# From coarse-grained to atomistic molecular modeling : how structure and dynamics shape intra-molecular communication and functional sites in proteins

Simon Aubailly

Centre Val de Loire UNIVERSITÉ Spirit of the Loire Valley

UNIVERSITÉ D'ORLÉANS

UNIVERSITE D'ORLEANS

*ÉCOLE DOCTORALE SCIENCES BIOLOGIQUES ET CHIMIE DU VIVANT*

Centre de Biophysique Moléculaire : Biophysique théorique et computationnelle

**Thèse** présentée par :

**Simon AUBAILLY**

soutenue le : **27 janvier 2017**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : **Biophysique théorique**

# From coarse-grained to atomistic molecular modeling : how structure and dynamics shape intra-molecular communication and functional sites in proteins

**Thèse dirigée par :**

| | |
|---|---|
| **Francesco PIAZZA** | Professeur des universités, Université d'Orléans |

**RAPPORTEURS :**

| | |
|---|---|
| **Yves-Henri SANEJOUAND** | Docteur, Université de Nantes |
| **Stefano LEPRI** | Docteur, Institute for complex systems (ISC/CNR) de Florence |

**JURY :**

| | |
|---|---|
| **Thierry DUDOK DE WIT** | Professeur des universités, Université d'Orléans |
| **Giuseppe FOFFI** | Professeur des universités, Université Paris-sud |

# Remerciements

Un peu plus de trois années se sont écoulées et ont permis l'élaboration de plusieurs projets scientifiques qui seront developpés dans la suite. Ces projets scientifiques sont assemblés d'une manière unique et dans un seul format, appellé communément *une thèse*, et cette dernière n'aurait pas vu le jour sans la présence de plusieurs personnes.

Je tiens donc d'abord dans un premier temps à remercier Yves-Henri Sanejouand et Stephano Lepri d'avoir accepté d'être les rapporteurs de ce manuscrit malgré le fait que cela impliquait une lecture lors des fêtes (j'en suis d'ailleurs désolé et je ramènerais des chocolats le jour J pour me faire pardonner) et Thierry Dudok de Wit et Giuseppe Foffi d'avoir aimablement accepté de juger ces travaux de recherches. J'espère que la lecture vous plaira à tous et que Thierry verra que j'ai progressé depuis ma dernière année de licence.

Dans un deuxième temps, je remercie l'équipe de Biophysique théorique et computationnelle du Centre de biophysique moléculaire de m'avoir accepté parmi eux et plus principalement Gerald Kneller qui m'a fait découvrir le domaine lors de mon stage de master 2, la biophysique, et Francesco Piazza de m'avoir encadré lors de ces trois dernières années. Francesco, tu m'as fait confiance et m'a laissé une grande part d'autonomie dans mes travaux de recherche ce que j'ai grandement apprécié. De plus, tu m'as permis de voyager : Chamonix, Florence, Aarhus (je remercie au passage Alberto Imparato de m'avoir encadré au Danemark) et cela a grandement contribué aux joies de la thèse. Je sais que l'écriture du manuscrit a été un peu compliqué mais maintenant que cela est fini je suis content et fier de présenter un tel manuscrit. Je ne serais pas "l'apprenti" chercheur que je suis maintenant sans ta contribution.

Les remerciements de l'équipe concerne bien sur également Marta Galanti et Stefano Iubini qui m'ont accompagné, pour Marta une bonne partie du début de thèse et pour Stefano quasiment tout son long. Stefano, tu étais toujours la pour apporter une réponse à mes questions (qui étaient parfois un peu naïves je le reconnais) et j'espère que la suite en tant que chercheur se passera bien pour toi.

Enfin, le projet "cooling" n'aurait pas autant avancé sans le soutien de Christian Kandt que je remercie grandement.

Mes remerciements ne seraient pas complets si j'oubliais mes camarades de café. Eric

Eveno bien sur qui m'a apporté une réponse à chacune des mes questions portant sur la biologie et qui a joué le rôle de conseiller d'orientation. Tu étais toujours de bonne humeur et ça m'a permis de garder le cap. Je n'oublierais jamais ces pauses café. Mais quand on parle de pauses café on ne peut pas ne pas parler de *la bande à Legrand* où la bonne humeur y est présente dès le matin et qui ont été mes deuxièmes mentors concernant cette matière assez expérimentale et "étrange" : la biologie.

La pause café de 10h est riche de monde et de conversations. Une conversation souvent récurrente avait pour sujet la gastronomie et je ne sais pas combien de matins, dès 10h, je rêvais de manger une côte de boeuf ou des noix de st Jacques. En tout cas, cette thèse n'aurait pas été la même sans Amandine, Anthony, Aline, Shalina (malgré qu'elle n'ait pas cuisiné assez de mi-cuits à mon goût) que j'ai pu connaitre en dehors de l'enceinte du laboratoire (Amandine ne t'inquiete pas tu auras un Vouvray dédicacé).

Il est difficile de citer tout les personnes et je m'excuse d'avance pour les personnes non-mentionnées du CBM mais sachez que j'ai passé 3 chouettes années dans ce laboratoire.

Une partie de ma vie ces trois dernières années était consacrée à la recherche mais le soir, quand je sortais du laboratoire, c'était souvent les (ex)-neuvillois et certains bougres que je connais depuis plus de 10 ans (eh oui ! on a bientôt 30 ans les gars !) qui me supportaient. Je commencerais par les fous de la fameuse "Coloc", Magret, Coco, Momo et Jinouf, je me souviendrais toujours des repas du dimanche et surtout du boeuf bourguignon... Il s'ensuit Lolo, Kiki, Biber, Alex et Clémence, tous toujours la pour me motiver au quotidien. Et bien sur, cette liste ne peut pas ne pas mentionner Pablo de la Muerta qui se destine à une grande carrière dans l'ingéni...la recherche pardon !!! Un grand merci à vous tous.

Mes loisirs "extra-scolaires" m'ont permis également de rencontrer une sacrée équipe de gay-lurons ! La DHCCT (Dream Hakka Coin-Coin Team) de Chilleurs basket qui tous les lundis et jeudis me mettent de bonne humeur. Toujours invaincu cette année j'espère que ça restera ainsi !!!

Enfin et pour finir, le soutien familial est l'un des plus important. Que ce soit Papa, Maman, Papy, Mamie, Aurel, Douardo, Alex, Gégé, Anisabelle, JP (prononcé Ji-Pi), Catherine, Jean-Pierre, Laurette, Lucie, et la génération qui suit, vous m'avez tous soutenus et je vous en suis grandement reconnaissant. Les week-end cousinades et couillinades sont toujours un moment super et une source de joie. Pourvu que ça dure !!!

Merci à tous !

# Résumé substantiel

## 1. Introduction

Les protéines sont de grandes macromolécules comprenant des milliers d'atomes et ayant une taille de 2-5 nm typiquement [12]. Ce sont les molécules les plus représentées dans la cellule après l'eau. Bien que l'acide désoxyribonucléique (ADN) est souvent reconnu comme étant l'information génétique des êtres vivants, les protéines sont les réels acteurs dans la cellule. Elles peuvent avoir un role moteur, un role structural, un role de transmission d'information entre l'environnement extracellulaire et la cellule ou même avoir un role catalytique [13–16]. Comprendre les propriétés générales de ces molécules, d'un point de vue structurelle et dynamique, est essentiel et est un sujet central dans la recherche en pharmacologie de nos jours [17, 18].

Les protéines sont crées par un processus complexe se situant dans la cellule. L'ADN est dans un premier temps transcrite en acide ribonucléique (ARN). Il existe trois différents ARN, les ARN messagers (ARNm), les ARN de transfert (ARNt) et les ARN ribosomiques (ARNr). Les ARNm, qui contiennent l'information permettant de synthétiser les protéines, entre en interaction avec une machine moléculaire complexe, le ribosome. Ce dernier *lit* l'ARNm codon par codon (un codon équivaut à trois bases) et assigne les anticodons correspondant appartenant à l'ARNt. À l'opposé de chaque anticodon de l'ARNt se trouve l'acide aminé correspondant à la séquence des trois bases. Cet acide aminé est ensuite ajouté aux premiers grâce à l'ARNr qui catalyse la réaction chimique. L'ARNm est ainsi lue par le ribosome et produit une protéine brique par brique, la brique fondamentale étant l'acide aminé.

Depuis les année 1960, l'apport de la cristallographie aux rayons X permet de connaitre la structure des protéines cristallisées à très bonne résolution. Plus tard, la résonnance magnétique nucléaire (RMN) permet à son tour de détecter la position des atomes des protéines avec ici l'avantage de déterminer la structure de protéines en solution.

L'objectif de ce projet de thèse est de comprendre les determinants structuraux qui régissent les comportements dynamique des protéines en utilisant des approches physiques et mathématiques. Nous avons principalement concentré nos recherches sur deux points différents, la prédiction d'endroits fonctionnels dans les protéines et la communication

intramoléculaire dans les protéines.

# 2. Methodes

Durant ce doctorat, nous avons étudié les protéines de deux façons : nous avons effectué des expériences numériques et observé comment la macromolécule évolue dans le temps et nous avons effectué des calculs basés sur les structures des protéines.

### Dynamique des protéines

Les premières simulations de dynamique moléculaire ont commencé dans les années 1970. La possibilité d'observer des processus moléculaires à haute précision a fait de la dynamique moléculaire un outil puissant. Les protéines étant des molécules ayant un nombre considérable d'atomes, l'utilisation d'algorithmes quantiques pour étudier ces macromolécules n'est pas envisageable du point de vue des ressources de calcul. Par conséquent, la dynamique des protéines est étudiée faisant appel à la mécanique classique. Une fois spécifié l'énergie potentielle du système (qui peut être décrit à l'échelle atomique ou via des modèles gros-grains), les équations de mouvement (Newton) sont intégrées numériquement à partir d'une condition initiale choisie.

À l'échelle atomique, le champ de forces décrivant la protéine prends en compte les interactions liées et non-liées. Les interactions liées sont représentées par les liaisons covalentes entre deux atomes, les angles entre trois atomes ou les angles diédraux entre les deux plans individués par groupes de quatre atomes consécutifs le long de la chaine polypeptidique. Les interactions non-liées prennent en compte les interactions électrostatiques et les liaisons de Van der Waals. Différent champs de forces existent [26–30] mais d'une manière générale, l'énergie potentielle totale du système peut être écrite sous la forme

$$
\begin{aligned}
U = & \sum_{bonds} k_{\alpha\beta}(r_{\alpha\beta} - r^0_{\alpha\beta})^2 \\
& + \sum_{angles} k_{\alpha\beta\gamma}(\psi_{\alpha\beta\gamma} - \psi^0_{\alpha\beta\gamma})^2 \\
& + \sum_{dihedrals} k_{\alpha\beta\gamma\delta}[1 + cos(n_{\alpha\beta\gamma\delta}\theta_{\alpha\beta\gamma\delta} - \delta_{\alpha\beta\gamma\delta})] \\
& + \sum_{pairs} 4\epsilon_{\alpha\beta}\big([\frac{\sigma_{\alpha\beta}}{r_{\alpha\beta}}]^{12} - [\frac{\sigma_{\alpha\beta}}{r_{\alpha\beta}}]^{6}\big) \\
& + \sum_{pairs} \frac{q_\alpha q_\beta}{4\pi\epsilon_0 r_{\alpha\beta}}
\end{aligned}
$$

où les trois premiers termes représentent les interactions liées et les deux derniers les interactions non-liées.

Les protéines peuvent aussi être décrites à une échelle plus *grossière*, où les atomes sont assemblés en *grains*. Durant ce doctorat, nous avons utilisé l'ANM, Anisotropic Network Model [51], où chaque acide-aminé est représenté par une boule centrée en son carbone $\alpha$ ($C_\alpha$) et chaque boule a la même masse étant la masse moyenne des acide-aminés, 110 amu. Les *grains* sont reliés par des ressorts de même raideur $k_2$ et seul les résidus ayant une interdistance inférieure à un rayon dit *cutoff* $R_c$, sont en intéraction. Nous pouvons représenter l'énergie potentielle du système par la formule suivante :

$$U = \frac{1}{2} k_2 \sum_{i<j} C_{ij} (r_{ij} - R_{ij})^2, \tag{1}$$

où $k_2$ est la raideur du ressort et est fixée de telle sorte à représenter les modes de vibrations lents [61] *i.e.* $k_2$=5 kcal mol$^{-1}$ Å$^{-2}$, $r_{ij} = |\mathbf{r}_{ij}|$ et $R_{ij} = |\mathbf{R}_{ij}|$ sont la distance instantanée et la distance d'équilibre des résidus $i$ et $j$ respectivement. $C_{ij}$ défini la matrice d'intéraction et dépend du cutoff $R_c$.

$$\begin{cases} C_{ij} = 1 \text{ pour } R_{ij} \leq R_c \\ C_{ij} = 0 \text{ pour } R_{ij} > R_c \end{cases} \tag{2}$$

Que ce soit en utilisant une description atomique ou une description à l'échelle de l'acide-aminé, une fois l'énergie potentielle du système définie, nous effectuons des expériences numériques où les particules évoluent dans le temps suite à l'application des différentes forces externes et internes. Nous pouvons ainsi suivre des mesures comme l'énergie locale des particules au cours du temps suite à différents stimulis, comme par exemple des perturbations locales du système à $t = 0$.

### Indicateurs structuraux

La dynamique des protéines permet d'étudier notamment comment leur énergie évolue à travers le temps en utilisant les lois de Newtons. Cependant, nous pouvons également inspecter les propriétés topologiques de la protéine et en dégager des indicateurs puissants qui sont reliés aux fonctions biologiques des protéines. Dans cette thèse, nous avons utilisé l'analyse des modes normaux (ou *normal mode analysis*, NMA, en anglais) et des indicateurs qui ont été dérivés de la théorie des réseaux complexes.

L'analyse des mode normaux consiste à étudier les déplacements des particules autour de leurs positions d'équilibre. Les équations du mouvements du système dans son ensemble sont insolvables analytiquement à moins que l'on considère uniquement les faibles

fluctuations des particules autour de leurs positions d'équilibre. En effet, pour de faible fluctuations, l'énergie potentielle du système peut être simplifiée via une expansion en série de Taylor,

$$U = U_{eq} + \sum_{i,\alpha} \Big(\frac{\partial U}{\partial x_{i\alpha}}\Big)\Big|_{\vec{x}^0} (x_{i\alpha} - x^0_{i\alpha}) + \frac{1}{2} \sum_{i,\alpha} \sum_{j,\beta} \Big(\frac{\partial^2 U}{\partial x_{i\alpha} \partial x_{j\beta}}\Big)\Big|_{\vec{x}^0} (x_{i\alpha} - x^0_{i\alpha})(x_{j\beta} - x^0_{j\beta}) + o(|x - x^0|^3) \tag{3}$$

où $U_{eq}$ est l'énergie potentielle du système à l'équilibre, $x_{i\alpha}$ et $x^0_{i\alpha}$ la coordonnée instantanée et à l'équilibre de la particule $i$ dans la direction $\alpha$, respectivement. $U_{eq}$ n'ayant aucun rôle physique, nous pouvons mettre $U_{eq} = 0$. De plus, étant évaluées à l'équilibre, les dérivées de premier ordre sont nulles, par conséquent, pour de faibles déplacements des particules autour de leurs positions d'équilibre, l'énergie potentielle du système prend la forme quadratique suivante

$$U = \frac{1}{2} \sum_{i,\alpha} \sum_{j,\beta} \Big(\frac{\partial^2 U}{\partial x_{i\alpha} \partial x_{j\beta}}\Big)\Big|_{\vec{x}^0} (x_{i\alpha} - x^0_{i\alpha})(x_{j\beta} - x^0_{j\beta}) \tag{4}$$

Puisque l'énergie potentielle du système a une forme quadratique, les équations du mouvements peuvent être résolues analytiquement et nous avons

$$x_{i\alpha}(t) = x^0_{i\alpha} + \frac{1}{\sqrt{m_i}} \sum_{k=1}^{3N} C_k \xi^k_{i\alpha} \cos(\omega_k t + \phi_k) \tag{5}$$

où les vecteurs de dimension $3N$, $\boldsymbol{\xi}^k$, sont les modes normaux. Ce sont les vecteurs propres du Hessien normalisé par la masse

$$\tilde{\mathbb{H}}^{\alpha\beta}_{ij} = \frac{1}{\sqrt{m_i m_j}} \mathbb{H}^{\alpha\beta}_{ij} \tag{6}$$

ayant pour valeur propre $\omega^2_k$.

Durant cette thèse, nous avons mis au point un indicateur, la rigidité locale (*local stiffness*) $\chi_i$ afin de prédire les sites catalytiques dans les enzymes. Ceci représente la contribution de chaque résidu $i$ aux déplacements atomique pour les cinq modes de plus hautes fréquences

$$\chi_i = \sum_{k=3N-4}^{3N} |\boldsymbol{\xi}^k_i|^2. \tag{7}$$

Une protéine peut aussi être représentée par un réseau complexe, où les résidus seraient les noeuds du réseau, reliés par des liens si leur distance respective est inférieure au rayon cutoff $R_c$. En utilisant la représentation abstraite des réseaux complexes, l'information

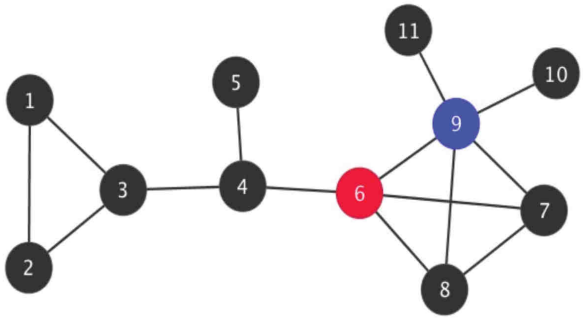



Figure 1.: Réseau de $N = 11$ noeuds. Le noeud ayant la plus haute connectivité est le noeud bleu et celui ayant la plus haute centralité de proximité est le noeud rouge

chimique et physique est perdue et nous nous retrouvons à manipuler un objet mathématique constitué uniquement de noeuds et de lien entre les différents noeuds. Nous pouvons ainsi utiliser toute la panoplie d'indicateurs des réseaux complexes. Lors de ce doctorat, nous avons utilisé la connectivité et la centralité de proximité.

La connectivité d'un résidu $i$ est simplement définie comme étant le nombre de voisins que ce résidu $i$ a. En équation,

$$c_i = \sum_{j \neq i} C_{ij}. \tag{8}$$

où $C_{ij}$ est la matrice de connectivité (equation 2). La centralité de proximité d'un résidu $i$ mesure à quel point ce résidu $i$ est proche des autres. Un résidu ayant une centralité de proximité non négligeable est un résidu qui minimise la distance qu'il a à parcourir pour atteindre un autre résidu $j$. Si $d(i, j)$ est cette distance, la centralité de proximité de $i$ est

$$CC_i = \sum_{j \neq i} \left[ d(i, j) \right]^{-1}. \tag{9}$$

La figure 1 représente un petit réseau où l'on a mis en évidence les noeuds ayant la plus haute connectivité et la plus haute centralité de proximité. Dans le chapitre 3, nous avons utilisés ces deux indicateurs dans un algorithme original conçu pour prédire les sites catalytiques dans les enzymes.

# 3. Cutoff lensing : prediction de sites catalytiques dans les enzymes

Le fait que le nombre de structures et séquences de protéines accessible est de plus en plus important offre la possibilité de tester un grand nombre de modèles et d'algorithmes afin d'identifier des nouveaux sites fonctionnels dans les macromolécules. La plupart des ces méthodes utilisent des informations sur la structure et/ou des critères de conservation dans la séquence [95–101]. Des approches purement séquentielles utilisent l'information phylogénetique et sont basées sur l'hypohtèse que les sites fonctionnels sont conservés à travers l'évolution. Typiquement, de tels algorithmes alignent un grand nombre de séquences et calculent différents type de score de conservation [94, 102–105].

Parmi les algorithmes basés sur des calculs structuraux, une classe prometteuse est celle utilisant des approches gros-grains basées sur un réseau de ressorts (ENM, Elastic Network Model). L'ENM [50] et sa version gros-grains [117, 118] sont des modèles simples et ayant un coût informatique très faible. Souvent, des outils provenant de la théorie des réseaux complexes en combinaison avec des approches ENM ont été utilisé pour identifier les points d'intérêts dans les protéines [83, 84, 119–124]. Dans ces méthodes, une protéine est calquée sur un réseau *via* différentes règles.

Dans ce chapitre, nous avons porté notre attention sur la prédiction de sites catalytiques dans les enzymes grâce à une stratégie originale appliquée sur un réseau élastique. Notre méthode combine trois différents indicateurs, deux provenant de la théorie des graphes et une mesure originale de raideur, la *stiffness*. Notre idée consiste à volontairement augmenter le cutoff $R_c$ dans les réseaux élastiques. Normalement, un cutoff excessif n'est plus physique, puisque la protéine devient globalement connectée quand $R_c$ est de l'ordre de la taille de la protéine. Cependant, nous avons remarqué que les sites fonctionnels des protéines peuvent être découverts en augmentant la connectivité $R_c$. Un résultat intéressant a été de trouver que chaque indicateur a une valeur optimale à laquelle le pouvoir de prédiction de sites catalytiques déjà connus expérimentalement dans les enzymes est maximum. Nous avons nommé notre méthode *cutoff lensing*.

## Indicateurs structuraux

Plusieurs études ont mis en évidence que les sites fonctionnels dans les protéines sont généralement situés dans des régions raides/rigides [129, 138–140]. De manière analogue, il a été montré que les sites fonctionnels ont tendance à se mouvoir indépendamment du reste de la structure, impliquant des vibrations localisées et à hautes fréquences [38, 132, 141, 142]. La rigidité peut être mésurée par différents indicateurs. Nous avons utilisé dans

cette étude une mesure de *stiffness*, $\chi_i$, pour chaque résidu $i$, qui calcule la contribution des cinq modes de plus hautes fréquences aux déplacements atomiques locaux, une mesure de connectivité, $c_i$ dans un réseau complexe et une mesure de centralité de proximité $CC_i$ (voir chapitre 2). Les trois indicateurs ci-dessus peuvent être vus comme différentes mesures de raideur. Alors que $\chi_i$ juge la raideur vibrationnelle des résidus *i.e.* leur capacité à vibrer aux hautes fréquences, $c_i$ et $CC_i$ mesurent la raideur topologique.

De manière générale, les mesures brutes de $\chi_i$, $c_i$ et $CC_i$ sont assez rugueuses et le résultat obtenus pour chaque indicateur est très irrégulier, avec un grand nombre de pics et de creux le long de la séquence. Notre objectif est d'extraire les pics les plus pertinents. Nous avons donc appliqué un filtre passe-haut et gardé seulement les valeurs des indicateurs criblés par ce filtre. Plus précisément :

$$\tilde{\chi_i} = \begin{cases} \frac{\chi_i}{\chi_i^{\max}} \text{ pour } \chi_i > 3\sigma_\chi \\ 0 \text{ sinon} \end{cases} \tag{10}$$

$$\tilde{CC_i} = \begin{cases} \frac{CC_i}{CC_i^{\max}} \text{ pour } CC_i > 3\sigma_{CC} \\ 0 \text{ sinon} \end{cases} \tag{11}$$

$$\tilde{c_i} = \begin{cases} \frac{c_i}{c_i^{\max}} \text{ pour } c_i > 5\sigma_c \\ 0 \text{ sinon} \end{cases} \tag{12}$$

où l'index "max" représente la valeur maximale de la mesure et $\sigma_I$ est l'écart type de l'indicateur $I$. Le cas de $CC$ est plus complexe puisque le résultat brut de cet indicateur est si dégénéré que plusieurs pics sont observés coalescer dans les structures étendues même après l'application du filtre. Nous appliquons donc une procédure de lissage afin que les pics dégénérés se rassemblent pour former un seul pic. Cette procédure est illustrée dans la figure 2.

## Résultats

Notre idée est de scruter les mesures filtrées de stiffness, connectivité et de centralité de proximité et de les relier aux sites fonctionnels. En faisant varier le cutoff $R_c$, on peut voir que les différentes mesures montrent un comportement différent à chaque valeur du cutoff (figure 3). La connectivité et la stiffness suggèrent un meilleur accord entre les sites catalytiques et les pics des indicateurs pour des valeurs intérmédiaire du cutoff. En revanche, la centralité, bien que prédisant relativement bien les sites catalytiques, parait

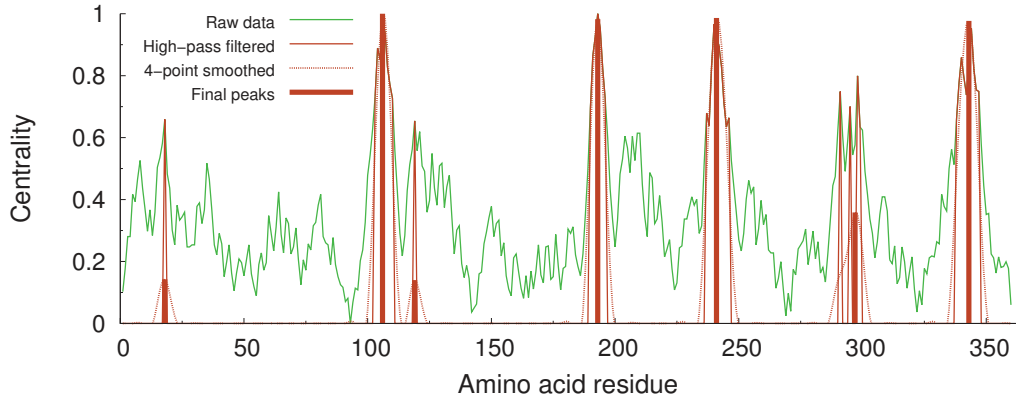



Figure 2.: Illustration des différentes étapes permettant de calculer la centralité de proximité normalisée par sa valeur maximale. Les pics finaux sont potentiellement situés sur les sites fonctionnels. Les calculs ont été réalisés sur la protéine Arginin Glycineaminotransferase (PDB code 1JDW).

assez stable et inchangée selon le cutoff. Il est important de noter que le nombre de pics $N_p$ varie également en fonction du cutoff. Cette information est évidemment à prendre en compte dans un souci de pouvoir de prédiction. Afin d'obtenir un indice statistique du pouvoir de prédiction de nos indicateurs, nous avons analysé 835 structures d'enzyme provenant d'une base de données, la Catalytic Site Atlas (CSA) [144]. Nous avons calculé la fraction de sites catalytiques trouvés à une distance $\Delta n$ près (en unité de résidus) le long de la séquence et calculé le nombre moyen de pics normalisé par la taille de chaque protéine. La figure 4 représente les différents résultats obtenus. Grâce à ces résultats, nous avons pu notamment obtenir une valeur optimale du cutoff pour chaque indicateur correspondant à la meilleure fraction de sites catalytiques prédits avec le plus petit nombre de pics. Ces valeurs sont de $R_c = 20$ Å pour la connectivité, $R_c = 22$ Å pour la stiffness et $R_c = 28$ Å pour la centralité.

Cette étude se termine par la formulation d'un stratégie optimale d'utilisation des indicateurs proposés. Dans un premier temps, utiliser la connectivité qui définit de larges régions dû à un nombre élevé de pics. Ensuite, la mesure de centralité, bien qu'ayant un nombre pics assez important, a un taux de prédiction d'environ 50% et permet de confirmer ou de proposer de nouveaux sites potentiellement fonctionnels. Enfin, la prédiction serait réajustée et peaufinée avec la mesure de stiffness.

## Conclusion

Dans ce chapitre, nous avons proposé un algorithme basé sur la structure de protéines ayant le rôle d'enzyme afin de prédire les sites fonctionnels dans ces dernières. Nous avons dérivé trois différents indicateurs, la connectivité, la centralité de proximité et la stiffness

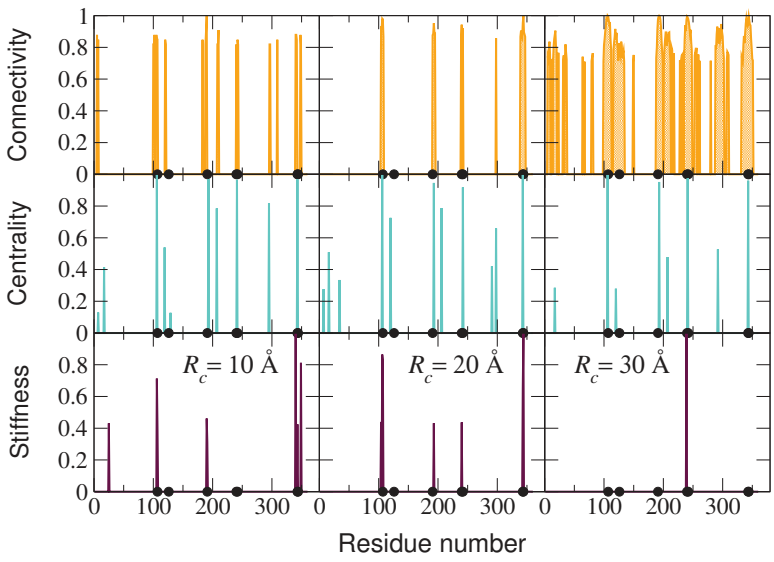



Figure 3.: Mesures filtrées de la connectivité, de la centralité de proximité et de la stiffness pour différentes valeurs du cutoff $R_c$. Ces calculs ont été effectués pour l'Arginin Glycineaminotransferase (PDB code 1JDW) dont les sites catalytiques sont affichés par des cercles noirs.

et optimisé leur seul paramètre variable, le rayon cutoff, afin d'obtenir le maximum de pouvoir de prédiction. Nous avons appelé cette procédure, *cutoff lensing*. Nous avons testé nos différentes mesures sur une base de donnée de 835 enzymes où les sites catalytiques ont été déterminés expérimentalement. Pour chaque indicateur, une valeur optimale du cutoff existe combinant le meilleur ratio fraction de prédiction / nombre de pics. Ces valeurs du cutoff sont : $R_c = 20$ Å pour la connectivité, $R_c = 22$ Å pour la stiffness et $R_c = 28$ Å pour la centralité. Avec ces valeurs, nous avons pu prédire presque 70% des sites catalytiques à une distance près de seulement deux acide-aminés le long de la séquence. Notre méthode a en plus l'avantage d'avoir un coût informatique faible. Nous avons aussi remarqué lors de notre étude que le pouvoir de prédiction était plus important pour des grandes protéines *i.e.* ayant un nombre de résidu supérieur à 540 acide-aminés avec un taux de prédiction de 45% pour la stiffness seule et plus de 60% pour la centralité seule également. Bien sur, notre méthode devrait être jointe à d'autres analyses utilisant la structure et/ou la séquence afin d'obtenir un taux de confiance plus élevé pour prédire des sites fonctionnels dans de nouvelles protéines.

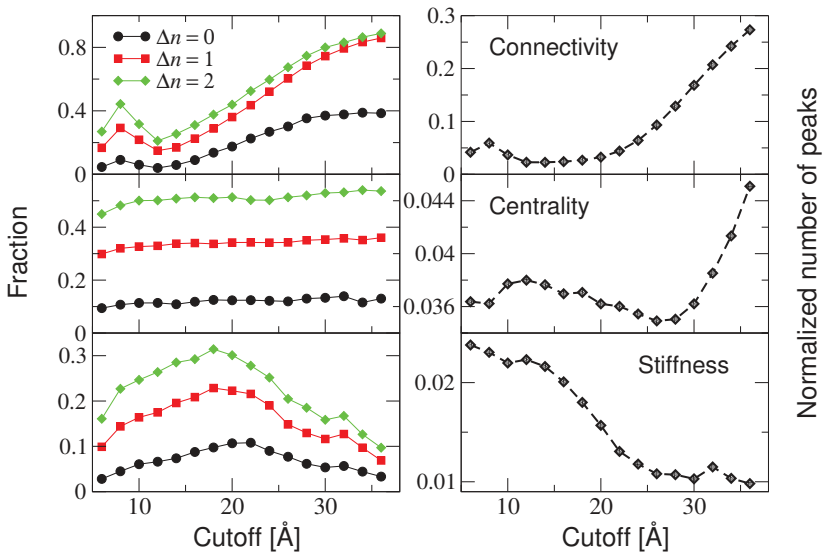



Figure 4.: Partie de gauche : fraction de sites catalytiques prédits à une distance le long de la séquence $\Delta n$ près en fonction du cutoff calculée sur l'ensemble d'enzymes de la base de données CSA. Partie de droite : nombre moyen de pics normalisé (nombre de pics divisé par nombre de résidus) calculé sur la base de donnée en fonction du cutoff.

# 4. Simulation de pompages et modes bimodaux, une nouvelle compréhension de la communication intramoléculaire dans les protéines

Les fonctions des protéines sont souvent liées à la transduction d'un signal le long de leur structure dû à une perturbation mécanique/chimique locale comme la liaison chimique d'un ligand ou l'hydrolyse de l'ATP [149–152]. Par exemple, plusieurs familles de répresseurs ont une affinité chimique pour l'ADN controllée par la liaison d'un ligand à une région de la protéine souvent très éloignée du site de liaison entre l'ADN et la protéine. [153–155]. On parle à ce moment là de *communication* intramoléculaire.

La communication peut être étudiée de différentes façons. Par exemple, l'étude de mouvements corrélés en dynamique moléculaire a permis de connecter des *hotspots* dans les protéines et des chemins structuraux [156–158]. La combinaison entre la théorie des réseaux complexes et la dynamique moléculaire a permis de mieux comprendre les points importants de la communication intramoléculaire. Cependant, bien que la structure des protéines et leurs fonctions sont reliées, nous avons généralement besoin de la dynamique moléculaire pour comprendre le rôle de la topologie en ce qui concerne la prédiction

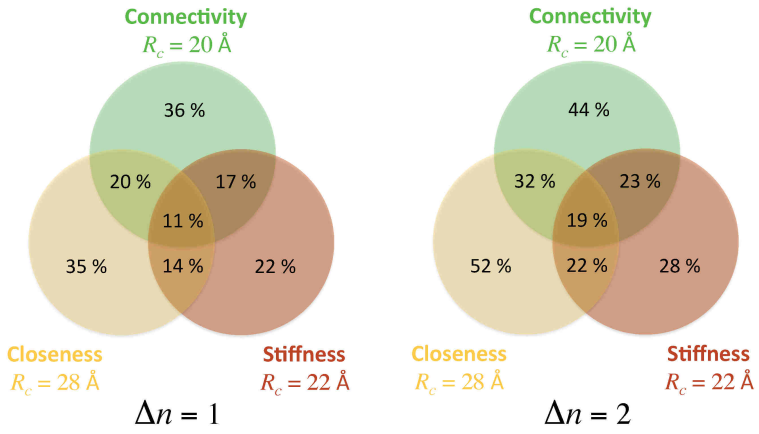



Figure 5.: Fraction de sites catalytiques prédits sur toute la base de données CSA pour chaque valeurs optimale du cutoff pour les différents indicateurs à une précision $\Delta n = 1$ et $\Delta n = 2$.

de sites fonctionnels. Des approches purement structurales peuvent prédire des sites fonctionnels [163–165] mais la reconstruction des cartes de communications entre plusieurs résidus est généralement difficile à obtenir. La prédiction de sites actifs par des approches uniquement structurales est généralement limitée à des méthodes basées sur l'analyse des mode normaux [56, 116, 166].

Par des expériences numériques de *kick* dans un réseaux non linéaire, des chercheurs ont montré que l'énergie est préférentiellement stockée dans des endroits *raides* [38, 57, 60, 167, 168]. Les transferts d'énergie dus à un apport local d'énergie dans les conditions initiales est une approche intéressante pour étudier la communication dans les protéines notamment puisque la communication s'effectue souvent suite à une liaison chimique d'un ligand avec la protéine, créant ainsi une perturbation mécanique dans une région précise de la structure. Il a déjà été démontré que les flux d'énergie peuvent conduire à une compréhension des différents chemins connectant des régions importantes dans les protéines [74, 169–171].

Dans ce chapitre, nous faisons des expériences numériques de pompage, où l'on force un résidu particulier à vibrer à une fréquence fixe et nous observons comment l'énergie se disperse dans la protéine suite à cette perturbation dans la même idée que l'analyse effectuée par Sharp et Skinner [67]. On observe typiquement qu'après une période transitoire, l'énergie se disperse dans la protéine suivant plusieurs règles de selection que nous pouvons relier à la structure.

## Protocole de simulations et illustration sur la Kinésine

Notre idée est donc d'appliquer une force oscillatoire à un résidu dans une direction particulière et à une fréquence précise et d'étudier comment l'énergie se propage suite à

ce stimuli en utilisant le modèle gros-grains avec un cutoff $R_c$ de 10 Å, une constante de raideur $k_2 = 5$ kcal mol$^{-1}$ Å$^{-2}$ et une masse moyenne $M = 110$ Da attribuée à chaque acide-aminé.

Toutes nos simulations ont été faites avec une amplitude de pompage fixe $F_0 = 0.5$ kcal mol$^{-1}$ Å$^{-1}$ et un coefficient de friction fixé $\gamma = 8 \times 10^{-3}$ ps$^{-1}$. Les fréquences et directions de pompages ont été choisies dans le but d'exciter localement un mode normal, soit $\omega_0 = \omega_k$ et $\hat{e}_n = \boldsymbol{\xi}_n^k / ||\boldsymbol{\xi}_n^k||$ où $k$ est le mode excité en question. Puisque le système évolue jusqu'à atteindre un état stationnaire, nous nous intéressons aux énergies moyennes de chaque résidu une fois l'état stationnaire atteint. Typiquement, nous mesurons pour chaque résidu $i$

$$\epsilon_i = \frac{\langle E_i \rangle}{\sum_m \langle E_m \rangle}, \tag{13}$$

où $\langle E_i \rangle$ est l'énergie du résidu $i$ moyennée sur l'état stationnaire.

Dans notre étude, nous travaillons avec les fréquences du spectre linéaire des modes normaux. Pour chaque mode $k$, on définit l'amplitude locale de vibration du résidu $i$ par

$$\chi_i^k = \sum_{\alpha=x,y,z} \left| \boldsymbol{\xi}_{i\alpha}^k \right|^2, \tag{14}$$

et évidemment $\sum_i \chi_i = 1, \forall k$.

Nous avons remarqué que d'un point de vue global, si un résidu à une part importante dans le spectre $\chi_i^k$ pour un mode particulier $k$, si l'on force le dit résidu à vibrer à la fréquence $\omega_k$ en question, l'énergie de la protéine est transférée principalement dans les résidus ayant une contribution importante dans le spectre $\chi^k$. Nous pouvons observer ce phénomène dans la figure 6 où après une inspection des modes normaux, nous avons remarqué que deux acide-aminés de la kinésine impliqués dans des fonctions biologiques [60] sont tous les deux représentés dans le mode $k = 1062$. Ainsi, en appliquant un pompage à la fréquence correspondante, $\omega_k = 92.58$ cm$^{-1}$, sur le résidu MET 96, on observe une redistribution de l'énergie se suivant le même motif que le mode normal correspondant. En particulier, ceci correspond à un transfert d'énergie à l'acide-aminé THR 298, qui serait impliqué dans la "marche" de la kinésine le long de la microtubule.

## Modes bilocalisés et transferts résonnants

Le cas de la kinésine nous a appris que la dispersion d'énergie lors d'une simulation de pompage a tendance à suivre le spectre $\chi_i^k$ (equation 14) si on force un résidu $i$ à vibrer à la fréquence $\omega_k$. Dans un souci d'efficacité en ce qui concerne les transferts d'énergie longue portée, nous avons donc cherché des modes qui étaient fortement localisés sur deux, trois ou quatre acide-aminés en cherchant à analyser un grand nombre de structures différentes.

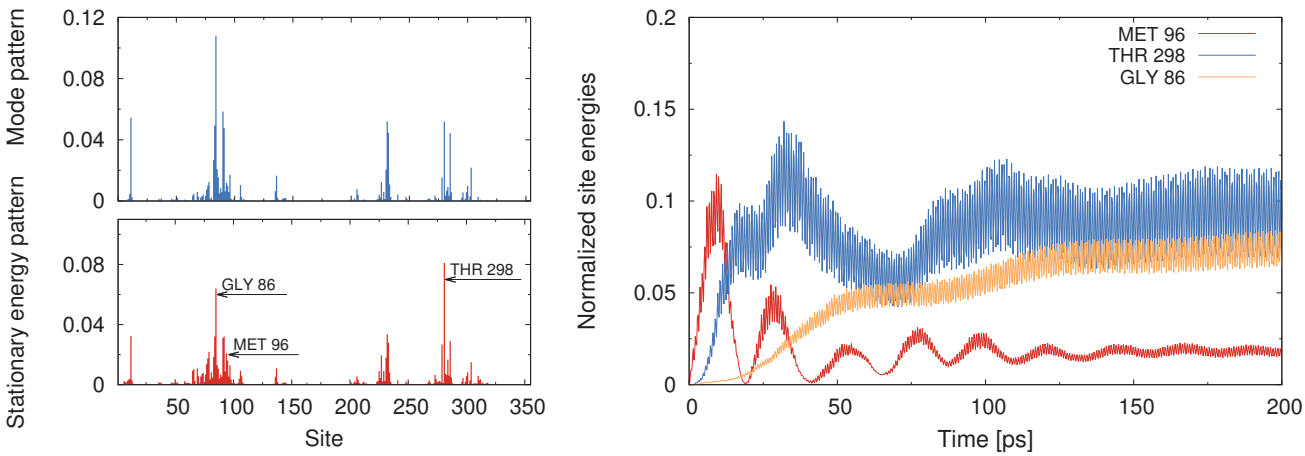



Figure 6.: Analyse d'une simulation de pompage de la kinésine (PDB 3KIN). En haut à gauche, motif $\chi_i^k$ pour $k = 1062$. En bas à gauche, énergie moyenne normalisée dans l'état stationnaire pour chaque résidu lorsque l'on force le résidu MET 96 à la fréquence $\omega_k = 92.58$ cm$^{-1}$ ($k = 1062$). À droite, energie locale des résidus pour la même simulation.

Pour ce faire, nous avons appliqué un filtre passe-haut et sélectionné les modes si le nombre de pics au dessus du filtre était compris entre deux et quatre. Parmi ces modes, nous avons ensuite pris en compte uniquement les deux acides aminés ayant la plus haute amplitude dans le spectre du mode. Dans la suite, on note par $i_1$ le pic ayant la plus haute amplitude dans le spectre $\chi_i^k$ du mode $k$ et $i_2$ le pic ayant la deuxième plus haute amplitude. Nous avons donc pour chaque protéine un nombre de modes $k$ localisés connectant deux résidus $i_1$ et $i_2$ et qui pourraient potentiellement être à l'origine d'un transport d'énergie important entre ces deux derniers résidus.

### Longues distances entre les résidus dans les modes de basses fréquences

Les modes de basse fréquence représentent généralement des mouvements collectifs, alors que les modes de haute fréquence sont plus localisés [56]. On pourrait naïvement s'attendre à ce que les modes potentiellement intéressants appartiennent à la population des modes de haute fréquence puisque le filtre utilisé sélectionne les modes par définition localisés. Cependant, nous observons un comportement plus subtil. Nos modes ont la particularité de connecter deux résidus. Lorsque nous avons calculé, sur notre base de données de structures, les distances connectant les deux acide-aminés ayant la plus haute contribution dans le déplacement des résidus des modes filtrés, nous remarquons que ces distances ont un spectre large de distribution, jusqu'à 180 Å. Nous avons groupé les modes par la distance entre les deux résidus $i_1$ et $i_2$ par tranche de 10 Å et laissé tous les modes connectant deux résidus à une distance supérieure à 50 Å dans la même tranche. La figure

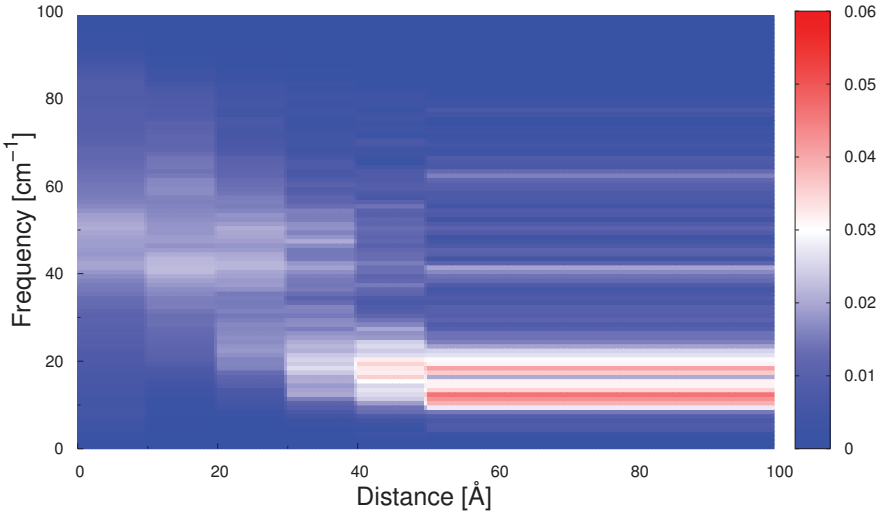



Figure 7.: Distribution des fréquences des modes filtrés en fonction de la distance des deux résidus ayant la plus grande contribution dans leurs patterns de déplacement. Cette analyse a été faite sur 1711 protéines.

7 représente leurs distributions en fonction de leurs fréquences. Il est clair que les modes connectant deux résidus proches s'étalent sur tout le spectre. Cependant, plus la distance entre les deux résidus augmente, plus la fréquence des modes pointe entre 10 et 20 cm$^{-1}$.

### Simulations de pompage

Afin de verifier l'abilité de transfert d'énergie associée à ces modes, nous avons effectué des simulations de pompage sur la même base de données. Pour chaque mode $k$ selectionné comme précedemment pour chaque protéine, nous avons choisi celui qui connecte les résidus $i_1$ et $i_2$ ayant la distance la plus grande. Nous avons procédé à deux simulations de pompage, une où l'on force le résidu $i_1$ à vibrer à la fréquence $\omega_k$ et une autre où l'on force le résidu $i_2$ à vibrer à la même fréquence. Pour chaque simulation, nous avons calculé l'énergie de chaque résidu dans l'état stationnaire (équation 13). La figure 8 représente la distribution d'énergie des résidus $i_1$, $i_2$ et de tous les autres. Si l'on force le résidu $i_1$ (*pump*), le résidu cible (*target*) est $i_2$ et réciproquement. On remarque que nos hypothèses initiales sont validées. En effet, la distribution d'énergie des résidus forcés pointe des énergies plus élevées que les autres résidus et la distribution d'énergie des résidus cibles est clairement séparée des autres. Étant donné que l'énergie est injectée dans la protéine au niveau des résidus forcés, cela n'est pas surprenant que les résidus forcés ont la majorité d'énergie totale des protéines. De plus, les résidus $i_1$ ayant une plus forte amplitude que les résidus $i_2$ dans le spectre $\chi_i^k$, l'hypothèse que l'énergie des

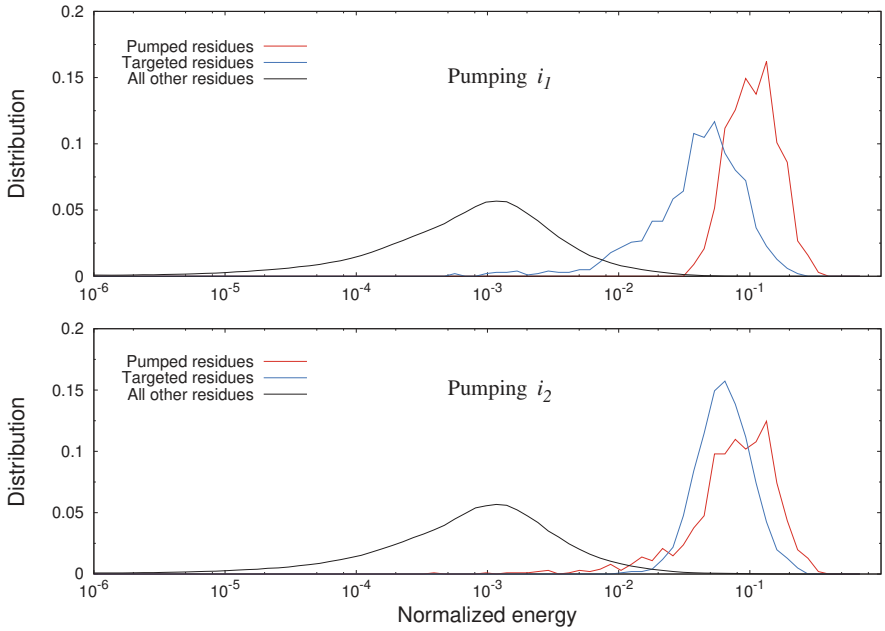


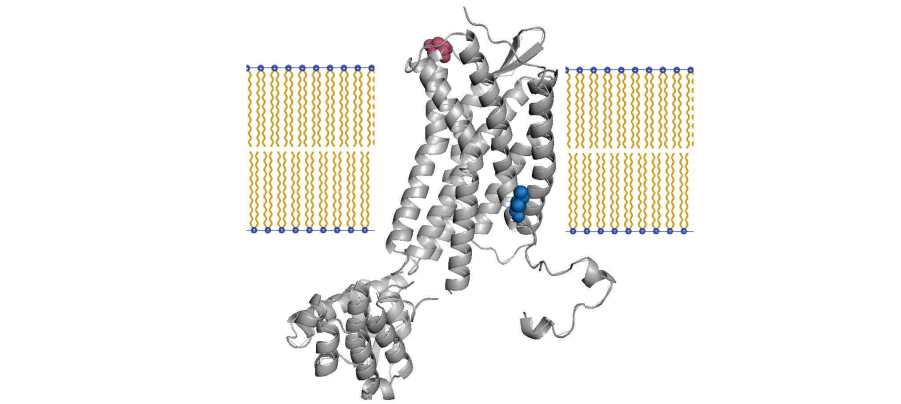

Figure 8.: Histogrammes des énergies moyennes des résidus dans l'état stationnaire normalisées. Rouge, résidus forcés, bleu, résidus ciblés, noir, tous les autres. En haut, résultats lorsque l'on force les résidus du type $i_1$. En bas, résultats lorsque l'on force les résidus du type $i_2$.

résidus dans l'état stationnaire suit le comportement *statique* $\chi_i^k$ est confirmé puisque les transferts de la sorte $i_2 \rightarrow i_1$ ont l'air plus efficace que les tranferts de la sorte $i_1 \rightarrow i_2$. Par conséquent, une plus grande contribution dans les déplacements du mode implique une capacité plus importante d'accumulation d'énergie.

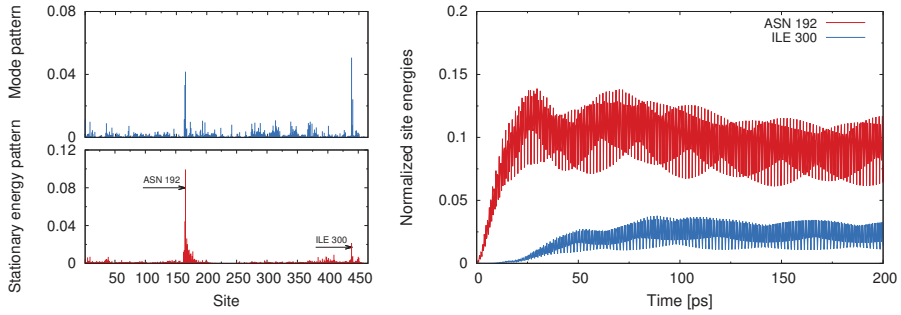
## Application aux transports transmembranaire et conclusion

Puisque les simulations de pompages ont permis d'observer des transferts d'énergie sur de grandes distances, nous avons voulu appliquer ce protocole de simulation sur des Recepteurs Couplés à la Protéine G (RCPG) et ainsi proposer une piste dans la direction d'expliquer le mécanisme de transduction allostérique chez les RCPG qui demeure toujours inexpliqué. Une étude a été faite sur 15 RCPG et nous avons remarqué que des transferts d'énergie entre les parties intracellulaire et extracelluaire ont pu être observés dans quatorze RCPG sur quinze. La figure 9 montre un transfert typique observé entre la partie extracellulaire et la partie intracellulaire de la protéine CXCR4 (récepteur de la chemiokine).

Pour conclure, une analyse des modes normaux permet de situer potentiellement des résidus allostériques connectant des acides-aminés fonctionnels dans les protéines. Nous avons prouvé par expériences numériques que si deux résidus sont connectés par un mode



(a)




(b)

Figure 9.: (a) Structure du recepteur CXCR4 de la chemiokine (PDB 3ODU) encrée dans la membrane cellulaire. Les sphères rouges représentent l'acide-aminé impliqué dans la liaison du ligand dans la partie extracellulaire (ASN 192), les sphères bleues représentent le résidu dans la cellule où le mode bilocalisé en question est centré (ILE 300). (b) Analyse d'une simulation du même recepteur. En haut à gauche, motif $\chi_i^k$ pour $k = 125$. En bas à gauche, énergie moyenne normalisée dans l'état stationnaire pour chaque résidu lorsque l'on force le résidu ASN 192 à la fréquence $\omega_k = 20.59$ cm$^{-1}$ ($k = 125$). À droite, energie locale des résidus ASN 192 et ILE 300 pour la même simulation.

de vibration bilocalisé, une communication peut s'établir entre ces deux derniers *via* un transfert d'énergie. Nous avons pu observer des transferts d'énergies entre deux acides-aminés ayant une distance de plus de 70 Å.

## 5. Refroidissement de surface, localisation spontannée d'énergie dans les protéines

Un système de plusieurs particules dans un état hors équilibre et couplé à son environnement à travers sa surface relaxera jusqu'à l'équilibre en suivant une hiérarchie complexe de taux de relaxations. L'introduction de la non linéarité donne lieu à des relax-

ations encore plus complexes. Lorsqu'un système non linéaire à plusieurs corps est refroidi à travers sa surface, une localisation spontannée d'énergie est observée de manière générale [181–184]. L'énergie est coincée dans le centre du systeme loin de la surface sous la forme de modes de vibration très localisés, les *discrete breathers* (DB) qui par définition vibrent à des fréquences qui n'appartiennent pas au spectre linéaire de vibration [38, 185, 186].

Il a été postulé que les DBs peuvent jouer un rôle dans les processus biologiques impliquant stockage et transfert d'énergie dans la protéine, comme l'allosterie par exemple [187]. De plus, il a été proposé que les DBs puissent jouer un rôle dans la catalyse enzymatique [132, 188, 189].

Une série récente d'études [38, 57, 60, 132] a examiné une extension du modèle gros-grains ENM, où un terme non-linéaire a été rajouté à l'énergie potentielle du réseau (NNM). Les expériences numériques illustrées dans la référence [38] montrent que refroidir une protéine à travers sa surface en utilisant un solvant implicite cause la formation spontanée de modes de vibration dans le coeur de la protéine. Cette dernière, dû au phénomène de refroidissement, voit sa temperature chuter mis à part certains *points chauds* (hotspots), où une fraction considérable de l'énergie initiale à été séquestrée loin de la surface. De manière intéressante, une étude statistique a montré que dans des situations de ce genre, l'énergie a tendance à se localiser dans des endroits *raides* (stiff) de la protéine.

Dans ce chapitre, nous avons étudié si un scénario similaire est observé lorsque l'on refroidit une protéine décrite à l'echelle atomique avec un champs de force réaliste et immergée dans un solvant explicite. Le champs de force gromos53a6 étant fortement non linéaire, nous nous attendons à observer phénoménologie similaire à celle observée dans le cadre du modèle NNM. Les questions que nous nous posons sont : (1) si une localisation d'énergie se produit ou non et (2) si c'est le cas, est ce que ce processus vise des acides aminés aux propriétés particulières en terme d'emplacement dans la structure de la protéine [38].

## Refroidissement à différentes vitesses

La relaxation de l'énergie totale d'un système de plusieurs corps vers l'équilibre lors d'une simulation de refroidissement de surface peut être vue comme la superposition de la relaxation de chaque mode de vibration. Si $E(t)$ est l'énergie totale du système et $\{\tau_1 < \tau_2 < \tau_3 .... < \tau_n\}$ les taux de relaxation associés aux $n = 3N - 6$ modes normaux ($N$ est

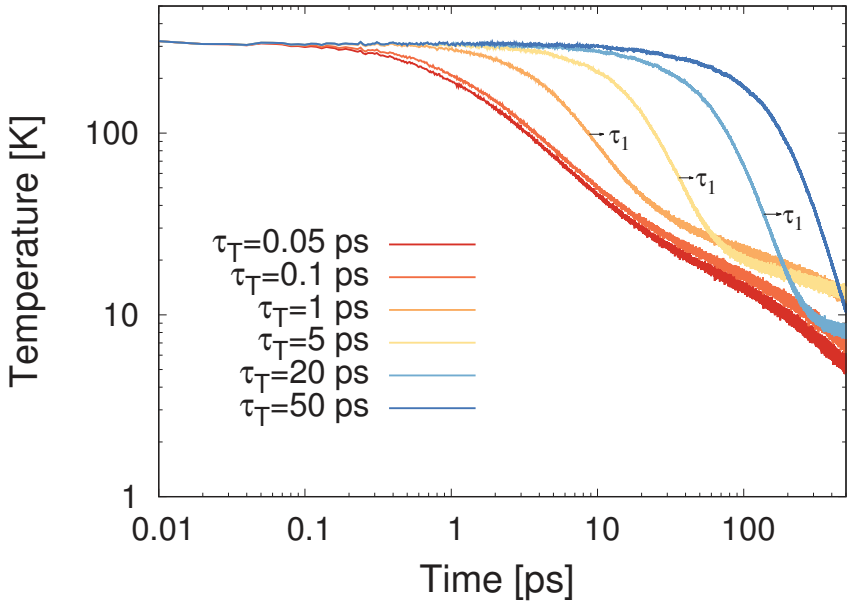



Figure 10.: Température de la citrate synthase (PSB 1IXE) lors de six différentes simulations de cooling avec des taux de refroidissement différents. On indique dans chaque cas l'échelle temporelle $\tau_1$ correspondant au changement de pente entre une exponentielle et une loi de puissance.

le nombre total d'atomes dans la protéine), nous pouvons noter [180]

$$E(t) \approx \begin{cases} c_1 e^{-t/\tau_1} & t \ll \tau_1 \\ \sum_\alpha c_\alpha e^{-t/\tau_\alpha} & t \gg \tau_1 \end{cases} \tag{15}$$

où $\tau_1$ est le taux de relaxation associé au mode normal qui relaxe le plus vite. On notera que les constantes $\tau_\alpha$ peuvent être calculées explicitement en calculant les modes de Langevin de la protéine [190].

