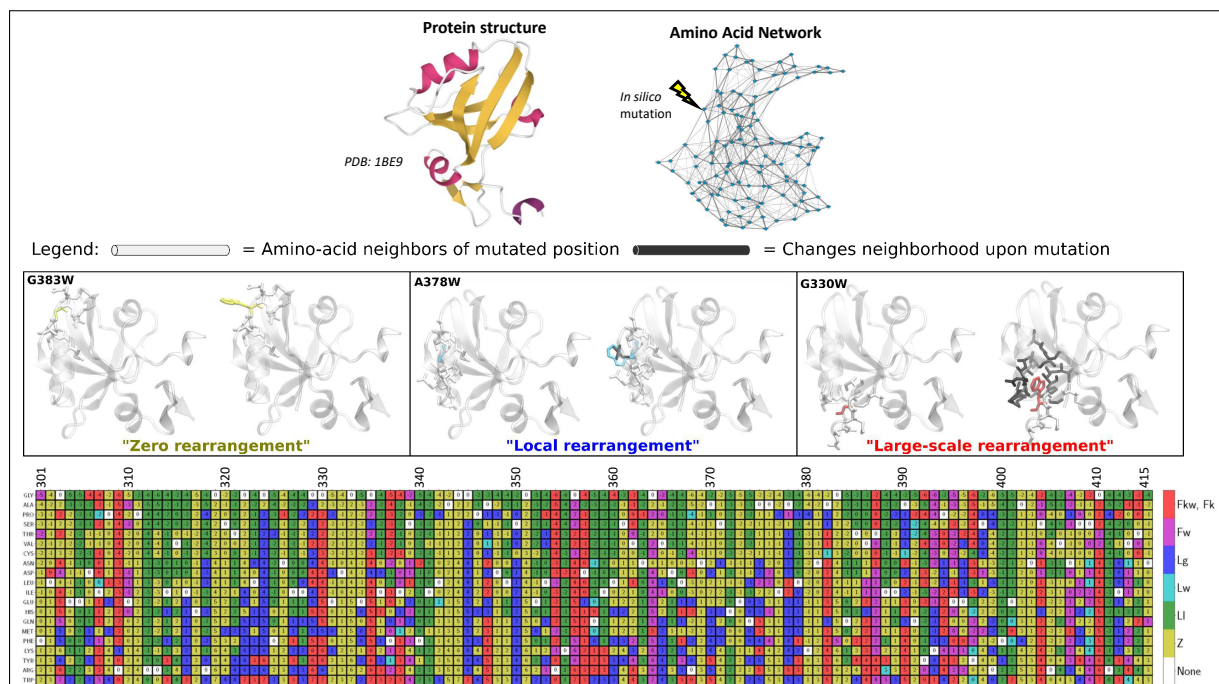
Case study: Third PDZ domain. Means to accommodate substitutions in the protein structure.

Highlights:

- *In-silico amino-acid mutations simulate atomic rearrangement of protein structures.*
- *In-silico amino-acid mutations lead to rearrangement at different scales.*
- *The scale of rearrangement depends on the neighborhood of the mutated amino acid.*
- *The scale of rearrangement also depends on the substituting amino acid.*

Abstract: Rearrangement of atomic interactions required in protein structures to accommodate substitutions is simulated using *in-silico* amino-acid mutations. Rearrangement is measured from changes in neighborhoods in the Amino Acid Network. Amino-acid mutations are accommodated with no need of rearrangement, with local-scale rearrangement or with large-scale rearrangement (above chemical neighbors) depending on the specific amino-acid mutation. This shows that the rearrangement depends on the neighborhood of the mutated amino-acid.

Graphical abstract:



Methods: Classification of *in-silico* amino acid mutations (Section 3.2.2).

6.1 Introduction

In Chapter 5 it was proposed that amino acid mutations can be accommodated in the protein structure even when the local packing is high, thanks to the potential for neighborhoods rearrangement allowed by the absence of adjacent high-weight links in the Amino Acid Network (AAN). Indeed, absence of adjacent high-weight links means that space is not congested at any scale.

In this chapter, the impact of amino acid mutations on a protein structure is studied to validate the existence of multiple scales of accommodation of mutations. To this goal, all amino acids of the third PDZ domain of the synaptic protein PDS-95 (PSD95^{pdz3}, PDB 1BE9) are mutated in silico by all nineteen other amino acid types.

In-silico acid-mutations are performed using FoldX [161], that produces an energetically-viable mutant structure starting from the WT structure. Even though in-silico mutants do not represent structures of experimentally-generated mutants, in-silico mutants produce a set of structures where the perturbation caused by a single mutation is not compensated by other mutations or by a global structural rearrangement, i.e. structural robustness is imposed. It follows that the comparison between the in-silico mutants and the WT structure directly pinpoints the network rearrangements that are necessary for structural robustness upon mutation. As a consequence, the in-silico approach allows distinguishing amino-acid neighborhoods that are suitable for any amino-acid type from the ones that need to be adjusted to the specific amino acid, under the constraint of structural robustness.

The mutants' structures are compared to the wild-type (WT) structure through the comparison of the corresponding Amino Acid Networks (AANs). Then, the mutations are classified based on the spatial scale of the rearrangement they provoke. Rearrangement is measured from changes in neighborhoods caused by the in-silico mutations, where a change in neighborhood means that an amino acid has gained or lost one or more links in the AAN. Changes in link weights without the addition or loss of links are not considered as neighborhood changes. This is an approximation, but also allows distinguishing degree-mechanisms from weight-mechanisms of perturbations (see next section).

The next section summarizes the classification procedure, while the computational details are reported in Chapter 3, Section 3.2.2.

6.2 Classification of in-silico amino-acid mutations based on the scale of neighborhoods rearrangement they provoke

Nineteen mutant structures are produced in silico for each amino acid of PSD95^{pdz3} (positions 301 to 415, resulting in a total of 2185 mutant structures). Then, the Amino Acid Network (AAN) of each mutant is built and compared with the WT AAN. The WT

AAN is built from the output of the RepairPDB FoldX command run on the 1BE9 PDB structure of PSD95^{pdz3}, after removal of the protein ligand. This choice is made because the mutations are performed on this “repaired” structure (Chapter 3, Section 3.2.1).

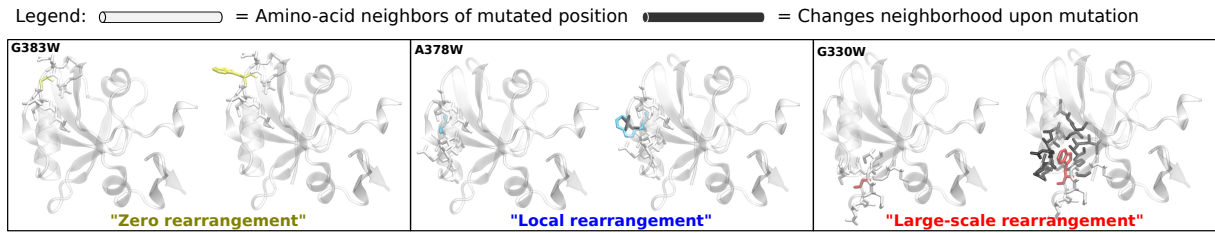


Figure 6.1: Examples of amino-acid mutations leading to no neighborhoods rearrangement (Z mutation), local rearrangement (L mutation) and large-scale rearrangement (F mutation). In each example, the WT structure is on the left and the mutant structure obtained in-silico is on the right. The mutated position is shown in color, the first-shell neighbors (distance $\leq 5\text{\AA}$) of the mutated position are shown in white and the amino acids that change neighborhood upon mutation (perturbed amino acids) are shown in black.

For each mutant structure resulting from the mutation of an amino acid at position i , the set of perturbed amino acid positions is defined as the set of amino acids whose nodes in the AAN have a different neighborhood compared to the WT AAN. The neighborhood of an amino acid is considered to be different if the amino acid has gained or lost one or more links (i.e. it has gained or lost neighbors). If an amino acid changes link weights with its neighbors but it does not change its set of neighbors, we consider its neighborhood to be unchanged, i.e. the amino acid is not perturbed by the mutation.

Based on the set of perturbed amino acids, the mutations $i \rightarrow i'$ are classified as follows (Fig. 6.1):

- Zero (Z) rearrangement if no amino acid changes neighborhood (Fig. 6.1, left panel);
- Local (L) rearrangement if only first-shell neighbors of i change neighborhood (Fig. 6.1, middle panel). These mutations are further classified as:
 - Ll if i loses first neighbors upon mutation;
 - Lg if i gains first neighbors upon mutation (or loses some neighbors and gains others);
 - Lw if i does not change neighbors but the connectivity among its first neighbors changes. This implies atomic rearrangement but without the gain or loss of amino-acid contacts at the i position.
- Far (F) rearrangement if nodes that are further from the first-shell neighborhood of i are perturbed (Fig. 6.1, right panel). These mutations are further classified as:
 - Fw if node i does not change neighborhood (the perturbation spreads through changes in link weights).
 - Fk if node i changes neighborhood and a path of neighborhood changes connects i with all perturbed nodes (the perturbation spreads through changes in neighborhoods);

- Fkw if node i changes neighborhood but not all perturbed nodes are connected by a path of perturbations to node i (the perturbation spreads through changes in neighborhoods and changes in weight).

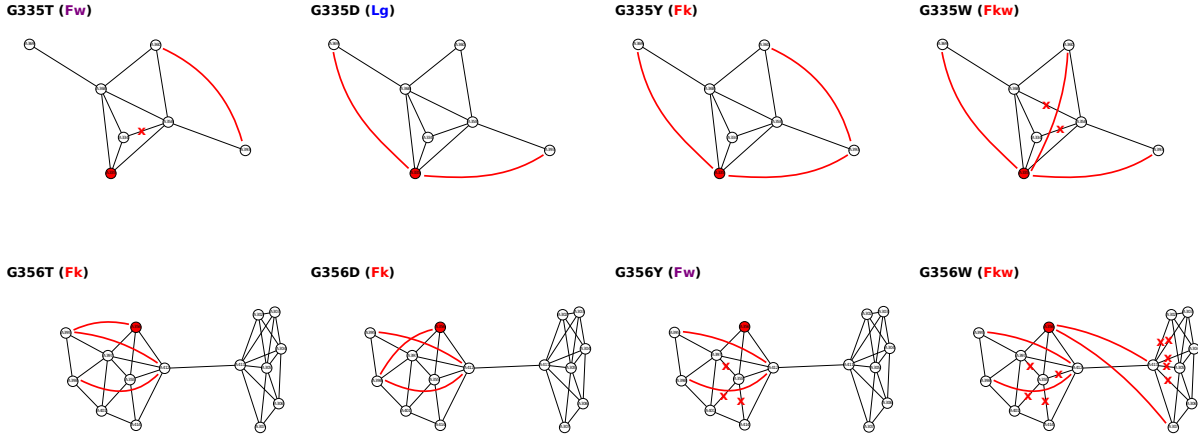


Figure 6.2: Examples of mutations leading to different neighborhoods rearrangement. Top: mutations of GLY335. Bottom: mutations of GLY356. The red node corresponds to the mutated amino acid. The other nodes correspond to perturbed nodes. For a given position, the largest set of perturbed nodes is shown, this corresponds to the mutation to W in both cases. Black links are present in the WT structure. Red links are added upon mutation and red crosses correspond to links lost upon mutation. From left to right, the mutations are in increasing order of side-chain length of the substituting amino acid.

To provide a better insight on the neighborhoods rearrangements associated to the local (L) and far (F) classes, Fig. 6.2 shows some examples of neighborhoods perturbations caused by the mutation of two GLY amino acids of PSD95^{pdz3}, in positions 335 (top) and 356 (bottom). Mutations of GLY amino acids are the easiest to analyze because any mutation of GLY amino acids adds atoms to the mutated position. The G335D mutation is an example of local perturbation: the mutated amino acid gains contacts with two other amino acids, while the contacts of all other amino acids are unchanged (Fig. 6.2, top, second panel). This means that the additional atoms of the ASP amino acid make interactions with the atoms of the two added neighbors, but no other neighborhoods rearrangement happens (Lg mutation). On the contrary, all other mutations considered in the example lead to a large-scale neighborhoods rearrangement (F mutations), meaning that amino acids further than the first neighbors of the mutated amino acids are perturbed.

The G335Y mutation causes the gain of the same interactions as the G335D mutation, but the perturbation is propagated further in the structure, as one of the added neighbors also gains interaction with a further amino acid (Fig. 6.2, top, third panel). This means that the position of these amino acids has changed, and the perturbation has propagated from the mutated position through changes in amino-acid contacts (Fk mutation). A similar mechanism is caused by the G356T and G356D mutations, where amino acids both in the neighborhood of the mutated position and further in the structure change amino-acid interactions, and a chain of changes in amino-acid interactions is visible from

the mutated position to the furthest perturbed amino acid (Fig. 6.2, bottom, first and second panel). We call this perturbation mechanism a degree-far-perturbation mechanism (Fk), because the mutated amino acid changes neighbors upon mutation.

A different mechanism leads to the perturbation of far amino acids upon the mutations G335T (Fig. 6.2, top, first panel) and G356Y (Fig. 6.2, bottom, third panel): as in the previous examples, the mutation provokes a change in contacts far from the mutated position (F mutation), but this time the changes are propagated without modification of the neighborhood of the mutated position. This means that the perturbation propagates in the protein structure through changes in atomic contacts between amino acids that were already interacting with the mutated position in the WT structure, i.e. the mutated position changes link weights with its neighbors but does not change neighborhood. Accordingly, we call such mechanism a weight-far-perturbation mechanism (Fw).

Finally, the degree- and weight-far-perturbation mechanisms may take place at the same time, as for the G335W and G356W mutations (Fig. 6.2). In this case, the mutation is said to follow a degree-and-weight-far-perturbation mechanism (Fkw). For simplicity, in the following, Fkw mutations are assimilated to the Fk class.

6.3 In-silico mutagenesis of PSD95^{pdz3}

Classes of mutations Fig. 6.3 shows the classification of all the in-silico mutations of PSD95^{pdz3} based on the neighborhoods rearrangement they provoke. Each column in Fig. 6.3 corresponds to an amino acid position and each row corresponds to an amino acid type. White cells correspond to the WT amino-acid type (row) at a given position (column). The number in each cell gives the difference in side-chain length between the mutated amino-acid type and the WT amino-acid type, measured in Å and rounded to the nearest integer. The values of side-chain lengths for all amino-acid types are reported in Table 6.1. It should be noted that the side-chains of the amino acids are dynamic and can fold in different conformations. We use the side-chain lengths corresponding to the fully-extended conformation of the amino acids, so that the differences in side-chain lengths in Fig. 6.3 correspond to the largest possible differences.

Table 6.1: Side-chain length of all amino-acid types. These values correspond to the fully extended conformation of the amino acid. Hydrogen atoms are not considered.

Amino acid	GLY	ALA	PRO	SER	THR	VAL	CYS	ASN	ASP	LEU	ILE	GLU	HIS	GLN	MET	PHE	LYS	TYR	ARG	TRP
Length [Å]	0.0	1.5	2.4	2.4	2.5	2.5	2.8	3.5	3.7	3.7	3.9	4.6	4.6	4.6	4.7	5.0	5.9	6.4	6.5	6.6

The results of Fig. 6.3 show that no rearrangement (Z), local (L) or large-scale (F) rearrangement take place upon amino-acid mutations. Moreover, both degree- (Ll, Lg, Fk) and weight-perturbation mechanisms (Lw, Fw) occur.

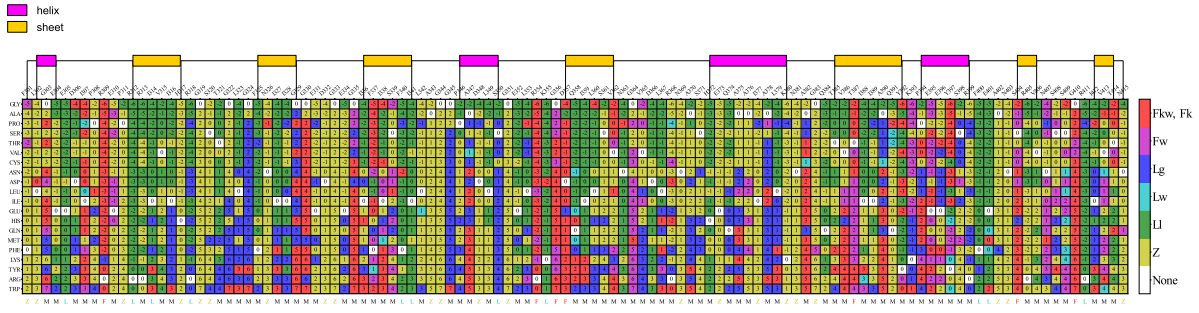


Figure 6.3: Classification of all the in-silico mutations of PSD95^{pdz3} based on the neighborhoods rearrangement they provoke. Each column corresponds to an amino acid position and each row corresponds to an amino acid type. The number in each cell gives the difference in side-chain length between the mutated amino-acid type and the WT amino-acid type, measured in Å and rounded to the nearest integer. The graphic on the top reports the secondary structure. The bottom line reports the classification of the amino-acid position as Z-, L-, F-, or M-class (see main text).

Classes of amino acids The mutation of some amino-acid positions provokes the same class of perturbation consistently, regardless the amino-acid type with which they are mutated: for example, ASP332, GLY333, ALA343, GLY344, ILE373, GLY383, ALA402 and ASN403 never provoke neighborhoods rearrangement when mutated (Z), while ARG309, GLY330, GLY356, ASP357 and GLY410 always or almost always provokes a Far-scale (F) perturbation. Notably, the scale of the neighborhoods rearrangement is not determined by the initial (WT) amino-acid type. For example, despite the fact that GLY does not have any side-chain, and thus its mutation to large amino acids such as TYR, ARG and TRP provoke the maximal changes in side-chain length, the neighborhoods rearrangement the mutation of GLY provokes ranges from zero (e.g. GLY333, GLY344 and GLY383) to large-scale (e.g. GLY330, GLY356 and GLY410) (Fig. 6.3). Importantly, the scale of the neighborhoods rearrangement is not always determined by the difference in side-chain length between the substituting amino acid and the WT amino acid neither across mutations at the same position. As an example, all mutations of the I241 amino acid lead to no neighborhoods rearrangement (Z) or loss of first neighbors (Ll), regardless on whether the substituting amino acid has a shorter or longer side-chain compared to ILE (Fig. 6.3).

The amino-acid positions that provoke a consistent response when mutated are classified as Z-class, L-class and F-class depending on the class of the perturbation they provoke. Precisely, amino acids are classified as Z-, L- or F-class if at least 17 of their mutations belong to the Z-, L- or F-class, respectively (Fig. 6.3, bottom line). For such classification, the distinction between degree- and weight-mechanisms is ignored as a first approximation.

While some amino-acid positions can be classified as Z, L or F, the mutation of other amino-acid positions provokes a rearrangement at different scales depending on the substituting amino acid. For example, the mutations of GLY303 provoke zero, local or far-scale rearrangement depending on the substituting amino-acid type. Importantly, also in this

case the scale of neighborhoods rearrangement is not determined by the side-chain length of the substituting amino acid: despite the fact that ASN, ASP, LEU and ILE have similar side-chain lengths, GLY303LEU is Z while GLY303ASN, GLY303ASP and GLY303ILE are F, and GLY303LYS is Z even though LYS has a longer side-chain compared to ASN, ASP, LEU and ILE (Fig. 6.3). A similar example is the one of VAL362, that provokes a F rearrangement when mutated to LEU (side-chain length difference = 1) but no rearrangement (Z class) when mutated to MET (side-chain length difference = 2). The amino-acid positions that do not provoke a consistent perturbation response when mutated are classified as M-class, where M stands for Mixed (Fig. 6.3, bottom line).

Among the 115 amino acids of PSD95^{pdz3}, 22 belong to the Z-class, 11 belong to the L-class, 7 belong to the F-class and 75 belong to the M-class. This shows that the neighborhoods of amino acids are highly diverse and tolerate mutations differently.

Different classes (Z, L, F or M) signify different spatial arrangements of neighbors. Amino acids that belong to different secondary structure elements are subject to different spatial constraints. Similarly, amino-acids that are on the surface of the protein structure are subject to different constraints compared to amino acids buried in the structure. We have investigated whether the class of amino acids (Z, L, F or M) is determined by the secondary structure or by the structural position of the amino-acid.

Secondary structure We have investigated whether the Z, L, F or M class is determined by the secondary structure to which the amino acid belongs. The secondary structure elements along the PSD95^{pdz3} are reported in Fig. 6.3 (top), where α -helix segments are depicted in pink and β -strand segments are depicted in yellow. We have classified amino acid positions based on the secondary structure to which they belong (α -helix, β -strand or coil) and verified whether the Z, L, F and M classes are populated differently depending on the secondary structure. Please note that amino acids that not belong to α -helices or β -strands are generally associated to coils, i.e. coils, loops and turns are not distinguished.

Table 6.2 reports the number and percentage of amino acids belong to each class (Z, L, F or M) for the three types of secondary structures.

Table 6.2: Number of amino acids in each class of structural perturbations (Z, L, F or M) for each type of secondary-structure element. In brackets, the percentage of amino acids belonging to a class among the amino acids of the same secondary-structure type, rounded to the nearest integer.

Secondary structure	Z-class	L-class	F-class	M-class	<i>total</i>
α -helix	5 (23%)	1 (5%)	0 (0%)	16 (73%)	22
β -strand	2 (6%)	2 (6%)	3 (10%)	24 (77%)	31
coil	15 (24%)	8 (13%)	4 (6%)	35 (56%)	62
<i>total</i>	22 (19%)	11 (10%)	7 (6%)	75 (65%)	115

None of the amino acids belonging to α -helices belong to the F-class, and on the

contrary the F-class is slightly over-populated by amino acids belonging to β -strands while the Z-class is under-populated with respect to the global statistics. The M-class is the most largely populated for all types of secondary structures, consistent with the fact that most amino acids belong to the M-class. Amino acids belonging to coils follow very similar statistics to the global statistics, apart from a slightly over-population of the Z-class to the detriment of the M-class (Table 6.2).

The results of Table 6.2 suggest that mutations in β -strands generally provoke neighborhoods rearrangement further in the protein structure compared to mutations in α -helices and coils. Nevertheless, the limited number of cases per class does not allow any firm conclusion and no obvious trend that relates the secondary structure to the amino-acid class is observed. Moreover, the statistical analysis should be extended to a larger dataset of mutations, including larger protein structures, to establish whether the observed tendency of longer-scale perturbations for mutations in β -strands is universal.

Buried versus surface-exposed amino acids Next, we have investigated whether the class of amino acid positions, Z, L, F or M is determined by the structural position of the amino acid, buried in the core or surface-exposed. To do so, we have calculated the relative Accessible Surface Area (rASA) of each amino acid and classified amino acids as surface-exposed if $\text{rASA} > 0.2$ and buried otherwise. Over the 115 amino acids of PSD95^{pdz3}, 64 are surface-exposed and 51 are buried. Fig. 6.4 shows the same results as Fig. 6.3 but selecting only surface-exposed amino acids. It shows 21 over the 22 amino acids belonging to the Z-class are surface exposed. However, not all surface-exposed amino acids belong to the Z-class: ARG309, ARG354 and THR387 belong to the F-class (their mutation provokes a large-scale neighborhoods rearrangement), 7 surface-exposed amino acids belong to the L-class (their mutation provokes a local-scale neighborhoods rearrangement) and 33 amino acids belong to the M-class (different scales are perturbed depending on the specific mutation).

Fig. 6.5 shows the same results as Fig. 6.3 but selecting only buried amino acids.

Compared to Fig. 6.4, more F mutations are present. However, not all buried amino acids provoke large-scale perturbations then mutated. ASP348 belongs to the Z-class, 4 amino acids belong to the L-class, 4 amino acids belong to the F-class and the remaining 42 buried amino acids belong to the M-class.

The results of Figs 6.4 and 6.5 show that surface-exposed amino acids are more likely to provoke no rearrangement when mutated (Z mutations), while buried amino acids are more likely to provoke large-scale rearrangement when mutated (F mutations). Nevertheless, all classes of mutations are observed for both surface-exposed and buried amino acids, meaning that the structural position of the amino acid does not strictly determine the scale of the rearrangement provoked by the mutations.

In the next paragraph, we verify whether the class of the amino acid (Z, L, F or M)

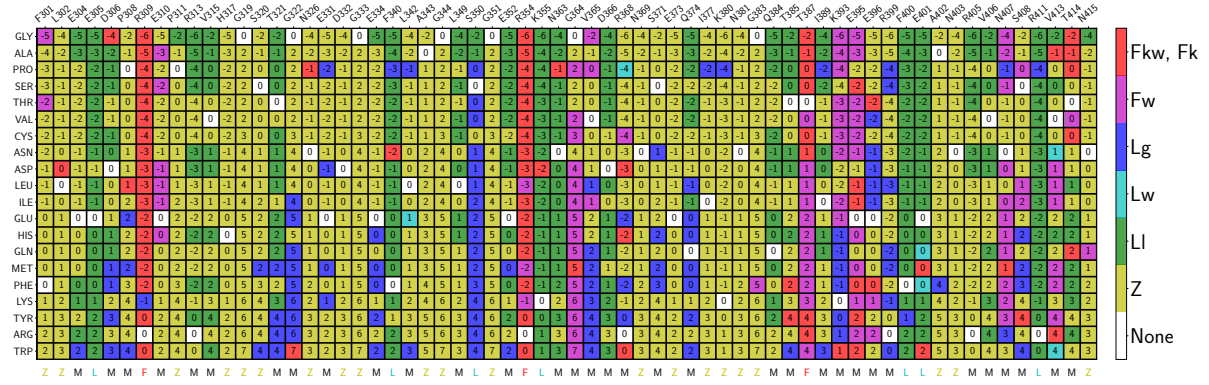


Figure 6.4: Classification of the in-silico mutations of the surface-exposed amino acids of PSD95^{pdz3} based on the neighborhoods rearrangement they provoke. The number in each cell gives the difference in side-chain length between the mutated amino-acid type and the WT amino-acid type, measured in Å and rounded to the nearest integer. The bottom line reports the classification of the amino-acid position as Z-, L-, F-, or M-class.

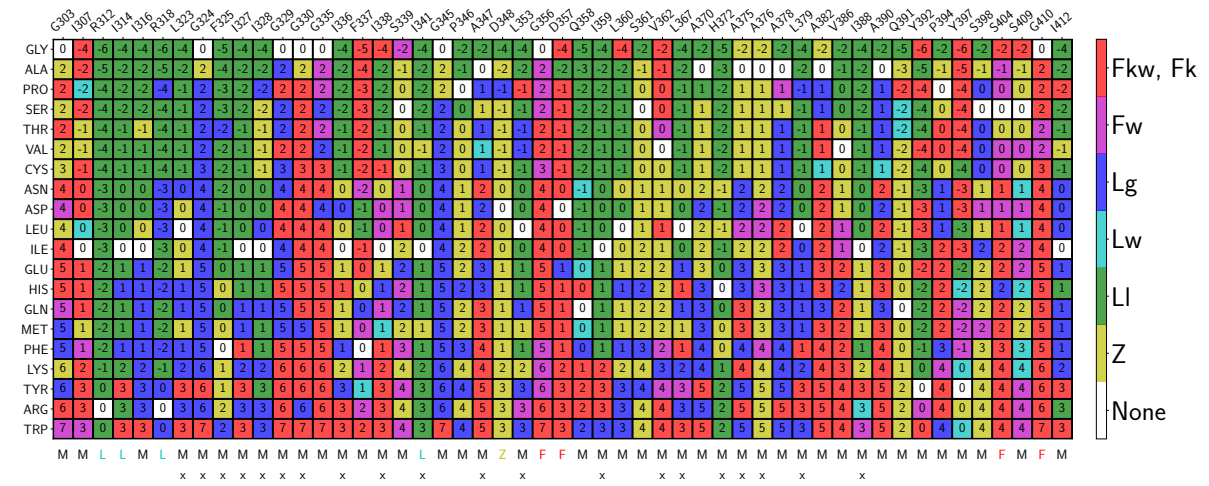


Figure 6.5: Classification of the in-silico mutations of the buried amino acids of PSD95^{pdz3} based on the neighborhoods rearrangement they provoke. The number in each cell gives the difference in side-chain length between the mutated amino-acid type and the WT amino-acid type, measured in Å and rounded to the nearest integer. The bottom line reports the classification of the amino-acid position as Z-, L-, F-, or M-class. Crosses indicate functionally-sensitive positions.

is related to the functional impact of its mutations, based on experimental evidence.

Functionally-sensitive positions In 2012, McLaughlin and collaborators have performed a comprehensive single mutagenesis analysis of PSD95^{pdz3}, where the functional impact of all the possible single-amino acid mutants was assessed [16]. They discovered that the mutation of 20 positions over the 83 analyzed sequence positions of PSD95^{pdz3} (amino acids 311 to 393) caused a loss of binding activity to the ligand of PSD95^{pdz3} compared to the WT protein. We refer to these position as functionally-sensitive positions. All functionally-sensitive positions are buried and they are indicated with a cross in Fig. 6.5. All functionally-sensitive positions belong to the M-class, apart from ILE341 that belongs to the L-class. This result shows that the functional impact of mutations is

not determined by the scale of the perturbation they provoke: for example, the positions GLY356 and ASP357 belong to the F-class, but are not functionally sensitive. Two conclusions can be drawn: first, the functional impact of mutation is not determined by how many amino acids are perturbed and how far these amino acids are from the mutated position. This suggests that the functional impact of the mutation depends on the specific atomic interactions that are gained or lost. This aspect is developed more in details in Chapter 11. Second, neighborhoods rearrangement does not necessarily inhibit the protein function, i.e. the protein is functionally-sustainable even upon large-scale neighborhoods perturbations. Nevertheless, these conclusions are based on in-silico mutants structures, so they should be validated based on experimental evidence.

6.4 Conclusion

As anticipated in Chapters 4 and 5, the in-silico mutagenesis of PSD95^{pdz3} shows that neighborhoods rearrangement is a viable mechanism of accommodation of mutations in protein structures and it involves local to global neighborhoods changes regardless of the secondary structure and the rASA of the mutated position. This supports the idea of using space-occupancy metrics to investigate the perturbation response of proteins.

Amino acids of the same type belong to different classes (Z, L, F and M), corresponding to different scales of rearrangement provoked by their mutation, meaning that the response to mutations is encoded by the neighborhood of the mutated amino acid.

If an amino acid position i belongs to the Z-class, it follows that the amino-acid neighborhood of i is suitable for all amino-acid types. The Z-class is populated mostly by surface-exposed amino acids, but it is accessible also to buried amino acids (D348). The neighborhoods of amino acid positions belonging to the Z-class encode for structural robustness, as mutations at those positions do not provoke any changes in neighborhoods.

If an amino acid position i belongs to the L-class, it means that local amino-acid neighborhood of i can be adjusted to accommodate different amino-acid types, by either adding neighbors (Lg), removing neighbors (Ll) or changing the connectivity among neighbors (Lw). The neighborhoods of L-class amino-acid positions may also be seen as encoding for structural robustness, as only a local modification of the neighborhood is needed to accommodate any amino-acid type.

If an amino acid position i belongs to the F-class, it means that the multi-scale neighborhood of i is to be adjusted to accommodate different amino-acid types under the constraint of structural robustness, multi-scale meaning that amino acids further from the first-shell chemical neighbors of i are involved in the rearrangement. This implies that not only the local neighborhood of the amino acid i encodes for the response to mutations of i , but also its larger-scale neighborhood. Moreover, it should be noted that

the existence of the M-class, that is in fact the most populated, implies that in general multi-scale amino acid neighborhoods do not encode for one response to perturbations, but for several responses.

The existence of F mutations shows that neighborhood changes may be provoked by mutations far from the perturbed positions. This is important for protein sustainability because it proves that the neighborhoods perturbations caused by a mutation can be corrected by rescue mutations at amino acids far in the protein structure and not only in the local neighborhood of the perturbed amino acid.

Moreover, none of the F-class amino-acid positions of PSD95^{pdz3} are functionally-sensitive and many M-class amino-acid positions containing F-mutations are non-functionally-sensitive (Fig. 6.5). This shows that large-scale rearrangement does not imply functional failure. The potential for large-scale structural change in proteins with or without functional consequences is in accordance with B.H. Walker's definition of resilient systems as systems able to change and adapt as a response to a perturbation [7] (Chapter 2, Section 2.1.1).

Importantly, large-scale neighborhoods perturbations (F mutations) may travel from the mutated position to further amino-acids by addition or loss of local neighbors that provoke further changes in neighborhoods elsewhere (degree mechanism, Fk), but also by local changes in atomic interactions (link weights) that leave the local neighborhood unchanged, but still provoke changes in neighborhoods elsewhere (weight mechanism, Fw). This highlights the role of link weights in controlling the protein's potential for neighborhoods rearrangement, as was proposed in Chapter 5.

If link weights (atomic interactions) control the response of the protein structure to amino-acid mutations, i.e. the rearrangement of atomic interactions following a perturbation, then it is reasonable to assume that they also control the protein dynamics, that also consist in rearrangements of atomic interactions. The difference is that the changes in atomic interactions during dynamical processes (e.g. folding, unfolding and functional dynamics) correspond to actual rearrangements happening within a single protein structure, while the neighborhoods rearrangement studied in this chapter is observed from the comparison of two protein structures: the wild-type and the mutant. In the next chapter, the role of link weights in controlling the protein dynamics is explored. Then, Chapters 8, 9 and 10 investigate the dynamical impact of rearrangements of atomic interactions caused by amino-acid mutations.

Chapter 7

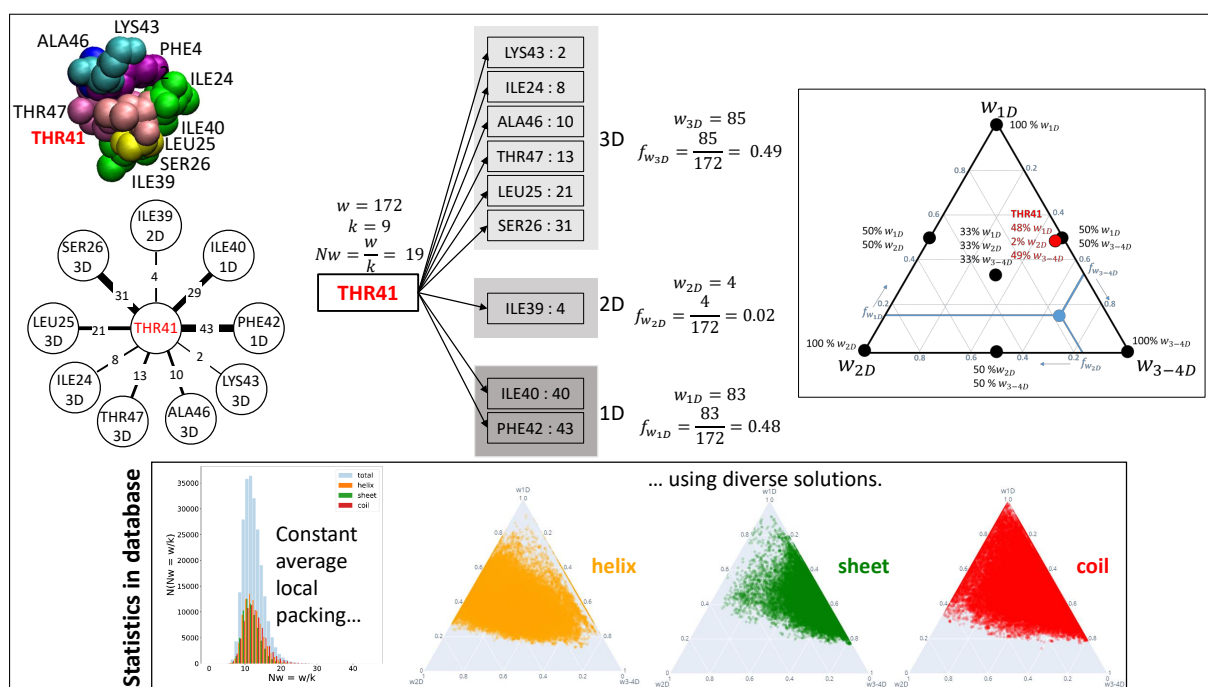
Database analysis: How to measure protein dynamics from a static structure?

Highlights:

- The space available for protein dynamics is assumed to be carved by the atomic packing (link weights).
- Atomic packing in primary, secondary or tertiary structural level is studied because their respective dynamics have different spatial and temporal scales.
- We show that diverse solutions in terms of allocation of atomic interactions (link weights) to primary, secondary and tertiary structural levels exist, under the only constraint of constant average atomic packing around amino acids.
- It is proposed that packing solutions encode for different dynamics.

Abstract: Protein dynamics consist in atomic motions, whose time and length scale depend on the protein structure level the atomic interactions belong to. We investigate whether the protein dynamics can be inferred from the allocation of atomic contacts to primary, secondary, and tertiary structural levels. The statistical analysis of a database of protein structures shows that similar Neighborhood watch (average local packing) around amino acids is obtained using diverse solutions in terms of atomic-contacts allocation to structural levels.

Graphical abstract:



Methods: Local structural-level allocation of the atomic interactions (Section 3.1.5).

7.1 Introduction

The previous chapters focused on the impact of amino acid mutations on the protein structure. Chapter 4 showed that atomic packing around amino acids in proteins is constant and moderate on average (Nw measure), and Chapter 5 showed that low-packing is observed in protein structures also at larger scales. Low packing means that empty space is always available around amino acids in protein structures and we suggested that low-packing provides the possibility to accommodate amino-acid mutations at any position of the protein structure. Chapter 6 showed that in-silico amino-acid mutations are accommodated in the protein structure thanks to the rearrangement of amino-acid interactions at different scales and it was proposed that different perturbation scales (Zero, Local or Far) and mechanisms (degree-mechanism and weight-mechanism) upon mutations should be encoded by different atomic packing (link-weights in the AAN) around the perturbed amino acids.

Structural rearrangement is made possible by low packing because atoms need empty space to move. In the same way, the empty space may be exploited for the atomic motions representing the protein folding, unfolding and functional dynamics. The goal of this chapter is the identification of a structural measure that relates the protein structure (atomic interactions) to the protein dynamics (atomic motions).

In Chapter 5, link weights in spatial networks were related to the potential for mobility in spatial systems using the analogy with granular materials, that are jammed (mobility is inhibited) or unjammed (mobility is allowed) depending on the link weights in their spatial network models. Following the analogy, the mobility in the urban systems was proposed to be determined by the link weights in the Building Network (BN), and such relation was verified by simulating random mobility in the urban systems.

Within the same the analogy, link weights in the AAN should encode for the protein dynamics. Consistently, we have shown that differences in link weights are associated to a change in dynamics of an allosteric protein upon binding, simulated using Molecular Dynamics [111]. Other examples from the literature between link weights in the AAN and protein dynamics are described in Chapter 2, Section 2.5.

The multi-scale protein dynamics, characterized by Molecular Dynamics simulation and experimental investigation, are depicted based on the different structural levels that exist in a protein: secondary and tertiary structural elements have fast and small collective motions (nanoseconds to tens of microseconds) while quaternary structural elements (interfaces) and large domains have slow and large collective motions (up to seconds) [57]. Based on this evidence, we make the hypothesis that the multi-scale protein dynamics is encoded at the amino-acid level by specifically allocating atomic interactions to structural levels (1D, 2D, 3D and 4D). Since link weights w_{ij} in the AAN measure atomic interactions, the hypothesis implies that the protein dynamics is encoded in the AAN by specific

link weights associated to structural levels.

A difficulty in validating the hypothesis is that the link weights w_{ij} in the AAN are expected to carry also purely structural information, e.g. the secondary structure to which the nodes i and j belong, the number of atoms of amino acids i and j , or their distance in the amino-acid sequence. Let us consider for example two α -helices belonging to two different proteins, and let us say that one α -helix is mobile and another one has more constrained dynamics because the two helices make a different ternary interactions. According to our hypothesis, we expect the links of the amino amino acids belonging to the two helices to show some common patterns that reflect the geometrical constraint of forming an helix, but also some divergent features, because of the different constraints in the tertiary structure, leading to different dynamics.

In this chapter, the same proteins database used in Chapters 4 and 5 is analyzed in terms of atomic interactions allocations to structural levels. The goal is to explore the degrees of similarity and diversity in atomic interactions allocation among amino acids that belong to very different proteins. Studying different proteins is necessary to explore the highest number of atomic-interactions solutions that are possible for amino-acids in protein structures.

Then, in Chapter 8 the allocation of atomic interaction in protein variants having similar structure but different dynamics is measured. Because of structural similarity, if the variants show differences in allocation then the hypothesis that the allocation encodes for the protein dynamics is validated.

7.2 Classification of atomic interactions into structural levels

The protein structure is defined by four structural levels, 1D, 2D, 3D and 4D, corresponding to the primary, secondary, tertiary and quaternary structure of the protein (Chapter 2, Section 2.2.2). These structural levels are subject to different spatial constraints (Fig. 7.1.):

- The 1D structural level is determined by the covalent bonds between first neighbors in the amino-acid sequence.
- The 2D structural level is constrained by the primary structure (1D structural level), as it involves atomic interactions between amino acids that are at short range in the amino-acid sequence. Depending on the spatial constraints fulfilled at the 2D level, different secondary structures are obtained. Here, three secondary-structure elements are considered: α -helices, β -strands and coils. Coils include all amino acids that neither belong to an α -helix nor to a β -strand, i.e. turns, loops and coils are not distinguished.
- The 3D structural level is constrained by the proximity between secondary-structure elements in the three-dimensional space.

- The 4D structural level is constrained by the proximity between the protein chains in the three-dimensional space.

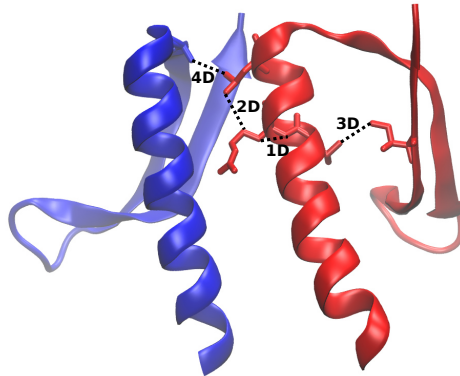


Figure 7.1: Classification of atomic interactions into four structural levels.

The atomic interactions (i.e. links in the AANs) can be classified according to the structural level (1D, 2D, 3D or 4D) to which they participate, as schematized in Fig. 7.1. As a reminder, links in the AAN are assigned when at least a pair of atoms, one belonging to one amino acid and one belonging to another, have distance $\leq 5\text{\AA}$. The following classification is proposed. A (i, j) link between amino acid i and amino acid j in the AAN of a protein is classified as: 1D if i and j belong to the same chain and are first neighbors in the protein amino-acids sequence; 2D if i and j belong to the same secondary structure element and i and j are distant by more than one and less than five positions in the protein amino-acids sequence; 3D if i and j belong to different secondary structure elements or are distant by at least five positions in the amino-acids sequence; 4D if i and j belong to two different chains. If an amino acid i makes a 1D (2D, 3D, 4D) link with its neighbor j , we say that j is a 1D (2D, 3D, 4D) neighbor of i . By symmetry, i is also a 1D (2D, 3D, 4D) neighbor of j . It should be noted that links between amino acids of two different β -strands that belong to the same β -sheet are classified as 3D links.

Amino acids are made of a finite number of atoms and the atomic distances between atoms of different amino acids have to respect steric hindrance. As seen in Chapter 4, this results in a bounded distribution of the node weights in the AANs of proteins (Fig. 4.7, top right) and a rough proportionality between the node weight and the size of the amino acid (Fig. 4.9, top right). In this sense, the atomic interactions may be seen as a finite resource that an amino acid shares with its neighbors. Because each atomic interaction belongs to one of the four structural levels, the “atomic-interactions resource” of an amino acid is allocated to the four structural levels.

As an example, the amino-acid neighborhood of amino acid THR41 of the 1B44 protein is reported in Fig. 7.2A. THR41 makes a total of 172 atomic interactions (measured by the node weight w in the AAN), shared among nine neighbors (measured by the node degree k). Among the 9 neighbors of THR41, two are 1D neighbors (ILE40 and PHE42), one is a 2D neighbor (ILE39) and six are 3D neighbors (ILE24, LEU25, SER26, LYS43,

ALA46 and THR47). By summing the weights of the links for each structural level, it is found that THR41 allocates $w_{1D} = 83$ atomic interactions to the 1D level, $w_{2D} = 4$ atomic interactions to the 2D level and $w_{3D} = 85$ atomic interactions to the 3D level.

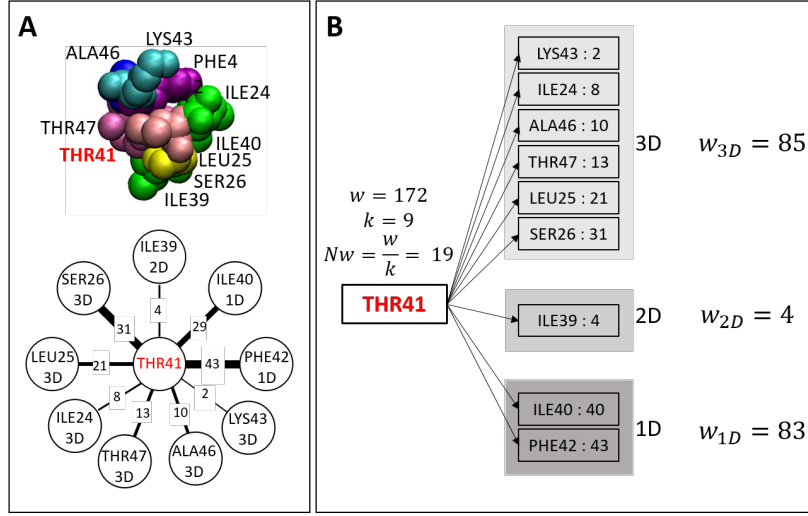


Figure 7.2: Allocation of atomic interactions into structural levels. A: Example of amino-acid neighborhood. B: Atomic interactions allocated to the 1D, 2D, and 3D structural levels.

To verify whether structural levels have characteristic link weights, the statistics of link weights in the proteins database are analyzed in the next section.

7.3 Statistics of link weights in the four structural levels in the proteins database

Figure 7.3a shows the distribution of the link weights in the proteins database, classified based on the structural level (1D, 2D, 3D or 4D) to which the link participates. Consistent with the example of the THR41 case, 1D-links have higher weights on average compared to the other links. Moreover, the shapes of the distributions of the link weight are different depending on the structural levels. The distribution of the weights of 1D-links is bell-shaped, the distribution of the weights of 2D-links is multi-modal, with a first peak at $w_{ij} \sim 2$ and a second a first peak at $w_{ij} \sim 8$, and the distributions of the weights of 3D- and 4D-links are right-skewed.

Of particular interest is the presence of two peaks in the 2D-link weight distribution. Figure 7.3b shows that the first peak is made by 2D-links between amino acids in β -strands and coils, while the second peak is made by 2D-links between amino acids in α -helices. This means that the number of atomic interactions in 2D amino-acid contacts depends on the secondary structure.

Figure 7.4 reports the two-dimensional histogram of the 2D-links weights (w_{ij}) and the distance in the sequence between the two amino acids ($|i - j|$). It shows that the peaks of the distributions of 2D-links in Figure 7.3b correspond to the weights w_{ij} of links

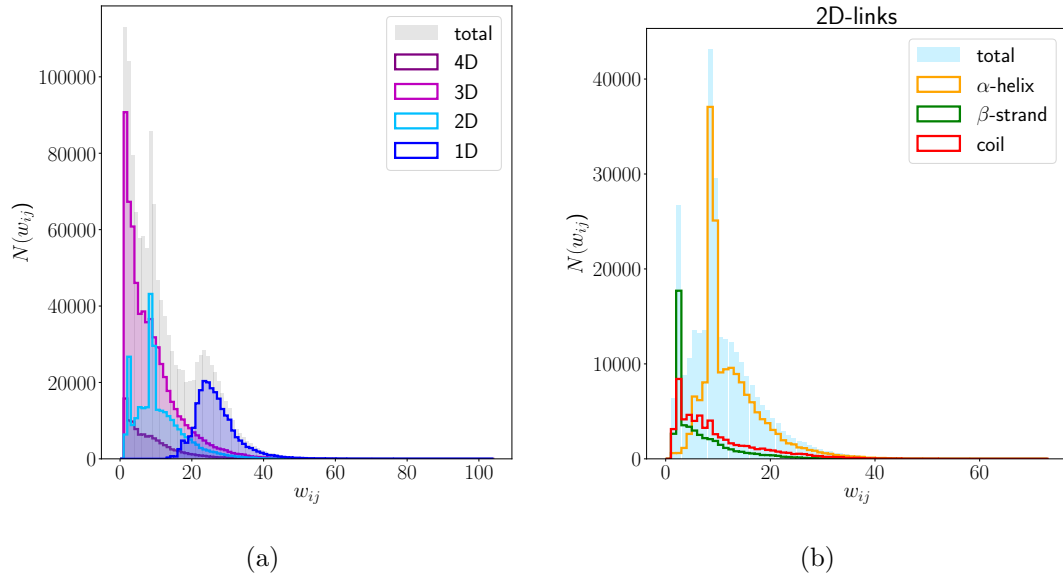


Figure 7.3: Distribution of the link weights $w_{i,j}$ in the proteins database after links classification. (a): the links are classified according to the structural level to which the link participates. (b): 2D-links only, classified based on the secondary structure of the nodes they connect.

between amino acids that are at distance $|i - j| = 2$ in the sequence. Moreover, 2D-links between amino acids at distances $|i - j| = 3, 4$ are rare in β -strands and coils, particularly in β -strands, but not in α -helices (Figure 7.3b, right column). Additionally, only in α -helices the 2D-link weights between residues at distance $|i - j| = 3$ are larger than the 2D-link weights between residues at distance $|i - j| = 2$, even though more heterogeneous. Finally, the 2D-link weights in coils at distance $|i - j| = 2$ are the most heterogeneous, consistent with the fact that coils are less constrained geometrically compared to α -helices and β -strands.

The results of Figures 7.3b and 7.4 show that the weights of 2D-links encode for the different geometrical constraints fulfilled by amino acids belonging to α -helices and β -strands. Precisely, the 2D-links weights between amino acids at distance $|i - j| = 2$ in the sequence are distinct in α -helices with respect to β -strands, while coils have broader range of admissible values.

Figure 7.5 shows that the difference in 2D-link weights between amino acids at distance $|i - j| = 2$ is due to the different conformation of the backbone: elongated in β -strands and more compact in α -helices. Coils have more variable conformations, most often elongated (peak at low 2D-link weight at distance $|i - j| = 2$ as in β -sheet, Fig. 7.4), but with a high number of compact configurations (higher values of 2D-link weight at distance $|i - j| = 2$ in Fig. 7.4).

In summary, due to different compactness of the protein backbone in α -helices versus β -strands and coils, amino acids in α -helices make more atomic interactions with their 2D neighbors compared to β -strands and coils. Fig. 7.6 reports the statistics of the node weight w and Neighborhood watch Nw for the amino acids of the proteins database,

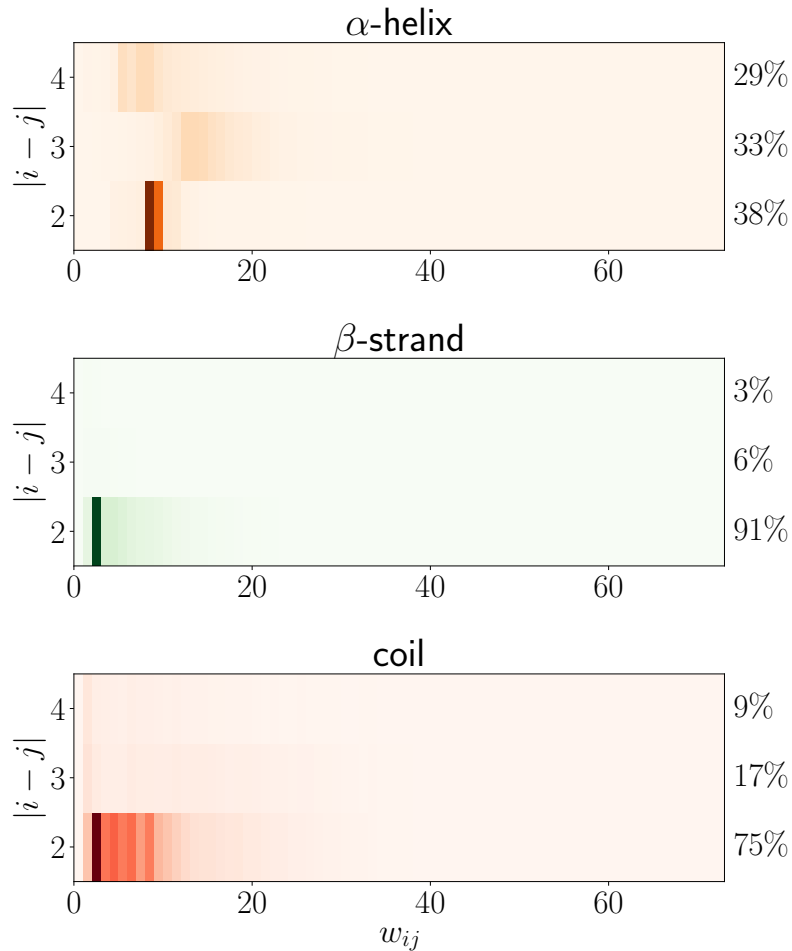


Figure 7.4: Distribution of the 2D link weights $w_{i,j}$ in the proteins database, classified based on the secondary structure of the nodes they connect and the distance between the two amino acids in the amino-acid sequence ($|i-j|$). Darker color means higher number of links for a given $(w_{i,j}, |i-j|)$ pair. On the right, the percentage of links at each distance $|i-j|$.

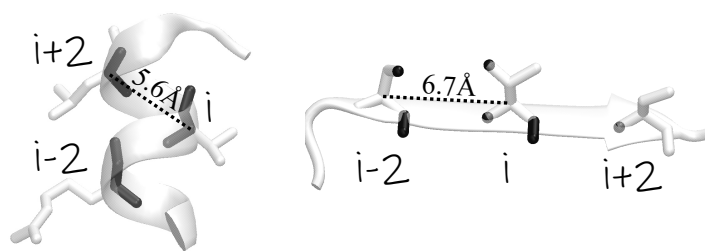


Figure 7.5: 2D atomic contacts between amino acids at distance $|i-j| = 2$ in the sequence in an α -helix (left) and in a β -strand (right). In white, the amino acids i , $i+2$ and $i-2$. In black, the atoms of these amino acids involved in the $(i, i+2)$ and $(i, i-2)$ contacts.

classified based on the type of secondary structure to which the amino acid belongs.

The $P(w)$ and $P(Nw)$ distributions in Fig. 7.6 show that despite the difference in weights of 2D-links made by α -helices with respect to β -strands and coils, the total (w) and the average (Nw) number of atomic interactions are similar regardless the secondary structure of the node. This means α -helices do not use more “atomic-interactions re-

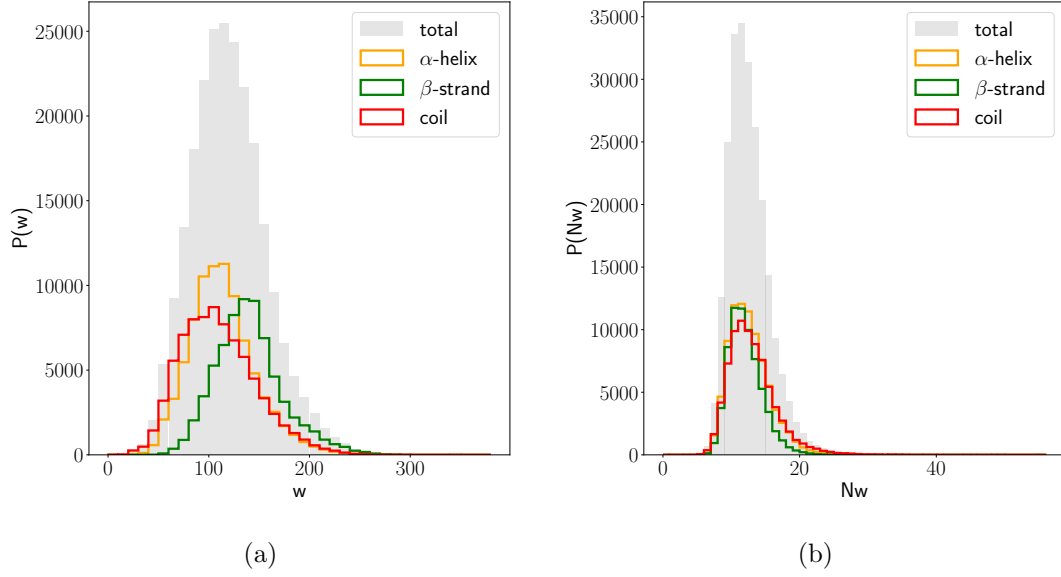


Figure 7.6: Distribution of the node weight (w) and Neighborhood watch (Nw) in the proteins database, classified based on the secondary structure. (a): Distribution of the node weight (w). (b): Distribution of the Neighborhood watch (Nw).

sources” but allocate the resources differently from β -strands and coils.

Consistently, Fig. 7.7 shows that statistically amino acids belonging to α -helices allocate more atomic interactions to the 2D structural level (w_{2D} , Fig. 7.7b) compared to β -strands and coils and inversely α -helices allocate fewer atomic interactions to the 3D and 4D structural levels (w_{3-4D} , Fig. 7.7c) compared to β -strands and coils. The number of atomic interactions allocated to the 1D structural level (w_{1D}) is independent on the secondary structure (Fig. 7.7a).

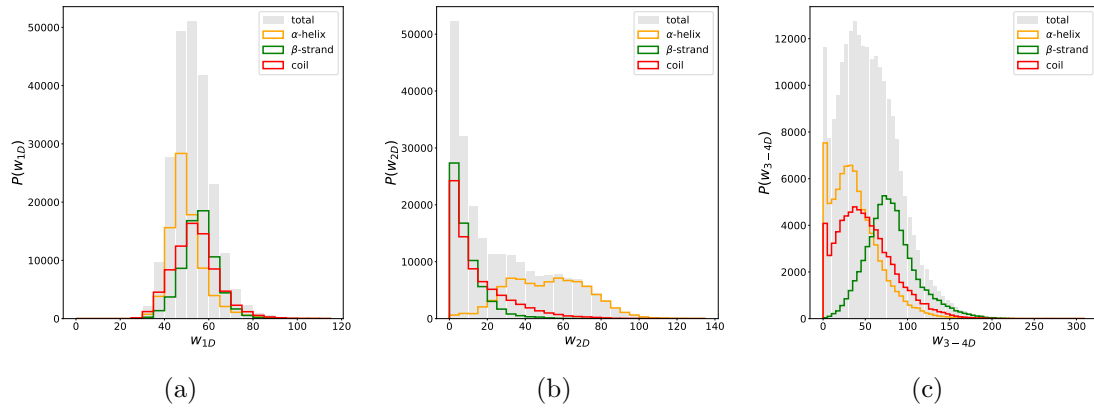


Figure 7.7: Distribution of the node weight (w) allocated to each structural levels in the proteins database, classified based on the secondary structure. (a): Distribution of the node weight in the 1D level (w_{1D}). (b): Distribution of the node weight in the 2D level (w_{2D}). (c): Distribution of the node weight in the 3D and 4D level (w_{3-4D}).

β -strands performing a large number of atomic contacts in the 3D structural level is not surprising, because β -strands assemble in the three-dimensional structures called β -sheets (Chapter 2, Fig. 2.5B). It is possible that β -sheet formation happens because

β -strands perform few interactions in the 2D structural level, leaving enough “atomic-interactions resource” available for 3D interactions. Following this direction, it would be interesting to verify whether the allocation to the 2D structural levels in α -helices is lower when the α -helix is part of an helix bundle, that is a three-dimensional arrangement of α -helices. Such analysis would require the comparative statistical analysis of two databases of proteins containing only α -helices as secondary structures, one database containing helix bundles and one database without helix bundles. This is not assessed in this work but proposed as a future direction of investigation.

Similarly, the formation of an interface during oligomerization, i.e. formation of 4D atomic interactions with a different protein chain, may be favored by a scarce number of atomic interactions in the 2D and 3D structural levels. Determining how little values make w_{2D} and w_{3D} scarce is a difficult question. Indeed, the threshold may depend on the types of amino-acid involved in the (potential) interface, and on their secondary structure (helical- versus β -interfaces). Moreover, the possibility for creating an interface depends on the geometry of the whole protein structure: the segment that would create the interface would need to be mobile enough to become exposed on the protein surface, and the rest of the protein structure on both sides of the interface should not collide. This aspect is developed more in detail in Chapter 9.

It should be also noted that drawing any causal conclusions on why a given amino acid performs a certain number of atomic interactions in each structural level is far from trivial. As an example, let us consider that an amino acid on a protein interface performs many amino acid interactions in the 4D level (high w_{4D}) and few in the 2D and 3D levels (low w_{2D} and w_{3D}). On the one hand, one may make the hypothesis that the interface was created because the w_{2D} and w_{3D} values were low, and thus the amino acid had spared atomic interactions to be made. On the other hand, one may also say that the w_{2D} and w_{3D} values are low because the w_{4D} values is high, and thus few atomic interactions could be made in the other structural levels. These two situations are not distinguishable if the folding mechanism of the protein is unknown, i.e. it is unknown whether oligomerization follows or precedes the folding of the secondary and tertiary structures.

Drawing hypotheses on the unfolding mechanism of the protein, rather than its folding mechanism, is easier. Let us take again the example of an amino acid on a protein interface. If it makes few atomic interactions in the 2D and 3D structural levels and many interactions in the 4D structural level, one may expect the secondary and tertiary structures to unfold before interface dissociation. Inversely, if it makes many atomic interactions in the 2D and 3D structural levels and few in the 4D structural level, one may expect the interface to dissociate before the monomer unfolds. This hypothesis will be verified in the next chapter, for the case of two variants of the B-subunit pentamers of AB₅ toxins that have similar structure but different experimental folding and unfolding mechanisms.

More in general, if two amino acids perform a different number of atomic interactions in the 1D, 2D, 3D and 4D structural levels, it means that the space is occupied differently around them. By difference, also the empty space available for atomic motions is carved differently around the two amino acids. We make the hypothesis that low values of w in a given structural level means that the atomic motions are less constrained in that structural level, and thus that different allocations of atomic interactions (link weights in the AAN) to structural levels result in different protein dynamics.

It should be noted that the total number of atomic interactions that may be performed by an amino acid depends on its size. Thus, to compare the allocation of atomic interactions to structural level made by amino acids of different type, relative quantities and not absolute number of interactions should be employed. In the next session, we propose to normalize the number of atomic interactions allocated to each structural level (w_{1D} , w_{2D} , etc.) by the total number of atomic interactions made by the amino acid (w).

Moreover, a data visualization technique is needed to represent how single amino acids allocate atomic interactions to structural levels. Indeed, Fig 7.3a shows that diverse solutions of link weights allocations to structural levels are possible, but it does not provide any information on the relations between structural levels. Similarly, Fig. 7.7 shows that statistically α -helices perform more 2D interactions and less 3-4D interactions than β -sheets, but it does not provide any information on the relation between the allocations of interactions to the 2D and 3-4D structural levels at the single amino-acid level. In the next section, ternary plots are proposed as a mean to visualize the structural-level allocation of atomic interactions at the single amino-acid level.

7.4 Measure and visualization of the local structural-level allocation of the atomic interactions

For an amino acid i and for a link category x , with $x = 1D, 2D, 3D$ or $4D$, we define the fraction of atomic interactions allocated by i to the category x as $f_{w_{x,i}} = \frac{w_{x,i}}{w_i}$, where $w_{x,i}$ is the number of atomic interactions of category x made by i (i.e. the sum of the weights $w_{i,j}$ of links of category x in the AAN, for all neighbors j of node i) and w_i is the total number of atomic interactions made by i (i.e., the sum of the weights $w_{i,j}$ in the AAN, for all neighbors j of node i). We define local structural-level allocation of the atomic interactions of an amino acid i the set of $(f_{1D,i}, f_{2D,i}, f_{3D,i}, f_{4D,i})$ values. By construction, $f_{1D,i} + f_{2D,i} + f_{3D,i} + f_{4D,i} = 1$ (details in Methods, Section 3.1.5).

For the example of THR41 (Fig. 7.2), the structural-level allocation of the atomic interactions is $(f_{1D,THR41} = 0.48, f_{2D,THR41} = 0.02, f_{3D,THR41} = 0.49, f_{4D,THR41} = 0)$: the THR41 amino acid allocates around half of its “atomic-interactions resource” to the 1D structural level, around half to the 3D structural level, and only a minimal part to the 2D structural level.

The local structural-level allocation of atomic interactions of amino acids in the AANs is visualized using ternary plots, as in Fig. 7.8.

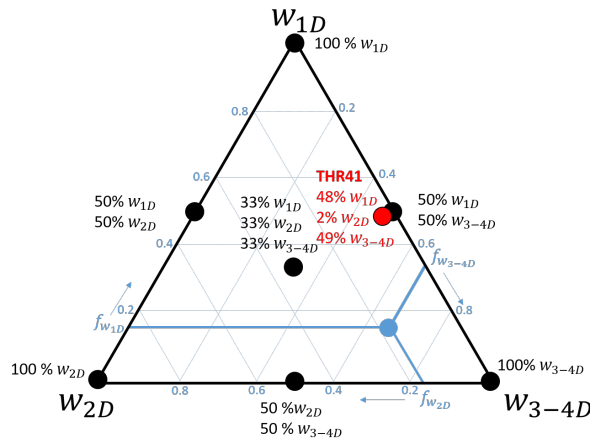


Figure 7.8: Ternary plot of the structural-level allocation of the atomic interactions to the 1D, 2D and 3-4D structural levels.

