Comparaison de surface épitopiques en fonction de différentes conditions et calculs d'affinité protéine-protéine par méthodes bio-informatiques

Laurent Berthier

N°d'ordre NNT : **2021LYSE1068**

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale N° 205
EDISS – Ecole Doctorale Interdisciplinaire Sciences-Santé

Spécialité de doctorat : Modélisation moléculaire, bioinformatique
Discipline : Biochimie

Soutenue publiquement le 20/07/2021, par :

Laurent BERTHIER

Comparaison de surface épitopiques en fonction de différentes conditions et calculs d'affinité protéine-protéine par méthodes bio-informatiques

Devant le jury composé de :

BARBE Sophie	DR	INRAE Toulouse	Rapporteur et présidente
DE BREVERN Alexandre	DR	INSERM Paris	Rapporteur
BRASS Olivier	CR	Sanofi-Pasteur Lyon	Examinateur
RADIX Sylvie	MC	Université Claude Bernard Lyon 1	Examinatrice
TERREUX Raphaël	PR	Université Claude Bernard Lyon 1	Directeur de thèse

Université Claude Bernard – LYON 1

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

COMPOSANTES SANTE

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

COMPOSANTES & DEPARTEMENTS DE SCIENCES & TECHNOLOGIE

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

Remerciements

Je remercie en premier lieu le Pr. Raphaël Terreux pour m'avoir donné l'opportunité de réaliser cette thèse et m'avoir fait l'honneur de la diriger. Merci de m'avoir aidé et conseillé tout au long de ce travail. J'adresse aussi ma profonde reconnaissance au Dr. Olivier Brass sans qui cette thèse n'aurait pas eu lieu et avec qui j'ai beaucoup échangé et partagé concernant les projets en collaboration avec Sanofi-Pasteur.

Je remercie chaleureusement mes rapporteurs et examinatrice le Dr. Sophie Barbe, Dr. Alexandre De Brevern et la Dr. Sylvie Radix d'avoir accepté de faire partie des membres de mon jury de thèse et de m'apporter leurs regards sur les travaux réalisé au cours de ma thèse.

Je remercie également les membres du comité de suivi des travaux de thèse le Pr. Sylvie Ricard Blum et le Dr. Pedro Da Silva pour m'avoir aiguillé et partagé leur expérience avec moi.

Je remercie aussi tous les membres de l'équipe, présents et passés, pour leur partage d'expérience et leur agréable compagnie : Shang, Jodi, Maylis, le Dr. Charlotte, le Dr Stéphanie, le Dr. Emmanuel Bettler ainsi que le Dr. Gilbert Deléage.

Je remercie aussi particulièrement le Dr. Christophe Vroland avec qui j'ai particulièrement collaboré sur le dernier projet de cette thèse, il m'a beaucoup appris sur les réseaux de neurones et sans lui ce projet n'aurait probablement pas été possible.

Enfin je remercie mes amis et ma famille qui ont concouru au bénéfice d'un environnement épanouissant tout au long de ma vie et de mes études en particulier mes Parents et grands-parents.

Table des Matières

Liste des abréviations

Par ordre d'apparition

BCG	Bacillus Calmette-Guérin
ROR	Rougeole Oreillons Rubéole
PPV	Pseudo-Particule Virale
ARNm	Acide RiboNucléique messager
ORL	Oto-Rhino-Laryngée
PVH	PapillomaVirus Humain
PRR	Pathogen Recognition Receptor
CMH	Complexe Majeur d'Histocompatibilité
HLA	Human Leukocyte Antigen
CD	Cluster de Différenciation
IgM	Immunoglobuline M
IgG	Immunoglobuline G
AMM	Autorisation de Mise sur le Marché
H1N1	Hémagglutinine 1 Neuraminidase 1
OMS	Organisation Mondiale de la Santé
EIG	Effets Indésirables Graves
VHB	Virus de l'Hépatite B
SEP	Sclérose En Plaques
IC-SNC	Inflammation Chronique du Système Nerveux Central
Cryo-ME	Cryo-Microscopie Electronique
RMN	Résonance Magnétique Nucléaire
PDB	Protein Data Bank
BMRB	Biological Magnetic Resonance Data Bank
VMD	Visual Molecular Dynamics
CAVEs	Cave Automatical Virtual Environments
BLAST	Basic Local Alignment Search Tool
I-TASSER	Iterative Threading ASSEmbly Refinement
GG	Gros Grains
UNRES	UNited RESidue