### Refroidissement lents, fraction de surface et dimension effective de la citrate synthase

Typiquement, l'équation 15 prédit que l'on devrait observer une changement entre une simple exponentielle et une relaxation plus complexe à une échelle de temps de l'ordre de $\tau_1$. La relaxation plus complexe nait de la superposition des relaxations exponentielles des autres modes et prend une forme intermédiare entre une loi de puissance et une exponentielle étirée [38, 182, 191]. La figure 10 montre ce changement entre une décroissance exponentielle de la temperature vers une relaxation qui tend vers une loi de puissance après un certain temps $\tau_1$ (équation 5.1). Nous avons effectué six simulations refroidissement de surface avec différentes vitesses de refroidissement $\tau_T$ allant de 0.05 ps jusqu'à

50 ps (plus $\tau_T$ est grand, plus le refroidissement est lent). Le changement abrupte décrit ci-dessus est observé pour les refroidissements $\tau_T = 1$ ps, $\tau_T = 5$ ps et $\tau_T = 20$ ps. Il est difficile à observer pour des refroidissements rapides ($\tau_T \leq 0.1$ ps) et completement absent pour un refroidissement très lent ($\tau_T = 50$ ps) ce qui suggère que la protéine relaxe sans atteindre un état stationnaire de non équilibre dans le lapse de temps sur lequel on a intégré la dynamique du système (500 ps).

Nous pouvons remarquer sur la figure 10 que le taux de relaxation associé au mode qui relaxe le plus vite, $\tau_1$, augmente lorsque $\tau_T$ augmente. Nous avons une relation linéaire de la forme

$$\tau_1 = \alpha\tau_T + \beta. \tag{16}$$

Nous savons que pour des refroidissements lents, le taux d'amortissement $\gamma_1 = 1/\tau_1$ est proportionnel à la fraction de surface de la protéine $f$ et inversement proportionnel à la dimension spatiale effective $d_{eff}$ [180]. Ainsi, en calculant la fraction de surface de la protéine, nous pouvons en calculer sa dimension effective. Un calcul effectué à l'aide d'outil de consensus donne $f = 0.26$ ce qui implique $d_{eff} \approx 1.75$. En considérant que la dimension effective des protéines est globalement du même ordre quelque soit cette dernière, ce résultat est en accord avec la référence [180] où les auteurs avaient trouvé une dimension effective pour la myoblogine $d \approx 1.45$.

## Refroidissements avec différentes conditions intiales

Dans la suite, nous voulons localiser et caractériser de manière statistique les résidus *chauds*. Nous avons appliqué un refroidissement rapide de surface $\tau_T = 0.05$ ps et avons effectué neuf simulations de cooling à partir de conditions initiales différentes correspondant au même état thermique $T = 310$ K. Les expériences numériques ont été réalisées avec le champs de force gromos53a6 en double précision.

### Localisation de la température et indicateurs structuraux

Pour chaque simulation, nous avons extrait les vitesses de chaque atome à chaque pas de temps. Pour chaque acide-aminé, nous avons ensuite calculé deux différentes températures, la température de ses atomes le long de la chaine ploypeptidique (bb) et la température de sa chaine latérale (sc) qui sont définies par

$$T_i^{bb} = \frac{1}{3k_B N_{bb_i}} \sum_{j \in \mathrm{bb_i}} \frac{|\mathbf{p}_j|^2}{m_j}, \tag{17}$$

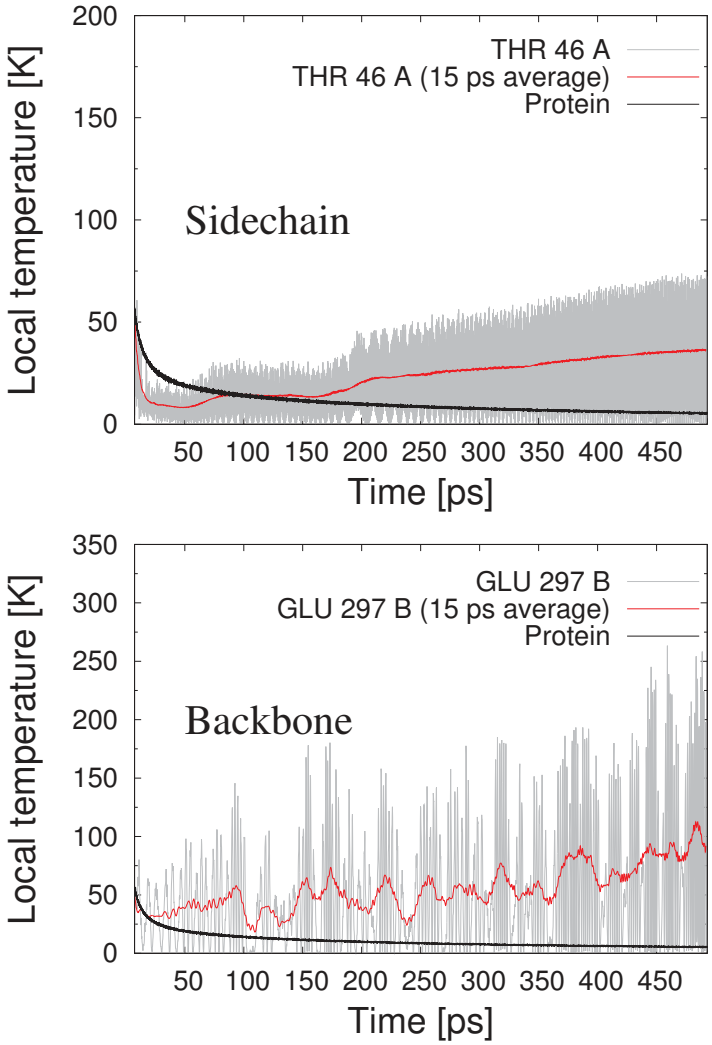



Figure 11.: Température globale de la protéine et température locale de deux résidus particuliers appartenant à la chaine latérale (en haut) et à la chaine principale (en bas) pour une simulation de 500 ps.

et

$$T_i^{sc} = \frac{1}{3k_B N_{sc_i}} \sum_{j \in \mathrm{sc_i}} \frac{|\mathbf{p}_j|^2}{m_j}, \tag{18}$$

où $\mathbf{p}_j$ est l'impulsion de l'atome $j$ et $m_j$ sa masse. La somme couvre tous les atomes le long de la chaine polypeptidique ($N_{bb_i}$) ou ceux de la chaine latérale ($N_{sc_i}$).

Une simulation typique de refroidissement de surface peut être vue sur la figure 11. On y remarque clairement que, bien que la protéine voit sa temperature diminuer au fur et à mesure, des résidus particuliers voient localement leur température augmenter. De tels phénomènes impliquent des transferts d'énergies intra-atomiques dans la protéine. Nous avons décidé, pour les neuf simulations, de sélectionner les dix chaines principales ayant la plus haute température et les dix chaines latérales ayant la plus haute température.

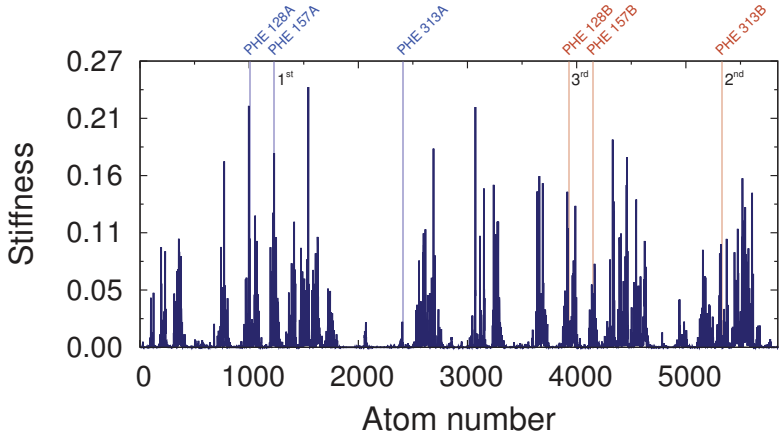



Figure 12.: Mesure de stiffness calculée à l'echelle atomique. Les phenylalanines qui sont représentées le plus parmis les dix chaines latérales les plus chaudes sont indiquées en bleu pour la chaine A et en jaune pour la chaine B.

De cette sélection, nous avons pu remarquer que globalement les chaines carbonnées ont une température plus élevée de 74 $\pm$ 3 comparée à 45 $\pm$ 2, la température moyenne des chaines latérales. De plus, alors que le type d'acide-aminé semble ne pas suivre des patterns spécifiques pour les parties sur la chaine polypeptidique, cela parait totalement différent en ce qui concerne les chaines latérales. Dans ce cas, on retrouve que deux tiers des chaines latérales les plus chaudes appartiennent à des phenylalanines (PHE). Notamment avec PHE 157, PHE 313 et PHE 128 que l'on retrouve (les chaines A et B confondues) parmi les chaines latérales les plus chaudes au moins dans deux tiers des expériences numériques.

De manière intéressante, nous remarquons que ces résidus se positionnent à des endroits *raides*, à savoir ils sont pour la plupart bien représentés dans le spectre de *stiffness* que nous avons calculé à l'échelle atomique (modèle ENM avec cutoff $R_c = 5$ Å [50]). La stiffness est calculée *via* l'équation

$$\chi_i^{aa} = \sum_{k=3N-39}^{3N} |\boldsymbol{\xi}_i^k|^2, \tag{19}$$

où la somme s'effectue sur les 40 modes de plus hautes fréquences (au lieu de 5 à l'échelle gros-grains, chapitre 3). La figure 12 montre que les PHEs ci-dessus correspondent bien à des endroits *raides* dans la citrate synthase ce qui rejoint les résultats à l'échelle gros-grains [38].

## Examination des mouvements et analyse en composantes principales

Nous avons isolé les mouvements des résidus chauds pour chaque simulation. En ce qui concerne les chaines principales, le seul mouvement que nous avons pu observer est associé aux déformations de l'angle impliquant les atomes $C_\alpha$-N-H. Une observation sur dix chaines principales a permis de constater que ces mouvements vibrent à une fréquence de 1439 $cm^{-1} \pm 6$ $cm^{-1}$.

Concernant les chaines latérales, on observe un grand nombre de mouvements différents. Nous nous somme intéressés dans un premier temps aux mouvements des PHEs puisque celles-ci sont très représentées dans les chaines latérales les plus chaudes. Le mouvement impliquant la part d'énergie cinétique la plus importante est un mouvement de déformation de l'anneau aromatique. Cette déformation fait vibrer les atomes de carbones à une fréquence allant de 868 $cm^{-1}$ to 872 $cm^{-1}$. Nous avons effectué une analyse en composantes principales afin d'étudier les mouvements des anneaux aromatiques que l'on retrouve excités suite au processus de cooling. De manière intéressante, l'analyse des phenylalanines lors des simulations de refroidissement donnent différents déplacements atomiques, ceux-ci correspondent à certains modes normaux d'un molécule de benzene isolée (figure 13).

## Conclusion et perspectives

Dans ce chapitre, nous avons effectué des expériences de refroidissement de surface à l'echelle atomique, afin d'approfondir le lien entre la structure des protéines et la localisation spontanée d'énergie dans des systèmes non linéaire. Nous avons pour cela modélisé la citrate synthase évoluant dans le champ de forces gromos53a6 et refroidi les molécules du solvant. Nous avons remarqué que malgré une décroissance assez générale de l'énergie des acide-aminés, certains d'entre eux voient tout de même leur énergie augmenter lors du refroidissement. Ceci est un effet exquisement non-linéaire.

Dans la première partie de cette étude, nous avons effectué six différentes simulations ayant des conditions initiales identiques mais un taux de refroidissment différent ($\tau_T$). Pour trois de ces simulations, nous avons pu observer un clair changement entre une décroissance exponentielle de la temperature de la protéine et une relaxation plus complexe. Ce changement nous a permis de définir la constante de temps typique, celle représentant la relaxation du mode normal qui relaxe le plus rapidement ($\tau_1$). La relation entre $\tau_1$ et $\tau_T$ nous a permis de calculer la dimension effective de la citrate synthase à partir de sa fraction de surface. En calculant cette dernière, nous avons obtenu une dimension effective $d_{eff} \approx 1.75$, en accord avec des études précédentes.

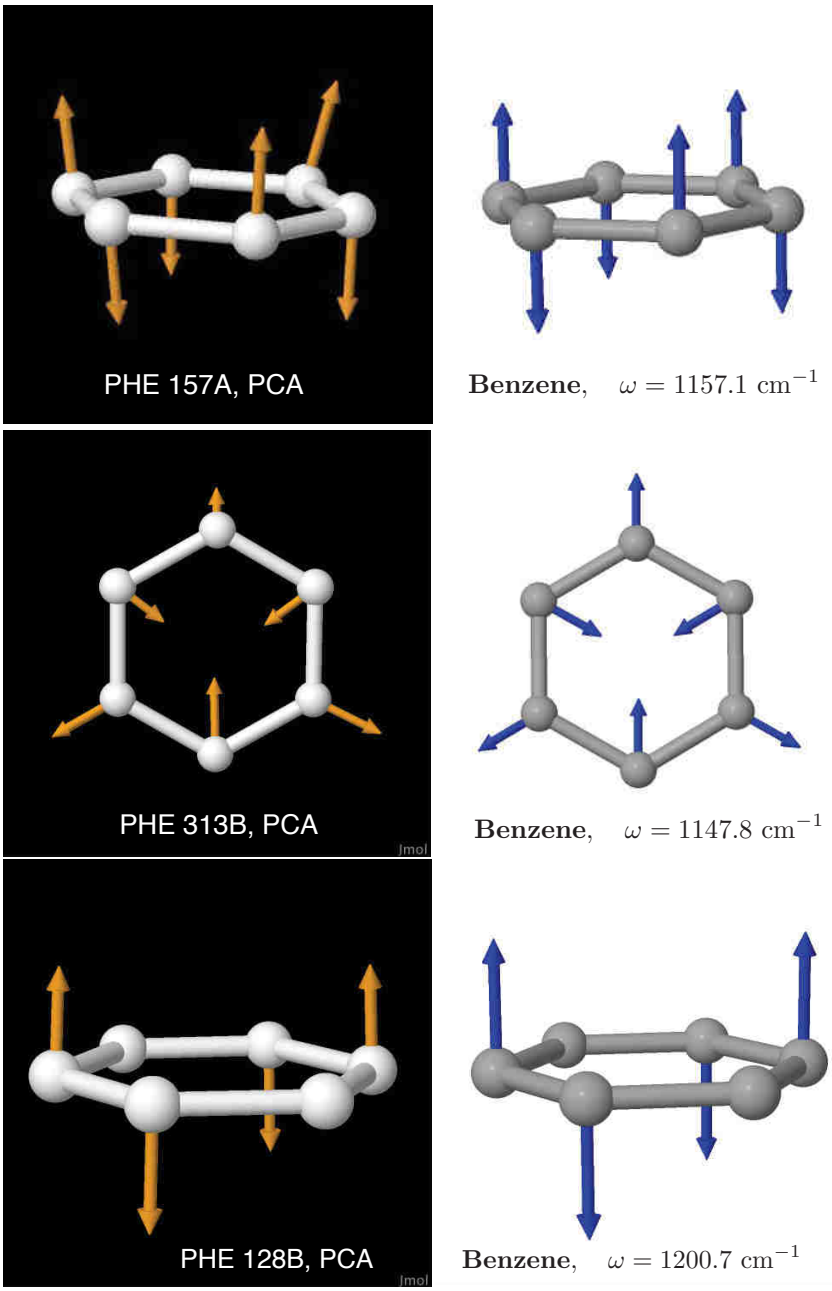



Figure 13.: Analyses en composantes principales de la déformation des anneaux aromatiques des PHEs dans l'état quasi-stationnaire. À gauche, déplacements du premier mode principal obtenue à partir des simulations de cooling, à droite, mode normal correspondant du benzene isolé [10].

Dans la deuxième partie de ce chapitre, nous avons effectué neuf simulations de refroidissement avec le même taux de cooling $\tau_T = 0.05$ ps mais à partir de conditions initiales différentes. Notre objectif étant de faire une étude statistique, nous avons pour chaque simulation extrait les dix segments de backbone les plus chauds ainsi que les dix chaines latérales les plus chaudes. Le premier résultat fut de remarquer l'importance du type d'acide-aminé dans les chaines latérales. En effet, nous avons pu observer une supré-

matie des phenylalanines parmi les chaines latérales les plus chaudes. Cette différentiation très marquée pour les chaines latérales n'a pas été repérée en ce qui concerne les segments le long du backbone. De manière intéressante, nous avons pu relier les phenylalanines qui sont parmi les dix chaines latérales les plus chaudes dans plus de deux tiers des expériences à une mesure de stiffness faite sur un réseau élastique à l'échelle atomique.

Enfin, nous avons isolé les vibrations des résidus chauds. Concernant les parties bacbone, l'énergie est stockée dans les déformation de l'angle $C_\alpha$-N-H, avec des fréquences de $(1439 \pm 6)$ cm$^{-1}$. Pour les PHEs, nous avons remarqué que l'énergie est principalement liée à la déformation de l'anneau aromatique, avec des fréquences variant de 868.1 cm$^{-1}$ to 872.4 cm$^{-1}$. Des analyses en composantes principales sur trois différentes phenylalines comparées à une analyse en modes normaux du benzene nous ont suggéré que l'énergie, lors de simulations de refroidissement, se stocke principalement dans les modes normaux de l'anneau aromatique.

Une étude complémentaire pourrait impliquer le calcul des modes de vibration de Langevin de la citrate synthase, modélisée par un réseau élastique et couplée à sa surface par un réservoir. En analysant les taux de relaxation associés à chaque mode de vibration, nous pourrions nous attendre que les modes qui relaxent les plus lentement soient localisés sur les phenylalanines que l'on retrouve systématiquement parmi les résidus les plus chauds.

# 6. Conclusion générale

Dans cette thèse, nous avons concentré nos recherches sur la relation qui existe entre les structures des protéines et leurs fonctions biologiques.

Dans notre premier projet, nous avons construit des indicateurs dans le but de prédire les sites actifs des enzymes. Nous avons testé notre analyse sur 835 enzymes dont les sites catalytiques sont connus expérimentallement. Nous avons utilisé trois différents indicateurs, deux provenant de la théorie des réseaux complexes, la connectivité et une mesure de centralité, et une mesure de raideur (stiffness) calculée en utilisant l'analyse des modes normaux de vibration. Ces trois indicateurs ne dépendent que d'un seul paramètre, le rayon *cutoff* $R_c$, qui défini comment les résidus intéragissent entre eux. Si deux acide-aminés ont une distance inférieure à ce rayon dans la structure de référence de la protéine, ils sont en interaction. En faisant varier ce rayon, nous avons déterminé une valeur optimale pour chaque indicateur afin d'avoir un maximum de pouvoir de prédiction. En combinant stratégiquement ces trois indicateurs à leur rayon cutoff optimal, nous avons pu prédire presque 70% des sites catalytiques de la base de données utilisée comme benchmarking à une distance près de seulement deux acide-aminés le long de la séquence.

Le deuxième projet sur lequel nous avons travaillé dans cette thèse avait pour but de comprendre la communication intramoléculaire dans les protéines, à la base de phénomènes importants comme l'allosterie par exemple. Pour ce faire, nous avons mis au point des simulations où l'on forçait localement la structure de la protéine à osciller à différentes fréquences et dans différentes directions, tout ceci en imposant une petite friction globale dans le but d'atteindre un état stationnaire asymptotiquement. Nous avons testé dans un premier temps ce protocole de simulation sur le moteur moléculaire kinésine, ce qui nous a permis de comprendre les traits principaux de ce type d'expérience numérique. Nous avons en particulier remarqué que certains modes de vibrations spécifiques ont la particularité d'être bilocalisés malgré leurs fréquences de vibrations assez basses (de l'ordre de 10-20 $cm^{-1}$). Ces modes ont la particularité de connecter principalement deux résidus et d'une manière générale, plus les deux résidus sont éloignés, plus la fréquence du mode qui les connecte est basse.

Nous avons ensuite élaboré une stratégie afin de séparer les modes bilocalisés potentiellement intéressants (fonctionnels) dans une grande base de structures de protéines. Pour chaque protéine, nous avons ensuite isolé, parmi ces modes bilocalisés, le mode qui connecte les deux résidus les plus distancés et effectué sur chaque résidu une simulation de pompage. Ceci nous a permis d'observer que les tranferts d'énergie entre les résidus dépendent de leurs interdistances. Si les deux résidus ont une interdistance inférieure à 20 Å, le transfert d'énergie sera plus efficace si leur contribution au déplacements dans le mode est importante. En revanche, pour des résidus ayant une distance supérieure à 20 Å, une contribution du *bruit* est importante, à savoir la contribution au déplacements des autres résidus dans le mode en question ne doit pas être trop faible pour que le transfert soit efficace. Nous avons pu observer des transferts d'énergie à des distances supérieures à 70 Å du résidu forcé. Au vu de ces observations, nous avons voulu savoir si des transferts d'énergie de cette sorte pourrait être à l'origine de la communication transmembranaire dans les cellules via la machinerie des récepteurs couplés à la protéine G (RCPG). Nous avons ainsi analysé quinze structures de RCPG et pu remarquer qu'un mode bilocalisé pour chaque récepteur existe, connectant la partie extracellulaire et la partie intracellulaire de la protéine. Des simulations de pompages *via* ces modes ont pu par la suite montrer un transfert d'énergie du milieu extracellulaire au milieu intracellulaire.

Dans la partie finale de cette thèse, nous avons présenté les principaux résultats de notre dernier projet, qui consiste à étudier le phénomène de transfert et localisation spontanés d'énergie dans des systèmes de plusieurs corps à une échelle atomique.

Pour ce faire, nous avons élaboré un protocol de relaxation où une protéine décrite

à l'échelle atomique et immergée dans un bain d'eau explicite est refroidie à travers le solvant.

Une étude statistique sur des simulations de refroidissement indépendantes a révélé qu'effectivement, bien que la protéine se refroidisse globalement, certains résidus voient spontanément leur énergie augmenter, que ce soit dans leur chaine principale ou dans leur chaine latérale. En ce qui concerne les parties sur le backbone, l'énergie est stockée dans le mouvement angulaire impliquant les atomes $C_\alpha$-N-H. Ces vibrations ne semblent pas viser spécifiquement des acide-aminés en particulier. Cela est en revanche différent quand on s'intéresse aux chaines latérales les plus chaudes. Dans ce cas, l'énergie a tendance à se stocker dans des phenylalanines, notamment dans les anneaux aromatiques avec des fréquences d'environ 870 $cm^{-1}$. Une analyse via un modèle de type réseau élastique à l'échelle atomique a confirmé que ces résidus sont pointés par une mesure de stiffness locale. Des analyses en composantes principales nous ont aussi suggéré que l'énergie du système lors de simulations de refroidissement avait tendance à se stocker dans des modes normaux des anneaux aromatiques. Enfin, en calculant les *vrais* modes normaux (ceux calculés en diagonalisant le Hessien atomique), on a découvert qu'un mode ayant une une fréquence d'environ 870 $cm^{-1}$ est principalement localisé sur une des PHEs les plus représentées dans nos simulations.

De futures recherches pourraient impliquer le calcul des modes de Langevin à l'échelle atomique et des analyses de type réseaux plus poussées afin de vérifier si les sites *chauds* peuvent être prédits par des indicateurs de type réseau complexe même à l'échelle atomique.

# Contents

# List of Figures

# List of Tables

# 0. Introduction

Proteins are large organic macromolecules comprising thousands of atoms and of typical size of the order of 2-5 nm [12]. In cells, proteins are the macromolecules that are the most represented. In mammalian cells, 18% of the total weight fraction is accounted for by proteins. They are the second most represented molecules after water, that accounts for about 70% of the total weight of a cell.

While Deoxyribonucleic acid (DNA) is generally referred to as the genetic code, containing all the information required to build and maintain a living organism, the proteins are the actual actors that play functional roles in cells [12]. For example, motor proteins transport chromosomes along microtubules during the mitosis [13], structural proteins such as tubulin and actin form the cytoskeleton that allows the cell to maintain its shape [14]. Other proteins, such as the G-protein-coupled receptors (GPCR), lie immersed in the bilayer plasma membrane and play a crucial role in mediating biochemical signaling upon ligand binding in their extra-cellular domains [15]. Enzymes, which lower the activation energy of chemical reactions in the cell making them thousands or even millions of times faster than without catalysis [16], are another important class of proteins.

Understanding the general properties of proteins and more specifically the subtle link between their complex structures, their vibrational and conformational dynamics and their function is a topic of fundamental importance and is nowadays a central theme in drug-discovery-related research [17, 18].

Given the swift pace at which new protein structures are resolved experimentally at atomic detail with a great resolution (now approaching the 0.5 Å), structure-based computational methods are becoming an important asset for computational biology.

The aim of this Ph.D. thesis is to contribute to these problems by providing new insight in the structure-rooted determinants of protein functional dynamics. Using different physical and mathematical approaches, we have focused on two main problems. 1) How one can predict functional sites in proteins where such information is lacking from experiments and 2) what are the main structure-based determinants of intra-molecular energy communication in proteins.

To provide answers to these questions, we have typically analyzed with custom-built

computational tools large databases of structures, with the aim of obtaining statistically sound results concerning the above problems.

In the first chapter of the thesis, we will introduce the basics of the complex three-dimensional arrangements that characterize protein structures, and also describe the way the community goes about organizing the enormous amount of structural information that keeps accumulating from experiments.

Chapter 2 is devoted to a detailed list of the computational tools that we have employed in the different studies that compose this Ph.D. project. The second part of this chapter will more specifically concentrate on the structure-based indicators that we have used to analyze protein function from the sheer point of view of the topology of their three-dimensional scaffolds. These will include normal mode analysis (NMA) and graph-theoretical measures.

Chapter 3, 4 and 5 are devoted to the presentation of our main results. Chapter 3 describes the toolkit that we have developed in order to predict catalytic sites in enzymes. This is based on an hitherto unknown effect that we have termed the *cutoff lensing* effect.

Chapter 4 relates the analyses performed via an original computational scheme, that we have designed to probe energy distribution and communication pathways in proteins. In this chapter we also report the discovery of an hitherto overlooked kind of normal modes, whose vibrational patterns are strongly localized at two distant locations in the protein scaffold, and that can be shown to be able to mediate efficient long-range energy transfer.

In chapter 5, we employ an original technique, surface cooling, to investigate the spontaneous localization of energy induced in a protein structure by a process where the solvent is cooled down progressively to zero temperature. The originality of the research reported in this chapter is the fact that this analysis is performed at atomic detail in explicit solvent. While this project still needs some analyses that we could not include in the present manuscript, the results reported here are intriguing and clearly suggest that surface effects linked to the interaction with solvent molecules are important in order to fully apprehend protein dynamics.

# 1. A brief account of the 3D structure of proteins

Proteins are polypeptides comprising linear chains of amino-acids folded into complex three-dimensional scaffolds, known as the protein *native* structures.

Amino-acids polymerize by forming a peptide bond (carboxyl moiety + amino moiety) between their common structural units. This identifies the linear skeleton of a protein, also referred to as the backbone. Amino-acids differ by their so-called side-chains, different chemical moieties attached to their common backbone segments (figure 1.1).

## 1.1. Protein synthesis

In living organisms, proteins are synthesized in the cell through a complex process involving many intermediates (see figure 1.2). The process starts in the cell nucleus where a given portion of the double-strand DNA is copied onto a single-strand identical molecule of messenger Ribonucleic acid (mRNA). Three different RNAs are needed in order to produce a protein, the messenger RNA (mRNA), the transport RNA (tRNA) and the ribosomal RNA (rRNA), which forms the so-called ribozyme within the ribosome complex, catalyzing peptide bond formation. A protein is created by the traduction of the mRNA into amino-acids. Once a mRNA is created in the cell nucleus, this mRNA moves from the nucleus to the cytoplasm and binds to the ribosome, a complex molecular machine. For each sequence of three mRNA nucleotide bases, a sequence of three complementary bases (a codon) belonging to the tRNA matches and bounds. At the other end of each tRNA, the specific corresponding amino-acid is attached. The mRNA slides through the ribosome, so that for each codon, an amino-acid is carried by the tRNA and added to the end of the growing chain of residues with the help of the rRNA until a codon stop is encountered, which stops the process.

Amino-acid chains can also be assembled in the test tube by chemical synthesis. In the 1960's, the first chemically assembled peptides were produced. In order to produce a chain of amino-acids to form a protein, chemists isolate different fragments of a protein and combine them to form the desired macromolecule. However, due to the complexity

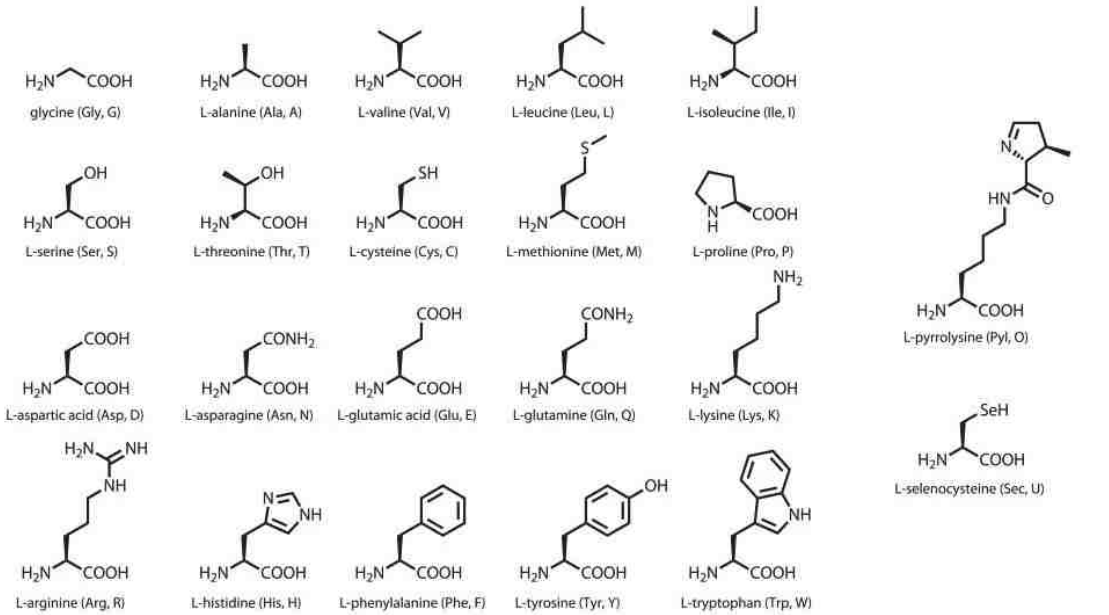



Figure 1.1.: List of the 22 amino-acids that make up all known protein structures. All amino-acids have a common amide and carboxyl group and differ by their side chains, originating from the $\alpha$ carbon [1]. Pyrrolysine and selenocysteine are non-standard amino-acids that are indirectly coded by a STOP codon in DNA. The majority of proteins is formed by the 20 standard amino-acids.

of such synthesis, only small proteins can be created by this process. In 2003, a protein of 166 amino-acids was chemically synthesized [19] but the difficulty to isolate fragments and assemble them properly confine generally the synthesis of smaller proteins to structures comprising 40 to 60 residues at most [20].

## 1.2. Local and global arrangements of protein atoms

The protein synthesis machinery creates a linear polypeptide chain, which subsequently folds to its native state. The three-dimensional structure of proteins is hierarchically arranged in four levels of structural organization (see figure 1.3). The linear sequence of amino-acids is known as the primary structure. A polypeptide chain is thus formed

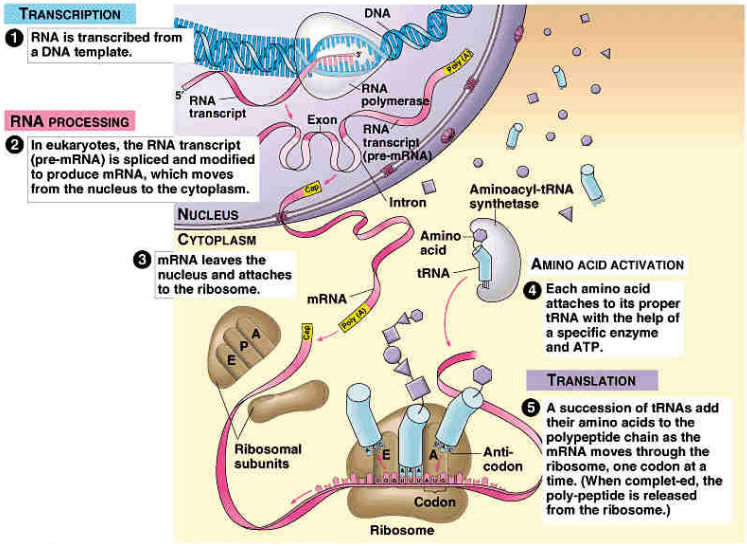



Figure 1.2.: Protein biosynthesis. The different steps are explained from the transcription of the DNA in the cell nucleus to the traduction of mRNA into polypeptide chains [2].

through a succession of peptide bonds, joining consecutive amino-acids through their carboxyl and amino groups (figure 1.4). The nature of the peptide bond is such that it lies in the plane formed by the bond with the preceding and successive $\alpha$-carbons. This peculiar local arrangement of the peptide unit also identifies two dihedral angles, $\psi$ and $\phi$, that turn out to be constrained by the local steric constraints to specific values for all known protein structures (see figure 1.4). The next level of structural organization beyond the primary structure corresponds to groups of residues that organize themselves into local motifs stabilized by hydrogen bonds. These are known as the secondary structure motifs.

While the secondary structure is a local arrangement of residues, the tertiary structure is the first level of global arrangement, as it combines the secondary structure motifs, $\alpha$ helices and $\beta$ sheets, in subunits. The latter are stabilized by different interactions. Disulfide bridges are specific covalent bonds between two cysteines, and are sometimes used in protein chemical synthesis to stabilize and fold the created peptide [21]. Electrostatic and hydrophobic interactions are also major driving forces that govern the stabilization of tertiary structures in proteins. Finally, the different subunits may assemble and form a larger complex comprising two or more monomers. The arrangement of the different subunits is known as the quaternary structure.

With 22 different amino-acids, a protein of $n$ residues can have $22^n$ distinct polypeptide sequences. The smallest known protein, the Trp-cage, is a miniprotein of 20 amino-acids [22]. $22^{20} > 10^{46}$ is far beyond the number of all the known proteins. Among

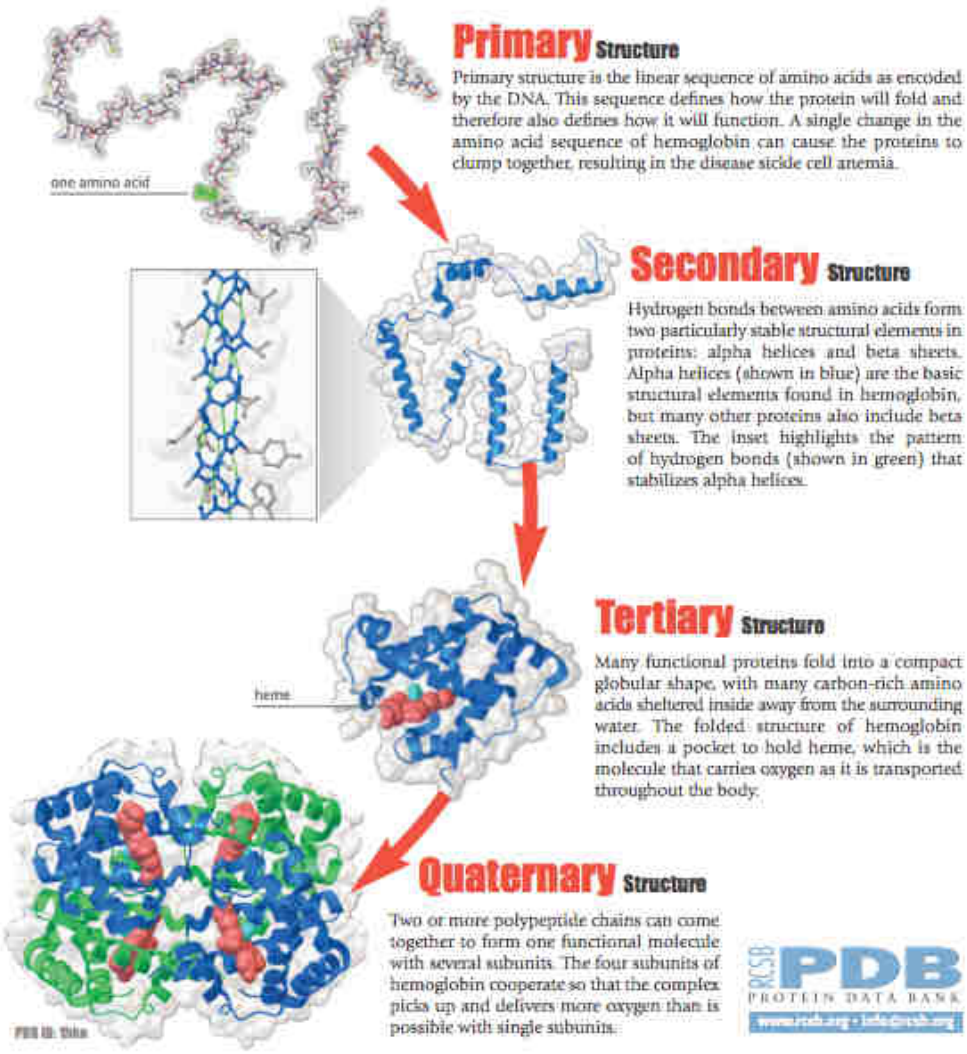


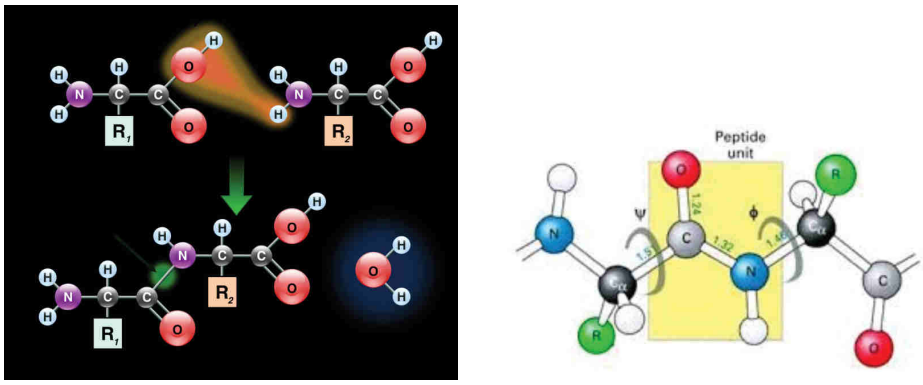

Figure 1.3.: Protein structures, from the amino-acids forming a linear chain to the local (secondary structure) and global arrangements forming the quaternary structure [3].

all these possibilities, evolution has selected a relatively small number of stable complex structural proteins, estimated nowadays to be about 8 x $10^6$ sequences [23].

To conclude, proteins are an ensemble of amino-acids which are assembled locally in $\alpha$ helices and $\beta$ sheets. The amino-acids chain goes up-and-down along the whole structure changing direction with a turn (loop) when it reaches the protein surface.

## 1.3. Experimental characterization of proteins

A fundamental input to computational biology comes from experiments that solve the structure of proteins and protein complexes, *i.e* determine the coordinates of all atoms that compose a stable form of the macromolecule. Of course proteins are dynamical machines,




(a) (b)

Figure 1.4.: (a) The formation of a peptide bond between two amino-acids. (b) Illustration of the two main torsional angles along the backbone of proteins.

that fluctuate about a given equilibrium structure due to the thermal fluctuations. This equilibrium structure is virtually inaccessible, as it depends on the local physico-chemical environment of the cell, whose effect cannot be predicted reliably. Hence, all the experimentally resolved structures have to be considered as the best guess to such unknown, functional equilibrium structures, within the specific limitations of the experimental technique employed to determine the structure. Two main experimental approaches exist to determine atomic positions within a protein, X-ray crystallography and Nuclear Magnetic Resonance (NMR).

### 1.3.1. X-ray crystallography

X-ray crystallography enables to obtain the structure of a protein by shining a X-ray beam on a protein crystal and observing the ensuing diffraction pattern.

Crystal manufacture is a complex and sophisticated procedure, as the quality of the crystal will directly impact the resolution of the protein structure. A protein crystal is a solid where proteins are periodically arranged within a three-dimensional (3D) lattice. High crystal quality involves good purity and a good protein periodicity. The more the crystal will be perfect, the less noise will be present in the diffraction pattern, and the higher the resolution of atomic positions. Each crystal is different and in order to achieve the crystallization, the crystallographer needs to take into account all the parameters such that the temperature, the pH, the ensemble of the solutes *i.e.* ions, water molecules and other small molecules.

Once the crystal has been manufactured, a X-ray beam is shone on the sample. The photons, scattered by the electrons in the crystals are collected to form a characteristic diffraction pattern. Analysis of different diffraction patterns is the key to build the electronic map of the protein of the crystal and consequently allows one to obtain the

position of atoms. However, only the position of heavy atoms is generally determined, hydrogen atoms being directly observable only for high-resolution structures (less than 1 Å). Moreover, the knowledge of protein chemistry and the substantial structural constraints imposed by the complex 3D arrangement of protein scaffolds allows one to determine hydrogen positions from the knowledge of the rest of the structure.

Myoglobin was the first protein to be crystallized and its structure was determined in 1958 [24]. In 2016, the Protein Data Bank (PDB), which gathers the structures of all known different biological molecules such as proteins, DNA, RNA and even bigger complexes had more than 100,000 structures referenced.

### 1.3.2. Nuclear Magnetic Resonance spectroscopy

X-ray crystallography was the Graal in the 1960s to solve protein structures. However, many problems can hinder the production of a good crystal. Nuclear Magnetic Resonance (NMR) is the alternative, as the protein structure can be determined in solution.

The sample to analyze is placed in a homogeneous magnetic field and radio frequency signals are sent through the specimen. A spectrum of absorption peaks enables then the separation of the different atoms of the amino-acids and their environment. This is a local information. Proteins being large molecules and each atom having a different environment, one-dimensional spectra are not enough, as the number of peak overlaps is too large. Two-dimensional, three-dimensional and even four-dimensional correlation spectra are usually needed in order to resolve the structure of a protein.

Analysis of these correlation spectra is the key to obtain protein structures. Although NMR experiments came almost twenty years after X-ray crystallography, the number of protein structures solved by this methods grow larger every year and more than 11,000 protein structures are available in the PDB repository as to the year 2016.

### 1.3.3. Protein Data Bank (PDB)

All protein structures that have been solved experimentally can be found in the Protein Data Bank. When a computational biologist needs the position of protein atoms, she/he can directly go on line [25] and download a PDB file that contains all the positions of the resolved atoms and all the information about the experiments that have been performed in order to determine the structure.

We should note that many redundancies exist in the database. For example, certain proteins have been resolved with different methods, in different states, for instance with a ligand attached to them, or as mutant proteins.

# 2. Methods

The purpose of this chapter is to lay down the basics of the physical models and the mathematical tools that we have used in this thesis to investigate how the structure of proteins influences and modulates their function of complex dynamical machines.

The first section will discuss protein dynamics, that is what models should be used to predict how particles (atoms or residues) evolve in time starting from a given initial condition within a given environment. We will introduce two descriptions of proteins, one at atomic scale and one at the scale of residues. For each one, the system potential energy will be described and we will present the simulation protocols that we have employed in this thesis for different purposes, including more technical issues, such as the integration schemes employed.

Starting from the mere topology, one can define structural indicators and isolate particular residues from the others on the basis of these measurements. Depending on the nature of the specific indicator, this approach may be employed to look for correlations between known functional information coming from experiments and specific structural properties. Furthermore, we will describe Normal Mode Analysis (NMA) of proteins, which is a powerful tool to explore fold-rooted functional motions.

## 2.1. Protein dynamics

Molecular simulations are performed since the advent of electronic computers in the 1970s. The ability to describe molecular processes with high resolution makes molecular simulations a powerful tool.

Although quantum chemistry calculations would be the correct tool to use in principle to examine protein functional dynamics, it is clearly unreasonably costly to do so from a computational point of view for macromolecules comprising thousands of atoms.

The tool of reference in this field is then classical molecular dynamics (also referred as molecular mechanics), where Newton's equations are integrated starting from a suitable (semi-empirical) formulation of the main interactions occurring among the basic constituents of a protein macromolecule. This is referred to as the specific *force field* of the protein. Of course, given an experimentally solved structure, there is little chance that

this also represent a minimum in the potential energy landscape of choice. Typically thus a minimization step constitutes the first stage of any molecular mechanics study, aimed at finding the global minimum of the system potential energy. As we will see, this is not always the case. In the class of models known as *Elastic Network Models* (ENM), the experimental structure is assumed from the beginning to be the reference equilibrium structure, corresponding to the only minimum in the potential energy landscape.

### 2.1.1. All-atom molecular mechanics

In all-atom molecular mechanics, one is interested in atomic displacements. Within the Born-Oppenheimer approximation, one can separate the motion of the electrons and that of the atomic nuclei by assuming the electron dynamics is much faster compared to the nuclear motions. Moreover, nuclei are typically treated as point-like masses with a fixed electric charge. The atoms evolve in a force field defining the different interactions each atom has with the others.

#### All-atom force fields

Different all-atom force fields exist such as CHARMM, GROMOS, AMBER, OPLS or MMFF94 [26–30], which typically comprise the same kind of potential energy terms and only differ with respect to their specific parametrization. In our work, we have used the GROMOS 53a6 force field [31]. This GROMOS force-field is an optimal trade-off between a good description accuracy and an optimization of computational time due to its relatively simple functional form. As all the other force fields, it comprises bonded and non-bonded interactions. Bonded interactions have different contributions, covalent bonds between two atoms, covalent angle-bending interactions referring to three-atom groups and dihedral angles involving four atoms (improper and torsional) (see figure 2.1). Non-bonded interactions are typically of the Van der Waals (VdW) form plus electrostatic (elec) interactions. The total potential $U$ can then be written as

$$U = U_{\text{bonds}} + U_{\text{angles}} + U_{\text{improper}} + U_{\text{torsional}} + U_{\text{VdW}} + U_{\text{elec}}. \tag{2.1}$$

Bonded interactions

$$U_{\text{bonds}} = \frac{1}{4}\sum_{b=1}^{N_b} K_b[r_b^2 - r_{0_b}^2]^2. \tag{2.2}$$

The sum runs over all the $N_b$ covalent bonds. $K_b$ is the bond constant and depends on the atoms involved in the bond, $r_b$ and $r_{0_b}$ are the instantaneous and the equilibrium bond

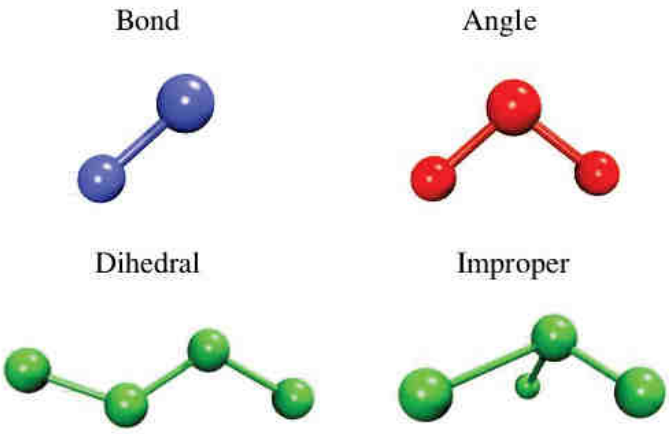



Figure 2.1.: Schematic representation of bonded interactions, two atoms for a bond, three atoms for an angle and four atoms for the torsional dihedral angle and for the improper angle [4].

length, respectively.

$$U_{\text{angles}} = \frac{1}{2}\sum_{a=1}^{N_a} K_a[\cos(\theta_a) - \cos(\theta_{0_a})]^2. \tag{2.3}$$

The sum runs over all the $N_a$ covalent angles. $K_a$ is the angle constant and depends on the atoms involved in the angle, $\theta_a$ and $\theta_{0_a}$ are the instantaneous and the equilibrium angle values respectively.

Typically, $K_b$, $K_a$ and $r_{0_b}$, $\theta_{0_a}$ are derived from spectroscopic data and X-ray diffraction data, respectively, for small molecules [32, 33].

$$U_{\text{improper}} = \frac{1}{2}\sum_{i=1}^{N_i} K_i[\xi_i - \xi_{0_i}]^2. \tag{2.4}$$

The sum runs over all the $N_i$ improper dihedral angles of constant parameter strength $K_i$. $\xi_i$ and $\xi_{0_i}$ are the instantaneous and the equilibrium improper dihedral angles, respectively.

$$U_{\text{torsional}} = \sum_{t=1}^{N_t} K_t[1 + \cos(\delta_t)\cos(m_t\phi_t)]. \tag{2.5}$$

The sum runs over all the $N_t$ torsional dihedral angles. $K_t$ is the constant strength parameter, $\delta_t$ is the phase shift (equal to 0 or $\pi$), $m_t$ is the multiplicity of the angle and

$\phi_t$ is the instantaneous value of the dihedral angle. The parametrization of the dihedral angles have been chosen so that the rotational energy computed by quantum mechanics calculations is reproduced for the torsional angles using the GROMOS force field [31].

Non-bonded interactions

In principle, non-bonded interactions should involve all atom pairs in the protein. However, to reduce computational cost, first and second neighbors are excluded from this contribution because of their high connectivity in bonded interactions. The third neighbors involved in aromatic rings are also typically excluded from long-range terms [31].

The Van der Waals potential between two atoms is highly repulsive when the distance between the two atoms is strongly reduced from the equilibrium and is attractive otherwise. This interaction is typically represented through a Lennard-Jones (LJ) potential, that is,

$$U_{\mathrm{VdW}} = \sum_{i<j} \left[ \frac{K_{12_{ij}}}{r_{ij}^{12}} - \frac{K_{6_{ij}}}{r_{ij}^{6}} \right]. \tag{2.6}$$

$K_{12_{ij}}$ and $K_{6_{ij}}$ are the parameters of the Van der Waals interaction potential and depend on the nature of the atoms $i$ and $j$. Figure 2.2 shows the functional form of the LJ potential for a pair of atoms. The potential displays a minimum with high repulsion at small distances and an attractive force at large separations from the minimum. Finally, the electrostatic interactions depend on the charge $q_i$ and $q_j$ of the two atoms $i$ and $j$, their interdistance $r_{ij}$ and the dielectric constant of the protein/solvent medium, $\epsilon_1$,

$$U_{\mathrm{elec}} = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1 r_{ij}}. \tag{2.7}$$

### Solvent cooling simulation protocol

We have used an atomic-detail representation of protein dynamics for the cooling simulations reported in chapter 5. In the following, we describe in more detail the steps that we followed in a typical cooling run.

Minimization

The first operation is a minimization step, that we have realized through a single Newton algorithm, where atoms move in the direction of the negative gradient until the total force falls below a specified zero threshold. Let $\boldsymbol{\nabla U}$ be the potential energy gradient. If $r_n^i$ is the position of atom $i$ at the $n-th$ integration, in order to minimize the potential energy,

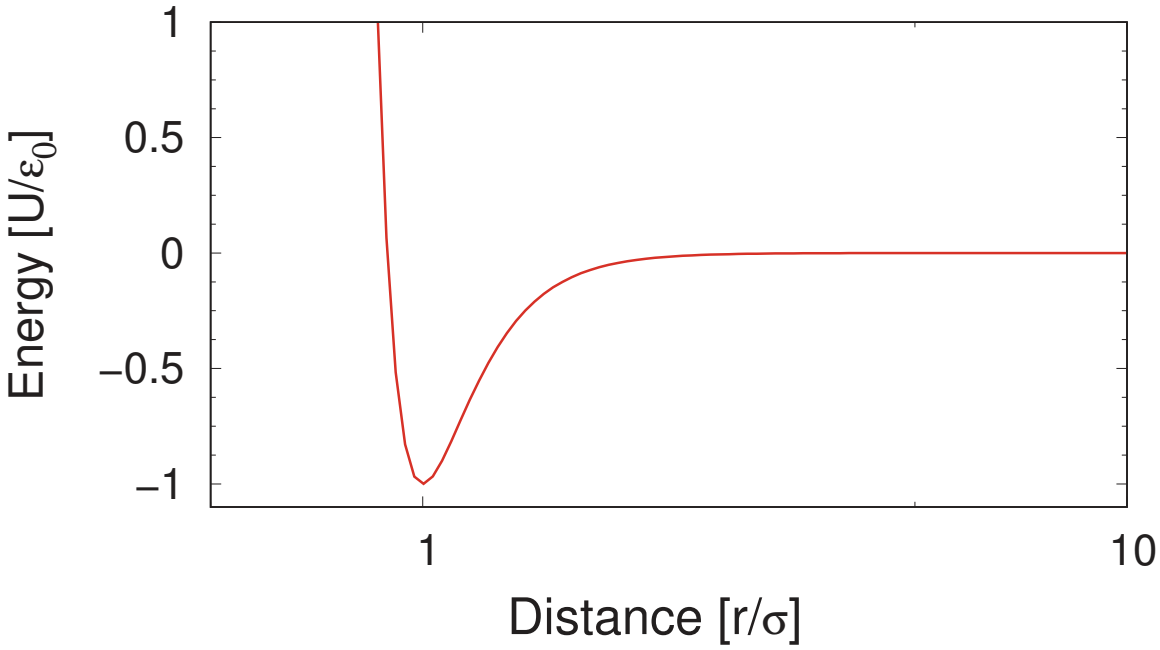



Figure 2.2.: Lennard-Jones potential of the form $U = \epsilon_0\left[\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6}\right]$ as a function of the distance (units of $\sigma$). As observed, the potential is highly repulsive at small distances and attractive at large distances, the minimum being at a separation $r = \sigma$.

the position of atom $i$ at the next step $n+1$ is computed as

$$r^i_{n+1} = r^i_n - \alpha_n \boldsymbol{\nabla U_i}(r_n), \tag{2.8}$$

where $\alpha_n$ denotes the length step. The parameter $\alpha_n$ has to be small enough to lower down the potential energy of the system without crossing the target minimum of the potential energy. However, $\alpha_n$ has to be large enough so that the *minimization* process converges quickly enough. Atoms move from their position in the direction opposite to the energy gradient until they reach the potential energy minimum within the fixed threshold for the total force. We have used a cutoff of 500 kJ $\text{mol}^{-1}$ $\text{nm}^{-1}$.

This procedure goes under the name of steepest descent. Using this algorithm, the potential energy decreases fast but does not converge easily. Typically, an energy minimization starts with a steepest descent algorithm while a more refined procedure is used to improve the convergence further, such as the conjugate gradient algorithm [34].

Equilibration

After the minimization procedure, the protein under study was equilibrated within the

NPT ensembles, at a temperature of 310 K and a pressure of 1 bar. This was accomplished by putting the protein in contact with a thermostat and a barostat, realized through the Berendsen algorithm [35]. Positions and velocities of all atoms are rescaled at each time step, $x \to \mu x$, $v \to \lambda v$, until the protein structure was stabilized at the required temperature and pressure. Rescaling velocities allows one to adjust temperature, while rescaling positions allows to adjust the pressure. To achieve this, the prescription is

$$\lambda = \left[1 + \frac{\Delta t}{\tau_T}\Big(\frac{T_0}{T} - 1\Big)\right]^{1/2}, \tag{2.9}$$

$$\mu = \left[1 + \frac{\Delta t}{\tau_P}\big(P_0 - P\big)\right]^{1/3}, \tag{2.10}$$

where $\Delta t$ is the integration step, and $\tau_T$ and $\tau_P$ the time coupling constants of the thermal bath for the temperature and the pressure, respectively. $P_0$ and $T_0$ are the reference pressure and temperature, respectively, while $P$ and $T$ are the instantaneous values. A direct inspection of equation 2.9 and equation 2.10 shows that the more the thermal bath is coupled to the system, the faster the rescaling ought to be. Other equilibration algorithm exist such as the Nose-Hoover thermostat [36] or the Parrinello and Rahman barostat [37].