In the ternary plot, the allocation to the 3D and 4D levels are merged into a unique level called 3-4D, because the atomic interactions belonging to the 3D and 4D levels are subject to similar constraints: they both represent interactions in the three-dimensional space between secondary-structure elements. Each vertex of the triangle in the ternary plot corresponds to a structural level (1D, 2D or 3-4D). In the ternary plot, each amino acid (i.e. each node in the AAN) is represented by a point. Amino acids allocating atomic interactions homogeneously to the three structural levels (1D, 2D and 3-4D) are close to the center of the triangle. Amino acid allocating most of their atomic interactions to one structural level are close to the corresponding vertex in the triangle and amino acids allocating interactions mostly to two structural levels are close to edges of the triangle. For example, the THR41 amino acid is close to the edge connecting the vertices w_{1D} and w_{3-4D} because it allocates atomic interaction mostly to the 1D and 3D structural levels and only in minimal part to the 2D structural level. Finally, the scheme in light blue in 7.8 shows how to read the $(f_{1D,i}, f_{2D,i}, f_{3D,i}, f_{4D,i})$ values of any point in the ternary plot.

The following section is dedicated to the analysis of the statistics of the allocation of atomic contacts to the structural levels by the amino acids in the proteins database.

7.5 Statistics of the local structural-level allocation of the atomic interactions in the proteins database

The local structural-level allocation of atomic interactions was calculated for all amino acids of the proteins database apart from the amino acids performing only one 1D link. These amino acids are the N- and C-terminus of all chains, plus the ones next to missing segments in the crystalline structure. These amino acids were removed because they could

introduce a bias towards lower values of f_{w1D} . After removing the amino acids making one 1D link, the database consists of 225590 amino acids.

Fig. 7.9 shows the ternary plot of the local structural-level allocation of the atomic interactions for the amino acids of the protein database.

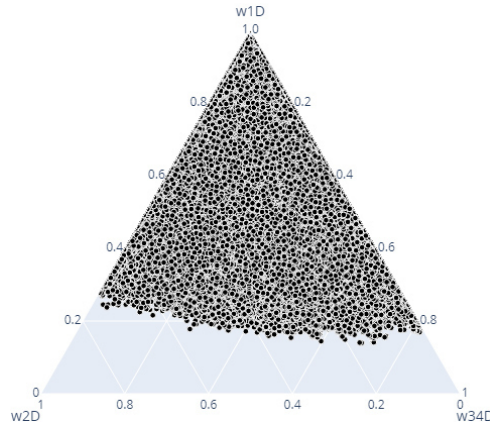


Figure 7.9: Allocation of the atomic packing into structural levels for the nodes of the proteins database.

The bottom part of the plot, far from the w_{1D} vertex, is forbidden because it corresponds to neighborhoods with minimal allocation of atomic interactions to the 1D structural level. This is consistent with the w_{1D} values being statistically higher compared to the other structural levels (Fig. 7.3a). The data points occupy all the remaining space of the plot, showing that the amino acids cover a wide range of solutions in terms of allocation of atomic interactions, despite the fact that the average number of atomic interactions performed by the amino acids is uniform (Fig. 7.6b).

Using fractions of atomic interactions allocated to the structural levels instead of absolute values was proposed as a normalization technique to allow the comparison of the allocation for different amino-acid types. We have verified whether the statistics of the allocation measure for amino acids of different types are similar, as expected from the normalization. Fig. 7.10 shows the ternary plots of the amino-acids in the database discriminating based on the amino-acid type. The amino acids are ordered from the smallest (GLY) to the largest (TRP) in terms of number of atoms.

The result shows that amino acids CYS, MET, PHE, TYR and TRP populate the upper and left-hand sides of the ternary-plot space less frequently than the other amino acids, implying that for these amino-acid types the allocation to the 1D and 2D structural levels is rarely predominant compared to the allocation to the 3-4D levels. However, these amino-acid types are also under-sampled respect to the other amino-acid types, suggesting that the observed differences may be due to under-sampling and not to an actual tendency to allocate more atomic interactions to the 3-4D structural level. This possibility is supported by the fact that some data points are present in the upper- and left-hand sides of the plots for these amino-acid types, meaning that there is no physical barrier to the population of these regions of the plots. It can be concluded that the

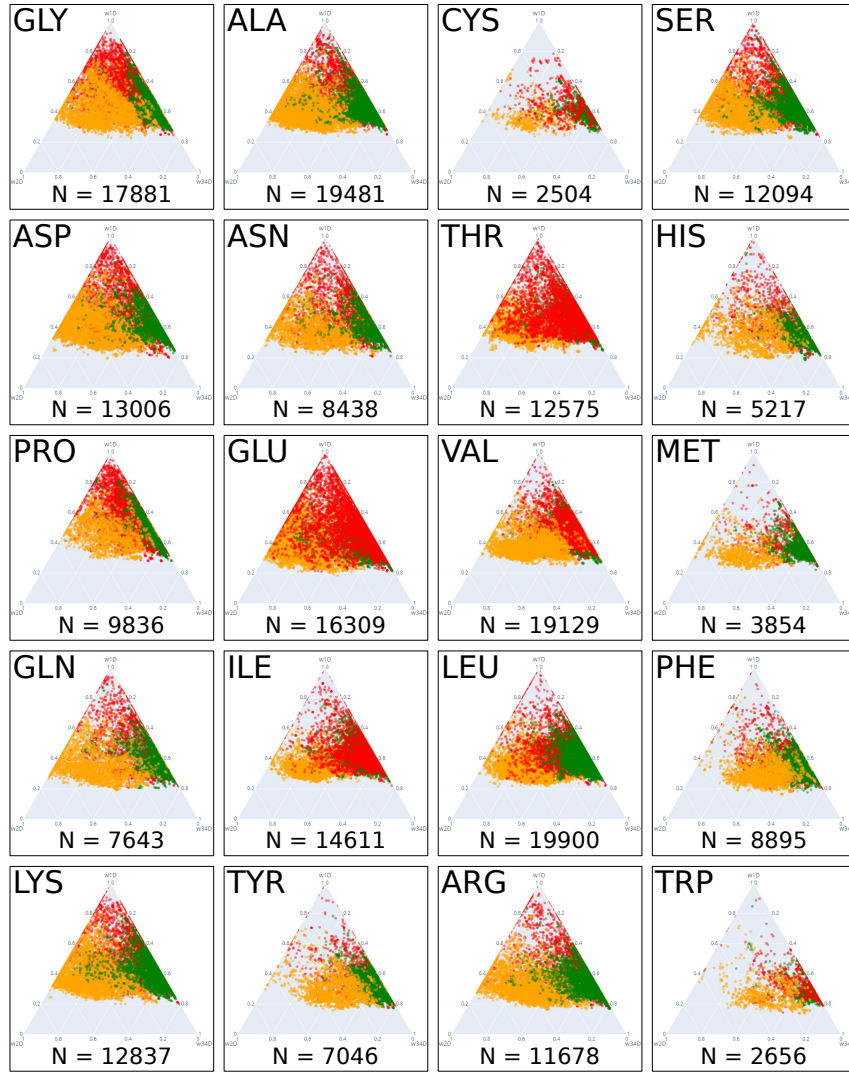


Figure 7.10: Allocation of the atomic packing into structural levels for the nodes of the proteins database, classified based on the amino-acid type. The amino acid types are ordered based on their size (number of atoms). The color code is the following: yellow for α -helices, green for β -strands and red for coils. For each amino-acid type, the number of entries N is reported.

allocation of atomic interactions to structural levels does not depend on the amino-acid type. Thus, it is possible to compare the allocation measure across amino acids of different type.

In the previous section it was found that the statistics of the number of atomic interactions allocated to the 2D and 3-4D structural levels are different depending on the secondary structure of the amino acid (Fig. 7.3). Fig. 7.11 shows the ternary plots of the local structural-level allocation of atomic interactions for the amino acids in the proteins database, classified based on the type of secondary-structure element they belong to (α -helix, β -strand or coil).

The comparison of the areas of the ternary plot populated by the amino acids belonging to each secondary-structure type shows that the region close to the w_{2D} vertex is reserved to α -helices: the maximum values of $f_{w_{2D}}$ are 0.73, 0.55 and 0.65 for α -helices, β -strands

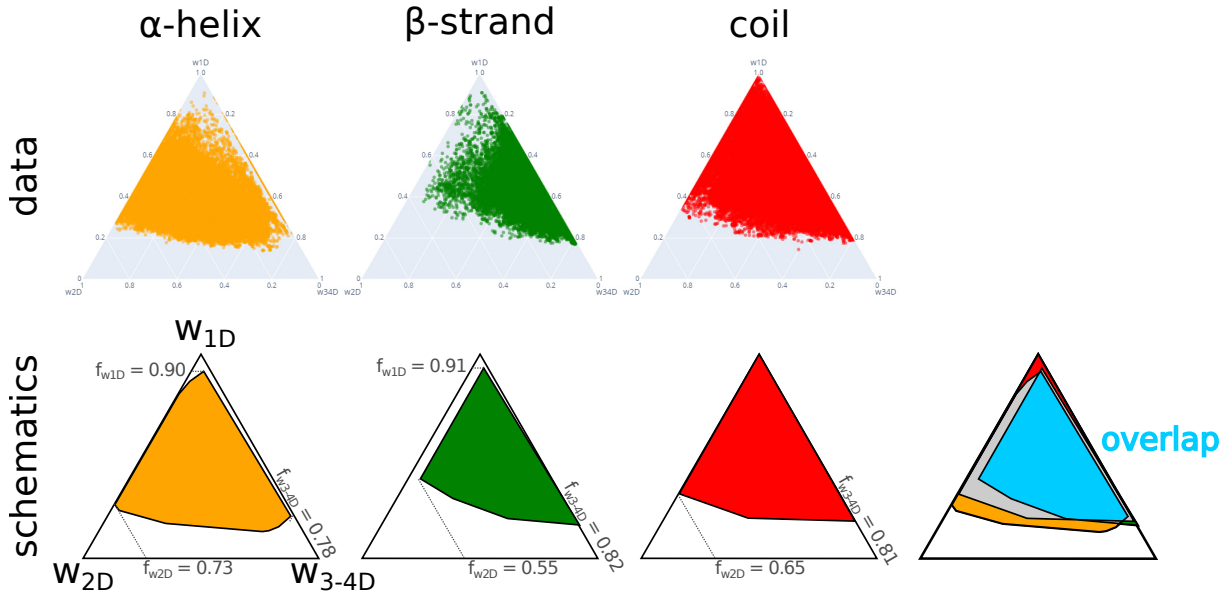


Figure 7.11: Allocation of the atomic packing into structural levels for the nodes of the proteins database, classified based on the secondary structure.

and coils, respectively (Fig. 7.11). This is consistent with the fact that α -helices have statistically higher w_{2D} compared to β -strands and coils (Fig. 7.3b).

For what concerns the 3-4D structural levels, all secondary-structure types have similar maximal values $f_{w_{3-4D}} \sim 0.8$, meaning that the percentage of atomic interactions allocated to the 3D and 4D structural levels can be up to 80% high independently on the secondary structure. We have checked the maximal values of $f_{w_{3D}}$ (i.e. without 4D interactions) and they remain similar to 0.8 for all secondary structures. Nevertheless, the average values of $f_{w_{3-4D}}$ are 0.3, 0.5 and 0.4 for α -helices, β -strands and coils, respectively. This is consistent with β -strand making on average more 3D and 4D atomic interactions compared to α -helices and coils (Fig. 7.7c).

An additional difference between the secondary-structure types is that only amino acids belonging to coils can have $f_{w_{1D}} > 0.91$ (Fig. 7.11). This means that α -helices and β -strands are possible only when some atomic interactions are performed in the 2D and 3D levels.

Despite the differences in statistics among the three types of secondary structures, it should be noted that the allocation measure covers a broad range of values for all secondary-structure types. Within the hypothesis that the allocation of atomic interaction encodes for the protein dynamics, the broadness of the point clouds in Fig. 7.11 would mean that a wide range of dynamics are possible for the same type of secondary-structure element. This is consistent with protein-specific functional dynamics. The point clouds of Fig. 7.11 have a large overlap (light blue in the figure), suggesting that similar dynamics may be performed by amino acids belonging to different types of secondary-structure elements.

While the diversity in atomic interactions allocation solutions adopted by amino acids

suggests a diversity in dynamics, the observed diversity may also be due to the fact that in Fig. 7.11 the position of the amino acid in the protein structure, buried in the protein core or surface-exposed, is not considered. To verify whether the structural position of the amino acid determines the allocation of atomic interactions to structural levels, the relative Accessible Surface Area (rASA) was calculated for the amino acids of the database using the DSSP method implemented in BioPython [183] and amino acids were classified as buried if their value of rASA is equal or inferior to 0.2 and as surface-exposed otherwise.

Figure 7.12 shows the ternary plots of the allocation, where amino acids are colored based on the rASA value. DSSP failed for 89 PDB structures^a over the 250 structures of the database. These 89 structures are discarded in Figure 7.12.

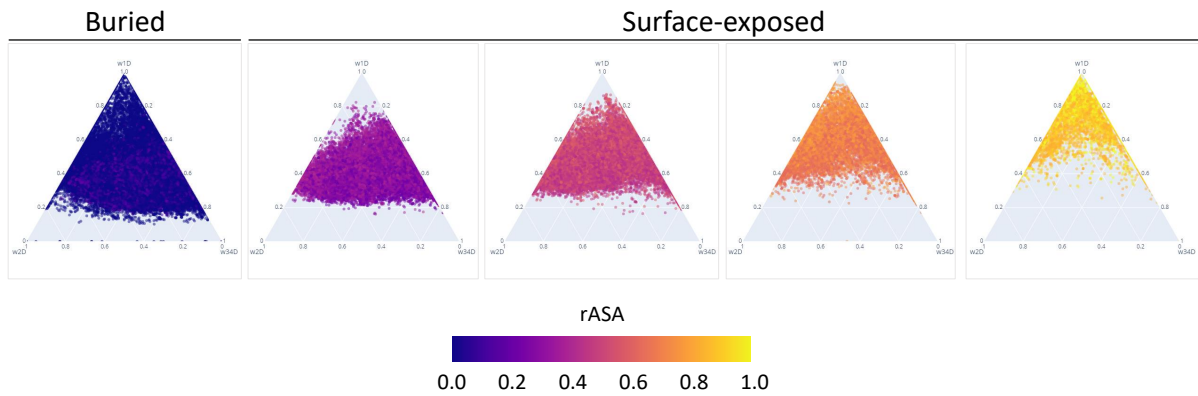


Figure 7.12: Allocation of the atomic packing into structural levels for the nodes of the proteins database, classified based on the relative Accessible Surface Area (rASA). Left: buried amino acids ($rASA \leq 0.2$). Right: surface-exposed amino acids ($rASA > 0.2$), subdivided into four intervals of rASA from 0.2 to 1.0.

Interestingly, the upper area of the plots, near the w_{1D} vertex, is occupied by amino acids that are either buried ($rASA \leq 0.2$) or surface-exposed with very large rASA ($rASA > 0.6$). Nevertheless, the point clouds of all the classes of rASA are broad and largely overlapping. It can be concluded that the allocation of atomic interaction to structural levels does not depend on whether the amino acid is buried or surface-exposed.

7.6 Conclusion

The measure of the local allocation of atomic interaction to structural levels shows that diverse solutions of atomic interactions are adopted by amino acids. These solutions represent different means to achieve a uniform average local atomic packing, measured by the Neighborhood watch Nw .

^a1E2Y, 1GKR, 1GX1, 1L3I, 1MW5, 1NJK, 1Q6W, 1SS4, 1T0A, 1TQ8, 1X7V, 1XEA, 1XRX, 1Y12, 1Y1O, 1Y9B, 1YLL, 2AP6, 2B6E, 2CV8, 2EIS, 2F2E, 2FS2, 2FYW, 2GW4, 2H6L, 2I7H, 2IEY, 2NWA, 2O2A, 2OBA, 2OIW, 2OQG, 2P0G, 2P5V, 2PD1, 2PDO, 2PKH, 2QDY, 2QL3, 2QM0, 2RA2, 2RAQ, 2VQ5, 3BDU, 3BID, 3BRC, 3BU2, 3D3R, 3D6X, 3DWA, 3E6Q, 3EUW, 3FIJ, 3FK3, 3H43, 3HE1, 3HXX, 3HYK, 3IO1, 3JW8, 3JYG, 3K12, 3KXE, 3L6E, 3MMZ, 3NJA, 3O6Q, 3OJY.

The proposed measure allows comparing the allocation of atomic interactions made by amino acids of different type. The measure was shown to be independent of the rASA of the amino acid and is able to pick up the differences in constraints associated to different secondary-structure types. Despite the differences, a broad range of solutions is adopted regardless the secondary-structure, and it was proposed that such diversity reflects the encoding of different protein dynamics at the amino-acid level.

The next chapter verifies the hypothesis that the allocation of atomic interactions to structural levels carries information on the protein dynamics. This is performed by comparison of the allocation measure in protein variants having similar structure but different experimental dynamics.

Chapter 8

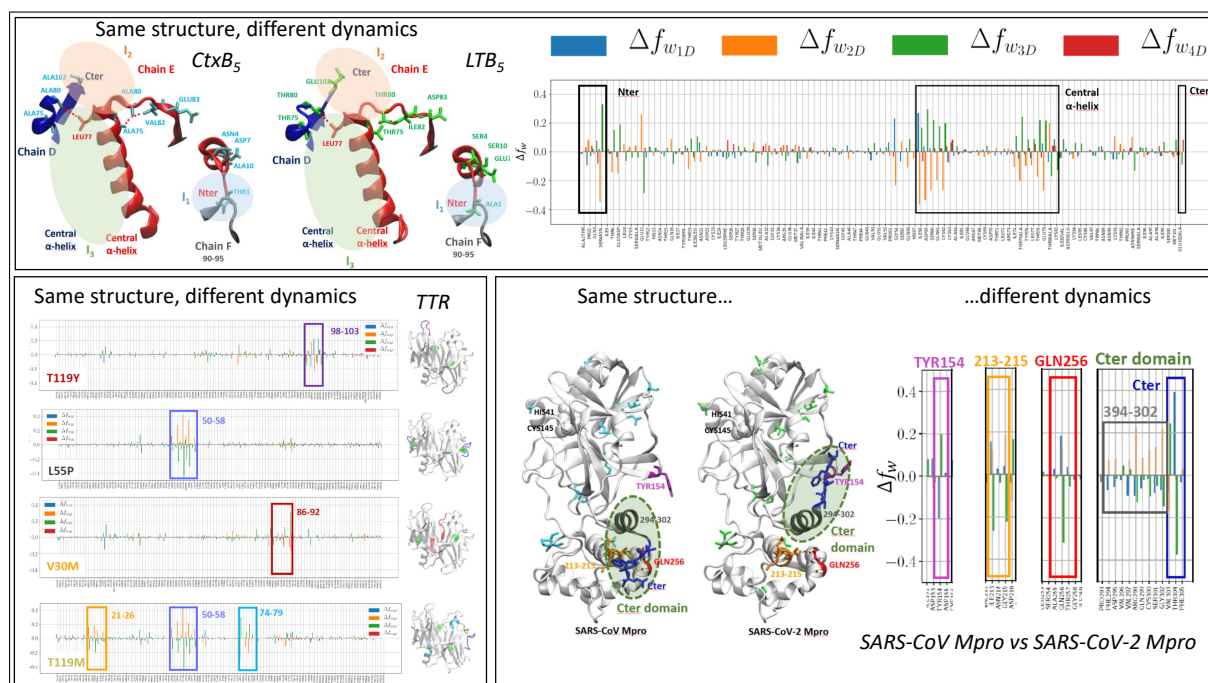
Case study: Transthyretin, B-subunits toxin pentamers, Main Proteases of SARS coronaviruses. Measure of the impact of amino-acid mutations on the protein dynamics.

Highlights:

- Protein structures are studied based on the allocation of atomic contacts to primary, secondary, tertiary and quaternary structural levels.
- Structurally-similar variants having different dynamics allocate atomic contacts differently.
- Differences in allocation are shown by variants with single- and multiple mutations.
- The dynamical perturbation depends on the structural level of the perturbed atomic contacts.

Abstract: Dynamic perturbations due to mutations are diagnosed by the measure of differences in the allocation of atomic interactions to structural levels. The methodology is validated using structurally-similar protein variants known to have different experimental dynamics (B-subunit pentamers of AB5 toxins and Transthyretin variants) and applied to proteins from SARS and COVID-19 coronaviruses, questioning the use of common drugs for their inhibition.

Graphical abstract:



Methods: Local structural-level allocation of the atomic interactions (Section 3.1.5).

Submitted publication: Pacini L. and Lesieur C. A computational methodology to diagnose sequence-variant dynamic perturbations by comparing atomic protein structures. *Under review in Bioinformatics*, January 2021

8.1 Introduction

In Chapter 7 it was proposed that the local allocation of atomic interactions to structural levels encodes for the protein dynamics. In this chapter, this hypothesis is validated through the comparison of the allocation for protein variants having superimposable structure but different folding dynamics. Structural integrity imposes that differences in the atomic-contacts allocation are due to dynamics perturbation and not structural changes.

We have selected case studies that are involved in human diseases. The first case study is the comparison of the B-pentamers of two AB₅ toxins, Cholera toxin (CtxB₅) and human heat labile (LTB₅), which have 82% of sequence identity, very similar structure (RMSD = 0.59 Å) and function but different folding/unfolding mechanisms: the formation of assembly intermediates is the rate-limiting step for CtxB₅ (fly-casting folding mechanism) while the folding of the monomers is the rate-limiting step for LTB₅ (induced-fit folding mechanism) [23, 184]. Since the differences in folding dynamics of CtxB₅ and LTB₅ are known, they represent a good case-study to test the hypothesis that differences in protein dynamics are encoded by different local structural-level allocation of atomic interactions in the protein structure.

The second case study is Transthyretin (TTR), that is involved in neurological and cardiac genetic diseases due to the precipitation of misfolded TTR in the form of amyloid fibrils and whose pathogenic variants show an increased tendency to amyloidogenesis [185–187]. The exact mechanism of amyloid formation of TTR is unknown and probably depends on the particular TTR variant and on the experimental conditions [188–192]. The study of TTR variants to explain their different folding pathways challenges the sensitivity of the proposed methodology to the detection of the impact of single amino acid mutations.

The third case study is the comparison of the main protease (Mpro) of two coronaviruses: SARS-CoV and SARS-CoV-2. SARS-CoV is responsible of the Severe Acute Respiratory Syndrome (SARS) pandemic that caused 774 deaths in 2003 (www.who.int/csr/sars/country/table2004_04_21/en/, last visited on 8th June 2020) and SARS-CoV-2 is responsible of the COVID-19 global pandemic that has affected 216 countries worldwide since December 2019, causing more than three million deaths to the present moment (www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/, last visited on 22nd April 2021). The inhibition of Mpro is one of the targets for the antiviral treatment of COVID-19 thanks to its functional importance in virus replication [193, 194]. Moreover, the structural similarity between the SARS-CoV Mpro and the SARS-CoV-2 Mpro has motivated the research for common drugs that could inhibit the Mpro activity of a broad spectrum of coronaviruses [195–197]. We apply our methodology, validated through the first two case studies, to the comparison of the local structural-level allocation of atomic interactions of SARS-CoV Mpro and the SARS-CoV-2 Mpro, to un-

derstand whether a different dynamical behavior for the two proteins is to be expected or not.

8.2 Comparison of local structural-level allocation of atomic interactions in protein variants

The PDB codes of the protein structures are the following. CtxB₅: 1EEI, LTB₅: 1LTR, WT TTR: 1F4I, T119Y TTR: 4TNE, L55P TTR: 3DJZ, V30M TTR: 3KGS, T119M TTR: 1BZE, SARS-CoV Mpro: 2H2Z, SARS-CoV-2 Mpro: 6Y2E. Chains D to H have been studied for CtxB₅ and LTB₅, from position 1 to position 102. Chains A and B have been studied for the TTR variants, from position 10 to position 124. Chain A has been studied for the Mpro of SARS-CoV and SARS-CoV-2, from position 1 to position 305.

Structural similarity between protein variants is assessed from the Root Mean Square Deviation (RMSD) measure, calculated with TopMatch [198].

The local structural-level allocation of atomic interactions $(f_{w_{1D,i}}, f_{w_{2D,i}}, f_{w_{3D,i}}, f_{w_{4D,i}})$ is calculated for all the amino acids of the protein structures as described in Chapter 7. The local structural-level allocation of atomic interactions in two protein variants is compared through the differences $(\Delta f_{w_{1D,i}}, \Delta f_{w_{2D,i}}, \Delta f_{w_{3D,i}}, \Delta f_{w_{4D,i}})$ for each amino acid i of the protein structure.

It should be noted that differently from Chapter 7, allocation of atomic interactions to the 3D- and 4D-levels are here measured separately for the AB₅ toxins and the TTR variants. This choice comes from the fact that folding and unfolding dynamics are assessed, and the folding/unfolding of the tertiary structure may be kinetically distinct from the association/dissociation of the quaternary structure. For the Mpro of SARS-CoV and SARS-CoV-2, the monomeric form is studied, so there is no 4D structural level.

8.2.1 B-subunit pentamers of AB₅ toxins

CtxB₅ and LTB₅ have 18 mutated residues over 102 (82% of sequence identity, Fig. 8.1A) and very similar structure (RMSD = 0.59 Å, Fig. 8.1A), but different folding/unfolding mechanisms. We have compared the local structural-level allocation of the atomic interactions in LTB₅ with respect to CtxB₅, for all amino acid positions, through the measure of $\Delta f_{w_{1D,i}} = f_{w_{1D,i}}^{LTB_5} - f_{w_{1D,i}}^{CtxB_5}$ (and similarly for $\Delta f_{w_{2D,i}}$, $\Delta f_{w_{3D,i}}$ and $\Delta f_{w_{4D,i}}$). The pentameric unit has been employed for the calculations, whereas the results (Fig. 8.1) are plotted only for chain E, for simplicity. The plots for all five chains are reported in Appendix B, Fig. B.3. The results that were not consistent in at least four chains over five have been considered as potential X-ray artifacts and not taken into consideration. Differences in $f_{w_{2D}}$, $f_{w_{3D}}$ and $f_{w_{4D}}$ are found at the N-terminal, the C-terminal and at the central α -helix of the pentamers (Fig. 8.1C, boxes), all involved in the interfaces of the pentamers (I₁, I₃

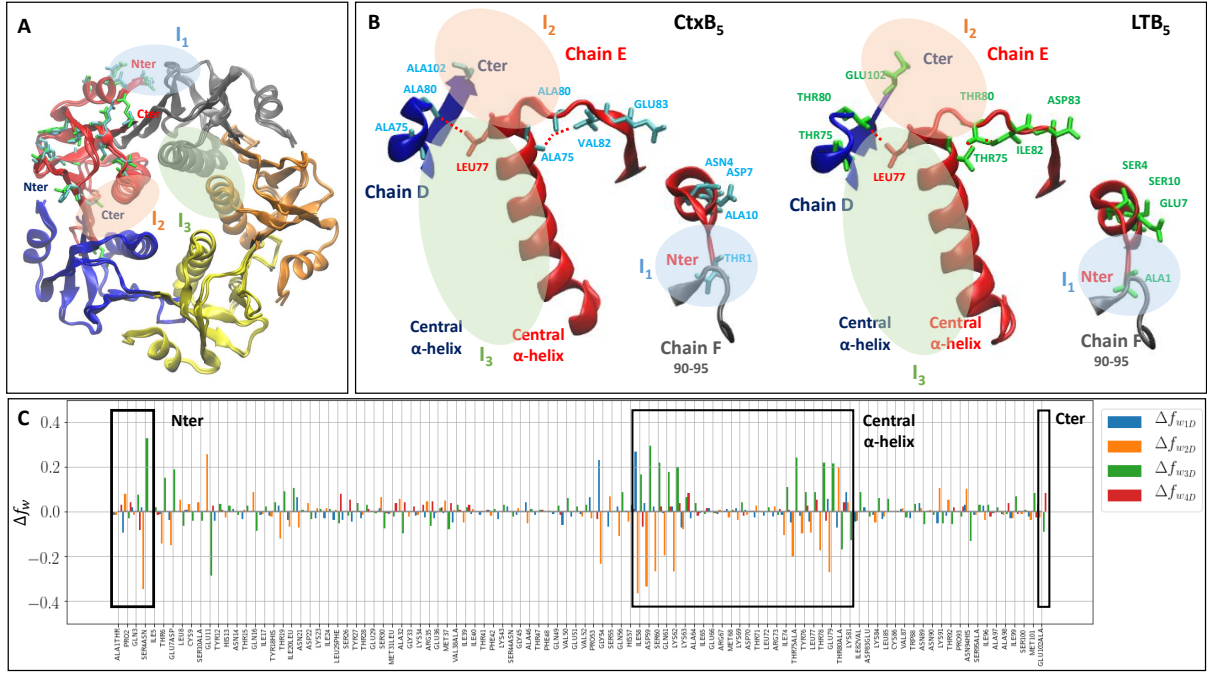


Figure 8.1: B subunits of AB₅ toxins. A: Superposition of the X-ray structures of the CtxB₅ pentamer (PDB: 1EEI) and the LTB₅ pentamer (PDB: 1LTR). The 18 mutated positions are highlighted in one chain (cyan amino acids for CtxB₅ and green amino acids for LTB₅). B: Details of the X-ray structures of CtxB₅ and the LTB₅. C: Changes in the local structural-level allocation of atomic interactions in LTB₅ compared to CtxB₅ (chain E), where $\Delta f_{w1D,i} = f_{w1D,i}^{LTB_5} - f_{w1D,i}^{CtxB_5}$ (and similarly for $\Delta f_{w2D,i}$, $\Delta f_{w3D,i}$ and $\Delta f_{w4D,i}$).

and I₂, respectively, Fig. 8.1A and 8.1B). LTB₅ allocates more atomic interactions to the 4D structural level at the I₂ and I₃ interfaces compared to CtxB₅ ($\Delta f_{w4D} > 0$), while it allocates less atomic interactions to the 2D structural level and more to the 3D structural level at the I₁ and I₃ interfaces compared to CtxB₅ ($\Delta f_{w2D} < 0$ and $\Delta f_{w3D} > 0$). The fact that more atomic interactions are allocated to the maintenance of the interface compared to the secondary-structure fold means that the secondary structure will be more susceptible to unfolding compared to the tertiary and quaternary structure: when the interfaces are disrupted, the monomers are already unfolded. This is consistent with the experimental evidence that assembly intermediates are not observed during unfolding of LTB₅, contrary to CtxB₅ [23, 184].

It must be underlined that only one mutation (ALA75THR) is present in the central α -helix, where most of the changes in allocation of atomic interactions between the two toxins are observed, making it difficult to anticipate the role of the α -helix in the protein dynamic differences from sequence comparison alone and proving the multi-scale nature of the dynamics perturbations caused by the 18 mutations.

The comparison of LTB₅ and CtxB₅ has demonstrated the validity of our hypothesis that differences in protein dynamics can be retrieved by differences in the local structural-level allocation of atomic interactions into the 1D, 2D, 3D and 4D classes. The following case study is then devolved to the assessment of the sensitivity of this measure to quantify

the impact of mild perturbations of the atomic interactions network of a protein (single amino acid mutations) on the protein folding dynamics.

8.2.2 Transthyretin variants

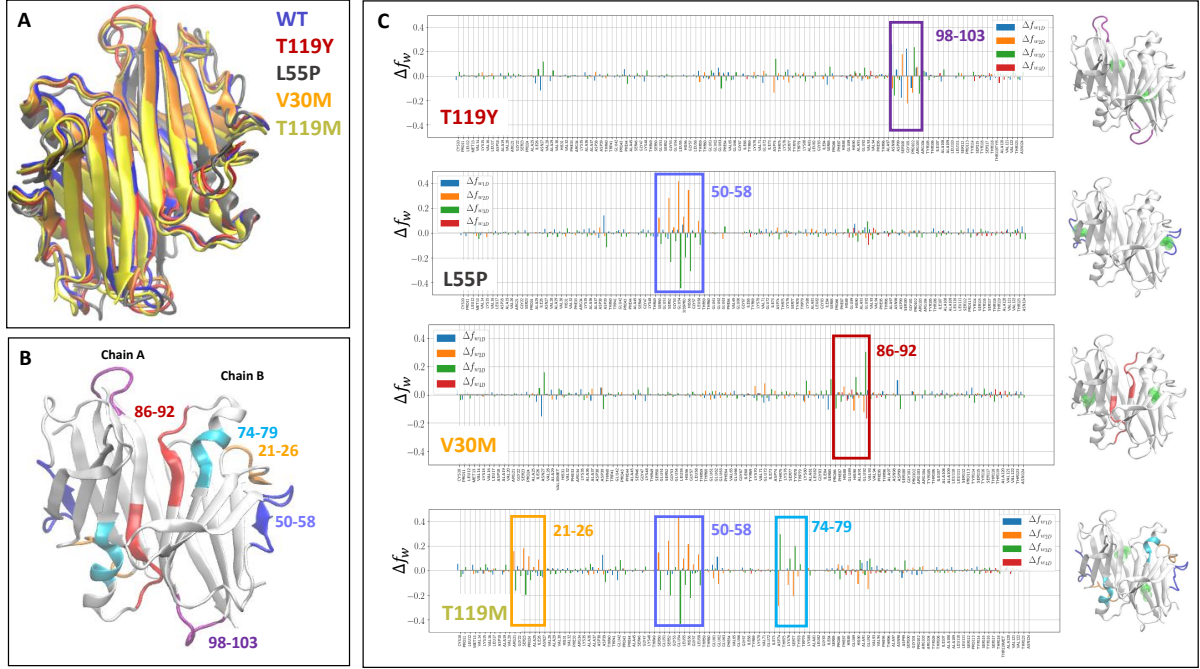


Figure 8.2: Transthyretin (TTR) variants. A: Superposition of the X-ray structures of the TTR dimer variants: WT (PDB: 1F41), T119Y (PDB: 4TNE), L55P (PDB: 3DJZ), V30M (PDB: 3KGS) and T119M (PDB: 1BZE). B: X-ray structure of the WT TTR dimer. The segments involved in changes in local structural-level allocation in the variants with respect to the WT are highlighted. C: Changes in the local structural-level allocation of atomic interactions in the TTR variants compared to WT TTR (chain A), where $\Delta f_{w1D,i} = f_{w1D,i}^{variant} - f_{w1D,i}^{WT}$ (and similarly for $\Delta f_{w2D,i}$, $\Delta f_{w3D,i}$ and $\Delta f_{w4D,i}$). Next to each plot, the segment involved in changes in local structural-level allocation of atomic interactions in the variants compared to the WT TTR are highlighted in the variant's X-ray structure, with the same color code as the boxes in the plots and panel B. The mutated position is highlighted in green.

Following the same procedure as for the AB₅ toxins, we investigate the local structural-level allocation of atomic interactions in functionally-robust (T119Y and T119M [187, 199]) and pathogenic (L55P and V30M, showing increased amyloidosis [187]) variants of TTR, compared to the WT TTR. Again, the four considered TTR variants have very similar crystalline structure compared to the WT TTR (RMSD values: 0.58 Å, 0.58 Å, 0.54 Å and 0.34 Å for L55P TTR, V30M TTR, T119Y TTR and T119M TTR, respectively, Fig. 8.2A). The dimeric unit has been employed for the calculations, whereas the results (Fig. 8.2C) are plotted only for chain A, for simplicity. The plots for both chains are reported in Appendix B, Fig. B.1 and Fig. B.2. The results that were not consistent for both chains have been considered as potential X-ray artifacts and not taken into further consideration.

The T119Y TTR variant shows very little changes in local structural-level allocation of atomic interactions compared to the WT TTR, apart from the 98-103 segment (Fig.

8.2B and 2C). Changes in this segment, which is part of a long coil, are observed for the four variants, with changes in chain A different from the changes in chain B (Appendix B, Fig. B.1 and Fig. B.2), suggesting a perturbation related to an X-ray bias rather than a dynamic fault. The absence of changes in the local allocation for the T119Y TTR is consistent with the fact that it does not show propensity to amyloid formation [199].

In contrast, the pathogenic L55P TTR variant shows large differences compared to the WT due to the reallocation of the atomic interactions of many residues within the segment 50-58 (Fig. 8.2C). The residues in the mutant devote more resources to 2D interactions and less to 3D interactions, this may increase the 3D-mobility of the segment 50-58 and make it available for intermolecular interactions with another monomer, leading to amyloid fiber formation. This is consistent with the L55P TTR amyloid fiber model we have recently proposed [24] and with the one proposed by Sebastião and collaborators [200]. This L55P TTR fiber model is presented in Chapter 9.

The changes in allocation of atomic interactions for the other pathogenic variant, V30M TTR, are different from the ones of L55P TTR (Fig. 8.2C). The values of Δf_w are lower compared to L55P TTR, but differences in $f_{w_{2D}}, f_{w_{3D}}$ and $f_{w_{4D}}$ are present at a dimer interface (residues 86-92, Fig. 8.2B and 2C). This is consistent with the fact that ex-vivo studies have shown that V30M TTR unfolds prior to fiber formation, contrary to WT TTR [188, 191].

We find little perturbation for the non-pathogenic T119Y variant and substantial perturbation for the pathogenic L55P and V30M variants. Moreover, the L55P and V30M variants exhibit allocation perturbations at different areas. V30M has interface perturbations suggesting that the mutation leads to dissociation prior fiber formation, in agreement with the diameter of the V30M fiber, higher than the size of the TTR tetramer [201]. The L55P variant has intra-molecular perturbations, suggesting a folding perturbation, in agreement with the L55P fiber diameter, similar to the TTR tetramer size [202], meaning that dissociation can happen but is not necessary for fiber formation.

The link between the quantity of perturbations and the pathogenicity of the mutation recalls what has been previously reported for another local measure, referred to as localized frustration [203, 204]. However, the same logic does not hold for the case of the non-pathogenic T119M variant, which shows even more perturbations compared to the pathogenic-variants (segments 74-79, 50-58 and 21-26, Fig. 8.2C).

The differences at the 50-58 segment are similar to the ones shown at the same segment in the pathogenic L55P TTR variant, but supplementary perturbations are observed at the 21-26 segment, which is spatially close to the 50-58 segment in the protein structure (Fig. 8.2C and Fig. 8.2B), suggesting that the perturbation in the 21-26 segment somehow compensates the perturbation of the 50-58 segment, preventing T119M TTR from amyloidosis. We calculated the number of atomic interaction between these two segments and found more in the T119M mutant compared to the WT TTR, supporting

the compensatory mechanism.

T119M is known to prevent amyloidogenesis when present together with the V30M mutation in heterozygous individuals [205] and to protect the V30M TTR tetramer from dissociation in vitro when some V30M chains are replaced by T119M chains [206], implying the existence of a rescue mechanism [12] by which the dynamics perturbation caused by the V30M mutation on one chain is corrected by a dynamics perturbation caused by the T119M mutation on another chain. A dynamic rescue mechanism requires changes in atomic interactions, consistent with the large perturbation caused by the T119M mutation despite its non-pathogenicity.

The depletion of the secondary structure contribution of the atomic interactions in the 74-79 segment ($\Delta f_{w_{2D}} < 0$ and shortening of the 74-82 α -helix compared the WT TTR, Fig. 8.2B and Fig. 8.2C) may be a key to the rescue mechanism: the 75-90 segment has been shown to initiate unfolding of WT TTR under acidic conditions [207], thus the weakening of the secondary structure of the 74-82 α -helix may lead the TTR tetramer carrying the V30M and T119M mutations on different chains to follow the same unfolding pathway of WT TTR, instead of the pathogenic unfolding pathway of V30M TTR.

The T119M TTR case highlights the necessity of a qualitative analysis of perturbations on top of their quantification to discriminate pathogenic from non-pathogenic mutations.

The differences in local structural-level allocation of the atomic interactions for the TTR variants compared to the WT show the sensitivity of this measure to single amino acid mutations to track the dynamic perturbation underlying robustness (T119Y variant), fragility (L55P and V30M variants) and rescue mechanism (T119M variant). Importantly, the rescue mechanism induced by the T119M mutation may be mimicked in the design of allosteric modulators that would inhibit amyloidosis in patients carrying the V30M TTR mutation.

The analysis of the B-subunit pentamers of AB₅ toxins and of TTR variants has proven that the comparison of local structural-level allocation of atomic interactions is a good measure of differences in protein folding dynamics. To further explore the issue of drug design strategy, we have applied our methodology to the comparison of the main proteases (Mpro's) of two coronaviruses, SARS-CoV and SARS-CoV-2, to determine whether they have different large-scale dynamics, questioning on the efficiency of using common drugs to treat both diseases, as recently proposed [195–197], rather than strain-specialized drugs.

8.2.3 Main proteases of SARS coronaviruses

The Mpro's of SARS-CoV and SARS-CoV-2 have very similar sequence (12 mutations over 305 residues, corresponding to 96% identity, Fig. 8.3B) and similar crystalline structure (RMSD = 0.71 Å for one Mpro chain, Fig. 8.3B). The percentage of sequence identity between the Mpro of SARS-CoV and SARS-CoV-2 is intermediate with respect to the previous case studies (82% and >99% of sequence identity for the B-pentamers of the

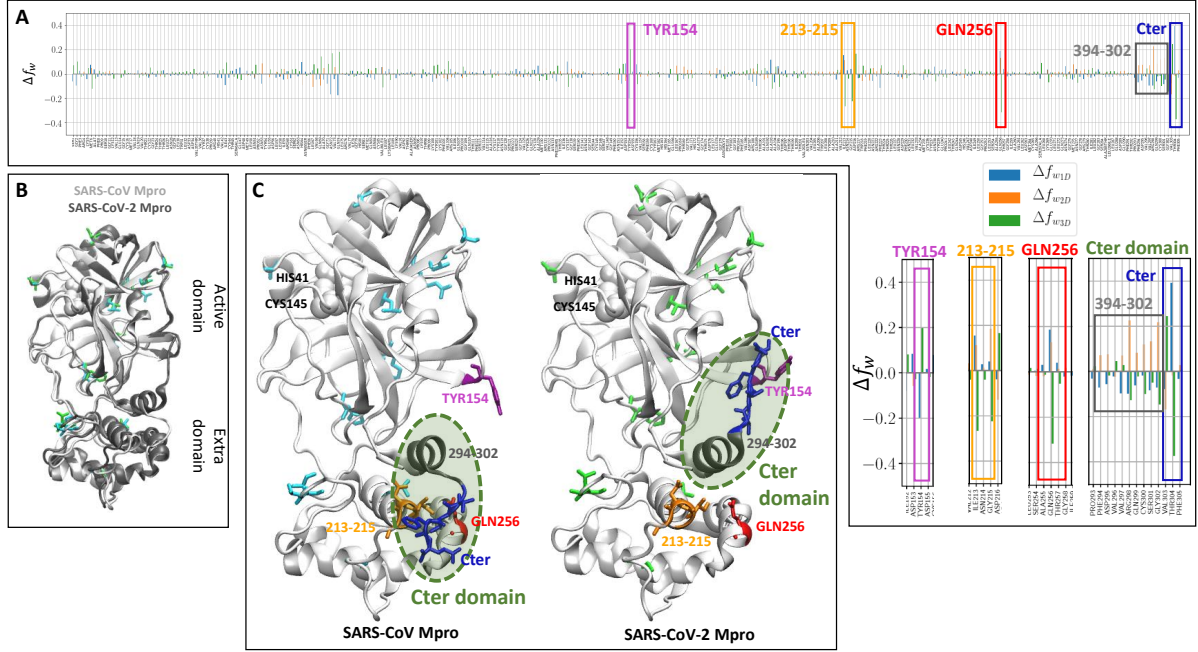


Figure 8.3: Main proteases (Mpro) of coronaviruses SARS-CoV and SARS-CoV-2. A: changes of local structural-level allocation of atomic interactions along the Mpro monomer, where $\Delta f_{w1D,i} = f_{w1D,i}^{SARS-CoV-2} - f_{w1D,i}^{SARS-CoV}$ (and similarly for $\Delta f_{w2D,i}$ and $\Delta f_{w3D,i}$). B: Superposition of the X-ray structures of the Mpro monomer of SARS-CoV (PDB: 2H2Z) and SARS-CoV-2 (PDB: 6Y2E). The 12 mutated positions are highlighted in one chain (cyan amino acids for SARS-CoV and green amino acids for SARS-CoV-2). C: Details of the X-ray structures of SARS-CoV Mpro and SARS-CoV-2 Mpro. The segments involved in changes in local structural-level allocation of atomic interactions are highlighted with different colors and the active sites (HIS 41 and CYS145) are represented in space-fill.

AB₅ toxins and for the TTR variants, respectively) and again the difference in crystalline structures is very low (RMSD < 1 Å). We can thus assume that differences in local structural-level allocation of atomic interactions will mostly measure differences in protein dynamics. The main differences in atomic interactions allocation are located at the C-terminal domain (Fig. 8.3A), where a higher fraction of atomic interactions is devolved to 2D and a lower fraction to 3D ($\Delta f_{w2D} > 0$, $\Delta f_{w3D} < 0$ and longer α -helix, Fig. 8.3C) in SARS-CoV-2 Mpro compared to SARS-CoV Mpro. Moreover, the C-terminal coil points towards the active domain in SARS-CoV-2 Mpro and towards the rest of the extra domain in SARS-CoV Mpro (Fig. 8.3C). This is reflected by $\Delta f_{w3D} < 0$ at residues 213 to 215 and GLN256 of the extra domain and $\Delta f_{w3D} > 0$ at residue TYR154 of the active domain (Fig. 8.3A and 8.3C).

Much effort is devolved to the design of drugs to inhibit the Mpro's of a broad spectrum of coronaviruses [195–197, 208], but our results suggest differences in the enzyme dynamics, particularly of the C-terminal dynamics, a region whose role in the enzymatic activity has been shown experimentally for SARS-CoV [209]. This result supported by the other cases of study advises against a broad-spectrum drug strategy due to the large-scale dynamics differences that are likely to prevent same-drug recognition. It therefore counsels on more strain-customized drugs to inhibit the function of the Mpro of coronaviruses.

8.3 Conclusion

From the presented results, we can deduce that the allocation of the atomic interactions into structural levels contains information on the protein dynamics. This could open-up novel directions towards the decoding of a protein dynamics from the static information contained in its crystalline structure.

Moreover, the comparison of the allocation of the atomic interactions into structural levels is a fast and computationally-light tool for the diagnostic of neutral, functionally-sensitive and rescue mutations in terms of their impact on the multi-scale protein dynamics. This information will allow digging into the molecular mechanisms underlying pathogenic mutations and error-correcting mutations, assisting drug design and the development of personalized therapy.

The next chapter analyzes more in details the perturbation in atomic interactions caused by the single mutations of Transthyretin and its impact on the protein dynamics, in turn causing large-scale conformational changes (amyloidosis). Then, Chapter 10 studies the dynamical differences between CTxB₅ and LTB₅ in details, using an integrative approach combining experimental measure with network modeling.

Chapter 9

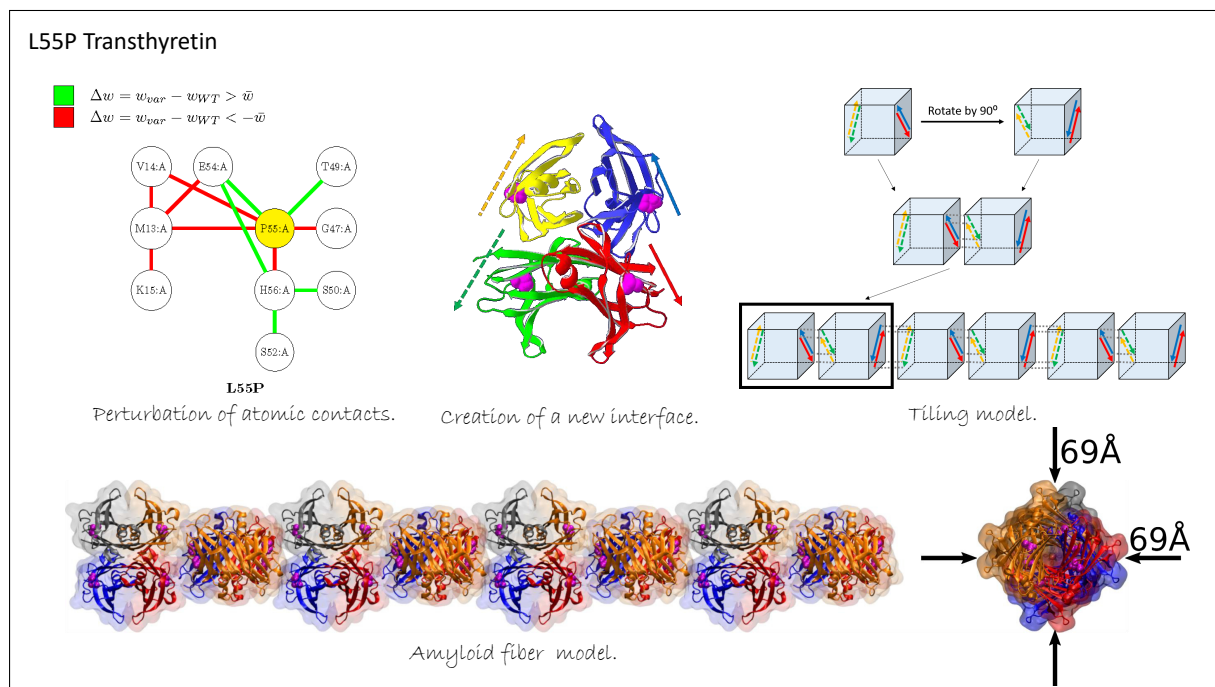
Case study: Transthyretin. Large scale changes in dynamics upon mutation leading to amyloid fiber formation.

Highlights:

- Atomic contacts in the structure of protein variants are compared using Induced Perturbation Networks.
- Pathogenic- and non-pathogenic Transthyretin variants have different Induced Perturbation Networks.
- Misfolding requires dynamical differences through changes in atomic contacts.
- A tiling model for the L55P Transthyretin amyloid fiber is proposed.

Abstract: Atomic contacts in the structures of pathogenic- and non-pathogenic Transthyretin variants, having different tendency to misfolding, are compared to the wild-type structure using the Induced Perturbation Network tool. The Induced Perturbation Network of the pathogenic L55P variants is used to infer a model of its amyloid fiber, consistent with experimental evidence and geometrical constraints.

Graphical abstract:



Methods: Perturbation Network and Induced Perturbation Network (Section 3.1.3).

Publication: Pacini L., Vuillon L. and Lesieur C. Induced Perturbation Network and tiling for modeling the L55P Transthyretin amyloid fiber. *Procedia Computer Science*, 2020, <https://doi.org/10.1016/j.procs.2020.11.002>

9.1 Introduction

In Chapter 8 the differences in local structural-level allocation of atomic interactions for Transthyretin (TTR) variants were related to dynamical differences. In this chapter, the pathogenic L55P TTR variant is studied more in detail. Using a tool called the Induced Perturbation Network (IPN), the detailed differences in atomic interactions between the wild-type (WT) and variant TTR structures are analyzed and related to dynamical differences. The method is validated by the comparison of the IPNs of the same TTR variants studied in Chapter 8: two pathogenic (L55P and V30M) and two non-pathogenic (T119Y and T119M). Then, a model of the L55P TTR amyloid fiber is proposed using the dynamical information extracted from the IPN and the application of a tiling approach.

9.2 Induced Perturbation Network

The amino acids interactions within a protein variant and the wild-type (WT) are compared through the analysis of the Perturbation Network (PN) [111] and Induced Perturbation Network (IPN) of threshold \bar{w} . The details of the construction of the PN and IPN are provided in Chapter 3, Section 3.1.3.

In brief, the PN is a network where links represent differences in link weight (number of atomic interactions) in the AANs of the WT- and variant-structure. The \bar{w} threshold determines the minimal difference in atomic interactions for a link to be in the PN. The IPN is defined for single amino-acid variants as the connected component of the PN that contains the mutation site. The IPN represents the network of changes in atomic contacts in the protein structure that have been caused in the protein structure by the single mutation. The links in the PN and IPN are colored in red if the number of atomic interactions has decreased and in green if the number of atomic interactions has increased after mutation.

In here, $\bar{w} = 4$ is employed, providing a compromise that allows showing perturbations while maintaining the size of the IPN network small enough to exhibit fewer but stronger differences between the variants. A threshold $\bar{w} = 4$ (at least four atomic interactions gained or lost for a link to be in the PN) is reasonably conservative, considering that the average link weight in the WT TTR AAN is $\langle w_{i,j} \rangle = 12.4$ with $\text{STD}(w_{i,j}) = 10.3$, over a total of 1175 links.

It is important to underline that while both the comparison of the local structural-level allocation of atomic interactions and the PN (or IPN) are used to measure link-weights differences leading to dynamical difference, the two measures provide different information. The comparison of the local structural-level allocation of atomic interactions gives information on what structural levels are impacted at all amino-acid positions by the mutation, while the PN provides information on which atomic interactions have changed.

To clarify the difference between the two measures, Fig. 9.1 shows two simplified examples of structural changes upon mutation of a residue i . Please note that protein structures reported in the examples to do correspond exactly to the measures proposed, but are used for illustration purpose only.

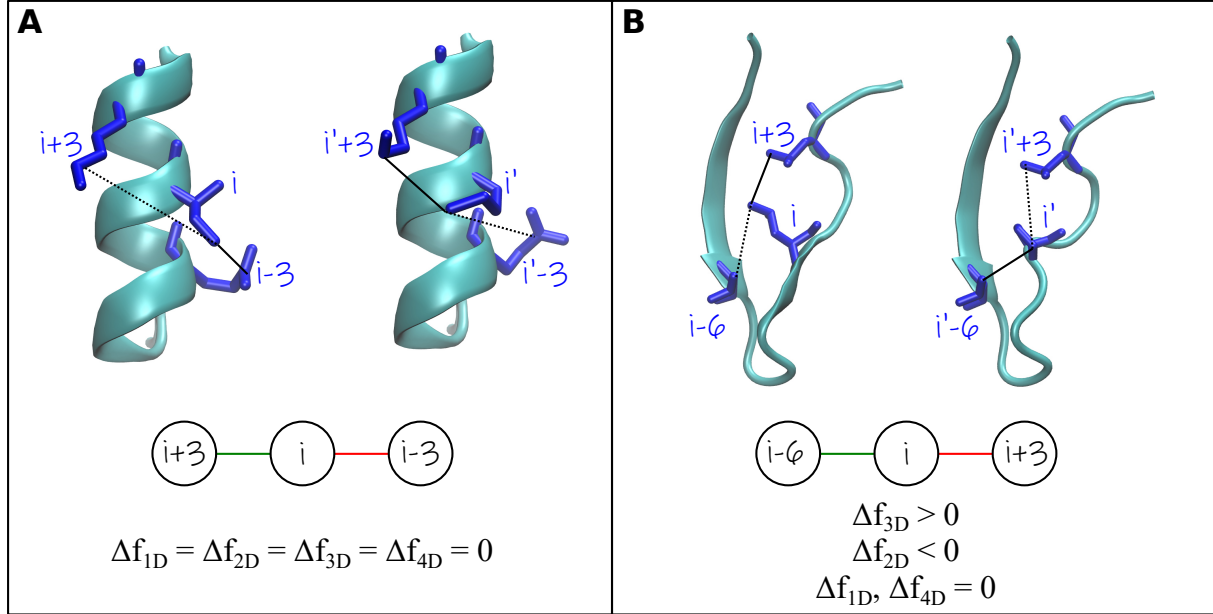


Figure 9.1: Comparison between the Induced Perturbation Network and local structural-level allocation of atomic interactions measure. In the Induced Perturbation Network, a red link correspond to a decrease in number of atomic interactions and a green link corresponds to an increase in number of atomic interactions in the mutant. A: the mutation of residue i causes the increase of the number of atomic interactions with residue $i + 3$ and the decrease of the number of atomic interactions with residue $i - 3$ of the same quantity. Both residues $i + 3$ and $i - 3$ are 2D-neighbors of i , so no changes in local structural-level allocation of atomic interactions are observed. B: the mutation of residue i causes the increase of the number of atomic interactions with residue $i - 6$ (3D-neighbor) and the decrease of the number of atomic interactions with residue $i + 3$ (2D-neighbor) of the same quantity. An increase in the allocation of atomic interactions to the 3D structural level and a decrease in the allocation of atomic interactions to the 2D structural level are observed. Please note that protein structures examples to do correspond exactly to the measures proposed, but are used for illustration purpose only.

In Fig. 9.1A, the mutation of residue i causes the increase of the number of atomic interactions with residue $i + 3$ and the decrease of the number of atomic interactions with residue $i - 3$. For simplicity, we assume that the number of atomic interactions gained with residue $i + 3$ is equal to the number of interactions lost with residue $i - 3$ and that no other atomic interaction changes happen. Both residues $i + 3$ and $i - 3$ are 2D-neighbors of i , so no changes in local structural-level allocation of atomic interactions are observed. In Fig. 9.1B, the mutation of residue i causes the increase of the number of atomic interactions with residue $i - 6$ (3D-neighbor) and the decrease of the number of atomic interactions with residue $i + 3$ (2D-neighbor). Again, we assume that the number of atomic interactions gained with residue $i - 6$ is equal to the number of interactions lost with residue $i + 3$ and that no other atomic interaction changes happen. An increase in the allocation of atomic interactions to the 3D structural level and a decrease in the

allocation of atomic interactions to the 2D structural level are observed. The IPNs do not allow the discrimination between the two cases, for which the comparison of the local structural-level allocation of atomic interactions is needed, but the IPN measure is necessary to trace back the exact changes in atomic interactions underlying changes in allocation.

In Section 9.4.2, the IPN of the L55P TTR variant is used to monitor the atomic interactions changes that lead to the observed differences in local structural-level allocation of atomic interactions (Chapter 8, Section 8.2.2).

9.3 Tiling model of fiber formation

The amyloid fiber structure of a protein can be seen as a result of a tiling process where a repeating unit interacts with N copies of itself using one or more interfaces. The position of the interface(s) in the repeating unit will determine the growth direction of the amyloid. As a result, no matter the exact molecular mechanisms creating the repeating unit, the process of amyloid formation can be seen as a transition from the initial oligomeric state according to the following steps [210]: (i) Rearrangement of the native oligomeric structure (with or without dissociation of the oligomer); (ii) Exposure of a novel interacting interface; (iii) Eventually, formation of an asymmetric repeating unit through the interaction via the novel interface; (iv) Growth of the amyloid by successive interactions of repeating units through the novel interacting interface.

The candidate interface is obtained from the IPN. Then, depending on the position of the candidate interface in the protein monomeric chain and on the relative position of the chains that build the quaternary structure of the protein, it is possible to understand whether the repetition of identical oligomers or identical sets of oligomers (repeating units), interacting through the candidate interface, allows the growth of a fiber without steric clashes. Figure 9.2 shows the schematics of two possible cases: on the left-hand side, a case where the repetition of identical oligomers leads to a fiber; on the right-hand side, a case where the repetition of identical oligomers leads to a steric clash. In the first case, the validity of the fiber model can be assessed by comparison of its diameter with some experimental evidence. In the second case, two explanations are possible: either the selected candidate interface is not the right choice, or the selected candidate interface is correct but the oligomer dissociates before the initiation of the fiber-forming process and no fiber model can be produced based on the sole knowledge of the oligomer's crystalline structure.

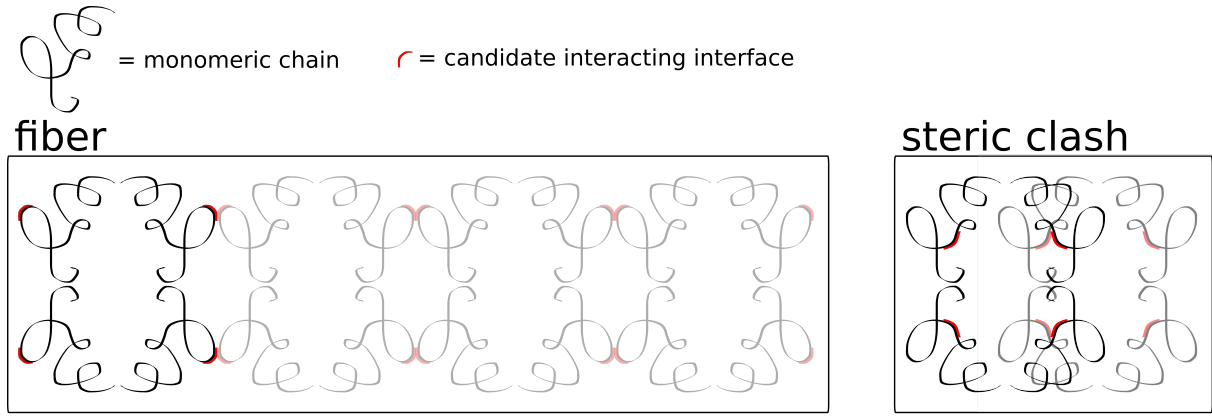


Figure 9.2: Schematics of two possible cases of position of the candidate interacting interface in the protein oligomer. Left: a case where the repetition of identical oligomers leads to a fiber. Right: a case where the repetition of identical oligomers leads to a steric clash.

9.4 L55P Transthyretin amyloid fiber model

The cause of the increased tendency to amyloid fiber formation of the L55P TTR variant is investigated according to the following hypotheses:

1. A mutation allows the reproduction of the native structure if the AANs of the variant and of the WT have similar node properties (node degree k_i , node weight w_i and node's Neighborhood watch Nw_i) for all nodes i [23] (Chapter 4).
2. The dynamics of atoms in the variant are similar to the dynamics of atoms in the WT protein if the link weights w_{ij} are similar in the AANs of the variant and of the WT (Chapter 5 and Chapter 8).

9.4.1 Structural comparison of Transthyretin variants from node properties in the AAN

The first hypothesis is verified by comparing the node properties of the AANs of WT and L55P TTR. For comparison, the same calculation is performed for another pathogenic (V30M) and two non-pathogenic (T119Y and T119M) variant structures. Figure 9.3 shows that the AANs of the WT, L55P, V30M, T119Y and T119M TTR variants share very similar node properties along the sequence. This means that all the variants are capable of reproducing the TTR native structure, and that the pathogenicity of the L55P and V30M variants is related to their dynamics, different from the WT TTR. This is consistent with a low Root Mean Square Deviation between the variants' structure and the WT structure (Chapter 8, Section 8.2.2).

It must be noted that the RMSD measure is performed considering only the position of the backbone atoms of the protein, while the comparison of the node degree k_i , node weight w_i and node's Neighborhood watch Nw_i for each amino acid in the AAN of the protein variant takes into consideration all atoms in the protein's structure (excluding hydrogen atoms since they are not detected by X-ray crystallography). As such, a similarity in

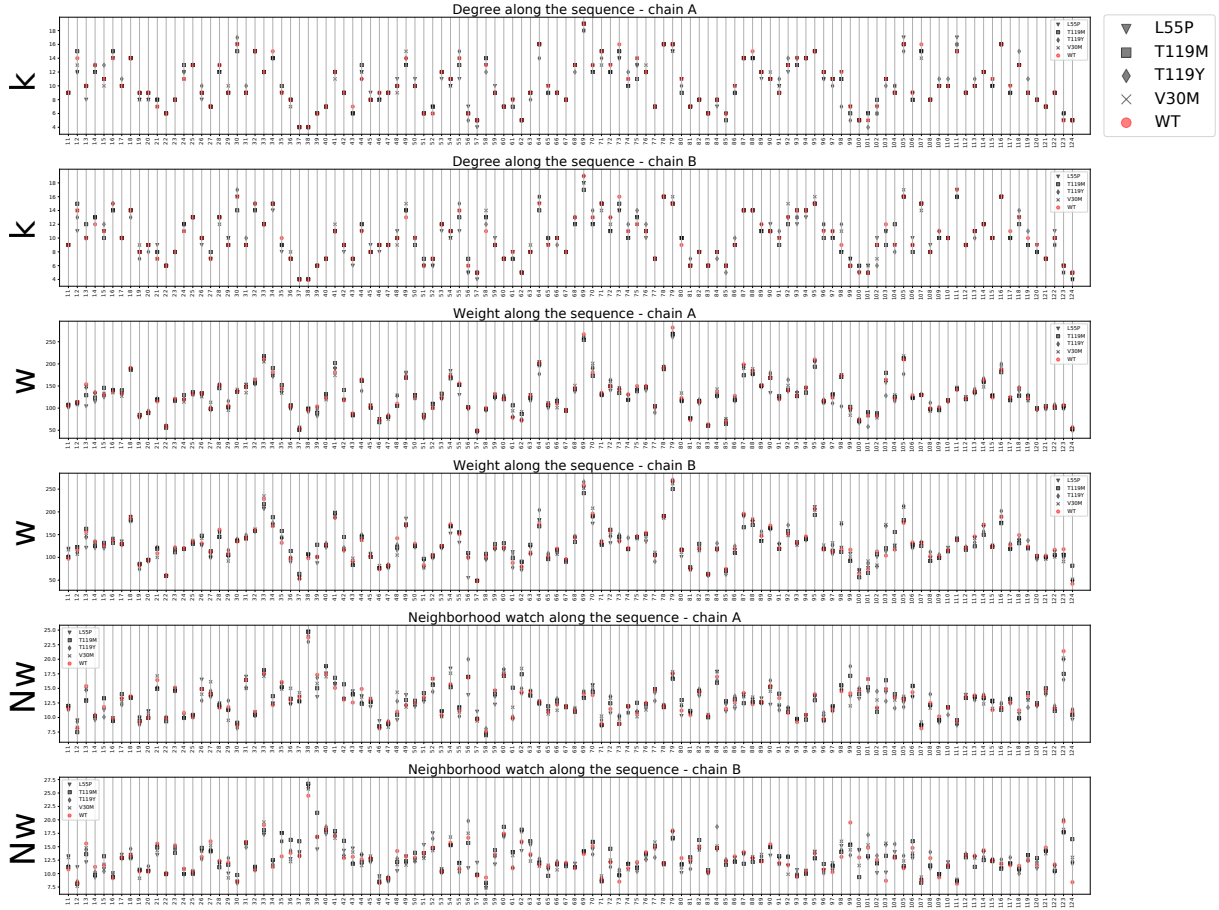


Figure 9.3: Node properties (degree k , weight w and Neighborhood watch $Nw = w/k$) of the amino acids of the TTR variants in their Amino Acid Networks, along the protein amino acid sequence. From top to bottom: degree k (chain A), degree k (chain B), weight w (chain A), weight w (chain B), Neighborhood watch $Nw = w/k$ (chain A), Neighborhood watch $Nw = w/k$ (chain B).

node properties in the AAN is a stricter criterion to assess protein structures similarity. While a low value of RMSD means that the shape of the protein is conserved among two variants, all nodes having the same degree, weight and Neighborhood watch in the AAN of two protein variants means that the number of contacts, that is, the amino acid and atomic packing around each amino acid is conserved. Nevertheless, an amino acid may have the same degree, weight and Neighborhood watch in the AAN of two protein variants but performing a different number of atomic contacts with its neighbors, resulting in a different link weight, if it moves closer to some neighbor(s) and further from other(s) (Chapter 8).