CASP	Critical Assessment of Structure Prediction
FFT	Fast Fourier Transform
CAPRI	Critical Assessment of PRediction of Interactions
QM	Mécanique Quantique
MM	Mécanique Moléculaire classique
GG	Gros Grains
NAMD	NAnoscale Molecular Dynamics
PME	Particle Mesh Ewald
NPT	Nombre de particules, Température et Pression constantes
NVE	Nombre de particules, Volume et Énergie constants
NVT	Nombre de particules, Volume et Température constantes
MDFF	Molécular Dynamique Flexible fitting
FEP	Free Energy Perturbation
CHARMM	Chemistry at HARvard Macromolecular Mechanics
AMBER	Assisted Model Building with Energy Refinement
GROMOS	GROningen MOlecular Simulation
LIE	Linear Interaction Energy
MM-PBSA	Molecular Mechanics, Poisson–Boltzmann Surface Area
SPC	Simple Point-Charge
TIP3P	Transferable Intermolecular Potential with 3 Points
SASA	Solvent-Accessible Surface Area
MM-GBSA	Molecular Mechanics, Generalized Born Surface Area
IA	Intelligence Artificielle
SVM	Support Vector Machine
GMM	Gaussian Mixture Models
ADL	Analyse Discriminante Linéaire
Tanh	Tangente hyperbolique
ReLU	Rectified Linear Unit
CUDA	Compute Unified Device Architecture
RNC	Réseau de Neurones Convolutifs
SGD	Stochastic Gradient Descent
NAG	Nesterov Accelerated Gradient
AdaGrad	Adaptive Gradient
RMSProp	Root Mean Square Propagation
Adam	Adaptive moment estimation
Nadam	Nesterov-accelerated Adaptive Moment Estimation
VHB	Virus de l’Hépatite B

PSVB	Particule sous virale de l'hépatite B
HBsAg	Hepatitis B surface Antigen
PLs	PhosphoLipides
KSCN	isothiocyanate de potassium
DOPC	Di-Oléoyl-PhosphatidylCholine
MFA	Microscopie à Force Atomique
RMSD	Root Mean Square Deviation
MOE	Molecular Operating Environment
BA	Boucle Antigénique
BC	Boucle Capsidique
EMD	Electron Microscopy Data
PPM	Positioning of Proteins in Membranes
NIAID	National Institute of Allergy and Infectious Diseases
OD	Densité Optique
FRODOCK	Fast Rotational DOCKing
PRODIGY	PROtein binDIng enerGY prediction
CDR	Complementarity determining regions
fHbp	factor H binding protein
DLPPS	Deep Learning Protein-Protein Scoring
PATTY	Programmable ATom TYping systems
IN2P3	Institut National de Physique Nucléaire et de Physique des Particules
GPU	Graphics Processing Unit
CPU	Central Processing Unit
CCharPPI	Computational Characterisation of Protein-Protein Interactions
SIPPER	Scoring by Intermolecular Pairwise Propensities of Exposed Residues

Résumé

La vaccination est l'outil de choix pour illustrer l'adage « mieux vaut prévenir que guérir ». Le mode de fonctionnement est simple, elle consiste à exposer l'organisme à un ou plusieurs antigènes d'un organisme cible afin d'entrainer le système immunitaire de l'hôte contre ce dernier. Les antigènes peuvent être présenté à l'organisme sous différentes formes selon le procédé de fabrication. Il peut s'agir d'un organisme entier à un simple fragment de protéine. A la suite d'une vaccination, le système immunitaire sera à même de reconnaitre et de neutraliser plus rapidement et efficacement l'agent pathogène grâce au phénomène de mémoire immunitaire.

Au cours de cette thèse, nous verront comment les techniques de mécanique moléculaire et de modélisation bioinformatique peuvent contribuer à la compréhension du mécanisme d'action de certains vaccins ainsi qu'à la conception de ces derniers. Pour cela nous nous pencherons sur différentes méthodes de modélisation de structure des protéines à travers des cas concret de vaccins préexistants. Nous aborderons en particulier les techniques bioinformatiques de prédiction de structure des complexes protéiques ainsi que celles capables de prédire les énergies d'interactions protéine-protéine. Nous constaterons les limitations des différentes méthodes existantes et explorerons l'usage d'une nouvelle approche pour prédire ces énergies d'interactions par l'apprentissage automatique en profondeur à l'aide de réseau de neurones convolutifs.

Mots-clés : Antigène vaccinale, Modélisation, Dynamique Moléculaire, Amarrage protéine-protéine, Prédiction d'énergie d'interaction, Apprentissage automatique, réseau de neurones convolutifs

INTRODUCTION

Un épitope, aussi appelé déterminant antigénique, est une partie d'un antigène qui peut être reconnue par un paratope (partie variable d'un anticorps). Cette reconnaissance épitope/paratope permet de déterminer si l'antigène en question appartient au domaine du soi ou au domaine du non-soi, elle est donc à la base de la réponse immunitaire spécifique. La caractérisation et l'isolation des épitopes spécifiques d'un pathogène constituent donc la clef de voûte de la recherche vaccinale. Pour assurer la vaccination, l'épitope identifié comme tel, doit être transmis à l'organisme tout en assurant la stabilité conformationnelle de ce dernier par rapport à son état natif. Ceci afin que la réponse immunitaire, engendrée par le vaccin, soit au plus près de la réponse immunitaire qu'engendrerai le pathogène sans pour autant que le patient ne subisse les effets de la maladie que provoquerait le pathogène. Les moyens pour arriver à cet objectif sont nombreux et sont présentés succinctement dans le chapitre 1 de cette thèse.