Solvent cooling simulations of a protein

In this thesis, we have focused on the relaxation dynamics of a specific protein, citrate synthase, which has been the subject of a similar study performed through a coarse-grained model [38]. The idea is to start from an equilibrated conformer of the protein and switch on a cooling procedure that extracts energy from the protein through the collisions with the solvent molecules.

To achieve this, we have used the Berendsen description with a reference temperature set at $T = 0$ K. By doing this, the solvent cools down until all water molecules are at rest and the protein is indirectly cooled down as well, with the important difference that no direct cooling force is exerted on the atoms in the protein. The results of these experiments will be discussed in chapter 5.

### 2.1.2. Coarse-grained representation of proteins

Although molecular mechanics describes the behavior of the protein at atomic level, one is considerably limited by the high computational costs associated with large structures or long time scales [39]. A way around this limitation is to decrease the number of degrees of freedom through the formulation of specific coarse-grained (CG) models. Depending

on the level of resolution, amino-acids can be replaced by one or more effective beads, that interact through a complex potential with many parameters such as the all-atom force fields or can interact through a simpler potential featuring a reduced number of parameters. One of the first examples of a coarse grained model of proteins was the Gō model [40] introduced with the aim of studying protein folding. Each amino-acid was replaced by a single bead, the beads interacting through non-bonded attractive or repulsive interactions. Since then, many models have been developed with more beads and models with two, four or even six beads per amino-acid have been introduced in the literature [41–46].

While many models use an all-atom-like force field with an interaction potential depending of bonded and non-bonded terms [47–49], the class of elastic network models (ENM), first introduced by Tirion [50], is much simpler, as it connects particles by harmonic springs with one and the same stiffness. In our work, we have used a coarse-grained version of the ENM, known as the Anisotropic Network Model (ANM) [51].The ENM and its CG versions are light and computationally inexpensive tools that have proved tremendously effective in dissecting function-related vibrational patterns in proteins, both embodied in low-frequency collective normal modes [52–55] and, more subtly, related to high-frequency localized vibrations [56–60].

### The Anisotropic Network Model (ANM)

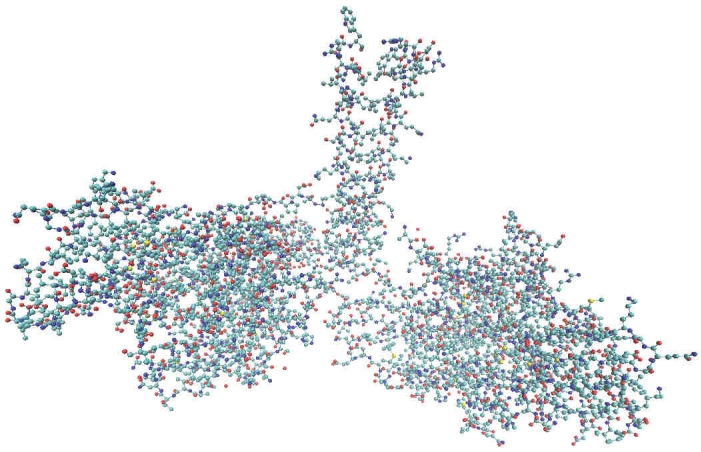
Within the ANM, a protein of $N$ residues is represented by $N$ beads centered on the $\alpha$-carbon of each residue (figure 2.3) connected by springs. Only the residues closer than a certain cutoff distance $R_c$ in the reference structure interact with each other. The system potential $U$ reads

$$U = \frac{1}{2} k_2 \sum_{i<j} C_{ij} (r_{ij} - R_{ij})^2, \tag{2.11}$$

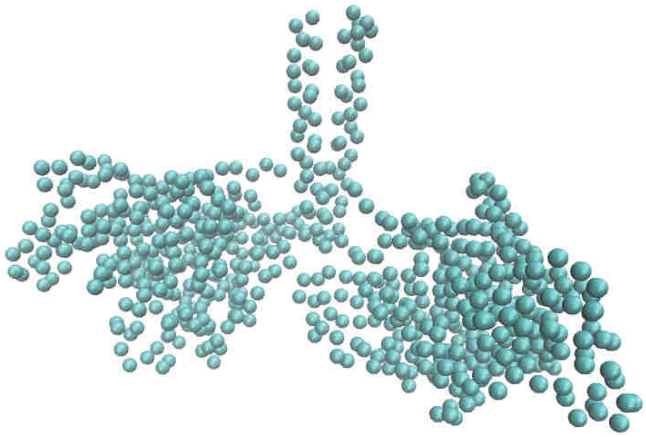
where $k_2$ is a constant fixed to fit the low frequency part of the linear spectrum of proteins [61] typically around $k_2$=5 kcal mol$^{-1}$ Å$^{-2}$, $r_{ij} = |\mathbf{r}_{ij}|$ and $R_{ij} = |\mathbf{R}_{ij}|$ are respectively the instantaneous distance and the equilibrium distance between particles $i$ and $j$. $C_{ij}$ defines the contact map and depends on the cutoff radius $R_c$,

$$\begin{cases} C_{ij} = 1 \text{ for } R_{ij} \leq R_c \\ C_{ij} = 0 \text{ for } R_{ij} > R_c \end{cases} \tag{2.12}$$

The two parameters here are $k_2$ and $R_c$. One can easily recognize that the more the protein is connected (large $R_c$), the more the stiffness of the springs should decrease in principle [51]. Differentiating the system potential defines the forces applied to each



(a)



(b)

Figure 2.3.: Residue-level coarse-grained representation of kinesin structure [5], (a) is the all-atom description and (b) is the coarse-grained description where each bead corresponds to one amino-acid.

residue $i$ along the Cartesian direction $\alpha$ ($\alpha = x, y, z$).

$$F_{i\alpha} = -\frac{\partial U}{\partial x_{i\alpha}} = -k_2 \sum_{j \neq i} C_{ij} \frac{(r_{ij} - R_{ij})(x_{i\alpha} - x_{j\alpha})}{r_{ij}}, \tag{2.13}$$

where $x_{i\alpha}$ is the coordinate of particle $i$ along the direction $\alpha$.

### Pump simulations

One of the main focuses of the present work is intra-molecular energy redistribution in proteins. With the growing number of structures available at atomic resolution and the increasing computational power available at reduced costs, people have developed many methods to explore this elusive problem [62–66].

Our contribution to these issues in this thesis is the development of an original procedure, that we have termed pumping simulations, inspired by the original work of Skinner and Sharp [67]. Within this scheme, a protein, described by the ANM, is subject to an external forcing applied locally (*i.e.* to a given particle) along a prescribed direction in space and at a specific frequency. Let $n$ denote the index of the forced residue, the system equations of motion then read

$$m_i \frac{\mathrm{d}^2 x_{i\alpha}}{\mathrm{d}t^2} = -\frac{\partial U}{\partial x_{i\alpha}} + F_0 \delta_{in} \hat{e}_\alpha \sin(\omega_0 t) - m_i \gamma \dot{x}_{i\alpha}. \tag{2.14}$$

Here $\hat{e}_\alpha$ are the direction cosines of the external pumping force. The presence of a weak distributed damping ($\gamma \ll \omega_0$) mimicks the presence of a solvent and ensures that a non-equilibrium steady state will be attained asymptotically.

In a typical pumping run, at time $t = 0$ all particles are at rest at their equilibrium position, as specified by the PDB coordinate file of the structure examined. After a transient stage, the system reaches its asymptotic steady state, characterized by a well-defined average field of local (residue) energies (see figure 2.4). Our idea was to employ such setting, with specific choices of the parameters $\omega_0$ and $\hat{e}_\alpha$, in order to pinpoint specific energy transduction pathways across the structure and rationalize the emergence of hot spots (energy pinning centers) in the light of the patterns of specific structural indicators. Our idea was to understand how the energy disperses across proteins depending on the frequency of the pump and whether the energy dispersion could be related to structural/topological measures. The results of these simulations will be discussed in chapter 4.

## 2.1.3. Integration algorithms

In order to solve numerically the equation of motions (2.14), we have used the velocity Verlet algorithm [68], which is a slightly different version of the algorithm originally introduced by Verlet in 1967 [69]. The velocity Verlet algorithm is similar to the Leapfrog scheme, where positions are taken at time $\Delta t$ and velocities at time $\Delta t/2$, with $\Delta t$ being the time step. This is a second-order, symplectic and time-reversible algorithm that is

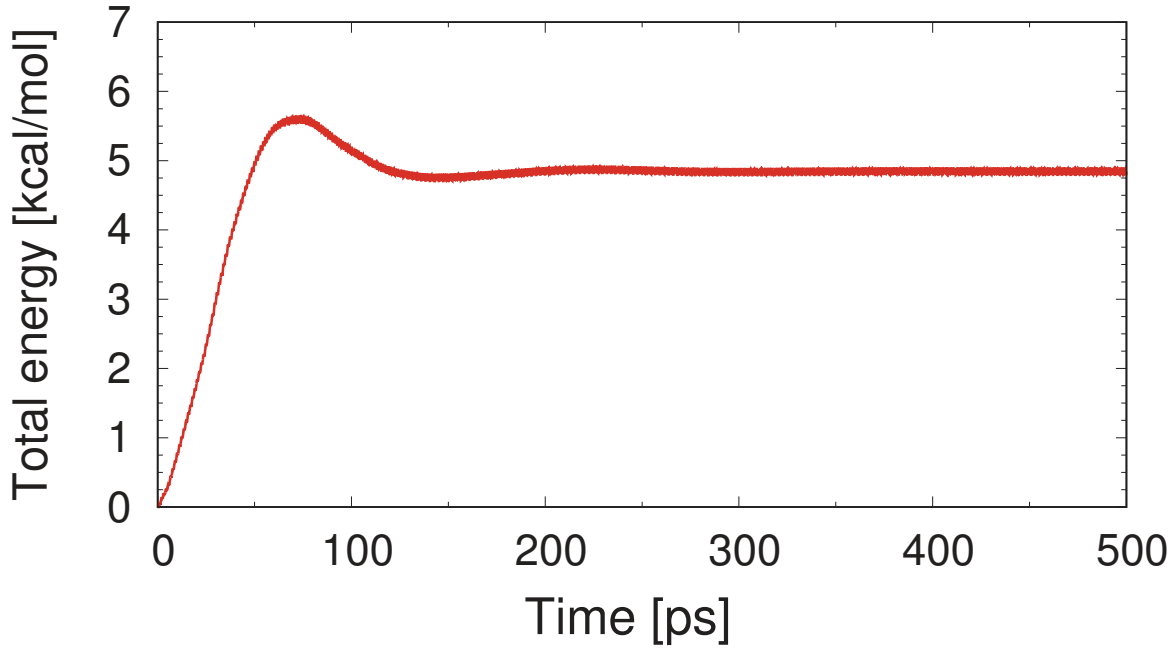



Figure 2.4.: Total energy of the human M2 muscarinic acetylcholine receptor [6] versus time during a pump simulation. After a transient period where the total energy increases due to the forcing, the total energy stabilizes due to the overall weak damping.

frequently used in molecular dynamics [70]. The iteration map reads

$$\begin{cases} x(t+\Delta t) = x(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 \\ v(t+\Delta t) = v(t) + \frac{1}{2}\big(a(t) + a(t+\Delta t)\big)\Delta t. \end{cases} \tag{2.15}$$

## 2.2. Structural indicators

Protein dynamics allows us to get time-resolved insight into intramolecular energy channeling and redistribution during a simulation where the protein evolves in time from a given initial condition. However, one can also introduce *static*, purely structural indicators, computed on the basis of the geometry of the protein fold, with the aim of pinpointing sites likely to behave as hot-spots [71,72]. Likewise, such measures can be used with some success to identify specific pathways connecting distinct regions within a given protein structure [73,74].

In this section, we will discuss Normal Mode Analysis (NMA) and the applications of this method explored in this thesis. Then, we shall explain how a protein structure can be mapped onto a graph and we will present the graph-theoretical measures that we have used in order to predict catalytic sites in enzyme structures. This is the subject of

chapter 3.

### 2.2.1. Normal Mode Analysis

Proteins are described by complex force fields involving many degrees of freedom. As it is customarily done in solid-state physics to study the vibrations of a crystal, it is expedient to consider small fluctuations of the structure around the equilibrium position. Hence, the total potential energy can be expanded as a power series in the vicinity of the native fold, that is,

$$U = U_{eq} + \sum_{i,\alpha} \Big(\frac{\partial U}{\partial x_{i\alpha}}\Big)\Big|_{\vec{x}^0} (x_{i\alpha} - x^0_{i\alpha}) + \frac{1}{2} \sum_{i,\alpha} \sum_{j,\beta} \Big(\frac{\partial^2 U}{\partial x_{i\alpha} \partial x_{j\beta}}\Big)\Big|_{\vec{x}^0} (x_{i\alpha} - x^0_{i\alpha})(x_{j\beta} - x^0_{j\beta}) + o(|x - x^0|^3) \tag{2.16}$$

where $U_{eq}$ is the system potential energy at equilibrium, $x_{i\alpha}$ and $x^0_{i\alpha}$ the instantaneous and equilibrium coordinates of the $i^{th}$ particle, respectively. Of course, by definition one has

$$\Big(\frac{\partial U}{\partial r_i}\Big)\Big|_{\vec{x}^0} = 0. \tag{2.17}$$

since the forces vanish in the equilibrium configuration. Taking $U_{eq} = 0$ with no loss of generality, one may write

$$U = \frac{1}{2} \sum_{i,\alpha} \sum_{j,\beta} \Big(\frac{\partial^2 U}{\partial x_{i\alpha} \partial x_{j\beta}}\Big)\Big|_{\vec{x}^0} (x_{i\alpha} - x^0_{i\alpha})(x_{j\beta} - x^0_{j\beta}) \tag{2.18}$$

Since the potential energy is a (positive-defined) quadratic form, the equations of motion of small fluctuations can be solved analytically. The equations of motion read

$$m_i u_{i\alpha} = -\sum_{j\beta} \mathbb{H}^{\alpha\beta}_{ij} u_{j\beta} \tag{2.19}$$

where $m_i$ is the mass of a residue (taken 110 a.m.u. in this work), $u_{i\alpha} = x_{i\alpha} - x^0_{i\alpha}$ are the components of the displacement field and we have defined the Hessian matrix elements

$$\mathbb{H}^{\alpha\beta}_{ij} = \frac{\partial^2 U}{\partial x_{i\alpha} \partial x_{j\beta}}\bigg|_{\vec{x}^0} \tag{2.20}$$

The solution to equations (2.19) can be written as :

$$x_{i\alpha}(t) = x^0_{i\alpha} + \frac{1}{\sqrt{m_i}} \sum_{k=1}^{3N} C_k \xi^k_{i\alpha} \cos(\omega_k t + \phi_k) \tag{2.21}$$

where the 3N-dimensional vectors $\boldsymbol{\xi}^k$ are the so-called normal modes. These are the eigenvectors of the mass-weighted Hessian matrix,

$$\tilde{\mathbb{H}}_{ij}^{\alpha\beta} = \frac{1}{\sqrt{m_i m_j}} \mathbb{H}_{ij}^{\alpha\beta} \tag{2.22}$$

corresponding to the eigenvalues $\omega_k^2$. In general, as can be seen from equation (2.21), atomic fluctuations are described by a superposition of many different frequency components, each weighted differently and contributing to the overall fluctuation pattern along the directions fixed by the normal modes.

In a thermal equilibrium state, each mode has the same energy $k_B T$. Hence at equilibrium one has

$$C_k = \frac{\sqrt{2k_B T}}{\omega_k} \tag{2.23}$$

where $k_B$ is the Boltzmann constant and $T$ the temperature.

In general, there are six degrees of freedom to which no energy change is associated according to the potential energy (2.18). Hence, the spectrum of normal modes always contain (for an expansion about a global minimum) six zero eigenfrequencies. These correspond to three rigid rotations and three rigid translations of the structure as a whole.

Generally speaking, low-frequency modes correspond to large-amplitude motions involving many residues of the protein, such as hinge-bending motions for example [75, 76], whereas high-frequency modes tend to be localized on a handful of residues [56, 59].

As an example, we have plotted the displacement pattern of the slowest and the fastest mode of vibration of the HIV-1 protease (PDB 1A30 [7]). Figure 2.5 shows that many residues are involved in the vibrational pattern of the slowest mode, whereas only few residues vibrate at the frequency of the fastest mode. In chapter 3, we have used the contribution to the 5 highest-frequency normal modes of vibration to build an indicator that gauges the local *stiffness* in order to predict catalytic sites in enzymes using a method that we have called *cutoff lensing*. This indicator measures the contribution to the fluctuations of the $i-th$ residue coming from the first 5 high-frequency normal modes, namely

$$\chi_i = \sum_{k=3N-4}^{3N} |\boldsymbol{\xi}_i^k|^2. \tag{2.24}$$

In chapter 4, we will discuss a class of normal modes that remarkably appear to be as general as largely overlooked. These modes are strongly localized, with two different and often widely separated (in 3D) localization centers. These modes, which we have termed bilocalized, appear to offer valuable clues as to the fold-rooted mechanisms of energy redistribution in proteins. This analysis will be the subject of chapter 4.

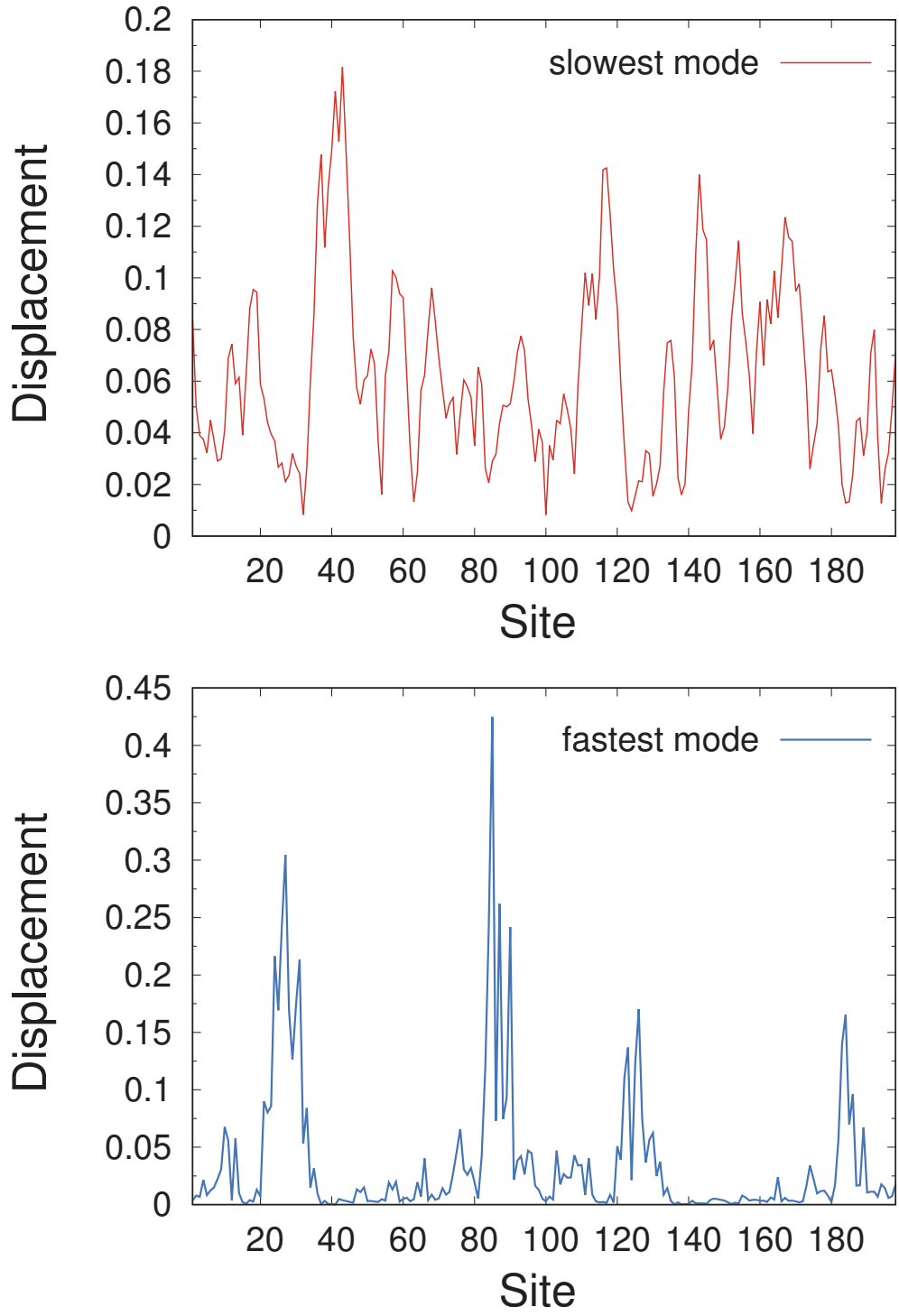



Figure 2.5.: Normal mode analysis of HIV-1 protease (PDB code : 1A30 [7]). Displacement field of the slowest (upper panel) and fastest (lower panel) normal mode. While the fastest mode is only represented by a few residues with a high contribution, the slowest mode involve many amino-acids with smaller displacements.

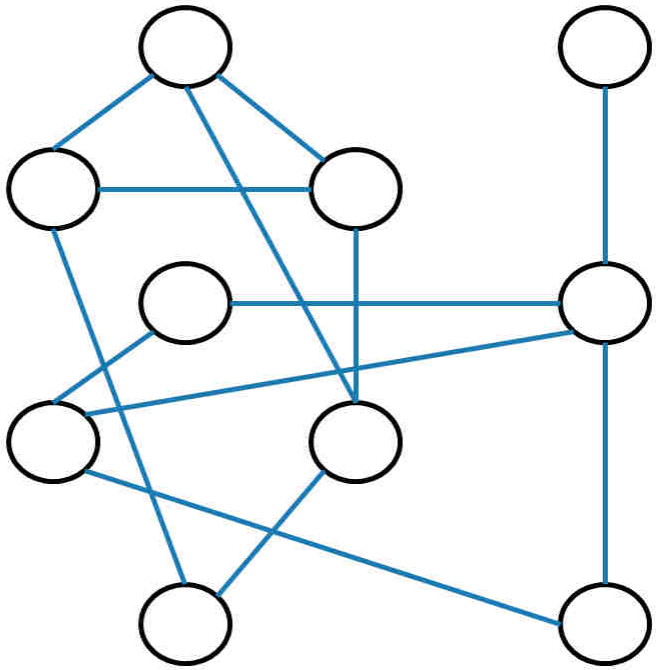


Figure 2.6.: Example of a small graph. A graph can be represented in two-dimensional space by nodes (black circles) and edges (blue lines) connecting the nodes.

### 2.2.2. Network representation of proteins

We have seen in the previous section that in coarse-grained elastic network model (CG ENM), each residue of a protein interacts with its neighbors within a certain distance, the cutoff radius $R_c$. Hence, a protein can be thought of as a collection of nodes (one for each residue) and edges (link between two interacting residues). In doing so, all the 3D protein structure is lost, all the distances between residues are in units of edges and we have an abstract representation of a protein as a connectivity graph.

By definition, a graph is a collection of nodes and edges. An edge connects two interacting nodes. If a graph is oriented, the edges that connect two nodes have a direction (from one node to the other). If a graph is weighted, the nodes and/or edges have different weights. A graph can be weighted and oriented.

In the following, we will consider graphs that are neither oriented nor weighted *i.e.* no preferential directions exist for the connections and the connections are all identical (no hierarchy among them). Since the protein CG ENM representation uses beads and springs between any two residues having an interdistance below the cutoff $R_c$, we have mapped a protein structure onto a graph with the following rules : each amino-acid is a node, an edge connects two residues if their $C_\alpha$ interdistance is smaller than the cutoff

radius.

Protein functions typically involves specific, well-defined restricted portions of the structure, such as in enzyme catalysis for example [77–79]. Furthermore, protein function can involve communication between different protein regions, such as for allostery [80–82]. Mapping a protein 3D structure on a graph enables one to use a large palette of interesting graph-theoretical measures. The prediction of hotspots in proteins can proceed by inspecting different measurements of localization on graphs, such as centrality measures [83–87]. Communication between nodes on a graph can be studied by using graph-theoretical tools that quantify specific paths across the edges, such as shortest-distance or betweeness measures [73, 74, 88, 89].

We will present here two indicators that we have used in order to predict catalytic sites in enzymes, connectivity and *closeness* centrality.

### Connectivity

Connectivity is one of the simplest features of nodes in a graph. The connectivity of a given node is the number of its neighbors, *i.e.* the number of other nodes with which it shares an edge. Hence, connectivity is a local measure. To compute this indicator, one only needs the contact map of the protein. Recalling the definition of the contact map $C_{ij}$ in coarse-grained elastic network (equation 2.12), the connectivity $c_i$ of the $i^{th}$ residue is simply given by

$$c_i = \sum_{j \neq i} C_{ij}. \tag{2.25}$$

### Closeness centrality

*Closeness* centrality of a node defines how *close* this node is from all the others. While the connectivity of a node is a local feature and depends only on the direct environment of the nodes, *closeness* centrality is a global measure as it involves the whole graph.

A node is *close* to another if their interdistance (in edge units) is short. Hence, taking into account the entire graph, the *closeness* centrality is a measurement of *how far* is a node from the other nodes. A node having a substantial *closeness* centrality minimizes the shortest path to the other nodes or, analogously, maximizes the inverse of the shortest path to the other nodes.

Let $d(i, j)$ be the shortest distance (in edge units) between residues $i$ and $j$. The *closeness* centrality $CC_i$ of node $i$ then reads,

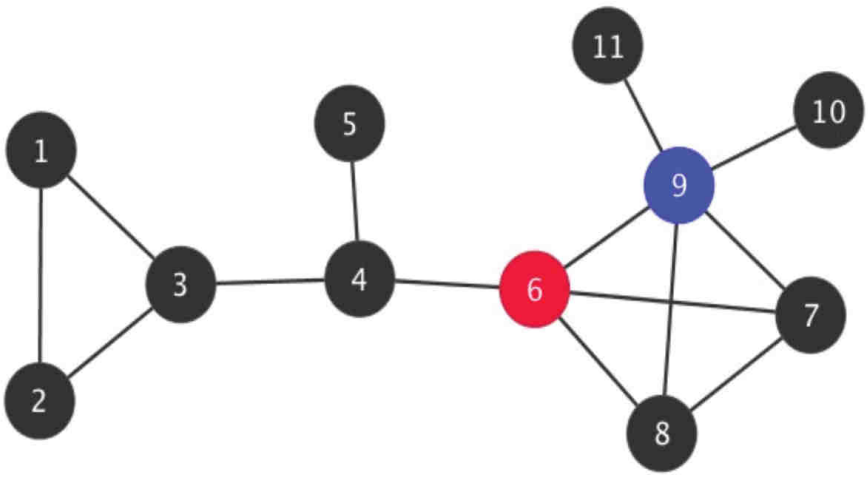



Figure 2.7.: Scheme of a graph with $N = 11$ nodes. The node with the highest connectivity is shown in blue while the node characterized by the highest *closeness* centrality is colored in red.

$$CC_i = \sum_{j \neq i} \left[ d(i,j) \right]^{-1}. \tag{2.26}$$

### An example

In order to illustrate the computation of the connectivity and the *closeness* centrality, let us examine a small graph consisting of 11 nodes and 13 edges (figure 2.7). The node that has the highest connectivity is the $9^{th}$ (in blue) whereas the node that has the highest *closeness* centrality is the $6^{th}$ (in red). The highest values of these indicators in this example single out two neighboring nodes. This appears rather logical, as higher connectivity also implies shorter distances from nodes around. The main difference between the two measures is that the connectivity is a *local* indicator, whereas the closeness centrality is a *global* measure *i.e.* one that depends on the graph topology as a whole.

## 2.3. Summary

In this chapter we have introduced the basic physical models underlying the analyses presented in the remaining parts of the thesis.

We have first detailed the all-atom method that we have used for the original cooling

simulations. In the following we have introduced the notion of elastic-network model and detailed the pump protocol that we have employed to investigate energy redistribution in proteins.

In the last part of this chapter we have dwelled on the concept of normal-mode analysis, which is a valuable tool which we resorted to frequently in our different studies. Finally, we have introduced the basic graph-theoretical measures that we have employed in our quest for catalytic sites in enzymes. This will be the subject of the next chapter.

# 3. Cutoff lensing: predicting catalytic sites in enzymes

With the rapid development and refinement of experimental techniques for protein structure determination at high resolution, predicting functional sites is a major issue in modern molecular biology in many protein families [71, 72, 90–94].

The swiftly growing amount of structural and sequence data poses big challenges and offers great opportunities to test automated prediction algorithms and platforms. Several approaches have been used to identify critical function-related sites (sometimes referred to as *hotspots*) in proteins. Most of these methods imply structural and/or sequence conservation information [95–101].

Purely sequence conservation approaches use phylogenetic information, relying on the idea that functional sites are conserved during evolution. Typically, such algorithms proceed through the alignment of a great number of different sequences and the ensuing computation of different conservation scores [94, 102–105].

Other approaches can be found in the literature, typically combining sequence-related information with structural data to achieve higher prediction rates [106–108].

Among the structure-based algorithms developed to identify and predict function-related sites in proteins, an appealing and promising class is that of coarse-grained (CG) [109, 110] approaches based on elastic-network models (ENM) [111–116].

The ENM [50]and its CG versions [117, 118] are light and computationally inexpensive tools that have proved tremendously effective in dissecting function-related vibrational patterns in proteins (see chapter Methods).

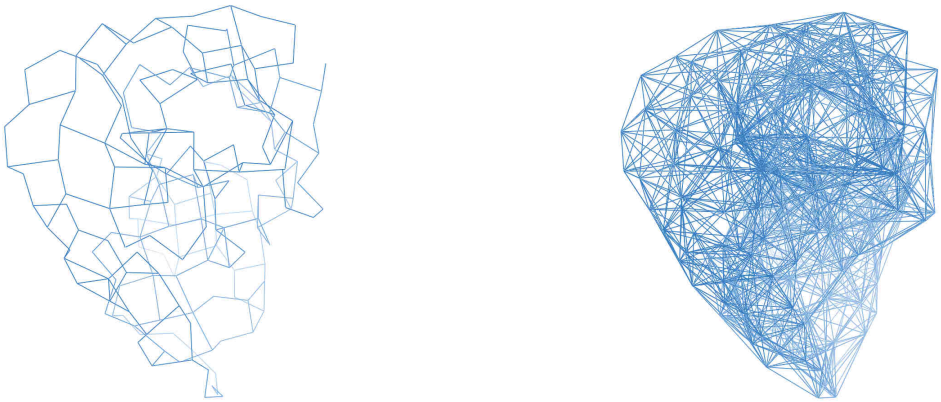
Often, graph-theoretical tools have been employed in combination with ENM-related approaches [83, 84, 119–124] to identify hotspots and binding interfaces. In these methods, a protein structure is mapped onto a network by means of some rule (see chapter Methods).

It is clear from the above discussion that a successful strategy to predict functional sites in proteins has to rely on a composite approach, combining information from sequence conservation with structure-based analyses. In turn, the latter should combine different indicators, related to the physical-chemical properties of amino acid environments and to patterns of chemical and topological connectivity.

In this chapter we focus on the prediction of catalytic sites in enzymes based on an original ENM-based strategy. Atomistic approaches devised to identify residues involved in catalysis in enzymes are not new [125]. More recently, approaches specifically relying on sophisticated electrostatic calculations have been introduced [126, 127]. Conversely, coarse-grained models have been relatively less exploited to solve this specific problem [116, 128, 129]. Yet, ENM-based tools are light (they can be applied to large databases of structures) and can be readily extended to perform all-residue searches in many structures. Moreover, CG topology-based methods have the advantage to strip the structure of most chemical details so as to bring to the surface purely topological features. This appears particularly important in the case of enzymes, as often sites that are involved in the catalytic action are intriguingly found far from the annotated catalytic sites [130, 131].

Our method combines three different indicators, two graph-theoretical measures with an original scale of local stiffness in a method that we termed *cutoff lensing*. The main idea is that catalytic sites can be spotlighted by employing elastic network models whose connectivity is increased beyond currently employed values. In ENMs, a spring is stretched between all pairs of residues that are separated in the equilibrium structure by a distance less than a specified cutoff length $R_c$. Typically employed values for protein models coarse-grained at the level of amino acids vary in the 10–13 Å range [38,118,132–134], even if values greater than 13 Å have also been considered episodically [135–137]. In principle, larger values of the cutoff are unphysical, as the connectivity graph becomes nearly fully connected as $R_c$ attains a value comparable with the protein size. Nevertheless, we have found that specific, function-related sites can be singled out in such regimes by using indicators associated with topological and structural measures of connectedness and stiffness. Remarkably, a scan of increasing values of the cutoff shows that there exists an optimum range where our structural indicators are the most sensitive in detecting catalytic sites known from experiments. This *lensing* effect can thus be used to predict the location of functional sites in unannotated proteins.

The chapter is organized as follows. In the next section we provide the description of the cutoff lensing method and introduce three structure-based indicators. In the following section we check the predictive power of our indicators against the pool of annotated catalytic sites in a large database of enzymes. Finally, we discuss our results and provide a working summary of our method.



(a) (b)

Figure 3.1.: HIV1-protease (PDB 1A30) represented as a 3D network of nodes and edges for two different values of the cutoff. (a) $R_c = 5$ Å, (b) $R_c = 10$Å.

## 3.1. Structural indicators

The basic idea of our method rests on the evidence reported by several studies that hotspot/functional sites in proteins are generally found in stiff/rigid regions [129,138–140]. Analogously, it has been shown that functional residues tend to move independently from the rest of the structure, involving high-frequency localized vibrations (the stiffer the bonds, the higher the frequency) [38, 132, 141, 142].

Structural rigidity can be gauged by many indicators, that assess the different flavors associated with it. The simplest and more intuitive method, albeit unsuitable for automated screening of large structure databases, would be to measure fluctuations directly via MD simulations, such as in reference [143]. Alternatively, but more indirectly, rigidity can be related to the local number of neighbors in the protein connectivity graph. A series of recent studies has demonstrated a rather surprising agreement between the location of catalytic sites in enzymes and the localization patterns of nonlinear vibrational modes known as discrete breathers (DB) [57, 59]. Such observations have been rationalized in terms of a *spectral measure* of *local stiffness*, based on the localization properties of high-frequency normal modes [132]. Here we introduce an original method based on a blend of suitable structural indicators combined with *cutoff lensing, i.e.* an analysis where the cutoff $R_c$ is let increase beyond physically realistic values. The key feature of this method is a selective sharpening of the predictive power of our indicators at specific intermediate values of $R_c$.

A *spectral* stiffness measure $\chi_i$ for each residue $i$ can be computed by looking at the contribution of a reduced set of high-frequency normal modes (NMs) (see chapter Methods). The rationale behind equation 2.24 comes from the observation that fast normal

modes tend to be localized at hotspot sites [132], *i.e.* sites that act as efficient energy storage and accumulation centers, typically flagging highly connected and buried regions. Along the same lines, fast modes have also been demonstrated to identify stability cores of proteins [56], adding to the meaningfulness of the *local stiffness* definition. Typically, in residue-based coarse-grained ENMs the last high-frequency NMs are localized around one, two sites at most. If one considers an average number of catalytic sites per enzyme around 5 (it is 2.5 in the Catalytic Site Atlas (CSA) [144]), it appears that the minimum number of high-frequency NMs to include in the definition 2.24 is five (adding a few more NMs does not change appreciably our results. Adding more results in useless blurred patterns).

Following a similar rationale, we shall also consider indicators referring to the connectivity graph, notably the local connectivity $c_i$ and the closeness centrality $CC_i$ for the $i^{th}$ residue (see chapter Methods).

The three above-defined indicators can be regarded as supplying different measures of stiffness. While $\chi_i$ gauges the vibrational stiffness of a given residue, *i.e.* its propensity to vibrate at high frequency with a space-localized pattern, $c_i$ and $CC_i$ exquisitely quantify the topological stiffness, in the sense of number of outgoing bonds ($c_i$) or shortest paths between two given locations flowing through $i$ ($CC_i$).

As a general rule, the raw measures of $\chi_i$, $c_i$ and $CC_i$ result in rather rugged and irregular patterns with many peaks and troughs for a given protein sequence. Our goal is to extract from such patterns the most relevant peaks as flags for potentially functional sites. To this aim, we apply a high-pass *filtering* procedure, by keeping for a given indicator pattern only the values above a specified number of standard deviations (computed over the whole sequence). More precisely,

$$\tilde{\chi_i} = \begin{cases} \frac{\chi_i}{\chi_i^{\max}} \text{ for } \chi_i > 3\sigma_\chi \\ 0 \text{ otherwise} \end{cases} \tag{3.1}$$

$$\tilde{CC_i} = \begin{cases} \frac{CC_i}{CC_i^{\max}} \text{ for } CC_i > 3\sigma_{CC} \\ 0 \text{ otherwise} \end{cases} \tag{3.2}$$

$$\tilde{c_i} = \begin{cases} \frac{c_i}{c_i^{\max}} \text{ for } c_i > 5\sigma_c \\ 0 \text{ otherwise} \end{cases} \tag{3.3}$$

with the index "max" denoting the maximum value of the measurement.

Our final site predictions are then obtained as the locations flagged by the peaks

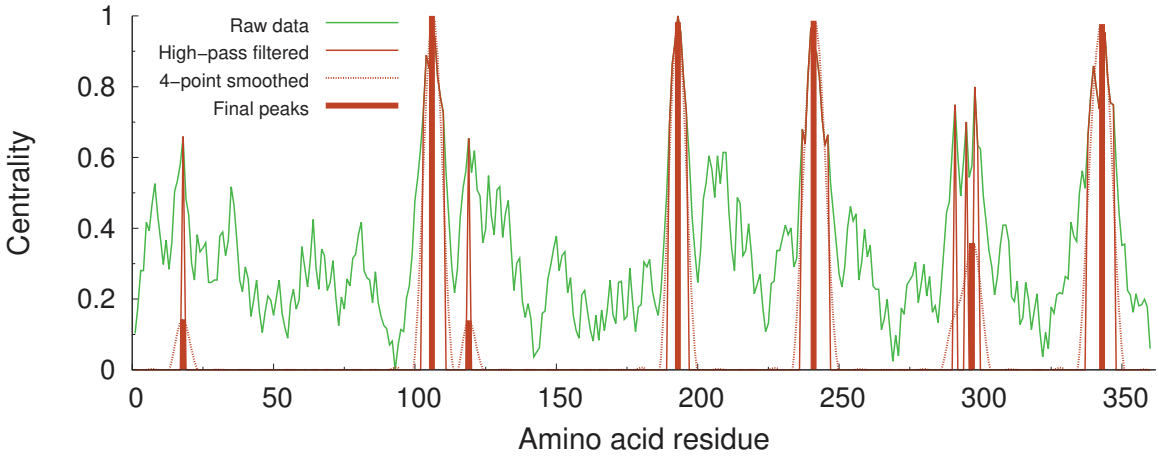



Figure 3.2.: Illustration of the computation of the reduced closeness centrality indicator through the different sequential steps described in the text. The patterns are normalized to the maximum value occurring in the sequence. The final peaks flag the potentially functional sites. The calculations refer to Arginin Glycineaminotransferase (PDB code 1JDW).

that survive in the pattern after the high-pass filtering. In the case of $CC$, the patterns showed overly rugged profiles (see figure 3.2), which resulted in a large number of close, quasi-degenerate peaks after the high-pass filtering. Accordingly, in order to eliminate the degeneracy associated with multiple-peak structures, we applied a 4-point smoothing procedure [145] to the filtered patterns, so as to automatically make the excessively degenerate structures coalesce in one single-peak prediction. The whole procedure is illustrated in figure 3.2.

## 3.2. Results

Our idea is to inspect reduced (filtered) patterns of local spectral and topological stiffnesses in search for hot spots. One of such patterns is reported in figure 3.3 for two values of the cutoff parameter $R_c$ used to construct the elastic network (see again equation 2.12). Interestingly, one may easily remark that there is a correspondence between the location of known catalytic sites and stiffness peaks. This finding agrees with observations made by other authors along the same lines [116,129]. However, if we now repeat the same analysis with a higher (even if less physical value of $R_c$), the surprising consequence is that the reduced pattern is sharpened down to a handful of peaks, which seem to much better pinpoint the known functional sites. Note that the observed sharpening of the predictive power implies both the evaporation and the relocation of some peaks. We term this effect altogether cutoff lensing. The logical questions to ask in view of such findings are (i)

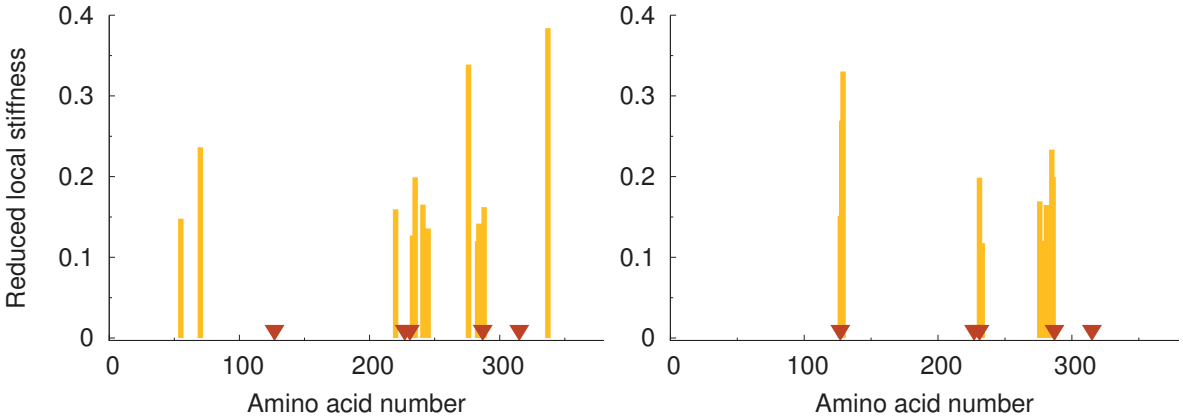



Figure 3.3.: Illustration of the cutoff lensing effect. Plot of the reduced stiffness pattern $\tilde{\chi}$, equation 3.1, for Arginin Kinase (PDB code 1BG0). Cutoff $R_c = 10$ Å (left) and $R_c = 20$ Å (right). The known catalytic sites are indicated by dark triangles. Note the disappearance of some irrelevant peaks and the appearance of a peak at one of the catalytic sites in going from $R_c = 10$ Å to $R_c = 20$ Å.

whether these effects also characterize the other indicators and (ii) whether there exists an optimum value of $R_c$, corresponding to the maximum overlap between (generalized) stiffness peaks and catalytic sites, beyond which the patterns get blurred again and one correspondingly loses predictive power. The latter possibility, in particular, seems highly realistic, as one expects sites to be no longer distinguishable (with respect to whatever measure) in nearly fully connected networks.

The results reported in figure 3.4 for a given enzyme seem to reply to the first question in the affirmative: intermediate values of the cutoff appear to be associated with increased predictive power. The reduced connectivity $\tilde{c}_i$ and spectral stiffness $\tilde{\chi}_i$ profiles suggest that intermediate values of $R_c$ yield a better match between the peaks of the indicator patterns and the annotated sites. The centrality, on the contrary, provides a good match but seems at the same time rather insensitive to changes in the cutoff. It is important to observe that the number of peaks $N_p$ is not constant as a function of $R_c$. Of course, this information has to be included in the picture if we want to provide a statistical assessment of the predictive power of our indicators as a function of $R_c$. On the one hand, $N_p$ is expected to increase at high values of $R_c$ for the connectivity and centrality measures, while it seems that stiffness patterns display less and less peaks as the cutoff is made larger. In order to shed further light on the above-described findings and proceed to a statistical assessment of the ability of our indicators to spotlight function-related sites, we have analyzed a pool of 835 enzyme structures from the Catalytic Site Atlas [144]. The CSA is a major resource in the field of structural biology, and provides up-to-date catalytic

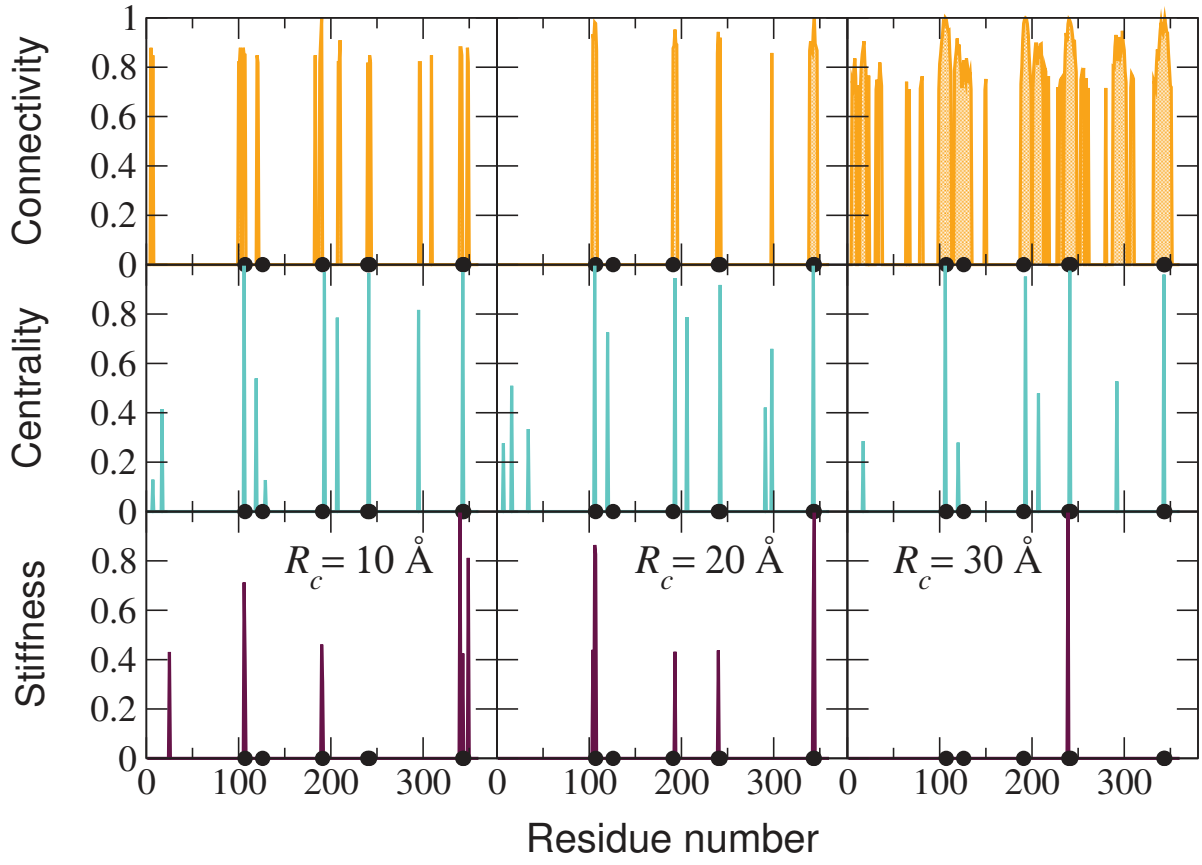



Figure 3.4.: Reduced and normalized connectivity, closeness centrality and stiffness patterns computed according to the prescription (equation 3.1, 3.2 and 3.3) for Arginin Glycineaminotransferase (PDB code 1JDW) for different values of the cutoff $R_c$. The annotated catalytic sites are indicated by black filled circles.

residue annotation for enzymes in the Protein Data Bank based on experimental data. The results of our ensemble analysis are reported in figure 3.5. For each indicator, we have calculated the fraction of catalytic sites that are found within a prescribed distance $\Delta n$ (in units of residues) along the sequence from a peak. For example, the curves at $\Delta n = 0$ indicate the fraction of catalytic sites that coincide with a peak for a given indicator. A number of interesting observations can be made by inspecting figure 3.5. The reduced connectivity $\tilde{c}_i$ increases its predictive power at increasing values of the cutoff. However, this is a trivial consequence of the fact the number of peaks also increases as the systems become more and more connected (top right panel). Therefore, the connectivity does

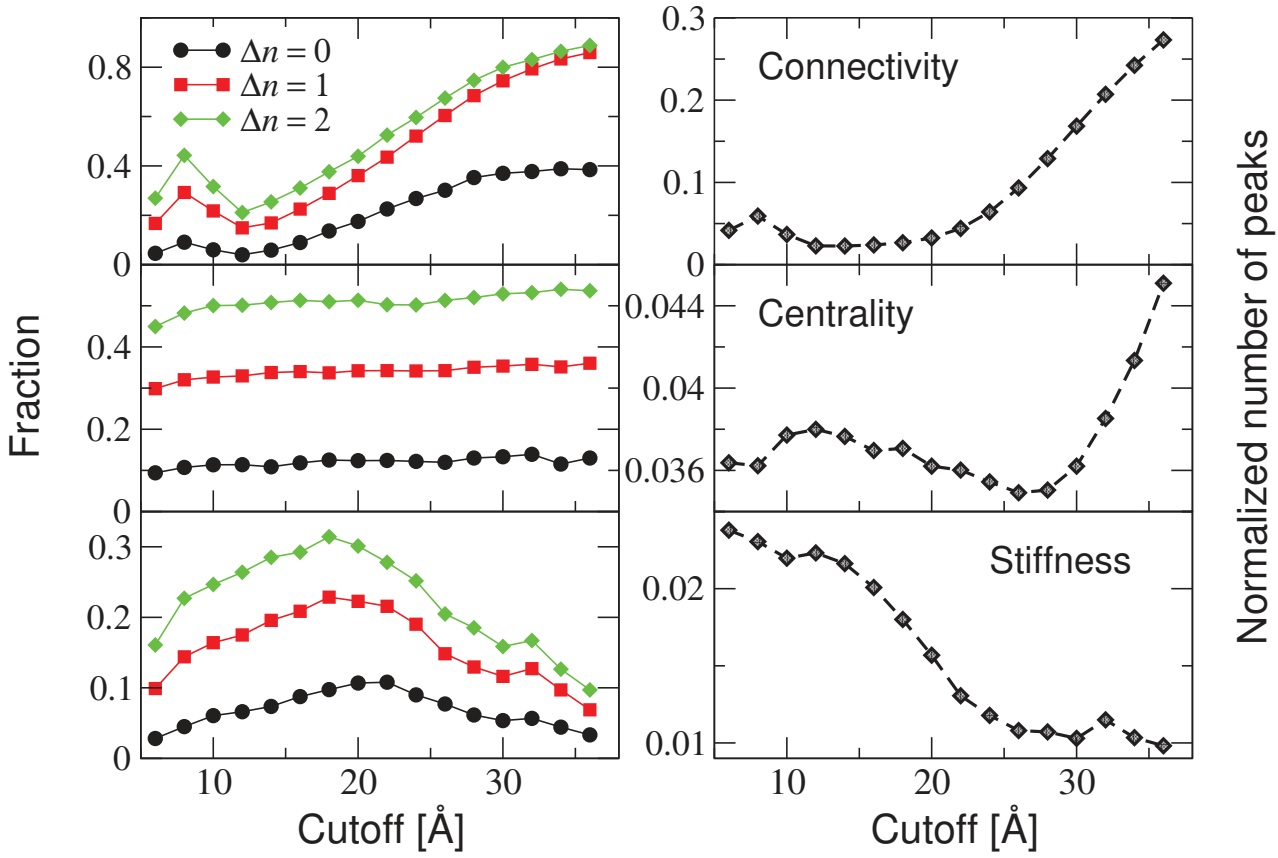



Figure 3.5.: Left panels: fraction of catalytic residues within $\Delta n$ sites from the nearest peak versus cutoff, as computed over the ensemble of enzymes from the CSA. Right panels: average peak fraction (number of peaks divided by number of residues) computed over the whole database versus cutoff.

not appear to provide a particularly insightful spotlighting tool. On the contrary, the reduced centrality $\tilde{CC}_i$ provides a comparatively more sensitive detection tool, with up to half of the whole pool of catalytic sites found at a separation of at most one residue (along the sequence) from a $\tilde{CC}$ peak. Furthermore, it is seen that the predictive power of this indicator is almost insensitive to the number of peaks, which increases of course as the structures become more and more connected (middle right panel). Interestingly, the average number of peaks in the $\tilde{CC}$ patterns displays a minimum (around $R_c = 28$ Å), which suggests that at this value of the cutoff the reliability of the observed predictive power of centrality is maximum.

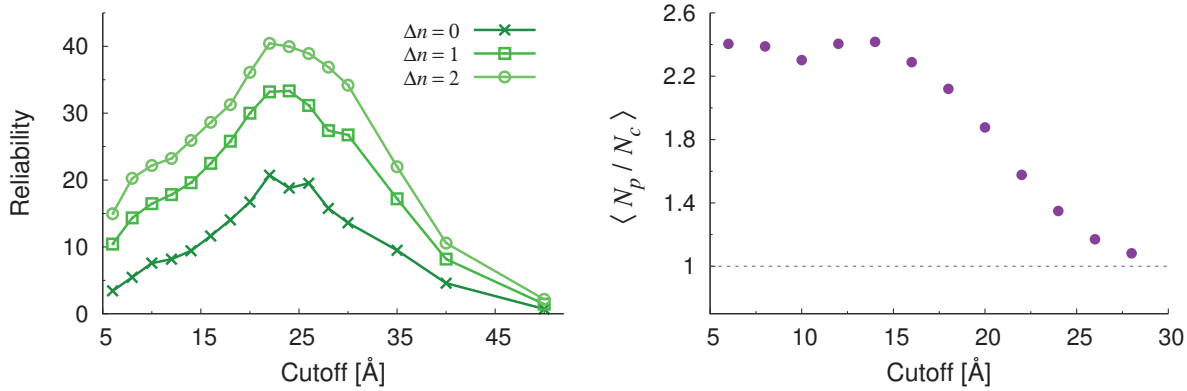



Figure 3.6.: Left panel: reliability of the predictive power of reduced stiffness patterns as a function of the cutoff $R_c$ (arbitrary units). The reliability is defined as the fraction of predicted catalytic sites (within $\Delta n$ amino acids along the sequence) divided by the fraction of stiffness peaks (number of peaks per amino acid). Right panel: Average number of peaks in the reduced stiffness patterns per catalytic site.

Of the three indicators, the reduced stiffness $\tilde{\chi}$ displays the most interesting behavior. The fraction of predicted sites shows a maximum at intermediate cutoff values (around 20 Å), with up to 30% of the known catalytic sites recovered at a distance of one amino acid from a peak of reduced stiffness. Interestingly, the number of such peaks decreases towards a nearly constant value as the cutoff is increased. Most remarkably, the maximum of predictive power clearly falls in a regime where the number of peaks has attained its minimum asymptotic value, which means that the statistical significance of the prediction at the maximum is also maximum. To make this observation more quantitative, one may introduce an intuitive measure of reliability, defined as the fraction of predicted sites divided by the number of peaks found at each value of the cutoff. This is illustrated in figure 3.6 (left panel). It is clear that the predictions made from the reduced stiffness patterns correspond to a maximum of reliability at the intermediate cutoff $R_c \cong 22$ Å. This suggests that the cutoff lensing effect can be effectively employed to predict the location of catalytic sites or to substantiate the predictions made by means of other methods based on different arguments. This is also confirmed by the observation that the highest number of predicted sites and maximum reliability corresponds to roughly one stiffness peak per catalytic site (see right panel in figure 3.6). This suggests that the condition of maximum predictive power is achieved with the least number of unassociated peaks, *i.e.* under conditions of highly reduced redundancy.

### 3.2.1. Analysis by size

It is interesting to ask the question whether the size of proteins matter for what concerns the predictive power of our toolkit. In fact, it is reasonable to assume that the same cutoff would correspond to different levels of *global connectedness* for structures that differ widely in size. As a result, the behavior of our structural indicators should not be homogeneous for structures whose sizes are markedly different. To investigate this aspect, we partitioned our pool of enzymes into three different classes, small, medium, large, depending on the number of residues. We then repeated the same analysis detailed in the previous sections on these new, size-resolved ensembles. The results are illustrated in figure 3.7. Small proteins have a number of residues between 60 and 300, medium sized proteins between 301 and 540, while large proteins have a number of residues higher than 540.

Closeness centrality remains rather constant for increasing values of the cutoff as we have seen in the previous section. However, an interesting finding is that the fraction of prediction is greater for large proteins than for small ones, with a difference of almost 40% for the prediction of catalytic sites within a distance of two amino-acids along the chain.