9.4.2 Dynamical comparison of Transthyretin variants from link weights in the AAN

To test the second hypothesis, IPNs are employed to compare the link weights in the AAN of each variant with the ones of the AAN of the WT. We used IPNs instead of PNs to focus on the atomic motions and rearrangement of atomic contacts resulting from the mutation

of the amino acid at the mutation site. The IPNs of the L55P, V30M, T119Y and T119M TTR variants are reported in Figure 9.4. Green links between nodes in the IPN mean that the corresponding amino acids are closer in the variant's structure compared to the WT (more atomic interactions); on the contrary, red links in the IPN mean that the two amino acids are further in the variant's structure compared to the WT.

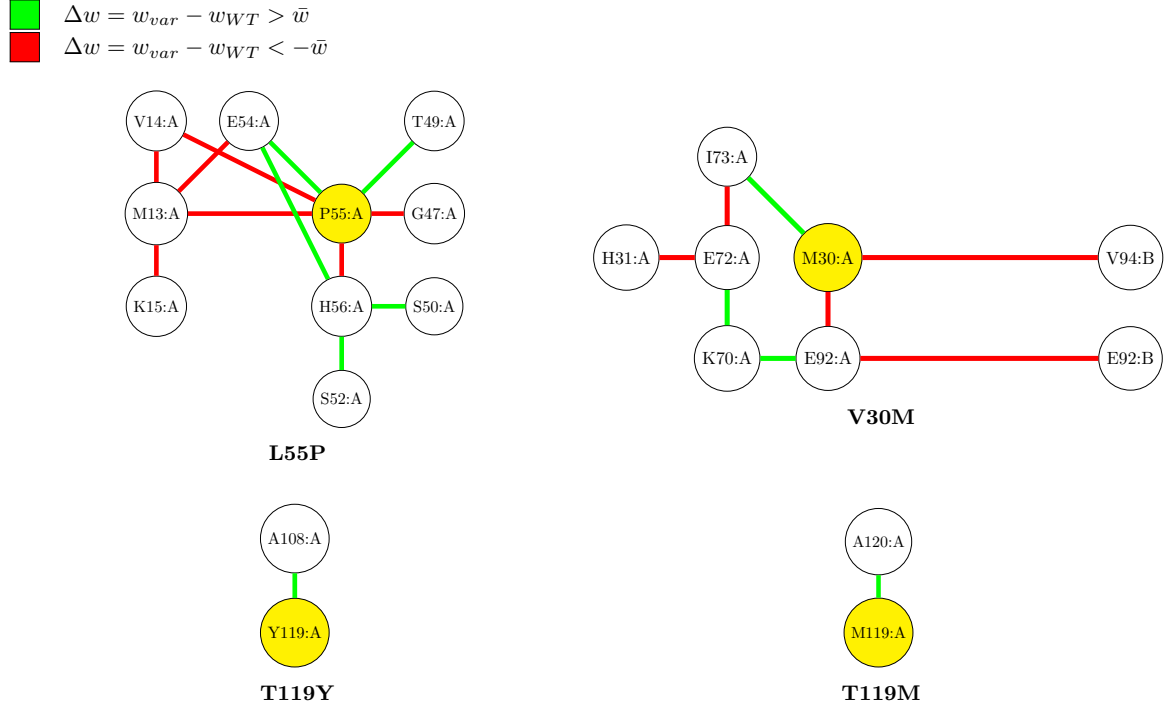


Figure 9.4: Induced Perturbation Networks (threshold $\bar{w} = 4$) for four TTR variants. Green links between nodes in the IPN mean that the corresponding amino acids are closer in the variant's structure compared to the WT (more atomic interactions); viceversa, red links in the IPN mean that the two amino acids are further in the variant's structure compared to the WT.

The IPN of L55P (Figure 9.4, top left) shows the presence of both red and green links, showing that a complex rearrangement of atomic positions, leading to atomic interaction changes, has taken place due to the mutation with respect to the native structure. In contrast, the IPNs of the non-pathogenic variants T119Y and T119M (Figure 9.4, bottom) have only one green link. It should be noted that the analysis of the local structural-level allocation of atomic interactions (Chapter 8, Section 8.2.2) has shown differences between the T119Y and T119M cases, with the T119M TTR having substantial dynamical differences with respect to the WT TTR. Thus, either a lower threshold \bar{w} would be needed to trace the changes in atomic interactions causing the dynamical changes due to the T119M mutation, or the atomic interaction changes are disconnected from the mutation site and the PN should be employed. Finally, the IPN of another pathogenic variant (V30M, Figure 9.4, top right) shows the presence of both red and green links (as the L55P IPN, contrarily to the non-pathogenic variants' IPNs), different from the IPN of L55P. These results show how the IPNs of pathogenic variants have different IPNs

compared to non-pathogenic variants. Moreover, the IPN allows recognizing differences among pathogenic variants (L55P and V30M), consistent with different fiber formation mechanisms for V30M and L55P TTR.

The IPN of L55P TTR (Figure 9.4, top left) shows that the loop containing residues S52 to E54 (red in Figure 9.5) loses interactions with the β -strand containing residues M13 to K15 (blue in Figure 9.5) and changes orientation angle with respect to the β -strand containing residues G47 to T49 (purple in Figure 9.5) since some atomic interactions are lost (red P55-G47 link) and some are gained (green P55-T49 link). Based on these observations, we make the hypothesis that the E54-T60 loop moves away from the rest of the structure, resulting from the loosening of atomic interactions with the M13, V14, K15 residues, and forms the new interacting interface that allows the fiber to grow (Figure 9.5).

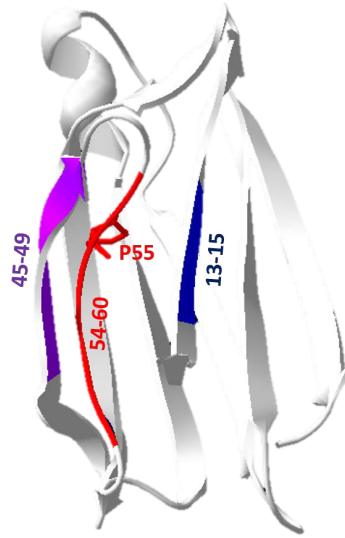


Figure 9.5: TTR L55P variant: position of the candidate new interacting interface (in red). For clarity, only one chain is shown.

The second pathogenic variant considered here, V30M, shows a different IPN compared to the IPN of the L55P TTR variant, suggesting that the pathways of fiber formation of these two variants are different. Contrary to the L55P mutation, the IPN of V30M (Figure 9.4, top-right panel) shows how the perturbation due to the single amino acid mutation in a chain of the TTR dimer reaches the second chain of the dimer: the nodes V94:B and E92:B are present in the IPN of V30M. Moreover, the links (E92:A, V94:B) and (E92:A, E92:B) that involve these nodes are red ($w_{V30M}(i, j) - w_{WT}(i, j) < -\bar{w}$), reflecting a loss of atomic interactions at the dimer interface due to the V30M mutation. This suggests that the V30M TTR tetramer probably dissociates to form the amyloid fiber, in agreement with the recent experimental analysis of the fibrils from a patient with V30M TTR amyloidosis [191].

The possibility for a specific segment to form the novel interface for the amyloid fiber

growth depends on two factors: the spatial occupancy of the protein structure and the chemical affinity between two copies of the segment. In the following, we focus on the first aspect, to determine whether the candidate interface for the L55P TTR inferred from its IPN satisfies the first necessary criterion, that is, allowing the growth of a fiber-like structure, without producing steric clashes.

9.4.3 Tiling model of the L55P TTR amyloid fiber

In order to understand whether the new candidate interface can allow the growth of an amyloid-like structure of repeating units of L55P TTR oligomers, the symmetry of the TTR tetramer is taken into consideration. The native TTR tetramer has a dihedral symmetry (Figure 9.6, left panel); as a consequence, the interacting interface on the four chains will follow the same symmetry if no dissociation of the tetramer happens (arrows in Figure 9.6, right panel).

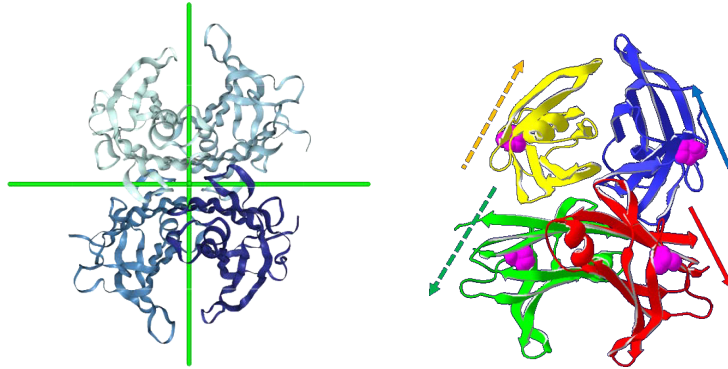


Figure 9.6: Tetrameric structure of TTR. Left: dihedral symmetry axes of the tetramer. Right: position of the P55 residue in the L55P variant (in pink) and relative orientation of the candidate interacting interfaces on each chain (arrows).

As schematized in Figure 9.7a, a tetramer with the interfaces following a dihedral symmetry will be able to interact with a copy of itself after a rotation of 90 degrees. Then, the obtained octamer will constitute a repeating unit for a tiling model, eventually resulting in an elongated fiber-like structure. In the following, we refer to a so-constructed fiber model as a *dihedral fiber model*. Alternatively, provided a sufficient mobility of the interfaces, they could align along a symmetry axis of the tetramer and degenerate to a central symmetry (Figure 9.7b, top). In this situation, the tetramer itself would represent the repeating unit for a tiling model, as schematized in Figure 9.7b (bottom). In the following, we refer to a so-constructed fiber model as a *central fiber model*.

Figure 9.8 shows the *dihedral* and *central* fiber models for L55P TTR. Due to the geometry of the TTR oligomer, the *dihedral* model produces a fiber of diameter of around 69 Å, while the *central* model produces a fiber with an ellipsoidal cross-section, with a major axis of around 69 Å and a minor axis of around 42 Å. Interestingly, a crystallographic study of L55P TTR (PDB id: 5TTR [200]) showed that the protein crystallized

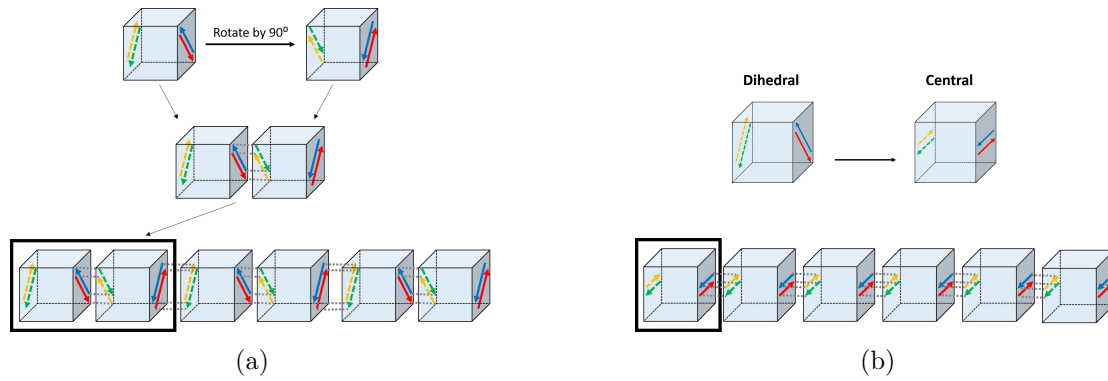


Figure 9.7: Model for fibers resulting from the creation of an active interface (arrows) on each chain of a tetrameric protein with dihedral symmetry. (a) *Dihedral fiber model*. A tetramer with dihedral symmetry can interact with a copy of itself rotated by 90 degrees through the interaction interfaces (arrows). The obtained octamer (in the box) constitutes the repeating unit for a fiber model. (b) *Central fiber model*. If the interaction interfaces (arrows) of a tetramer with central symmetry are mobile enough, they can align with an axis of symmetry of the tetramer, degenerating to a central symmetry. The degenerated tetramer (in the box) constitutes the repeating unit for a fiber model.

with the same symmetry as the the repeating unit of the *dihedral fiber model* proposed here. The observation of an asymmetric unit of this symmetry lead the authors of the study to propose that L55P exists in an amyloidogenic conformation, that would explain its higher propensity to the formation of amyloids compared to the WT. The *dihedral fiber model* perfectly matches the model proposed in reference [200], even though it was produced from a different crystalline structure and using different tools. Finally, electron micrographs of L55P TTR protofilaments produced *in vitro* show a diameter of 60 to 65 Å [202], consistent with the *dihedral fiber model*. The mentioned experimental evidences support the *dihedral fiber model* as a model for the amyloid fiber formed by the L55P TTR variant.

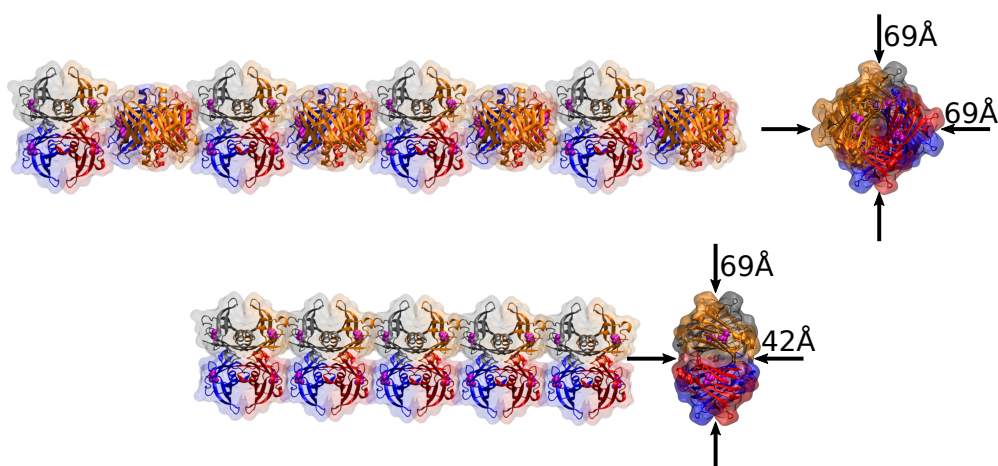


Figure 9.8: Fiber models for the TTR L55P variant. Top: *Dihedral fiber model*. Bottom: *Central fiber model*.

9.5 Conclusion

The analysis of the IPNs of TTR variants has shown how the IPN probes atomic motions and atomic links rearrangement that are consistent with the dynamics differences measured experimentally and is complementary to the comparison of the local structural-level allocation of atomic contacts. It is noticeable that few changes in atomic interactions (twelve links in the L55P TTR IPN over 228 amino acids in the TTR dimer and 1175 in the WT TTR dimer AAN) can lead to a drastic change in the protein's dynamics.

The tiling model proposed in this chapter takes into account for the geometrical constraints underlying fiber formation, necessary for the validation of the candidate novel interacting interfaces extracted from the analysis of the IPN. The coupling of the IPN and the tiling model has allowed the construction of a fiber model for the L55P TTR variant, that satisfies the geometrical constraints imposed by the symmetry of the TTR tetramer structure and is in agreement with experimental results. The proposed model does not require the dissociation of the tetramer prior to fiber formation. The methodology used in this study relies on the sole knowledge of the X-ray structure of protein variants and takes into consideration the space occupancy of whole protein structure for the construction of fiber models. This is necessary to ensure that the proposed novel interacting interface does not produce steric clashes when the aggregation of entire proteins is modeled.

Chapter 10

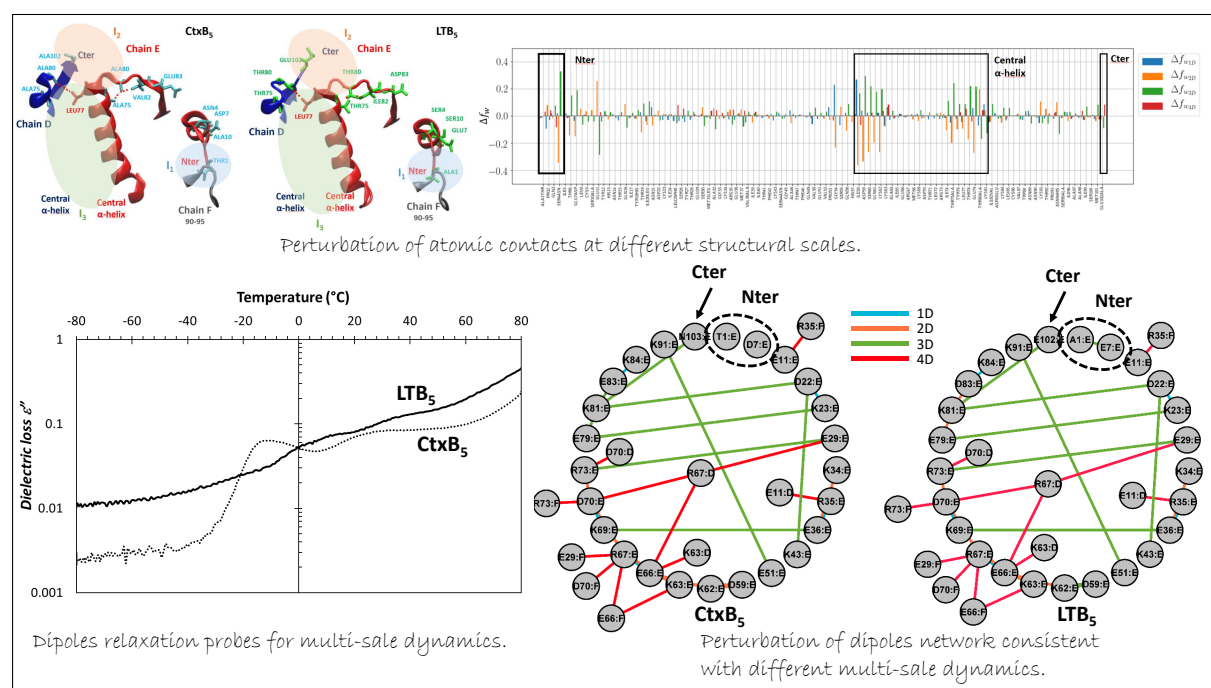
Case study: B-subunits toxin pentamers. Integrative approaches to protein dynamics perturbations related to mutations.

Highlights:

- Broadband Dielectric Spectroscopy measures protein multi-scale dynamics from dipoles relaxation.
- Dipoles involved in dielectric signal are modeled by Electrostatic Networks.
- The dynamical impact of mutations due to perturbations of different structural levels reflects into different experimental dynamics.
- The experimental dynamics differences are recovered by perturbations of the Electrostatic Network.

Abstract: Perturbation of the allocation of atomic contacts to structural levels caused by mutation results in different multi-scale dynamics. Multi-scale dynamics have been measured experimentally using Broadband Dielectric Spectroscopy. Differences in dielectric signal for protein variants measure dynamical perturbations. Electrostatic Networks are used to model dipoles involved in the dielectric signal and the experimental dynamics differences are recovered by perturbations of the Electrostatic Network.

Graphical abstract:



Methods: Electrostatic Network (Section 3.1.4).

Publications: Pacini L., Bourgeat L., Serghei A., and Lesieur C. Analysis of nanoconfined protein dielectric signals using charged amino acid network models. *Australian Journal of Chemistry*, 2020, DOI:10.1071/CH19502.

Bourgeat L., Pacini L., Serghei A., and Lesieur C. Experimental diagnostic of sequence-variant dynamic perturbations revealed by broadband dielectric spectroscopy. *Structure*, 2021, 10.1016/j.str.2021.05.005.

10.1 Introduction

In Chapter 8 Section 8.2.1 the differences in local structural-level allocation of atomic interactions between the B-subunits of CtxB₅ and LTB₅ were associated to differences in their experimental folding and unfolding dynamics. The experimental studies on which the discussion of Section 8.2.1 was developed are based on global measures of the protein folding and association kinetics.

In fact, protein dynamics involve multiple spatiotemporal scales, ranging in time-scale from femtoseconds (10^{-15} s, or 10^{15} Hz) to seconds (or 1 Hz) and in spatial scale from Angströms (10^{-10} m) to nanometers (10^{-9} m). Experimentally, measuring dynamics covering such a broad range of scales and at amino-acid level resolution is challenging.

In her PhD thesis, Laëtitia Bourgeat proposed Broadband Dielectric Spectroscopy (BDS) coupled with nanoconfinement as a novel experimental technique to measure protein dynamics at multiple scales (1 Hz to 10^6 Hz).

In this chapter, an integrative approach combining the experimental measure of protein dynamics with network models is proposed for the diagnostic of dynamics perturbations of the B-subunit of AB₅ toxins caused by mutations with high resolution. The results presented here are published in [27] and [28], that I have co-authored.

Even though I have not participated the experimental part of the study but only to the network modeling, I provide a summary of the experimental techniques and protocol in the next session, necessary to establish the relation with the network measures. The details can be found in the Laëtitia Bourgeat's PhD manuscript [25] and main publication [26].

10.2 Experimental measure of protein multi-scale dynamics

10.2.1 Nanoconfinement of proteins

One of the challenges of measuring protein dynamics and assigning the observed dynamics to specific atomic motions in the protein structure comes from the fact that proteins populate conformational ensembles (Chapter 2, Section 2.2). As an example, let us assume that an experimental measure allows detecting a molecular motion at a given frequency f in a protein sample and the goal is to determine which amino acids participate to the motion. If the protein structure is known, one could use physico-chemical considerations to try and assign the motion to specific sets of atomic interactions, e.g. based on the number and nature of atomic interactions existing in the protein structure. This procedure, that represents a sort of reverse-engineering of Molecular Dynamics simulations, is unfeasible with the current knowledge on protein dynamics, but is theoretically sound. On the contrary, if the protein sample contains several protein conformations, then the assignment of observed experimental dynamics to specific amino acids in the protein structure is

hopeless: indeed, it would be impossible to determine which protein conformation(s) participate to the signal. In the latter case, it would be impossible to “reverse-engineering the Molecular Dynamics”, because the input conformation is unknown. To circumvent this issue, it is necessary to reduce the protein’s conformational heterogeneity.

To reduce the conformational heterogeneity of the protein sample, the proteins are confined in nanopores of diameter equal to 40 nm and length equal to 10 μm . Because the dimensions of the nanopores are of the same order of magnitude of the proteins’ characteristic dimensions, few conformations of the protein are expected to be selected. For comparison, the diameter of the CtxB pentamer is around 5 nm. The diameter of the nanopores is around six times the diameter of the toxin; however, it should be noted that jamming of protein molecules at the entrance of the pore during the deposition of the protein solution reduce the effective pore diameter accessible to the other molecules. Experiments at variable nanoconfinement condition gave different results, confirming the role of the nanoconfinement in controlling the protein conformational state (See Laëtitia Bourgeat’s PhD manuscript [25]).

10.2.2 Broadband Dielectric Spectroscopy

Broadband Dielectric Spectroscopy (BDS) measures the dielectric response of a material upon application of an electric field. A textbook describing BDS theory and application in details is [211].

When an electric field $\vec{E}(t)$ is applied to a material, it creates a movement of electric charges. The movement of free electric charges creates a continuous current \vec{J}_f (eq. 10.1):

$$\vec{J}_f = \sigma \vec{E} \quad (10.1)$$

with σ the conductivity of the material. The movement of bound charges creates a displacement current \vec{J}_d (eq. 10.2):

$$\vec{J}_d = \frac{\partial \vec{D}}{\partial t}. \quad (10.2)$$

\vec{D} is the dielectric displacement, given by eq. 10.3

$$\vec{D} = \varepsilon^* \varepsilon_0 \vec{E} \quad (10.3)$$

where ε_0 is the dielectric permittivity of air and $\varepsilon^* = \varepsilon' - j\varepsilon''$ is the complex dielectric permittivity of the material. The polarization \vec{P} of the material is the component of the dielectric displacement caused by the electric field and is given by eq. 10.4

$$\vec{P} = \vec{D} - \vec{D}_0 = (\varepsilon^* - 1)\varepsilon_0 \vec{E}. \quad (10.4)$$

The polarization of the material is given by the sum of three components: electronic, atomic and orientational. In the following, the electronic and atomic polarization are discarded because they are dominant at high frequencies ($> 10^{12}$ Hz), out of the range studied here, and only the orientational polarization is considered. The orientational polarization results from the orientation of permanent dipoles with the direction of the applied electric field, where dipoles are couples of atoms of opposite charge in interaction.

Because of molecular fluctuations, the process of orientation of the dipoles is not instantaneous, but follows a relaxation process called Dielectric Relaxation. If an electric field $\vec{E}(t)$ is applied at time $t = 0$, the dielectric response of the material can be measured as $\varepsilon^*(t)$, the complex dielectric permittivity as a function of time.

In BDS experiments, an oscillating electric field $E(t) = E_0 \exp(j\omega t)$ is applied, and the complex dielectric permittivity ε^* is measured. ε' is the real part of ε^* and is proportional to the energy stored in the system per period. ε'' is the imaginary part of ε^* and is proportional to the energy dissipated per period, and hence is called dielectric loss. Either the real (ε') or the imaginary (ε'') parts of the dielectric permittivity are analysed, because they carry the same information. In here, ε'' is adopted, as usual in the BDS literature.

If $\varepsilon''(t)$ is measured, the Fourier Transform of the signal gives $\varepsilon''(\omega)$, the dependence of the dielectric loss on the frequency. Because the electric field is stationary, the difference of the time dependence between $\vec{E}(t)$ and $\vec{D}(t)$ is a phase shift. Moreover, for a small stationary disturbance (small amplitude of the applied electric field E_0), the Fluctuation Dissipation Theorem applies, meaning that the system reacts in the way it fluctuates, so the frequency of the peak is the frequency of the thermal fluctuations of the dipoles. As a consequence, the frequency of the peak is temperature-dependent.

Importantly, relaxation processes can be distinguished from conduction phenomena in BDS spectra from the shape of the $\varepsilon''(\omega)$ function, with ω the angular frequency of the applied electric field. Conduction phenomena show a decrease in $\varepsilon''(\omega)$ with increasing ω while relaxation processes show a peak in $\varepsilon''(\omega)$ at some angular frequency $\omega = \omega_P$. The position of the peak ω_P provides the relaxation time τ_P as $\tau_P = 2\pi/\omega_P$ or equivalently the characteristic relaxation rate $f_P = 1/\tau_P$ of the fluctuating dipoles that contribute to the relaxation process.

As a consequence, the relaxation times of the dipoles of the sample can be deduced from the position of the peaks of the BDS signal obtained at variable frequencies and fixed temperature.

The relaxation time τ_P of the relaxation process depends on the temperature T . The relation between τ_P and T may follow an Arrhenius-like relation with the temperature (Eq. 10.5 and Fig. 10.1) or a more complex relation (Vogel-Fulcher-Tammann). The Arrhenius-like dependency takes the form

$$\tau \sim \tau_0 \exp(-E_A/k_B T) \quad (10.5)$$

with τ_0 the relaxation time at infinite temperature, E_A the activation energy and k_B the Boltzmann constant. If the relaxation times follow an exact Arrhenius dependence on the temperature, then the dipoles fluctuating at different frequencies are independent from one another.

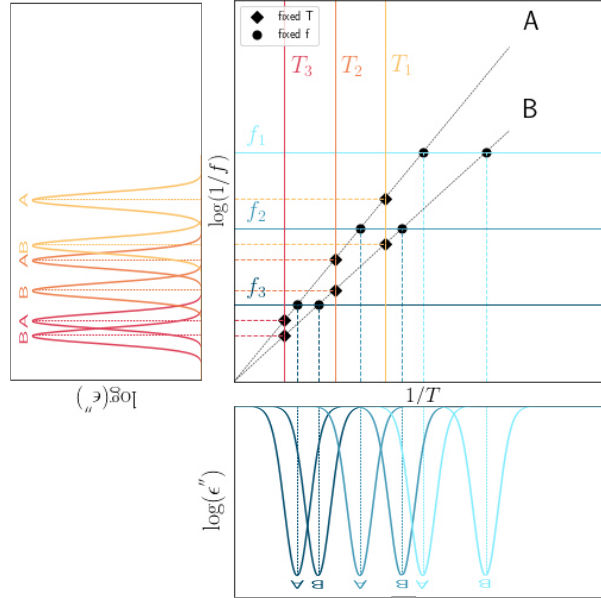


Figure 10.1: Relation between the BDS signal collected at fixed temperature and the BDS signal collected at fixed frequency. The signal is collected at temperatures T_1 , T_2 and T_3 , with $T_1 < T_2 < T_3$, as a function of the frequency f , or at frequencies f_1 , f_2 and f_3 , with $f_1 < f_2 < f_3$, as a function of the temperature T . Two relaxation processes A and B are detected, with A slower than B.

Because frequency and temperature are related (Eq. 10.5), the BDS signal can be obtained at variable temperature and fixed frequency. Specific dipoles relaxation processes correspond to peaks in $\varepsilon''(T)$ curves measured at constant frequency f , with faster relaxation processes (lower τ_P and higher f_P) associated to lower Temperature of maximal dielectric loss θ_{max} (position of the peak). The results presented in this chapter are based on $\varepsilon''(T)$ curves obtained at fixed frequencies.

The relation between the BDS signal collected at fixed temperature and the BDS signal collected at fixed frequency is schematized in Fig. 10.1, for two relaxation processes A and B, with A slower than B (larger relaxation time τ at fixed temperature).

Because peaks in $\varepsilon''(\omega)$ or $\varepsilon''(T)$ correspond to distinct dipoles relaxation processes and the dipoles relaxation time coincides with the thermal fluctuation time of the dipoles, the BDS measure is employed to characterize the molecular dynamics of protein samples. Specifically, proteins having different molecular dynamics are expected to provide different BDS signal.

10.2.3 Experimental protocol

The sample is prepared as follows: a drop of protein solutions is deposited on the nanomembrane. The sample is heated at 50°C for 15 minutes to evaporate the bulk

water and allow proteins to enter the pores and then cooled down to 30°C for 5 minutes (details in [26]). Then the sample is thermally treated at temperature T_t for three hours and then cooled down to -80°C at a rate of 2 K/min. The dielectric measurement is performed during the cooling, providing the $\varepsilon''(T)$ measure.

The following treatment-temperatures T_t are used: 60°C, 80°C, 100°C, 140°C and 180°C. The applied voltage is 0.2 V, at frequencies f of 1Hz, 10 Hz, 100 Hz, 1 kHz, 10 kHz, 100 kHz and 1 MHz.

The use of a broad range of thermal-treatment temperatures allows monitoring the changes in protein dynamics upon thermal denaturation, and the use of a broad range of frequencies allows detecting molecular dynamical processes at different time scales.

10.3 Electrostatic Network models of dipole interactions in the protein structure

The Electrostatic Network (EN) model is used to analyze the network of dipoles in a protein structure. The EN is a sub-graph of the Amino Acid Network (AAN) of a protein structure where only the links connecting charged amino acids of opposite charge are kept. The Intermolecular Electrostatic Network (4D-EN) is a sub-graph of the Electrostatic Network where only the links connecting two amino acids belonging to different chains are kept. Inversely, the Intramolecular Electrostatic Network is a sub-graph of the Electrostatic Network where only the links connecting two amino acids belonging to the same chain are kept. The Induced (Intermolecular) Electrostatic Network is given by the (Intermolecular) Electrostatic Network plus all the first neighbors of the charged nodes in the AAN, with the respective links. For the details, please refer to Chapter 3, Section 3.1.4.

It should be noted that the strength and direction of the dipole moments between the atoms of the protein structure are not calculated. This is because the exact conformational state(s) of the protein in the experimental conditions is not accessible, and thus the Electrostatic Networks are built based on the X-ray structure of the protein. Moreover, the BDS signal is collected after thermal treatment, when the protein is partially denaturated. It would be unreasonable to assume that the atoms participating to the dipole moments have exactly the same position compared to the X-ray structure.

As a consequence, Electrostatic Networks are expected to provide an indication of the main electrostatic interactions expected in the protein sample, at the coarse-grained amino-acid level. As an approximation, the links in the Electrostatic Network are called dipoles and links in the Induced Electrostatic Network are called induced dipoles.

Electrostatic Network models are used to validate the interpretation of the BDS signal of CtxB₅ obtained after thermal treatment (Section 10.4) and to compare the experimental dynamics of CtxB₅ and LTB₅ measured by BDS (Section 10.5).

10.4 Integrative approach to probe multi-scale dynamics of the CtxB₅ toxin

The thermal unfolding dynamics of CtxB₅ was detected using BDS after thermal treatment of the protein [26, 27]. The following section summarizes the main experimental results. Then, Electrostatic Network are used to infer the protein dipoles present in the native toxin pentamer and validate the interpretation of the experimental results.

10.4.1 Experimental measure: Broadband Dielectric Spectroscopy

Fig. 10.2 shows the dielectric loss ε'' as a function of temperature of the CtxB₅ toxin during cooling after 3 hours of thermal treatment at 80°C, at different frequencies. Three distinct peaks are detected by BDS. These peaks are named P_{1C} , P_{2C} and P_{3C} , with the index C standing for CtxB₅. The peaks are ordered based on their position in the temperature axis. Fig. 10.3 shows the dependence of the relaxation time $\tau = 1/f$ on the temperature at maximum signal θ_{max} for the three processes P_{1C} , P_{2C} and P_{3C} for the BDS signal collected after treatment at 80°C. The Arrhenius dependence of $\tau(\theta_{max})$ means that the dipoles participating to the three processes are independent. The relaxation process detected by P_{1C} is faster than the process detected by P_{3C} , in turn faster than process detected by P_{2C} [27].

The results for the other thermal treatment temperatures were published in [26]. It is found that the P_{1C} peak is present after thermal treatment at all temperatures from 60°C to 180°C, the P_{3C} peak is present after thermal treatment at temperatures from 60°C to 140°C, and the P_{2C} peak is present after thermal treatment at temperatures from 80°C to 180°C. Thus the peaks correspond to molecular fluctuations present at different unfolding stages of the toxin.

In macroscopic conditions, thermal denaturation of CtxB₅ is observed at 80°C [212]. Because the treatment temperatures are up to 180°C, the molecular dynamics observed from BDS peaks following different treatment temperature are expected to probe the molecular dynamics of the toxin at different stages of thermal unfolding. Interestingly, peaks in the BDS signal are observed after all treatment temperatures, implying that the nanoconfinement stabilized the protein.

In [26] and [27], the results are interpreted as follows:

- P_{1C} is associated with two sets of dipoles, one with slow fluctuations and the other one with fast fluctuations.
- The set with fast motions is also detected in P_{3C} .
- A third set of dipoles with slow motions is detected in P_{2C} .

Based on the frequencies of detection and the fact that the position of the peaks varies with the treatment temperature, the peaks of the BDS spectrum were assigned to interdomain motions at different stages of the toxin unfolding. The following model was proposed:

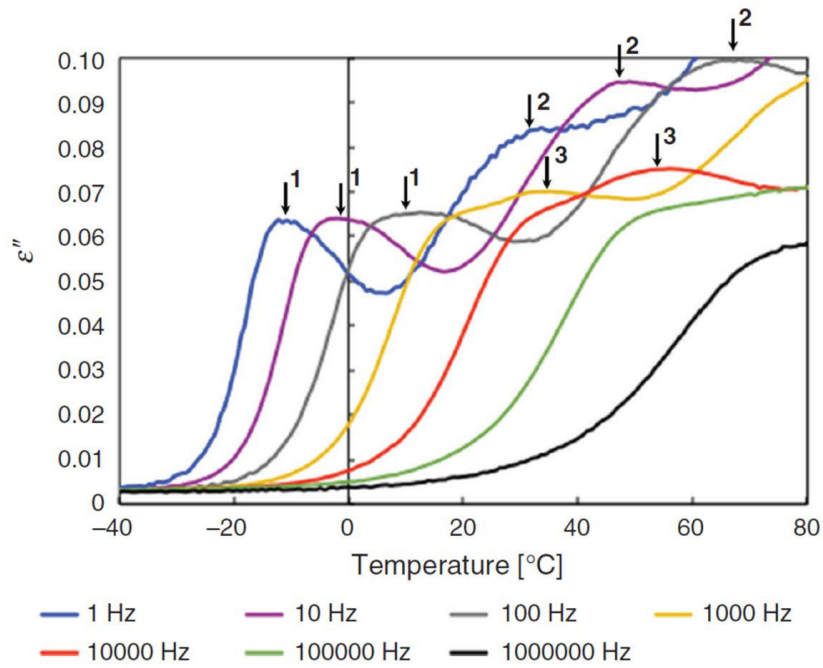


Figure 10.2: Dielectric loss as a function of temperature after 3 hours of thermal treatment at 80°C for the cholera toxin B pentamer at frequencies from 1 to 106 Hz. The three molecular dynamics peaks (P_{1C} , P_{2C} and P_{3C}) are observed corresponding to three relaxation processes. The relaxation processes corresponding to the molecular dynamics peaks P_{1C} , P_{2C} and P_{3C} are indicated by arrows.

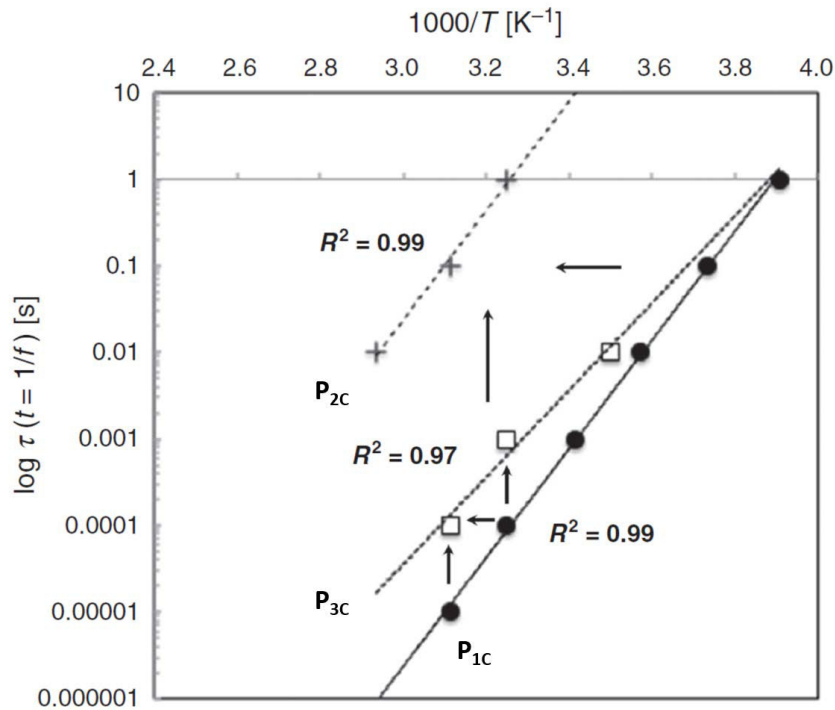


Figure 10.3: Temperature dependencies of the relaxation times for the three relaxation processes P_{1C} , P_{2C} and P_{3C} identified for the cholera toxin B pentamer after thermal treatments at 80°C. On the x axis the temperature (T) is θ_{max} . Dipoles changing frequencies across relaxation processes are highlighted with vertical arrows. Dipoles changing environment but not relaxation times are indicated with horizontal arrows. The data fit with exponential curves that are indicated by straight lines with their respective R^2 values.

- P_{1C} and P_{3C} are associated with non-native pentamers having different conformations.
- P_{2C} is associated with assembly intermediates, implying that the toxin interfaces are probed by the BDS signal.

For the model to be valid, the following conditions must hold:

1. The toxin interfaces must be thermally sensitive;
2. The toxin interfaces must be dielectrically detectable;
3. The toxin interfaces must be composed of three different sets of dipolar relaxations with distinct thermal sensitivity.

The Intramolecular Amino Acid Network, Intermolecular Electrostatic Network and Induced Intermolecular Electrostatic Network models of CtxB₅ are used to investigate whether these conditions are valid.

10.4.2 Electrostatic Network model to validate the peak-assignment of BDS signal

Condition 1. The interfaces in the CtxB₅ pentamer are shown in Fig. 10.4 (left), the same notation (I_1 , I_2 and I_3) as in Chapter 8, Section 8.2.1 is used. Moreover, the sub-interfaces I_{1a} , I_{1b} , I_{2a} , I_{2b} and I_{2c} are defined. The Intermolecular Amino Acid Network (4D-AAN) is the subnetwork of the AAN that contains only 4D-links, i.e. links connecting amino acids belonging to different chains. As an approximation of the interface thermal stability, the number of amino acid and atomic interactions (link weights) involved in the 4D-AAN are computed. In Table 10.1 column Interface parameter, the number of amino acid pairs (number of links in the 4D-AAN) and number of atomic interactions (sum of the link weights in the 4D-AAN) are reported (first two parameters). Accordingly, the interfaces are ranked based on their stability as follows: I_{1a} and I_{2c} are the weakest, I_{1b} and I_{2b} are moderate, and I_{2a} and I_3 are the strongest.

In summary, the interfaces are thermally sensitive and are expected to be perturbed in the following order upon thermal denaturation: I_{1a} and I_{2c} first, then I_{1b} and I_{2b} , and I_{2a} and I_3 last.

Condition 2. In order to assess the dielectric properties of the toxin interfaces, the Intermolecular Electrostatic Network (4D-EN) and the Induced Intermolecular Electrostatic Network (4D-IEN) of CtxB₅ are calculated. Fig. 10.4 (right) shows the 4D-EN and the 4D-IEN of CtxB₅. In the 4D-EN the nodes are charged amino acids that belong to different chains and in the 4D-IEN the first neighbors in the AAN of the nodes of the 4D-IEN are added. Links in the 4D-EN are ionic dipoles and links in the 4D-IEN are induced dipoles. The visibility of an interface in a BDS experiment depends on the number of dipoles present at the interface. Table 10.1 reports the ionic dipoles (links in the 4D-EN) and induced dipoles (links in the 4D-IEN) for all the interfaces. I_{1a} and I_{2c} are

Thermal strength	Interfaces	Chain E	Chain F	Interface parameter [^]	Chain D	Interface parameter [^]	Ionic dipoles	Induced dipoles
Weak interfaces	I _{1a}	1–3	92–93	5, 46, 0, 0			—	—
	I _{2c}	101–103	73–77	5, 22, 0, 2			—	M101–R73 N103–Y76
Moderate interfaces	I _{1b}	1–12	28–39	19, 137, 2, 3			T1–R35 E11–R35	P2–R35 L8–R35 Y12–R35
	I _{2b}	28–39			58–68	19, 136, 2, 9	E29–R67 E36–K63	E29–A64 E29–M68 K34–I58 K34–S60 E36–I58 E36–S60 E36–Q61 E36–A64
Strong interfaces	I _{2a}	25–32			88, 96–103	26, 198, 0, 4	—	A97–E29 A98–E29 I99–E29 S100–E29
	I ₃ Central α -helix: 63 to 81	63–77			63–81	12, 97, 7, 2	K63–E66	I65–R67
							E66–K63 E66–R67 R67–R67 K69–K63 K69–R67 D70–R67 R67–R67 R67–E66 R67–K69 R67–D70 R67–R73 D70–R73	Y76–K81 T71–R73 I74–R73 K81–Y76
		67–81	66–77	12, 117, 6, 3				

[^]Parameters characterising interfaces: number of amino acid pairs, number of atomic interactions, number of ionic dipoles (pairs of charged residues), and number of induced dipoles (pairs of charged and non-charged residues).

Table 10.1: Interface interaction and dipole features inferred from amino acid network based models.

dielectrically invisible and weakly visible, respectively, as they have no charged residues but I_{2c} has two induce dipoles. I_{1b} and I_{2b} have two ionic dipoles and some induced dipoles. I_{2a} is invisible dielectrically as it has no charged residues but it can be detected through the induced dipoles in the environment of the charged residue Glu29. I₃ is the strongest dielectric interface domain with 13 ionic dipoles and 5 induced dipoles.

In summary, the detectability of the interfaces follows the following order: I_{1a} < I_{2a} and I_{2c} < I_{1b} and I_{2b} < I₃.

Condition 3. I_{1a} and I_{2c} are the most thermally-sensitive but cannot be detected by BDS or only weakly through induced dipoles. The thermal perturbation of I_{1a} will be detected when I_{1b} is thermally perturbed as well through the ionic dipoles T1-R35 and E11-R35 (Table 10.1) (Fig. 10.5). Likewise, the thermal perturbation of the I_{2c} domain will be detected via the perturbation of the induced dipoles and the ionic dipoles involving Glu29 when I_{2a} and I_{2b} are thermally disturbed (Table 10.1) (Fig. 10.5).

Because I_{1b} is a moderate interface in terms of amino acid and atomic interactions compared with I_{2a}, I_{1b} unfolding is likely to occur before the unfolding of I_{2a} and so I_{1b} is associated with the slow motion dipoles detected in P_{1C} but lost in P_{3C}. It follows that I_{2a} and I_{2b} are associated with the fast motion dipoles observed in P_{1C} and P_{3C}. The

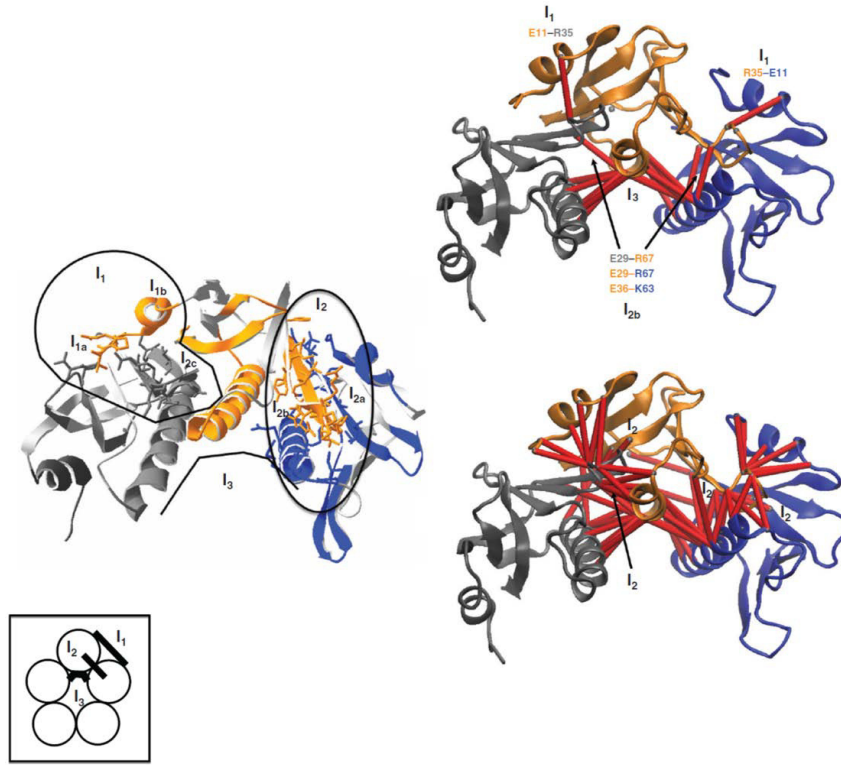


Figure 10.4: The toxin interface. The cholera toxin B pentamer has a complex chain-interface made of mainly three sub-interfaces: I_1 , I_2 , and I_3 , that are shown on a trimer of the X-ray structure of the toxin for simplicity (left). The chain F is in dark grey, the chain E in orange and the chain D in blue. I_1 is sub-divided into two interfaces: I_{1a} made of residues 1 to 3 of chain E interacting with residues 92 and 93 on the adjacent chain F, and I_{1b} made of residues 1 to 12 of chain E interacting with residues 28 to 39 on chain F. I_2 is also sub-divided into three interfaces: I_{2a} made of residues 25 to 32 of chain E interacting with residues 88 and 96 to 103 on chain D, I_{2b} made of residues 28 to 39 of chain E interacting with residues 58 to 68 on the adjacent chain D, and I_{2c} made of residues 101 to 103 of chain E interacting with residues 73 to 77 on the adjacent chain F. I_3 is a tripartite interface composed of residues 58 to 81 of chain M (e.g. E) interacting with residues 58 to 81 on chain M - 1 (D) and M + 1 (F). The Electrostatic Networks, made using VMD 3D-representations, are on the right, with the Inter-molecular Electrostatic Network at the top and the Induced Inter-molecular Electrostatic Network at the bottom. The dipoles are represented by red sticks. A schematic of the three interfaces in the native pentamer are shown in the left side bottom corner.

slow dipoles of P_{2C} are associated with the I_3 interface sub-domain as this domain has the strongest stability and the highest number of ionic dipoles and would be detected in the most unfolded stage of the toxin. In P_{2C} , the toxin conformations would have lost the I_1 and I_2 sub-domain interfaces, their dipoles being no more detected by BDS such that the chain dissociation would occur as only an I_3 destabilised interface remains (slow motion = large fluctuation).

The thermal unfolding perturbation and the dipoles detected in P_{1C} are schematized in Fig. 10.5. The weak interfaces unfold, leading to the perturbation of the I_1 N-terminal interface detected by dipoles with slow motions and to the perturbation of the I_2 β -interface detected by dipoles with fast motions. The dipoles detected in P_{3C} and P_{2C} are shown on the 4D-EN and the 4D-IEN, respectively on Fig. 10.6. The interface I_2 further unfolds in P_{3C} where the dipoles involving the residues Glu29 and Glu36 and the residues

Lys63 and Arg67 on the adjacent chains are detected. In P_{2C} , the N-terminal I_1 interface and the I_2 interface are unfolded and not detected anymore, while only dipoles from the central helix I_3 interface are still detected.

In summary, P_{1C} detects the unfolding of interface I_{1b} and interface I_{2b} , P_{3C} detects the further unfolding of interface I_{2b} , and P_{2C} detects the unfolding of interface I_3 .

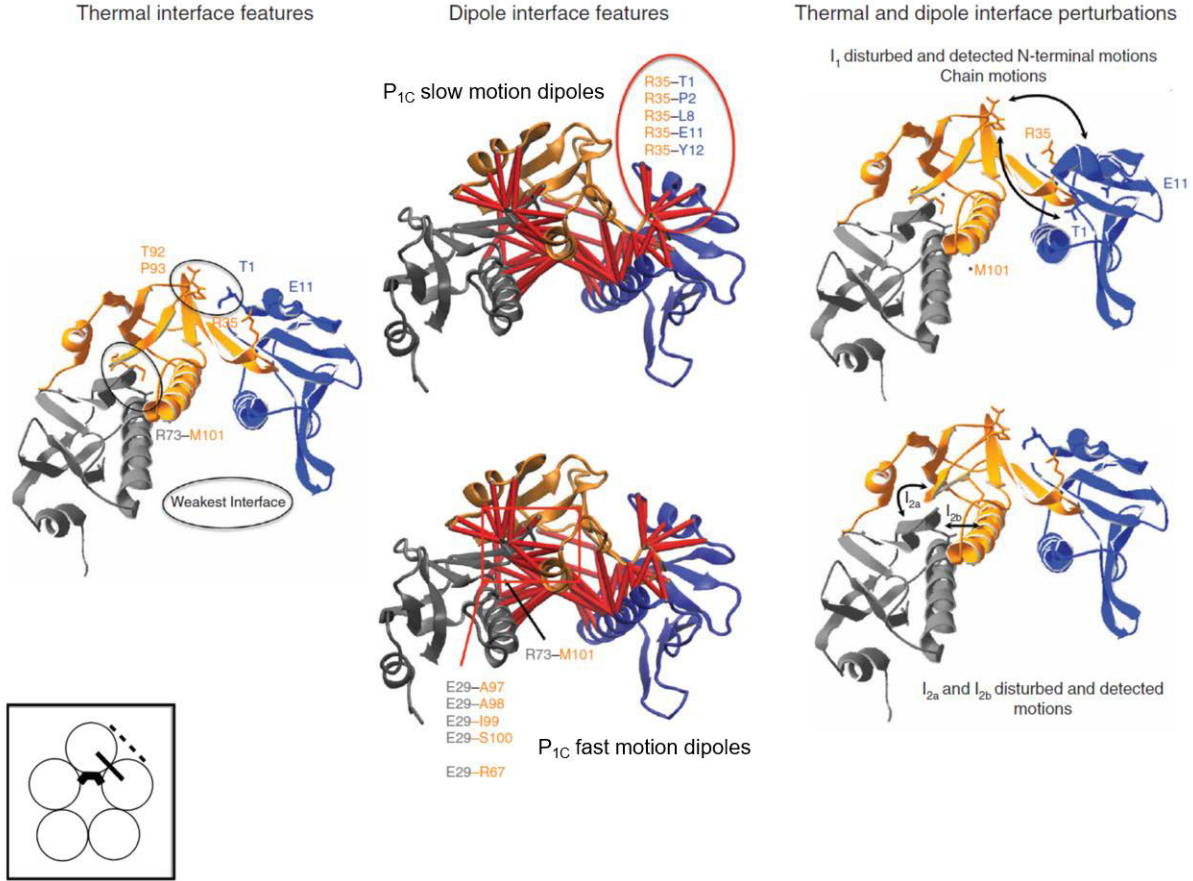


Figure 10.5: Toxin interface state and dipole detected in P_{1C} . P_{1C} is associated with the interface state closest to the native pentamer. The weakest interfaces I_{1a} and I_{2c} are thermally disturbed leading to I_{2a} and I_1 interface fluctuations (left and middle structures). I_{1b} is weaker than I_{2a} so it is associated with the slowest relaxation times of P_{1C} (ten milliseconds to second) while I_{2a} and I_{2b} 's fluctuations are associated with the fastest relaxation times of P_{1C} (microsecond to millisecond). The charged dipoles participating in the P_{1C} signals are shown on the right structures. The box in the left bottom corner is a schematic of the toxin interface state in P_{1C} .

Conclusion The results presented in this section show that Electrostatic Network model is an appropriate tool to guide the interpretation of the BDS signal, probing for multi-scale protein dynamics. The advantage of this integrative approach is the possibility of pinpointing specific dipoles responsible for the observed experimental dynamics and governing the unfolding of the toxin. Thanks to the high resolution of the measure, linking local interactions (dipoles, i.e. links in the Electrostatic Network) to global motions (unfolding dynamics), it is a promising technique to study global dynamical differences in protein variants caused by local perturbations (amino acid mutations, resulting in links

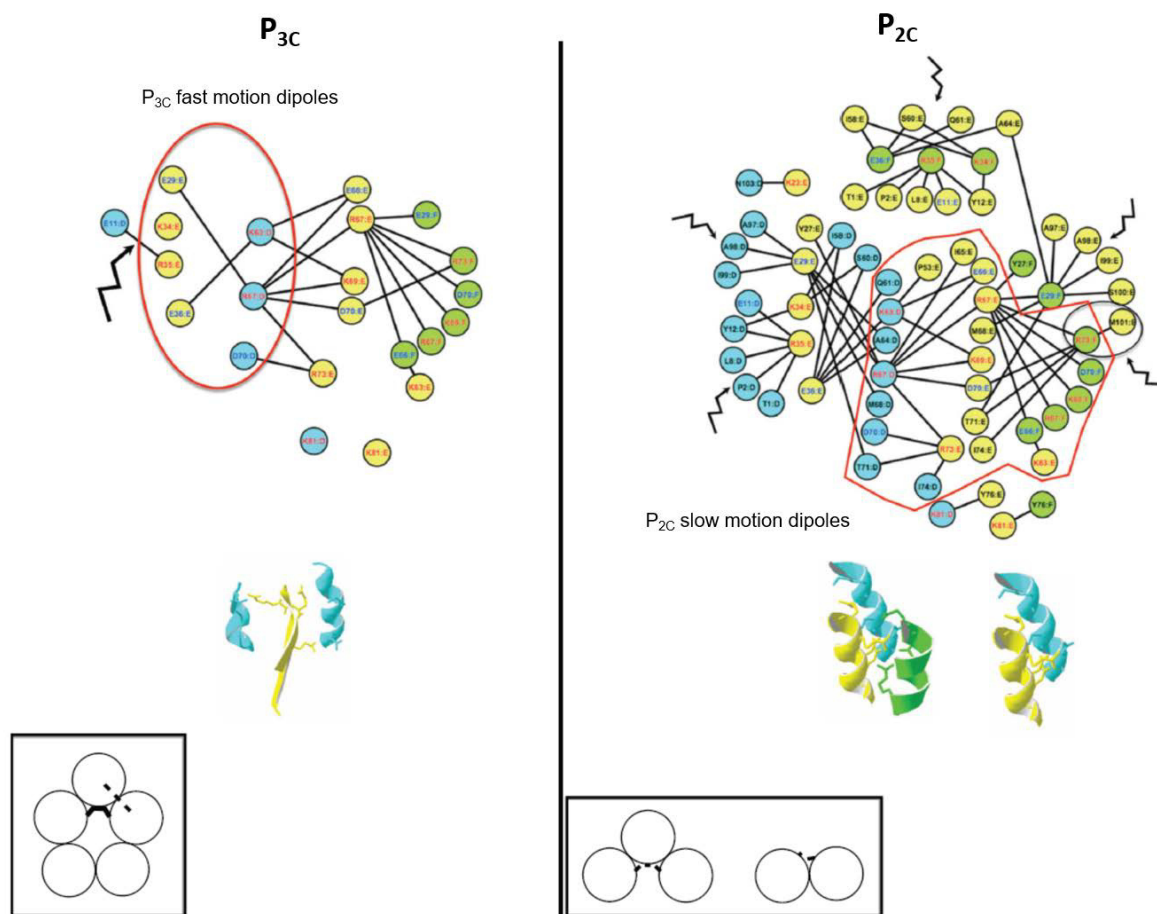


Figure 10.6: Toxin interface state and dipoles detected in P_{2C} and P_{3C} . The P_{3C} interface state is shown on the left panel. The 4D-EN of the whole interface is shown for a toxin trimer with the nodes (charged residues) of chain D in cyan, of chain E in yellow, and of chain F in green. The residue name and sequence positions are indicated within the nodes, with non-charged residues written in black, positively charged residues in red, and negatively charged residues in blue. P_{3C} has lost the slow motion dipoles associated with the N-terminal interface, which is considered damaged and not detected (jagged arrows). Consequently, the I_{2b} interface becomes destabilised and detected (red box) with dipoles E29-R67 and E36-K63 shown as linked on the 4D-EN. The structure of the I_{2a} and I_{2b} interfaces detected in P_{3C} are shown below the network. The box in the left bottom corner is a schematic of the toxin interface state in P_{3C} . The P_{2C} interface state is shown on the right panel. P_{2C} is associated with the most destabilised interface state of the toxin. Only the central α -helix interface I_3 is assumed to be still present. The 4D-IEN of the whole interface is shown for a toxin trimer with the nodes (charged residues) of chain D in cyan, of chain E in yellow, and of chain F in green. The damage interfaces are indicated with jagged arrows while the red box indicates the destabilised I_3 interface and the maximum dipoles detected in P_{2C} . The structure of the I_3 interfaces detected in P_{2C} is shown below the network. The box in the left bottom corner is a schematic of the toxin interface state in P_{2C} .

perturbations). To investigate this possibility, in the next session the dynamics of the two AB_5 toxin variants, CtxB₅ and LTB₅, are compared using network analysis coupled with the experimental measure from BDS.

10.5 Measure of multi-scale dynamical perturbation upon mutation: CtxB₅ versus LTB₅ toxins

Seventeen amino acid substitutions exist in LTB₅ compared to CtxB₅: A1T S4N, E7D, S10A, Y18H, I20L, L25F, M31L, V38A, S44N, T75A, T80A, I82V, D83E, N94H, S95A and E102A. These mutations do not impact the structure of the two toxins but make LTB₅ more stable and bring the two toxins to unfold and refold by different mechanisms. In Chapter 8 (Section 8.2.1), these differences were associated to different allocation of atomic interactions at the central α -helix, the C-terminal domain and the N-terminal domain.

In this section, the experimental dynamical of the two toxins upon thermal denaturation are compared based on the BDS experimental measure and the Electrostatic Network models of the two toxins are used to interpret the differences in dielectric signal.

10.5.1 Differences in experimental molecular dynamics

The molecular dynamics of LTB₅ were studied with BDS following the same experimental protocol used for CtxB₅ (Fig. 10.7). In the following, the main experimental results are summarized. The details can be found in the publication [28].

Fig. 10.7 shows that the BDS signal of LTB₅ is different from the signal of CtxB₅. Four peaks of BDS signal are observed for LTB₅ after different thermal treatments and at different frequencies. The peaks are named P_{1L} , P_{2L} , P_{3L} and P_{4L} , ordered from the left-most to the right-most position in the temperature axis.

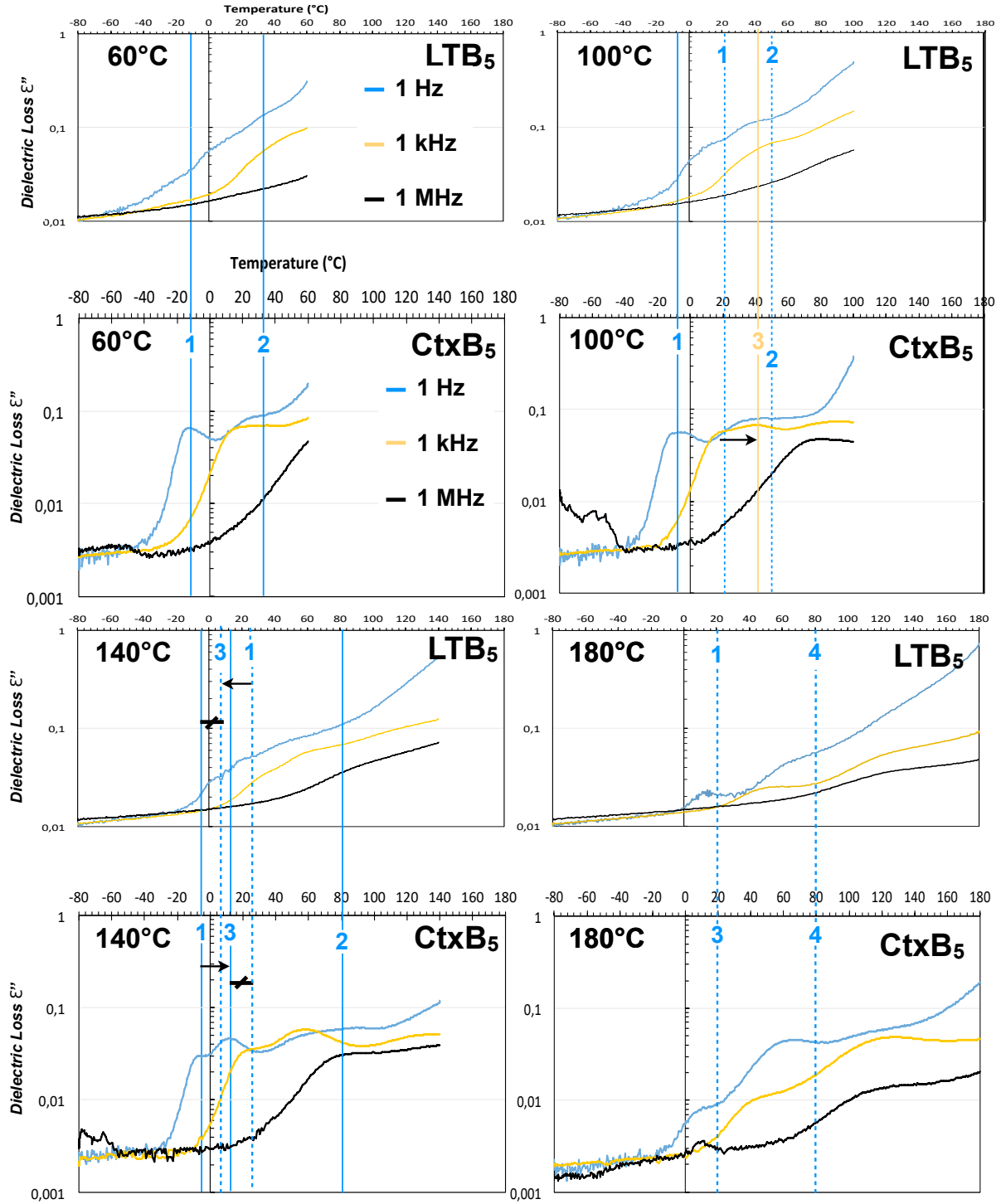


Figure 10.7: LTB₅ and CtxB₅ thermal dynamics changes measured by BDS. Temperature dependencies of the LTB₅ and CtxB₅ dielectric signals after the thermal treatments at different temperatures and for three different frequencies. The vertical lines underline the temperature positions of the different peaks for 1 Hz (cyan) and 1 kHz frequencies (orange), identified by their respective positions. The dotted lines are for LTB₅ peaks while the continuous lines are for CtxB₅ peaks. The arrows illustrate the shift of the position in temperature of the new peaks emerging at higher temperatures of the thermal treatments (Peaks 3) compared to the peaks detected at lower temperatures of the thermal treatments (Peaks 1). The differences between the two toxin B-subunit peaks 1 and 3 which exhibit the presence of additional dipoles is illustrated with a bar symbol.