Aux travers d'exemples d'intérêt pour Sanofi-Pasteur nous verrons comment des méthodes de bioinformatiques structurales peuvent servir à caractériser et évaluer certains de ces épitopes. Pour cela, un tour d'horizon des principales méthodes de modélisation des protéines et de mécanique moléculaire sera effectué au chapitre 2. Puis nous explorerons dans le chapitre 3, les bases d'une famille d'algorithme dont l'émergence à bouleverser la recherche scientifique ainsi que les industries dans de nombreux domaines. Ce sont les algorithmes d'apprentissages automatiques, en particulier ceux utilisant des réseaux de neurones à multiples couches que nous appelons généralement algorithme d'apprentissage profond (Deep Learning).

Les travaux de cette thèse commencent au chapitre 4 et porteront en premier lieu sur la construction de modèles structuraux à l'échelle atomiques de particules sous virales de l'hépatite B (PSVB). En effet, cette protéine recombinante, largement utilisée pour la vaccination contre le virus de l'hépatite B n'a, à ce jour, toujours pas été résolue expérimentalement. Nous chercherons donc à relever le défi de la modéliser en nous servant de diverses méthodes bioinformatiques telles que la modélisation *ab initio*, l'amarrage protéine-protéine ainsi que plusieurs méthodes de dynamiques moléculaires.

Dans le chapitre 5, nous nous attarderons sur le calcul d'affinité protéine-protéine par méthodes bio-informatiques. D'une part, nous évaluerons la robustesse de différentes approches pour réaliser un criblage virtuel des épitopes de l'hémagglutinine (protéine de surface du virus de la grippe saisonnière). D'autre part, nous appliquerons l'approche retenue

afin d'évaluer l'effet que peuvent avoir des mutations sur l'affinité de liaison entre le fHbp (une des lipoprotéines immunogènes de Neisseria meningitidis de sérogroupe B utilisée dans les vaccins contre les méningocoques) et le facteur H du complément. Le facteur H est une protéine humaine qui joue un rôle dans l'écrantage de l'épitope vaccinal, réduisant ainsi son efficacité.

Enfin, après avoir constaté certaines limites des méthodes précédemment testées dans le calcul d'affinité protéine-protéine, nous proposerons dans le chapitre 6, une nouvelle méthode de prédiction de score d'interaction protéine-protéine à partir de leur structure, basée sur l'apprentissage automatique. Pour cela nous emploierons un type de réseau de neurone spécialisé dans le traitement des images, les réseaux de neurones convolutifs.

PARTIE 1 : ETAT DE L'ART

Chapitre 1 : La vaccination et ses problématiques

1. Généralités

La vaccination implique l'introduction chez un individu d'une préparation antigénique issu ou du moins suffisamment proche d'un agent infectieux déterminé. Ceci doit permettre de créer une réponse immunitaire capable de protéger l'individu contre la survenue d'une maladie liée à cet agent infectieux. La pratique de la vaccination au sein d'une collectivité ou une population peut permettre le contrôle voire l'élimination de certaines infections contagieuses. C'est pourquoi la vaccination constitue un instrument essentiel pour la santé publique.

Les premiers vaccins étaient en premier lieu destinés à prévenir les infections virales : la variole, Jenner 1796, et la rage, Pasteur 1885 (1). Depuis lors, divers autres vaccins antiviraux et antibactériens ont été inventé et certains ont mêmes permis l'éradication de certaines infections comme le vaccin antivariolique qui provoqua la disparition total de la maladie en 1980 (2).

Les vaccins peuvent être composés de la totalité d'une bactérie (ex : le BCG : Bacillus Calmette-Guérin) ou d'un virus (ex : le ROR : Rougeole Oreillons Rubéole). Ils peuvent aussi n'être qu'un fragment de ces derniers (ex : la coqueluche acellulaire), voire simplement un fragment d'une protéine (ex : l'anatoxine tétanique, exotoxine élaborée par Clostridium tetanii). Des antigènes peuvent également être obtenus par génie génétique et ensuite purifiés (vaccins contre certains types de papillomavirus humains).