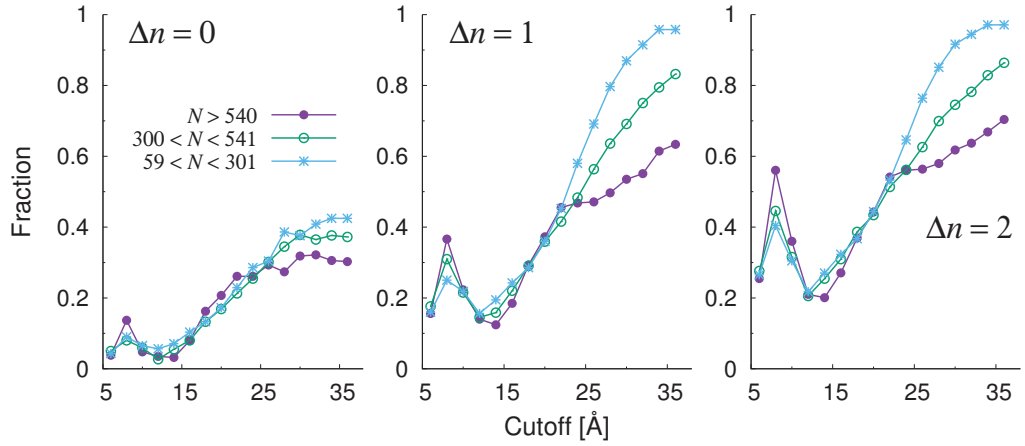
The analysis of stiffness yields even more interesting results. When we observe the last row of figure 3.7, we remark that the fraction of prediction increases with size, and the cutoff corresponding to the optimum also shifts towards greater values, depending on the size of the proteins. If we look at the bottom right panel, we observe that optimal prediction for small enzymes is achieved for $R_c = 16$ Å, for medium for $R_c = 20$ Å while we are able to predict 45% of the catalytic sites with the sole use of the stiffness indicator for large proteins with a cutoff $R_c = 24$ Å.

Overall, the above observations demonstrate that our indicators have a higher power of prediction for large proteins. For small ones, the stiffness measure can be computed with $R_c = 16$ Å in order to yield a better prediction than the overall optimal value, *i.e.* $R_c = 22$ Å. In general, also the fraction of catalytic sites recovered by reduced closeness and connectivity profiles is greater for enzymes of larger sizes.

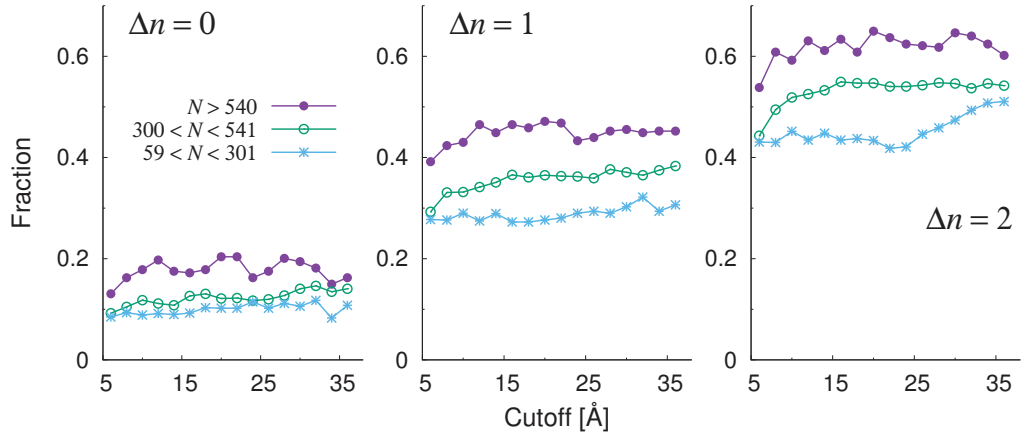
### 3.2.2. Combined analysis

With three different indicators, the question remains whether we can construct a better-performing single indicator through a combination of the three individual figures or whether a suitable sequence of separate analyses is best.

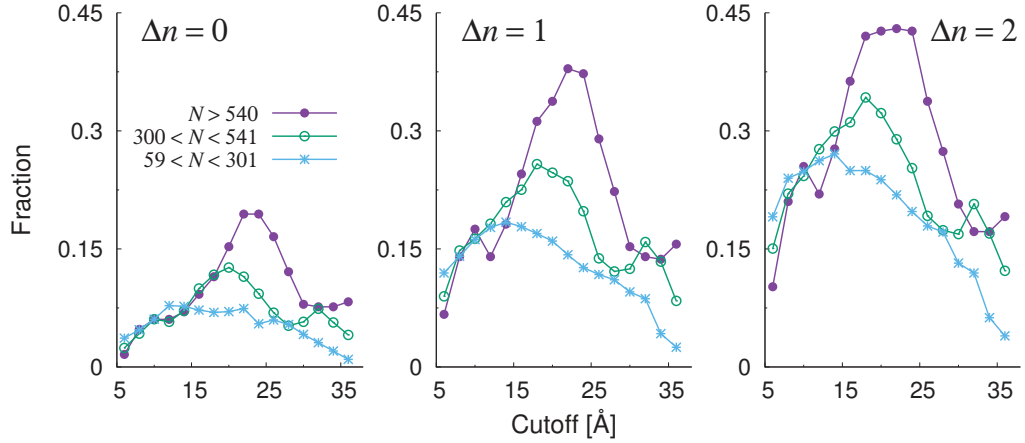
The idea is to use the indicators with their respective optimal cutoff value, 20 Å for the connectivity, 28 Å for the closeness centrality and 22 Å for the stiffness, and




(a)




(b)




(c)

Figure 3.7.: Fraction of catalytic sites within $\Delta n$ sites from the nearest peak of the three reduced patterns computed over three different size classes in the CSA database versus cutoff. (a) Connectivity, (b) closeness, (c) stiffness.

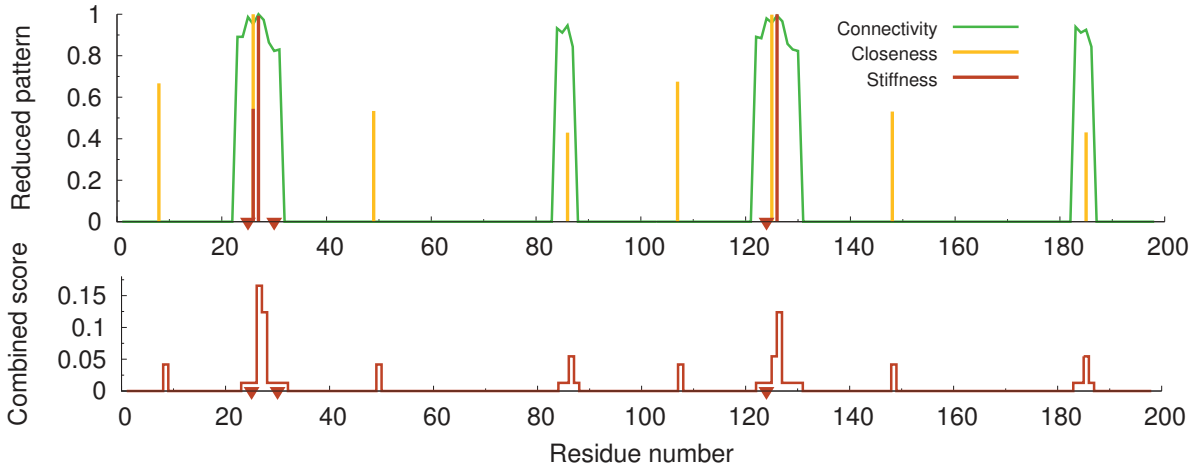



Figure 3.8.: Analysis of HIV-1 protease (PDB 1A30). The upper plot shows the three reduced indicator patterns. The bottom panel illustrates the combined site score given by equation 3.4.

construct a single combined measurement in order to obtain a maximum of prediction power. The simplest combined analysis would be to add the three indicators normalized by the corresponding number of peaks. In formulae we have for each particle $i$,

$$S_i = \frac{1}{3}\left[\sigma_i^{\tilde{\chi}} + \sigma_i^{\tilde{CC}} + \sigma_i^{\tilde{c}}\right], \tag{3.4}$$

with $\sigma_i^{\tilde{I}} = \frac{1}{N_p^{\tilde{I}}}$, $N_p^{\tilde{I}}$ being the number of peaks of the indicator $\tilde{I}$, if $\tilde{I}_i > 0$ and 0 elsewhere. By construction, we have $\sum_i \sigma_i^{\tilde{I}} = 1$ and $\sum_i S_i = 1$. The meaning of $S_i$ is to gauge the local prediction by adding all peaks derived from the three individual indicators but normalized by their respective number of peaks. Indeed, the more the peaks, the smaller its predictive power. For the sake of simplicity, we treated all the indicators with the same global *weight*, namely 1/3 for each.

In figure 3.8 we illustrate the performance of the combined indicator for a protein of choice, alongside with the three individual indicator patterns computed at their respective optimal cutoff. While all the known catalytic sites can be flagged using our algorithm, the combined indicator does not seem to provide enhanced predictive power compared to the results of the separate measurements.

We can construct an overall score for each enzyme by adding up each score $S_i$ within a distance $\Delta n$ from each of the $N_c$ catalytic sites. The indicator $S_{\Delta n}$ can be computed as

$$S_{\Delta n} = \sum_{i_c=1}^{N_c} \sum_{i=i_c-\Delta n}^{i=i_c+\Delta n} S_i. \tag{3.5}$$

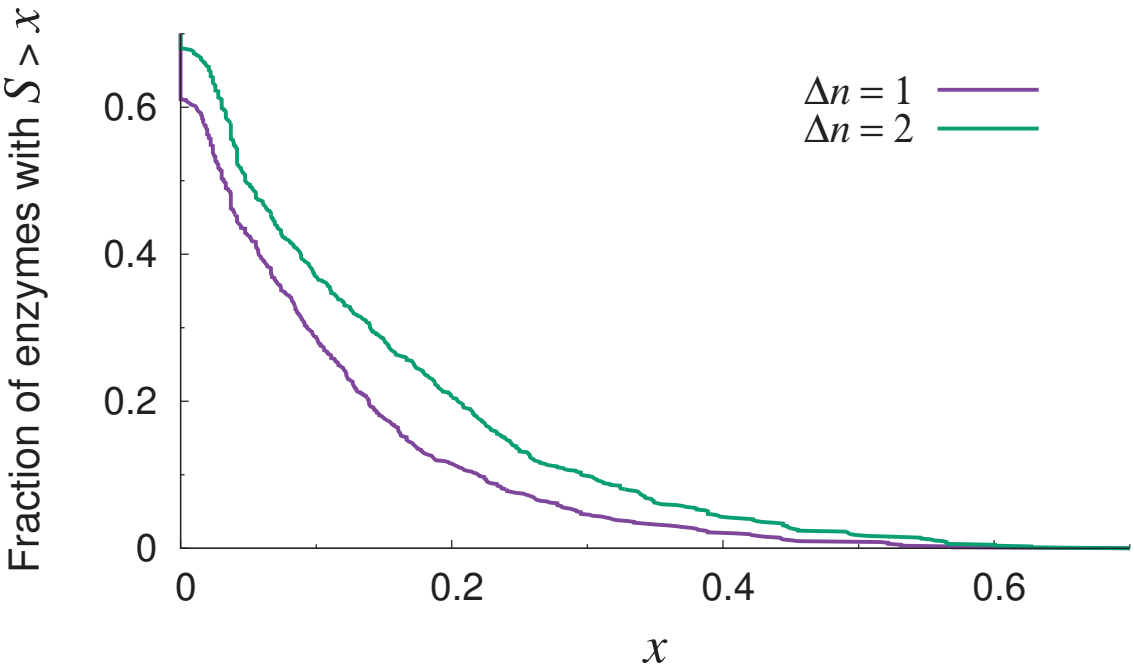



Figure 3.9.: Complementary cumulative distributions of global enzyme scores computed over the whole CSA database through equation 3.5. A positive score signals that a prediction has been made. The actual value of the score is a measure of the relative number of orphan peaks (putative false positives). As a general rule, the larger the score, the less in number and/or the smallest in height were the orphan peaks.

If $S_{\Delta n} > 0$ for a given structure, our algorithm is able to provide at least one prediction. An analysis performed over the whole CSA database shows that the fraction of structures where the combined algorithm returns a prediction is 0.61 for $\Delta n = 1$ and 0.68 for $\Delta n = 2$ (see figure 3.9). In order to elucidate the meaning of the site scores $S_i$ and global score $S_{\Delta n}$, it proves useful to concentrate on a specific enzyme. In figure 3.8, we consider the classic case of HIV-1 protease. Let us first concentrate on the profile of the combined score (3.4). Two facts are immediately apparent: (i) the catalytic sites appear to be all captured but (ii) there are a number of *orphan* peaks. The global scores for this enzyme are $S_{\Delta n=1} = 0.27$ and $S_{\Delta n=2} = 0.29$. Thus, despite the algorithm flags correctly all the catalytic sites, it does so with some degree of over-prediction (incidentally, we observe that the orphan peaks shown in the combined score profile in figure 3.8 might as well spotlight some hitherto unknown functional sites of HIV-1 protease). Of course, when applying the algorithm to unannotated structures, one does not know a priori which peak in the combined score is more likely to point to a catalytic site. This shows the limitations of using only a combined score. Analogous conclusions can be drawn by looking at the fraction of catalytic sites predicted by one or more indicators (see figure 3.10). For

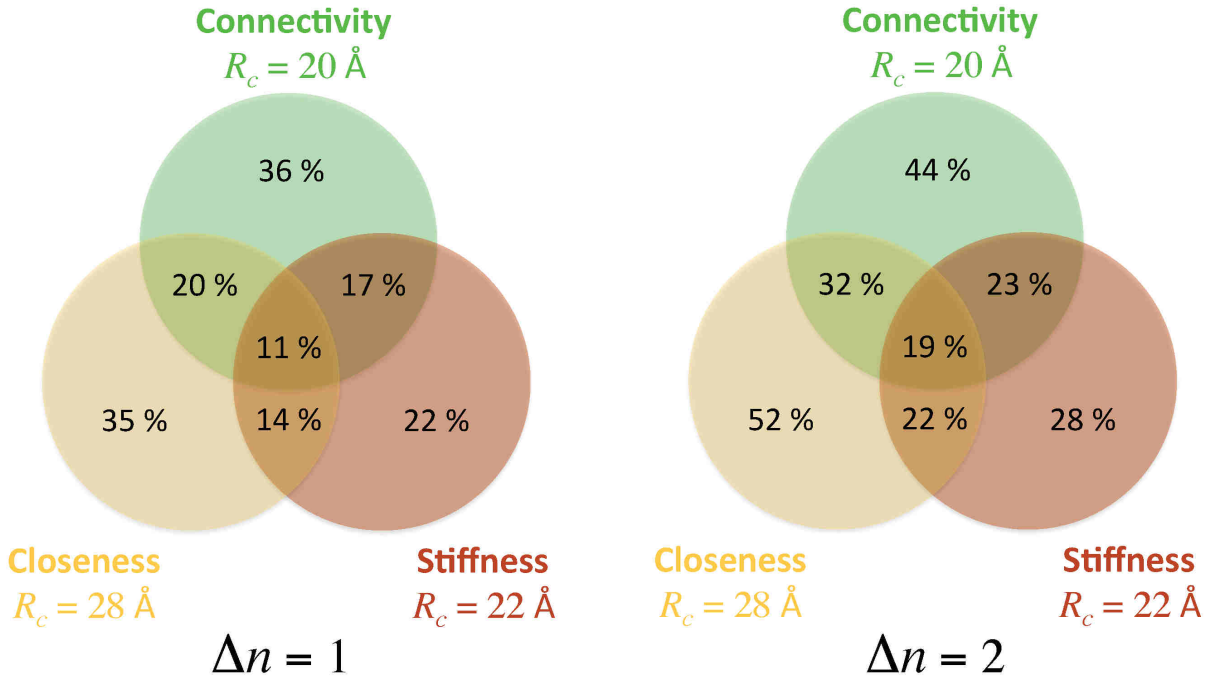



Figure 3.10.: Synoptic representation of the fraction of predicted catalytic sites over the CSA database at the individual optimal cutoff values for $\Delta n = 1$ and $\Delta n = 2$.

example, closeness and stiffness reduced patterns predict 52% and 28%, respectively, of the catalytic sites within a range $\Delta n = 2$. However only a fraction of 22% is predicted by both. Our conclusion is that each indicator has its specific predictive power, which should be exploited independently, while combined scores should be checked to gauge the confidence associated with multiple-indicator predictions.

Looking again at the example of HIV protease will make our point more clear (figure 3.8). It is not difficult to realize that a sequential inspection of the three separate indicator profiles at their respective optimal cutoff values is more likely to point to the known catalytic sites first. By inference, we propose that the same inspection sequence be adopted for hitherto unannotated proteins. The connectivity profiles should be examined first. These are the ones with the largest number of peaks, often coalescing to highlight extended regions. The search should be subsequently narrowed down with the corresponding closeness profile, typically featuring more localized peaks, albeit many of them likely to be orphan ones. The prediction should then be refined through the reduced stiffness patterns, the ones with the least number of peaks. Of course, extra information coming from other structure- and/or sequence-based algorithms should be used at each step in conjunction with our algorithm, if possible, to single out interesting sites.

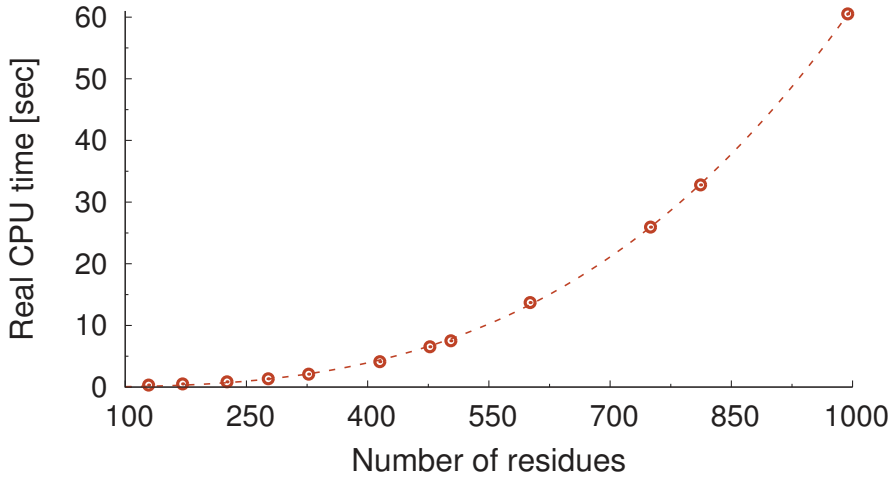



Figure 3.11.: CPU time required to compute reduced patterns for the three indicators illustrated in the text as a function of the number of amino acids in the enzymes. The computation refer to an ordinary desktop workstation equipped with an Intel(R) Xeon(R) CPU E5-1620 at 3.60 GHz. The dashed line is a fit with a cubic polynomial, $t = (N/N_0)^3$ , which gives $N_0 = 253.2$. As expected, the overall time is dominated by the time needed to diagonalize the Hessian matrix (operation which scales as the cube of the matrix dimension).

## 3.3. Conclusion and perspectives

In this chapter we have investigated the ability of different structure-related indicators to pinpoint the location of known catalytic sites in a large number of enzyme structures in the framework of the elastic network model. More precisely, we defined reduced peak patterns of (i) local connectivity, (ii) closeness centrality and (iii) structural stiffness, where the peaks retained along the protein sequence are assumed to flag potentially interesting sites. Our method is general and computationally inexpensive (see figure 3.11 for a benchmark test). Our analysis shows that all three considered indicators display a considerable predictive power (up to 50% of the catalytic sites recovered within a distance of two amino acids along the sequence), when the computed peak structures are compared with the location of annotated catalytic sites in a large database of enzymes (the Catalytic Site Atlas [144]). This suggests that the three indicators can be employed in some suitable combination/sequence to make predictions in unannotated enzyme structures.

In order to find the optimal procedure to combine the three indicators, we have investigated their behavior as a function of the cutoff $R_c$ used to construct the elastic networks, while monitoring in parallel the number of peaks per amino acid present in the indicator patterns. We have termed this procedure *cutoff lensing*. This analysis has revealed that

optimal values of the cutoff exist in all cases. For the connectivity, the fraction of recovered known catalytic sites trivially (and uninformatively) increases with the cutoff, as the number of high-connectivity peaks retained also increases. For this reason, we argue that the optimal cutoff corresponds to the least number of peaks per amino acid (about 40% of the catalytic sites recovered within a distance of two amino acids along the sequence), which means $R_c \approx 20$ Å. By contrast, somewhat surprisingly, centrality patterns display nearly cutoff-invariant predictive power. However, the specific number of peaks displays a minimum around $R_c = 28$ Å. Therefore, we conclude that $R_c = 28$ Å can be taken as the optimality condition, reflecting the idea that for equal fractions of recovered catalytic sites the most reliable prediction is the one made with the least number of peaks.

The study of reduced stiffness patterns has led us to uncover an interesting effect, that we termed cutoff lensing: when the cutoff is increased, the fraction of catalytic sites spotlighted by the stiffness peak patterns displays a maximum at around $R_c = 22$ Å. Remarkably, this is achieved with a minimum degree of redundancy, as the number of peaks in the patterns (pointing to potentially interesting sites) is a minimum for $R_c > 22$ Å, while at the same time the average number of peaks per catalytic sites is about 1 on average in this range of cutoff values. We conclude that $R_c \cong 22$ Å is the value of choice for predictions of catalytic sites made through stiffness patterns.

Remarkably, we found that the fraction of catalytic sites recovered by our indicators at the optimal cutoff is larger the larger the protein (see again figure 3.7). Connectivity patterns are an exception, as at the optimal cutoff the fraction of catalytic sites recovered is nearly the same independently of the size of the enzymes.

Of course, our algorithm can be combined to other structural and/or sequence based algorithms in order to find *hotspots* in proteins.

As a final observation, we note that our choice to attribute an equal weight to the three indicators in constructing the combined score $S_i$ in equation (3.4) is arbitrary. It would be interesting to inquire whether there exists an optimal combination of weights $W_I$ defining better generalized scores, namely

$$S_i = W_{\tilde{\chi}}\sigma_i^{\tilde{\chi}} + W_{\tilde{CC}}\sigma_i^{\tilde{CC}} + W_{\tilde{c}}\sigma_i^{\tilde{c}}, \tag{3.6}$$

with $\sum_{\tilde{I}} W_{\tilde{I}} = 1$.

For example, one may imagine to use standard optimization techniques [146, 147] or genetic algorithms [148] to efficiently determine an optimal set of weights, by training our algorithm on the CSA and other databases.

# 4. Pump simulations and bilocalized modes, a new understanding of protein communication

Proteins often perform their biological functions by transducing signals across their structures, initiated as a result of mechanical/chemical stimuli, such as ligand binding or ATP hydrolysis [149–152]. For example, many families of repressors have their affinities for DNA controlled by ligand binding at locations often very distant from the DNA-binding motifs [153–155]. Such phenomena are often referred to as intramolecular *communication*.

Communication can be studied using different approaches. For example, the study of correlated movements in molecular dynamics simulations has allowed to connect *hotspots* and structural pathways in proteins [156–158]. This kind of analysis generally proceeds by extracting a measurement from a molecular dynamics simulation such as mutual information or the matrix of position cross-correlations, to build a map of connections or construct a weighted network based on the specific indicator [73, 89, 159–162]. Combining graph theory and molecular dynamics simulations has allowed researchers to advance in the understanding of structural and dynamical determinants of intramolecular communication. However, even though structure and function are related, the need of molecular dynamics is generally necessary to get insight into the role of topology for the prediction of functional sites. Purely structural analyses can be performed to predict *hotspots* in proteins [163–165] but it is still hard to reconstruct the map of communication pathways involving several residues. Prediction of active sites and communication pathways through purely structure-based methods is generally restricted to NMA-based methods [56, 116, 166].

Using single-bead kick within a non-linear network model, previous authors remarked that energy is preferentially stored in stiff parts of proteins and is typically carried by a set of a few high-frequency normal modes [38, 57, 60, 167, 168]. Energy transfer following local impulse-type initial conditions is an instructive approach to work on communication for several reasons, the main one being that a structural communication process is often initiated by a ligand interacting with the protein and creating a mechanical perturbation

in a very specific region of the protein scaffold. Proteins are complex systems and the energy follows complicated patterns of redistribution upon such local perturbations. It has already been demonstrated that energy flows can map out pathways connecting important protein regions [74, 169–171].

In this chapter, instead of performing single-bead kick simulations to inject energy impulsively into the system, we apply an oscillating force to particular residues along specific directions with the same spirit as the frequency resolved pump-probe simulations described by Sharp and Skinner [67]. We typically observed that, after a transient, energy spreads out into the molecule following particular structure-encoded selection rules. In the following, we will present these selection rules and show how they allow the communication between amino-acids even at distances higher than 70 Å. One important result is that we can predict how the system will evolve during a pump simulation by a simple structural analysis.

The first section describes the protocol of pump simulations. As an example, we will then move to illustrating the methods and the main results on a test system, kinesin. In the following section, we will identify and discuss the typical structure-encoded selection rules. This work has been conducted on a database of 1711 protein structures, comprising 1177 scaffolds preferentially featuring $\alpha$-helix motifs and 534 structures where $\beta$-sheets were the dominant secondary structure elements. Finally, we will report on a pump analysis performed on a set of 15 G-protein-coupled receptors (GPCR), which will allow us to surmise concerning a possible mechanism of signal transduction across the bilayer cell membrane.

## 4.1. Pump protocol and illustration on kinesin

Our idea here is to apply an oscillatory force on a particular residue along a specific direction with a given frequency and study how the injected energy propagates across the protein structure. For this analysis we will use the coarse grained elastic network model (see equation 2.14 in chapter 2) with fixed residue mass $M = 110$ Da (*i.e.* the average amino-acid mass), $R_c = 10$ Å and $k_2 = 5$ kcal mol$^{-1}$ Å$^{-2}$.

Recalling equation 2.14, all our simulations were performed with $F_0 = 0.5$ kcal mol$^{-1}$ Å$^{-1}$ and $\gamma = 8 \times 10^{-3}$ ps$^{-1}$. The frequencies and directions of pumping were chosen so as to excite locally a given normal mode. More precisely, we systematically chose $\omega_0 = \omega_k$ and $\hat{e}_n = \boldsymbol{\xi}_n^k / ||\boldsymbol{\xi}_n^k||$ where $k$ specifies the chosen normal mode. The guiding principle is to investigate whether the specific, fold-rooted pattern of vibration embodied in certain normal modes allows the injected energy to be efficiently and directionally channeled towards specific locations, possibly far away from the pumped region.

A typical pumping simulation starts will all particles at their equilibrium position and with zero velocity. At time $t = 0$, the sinusoidal force is applied to a residue and the system starts evolving in time. All the residues start to oscillate and the system quickly reaches a steady state due to the presence of the weak distributed damping force.

It is interesting to compute the average energy of each residue $i$ once the steady state is attained. Typically, we will compute normalized asymptotic energies, defined as

$$\epsilon_i = \frac{\langle E_i \rangle}{\sum_m \langle E_m \rangle}. \tag{4.1}$$

where $\langle E_i \rangle$ is the average local total energy averaged over an extended lapse of time into the stationary state (typically from $t = 180$ ps to $t = 200$ ps, corresponding to 50000 time steps).

In our study, we deal with frequencies belonging to the linear spectrum of normal modes. For each mode, we define the local amplitude of vibration

$$\chi_i^k = \sum_{\alpha=x,y,z} |\boldsymbol{\xi}_{i\alpha}^k|^2. \tag{4.2}$$

where we have assumed normalized mode patterns, so that $\sum_i \chi_i = 1$.

Interestingly, as a first general observation, we remarked that, provided a residue has a sizeable $\chi_i$ measure for a particular mode $\tilde{k}$, then if we force this residue to oscillate at the frequency $\omega_{\tilde{k}}$, the energy of the protein is preferentially transferred to the residues which are most represented in the displacement pattern $\chi^{\tilde{k}}$.

## An example, kinesin

Kinesin is a molecular motor which has the ability to *walk* on microtubules [172, 173]. A previous study has shown that using single-bead kick dynamics on a nonlinear network, one can observe an energy transfer between a residue close to the pocket involved in the ATP hydrolysis (MET 96) and residues in the neck linker docking site (CYS 296 and THR 298) [60].

Curiously, an accurate inspection of the normal mode spectrum shows that a particular normal mode ($k = 1062$) of kinesin connects MET 96 and THR 298. Applying a periodic force to the first residue at its frequency $\omega_k$ confirms that these peculiar locations are connected and we remark that the normalized energy pattern for each residue once we reach the steady state follows to a large extent the pattern of $\chi^k$ for the mode $k$ (figure 4.1). Looking at panel (b) of figure 4.1 (right), we remark that different frequencies are excited during the simulation. There is a fast oscillation which corresponds to the pump frequency

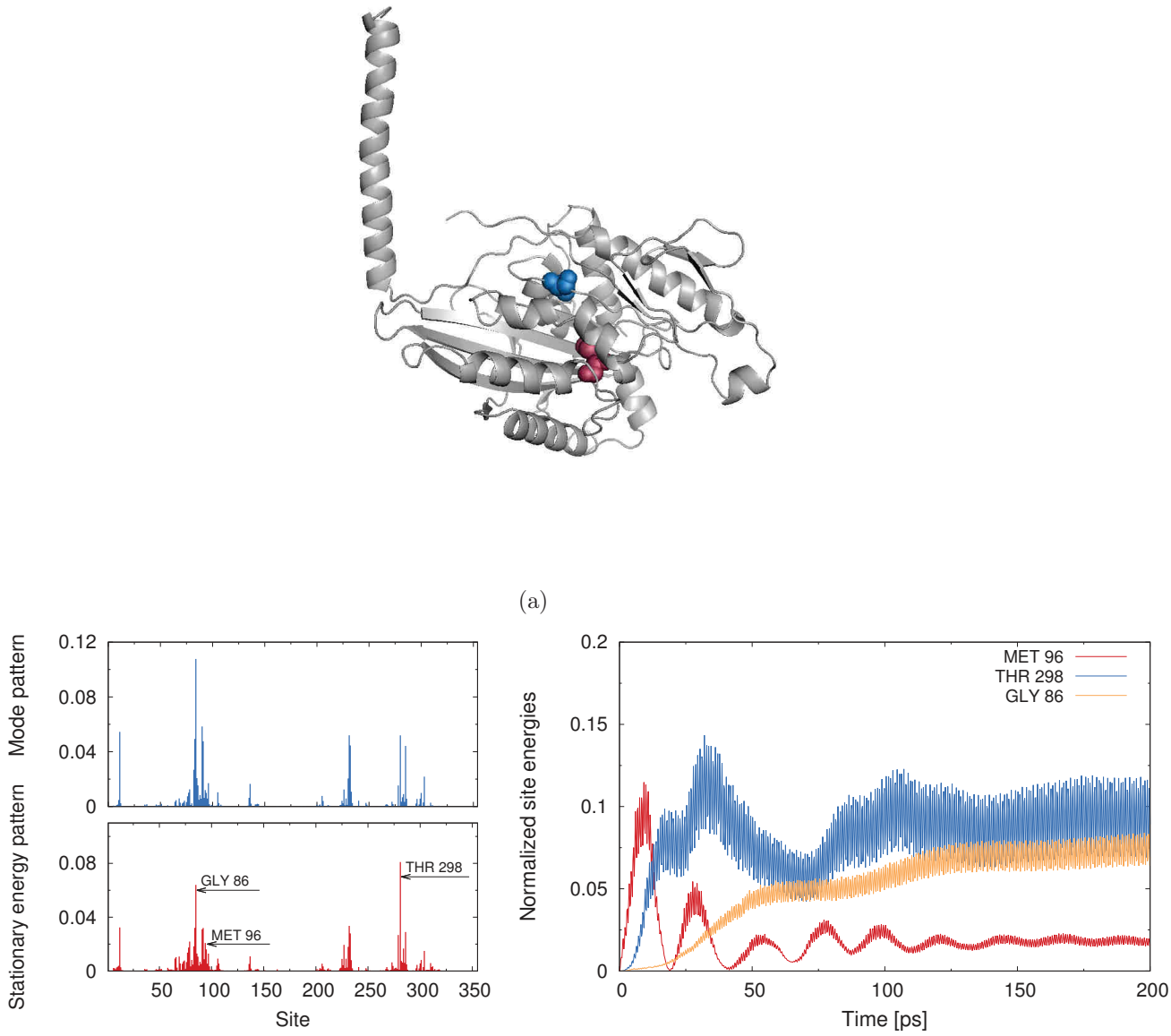



Figure 4.1.: (a) Structure of kinesin (PDB 3KIN [5]). Blue spheres represent the amino-acid involved in the walk on microtubules (THR 298), red spheres highlight the residue near the ATP pocket (MET 96). (b) Pumping analysis of kinesin (PDB 3KIN). Top left, pattern of $\chi_i^k$ for $k = 1062$. Bottom left, average normalized energy of the steady state for each bead when we force the residue MET 96 at the frequency $\omega_k = 92.6$ cm$^{-1}$ for $k = 1062$. Right, local total energy of residues versus time for the same pumping simulation of residue MET 96. Red shows the total energy of MET 96. Blue and yellow highlight the residues that have the highest energy in the steady state *i.e.* THR 298 and GLY 86 respectively.

but slower frequencies are also excited. A Fourier transform helps separate the slow frequency component (figure 4.2). This slow frequency can be simply rationalized through

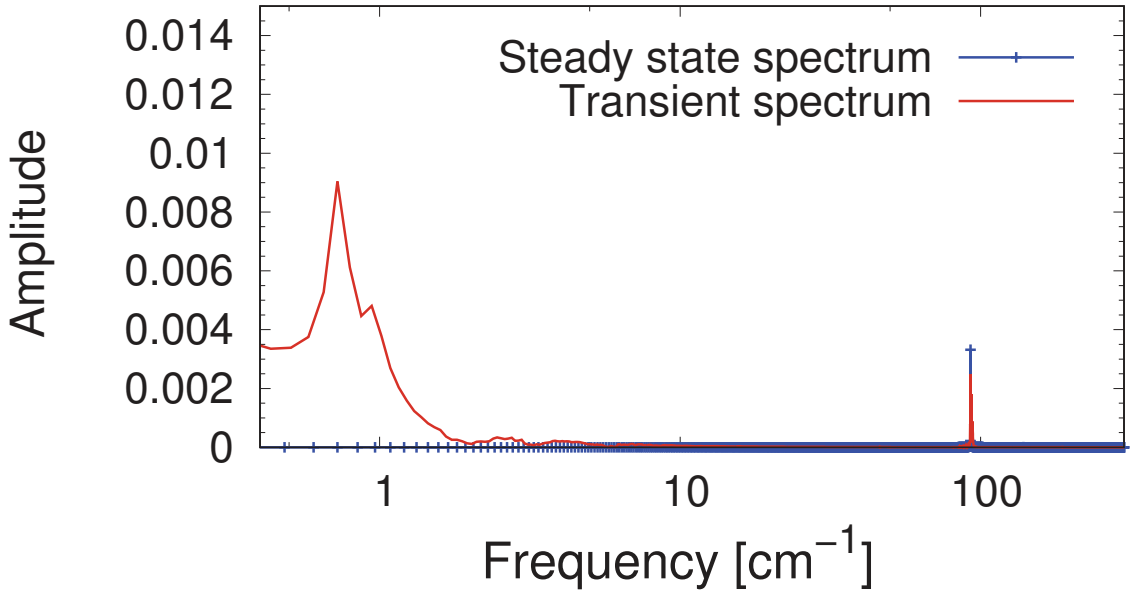



Figure 4.2.: Fourier transform amplitude of the displacement time series of MET 96 (figure 4.1) during the transient (red curve) and during the steady state (blue curve). Once the steady state is reached, the only frequency present in the spectrum of the forced particle is the pump frequency, $\omega_k = 92.6$ cm$^{-1}$.

a simple selection rule. In a typical pumping simulation, there is another frequency which is an important piece of the puzzle beyond the pumping frequency. This is the frequency of the mode $k$ such that $\chi_n^k$ is the largest for $k = 7, 8 \ldots 3N$ at the pumped residue $n$. We refer to such specific frequency in the following as the *resonance frequency*, $\omega_n^R$. The slow frequency component $\omega_{\text{slow}}$ that one observes during the pumping simulation illustrated in figure 4.1 is a consequence of the frequency mismatch between the local pumping and resonance frequency. Namely,

$$\omega_{\text{slow}} = \omega_n^R - \omega_0. \tag{4.3}$$

However, such frequency occurs only during the transient *i.e.* before the steady state is attained. Indeed, once we reach the steady state, a Fourier transform performed in the non-equilibrium state clearly shows that the only frequency present in the spectrum of the forced residue is the pump frequency (figure 4.2). In general, the phenomenology can be complex during the transient state before the steady-state sets in, notably with the concerted presence of several frequency components. However, the general robust finding is that when we force a residue at the frequency of a given mode $k$, the energy spreads out among the residues that are well represented in the pattern of this mode. Figure 4.1 illustrates the typical steady-state energy pattern.

However, localized energy transfer does not occur if we force a residue which has a

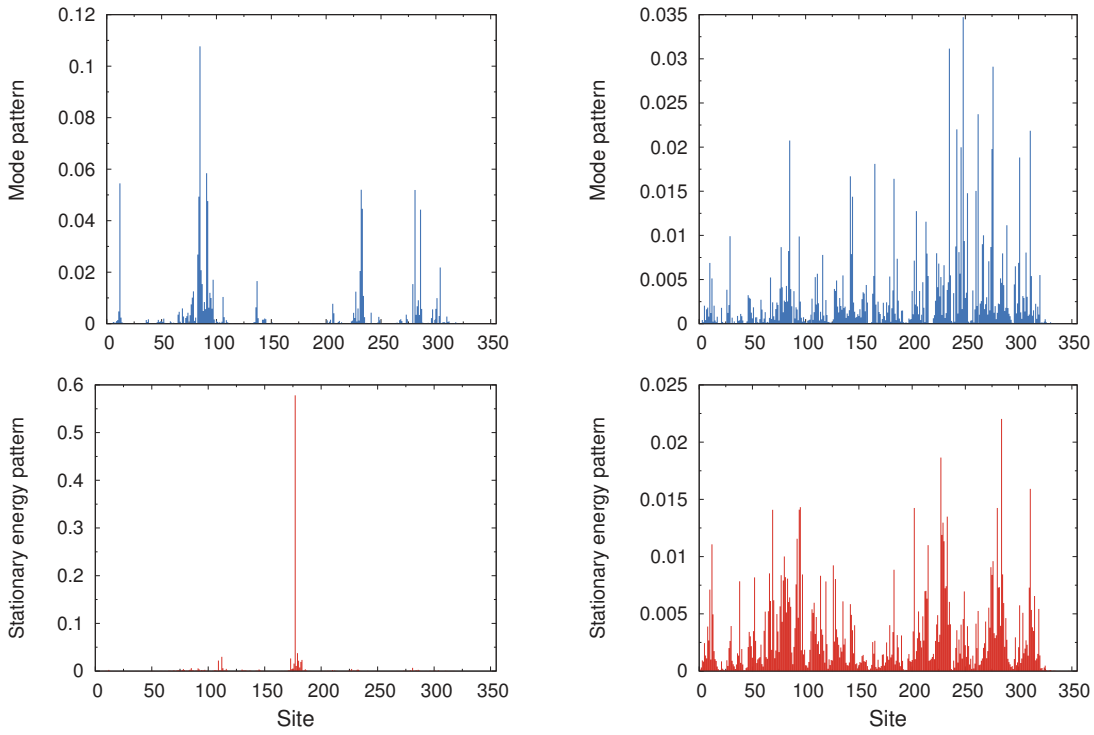



Figure 4.3.: (a) Pumping analysis of kinesin (PDB 3KIN). Top, pattern of $\chi_i^k$ for $k = 1062$. Bottom, average normalized energy in the steady state for each bead when we force the residue GLU 178 at the frequency $\omega_k = 92.6$ cm$^{-1}$ for $k = 1062$. (b) Top, pattern of $\chi_i^k$ for $k = 1002$. Bottom, average normalized energy in the steady state for each bead when we force the residue MET 96 at the frequency $\omega_k = 80.7$ cm$^{-1}$ for $k = 1002$. We observe that for both numerical experiments, no energy transfers occur.

negligible contribution to the normal mode along which we locally pump or if too many residues contribute to the mode. This difference is illustrated in figure 4.3 where we have compared the stationary energy pattern and the mode displacement pattern for two pump simulations. In the first numerical experiment (part (a) of figure 4.3), we have forced the residue GLU 178 at frequency $\omega_k = 92.6$ cm$^{-1}$ ($k = 1062$). This amino-acid has a negligible displacement in the mode pattern (upper panel). As a result, the energy injected in the protein stays on the pumped residue and we do not observe any energy transfer. For the second numerical experiment (part (b) of figure 4.3), we have forced the residue MET 96 at frequency $\omega_k = 80.7$ cm$^{-1}$ ($k = 1002$). The mode $k = 1002$ has the particularity to be delocalized on many residues, each contributing a small displacement to the mode pattern. As a consequence, the stationary energy pattern is *noisy* and we clearly do not observe any *localized* energy transfer.

## 4.2. Bilocalized modes and resonant transfer

The kinesin case study has taught us that dispersion of energy during a pump simulation follows frequency selection rules, arising from the mismatch between the pump and local *resonance* frequencies. Moreover, a general tendency exists for the stationary energy pattern to follow the normal mode displacement pattern, the more so the more we are forcing at a residue whose local vibrational amplitude is sizeable for such normal mode. In the quest for long-range energy transfer, it is thus natural to look for normal modes whose pattern would be strongly localized but at the same time reaching out to rather distant locations. In order to scout out for such modes automatically, we set up an automatic high-pass filtering procedure. More precisely, for each mode $k$, we calculate the contribution of each residue to atomic fluctuation $\chi_i^k$ and the corresponding standard deviation $\sigma_k$,

$$\sigma_k = \sqrt{\langle {\chi^k}^2 \rangle - \langle \chi^k \rangle^2}. \tag{4.4}$$

We then introduce a new filtered pattern $\tilde{\chi}_i^k$ defined as

$$\begin{cases} \tilde{\chi}_i^k = \chi_i^k \text{ for } \chi_i^k \geq 3\sigma_k \\ \tilde{\chi}_i^k = 0 \text{ for } \chi_i^k < 3\sigma_k \end{cases} \tag{4.5}$$

For a given protein, we compute the number of times that $\tilde{\chi}_i^k$ is greater than zero and this gives for each mode $k$ a number of peaks $N_p^k$. Moreover, since the patterns $\chi_i^k$ of normal modes are normalized ($\sum_i \chi_i^k = 1, \forall k$) the overall sum of the peaks retained in the filtered patterns carry an information on the degree of localization of the mode. Since we want to single out modes which are strongly localized, we concentrate our study on those modes that exhibit at most four peaks in the filtered pattern. Since we are interested in energy transfer, we also obviously require that the number of peaks in the filtered pattern be at least two. For the modes singled out through this procedure, we isolate the two residues whose peaks are the highest and the second-highest in the pattern. This procedure is illustrated in figure 4.4 for a protein of choice.

Our analysis was performed on a database of 1711 proteins, for which we have selected specific bilocalized modes for each protein through the procedure described above.

### 4.2.1. Large distances between localized displacements in low-frequency modes

Since slow modes generally represent collective motions, whereas fast modes are more localized [56], one would naively expect that the potentially interesting modes belong

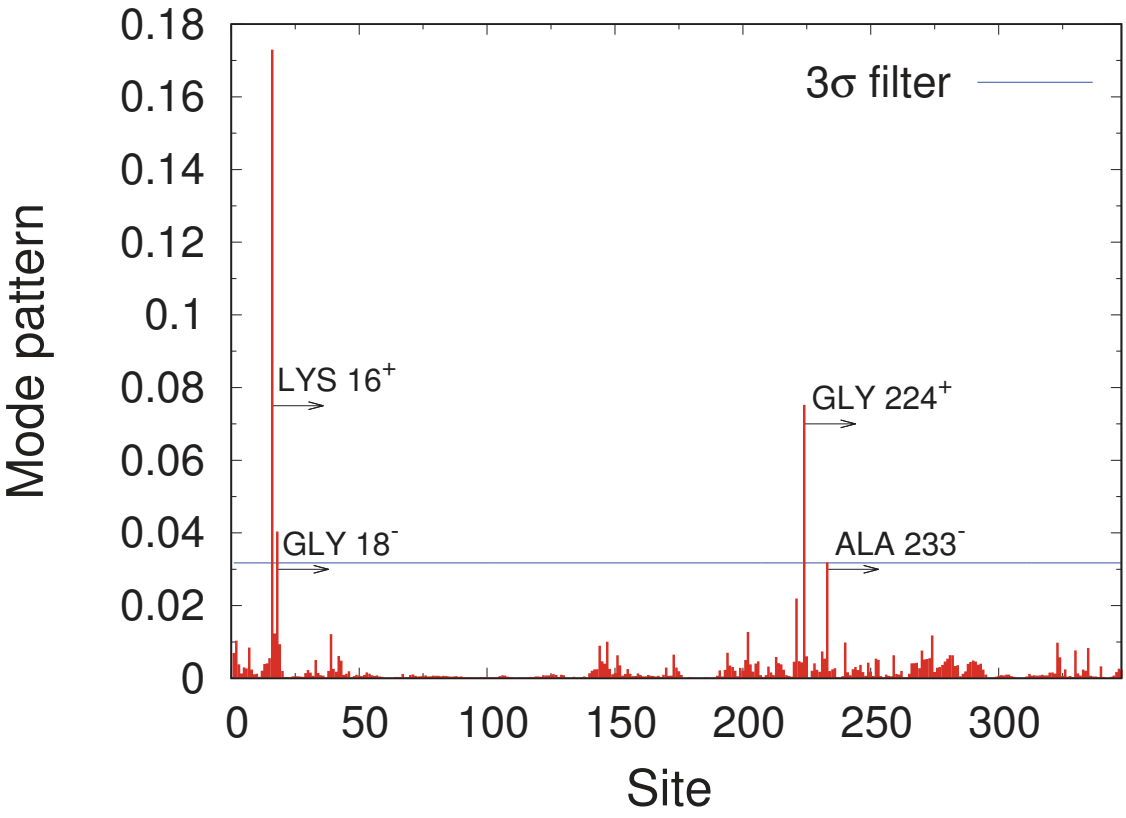



Figure 4.4.: Red, displacement pattern of the mode $k = 998$ for Bovine Rhodopsin (PDB 1U19 [8]). Blue line : $3\sigma$ filter. We remark that 4 residues have an amplitude greater than the filter (residues LYS 16, GLY 18, GLY 224 and ALA 233). However, our study will focus on the two that have the highest amplitude (with the index $+$) *i.e.* residues LYS 16 and GLY 224 (see text).

to the high-frequency region as the filter we used by definition isolates localized modes. However, it turned out that the emerging picture is much subtler. Our modes have the particularity of connecting two distinct residues or small groups of residues. When computed over a large database of structures, the distances between the two highest peaks in the displacement patterns of the filtered modes appear to be distributed over a broad range, up to 180 Å. We grouped the modes by inter-peak distances in slices 10 Å wide and let all the modes connecting two residues at distance larger than 50 Å apart in the same slice. Figure 4.5 illustrates how these modes are distributed as to their frequency. It is apparent that, whereas modes connecting close residues spread along almost all the spectrum, when the peak interdistance increases the frequencies narrow down to between 10 and 20 $cm^{-1}$. This represents a rather unstudied category of modes. People focused their attention on slow modes because of their ability to reconstruct observed large-scale conformation changes [174,175]. Moreover, hinge regions flagging the nodes of large-scale motions spotlight catalytic residues [116, 176]. Alternatively, great

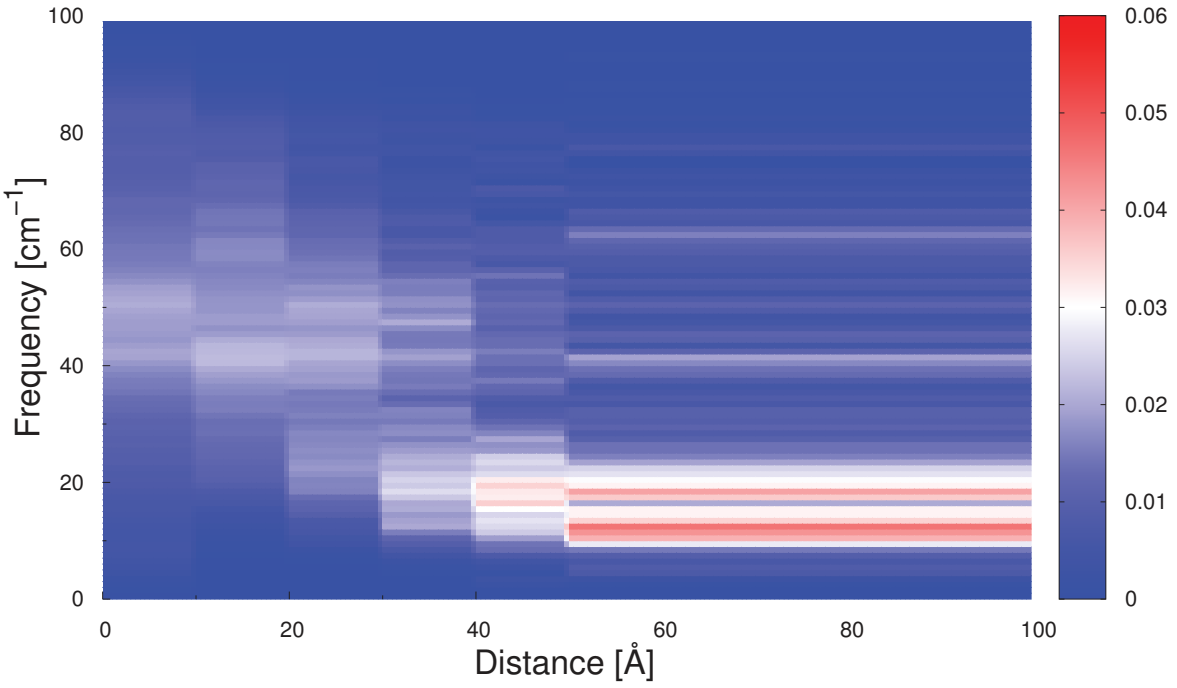



Figure 4.5.: Distribution of frequencies of the filtered modes as a function of the distance between the two highest peaks in the displacement pattern. This calculation has been performed on the entire pool of 1711 structures.

attention has been devoted to high-frequency normal modes because residues in these modes seem to be able to act as energy collecting centers [60, 168]. Here, we are dealing with a somewhat intermediate and relatively unexplored category of modes, low-frequency bilocalized modes.

### 4.2.2. Pump simulations

In order to probe the energy transfer ability associated with these modes, we performed pump simulations on the same protein database. For each structure, we selected from the pool of filtered modes the one that connects the two most separated residues. We performed then systematically two simulations, one where the forced residue is the one having the highest amplitude in the displacements pattern (we term this residue $i_1$) and the second one where the forced residue has the second highest amplitude (we term it $i_2$).

We applied the same oscillating force at the mode frequency $\omega_k$ for the two simulations and monitored the energy spreading process as a stationary energy profile sets in as a result of the combined action of the forcing and weak damping (see chapter 2). Since we chose the modes such that only few residues are represented, we expected that for each

simulation, the residues $i_1$ and $i_2$ would accumulate most of the total energy as compared to the others. Figure 4.6 confirms the expectations. Indeed, the distribution of energies of the pumped residues peaks at higher values than that of the other residues, while the distribution of the energies of the targeted residues is clearly separated from the others. The energy is injected in the system at the forced residue, it is then natural that pumped residues carry generally most of the total energy. Looking carefully to figure 4.6, we remark that there is a clear shift to higher energies for the population of the targeted residues when the pumped site is $i_2$. Energy transfers from $i_2$ to $i_1$ seem more efficient than from $i_1$ to $i_2$. We recall that residues of kind $i_1$ are those with the highest amplitude in the static analysis $\chi_i^k$. This confirms our hypothesis that the stationary energy patterns are governed by the *static* profiles of the excited modes. A higher amplitude implies a higher accumulation of energy. Hence, energy transfer mediated by bilocalized modes identifies a privileged direction of energy flow, $i_2$ to $i_1$ being more efficient than $i_1$ to $i_2$. Biological processes involving intra-molecular communication do involve directionality. For example, the binding of a ligand to a repressor can strongly decrease its affinity for DNA, thus reinstating transcription. In this case, there is clearly a *flow of information* between the amino-acids involved in the binding pocket and the amino-acids that bind DNA on the other end of the structure.

The Tet Repressor (TetR) binds to DNA in bacteria hindering the transcription leading to the synthesis of an efflux pump, TetA. However, when the antibiotic tetracycline is present in the bacterial cell, it binds to TetR, with the consequence of lowering drastically its affinity for DNA. Therefore, the efflux pump protein TetA can be synthesized, providing an efficient means for getting rid of the antibiotic [153]. The structure of TetR has been isolated in the apo state and also with many different ligands bound. The binding site of tetracycline with TetR involves principally two residues, HIS 64 and HIS 100, while TetR binds to DNA with a H-T-H (Helix-Turn-Helix) motif starting at residue 26 and ending at residue 45. All the residues involved in the binding pocket and in the H-T-H motif are highly conserved across the seven classes of TetR (A to E, G, and H) [153]. The apo TetR (PDB 1A6I [177]) and the protein with tetracycline bound (PDB 2TRT [9]) were analyzed with the standard pumping procedure described above. First, a high-pass filtering procedure has allowed us to identify the potentially interesting bilocalized modes and each of them has been analyzed. Interestingly, while no mode selected from the filtering procedure connects residues HIS 64 or HIS 100 to the DNA binding region in the apo protein, one mode ($k = 123$) of frequency $\omega_k = 32.03$ cm$^{-1}$ connects HIS 64 (ligand-binding pocket) to GLY 35 (H-T-H motif) in the conformation with tetracycline bound. The mode pattern corresponding to this frequency shows that the amplitude of GLY 35 is the highest and HIS 64 is the second highest. Pump simulations on residue HIS

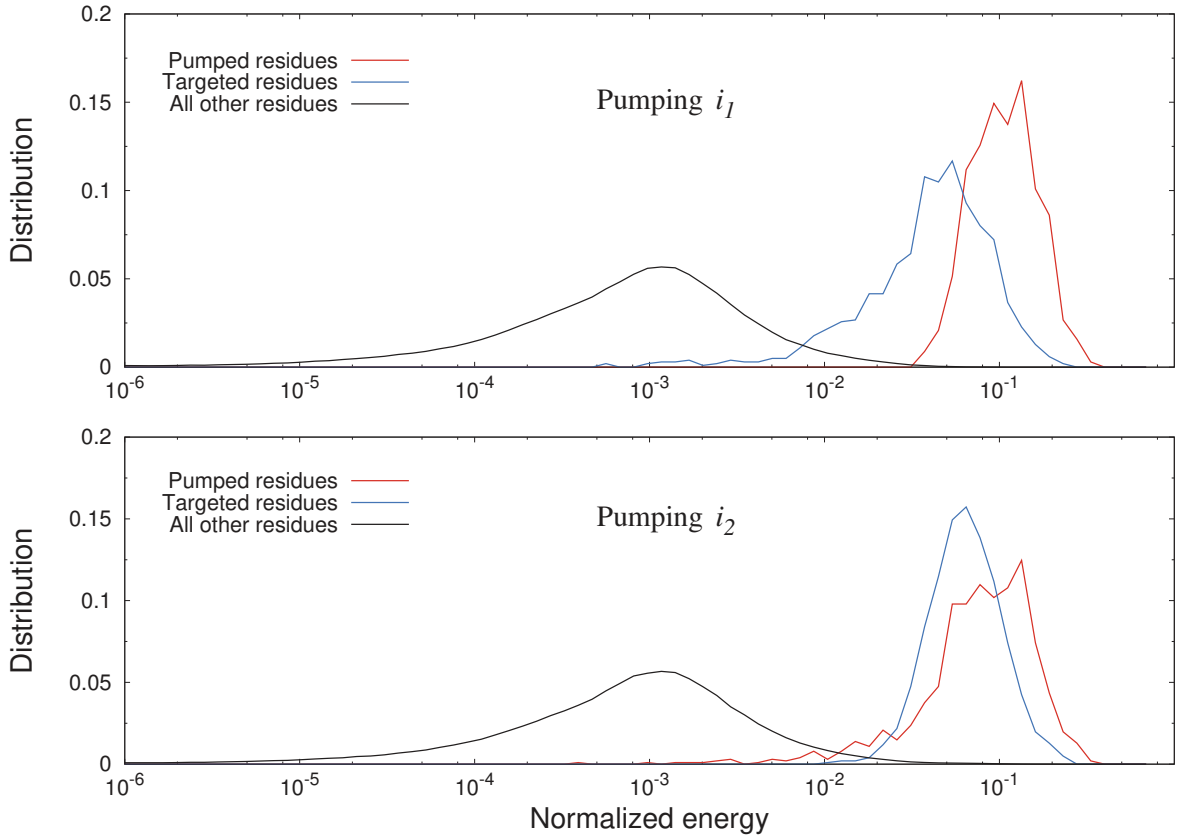
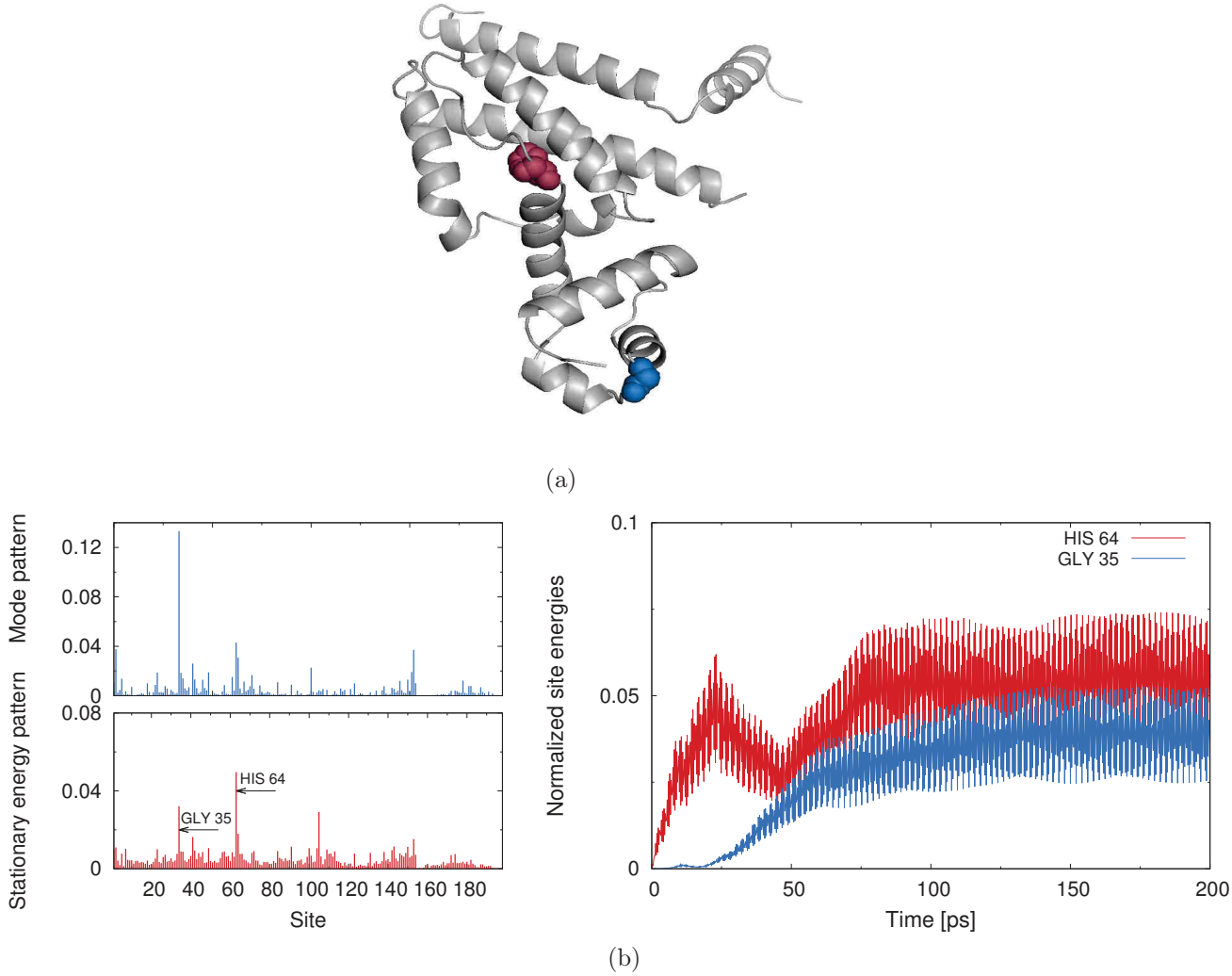



Figure 4.6.: Histograms of the average site energies in the stationary state. All average energies for a given protein are normalized to a total unit energy. Red, forced residues, blue, targeted residues, black, all other residues. The upper panel represents the simulations where the pumped residues was the residue $i_1$ (highest amplitude in the mode) while the lower panel represents the simulations where the pumped residue was the residue $i_2$ (second highest amplitude in the mode). Note the logarithmic scale on the $x$ axis.