The position of the peaks as a function of the thermal-treatment temperature for LTB₅ and CtxB₅ (Fig. 10.7) were compared to infer if there exists a correspondence between the peaks observed in LTB₅ and the ones observed in CtxB₅. A correspondence means that the same set of dipoles is probed. The result is summarized in Fig. 10.8.

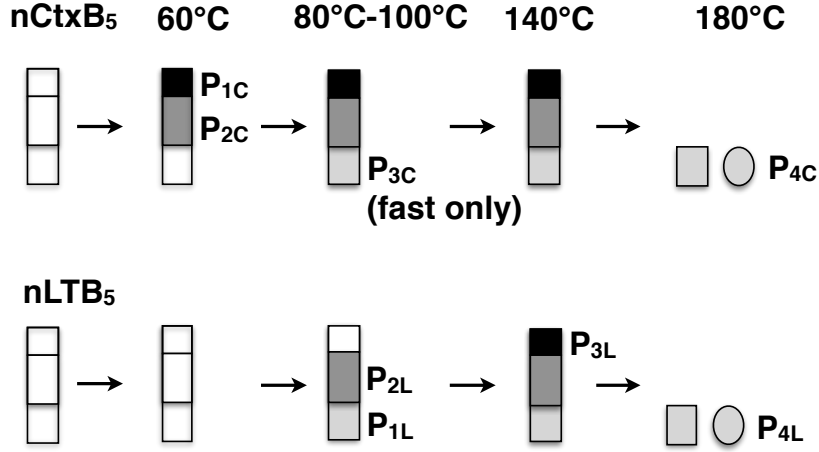


Figure 10.8: Schematics of the unfolding mechanisms based on the different relaxation peaks detected by BDS along the thermal treatments. The whole box represents the toxin and each smaller colored box schematizes the toxin area with the dipole fluctuations contributing to one relaxation peak. The white boxes depict the native state of the toxin pentamers, namely the structural state before the thermal unfolding starts, which is not detected in the experiments. The absence of boxes indicates a relaxation process no more detected, implying a structural state that cannot be identified because its dynamics fails outside the frequency and temperature windows of the experiments. The P_4 relaxation is illustrated as a conformational change resulting from the P_{3C} (CtxB₅) or P_{1L} (LTB₅) dipoles having different local environment after the thermal treatment at 180°C. Alternatively, the P_4 relaxation could arise from the fluctuations of new dipoles not destabilized at 140°C, in which case the toxin native state should be represented by four boxes instead of three.

Two main conclusions can be drawn: first, dielectric relaxation processes are observed after thermal treatment at higher temperature in LTB₅ compared to CtxB₅. This is consistent with the higher thermal resistance of LTB₅ and it confirms that the BDS signal at the frequencies employed here does not probe the native conformation of the toxin, but only the dipoles fluctuations resulting from the thermal perturbation of the protein structure. Second, the order of appearance of the dipoles fluctuations is not the same in the two toxins, consistent with different unfolding mechanisms.

As a reminder, in the previous section the following assignment of the peaks was proposed: P_{1C} detects the unfolding of interface I_{1b} and interface I_{2b} , P_{3C} detects the further unfolding of interface I_{2b} , and P_{2C} detects the unfolding of interface I_3 . If the correspondence of the peaks proposed in Fig. 10.8 holds, then P_{1L} detects the unfolding of interface I_{2b} , P_{2L} detects the unfolding of interface I_3 , and P_{3C} detects the unfolding of interface I_{1b} and interface I_{2b} . It follows that the I_2 and I_3 interfaces are perturbed before the I_3 interface in CtxB₅, while the I_3 interface is perturbed before the I_1 and I_2 interfaces in LTB₅. The P_{4C} and P_{4L} peaks appear after thermal treatment at high temperature (180°C) and would correspond to unfolded states.

Importantly, the peaks assignment in Fig. 10.8 is based on the position of the peaks, that is an indicator of the characteristic time-scale of the fluctuation. Nevertheless, the broadness of the peaks is also an important feature, as broader peaks correspond to more heterogeneous relaxation processes. P_{2C} and P_{2L} are assigned to the same sets of dipoles (interface I_3), but the P_{2C} peak is broader compared to the P_{2L} peaks. It is proposed that P_{2C} is broader because it probes the dynamics of the I_3 interface in assembly intermediates, while P_{2L} probes the dynamics of the I_3 interface in pentamers.

In the following, the Electrostatic Networks of the two toxins are compared to infer which atomic dipoles changes, caused by the mutations, are responsible for the observed differences in the observed unfolding dynamics.

10.5.2 Electrostatic Network models

The two toxin pentamer interfacial domains I_{1a} , I_{1b} , I_{2a} , I_{2b} , I_{2c} and I_3 , have same stability ranking and intermolecular electrostatic dipole content, inconsistently with the two different B subunit dielectric signals (Table 10.2).

This suggests that intramolecular electrostatic dipoles contribute to the dielectric signals as well, and are responsible for the differences. The Electrostatic Network (EN) containing both intermolecular and intramolecular electrostatic dipoles was then built for the two toxin pentamers to investigate such possibility (Fig. 10.9).

The residue 103, a lysine in the X-ray structure PDB 1LTR of LTB₅, was ignored in the EN of LTB₅. Nevertheless, the dipole (K81, N103) is considered to be present in the BDS measurement because the residue 103 is an asparagine in the experimental LTB₅ and because the carboxyl group of the C-terminal is in the same position in the two toxin structures. There is no X-ray structure available in the Protein Data Bank (<https://www.rcsb.org/>) for the human LTB₅ with the experimental C-terminal.

The comparison of the ENs shows that LTB₅ has two supplementary intramolecular electrostatic dipoles (A1, E7) and (K81, E102) compared to CtxB₅, due to the mutations A1T, E7D and E102A at positions 1, 7 and 102. The ionized amino group of the N-terminal of the residue 1 is involved in the (A1, E7) dipoles. The electrostatic dipole (K81, E102) is a backup of the electrostatic dipole (K81, N103), present in both toxin pentamers whereas the dipole (A1, E7) has no backup in CtxB₅. The (K81, N103) dipole is also present in LTB₅ although not shown on the LTB₅ network.

The (A1, E7) and (K81, E102) dipoles can be expected to increase the stability of the LTB₅ N-terminal and of the 3D-contacts between the β_6 (96-103) and the large central α -helix (59-79), respectively (Fig. 10.10), as salt bridges stabilize protein conformations [213]. Moreover, the N-terminal is involved in the I_{1a} and I_{1b} interfaces (Table 10.2), so both interfaces can be expected to be more stable and less susceptible to dissociation in LTB₅ (Fig. 10.10). The (K81, E102) locks the lower end of the β_6 (96-103) into position with the large central α -helix (59-79) and the (A1, P93) locks the upper end of the β_6 into

Thermal strength	Interfaces	Chain E ^a	Chain F ^a	Interface Parameter ^b	Chain D ^a	Interface Parameter ^b	Electrostatic dipoles
Weak interfaces	I _{1a}	1-3	92-93	5, 50 <u>5, 46</u>			-
	I _{2c}	101-103	73-79 <u>73-77</u>	5, 65 <u>5, 22</u>			-
Moderate interfaces	I _{1b}	1-12	28-39	18, 139 <u>19, 137</u>			E11-R35
	I _{2b}	28-39			57-73 <u>58-68</u>	19, 157 <u>19, 136</u>	E29-R67
Strong interfaces	I _{2a}	25-32			88, 96-103	24, 166 <u>26, 198</u>	-
	I ₃ Central α -helix: 61 to 81	61-77 <u>63-77</u>			63-81	12, 120 <u>12, 97</u>	E66-K63 E66-R67 D70-R67 R73-D70
		67-81	66-77	13, 90 <u>12, 117</u>			R67-E66 R67-D70 D70-R73

^a The numbers refer to the first and last residue of each domain that constitutes the interface, ^bInterface Parameters: number of amino acid pairs, number of atomic interactions.

Table 10.2: LTB₅ and CtxB₅ (underlined if different) interface interactions and intermolecular electrostatic dipoles inferred from network-based models.

position with the I_{1a} interface (Fig. 10.10). Thus, the 3D-mobility of β_6 , linked to the fluctuations of the dipoles (A1, E7) and (K81, E102) can be expected to be reduced in LTB₅. Likewise, the (K81, D22) dipole locks the upper end of the β_2 (25-31) into position with the large central α -helix (59-79) while the intermolecular dipole (D11, R35) locks the lower end of the β_2 into position with the I_{1b} interface, making the 3D mobility of the β_2 dependent as well on the dynamics of the additional dipoles (A1, E7) and (K81, E102) (Fig. 10.10). Since β_6 and β_2 on the adjacent chain compose the main β -interface (I_{2a}), the two supplementary dipoles are likely to also make I_{2a} less susceptible to dissociation

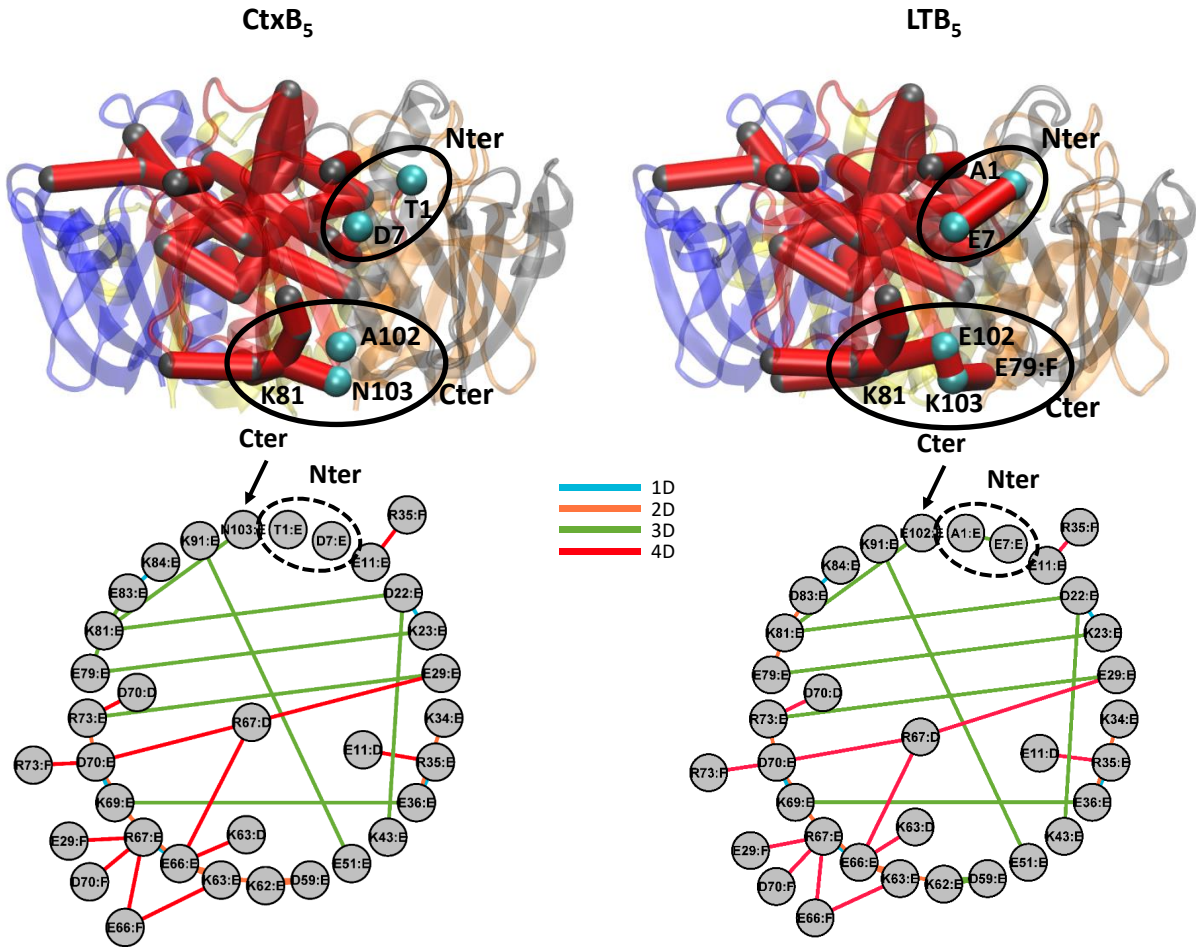


Figure 10.9: Electrostatic Networks of *CtxB₅* and *LTB₅*. In the *LTB₅* structure (PDB 1LTR), K103 makes a dipole with E102 of the same chain and E79 of the neighboring chain. However, residue 103 is an asparagine in the experimental *LTB₅*, as in *CtxB₅*. For this reason, dipoles (K103:E, E102:E) and (K103:E, E79:F) of *LTB₅* are discarded in the network analysis, and a dipole (N103:E, K81:E) is assumed to be present.

in *LTB₅* (Fig. 10.10). Finally, β_6 is connected intramolecularly to the α -helix (59-79) and the N-terminal such that a lower β_6 3D-mobility will also protect the I_3 interface (Fig. 10.10).

The increase in dipole interactions at the N- and C-terminal (I_1 and I_2 interfaces) of *LTB₅* compared to *CtxB₅* is consistent with the differences in structural allocation of atomic contacts observed in Chapter 8, that showed that *LTB₅* allocates more interactions to the maintenance of interfaces and less to the maintenance of the secondary structure in *LTB₅* compared to *CtxB₅*.

The network and structural analysis pinpoints that the two additional dipoles can stabilize the N-terminal, the C-terminal and the interfaces of *LTB₅*, consistently with a higher resistance to dissociation for *LTB₅* and the attribution of the narrow peak P_{2L} to non-native-pentamers with destabilized interfaces but not enough for dissociation, and the attribution of the broad peak P_{2C} to assembly intermediates in *CtxB₅*. The network model also shows that the difference in the two toxin B subunit dielectric signals is due

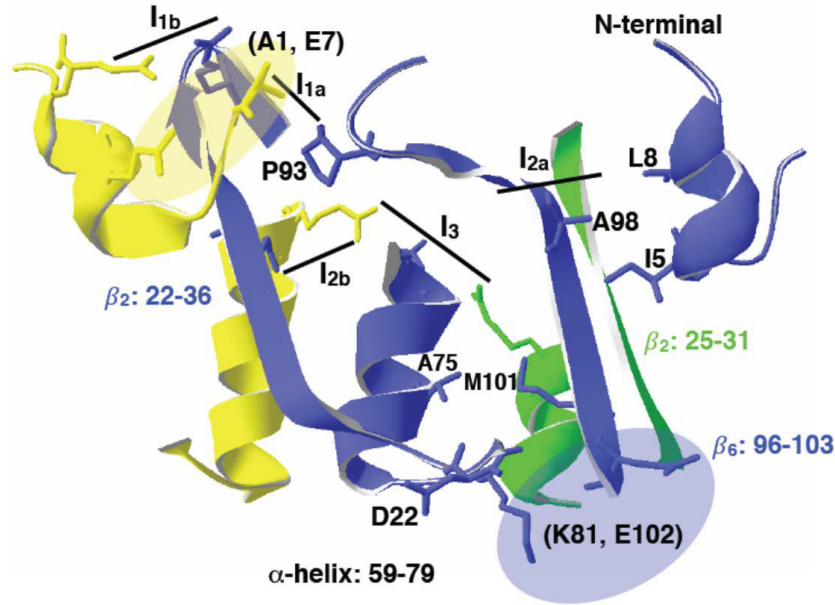


Figure 10.10: Structural analyses of the dynamic differences between LTB₅ and CtxB₅. Positions of the additional dipoles (A1, E7) and (K81, E102) in the LTB₅ X-ray structure. The dipole (A1, E7) is located on the N-terminal domain which is involved in the I₁ interfaces (I_{1a} and I_{1b}). The dipole (K81, E102) connects the β_6 from the main β -interface (I_{2a}, β_2 : 25 to 31 and β_6 : 96 to 103) to the tip of the central α -helix (residues 59 to 79) involved in the I₃ interface.

to differences in the intramolecular dipole contents.

Because intramolecular dipoles contribute to the signal, monomers could also be detected. This required revisiting of the model proposed in Section 10.4, as reported in [28]. In particular, the P_{4C} and P_{4L} were proposed to monitor the dynamics of the B-subunit monomers. Moreover, the following possibility was proposed: P_{2C} could probe the destabilization of the I₃ interface in CtxB₅, leading to assembly intermediates, while P_{2L} would probe unfolding of the I₃ interface within pentamers. This possibility was supported by the fact that P_{2C} is broad while P_{2L} is narrow [28]. Importantly, this possibility is also consistent with the results of Chapter 8. Indeed, the decrease in the allocation of atomic interactions to the 2D-structural level in the central α -helix in LTB₅ compared to CtxB₅ suggested that the α -helix unfolds before the I₃ interface dissociates in LTB₅, while the I₃ interface dissociates in CtxB₅ before the α -helix unfolds.

10.6 Conclusion

The combination of the experimental results and the network analysis reveal the potential role of the two LTB₅-specific dipoles (A1, E7) and (K81, E102) and the mutations A1T, E7D and E102A in deviating LTB₅ to unfolding paths that make LTB₅ more heat resistant and less prompt to thermal dissociation through assembly intermediates. This recalls what has been shown for the B-subunit in vitro reassembly under macroscopic conditions where the rate-limiting step for LTB₅ is a folding step involving the N-terminal but is interface

formation involving the main β -interface (I_{2a}) for CtxB₅ [184, 214–218]. The concordance between the nanoscopic and the macroscopic investigations highlights the encoding of the dynamics (here the slow large-scale dynamics) in the sequence of proteins as does the distinct dielectric responses of the two toxin pentamers, and as shown by Anfinsen experiments in the early seventies [219].

More in general, the results presented in Chapters 7 to 10 have proven that the information on the protein dynamics is encoded in the Amino Acid Network (AAN) model of the protein structure. Even though the current knowledge does not allow decoding the dynamics from the protein structure or its AAN, the case studies presented in these chapters show that the scale of the protein dynamics are controlled by the allocation of atomic interactions (links in the AAN) to structural levels (1D, 2D, 3D and 4D). As a consequence, the impact of a mutation on the protein structure depends not only on the number of atomic interactions that are perturbed (Chapter 9) or the spatial scale covered by the perturbation (Chapter 6), but also on the structural level of the perturbed interactions (Chapter 8 and current chapter). This implies that the consequence of mutations depends on the direction of the perturbation they provoke, towards the secondary, tertiary or quaternary structural levels. In other words, to assess the impact of an amino acid mutation on the protein dynamics, it is important to determine which other amino acids are perturbed.

The next Chapter proposes a directed network model of perturbations caused by in-silico mutations as a tool to gather insights on the directionality of the perturbations encoded in the protein structure.

Chapter 11

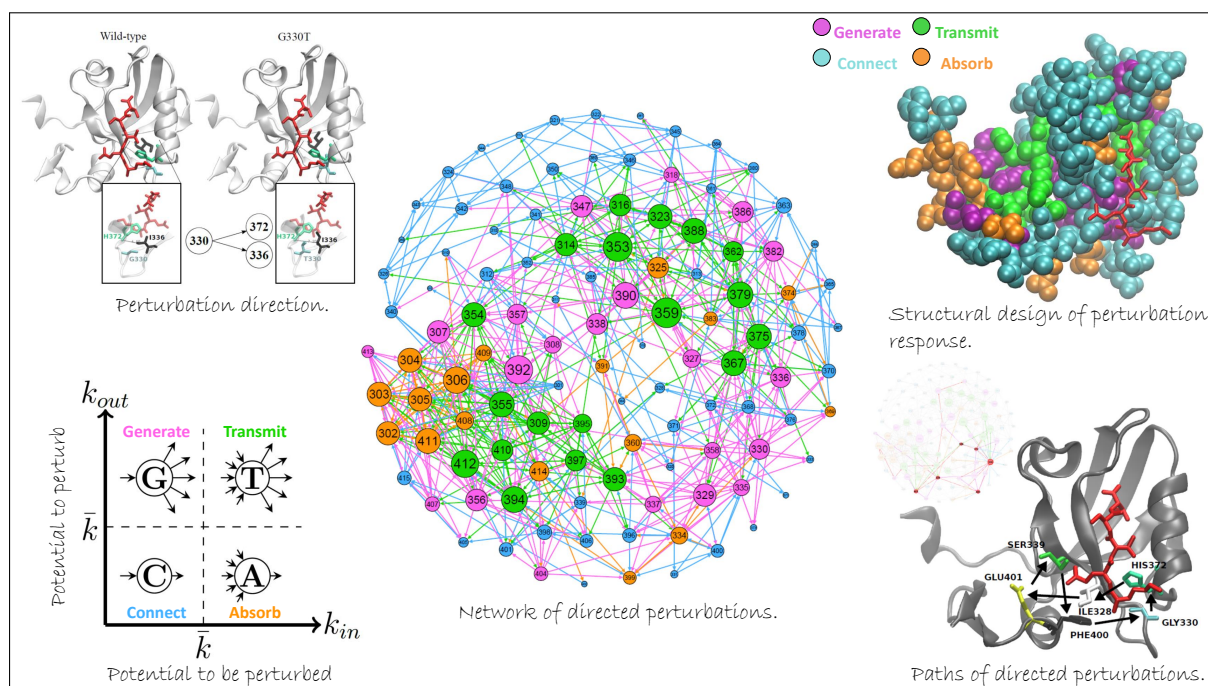
Proof of concept: Third PDZ domain. Directions of perturbations caused by amino-acid mutations in the protein structure measured by a directed network.

Highlights:

- The direction of perturbations caused by amino-acid mutations is modeled using a directed network.
- The out- and in-degree of nodes measure potentials of amino-acid positions to perturb if mutated and to be perturbed upon mutation elsewhere, respectively.
- 4 classes of amino acid positions are found based on their potential to perturb or to be perturbed.
- The asymmetry of perturbations caused by mutations is picked up by the GCAT model.

Abstract: Amino-acid-neighborhood perturbations caused by amino-acid mutations are modeled using a directed network called GCAT. Four classes of amino acid positions are found based on their potential to perturb or to be perturbed upon mutations (A: likely perturbed, G: likely perturbs, T: both, C: none). The likelihood to perturb or to be perturbed are measured by the out- and in-degree of nodes in the GCAT network, respectively. The position of the GCAT classes in the protein structure suggests an embedding of the response to perturbations in the protein structure design. The asymmetry of perturbations caused by mutations is picked up by the GCAT model. Oriented paths in the GCAT network are a tool to investigate the impact of multiple mutations.

Graphical abstract:



Methods: GCAT Network (Section 3.2.3).

11.1 Introduction

The previous chapters have shown that amino-acid mutations cause changes in the atomic interactions in the protein structure, probed by changes in link weights in the corresponding Amino Acid Network (AAN).

Chapters 8, 9 and 10 showed that the impact of mutations on the protein dynamics depends on the structural level (1D, 2D, 3D or 4D) to which the perturbed atomic interactions belong. It follows that if a mutation $i \rightarrow i'$ causes the gain or loss of a (j, k) contact (i.e. amino acids j and k are perturbed, as defined in Chapter 6), then the impact of the $i \rightarrow i'$ mutation on the protein dynamics depends on whether j and k are 2D-, 3D- or 4D-neighbors. Please note that $j = i$ in the case of a Local (L) mutation (Chapter 6).

We are still far from decoding the impact of amino acid mutation on the protein dynamics from the sole knowledge of the wild-type AAN. Nevertheless, as proposed in Chapter 6, the changes in atomic contacts needed to accommodate mutations by the protein structure can be simulated using in-silico mutations. Then, the Induced Perturbation Network (IPN, Chapter 9) and the comparison of the allocation of atomic interaction to structural levels (Chapter 8) have been proven efficient tools to deduce the dynamical impact of the mutation from atomic-contact changes.

However, two main problems persist if one wants to understand the consequences of any mutations made on a wild-type protein structure. The first problem is practical. Let us say that we want to investigate the impact of any mutation at any amino acid position of a protein structure made of N amino acids. IPNs describe the perturbation caused by a mutation in terms of changes in atomic interactions (link weights in the AAN). To investigate the perturbations caused by all possible single mutations in a protein structure made on N amino acids, $19N$ IPNs should be analyzed. This would be extremely laborious even for small proteins. A tool is needed to visualize the perturbation caused by all $19N$ single amino-acid mutations at the same time.

The second problem concerns multiple mutations. The IPN is not defined for multiple mutations (Chapter 3, Section 3.1.3) but the Perturbation Network (PN) and the comparison of the allocation of atomic interactions to structural levels provide information on the changes in atomic interactions caused by the mutations (Chapter 8). However, both methods allow to determine the global effect of set of mutations, while the contribution of each single mutation is unknown as well as how they influence one another.

Determining the contribution of each single mutation to the global impact of multiple mutations is a complicate task because the perturbation effects of mutations are not always additive, as observed experimentally and from simulations [71–73]. This means that the effect of a double mutation at two positions i and j cannot be reduced to the superposition of the effects of the mutations at positions i and j taken individually. On the one hand, as a consequence of the non-additivity, a deconvolution of the role of each

mutation in protein variants is a challenging task. On the other hand, unraveling the inter-dependence between mutations is necessary to explain compensatory mutations and functional change (evolvability) through subsequent mutations [13, 70], key ingredients for the sustainability of proteins (Chapter 2, Section 2.3).

The existence of non-additivity, compensatory mechanisms and functional change through subsequent mutations highlights the complexity of the perturbation response of protein structures: amino-acid mutations act cooperatively, such that the effects of amino-acid mutations are not always independent. If the impact of the mutations of two amino acids i and j are not independent, we say that i and j are in cooperative interaction.

In this chapter, a computational tool named the GCAT network is proposed to describe the perturbation caused by all in-silico amino acid mutations of a protein in a single representation (first problem). An exploratory analysis of the GCAT network of a protein structure is proposed. Then, the relevancy of the GCAT network tool to investigate the inter-dependence between mutations (second problem) is discussed.

The GCAT network of the third PDZ domain of the synaptic protein PDS-95 (PSD95^{pdz3}, PDB 1BE9) is presented. This case of study was chosen as in Chapter 6 because the functional impact of most of its single amino-acid mutations is known from experiments [16].

11.2 The GCAT network tool

The tool presented here, called the GCAT network, represents the perturbations caused by in-silico single-amino acid mutations in the form of a directed network. The same in-silico mutants as in Chapter 6 are used. Each node in the GCAT network is an amino acid position and a directed arc^a connects a node i to a node j if there exists at least one in-silico mutation of amino acid i that causes a change in the amino acid neighborhood of amino acid j , compared to the wild-type (WT) structure. In this case, as defined in Chapter 6, we say that amino acid i perturbs amino acid j . It should be noted that this procedure aggregates the information on the perturbation caused by all the nineteen mutations of an amino-acid position. This choice is made to reduce the complexity of the GCAT network in terms of number of arcs. The amino acid neighborhoods of all the amino acid positions in the in-silico mutant and in the WT structure are extracted from their respective Amino Acid Networks (AANs). The details of the methods are reported in Chapter 3, Section 3.2.3.

The following example illustrates how arcs are added to the GCAT network. The in-silico mutation G330T of PSD95^{pdz3} causes the loss of all atomic interactions between the amino acids H372 and I336, that are present in the WT structure (Fig. 11.1). Because node 372 has changed neighborhood due to a mutation of node 330 (i.e. 372 is perturbed by 330), an arc from node 330 to node 372 is present in the GCAT network. Similarly, also

^aLinks in a directed network are called arcs.

node 336 has changed neighborhood due to a mutation of node 330 (i.e. 336 is perturbed by 330), so an arc from node 330 to node 336 is present in the GCAT network. None of the 19 in-silico mutations of the H372 residue have an impact on the neighborhood of the G330 residue, and thus there is no link in the GCAT network from position 372 to position 330. Similarly there is no arc from node 336 to node 330 because none of the 19 mutations of residue I336 perturbs the residue G330.

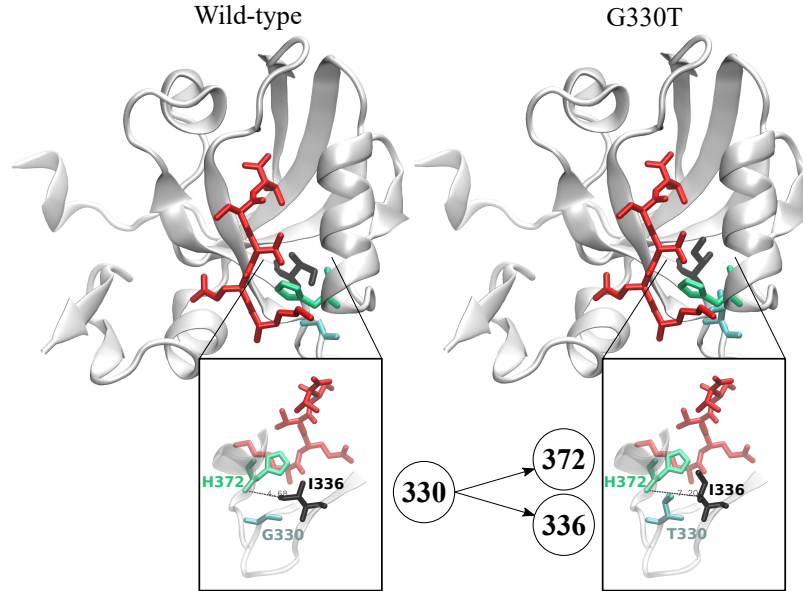


Figure 11.1: Perturbation of amino acids H372 and I336 due to the G330T mutation in PSD95^{pdz3}. The position of the ligand is shown in red. *Left*: the wild-type structure (PDB: 1BE9). *Right*: the G330T in-silico mutant produced with FoldX [161]. The images are produced with VMD [38].

As illustrated by the example, an arc $i \rightarrow j$ in the GCAT network reads “ i can perturb j ” or equivalently “ j can be perturbed by i ”. We use the expression “can perturb” and not simply “perturbs” because an arc $i \rightarrow j$ is present if there exists some mutation of i that perturbs j , but not necessarily all mutations of i perturb j . It follows that the number of arcs leaving a node i , i.e. the out-degree $k_{out,i}$, represents the likelihood that i will perturb some other amino acid j (considering all nineteen $i \rightarrow i'$ mutations together). Conversely, the number of arcs entering a node i , i.e. the in-degree $k_{in,i}$, represents the likelihood that amino acid i will be perturbed by some other amino acid j (considering all nineteen $j \rightarrow j'$ mutations together).

We have classified the nodes (amino acid positions) in the GCAT network according to their likelihood to perturb or to be perturbed, measured by the in- and out-degree. Four classes - G for Generate, C for Connect, A for Absorb and T for Transmit - are defined as follows: G if $k_{out,i} > \bar{k}$ and $k_{in,i} \leq \bar{k}$; C if $k_{out,i} \leq \bar{k}$ and $k_{in,i} \leq \bar{k}$; A if $k_{out,i} \leq \bar{k}$ and $k_{in,i} > \bar{k}$; T if $k_{out,i} > \bar{k}$ and $k_{in,i} > \bar{k}$, with $\bar{k} = \bar{k}_{in} = \bar{k}_{out}$ the average in- and out-degree in the network (the \bar{k} notation stands for average of k). A scheme of the GCAT classification is reported in Fig. 11.2. In the GCAT network of PSD95^{pdz3}, $\bar{k} = 7$.

It should be noted that the neighboring nodes in the GCAT network are not necessarily

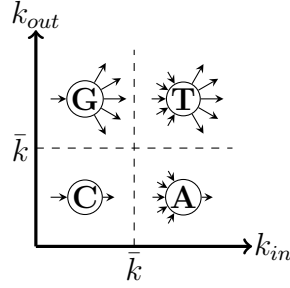


Figure 11.2: Classification of nodes as G (Generate), C (Connect), A (Absorb) and T (Transmit) in the GCAT network, based on their in- and out-degree.

neighboring nodes in the Amino Acid Network (AAN) of the protein structure. This is because the perturbation caused by mutations can reach amino-acids further than the chemical neighbors of the mutated position, as was measured in Chapter 6 (Far class of mutations). We measure distances between two amino acid i and j by their geodesic distance in the AAN ($d_{AAN}(i, j)$), that is equivalent to the shortest path between the two nodes. As a reminder, the shortest path between two nodes is equal to the number of “hops” necessary to travel from one node to the other, where “hops” are only allowed between nodes connected by a link. $d_{AAN}(i, j) = 1$ means that i and j are connected, i.e. i and j are within chemical distance ($\leq 5\text{\AA}$), while $d_{AAN}(i, j) > 1$ means that i and j are at distance $> 5\text{\AA}$. $d_{AAN}(i, j) = 2$ means that i and j have at least one neighbor in common, while $d_{AAN}(i, j) = 3$ means that i and j do not share any neighbors, but at least two of their neighbors are connected, etc.

In Chapter 6, in-silico mutations were classified as Z, Ll, Lg, Lw, Fk, Fw based on the spatial scale of the perturbation and the perturbation mechanism (Z: no perturbation, L: local perturbation, F: large-scale perturbation, k: degree-mechanism, w: weight-mechanism). The relation between the mutations classification and links in the GCAT network is as follows. Z mutations create no link in the GCAT network; Ll and Lw mutations create one or more link in the GCAT network with $d_{AAN}(i, j) = 1$; Lg mutations create one or more links in the GCAT network with $d_{AAN}(i, j) = 2$; Fk mutations create one or more links in the GCAT network with $d_{AAN}(i, j) = 1$, one or more links with $d_{AAN}(i, j) = 2$ and possibly links with $d_{AAN}(i, j) > 2$. Fw mutations create no link with $d_{AAN}(i, j) = 1$, but some links with $d_{AAN}(i, j) \geq 2$.

However, it should be noted that the classes of mutations are not directly readable from the GCAT network, because the information on all 19 mutations of an amino acid is aggregated in its out-arcs. A possibility to incorporate the information on the effect of each single mutation of the amino acids would be to employ multiple arcs between nodes, each arc labeled based on the exact mutation that has caused the perturbation. This procedure would highly increase the complexity of the GCAT model, and is not adopted in this first study.

In the next section, the properties of the GCAT network of PSD95^{pdz3} are explored.

11.3 Exploratory analysis of the GCAT network of PSD95^{pdz3}

Global network properties. The GCAT network of PSD95^{pdz3} is reported in Fig. 11.3. Only two amino acids, GLY351 and ASN403, have $k_{in} = k_{out} = 0$ and are removed from the GCAT network. GLY351 and ASN403 are two surface-exposed residues (relative Surface Accessible Areas equal to 0.76 and 0.69, respectively) whose mutations never perturb any amino acids, i.e. they were classified as Zero in Chapter 6, explaining why $k_{out} = 0$. An additional information provided by the GCAT network is that these two amino acids are also never perturbed by any other mutations ($k_{in} = 0$). Because the positions 351 and 403 are never perturbed and never perturb, they could be considered as independent from the rest of the structure. After removal of the nodes 351 and 403, the GCAT network is a connected network, i.e. made by a single connected component.

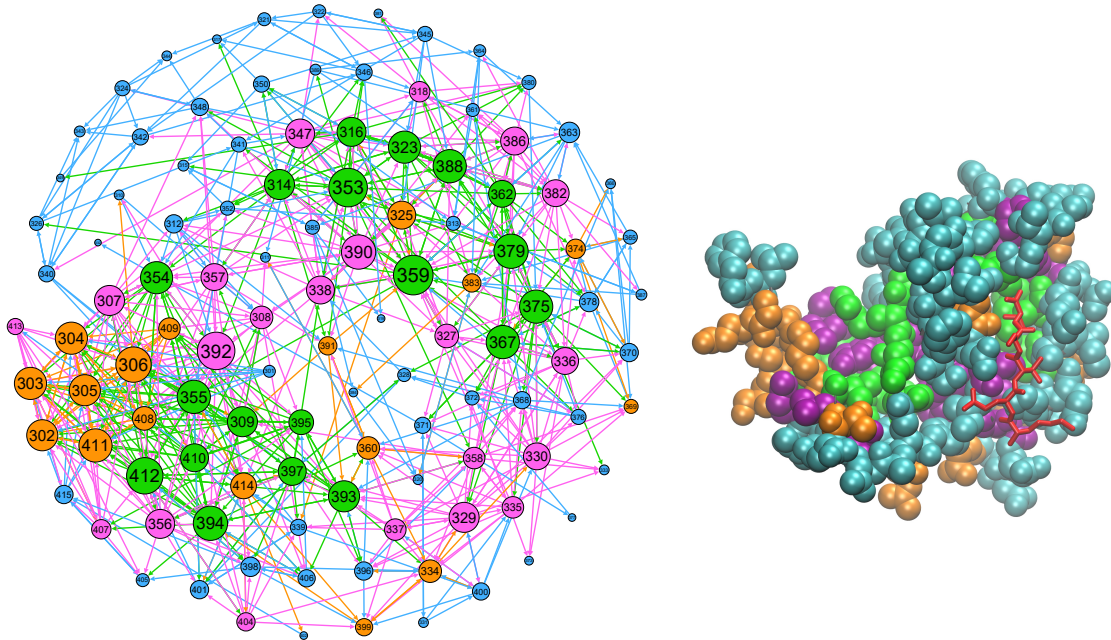


Figure 11.3: The GCAT network of PSD95^{pdz3}. The size of each node is proportional to its degree (sum of in-degree and out-degree). The nodes are colored based on the GCAT classification (G: pink, C: blue, A: orange, T: green). The amino acids are colored accordingly in the protein structure (PDB: 1BE9). The ligand is represented in red in the protein structure.

The network density is a measure of how well a network is connected, as it provides the fraction of links (arcs in a directed network) that are present in the network among all the possible links (arcs) that correspond to all possible pairs of nodes. In an undirected network, the maximal number of links is $N(N-1)/2$, with N the number of nodes in the network, and the density is calculated as (Eq. 11.1):

$$\rho_{undir} = \frac{2m}{N(N-1)} \quad (11.1)$$

with m the number of links in the network. In a directed network, the maximal number

of arcs is $N(N - 1)$ and the density is calculated as (Eq. 11.2):

$$\rho_{dir} = \frac{m}{N(N - 1)}. \quad (11.2)$$

The density of the GCAT network of PSD95^{pdz3}, calculated using N equals to the number of amino acids, is $\rho_{dir,GCAT} = 0.06$. If the direction of the arcs is discarded, the density is $\rho_{undir,GCAT} = 0.12$. For comparison, of the AAN of PSD95^{pdz3} is $\rho_{undir,AAN} = 0.09$. This shows that the GCAT network and the AAN have similar connectivity, with the GCAT network slightly more connected than the WT network.

The fact that $\rho_{undir,GCAT} > \rho_{undir,AAN}$ is consistent with the fact that arcs in the GCAT network can connect amino acids that are not linked in the AAN. More in detail, the geodesic distance $d_{AAN}(i, j)$ between nodes that are neighbors in the GCAT network ranges from 1 to 3, consistent with short-range to long-range perturbation effects (Chapter 6). Among the 796 $i \rightarrow j$ arcs, 48% have $d_{AAN}(i, j) = 1$, 45% have $d_{AAN}(i, j) = 2$ and 7% have $d_{AAN}(i, j) = 3$.

The fact that the difference between $\rho_{undir,GCAT}$ and $\rho_{undir,AAN}$ is small implies that not all amino acids linked in the AAN are also linked in the GCAT network. Consistently, only 48% of the links in the AAN are present as arcs in the GCAT network, calculated discarding the direction of the arc. This means that when an amino acid i is mutated, its chemical neighbors are not necessarily perturbed, consistent with the existence of Z mutations and weight-perturbation mechanisms (Chapter 6).

We have calculated the fraction of chemical neighbors of an amino acid i that are perturbed upon mutations of i , measured by $k_{out}^{d=1}/k_{AAN}$, where $k_{out}^{d=1}$ is the number of out-neighbors of a node that are also neighbors of the node in the AAN and k_{AAN} is the number of neighbors of the node in the AAN. The distribution of the $k_{out}^{d=1}/k_{AAN}$ ratio among all the nodes in the GCAT network is plotted in Fig. 11.4 and it shows that indeed few amino acids perturb more than 60% of their chemical neighbors when they are mutated.

In summary, the perturbation relations (“ i can perturb j ”) modeled by the GCAT network go beyond chemical interactions (links in the AAN) and chemical interactions do not imply perturbation relations. The next question is whether perturbation relations are symmetrical, i.e. whether “ i can perturb j ” implies that “ j can perturb i ”.

Asymmetric interactions An important difference between the GCAT network and the other structure-based networks presented in the previous chapters is that the GCAT network is directed. Among the 656 pairs of nodes of the GCAT network that are connected by an arc, only 140 (around 20%) are connected by a bidirected arc ($i \leftrightarrow j$). By difference, around 80% of the arcs are not bidirected. This means that if the mutation of an amino acid i perturbs and amino acid j , in 80% of the cases the mutation of j does

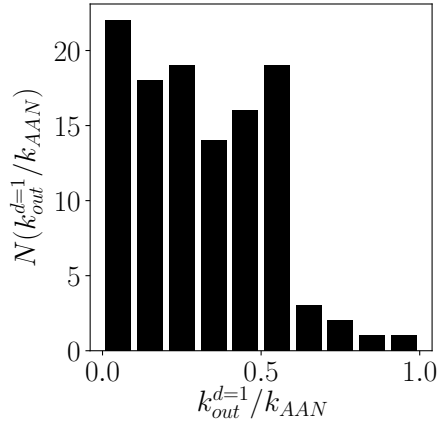


Figure 11.4: Distribution of $k_{out}^{d=1}/k_{AAN}$, that measures the fraction of chemical neighbors of an amino acid i that are perturbed upon mutations of i .

not perturb amino acid i . Among the 140 symmetric arcs, 72% have $d_{AAN}(i, j) = 1$, 26% have $d_{AAN}(i, j) = 2$ and only two arcs have $d_{AAN}(i, j) = 3$. Thus, symmetric arcs are mostly within chemical neighbors. Among all the arcs only half are between first neighbors, meaning that chemical interaction makes a symmetric perturbation relation more likely.

The directionality of the perturbations probed by the arcs in the GCAT network is consistent with the fact that amino acids are anisotropic: upon mutation, only the side-chain atoms of the amino acid are replaced while the backbone atoms are unchanged, and thus the direction of the perturbation is expected to follow the direction of the side-chain. The exact direction of the substituting side-chain is not necessarily the same as the original side-chain. However, due to the steric hindrance caused by the backbone and the side-chains of the sequence first neighbors, the possible side-chain directions span only half of a sphere. Thus, the perturbation caused by the mutation is expected to only propagate to the amino-acids that lie in such half sphere. This is consistent with the fact that in general not all chemical neighbors of an amino acid i are perturbed by the $i \rightarrow i'$ mutation (Fig. 11.4).

Thanks to the directionality introduced in the GCAT network, the mutual amino-acid interactions represented by links in the Amino Acid Network (AAN) are here replaced by asymmetric interactions. Because of this asymmetry, the paths between nodes in the GCAT network are also directed. As a consequence, two distances between two nodes i and j are defined: the shortest path from i to j and the shortest path from j to i . As an example, Fig. 11.5 reports the shortest paths between nodes 330 and 372. The arc $330 \rightarrow 372$ exists in the GCAT network (Fig. 11.1), so the node 372 is at distance equal to one from node 330. On the contrary, the shortest path from node 372 to node 330 is of length five (i.e. 330 is at geodesic distance five from 372), and it involves amino acids far from the GLY330 and HIS372 amino acids in the protein structure (Fig. 11.5). The

biological interpretation of paths in the GCAT network is discussed later in the chapter.

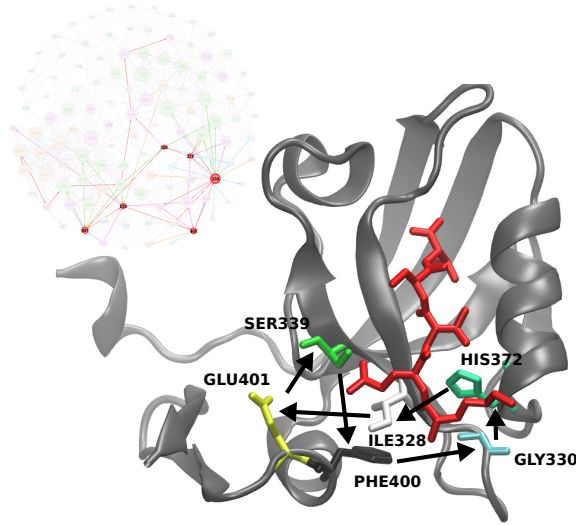


Figure 11.5: The directed shortest paths from position 330 to position 372 and vice-versa in the GCAT network.

Degrees of the GCAT network The average in- and out-degree in the GCAT network is $\bar{k} = 7$. Fig. 11.6 reports the in- and out-degree for each node in the GCAT network, showing that the in- and out-degree spread a broad range of values, from 0 to 23. Moreover, the in- and out-degree of a node are not correlated ($R^2 = 0.07$).

The absence of correlation between the in- and out-degrees is not surprising, because of the directionality of the perturbations. Let us say that the mutation of an amino acid i can only perturb amino acids that lie in half a sphere denoted $HS(i)$. The number of amino acids that are perturbed by i ($k_{out,i}$) is expected to be proportional to the number of amino acids that lie in $HS(i)$. Conversely, the number of amino acids that can perturb i ($k_{in,i}$) is expected to be proportional to the number of half-spheres $HS(j)$ in which i lies. There is no reason to assume that these two quantities are correlated.

The directions of the side-chains of amino acids are not modeled by the AAN. Nevertheless, the AAN measures packing in the protein structure, and thus it may potentially provide information on the number of amino acids that can perturb an amino acid i ($k_{in,i}$) and the number of amino acids that the amino acid i perturbs ($k_{out,i}$). We have investigated whether the in-degree k_{in} , the out-degree k_{out} and the total degree $k_{tot} = k_{in} + k_{out}$ in the GCAT network are determined by the degree in the AAN (k_{AAN}). Fig. 11.7 reports the histograms of (k_{in}, k_{AAN}) , (k_{out}, k_{AAN}) and (k_{tot}, k_{AAN}) , as well as the linear regressions of k_{in} , k_{out} and k_{tot} with k_{AAN} . It shows that the degrees in the GCAT network are only weakly correlated with the degree in the AAN, discarding the hypothesis that the amino-acid packing around an amino acid determines how many amino acids it can perturb or how many amino acids can perturb it. Even lower correlations are obtained

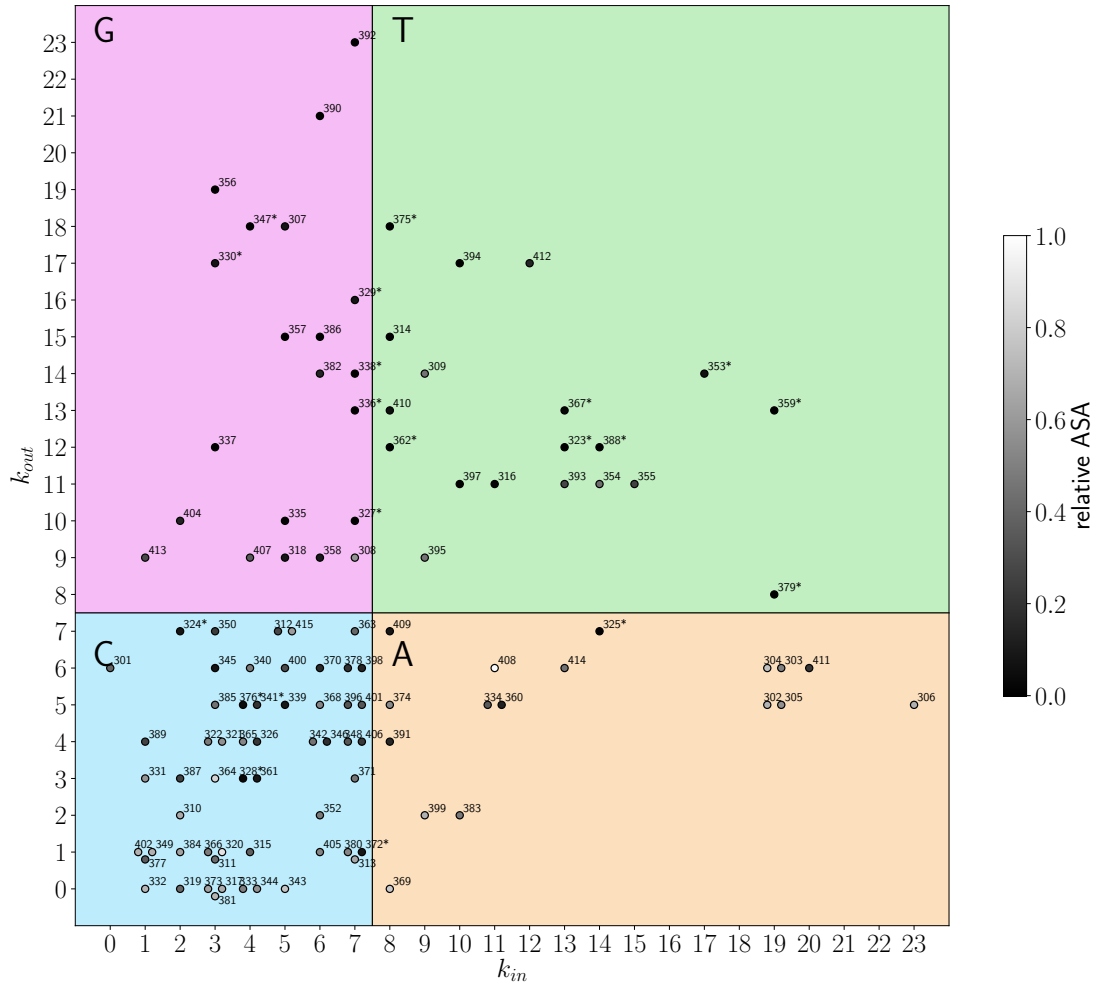


Figure 11.6: In- and out-degree for all the nodes of the GCAT network of the PSD95^{pdz3} protein. The points are colored according to the the Accessible Surface Area (ASA) of the amino acid. Functionally-sensitive positions (see main text) are indicated with a star.

between the atomic packing of the amino acid (w_i in the AAN) and k_{in} , k_{out} and k_{tot} ($R^2 = 0.03$, 0.20 and 0.16 , respectively, plots not shown).

The results of Fig. 11.7 show that measuring packing at chemical distance is not sufficient to determine the perturbative potential of amino-acid mutations. This is consistent with the view of perturbations caused by amino-acid mutations as multi-scale phenomena, as developed in the previous chapters, suggesting that the impact of mutations is determined by the multi-scale structural arrangement of atoms around amino acids, i.e. from the chemical neighborhood to further distances. A possible strategy towards the prediction of the impact of mutations from the WT protein structure only would then be measuring the multi-scale structural arrangement of amino acids in protein structures from AANs created at variable cutoff. However, the asymmetry of arcs in the GCAT network and the fact that the in-degree and out-degree are uncorrelated suggest that the direction of side-chains (i.e. the anisotropy of the amino-acids) should also be taken into consideration as an indication of the possible perturbation direction. This requires con-

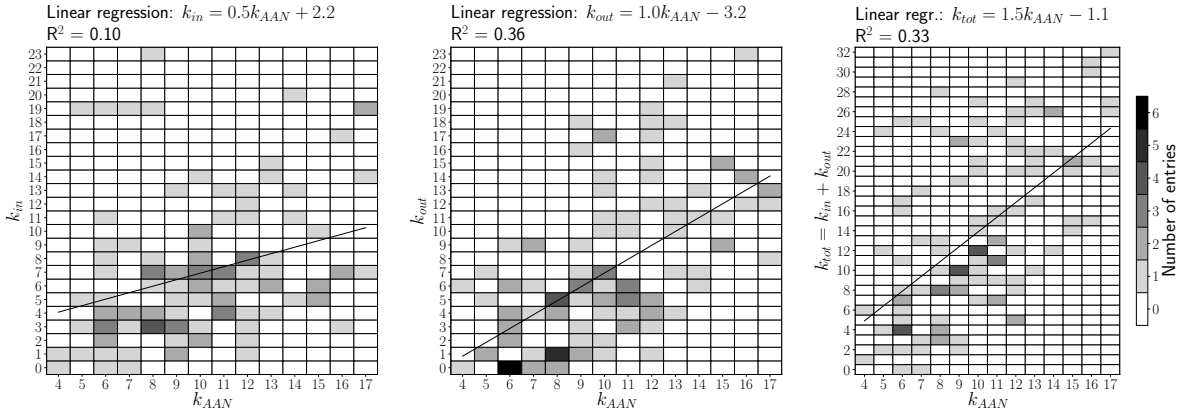


Figure 11.7: Relation between the in-degree k_{in} , out-degree k_{out} and total degree $k_{tot} = k_{in} + k_{out}$ in the GCAT network with the degree in the Amino Acid Network k_{AAN} . From left to right, the histograms of (k_{in}, k_{AAN}) , (k_{out}, k_{AAN}) and (k_{tot}, k_{AAN}) , respectively. The black lines represent linear regressions.

ceiving a new model for protein structures that keeps the information on the side-chain directions, and will not be assessed in this study.

The GCAT classes Because the in- and out-degree of the nodes are not correlated, all the GCAT classes can be populated. Among the 113 nodes of the GCAT network, 21 belong to the G class, 56 belong to the C class, 17 belong to the A class and 19 belong to the T class (Fig. 11.3). This result means that around half of the amino acids in the PSD95^{pdz3} structure perturb rarely and are rarely perturbed (C class). In the other half, the nodes distribute uniformly in the three remaining classes (G, A and T).

Surface-exposed versus buried positions We have investigated whether the GCAT class of an amino acid is determined by its position in the protein structure, buried^b or surface-exposed. In Fig. 11.6 the points correspond to the amino acids of PSD95^{pdz3} colored based on their relative Accessible Surface Area (rASA), with lighter colors corresponding to surface-exposed amino acid positions. Table 11.1 reports the partitioning of the surface-exposed and buried amino acids in the GCAT classes, where amino acids are classified as surface-exposed if $rASA > 0.2$ and buried otherwise. According to this classification, 62 positions are surface-exposed and 51 positions are buried in PSD95^{pdz3}.

Surface-exposed amino acids populate mostly the C class (67.7% of surface-exposed amino acids are C) and most of the C nodes correspond to surface-exposed amino acids (75.0%). Only 19.4% of the surface-exposed amino acids are A, but 70.6% of the A nodes correspond to surface-exposed amino acids. Only 4.8% and 8.1% of the surface-exposed amino acids belong to the G and T classes, respectively.

The distribution of buried amino acids in the GCAT classes is different compared to the surface-exposed amino acids. Buried amino acids populate rather homogeneously

^bBuried amino acids are also called core amino acids in the literature. We do not use the term core amino acids to prevent confusion with the core of the GCAT network.

Table 11.1: Partition of the surface-exposed and buried amino acids into the GCAT classes. For each GCAT class (G, C, A, T) and each category (surface-exposed or buried), the number of occurrences (n), the percentage in the GCAT class (%↓) and the percentage in the category (%→) are reported.

	G			C			A			T			tot		
	n	%↓	%→	n	%↓	%→	n	%↓	%→	n	%↓	%→	n	%↓	%→
surface	3	14.3	4.8	42	75.0	67.7	12	70.6	19.4	5	26.3	8.1	62	54.8	100
buried	18	85.7	35.3	14	25.0	27.5	5	29.4	9.8	14	73.7	27.5	51	45.1	100
tot	21	100	18.6	56	100	49.6	17	100	15.0	19	100	16.8	113	100	100

the G, C and T classes (35.5%, 27.5% and 27.5%, respectively), with a slightly higher representation of the G class. However, concerning the C and G classes, it should be noted that buried amino acids correspond to only 25% of the C nodes, while buried amino acids correspond to 86% of the G nodes. The A class is under-represented in buried amino acids, with only 9.8% of buried amino acids classified as A.

In summary, the C and A classes, corresponding to amino acids that perturb few amino acids, are roughly associated with surface-exposed amino acids. Inversely, the G and T classes, corresponding to amino acids that perturb many amino acids, are roughly associated to buried amino acids. However, all the GCAT classes are populated by both surface-exposed and buried amino acids. This means that the potential of an amino acid to perturb or to be perturbed is not strictly determined by its position in the protein structure.

The GCAT core To investigate whether the buried amino acids are more well-connected to each other compared to surface amino acids, we have extracted the k -cores of the GCAT network. K -cores are the maximal subsets of nodes such that each node in the subset is connected to at least k other nodes in the subset [79]. For a k -core, k is called the degree of the core. By definition, the nodes belonging to a core of degree k have degree greater or equal to k . Please note that the direction of the arc is not taken into consideration, so that the nodes belonging to a k -core have $k_{tot} = k_{in} + k_{out} \geq k$. The maximal core, or inner core, of the network is the core of highest degree k . Fig. 11.8 shows the GCAT with nodes colored based on the k -core to which they belong. Darker color corresponds to larger k . The presence of a core-periphery structure is visible, with a subset of nodes (dark color in Fig. 11.8) corresponding to the core of the network, and a subset of nodes (light color in Fig. 11.8) corresponding to the periphery. The inner core of the GCAT network of PSD95^{pdz3} is the 13-core.

It should be noted that a core-periphery structure is not a direct consequence of the presence of high-degree nodes. Indeed, if high-degree nodes were far from each other in the network, the maximal core would have a low degree k , sign of the absence of a core-periphery structure. However, the maximal possible distance between high-degree nodes depends on the fraction of high-degree nodes in the network: if a high fraction

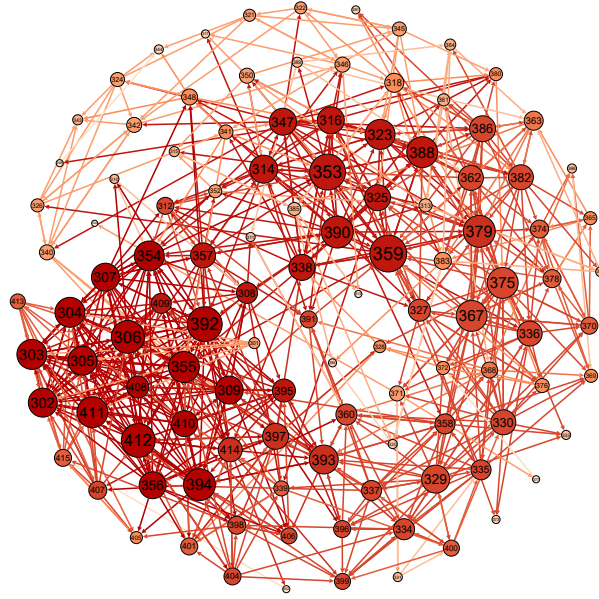


Figure 11.8: Core-periphery structure of the GCAT network of PSD95^{pdz3}. The nodes are colored based on the k -core to which they belong, with darker color corresponding to larger k . The size of each node is proportional to its degree (sum of in-degree and out-degree).

of high-degree nodes is present, then high-degree nodes are necessarily well-connected in the network. To verify whether the core-periphery structure of the GCAT network is a consequence of the number of high-degree nodes, we have calculated the k -cores of a randomized version of the GCAT network created by keeping the total degree of the nodes unchanged but shuffling the links. For the randomized GCAT network, the inner core resulted to be made of 52 nodes and had degree $k = 11$, relatively high but lower than the degree of the inner core of the GCAT network. This means that a core-periphery structure for the GCAT network is favored by the presence of high-degree nodes, but is not entirely explained by the distribution of the degrees of the nodes.

The inner core of the GCAT network (13-core) is composed by seven A nodes, six T nodes and three G nodes (Fig. 11.9). These nodes are part of the N- and C-terminal domains, the 354-356 segment, plus the amino acids 392 and 394. Among the sixteen amino acids belonging to the 13-core, seven are buried (303, 307, 356, 392, 394, 410 and 412) and nine are surface-exposed (302, 304, 305, 306, 309, 354, 355, 408 and 411). This shows that the GCAT core does not correspond to the structural core (buried amino acids) of the protein. Thus, despite the fact that the G and T classes (high out-degree) are roughly associated to buried amino acids, the subset of most-connected nodes measured by the inner core of the GCAT does not correspond to buried amino acids.

Functionally-sensitive positions In 2012, McLaughlin and collaborators have performed a comprehensive single mutagenesis analysis of the PSD95^{pdz3} protein, where the functional impact of all the possible single-amino acid mutants was assessed [16]. They discovered that the mutation of 20 positions over the 83 analyzed sequence positions of

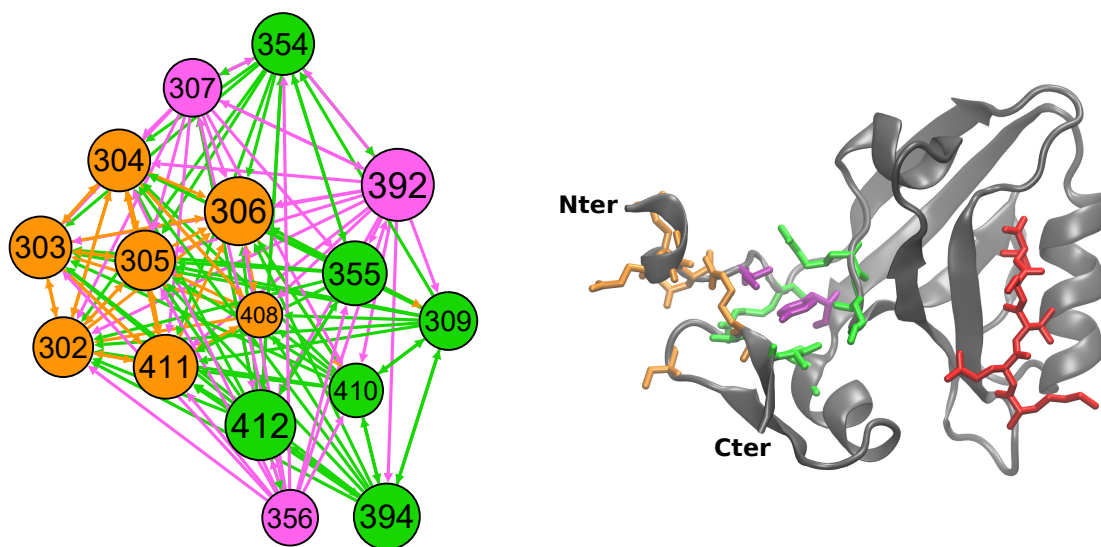


Figure 11.9: The inner core (13-core) of the GCAT network of PSD95^{pdz3}. The size of each node is proportional to its degree (sum of in-degree and out-degree). The nodes are colored based on the GCAT classification (G: pink, C: blue, A: orange, T: green). In the protein structure, the red molecule is the ligand.

PSD95^{pdz3} (amino acids 311 to 393) caused a loss of binding activity to the ligand of PSD95^{pdz3} compared to the WT protein. As in Chapter 6, we refer to these positions as functionally-sensitive positions. Functionally-sensitive positions are indicated with a star in Fig. 11.6.

Fig. 11.10 shows the sub-network of the GCAT network involving functionally-sensitive positions, colored according to the GCAT class they belong to.

The sub-network of functionally-sensitive positions is connected in the GCAT network, meaning that functionally-sensitive positions tend to perturb each other when mutated. Functionally-sensitive positions populate all the four GCAT classes: six are G, five are C, one is A and eight are T. The fact that functionally-sensitive positions may be of any GCAT class means that the functional impact of mutations is not determined by the amount of perturbations provoked by the mutation (otherwise all functionally-sensitive positions would be G or T), but by which amino acids are perturbed by the mutation. This is consistent with the finding that functionally-sensitive positions belong to the M class (their mutations impact from the local scale to the large-scale, depending on the specific mutation), as described in Chapter 6. However, it should be noted that the only A node and four over the five C nodes are hotspots, i.e. directly bound to the ligand, suggesting that the mechanisms of functional-change of A and C functionally-sensitive positions is different from the one of G and T functionally-sensitive positions. Hotspots are indicated with a black arrow in Fig. 11.10.