Le vaccin doit conserver un pouvoir immunogène suffisant pour susciter des réactions immunitaires qui induiront la formation d'anticorps par activation des lymphocytes B et la mise en alerte de lymphocytes T. Il doit malgré tout être dénué d'un quelconque pouvoir infectieux ou toxique. L'inactivation peut être obtenue soit en tuant la bactérie ou le virus, soit en atténuant son pouvoir de transmission et de reproduction. Il s'agit, dans ce dernier cas, d'un vaccin vivant atténué comme le BCG, les vaccins contre la rubéole et la fièvre jaune.

La défense de l'organisme contre le milieu extérieur comporte une immunité dite innée, existant en l'absence de tout contact avec un antigène, ainsi qu'une immunité dite adaptative/acquise, apparaissant après contact de l'organisme avec des molécules étrangères qui sont des antigènes. Les défenses immunitaires acquises sont activées par la présence d'un antigène, elles concernent divers types de cellules comme les macrophages, les cellules de Langerhans, les cellules dendritiques cutanées, les lymphocytes B à l'origine des anticorps ou immunoglobulines, lymphocytes T et diverses cytokines.

Le système immunitaire a deux caractéristiques essentielles :

- La mémoire : un seul contact avec un antigène, comme dans le cas de certaines vaccinations, peut protéger l'organisme pendant toute la durée de sa vie.
- L'importance des contacts intercellulaires et intermoléculaires nécessitant une complémentarité stéréochimique comme dans le cas des interactions antigène-anticorps.

La vaccination joue sur la mémoire immunitaire (immunité acquise), elle permet la mise en place rapide de moyens de défense spécifiques (réponse mémoire), plus efficaces pour contrôler l'infection car se mettant en œuvre rapidement. L'efficacité d'un vaccin dépend de la réceptivité de l'hôte à l'immunogène, de la capacité du vaccin à stimuler les moyens de défense de l'organisme mais aussi de la capacité de la réponse immune ainsi produite à neutraliser l'agent infectieux.

En plus de ces critères d'efficacité, un vaccin doit répondre au cahier des charges suivant, i doit être :

- Sûr, donc n'entrainant pas d'effet indésirable grave.
- Protecteur au long terme.
- Pratique d'utilisation (faibles coûts, stabilité, facilité d'administration).

Nous pouvons alors classer les vaccins selon leur sûreté, leur nature ainsi que leur immunogénicité (figure 1).

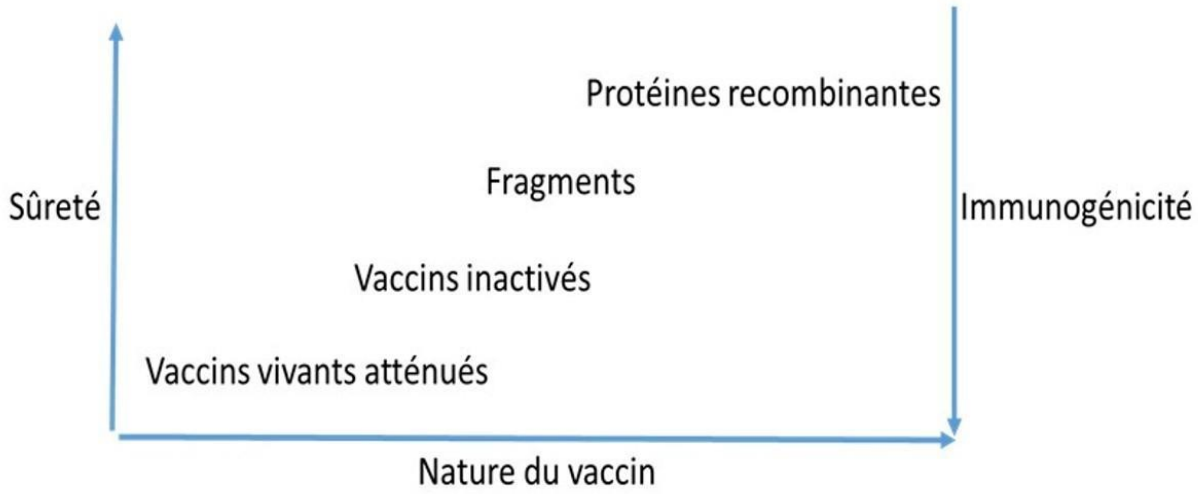


Figure 1 : Représentation de la sûreté et de l'immunogénicité des vaccins en fonction de leur nature.

Les premiers vaccins obtenus historiquement sont des vaccins vivants atténués et des vaccins inactivés. Ce qui signifie que les micro-organismes sont entiers et sont donc proche de l'infection provoquée par les micro-organismes d'origine. Ces derniers vaccins sont en général très immunogènes mais comportent des risques. A l'inverse, les protéines recombinantes des micro-organismes, qui entrent dans la composition des nouveaux vaccins (vaccin contre les papillomavirus humains par exemple), sont peu immunogènes seules mais

très sûres. Il est donc possible d'avoir recours à des adjuvants pour favoriser la mise en place d'une réponse immunitaire suffisamment efficace (Figure 2) (3).