64 at frequency $\omega_k = 32.03$ cm$^{-1}$ confirm that energy transfer between these two residues does occur, connecting a residue involved in the binding to tetracycline to a residue that belongs to the DNA binding motif (figure 4.7).

### 4.2.3. Selection rules

Energy seems to accumulate across a given forced structure following the specific displacement pattern of the forced *bilocalized* mode. We have learnt so far that the more a given residue is represented in the displacement pattern of the mode, the more it will act as an energy-accumulation center. By pursuing this line of reasoning further, one would be tempted to look for modes where the two distant residues correspond to amplitudes gathering most of the overall normal mode displacement pattern (normalized). This *low-noise* situation should lead to efficient energy transfer. However, we recognized

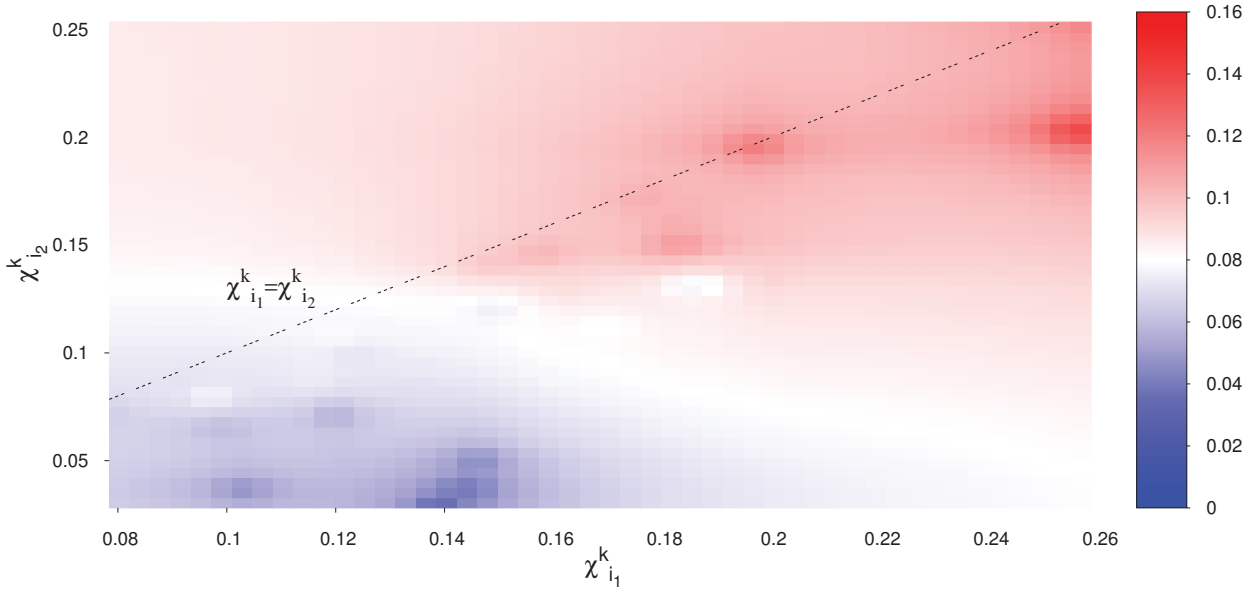



(a)

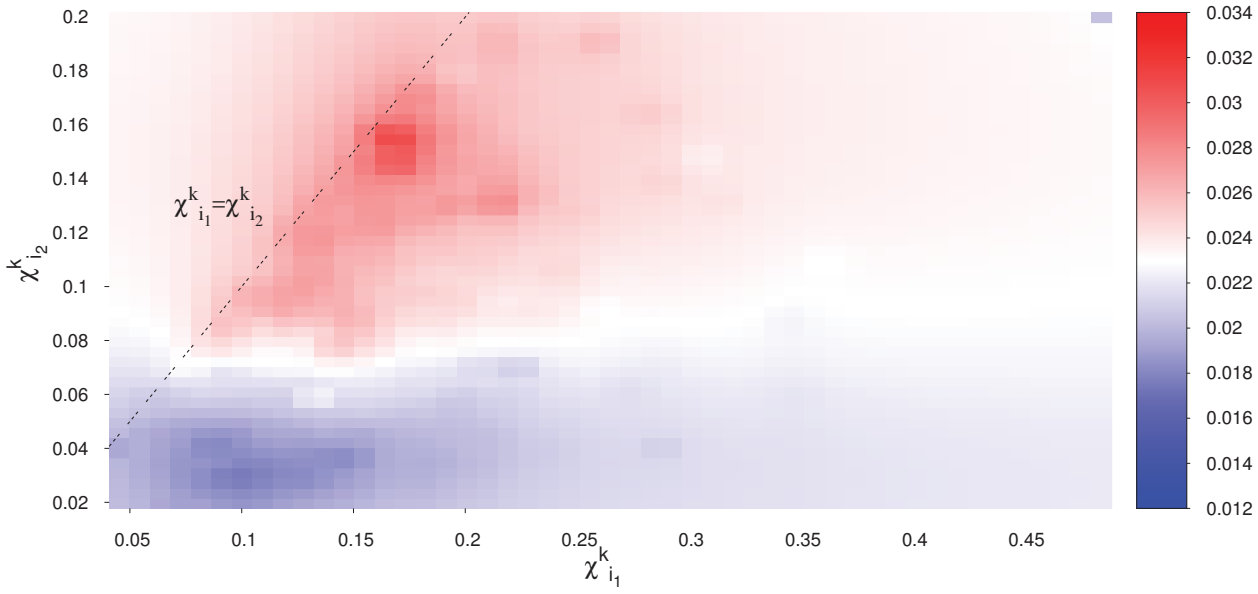
(b)

Figure 4.7.: (a) Structure of TetR (PDB 2TRT [9]). The red spheres represent the amino-acid involved in the binding of tetracycline (HIS 64), while blue spheres highlight the residue in the binding DNA motif (GLY 35) a distance of 31.8 Å away. (b) Pumping analysis of TetR (PDB 2TRT). Top left, pattern of $\chi_i^k$ for $k = 123$. Bottom left, average normalized energy of the steady state for each bead when we force the residue HIS 64 at the frequency $\omega_k = 32.03$ cm$^{-1}$. Right, local total energy of residues versus time. Red shows the total energy of HIS 64 and blue highlights the total energy of GLY 35.

that excitation of one of the two sites for such modes do not result in particularly efficient energy transfers. Rather, in this case the whole process is regulated by the distance between the two centers. We observed that if the mode is too localized and the residues are too far from each other, no energy transfer occurs. To observe a long-range energy transfer, some *noise* is required in the mode pattern, that is, a non-negligible fraction of overall displacement pattern involving residues other than the two localization centers.



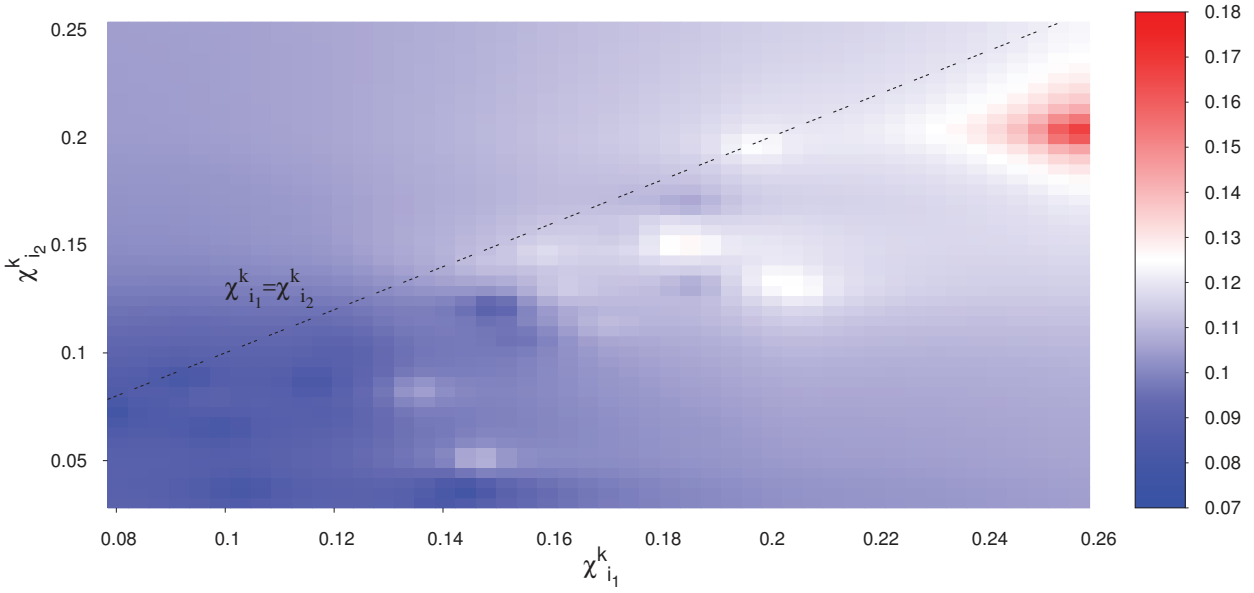

(a)



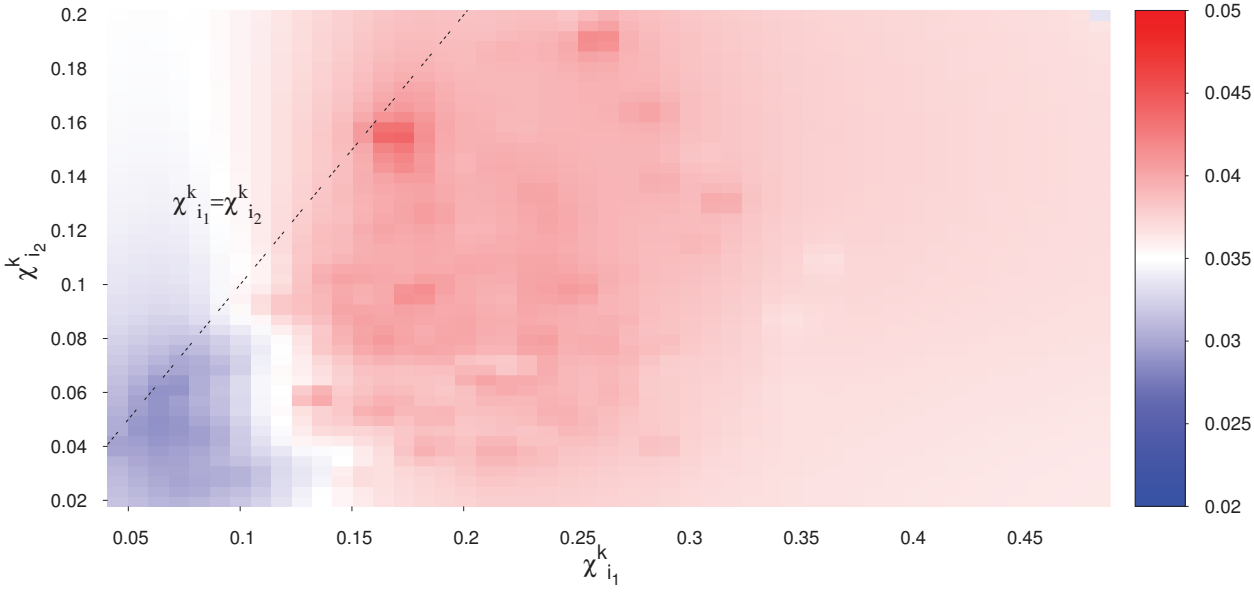

(b)

Figure 4.8.: Pumping $i_1$. Density plot of the average normalized energy of the target residue $(i_2)$ in the $(\chi^k_{i_1}, \chi^k_{i_2})$ plane. (a) $\mathrm{d}(i_1, i_2) < 20$ Å, (b) $\mathrm{d}(i_1, i_2) \geq 20$ Å.

In our elastic network, each residue interacts with other residues within a radius of 10 Å. If we want to transfer energy between two residues far from each other, the need of one or more intermediates is indispensable. Therefore, if two residues are too represented in a mode, this means that other residues cannot form an energy transduction pathway.




(a)




(b)

Figure 4.9.: Pumping $i_2$. Density plot of the average normalized energy of the target residue ($i_1$) in the $(\chi^k_{i_1}, \chi^k_{i_2})$ plane. (a) $\mathrm{d}(i_1, i_2) < 20$ Å, (b) $\mathrm{d}(i_1, i_2) \geq 20$ Å.

Since residues of kind $i_1$ by definition have a greater amplitude than those of kind $i_2$ ($\chi^k_{i_1} > \chi^k_{i_2}$), energy transfers of the kind $i_1 \rightarrow i_2$ seem to be more difficult. However, if $\chi^k_{i_2}$ is beyond a critical threshold of about 0.15, seemingly irrespective of $\chi^k_{i_1}$ at short $i_1 - i_2$ separations some transfer efficiency is recovered (figure 4.8 (a)). This is because for

short interdistances, the possible intermediate sources of *noise* are reduced. As a result, the transfer process depends critically (threshold effect) on the contribution of the target residue to the displacement pattern of the mode under scrutiny ($\chi^k_{i_2}$).

The emerging picture for transfers involving modes with inter-peak distances greater than 20 Å is rather intriguing (figure 4.8 (b)). In this case, the need for an adequate number of intermediate energy-passing residues is more critical. This implies that the displacements of the pumped and target residues in a mode $k$ need to be balanced with all the other residues. Indeed, there appears to exist an optimum area around the characteristic $\chi^k_{i_1} = \chi^k_{i_2}$, where transfers are more efficient. This region is defined by the requirement that the displacements of the two residues in the mode be not the highest (that would imply no intermediate, hence no transfer) but also not too small (*noise* still needs to be limited to allow for a well-focused transfer).

The picture emerging from the simulations where residues of the kind $i_2$ are forced is somewhat different. For short inter-peak distances, the higher the displacement of the target residues (here $i_1$), the more efficient the energy transfer. This confirms the previous hypothesis that energy accumulates to residues that have a high contribution to the displacements of the forced mode. Moreover, transfers are also more efficient as $\chi^k_{i_2}$ increases. This is rather logical and is a direct consequence of the relation $\chi^k_{i_1} > \chi^k_{i_2}$ (high $\chi^k_{i_2}$ involves high $\chi^k_{i_1}$) and of the limited number of intermediates for short inter-peak distances. Figure 4.9 (a) displays in a clear fashion this behavior and one can observe an optimum for $\chi^k_{i_1} \sim \chi^k_{i_2} \sim 0.2 \div 0.25$. In the short-distance case, $\chi^k_{i_2}$ does not play a crucial role, except counterbalancing the *noise*. The larger $\chi^k_{i_2}$, the lower the contribution of the other residues to the mode displacement field and as a consequence the more efficient the transfer. Finally, long-range energy transfers where residues of the kind $i_2$ are forced confirm the hypothesis that $\chi^k_{i_2}$ plays a *secondary role*. One can observe that transfers are efficient at intermediate values of $\chi^k_{i_1}$ *i.e.* $\chi^k_{i_1} \sim 0.15 \div 0.30$ irrespective of $\chi^k_{i_2}$ (figure 4.9 (b)).

Overall, the conclusion is that long-range energy transfer is promoted by the presence of *noise* in the displacement pattern, that is, a non-negligible contribution from the other residues besides the peaks of the pumped and the target centers. More precisely, the most efficient transfers are observed at intermediate values of the target residue contribution to the displacement patterns. Moreover, one can observe that in general transfers of the kind $i_2 \rightarrow i_1$ are more efficient than those of the kind $i_1 \rightarrow i_2$, as the energies of the target residues are generally higher when the target is $i_1$. This feature has already been spotlighted in figure 4.6.

### 4.2.4. Application to signal transduction in the GPCR family

G-Protein Coupled Receptors (GPCRs) are a family of membrane proteins that have the ability to transduct chemical information across the cell membrane and are an important area of research for drug-design [159, 178, 179]. When a ligand binds to their extracellular domain, allosteric response through the membrane allows to release the G-protein at the juxta-membrane interface in the cell interior, which in turn generates a biochemical signaling cascade. This phenomenon implies long-range intramolecular *communication*, as some kind of information has to cross the cell membrane, corresponding to a thickness of more than 40 Å. Since our pump simulation scheme has shown the emergence of efficient long-range energy transfers, it is natural to inquire whether some light can be shed on allosteric communication in GPCRs through our method.

A signal transduction process starts with the binding of a ligand to the extracellular part of GPCR. Once the ligand binds, a mechanistic allosteric response enables the separation of the subunit-$\alpha$ of G-protein that is bound to GPCR at the proximal membrane region in the cytoplasm. The subunit alpha of G-protein is linked to the GPCR at three different places : between helix 3 and helix 4, between helix 5 and helix 6 and at the end of helix 7 (see figure 4.10). Potentially interesting modes in signal transduction should involve two residues, at opposite sides of the cell membrane, *i.e.* one residue in the extracellular part and the other in the intracellular part, with a preference for the interface region between subunit-$\alpha$ and the GPCR in the intracellular region. Applying the same high-pass filtering procedure detailed in section 4.2. to fifteen different GPCRs (see table 4.1), we have been able to identify for each protein a particular mode that connects two residues at either end of the membrane. We performed systematically two pump simulations for each protein at the frequency of the specific mode $k$ connecting the two residues across the membrane. A first simulation where the forced residue is the extracellular one and the second where we force the intracellular one. Out of the fifteen proteins, we have been able to observe a clear energy transfer between the two sides of the membrane in fourteen proteins.

The results are collected in table 4.2, 4.3 and 4.4, where we have separated the transfers of the kind *extracellular* $\rightarrow$ *intracellular* (table 4.2), *intracellular* $\rightarrow$ *extracellular* (table 4.3) and no transfers at all (table 4.4). For the cases where energy transfers do occur, we have computed two indicators to gauge the quality of transfers, the energy ratio between the target and the pumped residues in the steady-state (*efficiency* of transfer) and the energy ratio between the target and all other residues except the pumped one

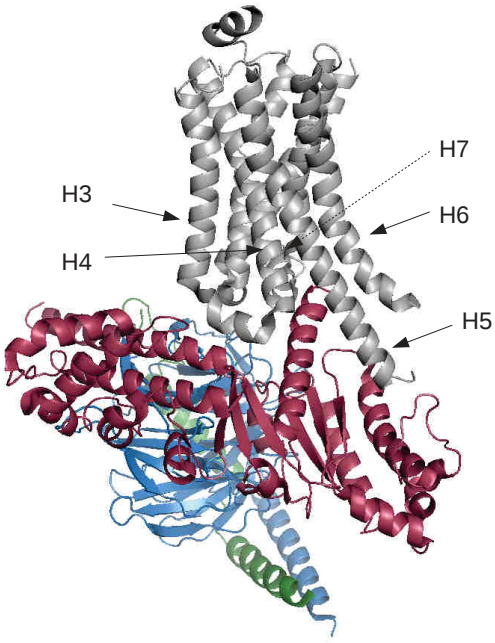



Figure 4.10.: GPCR coupled to the different subunits of the G-protein. The color code helps identify the GPCR and each subunit. Gray stands for the GPCR, red represents the $\alpha$-subunit, blue the $\beta$-subunit and green the $\gamma$-subunit. We observe that the GPCR is linked to the subunit-$\alpha$ at three different places : between helix 3 and helix 4, between helix 5 and helix 6 and at the end of helix 7.

(*directionality* of transfer). We have thus

$$\text{Efficiency} = \frac{\langle E_{\text{target}} \rangle}{\langle E_{\text{pump}} \rangle} \tag{4.6}$$

and

$$\text{Directionality} = \frac{\langle E_{\text{target}} \rangle}{\frac{1}{N-2} \sum_{i \neq (\text{pump,target})} \langle E_i \rangle}. \tag{4.7}$$

Again, the averages refer to a 20 ps window well into the steady state.

The tables also specify whether the residues are outside the cell (Ext. cell.) or inside the cell (Int. cell.), the contribution to the displacement pattern of the forced mode of the pump ($\chi_{i_p}$) and the target ($\chi_{i_t}$) residues, alongside the inter-distance between the two connected residues (Dist.). Rows with red entries correspond to proteins where the intracellular residue involved in the mode connecting the two sides of the membrane lies within the interface region between the GPCR and the $\alpha$-subunit of G- protein. As a

| PDB | Name (Family) | Class | Species |
|---|---|---|---|
| 3ODU | CXCR4 (Chemokine receptors) | Class A | Human |
| 4PXZ | P2Y12 (P2Y receptors) | Class A | Human |
| 3RZE | H1 receptor (Histamine receptors) | Class A | Human |
| 4OR2 | mGlu1 receptor (Metabotropic glutamate receptors) | Class C | Human |
| 1U19 | Bovine rhodopsin | Class A | Bos Taurus |
| 4EIY | A2A receptor (Adenosine receptors) | Class A | Human |
| 4JKV | SMO (Frizzled) | Class F | Human |
| 4MQS | M2 receptor (Acetylcholine receptors (muscarinic)) | Class A | Human |
| 4Z35 | LPA1 receptor (Lysophospholipid (LPA) receptors) | Class A | Human |
| 4PHU | FFA1 receptor (Free fatty acid receptors) | Class A | Human |
| 4XNV | P2Y1 receptor in complex with an antagonist (P2Y receptors) | Class A | Human |
| 4YAY | AT1 receptor (Angiotensin receptors) | Class A | Human |
| 4BUO | NTS1 receptor | Class A | Rattus norvegicus |
| 4DKL | $\mu$ receptor | Class A | Mus musculus |
| 4N6H | $\delta$ receptor (Opioid receptors) | Class A | Human |

Table 4.1.: List of the fifteen GPCRs analyzed with information about their family, class and species.

| Trans. | PDB | Ext. cell. | Int. cell. | Eff. | Dir. | $\chi_{i_p}$ | $\chi_{i_t}$ | Dist. [Å] |
|---|---|---|---|---|---|---|---|---|
| ✓ | 3ODU | ASN 192 | ILE 300 | 0.25 | 11.97 | 0.041 | 0.050 | 43.40 |
| ✓ | 4PXZ | THR 264 | ARG 224 | 0.28 | 8.68 | 0.057 | 0.059 | 64.8 |
| ✓ | 3RZE | HIS 448 | ASP 1061 | 0.06 | 2.80 | 0.130 | 0.048 | 59.6 |
| ✓ | 4PHU | VAL 81 | SER 1136 | 0.11 | 4.68 | 0.082 | 0.062 | 70.5 |
| ✓ | 4OR2 | ALA 687 | LYS 740 | 0.09 | 3.66 | 0.176 | 0.032 | 63.5 |
| ✓ | 4YAY | LYS 20 | LYS 224 | 0.53 | 9.58 | 0.049 | 0.179 | 55.6 |
| ✓ | 4BUO | ALA 69 | THR 179 | 0.31 | 5.87 | 0.041 | 0.116 | 45 |
| ✓ | 1U19 | LYS 16 | GLY 224 | 0.23 | 7.20 | 0.172 | 0.075 | 52.9 |
| ✓ | 4EIY | PRO 266 | SER 35 | 0.80 | 9.20 | 0.049 | 0.079 | 48.9 |
| ✓ | 4JKV | ASP 287 | THR 349 | 0.32 | 13.80 | 0.144 | 0.076 | 57.6 |
| ✓ | 4MQS | PRO A 415 | PRO B 44 | 0.77 | 19.75 | 0.141 | 0.083 | 70.9 |
| ✓ | 4Z35 | HIS 40 | THR 1096 | 0.22 | 6.63 | 0.070 | 0.080 | 73.6 |

Table 4.2.: Energy transfer of the kind *extracellular* $\rightarrow$ *intracellular*. Efficiency (Eff.) and directionality (Dir.) are computed through equations 4.6 and 4.7, respectively. Mode displacements of the pump and targeted residue, $i_p$ and $i_t$ ($\chi_{i_p}$ and $\chi_{i_t}$ respectively) and their respective interdistance (Dist.) are also reported explicitly.

| Trans. | PDB | Int. cell. | Ext. cell. | Eff. | Dir. | $\chi_{i_p}$ | $\chi_{i_t}$ | Dist. [Å] |
|---|---|---|---|---|---|---|---|---|
| ✓ | 3ODU | ILE 300 | ASN 192 | 0.35 | 14.52 | 0.050 | 0.041 | 43.40 |
| ✓ | 4PXZ | ARG 224 | THR 264 | 0.45 | 6.97 | 0.059 | 0.057 | 64.8 |
| ✓ | 3RZE | ASP 1061 | HIS 448 | 0.44 | 2.80 | 0.048 | 0.130 | 59.6 |
| ✓ | 4PHU | SER 1136 | VAL 81 | 0.20 | 6.32 | 0.062 | 0.082 | 70.5 |
| ✓ | 4OR2 | LYS 740 | ALA 687 | 0.76 | 20.01 | 0.032 | 0.176 | 63.5 |
| ✓ | 4BUO | THR 179 | ALA 69 | 0.13 | 3.41 | 0.116 | 0.041 | 45 |
| ✓ | 4DKL | ILE 1050 | GLN 212 | 3.68 | 24.7 | 0.026 | 0.111 | 80.1 |
| ✓ | 4N6H | GLY 248 | ALA 195 | 0.30 | 8.65 | 0.052 | 0.145 | 64.5 |
| ✓ | 1U19 | GLY 224 | LYS 16 | 2.29 | 31.08 | 0.075 | 0.172 | 52.9 |
| ✓ | 4EIY | SER 35 | PRO 266 | 0.37 | 13.99 | 0.079 | 0.049 | 48.9 |
| ✓ | 4JKV | THR 349 | ASP 287 | 0.65 | 18.60 | 0.076 | 0.144 | 57.6 |
| ✓ | 4MQS | PRO B 44 | PRO A 415 | 0.36 | 10.53 | 0.083 | 0.141 | 70.9 |
| ✓ | 4Z35 | THR 1096 | HIS 40 | 0.27 | 7.15 | 0.080 | 0.070 | 73.6 |

Table 4.3.: Energy transfer of the kind *intracellular* → *extracellular*. Efficiency (Eff.) and directionality (Dir.) are computed with equations 4.6 and equation 4.7 respectively. Mode displacements of the pump and targeted residue, $i_p$ and $i_t$ ($\chi_{i_p}$ and $\chi_{i_t}$ respectively) and their respective interdistance (Dist.) are also reported explicitly.

| Trans. | PDB | Int. cell. | Ext. cell. | $\chi_{i_1}$ | $\chi_{i_2}$ | Dist. [Å] |
|---|---|---|---|---|---|---|
| no | 4XNV | VAL 1024 | GLY 193 | 0.118 | 0.026 | 76.6 |

Table 4.4.: Protein where a mode connecting either end of the cell membrane does exist but does not lead to any energy transfer when forced.

general observation, energy transfers occur in all the numerical experiments except one with a good efficiency (more than 30%) and a good directionality (target residues have an energy more than 6 times higher than the mean of the other residues) and connect residues separated by remarkably large inter-distances, between 43.4 Å and 80.1 Å (see Appendix A.2. for the figures representing the mode pattern and the average energy pattern in the stationary state for each simulation). Moreover, for what concerns transfers of the kind *extra* $\rightarrow$ *intra*, for seven proteins the transfer reaches to an intracellular residue belonging to the interaction region between the GPCR and the $\alpha$-subunit of G-protein. This suggests that such vibrational modes could be involved in the allosteric response that leads to the release of the $\alpha$-subunit of G-protein. An inspection of the displacement pattern of the two residues $i_1$ and $i_2$ in table 4.4 (no transfer observed) shows that the value of displacement pattern $\chi_{i_2}$ is the lowest of all. Moreover, the interdistance between residues $i_1$ and $i_2$ is one of the highest, implying a high number of intermediate centers. However, $\chi_{i_1}$ is large (greater than 0.1), which suggests that the intermediate residues are not able to transfer efficiently the energy and consequently an overall flow across the protein structure cannot be observed.

We have seen that, in order to make long-range energy transfer possible, the displacement of residues $i_1$ and $i_2$ in the bilocalized mode should reflect a trade-off between high localization and high contribution of *noise*. We observe in tables 4.2 and 4.3 that transfers occur generally for intermediate values of $\chi_{i_1}$ and $\chi_{i_2}$, *i.e.* between 0.05 and 0.1, which confirms the above conjecture. A pictorial representation of one of the transfer events reported in table 4.2 is illustrated in figure 4.11 for an event of the kind *extracellular* $\rightarrow$ *intracellular* (CXCR4 chemokine receptor). By applying a pump at a frequency 20.6 $\text{cm}^{-1}$ ($k = 125$) to ASN 192 (outside of the cell), one can observe an energy transfer to ILE 300 at the end of Helix 7 in the intracellular region. The corresponding values of the efficiency and the directionality are 0.25 ans 11.97, respectively. Interestingly, we remark that $\chi_{i_p}$ and $\chi_{i_t}$ are between 0.04 and 0.06, thus relatively small but large enough to cover the *noise* and allow for a directed energy transfer. The other energy transfers can be examined in the figures reported in Appendix A.2.

## 4.3. Conclusion

In chapter 3, we have proposed a protocol to predict catalytic sites in enzymes. Along similar methodological lines, the aim of this chapter was to propose a mechanistic model of how specific, possibly functional sites, are able to *communicate* between them as a global property of the scaffold they belong to. We observed that a vibrational excitation at one given site, in the form of a forcing at a frequency within the normal mode

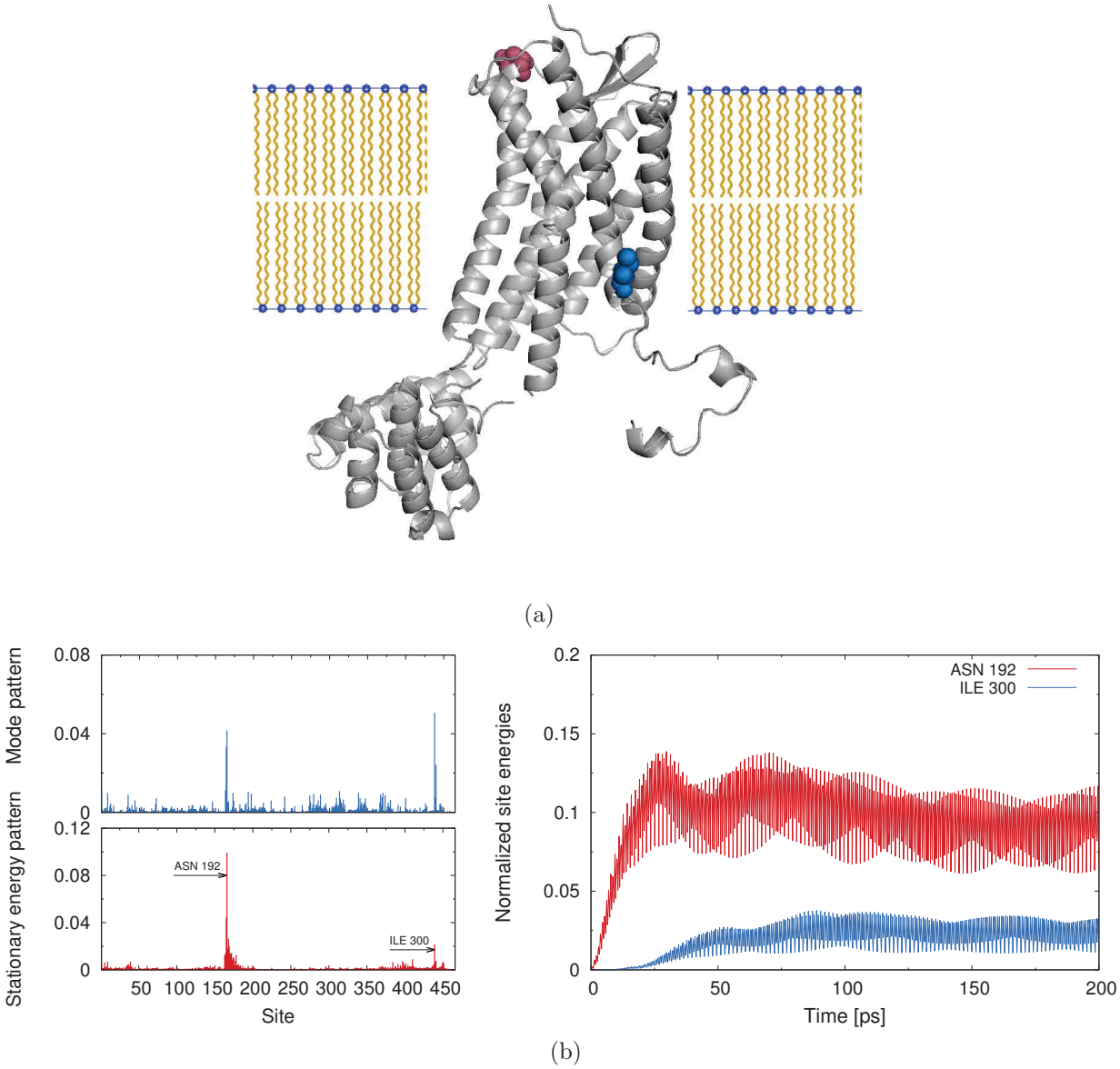



Figure 4.11.: Summary of pumping simulation for the CXCR4 chemokine receptor. (a) Structure of the receptor (PDB 3ODU) schematically immersed in the cell membrane. Red spheres represent the amino-acid involved in the binding of the ligand in the extracellular region (ASN 192), blue spheres highlight the residue in the intracellular region (ILE 300). (b) Pumping analysis. Top left, $\chi_i^k$ pattern ($k = 125$). Bottom left, average normalized energy pattern in the steady state when we force ASN 192 at a frequency $\omega_k = 20.6$ cm$^{-1}$. Right, local total energy of residues versus time.

spectrum, may trigger an energy flow to a distant location, following specific selection rules. As a general finding, energy accumulates across the structure following the specific displacement pattern of the forced mode, even though residues displaying a non-negligible displacement pattern are at considerable distance from the energy injection center.

We analyzed a database of 1711 proteins in order to draw a statistical picture of this phenomenon. In the quest for efficient energy transfers, we decided to perform a high-pass filtering procedure and select only the modes $k$ where the number of peaks $N_p^k$ in the displacement pattern was between $N_p^k = 2$ and $N_p^k = 4$. Among these modes, we further selected the specific mode where the two residues with the highest contribution were the more distant. A structural analysis on these bilocalized modes with widely separated localization centers revealed that these correspond to frequencies in the range 10-20 $\text{cm}^{-1}$ for inter-center distances lager than 50 Å.

A set of pump simulations were performed for all the proteins of the database. For each protein, we selected the mode $k$ that connects the two most separated residues and performed two numerical experiments, first by forcing residue $i_1$ (the one with the highest contribution to the mode pattern) and then residue $i_2$ (the second-highest contribution). The stationary energy patterns confirm that residues $i_1$ and $i_2$ indeed act as accumulation centers as their energy distribution in the stationary state peaks at higher values than that referring to other residues. Moreover, we observed that energy transfers of the kind $i_2 \to i_1$ were more efficient than those of the kind $i_1 \to i_2$.

More detailed analyses allowed us to make the intriguing observation that the inter-distance between residues $i_1$ and $i_2$ plays a subtle role. We remarked that the larger the inter-distance between these residues, the more *noise* is required for a transfer to occur, that is, a non-negligible contribution besides the peaks of the pumped and the target residues to the normal mode displacement pattern is important, as the energy needs to be channeled through intermediate residues. If the pumped and target residues contribute to an excessive extent (more than 40%) to the mode pattern, the energy cannot leave the excitation point and stays stored at that site. However, for short inter-distances, the number of intermediate structural points being limited, the higher the contribution of the pumped and target residues to the displacement pattern of the pumped mode, the more efficient the energy transfer.

To conclude, NMA allows us to isolate potentially allosteric residues that connect to functional residues in proteins. We proved by numerical experiments that if two amino-acids are connected by one mode, depending on their interdistance and their contribution to the displacement pattern of the mode, they are able to communicate *via* energy transfer. We have observed neat energy transfers for residues featuring distances of more than 70 Å. Moreover, a direct inspection of the pattern of bilocalized modes featuring highly separated localization centers allows one to gather information on the energy transduction pathways. To this end it is enough to inspect the mode pattern and isolate the additional localization centers besides the two main ones, that will act as intermediates (pathway) in the energy channeling process.

# 5. Surface cooling : spontaneous localization of energy in proteins

A many-body system in an out-of-equilibrium state coupled to its environment through its surface will relax to equilibrium following a complex hierarchy of relaxation rates. This is a consequence of the spatially inhomogeneous coupling with the environment [180].

The introduction of nonlinearity leads to even more complex relaxation patterns. In fact, when a nonlinear many-body system is cooled from the surface, spontaneous energy localization emerges rather generally [181–184]. Energy gets trapped in the bulk of the system, far from the colder surface, in the form of localized vibrational modes. Such modes, whose spatial envelope is localized exponentially, are known as discrete breathers (DB) and can be considered as collective (even if localized) vibrations whose frequency by definition must not overlap with the linear spectrum of the system [38, 185, 186].

It has been speculated that DBs may play a role in certain biological processes involving energy storage and transfer across protein structures, such as allostery [187]. Moreover, it has been also proposed that DBs may play a role in enzyme catalysis [132, 188, 189], as they tend to localize at the stiffest sites in enzymes. Interestingly, as we have shown in chapter 3, local stiffness, measured as the contribution of high-frequency normal modes to residue fluctuations, is able to flag catalytic sites in enzymes, which strengthens the conjecture that DBs may be involved in local storage of energy during enzyme catalysis.

A recent series of studies [38, 57, 60, 132] has investigated an extension of the coarse-grained ENM, the coarse-grained Nonlinear Network Model (NNM), where a nonlinear term has been added to the system potential energy. The numerical experiments reported in reference [38] have shown that cooling a protein through its surface within an implicit solvent model leads to the spontaneous formation of localized vibrations in the bulk. The protein as a result of the cooling process was completely frozen, except for such *hotspots*, where a considerable fraction of the original energy had been pinned down far from the surface. Interestingly, a thorough statistical analysis has shown that energy localization targeted specific residues, typically the stiffest ones.

In this chapter, we investigate the question whether the same or a similar scenario is observed when cooling a protein described at atomic level, with a realistic force-field,

and immersed in an explicit solvent. Since the force-field that we used (Gromos53a6, see chapter Methods) is highly non-linear, we expect to observe spontaneous energy localization. The basic questions that we want to address are : (1) whether energy localization appears at all and (2) whether in this case it is a site-specific process as in the case of the coarse-grained scheme employed in reference [38]. We anticipate that energy localization can be observed indeed and there seems to be a similar match between *hot residues* and stiff regions. Moreover, one interesting finding, as we will show in the following, is that specific vibrational patterns tend to emerge from the surface cooling, typically involving specific deformations of phenylalanine aromatic rings.

## 5.1. Cooling at different damping rates

The solvent cooling procedure consists in progressively freezing the solvent. In order to cool down the water, we rescale velocities of the solvent molecules at each step of the simulation (see chapter Methods), $\tau_T$ being the time constant of the temperature coupling. Hence, recalling equation 2.9, the higher $\tau_T$, the slower the cooling.

We observe that, depending on the rapidity of the cooling process ($\tau_T$) with respect to the hierarchy of vibrational time scales in the protein, typically the asymptotic dynamics will drive the protein to a different local minimum of the rugged potential energy landscape. In the case of very slow (*i.e.* adiabatic) cooling, one may expect the protein to reach its (force-field dependent) global minimum. Incidentally, this procedure might as well be employed as a general minimization technique to reach force-field dependent global minima. However, in this case, no localization is expected to appear. This is an obvious difference with respect to the coarse-grained scheme employed in the study [38], where by definition there are no local minima.

We have performed simulations on citrate synthase starting from several independent realizations of a thermal equilibrium state at $T = 310$ K. As a first study, we have investigated the role of the time constant $\tau_T$ by considering six values, ranging from $\tau_T = 0.05$ ps to $\tau_T = 50$ ps. The typical observable that we have examined is the field of local temperatures, which we have computed residue by residue by adding up the kinetic energies of all the atoms in the residue and separately for the side-chains and for the backbone segments. The relaxation of the total energy of a many-body system to its asymptotic state during a typical surface cooling simulation can be understood in the following way. Let $E(t)$ denote the total system energy, and let $\{\tau_1 < \tau_2 < \tau_3 \cdots < \tau_n\}$ denote the relaxation rates associated with the $n = 3N - 6$ normal modes ($N$ is the

number of atoms in the system). Then, one should expect [180]

$$E(t) \approx \begin{cases} c_1 e^{-t/\tau_1} & t \ll \tau_1 \\ \sum_\alpha c_\alpha e^{-t/\tau_\alpha} & t \gg \tau_1 \end{cases} \tag{5.1}$$

Here $\tau_1$ is the relaxation rate associated with the normal mode that relaxes the faster. We note that the time constants $\tau_\alpha$ can in principle be estimated by computing the Langevin modes of the protein [190].

### 5.1.1. Slow cooling, surface fraction and effective dimension of citrate synthase

Typically, equation (5.1) predicts that, as a result of the inhomogeneous coupling with the solvent, one should observe a crossover from a single exponential to a slower, more complex relaxation law, on a time scale of the order of $\tau_1$. The latter originates from the superposition of many exponentially decaying relaxation channels and often appears as something intermediate between a power-law and a stretched exponential [38, 182, 191]. Figure 5.1 illustrates the crossover from an exponential temperature decay to a slower relaxation, which tends to a power-law beyond a certain time $\tau_1$ (equation 5.1). We performed six cooling simulations with identical initial conditions but with different cooling rates, with $\tau_T$ varying from 0.05 ps to 50 ps. The crossover mentioned above is observed for cooling simulations with time constant $\tau_T = 1$ ps, $\tau_T = 5$ ps and $\tau_T = 20$ ps, whereas it is more difficult to observe for faster coolings ($\tau_T \leq 0.1$ ps) where an integrated relaxation law seems to set in already at an early stage. The slowest cooling simulation ($\tau_T = 50$ ps) do not show a similar transition in the analyzed integration window. According to the analysis performed in Appendix B, we might expect a transition to the *integrated* behavior at times longer than about 340 ps.

One can observe on figure 5.1 that $\tau_1$, the relaxation rate associated with the normal mode that relaxes the fastest, increases as $\tau_T$ increases. The relation between $\tau_1$ and $\tau_T$ for $\tau_T = 1$ ps, $\tau_T = 5$ ps and $\tau_T = 20$ ps is linear (see Appendix B), namely

$$\tau_1 = \alpha \tau_T + \beta. \tag{5.2}$$

In analogy with the previous surface cooling studies [180], we expect that, for slow cooling, the damping rate $\gamma_1 = 1/\tau_1$ be : (1) proportional to the surface fraction $f$ of the protein, which is the number of atoms interacting with the solvent normalized to the total number of atoms and (2) inversely proportional to the effective spatial dimension $d_{eff}$ of the protein. Then, by computing the surface fraction through a consensus tool [192] with a

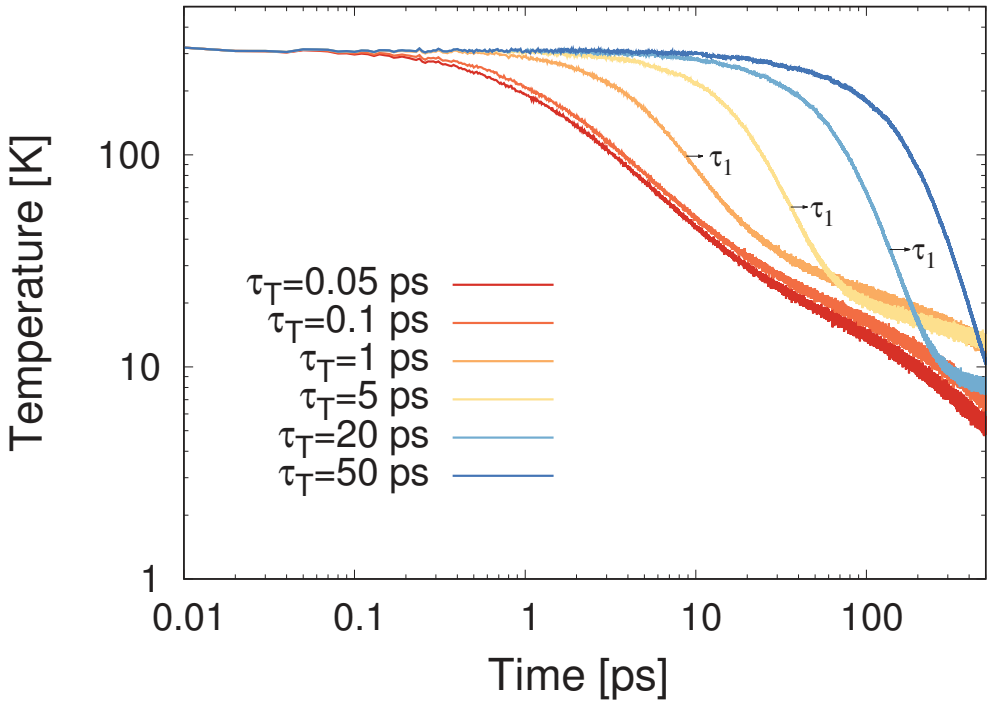



Figure 5.1.: Citrate synthase (PDB code : 1IXE) temperature for six cooling simulations with different cooling rates ($\tau_T$) in log-log scale. We remark that, for $\tau_T = 1$ ps, $\tau_T = 5$ ps and $\tau_T = 20$ ps, a change of relaxation behavior is observed from an exponential decay to a slower, power-law-like decay. We indicate with $\tau_1$ the different times corresponding to the change of slope from an exponential decay to a power law.

cutoff on the local exposed surface set at 2.5 Å$^2$ (atoms are considered in interaction with the solvent if their surface exposed to the solvent is greater than 2.5 Å$^2$), we can relate the observed relaxation rate $\gamma_1$ to the effective dimension $d_{eff}$ of citrate synthase. This yields $f = 0.26$, which gives $d_{eff} \approx 1.75$. Assuming that the effective dimension of proteins is an intrinsic property of protein scaffolds,, this result is in agreement with reference [180], where authors have found for myoglogin an effective dimension $d_{eff} \approx 1.45$.

### 5.1.2. Asymptotic temperatures as a function of the cooling rate

In order to quantify the extent of spontaneous energy localization induced by the cooling process, we followed the local temperature of each residue during the different independent cooling simulations, that is,

$$T_i = \frac{1}{3k_B N_i} \sum_{j \in R_i} \frac{|\mathbf{p}_j|^2}{m_j}, \tag{5.3}$$

where $\mathbf{p}_j$ is the momentum of the $j^{th}$ atom, $m_j$ its mass and $k_B$ the Boltzmann constant. The sum runs over all the atoms belonging to residue $R_i$, possessing $N_i$ atoms.

| $\tau_T$ [ps] | Average temp. (10 hottest res.) [K] | Average asymptotic temp. (entire protein) [K] | Ratio |
|---|---|---|---|
| 0.05 | $43 \pm 3$ | $5.4 \pm 0.2$ | $8.0 \pm 0.1$ |
| 0.1 | $54 \pm 4$ | $7.1 \pm 0.3$ | $7.6 \pm 0.1$ |
| 1 | $72 \pm 7$ | $12.8 \pm 0.4$ | $5.6 \pm 0.1$ |
| 5 | $75 \pm 3$ | $13.3 \pm 0.4$ | $5.6 \pm 0.1$ |
| 20 | $47 \pm 4$ | $8.2 \pm 0.2$ | $5.7 \pm 0.1$ |
| 50 | $31 \pm 3$ | $11.1 \pm 0.3$ | $2.8 \pm 0.1$ |

Table 5.1.: Average temperature of the ten hottest residues (with the corresponding statistical errors on the mean) computed by averaging over the last 100 ps long stretch of the simulations and final temperature for each simulation averaged over the last 20 ps of the simulations. The cooling runs start from the same initial conditions and differ by their temperature coupling time constant.

The typical outcome of a relaxation experiment is illustrated in figures 5.2 and 5.3 for two different damping rates, where we have represented the average local temperature of the backbone segments and sidechains of all residues (figure 5.2) and the local temperature of two chosen residues (figure 5.3). While most residues cool down completely, spontaneous self-excitation of energy is observed at certain sites, both along the backbone and in the side-chains. In the following, for each simulation we have typically focused on the ten hottest residues in the quasi-stationary state. In turn, this has been defined as the last 100 ps-long stretch of the cooling simulations.

Table 5.1 reports an ensemble analysis of the six different relaxation experiments, identical except for the choice of the relaxation rate $\tau_T$. A number of interesting observations can be made. First of all, it is apparent that at intermediate cooling rate the average asymptotic temperatures, both local and global are higher. Moreover, as expected, the slower the cooling, the less prominent is the local energy pinning, gauged by the ratio of the asymptotic average local temperature (10 hottest sites) and global protein temperature. These results agree with the behavior shown in figure 5.1, where we observe a change of slope in the temperature decay at intermediate values of $\tau_T$. Finally, even though the final protein temperature for the simulation with $\tau_T = 50$ ps is higher than the one for $\tau_T = 0.05$ ps, $\tau_T = 0.1$ ps and $\tau_T = 20$ ps, we observe that energy does not localize as much as in the other simulations (see Appendix B). Indeed, the average temperature over the ten hottest residues when $\tau_T = 50$ ps is the lowest value of all. Hence, we note that a 500 ps simulation is probably too short to attain the quasi-stationary state in this case, if such a state is ever attained.

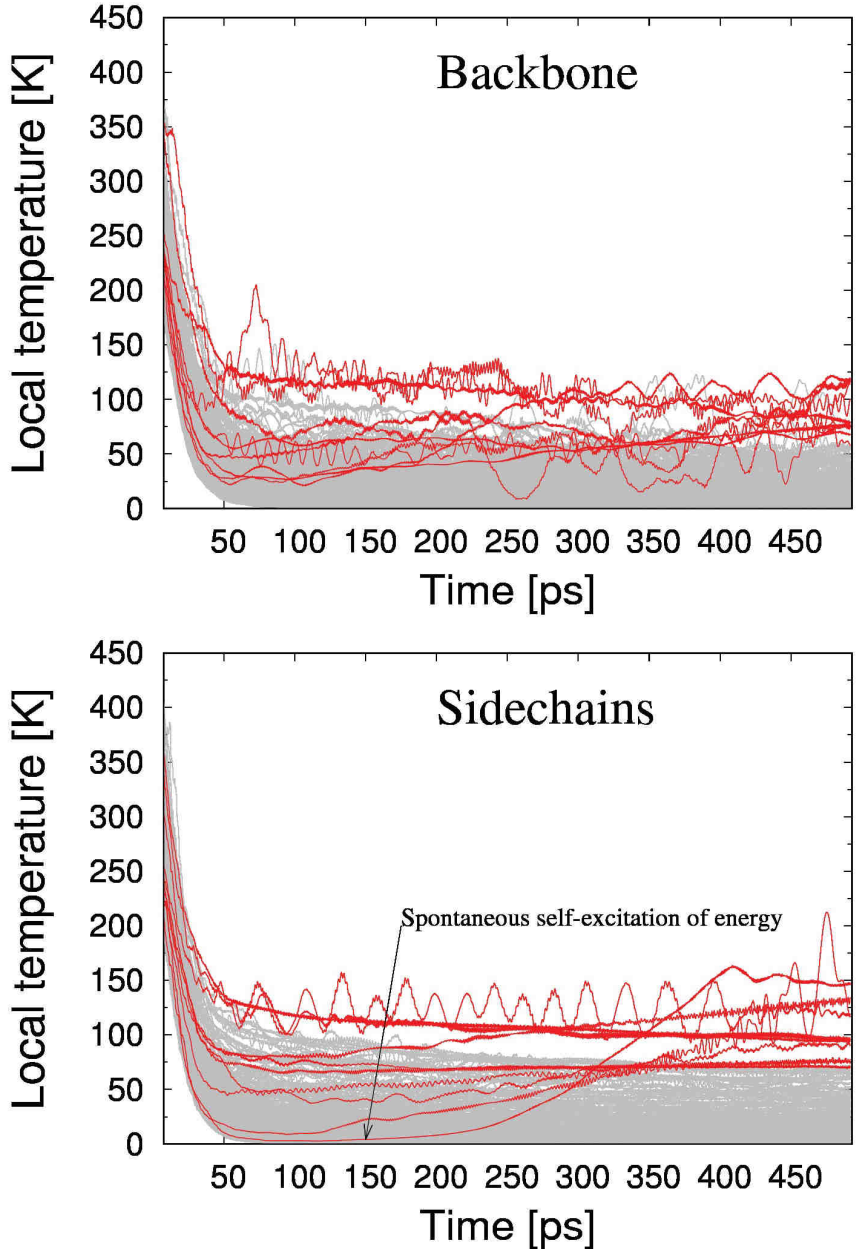



Figure 5.2.: Average local temperature (computed over a set of 100 points, 15 ps average) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 5$ ps. While we observe a generalized decay of temperature, we remark that, despite the steady cooling, some residues (highlight in red) see their local temperature increase during the relaxation process.