In the sub-network of functionally-sensitive positions of the GCAT network, all non-hotspot nodes apart from node 341 have an arc that points directly to a hotspot position

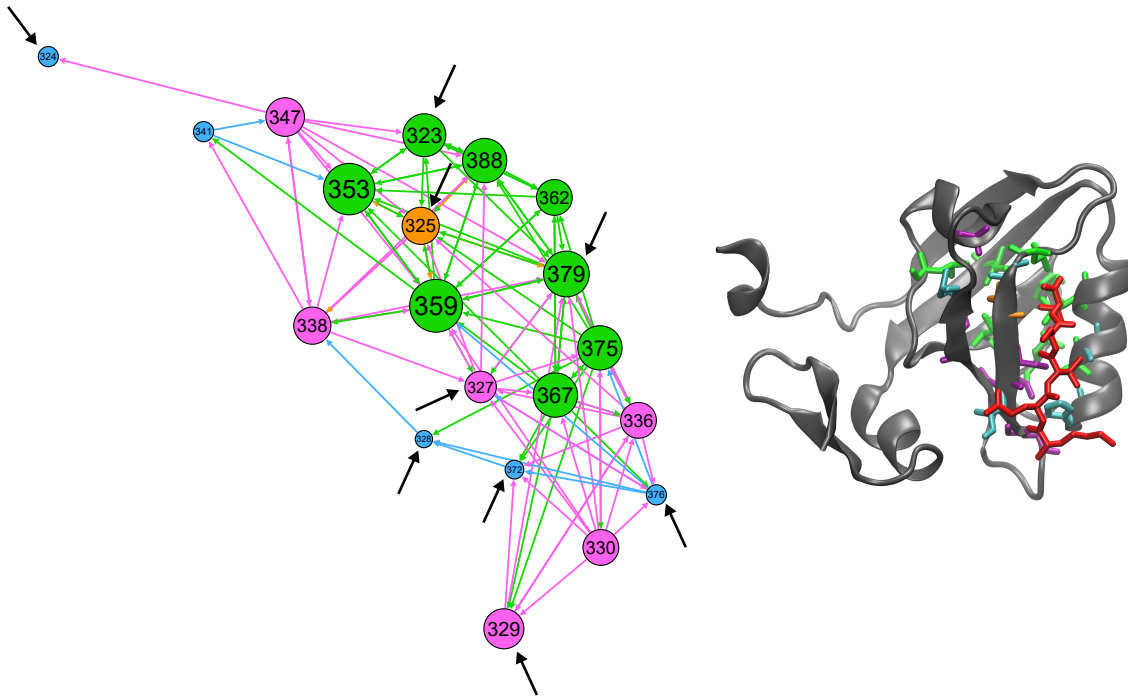


Figure 11.10: The sub-network of functionally-sensitive positions (functionally-sensitive positions) of the GCAT network of PSD95^{pdz3}. Black arrows indicate hotspots, i.e. nodes corresponding to amino acids directly bound to the ligand in the protein structure. The nodes are colored based on the GCAT classification (G: pink, C: blue, A: orange, T: green).

(Fig. 11.10). This means that mutating the amino acids corresponding to these nodes causes a change in neighborhood of the hotspot positions, explaining the change in ligand-binding activity, as was verified for the G330T mutation [72] (Fig. 11.1).

Paths in the GCAT network and allostery The non-hotspot functionally-sensitive 341 node has no arc that points directly to a hotspot node (Fig. 11.10). However, it is connected to the 323, 324 and 325 hotspot nodes through the 347 and 353 nodes, according to the directed paths $341 \rightarrow 347 \rightarrow 323$, $341 \rightarrow 347 \rightarrow 324$ and $341 \rightarrow 353 \rightarrow 325$ (Fig. 11.10). These paths are represented in the protein structure in Fig. 11.11. This example suggests that paths in the GCAT network may highlight allosteric paths in the protein structure, i.e. chains of sequential rearrangements of atomic interactions that dynamically lead to a change in protein function. Such rearrangements of atomic interactions leading to the perturbation of the functionally-sensitive hotspots from the mutation of residue 324 need to be dynamic, otherwise the arcs from node 324 to the functionally-sensitive hotspots would be present in the GCAT network. The underlying hypothesis is that the perturbations probed by the amino acid mutations, e.g. the perturbation of residue 323 caused by the mutation of residue 347, also represent paths of atomic motions. This would explain why 341 is a functionally-sensitive position. This hypothesis is supported by the increasing evidence that amino acid mutations perturb protein structures along the same vibrational modes that control the native protein structural dynamics [220, 221].

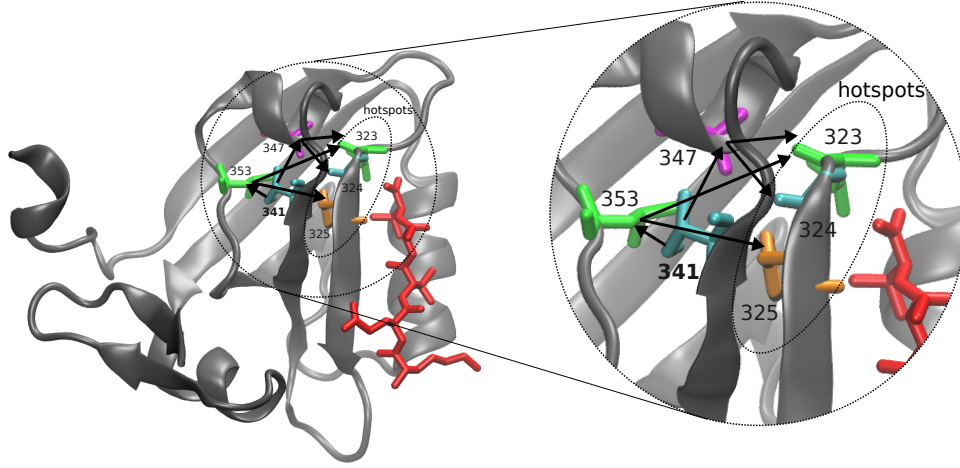


Figure 11.11: The paths from the functionally-sensitive position 341 to functionally-sensitive hotspots in the GCAT network.

If arcs and paths in the GCAT network highlight allosteric paths in the protein structure, then the inner core of the GCAT network, that is highly connected, should be involved in allosteric mechanisms. Consistently, the C-terminal and N-terminal domains of PSD95^{pdz3}, that are part of the GCAT inner core (Fig. 11.9) were shown to control the protein function via dynamic allostery [222–224]. It is possible that the C-terminal and N-terminal control the dynamics of the binding site of PSD95^{pdz3} through the 354, 355, 356, 392 and 394 amino acids, that are also in the GCAT core, are spatially closer to the binding site compared to the C-terminal and belong to the G and T classes (high out-degree). Obviously, this hypothesis should be verified on other cases of dynamic allosteric control of the protein dynamics and function.

Paths in the GCAT network and inter-dependent mutations According to the hypothesis that arcs in the GCAT network correspond to paths of atomic motions, paths in the GCAT network have been interpreted as possible allosteric paths in the protein structure. However, another interpretation is possible and is described below.

If an arc $i \rightarrow j$ is present in the GCAT network, it means that when the amino acid i is mutated, then amino acid j is in a perturbed state. Thus, an arc $i \rightarrow j$ in the GCAT network means that the amino-acid neighborhood of j is determined by the amino-acid type at position i , and the arc $i \rightarrow j$ may be interpreted as node i controlling the state of node j . Let us now consider two arcs $i \rightarrow j$ and $k \rightarrow l$. It means that when i is mutated, then j is in a perturbed state j^* and when k is mutated, then l is in a perturbed state l^* . If $i \rightarrow j$ and $k \rightarrow l$ are disjoint in the GCAT network, then if both i and k are mutated we expect j and l to be in their perturbed states j^* and l^* .

Let us now consider the case where k coincides with j , that is the path $i \rightarrow j \rightarrow l$ is present in the GCAT network. It means that when i is mutated, then j is in a perturbed state j^* and when j is mutated, then l is in a perturbed state l^* . If both amino acids i and

j are mutated, then what is the state of amino acid l ? l may be in the perturbed state l^* , in a different perturbed state l^{**} , or even in its unperturbed state l , if the mutation of i corrects the perturbation caused by the mutation of j . If l is in the unperturbed state l or in the perturbed state $l^{**} \neq l^*$, we say that i and j are inter-dependent, as the state of l depends on the amino-acid type at both positions i and j .

Based on this reasoning, arcs in the GCAT network may underline inter-dependence between amino-acid mutations. As defined in the introduction, we say that two amino-acid position are in cooperative interaction if the impacts of their mutations are inter-dependent.

Salinas and Ranganathan have measured cooperativity between the amino acids belonging to the second helix of PSD95^{pdz3}, that is part of the binding pocket (helix2, amino acid positions 372 to 379) [225]. Cooperative interactions between all pairs of these amino acids were measured experimentally based on the impact of single- and double-mutations on the binding free energy. They found cooperative interactions for all two-by-two combinations of amino acids in positions 372, 375, 376 and 379, plus the pairs 374-375 and 378-379.

To investigate the hypothesis that arcs in the GCAT network underline cooperative interactions, we have analyzed the sub-network of the GCAT network involving amino acids of the helix2. This sub-network is represented in Fig. 11.12.

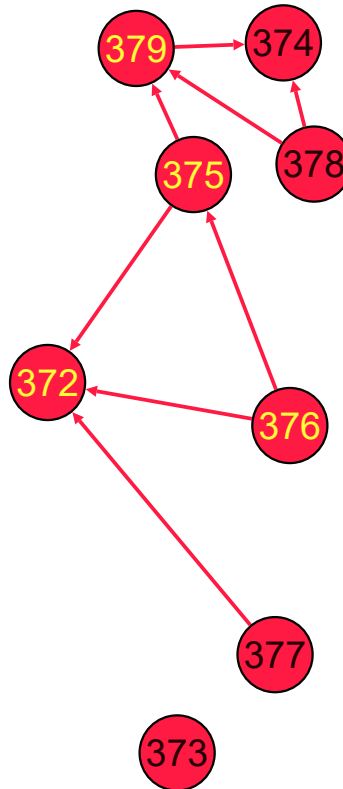


Figure 11.12: Sub-network of the GCAT network of the amino acids of the α -helix of the binding pocket of PSD95^{pdz3}. Functionally-sensitive positions are highlighted in yellow.

Consistent with the hypothesis, most cooperative interactions $i - j$ measured experimentally correspond to arcs $i \rightarrow j$ or $j \rightarrow i$ in the GCAT network. Amino acids 374 and 375, that are in cooperative interaction, are not connected by an arc, but are connected through amino acid 379 via the $375 \rightarrow 379 \rightarrow 374$ path. Similarly, the amino acids 376 and 379 are not connected by an arc, but are connected via the $376 \rightarrow 375 \rightarrow 379$ path. This suggests that cooperative interactions may be picked up by arcs and paths in the GCAT network. Only the cooperative interaction between amino acids 372 and 378 is not picked up by any path. However, 372 is a functionally-sensitive hotspot (Fig. 11.10) and amino acid 378 perturbs the functionally-sensitive hotspot 379 (Figs 11.10 and 11.12). It is possible that the cooperative interaction observed between amino acids 372 and 378 derives from the perturbation of the interaction between the two hotspots 372 and 379 with the ligand, not modeled by the GCAT network.

Amino acid 373, that is not involved in any in cooperative interaction with the other amino acids of the helix, is disconnected from the rest of the sub-network of Fig. 11.12.

Some “extra” arcs and paths connecting amino acids that are not in cooperative interaction are present in the GCAT network, e.g. the arcs $378 \rightarrow 374$, $379 \rightarrow 374$ and $377 \rightarrow 372$. It should be noted that in [225] cooperativity was measured based on the functional impact of mutations. These “extra arcs” involve two non-functionally sensitive positions (378 and 374) and only one functionally sensitive position (372). If the mutations of amino acids 378 and 374 do not have an impact of the protein function, then measuring binding energy would not pick up any cooperative interactions involving amino acids 378, 374 and 372. Yet, the structural or dynamical impact of the double mutations may be different from the sum of the structural or dynamical impact of the single mutations involving amino acids 378, 374 and 372. If this is the case, the couples 378-374, 379-374 and 377-372 could have inter-dependent impacts on the protein structure or dynamics, i.e. be in cooperative interaction, but do not impact the protein function.

Paths in the GCAT network and evolvability Cooperative interactions between amino-acid mutations are a mechanism for functional change, or evolvability, of proteins. As an example, let us consider the mutations G330T and H372A of PSD95^{pdz3}, that are functionally-sensitive. H372 is a hotspot while G330 is not. The arc $330 \rightarrow 372$ is present in the GCAT network (Fig. 11.1). In [16], the binding affinity of the WT protein, the G330T and H372A single mutants and G330T/H372A double mutant to two ligands were measured. The two ligands are the ligand of the WT protein and a mutated version. The WT protein binds to the WT ligand with high affinity and to the mutated ligand with low affinity. It was shown that G330T makes the protein lose the specificity of ligand binding (the protein binds to the two ligands with high affinity), while H372A makes the protein lose the affinity with the WT ligand and gain affinity with the mutated ligand (partial specificity, 14-fold preference for the mutated ligand). The double mutation

G330T/H372A was shown to provide a complete specificity switch: the double mutant has high affinity to the mutated ligand and low affinity to the WT ligand (45-fold preference for the mutated ligand) [16]. In a later study, it was proposed that because the G330T mutation acts allosterically on the binding site of PSD95^{pdz3} and serves as an intermediate for adaptation (full adaptation is obtained upon the subsequent H372A mutation), then allostery may have its origin in evolvability [72].

In the previous paragraphs, paths in the GCAT network were proposed to highlight allosteric paths and cooperative interactions among the amino acids. These two interpretations fit well with the view of allostery as evolutionary-driven proposed in [72]: paths in the GCAT network could represent the inter-dependence between amino-acid mutations (cooperativity between amino acid positions), that exploits allosteric mechanisms and controls the protein evolvability when the protein function is impacted by the mutations.

Paths in the GCAT network and co-evolution Ranganathan and collaborators have shown that the functionally-sensitive positions of PSD95^{pdz3} are part of the protein sector of PSD95^{pdz3}, i.e. a sparse network of co-evolving amino-acid positions [16, 72, 226]. In [16] it was concluded that the sector architecture may correspond to a design for adaptability to environmental changes, because the mutation of some functionally-sensitive positions leads to a change in binding affinity from the WT ligand to its mutated version.

If arcs and paths in the GCAT network correspond to cooperative interactions between amino acids and protein evolvability exploits multiple mutations at cooperatively-interacting positions, then co-evolving amino acid positions should be connected in the GCAT network. Consistently, the sub-network of functionally-sensitive positions is connected (Fig. 11.10).

Thus, an additional interpretation of paths in the GCAT network might be that paths connect co-evolving amino acids. Again, this interpretation fits well with the previous interpretations of paths underlying allosteric paths or cooperative interactions, because allostery, cooperativity and co-evolution are not independent.

The fact that surface-exposed amino acids are mostly C and A (low out-degree) is consistent with the observation that mutations at surface-exposed residues cumulate during the evolution of protein structures and prepare the protein structure for subsequent mutations of buried amino acids, i.e. these surface and buried positions are in cooperative interaction: the outcome of the mutation at the buried positions is dependent on the mutations that have taken place on the surface [227]. Indeed, mutations of A and C nodes are not expected to provoke large perturbations, so they could appear as neutral mutations. However, if these positions have some arc or path pointing to highly-perturbative buried G and T amino acids, these highly perturbative positions would be in a different state compared to the original structure. Then, subsequent mutation of these amino acids

would have a different impact on the protein structure, dynamics or function compared to when performed in the original-sequence background. Such cooperative interaction between surface-exposed and buried amino acids could result in functional change (evolvability) or in functional robustness, depending on the effect of the mutation on the protein function.

Paths in the GCAT network and error-correction While co-evolution of amino acid positions may be driven by the adaptability to a novel environment (functional change or evolvability, as the G330T/H372A double mutation example), co-evolution may also be a manifestation of error-correction mechanisms. In fact, error-correction is another manifestation of cooperative interactions between amino acids. If a single amino-acid mutation perturbs the protein structure, dynamics or function, a rescue mutation at a different position may correct the error introduced by the first mutation and restore the original protein structure, dynamics or function.

Let us say that a mutation $i \rightarrow i'$ perturbs the protein structure at some amino acid positions $J = \{j | j \text{ is perturbed by } i \rightarrow i'\}$. This means that upon the $i \rightarrow i'$ mutation, the amino acids belonging to J lose or gain neighbors in the Amino Acid Network (AAN). By definition, J is the set of out-neighbors of node i in the GCAT network of the protein structure. We denote it $N_{out}(i) = J = \{j | j \text{ is perturbed by } i \rightarrow i'\}$, where $N_{out}(i)$ is the out-neighborhood of node i . How could the perturbation $i \rightarrow i'$ be compensated by a rescue mutation $k \rightarrow k'$ at some amino acid position k ? Some hypotheses can be drawn:

1. The set of perturbed amino acids $N_{out}(i)$ corresponds to the amino acid k and its neighbors. Mutating k allows restoring the amino-acid interactions that were lost or gained upon the mutation $i \rightarrow i'$. This requires an arc $i \rightarrow k$ in the GCAT network.
2. The mutation $k \rightarrow k'$ impacts the same set of amino acids $N_{out}(i)$ and restores the amino-acid interactions that were lost or gained upon the mutation $i \rightarrow i'$ among the amino acids in $N_{out}(i)$. This requires $N_{out}(i) \cap N_{out}(k) \neq \emptyset$ in the GCAT network. This is measured by the Jaccard similarity coefficient of the out-neighborhoods of i and k , that is equal to 1 when i and k have the same neighbors ($N_{out}(i) \equiv N_{out}(k)$) and is equal to 0 when i and k do not share any neighbors ($N_{out}(i) \cap N_{out}(k) = \emptyset$).
3. The mutation $k \rightarrow k'$ impacts the amino acid i so that i is in a perturbed state i^* and the mutation $i^* \rightarrow i'$ does not have the same impact as the mutation $i \rightarrow i'$, i.e. i and k are in epistatic interaction. This requires an arc $k \rightarrow i$ in the GCAT network.

Obviously, the three conditions (arc $i \rightarrow k$, or $N_{out}(i) \cap N_{out}(k) \neq \emptyset$, or arc $k \rightarrow i$) are not sufficient for a rescue mechanism, because rescue requires either the recovery of the wild-type atomic interactions lost upon the mutation $i \rightarrow i'$ or that an alternative solution that encodes for the same structure, dynamics or function is obtained. The IPN of the $i \rightarrow i'$ and $k \rightarrow k'$ mutants should be investigated to determine whether

a rescue mechanism that restores the wild-type atomic interactions is possible. Rescue mechanisms involving alternative solutions would be harder to pick up, because the set of solutions that provide the same structure, dynamics of function to the protein is unknown. Nevertheless, the GCAT network may be exploited to filter out candidate positions k that have the potential to provide a rescue mechanism for the $i \rightarrow i'$ mutation when mutated. According to the hypotheses drawn above, these candidate positions would be:

1. All out-neighbors of i (an arc $i \rightarrow k$ is present).
2. All nodes that share out-neighbors with i ($N_{out}(i) \cap N_{out}(k) \neq \emptyset$).
3. All in-neighbors of i (an arc $k \rightarrow i$ is present).

A schematic of these conditions is presented in Fig. 11.13a, where the nodes are labeled based on the condition they fulfill.

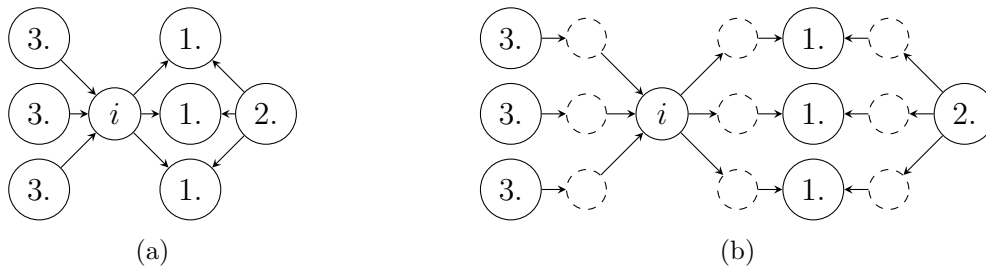


Figure 11.13: Candidate positions whose mutation may provide a rescue mechanism for the mutation of position i . (a): Correction of neighborhood perturbation. The nodes are labeled based on the condition they fulfill: 1. an arc $i \rightarrow k$ is present; 2. $N_{out}(i) \cap N_{out}(k) \neq \emptyset$; 3. an arc $k \rightarrow i$ is present. (b): Correction of dynamics perturbation. The nodes are labeled based on the condition they fulfill: 1. a directed path from node i to node k exists; 2. if a directed path from node i to node j exists, then a directed path from node k to node j exists; 3. a directed path from node k to node i exists.

Importantly, because arcs in the GCAT network involve amino acids that may be far in the protein structure (geodesic distance up to 3 in the AAN), the candidate amino acid positions k that may provide a rescue mechanism for the $i \rightarrow i'$ mutation are not restricted to the amino-acid neighbors of i in the AAN. Moreover, amino acid proximity (link in the AAN) does not imply mutual perturbation, as proven by the fact that 80% of the links in the GCAT network are asymmetric. Thus, the GCAT network seems a more appropriate tool than the AAN to explore error-correction and rescue mechanisms for mutations that cause changes in amino-acids neighborhoods.

Drawing hypotheses on rescue mechanisms for dynamical perturbations from the GCAT network is more complicated. If the $i \rightarrow i'$ mutation perturbs the dynamics of the protein structure, it means that the dynamical atomic interactions between amino acids are perturbed. As shown in the previous chapters, dynamical perturbations are tracked by changes in link weights in the amino acid network. Because the GCAT network models changes in neighborhoods caused by mutations and not changes in link weights, arcs in the GCAT network do to directly pinpoint dynamical perturbations that do not provoke any changes in amino-acid neighborhoods. However, it was proposed above that paths in the GCAT network correspond to possible allosteric paths in the protein structure,

i.e. dynamical atomic interactions between amino acids. If this hypothesis holds, then dynamical perturbations caused by a mutation may be picked up by paths in the GCAT network. A rescue mutation for a dynamical perturbation would need to recover the dynamical atomic interactions lost or gained upon the $i \rightarrow i'$ mutation. Then, following the same reasoning as for structural perturbations, candidate rescue mutation positions k may be selected as follows:

1. All nodes that are at finite distance from i (a directed path from node i to node k exists).
2. All nodes from which there exists a path to all nodes that are at finite distance of i (if a directed path from node i to node j exists, then a directed path from node k to node j exists).
3. All nodes for which i is at finite distance (a directed path from node k to node i exists).

A schematic of these conditions is presented in Fig. 11.13b, where the nodes are labeled based on the condition they fulfill.

However, more complex error-correction mechanisms may take place, for example acting on amino acids that lie in the path between amino acid i and a dynamically-perturbed amino acid j . Moreover, a maximal length for paths in the GCAT network to be relevant for dynamical perturbations probably exists.

Validating the proposed rescue-mechanism hypotheses for structural and dynamical perturbations would require an extensive analysis of the structure and dynamics of protein variants with single and multiple amino-acid mutations, and is out of the scope of this study. However, focusing on the simpler case of mutations that cause changes in amino-acid neighborhoods, the GCAT network of PSD95^{pdz3} has been explored in terms of Jaccard similarity between out-neighborhoods.

Fig. 11.14 reports the Jaccard similarity between the out-neighborhoods in the GCAT network of PSD95^{pdz3}. If the set of out neighbors of a node i is $N_{out}(i)$, the Jaccard similarity between two nodes i and j is defined as (Eq.11.3):

$$Jaccard(i, j) = \frac{|N_{out}(i) \cap N_{out}(j)|}{|N_{out}(i) \cup N_{out}(j)|} \quad (11.3)$$

with $|N_{out}(i) \cap N_{out}(j)|$ the number of common out-neighbors between i and j and $|N_{out}(i) \cup N_{out}(j)|$ the size of the union of $N_{out}(i)$ and $N_{out}(j)$.

High values of Jaccard similarity (dark blue in Fig. 11.14) are observed between the nodes that belong to the inner core of the GCAT network (N-terminal domain, C-terminal domain, and amino acid positions 354, 355, 356, 392, 394, Fig. 11.9). This is not surprising, because the inner core of the network is a sub-network of highly-connected nodes, so there is a high probability that nodes of the inner core share a large number of neighbors.

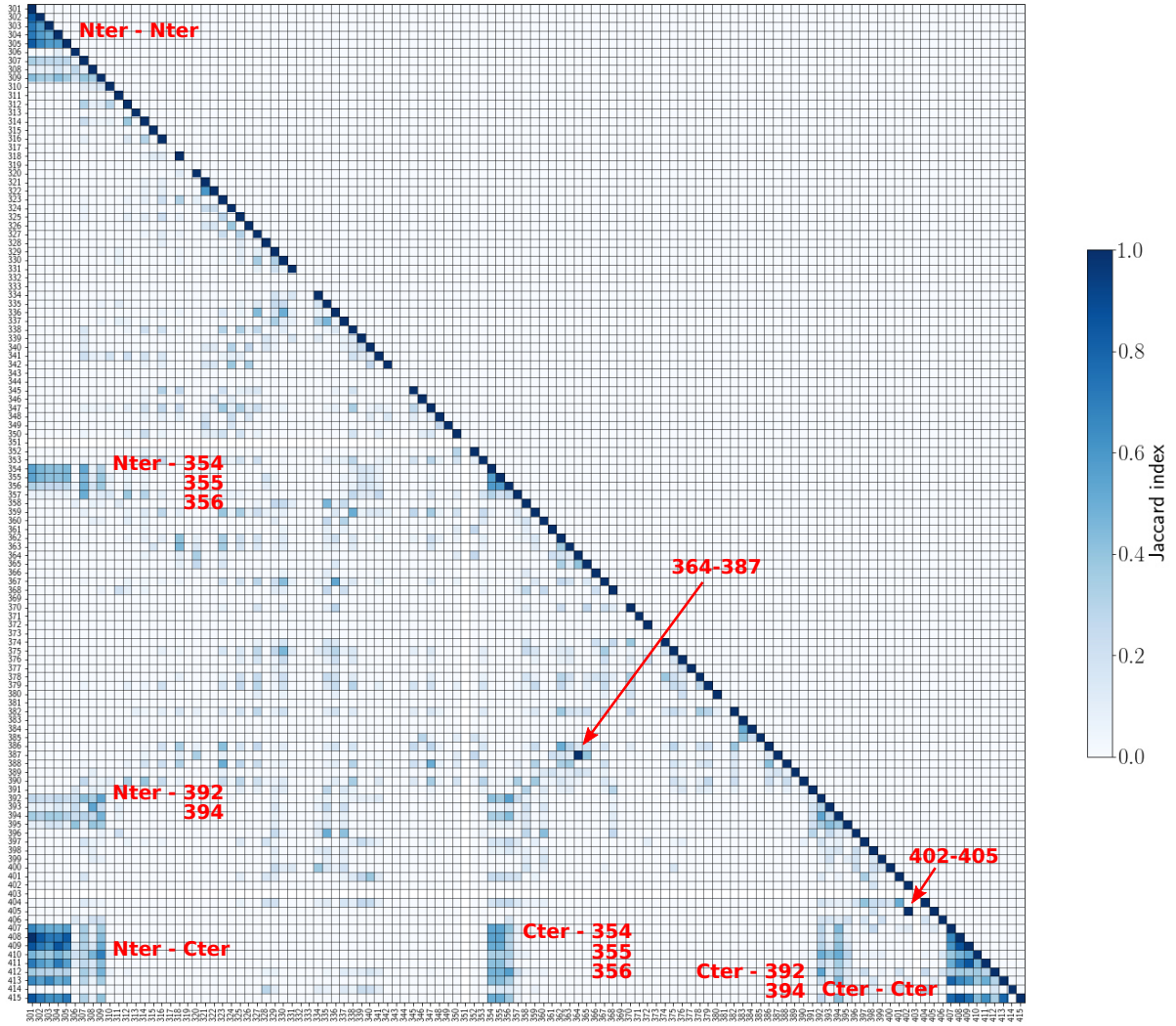


Figure 11.14: Jaccard similarity between out-neighborhoods in the GCAT network of PSD95^{pdz3}. Darker blue is associated to high Jaccard similarity. The Jaccard similarity is equal to 1 in the diagonal (a node has 100% similarity with itself). White entries in the diagonal correspond to nodes that have $k_{out} = 0$, so the Jaccard similarity is not defined.

More surprising is the Jaccard similarity equal to one for two node pairs, (364, 387) and (402, 405), that are not in the inner core of the GCAT network. All these four nodes belong to the C class (low in-degree and low-out degree). Nodes 364 and 387 have $k_{out,364} = k_{out,387} = 3$ and their out-neighbors are the nodes 313, 363 and 383. Nodes 402 and 405 have $k_{out,402} = k_{out,405} = 1$ and their out-neighbor is node 415.

To get an insight on whether the Jaccard similarity measure pinpoints possible error correction mechanisms, we have inspected the structural perturbations caused by the mutations of the G364, T387, R405 and A402 amino acids by comparing the mutants' and WT AANs.

Most amino acid mutations of T387 cause the loss of the (363, 383) contact in the AAN, and the mutations to G, A, C, N, H, Y and K cause the additional loss of the (313, 387) contact in the AAN. Most of the mutations of G364 also cause the loss of the (363,

383) contact. However, the G364M mutation additionally causes the gain of the (313, 364) contact. If the loss of the (313, 387) contact introduced by a mutation of T387 destabilizes the β -sheet tertiary structure, the error may be corrected by the G364M mutation, that would keep the β -strand containing amino acid R313 in place (Fig. 11.15A). Additional mutations would be needed to correct the loss of the (363, 383) 3D contact.

Similarly, the mutations of R405 to G, A, S, D, N and T cause the loss of the (405, 415) contact in the AAN, while the A402F mutation causes the gain of the (402, 405) contact. The loss of interaction between N415 and R405 caused by the mutation of R405 may destabilize the position of the C-terminal domain. The A402F mutation may correct the error by restoring interaction with N415 (Fig. 11.15B). This is particularly relevant, because the C-terminal domain controls the dynamics and binding activity of PSD95^{pdz3} [222], so a structural change at the C-terminal may impact the protein's function, similarly to what proposed for the Mpro of SARS-CoV-2 respect to SARS-CoV (Chapter 8, Section 8.2.3).

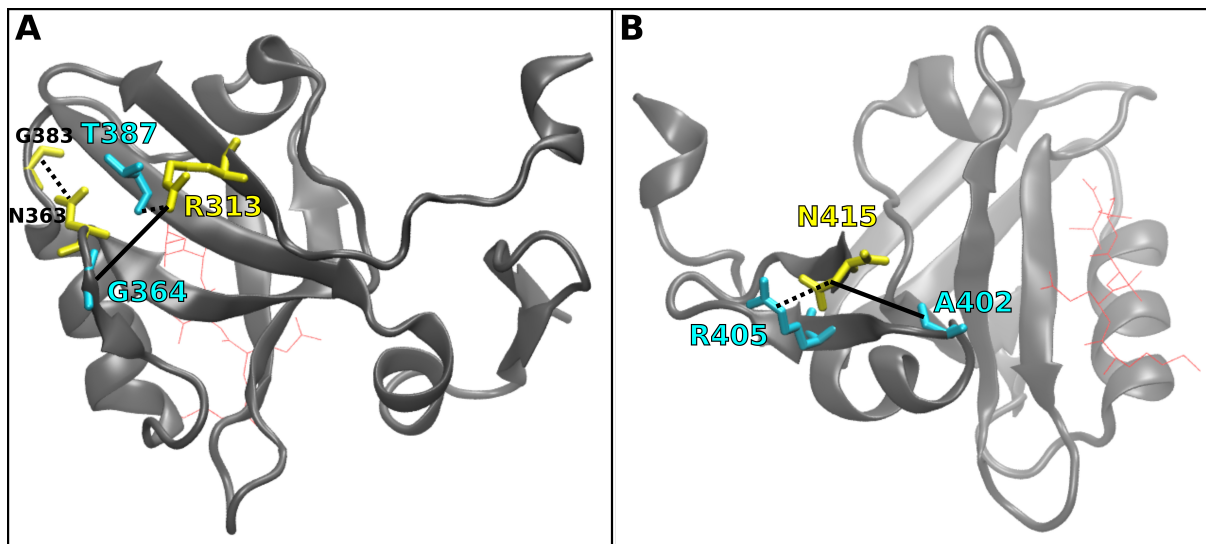


Figure 11.15: Two possible error-correction mechanisms of structural perturbations of PSD95^{pdz3}. Mutated amino acids are in cyan and perturbed amino acids are in yellow. A: the G364M mutation could correct the loss of interaction of R313 with the rest of the β -sheet caused by the mutation of T387. B: the A402F mutation could correct the loss of interaction of the C-terminus (N415) with R405 caused by the mutation of R405.

The two examples of the error-correction mechanisms involving the (364, 387) and (402, 405) pairs show the potential of the GCAT network as a tool to guide the investigation of error-correction mechanisms in protein structures through rescue mutations, at least concerning changes in amino-acid neighborhoods.

11.4 Conclusion

In summary, a novel tool called the GCAT network has been presented in this chapter to model the directional perturbations caused by all possible amino-acid mutations in a

protein structure.

An important finding is that amino acids have different potential to perturb and to be perturbed one from another, measured by the connectivity in the GCAT network and resulting in the population of the four GCAT classes. Moreover, the GCAT network shows a core-periphery topology, where the inner core does not correspond to the structural core of the protein structure. This suggests a complex structural design of the protein governing the response to perturbations.

A relation between paths in the GCAT network and allosteric mechanisms, cooperative interactions between amino acids and co-evolution has been proposed. Moreover, a possible strategy to extract information on error-correcting multiple mutations has been suggested. These hypotheses remain speculative at the moment, but they could be verified using evolutionary information and experimental or simulated dynamics of protein variants with single and multiple amino acid mutations.

As future directions, the statistical analysis of GCAT network properties of proteins with different structure, dynamics and function should be performed. Importantly, this statistical analysis should aim to the definition of the in-degree and out-degree threshold to use for the classification of nodes into the GCAT classes. In this proof of concept, the average degree $\bar{k} = 7$ has been employed, but there is no guarantee that such average degree is universal across GCAT networks of different protein structures. Similarly, the statistical analysis should assess whether the core-periphery structure of the GCAT network is a universal property of globular protein structures or whether it is protein-dependent.

Moreover, a threshold for the the relevancy of path lengths for cooperative interactions and/or allostery should be determined based on experimental evidence. This is a difficult task, because different phenomena may exploit paths of different lengths. Moreover, such thresholds may depend on the specific protein and its role in protein-protein interactions. For example, the analysis of mutations related to the development of a brain tumor revealed that isolated mutations on hub proteins are more disease-causing compared to mutations belonging patches of spatially close-mutations, that may correspond to paths in the GCAT network. On the contrary, mutations belonging to patches are more disease-causing in other proteins [228].

The possibility of weighting the arc of the GCAT network based on the amount of perturbation provoked by the mutations could also be explored to filter out the most relevant information, again relying on experimental data.

Chapter 12

Conclusion and perspectives

The goal of this manuscript was the investigation of the sustainability of proteins and the assessment of the possibility of using protein structures as a model for the design of bio-inspired sustainable urban structures.

The sustainability of a system is here defined as the capacity to maintain the functionality in time. This requires the implementation of two responses to perturbations: maintenance of the functionality (the system is robust) and adaptation of the function to different environments (the system is evolvable).

We focused on spatial systems, i.e. material systems whose functionality relies on the arrangement of the system's components in space. In these cases, the system's structure is defined by the size and shapes of the components, the spatial proximity between components and their relative orientation. These parameters describe how the space is occupied in the structure, i.e. what regions of the space are filled by the system components and what regions are left available for the system's dynamics and to accommodate changes in the system components.

Using a Network Modeling approach, this work contributes to the issue of determining what structural designs provide sustainability to proteins and urban structures. This issue was assessed from the point of view of the system's response to internal perturbation, here limited to substitutions of the system components.

We have proposed to measure space occupancy as a mean to diagnose spatial sustainability, locally (Chapter 4) and at the multi-scale level (Chapter 5). We find that the design of protein structures, contrary to urban structures, always provides empty space. We propose that this is a spatial sustainability feature as this space can be exploited to accommodate substitutions of the components (amino-acid mutations) and dynamics (atomic motions). Dynamical and structural impacts of mutations were measured from changes in interactions between amino-acids, probed by links in the network models. Using an in-silico approach, it was shown that some amino-acid mutations can be accommodated with no need of rearrangement of interactions, while other amino-acid mutations require changes in amino-acid neighborhoods from the local scale to the large scale to maintain structural integrity (Chapter 6). To diagnose the impact of amino-acid mutations on the protein dynamics, it was proposed to measure the changes in allocations of atomic contacts to the four levels of descriptions of protein structures (Chapter 7). Thanks to this measure, it was possible to recover known dynamical differences between B-subunits of AB₅ toxins and Transthyretin variants and to investigate possible

dynamical differences between variants of Main proteases of coronaviruses (Chapter 8). The case of the L55P pathogenic Transthyretin variant was studied more in detail and the analysis of the perturbation of atomic contacts caused by the mutation coupled with a tiling approach allowed the inference of a model of amyloid fiber consistent with experimental evidence and geometrical constraints (Chapter 9). Concerning the B-subunits of AB₅ toxins, comparing the networks of dipoles of the protein structures was successfully applied to the interpretation of the experimental dielectric signal, that probes for the different molecular dynamics of the two toxins (Chapter 10). Finally, a novel tool called the GCAT network was defined to describe the directional perturbation of amino-acid neighborhoods caused by in-silico mutations, opening the possibility to study the contribution of each amino-acid mutation in protein variants where multiple mutations act collectively and decipher error-correction mechanisms and evolvability of protein structures (Chapter 11).

In the following, we discuss the advantages and possible improvements of the network modeling approach and we draw some perspectives in the context of protein sustainability assessment and urban biomimicry.

Advantages and limitations of the network modeling approach Modeling spatial systems as spatial networks - here, the Amino Acid Network (AAN) of protein structures and the Building Network (BN) of urban structures- allows quantifying space occupancy in the system's structure at multiple scales. The analysis of the systems is eased by the availability of a wide range of well-known metrics and graph algorithms implemented in publicly-available libraries and softwares such as NetworkX [159] and Gephi [160]. Moreover, these metrics can be coupled with ad-hoc measures adapted to the specific system, e.g. the allocation of atomic interactions to structural levels, that is relevant to the study of the encoding of protein dynamics in the protein structure (Chapters 7 and 8).

An additional strength of the network approach is that the high level of abstraction of the models allows comparing structures of extremely different systems, such as proteins, cities and granular materials (Chapters 4 and 5).

The network approach requires choosing the granularity that defines the nodes of the network and the cutoff distance that defines the connectivity between the nodes. This is a critical choice, as different choices of these parameters may lead to different network topologies, as reviewed for AANs in Chapter 2. However, the possibility of modulating the granularity of the model and the cutoff for connectivity also provides the freedom to choose the spatial scale to be analyzed. For illustration, Fig. 12.1 shows BNs with different granularity, where high granularity (left) is obtained with the buildings merging procedure (Chapter 3, Section 3.3.1) and low granularity (right) is obtained without merging.

The BN at high granularity (Fig. 12.1, left) ignores the internal cavities of the buildings

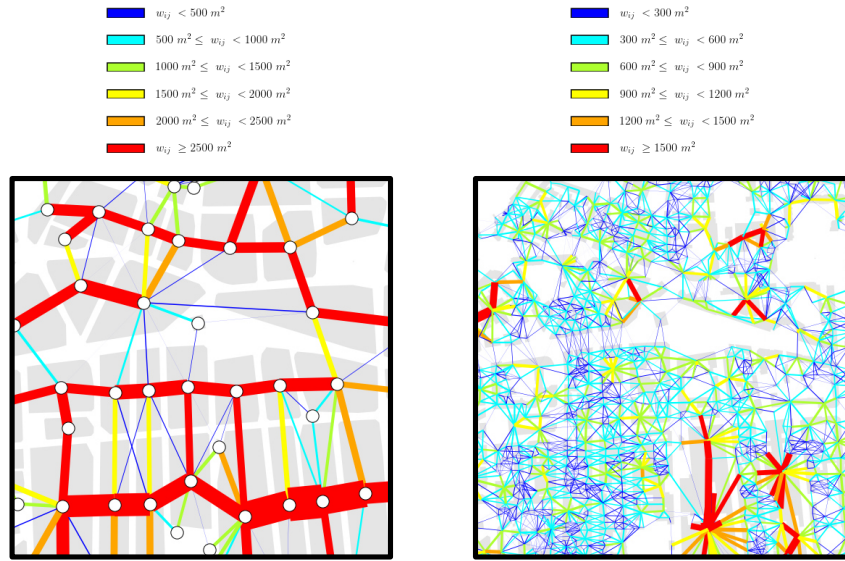


Figure 12.1: Example of Building Networks that model the urban structure at different granularity. Left: a part of the BN of Charpennes (Metropolitan area of Lyon, France) built as in Chapters 4 and 5 using the buildings merging procedure. Right: BN of the same area, built without merging buildings.

and approximates the shape of adjacent buildings to their convex hull. As such, it is more relevant to study the overall space occupancy in urban areas and their impact on the possibility for large changes of urban structure (e.g. the reconstruction of a whole block of buildings) and the accommodation of traffic, that cannot exploit narrow spaces within buildings. On the contrary, the BN at low granularity (Fig. 12.1, right) takes into account the actual shape of buildings, that is relevant for the assessment of the space available for pedestrian mobility and for subtle changes in buildings geometries. Obviously, the higher resolution of the BN at low granularity comes at the expense of a higher complexity (higher number of nodes and links).

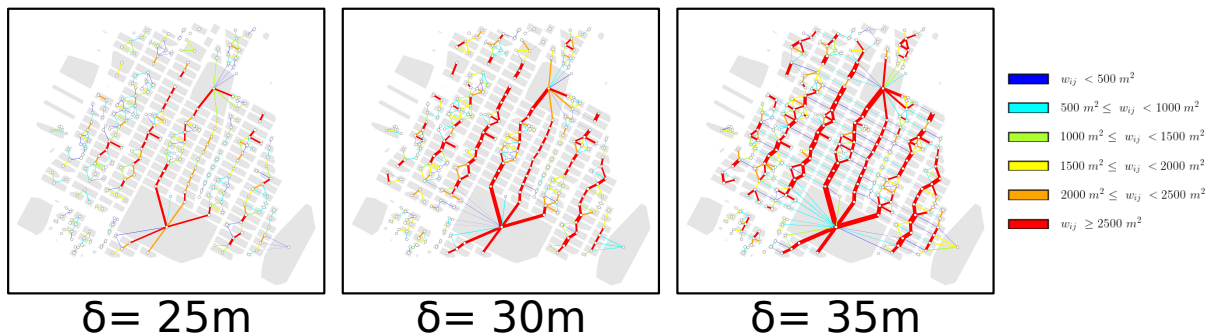


Figure 12.2: Building Networks of Manhattan (New York City, USA) with different cutoff distance δ .

Fig. 12.2 shows BNs with different cutoff distance δ . It shows that tuning the δ parameter allows choosing the scale of buildings packing to analyze: from the closest vicinity (low δ) to further neighbors (higher δ).

Similarly, as illustrated in Fig. 12.3, varying the cutoff parameter of the AAN model allows measuring the atomic and amino-acid packing around amino acids at different

scales: from the nearest neighbors (low cutoff) to further neighbors (higher cutoff).

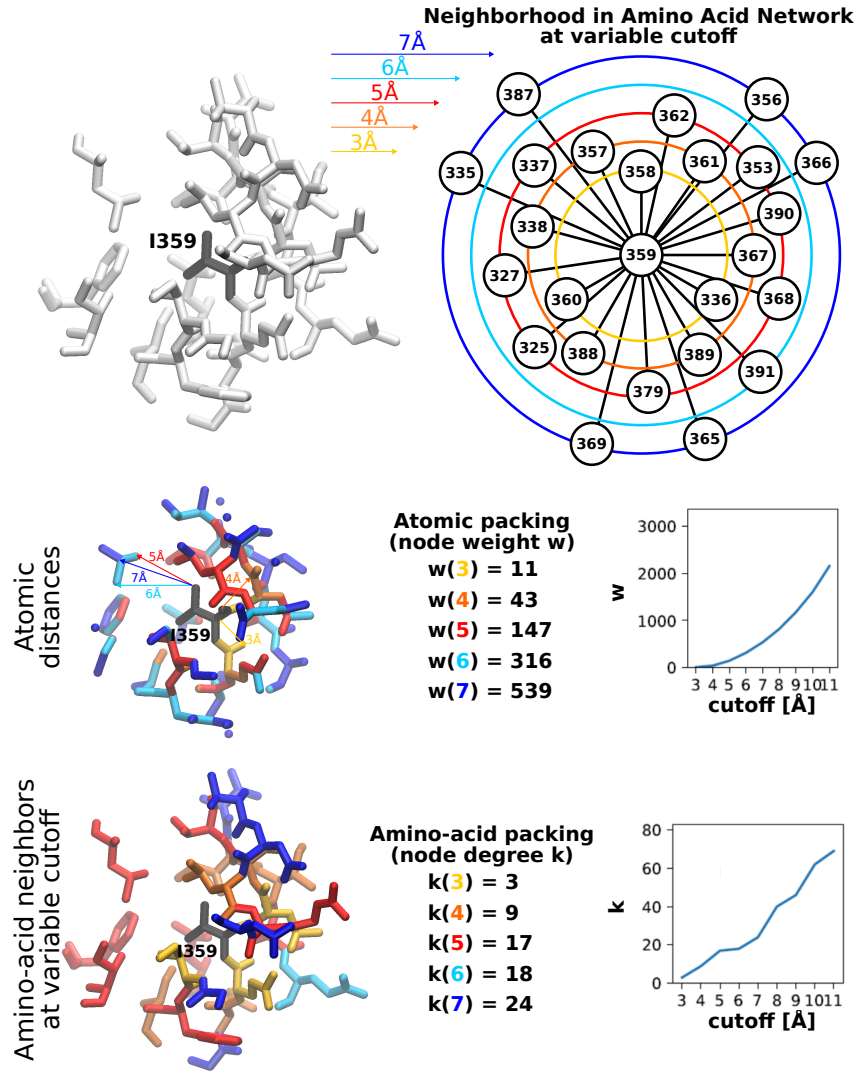


Figure 12.3: Amino-acid neighborhood of the I359 amino acid of PSD95^{pdz3} for different cutoff distances of the Amino Acid Network model.

In Chapters 4 and 5 it was proposed that the multi-scale packing around the components of spatial networks determines how much space is available for accommodating substitutions of components with or without structural rearrangement. As a consequence, the multi-scale packing around amino acids measured by the AANs at variable cutoff should encode for different scales of rearrangement needed to accommodate amino-acid mutations, corresponding to the different classes of amino-acid positions defined in Chapter 6. These classes were: Zero (Z, no rearrangement), Local (L, local rearrangement), Far (F, large-scale rearrangement), Mixed (M, the scale of rearrangement depends on the specific mutations). Consistent with the diversity of responses to mutations, the amino acids of PSD95^{pdz3} show a high diversity of dependencies of amino-acid packing with the cutoff, measured by k -versus-cutoff curves, as illustrated in Fig. 12.4 where some examples of k -versus-cutoff curves of ILE amino-acids are reported. I341 shows a plateau between

cutoffs 5Å and 6Å (oval in Fig. 12.4), I328 has an inflection point at cutoff 7Å (circle in Fig. 12.4), I336 has an almost linear increase of k with the cutoff and I359 has a plateau and several inflection points. These types of behavior are observed for all amino acid types, regardless the position in the protein structure (buried or surface-exposed).

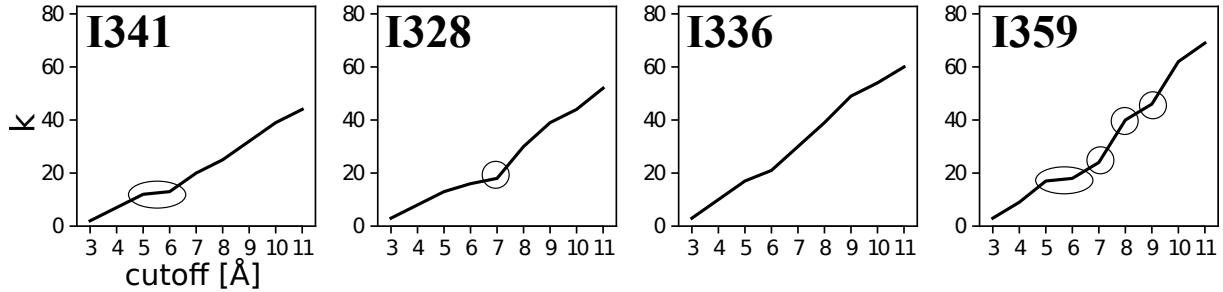


Figure 12.4: Examples of dependencies of ILE amino-acid packings (k) with the cutoff of the Amino Acid Network of PSD95^{pdz3}. The amino-acid packing at a given cutoff is measured by the node degree k . Circles indicate inflection points and ovals indicate plateaux.

As schematized in Fig. 12.5, a linear behavior versus a plateau of the k -versus-cutoff curve suggests different sizes of the neighbors of the amino acid. A linear increase means that neighbors are added at all cutoff distances (Fig. 12.5a). The number of neighbors added at each cutoff is constant or not depending on whether inflection points are present or not (e.g. I336 versus I328 in Fig. 12.4). Instead, the presence of a plateau between two cutoffs c_1 and c_2 means that no additional neighbors are added at cutoff c_2 . This suggests the presence of large amino-acid neighbors that occupy all available space in the annulus of radii c_1 and c_2 around the amino acids (Fig. 12.5b, where $c_1 = 3\text{Å}$ and $c_2 = 5\text{Å}$). Alternatively, a plateau may be due to the presence of empty space between c_1 and c_2 . However, this last option would be reflected by a plateau also in the w -versus-cutoff curve (Fig. 12.5c) that is never observed for the Amino Acid Network of PSD95^{pdz3}.

Because different k -versus-cutoff curves mean that the amino-acid neighbors occupy space differently, it is reasonable to assume that they encode for different responses to perturbations. However, we have not been yet successful in determining any correspondence between the k -versus-cutoff curves and the class of the amino acid (Z, L, F or M). A possible explanation is that the encoding of the perturbation response by the multi-scale amino-acid packing depends on the chemistry of the neighboring amino acids. Additionally, such encoding may depend on the chemistry of the amino-acid itself and its position in the protein structure. If this is the case, the statistical analysis of a larger database of in-silico perturbations would be necessary to verify any relation between the k -versus-cutoff curves and the perturbation classes. Such database would need to include a statistically-relevant number of cases of amino acids with similar and different neighborhoods, for all amino-acid types, both for buried and surface-exposed positions.

An additional challenge is that the perturbations caused by the mutations follow specific directions depending on the orientation of the side chain of the mutated amino

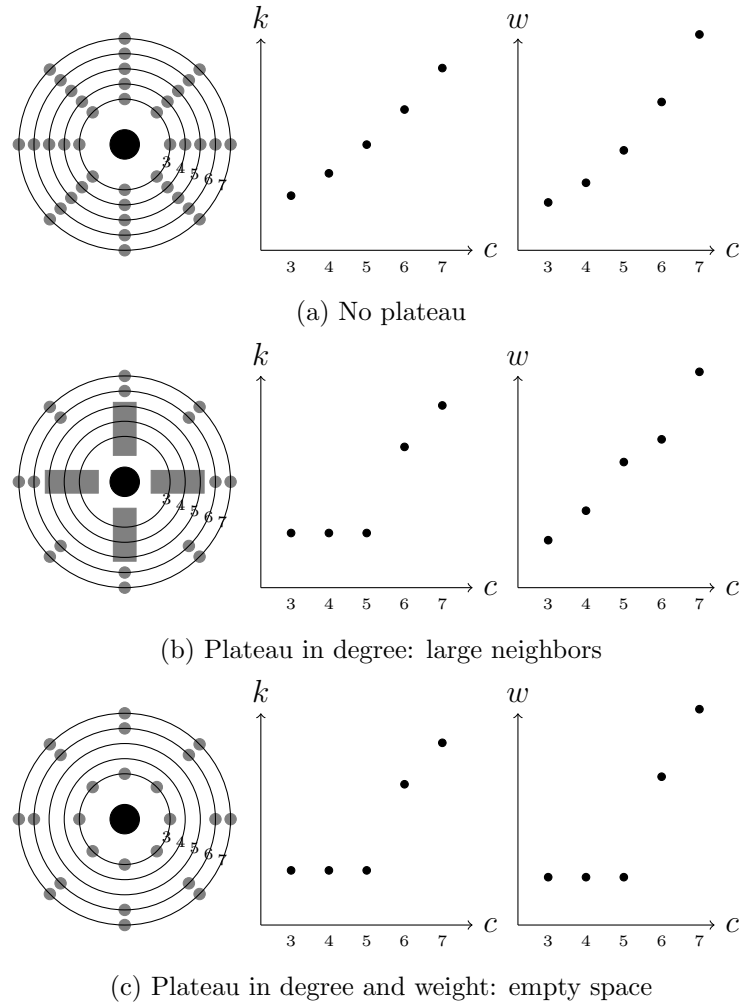


Figure 12.5: Scheme of the evolution of degree and weight as a function of the cutoff distance for different neighbors sizes.

acids, as highlighted by the asymmetry of the arcs in the GCAT network (Chapter 11). The AAN measures packing along all the directions around the amino acid and as a consequence cannot pick up the directionality of amino-acid interactions. To decode the perturbation response of an amino-acid position from the wild-type structure only, the AAN model may need to be refined to discriminate atomic contacts involving amino-acid backbones from the ones that involve side-chains. It should be noted that a refinement of the model in this direction is far from trivial, because GLY amino-acids do not have any side-chain atoms, while they acquire a side-chain when mutated. Moreover, even though backbone atoms are not involved in the atomic rearrangement upon in-silico mutations, they play a role in shaping the space available for the rearrangement of side-chain atoms, and thus they cannot be discarded by the model.

Perspectives and future challenges in the study of proteins sustainability

Overall, the results presented in this manuscript show the importance of atomic interactions, measured by link weights in the AAN, in encoding the perturbation response of

have an impact on the protein function or not. The comparison of the allocation measure in protein variants having similar structure and folding/unfolding dynamics but different function should be performed to make sure that the allocation measure can pick-up differences in functional dynamics. Then, the challenge would be to identify a methodology to distinguish changes in allocation that lead to perturbations of the folding/unfolding dynamics from the ones that lead to perturbations of the functional dynamics.

A third difficulty is that changes in protein structure or functional dynamics may not result in functional change, because of the redundancy of protein structure and dynamics that provide the same function (Chapter 2, Section 2.2). To determine whether changes in atomic interactions leading to structural and dynamical perturbations have an impact on the protein function, it is first necessary to determine the ranges of functional plasticity of protein structures.

Finally, the ultimate goal of assessing the functional impact of perturbations is to determine how pathogenic faults can be corrected and how the protein function is modulated by rescue mutations (robustness) versus mutation leading to functional change (evolvability). To this goal, the mechanisms of error-correction and functional change through multiple mutations should be deciphered. The GCAT network tool presented in Chapter 11 seems promising in guiding the extraction of cooperative interactions between amino-acid mutations. Nevertheless, the analysis of the GCAT networks of proteins having diverse structures, dynamics and function, coupled with experimental measures, is necessary to draw any firm conclusions on the performance of this tool in picking up error-correction and functional-change mechanisms. Moreover, as discussed in the conclusions of Chapter 11, the GCAT network model may be refined to extract the most relevant information, for example by weighting and labeling the arcs based on the perturbation they represent (e.g., specific mutation, number of impacted atomic interactions, spatial scale of the perturbation).

Perspectives towards bio-inspired sustainable cities The analysis of BN models of urban structures has revealed the existence of a high diversity of building neighborhoods solutions in terms of spatial occupancy. Building neighborhoods were classified as spatially sustainable or unsustainable based on the quantity of space that is occupied by buildings versus the quantity of empty space available for accommodating perturbations, e.g. changes in buildings geometries and changes in the demand for mobility.

It was found that sustainable solutions that resemble the sustainable neighborhoods of amino acids are possible regardless the size and shape of the buildings. This means that it is always possible to generate bio-inspired building neighborhoods where space is occupied such that the urban structure can accommodate perturbations. Future work may be dedicated to the creation of tools to automatically propose sustainable solutions under some constraints, e.g. the geometries of the buildings and their maximal and minimal

distance. Such tools could be employed at the early stages of urban design.

Moreover, the BN model could be refined to take into consideration the height of the buildings, discarded in the present work but fundamental to determine the impact of the multi-scale urban structure on the flow of air and pollution in the city. Then, the relation between air quality and the network characteristics of such three-dimensional BN models could be investigated. If any relation exists, it may be possible to extract the criteria that govern environmental sustainability (good air quality), to be coupled with the criteria for spatial sustainability (low average space occupancy and absence of chains of high-weight links).

An additional layer of information that could be added to the BN model, in its current form or in a three-dimensional form, is the function of the buildings and the quantity of people that access the building or live in it. This is particularly relevant for assessing the relation between BN measures and the potential for accommodating traffic in the city. Indeed, the usage of the building could provide an indication on the expected demand for pedestrian and/or car mobility.

Finally, the buildings merging procedure could be investigated as a urban-features-extraction measure per se. Indeed, large nodes of merged building in the BN, as the example of the city-center of Lyon (Chapter 4, Fig. 4.12), represent areas where the buildings geometry is irregular and is not expected to favour the mobility of vehicles. It may be interesting to investigate the possibility of using the buildings merging procedure as a mean to pinpoint areas of the city that would be best suited to be reserved to pedestrian mobility, versus the areas towards which the car traffic should be redirected.