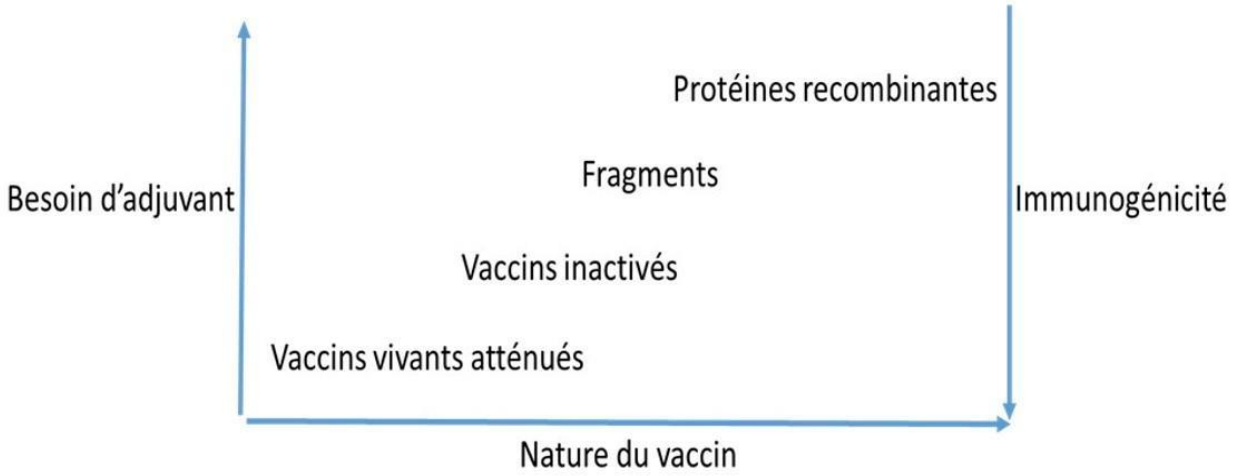


Figure 2 : Représentation de l'immunogénicité des vaccins en fonction de leur nature et de la nécessité d'un adjuvant.

2. Les différents types de vaccins

2.1. Vaccins vivants atténués

Dans le cadre des vaccins vivants atténués, l'agent virulent obtenu à partir d'un sujet infecté est affaibli par passage sur un hôte non naturel ou d'un milieu peu favorable à son développement. De cette manière l'agent infectieux se multiplie ensuite chez l'hôte naturel sans provoquer de maladie. Ainsi fut développé de BCG dans les années 1920. A l'exception du BCG, les vaccins atténués sont aujourd'hui tous faits avec des virus atténués (4). L'un des principaux risques lié à ce procédé est la possibilité de réversion à des formes virulentes comme ce fut le cas par exemple pour le vaccin oral contre la poliomyélite (5). Il est difficile de maintenir un germe vivant sans modification de son ineffectivité, tout en préservant son innocuité.

Ci-dessous les principaux exemples qui rentrent dans cette catégorie (4) : les vaccins anti- :

- Poliomyélite, vaccin oral
- Adénovirus, vaccin oral
- BCG

- Rubéole
- Oreillons
- Rougeole
- Fièvre jaune
- Varicelle
- Variole

2.2. Vaccins inactivés

Ce sont des vaccins complets où l'agent bactérien ou viral est inactivé par différents procédés chimiques mais dans des conditions telles que son immunogénicité est préservée mais ne provoque pas la maladie.

Ci-dessous les principaux exemples qui rentrent dans cette catégorie : les vaccins anti- :

- grippe (6)
- rage
- poliomyélite (7)
- coqueluche
- leptospirose
- encéphalite à tiques

2.3. Les fragments de micro-organismes

Ce sont des fractions antigéniques ou sous-unités vaccinantes qui sont soit des particules virales, fractionnées, soit des toxines naturelles détoxifiées (anatoxines), soit des antigènes capsulaires (polysaccharides de pneumocoques ou de méningocoques) ou membranaires (protéines bactériennes ou virales) (8).

Ci-dessous les principaux exemples qui rentrent dans cette catégorie : les vaccins anti- :

- tétanos : anatoxine
- diphtérie : anatoxine
- coqueluche acellulaire
- Haemophilus : polysaccharide adsorbé et conjugué
- méningocoque : polysaccharide seul ou conjugué
- pneumocoque : polysaccharide seul ou conjugué
- typhoïde : polysaccharide
- hépatite A
- grippe
- encéphalite japonaise.