## 5.2. Cooling from independent initial conditions

In this section, we investigate whether the cooling process tends to direct energy systematically to specific regions of the protein. To this end, we examine the outcome of identical relaxation experiments performed by starting from different initial configurations, sam-

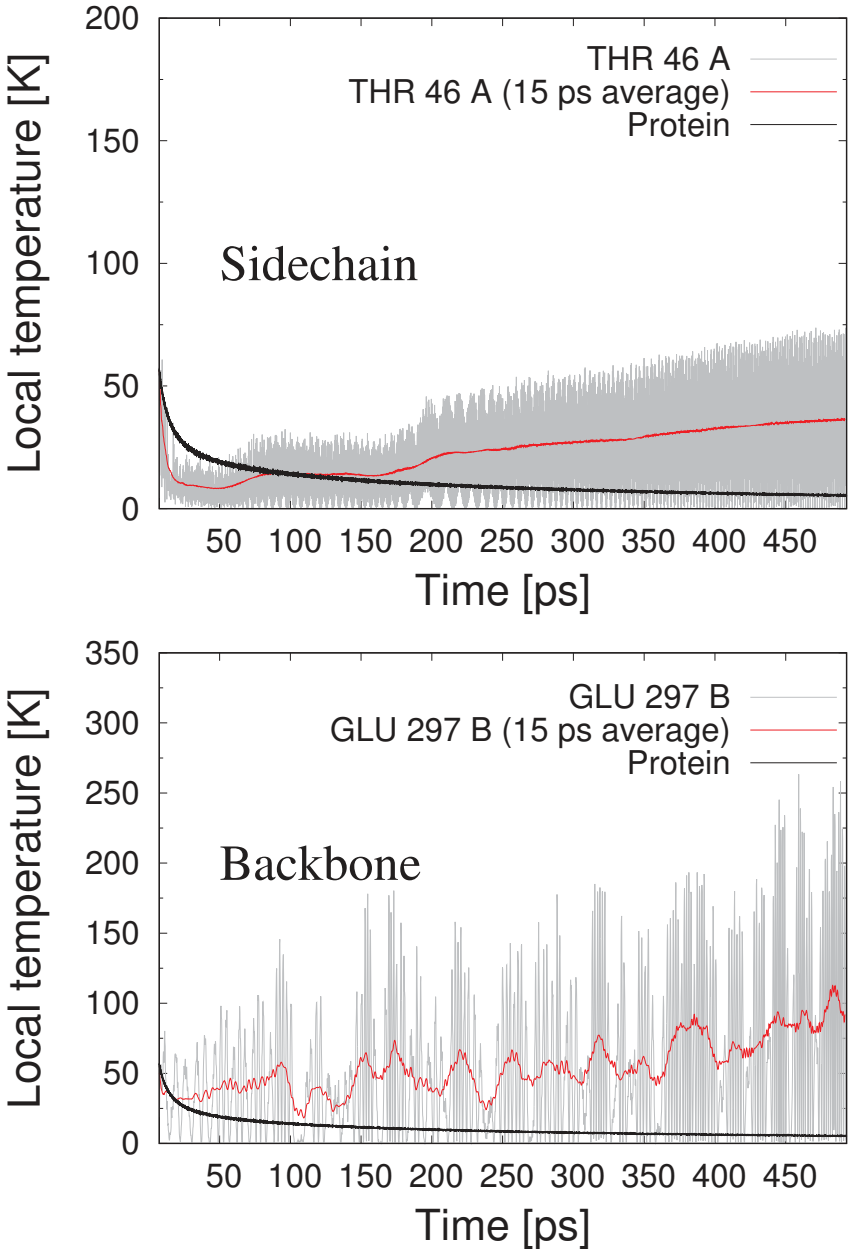



Figure 5.3.: Local temperature and average local temperature (computed over a set of 100 points, 15 ps average) of sidechain (upper panel) and backbone segments (lower panel) of two different residues of citrate synthase for a solvent cooling simulation of 500 ps with $\tau_T = 0.05$ ps. While the temperature of the protein decreases, some particular residues, after a transient, become hotter as time elapses. Hence, we observe spontaneous transfer of energy to particular residues while the others get colder and colder as the cooling proceeds.

pled from an equilibrium simulation at $T = 310$ K. As the local to global asymptotic energy ratio is the highest, we have chosen a damping rate $\tau_T = 0.05$ ps (see table 5.1). We have analyzed nine independent initial conditions, obtained by sampling an equilib-

rium simulation at $T = 310$ K (NPT ensemble) every nanosecond after a first equilibration stage of 10 ns. Simulations were performed in double precision with the gromos53a6 force field.

### 5.2.1. Temperature localization and structural indicators

During each cooling simulation, we have extracted the velocities of each atom and stored two time series for each residue: the local temperature of its backbone atoms (bb) and the local temperature of its sidechain atoms (sc). These are simply defined as

$$T_i^{bb} = \frac{1}{3k_B N_{bb_i}} \sum_{j \in \mathrm{bb_i}} \frac{|\mathbf{p}_j|^2}{m_j}, \tag{5.4}$$

and

$$T_i^{sc} = \frac{1}{3k_B N_{sc_i}} \sum_{j \in \mathrm{sc_i}} \frac{|\mathbf{p}_j|^2}{m_j}, \tag{5.5}$$

where $\mathbf{p}_j$ is the momentum of the atom $j$, $m_j$ its mass and $k_B$ the Boltzmann constant. The sum runs over all the atom included in the backbone segments ($N_{bb_i}$) or in the different side-chains ($N_{sc_i}$), including hydrogens in both cases.

The time evolution of a typical cooling simulation is shown in figure 5.3. It is apparent that, while the whole protein sees its temperature decrease, particular residues become hotter during the cooling procedure. Figure 5.3 shows that intra-molecular energy transfers occur, so that the temperature at a given location starts growing as the corresponding amino-acid gathers energy from the background. More precisely, we observe that after a relaxing transient, while the temperature of the protein is steadily decreasing, some residues see their local temperature increase during the simulation. Such phenomenon would involve energy transfers between residues that become quenched and other residues, such as THR 46 A and GLU 297 B (figure 5.3) whose local temperature increases during the cooling process. In the following, we will focus on the ten hottest residues (backbone and sidechains) for each simulation. From this selection, we have observed that the temperatures of the backbone segments are in general higher than the temperatures of the sidechains, with an average temperature of 74 ± 3 K and 45 ± 2 K, respectively (table 5.2). Moreover, while the nature of the targeted residues does not seem to follow any specific pattern for what concerns their backbone segments, it is the opposite when considering the sidechains, where one particular residue, *i.e.* phenylalanine (PHE) seems to constitute a preferential energy pinning target. Table 5.3 reports an ensemble analysis of the nine different cooling simulations. More precisely, we find that PHE 157 (either in chain A or B of the citrate synthase dimer) is found among the ten hottest residues

| Conformation | Average temp. (10 hottest bb.) [K] | Average temp. (10 hottest sc.) [K] |
|---|---|---|
| 1 | 93 ± 11 | 36 ± 1 |
| 2 | 85 ± 11 | 54 ± 9 |
| 3 | 68 ± 8 | 43 ± 1 |
| 4 | 65 ± 4 | 40 ± 3 |
| 5 | 68 ± 4 | 48 ± 4 |
| 6 | 65 ± 7 | 50 ± 6 |
| 7 | 84 ± 8 | 44 ± 5 |
| 8 | 61 ± 5 | 51 ± 4 |
| 9 | 77 ± 10 | 43 ± 4 |

Table 5.2.: Average temperature of the ten hottest backbone segments and ten hottest sidechains for each simulation. Simulations start from different initial conditions (thermal state at $T = 310$ K) and have been performed with a cooling time constant of $\tau_T = 0.05$ ps.

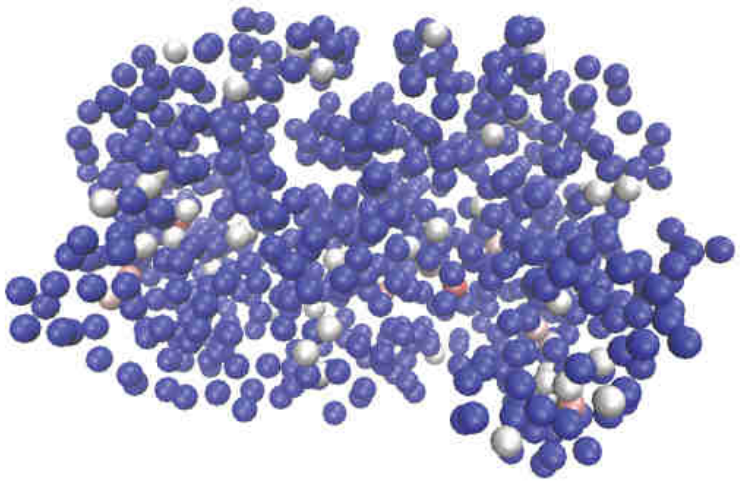
ten times over the entire pool of nine independent relaxation experiments (*i.e.* in one simulation, the same residue in both chains is a center of energy localization). PHE 313 and PHE 128 are also highly represented (found among the hottest residues in eight and seven independent cooling runs, respectively). In order to provide a visual representation of the distribution of *hottest* sites, it is instructive to show the 3D structure of the citrate synthase dimer in a residue representation, where each amino-acid is colored with a color code matching the number of times it is found among the ten hottest residues over the whole pool of independent relaxation experiments. Blue indicates a cold residue (never found among the hottest sites), red means hot residue (preferential target for cooling-induced energy localization). Surprisingly, many of the hot residues are found close to the protein surface, hence close to the solvent. However, due to the complexity of the protein surface at atomic scale, a residue that is close to the surface does not imply that all atoms belonging to it are also close to the solvent. In summary, the picture emerging from atomic-scale surface cooling, despite bearing some similarities with the results of coarse-grained simulations, also shows a substantially higher degree of complexity.

### Shedding some light on the observed relaxation pattern : the lense of structural indicators

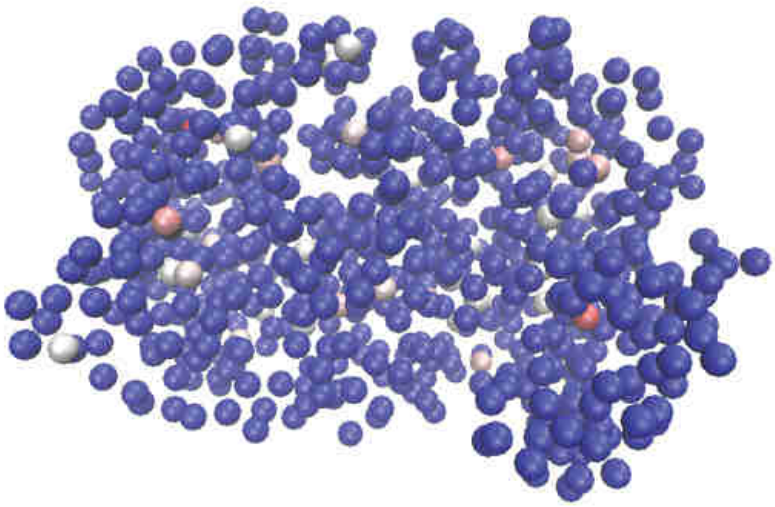
We have seen that certain specific residues tend to appear systematically among the pool of preferential energy-pinning centers. We would like to address the question whether such locations can be singled out independently of the cooling process, on the basis of specific local structural indicators. In the same spirit of the analysis reported in chapter 3, it is natural to turn to the same indicators examined with the purpose of identifying

| Backbone | | Sidechains | |
|---|---|---|---|
| TYR 314 A (5) | LYS 361 A (1) | **PHE 157 A (7)** | ARG 71 B (1) |
| HIS 219 B (4) | GLY 25 B (1) | **PHE 313 B (6)** | THR 46 A (1) |
| LEU 37 A (3) | GLU 45 A (1) | **PHE 128 B (4)** | **PHE 186 B (1)** |
| ALA 298 B (3) | LYS 83 B (1) | **PHE 64 A (4)** | THR 190 B (1) |
| ASP 352 B (3) | SER 100 B (1) | **PHE 128 A (3)** | TYR 204 B (1) |
| GLY 25 A (2) | LYS 120 B (1) | **PHE 191 A (3)** | HIS 219 B (1) |
| VAL 243 A (2) | VAL 132 B (1) | **PHE 324 A (3)** | ASN 223 B (1) |
| GLU 36 A (2) | ALA 133 B (1) | TYR 362 A (3) | ARG 71 A (1) |
| LEU 37 B (2) | ILE 145 B (1) | HIS 153 B (3) | **PHE 329 B (1)** |
| LEU 50 B (2) | MET 174 B (1) | **PHE 157 B (3)** | TRP 341 B (1) |
| THR 190 B (2) | ASN 187 B (1) | **PHE 191 B (3)** | ARG 376 B (1) |
| ASN 310 B (2) | SER 189 B (1) | **PHE 324 B (3)** | HIS 77 A (1) |
| THR 13 A (1) | SER 198 B (1) | **PHE 102 A (2)** | **PHE 82 A (1)** |
| LEU 158 A (1) | SER 213 B (1) | **PHE 197 A (2)** | |
| GLY 163 A (1) | LYS 215 B (1) | TYR 261 A (2) | |
| MET 174 A (1) | GLU 62 A (1) | **PHE 313 A (2)** | |
| ASP 175 A (1) | GLU 241 B (1) | **PHE 333 A (2)** | |
| LEU 178 A (1) | VAL 243 B (1) | **PHE 64 B (2)** | |
| ILE 179 A (1) | ALA 248 B (1) | **PHE 48 A (2)** | |
| ALA 193 A (1) | SER 65 A (1) | **PHE 197 B (2)** | |
| THR 235 A (1) | ALA 263 B (1) | TYR 261 B (2) | |
| ALA 239 A (1) | ASP 265 B (1) | **PHE 12 A (1)** | |
| ARG 259 A (1) | ALA 275 B (1) | **PHE 186 A (1)** | |
| ALA 268 A (1) | LYS 292 B (1) | SER 189 A (1) | |
| ALA 279 A (1) | VAL 294 B (1) | ARG 218 A (1) | |
| TYR 288 A (1) | TYR 308 B (1) | HIS 219 A (1) | |
| ILE 290 A (1) | **PHE 313 B (1)** | HIS 258 A (1) | |
| GLU 297 A (1) | TYR 314 B (1) | TYR 288 A (1) | |
| ALA 298 A (1) | SER 320 B (1) | TYR 314 A (1) | |
| ASN 310 A (1) | ASP 321 B (1) | TRP 341 A (1) | |
| TYR 319 A (1) | ALA 334 B (1) | TYR 370 A (1) | |
| ARG 337 A (1) | ALA 73 A (1) | TYR 29 B (1) | |
| SER 339 A (1) | ARG 357 B (1) | SER 41 B (1) | |
| GLN 349 A (1) | LEU 78 A (1) | **PHE 43 B (1)** | |
| GLU 350 A (1) | LEU 79 A (1) | **PHE 48 B (1)** | |

Table 5.3.: Classification of the ten hottest backbone segments and sidechains by their abundance (in brackets) in the ten hottest pool over the nine independent numerical experiments. Phenylalanine (PHE) residues are colored in red.



(a)

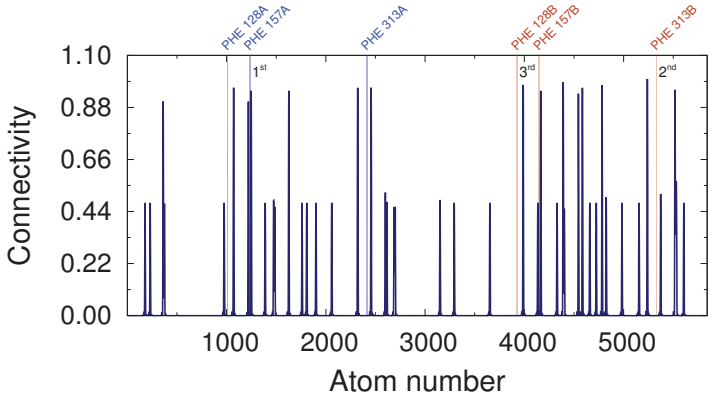


(b)

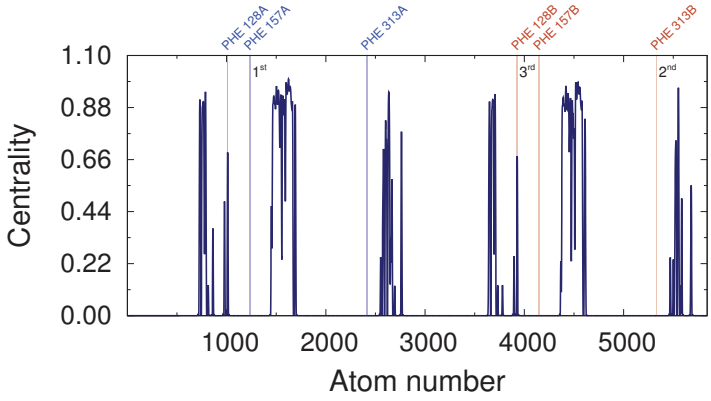
Figure 5.4.: 3D structure of citrate synthase (PDB 1IXE) represented by a sphere per amino-acid. The color scale defines the relative abundance of the residues included among the ten hottest over the entire pool of cooling runs. Blue means that the residue is never found, while red means high abundance. (a) hot backbone segments and (b) hot sidechains. Many of the *hot* backbone residues are at the surface compared to the *hot* sidechain residues, which are more in the bulk.

catalytic residues in enzymes.

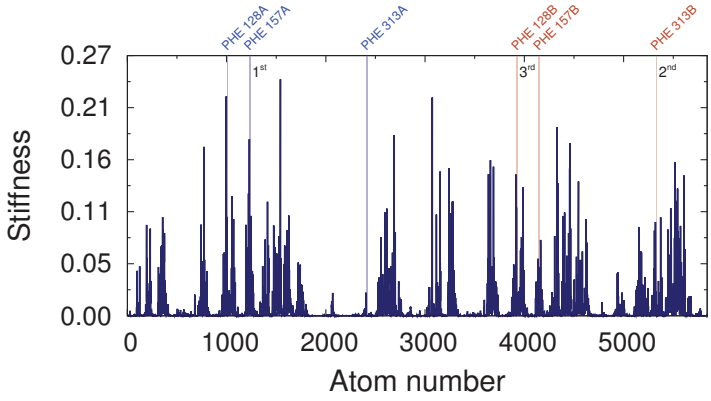
To this aim, we constructed an elastic network model at atomic level (à la Tirion) [50]. All atoms, except hydrogens, are modeled by a bead of the same average mass (14 amu). A




(a)




(b)




(c)

Figure 5.5.: Structural indicators computed for citrate synthase (PDB 1IXE) through an atomistic elastic network analysis. (a) Connectivity, (b) closeness, (c) stiffness. Phenylalanines that feature among the hottest side-chains in more than two thirds of the independent cooling simulations are highlighted in blue (chain A) and yellow (chain B).

network of springs is then constructed as usual by imposing an interaction cutoff $R_c$ on the equilibrium structure (PDB 1IXE). With 5842 heavy atoms compared to 741 amino-acids,

we have adapted our stiffness measure, by including a higher number of highest-frequency modes. As we have approximately eight times more beads at atomic level, we retained a set comprising the 40 highest-frequency normal modes, that is

$$\chi_i^{aa} = \sum_{k=3N-39}^{3N} |\boldsymbol{\xi}_i^k|^2 \tag{5.6}$$

where $\chi_i^{aa}$ is the stiffness of atom $i$ and $\boldsymbol{\xi}_i^k$ the contribution to atomic displacements of atom $i$ to the normal mode $k$. We have used a cutoff radius $R_c = 5$ Å in the system potential energy (see chapter 2). In addition to the stiffness, we also computed two other indicators, namely the connectivity and the closeness centrality, computed as described in chapter 3. Following the findings of chapter 3, we rescaled the optimal cutoff radius for the connectivity and closeness centrality measures following similar arguments. In particular, we computed connectivity patterns with a value of $R_c$ identical to that used for calculating the stiffness pattern, *i.e.* $R_c = 5$ Å. Concerning centrality, the optimal value determined in chapter 3 at the scale of amino-acids, *i.e.* $R_c = 28$ Å, translates here to $R_c = 6.4$ Å. Moreover, we applied the same high-pass filtering procedure, with correspondingly greater values for the thresholds, 5.5 standard deviations for the connectivity and 3.5 standard deviations for the centrality.

Figure 5.5 illustrates the different indicators computed at their respective cutoff radius. The locations of the relevant residues highlighted by the cooling process are indicated explicitly. Remarkably, each indicator spotlights at least one of the three most represented PHEs and the corresponding twin (*i.e.* the same residue in the other monomer). The stiffness pattern is clearly the more interesting, with a match for all the six PHEs that are among the ten hottest residues in more than two thirds of the nine independent numerical experiments. Within the limits of more complex stiffness patterns, these results confirm and shed some additional light on the findings reported in reference [38].

### 5.2.2. Detailed examination of displacement patterns

Our analyses have allowed us to isolate the *hottest* residues for each simulation and analyze their movements. For what concerns vibrations developing as a result of cooling within the backbone, it turns out that energy is systematically stored in the angle-bending degrees of freedom involving the atoms $C_\alpha - N - H$. A typical vibration of this kind is shown in figure 5.6. A Fourier analysis on a set of ten backbone angle-bending vibrations yield frequencies of (1439 $\pm$6) cm$^{-1}$. Concerning side-chain segments, we mainly focused our attention on PHEs displacements in view of their over-representation among the *hottest* residues in the nine independent cooling simulations (see table 5.3). Inspecting the nine

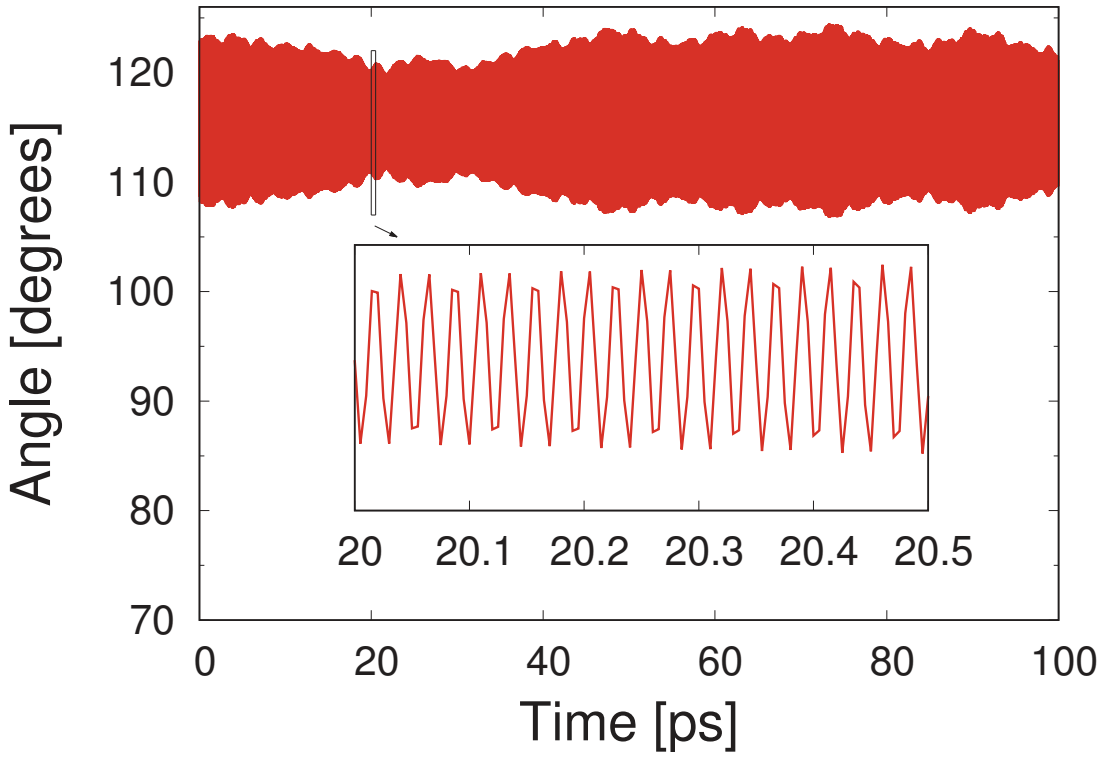



Figure 5.6.: Amplitude of the backbone angle bending involving the atoms [$C_\alpha$-N-H] of residue ASN 187 B in the quasi-stationary state. $t = 0$ means 400 ps after switching on the solvent cooling.

independent trajectories, we remark that all PHEs display typical movements involving the six carbon in the aromatic ring. A Fourier analysis of such ring deformations on a set of more than forty PHEs yielded frequencies varying from 868.1 $cm^{-1}$ to 872.4 $cm^{-1}$ (see Appendix B.2).

### Principal component analysis of PHE ring deformations

We observed that the typical movements of hot PHEs involve the six ring carbon atoms and that the frequencies extracted from the quasi-stationary time series of angle-bending motions (last 100 ps) varied from 868.1 $cm^{-1}$ to 872.4 $cm^{-1}$. To shed further light on the typical movements of the aromatic rings excited as a result of cooling, we performed a principal component analysis (PCA) on three trajectories of the six ring carbon atoms, the first corresponding to the trajectory of PHE 157 A, the second to the trajectory of PHE 313 B and the third to the trajectory of PHE 128 B, which are all found among the hottest residues several times in the nine independent cooling runs. In order to extract a frequency from the PCA, we further projected the trajectory of the rings for each PHE on the first principal mode and performed a Fourier analysis of these time series. Let $\mathbf{u}_i$

be the displacement of atom $i$ from its average position in the trajectory and $\boldsymbol{\xi}_i^k$ be the normalized displacements of atom $i$ in the principal mode $k$. Then, the projection $P_k$ of the trajectory on the mode $k$ is

$$P_k(t) = \sum_i \mathbf{u}_i(t) \cdot \boldsymbol{\xi}_i^k. \tag{5.7}$$

Figure 5.7 displays the projection of the six-C trajectory on the first principal component (upper panel) of PHE 157 A and the associated Fourier analysis (lower panel). We observe that the spectrum peaks at 868 $\text{cm}^{-1}$, which matches the frequencies of the ring deformation of the *hot* PHEs extracted from the angle-bending time series.

Since the observed movements involve the aromatic ring of PHE, it is instructive to compare the displacement patterns of the six carbon atoms obtained from PCA to the modes of vibration of an isolated benzene molecule. Interestingly, the analyzed PHEs during the cooling process display different displacement patterns, all matching the displacement pattern of a normal mode of benzene (see figure 5.8). We note that the frequencies of the matching benzene modes (1147.8 ,1157.1 and 1200.7 $\text{cm}^{-1}$) are higher than the ones observed during the cooling process. This shift in frequency can be understood by considering that PHEs in the bulk of citrate synthase are subject to interactions with the surrounding atoms, which can be interpreted as an increased effective mass of the ring carbons. From the observed frequencies, we can estimate that the relative mass difference between the free and constrained vibrations is about 10%.

### 5.2.3. Normal mode analysis, Langevin modes and first leads on the high implication of phenylalanines among the *hot* residues

Whether the movements involve the backbone segments or the ring deformations, the frequencies extracted from the cooling simulations fall within the linear spectrum (see Appendix B.3). It is then natural to ask whether the cooling process results in the excitation of a specific normal mode of the protein, localized strongly at one of the observed hot residues. To answer this question, we computed the normal modes of the protein at atomic detail within the gromos53a6 force field. We then looked for normal modes with frequencies in the range of those observed in the quasi-stationary states during the cooling. Interestingly, a normal mode exists with a frequency $\omega = 872.4$ $\text{cm}^{-1}$ matching to a very good extent the frequencies computed in the cooling simulations and localized at PHE 313 B, *i.e.* one of the most represented hot residues. One possible explanation of energy localization would be that certain modes of vibration subsist during the cooling

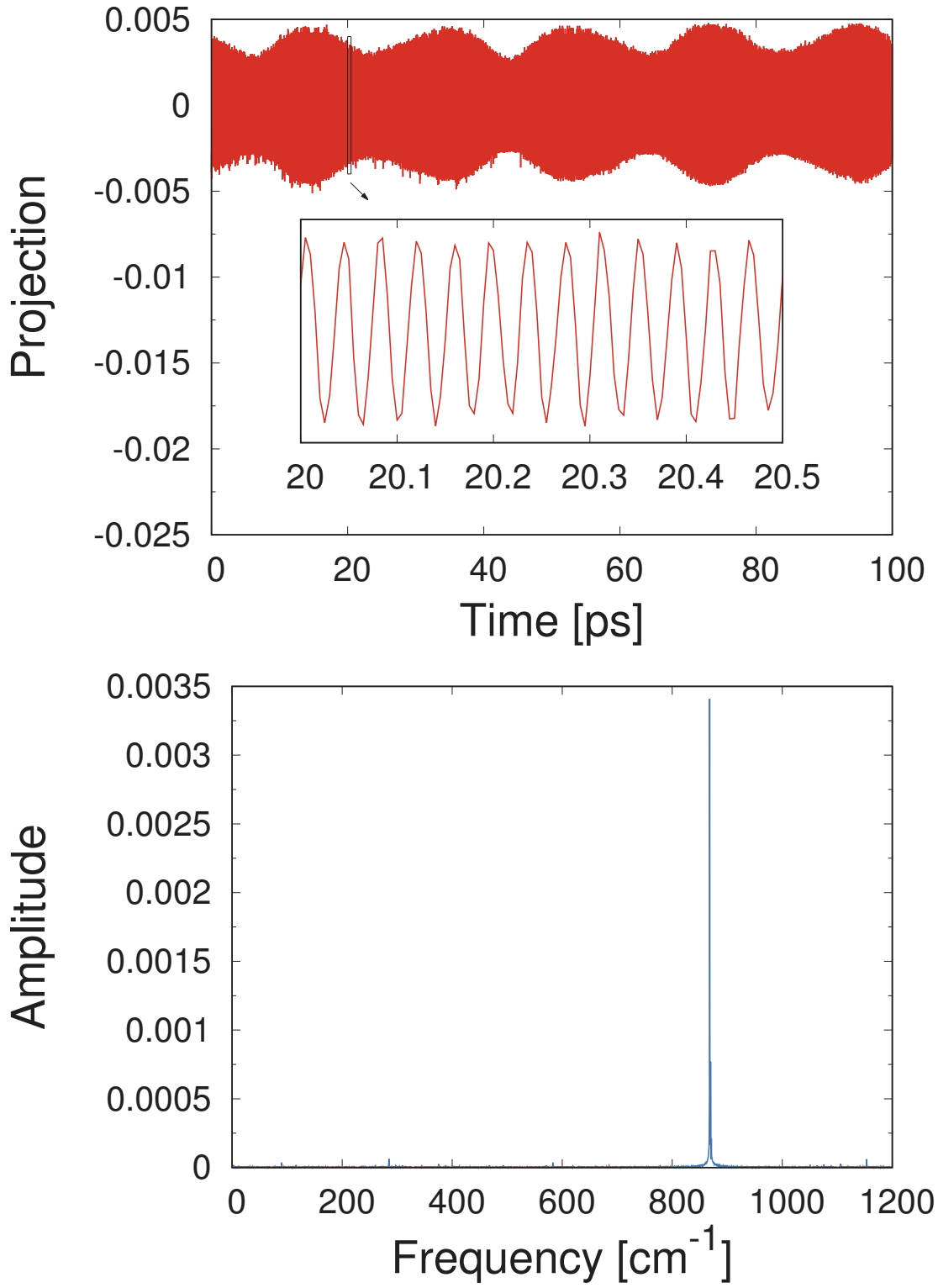



Figure 5.7.: Projection of the six-C-trajectory of the sidechain of PHE 157 A on the first principal mode (upper panel) and the associated Fourier spectrum (lower panel)

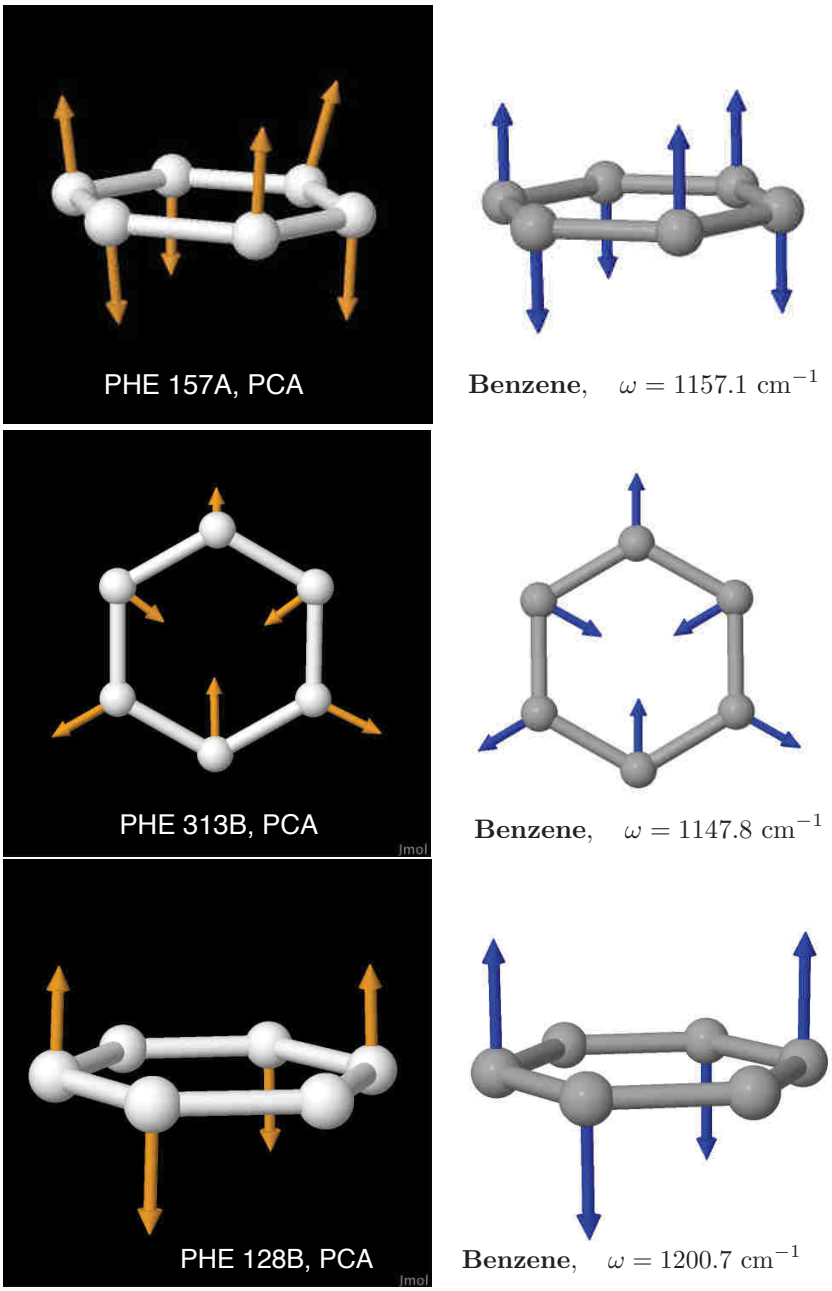



Figure 5.8.: Principal component analysis of PHE ring deformation in the quasi-stationary state. The left panels show the displacement pattern of the first principal mode, while the right panels show the corresponding normal mode pattern of an isolated benzene ring [10]. From top to bottom: PHE 157 A (a), PHE 313 B (b) and PHE 128 B (c).

and the atoms that vibrate at the frequency of the *survival* modes still move and carry the energy. However, we have not been able to find other modes that are localized on the other *hot* residues.

Another approach would be to compute the Langevin modes of vibration. Each Langevin mode has a different relaxation rate $\tau_\alpha$ (see equation 5.1). Modeling Citrate

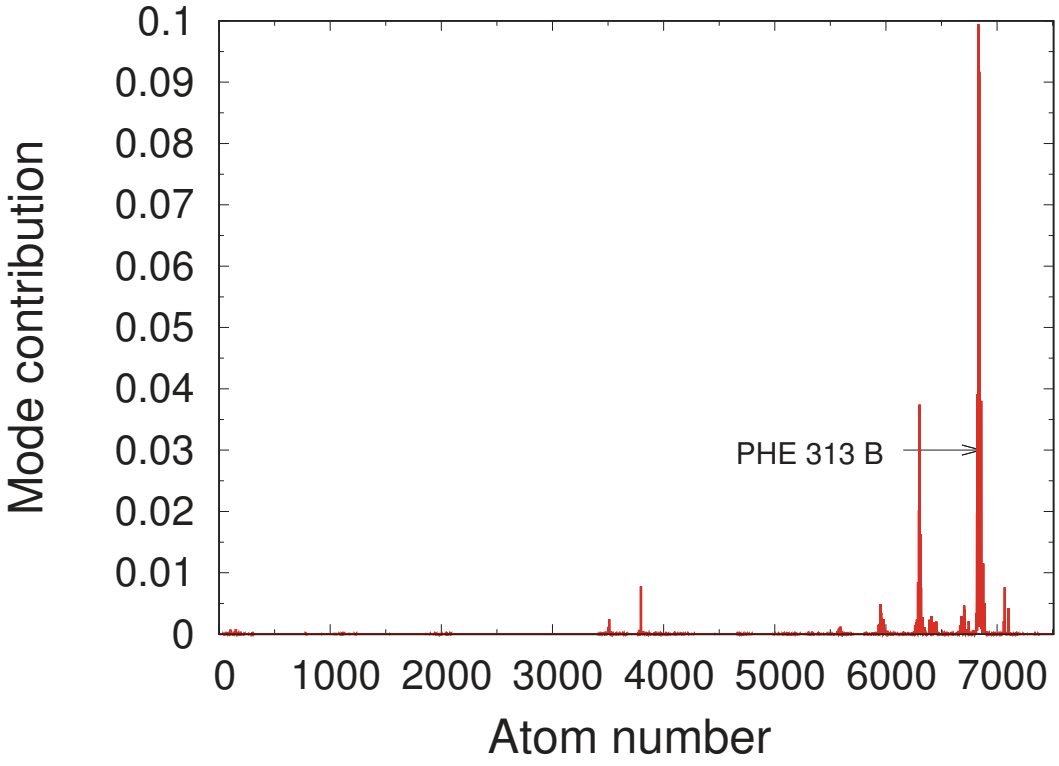



Figure 5.9.: Contribution to atomic displacements for a normal mode with a frequency of 872.4 $cm^{-1}$ (Analysis with the gromos53a6 force-field). We observe that this mode is mainly centered on PHE 313.

Synthase as an elastic network at atomic scale and adding a dissipation force on residues belonging to the protein surface, we could compute the atomic-level Langevin modes. By separating the modes of vibration that relax slowly from the others, we could expect that such modes be localized preferentially at the same hot regions that we singled out through our cooling simulation protocol.

## 5.3. Conclusion and perspectives

In this chapter, we have performed solvent cooling numerical experiments at atomic scale in order to further investigate the structural determinants of the spontaneous localization of energy observed in nonlinear coarse grained simulations [38,184]. Using the gromos53a6 force-field, we have cooled down a protein immersed in explicit solvent by extracting energy steadily from the water bath. We observed that, despite the global decay of the protein energy, some particular residues see their local temperature increase during the cooling process.

In the first part of this study, we have performed six cooling simulations starting

from the same initial condition but with different cooling rates ($\tau_T$). For three of these simulations we have spotlighted a clear crossover between an exponential decay and a more complex relaxation law in the system temperature. This crossover allows us to define a typical decay constant, that is the time constant of the relaxation of the normal mode that is damped the fastest. The relation between $\tau_1$ and $\tau_T$ allowed us to compute the effective dimension of citrate synthase from the knowledge of the surface fraction of the protein. By doing this, we obtain an effective dimension $d_{eff} \approx 1.75$, in agreement with previous studies and pointing to the existence of an universal effective reduced dimension of protein structures.

In the second part of this chapter, we have analyzed nine cooling simulations performed from as many independent initial conditions with the same cooling rate $\tau_T = 0.05$ ps. Our aim was to perform a statistical analysis of the *pool* of observed hot residues. We have thus focused our attention on the 10 *hottest* residues for each of the nine independent cooling runs. The first interesting finding was the systematic implication of phenylalanines among the hot sidechains. In particular, PHE 313 and PHE 157 showed up among the hottest residues in almost all the nine numerical experiments. Furthermore, an all-atom elastic-network analysis confirmed a result reported in reference [38] where the authors discovered that DBs tend to self-focus in the stiffest portions of citrate synthase. A measure of stiffness at atomic level allowed us to show that PHEs 313 and 157 can be indeed singled out by a purely structural indicator. Concerning the hot backbone segments, we have not been able to isolate residues that are more represented than others among the hot centers, suggesting that the local vibrations that set in on the backbone segments do not target any preferential locations in the protein scaffold.

We have then isolated the vibrational patterns that characterize the motion of the observed hot residues. Concerning the backbone segments, the kinetic energy turns out to be stored in the angle bending $C_\alpha$-N-H motions, that vibrate at an average frequency of (1439 $\pm$6) cm$^{-1}$. Concerning PHEs, the main contribution to the kinetic energy turned out to be due to the deformation of its aromatic ring at frequencies varying from 868.1 cm$^{-1}$ to 872.4 cm$^{-1}$. A principal component analysis allowed us to unveil that the modes that set in spontaneously as a result of a cooling match specific normal modes of an isolated benzene ring. However, the frequencies of these self-localizing modes are slower with respect to the corresponding modes of benzene, suggesting that the local interactions with the surrounding atoms in the bulk of the protein result in an increased effective mass (about 10% heavier).

In order to understand the high abundance of phenylalanines, we performed a normal mode analysis with the aim of identifying modes that vibrate at frequencies close to the observed frequencies of ring deformations. Interestingly, we observed that mode 13250

with a frequency of 872.4 $cm^{-1}$ is strongly localized on PHE 313 B, that is a residue well represented among the *hottest* residues. However, we have not been able to isolate other modes of vibration that would flag other *hot* residues in the range $863 < \omega < 877$ $cm^{-1}$.

Further studies could involve the computation of the Langevin modes of vibration of citrate synthase modeled as an elastic network (à la Tirion) coupled to the protein surface through an external bath. By analyzing the relaxation rates associated with each mode of vibration, we could expect that the modes that relax the slower are localized on the particular phenylalanines that we find among the *hot* residues. If this is the case, this would suggest that a cooling simulation at atomic scale performed with a realistic force-field and an explicit solvent can be mimicked by an elastic network model coupled to an external bath at the protein surface.

# 6. General conclusions

In this thesis we have focused on the elusive relation that exists in proteins between their complex structures and the even more complex and sophisticated functions that they perform.

Our first project was to construct indicators with the aim of predicting active sites as *hotspots* in enzymes. We have tested our analysis on a set of 835 enzymes where the catalytic sites were known experimentally. We have used for this purpose three different structural indicators, the connectivity and the closeness centrality computed from the connectivity graph and the stiffness, an indicator calculated through a normal mode analysis of protein fluctuations. These three different figures of merit depend on a single parameter, the cutoff radius that defines which residues interact with which in the network of beads and springs associated with the protein scaffold. By varying this parameter, we have remarked that there exist indicator-dependent optimal values corresponding to a maximum of prediction power. These cutoff values were 20 Å for the connectivity, 22 Å for the stiffness and 28 Å for the closeness centrality. As a combined strategy, we have proposed a specific algorithm that makes use of the three indicators in a given sequential manner. By doing this, we have been able to predict almost 70% of the catalytic sites of the database within a distance of 2 amino-acids along the sequence. Our method should of course be used in conjunction with other existing methods derived from structural and sequence analysis. One of the obvious advantages of our strategy is that it has a strongly reduced computational cost.

The second project illustrated in this thesis had the aim of elucidating intramolecular *communication* in proteins such as observed in allostery. To this aim, we have set up a simulation scheme where a given protein structure is subject to a local oscillating force, whose frequency and direction of application can be varied, while a tiny damping rate applied uniformly across the structure ensures that a steady state be reached asymptotically.

A case study where we tested this simulation protocol on the molecular motor kinesin allowed us to apprehend the salient features of such non-equilibrium process. In particular, this clearly pointed us to the existence of a specific type of normal modes that seem to

have been largely overlooked by previous analyses. Such modes, that we termed *bilocalized* are typically intermediate-frequency modes ($\omega$ of the order of 10-20 $cm^{-1}$) and yet their vibrational patterns are strongly localized, with significant components typically involving two distinct and often widely separated portions of the protein scaffolds. As a general observation, the widest the separation between two distinct localization centers, the slower the frequency of the bilocalized mode.

We have then elaborated an original strategy to isolate potentially interesting (functional) bilocalized modes in a large pool of structures. For each structure, we have isolated a single bilocalized mode among the pool of normal modes localized on few residues, namely the one featuring the highest interdistance (in real space) between the two major localized components.

Pump simulations performed at either localized component along these normal modes allowed us to show that long energy transfer occur mediated by bilocalized modes rather generally and are distance-dependent. An analysis on modes with localization centers lying at less than 20 Å from each other showed that energy transfers mediated by these modes are more efficient the greater the relative contribution of the two peaks to the overall (normalized) displacement pattern. The scenario turned out to be different for bilocalized modes with peaks separated by more than 20 Å. In this case, a certain amount of *noise* in the mode pattern is required for an efficient transfer along the mode, *i.e.* the relative contribution to the mode pattern of non-peak (intermediate) regions cannot be vanishingly small. We have found that energy can be channeled to distances as great as 70 Å away from the pumped residue through this kind of bilocalized modes. Incidentally, the analysis of these *noisy* components also points to an operative strategy to pinpoint energy transduction pathways connecting distant localization centers.

In view of these observations, we turned our analysis to inquire whether a mechanism of this sort might shed some light on the open problem of intramolecular communication across the cell membrane in GPCR structures. An analysis of fifteen different GPCRs belonging to different functional classes revealed that a specific bilocalized mode exists for each candidate, connecting through the associated correlated displacement pattern two residues lying close to either end of the bilayer membrane. Our pump simulation scheme revealed that such modes are effectively able to channel energy from the extracellular side to the intracellular side.

In the final part of this thesis, we have illustrated the main results of our last project, aimed at elucidating at the atomic level the intriguing phenomenon of spontaneous energy localization in many-body systems. To this aim, we have set up a relaxation protocol where a protein, immersed in explicit water, and described through an atomistic force-

field, is cooled down *through* the solvent. This has been achieved by a simple algorithm where the velocity of the solvent molecules are gradually reduced with a prescribed time constant, $\tau_T$. By doing this, the protein is also cooled down, although in an indirect way, *i.e.* through the glassification of its exterior hydration shells.

A statistical analysis of different independent cooling simulations has revealed an intriguing feature of the relaxation process. First of all, our atomistic simulations show that spontaneous energy localization occurs generically as a result of surface cooling at the atomistic level, confirming the results of similar simulations performed with much simpler models. However, the scenario that we observe is a much subtler one. We find energy localization both at the level of backbone segments and in the sidechains. Typically, in the backbone angle-bending modes involving $C_\alpha$-N- H groups are excited, with frequencies of the order of 1400 $cm^{-1}$. These vibrations do not seem to be site-specific and we find them localized at many different locations over the ensemble of our independent cooling runs. The situation is dramatically different when we look at side-chains. In this case, we find a strong bias towards the excitation of ring modes of the aromatic rings of a few phenylalanine residues at about 870 $cm^{-1}$. A principal component analysis performed on selected PHE residues in the quasi-stationary state revealed that the excited ring motions match specific normal modes of an isolated benzene molecule, although with a slower frequency that can be interpreted as an increased effective mass of the carbon atoms in the ring

An elastic-network model analysis performed at atomic level confirmed that such residues were among those flagged by a local stiffness analysis. When we looked at the *true* normal modes (*i.e.* the ones obtained by diagonalizing the atomistic Hessian) we found a normal mode with the same frequency indeed localized at one of the most represented PHEs in our ensemble of simulations.

Further developments of this work might include a Langevin mode analysis within the framework of an elastic network model at the atomistic level and more careful network analysis to inquire into the predictive power of specific graph-associated indicators.

# A. About the pump simulations

## A.1. Pump databases

In chapter 4, we have reported the analysis of pump simulations performed on a selection of structures with specified content of secondary structure. Our databases were constructed so as to contain structures with less than 10% of $\beta$ sheets (ALL-$\alpha$ database) and with less than 10% of $\alpha$ helix content (ALL-$\beta$ database). Additionally, we required that the structures within each data set contain representatives at more than 95% sequence identity. This means that multiple structures whose sequences have at least 95% sequence identity are represented by a single structure in our database set. The actual content of secondary structure in both databases is shown in figure A.1.

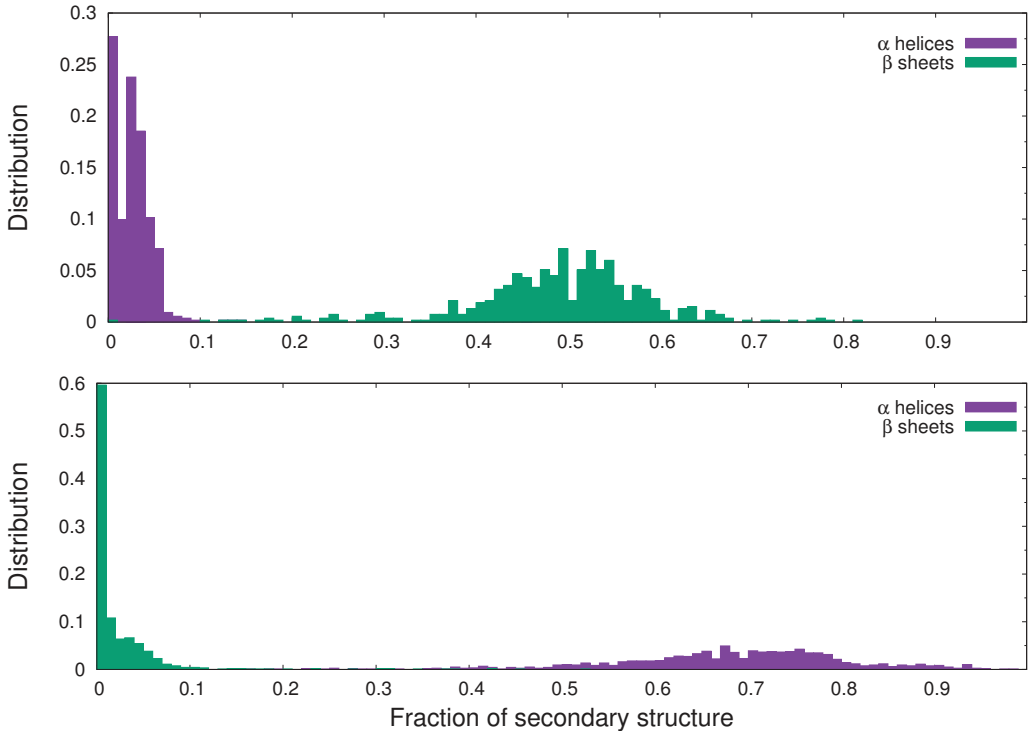



Figure A.1.: Histograms of the fraction of secondary structure in the *all* $\beta$ database (upper panel) and in the *all* $\alpha$ database (lower panel). The secondary structure content has been computed through the DSSP algorithm [11] with an energy threshold of 0.5 kcal/mol.

We observe that the fraction of $\alpha$ helices in the *all* $\beta$ database is smaller than 0.1 and that the fraction of $\beta$ sheets in the *all* $\alpha$ database in also smaller than 0.1. Moreover, we observe clearly for the *all* $\beta$ database that on average more than half the proteins have $\beta$ sheets in more than 50% of the structure. This proportion is even higher for the $\alpha$ helices in the *all* $\alpha$ database.

## A.2. GPCR simulations

In this appendix, we illustrate in detail the pump simulations on the GPCR structures whose results are presented in tables 4.2, 4.3 and 4.4. For each GPCR structure, we show the displacement pattern of the particular mode connecting the two residues lying at either ends of the bilayer membrane and the normalized average local energy field in the steady state (equation 4.1).