Appendix A

Proteins database

PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length
1B44	D	106	1SNR	A	336	2F2E	A	140	3BID	A	57
	E	107		B	337		B	138		B	56
	F	106		C	336		A	154		C	57
	G	106		A	120		B	166		D	56
	H	106		B	121	2FS2	A	128		E	60
1B9L	A	119	1SQL	C	122		B	134		F	57
	B	119		D	120		A	261		G	59
	C	119		E	120		B	261		H	56
	D	119		F	121		C	260	3BRC	A	148
	E	119		G	121	2GJV	A	136		B	149
1BAZ	F	119		H	122		B	136		A	189
	G	119		I	121		C	136		B	189
	H	119		J	123		D	136		C	189
	A	49		K	120		E	136		D	189
	B	41		L	122	2GW4	F	136	3BU7	A	355
1BND	C	46		M	123		A	61		B	358
	D	40		N	121		B	156		A	465
	A	109		O	120		C	61		B	465
	B	108		P	122		D	156		C	468
1CUL	A	189	1SS4	A	143	2H4U	A	121		D	466
	B	190		B	140		B	121	3CG0	A	126
	C	328		A	167		C	121		B	125
	A	140		B	166		D	130		C	126
	B	140	1T0A	A	155	2H6L	A	137		D	126
1DKU	A	295		B	155		B	137	3CGI	A	112
	B	295		C	155		C	137		B	112
	A	72		A	210		A	77		C	119
	B	72		B	212		B	80		D	117
	A	161		C	167	2I7H	A	183	3D3R	A	80
1E2Y	B	165	1T90	A	484		B	183		B	77
	C	164		B	484		C	183		A	130
	D	160		C	484		D	183	3D6X	B	133
	E	159		D	484		E	181		C	132
	F	155	1TBU	A	96		F	183		D	132
1F5M	G	167		B	98	2IEY	A	159		E	132
	H	163		C	96		B	159		F	131
	I	164		D	97	2IG8	A	142	3DGG	A	217
	J	160		A	114		B	143		B	212
	A	176	1TFP	B	114		C	142		C	215
1F86	B	177		A	126	2IVO	A	325		D	215
	A	115		B	126		B	325	3DM3	A	98
	B	115		C	126		C	325		B	98
	A	450		D	126		D	325		C	98
	B	450	1TQ8	E	125		A	509	3DNM	A	297
1GKR	C	450		F	126	2JG7	B	509		B	296
	D	450		A	64		C	509		C	291

Continued on next page

PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length
1GX1	A	153		B	71		D	509	3DWA	D	291
	B	153		C	64		E	509		A	116
	C	153		D	71		F	509		B	115
1H2I	A	186	1U1S	A	66	2NVW	G	509	3E6Q	C	118
	B	186		B	66		H	509		D	115
	C	186		C	65		A	413		E	114
	D	186		D	66		B	405		A	124
	E	186		E	66	2NWA	A	76		B	124
	F	186		F	65		B	76		C	124
	G	186	1UIR	A	309		C	76		D	126
	H	186		B	313		D	76		E	125
	I	186	1ULQ	A	398		E	76		F	126
	J	186		B	398		F	76		G	124
	K	186		C	399		G	76		H	124
	L	186		D	399		H	76		I	124
	M	186		E	399	2O2A	A	123		J	125
	N	186		F	398		B	124		K	125
	O	186		G	399		C	124		L	125
	P	186		H	399		D	122	3EJ6	A	681
	Q	186	1UT1	A	140	2O66	A	107		B	681
	R	186		B	139		B	108		C	681
	S	186		C	139		C	106		D	681
	T	186		D	139	2O97	A	70	3EJ9	A	62
	U	186		E	140		B	71		B	59
	V	186		F	139	2OBA	A	119		C	63
1HDM	A	184	1UUN	A	184		B	120		D	58
	B	183		B	184		C	120		E	64
1HI9	A	274	1V7Z	A	257	2OBK	D	119		F	58
	B	274		B	257		E	120	3EUW	A	328
	C	274		C	257		F	120		B	328
	D	274		D	257		A	84		C	328
	E	274		E	257		B	85	3EZW	D	328
1HXM	A	206	1VDD	F	257		C	82		A	498
	B	230		A	199		D	84		B	496
	C	206		B	198		E	80		C	489
	D	230	1W7W	C	196		F	86		D	482
	E	206		D	198		G	84		E	492
	F	230		A	151		H	83		F	492
	G	206		B	150	2OIW	A	129		G	498
	H	230		C	151		B	129	3F0D	H	492
1I4K	1	72	1WE0	D	150	2OKA	C	129		A	155
	2	72		E	151		D	126		B	144
	A	72		F	150		A	84		C	148
	B	72		A	166		B	84		D	157
	C	71		B	166		C	84		E	145
	D	72		C	166		D	84		F	148
	E	72		D	166	2OQG	A	107	3FIJ	A	220
	F	71		E	166		B	107		B	219
	G	72		F	166		C	105		C	218
	H	71		G	166		D	105		D	220
	I	73		H	166	2P0G	A	77		E	220
	J	71		I	166		B	79		F	220
	K	72		J	166	2P1B	C	79		G	219
	L	71	1WWW	V	108		D	79		H	220
	M	72		W	109		A	100	3FK3	A	138
	N	71		X	101		B	100		B	138

Continued on next page

PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length
	O	71		Y	101		C	100		C	138
	P	72	1X7V	A	96		D	100	3FP9	A	150
	Q	71		B	96		E	100		B	141
	R	72		C	97		F	100		C	141
	S	71	1XCK	A	524		G	100		D	142
	T	71		B	524		H	100		E	130
	U	71		C	524		I	100		F	140
	V	71		D	524	2P5V	J	100		G	142
	W	71		E	524		A	155		H	141
	X	71		F	524		B	155		I	141
	Y	71		G	524		C	153		J	142
	Z	71		H	524		D	155		K	140
1I7N	A	308		I	524		E	155		L	134
	B	308		J	524		F	153	3G5W	A	318
1IOK	A	482		K	524		G	156		B	318
	B	482		L	524		H	153		C	318
	C	482		M	524	2PD1	A	101		D	318
	D	482		N	524		B	98		E	318
	E	482	1XEA	A	304		C	99		F	318
	F	482		B	304		D	99	3GFB	A	347
	G	482		C	304	2PDO	A	134		B	347
1IUJ	A	102		D	304		B	135		C	347
	B	103	1XRX	A	33		C	134		D	347
1J20	A	386		B	33		D	136	3GW6	A	256
	B	386		C	33		E	134		B	255
	C	386		D	33		F	135		C	254
	D	386	1XXA	A	71		G	132		D	252
1J2G	A	306		B	72		H	134		E	257
	B	306		C	73	2PKH	A	139		F	257
	C	306		D	71		B	140	3H43	A	74
	D	306		E	71		C	140		B	75
1J3W	A	134		F	71		D	139		C	74
	B	135	1Y12	A	154		E	140		D	74
	C	133		B	149		F	139		E	74
	D	133		C	155		G	139		F	74
1KQ1	A	60	1Y1O	A	166		H	140		G	74
	B	61		B	166	2Q2B	A	141		H	76
	H	66		C	166		B	141		I	77
	I	60		D	166	2QAP	A	358		J	74
	K	61	1Y9B	A	79		B	358		K	76
	M	61		B	76		C	358		L	77
	N	61	1YA5	A	198		D	358	3HE1	A	144
	R	61		B	197	2QDY	A	195		B	140
	S	61		T	89		B	211		C	144
	T	61	1YLL	A	188	2QL3	A	200		D	140
	W	61		B	187		B	200		E	144
	Y	60		C	192		C	200		F	140
1L3I	A	178		D	193		D	200	3HFO	A	65
	B	178	1YT5	A	256		E	200		B	66
	C	178		B	256		F	200		C	66
	D	178		C	252		G	200	3H XK	A	249
	E	178		D	256		H	200		B	249
	F	178	1YZW	A	220		I	200		C	249
1LTR	D	108		B	220		J	200		D	249
	E	105		C	220		K	200	3HYK	A	114
	F	109		D	220		L	200		B	112

Continued on next page

PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length
1MBY	G	105	1Z6B	A	142	2QM0	A	259	3I71	C	113
	H	105		B	132		B	265		A	56
	A	75		C	138	2QTX	A	57		B	58
1MVK	B	75	1ZOF	D	132		B	58	3I96	A	109
	A	46		E	137		C	57		B	111
	B	46		F	142	2QW7	D	56		C	109
	C	46		A	170		E	57	3IEY	A	147
	D	45		B	175		F	56		B	152
	E	45		C	170		G	57	3IIE	A	389
	F	46		D	175		H	57		B	390
	G	47		E	170		I	57	3IO1	A	384
	H	50		F	175		J	57		B	380
	I	48		G	170		K	56	3ISM	A	245
	J	46		H	175		L	57		B	247
	K	46		I	170		A	96		C	286
1MW5	L	46		J	175		B	96	3JVV	A	331
	A	159	1ZYE	A	162		C	95		B	328
	B	159		B	162		D	96		C	337
1MWW	A	120		C	162		E	96	3JW8	A	282
	B	118		D	162		F	96		B	270
	C	120		E	162		G	96	3JYG	A	178
1NAQ	A	106		F	162		H	98		B	178
	B	103		G	162		I	96		C	178
	C	105		H	162	2R5O	J	96		D	178
	D	104		I	162		A	167	3K12	E	178
	E	105		J	162		B	166		F	178
1NHK	F	105		K	162	2RA2	A	52		A	113
	L	143	2A6M	L	162		B	55		B	112
	R	144		A	130		C	52		C	113
1NJK	A	130		B	150		D	52		D	118
	B	130	2AAL	A	129		E	50		E	113
	C	130		B	130	2RAQ	F	50		F	112
	D	130		C	129		A	91	3K2W	A	475
	A	133		D	129		B	92		B	488
	B	133		E	129		C	91		C	484
	C	133		F	129		D	91		D	483
	D	132	2ADV	A	161		E	91		E	486
	E	133		B	28		F	91		F	483
	F	133		C	494	2UYW	G	91		G	484
	A	136	2AHB	A	334		A	119	3KER	H	486
	B	134		B	334		D	119		A	117
	C	134	2AP6	A	100	2UZH	A	153		B	117
	D	134		B	100		B	153		C	117
	E	134		C	100		C	155		D	117
	F	134		D	100	2V6U	A	103	3KXE	A	91
	G	134		E	100		B	99		B	91
	H	135		F	100	2V78	A	311		C	75
	I	135		G	100		B	311	3L6E	D	74
	A	116		H	100		C	311		A	204
	B	116	2B3Z	A	359	2V8Q	A	102		B	204
	C	116		B	359		B	73	3LQ1	A	513
	D	116		C	359		E	304		B	518
1OTF	A	59	2B4R	D	360	2VGB	A	517	3LX2	A	247
	B	59		O	334		B	491		B	246
	C	59		P	334		C	517		C	246
	D	59		Q	334		D	512	3M6I	A	358

Continued on next page

PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length
1OTG	E	59	2B6E	R	334	2VHH	A	345	3MAE	B	358
	F	59		A	134		B	379		A	234
	A	125		B	136		C	382		B	240
1OYP	B	125		C	136	2VL6	D	379	3MB2	C	239
	C	125		D	143		A	259		A	61
	A	226		E	134		B	259		B	59
	B	226		F	135	2VQ5	C	259		C	62
	C	226		G	135		A	169		D	58
	D	226		H	134		B	151		E	61
	E	226	2BCM	A	138	2WKN	A	408		F	54
	F	226		B	138		B	408		G	61
	A	294		C	137		C	408		H	58
1P9E	B	294	2BT9	A	90		D	408		I	60
	A	182		B	88		E	408		J	59
1PKH	B	175		C	88		F	408		K	62
	A	149		D	92	2WQK	G	408		L	54
1PKU	B	149	2COV	E	92		H	408	3MMZ	A	162
	C	149		F	87		A	248		B	162
	D	149		G	88		B	248		C	161
	E	149		H	87	2X3H	A	498		D	173
	F	149		I	88		B	498	3MN1	A	183
	G	149		A	156		C	498		B	174
	H	149	2D3A	B	176		A	128		C	174
	I	149		A	353		B	124		D	176
	J	149		B	353		C	126		E	174
	K	149		C	353		D	123		F	176
	L	149		D	353		A	133		G	173
	A	214		E	353	2Z77	B	131		H	173
1Q23	B	217		F	353		C	133		I	174
	C	213		G	353		D	130		J	176
	D	213		H	353		A	169		K	176
	E	215		I	353	2Z8U	B	175		L	173
	F	212		J	353		P	173	3NBK	A	158
	G	213		A	99		Q	175		B	158
	H	216	2DCL	B	98		O	494		C	162
	I	215		C	97		Y	494		D	163
	J	212	2DCN	A	308		A	146	3NCQ	A	112
	K	213		B	308		B	146		B	112
	L	210		C	308		C	144		C	115
	A	517		D	308		D	146	3NJA	A	104
1Q3S	B	517		E	308		E	146		B	106
	C	517		F	308		F	146		C	106
	D	517		G	308		G	146		D	103
	E	517		H	308		H	146	3O1K	A	117
	F	517		I	308		A	149		B	118
	G	517		J	308	2ZQB	B	147		C	117
	H	517		K	308		C	151		D	118
	A	140		L	308		D	152	3O6Q	A	141
1Q4U	B	139		A	398		A	317		B	43
1Q55	A	540	2DPH	B	398		B	316		C	140
	B	540		A	354		C	316		D	44
	C	540		B	353	2ZYZ	A	93	3O6X	A	638
	D	540		C	353		B	183		B	638
1Q5H	A	127		D	353		C	96		C	638
	B	136		A	223		D	183		D	638
	C	142	2DVY	A	215	3A2V	A	243		E	638
				B							

Continued on next page

PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length	PDB	Chain	Length
1Q6W	A	149		C	218		B	242		F	638
	B	149		D	214		C	243	3O7M	A	166
	C	151		E	223		D	240		B	171
	D	149		F	211		E	242		C	164
	E	149	2DYA	A	154		F	244		D	164
	F	149		B	155		G	243	3OJY	A	478
	G	147	2E1M	A	356		H	242		B	503
	H	147		B	90		I	242		C	165
	I	147		C	151		J	243	3OL0	A	43
	J	147	2E4M	A	285	3A76	A	147		B	41
	K	147		B	285		B	153		C	41
	L	147		C	143		C	152	3PQA	A	456
1Q7L	A	192	2E6G	A	221	3B54	A	145		B	458
	B	88		B	225		B	145		C	455
	C	190		C	225	3BAL	A	149		D	456
	D	85		D	228		B	149	3ULL	A	106
1QYN	A	134		E	233		C	149		B	103
	B	133		F	228		D	149	4OOQ	A	131
	C	132		G	229	3BDU	A	51		B	130
	D	136		H	229		B	50		C	127
1R61	A	205		I	227		C	50			
	B	205		J	222		D	50			
1RJ8	A	140		K	226		E	51			
	B	140		L	229		F	50			
	D	140	2E7D	A	116		G	50			
	E	140		B	112	3BGH	A	187			
	F	140	2EIG	A	230		B	186			
	G	140		B	230	3BH3	A	244			
1S1G	A	107		C	230		B	244			
	B	110		D	230		C	244			
1S5U	A	129	2EIS	A	117		D	245			
	B	128		B	117	3D9X	A	114			
	C	130	2ESN	A	299		B	111			
	D	129		B	300		C	114			
	E	136		C	298	3DDV	A	121			
	F	130		D	298		B	139			
	G	130	2F1F	A	163		C	120			
	H	130		B	158		D	138			

Appendix B

Supplementary Figures

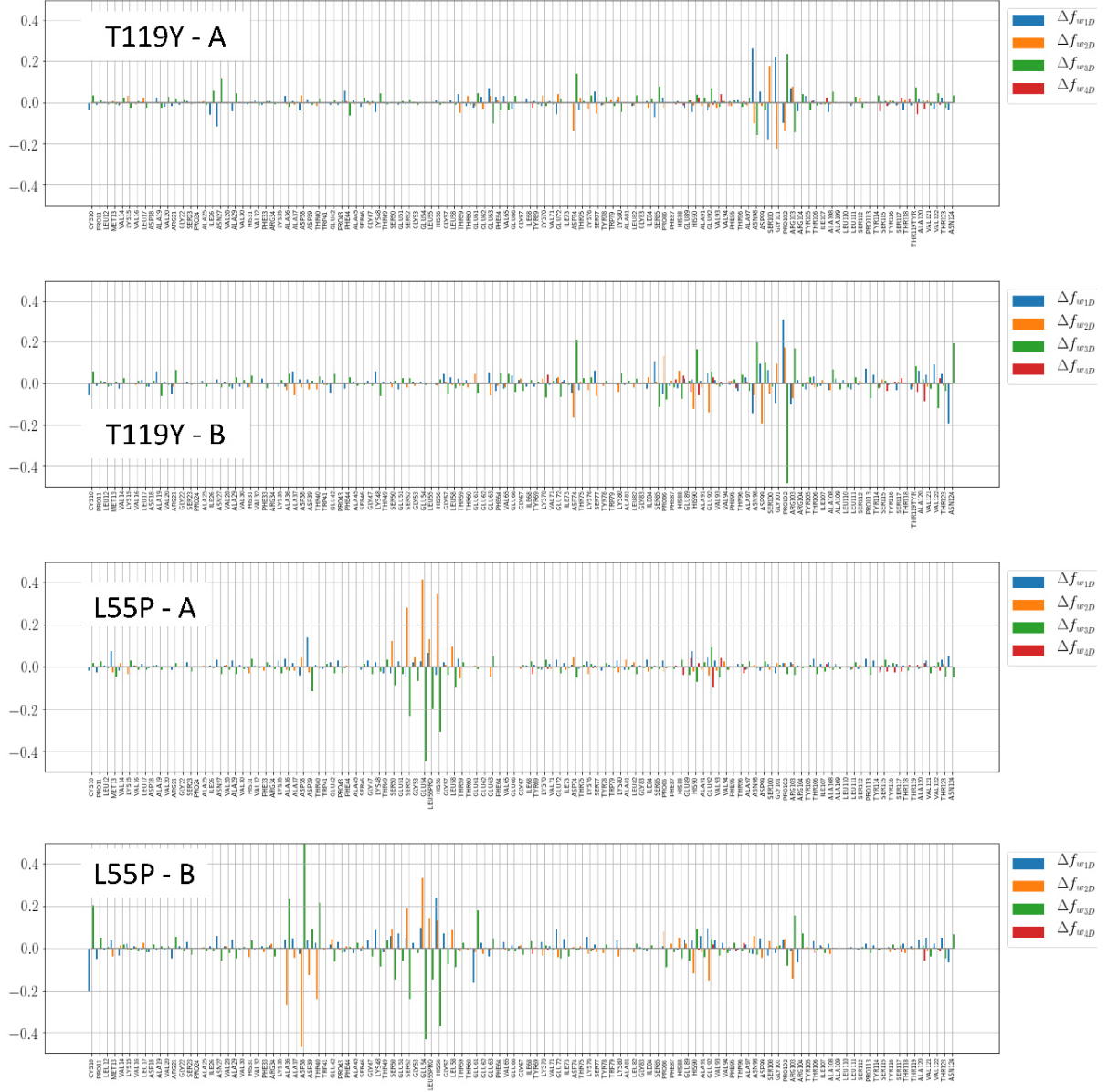


Figure B.1: Comparison of the local structural-level allocation of atomic interactions in Transthyretin T119Y and L55P variants, where $\Delta f_{w1D,i} = f_{w1D,i}^{variant} - f_{w1D,i}^{WT}$ (and similarly for $\Delta f_{w2D,i}$, $\Delta f_{w3D,i}$ and $\Delta f_{w4D,i}$)



Figure B.2: Comparison of the local structural-level allocation of atomic interactions in Transthyretin V30M and T119M variants, where $\Delta f_{w1D,i} = f_{w1D,i}^{variant} - f_{w1D,i}^{WT}$ (and similarly for $\Delta f_{w2D,i}$, $\Delta f_{w3D,i}$ and $\Delta f_{w4D,i}$)

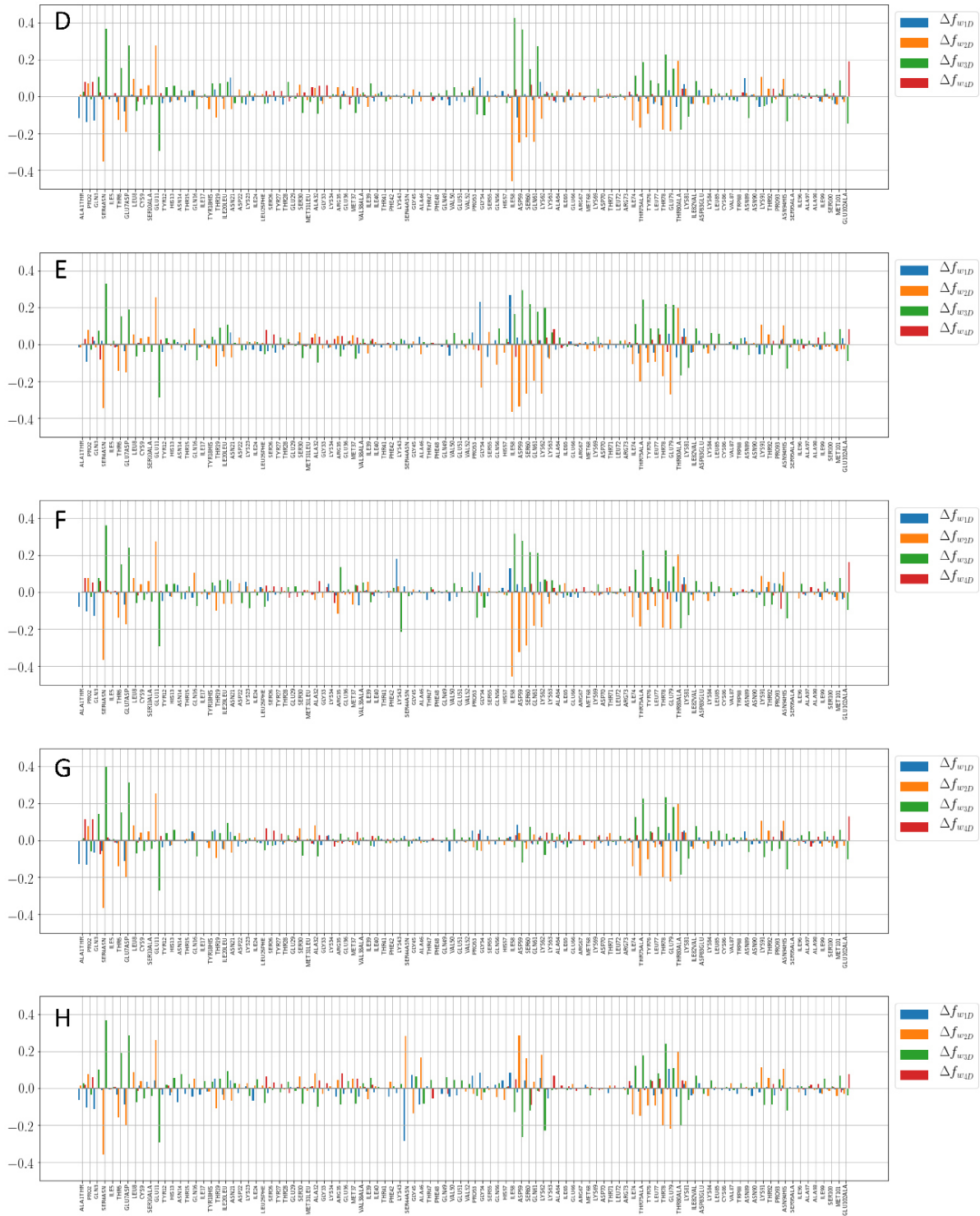


Figure B.3: Comparison of local structural-level allocation of atomic interactions in LTB₅ and CtxB₅, where $\Delta f_{w_{1D},i} = f_{w_{1D},i}^{LTB_5} - f_{w_{1D},i}^{CtxB_5}$ (and similarly for $\Delta f_{w_{2D},i}$, $\Delta f_{w_{3D},i}$ and $\Delta f_{w_{4D},i}$)

Appendix C

Résumé étendu

C.1 Introduction

Les protéines remplissent les fonctions biologiques dans les organismes sur Terre depuis plus de trois milliards d'années. Cela signifie que les protéines conservent leur fonction dans le temps et dans différentes conditions environnementales, c'est-à-dire que les protéines sont des systèmes durables.

Ce travail étudie comment la durabilité est atteinte dans les protéines du point de vue de leur conception structurelle et explore la durabilité de la structure d'autres systèmes, complètement indépendants : les structures urbaines. Cette deuxième tâche est rendue possible par l'utilisation de stratégies de modélisation similaires, de sorte que les deux systèmes peuvent être comparés malgré leurs différences.

Le terme "durabilité" prend diverses significations selon les disciplines [1]. Ici, en utilisant un point de vue du design industriel [5, 6], un système durable est défini comme un système qui reste fonctionnel dans le temps et dans un environnement changeant.

Dans le temps, un système est soumis à des perturbations, internes et externes. Dans la ville, une perturbation interne est la démolition et la reconstruction d'un bâtiment et une perturbation externe est une catastrophe naturelle. Dans la protéine, une perturbation interne est la mutation d'un acide aminé (les acides aminés sont les composants de la protéine) avec un acide aminé d'un type différent, et une perturbation externe est un changement dans l'écosystème dans lequel vit l'organisme.

Le fait que les systèmes soient soumis à des perturbations internes et externes implique deux exigences pour qu'un système conserve sa fonctionnalité dans le temps, c'est-à-dire qu'il soit durable.

La première exigence est qu'en l'absence de perturbations externes, les perturbations internes n'inhibent pas la fonctionnalité du système. Ainsi, soit la fonctionnalité du système n'est pas affectée par la perturbation (le système est robuste à la perturbation car le système modifié est une solution alternative pour l'accomplissement de la fonction), soit il doit implémenter un mécanisme de correction d'erreur qui permet de restaurer la fonctionnalité après perturbation.

La deuxième exigence est que le système doit pouvoir évoluer avec son environnement, c'est-à-dire que le système doit adapter sa fonction à des nouveaux environnements. Il convient de noter que les perturbations internes sont un moyen de moduler la fonction du système, offrant une adaptabilité aux futurs changements environnementaux. Ainsi,

les perturbations internes sont à la fois des défis auxquels le système est confronté dans le temps mais aussi des moyens de s'adapter.

Nous pouvons ainsi lister quelques ingrédients clés pour la conception de systèmes durables : robustesse aux perturbations par la redondance des solutions, robustesse aux perturbations par des mécanismes de correction d'erreurs et adaptabilité par des perturbations internes. Ce manuscrit étudie la durabilité selon les ingrédients proposés en se concentrant sur les perturbations internes et leurs résultats possibles : robustesse ou adaptabilité. Les perturbations externes ne sont pas directement prises en compte dans ce travail. Au lieu de cela, dans le contexte de l'adaptabilité aux changements environnementaux, les perturbations internes sont évaluées comme un moyen de changement fonctionnel. Les perturbations internes sont ici considérées comme des substitutions des composants du système (acides aminés dans les protéines et bâtiments dans les villes).

Considérer l'impact des substitutions des composants du système illustre comment la durabilité nécessite d'équilibrer deux critères contradictoires : d'une part, la robustesse nécessite que le système ne soit pas impacté par la substitution ; et d'autre part, l'adaptabilité nécessite que le système soit réactif à la substitution, afin d'évoluer. Suivant la définition proposée par Brian H. Walker qui dit "La résilience est [...] la capacité de s'adapter et de changer, de se réorganiser, tout en faisant face aux perturbations." [7], les systèmes durables sont résilients et robustes aux substitution de leurs composants.

De plus, la possibilité de corriger les erreurs implique que l'impact des substitutions dépend de l'environnement du composant substitué. Cela signifie que les systèmes durables mettant en œuvre des mécanismes de correction d'erreurs sont par définition des systèmes complexes, c'est-à-dire des systèmes dont le comportement global ne peut être inféré de la connaissance du comportement de ses composants pris individuellement [8].

Les composants des protéines sont des acides aminés. L'ensemble des acides aminés composant une protéine est déterminé par la séquence protéique et la façon dont les acides aminés de la protéine sont disposés dans l'espace tridimensionnel constitue la structure de la protéine. Étant donné que la position des acides aminés dans la structure de la protéine détermine quelles interactions atomiques seront présentes, alors la disposition spatiale des acides aminés dans l'espace contrôle la dynamique atomique dans la structure de la protéine, contrôlant à son tour la fonction biologique.

Les protéines sont des systèmes durables car :

1. La substitution de leurs composants (mutations d'acides aminés) n'impacte que rarement leur fonction [10], c'est-à-dire que les protéines sont robustes aux perturbations internes. Cette propriété signifie qu'il existe plusieurs solutions fonctionnelles en termes de composition en acides aminés pour qu'une protéine soit fonctionnelle. Une conséquence directe est que, malgré le fait que le code génétique contenu dans l'ADN est unique pour chaque individu, c'est-à-dire que plusieurs mutations d'acides aminés existent parmi les séquences d'une protéine chez différents individus [11], tous

les corps humains sont fonctionnels.

2. Les mutations des acides aminés n’agissent pas indépendamment, permettant des mécanismes de correction d’erreur. Par exemple, les mutations dites *rescue* restaurent l’activité de la protéine après qu’elle ait été perturbée par une première mutation d’acide aminé [12].
3. Les rares mutations d’acides aminés ayant un impact sur la fonction ont été exploitées pour l’évolution, c’est-à-dire que les protéines sont adaptables à des nouveaux environnements [13].

Étudier comment la durabilité est conçue dans les structures protéiques a un double intérêt. Le premier intérêt est biomédical. Malgré la durabilité fonctionnelle générale des protéines, certaines mutations d’acides aminés peuvent conduire au développement de maladies [14]. Comprendre comment robustesse ou changement fonctionnel lors des mutations d’acides aminés sont déterminés par la structure de la protéine aidera à déchiffrer les mécanismes moléculaires des maladies causées par des mutations d’acides aminés [15–19]. De plus, comprendre l’impact des mutations d’acides aminés permettra de développer une médecine personnalisée, qui vise à évaluer le risque de développement de la maladie et à sélectionner des médicaments adaptés en fonction du bagage génétique des patients, c’est-à-dire en considérant les mutations d’acides aminés existant dans les protéines d’un patient comparé à un autre [20]. Enfin, si les mécanismes de correction d’erreur se produisant dans les protéines lors de mutations simples et multiples sont déchiffrés, ils peuvent alors être imités pour le développement de thérapies contre les mutations pathogènes.

Le deuxième intérêt concerne la conception de systèmes durables artificiels. Étant donné que les protéines sont durables, elles constituent un bon modèle à suivre pour concevoir des systèmes spatiaux durables en général, à condition que la fonctionnalité du système repose sur les contraintes géométriques entre les composants du système. Dans cette étude, la possibilité d’appliquer cette approche biomimétique à la conception de structures urbaines durables est explorée.

Nous nous concentrons sur l’impact des mutations des acides aminés sur la structure et la dynamique des protéines. C’est un pré-requis pour la prédiction de l’impact fonctionnel des mutations : prédire si la fonction protéique est impactée par la mutation nécessite d’abord d’évaluer si la structure protéique est modifiée et à quelle échelle, puis si la dynamique est impactée, et enfin si les changements structuraux et dynamiques entraînent ou non des changements fonctionnels.

Une difficulté est que nous n’avons accès qu’à des séquences protéiques qui correspondent à des solutions fonctionnelles, sinon elles ne seraient pas produites par un organisme. En conséquence, il n’est pas possible d’extraire les “caractéristiques de durabilité” en comparant les cas durables et non durables, car les cas non durables sont inaccessibles. De plus, les structures protéiques ont une structure tridimensionnelle complexe, composée de plusieurs niveaux structuraux, ce qui rend difficile d’anticiper si les mutations

d'acides aminés peuvent être accommodées avec ou sans réarrangement structurel, et si oui, à quelle échelle. Ces difficultés sont atténuées en étudiant, en parallèle aux structures protéiques, des systèmes spatiaux plus simples : les structures urbaines. Étant donné que les structures urbaines sont modélisées comme des systèmes bidimensionnels, elles sont facilement visualisées, ce qui permet d'évaluer plus facilement si la substitution des composants du système (bâtiments) peut être acceptée par la géométrie du système. De plus, les structures urbaines ne sont pas garanties d'être durables. Si des structures urbaines non durables sont identifiées, elles peuvent être utilisées comme modèles de systèmes spatiaux non durables, inaccessibles lors de l'étude de structures protéiques seules.

D'un côté le défi d'évaluer la durabilité des protéines est facilité par la comparaison avec les structures urbaines, et d'autre part la comparaison entre les structures protéiques durables et les structures urbaines permet de diagnostiquer les structures urbaines durables et non durables. Lorsque les règles de conception de systèmes durables seront complètement comprises à partir de structures protéiques, elles pourront être utilisées pour guider la conception de structures urbaines durables bio-inspirées. Cependant, cela reste une perspective à l'heure actuelle.

C.2 État de l'art

Durabilité

Le terme “durabilité” est utilisé dans divers contextes, et malgré les tentatives de fournir une définition unifiée de la durabilité [29, 30], sa signification dépend fortement de la discipline et des limites du système considéré [1]. A titre d'exemple, la durabilité écologique et environnementale a été définie en relation avec l'impact de l'activité humaine sur la santé des écosystèmes [31, 32]. Les approches dites systémiques incluent également des critères économiques, sociaux et culturels, où le terme “systémique” souligne la nécessité de prendre en considération toutes les facettes d'une société durable [33–35]. D'autres approches se réfèrent plutôt à la définition littérale de la durabilité comme “la qualité de pouvoir continuer sur une période de temps” (Oxford Dictionary, <https://dictionary.cambridge.org/dictionnaire/anglais/sustainability>, consulté le 23 février 2021), comme dans le cadre des architectures logicielles durables [4] et de la conception de produits durables [5, 6]. Je propose que différentes approches de durabilité reflètent différentes échelles du système prises en considération, tandis que l'objectif est toujours le maintien de la fonctionnalité du système au fil du temps, impliquant que le système est capable de faire face aux perturbations. Je propose de classer grosso modo les approches de durabilité en trois classes qui suivent une structure hiérarchique : “Durabilité fonctionnelle” incluse dans “Durabilité écologique”, à son tour incluse dans “Durabilité globale”. Dans ce travail, l'approche “Durabilité fonctionnelle” est utilisée, en se concentrant sur la durabilité des systèmes spatiaux. Deux systèmes sont considérés : protéines et villes.

Protéines : systèmes spatiaux fonctionnels

Les composants des protéines sont les acides aminés, qui sont des molécules organiques composées d'une partie commune (atomes du squelette) et une chaîne latérale spécifique à chaque acide aminé [37]. Il existe vingt-deux types d'acides aminés, dont vingt abondants dans la nature, de tailles et de formes différentes. Les acides aminés d'une protéine sont liés de manière covalente pour former une chaîne appelée séquence d'acides aminés qui encode la structure, la dynamique et la fonction de la protéine.

La structure de la protéine est caractérisée selon quatre niveaux structuraux, représentant quatre types de contraintes spatiales entre les acides aminés [37] :

- La structure primaire est la séquence d'acides aminés : elle définit les deux premiers voisins (gauche et droit) d'un acide aminé. Les acides aminés sont liés de manière covalente à leurs voisins de séquence via les atomes du squelette par des liens peptidiques.
- La structure secondaire représente les conformations de portions de la séquence d'acides aminés résultant du modèle d'interactions des atomes de squelette entre les acides aminés qui sont proches dans la séquence. Sur la base du type de contraintes remplies par les acides aminés qui appartiennent à un segment de la séquence, les éléments de structure secondaire comme les α -helices, les β -strands et les *coils* (sans géométrie fixe) sont reconnus.
- La structure tertiaire correspond à l'agencement des éléments de la structure secondaire dans l'espace tridimensionnel. Une forme particulière de structure tertiaire est la feuille β , constituée par l'alignement planaire de brins β .
- La structure quaternaire n'est définie que pour les protéines constituées de plusieurs chaînes (oligomères) et correspond à l'arrangement tridimensionnel des chaînes dans l'espace.

Alors que la structure primaire est la première à se former, l'ordre dans lequel les autres niveaux structuraux sont repliés est dépendant des protéines [39].

Les techniques expérimentales les plus courantes pour la détermination de la structure des protéines au niveau atomique sont la cristallographie aux rayons X, la spectroscopie de résonance magnétique nucléaire et la microscopie électronique. Lorsque les structures de nouvelles protéines sont évaluées et validées, elles sont ajoutées aux archives internationales, la plus importante étant la Protein Data Bank (PDB, <https://www.rcsb.org/>). La grande majorité des entrées dans la PDB proviennent d'expériences de cristallographie aux rayons X.

Les protéines sont des objets dynamiques : elles se replient et déplient et la fonction biologique des protéines repose également sur des mouvements atomiques multi-échelles contrôlés. La relation entre la structure, la dynamique et la fonction de la protéine fait des protéines des systèmes spatiaux : la fonctionnalité de la protéine est déterminée par

la façon dont ses composants (les acides aminés) sont disposés dans l'espace (la structure de la protéine).

D'une part, la structure de la protéine détermine les interactions entre les atomes de la protéine, et donc les forces auxquelles chaque atome est soumis. En conséquence, la dynamique atomique devrait être entraînée par les forces d'interaction entre les atomes. Sur la base de cette relation entre la structure et la dynamique des protéines, des simulations de dynamique moléculaire (*Molecular Dynamics*, MD) sont utilisées pour prédire la dynamique des protéines à partir de la structure et même à partir de la chaîne d'acides aminés dépliée [43, 44]. D'autre part, la structure de la protéine détermine également l'espace vide disponible pour les mouvements atomiques, comme a été formulé en 1977 par Frederic M. Richards [49]. De ce point de vue, les interactions atomiques représentent tout l'espace qui contraint les mouvements atomiques. Ce travail adopte la deuxième approche, qui se concentre sur le problème géométrique d'inférer des dynamiques possibles à partir de l'espace vide sculpté par l'espace plein dans la structure.

La séquence d'acides aminés détermine la structure des protéines, la structure des protéines détermine la dynamique des protéines et la dynamique des protéines contrôle la fonction des protéines. Cependant, le problème de l'évaluation de la fonction protéique à partir de la séquence est complexe et ne peut être réduit à la prédiction de la structure de la protéine en raison des problèmes suivants. Premièrement, une séquence ne correspond pas à une structure, mais à un ensemble de conformations (l'ensemble conformationnel protéique) qui sont accessibles à la séquence protéique. Ceci est bien décrit par le concept de "paysage énergétique" (*energy landscape*) des protéines, où les conformations stables des protéines correspondent aux minima d'énergie libre d'un espace de haute dimension de coordonnées conformationnelles [59–62]. Deuxièmement, la même séquence protéique peut accomplir plusieurs fonctions, associées à différentes conformations structurales [63]. Troisièmement, différentes séquences et structures peuvent accomplir la même fonction [64].

Le défi consiste à déterminer à quel point les ensembles conformationnels devraient être différents pour encoder différentes dynamiques, et à quel point la dynamique devrait être différente pour fournir différentes fonctions. Enfin, pour les applications biomédicales, un défi supplémentaire serait de déterminer quels changements dans la structure, la dynamique et la fonction conduisent à une défaillance fonctionnelle et au développement de maladies.

Durabilité des protéines

Des solutions alternatives, des mécanismes de correction d'erreurs et la modularité sont des ingrédients pour la durabilité de la structure des protéines. Cependant, l'observation de ces propriétés ne suffit pas à expliquer les mécanismes par lesquels la durabilité est obtenue dans les protéines. Dans le but ultime de prédire l'impact des mutations d'acides

aminés, d’exploiter les mécanismes de correction d’erreurs pour guérir les pathologies et d’imiter la durabilité des protéines dans la conception de systèmes durables artificiels, il est nécessaire de déchiffrer comment la possibilité de solutions alternatives à tous les niveaux (séquence, structure, dynamique et fonction) et les mécanismes de correction d’erreurs sont conçus dans la structure de la protéine.

Les protéines en tant que systèmes complexes

Au cours des vingt dernières années, les modèles de réseau ont été utilisés pour modéliser les structures des protéines, comme est courant pour la modélisation des systèmes complexes [75, 76]. Les modèles basés sur des contacts atomiques entre les acides aminés sont appelés réseaux d’acides aminés (*Amino Acid Network*, AAN). Les nœuds sont des acides aminés ou des atomes et les liens représentent la proximité spatiale.

Suite à la reconnaissance des nœuds centraux dans les réseaux, les positions d’acides aminés ayant une centralité élevée dans l’AAN ont été supposées être fonctionnellement importantes pour la protéine [91, 94, 95]. Cependant, il existe une corrélation positive entre la centralité, la conservation évolutive (l’acide aminé est rarement muté) et la proximité du centre de masse des acides aminés de la protéine [91, 97]. Cela rend difficile de tirer des conclusions sur la causalité entre la centralité et l’importance fonctionnelle.

Une difficulté supplémentaire dans l’évaluation de l’impact fonctionnel des mutations des acides aminés à partir des structures protéiques vient du fait que les mutations des acides aminés peuvent perturber la fonctionnalité des protéines pour différentes raisons. Ainsi, la quantité de contacts atomiques établis par un acide aminé n’est pas un bon indicateur de la sensibilité fonctionnelle aux mutations en soi, ce qui compte est plutôt quels contacts atomiques sont gagnés ou perdus lors de la mutation, et quel est l’impact des changements des contacts atomiques sur la dynamique des protéines.

Les AAN ont été utilisés seuls ou couplés à des simulations ou des expériences pour étudier la dynamique des protéines et l’impact des mutations des acides aminés sur la dynamique de la protéine [99–103, 105, 106]. L’utilisation de simulations MD en plus de l’analyse de structure statique a l’avantage d’incorporer directement les informations dynamiques [107]. Ces approches ont été utilisées pour étudier la dynamique fonctionnelle des protéines à l’aide de réseaux où les liens entre les nœuds représentent soit des contacts atomiques [108–111] ou des mouvements corrélés obtenu à partir des résultats du MD [112, 113]. Des méthodes exploitant à la fois des informations sur les contacts atomiques et les mouvements corrélés sont aussi utilisés [90, 114]. De même, les réseaux d’échange d’énergie (*Energy Exchange Network*, EEN) modélisent le transport d’énergie dans les réseaux de contact de protéines en fonction à la fois de la structure et de la dynamique des protéines simulées à l’aide de MD [115]. Par exemple, la comparaison des EEN des formes apo et holo d’une protéine allostérique a permis d’identifier des résidus impliqués dans la dynamique fonctionnelle [116].

Le repliement simulé de variants de protéines ayant une structure similaire a montré que l'impact des mutations sur la dynamique de repliement dépend du voisinage des positions mutées [117]. L'impact des mutations s'explique en termes de “frustration locale” des contacts atomiques induits par la mutation.

La principale limitation des méthodes basées sur les simulations MD est le coût de calcul élevé. Des limitations supplémentaires sont la forte dépendance des résultats de simulation sur le choix du champ de force auquel les atomes sont soumis [119] et sur la longueur de la période simulée [120].

Une alternative aux simulations MD pour incorporer des informations dynamiques au-dessus des AAN est représentée par le modèle de réseau élastique, ou réseau Gaussien [121]. La différence avec l'AAN est que les liens du réseau Gaussien ne sont pas des entités rigides mais des ressorts avec une certaine constante de force. Les nœuds du réseau sont soumis à des fluctuations, simulées à l'aide de l'analyse des mode normaux et caractérisées en termes d'amplitude, d'échelle temporelle et de corrélations [122]. Des mouvements lents (basse fréquence) ont été associés à des mouvements collectifs (mouvements globaux, impliquant de nombreux résidus) et inversement des mouvements rapides (haute fréquence) ont été associés à des mouvements locaux [122]. Les modèles de réseau élastique ont été utilisés pour identifier des différences dynamiques dans le régime de fréquence basse à intermédiaire entre des protéines ayant une structure monomère similaire mais des fonctions et/ou des états d'oligomérisation différents [124, 125]. Cependant, le modèle de réseau élastique n'est pas adapté pour étudier les mouvements de domaine à grande échelle avec des barrières à haute énergie [125].

Les villes en tant que systèmes complexes

L'objectif 11 des Objectifs de Développement Durable des Nations Unies est de “rendre les villes inclusives, sûres, résilientes et durables” en vue d'une augmentation attendue de la population urbaine (60% de la la population mondiale devrait vivre dans les villes en 2030) et sur la base de l'observation que les villes représentent 70% des émissions mondiales de carbone et plus de 60% de l'utilisation des ressources (<https://www.un.org/sustainabledevelopment/cities/>, consulté le 6 avril 2021). Cet objectif requiert des méthodes pour identifier la durabilité urbaine, la résilience et la capacité de transformation [128].

Les villes ont été reconnues comme des systèmes complexes en raison des réseaux complexes d'interactions existant dans le système urbain à plusieurs niveaux (par exemple, les transports, la mobilité, les activités économiques, etc.), qui génèrent des comportements émergents [129–132]. Des exemples de comportements émergents sont les modes de transport qui dépendent de l'emplacement des commodités et de la densité de population et la ségrégation raciale ou sociale dramatique qui peut résulter d'un léger préjugé préférentiel des personnes pour être entourées d'autres personnes de la même origine ethnique ou du

même statut social, tel que revu par Batty en 2009 [130].

Du point de vue spatial, les villes sont des structures constituées de bâtiments (“espace rempli”) qui occupent le territoire urbain et façonnent “l’espace vide” (routes, parcs, etc.) disponible pour la mobilité des piétons, voitures, vélos, transports en commun, etc.

La topologie de “l’espace vide” a été largement étudiée pour investiguer la mobilité urbaine, principalement dans le cadre des réseaux routiers urbains [133–140], la théorie de la syntaxe spatiale [141, 142] et les réseaux de rues nommées [143], qui se concentrent sur l’agencement spatial des routes dans la ville.

En parallèle, la Morphologie Urbaine étudie les caractéristiques géométriques des formes urbaines, notamment les bâtiments, les îlots et les rues. Ainsi, les études de morphologie urbaine intègrent à la fois l’espace “rempli” et “vide” de la structure urbaine. La morphologie urbaine a été appliquée à l’étude de la croissance urbaine au cours du temps [144] et à la comparaison de différentes villes [145, 146]. Ces méthodes reposent sur des statistiques multivariées impliquant un grand nombre d’indicateurs (par exemple, compacité du bâti, dimension fractale de la surface bâtie, autocorrélation spatiale de la surface bâtie, surface d’îlots, longueurs de rues, etc.). Les résultats des études de morphologie urbaine fournissent soit une description globale des systèmes urbains [144, 146] soit une classification des formes urbaines uniques [147].

De plus, la morphologie urbaine couplée à des simulations physiques a montré le rôle de la géométrie de construction (l’espace “rempli”) dans la modulation du microclimat urbain [148, 149], y compris la qualité de l’air [150–152] et la quantité d’effet d’îlot de chaleur urbain [153], ainsi que la consommation énergétique des bâtiments [149], tous facteurs influençant la durabilité environnementale du système urbain.

C.3 Méthodes

Protéines et villes sont modélisées par des réseaux spatiaux pondérés.

Analyse des structures protéiques

Donées Toutes les données sur les structures des protéines ont été téléchargées à partir de la Protein Data Bank (PDB, <https://www.rcsb.org/>). Dans la PDB, les données sur la structure des protéines sont codées sous la forme d’un fichier texte au format .pdb, où sont indiquées les coordonnées spatiales des atomes de la protéine, ainsi que la séquence d’acides aminés et l’affectation de la structure secondaire. Chaque structure protéique de la PDB est identifiée par un code unique à 4 caractères.

Réseau d’acides aminés À partir des données PDB, les structures des protéines sont modélisées par le réseau d’acides aminés (*Amino Acid Network*, AAN) [23], un modèle établi en biologie computationnelle. L’AAN est un graphe $G = (V, E)$, avec V l’ensemble

des N nœuds du réseau (sommets du graphe) et E l'ensemble des liens du réseau (arêtes du graphe). Chaque nœud de l'AAN correspond à un acide aminé de la séquence de la protéine. Un lien est une interaction atomique définie par la proximité atomique : deux acides aminés i et j sont connectés s'il existe au moins un couple d'atomes, un appartenant à i et un appartenant à j , dont la distance est inférieure ou égale à un seuil de 5\AA , car il s'agit d'une limite pour les interactions chimiques. Chaque lien est pondéré en fonction du nombre de couples atomiques respectant la condition de distance inférieure ou égale à 5\AA . En conséquence, les poids de liens mesurent le nombre d'interactions atomiques entre deux acides aminés, c'est-à-dire le nombre de couples atomiques présents dans le volume d'intersection de l'entourage à 5\AA de chaque acide aminé.

Dans l'AAN, le degré de nœud k_i , défini comme le nombre de voisins d'un nœud i , mesure l'encombrement d'acides aminés autour de l'acide aminé i . Le poids du nœud w_i est défini comme la somme de tous les poids des liens qui relient le nœud i à ses voisins ($w_i = \sum_{j \in N(i)} w_{ij}$, avec $N(i)$ l'ensemble des voisins du nœud i) et il mesure l'encombrement d'atomes autour de l'acide aminé i . Enfin, le *Neighborhood watch* du nœud Nw_i est défini comme le rapport entre le poids du nœud et le degré du nœud ($Nw_i = w_i/k_i$) [23]. Nw_i représente le poids de lien moyen que le nœud i effectue avec ses voisins, c'est-à-dire le nombre moyen d'interactions atomiques que i effectue avec ses voisins. La mesure Nw permet de comparer l'encombrement atomique autour d'acides aminés de tailles différentes, car le nombre total d'interactions atomiques (w_i) est normalisé par le nombre de voisins de l'acide aminé (k_i).

Réseau des perturbations et Réseau de perturbations induites Le réseau des perturbations (*Perturbation Network*, PN) est utilisé pour comparer les AAN de structures protéiques avec des séquences superposables. Les interactions en acides aminés du variant (*var*) et du type sauvage (*wild-type*, WT) sont comparées à travers l'analyse du PN de seuil \bar{w} , représenté par un graphique $G_p = (V_p, E_p)$ avec des liens $E_p = \{(i, j) \in E_{WT} \cup E_{var} \text{ avec } |w_{var}(i, j) - w_{WT}(i, j)| > \bar{w}\}$, poids des liens $w_p(i, j) = |\Delta w(i, j)| = |w_{var}(i, j) - w_{WT}(i, j)|$ et couleur du lien rouge si $w_{var}(i, j) - w_{WT}(i, j) < -\bar{w}$ et vert si $w_{var}(i, j) - w_{WT}(i, j) > \bar{w}$. L'ensemble de nœuds V_p est un sous-ensemble de V , comprenant tous les nœuds pour lesquels au moins un lien a un poids différent dans l'AAN de la variante par rapport à l'AAN WT selon le seuil \bar{w} (c'est-à-dire que les nœuds de degré zéro sont supprimés du PN). Le réseau de perturbations induites (*Induced Perturbation Network*, IPN) d'un variant simple (un seul acide aminé muté) est le composant connecté du PN qui contient le site de mutation. L'IPN d'une mutation permet de ne conserver que l'information sur les perturbations qui se sont propagées du point de mutation vers d'autres zones de la structure protéique. Dans cette étude, $\bar{w} = 4$ est utilisé, fournissant un compromis qui permet de montrer des perturbations tout en maintenant la taille du réseau IPN suffisamment petite pour présenter des différences moins nombreuses mais

plus fortes entre les variantes.

Réseau électrostatique Le réseau électrostatique (*Electrostatic Network*, EN) est utilisé pour interpréter le signal expérimental de dynamique des protéines obtenu avec la spectroscopie diélectrique à large bande (*Broadband Dielectric Spectroscopy*, BDS) [27]. L'EN est un sous-graphe de l'AAN où seuls les liens reliant les acides aminés chargés de charge opposée sont conservés. L'EN Intermoléculaire (4D-EN) est un sous-graphe du EN où seuls les liens reliant deux acides aminés appartenant à des chaînes différentes sont conservés. Inversement, l'EN Intramoléculaire est un sous-graphe du EN où seuls les liens reliant deux acides aminés appartenant à la même chaîne sont conservés. L'EN Induit est donné par l'EN plus tous les premiers voisins des nœuds chargés dans l'AAN, avec les liens respectifs.

Allocation locale des interactions atomiques aux niveaux structurels L'allocation locale des interactions atomiques aux niveaux structurels quantifie la fraction d'interactions atomiques qu'un acide aminé i alloue aux différents niveaux structurels (structure primaire, secondaire, tertiaire et quaternaire), extraite de l'AAN de la protéine.

Mutagenèse in-silico

Les 19 mutations possibles de tous les acides aminés d'une protéine sont produites in-silico à l'aide de FoldX [161] version 5, produisant 19 mutants par position d'acide aminé.

Classification des mutations Nous classons les mutations in-silico en fonction de l'impact qu'elles ont sur les voisinages d'acides aminés dans la structure de la protéine, tel que modélisé par l'AAN. Nous disons qu'un acide aminé est perturbé par une mutation d'acide aminé si son ensemble de voisins dans l'AAN du mutant est différent de son ensemble de voisins dans l'AAN du type sauvage. Nous définissons source la position mutée. Il est à noter que le nombre d'interactions atomiques (c'est-à-dire les poids de liens dans l'AAN) n'est pas pris en considération : si une mutation d'acide aminé provoque une modification des poids des liens faits par un acide aminé avec ses voisins mais la liste des voisins de l'acide aminé n'est pas modifiée, alors l'acide aminé est dit non perturbé par la mutation.

Les classes des mutations d'acides aminé sont :

- **Zéro (Z)** si aucun nœud n'est perturbé;
- **Local (L)** si seuls les premiers voisins (distance ≤ 5 Å) du point de mutation sont perturbés;
- **Loin (F)** si les acides aminés à distance > 5 Å du point de mutation sont perturbés.

Une classification additionnelle des perturbations **L** est faite selon que la perturbation passe de la source aux voisins via un réarrangement des atomes n'impliquant pas de

changement dans les voisins de la source (**Lw**, mécanisme poids) ou la source perd (**Ll**) ou gagne (**Lg**) des voisins (mécanisme degré). Si la source perd certains voisins mais en gagne d'autres, la perturbation est classée comme **Lg**. En effet, cela signifie que l'acide aminé mutant est capable d'atteindre des voisins qui n'étaient pas accessibles auparavant.

De même, une perturbation **F** est définie comme **Fk** (mécanisme degré) si les perturbations se propagent par des changements de voisinages de la source aux autres nœuds perturbés, **Fw** (mécanisme poids) si la source ne gagne pas ou ne perd pas de voisins, ou **Fkw** si la source gagne ou perd des voisins mais qu'un mécanisme poids intervient ensuite.

Le réseau GCAT Le réseau dirigé GCAT modélise les directions des perturbations causées par des mutations in-silico d'acides aminés dans une structure protéique. Une fois que tous les AAN de la protéine WT et de tous ses mutant à un seul acide aminé (19 fois la longueur de la séquence d'acides aminés) sont créés, nous comparons le voisinage de chaque nœud dans chaque AAN mutant par rapport à l'AAN WT et nous utilisons cette information pour construire le réseau GCAT : si un nœud j a changé de voisinage dans l'AAN d'un mutant $i \rightarrow i'$ par rapport à l'AAN WT, alors nous disons que la mutation de i a perturbé le l'acide aminé j , et nous ajoutons un arc $i \rightarrow j$ dans le réseau GCAT.

Dans le réseau GCAT, un arc sortant d'un nœud i (correspondant à l'acide aminé i) représente la perturbation potentielle d'un autre acide aminé j causée par la mutation de i , tandis qu'un arc entrant dans le nœud i représente une perturbation potentielle de i causée par la mutation d'un autre acide aminé. Ainsi, le degré de sortie $k_{out,i}$ d'un nœud i (c'est-à-dire le nombre d'arcs qui partent du nœud i) représente son potentiel à perturber, tandis que le degré d'entrée $k_{in,i}$ (c'est-à-dire le nombre d'arcs qui entrent dans le nœud i) représente son potentiel à être perturbé. Nous classons les nœuds (positions d'acides aminés) dans le réseau GCAT en quatre classes en fonction de leur degré d'entrée et de sortie :

- Classe **Générer (G)** : nœuds avec $k_{in} \leq \bar{k}_{in}$ et $k_{out} > \bar{k}_{out}$;
- Classe **Connecter (C)** : nœuds avec $k_{in} \leq \bar{k}_{in}$ et $k_{out} \leq \bar{k}_{out}$;
- Classe **Absorber (A)** : nœuds avec $k_{in} > \bar{k}_{in}$ et $k_{out} \leq \bar{k}_{out}$;
- Classe **Transmettre (T)** : nœuds avec $k_{in} > \bar{k}_{in}$ et $k_{out} > \bar{k}_{out}$.

Avec \bar{k}_{in} et \bar{k}_{out} les degrés moyens entrant et sortant, respectivement. Il faut noter que par définition $\bar{k}_{in} = \bar{k}_{out} = \bar{k} = \frac{|E|}{N}$, avec $|E|$ le nombre d'arcs et N le nombre de nœuds, pour tout réseau orienté.

Analyse des structures urbaines

Données L'entrée du modèle des structures urbaines est l'empreinte des bâtiments disponible dans OpenStreetMaps (<https://www.openstreetmap.org/>).

Réseau de Bâtiments Le Réseau de Bâtiments (*Building Network*, BN) modélise les structures urbaines. Le BN vise à quantifier l'espace occupé dans les systèmes urbains et extraire - par différence - l'espace disponible pour la mobilité et pour la modification de ses composants. Les nœuds du réseau représentent les composants du système, c'est-à-dire les entités géométriques qui occupent l'espace, de la même manière que les acides aminés qui occupent l'espace dans la structure protéique. Ainsi, les bâtiments adjacents sont fusionnés, car ils représentent un élément géométrique unique, même s'ils sont administrativement distincts. Ensuite, les bâtiments fusionnés sont remplacés par leur enveloppe convexe. Ceci est fait parce que les cavités dans la forme du bâtiment (par exemple les cours) ne peuvent pas être exploitées pour la mobilité et la reconstruction. Là encore, les enveloppes convexes qui se chevauchent sont considérées comme une entité géométrique unique et fusionnées et remplacées par l'enveloppe convexe de leur union. Cette procédure est répétée itérativement jusqu'à ce que toutes les enveloppes convexes soient disjointes. Les enveloppes convexes résultantes des bâtiments fusionnés représentent les nœuds du BN. Par souci de simplicité, dans ce qui suit, les enveloppes convexes des bâtiments fusionnés seront appelées bâtiments.

Deux nœuds sont liés lorsque deux bâtiments i et j sont à distance inférieure ou égale à un seuil de distance δ égal à 30m et n'ont aucun autre bâtiment entre eux (aucun autre bâtiment ne traverse le segment joignant les centres de bâtiments i et j). Plus précisément, deux nœuds sont connectés lorsqu'au moins deux points, l'un du bâtiment i et du bâtiment j sont à une distance ≤ 30 m.

Compter le nombre de voisins d'un bâtiment dans le BN (c'est-à-dire le degré du nœud k_i) n'est pas suffisant pour mesurer l'espace occupé dans l'espace urbain. En effet, l'occupation de l'espace dépend de la distance relative et de l'orientation entre un bâtiment et ses voisins, ainsi que la taille du bâtiment et la taille de ses voisins. Ainsi, les liens doivent être pondérés par une mesure qui prend en compte tous ces paramètres. Ceci est accompli en utilisant la procédure suivante. La surface B_i d'un bâtiment i est augmentée d'une longueur $\delta = 30$ m pour donner une nouvelle surface diluée BD_i . La même chose est effectuée sur le bâtiment voisin j pour donner une surface BD_j diluée de la surface originale B_j . La zone d'intersection A_i entre la zone BD_i et la zone B_j et la zone d'intersection A_j entre la zone BD_j et B_i sont calculées et additionnées pour donner le poids du lien w_{ij} .

De la même manière que pour la comparaison du encombrement atomique autour des acides aminés avec l'AAN, la comparaison de l'occupation de l'espace autour des bâtiments i dans le BN se fait à l'aide du *Neighborhood watch* $Nw_i = w_i/k_i$.

Modélisation par agents de la mobilité aléatoire dans les systèmes urbains

Pour quantifier la relation entre les poids des liens w_{ij} dans le BN et le potentiel de mobilité dans le système urbain, des simulations par agents de la mobilité aléatoire dans

le système urbain ont été effectuées avec GAMA [166] en utilisant la procédure qui suit.

1. Création du BN. Le BN d'une zone de taille fixe (2km x 2km) est produit.
2. Mise en place de l'espace de simulation. L'espace de simulation consiste en un carré de 2 km x 2 km rempli par les nœuds BN. Ainsi, l'espace de simulation se compose d'un espace plein (occupé par des bâtiments) et d'un espace vide. L'espace de simulation est subdivisé en cellules (résolution = $n_{cell} \times n_{cell}$), les cellules peuvent être complètement remplies par des bâtiments, complètement vides, ou partiellement remplies.
3. Initialisation des agents. N_{agents} agents sont initialisés. Chaque agent est initialisé avec une position initiale aléatoire dans l'espace vide et une direction aléatoire initiale.
4. A chaque pas-de-temps, chaque agent :
 - Essaie de se déplacer dans sa direction ;
 - S'il ne peut pas se déplacer (il est coincé par un bâtiment ou par les frontières de l'espace), il choisit une nouvelle direction aléatoire ;
 - est réinitialisé avec probabilité $p_{restart}$.

La simulation est arrêtée après N_t pas-de-temps. La probabilité de redémarrage $p_{restart}$ est incluse pour assurer l'exploration de tout l'espace de simulation.

A chaque pas-de-temps de la simulation, le nombre d'agents occupant chaque cellule de l'espace de simulation est compté. À la fin de la simulation, le temps de sortie moyen par cellule est calculé, défini comme le temps moyen passé par les agents à l'intérieur de la cellule avant de la quitter.

Les valeurs suivantes pour les paramètres ont été utilisées : $N_t = 5000$, $N_{agents} = 50$, $p_{restart} = 0,05$, $n_{cell} = 100$.

C.4 Analyse de la base des données : Les structures protéiques et urbaines peuvent-elles s'adapter aux substitutions de leurs composants sans réarrangement structurel ?

Introduction

Faire une mutation dans un système spatial signifie changer la taille, la forme et l'orientation d'un composant. Pour qu'un système spatial conserve sa fonctionnalité lors de la mutation d'un de ses composants, il est nécessaire (mais pas suffisant) que la structure du système puisse accueillir géométriquement le nouveau composant. Nous définissons spatialement durable un système dont la structure laisse suffisamment d'espace pour accueillir spatialement les mutations de ses composants, quel que soit l'impact que cela aura sur le fonctionnement du système.

La mutation d'un composant peut être accueillie avec ou sans réarrangement de la

structure du système. La mutation d'un composant est possible sans besoin de réarrangement structurel du système s'il y a suffisamment d'espace autour du composant pour s'adapter à un changement de taille, de forme ou d'orientation. L'objectif de ce chapitre est la définition d'une mesure spatiale de l'encombrement local dans les systèmes spatiaux qui permet de distinguer les cas dans lesquels la mutation peut être accueillie sans besoin de réarrangement structurel (durabilité spatiale locale) de ceux où un réarrangement structurel est nécessaire pour accueillir la mutation (insoutenabilité spatiale locale). La mesure est appliquée au diagnostic des structures protéiques et urbaines.

Les modèles réseaux des protéines et des structures urbaines sont utilisés comme moyen de mesurer l'occupation de l'espace. La mesure appelée *Neighborhood watch* est proposée comme mesure de diagnostic de l'encombrement local des voisinage en termes de durabilité spatiale locale.

Études de cas de structures urbaines

Nous avons étudié les réseaux de bâtiments (BNs) de trois zones urbaines de taille 2km x 2km : Monplaisir à Lyon (France), Charpennes à Villeurbanne (Métropole de Lyon, France) et Manhattan à New York City (USA). Les trois cas ont une densité de surface bâtie similaire mais une morphologie urbaine différente. Manhattan est une zone urbaine plus moderne avec des bâtiments de surfaces plus élevées et moins de bâtiments dans la zone de 4 km², des degrés de nœuds inférieurs et des poids de nœuds plus importants par rapport à Charpennes et Monplaisir. Les valeurs moyennes de *Neighborhood watch* suivent l'ordre Monplaisir < Charpennes < Manhattan. Alors que les bâtiments avec *Nw* maximale sont rares dans le cas de Monplaisir, un groupe de bâtiments avec *Nw* maximale est observé dans le quartier des Charpennes et des bâtiments avec *Nw* maximale dans tout le quartier de Manhattan sont observés.

La comparaison des trois études de cas montre que les structures urbaines exploitent des solutions diverses en termes de tailles et de formes de bâtiments, de proximités de bâtiments (k) et d'encombrements de bâtiments (Nw). En particulier, les valeurs de Nw sont très hétérogènes, montrant que la mesure Nw est sensible aux différences de formes urbaines.

Études de cas de structures protéiques

Nous avons étudié les réseaux d'acides aminés de trois protéines (PDB 1B44, 1CUL et 1B9L). Par rapport aux cas urbains, les valeurs de Nw sont homogènes parmi tous les acides aminés dans les trois structures protéiques et prennent des valeurs modérées. Peu de valeurs extrêmes de Nw existent dans les protéines 1CUL et 1B9L, mais elles sont rares par rapport au nombre d'acides aminés avec des Nw modérés.

Statistiques d'occupation de l'espace dans les protéines et les villes

Base de données de protéines Les AANs d'une base de données de 250 structures de protéines globulaires ont été analysés par Rodrigo Dorantes Gilardi dans son travail de doctorat et publiés dans la réf. [23]. En considérant tous les AAN ensemble, la base de données comprend 230528 nœuds.

Les distributions $P(k)$, $P(w)$ et $P(Nw)$ pour la base de données de protéines présentent un bon ajustement avec des distributions normales, ce qui signifie que les valeurs exceptionnellement élevées et exceptionnellement basses de k , w et Nw sont rares. Au contraire, la distribution $P(w_{ij})$ n'est pas normale. Le fait que le degré de nœud k soit borné est attendu du fait que l'AAN est spatial. De même, le fait que le w soit borné est attendu du fait que le nombre d'interactions atomiques faites par un acide aminé, mesuré par la valeur w , est contraint par le nombre d'atomes de l'acide aminé et par l'encombrement stérique. Cependant, le fait que la distribution Nw suive une distribution normale ne s'explique pas facilement, car le Nw représente la moyenne des poids de liens faits par un nœud, et les poids de liens $w(i, j)$ ne suivent pas une distribution normale.

La moyenne et l'écart type du Nw dans les AANs pris individuellement sont similaires à la moyenne et à l'écart type du Nw dans l'ensemble de la base de données, montrant que la mesure Nw est uniforme parmi les nœuds des AANs également au niveau de la protéine isolée.

Le nombre d'acides aminés voisins (k) ne dépend pas fortement du type d'acide aminé, alors qu'en moyenne les acides aminés plus gros font un plus grand nombre de contacts atomiques (w) par rapport aux acides aminés plus petits. Cependant, le nombre moyen de contacts atomiques effectués par l'acide aminé ($Nw = w/k$) est indépendant du type d'acide aminé. Ainsi, l'observation d'une distribution normale de la mesure Nw autour d'une valeur moyenne reste valable lorsque le type d'acide aminé est pris en compte.

Nous proposons que des valeurs modérées et similaires de Nw pour tous acides aminés signifie que de l'espace vide est disponible dans les structures protéiques pour accueillir les substitutions des composants du système (mutations d'acides aminés), c'est-à-dire que les protéines sont localement spatialement durables.

De plus, nous avons vérifié que des valeurs modérées de Nw sont observées aussi dans l'échelle quasi-2D, c'est-à-dire quand seulement les liens correspondants à la structure secondaire de la protéine sont pris en compte. Cela permet de comparer l'occupation de l'espace dans des systèmes bidimensionnels (structures urbaines) et tridimensionnels (protéines) par le biais de la mesure Nw .

Base de données de bâtiments : la ville de Lyon Une base de données des bâtiments a été obtenue en créant le BN de toute la ville de Lyon (France). La base de données comprend 12353 bâtiments, avec une surface allant de 0,8 m² à 0,8 km² (moyenne : 1339 m², écart type : 8178 m², médiane : 218 m²). La diversité des formes

des bâtiments a été vérifiée à partir de la mesure du facteur de forme ϕ des bâtiments. Un cercle parfait a $\phi = 1$ et ϕ diminue à mesure que l'allongement le long d'une direction augmente. La distribution $N(\phi)$ pour la base de données est large, montrant une forte diversité de formes. De manière similaire à la base de données de protéines, la distribution de degrés $P(k)$ est bien ajustée à une distribution normale, cohérent avec le fait que le BN est un réseau spatial et donc son degré est contraint. Cependant, aucune des autres quantités (w , Nw et w_{ij}) ne suit une distribution normale. En particulier, la distribution $P(Nw)$ montre l'existence de nœuds de valeur très élevée de Nw et une plus grande hétérogénéité par rapport aux structures protéiques.

Dans la base de données sur les protéines, il a été constaté que la mesure Nw est indépendante du type d'acide aminé, ce qui signifie que les acides aminés de taille et de forme différentes ont des voisinages d'encombrement similaire. Pour déterminer si cela serait également possible dans le système urbain, nous avons mesuré la corrélation entre les caractéristiques géométriques de trois bâtiments - la surface du bâtiment, le périmètre et le facteur de forme - et leur valeur Nw . Les coefficients de corrélation entre le Nw et la surface du bâtiment, le périmètre et le facteur de forme sont respectivement de 0,26, 0,67 et -0,13. Ainsi, le Nw n'est pas corrélé à la surface du bâtiment et au facteur de forme du bâtiment, alors qu'il est partiellement corrélé positivement avec le périmètre. Néanmoins, il existe des cas de bâtiments de périmètre élevé ($2p > 1000m^2$) et de Nw modérés. Cela signifie que le voisinage d'un bâtiment peut être choisi pour donner une valeur modérée de Nw , quelle que soit la géométrie du bâtiment lui-même. La gestion durable de l'espace (Nw faible ou modérée) est obtenue en ayant des dimensions de bâtiments voisins compensant la taille du bâtiment central. Les petits bâtiments peuvent accueillir des bâtiments de grande et moyenne taille dans leurs voisinage. Les bâtiments de taille moyenne peuvent également accueillir de grands bâtiments dans leurs voisinage, mais uniquement s'ils sont associés à de petits voisins ou à des voisins de taille moyenne plus éloignés. Les grands bâtiments doivent avoir des voisins de petite ou moyenne taille. Si l'espace est vu comme une ressource à répartir entre les bâtiments d'un voisinage, des valeurs modérées de Nw sont obtenues par une utilisation frugale de la ressource spatiale : si le bâtiment central est grand (c'est-à-dire qu'il consomme beaucoup de ressources), alors ses voisins doivent être petits (c'est-à-dire qu'ils consomment peu de ressources).

Conclusion

La métrique Nw , définie pour l'analyse des structures protéiques et appliquée ici aux structures urbaines, est une mesure diagnostique appropriée pour la durabilité spatiale locale dans les systèmes spatiaux modélisés à l'aide de réseaux spatiaux. Les valeurs modérées de Nw correspondent à des voisinages qui peuvent tolérer une mutation de la composante centrale sans avoir besoin de réaménagement structurel, car suffisamment d'espace est épargné entre les composantes (durabilité spatiale locale). A l'inverse, les

valeurs extrêmes de Nw correspondent à des voisinages où la mutation de la composante centrale nécessiterait un réaménagement structurel, car l'encombrement du voisinage est trop important (non-durabilité spatiale locale).

L'encombrement local est uniforme et modéré dans les structures protéiques, assurant une durabilité spatiale locale. Cela signifie qu'en général les acides aminés peuvent être remplacés par des acides aminés de différents types. Au contraire, des solutions d'encombrement non durables sont observées dans les structures urbaines. Néanmoins, un faible encombrement autour des bâtiments est possible dans le système urbain quelle que soit la caractéristique géométrique du bâtiment lui-même, à condition que les bâtiments voisins soient bien choisis. C'est ce que l'on observe dans les voisinages d'acides aminés dans les structures protéiques : les voisins d'acides aminés sont adaptés à l'acide aminé central, de sorte qu'il n'y a pas trop de contacts atomiques et pas trop peu.

Le modèle BN proposé ici s'est avéré être un outil viable pour mesurer l'encombrement local dans les systèmes urbains. La mesure Nw peut être utilisée aux premiers stades de la conception urbaine pour vérifier la durabilité de la solution proposée : les solutions de conception urbaine aboutissant à de faibles valeurs de Nw laissent de la place à la croissance urbaine, qui nécessite la création de nouveaux bâtiments et le remplacement de bâtiments avec des bâtiments de plus grande taille.

C.5 Analyse de la base de données : Si nécessaire, les structures protéiques et urbaines peuvent-elles s'adapter aux substitutions de leurs composants par le biais d'un réarrangement structurel ?

Introduction

Dans le chapitre précédent, la durabilité spatiale locale des systèmes spatiaux a été définie comme la possibilité de substituer des composants par des composants de taille et de forme différentes sans avoir besoin de réarrangement structurel du système. Dans ce chapitre, la possibilité d'accommoder les substitutions des composants du système par un réarrangement structurel de la structure du système est évaluée. Nous appelons cette propriété durabilité spatiale multi-échelle, car le réarrangement peut impliquer plusieurs échelles spatiales, du voisinage des composants mutés (échelle locale) à un réarrangement global du système.

La durabilité spatiale multi-échelle dépend de l'occupation de l'espace à plusieurs échelles dans le système spatial. L'échelle jusqu'à laquelle le réarrangement des composants est possible dépend de l'échelle jusqu'à laquelle l'encombrement n'est pas trop élevé pour permettre la mobilité des composants. Étant donné que les poids des liens dans les réseaux spatiaux des protéines et des villes (modèles AAN et BN, respective-

ment) mesurent l'occupation de l'espace, liens de poids élevé adjacents désignent des zones très denses. Les séquences de liens de poids élevé adjacents correspondent à des zones encombrées à plusieurs échelles.

Nous utilisons une analogie avec les matériaux granulaires pour clarifier la relation entre la taille des composants, les poids des liens et la possibilité de réarrangement, c'est-à-dire la mobilité des composants du système. Cette analogie vient du fait que les matériaux granulaires peuvent être dans un état *jammed* (coincé, aucun réarrangement des grains n'est possible) ou dans un état *unjammed* (non coincé, les grains sont mobiles).