2.4. Les protéines recombinantes

Ce sont des vaccins dont les protéines ont été produites par une cellule qui a vu son le matériel génétique modifié par recombinaison génétique. Les représentants de cette catégorie sont les vaccins contre l'hépatite B (9) ou contre les papillomavirus humains (10). Ces derniers vaccins ne sont d'ailleurs pas de simple protéine produite en culture cellulaire mais elles forment des autoassemblages formant des pseudo-particules virales (PPV) non infectieuse car ne contenant pas de matériel génétique (11).

2.5. Les vaccins à vecteurs viraux

Les vaccins vecteurs viraux sont des virus recombinants qui codent pour des antigènes d'intérêt dans un virus modifié non apparenté. Ils délivrent l'antigène dans les cellules imitant l'infection naturelle, de sorte qu'ils induisent en soi de fortes réponses immunitaires cellulaires et humorales spécifiques de l'antigène, évitant ainsi le besoin d'adjuvants supplémentaires (11). De plus, les vecteurs viraux sont capables d'accepter de grandes insertions dans leur génome, fournissant une plate-forme flexible pour la conception d'antigènes. Les vecteurs viraux les plus fréquemment utilisés sont les adénovirus (12) et les rétrovirus.

2.6. Les vaccins à ARNm

Dans le cas des vaccins à ARNm (acide ribonucléique messager) ce n'est pas l'antigène de l'agent pathogène qui est directement injecté dans l'organisme du patient mais seulement l'information génétique permettant aux cellules de le produire. La difficulté de conception de ces vaccins réside dans la fragilité de l'ARNm dans l'environnement et la nécessité de faire pénétrer cet ARN dans la cellule pour être traduit en protéine (13). Il est donc nécessaire de vectorisé l'ARNm dans des liposomes et de le conserver à des températures fortement négatives (11). Cette technologie vaccinale est très récente et a beaucoup fait parler de lui puisque qu'il a permis la commercialisation du premier vaccin contre le SARS-Cov2 en décembre 2020. Ce fut aussi la première fois que ce type de vaccin fut utilisé à grande échelle.

3. Les effets recherchés

3.1. Effets utiles en clinique

Le but est d'obtenir une protection efficace et durable contre des maladies graves dont un grand nombre ne bénéficie pas de traitements médicamenteux de type antibiotiques ou antiviraux (ex. poliomyélite, rougeole, rubéole, oreillons, tétanos). Cette protection par la vaccination est plus particulièrement utile à certaines périodes de la vie où certaines affections sont graves en particulier chez les nourrissons (Haemophilus, pneumocoques, méningocoques) ou, dans le cas de la grippe saisonnière, pour les personnes âgées et celles qui ont des pathologies cardio-respiratoires.

A cette protection, à visée individuelle, correspond aussi une préoccupation de santé publique qui vise à diminuer (voire éteindre) les affections contagieuses les plus fréquentes et les plus graves. Cet objectif de protection populationnelle est capital car lorsque nous sommes effectivement vaccinés contre une maladie infectieuse, nous évitons de développer cette maladie et, très souvent, cela limite la transmission de l'agent pathogène aux autres. D'autant plus qu'en se protégeant soi-même par la vaccination, nous protégeons également toutes les personnes qui ne peuvent se faire vacciner, comme les personnes malades

(immunodéprimés), les femmes enceintes ou les nourrissons car nous limitons la circulation de l'agent infectieux. Plus la maladie est contagieuse, comme la rougeole ou la grippe, plus la vaccination protège les autres personnes. C'est dans cette optique de prévention de certains cancers (col de l'utérus, cancers ORL : Oto-Rhino-Laryngée) qu'on développe aussi la vaccination préalable à l'exposition aux virus à pouvoir oncogène (pour l'instant ne concerne que certains virus PVH : PapillomaVirus Humain) (14).

3.2. Pharmacodynamies des effets utiles en clinique

L'immunisation nécessite l'intervention concertée de 2 types de cellules immunitaires, les lymphocytes T et B :

- Les lymphocytes T ne sont activés que par le peptide antigénique « préparé » par une cellule présentatrice d'antigène.
- Les lymphocytes B, quant à eux, reconnaissent directement les déterminants antigéniques de la molécule.

En matière de vaccins, il est possible de juger l'efficacité de la vaccination par différentes approches :

- La fréquence de l'infection (ou du cancer) dans une population donnée pendant les années qui suivent la vaccination.
- La gravité des cas d'infection
- La mesure de critères intermédiaires qui prouvent une immunisation contre l'agent infectieux tel que le titre des anticorps (immunité humorale) ou dans certains cas une intradermo-réaction (immunité cellulaire).