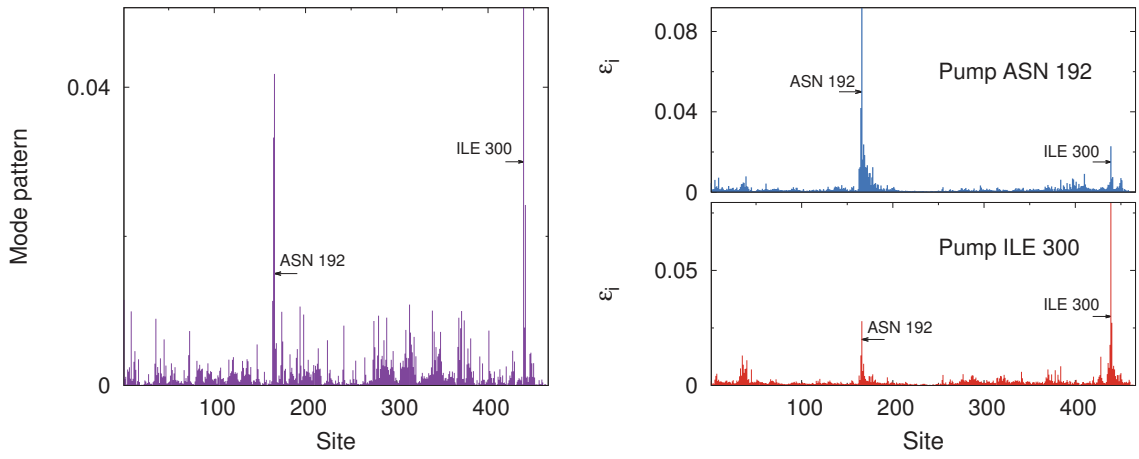



Figure A.2.: CXCR4 (Chemokine receptors) (PDB code : 3ODU)

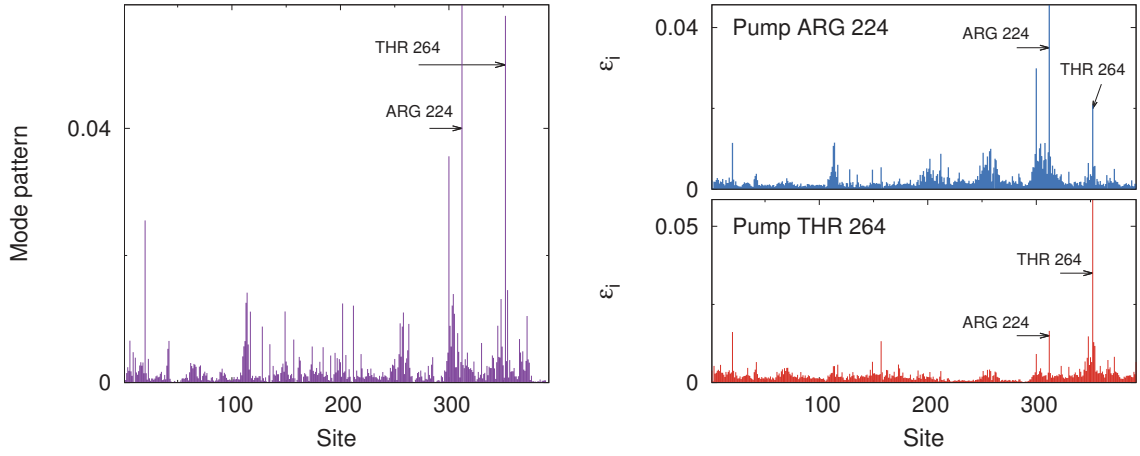



Figure A.3.: P2Y12 (P2Y receptors) (PDB code : 4PXZ)

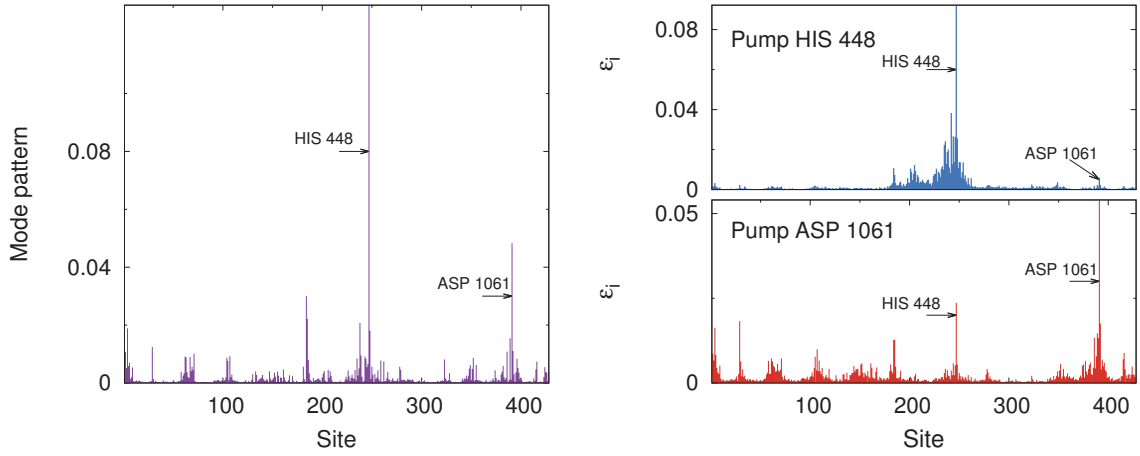



Figure A.4.: H1 receptor (Histamine receptors) (PDB code : 3RZE)

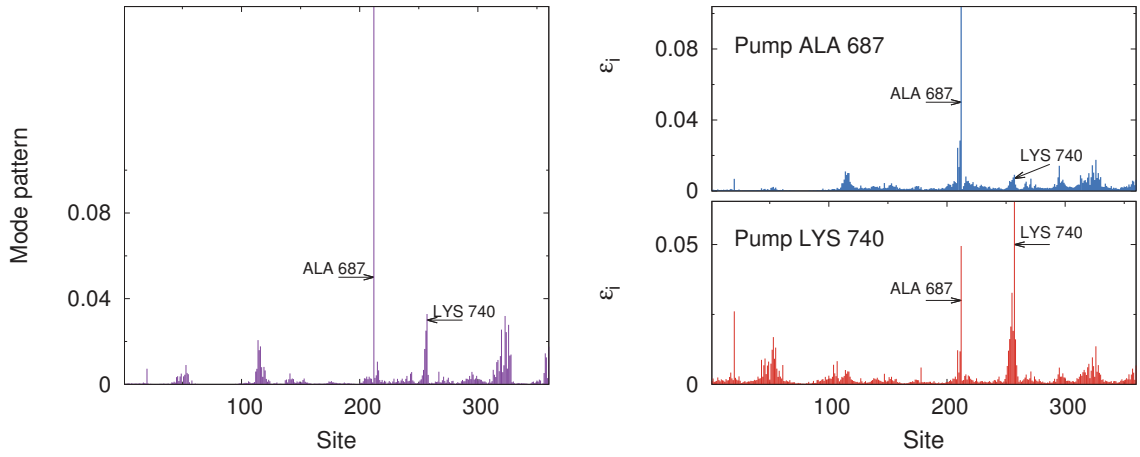



Figure A.5.: mGlu1 receptor (Metabotropic glutamate receptors) (PDB code : 4OR2)

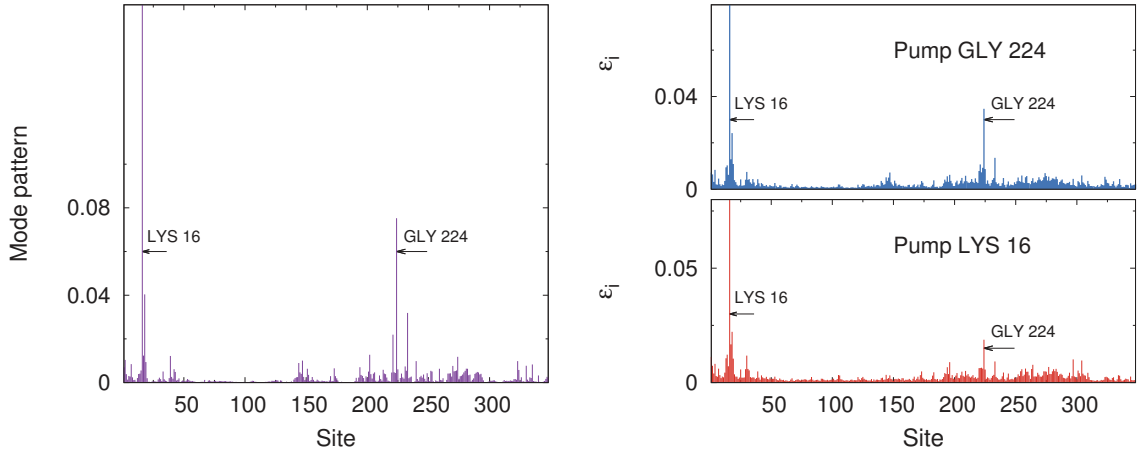



Figure A.6.: Bovine rhodopsin (PDB code : 1U19)

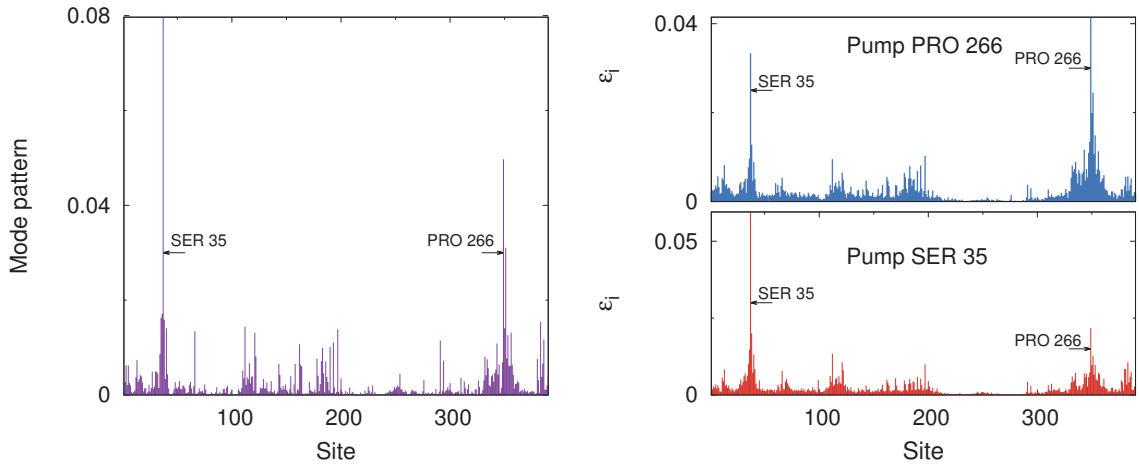



Figure A.7.: A2A receptor (Adenosine receptors) (PDB code : 4EIY)

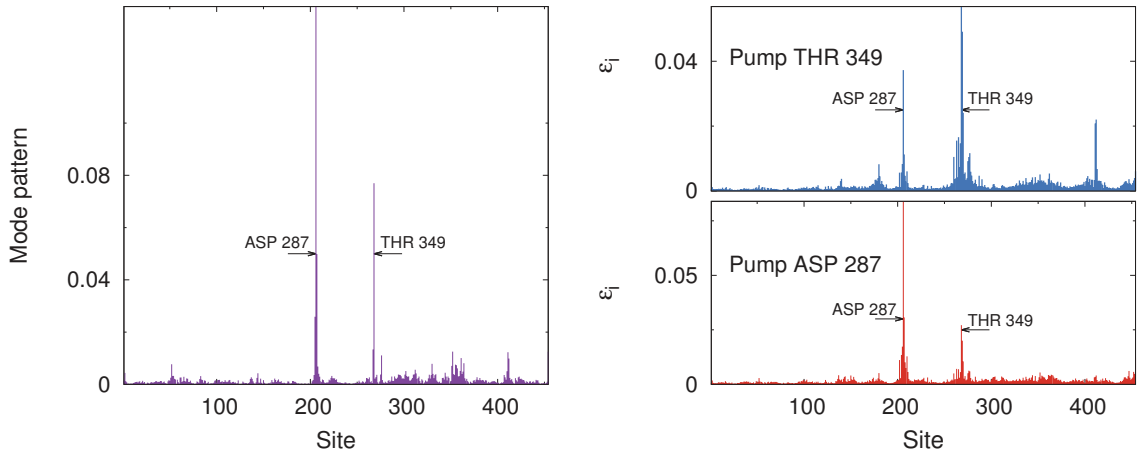



Figure A.8.: SMO (Frizzled) (PDB code : 4JKV)

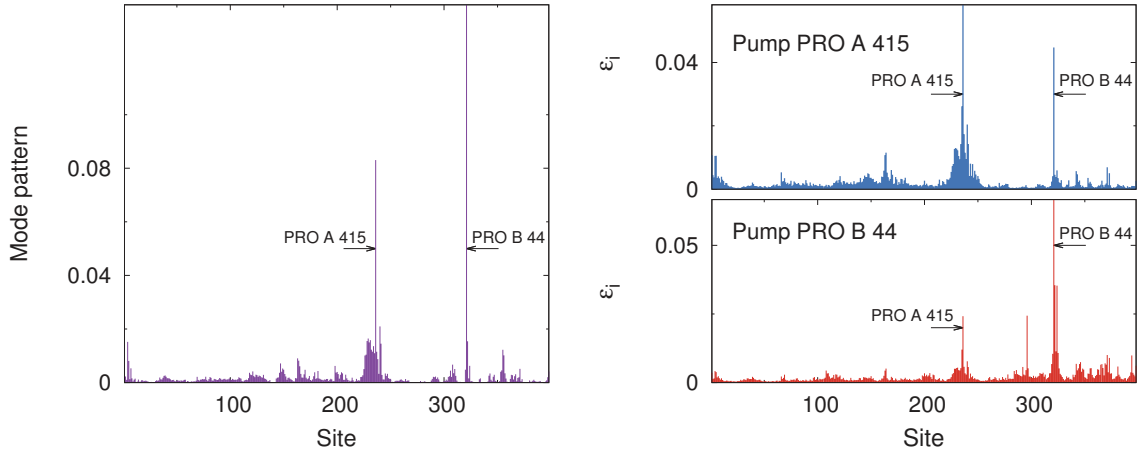



Figure A.9.: M2 receptor (Acetylcholine receptors (muscarinic)) (PDB code : 4MQS)

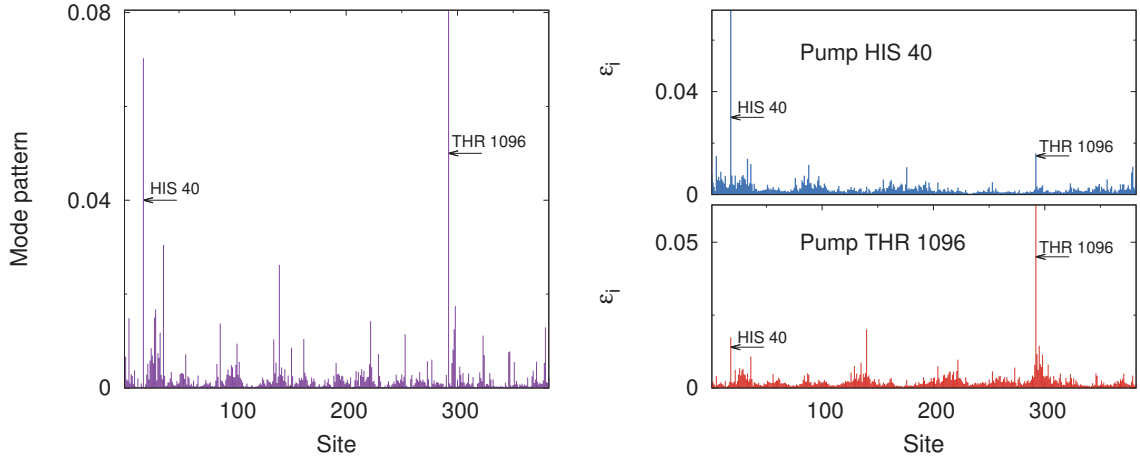



Figure A.10.: LPA1 receptor (Lysophospholipid (LPA) receptors) (PDB code : 4Z35)

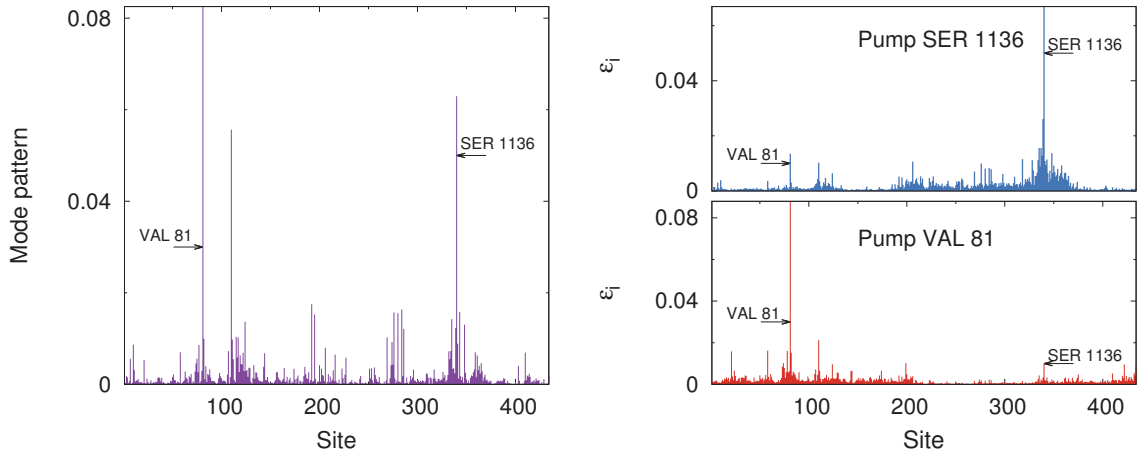



Figure A.11.: FFA1 receptor (Free fatty acid receptors) (PDB code : 4PHU)

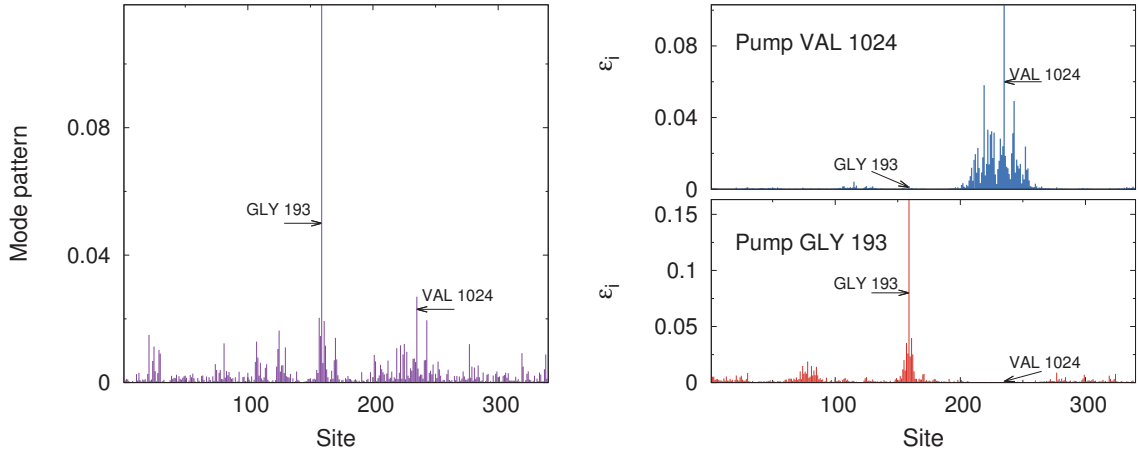



Figure A.12.: P2Y1 receptor in complex with an antagonist (P2Y receptors) (PDB code : 4XNV)

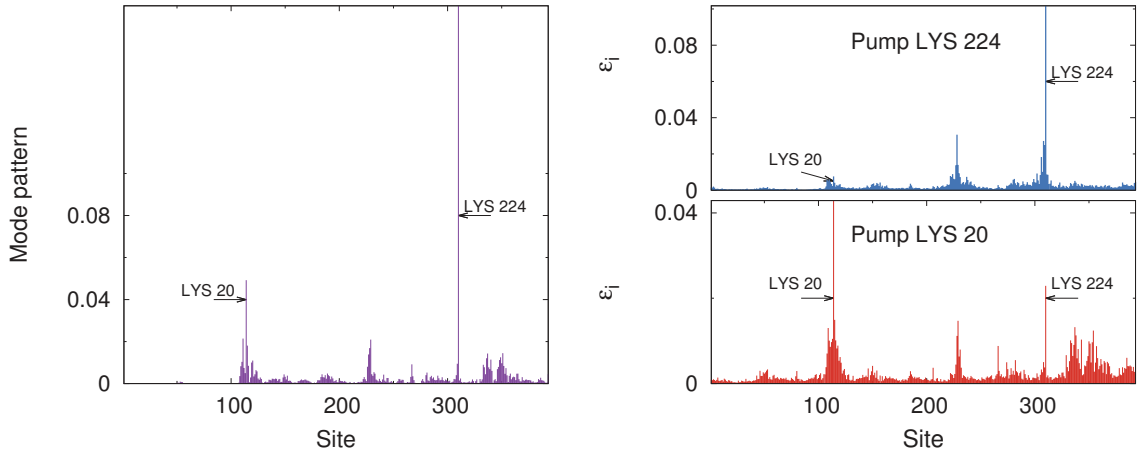



Figure A.13.: AT1 receptor (Angiotensin receptors) (PDB code : 4YAY)

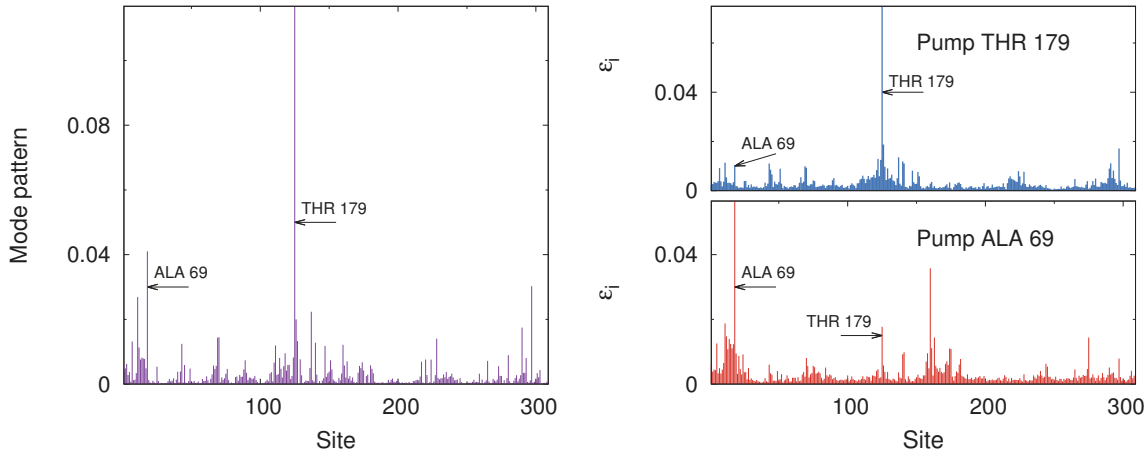



Figure A.14.: NTS1 receptor (PDB code : 4BUO)

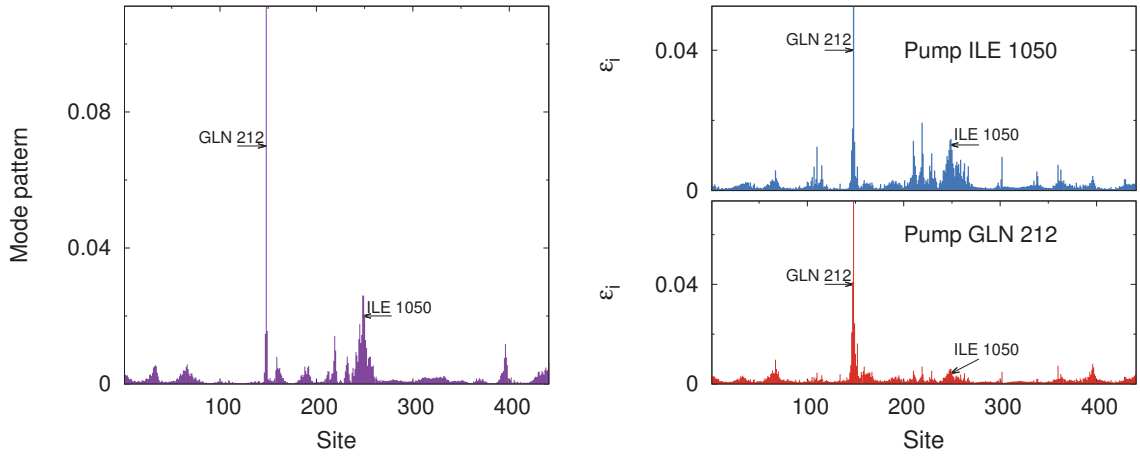



Figure A.15.: $\mu$ receptor (PDB code : 4DKL)

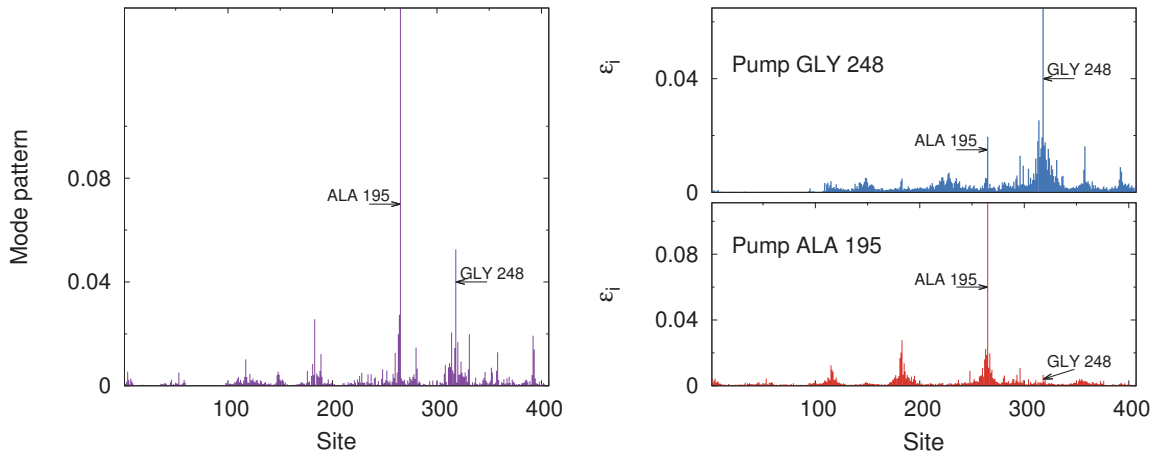



Figure A.16.: $\delta$ receptor (PDB code : 4N6H)

# B. About cooling simulations

## B.1. Surface fraction and effective dimension

In the first part of chapter 5, we have analyzed the effect of varying the cooling rate on the global temperature relaxation. Figure B.1 shows the time scale of the crossover from an exponential decay to a power-law relaxation extracted from the simulations as a function of the cooling time constant.

A linear fit of the form $\tau_1 = \alpha\tau_T + \beta$ yields $\alpha = 6.76 \pm 0.02$ and $\beta = 2.07 \pm 0.09$ ps.

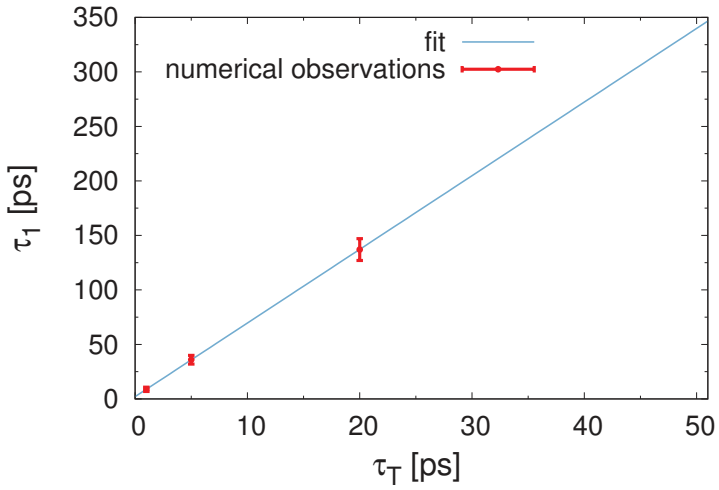



Figure B.1.: Exponential-to-power-law crossover time for the energy relaxation of citrate synthase immersed in water cooled down at a rate $\tau_1^{-1}$ (see figure 5.1).

## B.2. Phenylalaline vibrations

In chapter 5, we have shown that cooling citrate synthase at a given rate through the solvent from different initial configurations resulted in an over-representation of phenylalanine rings among the self-localizing hot spots. Figure B.2 shows the frequency of the ring-buckling mode extracted from the simulations over the entire pool of nine different initial conditions.

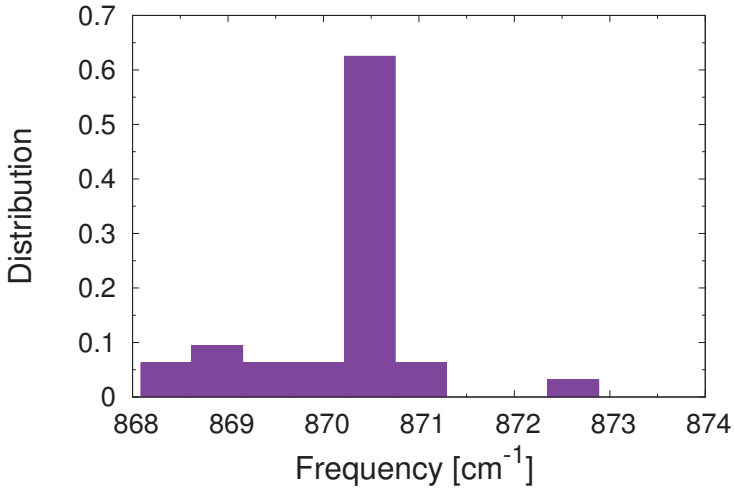



Figure B.2.: Histograms of the frequencies of the phenylalanine ring-buckling modes extracted from the simulations for all the nine independent cooling runs. The frequencies were extracted from the time series of angle-bending motions in the aromatic rings.

## B.3. Linear spectrum

In this appendix, we show the linear spectrum of citrate synthase (PDB 1IXE) obtained by diagonalizing the atomic-level Hessian matrix computed with the gromos force-field. We remark that the frequencies observed during our analyses indeed belong to the linear spectrum (figure B.3).

## B.4. Cooling simulation representations

In this appendix, we illustrate the relaxation of local temperature for the backbone segments and sidechains for all simulations discussed in chapter 5. For the runs performed with $\tau_T = 0.05$ ps from different initial configurations we only show the time evolution in a window corresponding to the quasi-stationary state, *i.e.* for time longer than 400 ps.

### B.4.1. Different damping rates

### B.4.2. Independent initial conditions

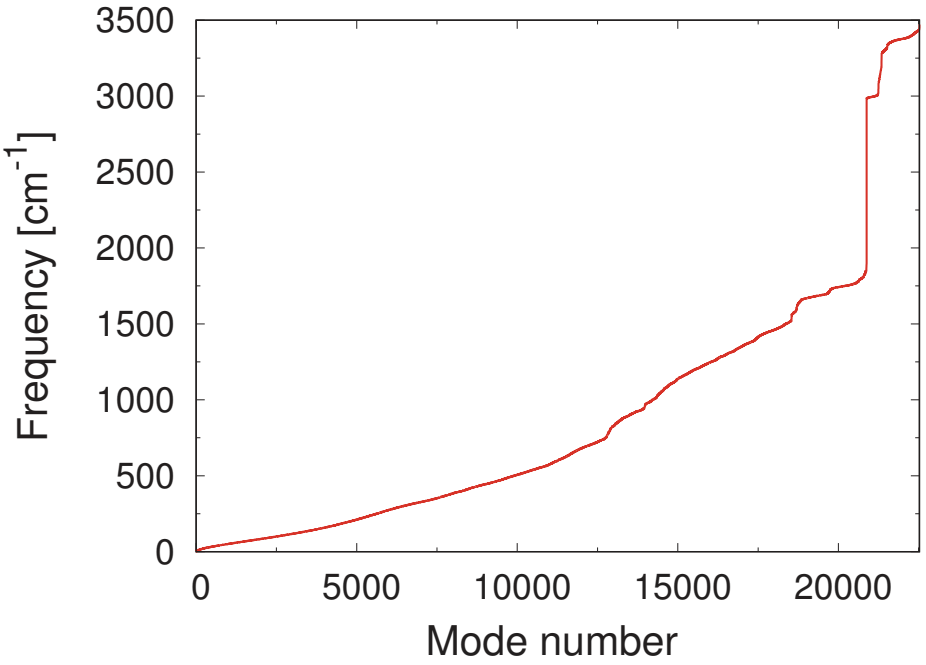



Figure B.3.: Linear spectrum of citrate synthase computed by diagonalizing the atomic Hessian matrix.

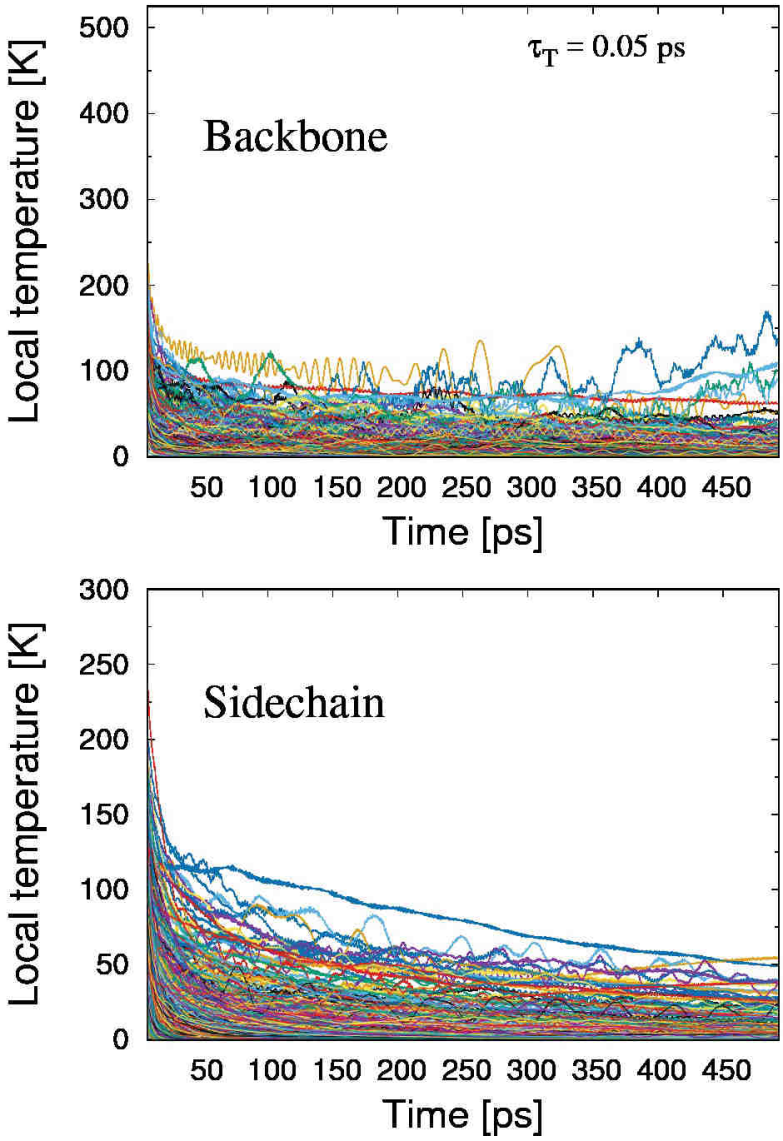



Figure B.4.: Average local temperature (computed over a stretch of 15 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps.

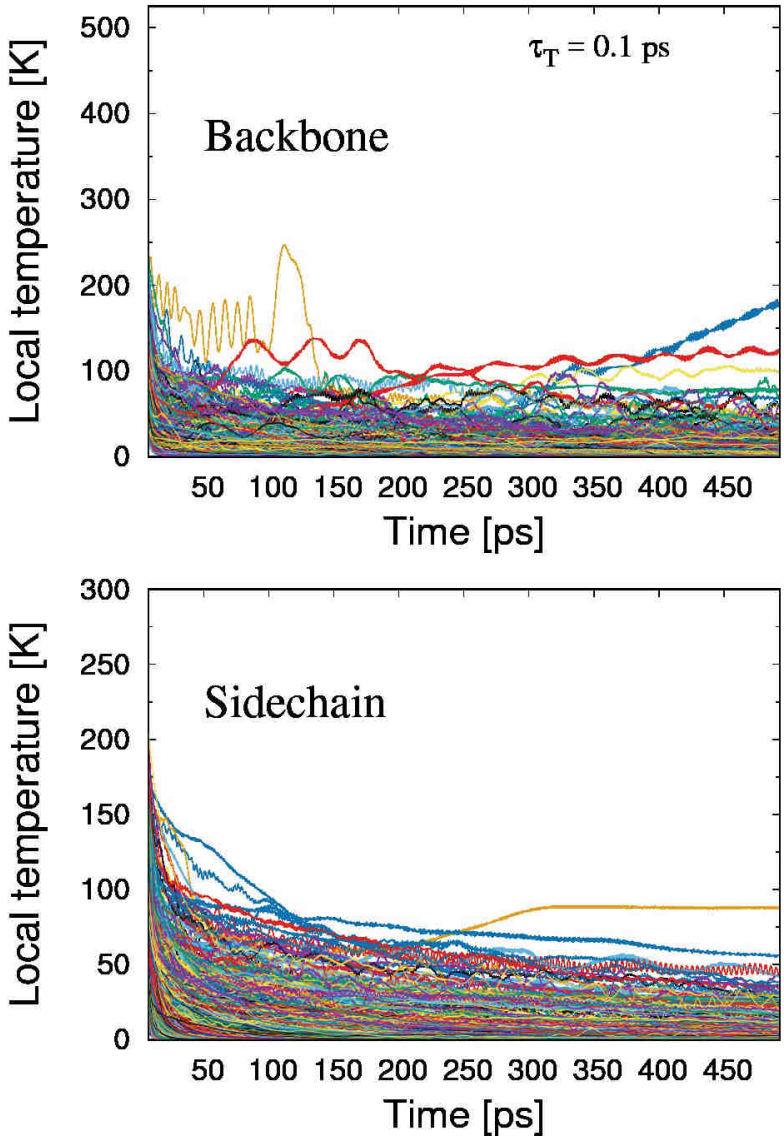



Figure B.5.: Average local temperature (computed over a stretch of 15 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.1$ ps.

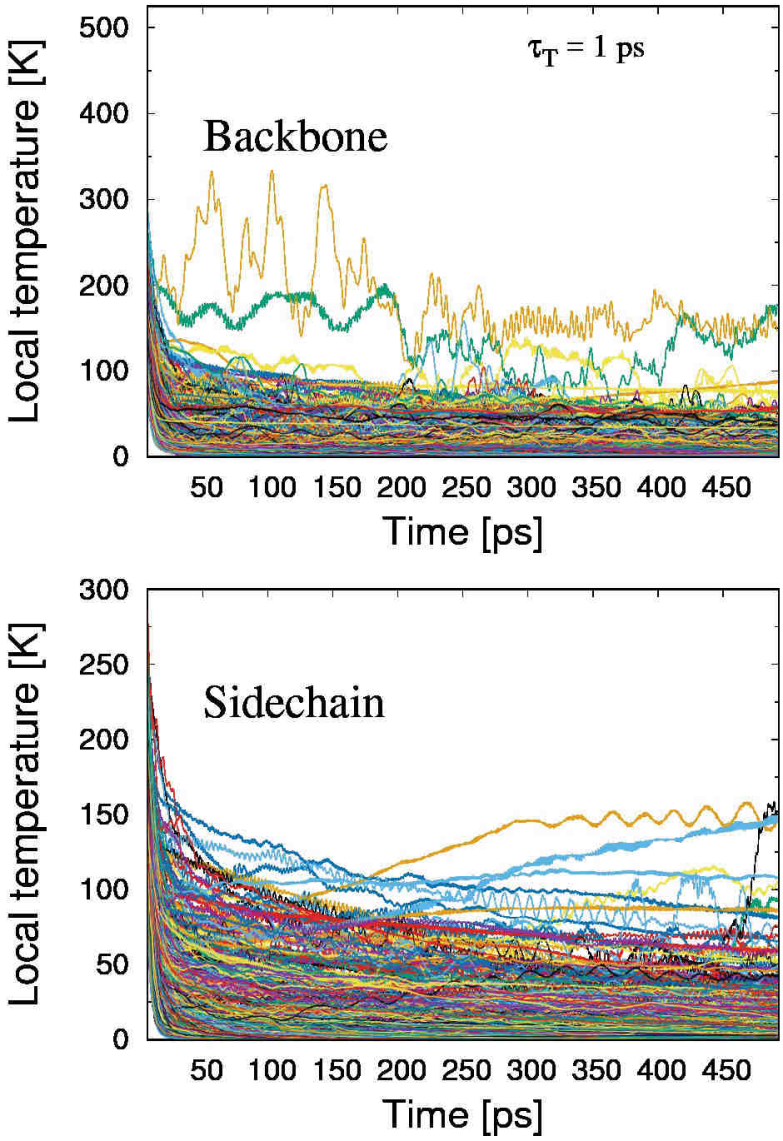



Figure B.6.: Average local temperature (computed over a stretch of 15 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 1$ ps.

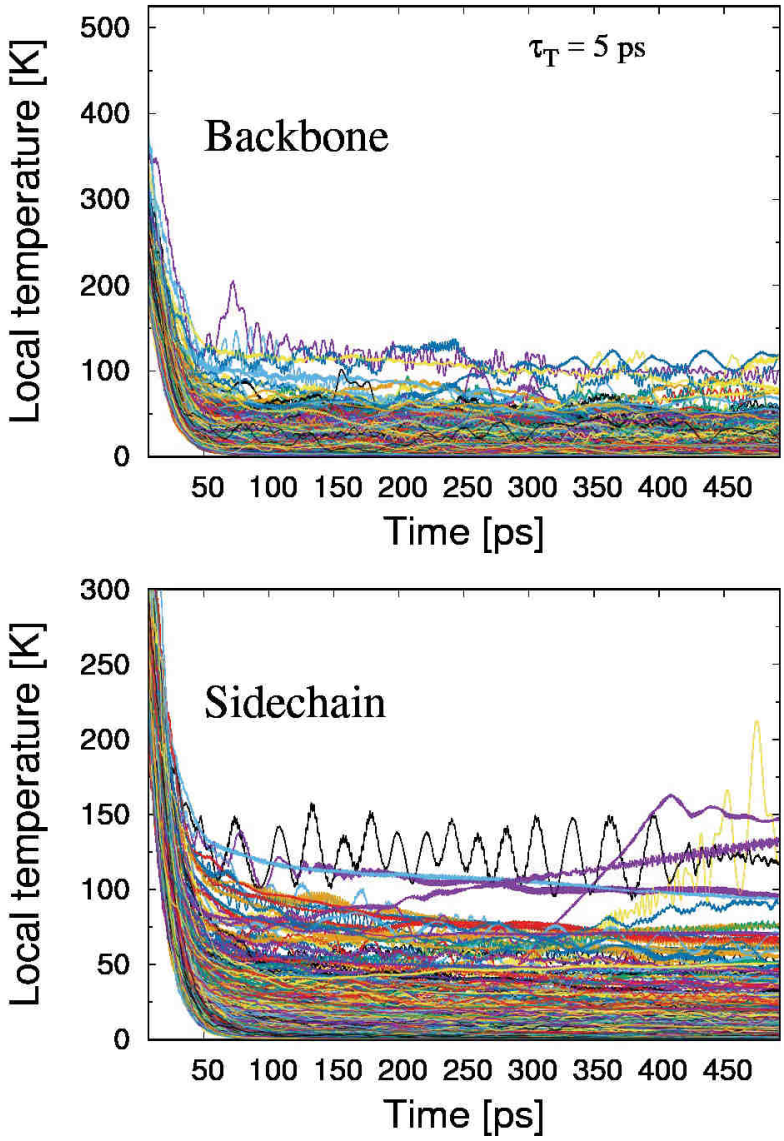



Figure B.7.: Average local temperature (computed over a stretch of 15 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 5$ ps.

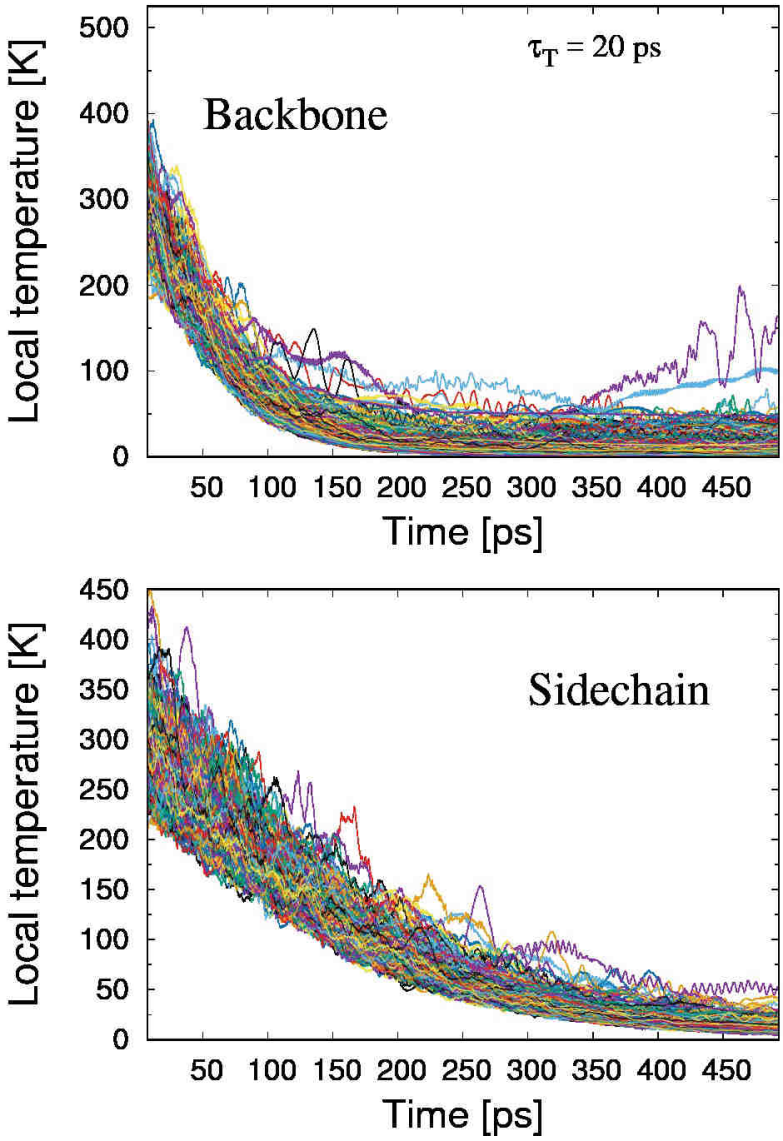



Figure B.8.: Average local temperature (computed over a stretch of 15 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 20$ ps.

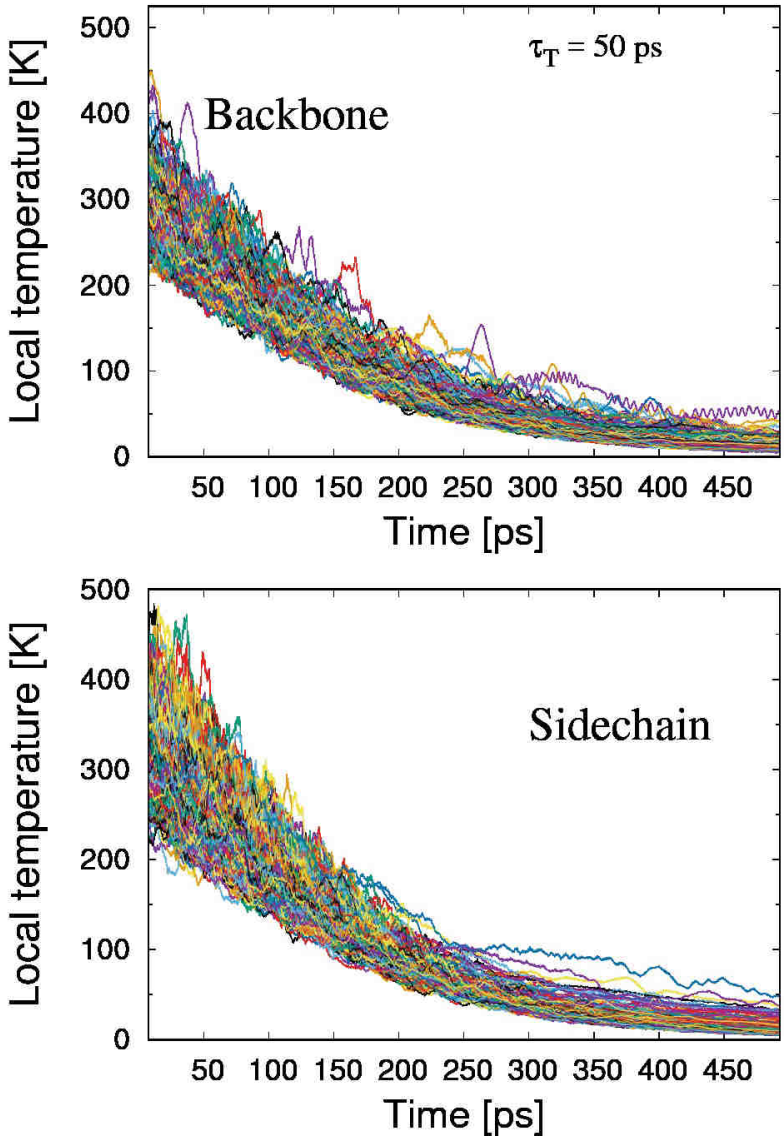



Figure B.9.: Average local temperature (computed over a stretch of 15 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 50$ ps.

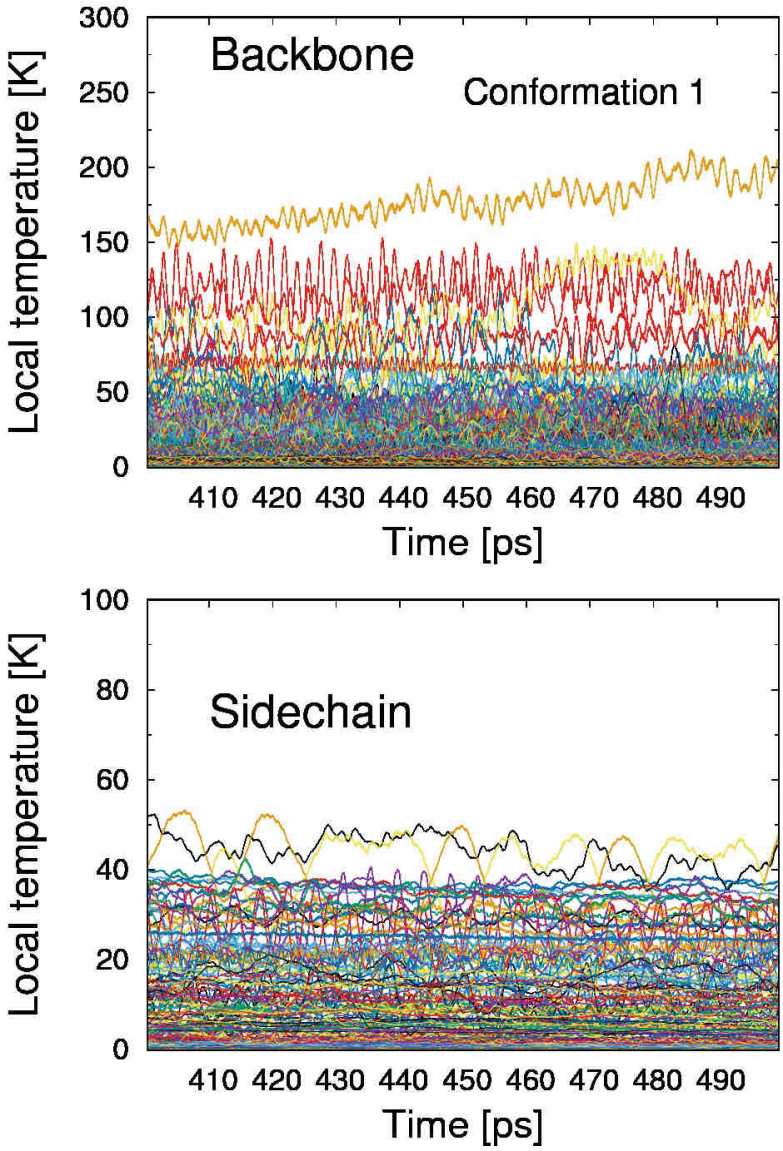



Figure B.10.: Average local temperature (computed over a stretch of 0.5 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps starting from conformation 1. Only the time evolution in the quasi-stationary state is shown here.

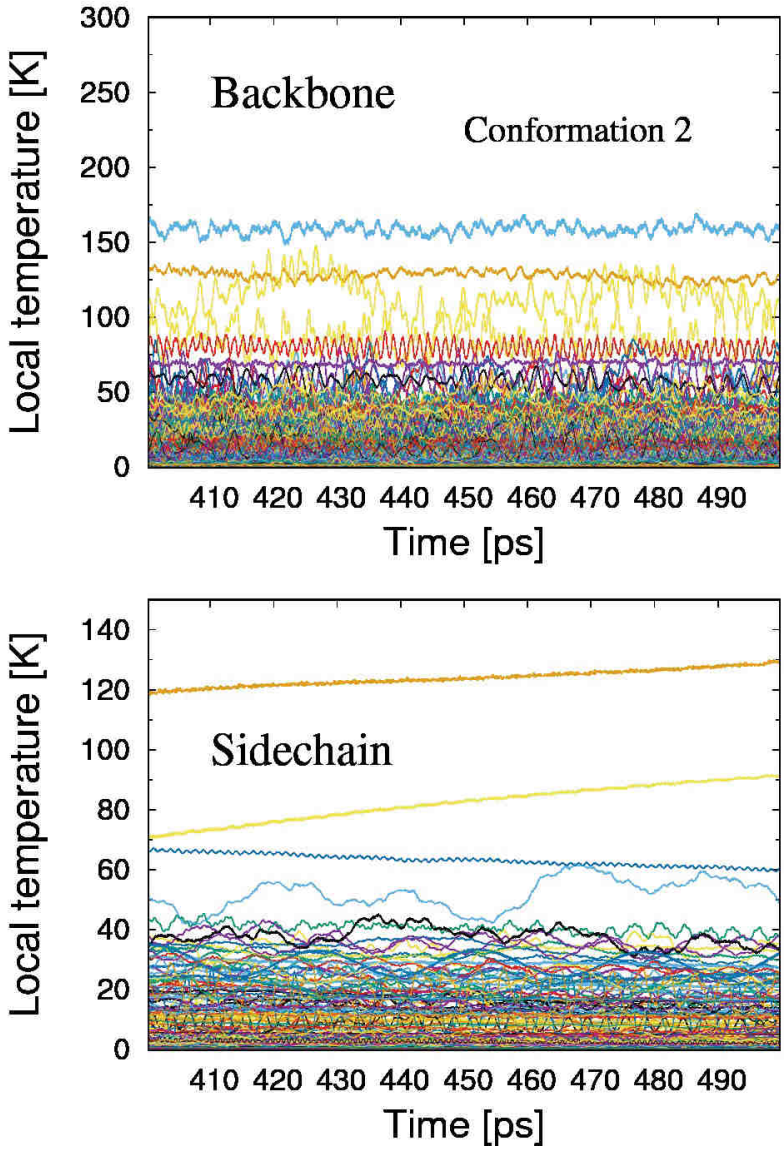



Figure B.11.: Average local temperature (computed over a stretch of 0.5 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps starting from conformation 2. Only the time evolution in the quasi-stationary state is shown here.

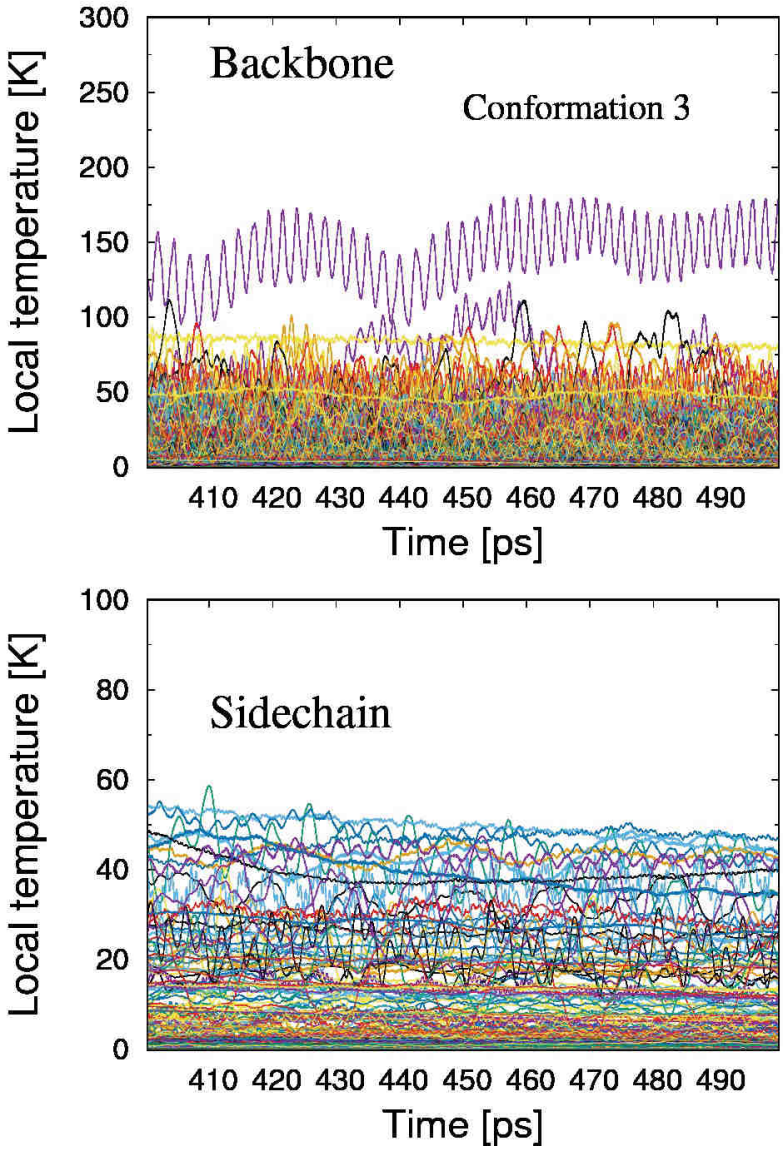



Figure B.12.: Average local temperature (computed over a stretch of 0.5 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps starting from conformation 3. Only the time evolution in the quasi-stationary state is shown here.

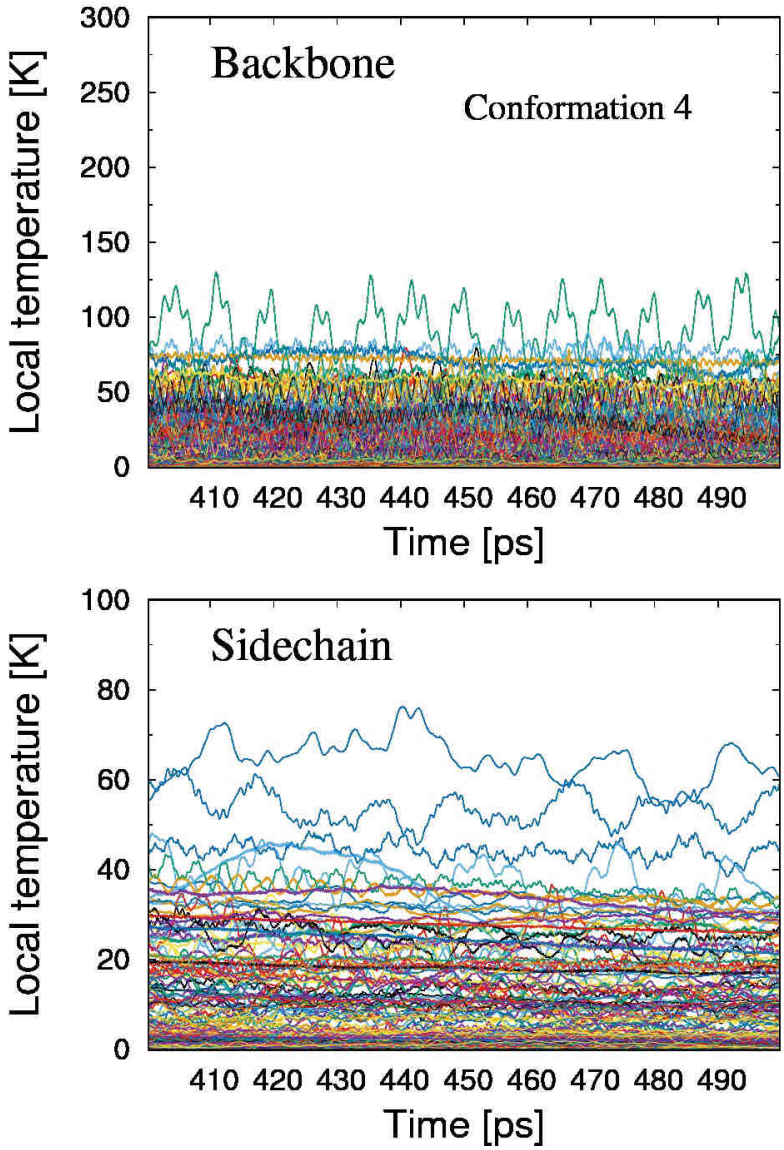



Figure B.13.: Average local temperature (computed over a stretch of 0.5 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps starting from conformation 4. Only the time evolution in the quasi-stationary state is shown here.