***Jamming* des matériaux granulaires**

Les matériaux granulaires sont constitués de grains (ou particules) en contact et de vides autour des grains [168]. Il a été démontré que ces matériaux se comportent comme des fluides complexes : leur mécanique peut être de type liquide ou solide selon le coefficient de frottement entre les particules, la densité d'empilement et la contrainte appliquée et la distribution granulométrique des grains [169, 170]. La transition du comportement de type liquide (système *unjammed*) au comportement de type solide (système *jammed*) est appelée transition de *jamming* [171–173].

La capacité du système *jammed* à résister au cisaillement ou aux contraintes isotropes a été expliquée par la formation de chaînes de force au sein du réseau de contact des particules [174]. Le réseau de contact des particules peut être représenté sous la forme d'un graphe où chaque particule correspond à un nœud et un lien pondéré existe entre les particules qui sont en contact physique, avec le poids du lien l'amplitude de la force de contact entre les deux particules. Expérimentalement, le réseau de contact d'un système granulaire peut être obtenu à l'aide de matériaux granulaires photoélastiques. Alternativement des simulations de matériaux granulaires peuvent être effectuées et le réseau de contact synthétique peut être obtenu.

Les chaînes de force sont des sous-réseaux connectés du réseau de contact où les poids des liens sont supérieurs à la moyenne [180]. Ils peuvent être identifiés dans le réseau de contacts en filtrant les liens en fonction de leur poids et en analysant la topologie du graphe restant [181] ou en appliquant une analyse de détection de communauté [177, 182].

Il a été montré qu'une plus grande hétérogénéité dans la taille des grains entraîne une plus grande mobilité (le système est moins *jammed*) [171, 173].

L'analogie entre les matériaux granulaires, les protéines et les villes

Les structures protéiques et les villes peuvent être décrites comme des systèmes granulaires où les grains sont respectivement des acides aminés et des bâtiments. L'analogie est justifiée par le fait que dans les deux cas les composants du système (acides aminés et bâtiments) s'arrangent dans l'espace pour constituer la structure du système, de la

même manière que les grains s’arrangent dans l’espace pour constituer la structure du matériau granulaire. Par analogie avec les matériaux granulaires, les protéines et les villes sont considérées comme *jammed* (spatialement non durables) ou *unjammed* (spatialement durables au niveau multi-échelle) selon que des “chaînes de force” de liens de poids élevé adjacents sont présentes dans l’AAN et BN, respectivement. La question est alors de savoir comment mesurer si les AAN et les BN contiennent des “chaînes de force” ou non.

Détection des “chaîne de force” dans les réseaux spatiaux

Un réseau est dit connecté si toutes les distances entre les nœuds du réseau sont finies, c’est-à-dire qu’il est possible de “se déplacer” de n’importe quel nœud du réseau à n’importe quel autre nœud du réseau en utilisant un nombre fini de pas. Si un réseau est déconnecté, alors il est constitué d’un nombre fini de composants connectés. Le plus grand composant connecté (*Largest Connected Component*, LCC) du réseau est le composant connecté qui contient le plus grand nombre de nœuds. La taille du plus grand composant connecté est donnée par le nombre de nœuds qu’il contient.

Pour détecter les “chaînes de force” dans les réseaux spatiaux, l’approche suivante est adoptée. Les liens sont classés en poids faible ou élevé selon que le poids du lien est respectivement inférieur ou supérieur à $\mu_{w_{ij}} + 2\sigma_{w_{ij}}$, avec $\mu_{w_{ij}}$ le poids moyen des liens et $\sigma_{w_{ij}}$ l’écart type, basé sur la statique des bases de données des protéines et des bâtiments. Ensuite, les liens de faible poids sont supprimés du réseau spatial et le nombre de composants connectés et la taille du LCC du sous-réseau restant sont calculés. Un réseau avec une chaîne de force aura moins de composants connectés et une plus grande taille de LCC dans le sous-réseau de liens de poids élevé par rapport à un réseau sans chaînes de force avec le même nombre de nœuds, de liens et de liens à poids élevé.

Les protéines et les structures urbaines sont-elles *jammed* ?

Les liens avec des valeurs maximales de poids sont dispersés dans les structures protéiques et les sous-réseaux de liens de poids élevé des AAN de structures protéiques ont de petites composantes connectées par rapport à la taille du réseau ($LCC/N \sim 2\%$). Ainsi, les structures protéiques sont des systèmes spatiaux *unjammed*. Cela est cohérent avec l’observation que les protéines sont des objets dynamiques et que leur fonction biologique repose sur les mouvements atomiques. De plus, si les acides aminés sont *unjammed* dans la structure de la protéine, alors les mutations des acides aminés peuvent être accommodées par un réarrangement structurel à plusieurs échelles, si la structure locale autour de l’acide aminé à muter est trop compacte : les protéines sont spatialement durables au niveau multi-échelle.

Par rapport aux structures protéiques, les études de cas urbaines présentent une plus grande diversité dans le pourcentage de nœuds appartenant au LCC du sous-réseau de

liens de poids élevé (LCC/N). Monplaisir est la zone la moins bloquée et a des valeurs de LCC/N et $\%m_{\text{high}}$ similaires à celles des protéines. Au contraire, Charpennes et Manhattan ont des valeurs plus élevées de LCC/N et $\%m_{\text{high}}$, Manhattan étant la zone la plus encombrée. En effet, des “chaînes de force” de groupes de liens à fort poids sont observées dans le BN de Manhattan et dans une partie du BN de Charpennes, tandis que les liens à fort poids sont dispersés dans le BN de Monplaisir. Il est à noter que Monplaisir est la zone la plus hétérogène en termes de géométries de bâtiments et Manhattan est la plus homogène, conforme à l’analogie avec les matériaux granulaires, où l’hétérogénéité plus élevée est associée à une mobilité plus élevée.

Dans la ville de Lyon, le pourcentage de nœuds appartenant au LCC est faible (LCC/N = 1%), ce qui conduit à classer la ville comme *unjammed*. Cependant, il faut être attentif au fait que la répartition des liens à poids élevé n’est pas uniforme dans la BN de Lyon : les liens à poids élevé sont concentrés et contigus en centre-ville, alors qu’elles sont rares en banlieue. Nous pouvons en conclure que la ville de Lyon est partiellement *jammed*.

Suivant le parallèle avec les mutations d’acides aminés, les zones urbaines *jammed* ne peuvent pas supporter la substitution de bâtiments, car les voisinages locaux sont fortement encombrés (Nw élevé) et l’encombrement multi-échelle est également élevé (des “chaînes de force” sont présentes).

Evidemment, même dans des systèmes urbains *unjammed* (par exemple Monplaisir), le réaménagement structurel à plusieurs échelles est pratiquement infaisable, car il nécessite de détruire et de reconstruire plusieurs bâtiments. Cependant, un encombrement multi-échelle élevé, correspondant à des systèmes *jammed*, signifie également que peu d’espace vide est disponible sur des grandes distances. La mobilité urbaine exploite cet espace vide, ainsi un encombrement multi-échelle élevé (c’est-à-dire la présence de liens de poids élevé adjacentes dans le BN) suggère une mobilité entravée et un potentiel plus élevé d’accumulation de trafic.

Modélisation par agents de la mobilité aléatoire dans les systèmes urbains

Pour étudier la relation entre la mobilité et l’encombrement multi-échelle dans le système urbain, mesuré par les poids de lien dans le BN, une simulation de la mobilité aléatoire dans les trois études de cas de systèmes urbains est réalisée. La mobilité aléatoire simulée ici n’est évidemment pas représentative du trafic réel dans la ville, qui dépend de l’origine et de la destination des déplacements, de la largeur des rues, de la gestion du trafic, etc. Cependant, la modélisation de la mobilité aléatoire permet de se concentrer sur un seul aspect de la mobilité urbaine, c’est-à-dire l’influence de la géométrie du bâtiment en termes de dimensions, de formes et de proximité.

Les résultats des simulations montrent que la dynamique de la mobilité aléatoire dans le système urbain, favorisée par un faible encombrement et entravée par un emballage élevé à plusieurs échelles, peut être déduite du modèle BN. Plus en détail, les liens de poids élevé

dans la BN correspondent à des endroits où la mobilité est entravée, car l'encombrement des bâtiments est élevé. Les “chaînes de force” du système urbain provoquent une mobilité entravée à plusieurs échelles. Une perturbation dans ces zones *jammed*, par exemple la fermeture d’une rue, sera plus difficilement soutenue qu’une perturbation dans les zones *unjammed*, car les bâtiments sont serrés à plusieurs échelles, c’est-à-dire qu’ils laissent peu d’espace vide disponible pour accueillir la mobilité.

Conclusion

Contrairement aux systèmes urbains, les structures protéiques sont toujours *unjammed* : les liens de poids élevé sont dispersés dans les structures protéiques et ne se regroupent pas en “chaînes de force” de liens de poids élevé adjacents. L’absence de “chaînes de force” dans les structures protéiques signifie que l’espace vide est disponible pour la mobilité atomique à toutes les échelles. Cette propriété rend la protéine spatialement durable au niveau multi-échelle : les mutations d’acides aminés peuvent être tolérées même lorsque l’encombrement local est élevé (Nw élevé) par un réarrangement structural du système à plusieurs échelles.

C.6 Étude de cas : Troisième domaine PDZ. Moyens pour s’adapter aux substitutions dans la structure protéique.

Introduction

Dans ce chapitre, l’impact des mutations des acides aminés sur la structure d’une protéine est étudié pour valider l’existence de plusieurs échelles d’accommodation des mutations. À cet objectif, tous les acides aminés du troisième domaine PDZ de la protéine synaptique PDS-95 (PSD95^{pdz3}) sont mutés *in silico* par les dix-neuf autres types d’acides aminés. L’approche *in silico* permet de distinguer les voisinages d’acides aminés qui conviennent à tout type d’acide aminé de ceux qui doivent être ajustés à l’acide aminé spécifique, sous la contrainte de la robustesse structurelle.

Les structures des mutants sont comparées à la structure de type sauvage (*wild-type*, WT) grâce à la comparaison des réseaux d’acides aminés (AAN) correspondants. Ensuite, les mutations sont classées en fonction de l’échelle spatiale du réarrangement qu’elles provoquent. Le réarrangement est mesuré à partir des changements de voisinage provoqués par les mutations *in-silico*, où un changement de voisinage signifie qu’un acide aminé a gagné ou perdu un ou plusieurs liens dans l’AAN.

Mutagenèse *in-silico* de PSD95^{pdz3}

Nous constatons qu’aucun réarrangement (Z), réarrangement local (L) ou réarrangement à grande échelle (F) ont lieu lors de mutations d’acides aminés. De plus, des mécanismes

de perturbation de degré (Ll, Lg, Fk) et de poids (Lw, Fw) se produisent.

La mutation de certaines positions d'acides aminés provoque systématiquement la même classe de perturbation, quel que soit le type d'acide aminé avec lequel elles sont mutées. Il est intéressant de noter que l'échelle du réarrangement des quartiers n'est pas déterminée par le type d'acide aminé initial (WT). Les positions d'acides aminés qui provoquent une réponse cohérente lorsqu'elles sont mutées sont classées en classe Z, classe L et classe F en fonction de la classe de la perturbation qu'elles provoquent. Précisément, les acides aminés sont classés dans la classe Z, L ou F si au moins 17 de leurs mutations appartiennent respectivement à la classe Z, L ou F.

Alors que certaines positions d'acides aminés peuvent être classées en Z, L ou F, la mutation d'autres positions d'acides aminés provoque un réarrangement à différentes échelles en fonction de l'acide aminé de substitution. Les positions d'acides aminés qui ne provoquent pas de réponse de perturbation cohérente lorsqu'elles sont mutées sont classées dans la classe M, où M signifie Mixte.

Parmi les 115 acides aminés de PSD95^{pdz3}, 22 appartiennent à la classe Z, 11 appartiennent à la classe L, 7 appartiennent à la classe F et 75 appartiennent à la classe M. Cela montre que les voisinages d'acides aminés sont très divers et tolèrent les mutations différemment.

Nous avons investigué si la classe Z, L, F ou M est déterminée par la structure secondaire à laquelle appartient l'acide aminé. Nous avons trouvé qu'aucun des acides aminés appartenant aux hélices α n'appartient à la classe F et au contraire la classe F est légèrement surpeuplée par les acides aminés appartenant aux brins β , tandis que la classe Z est sous-peuplée par rapport aux statistiques globales. La classe M est la plus peuplée pour tous les types de structures secondaires, ce qui est cohérent avec le fait que la plupart des acides aminés appartiennent à la classe M. Les acides aminés appartenant aux *coils* suivent des statistiques très similaires aux statistiques globales. Les résultats suggèrent que les mutations dans les brins β provoquent généralement un réarrangement des voisinages à plus logue échelle par rapport aux mutations dans les hélices et les *coils*. Néanmoins, le nombre limité de cas par classe ne permet aucune conclusion ferme et aucune tendance évidente reliant la structure secondaire à la classe des acides aminés n'est observée.

Ensuite, nous avons cherché à savoir si la classe des positions d'acides aminés est déterminée par la position structurelle de l'acide aminé, dans le noyau structurel ou exposé en surface. Pour ce faire, nous avons calculé la surface accessible relative (*relative Accessible Surface Area*, rASA) de chaque acide aminé et classé les acides aminés comme exposés en surface si $rASA > 0,2$ et dans le noyau sinon. Sur les 115 acides aminés de PSD95^{pdz3}, 64 sont exposés en surface et 51 sont dans le noyau.

21 sur les 22 acides aminés appartenant à la classe Z sont exposés en surface. Cependant, tous les acides aminés exposés en surface n'appartiennent pas à la classe Z : 3

appartiennent à la classe F, 7 appartiennent à la classe L et 33 appartiennent à la classe M.

Parmi les acides aminés du noyau, plus de mutations F sont présentes. Cependant, tous les acides aminés du noyau ne provoquent pas de perturbations à grande échelle quand mutés. Un acide aminé du noyau appartient à la classe Z, 4 acides aminés appartiennent à la classe L, 4 appartiennent à la classe F et les 42 restants appartiennent à la classe M.

Les résultats montrent que les acides aminés exposés en surface sont plus susceptibles de provoquer aucun réarrangement lorsqu'ils sont mutés (mutations Z), tandis que les acides aminés du noyau sont plus susceptibles de provoquer un réarrangement à grande échelle lorsqu'ils sont mutés (mutations F). Néanmoins, toutes les classes de mutations sont observées pour les acides aminés exposés en surface et du noyau, ce qui signifie que la position structurelle de l'acide aminé ne détermine pas strictement l'ampleur du réarrangement provoqué par les mutations.

Ensuite, nous avons vérifié si la classe de l'acide aminé (Z, L, F ou M) est liée à l'impact fonctionnel de ses mutations, sur la base de preuves expérimentales qui ont montré que vingt acides aminés de PSD95^{pdz3} sont fonctionnellement sensibles aux mutations [16]. Nous avons trouvé que l'impact fonctionnel des mutations n'est pas déterminé par l'ampleur de la perturbation qu'elles provoquent (classe Z, L, F ou M).

Conclusion

Les résultats montrent que le réarrangement des voisinages est un mécanisme viable d'accommodation des mutations dans les structures protéiques et qu'il implique des changements de voisinage locaux à globaux indépendamment de la structure secondaire et du rASA de la position mutée. Cela soutient l'idée d'utiliser des métriques d'occupation de l'espace pour étudier la réponse aux perturbations des protéines. De plus, nous avons montré qu'un réarrangement à grande échelle n'implique pas d'échec fonctionnel.

Du fait que des acides aminés du même type appartiennent à des classes différentes (Z, L, F et M), correspondant à des échelles de réarrangement différentes provoquées par leur mutation, nous pouvons conclure que la réponse aux mutations est codée par le voisinage de l'acide aminé muté. Si une position d'acide aminé i appartient à la classe Z, alors le voisinage d'acides aminés de i convient à tous les types d'acides aminés. Les voisinages des positions d'acides aminés appartenant à la classe Z codent pour la robustesse structurelle, car les mutations à ces positions ne provoquent aucun changement dans les voisinages. Si une position d'acide aminé i appartient à la classe L, cela signifie que le voisinage local d'acides aminés de i peut être ajusté pour s'adapter à différents types d'acides aminés, soit en ajoutant des voisins (Lg), en supprimant des voisins (Ll) ou en changeant la connectivité entre voisins (Lw). Les voisinages des positions d'acides aminés de classe L peuvent également être considérés comme codant pour la robustesse structurelle, car seule une modification locale du voisinage est nécessaire pour s'adapter à tout type d'acide

aminé. Si une position d'acide aminé i appartient à la classe F, cela signifie que le voisinage multi-échelle de i doit être ajusté pour accueillir différents types d'acides aminés sous la contrainte de robustesse structurelle, multi-échelle signifiant que les acides aminés plus éloignés des voisins chimiques de i sont impliqués dans le réarrangement. Cela implique que non seulement le voisinage local de l'acide aminé i code pour la réponse aux mutations de i , mais aussi son voisinage à plus grande échelle. De plus, il est à noter que l'existence de la classe M, qui est en fait la plus peuplée, implique qu'en général les voisinages d'acides aminés multi-échelles ne codent pas pour une réponse aux perturbations, mais pour plusieurs réponses.

C.7 Analyse de base de données : Comment mesurer la dynamique des protéines à partir d'une structure statique ?

Introduction

De la même manière que le réarrangement structurel est rendu possible par un faible encombrement, l'espace vide peut être exploité pour les mouvements atomiques représentant le repliement, le dépliement et la dynamique fonctionnelle de la protéine. L'objectif de ce chapitre est l'identification d'une mesure structurelle qui relie la structure de la protéine (interactions atomiques) à la dynamique de la protéine (mouvements atomiques).

Dans le Chapitre 5, les poids des liens dans les réseaux spatiaux ont été mis en relation avec le potentiel de mobilité des systèmes spatiaux en utilisant l'analogie avec les matériaux granulaires, qui sont *jammed* (la mobilité est inhibée) ou *unjammed* (la mobilité est possible) en fonction des poids des liens dans leurs modèles réseaux spatiaux. En suivant l'analogie, il a été proposé que la mobilité dans les systèmes urbains soit déterminée par les poids des liens dans le réseau de bâtiments (BN), et cette relation a été vérifiée en simulant une mobilité aléatoire dans les systèmes urbains.

Dans le cadre de la même analogie, les poids des liens dans le AAN devraient coder pour la dynamique des protéines. En effet, nous avons montré que les différences de poids des liens sont associées à un changement de dynamique fonctionnelle d'une protéine allostérique simulée par dynamique moléculaire [111].

La dynamique des protéines à plusieurs échelles, caractérisée par la simulation de la dynamique moléculaire et l'investigation expérimentale, est décrite par les différents niveaux structuraux qui existent dans une protéine : les éléments structuraux secondaires et tertiaires ont des mouvements collectifs rapides et à petite échelle (nanosecondes à dizaines de microsecondes) tandis que les éléments structurels quaternaires (interfaces) et les grands domaines ont des mouvements collectifs lents et à large échelle (jusqu'à quelques secondes) [57]. Sur la base de ces observations, nous émettons l'hypothèse que la dynamique des protéines multi-échelles est codée au niveau des acides aminés par

une allocation spécifique des interactions atomiques aux niveaux structuraux (1D, 2D, 3D et 4D). Étant donné que les poids des liens w_{ij} dans l'AAN mesurent les interactions atomiques, l'hypothèse implique que la dynamique des protéines est codée dans l'AAN par des poids de liens spécifiques associés aux niveaux structuraux.

Une difficulté pour valider l'hypothèse est que les poids des liens w_{ij} dans l'AAN sont censés porter également des informations purement structurales, par exemple la structure secondaire à laquelle appartiennent les nœuds i et j , le nombre d'atomes des acides aminés i et j , ou leur distance dans la séquence d'acides aminés.

Dans ce chapitre, la même base de données de protéines utilisée dans les Chapitres 4 et 5 est analysée en termes d'allocation des interactions atomiques aux niveaux structuraux. L'objectif est d'explorer les degrés de similarité et de diversité dans l'allocation des interactions atomiques parmi les acides aminés qui appartiennent à des protéines très différentes.

Statistiques des poids des liens dans les quatre niveaux structuraux de la base de données des protéines

Les liens 1D ont des poids plus élevés en moyenne par rapport aux autres liens. De plus, les formes des distributions du poids des liens sont différentes selon les niveaux structuraux. La distribution des poids des liens 1D est en cloche, la distribution des poids des liens 2D est multimodale, avec un premier pic à $w_{ij} \sim 2$ et un second un premier pic à $w_{ij} \sim 8$, et les distributions des poids des liens 3D et 4D sont asymétriques à droite. De plus, la distribution des liens 2D montre deux pics. Ces deux pics correspondent aux poids des liens entre acides aminés à distance $|i - j| = 2$ dans la séquence. De plus, nous avons trouvé que les liens 2D à distances $|i - j| = 3, 4$ sont rares dans les brins β et les *coils* mais pas dans les hélices α . De plus, ce n'est que dans les hélices α que les poids des liens 2D entre les résidus à distance $|i - j| = 3$ sont plus grands que les poids des liens 2D entre les résidus à la distance $|i - j| = 2$, même si plus hétérogènes. Enfin, les poids des liens 2D à distance $|i - j| = 2$ dans les coils sont les plus hétérogènes, cohérent avec le fait que les coils sont moins contraints géométriquement par rapport aux hélices α et les brins β .

Ces résultats montrent que les poids des liens 2D encodent les différentes contraintes géométriques accomplies par les acides aminés des hélices α , brins β et *coils*. De manière cohérente, les hélices α allouent plus d'interactions atomiques au niveau structural 2D (w_{2D}) et moins aux niveaux structuraux 3D et 4D (w_{3-4D}) par rapport aux brins β et aux *coils*. Le nombre d'interactions atomiques allouées au niveau 1D est indépendant de la structure secondaire.

Si deux acides aminés effectuent un nombre différent d'interactions atomiques aux niveaux structuraux 1D, 2D, 3D et 4D, cela signifie que l'espace est occupé différemment autour d'eux. Par différence, l'espace vide disponible pour les mouvements atomiques est également sculpté différemment autour des deux acides aminés. Nous faisons l'hypothèse

que des faibles valeurs de w dans un niveau structurel donné signifient que les mouvements atomiques sont moins contraints dans ce niveau structurel, et donc que différentes allocations d'interactions atomiques (poids de lien dans l'AAN) aux niveaux structurels entraînent différentes dynamique des protéines.

Pour pouvoir comparer l'allocation pour des cas d'acides aminés de type différent, nous proposons de normaliser le nombre d'interactions atomiques allouées à chaque niveau structurel (w_{1D} , w_{2D} , etc.) par le nombre total d'interactions atomiques faites par l'acide aminé (w). Cela donne les mesures $f_{w_{1D}} = w_{1D}/w$, $f_{w_{2D}} = w_{2D}/w$, $f_{w_{3D}} = w_{3D}/w$ et $f_{w_{4D}} = w_{4D}/w$. Nous utilisons des représentations ternaires pour visualiser l'allocation des interactions atomiques aux niveaux 1D, 2D et 3-4D ($f_{w_{1D}}$, $f_{w_{2D}}$ et $f_{w_{3-4D}}$) à l'échelle de l'acide aminé isolé, avec les allocations aux niveaux 3D et 4D unies dans une seule mesure $f_{w_{3-4D}} = (w_{3D} + w_{4D})/w$.

Statistiques de l'allocation au niveau structurel local des interactions atomiques dans la base de données des protéines

L'allocation au niveau structurel local des interactions atomiques a été calculée pour tous les acides aminés de la base de données de protéines, à l'exception des acides aminés n'effectuant qu'un seul lien 1D. Nous constatons que les acides aminés couvrent un large éventail de solutions en termes d'allocation d'interactions atomiques, malgré le fait que le nombre moyen d'interactions atomiques (Nw) effectuées par les acides aminés est uniforme. Nous constatons également que l'allocation des interactions atomiques aux niveaux structurels ne dépend pas du type d'acide aminé. Ainsi, il est possible de comparer la mesure d'allocation entre les acides aminés de type différent.

Nous trouvons que les différences de statistiques des poids des liens 2D et 3-4D entre les types des structures secondaires se traduisent par des différences dans les solutions d'allocations accessibles aux acides aminés des hélices α , brins β et *coils*. Néanmoins la mesure d'allocation couvre une large gamme de valeurs pour tous les types de structures secondaires. Dans l'hypothèse que l'allocation des interactions atomiques code pour la dynamique des protéines cela signifierait qu'un large éventail de dynamiques sont possibles pour le même type de élément de structure secondaire. Ceci est cohérent avec une dynamique fonctionnelle spécifique aux protéines.

Pour vérifier si la position structurelle de l'acide aminé détermine l'allocation des interactions atomiques aux niveaux structurels, nous avons comparé les solutions d'allocations existantes pour les acides aminés dans le noyau structural ou exposé en surface ($rASA \leq 0.2$ et $rASA > 0.2$, respectivement). Nous nous n'avons pas retrouvé de différences importantes entre les allocations faites par les acides aminés des deux classes, signifiant que l'allocation ne dépend pas de la position de l'acide aminé dans la structure.

Conclusion

La mesure de l'allocation locale des interactions atomiques aux niveaux structuraux montre que diverses solutions sont adoptées par les acides aminés. Ces solutions représentent différents moyens d'obtenir un encombrement atomique local moyen uniforme, mesuré par Nw .

La mesure proposée permet de comparer l'allocation des interactions atomiques faites par les acides aminés de type différent. La mesure s'est avérée indépendante du rASA du type d'acide aminé et permet de détecter les différences contraintes associées aux différents types de structure secondaire. Malgré les différences, une large gamme de solutions est adoptée quelle que soit la structure secondaire et il a été proposé qu'une telle diversité reflète le codage de différentes dynamiques de protéines au niveau des acides aminés.

Le chapitre suivant vérifie l'hypothèse selon laquelle l'allocation des interactions atomiques aux niveaux structuraux porte des informations sur la dynamique des protéines. Ceci est effectué par comparaison de la mesure d'allocation dans des variantes de protéines ayant une structure similaire mais une dynamique expérimentale différente.

C.8 Étude de cas : Transthyrétine, pentamères des sous-unités B de toxines, protéases principales des coronavirus du SRAS. Mesure de l'impact des mutations des acides aminés sur la dynamique des protéines.

Introduction

Dans le Chapitre 7, il a été proposé que l'allocation locale des interactions atomiques aux niveaux structuraux code pour la dynamique des protéines. Dans ce chapitre, cette hypothèse est validée par la comparaison de l'allocation de variants protéiques ayant une structure superposable mais des dynamiques de repliement différentes. L'intégrité structurale impose que les différences dans l'allocation des contacts atomiques soient dues à des perturbations dynamiques et non à des changements structuraux. Nous avons sélectionné des études de cas qui sont impliquées dans les maladies humaines.

Les pentamères de la sous-unité B des toxines AB₅

La première étude de cas est la comparaison des pentamères B de deux toxines AB₅, la toxine cholérique (CtxB₅) et la thermolabile humaine (LTB₅), qui ont 82% d'identité de séquence, structure très similaire (RMSD = 0,59 Å) et même fonction mais mécanismes de repliement/dépliment différents : la formation d'intermédiaires d'assemblage est l'étape limitante pour CtxB₅ (mécanisme *fly-casting*) tandis que le repliement des monomères est l'étape limitante pour LTB₅ (mécanisme *induced fit*) [23, 184]. Étant donné que les

différences de dynamique de repliement de CtxB₅ et LTB₅ sont connues, elles représentent une bonne étude de cas pour tester l'hypothèse selon laquelle les différences de dynamique des protéines sont codées par différentes allocations locales d'interactions atomiques aux niveaux structurels.

Nous avons comparé l'allocation dans LTB₅ par rapport à CtxB₅, pour toutes les positions d'acides aminés, par la mesure de $\Delta f_{w_{1D,i}} = f_{w_{1D,i}}^{LTB_5} - f_{w_{1D,i}}^{CtxB_5}$ (et de même pour $\Delta f_{w_{2D,i}}$, $\Delta f_{w_{3D,i}}$ et $\Delta f_{w_{4D,i}}$). Des différences entre $f_{w_{2D}}$, $f_{w_{3D}}$ et $f_{w_{4D}}$ se retrouvent au N-terminal, au C-terminal et dans l'hélice α centrale, tous impliqués dans les interfaces des pentamères.

LTB₅ alloue plus d'interactions atomiques au niveau structurel 4D aux interfaces I₂ et I₃ par rapport à CtxB₅ ($\Delta f_{w_{4D}} > 0$), alors qu'elle alloue moins d'interactions atomiques au niveau structurel 2D et plus au niveau structurel 3D aux interfaces I₁ et I₃ par rapport à CtxB₅ ($\Delta f_{w_{2D}} < 0$ et $\Delta f_{w_{3D}} > 0$). Le fait que plus d'interactions atomiques soient allouées au maintien de l'interface par rapport au repliement de structure secondaire signifie que la structure secondaire sera plus susceptible au dépliement par rapport aux structures tertiaire et quaternaire : lorsque les interfaces sont perturbées, les monomères sont déjà déplié. Ceci est cohérent avec la preuve expérimentale que les intermédiaires d'assemblage ne sont pas observés lors du dépliement de LTB₅, contrairement à CtxB₅ [23, 184].

La comparaison de LTB₅ et CtxB₅ a démontré la validité de notre hypothèse selon laquelle les différences dans la dynamique des protéines peuvent être récupérées par des différences dans l'allocation locale des interactions atomiques aux niveaux structurels 1D, 2D, 3D et 4D.

Variantes de la Transthyrétine

La deuxième étude de cas est la Transthyrétine (TTR), qui est impliquée dans les maladies génétiques neurologiques et cardiaques dues à la précipitation de la TTR mal repliée sous forme de fibrilles amyloïdes et dont les variantes pathogènes montrent une tendance augmentée à l'amyloïdogenèse [185–187]. Le mécanisme exact de la formation d'amyloïde de la TTR est inconnu et dépend probablement de la variante et des conditions expérimentales [188–192]. L'étude des variants TTR pour expliquer leurs différentes voies de repliement évalue la sensibilité de la méthodologie proposée à la détection de l'impact des mutations d'un seul acide aminé.

En suivant la même procédure que pour les toxines AB₅, nous étudions l'allocation dans des variantes de la TTR fonctionnellement robustes (T119Y et T119M [187, 199]) et pathogènes (L55P et V30M [187]), comparées à la TTR de type sauvage (*wild-type*, WT).

La variante TTR T119Y montre très peu de changements dans l'allocation par rapport à la TTR WT, consistant avec sa non-pathogénicité.

Le variant pathogène TTR L55P montre de grandes différences d'accolation par rapport à la TTR WT dans le segment 50-58. Les résidus dans le mutant consacrent plus de ressources aux interactions 2D et moins aux interactions 3D, ce qui peut augmenter la mobilité 3D du segment 50-58 et le rendre disponible pour des interactions intermoléculaires avec un autre monomère, conduisant à la formation de fibres amyloïdes. Cette possibilité est développée plus dans les détails dans le Chapitre 9.

Les changements dans l'allocation des interactions atomiques pour l'autre variante pathogène, TTR V30M, sont différents de ceux de TTR L55P. Les valeurs de Δf_w sont inférieures à celles du TTR L55P, mais des différences entre $f_{w_{2D}}, f_{w_{3D}}$ et $f_{w_{4D}}$ sont présentes à une interface du dimère. Ceci est cohérent avec le fait que des études ex-vivo ont montré que TTR V30M se déplie avant la formation de fibres, contrairement à la TTR WT [188, 191].

Nous trouvons peu de perturbation pour la variante non pathogène T119Y et une perturbation substantielle pour les variantes pathogènes L55P et V30M. Le lien entre la quantité de perturbations et la pathogénicité de la mutation rappelle ce qui a été rapporté précédemment pour une autre mesure locale, appelée frustration locale [203, 204]. Cependant, la même logique ne vaut pas pour le cas du variant non pathogène T119M, qui montre encore plus de perturbations par rapport aux variants pathogènes (segments 74-79, 50-58 et 21-26). Les différences au segment 50-58 sont similaires à celles montrées au même segment dans la variante pathogène TTR L55P, mais des perturbations supplémentaires sont observées au segment 21-26, qui est spatialement proche du segment 50-58 dans la protéine structure, suggérant que la perturbation dans le segment 21-26 compense en quelque sorte la perturbation du segment 50-58, empêchant l'amyloïdogenèse de la TTR T119M. T119M empêche l'amyloïdogenèse lorsqu'il est présent avec la mutation V30M chez les individus hétérozygotes [205], ce qui implique l'existence d'un mécanisme de correction [12] par lequel la perturbation dynamique causée par la mutation V30M sur une chaîne est corrigée par une perturbation dynamique causée par la mutation T119M sur une autre chaîne. Un mécanisme de correction dynamique nécessite des changements dans les interactions atomiques, compatibles avec la grande perturbation causée par la mutation T119M malgré sa non-pathogénicité. Le cas TTR T119M met en évidence la nécessité d'une analyse qualitative des perturbations en plus de leur quantification pour discriminer les mutations pathogènes des mutations non pathogènes.

Les différences d'allocation locale des interactions atomiques aux niveaux structuraux pour les variantes TTR par rapport au WT montrent la sensibilité de cette mesure aux mutations d'un seul acide aminé pour détecter les perturbations dynamiques sous-jacente à la robustesse (variante T119Y), à la fragilité (variantes L55P et V30M) et aux mécanismes de correction (variante T119M).

Protéases principales des coronavirus du SRAS

La troisième étude de cas est la comparaison des protéases principales (Mpro) de deux coronavirus : SARS-CoV et SARS-CoV-2. Le SARS-CoV est responsable de la pandémie du syndrome respiratoire aigu sévère (SRAS) qui a causé 774 décès en 2003 (www.who.int/csr/sars/country/table2004_04_21/en/, dernière visite le 8 juin 2020) et le SARS-CoV-2 est responsable de la pandémie mondiale de COVID-19 qui a touché 216 pays dans le monde depuis décembre 2019, causant plus de trois millions de décès à l'heure actuelle (www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/, dernière visite le 22 avril 2021). L'inhibition de Mpro est l'une des cibles du traitement antiviral du COVID-19 grâce à son importance fonctionnelle dans la réplication du virus [193, 194]. De plus, la similitude structurelle entre la Mpro du SARS-CoV et la Mpro du SARS-CoV-2 a motivé la recherche de médicaments qui pourraient inhiber l'activité des Mpro d'un large spectre de coronavirus [195–197]. Nous appliquons notre méthodologie, validée par les deux premières études de cas, à la comparaison de l'allocation au niveau structurel local des interactions atomiques des Mpro du SARS-CoV et SARS-CoV-2, pour comprendre si un comportement dynamique différent pour les deux protéines est à prévoir ou non.

Les Mpro du SARS-CoV et du SARS-CoV-2 ont une séquence très similaire (12 mutations sur 305 résidus, correspondant à 96% d'identité) et une structure cristalline similaire (RMSD = 0,71 Å pour une chaîne).

Les principales différences dans l'allocation des interactions atomiques se situent au niveau du domaine C-terminale, où une fraction plus élevée des interactions atomiques est dévolue à la 2D et une fraction inférieure à la 3D dans la Mpro du SARS-CoV-2 comparé à la Mpro du SARS-CoV. De plus, le C-terminale pointe vers le domaine actif dans la Mpro du SARS-CoV-2 et vers le reste du domaine supplémentaire dans la Mpro du SARS-CoV.

Beaucoup d'efforts sont consacrés à la conception de médicaments pour inhiber les Mpro d'un large spectre de coronavirus [195–197, 208], mais nos résultats suggèrent des différences dans la dynamique enzymatique, en particulier de la dynamique C-terminale, une région dont le rôle dans l'activité enzymatique a été démontré expérimentalement pour le SARS-CoV [209]. Ce résultat soutenu par les autres cas d'étude déconseille une stratégie médicamenteuse à large spectre en raison des différences en dynamiques à grande échelle susceptibles d'empêcher la reconnaissance d'un même médicament. Il conseille donc des médicaments plus adaptés à la souche pour inhiber la fonction du Mpro des coronavirus.

Conclusion

A partir des résultats présentés, nous pouvons déduire que l'allocation des interactions atomiques aux niveaux structurels contient des informations sur la dynamique des protéines. Cela pourrait ouvrir de nouvelles voies vers le décodage de la dynamique d'une protéine

à partir de l'information statique contenue dans sa structure cristalline.

De plus, la comparaison de l'allocation des interactions atomiques en niveaux structuraux est un outil rapide et léger pour le diagnostic des mutations neutres, fonctionnellement sensibles et de *rescue* (“secours”) en termes d'impact sur la dynamique des protéines à plusieurs échelles. Ces informations permettront d'explorer les mécanismes moléculaires sous-jacents aux mutations pathogènes et aux mutations correctrices d'erreurs, en aidant la conception de médicaments et le développement de thérapies personnalisées.

C.9 Étude de cas : Transthyréline. Changements à grande échelle de la dynamique lors de la mutation conduisant à la formation de fibres amyloïdes.

Introduction

Dans le Chapitre 8, les différences d'allocation aux niveaux structuraux des interactions atomiques pour les variantes de la Transthyréline (TTR) ont été liées à des différences de dynamique. Dans ce chapitre, le variant pathogène TTR L55P est étudié plus en détail. À l'aide du réseau de perturbations induites (IPN), les différences détaillées dans les interactions atomiques entre les structures de type sauvage (WT) et les variantes de la TTR sont analysées et liées aux différences dynamiques. La méthode est validée par la comparaison des IPN des mêmes variants TTR étudiés au Chapitre 8 : deux pathogènes (L55P et V30M) et deux non pathogènes (T119Y et T119M). Ensuite, un modèle de la fibre amyloïde de la TTR L55P est proposé en utilisant les informations dynamiques extraites de l'IPN et l'application d'une approche de pavage.

Modèle de fibre amyloïde de la Transthyréline L55P

La cause de la tendance à la formation de fibres amyloïdes du variant L55P de la TTR est étudiée selon les hypothèses suivantes :

1. Une mutation permet la reproduction de la structure native si les AAN du variant et du WT ont des propriétés de nœud similaires (degré de nœud k_i , poids de nœud w_i et *Neighborhood watch* Nw_i) pour tous les nœuds i [23] (Chapitre 4).
2. La dynamique des atomes dans le variant est similaire à la dynamique des atomes dans la protéine WT si les poids des liens w_{ij} sont similaires dans les AAN du variant et du WT (Chapitre 5 et le Chapitre 8).

La première hypothèse a été vérifiée en comparant les propriétés des nœuds des AAN des TTR WT et L55P. A titre de comparaison, le même calcul est effectué pour une autre variable pathogène (V30M) et deux variables non pathogènes (T119Y et T119M). Nous montrons que les AAN des variantes de la TTR WT, L55P, V30M, T119Y et T119M partagent des propriétés de nœud très similaires le long de la séquence.

Pour tester la deuxième hypothèse, les IPNs sont utilisés pour comparer les poids des liens dans le AAN de chaque variante avec ceux de l'AAN du WT. L'IPN de L55P a des liens rouges et verts, montrant qu'un réarrangement complexe des positions atomiques, conduisant à des changements d'interaction atomique par rapport à la TTR WT. En revanche, les IPN des variants non pathogènes T119Y et T119M n'ont qu'un seul lien vert. Enfin, l'IPN d'un variant pathogène V30M a des liens rouges et verts (comme l'IPN L55P, contrairement aux IPN des variants non pathogènes), différents de l'IPN de L55P. Ces résultats montrent comment les variants pathogènes ont des IPN différents par rapport aux variants non pathogènes. De plus, l'IPN permet de reconnaître les différences entre les variants pathogènes (L55P et V30M), compatibles avec différents mécanismes de formation de fibres pour V30M et TTR L55P.

L'IPN de TTR L55P montre que le *coil* contenant les résidus S52 à E54 perd d'interactions avec le brin β contenant résidus M13 à K15 et change d'angle d'orientation par rapport au brin β contenant les résidus G47 à T49. Sur la base de ces observations, nous faisons l'hypothèse que le *coil* E54-T60 s'éloigne du reste de la structure, résultant du relâchement des interactions atomiques avec les résidus M13, V14, K15, et forme la nouvelle interface d'interaction qui permet à la fibre amyloïde de grandir.

La possibilité pour un segment spécifique de former la nouvelle interface pour la croissance des fibres amyloïdes dépend de deux facteurs : l'occupation spatiale de la structure protéique et l'affinité chimique entre deux copies du segment. Nous nous concentrons sur le premier aspect, pour déterminer si l'interface candidate pour la TTR L55P déduite de son IPN permet la croissance d'une structure fibreuse, sans produire de collisions stériques.

Afin de comprendre si la nouvelle interface candidate peut permettre la croissance d'une structure de type amyloïde d'unités répétitives d'oligomères de TTR L55P, la symétrie du tétramère de la TTR est prise en considération. Le tétramère natif de la TTR a une symétrie dièdre ; en conséquence, l'interface d'interaction sur les quatre chaînes suivra la même symétrie si aucune dissociation du tétramère ne se produit.

Un tétramère dont les interfaces suivent une symétrie dièdre pourra interagir avec une copie de lui-même après une rotation de 90 degrés. Ensuite, l'octamère obtenu constituera une unité répétitive pour un modèle de pavage, aboutissant finalement à une structure de type fibre allongée. Nous nous référons à un modèle de fibre ainsi construit en tant que *modèle de fibre dièdre*. Alternativement, à condition d'une mobilité suffisante des interfaces, elles pourraient s'aligner le long d'un axe de symétrie du tétramère et dégénérer en une symétrie centrale. Dans cette situation, le tétramère lui-même représenterait l'unité répétitive pour un modèle de pavage. Nous nous référons à un modèle de fibre ainsi construit en tant que *modèle de fibre central*.

En raison de la géométrie de l'oligomère TTR, le modèle *dièdre* produit une fibre de diamètre d'environ 69 Å, tandis que le modèle *central* produit une fibre de section ellipsoïdale, avec l'axe majeur d'environ 69 Å et un axe mineur d'environ 42 Å.

Le *modèle de fibre dièdre* correspond parfaitement au modèle proposé dans la référence [200], même s'il a été réalisé à partir d'une structure cristalline différente et à l'aide d'outils différents. De plus, les micrographies électroniques des protofilaments de la TTR L55P produits *in vitro* montrent un diamètre de 60 à 65 Å [202], cohérent avec le *modèle de fibre dièdre*. Les preuves expérimentales mentionnées soutiennent le *modèle de fibre dièdre* comme modèle pour la fibre amyloïde formée par la variante TTR L55P.

Conclusion

L'analyse des IPN des variants TTR a montré comment l'IPN sonde les mouvements atomiques et les réarrangements des liens atomiques qui sont cohérents avec les différences dynamiques mesurées expérimentalement et est complémentaire à la comparaison de l'allocation locale des interactions atomiques aux niveaux structuraux. Il est à noter que peu de changements dans les interactions atomiques peuvent conduire à un changement drastique de la dynamique de la protéine.

Le modèle de pavage proposé dans ce chapitre prend en compte les contraintes géométriques sous-jacentes à la formation des fibres, nécessaires à la validation des potentielles interfaces d'interaction extraites de l'analyse de l'IPN. Le couplage de l'IPN et du modèle de pavage a permis la construction d'un modèle de fibre pour la variante TTR L55P, qui satisfait les contraintes géométriques imposées par la symétrie de la structure tétramère TTR et est en accord avec les résultats expérimentaux. Le modèle proposé ne nécessite pas la dissociation du tétramère avant la formation des fibres.

C.10 Étude de cas : pentamères des sous-unités B de toxines. Approches intégratives des perturbations de la dynamique des protéines liées aux mutations.

Introduction

Dans le Chapitre 8, les différences d'allocation locale des interactions atomiques aux niveaux structuraux entre les sous-unités B de CtxB₅ et LTB₅ ont été associées à des différences dans leurs dynamiques expérimentales de repliement et de dépliement. Dans sa thèse de doctorat, Laëtitia Bourgeat a proposé la Spectroscopie Diélectrique à Large Bande (*Broadband Dielectric Spectroscopy*, BDS) couplée au nanoconfinement comme nouvelle technique expérimentale pour mesurer la dynamique des protéines à plusieurs échelles (1 Hz à 10⁶ Hz). Dans ce chapitre, une approche intégrative combinant la mesure expérimentale de la dynamique des protéines avec des modèles de réseau est proposée pour le diagnostic des perturbations dynamiques de la sous-unité B des toxines AB₅ causées par des mutations à haute résolution.

Mesure expérimentale de la dynamique multi-échelle des protéines

La spectroscopie diélectrique à large bande (BDS) mesure la réponse diélectrique d'un matériau lors de l'application d'un champ électrique. Les processus de relaxation dans l'échantillon sont détectés dans les spectres BDS par des pics de $\varepsilon''(\omega)$, avec ε'' la partie imaginaire de la permittivité diélectrique complexe de l'échantillon et ω la fréquence angulaire du champ électrique appliqué. La fréquence angulaire de perte maximale ω_P (position du pic) fournit le temps de relaxation τ_P comme $\tau_P = 2\pi/\omega_P$ ou de manière équivalente le taux de relaxation caractéristique $f_P = 1/\tau_P$ des dipôles fluctuants qui contribuent au processus de relaxation.

De la même manière, les processus de relaxation des dipôles spécifiques correspondent aux pics des courbes $\varepsilon''(T)$ mesurés à fréquence constante f et température variable T . Des processus de relaxation plus rapides (τ_P plus petit et f_P plus élevée) sont associés à une température plus basse de perte diélectrique maximale θ_{max} (position du pic). Les résultats présentés dans ce chapitre sont basés sur des courbes $\varepsilon''(T)$ obtenues à fréquences fixes.

Le protocole expérimental suivi par Laëtitia Bourgeat est le suivant [25, 26]. Une goutte de solutions protéique est déposée sur une nanomembrane. L'échantillon est chauffé à 50°C pendant 15 minutes pour évaporer l'eau libre et permettre aux protéines d'entrer dans les nanopores, puis refroidi à 30°C pendant 5 minutes. En suite l'échantillon est traité thermiquement à une température T_t pendant trois heures puis refroidi à -80°C à une vitesse de 2 K/min. La mesure diélectrique est effectuée pendant le refroidissement, fournissant la mesure $\varepsilon''(T)$.

Modèles de réseaux électrostatiques d'interactions dipolaires dans la structure protéique

Le modèle Réseau Électrostatique (EN) est utilisé pour analyser le réseau de dipôles dans une structure protéique. Les réseaux électrostatiques fournissent une indication des principales interactions électrostatiques attendues dans l'échantillon de protéine, à la granulosité des acides aminés. Par approximation, les liens dans le réseau électrostatique sont appelés dipôles et les liens dans le réseau électrostatique induit sont appelés dipôles induits. Des modèles de réseaux électrostatiques sont utilisés pour valider l'interprétation du signal BDS de CtxB₅ obtenu après traitement thermique et pour comparer la dynamique expérimentale de CtxB₅ et LTB₅ mesuré par BDS.

Approche intégrative pour sonder la dynamique multi-échelle de la toxine CtxB₅

La dynamique de dépliement thermique de CtxB₅ a été détectée à l'aide de BDS après traitement thermique de la protéine [26, 27]. Trois pics sont observés dans la courbe

de perte diélectrique $\varepsilon''(T)$. Ces pics sont nommés P_{1C} , P_{2C} et P_{3C} . Le processus de relaxation détecté par P_{1C} est plus rapide que le processus détecté par P_{3C} , à son tour plus rapide que le processus détecté par P_{2C} [27]. Il a été trouvé que le pic P_{1C} est présent après traitement thermique à toutes températures de 60°C à 180°C, le pic P_{3C} est présent après traitement thermique à des températures de 60°C à 140°C, et le pic P_{2C} est présent après traitement thermique à des températures de 80°C à 180°C. Ainsi, les pics correspondent aux fluctuations moléculaires présentes à différents stades de dépliement de la toxine [26].

Dans [26] et [27], les résultats sont interprétés comme suit :

- P_{1C} est associé à deux ensembles de dipôles, l'un à fluctuations lentes et l'autre à fluctuations rapides.
- L'ensemble avec des mouvements rapides est également détecté dans P_{3C} .
- Un troisième ensemble de dipôles avec des mouvements lents est détecté dans P_{2C} .

Sur la base des fréquences de détection et du fait que la position des pics varie avec la température de traitement, les pics du spectre BDS ont été attribués aux mouvements interdomaines à différents stades du dépliement de la toxine. Le modèle suivant a été proposé :

- P_{1C} et P_{3C} sont associés à des pentamères non natifs ayant des conformations différentes.
- P_{2C} est associé à des intermédiaires d'assemblage, ce qui implique que les interfaces de la toxine sont sondées par le signal BDS.

Pour que le modèle soit valide, les conditions suivantes doivent être remplies :

1. Les interfaces des toxines doivent être thermosensibles ;
2. Les interfaces des toxines doivent être détectables par diélectriquement ;
3. Les interfaces de la toxine doivent être composées de trois ensembles différents de relaxations dipolaires avec une sensibilité thermique distincte.

Les modèles réseau d'acides aminés intermoléculaires (4D-AAN), réseau d'acides aminés intermoléculaires (4D-EN) et réseau électrostatique induit intermoléculaire (Induced 4D-EN) de CtxB₅ sont utilisés pour déterminer si ces conditions sont valides.

Les interfaces I_{1a} , I_{1b} , I_{2a} , I_{2b} , I_{2c} et I_{3c} de CtxB₅ sont classés en fonction du nombre de liens dans le 4D-AAN et de la somme des poids des liens dans le 4D-AAN. Il est trouvé que I_{1a} et I_{2c} sont les plus faibles, I_{1b} et I_{2b} sont modérés, et I_{2a} et I_3 sont les plus forts. Donc, les interfaces sont thermosensibles (condition 1) et devraient être perturbées dans l'ordre suivant lors de la dénaturation thermique : I_{1a} et I_{2c} d'abord, puis I_{1b} et I_{2b} , et I_{2a} et I_3 en dernier.

Les liens dans le 4D-EN et le 4D-EN induit indiquent si des dipôles et des dipôles induits sont présents aux interfaces, et donc si les interfaces sont détectables diélectriquement. La détectabilité des interfaces (condition 2) suit l'ordre suivant : $I_{1a} < I_{2a}$ et $I_{2c} < I_{1b}$ et $I_{2b} < I_3$.

Nous trouvons que I_{1a} et I_{2c} sont les plus sensibles thermiquement mais ne peuvent pas être détectés par BDS ou seulement faiblement par les dipôles induits. La perturbation thermique de I_{1a} sera détectée lorsque I_{1b} est également perturbé thermiquement. De même, la perturbation thermique du domaine I_{2c} sera détectée via la perturbation des dipôles induits et des dipôles ioniques lorsque I_{2a} et I_{2b} sont thermiquement perturbés.

Basé sur ces observations, nous proposons que P_{1C} détecte le dépliement de l'interface I_{1b} et de l'interface I_{2b} , P_{3C} détecte le dépliement ultérieur de l'interface I_{2b} et P_{2C} détecte le dépliement de l'interface I_3 (condition 3).

Les résultats présentés dans cette section montrent que le modèle EN est un outil approprié pour guider l'interprétation du signal BDS, en sondant la dynamique des protéines à plusieurs échelles. L'avantage de cette approche intégrative est la possibilité de localiser des dipôles spécifiques responsables de la dynamique expérimentale observée et contrôlant le dépliement de la toxine. Grâce à la haute résolution de la mesure, reliant les interactions locales (dipôles, c'est-à-dire les liens dans le réseau électrostatique) aux mouvements globaux (dynamique de dépliement), c'est une technique prometteuse pour étudier les différences dynamiques globales dans les variants de protéines causées par des perturbations locales (mutations d'acide aminés entraînant des perturbations des liens). Pour investiguer cette possibilité, les dynamiques de CtxB₅ et LTB₅ sont comparées à l'aide d'une analyse de réseau couplée à la mesure expérimentale de BDS.

Mesure de la perturbation dynamique multi-échelle lors des mutations : toxines CtxB₅ versus LTB₅

Dix-sept mutations d'acides aminés existent dans LTB₅ par rapport à CtxB₅. Ces mutations n'impactent pas la structure des deux toxines mais rendent LTB₅ plus stable et amènent les deux toxines à se déplier et à se replier par des mécanismes différents. Dans le Chapitre 8, ces différences étaient associées à différentes allocations d'interactions atomiques au niveau de l'hélice α centrale, le domaine C-terminal et le domaine N-terminal. Ici, la dynamique expérimentale des deux toxines lors de la dénaturation thermique est comparée sur la base de la mesure expérimentale BDS et les modèles de réseau électrostatique des deux toxines sont utilisés pour interpréter les différences de signal diélectrique.

Quatre pics de signal BDS ont été observés pour LTB₅ après différents traitements thermiques et à différentes fréquences [28]. Les pics sont nommés P_{1L} , P_{2L} , P_{3L} et P_{4L} , classés de la position la plus à gauche à la plus à droite dans le axe de température. La position des pics en fonction de la température de traitement thermique pour LTB₅ et CtxB₅ a été comparée pour déduire s'il existe une correspondance entre les pics observés dans LTB₅ et ceux observés dans CtxB₅. Deux conclusions principales sont tirées : tout d'abord, des processus de relaxation diélectrique sont observés après traitement thermique à une température plus élevée en LTB₅ par rapport à CtxB₅. Ceci est cohérent avec la résistance thermique plus élevée de LTB₅ et confirme que le signal BDS aux fréquences

utilisées ici ne sonde pas la conformation native de la toxine, mais uniquement les fluctuations dipolaires résultant de la perturbation thermique de la structure de la protéine. Deuxièmement, l'ordre d'apparition des fluctuations des dipôles n'est pas le même dans les deux toxines, ce qui correspond à des mécanismes de dépliement différents.

Il est proposé que P_{1L} détecte le dépliement de l'interface I_{2b} , P_{2L} détecte le dépliement de l'interface I_3 , et P_{3C} détecte le dépliement de l'interface I_{1b} et de l'interface I_{2b} . Les pics P_{4C} et P_{4L} apparaissent après traitement thermique à haute température (180°C) et correspondraient à des états dépliés.

Il est important de noter que l'affectation des pics est basée sur la position des pics, c'est-à-dire un indicateur de l'échelle de temps caractéristique de la fluctuation. Néanmoins, la largeur des pics est également une caractéristique importante, car des pics plus larges correspondent à des processus de relaxation plus hétérogènes. P_{2C} et P_{2L} sont affectés aux mêmes ensembles de dipôles (interface I_3), mais le pic P_{2C} est plus large par rapport au P_{2L} pics. Il est proposé que P_{2C} soit plus large car il sonde la dynamique de l'interface I_3 dans les intermédiaires d'assemblage, tandis que P_{2L} sonde la dynamique de l'interface I_3 dans les pentamères.

Les domaines interfaciaux des deux pentamères I_{1a} , I_{1b} , I_{2a} , I_{2b} , I_{2c} et I_3 ont le même classement de stabilité et le même contenu de dipôle électrostatique intermoléculaire, inconsistant avec les deux signaux diélectriques différents. Cela suggère que les dipôles électrostatiques intramoléculaires contribuent également aux signaux diélectriques et sont responsables des différences. La comparaison des EN montre que LTB_5 possède deux dipôles électrostatiques intramoléculaires supplémentaires (A1, E7) et (K81, E102) par rapport à $CtxB_5$, en raison des mutations A1T, E7D et E102A aux positions 1, 7 et 102. Nous pouvons nous attendre à ce que les dipôles (A1, E7) et (K81, E102) augmentent la stabilité du N-terminal de LTB_5 et des contacts 3D entre le β_6 (96-103) et l' α -hélice centrale (59-79), respectivement. De plus, le N-terminal est impliqué dans les interfaces I_{1a} et I_{1b} , donc nous pouvons nous attendre à ce que les deux interfaces soient plus stables et moins sensibles à la dissociation en LTB_5 .

L'augmentation des interactions dipolaires aux terminales N et C (interfaces I_1 et I_2) de LTB_5 par rapport à $CtxB_5$ est cohérente avec les différences d'allocation des interactions atomiques observées au Chapitre 8, qui a montré que LTB_5 alloue plus d'interactions à la maintenance des interfaces et moins à la maintenance de la structure secondaire dans LTB_5 par rapport à $CtxB_5$.

Conclusion

La combinaison des résultats expérimentaux et de l'analyse du réseau révèle le rôle potentiel des deux dipôles spécifiques de LTB_5 (A1, E7) et (K81, E102) et des mutations A1T, E7D et E102A dans la déviation de LTB_5 vers des chemins de déploiement qui rendent LTB_5 plus résistant à la chaleur et moins propice à la dissociation thermique par

les intermédiaires d'assemblage.

Plus généralement, les résultats présentés dans les Chapitres 7 à 11 ont prouvé que les informations sur la dynamique des protéines sont codées dans le modèle AAN de la structure de la protéine. Même si les connaissances actuelles ne permettent pas de décoder la dynamique de la structure de la protéine ou de son AAN, les études de cas présentées dans ces chapitres montrent que l'échelle de la dynamique de la protéine est contrôlée par l'allocation des interactions atomiques (liens dans l'AAN) aux structures niveaux (1D, 2D, 3D et 4D). Par conséquent, l'impact d'une mutation sur la structure de la protéine ne dépend pas seulement du nombre d'interactions atomiques qui sont perturbées (Chapitre 9) ou de l'échelle spatiale couverte par la perturbation (Chapitre 6), mais aussi au niveau structurel des interactions perturbées (Chapitre 8 et chapitre actuel). Ceci implique que la conséquence des mutations dépend du sens de la perturbation qu'elles provoquent, vers les niveaux structuraux secondaire, tertiaire ou quaternaire. C'est-à-dire, pour évaluer l'impact d'une mutation d'acide aminé sur la dynamique des protéines, il est important de déterminer quels autres acides aminés sont perturbés.

C.11 Preuve de concept : Troisième domaine PDZ. Directions des perturbations causées par les mutations des acides aminés dans la structure des protéines mesurées par un réseau dirigé.

Introduction

Dans ce chapitre, un outil de calcul nommé le réseau GCAT est proposé pour décrire la perturbation causée par toutes les mutations in-silico d'acides aminés d'une protéine dans une seule représentation. Une analyse exploratoire du réseau GCAT d'une structure protéique est proposée. Ensuite, la pertinence de l'outil réseau GCAT pour étudier l'interdépendance entre les mutations est discutée.

Le réseau GCAT de PSD95^{pdz3} est présenté. Ce cas d'étude a été choisi comme dans le Chapitre 6 car l'impact fonctionnel de la plupart de ses mutations d'acides aminés simples est connu à partir des expériences [16].

Dans le réseau GCAT, un arc lie les nœuds i au nœuds j s'il existe une mutation in silico de i qui cause le changement du voisinage de j dans le réseau d'acides aminés (AAN). Par conséquent, un arc $i \rightarrow j$ dans le réseau GCAT indique " i peut perturber j " ou de manière équivalente " j peut être perturbé par i ". Nous utilisons l'expression "peut perturber" et pas simplement "perturbe" car un arc $i \rightarrow j$ est présent s'il existe une mutation de i qui perturbe j , mais pas nécessairement toutes les mutations de i perturbent j . Donc le nombre d'arcs quittant un nœud i , c'est-à-dire le degré de sortie $k_{out,i}$, représente le potentiel que i perturbe un autre acide aminé j (en considérant toutes les dix-neuf $i \rightarrow i'$

mutations ensemble). Inversement, le nombre d'arcs entrant dans un nœud i , c'est-à-dire le degré $k_{in,i}$, représente le potentiel que l'acide aminé i soit perturbé par un autre acide aminé j (en considérant toutes dix-neuf $j \rightarrow j'$ mutations ensemble).

Nous avons classé les nœuds (positions d'acides aminés) dans le réseau GCAT selon leur potentiel de perturber ou d'être perturbé, mesurée par le degré d'entrée et de sortie. Quatre classes - G pour Generate, C pour Connect, A pour Absorb et T pour Transmit - sont définies comme suit : G if $k_{out,i} > \bar{k}$ et $k_{in,i} \leq \bar{k}$; C si $k_{out,i} \leq \bar{k}$ et $k_{in,i} \leq \bar{k}$; A si $k_{out,i} \leq \bar{k}$ et $k_{in,i} > \bar{k}$; T si $k_{out,i} > \bar{k}$ et $k_{in,i} > \bar{k}$, avec $\bar{k} = \bar{k}_{in} = \bar{k}_{out}$ le degré moyen d'entrée et de sortie dans le réseau (la notation \bar{k} correspond à une moyenne de k).

Il convient de noter que les nœuds voisins du réseau GCAT ne sont pas nécessairement des nœuds voisins du AAN de la structure protéique.

Analyse exploratoire du réseau GCAT de PSD95^{pdz3}

Nous avons investigué les propriétés du réseau GCAT de PSD95^{pdz3} comparé au AAN.

Le réseau GCAT et l'AAN ont une connectivité similaire, le réseau GCAT étant légèrement plus connecté que le réseau WT. La distance géodésique $d_{AAN}(i, j)$ entre les nœuds voisins dans le réseau GCAT varie de 1 à 3, ce qui est cohérent avec les effets de perturbation de courte à longue échelle spatiale (Chapitre 6). Nous avons trouvé que peu d'acides aminés perturbent plus de 60% de leurs voisins chimiques (voisins dans l'AAN) lorsqu'ils sont mutés.

Une différence importante entre le réseau GCAT et les autres réseaux structurés présentés dans les chapitres précédents est que le réseau GCAT est dirigé. Parmi les 656 paires de nœuds du réseau GCAT qui sont reliées par un arc, seulement 140 (environ 20%) sont reliées par un arc bidirectionnel ($i \leftrightarrow j$). Par différence, environ 80% des arcs ne sont pas bidirectionnels. Cela signifie que si la mutation d'un acide aminé i perturbe l'acide aminé j , dans 80% des cas la mutation de j ne perturbe pas l'acide aminé i .

Le degré moyen d'entrée et de sortie dans le réseau GCAT est $\bar{k} = 7$. Les degrés d'entrée et de sortie s'étendent sur une large gamme de valeurs, de 0 à 23, et ne sont pas corrélés. L'absence de corrélation entre les degrés d'entrée et de sortie n'est pas surprenante, en raison de la directionnalité des perturbations. Les degrés dans le réseau GCAT ne sont que faiblement corrélés avec le degré dans l'AAN, écartant l'hypothèse selon laquelle le nombre d'acides aminés autour d'un acide aminé détermine combien d'acides aminés il peut perturber ou combien d'acides aminés peuvent le perturber. Cela montre que la mesure de l'encombrement à distance chimique n'est pas suffisante pour déterminer le potentiel perturbateur des mutations des acides aminés.

Étant donné que les degrés d'entrée et de sortie des nœuds ne sont pas corrélés, toutes les classes GCAT peuvent être peuplées. Parmi les 113 nœuds du réseau GCAT, 21 appartiennent à la classe G, 56 appartiennent à la classe C, 17 appartiennent à la classe A et 19 appartiennent à la classe T.

Les classes C et A, correspondantes à des acides aminés qui perturbent peu d’acides aminés, sont grossièrement associées aux acides aminés exposés en surface. À l’inverse, les classes G et T, correspondant à des acides aminés qui perturbent de nombreux acides aminés, sont grossièrement associées à des acides aminés du noyau structurel. Cependant, toutes les classes GCAT sont peuplées d’acides aminés exposés en surface et enterrés. Cela signifie que le potentiel d’un acide aminé à perturber ou à être perturbé n’est pas strictement déterminé par sa position dans la structure de la protéine.

Pour déterminer si les acides aminés enfouis sont plus bien connectés les uns aux autres par rapport aux acides aminés de surface, nous avons extrait les *k*-cores (noyaux *k*) du réseau GCAT. Les *k*-cores sont les sous-ensembles maximaux de nœuds tels que chaque nœud du sous-ensemble est connecté à au moins *k* autres nœuds du sous-ensemble [79]. Le *core* interne du réseau GCAT de PSD95^{pdz3} est le 13-cores (noyaux de degré 13). Il est composé de sept nœuds A, six nœuds T et trois nœuds G. Ces nœuds font partie des domaines terminales N et C, le segment 354-356, plus les acides aminés 392 et 394. Parmi les seize acides aminés appartenant au 13-core, sept font partie du noyau structurel et neuf sont exposés en surface. Cela montre que le noyau du GCAT ne correspond pas au noyau structurel de la protéine. Ainsi, malgré le fait que les classes G et T (degré élevé) soient grossièrement associées aux acides aminés du noyau structurel, le sous-ensemble des nœuds les plus connectés mesurés par le noyau interne du GCAT ne correspond pas aux acides aminés du noyau structurel.

Vingt positions de PSD95^{pdz3} provoquent expérimentalement une perte d’activité de liaison au ligand de PSD95^{pdz3} par rapport à la protéine WT [16]. Nous appelons ces positions fonctionnellement sensibles. Le sous-réseau des positions fonctionnellement sensibles est connecté dans le réseau GCAT, ce qui signifie que les positions fonctionnellement sensibles ont tendance à se perturber entre elles lorsqu’elles sont mutées. Il est à noter que les positions fonctionnellement sensibles peuplent toutes les classes GCAT, donc l’impact fonctionnel des mutations n’est pas déterminé par la quantité de perturbation provoquée.