3.3. Les adjuvants

Les adjuvants permettent d'augmenter l'immunogénicité des préparations vaccinales. Les adjuvants historiques sont des dérivés de l'adjuvant de Freund (émulsion d'huile dans l'eau), de l'alun et des sels d'aluminium. Leur utilisation est avant tout empirique, mais nous avons aujourd'hui montré qu'ils agissent en permettant un relargage progressif de l'antigène, cela permet une meilleure prise en charge des antigènes par les cellules présentatrices

d'antigène et donc une réponse immunitaire plus efficace (8). Certains adjuvants plus spécifiques sont aujourd'hui à l'étude : en utilisant des dérivés de ligands de PRR (Pathogen Recognition Receptor), nous cherchons à stimuler le système immunitaire de façon sélective, et ainsi à orienter la réponse adaptative vers une réponse efficace.

4. Analyse de la réponse immune

4.1. Les anticorps

Les vaccins doivent induire la production d'anticorps protecteurs par l'individu vacciné. La neutralisation des effets pathogènes de l'agent infectieux se fait par différents mécanismes. Certains anticorps agissent sur les épitopes essentiels à l'expression du pouvoir pathogène (comme la pénétration dans la cellule). Certains s'associent au complément pour agglutiner et lyser les bactéries. D'autres vont armer des phagocytes ou des lymphocytes et les rendre capables de reconnaître et de lyser des cellules infectées par des agents à développement intracellulaire (virus).

Cette immunité humorale est transférable par le sérum. La mesure du titre de certains anticorps neutralisant est le moyen le plus simple et le plus utilisé en pratique pour évaluer la réponse immunitaire induite par un vaccin. Les anticorps ne sont en fait que l'expression finale de la réponse immunitaire. Ils sont produits par les plasmocytes et les lymphocytes B après une succession de réactions cellulaires et tissulaires provoquées par la présence et la stimulation que provoque l'antigène.

4.2. Evènements cellulaires

Les antigènes vaccinaux doivent franchir les barrières naturelles qui isolent l'organisme du milieu extérieur tel que la peau, les muqueuses ainsi que les facteurs de défense non spécifiques susceptibles de détruire les corps étrangers avant que le système immunitaire spécifique ne soit mis en jeu.

Les événements cellulaires font intervenir :

- Les cellules présentatrices d'antigène : macrophages, cellules dendritiques, faisant intervenir soit le complexe majeur d'histocompatibilité de classe II (CMH2 : protéines antigéniques, bactérie à développement extracellulaire) ou, au contraire, les complexes majeurs d'histocompatibilité de classe I pour les virus ou bactéries qui infectent les cellules phagiques (CMH1).
- Les lymphocytes T auxiliaires CD4 (Cluster de Différenciation 4) sont activés précocement soit par des peptides antigéniques associés à des molécules HLA (Human Leukocyte Antigen) de classe II, soit par l'interleukine 1 produite par les macrophages sensibilisés. Il s'ensuit une production autocrine d'interleukines, notamment d'interleukine 2 et d'interféron gamma qui jouent un rôle important dans le développement de la réponse immune.
- Les lymphocytes T cytotoxiques CD8 reconnaissent les fragments protéiques d'origine virale présentés par les molécules du CMH de classe I. Les lymphocytes T sont porteurs d'un récepteur pour l'antigène ; ils sont susceptibles de détruire in vitro comme in vivo des cellules infectées par des virus ou des bactéries à développement intracellulaire. Les lymphocytes CD4 sécrétants de l'interleukine 2 et de l'interféron gamma stimulent la réponse aux antigènes viraux et le potentiel cytolytique de ces lymphocytes CD8.
- Les lymphocytes B comportent des immunoglobulines de surface qui sont capables de distinguer la conformation spatiale des antigènes. Le complexe antigène-immunoglobuline est internalisé par endocytose. Puis ces lymphocytes vont exprimer à leur surface un peptide associé au récepteur du CMH de classe II. La présence de ces complexes est reconnue par certains lymphocytes T auxiliaires qui contribuent à la différenciation de ces lymphocytes B en plasmocytes sécrétant des anticorps par l'intermédiaire des lymphokines. Des cellules B mémoire sont également produites : elles expriment des récepteurs IgG et IgA très spécifiques et spécialisés permettant une réponse secondaire plus adaptée et plus rapide.

La réponse immunitaire implique donc dans tous les cas une étroite coopération cellulaire. Elle est fortement dépendante du complexe majeur d'histocompatibilité (CMH) et, par conséquent, dépend des caractéristiques génétiques de chaque individu. Cela expliquerait en partie la variabilité des réponses immunitaires obtenues après inoculation d'un même vaccin chez différents patients.

En résumé, la vaccination induit deux éléments qui contribuent à la défense :
- les anticorps qui neutralisent les toxines ou agents pathogènes ou favorisent la phagocytose,
- les cellules T cytotoxiques qui vont détruire les cellules infectées.