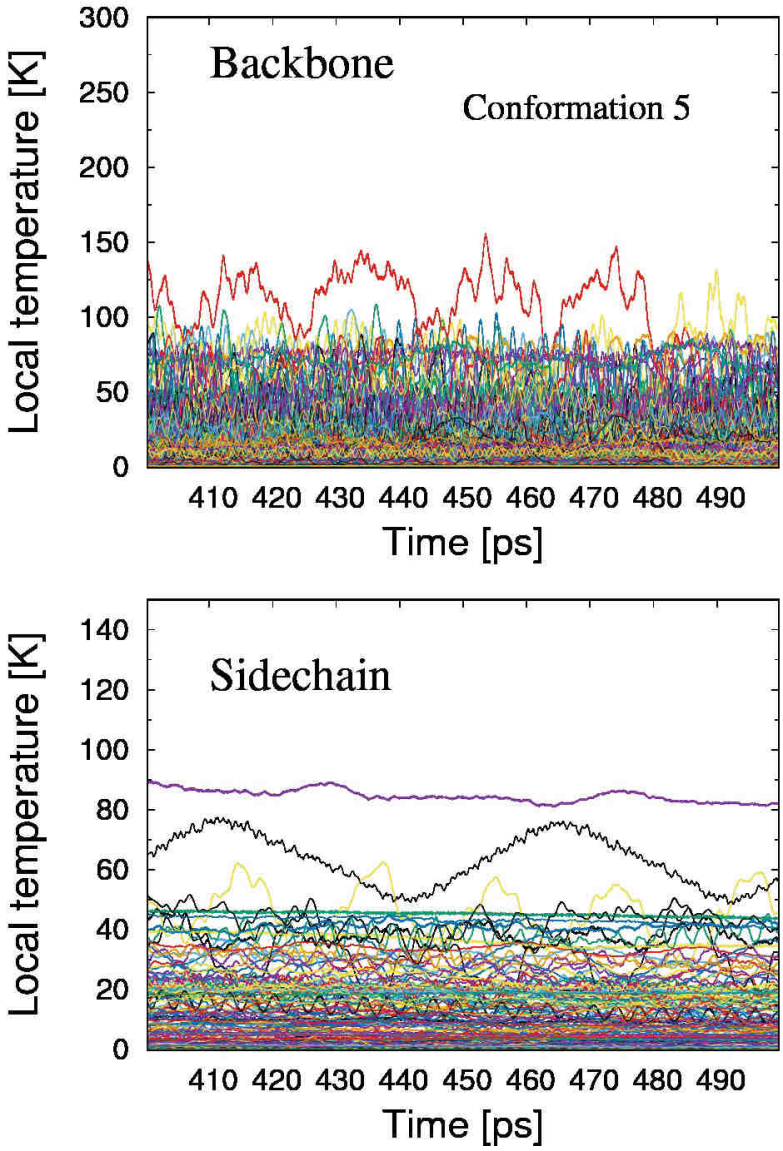



Figure B.14.: Average local temperature (computed over a stretch of 0.5 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps starting from conformation 5. Only the time evolution in the quasi-stationary state is shown here.

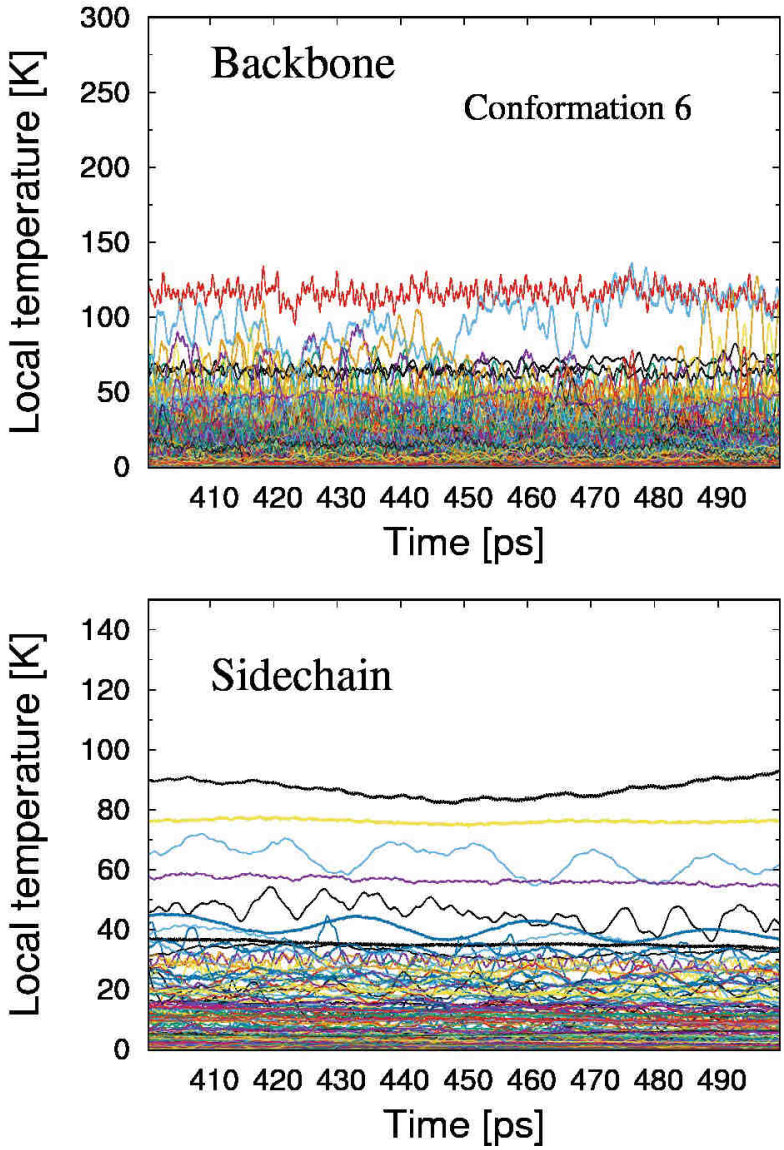



Figure B.15.: Average local temperature (computed over a stretch of 0.5 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps starting from conformation 6. Only the time evolution in the quasi-stationary state is shown here.

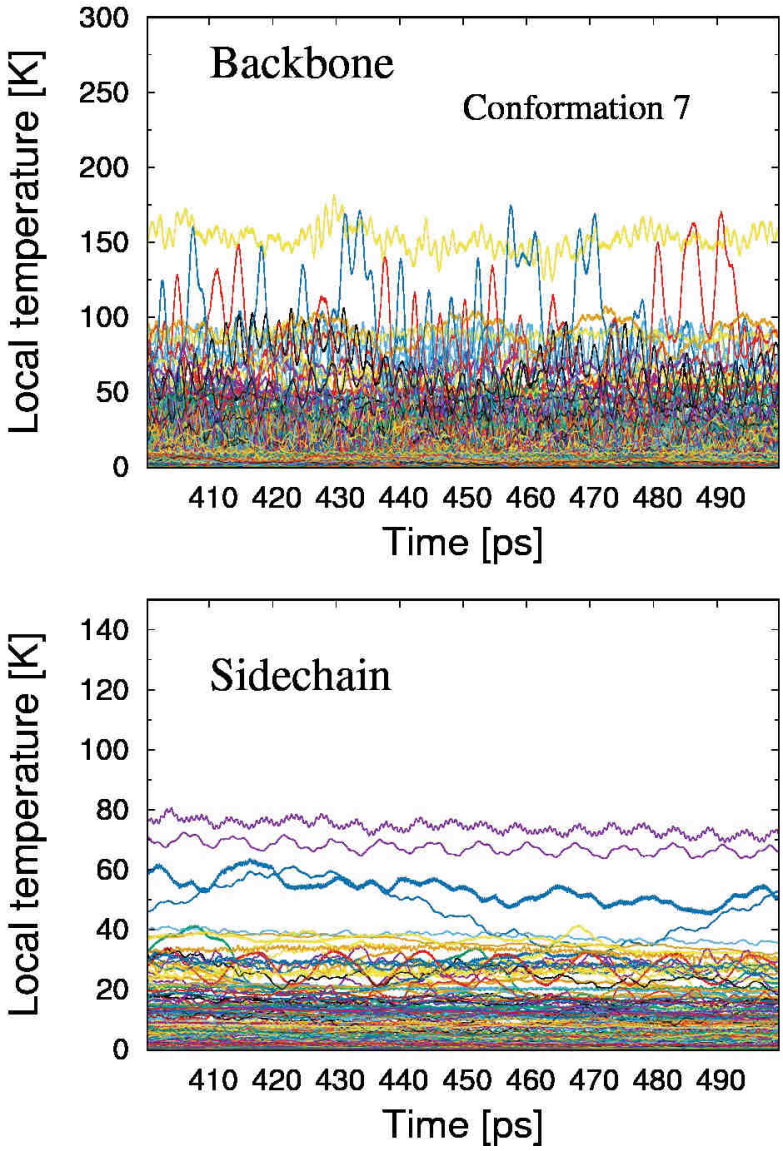



Figure B.16.: Average local temperature (computed over a stretch of 0.5 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps starting from conformation 7. Only the time evolution in the quasi-stationary state is shown here.

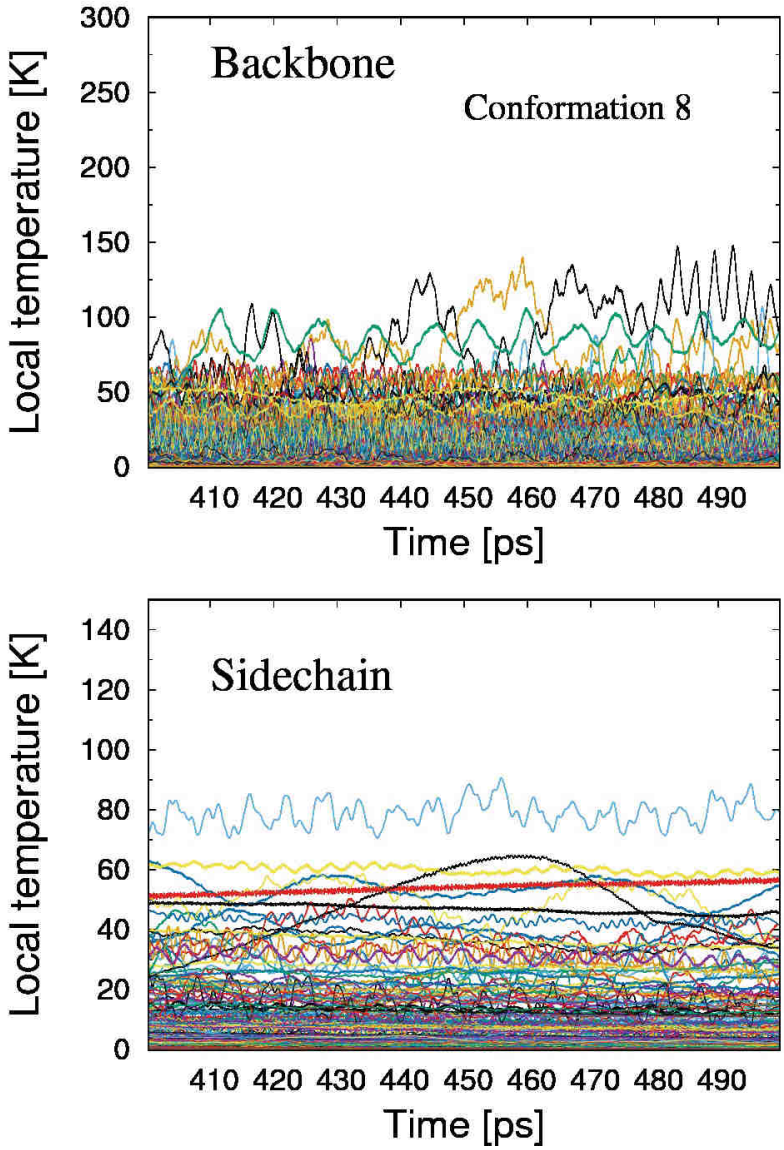



Figure B.17.: Average local temperature (computed over a stretch of 0.5 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps starting from conformation 8. Only the time evolution in the quasi-stationary state is shown here.

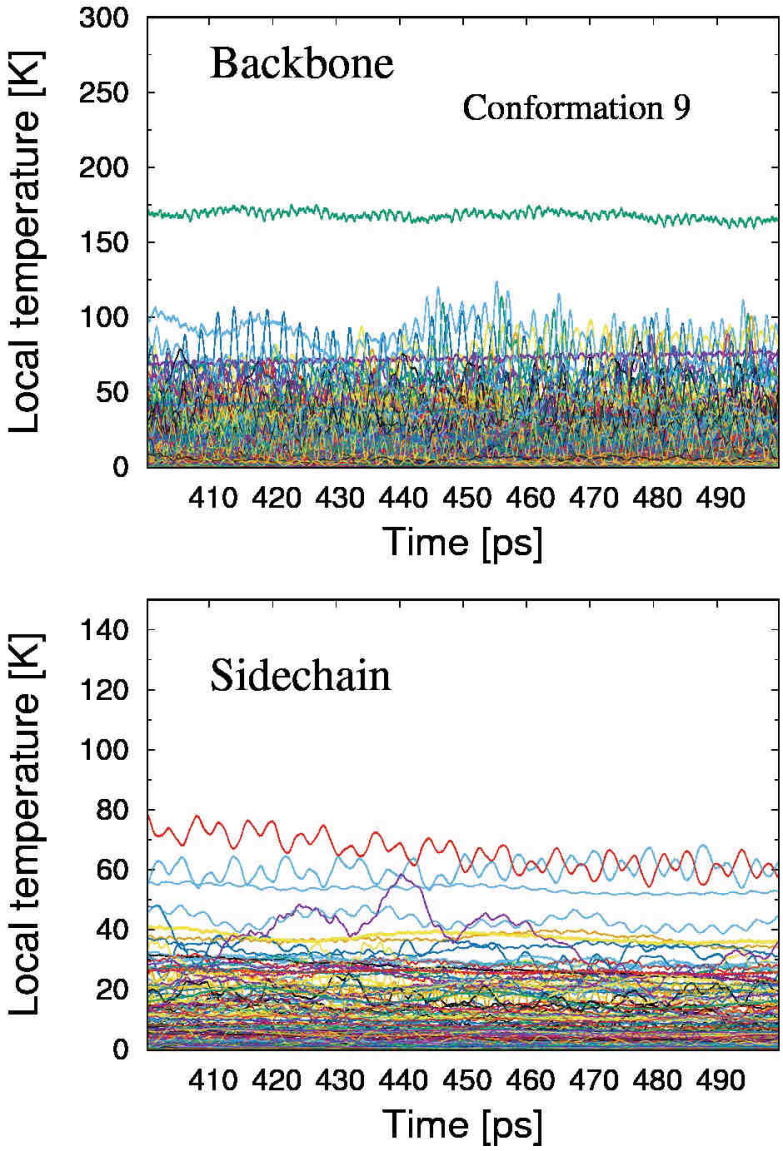



Figure B.18.: Average local temperature (computed over a stretch of 0.5 ps) of backbone segments (upper panel) and sidechains (lower panel) of citrate synthase for a cooling simulation of 500 ps with $\tau_T = 0.05$ ps starting from conformation 9. Only the time evolution in the quasi-stationary state is shown here.

# Publications

1. Simon Aubailly and Francesco Piazza, Cutoff lensing: predicting catalytic sites in enzymes, *Scientific Reports* **5**, Article number: 14874

2. Simon Aubailly, Stefano Iubini, Alberto Imparato and Francesco Piazza, Bilocalized normal modes, a new understanding of intramolecular communication in proteins, submitted to *Phys. Biol.* (2016)

3. Simon Aubailly, Stefano Iubini, Christian Kandt and Francesco Piazza, Plunging proteins in a cool explicit water bath: relaxation and spontaneous self-localization of energy, in preparation

# Bibliography


[1] http://images.wisegeek.com/diagram-of-the-22-amino-acids.jpg.

[2] https://jeanzin.fr/ecorevo/sciences/proteins.gif.

[3] http://cdn.rcsb.org/pdb101/learn/resources/what_is_a_protein.pdf.

[4] http://image.slidesharecdn.com/1-force-field-parameters-ppt-s-110414114515-phpapp02/95/parameters-for-classical-force-fields-e-tajkhorshid-4-728.jpg?cb=1325845507.

[5] F. Kozielski, S. Sack, A. Marx, M. Thormählen, E. Schönbrunn, V. Biou, A. Thompson, E.-M. Mandelkow, and E. Mandelkow, "The Crystal Structure of Dimeric Kinesin and Implications for Microtubule-Dependent Motility," *Cell*, vol. 91, no. 7, pp. 985–994, 1997.

[6] K. A. C., R. A. M., M. Aashish, H. Jianxin, H. Kelly, E. Katrin, H. Harald, P. Els, V. Celine, S. P. M., C. Arthur, F. C. C., G. Peter, S. Jan, W. W. I., G. K. Christopher, W. Jurgen, and K. B. K., "Activation and allosteric modulation of a muscarinic acetylcholine receptor," *Nature*, vol. 504, pp. 101–106, dec 2013.

[7] J. M. Louis, F. Dyda, N. T. Nashed, Alan R. Kimmel, and D. R. Davies, "Hydrophilic Peptides Derived from the Transframe Region of Gag-Pol Inhibit the HIV-1 Protease," *Biochemistry*, vol. 37, no. 8, pp. 2105–2110, 1998. PMID: 9485357.

[8] T. Okada, M. Sugihara, A.-N. Bondar, M. Elstner, P. Entel, and V. Buss, "The Retinal Conformation and its Environment in Rhodopsin in Light of a New 2.2 Å Crystal Structure?," *Journal of Molecular Biology*, vol. 342, no. 2, pp. 571–583, 2004.

[9] W. Hinrichs, C. Kisker, M. Duvel, A. Muller, K. Tovar, W. Hillen, and W. Saenger, "Structure of the Tet repressor-tetracycline complex and regulation of antibiotic resistance," *Science*, vol. 264, no. 5157, pp. 418–420, 1994.

[10] "http://www.chemtube3d.com/vibrationsc6h6.htm.".

[11] "http://www.cmbi.ru.nl/dssp.html.".

[12] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. Garland Science, 5 ed., 2007.

[13] N. R. Barton and L. S. Goldstein, "Going mobile: microtubule motors and chromosome segregation," *Proceedings of the National Academy of Sciences*, vol. 93, no. 5, pp. 1735–1742, 1996.

[14] J. Avila, "Microtubule dynamics.," *The FASEB Journal*, vol. 4, no. 15, pp. 3284–90, 1990.

[15] J. Wess, "G-protein-coupled receptors: molecular mechanisms involved in receptor activation and selectivity of G-protein recognition.," *The FASEB Journal*, vol. 11, no. 5, pp. 346–54, 1997.

[16] A. Fersht, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W. H. Freeman, 1999.

[17] R. Nussinov and C.-J. Tsai, "Unraveling structural mechanisms of allosteric drug action," *Trends in Pharmacological Sciences*, vol. 35, no. 5, pp. 256–264, 2014.

[18] S. Lu, S. Li, and J. Zhang, "Harnessing Allostery: A Novel Approach to Drug Discovery," *Medicinal Research Reviews*, vol. 34, no. 6, pp. 1242–1285, 2014.

[19] C. F. W. Becker, C. L. Hunter, R. Seidel, S. B. H. Kent, R. S. Goody, and M. Engelhard, "Total chemical synthesis of a functional interacting protein pair: The protooncogene H-Ras and the Ras-binding domain of its effector c-Raf1," *Proceedings of the National Academy of Sciences*, vol. 100, no. 9, pp. 5075–5080, 2003.

[20] M. Engelhard, "Quest for the chemical synthesis of proteins," *Journal of Peptide Science*, vol. 22, no. 5, pp. 246–251, 2016. PSC-15-0163.R2.

[21] G. Bulaj, "Formation of disulfide bonds in proteins and peptides," *Biotechnology Advances*, vol. 23, no. 1, pp. 87–92, 2005.

[22] Neidigh Jonathan W., Fesinmeyer R. Matthew, and Andersen Niels H., "Designing a 20-residue protein," *Nat Struct Mol Biol*, vol. 9, pp. 425–430, jun 2002. 10.1038/nsb798.

[23] M. Levitt, "Nature of the protein universe," *Proceedings of the National Academy of Sciences*, vol. 106, no. 27, pp. 11079–11084, 2009.

[24] KENDREW J. C., BODO G., DINTZIS H. M., PARRISH R. G., WYCKOFF H., and PHILLIPS D. C., "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis," *Nature*, vol. 181, pp. 662–666, mar 1958. 10.1038/181662a0.

[25] " http://www.rcsb.org/pdb/home/home.do.".

[26] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[27] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: Fast, flexible, and free," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[28] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The Amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.

[29] T. A. Halgren, "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94," *Journal of Computational Chemistry*, vol. 17, no. 5-6, pp. 490–519, 1996.

[30] . William L. Jorgensen, David S. Maxwell, and J. Tirado-Rives, "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids," *Journal of the American Chemical Society*, vol. 118, no. 45, pp. 11225–11236, 1996.

[31] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.

[32] W. F. V. Gunsteren and M. Karplus, "Protein dynamics in solution and in a crystalline environment: a molecular dynamics study," *Biochemistry*, vol. 21, no. 10, pp. 2259–2274, 1982.

[33] G. V. Gurskaya, *The molecular structure of amino acids ; determination by X-ray diffraction analysis.* Springer, 2012.

[34] R. M. Kini and H. J. Evans, "Molecular Modeling of Proteins: A Strategy for Energy Minimization by Molecular Mechanics in the AMBER Force Field," *Journal of Biomolecular Structure and Dynamics*, vol. 9, no. 3, pp. 475–488, 1991. PMID: 1687724.

[35] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[36] D. J. Evans and B. L. Holian, "The Nose–Hoover thermostat," *The Journal of Chemical Physics*, vol. 83, no. 8, pp. 4069–4074, 1985.

[37] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *Journal of Applied Physics*, vol. 52, no. 12, pp. 7182–7190, 1981.

[38] B. Juanico, Y.-H. Sanejouand, F. Piazza, and P. De Los Rios, "Discrete Breathers in Nonlinear Network Models of Proteins," *Phys. Rev. Lett.*, vol. 99, p. 238104, Dec 2007.

[39] M. Vendruscolo and C. M. Dobson, "Protein Dynamics: Moore's Law in Molecular Biology," *Current Biology*, vol. 21, no. 2, pp. R68–R70, 2011.

[40] Y. Ueda, H. Taketomi, and N. Go, "Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme," *Biopolymers*, vol. 17, no. 6, pp. 1531–1548, 1978.

[41] I. Bahar and R. Jernigan, "Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation 1," *Journal of Molecular Biology*, vol. 266, no. 1, pp. 195–214, 1997.

[42] A. Mukherjee and B. Bagchi, "Contact pair dynamics during folding of two small proteins: Chicken villin head piece and the Alzheimer protein beta-amyloid," *The Journal of Chemical Physics*, vol. 120, no. 3, pp. 1602–1612, 2004.

[43] A. Voegler Smith and C. K. Hall, "alpha-Helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model," *Proteins: Structure, Function, and Bioinformatics*, vol. 44, no. 3, pp. 344–360, 2001.

[44] A. V. Smith and C. K. Hall, "Assembly of a tetrameric alpha-helical bundle: Computer simulations on an intermediate-resolution protein model," *Proteins: Structure, Function, and Bioinformatics*, vol. 44, no. 3, pp. 376–391, 2001.

[45] Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, and P. G. Wolynes, "Optimizing physical energy functions for protein folding," *Proteins: Structure, Function, and Bioinformatics*, vol. 54, no. 1, pp. 88–103, 2004.

[46] S. Yup Lee, Y. Fujitsuka, D. H. Kim, and S. Takada, "Roles of physical interactions in determining protein-folding mechanisms: Molecular simulation of protein G and alpha spectrin SH3," *Proteins: Structure, Function, and Bioinformatics*, vol. 55, no. 1, pp. 128–138, 2004.

[47] Marrink Siewert J., Risselada H. Jelger, Yefimov Serge, Tieleman D. Peter, and d. V. A. H., "The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations," *The Journal of Physical Chemistry B*, vol. 111, no. 27, pp. 7812–7824, 2007. doi: 10.1021/jp071097f.

[48] P. Kar, S. M. Gopal, Y.-M. Cheng, A. Predeus, and M. Feig, "PRIMO: A Transferable Coarse-Grained Force Field for Proteins," *Journal of Chemical Theory and Computation*, vol. 9, no. 8, pp. 3769–3788, 2013. PMID: 23997693.

[49] M. Pasi, R. Lavery, and N. Ceres, "PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties," *Journal of Chemical Theory and Computation*, vol. 9, no. 1, pp. 785–793, 2013. PMID: 26589071.

[50] M. M. Tirion, "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis," *Phys. Rev. Lett.*, vol. 77, pp. 1905–1908, Aug 1996.

[51] A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar, "Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model," *Biophysical Journal*, vol. 80, no. 1, pp. 505–515, 2001.

[52] F. Piazza, P. De Los Rios, and F. Cecconi, "Temperature Dependence of Normal Mode Reconstructions of Protein Dynamics," *Phys. Rev. Lett.*, vol. 102, p. 218104, May 2009.

[53] P. D. L. Rios, F. Cecconi, A. Pretre, G. Dietler, O. Michielin, F. Piazza, and B. Juanico, "Functional Dynamics of {PDZ} Binding Domains: A Normal-Mode Analysis," *Biophysical Journal*, vol. 89, no. 1, pp. 14–21, 2005.

[54] M. Delarue and Y.-H. Sanejouand, "Simplified Normal Mode Analysis of Conformational Transitions in DNA-dependent Polymerases: the Elastic Network Model," *Journal of Molecular Biology*, vol. 320, no. 5, pp. 1011–1024, 2002.

[55] F. Tama and Y.-H. Sanejouand, "Conformational change of proteins arising from normal mode calculations," *Protein Engineering*, vol. 14, no. 1, pp. 1–6, 2001.

[56] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, "Vibrational Dynamics of Folded Proteins: Significance of Slow and Fast Motions in Relation to Function and Stability," *Phys. Rev. Lett.*, vol. 80, pp. 2733–2736, Mar 1998.

[57] F. Piazza and Y.-H. Sanejouand, "Long-range energy transfer in proteins," *Physical Biology*, vol. 6, no. 4, p. 046014, 2009.

[58] S. Nicolay and Y.-H. Sanejouand, "Functional Modes of Proteins Are among the Most Robust," *Phys. Rev. Lett.*, vol. 96, p. 078104, Feb 2006.

[59] F. Piazza and Y.-H. Sanejouand, "Energy transfer in nonlinear network models of proteins," *EPL (Europhysics Letters)*, vol. 88, no. 6, p. 68001, 2009.

[60] M. Caraglio and A. Imparato, "Energy transfer in molecular devices," *Phys. Rev. E*, vol. 90, p. 062712, Dec 2014.

[61] Q. Cui and I. Bahar, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. Chapman and Hall/CRC, 2005.

[62] G. Collier and V. Ortiz, "Emerging computational approaches for the study of protein allostery," *Archives of Biochemistry and Biophysics*, vol. 538, no. 1, pp. 6–15, 2013.

[63] D. K. Chakravorty and K. M. M. Jr., "Chapter Six - Studying Allosteric Regulation in Metal Sensor Proteins Using Computational Methods," in *Biomolecular Modelling and Simulations* (T. Karabencheva-Christova, ed.), vol. 96 of *Advances in Protein Chemistry and Structural Biology*, pp. 181–218, Academic Press, 2014.

[64] J. A. Mackinnon, N. Gallastegui, D. J. Osguthorpe, A. T. Hagler, and E. Estébanez-Perpiñá, "Allosteric mechanisms of nuclear receptors: insights from computational simulations," *Molecular and Cellular Endocrinology*, vol. 393, no. 1–2, pp. 75–82, 2014.

[65] Y. Li and H. Gong, "Theoretical and simulation studies on voltage-gated sodium channels," *Protein & Cell*, vol. 6, no. 6, pp. 413–422, 2015.

[66] A. Christopoulos, "Advances in GPCR Allostery: From Function to Structure," *Molecular Pharmacology*, 2014.

[67] K. Sharp and J. J. Skinner, "Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 2, pp. 347–361, 2006.

[68] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters," *The Journal of Chemical Physics*, vol. 76, no. 1, pp. 637–649, 1982.

[69] Verlet Loup, "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules," *Phys. Rev.*, vol. 159, pp. 98–103, jul 1967.

[70] A. K. Mazur, "Common Molecular Dynamics Algorithms Revisited: Accuracy and Optimal Time Steps of Störmer–Leapfrog Integrators," *Journal of Computational Physics*, vol. 136, no. 2, pp. 354–365, 1997.

[71] Dukka B KC, "Structure-based Methods for Computational Protein Functional Site Prediction," *Computational and Structural Biotechnology Journal*, vol. 8, p. e201308005, nov 2013.

[72] H. Neuvirth, R. Raz, and G. Schreiber, "ProMate: A Structure Based Prediction Program to Identify the Location of Protein–Protein Binding Sites," *Journal of Molecular Biology*, vol. 338, no. 1, pp. 181–199, 2004.

[73] A. A. S. T. Ribeiro and V. Ortiz, "Determination of Signaling Pathways in Proteins through Network Theory: Importance of the Topology," *Journal of Chemical Theory and Computation*, vol. 10, no. 4, pp. 1762–1769, 2014. PMID: 26580384.

[74] T. Ishikura, Y. Iwata, T. Hatano, and T. Yamato, "Energy exchange network of inter-residue interactions within a thermally fluctuating protein molecule: A computational study," *Journal of Computational Chemistry*, vol. 36, no. 22, pp. 1709–1718, 2015.

[75] B. Brooks and M. Karplus, "Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme," *Proceedings of the National Academy of Sciences*, vol. 82, no. 15, pp. 4995–4999, 1985.

[76] O. Marques and Y.-H. Sanejouand, "Hinge-bending motion in citrate synthase arising from normal mode calculations," *Proteins: Structure, Function, and Bioinformatics*, vol. 23, no. 4, pp. 557–560, 1995.

[77] B. Svensson and M. Søgaard, "Mutational analysis of glycosylase function," *Journal of Biotechnology*, vol. 29, no. 1, pp. 1–37, 1993.

[78] Daniel A. Kraut, Kate S. Carroll, and D. Herschlag, "Challenges in Enzyme Mechanism and Energetics," *Annual Review of Biochemistry*, vol. 72, no. 1, pp. 517–571, 2003. PMID: 12704087.

[79] N. Nagahara, T. Matsumura, and R. Kajihara, "Protein Cysteine Modifications: Medicinal Chemistry for Proteomics," 2009.

[80] B. Stephens and T. M. Handel, "Chapter nine - chemokine receptor oligomerization and allostery," in *Oligomerization and Allosteric Modulation in G-Protein Coupled Receptors* (T. Kenakin, ed.), vol. 115 of *Progress in Molecular Biology and Translational Science*, pp. 375–420, Academic Press, 2013.

[81] R. Nussinov and C.-J. Tsai, "Unraveling structural mechanisms of allosteric drug action," *Trends in Pharmacological Sciences*, vol. 35, no. 5, pp. 256–264, 2014.

[82] S. Raman, N. Taylor, N. Genuth, S. Fields, and G. M. Church, "Engineering allostery," *Trends in Genetics*, vol. 30, no. 12, pp. 521–528, 2014.

[83] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski, "Network Analysis of Protein Structures Identifies Functional Residues," *Journal of Molecular Biology*, vol. 344, no. 4, pp. 1135–1146, 2004.

[84] E. Chea and D. R. Livesay, "How accurate and statistically robust are catalytic site predictions based on closeness centrality?," *BMC Bioinformatics*, vol. 8, no. 1, pp. 1–14, 2007.

[85] Y.-R. Tang, Z.-Y. Sheng, Y.-Z. Chen, and Z. Zhang, "An improved prediction of catalytic residues in enzyme structures," *Protein Engineering Design and Selection*, vol. 21, no. 5, pp. 295–302, 2008.

[86] P. Slama, I. Filippis, and M. Lappe, "Detection of protein catalytic residues at high precision using local network properties," *BMC Bioinformatics*, vol. 9, no. 1, pp. 1–13, 2008.

[87] J. E. Fajardo and A. Fiser, "Protein structure based prediction of catalytic residues," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–11, 2013.

[88] M. Bhattacharyya and S. Vishveshwara, "Probing the Allosteric Mechanism in Pyrrolysyl-tRNA Synthetase Using Energy-Weighted Network Formalism," *Biochemistry*, vol. 50, no. 28, pp. 6225–6236, 2011. PMID: 21650159.

[89] A. Allain, I. Chauvot de Beauchene, F. Langenfeld, Y. Guarracino, E. Laine, and L. Tchertanov, "Allosteric pathway identification through network analysis: from molecular dynamics simulations to interactive 2D and 3D graphs," *Faraday Discuss.*, vol. 169, pp. 303–321, 2014.

[90] Roy Ambrish, Kucukural Alper, and Zhang Yang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nature protocols*, vol. 5, pp. 725–738, mar 2010.

[91] A. T. R. Laurie and R. M. Jackson, "Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites," *Bioinformatics*, vol. 21, no. 9, pp. 1908–1916, 2005.

[92] P. Puntervoll, R. Linding, C. Gemünd, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D. M. A. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferrè, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Küster, M. Helmer-Citterich, W. N. Hunter, R. Aasland, and T. J. Gibson, "ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3625–3630, 2003.

[93] D. Devos and A. Valencia, "Practical limits of function prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 41, no. 1, pp. 98–107, 2000.

[94] O. Lichtarge, H. R. Bourne, and F. E. Cohen, "An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families," *Journal of Molecular Biology*, vol. 257, no. 2, pp. 342–358, 1996.

[95] T. Zhang, H. Zhang, K. Chen, S. Shen, J. Ruan, and L. Kurgan, "Accurate sequence-based prediction of catalytic residues," *Bioinformatics*, vol. 24, no. 20, pp. 2329–2338, 2008.

[96] J. D. Fischer, C. E. Mayer, and J. Söding, "Prediction of protein functional residues from sequence by probability density estimation," *Bioinformatics*, vol. 24, no. 5, pp. 613–620, 2008.

[97] J. A. Capra and M. Singh, "Predicting functionally important residues from sequence conservation," *Bioinformatics*, vol. 23, no. 15, pp. 1875–1882, 2007.

[98] N. V. Petrova and C. H. Wu, "Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties," *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–12, 2006.

[99] J. D. Watson, R. A. Laskowski, and J. M. Thornton, "Predicting protein function from sequence and structural data," *Current Opinion in Structural Biology*, vol. 15, no. 3, pp. 275–284, 2005. Sequences and topology/Nucleic acids.

[100] A. R. Panchenko, F. Kondrashov, and S. Bryant, "Prediction of functional sites by analysis of sequence and structure conservation," *Protein Science*, vol. 13, no. 4, pp. 884–892, 2004.

[101] C. Innis, A. Anand, and R. Sowdhamini, "Prediction of Functional Sites in Proteins Using Conserved Functional Group Analysis," *Journal of Molecular Biology*, vol. 337, no. 4, pp. 1053–1068, 2004.

[102] S. Sankararaman, F. Sha, J. F. Kirsch, M. I. Jordan, and K. Sjölander, "Active site prediction using evolutionary and structural information," *Bioinformatics*, vol. 26, no. 5, pp. 617–624, 2010.

[103] B. Thibert, D. E. Bredesen, and G. del Rio, "Improved prediction of critical residues for protein function based on network and phylogenetic analyses," *BMC Bioinformatics*, vol. 6, no. 1, pp. 1–15, 2005.

[104] G. Cheng, B. Qian, R. Samudrala, and D. Baker, "Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design," *Nucleic Acids Research*, vol. 33, no. 18, pp. 5861–5867, 2005.

[105] S. Madabushi, H. Yao, M. Marsh, D. M. Kristensen, A. Philippi, M. E. Sowa, and O. Lichtarge, "Structural clusters of evolutionary trace residues are statistically significant and common in proteins1," *Journal of Molecular Biology*, vol. 316, no. 1, pp. 139–154, 2002.

[106] W. Tong, R. J. Williams, Y. Wei, L. F. Murga, J. Ko, and M. J. Ondrechen, "Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines," *Protein Science*, vol. 17, no. 2, pp. 333–341, 2008.

[107] J. Ko, L. F. Murga, P. André, H. Yang, M. J. Ondrechen, R. J. Williams, A. Agunwamba, and D. E. Budil, "Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves," *Proteins: Structure, Function, and Bioinformatics*, vol. 59, no. 2, pp. 183–195, 2005.

[108] A. Gutteridge, G. J. Bartlett, and J. M. Thornton, "Using A Neural Network and Spatial Clustering to Predict the Location of Active Sites in Enzymes," *Journal of Molecular Biology*, vol. 330, no. 4, pp. 719–734, 2003.

[109] S. Riniker, J. R. Allison, and W. F. van Gunsteren, "On developing coarse-grained models for biomolecular simulation: a review," *Phys. Chem. Chem. Phys.*, vol. 14, pp. 12423–12430, 2012.

[110] V. Tozzini, "Coarse-grained models for proteins," *Current Opinion in Structural Biology*, vol. 15, no. 2, pp. 144–150, 2005. Theory and simulation/Macromolecular assemblages.

[111] W. Zheng and M. Tekpinar, "Large-scale evaluation of dynamically important residues in proteins predicted by the perturbation analysis of a coarse-grained elastic model," *BMC Structural Biology*, vol. 9, no. 1, pp. 1–17, 2009.

[112] O. N. A. Demerdash, D. M. D., and M. J. C., "Structure-Based Predictive Models for Allosteric Hot Spots," *PLoS Comput Biol*, vol. 5, pp. 1–25, 10 2009.

[113] T. Haliloglu, E. Seyrek, and B. Erman, "Prediction of Binding Sites in Receptor-Ligand Complexes with the Gaussian Network Model," *Phys. Rev. Lett.*, vol. 100, p. 228102, Jun 2008.

[114] U. Emekli, D. Schneidman-Duhovny, H. J. Wolfson, R. Nussinov, and T. Haliloglu, "HingeProt: Automated prediction of hinges in protein structures," *Proteins: Structure, Function, and Bioinformatics*, vol. 70, no. 4, pp. 1219–1227, 2008.

[115] A. Ertekin, R. Nussinov, and T. Haliloglu, "Association of putative concave protein-binding sites with the fluctuation behavior of residues," *Protein Science*, vol. 15, no. 10, pp. 2265–2277, 2006.

[116] L.-W. Yang and I. Bahar, "Coupling between Catalytic Site and Collective Dynamics: A Requirement for Mechanochemical Activity of Enzymes," *Structure*, vol. 13, no. 6, pp. 893–904, 2005.

[117] T. Haliloglu, I. Bahar, and B. Erman, "Gaussian Dynamics of Folded Proteins," *Phys. Rev. Lett.*, vol. 79, pp. 3090–3093, Oct 1997.

[118] J. Kuriyan, G. A. Petsko, R. M. Levy, and M. Karplus, "Effect of anisotropy and anharmonicity on protein crystallographic refinement," *Journal of Molecular Biology*, vol. 190, no. 2, pp. 227–254, 1986.

[119] L. Soltan Ghoraie, F. Burkowski, and M. Zhu, "Sparse networks of directly coupled, polymorphic, and functional side chains in allosteric proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 83, no. 3, pp. 497–516, 2015.

[120] D. I. Flores, S.-M. R. R., and B. C. A., "A Simple Extension to the CMASA Method for the Prediction of Catalytic Residues in the Presence of Single Point Mutations," *PLoS ONE*, vol. 9, pp. 1–13, 09 2014.

[121] A. J. Gonzalez, L. Liao, and C. H. Wu, "Predicting Ligand Binding Residues and Functional Sites Using Multipositional Correlations with Graph Theoretic Clustering and Kernel CCA," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, pp. 992–1001, July 2012.

[122] C. Pons, F. Glaser, and J. Fernandez-Recio, "Prediction of protein-binding areas by small-world residue networks and application to docking," *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–10, 2011.

[123] Vacic Vladimir, Iakoucheva Lilia M., Lonardi Stefano, and Radivojac Predrag, "Graphlet Kernels for Prediction of Functional Residues in Protein Structures," *Journal of Computational Biology*, vol. 17, no. 1, pp. 55–72, 2010. doi: 10.1089/cmb.2009.0029.

[124] G.-H. Li and J.-F. Huang, "CMASA: an accurate algorithm for detecting local protein structural similarity and its application to enzyme catalytic site annotation," *BMC Bioinformatics*, vol. 11, no. 1, pp. 1–13, 2010.

[125] M. J. Zvelebil and M. J. Sternberg, "Analysis and prediction of the location of catalytic residues in enzymes," *Protein Engineering*, vol. 2, no. 2, pp. 127–138, 1988.

[126] P. Bate and J. Warwicker, "Enzyme/Non-enzyme Discrimination and Prediction of Enzyme Active Site Location Using Charge-based Methods," *Journal of Molecular Biology*, vol. 340, no. 2, pp. 263–276, 2004.

[127] A. H. Elcock, "Prediction of functionally important residues based solely on the computed energetics of protein structure," *Journal of Molecular Biology*, vol. 312, no. 4, pp. 885–896, 2001.

[128] S. Sacquin-Mora, O. Delalande, and M. Baaden, "Functional Modes and Residue Flexibility Control the Anisotropic Response of Guanylate Kinase to Mechanical Stress," *Biophysical Journal*, vol. 99, no. 10, pp. 3412–3419, 2010.

[129] S. Sacquin-Mora, É. Laforet, and R. Lavery, "Locating the active sites of enzymes using mechanical properties," *Proteins: Structure, Function, and Bioinformatics*, vol. 67, no. 2, pp. 350–359, 2007.

[130] H. R. Brodkin, N. A. DeLateur, S. Somarowthu, C. L. Mills, W. R. Novak, P. J. Beuning, D. Ringe, and M. J. Ondrechen, "Prediction of distal residue participation in enzyme catalysis," *Protein Science*, vol. 24, no. 5, pp. 762–778, 2015.

[131] Lee Jeeyeon and Goodey Nina M., "Catalytic Contributions from Remote Regions of Enzyme Structure," *Chemical Reviews*, vol. 111, no. 12, pp. 7595–7624, 2011. doi: 10.1021/cr100042n.

[132] F. Piazza and Y.-H. Sanejouand, "Discrete breathers in protein structures," *Physical Biology*, vol. 5, no. 2, p. 026001, 2008.

[133] D. A. Kondrashov, Q. Cui, and G. N. P. Jr., "Optimization and Evaluation of a Coarse-Grained Model of Protein Motion Using X-Ray Crystal Data," *Biophysical Journal*, vol. 91, no. 8, pp. 2760–2767, 2006.

[134] K. Suhre and Y.-H. Sanejouand, "ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement," *Nucleic Acids Research*, vol. 32, no. suppl 2, pp. W610–W614, 2004.

[135] J. Hafner and W. Zheng, "Optimal modeling of atomic fluctuations in protein crystal structures for weak crystal contact interactions," *The Journal of Chemical Physics*, vol. 132, no. 1, 2010.

[136] D. Riccardi, Q. Cui, and G. N. P. Jr., "Application of Elastic Network Models to Proteins in the Crystalline State," *Biophysical Journal*, vol. 96, no. 2, pp. 464–475, 2009.

[137] E. Eyal, L.-W. Yang, and I. Bahar, "Anisotropic network model: systematic evaluation and a new web interface," *Bioinformatics*, vol. 22, no. 21, pp. 2619–2627, 2006.

[138] A. J. Rader and S. M. Brown, "Correlating allostery with rigidity," *Mol. BioSyst.*, vol. 7, pp. 464–471, 2011.

[139] M. Z. Kamal, M. T. A. Shamim, K. G., and R. N. Madhusudhana, "Role of Active Site Rigidity in Activity: MD Simulation and Fluorescence Study on a Lipase Mutant," *PLoS ONE*, vol. 7, pp. 1–8, 04 2012.

[140] X. Guo, D. He, L. Huang, L. Liu, L. Liu, and H. Yang, "Strain energy in enzyme–substrate binding: An energetic insight into the flexibility versus rigidity of enzyme active site," *Computational and Theoretical Chemistry*, vol. 995, pp. 17–23, 2012.

[141] R. Brandman, J. N. Lampe, Y. Brandman, and P. R. O. de Montellano, "Active-site residues move independently from the rest of the protein in a 200 ns molecular dynamics simulation of cytochrome {P450} {CYP119}," *Archives of Biochemistry and Biophysics*, vol. 509, no. 2, pp. 127–132, 2011.

[142] P. Csermely, R. Palotai, and R. Nussinov, "Induced fit, conformational selection and independent dynamic segments: an extended view of binding events," *Trends in Biochemical Sciences*, vol. 35, no. 10, pp. 539–546, 2010.

[143] B. Halle, "Flexibility and packing in proteins," *Proceedings of the National Academy of Sciences*, vol. 99, no. 3, pp. 1274–1279, 2002.

[144] C. T. Porter, G. J. Bartlett, and J. M. Thornton, "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data," *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D129–D133, 2004.

[145] C. Lanczos, *Applied Analysis*. Dover Publications, 2010.

[146] U. Diwekar, *Introduction to Applied Optimization*, vol. 22. Springer US, 2008.

[147] B. D. and T. J., *Introduction to Linear Optimization*. Athena Scientific, 1997.

[148] C. Reeves and J. E. Rowe, *Genetic Algorithms : Principles and Perspectives*. Springer US, 2002.

[149] S. Sattin, J. Tao, G. Vettoretti, E. Moroni, M. Pennati, A. Lopergolo, L. Morelli, A. Bugatti, A. Zuehlke, M. Moses, T. Prince, T. Kijima, K. Beebe, M. Rusnati, L. Neckers, N. Zaffaroni, D. A. Agard, A. Bernardi, and G. Colombo, "Activation of Hsp90 Enzymatic Activity and Conformational Dynamics through Rationally Designed Allosteric Ligands," *Chemistry – A European Journal*, vol. 21, no. 39, pp. 13598–13608, 2015.

[150] B. Srinivasan, F. Forouhar, A. Shukla, C. Sampangi, S. Kulkarni, M. Abashidze, J. Seetharaman, S. Lew, L. Mao, T. B. Acton, R. Xiao, J. K. Everett, G. T. Montelione, L. Tong, and H. Balaram, "Allosteric regulation and substrate activation in cytosolic nucleotidase II from Legionella pneumophila," *FEBS Journal*, vol. 281, no. 6, pp. 1613–1628, 2014.

[151] A. S. Tora, X. Rovira, I. Dione, H.-O. Bertrand, I. Brabet, Y. De Koninck, N. Doyon, J.-P. Pin, F. Acher, and C. Goudet, "Allosteric modulation of metabotropic glutamate receptors by chloride ions," *The FASEB Journal*, vol. 29, no. 10, pp. 4174–4188, 2015.

[152] K. E. Livingston and J. R. Traynor, "Disruption of the Na+ ion binding site as a mechanism for positive allosteric modulation of the mu-opioid receptor," *Proceedings of the National Academy of Sciences*, vol. 111, no. 51, pp. 18369–18374, 2014.

[153] W. Saenger, P. Orth, C. Kisker, W. Hillen, and W. Hinrichs, "The Tetracycline Repressor—A Paradigm for a Biological Switch," *Angewandte Chemie International Edition*, vol. 39, no. 12, pp. 2042–2052, 2000.

[154] Arrowsmith C. H., Czaplicki J., Iyer S. B., and Jardetzky O., "Unusual dynamic features of the trp repressor from Escherichia coli," *Journal of the American Chemical Society*, vol. 113, no. 10, pp. 4020–4022, 1991. doi: 10.1021/ja00010a070.

[155] C. Wen-I, B. Pamela, and M. K. Shive, "Identification and characterization of aspartate residues that play key roles in the allosteric regulation of a transcription factor: aspartate 274 is essential for inducer binding in lac repressor," *Biochemistry*, vol. 33, no. 12, pp. 3607–3616, 1994. doi: 10.1021/bi00178a018.

[156] A. Das, M. Ghosh, and J. Chakrabarti, "Time dependent correlation between dihedral angles as probe for long range communication in proteins," *Chemical Physics Letters*, vol. 645, pp. 200–204, 2016.

[157] R. Kalescky, J. Liu, and P. Tao, "Identifying Key Residues for Protein Allostery through Rigid Residue Scan," *The Journal of Physical Chemistry A*, vol. 119, no. 9, pp. 1689–1700, 2015. PMID: 25437403.

[158] C. L. McClendon, G. Friedland, D. L. Mobley, H. Amirkhani, and M. P. Jacobson, "Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles," *Journal of Chemical Theory and Computation*, vol. 5, no. 9, pp. 2486–2502, 2009. PMID: 20161451.

[159] Y. Lee, S. Choi, and C. Hyeon, "Mapping the intramolecular signal transduction of G-protein coupled receptors," *Proteins: Structure, Function, and Bioinformatics*, vol. 82, no. 5, pp. 727–743, 2014.

[160] A. T. VanWart, J. Eargle, Z. Luthey-Schulten, and R. E. Amaro, "Exploring Residue Component Contributions to Dynamical Network Models of Allostery," *Journal of Chemical Theory and Computation*, vol. 8, no. 8, pp. 2949–2961, 2012. PMID: 23139645.

[161] A. Ghosh and S. Vishveshwara, "A study of communication pathways in methionyl-tRNA synthetase by molecular dynamics simulations and structure network analy-

sis," *Proceedings of the National Academy of Sciences*, vol. 104, no. 40, pp. 15711–15716, 2007.

[162] A. A. S. T. Ribeiro and V. Ortiz, "Energy Propagation and Network Energetic Coupling in Proteins," *The Journal of Physical Chemistry B*, vol. 119, no. 5, pp. 1835–1846, 2015. PMID: 25569787.

[163] J. Zhou, W. Yan, and G. Shen, "Amino acid network for prediction of catalytic residues in enzymes: a comparison survey," 2016.

[164] F. Radivojac, "Computational Methods for Identification of Functional Residues in Protein Structures," 2011.

[165] S.-W. Huang, S.-H. Yu, C.-H. Shih, H.-W. Guan, and T. Hwang, "On the Relationship Between Catalytic Residues and their Protein Contact Number," 2011.

[166] C. Chennubhotla and B. Ivet, "Signal Propagation in Proteins and Relation to Equilibrium Fluctuations," *PLoS Comput Biol*, vol. 3, pp. 1–11, 09 2007.

[167] S. Luccioli, A. Imparato, S. Lepri, F. Piazza, and A. Torcini, "Discrete breathers in a realistic coarse-grained model of proteins," *Physical Biology*, vol. 8, no. 4, p. 046008, 2011.

[168] F. Piazza and Y.-H. Sanejouand, "Breather-mediated energy transfer in proteins," *Discrete and Continuous Dynamical Systems - Series S*, vol. 4, no. 5, pp. 1247–1266, 2011.

[169] S. W. Lockless and R. Ranganathan, "Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families," *Science*, vol. 286, no. 5438, pp. 295–299, 1999.

[170] N. Ota and D. A. Agard, "Intramolecular Signaling Pathways Revealed by Modeling Anisotropic Thermal Diffusion," *Journal of Molecular Biology*, vol. 351, no. 2, pp. 345–354, 2005.

[171] X. Ma, Y. Qi, and L. Lai, "Allosteric sites can be identified based on the residue–residue interaction energy difference," *Proteins: Structure, Function, and Bioinformatics*, vol. 83, no. 8, pp. 1375–1384, 2015.

[172] J. M. Scholey, "Intraflagellar Transport," *Annual Review of Cell and Developmental Biology*, vol. 19, no. 1, pp. 423–443, 2003. PMID: 14570576.

[173] R. D. Vale, "The Molecular Motor Toolbox for Intracellular Transport," *Cell*, vol. 112, no. 4, pp. 467–480, 2003.

[174] Ramanathan Arvind and Agarwal Pratul K., "Computational Identification of Slow Conformational Fluctuations in Proteins," *The Journal of Physical Chemistry B*, vol. 113, no. 52, pp. 16669–16680, 2009. doi: 10.1021/jp9077213.

[175] S. Mahajan and Y.-H. Sanejouand, "On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins," *Archives of Biochemistry and Biophysics*, vol. 567, pp. 59–65, 2015.

[176] A. K. Srivastava, L. R. McDonald, A. Cembran, J. Kim, L. R. Masterson, C. L. McClendon, S. S. Taylor, and G. Veglia, "Synchronous Opening and Closing Motions Are Essential for cAMP-Dependent Protein Kinase A Signaling," *Structure*, vol. 22, no. 12, pp. 1735–1743, 2014.

[177] P. Orth, F. Cordes, D. Schnappinger, W. Hillen, W. Saenger, and W. Hinrichs, "Conformational changes of the Tet repressor induced by tetracycline trapping1," *Journal of Molecular Biology*, vol. 279, no. 2, pp. 439–447, 1998.

[178] S. Maudsley, S. A. Patel, S.-S. Park, and L. Martin, "Functional Signaling Biases in G Protein-Coupled Receptors: Game Theory and Receptor Dynamics," 2012.

[179] Z. Chilmonczyk, A. J. Bojarski, and I. Sylte, "Ligand-directed trafficking of receptor stimulus," *Pharmacological Reports*, vol. 66, no. 6, pp. 1011–1021, 2014.

[180] F. Piazza, P. De Los Rios, and Y.-H. Sanejouand, "Slow Energy Relaxation of Macromolecules and Nanoclusters in Solution," *Phys. Rev. Lett.*, vol. 94, p. 145502, Apr 2005.

[181] A. Bikaki, N. K. Voulgarakis, S. Aubry, and G. P. Tsironis, "Energy relaxation in discrete nonlinear lattices," *Phys. Rev. E*, vol. 59, pp. 1234–1237, Jan 1999.

[182] F. Piazza, S. Lepri, and R. Livi, "Slow energy relaxation and localization in 1D lattices," *Journal of Physics A: Mathematical and General*, vol. 34, no. 46, p. 9803, 2001.

[183] R. Reigada, A. Sarmiento, and K. Lindenberg, "Energy relaxation in nonlinear one-dimensional lattices," *Phys. Rev. E*, vol. 64, p. 066608, Nov 2001.

[184] F. Piazza, S. Lepri, and R. Livi, "Cooling nonlinear lattices toward energy localization," *Chaos*, vol. 13, no. 2, pp. 637–645, 2003.

[185] S. Flach and A. Gorbach, "Discrete breathers in Fermi–Pasta–Ulam lattices," *Chaos*, vol. 15, no. 1, 2005.

[186] S. Flach and C. Willis, "Discrete breathers," *Physics Reports*, vol. 295, no. 5, pp. 181–264, 1998.

[187] F. Piazza, "Nonlinear excitations match correlated motions unveiled by NMR in proteins: a new perspective on allosteric cross-talk," *Physical Biology*, vol. 11, no. 3, p. 036003, 2014.

[188] A. Sitnitsky, "Dynamical contribution into enzyme catalytic efficiency," *Physica A: Statistical Mechanics and its Applications*, vol. 371, no. 2, pp. 481–491, 2006.

[189] V. Dubinko and F. Piazza, "On the role of disorder in catalysis driven by discrete breathers," *Letters on materials*, vol. 4, no. 4, pp. 273–278, 2014.

[190] G. Lamm and A. Szabo, "Langevin modes of macromolecules," *The Journal of Chemical Physics*, vol. 85, no. 12, pp. 7334–7348, 1986.

[191] M. Eleftheriou, S. Lepri, R. Livi, and F. Piazza, "Stretched-exponential relaxation in arrays of coupled rotators," *Physica D: Nonlinear Phenomena*, vol. 204, no. 3–4, pp. 230–239, 2005.

[192] "http://cib.cf.ocha.ac.jp/bitool/ASA/.".

Simon Aubailly

# Du gros-grains à la modélisation moléculaire tout-atome : comment la structure/dynamique façonnent la communication intramoléculaire et les sites fonctionnels dans les protéines

Résumé :
Dans cette thèse, nous nous sommes intéressés à la relation subtile qui existe entre les structures complexes des protéines et leurs fonctions encore plus raffinées que ces dernières effectuent. Basés sur deux descriptions différentes des protéines, à l'échelle de acide-aminé et à l'echelle atomique, un de nos objectifs était de connecter des indicateurs structuraux calculés à partir de la topologie des protéines à des sites fonctionnels tels que les sites catalyiques dans les enzymes. Un autre pan de la recherche de cette thèse était d'utiliser nos outils basés sur la structure et de mettre au point de nouvelles simulations numériques pour étudier les déterminants basiques structuraux et dynamiques de la communication intramoléculaire dans les protéines. Une première découverte fut de montrer comment l'analyse des modes normaux et la théorie des reseaux complexes conduisent à la prédiction des sites catalytiques dans les enzymes. De plus, nous avons travaillé sur un groupe relativement peu connu de modes normaux qui ont la particularité d'être localisés à deux endroits très eloignés dans la structure des protéines. Ces modes bilocalisés ont permis de réaliser des transferts d'énergie à des distances considérables (plus de 70 Å). Finalement, des expériences de refroidissement effectuées sur un système protéine-eau décrit à l'échelle atomique ont dévoilé que le refroidissement induit une localisation spontanée d'énergie, indiquant certaines déformations des anneaux du benzene comme possible centres de stockage de l'énergie dans les protéines.

Mots clés : protéine, fonction, structure, prédiction, communication, gros-grains, tout-atome

# From coarse-grained to atomistic molecular modeling : how structure and dynamics shape intra-molecular communication and functional sites in proteins

Abstract :
In this thesis we have focused on the elusive relation that exists in proteins between their complex structures and the even more complex and sophisticated functions that they perform. Based on two different descriptions of proteins, at residue and atomistic scale, one of our aims was to connect structural indicators computed from the topology of protein scaffolds to hot spots in proteins such as catalytic sites in enzymes. Another goal of this thesis was to employ our structure-based tools and set up original simulation scheme to investigate the basic structural and dynamical determinants of intramolecular communication in proteins. As a first important finding, we have shown how normal mode analysis and specific graph-theoretical approaches lead to the prediction of catalytic sites in enzymes. Moreover, we have concentrated our attention on an overlooked class of normal modes, that are strongly localized at two widely separated locations in protein scaffolds. These bilocalized modes turned out to efficiently mediate energy transfer even across considerable distances (more than 70 Å). Finally, cooling experiments performed on a protein-water system described at atomic level have unveiled complex cooling-induced spontaneous energy localization patterns, pointing to specific deformation modes of benzene rings as potential energy-storage centers.

Keywords : protein, function, structure, prediction, communication, coarse-grained, all-atom.

**Centre de Biophysique Moléculaire (CBM),**
**Rue Charles Sadron, 45071 Orléans**