Dans le sous-réseau des positions fonctionnellement sensibles du réseau GCAT, tous les nœuds non hotspot (non liées au ligand de la protéine) à l’exception du nœud 341 ont un arc qui pointe directement vers une position hotspot (liée au ligand de la protéine). Cela signifie que la mutation des acides aminés correspondant à ces nœuds provoque un changement de voisinage des positions des hotspots, expliquant le changement d’activité de liaison de ligand. Le nœud 341 est connecté à des positions hotspot non pas par un arc mais par un chemin dans le réseau GCAT. Cela suggère que les chemins dans le réseau GCAT peuvent mettre en évidence des chemins allostériques dans la structure de la protéine, c’est-à-dire des chaînes de réarrangements séquentiels d’interactions atomiques qui conduisent dynamiquement à un changement dans la fonction des protéines. L’hypothèse sous-jacente est que les perturbations sondées par les mutations des acides aminés représentent également des chemins de mouvements atomiques.

Selon l'hypothèse que les arcs du réseau GCAT correspondent à des chemins de mouvements atomiques, les chemins du réseau GCAT ont été interprétés comme des chemins allostériques possibles dans la structure de la protéine. Cependant, une autre interprétation est possible: les arcs du réseau GCAT peuvent souligner l'interdépendance entre les mutations des acides aminés. Pour étudier l'hypothèse selon laquelle les arcs du réseau GCAT soulignent les interactions coopératives, nous avons analysé le sous-réseau du réseau GCAT impliquant les acides aminés de l'hélice2 car les interactions coopératives entre ces acides aminés ont été mesurées [225]. Conformément à l'hypothèse, la plupart des interactions coopératives $i - j$ mesurées expérimentalement correspondent à des arcs ou des chemins $i \rightarrow j$ ou $j \rightarrow i$ dans le réseau GCAT.

Ranganathan et ses collaborateurs ont montré que les positions fonctionnellement sensibles de PSD95^{pdz3} font partie du secteur protéique de PSD95^{pdz3}, c'est-à-dire un réseau de positions qui co-évolvent [16, 72, 226]. Dans [16], il a été conclu que l'architecture du secteur peut correspondre à une conception d'adaptabilité aux changements environnementaux. Une interprétation supplémentaire des chemins dans le réseau GCAT pourrait être que les chemins connectent des acides aminés en co-évolution. Cette interprétation s'accorde bien avec les interprétations précédentes des chemins sous-jacents aux chemins allostériques ou aux interactions coopératives, car l'allostérie, la coopérativité et la co-évolution ne sont pas indépendantes.

Alors que la co-évolution des positions des acides aminés peut être entraînée par l'adaptabilité à un nouvel environnement (changement fonctionnel ou évolutivité), la co-évolution peut également être une manifestation de mécanismes de correction d'erreur. En fait, la correction d'erreurs est une autre manifestation des interactions coopératives entre les acides aminés. Le réseau GCAT peut être exploité pour filtrer les positions candidates k qui ont le potentiel de fournir un mécanisme de sauvetage pour une mutation $i \rightarrow i'$. Nous proposons les critères suivants à explorer dans cette direction :

1. Tous les voisins sortants de i (un arc $i \rightarrow k$ est présent).
2. Tous les nœuds qui partagent des voisins sortants avec i ($N_{out}(i) \cap N_{out}(k) \neq \emptyset$).
3. Tous les voisins entrants de i (un arc $k \rightarrow i$ est présent).

Cependant, des critères plus complexes peuvent être nécessaires.

La validation des hypothèses de mécanisme de correction d'erreur proposées pour les perturbations structurelles et dynamiques nécessiterait une analyse approfondie de la structure et de la dynamique des variantes protéiques avec des mutations d'acides aminés simples et multiples, et est hors de la portée de cette étude. Cependant, en se concentrant sur le cas plus simple des mutations qui provoquent des changements dans les voisinages d'acides aminés, le réseau GCAT de PSD95^{pdz3} a été exploré en termes de similarité Jaccard entre les voisinages.

Nous avons trouvé voisinages sortants égaux pour les positions 364 avec 387 et 402 avec 405. Nous avons inspecté les perturbations structurelles causées par les mutations

des acides aminés G364, T387, R405 et A402 en comparant les AAN des mutants et WT. Cela nous a permis d'indiquer des possibles mécanismes de correction d'erreur par doubles mutations.

Conclusion

En résumé, un nouvel outil appelé le réseau GCAT a été présenté dans ce chapitre pour modéliser les perturbations directionnelles causées par toutes les mutations possibles des acides aminés dans une structure protéique.

Une découverte importante est que les acides aminés ont un potentiel différent de perturber et d'être perturbé les uns par rapport aux autres, mesuré par la connectivité dans le réseau GCAT et résultant en la population des quatre classes GCAT. De plus, le réseau GCAT montre une topologie noyau-périphérie, où le noyau interne ne correspond pas au noyau structurel de la structure protéique. Cela suggère une conception structurelle complexe de la protéine régissant la réponse aux perturbations.

Une relation entre les chemins dans le réseau GCAT et les mécanismes allostériques, les interactions coopératives entre les acides aminés et la co-évolution a été proposée. De plus, une stratégie possible pour extraire des informations sur la possibilité de correction d'erreurs par mutations multiples a été suggérée. Ces hypothèses restent spéculatives pour le moment, mais elles pourraient être vérifiées à l'aide d'informations évolutives et de dynamiques expérimentales ou simulées de variantes de protéines avec des mutations d'acides aminés simples et multiples.

C.12 Conclusion et perspectives

L'objectif de ce manuscrit était l'étude de la durabilité des protéines et l'évaluation de la possibilité d'utiliser des structures protéiques comme modèle pour la conception de structures urbaines durables bio-inspirées.

La durabilité d'un système est ici définie comme la capacité à maintenir la fonctionnalité dans le temps. Cela nécessite la mise en œuvre de deux réponses aux perturbations : le maintien de la fonctionnalité (le système est robuste) et l'adaptation de la fonction aux différents environnements (le système est évolutif).

Nous nous sommes concentrés sur les systèmes spatiaux, c'est-à-dire systèmes matériels dont la fonctionnalité repose sur l'agencement des composants du système dans l'espace. Dans ces cas, la structure du système est définie par la taille et les formes des composants, la proximité spatiale entre les composants et leur orientation relative. Ces paramètres décrivent comment l'espace est occupé dans la structure, c'est-à-dire quelles régions de l'espace sont remplies par les composants du système et quelles régions sont laissées disponibles pour la dynamique du système et pour s'adapter aux changements dans les composants du système.

En utilisant une approche de modélisation de réseau, ce travail contribue à la question de déterminer quelles conceptions structurales assurent la durabilité des protéines et des structures urbaines. Cette question a été évaluée du point de vue de la réponse du système aux perturbations internes, ici limitée aux substitutions des composants du système.

Nous avons proposé de mesurer l'occupation de l'espace comme moyen de diagnostiquer la durabilité spatiale, localement (Chapitre 4) et au niveau multi-échelle (Chapitre 5). Nous constatons que la conception des structures protéiques, contrairement aux structures urbaines, fournit toujours un espace vide. Nous proposons qu'il s'agisse d'une caractéristique de durabilité spatiale car cet espace peut être exploité pour accueillir des substitutions de composants (mutations d'acides aminés) et de dynamiques (mouvements atomiques). Les impacts dynamiques et structurels des mutations ont été mesurés à partir des changements d'interactions entre les acides aminés, sondés par des liens dans les modèles de réseau. En utilisant une approche *in-silico*, il a été montré que certaines mutations d'acides aminés peuvent être adaptées sans avoir besoin de réarrangement des interactions, tandis que d'autres mutations d'acides aminés nécessitent des changements dans les voisinages d'acides aminés de l'échelle locale à la grande échelle pour maintenir l'intégrité structurale (Chapitre 6). Pour diagnostiquer l'impact des mutations des acides aminés sur la dynamique des protéines, il a été proposé de mesurer les changements d'allocation des contacts atomiques aux quatre niveaux de descriptions des structures protéiques (Chapitre 7). Grâce à cette mesure, il a été possible de retrouver les différences dynamiques connues entre les sous-unités B des toxines AB₅ et les variants de la Transthyréline et d'étudier d'éventuelles différences dynamiques entre les variants des protéases principales des coronavirus (Chapitre 8). Le cas du variant pathogène de la Transthyréline L55P a été étudié plus en détail et l'analyse de la perturbation des contacts atomiques provoquée par la mutation couplée à une approche de pavage a permis l'inférence d'un modèle de fibre amyloïde cohérent avec les preuves expérimentales et les contraintes géométriques (Chapitre 9). Concernant les sous-unités B des toxines AB₅, la comparaison des réseaux de dipôles des structures protéiques a été appliquée avec succès à l'interprétation du signal diélectrique expérimental, qui sonde les différentes dynamiques moléculaires des deux toxines (Chapitre 10). Enfin, un nouvel outil appelé le réseau GCAT a été défini pour décrire la perturbation directionnelle des voisinages d'acides aminés causée par des mutations *in-silico*, ouvrant la possibilité d'étudier la contribution de chaque mutation d'acide aminé dans des variantes de protéines où de multiples mutations agissent collectivement et décrypter les mécanismes de correction d'erreurs et d'évolutivité des structures protéiques (Chapitre 11).

L'approche réseau nécessite de choisir la granularité qui définit les nœuds du réseau et la distance de coupure qui définit la connectivité entre les nœuds. Il s'agit d'un choix critique, car différents choix de ces paramètres peuvent conduire à différentes topologies de réseau. Cependant, la possibilité de moduler la granularité du modèle et le seuil de

connectivité offre également la liberté de choisir l'échelle spatiale à analyser.

Dans l'ensemble, les résultats présentés dans ce manuscrit montrent l'importance des interactions atomiques, mesurées par des poids de liaison dans l'AAN, dans le codage de la réponse aux perturbations des structures protéiques où la perturbation est une mutation d'acides aminés.

Un corollaire de la découverte que les changements dans les interactions atomiques déterminent les différences de dynamique entre les variantes est que la dynamique des protéines est codée par l'allocation des interactions atomiques aux niveaux structuraux. Ainsi, la mesure d'allocation pourrait représenter une "empreinte digitale" de la dynamique des protéines. Si une telle "empreinte dynamique" de la structure de la protéine pouvait être décodée, à l'avenir, il pourrait être possible de déduire le mécanisme de repliement et de dépliement et/ou la dynamique fonctionnelle de la protéine à partir de sa structure cristalline.

Concernant les systèmes urbains, les travaux futurs pourraient être consacrés à la création d'outils pour proposer automatiquement des solutions durables sous certaines contraintes, par exemple les géométries des bâtiments et leur distance maximale et minimale. De tels outils pourraient être utilisés dès les premières étapes de la conception urbaine. De plus, le modèle BN pourrait être affiné pour prendre en considération la hauteur des bâtiments, écartée dans le présent travail mais fondamentale pour déterminer l'impact de la structure urbaine multi-échelle sur les flux d'air et la pollution dans la ville. Une couche d'information supplémentaire qui pourrait être ajoutée au modèle BN, sous sa forme actuelle ou sous une forme tridimensionnelle, est la fonction des bâtiments et la quantité de personnes qui accèdent au bâtiment ou y habitent. Ceci est particulièrement pertinent pour évaluer la relation entre les mesures de BN et le potentiel d'adaptation du trafic dans la ville. De plus, il peut être intéressant d'étudier la possibilité d'utiliser la procédure d'union des bâtiments *per se* comme moyen de repérer les zones de la ville qui seraient les mieux adaptées pour être réservées à la mobilité piétonne, par rapport aux zones vers lesquelles le trafic automobile devrait être redirigé.

Bibliography

- [1] Robert Costanza and Bernard C Patten. “Defining and predicting sustainability”. In: *Ecological economics* 15.3 (1995), pp. 193–196.
- [2] Chang-Chun Zhou, Guo-Fu Yin, and Xiao-Bing Hu. “Multi-objective optimization of material selection for sustainable products: artificial neural networks and genetic algorithm approach”. In: *Materials & Design* 30.4 (2009), pp. 1209–1215.
- [3] Yosef Rafeq Jabareen. “Sustainable urban forms: Their typologies, models, and concepts”. In: *Journal of planning education and research* 26.1 (2006), pp. 38–52.
- [4] Heiko Koziolk. “Sustainability Evaluation of Software Architectures: A Systematic Review”. In: *Proceedings of the joint ACM SIGSOFT conference–QoSA and ACM SIGSOFT symposium–ISARCS on Quality of software architectures–QoSA and architecting critical systems–ISARCS* (2011), p. 10. DOI: 10.1145/2000259.2000263.
- [5] Mary E Kasarda, Janis P Terpenney, Dan Inman, Karl R Precoda, John Jelesko, Asli Sahin, and Jaeil Park. “Design for adaptability (DFAD)—a new concept for achieving sustainable design”. In: *Robotics and Computer-Integrated Manufacturing* 23.6 (2007), pp. 727–734.
- [6] M Mörtl. ““Design for Upgrading” of machines and production processes: A guideline based on actual demands of industry and sustainable design”. In: *14th International Conference on Engineering Design ICED’03*. 2003, p. 10.
- [7] Brian Walker. “Resilience: what it is and is not”. In: *Ecology and Society* 25.2 (2020).
- [8] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge university press, 2016.
- [9] Katherine Henzler-Wildman and Dorothee Kern. “Dynamic personalities of proteins”. In: *Nature* 450.7172 (2007), pp. 964–972.
- [10] Lydia Robert, Jean Ollion, Jérôme Robert, Xiaohu Song, Ivan Matic, and Marina Elez. “Mutation dynamics and fitness effects followed in single cells”. In: *Science* 359.6381 (2018), pp. 1283–1286.
- [11] 1000 Genomes Project Consortium et al. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (2012), p. 56.
- [12] Özlem Demir, Roberta Baronio, Faezeh Salehi, Christopher D Wassman, Linda Hall, G Wesley Hatfield, Richard Chamberlin, Peter Kaiser, Richard H Lathrop, and Rommie E Amaro. “Ensemble-based computational approach discriminates functional activity of p53 cancer and rescue mutants”. In: *PLoS Comput Biol* 7.10 (2011), e1002238.
- [13] G Brian Golding and Antony M Dean. “The structural basis of molecular adaptation.” In: *Molecular biology and evolution* 15.4 (1998), pp. 355–369.
- [14] Janita Thusberg and Mauno Vihinen. “Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods”. In: *Human mutation* 30.5 (2009), pp. 703–714.

- [15] Mounia Achoch, Rodrigo Dorantes-Gilardi, Chris Wymant, Giovanni Feverati, Kave Salamatian, Laurent Vuillon, and Claire Lesieur. “Protein structural robustness to mutations: an in silico investigation”. In: *Physical Chemistry Chemical Physics* 18.20 (2016), pp. 13770–13780. DOI: 10.1039/C5CP06091E.
- [16] Richard N. McLaughlin Jr, Frank J. Poelwijk, Arjun Raman, Walraj S. Gosal, and Rama Ranganathan. “The spatial architecture of protein function and adaptation”. In: *Nature* 491.7422 (2012), pp. 138–142. DOI: 10.1038/nature11500.
- [17] Athi N Naganathan. “Modulation of allosteric coupling by mutations: from protein dynamics and packing to altered native ensembles and function”. In: *Current Opinion in Structural Biology* 54 (2019), pp. 1–9. DOI: 10.1016/j.sbi.2018.09.004.
- [18] Laurent Vuillon and Claire Lesieur. “From local to global changes in proteins: a network view”. In: *Current Opinion in Structural Biology* 31 (2015), pp. 1–8. DOI: 10.1016/j.sbi.2015.02.015.
- [19] Francesco Bemporad and Fabrizio Chiti. “Protein Misfolded Oligomers: Experimental Approaches, Mechanism of Formation, and Structure-Toxicity Relationships”. In: *Chemistry & Biology* 19.3 (2012), pp. 315–327. DOI: 10.1016/j.chembiol.2012.02.003.
- [20] Robin Burnette, Leigh Ann Simmons, and Ralph Snyderman. “Personalized health care as a pathway for the adoption of genomic medicine”. In: *Journal of personalized medicine* 2.4 (2012), pp. 232–240.
- [21] Kristina Wanieck, Pierre-Emmanuel Fayemi, Nicolas Maranzana, Cordt Zollfrank, and Shoshanah Jacobs. “Biomimetics and its tools”. In: *Bioinspired, Biomimetic and Nanobiomaterials* 6.2 (2017), pp. 53–66.
- [22] Rodrigo Dorantes Gilardi. “Bio-mathematical aspects of the plasticity of proteins”. PhD thesis. 2018. URL: <https://theses.fr/2018GREM092>.
- [23] Rodrigo Dorantes-Gilardi, Laëtitia Bourgeat, Lorenza Pacini, Laurent Vuillon, and Claire Lesieur. “In proteins, the structural responses of a position to mutation rely on the Goldilocks principle: not too many links, not too few”. In: *Physical Chemistry Chemical Physics* 20.39 (2018), pp. 25399–25410.
- [24] Lorenza Pacini, Laurent Vuillon, and Claire Lesieur. “Induced Perturbation Network and tiling for modeling the L55P Transthyretin amyloid fiber”. In: *Procedia Computer Science* 178 (2020), pp. 8–17.
- [25] Laëtitia Bourgeat. “Etude expérimentale de la dynamique moléculaire des protéines : Nouvelle méthode de criblages de conformations protéiques”. PhD thesis. 2020. URL: <https://theses.fr/s200911>.
- [26] Laëtitia Bourgeat, Anatoli Serghei, and Claire Lesieur. “experimental protein Molecular Dynamics: Broadband Dielectric Spectroscopy coupled with nanoconfinement”. In: *Scientific reports* 9.1 (2019), pp. 1–12.
- [27] Lorenza Pacini, Laetitia Bourgeat, Anatoli Serghei, and Claire Lesieur. “Analysis of Nanoconfined Protein Dielectric Signals Using Charged Amino Acid Network Models”. In: *Australian Journal of Chemistry* 73.8 (2020), pp. 803–812.
- [28] Laëtitia Bourgeat, Lorenza Pacini, Anatoli Serghei, and Claire Lesieur. “Experimental diagnostic of sequence-variant dynamic perturbations revealed by broadband dielectric spectroscopy”. In: *Structure* (2021).
- [29] Heriberto Cabezas and Brian D Fath. “Towards a theory of sustainable systems”. In: *Fluid phase equilibria* 194 (2002), pp. 3–14.

- [30] Heriberto Cabezas, Christopher W Pawlowski, Audrey L Mayer, and N Theresa Hoagland. “Sustainable systems theory: ecological and other aspects”. In: *Journal of Cleaner Production* 13.5 (2005), pp. 455–467.
- [31] J Baird Callicott and Karen Mumford. “Ecological sustainability as a conservation concept: Sustentabilidad ecologica como concepto de conservacion”. In: *Conservation biology* 11.1 (1997), pp. 32–40.
- [32] John Morelli. “Environmental sustainability: A definition for environmental professionals”. In: *Journal of environmental sustainability* 1.1 (2011), p. 2.
- [33] Markus Vogt and Christoph Weber. “Current challenges to the concept of sustainability”. In: *Global Sustainability* 2 (2019).
- [34] Michael U Ben-Eli. “Sustainability: definition and five core principles, a systems perspective”. In: *Sustainability Science* 13.5 (2018), pp. 1337–1343.
- [35] Johan Rockström, Xuemei Bai, and Bert deVries. “Global sustainability: the challenge ahead”. In: *Global Sustainability* 1 (2018).
- [36] Tim Cooper. “Beyond recycling: The longer life option”. In: (1994).
- [37] Clare M O’Connor, Jill U Adams, and Jennifer Fairman. *Essentials of cell biology*. Cambridge, MA: NPG Education, 2010.
- [38] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD: visual molecular dynamics”. In: *Journal of molecular graphics* 14.1 (1996), pp. 33–38.
- [39] Robert H Callender, R Brian Dyer, Rudolf Gilmanshin, and William H Woodruff. “Fast events in protein folding: the time evolution of primary processes”. In: *Annual review of physical chemistry* 49.1 (1998), pp. 173–202.
- [40] Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. “Atomic-resolution protein structure determination by cryo-EM”. In: *Nature* 587.7832 (2020), pp. 157–161.
- [41] Sumit Prakash and Andreas Matouschek. “Protein unfolding in the cell”. In: *Trends in biochemical sciences* 29.11 (2004), pp. 593–600.
- [42] Sebastian Buchenberg, Norbert Schaudinnus, and Gerhard Stock. “Hierarchical biomolecular dynamics: Picosecond hydrogen bonding regulates microsecond conformational transitions”. In: *Journal of chemical theory and computation* 11.3 (2015), pp. 1330–1336.
- [43] Juan R Perilla, Boon Chong Goh, C Keith Cassidy, Bo Liu, Rafael C Bernardi, Till Rudack, Hang Yu, Zhe Wu, and Klaus Schulten. “Molecular dynamics simulations of large macromolecular complexes”. In: *Current opinion in structural biology* 31 (2015), pp. 64–74.
- [44] Isabella Daidone, Andrea Amadei, Danilo Roccatano, and Alfredo Di Nola. “Molecular dynamics simulation of protein folding by essential dynamics sampling: folding landscape of horse heart cytochrome c”. In: *Biophysical journal* 85.5 (2003), pp. 2865–2871.
- [45] Michael Landreh, Erik G Marklund, Povilas Uzdaviny, Matteo T Degiacomi, Mathieu Coincon, Joseph Gault, Kallol Gupta, Ildir Liko, Justin LP Benesch, David Drew, et al. “Integrating mass spectrometry with MD simulations reveals the role of lipids in Na⁺/H⁺ antiporters”. In: *Nature communications* 8.1 (2017), pp. 1–9.
- [46] Martin Karplus and J Andrew McCammon. “Molecular dynamics simulations of biomolecules”. In: *Nature structural biology* 9.9 (2002), pp. 646–652.
- [47] Claire Lesieur and Klaus Schulten. “Editorial overview: Theory and simulation.” In: *Current opinion in structural biology* 31 (2015), pp. v–vi.

- [48] Scott A Hollingsworth and Ron O Dror. “Molecular dynamics simulation for all”. In: *Neuron* 99.6 (2018), pp. 1129–1143.
- [49] Frederic M Richards. “Areas, volumes, packing, and protein structure”. In: *Annual review of biophysics and bioengineering* 6.1 (1977), pp. 151–176.
- [50] Jie Liang and Ken A Dill. “Are proteins well-packed?” In: *Biophysical journal* 81.2 (2001), pp. 751–766.
- [51] T Andrew Binkowski, Larisa Adamian, and Jie Liang. “Inferring functional relationships of proteins from local sequence and spatial surface patterns”. In: *Journal of molecular biology* 332.2 (2003), pp. 505–526.
- [52] Xiang Li, Ozlem Keskin, Buyong Ma, Ruth Nussinov, and Jie Liang. “Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking”. In: *Journal of molecular biology* 344.3 (2004), pp. 781–795.
- [53] Maurizio Brunori. “Structural dynamics of myoglobin”. In: *Biophysical chemistry* 86.2-3 (2000), pp. 221–230.
- [54] Rajesh Ramachandran, Rodney K Tweten, and Arthur E Johnson. “Membrane-dependent conformational changes initiate cholesterol-dependent cytolysin oligomerization and intersubunit β -strand alignment”. In: *Nature structural & molecular biology* 11.8 (2004), pp. 697–705.
- [55] JM Passner, SC Schultz, and TA Steitz. “Modeling the cAMP-induced allosteric transition using the crystal structure of CAP-cAMP at 2.1 Å resolution”. In: *Journal of molecular biology* 304.5 (2000), pp. 847–859.
- [56] Nataliya Popovych, Shangjin Sun, Richard H Ebright, and Charalampos G Kalodimos. “Dynamically driven protein allostery”. In: *Nature structural & molecular biology* 13.9 (2006), pp. 831–838.
- [57] Victor Munoz and Michele Cerminara. “When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches”. In: *Biochemical Journal* 473.17 (2016), pp. 2545–2559.
- [58] Józef R Lewandowski, Meghan E Halse, Martin Blackledge, and Lyndon Emsley. “Direct observation of hierarchical protein dynamics”. In: *Science* 348.6234 (2015), pp. 578–581.
- [59] Hans Frauenfelder, Stephen G Sligar, and Peter G Wolynes. “The energy landscapes and motions of proteins”. In: *Science* 254.5038 (1991), pp. 1598–1603.
- [60] Konstantin Röder, Jerelle A Joseph, Brooke E Husic, and David J Wales. “Energy landscapes for proteins: from single funnels to multifunctional systems”. In: *Advanced Theory and Simulations* 2.4 (2019), p. 1800175.
- [61] Gozde Kar, Ozlem Keskin, Attila Gursoy, and Ruth Nussinov. “Allostery and population shift in drug discovery”. In: *Current opinion in pharmacology* 10.6 (2010), pp. 715–722.
- [62] Stefano Gianni, María Inés Freiburger, Per Jemth, Diego U Ferreira, Peter G Wolynes, and Monika Fuxreiter. “Fuzziness and frustration in the energy landscape of protein folding, function, and assembly”. In: *Accounts of chemical research* 54.5 (2021), pp. 1251–1259.
- [63] Leo C James and Dan S Tawfik. “Conformational diversity and protein evolution—a 60-year-old hypothesis revisited”. In: *Trends in biochemical sciences* 28.7 (2003), pp. 361–368.

- [64] Leo C James and Dan S Tawfik. “Catalytic and binding poly-reactivities shared by two unrelated proteins: The potential role of promiscuity in enzyme evolution”. In: *Protein Science* 10.12 (2001), pp. 2600–2607.
- [65] Amir Aharoni, Leonid Gaidukov, Olga Khersonsky, Stephen McQ Gould, Cintia Roodveldt, and Dan S Tawfik. “The ‘evolvability’ of promiscuous protein functions”. In: *Nature genetics* 37.1 (2005), pp. 73–76.
- [66] Hiroaki Kitano. “Biological robustness”. In: *Nature Reviews Genetics* 5.11 (2004), pp. 826–837.
- [67] Marie E Csete and John C Doyle. “Reverse engineering of biological complexity”. In: *science* 295.5560 (2002), pp. 1664–1669.
- [68] Yamir Moreno, Romualdo Pastor-Satorras, and Alessandro Vespignani. “Epidemic outbreaks in complex heterogeneous networks”. In: *The European Physical Journal B-Condensed Matter and Complex Systems* 26.4 (2002), pp. 521–529.
- [69] Anna S Garza, Nihal Ahmad, and Raj Kumar. “Role of intrinsically disordered protein regions/domains in transcriptional regulation”. In: *Life sciences* 84.7-8 (2009), pp. 189–193.
- [70] Jay F Storz. “Compensatory mutations and epistasis for protein function”. In: *Current opinion in structural biology* 50 (2018), pp. 18–25.
- [71] Erik D Nelson and Nick V Grishin. “Long-range epistasis mediated by structural change in a model of ligand binding proteins”. In: *PLoS One* 11.11 (2016), e0166739.
- [72] Arjun S Raman, K Ian White, and Rama Ranganathan. “Origins of allostery and evolvability in proteins: a case study”. In: *Cell* 166.2 (2016), pp. 468–480.
- [73] Frank J Poelwijk, Michael Socolich, and Rama Ranganathan. “Learning the pattern of epistasis linking genotype and phenotype in a protein”. In: *Nature communications* 10.1 (2019), pp. 1–11.
- [74] Martin Werner, Vytautas Gapsys, and Bert L. de Groot. “One Plus One Makes Three: Triangular Coupling of Correlated Amino Acid Mutations”. In: *The Journal of Physical Chemistry Letters* 12.XXX (2021), pp. 3195–3201.
- [75] Canan Atilgan, Osman Burak Okan, and Ali Rana Atilgan. “Network-based models as tools hinting at nonevident protein functionality”. In: *Annual review of biophysics* 41 (2012), pp. 205–225.
- [76] Wenying Yan, Jianhong Zhou, Maomin Sun, Jiajia Chen, Guang Hu, and Bairong Shen. “The construction of an amino acid network for understanding protein structure and function”. In: *Amino acids* 46.6 (2014), pp. 1419–1439.
- [77] Carlos H da Silveira, Douglas EV Pires, Raquel C Minardi, Cristina Ribeiro, Caio JM Veloso, Julio CD Lopes, Wagner Meira Jr, Goran Neshich, Carlos HI Ramos, Raul Habesch, et al. “Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 74.3 (2009), pp. 727–743.
- [78] Juan Salamanca Vilorio, Maria Francesca Allega, Matteo Lambrugh, and Elena Papaleo. “An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass”. In: *Scientific reports* 7.1 (2017), pp. 1–11.
- [79] Mark Newman. *Networks: An Introduction*. Oxford University Press, 2010.
- [80] Paul Erdős and Alfréd Rényi. “On the evolution of random graphs”. In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1 (1960), pp. 17–60.

- [81] Réka Albert and Albert-László Barabási. “Statistical mechanics of complex networks”. In: *Reviews of modern physics* 74.1 (2002), p. 47.
- [82] Paolo Crucitti, Vito Latora, and Sergio Porta. “Centrality measures in spatial networks of urban streets”. In: *Physical Review E* 73.3 (2006), p. 036125.
- [83] Luiz GA Alves, Alberto Aleta, Francisco A Rodrigues, Yamir Moreno, and Luís A Nunes Amaral. “Centrality anomalies in complex networks as a result of model over-simplification”. In: *New Journal of Physics* 22.1 (2020), p. 013043.
- [84] Lesley H Greene and Victoria A Higman. “Uncovering network systems within protein structures”. In: *Journal of molecular biology* 334.4 (2003), pp. 781–791.
- [85] Md Aftabuddin and Sudip Kundu. “Weighted and unweighted network of amino acids within protein”. In: *Physica A: Statistical Mechanics and its Applications* 369.2 (2006), pp. 895–904.
- [86] Claire Lesieur and Laurent Vuillon. “Topology Results on Adjacent Amino Acid Networks of Oligomeric Proteins”. In: *Allostery*. Springer, 2021, pp. 113–135.
- [87] Erzsébet Ravasz and Albert-László Barabási. “Hierarchical organization in complex networks”. In: *Physical review E* 67.2 (2003), p. 026112.
- [88] Tao Yu, Xiaoqin Zou, Sheng-You Huang, and Xian-Wu Zou. “Cutoff variation induces different topological properties: a new discovery of amino acid network within protein”. In: *Journal of theoretical biology* 256.3 (2009), pp. 408–413.
- [89] Réka Albert, Hawoong Jeong, and Albert-László Barabási. “Error and attack tolerance of complex networks”. In: *nature* 406.6794 (2000), pp. 378–382.
- [90] Francesca Fanelli, Angelo Felling, Francesco Raimondi, and Michele Seeber. “Structure network analysis to gain insights into GPCR function”. In: *Biochemical Society Transactions* 44.2 (2016), pp. 613–618.
- [91] I Arnold Emerson and KM Gothandam. “Residue centrality in alpha helical polytopic transmembrane protein structures”. In: *Journal of theoretical biology* 309 (2012), pp. 78–87.
- [92] Réka Albert, István Albert, and Gary L Nakarado. “Structural vulnerability of the North American power grid”. In: *Physical review E* 69.2 (2004), p. 025103.
- [93] Ben Derudder, Lomme Devriendt, and Frank Witlox. “Flying where you don’t want to go: An empirical analysis of hubs in the global airline network”. In: *Tijdschrift voor economische en sociale geografie* 98.3 (2007), pp. 307–324.
- [94] Gil Amitai, Arye Shemesh, Einat Sitbon, Maxim Shklar, Dvir Netanel, Ilya Venger, and Shmuel Pietrokovski. “Network analysis of protein structures identifies functional residues”. In: *Journal of molecular biology* 344.4 (2004), pp. 1135–1146.
- [95] Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. “Residues crucial for maintaining short paths in network communication mediate signaling in proteins”. In: *Molecular systems biology* 2.1 (2006), pp. 2006–0019.
- [96] Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. “Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families”. In: *Protein Science* 15.9 (2006), pp. 2120–2128.
- [97] I Arnold Emerson and Preeti Tabitha Louis. “Detection of active site residues in bovine rhodopsin using network analysis”. In: *Trends in Bioinformatics* 8.2 (2015), p. 63.

- [98] Søren W Gersting, Kristina F Kemter, Michael Staudigl, Dunja D Messing, Marta K Danecka, Florian B Lagler, Christian P Sommerhoff, Adelbert A Roscher, and Ania C Muntau. “Loss of function in phenylketonuria is caused by impaired molecular motions and conformational instability”. In: *The American Journal of Human Genetics* 83.1 (2008), pp. 5–17.
- [99] Jennifer M Axe, Eric M Yezdimer, Kathleen F O’Rourke, Nicole E Kerstetter, Wanli You, Chia-en A Chang, and David D Boehr. “Amino acid networks in a (β/α) 8 barrel enzyme change during catalytic turnover”. In: *Journal of the American Chemical Society* 136.19 (2014), pp. 6818–6821.
- [100] Jennifer M Axe, Kathleen F O’Rourke, Nicole E Kerstetter, Eric M Yezdimer, Yan M Chan, Alexander Chasin, and David D Boehr. “Severing of a hydrogen bond disrupts amino acid networks in the catalytically active state of the alpha subunit of tryptophan synthase”. In: *Protein Science* 24.4 (2015), pp. 484–494.
- [101] Michael D Daily, Tarak J Upadhyaya, and Jeffrey J Gray. “Contact rearrangements form coupled networks from local motions in allosteric proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 71.1 (2008), pp. 455–466.
- [102] Micol De Ruvo, Alessandro Giuliani, Paola Paci, Daniele Santoni, and Luisa Di Paola. “Shedding light on protein–ligand binding by graph theory: the topological nature of allostery”. In: *Biophysical Chemistry* 165 (2012), pp. 21–29.
- [103] Nikolay V Dokholyan, Lewyn Li, Feng Ding, and Eugene I Shakhnovich. “Topological determinants of protein folding”. In: *Proceedings of the National Academy of Sciences* 99.13 (2002), pp. 8637–8641.
- [104] Jie Li, Jun Wang, and Wei Wang. “Identifying folding nucleus based on residue contact networks of proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 71.4 (2008), pp. 1899–1907.
- [105] Michele Vendruscolo, Nikoley V Dokholyan, Emanuele Paci, and Martin Karplus. “Small-world view of the amino acids that play a key role in protein folding”. In: *Physical Review E* 65.6 (2002), p. 061910.
- [106] M Michael Gromiha. “Multiple contact network is a key determinant to protein folding rates”. In: *Journal of chemical information and modeling* 49.4 (2009), pp. 1130–1135.
- [107] Elena Papaleo. “Integrating atomistic molecular dynamics simulations, experiments, and network analysis to study protein dynamics: strength in unity”. In: *Frontiers in molecular biosciences* 2 (2015), p. 28.
- [108] Broto Chakrabarty, Dibyajyoti Das, Navneet Bung, Arijit Roy, and Gopalakrishnan Bulusu. “Network analysis of hydroxymethylbilane synthase dynamics”. In: *Journal of Molecular Graphics and Modelling* 99 (2020), p. 107641.
- [109] Kristin Blacklock and Gennady M Verkhivker. “Computational modeling of allosteric regulation in the hsp90 chaperones: a statistical ensemble analysis of protein structure networks and allosteric communications”. In: *PLoS Comput Biol* 10.6 (2014), e1003679.
- [110] Quan Li, Ray Luo, and Hai-Feng Chen. “Dynamical important residue network (DIRN): network inference via conformational change”. In: *Bioinformatics* 35.22 (2019), pp. 4664–4670.
- [111] Aria Gheeraert, Lorenza Pacini, Victor S Batista, Laurent Vuillon, Claire Lesieur, and Ivan Rivalta. “Exploring allosteric pathways of a v-type enzyme with dynamical perturbation networks”. In: *The Journal of Physical Chemistry B* 123.16 (2019), pp. 3452–3461.

- [112] Christian FA Negre, Uriel N Morzan, Heidi P Hendrickson, Rhitankar Pal, George P Lisi, J Patrick Loria, Ivan Rivalta, Junming Ho, and Victor S Batista. “Eigenvector centrality for characterization of protein allosteric pathways”. In: *Proceedings of the National Academy of Sciences* 115.52 (2018), E12201–E12208.
- [113] Adam T VanWart, John Eargle, Zaida Luthey-Schulten, and Rommie E Amaro. “Exploring residue component contributions to dynamical network models of allostery”. In: *Journal of chemical theory and computation* 8.8 (2012), pp. 2949–2961.
- [114] Amit Ghosh and Saraswathi Vishveshwara. “A study of communication pathways in methionyl-tRNA synthetase by molecular dynamics simulations and structure network analysis”. In: *Proceedings of the National Academy of Sciences* 104.40 (2007), pp. 15711–15716.
- [115] David M Leitner and Takahisa Yamato. “Recent developments in the computational study of protein structural and vibrational energy dynamics”. In: *Biophysical reviews* 12.2 (2020), pp. 317–322.
- [116] Kunitaka Ota and Takahisa Yamato. “Energy exchange network model demonstrates protein allosteric transition: an application to an oxygen sensor protein”. In: *The Journal of Physical Chemistry B* 123.4 (2019), pp. 768–775.
- [117] Yunxiang Sun and Dengming Ming. “Energetic frustrations in protein folding at residue resolution: A homologous simulation study of Im9 proteins”. In: *PLoS One* 9.1 (2014), e87719.
- [118] Rebecca N D’Amico, Alec M Murray, and David D Boehr. “Driving Protein Conformational Cycles in Physiology and Disease: “Frustrated” Amino Acid Interaction Networks Define Dynamic Energy Landscapes: Amino Acid Interaction Networks Change Progressively Along Alpha Tryptophan Synthase’s Catalytic Cycle”. In: *BioEssays* 42.9 (2020), p. 2000092.
- [119] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, Michael P Eastwood, Ron O Dror, and David E Shaw. “Systematic validation of protein force fields against experimental data”. In: *PloS one* 7.2 (2012), e32131.
- [120] Tod D Romo and Alan Grossfield. “Validating and improving elastic network models with molecular dynamics simulations”. In: *Proteins: Structure, Function, and Bioinformatics* 79.1 (2011), pp. 23–34.
- [121] AJ Rader, Chakra Chennubhotla, Lee-Wei Yang, Ivet Bahar, and Q Cui. “The Gaussian network model: Theory and applications”. In: *Normal mode analysis: Theory and applications to biological and chemical systems* 9 (2006), pp. 41–64.
- [122] Yin hao Wu, Xianzhang Yuan, Xia Gao, Haiping Fang, and Jian Zi. “Universal behavior of localization of residue fluctuations in globular proteins”. In: *Physical Review E* 67.4 (2003), p. 041909.
- [123] Weitao Sun. “The relationship between low-frequency motions and community structure of residue network in protein molecules”. In: *Journal of Computational Biology* 25.1 (2018), pp. 103–113.
- [124] She Zhang, Hongchun Li, James M Krieger, and Ivet Bahar. “Shared signature dynamics tempered by local fluctuations enables fold adaptability and specificity”. In: *Molecular biology and evolution* 36.9 (2019), pp. 2053–2068.
- [125] Luca Ponzoni, She Zhang, Mary Hongying Cheng, and Ivet Bahar. “Shared dynamics of LeuT superfamily members and allosteric differentiation by structural irregularities and multimerization”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1749 (2018), p. 20170177.

- [126] Sebastian Buchenberg, Florian Sittel, and Gerhard Stock. “Time-resolved observation of protein allosteric communication”. In: *Proceedings of the National Academy of Sciences* 114.33 (2017), E6804–E6811.
- [127] Kannan Sankar, Sambit K Mishra, and Robert L Jernigan. “Comparisons of protein dynamics from experimental structure ensembles, molecular dynamics ensembles, and coarse-grained elastic network models”. In: *The Journal of Physical Chemistry B* 122.21 (2018), pp. 5409–5417.
- [128] Thomas Elmqvist, Erik Andersson, Niki Frantzeskaki, Timon McPhearson, Per Olsson, Owen Gaffney, Kazuhiko Takeuchi, and Carl Folke. “Sustainability and resilience for transformation in the urban century”. In: *Nature Sustainability* 2.4 (2019), pp. 267–273.
- [129] Matthias Ruth and Dana Coelho. “Understanding and managing the complexity of urban systems under climate change”. In: *Climate Policy* 7.4 (2007), pp. 317–336.
- [130] Michael Batty. “Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies.” In: (2009).
- [131] Jo Da Silva, Sam Kernaghan, and Andrés Luque. “A systems approach to meeting the challenges of urban climate change”. In: *International Journal of Urban Sustainable Development* 4.2 (2012), pp. 125–145.
- [132] Taylor Anderson and Suzana Dragičević. “Complex spatial networks: Theory and geospatial applications”. In: *Geography Compass* 14.9 (2020), e12502.
- [133] Marc Barthélemy. “Spatial networks”. In: *Physics Reports* 499.1-3 (2011), pp. 1–101.
- [134] Paolo Crucitti, Vito Latora, and Sergio Porta. “Centrality in networks of urban streets”. In: *Chaos: an interdisciplinary journal of nonlinear science* 16.1 (2006), p. 015113.
- [135] Jorge Gil. “Street network analysis “edge effects”: Examining the sensitivity of centrality measures to boundary conditions”. In: *Environment and Planning B: Urban Analytics and City Science* 44.5 (2017), pp. 819–836.
- [136] Nahid Mohajeri and Agust Gudmundsson. “The evolution and complexity of urban street networks”. In: *Geographical Analysis* 46.4 (2014), pp. 345–367.
- [137] Yikang Rui, Yifang Ban, Jiechen Wang, and Jan Haas. “Exploring the patterns and evolution of self-organized urban street networks through modeling”. In: *The European Physical Journal B* 86.3 (2013), p. 74.
- [138] Alessio Cardillo, Salvatore Scellato, Vito Latora, and Sergio Porta. “Structural properties of planar graphs of urban street patterns”. In: *Physical Review E* 73.6 (2006), p. 066107.
- [139] Salvatore Scellato, Alessio Cardillo, Vito Latora, and Sergio Porta. “The backbone of a city”. In: *The European Physical Journal B-Condensed Matter and Complex Systems* 50.1-2 (2006), pp. 221–225.
- [140] Sonic HY Chan, Reik V Donner, and Stefan Lämmer. “Urban road networks—spatial networks with universal geometric features?” In: *The European Physical Journal B* 84.4 (2011), pp. 563–577.
- [141] Sonit Bafna. “Space syntax: A brief introduction to its logic and analytical techniques”. In: *Environment and behavior* 35.1 (2003), pp. 17–29.
- [142] Bill Hillier. “The hidden geometry of deformed grids: or, why space syntax works, when it looks as though it shouldn’t”. In: *Environment and Planning B: planning and Design* 26.2 (1999), pp. 169–191.
- [143] Bin Jiang and Christophe Claramunt. “Topological analysis of urban street networks”. In: *Environment and Planning B: Planning and design* 31.1 (2004), pp. 151–162.

- [144] Silvia Pili, Efstathios Grigoriadis, Margherita Carlucci, Matteo Clemente, and Luca Salvati. “Towards sustainable growth? A multi-criteria assessment of (changing) urban forms”. In: *Ecological Indicators* 76 (2017), pp. 71–80.
- [145] Jingliang Hu, Yuanyuan Wang, Hannes Taubenböck, and Xiao Xiang Zhu. “Land consumption in cities: A comparative study across the globe”. In: *Cities* 113 (2021), p. 103163.
- [146] Apostolos Lagarias and Poulicos Prastacos. “Comparing the urban form of South European cities using fractal dimensions”. In: *Environment and Planning B: Urban Analytics and City Science* 47.7 (2020), pp. 1149–1166.
- [147] Jorge Gil, José Nuno Beirão, Nuno Montenegro, and José Pinto Duarte. “On the discovery of urban typologies: data mining the many dimensions of urban form”. In: *Urban morphology* 16.1 (2012), p. 27.
- [148] Marco Maretti, Barbara Gherri, Anthea Chiovitti, Greta Pitanti, Francesco Scattino, Nicolò Boggio, et al. “Morphology and sustainability in the project of public spaces”. In: *The Journal of Public Space* 5.2 (2020), pp. 23–44.
- [149] Kavan Javanroodi, Mohammadjavad Mahdavejad, and Vahid M Nik. “Impacts of urban morphology on reducing cooling load and increasing ventilation potential in hot-arid climate”. In: *Applied energy* 231 (2018), pp. 714–746.
- [150] Zhiwen Jiang, Haomiao Cheng, Peihao Zhang, and Tianfang Kang. “Influence of urban morphological parameters on the distribution and diffusion of air pollutants: A case study in China”. In: *Journal of Environmental Sciences* 105 (2021), pp. 163–172.
- [151] Asmaa M Hassan, Ashraf A ELMokadem, Naglaa A Megahed, and Osama M Abo Eleinen. “Urban morphology as a passive strategy in promoting outdoor air quality”. In: *Journal of Building Engineering* 29 (2020), p. 101204.
- [152] Joanna Badach, Dimitri Voordeckers, Lucyna Nyka, and Maarten Van Acker. “A framework for Air Quality Management Zones-Useful GIS-based tool for urban planning: Case studies in Antwerp and Gdańsk”. In: *Building and Environment* 174 (2020), p. 106743.
- [153] JM Sobstyl, T Emig, MJ Abdolhosseini Qomi, F-J Ulm, and RJ-M Pellenq. “Role of city texture in urban heat islands at nighttime”. In: *Physical review letters* 120.10 (2018), p. 108701.
- [154] David Schrank, Bill Eisele, Tim Lomax, and Jim Bak. “2015 urban mobility scorecard”. In: (2015).
- [155] Jan Hugo and Chrisna Du Plessis. “A quantitative analysis of interstitial spaces to improve climate change resilience in Southern African cities”. In: *Climate and Development* 12.7 (2020), pp. 591–599.
- [156] Hamil Pearsall. “Staying cool in the compact city: vacant land and urban heating in Philadelphia, Pennsylvania”. In: *Applied geography* 79 (2017), pp. 84–92.
- [157] Haiyan Chen, Beisi Jia, and SSY Lau. “Sustainable urban form for Chinese compact cities: Challenges of a rapid urbanized economy”. In: *Habitat international* 32.1 (2008), pp. 28–40.
- [158] Petter Næss. “Urban form, sustainability and health: the case of greater Oslo”. In: *European Planning Studies* 22.7 (2014), pp. 1524–1543.
- [159] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

- [160] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. “Gephi: an open source software for exploring and manipulating networks”. In: *Third international AAAI conference on weblogs and social media*. 2009.
- [161] Joost Schymkowitz, Jesper Borg, Francois Stricher, Robby Nys, Frederic Rousseau, and Luis Serrano. “The FoldX web server: an online force field”. In: *Nucleic acids research* 33.suppl.2 (2005), W382–W388.
- [162] Geoff Boeing. “OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks”. In: *Computers, Environment and Urban Systems* 65 (2017), pp. 126–139.
- [163] K Jordahl. “GeoPandas: Python tools for geographic data”. In: *URL: <https://github.com/geopandas/geopandas>* (2014).
- [164] Sean Gillies, A Bierbaum, K Lautaportti, and O Tonnhofer. “Shapely: manipulation and analysis of geometric objects”. In: *Available online: github.com/Toblerity/Shapely (accessed on 15 June 2019)* (2007).
- [165] Vandroux B Soares I and Magalon N. *Cohérence des dimensions. Référentiel conception et gestion des espaces publics*. Tech. rep. Grand Lyon, 2008.
- [166] Patrick Taillandier, Benoit Gaudou, Arnaud Grignard, Quang-Nghi Huynh, Nicolas Marilleau, Philippe Caillou, Damien Philippon, and Alexis Drogoul. “Building, composing and experimenting complex spatial models with the GAMA platform”. In: *GeoInformatica* 23.2 (2019), pp. 299–322.
- [167] Stefano Battiston, Michelangelo Puliga, Rahul Kaushik, Paolo Tasca, and Guido Caldarelli. “Debtrank: Too central to fail? financial networks, the fed and systemic risk”. In: *Scientific reports* 2.1 (2012), pp. 1–6.
- [168] K Iwashita and Masaya Oda. *Mechanics of granular materials: an introduction*. CRC press, 1999.
- [169] TS Majmudar, M Sperl, Stefan Luding, and Robert P Behringer. “Jamming transition in granular systems”. In: *Physical review letters* 98.5 (2007), p. 058001.
- [170] Igor S Aranson and Lev S Tsimring. “Patterns and collective behavior in granular media: Theoretical concepts”. In: *Reviews of modern physics* 78.2 (2006), p. 641.
- [171] NH Minh and YP Cheng. “A DEM investigation of the effect of particle-size distribution on one-dimensional compression”. In: *Géotechnique* 63.1 (2013), p. 44.
- [172] Dengming Wang and Youhe Zhou. “Statistics of contact force network in dense granular matter”. In: *Particuology* 8.2 (2010), pp. 133–140.
- [173] NH Minh and YP Cheng. “On the contact force distributions of granular mixtures under 1D-compression”. In: *Granular Matter* 18.2 (2016), p. 18.
- [174] Daniel Howell, RP Behringer, and Christian Veje. “Stress fluctuations in a 2D granular Couette experiment: a continuous transition”. In: *Physical Review Letters* 82.26 (1999), p. 5241.
- [175] Trushant S Majmudar and Robert P Behringer. “Contact force measurements and stress-induced anisotropy in granular materials”. In: *Nature* 435.7045 (2005), p. 1079.
- [176] David M Walker, Antoinette Tordesillas, Jie Zhang, Robert P Behringer, Edward Andò, Gioacchino Viggiani, Andrew Druckrey, and Khalid Alshibli. “Structural templates of disordered granular media”. In: *International Journal of Solids and Structures* 54 (2015), pp. 20–30.

- [177] Chad Giusti, Lia Papadopoulos, Eli T Owens, Karen E Daniels, and Danielle S Bassett. “Topological and geometric measurements of force-chain structure”. In: *Physical Review E* 94.3 (2016), p. 032909.
- [178] Karen E Daniels, Jonathan E Kollmer, and James G Puckett. “Photoelastic force measurements in granular materials”. In: *Review of Scientific Instruments* 88.5 (2017), p. 051808.
- [179] Yuming Huang and Karen E Daniels. “Friction and pressure-dependence of force chain communities in granular materials”. In: *Granular Matter* 18.4 (2016), p. 85.
- [180] C-h Liu, Sydney R Nagel, DA Schecter, SN Coppersmith, Satya Majumdar, Onuttom Narayan, and TA Witten. “Force fluctuations in bead packs”. In: *Science* 269.5223 (1995), pp. 513–515.
- [181] JF Peters, M Muthuswamy, J Wibowo, and A Tordesillas. “Characterization of force chains in granular material”. In: *Physical review E* 72.4 (2005), p. 041307.
- [182] Danielle S Bassett, Eli T Owens, Mason A Porter, M Lisa Manning, and Karen E Daniels. “Extraction of force-chain network architecture in granular materials using community detection”. In: *Soft Matter* 11.14 (2015), pp. 2731–2744.
- [183] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (2009), pp. 1422–1423.
- [184] Jihad Zrimi, Alicia Ng Ling, Ernewati Giri-Rachman Arifin, Giovanni Feverati, and Claire Lesieur. “Cholera Toxin B Subunits Assemble into Pentamers - Proposition of a Fly-Casting Mechanism”. In: *PLoS ONE* 5.12 (2010). Ed. by Andreas Hofmann, e15347. DOI: 10.1371/journal.pone.0015347.
- [185] Laura Cendron, Antonio Trovato, Flavio Seno, Claudia Folli, Beatrice Alfieri, Giuseppe Zanotti, and Rodolfo Berni. “Amyloidogenic Potential of Transthyretin Variants: INSIGHTS FROM STRUCTURAL AND COMPUTATIONAL ANALYSES”. In: *Journal of Biological Chemistry* 284.38 (2009), pp. 25832–25841. DOI: 10.1074/jbc.M109.017657.
- [186] Frederick L. Ruberg and John L. Berk. “Transthyretin (TTR) cardiac amyloidosis”. In: *Circulation* 126.10 (2012), pp. 1286–1300.
- [187] Maria João Mascarenhas Saraiva. “Transthyretin mutations in hyperthyroxinemia and amyloid diseases”. In: *Human mutation* 17.6 (2001), pp. 493–503.
- [188] Anvesh K. R. Dasari, Ivan Hung, Brian Michael, Zhehong Gan, Jeffery W. Kelly, Lawreen H. Connors, Robert G. Griffin, and Kwang Hun Lim. “Structural Characterization of Cardiac Ex Vivo Transthyretin Amyloid: Insight into the Transthyretin Misfolding Pathway In Vivo”. In: *Biochemistry* 59.19 (2020), pp. 1800–1803. DOI: 10.1021/acs.biochem.0c00091.
- [189] Anvesh K.R. Dasari, Ivan Hung, Zhehong Gan, and Kwang Hun Lim. “Two distinct aggregation pathways in transthyretin misfolding and amyloid formation”. In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1867.3 (2019), pp. 344–349. DOI: 10.1016/j.bbapap.2018.10.013.
- [190] J. A. Hamilton and M. D. Benson. “Transthyretin: a review from a structural perspective”. In: *Cellular and Molecular Life Sciences CMLS* 58.10 (2001), pp. 1491–1521.

- [191] Matthias Schmidt, Sebastian Wiese, Volkan Adak, Jonas Engler, Shubhangi Agarwal, Günter Fritz, Per Westermark, Martin Zacharias, and Marcus Fändrich. “Cryo-EM structure of a transthyretin-derived amyloid fibril from a patient with hereditary ATTR amyloidosis”. In: *Nature Communications* 10.1 (2019), p. 5008. DOI: 10.1038/s41467-019-13038-z.
- [192] Ai Woon Yee, Matteo Aldeghi, Matthew P. Blakeley, Andreas Ostermann, Philippe J. Mas, Martine Moulin, Daniele de Sanctis, Matthew W. Bowler, Christoph Mueller-Dieckmann, Edward P. Mitchell, Michael Haertlein, Bert L. de Groot, Elisabetta Boeri Erba, and V. Trevor Forsyth. “A molecular mechanism for transthyretin amyloidogenesis”. In: *Nature Communications* 10.1 (2019), p. 925. DOI: 10.1038/s41467-019-08609-z.
- [193] Wenhao Dai, Bing Zhang, Haixia Su, Jian Li, Yao Zhao, Xiong Xie, Zhenming Jin, Fengjiang Liu, Chunpu Li, You Li, Fang Bai, Haofeng Wang, Xi Cheng, Xiaobo Cen, Shulei Hu, Xiuna Yang, Jiang Wang, Xiang Liu, Gengfu Xiao, Hualiang Jiang, Zihe Rao, Lei-Ke Zhang, Yechun Xu, Haitao Yang, and Hong Liu. “Structure-based design of antiviral drug candiyeas targeting the SARS-CoV-2 main protease”. In: *Science* (2020), eabb4489. DOI: 10.1126/science.abb4489.
- [194] H. Yang, M. Yang, Y. Ding, Y. Liu, Z. Lou, Z. Zhou, L. Sun, L. Mo, S. Ye, H. Pang, G. F. Gao, K. Anand, M. Bartlam, R. Hilgenfeld, and Z. Rao. “The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor”. In: *Proceedings of the National Academy of Sciences* 100.23 (2003), pp. 13190–13195. DOI: 10.1073/pnas.1835675100.
- [195] Stephen K. Burley. “How to help the free market fight coronavirus”. In: *Nature* 580.7802 (2020), pp. 167–167. DOI: 10.1038/d41586-020-00888-7.
- [196] Zhenming Jin, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, Xiaofeng Li, Leike Zhang, Chao Peng, Yinkai Duan, Jing Yu, Lin Wang, Kailin Yang, Fengjiang Liu, Rendu Jiang, Xinglou Yang, Tian You, Xiaoce Liu, Xiuna Yang, Fang Bai, Hong Liu, Xiang Liu, Luke W. Guddat, Wenqing Xu, Gengfu Xiao, Chengfeng Qin, Zhengli Shi, Hualiang Jiang, Zihe Rao, and Haitao Yang. “Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors”. In: *Nature* (2020). DOI: 10.1038/s41586-020-2223-y.
- [197] Haitao Yang, Weiqing Xie, Xiaoyu Xue, Kailin Yang, Jing Ma, Wenxue Liang, Qi Zhao, Zhe Zhou, Duanqing Pei, John Ziebuhr, Rolf Hilgenfeld, Kwok Yung Yuen, Luet Wong, Guangxia Gao, Saijuan Chen, Zhu Chen, Dawei Ma, Mark Bartlam, and Zihe Rao. “Design of Wide-Spectrum Inhibitors Targeting Coronavirus Main Proteases”. In: *PLoS Biology* 3.10 (2005), e324. DOI: 10.1371/journal.pbio.0030324.
- [198] Manfred J. Sippl and Markus Wiederstein. “Detection of Spatial Correlations in Protein Structures and Molecular Complexes”. In: *Structure* 20.4 (2012), pp. 718–728. DOI: 10.1016/j.str.2012.01.024.
- [199] Lorena Saelices, Lisa M Johnson, Wilson Y Liang, Michael R Sawaya, Duilio Cascio, Piotr Ruchala, Julian Whitelegge, Lin Jiang, Roland Riek, and David S Eisenberg. “Uncovering the mechanism of aggregation of human transthyretin”. In: *Journal of Biological Chemistry* 290.48 (2015), pp. 28932–28943.
- [200] Maria Paula Sebastião, Maria João Saraiva, and Ana Margarida Damas. “The crystal structure of amyloidogenic Leu55→Pro transthyretin variant reveals a possible pathway for transthyretin polymerization into amyloid fibrils”. In: *Journal of Biological Chemistry* 273.38 (1998), pp. 24715–24722.

- [201] LC Serpell, M Sunde, PE Fraser, PK Luther, EP Morris, O Sangren, E Lundgren, and CCF Blake. *Examination of the structure of the transthyretin amyloid fibril by image reconstruction from electron micrographs*. 1995.
- [202] Hilal A Lashuel, Christine Wurth, Linda Woo, and Jeffery W Kelly. “The most pathogenic transthyretin variant, L55P, forms amyloid fibrils under acidic conditions and protofilaments under physiological conditions”. In: *Biochemistry* 38.41 (1999), pp. 13560–13573.
- [203] Diego U Ferreira, Joseph A Hegler, Elizabeth A Komives, and Peter G Wolynes. “Localizing frustration in native proteins and protein assemblies”. In: *Proceedings of the National Academy of Sciences* 104.50 (2007), pp. 19819–19824.
- [204] Sushant Kumar, Declan Clarke, and Mark Gerstein. “Localized structural frustration for evaluating the impact of sequence variants”. In: *Nucleic acids research* 44.21 (2013), gkw927.
- [205] IL Alves, DR Jacobson, MF Torres, T Coelho, G Holmgren, and MJM Saraiva. “Compound heterozygosity in patients with TTR-related amyloidosis”. In: *Neuromuscular Disorders* 6 (1996), S19.
- [206] P. Hammarstrom. “Trans-Suppression of Misfolding in an Amyloid Disease”. In: *Science* 293.5539 (2001), pp. 2459–2462. DOI: 10.1126/science.1062245.
- [207] Satheesh K Palaninathan, Nilofar N Mohamedmohaideen, William C Snee, Jeffery W Kelly, and James C Sacchettini. “Structural insight into pH-induced conformational changes within the native human transthyretin tetramer”. In: *Journal of molecular biology* 382.5 (2008), pp. 1157–1167.
- [208] Matteo Carli, Giulia Sormani, Alex Rodriguez, and Alessandro Laio. “Candidate Binding Sites for Allosteric Inhibition of the SARS-CoV-2 Main Protease from the Analysis of Large-Scale Molecular Dynamics Simulations”. In: *The Journal of Physical Chemistry Letters* 12 (2020), pp. 65–72.
- [209] Jiahai Shi and Jianxing Song. “The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain”. In: *FEBS Journal* 273.5 (2006), pp. 1035–1045. DOI: 10.1111/j.1742-4658.2006.05130.x.
- [210] C Lesieur and L Vuillon. “From Tilings to Fibers—Bio-mathematical Aspects of Fold Plasticity”. In: *Oligomerization of Chemical and Biological Compounds* (2014), p. 395.
- [211] Friedrich Kremer and Andreas Schönhal. *Broadband dielectric spectroscopy*. Springer Science & Business Media, 2002.
- [212] Beth Goins and Ernesto Freire. “Thermal stability and intersubunit interactions of cholera toxin in solution and in association with its cell-surface receptor ganglioside GM1”. In: *Biochemistry* 27.6 (1988), pp. 2046–2052.
- [213] Yuan-Ling Xia, Jian-Hong Sun, Shi-Meng Ai, Yi Li, Xing Du, Peng Sang, Li-Quan Yang, Yun-Xin Fu, and Shu-Qun Liu. “Insights into the role of electrostatics in temperature adaptation: a comparative study of psychrophilic, mesophilic, and thermophilic subtilisin-like serine proteases”. In: *RSC advances* 8.52 (2018), pp. 29698–29713.
- [214] Marc JS De Wolf, Guido AF Van Dessel, Albert R Lagrou, Herwig JJ Hilderson, and Wilfried SH Dierick. “pH-induced transitions in cholera toxin conformation: a fluorescence study”. In: *Biochemistry* 26.13 (1987), pp. 3799–3806.
- [215] Claire Lesieur, Matthew J Cliff, Rachel Carter, Roger FL James, Anthony R Clarke, and Timothy R Hirst. “A kinetic model of intermediate formation during assembly of cholera toxin B-subunit pentamers”. In: *Journal of Biological Chemistry* 277.19 (2002), pp. 16697–16704.

- [216] Lloyd W Ruddock, Jeremy JF Coen, Caroline Cheesman, Robert B Freedman, and Timothy R Hirst. “Assembly of the B subunit pentamer of Escherichia coli heat-labile enterotoxin: Kinetics and molecular basis of rate-limiting steps in vitro”. In: *Journal of Biological Chemistry* 271.32 (1996), pp. 19118–19123.
- [217] Lloyd W Ruddock, Stephen P Ruston, Sharon M Kelly, Nicholas C Price, Robert B Freedman, and Timothy R Hirst. “Kinetics of acid-mediated disassembly of the B subunit pentamer of Escherichia coli heat-labile enterotoxin: molecular basis of pH stability”. In: *Journal of Biological Chemistry* 270.50 (1995), pp. 29953–29958.
- [218] Lloyd W Ruddock, Helen M Webb, Stephen P Ruston, Caroline Cheesman, Robert B Freedman, and Timothy R Hirst. “A pH-dependent conformational change in the B-subunit pentamer of Escherichia coli heat-labile enterotoxin: structural basis and possible functional role for a conserved feature of the AB5 toxin family”. In: *Biochemistry* 35.50 (1996), pp. 16069–16076.
- [219] Christian B Anfinsen. “Principles that govern the folding of protein chains”. In: *Science* 181.4096 (1973), pp. 223–230.
- [220] Turkan Haliloglu and Ivet Bahar. “Adaptability of protein structures to enable functional interactions and evolutionary implications”. In: *Current opinion in structural biology* 35 (2015), pp. 17–23.
- [221] Kabir Husain and Arvind Murugan. “Physical constraints on epistasis”. In: *Molecular Biology and Evolution* 37.10 (2020), pp. 2865–2874.
- [222] Chad M Petit, Jun Zhang, Paul J Sapienza, Ernesto J Fuentes, and Andrew L Lee. “Hidden dynamic allostery in a PDZ domain”. In: *Proceedings of the National Academy of Sciences* 106.43 (2009), pp. 18249–18254.
- [223] Tandac F Guclu, Nazli Kocatug, Ali Rana Atilgan, and Canan Atilgan. “N-Terminus of the Third PDZ Domain of PSD-95 Orchestrates Allosteric Communication for Selective Ligand Binding”. In: *Journal of Chemical Information and Modeling* (2020).
- [224] Tandac F Guclu, Ali Rana Atilgan, and Canan Atilgan. “Dynamic Community Composition Unravels Allosteric Communication in PDZ3”. In: *The Journal of Physical Chemistry B* 125.9 (2021), pp. 2266–2276.
- [225] Victor H Salinas and Rama Ranganathan. “Coevolution-based inference of amino acid interactions underlying protein function”. In: *elife* 7 (2018), e34300.
- [226] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. “Protein sectors: evolutionary units of three-dimensional structure”. In: *Cell* 138.4 (2009), pp. 774–786.
- [227] Ágnes Tóth-Petróczy and Dan S Tawfik. “Slow protein evolutionary rates are dictated by surface–core association”. In: *Proceedings of the National Academy of Sciences* 108.27 (2011), pp. 11151–11156.
- [228] Cansu Dincer, Tugba Kaya, Ozlem Keskin, Attila Gursoy, and Nurcan Tuncbag. “3D spatial organization and network-guided comparison of mutation profiles in Glioblastoma reveals similarities across patients”. In: *PLoS computational biology* 15.9 (2019), e1006789.

Acknowledgements

I thank the École Normale Supérieure de Lyon for financing my PhD through the allocation of a Contrat Doctoral Spécifique pour Normaliens.

I would like to thank the reviewers and the members of the jury of taking the time to read and evaluate this manuscript and participating to my PhD defense.

I am extremely grateful to Claire for everything she has thought me, for the passion and effort she put daily in supervising this work and for encouraging my development as a scientist.

I want to thank Laurent for co-supervising this work, for his trust and for his help.

I thank Matteo Degiacomi for his warm welcome in Durham and hosting me in his team, where I have learn a lot.

I am grateful to Rémy Cazabet for his help in developing the Building Network model and for his advices. I want to thank Anatoli Sergei for introducing me to the BDS and for proof-reading Chapter 10.

Thanks to the members and fellow PhD students of the Ampère laboratory for welcoming me and for the good times spent together. Laëtita, it was a pleasure working with you and sharing the adventure of the doctoral studies.

I want to thank to the fellow PhD students and PostDocs of the IXXI and our neighbors of the CBP, with whom I have spent countless moments of discussion, support, and fun.

Finally, *grazie Mamma e Nonna* for believing in me and for having always supported my passions and choices. And *merci* Shanoor for being there.