Si généralement le titre des anticorps caractérise la qualité de la réaction immunitaire, pour certains agents comme le BCG il n'y a pas de réaction humorale mesurable ; en

revanche, il est possible de mesurer les effets de l'immunité cellulaire de manière qualitative ou semi quantitative en mesurant l'inflammation et l'induration sous cutanée après injection intradermique d'une solution de tuberculine.

4.3. Phénomène de rappel

Lors de la première exposition à un antigène vaccinal, la réponse immune est lente, peu spécifique et s'exprimant initialement par des IgM (Immunoglobuline M) car la première vaccination mime une primo-infection. Lors de nouveaux contacts d'antigène, le délai de réponse se raccourcit drastiquement et les anticorps atteignent des titres beaucoup plus élevés. Il s'agit alors essentiellement d'IgG (Immunoglobuline G) dont la spécificité et l'avidité à l'antigène est plus grande. La réaction cellulaire est ainsi accélérée et intensifiée (Figure 3). Le temps de réaction du système immunitaire peut alors être suffisamment raccourci pour empêcher l'apparition de manifestations cliniques de l'infection. Cela permet d'assurer la protection du sujet mais aussi de limiter le taux de transmissibilité du virus par l'individu.

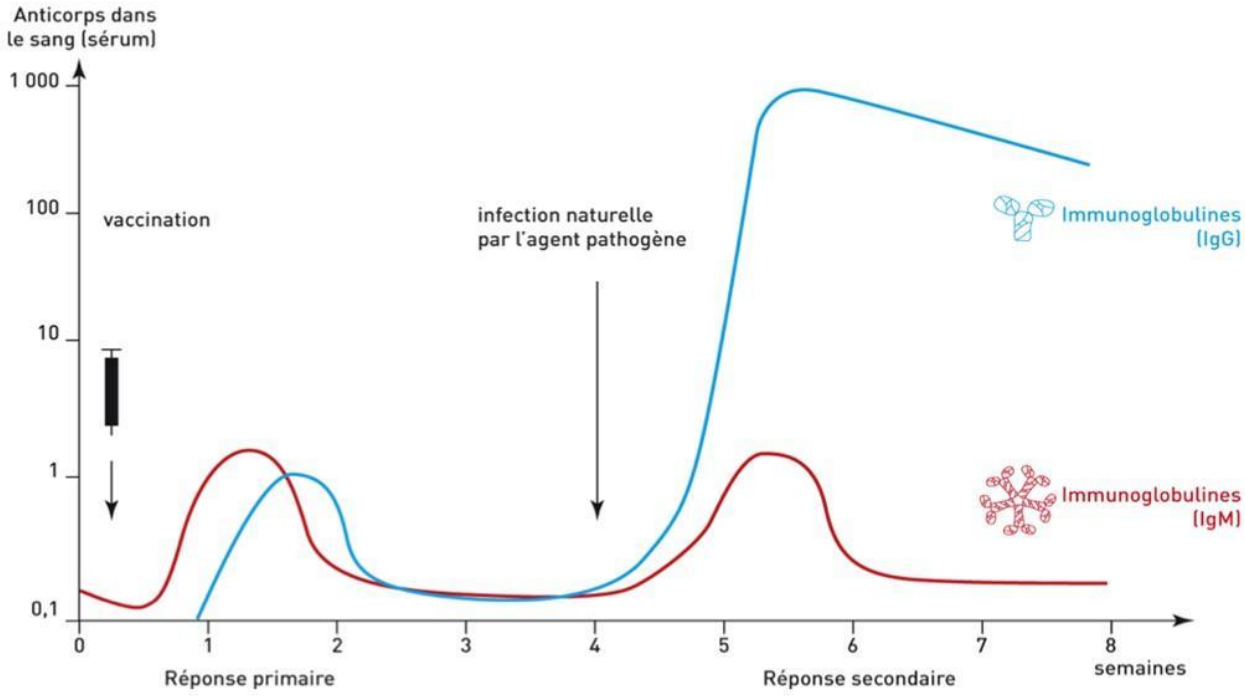


Figure 3 : Vaccination et effet mémoire au cours du temps (Source : Banque de Schémas-ENS Lyon).

Ce phénomène de rappel repose sur les cellules mémoire. Les lymphcytes T atteignent leur niveau de titre le plus élevé deux à six semaines après l'inoculation. Les cellules productrices d'anticorps augmentent lentement jusqu'à la 6ème semaine puis décroissent progressivement. Les lymphocytes B mémoire atteignent leur maximum au bout de dix à quinze semaines, avant de décroître eux aussi progressivement. Ces cellules à mémoire contribuent à la production rapide d'anticorps lors des stimulations antigéniques ultérieures (phénomène de rappel) (Figure 4). La réactivité de l'hôte à un vaccin dépend donc en partie de ses antécédents de stimulation antigénique antérieure et de l'état de son système immunitaire à l'instant de la vaccination.

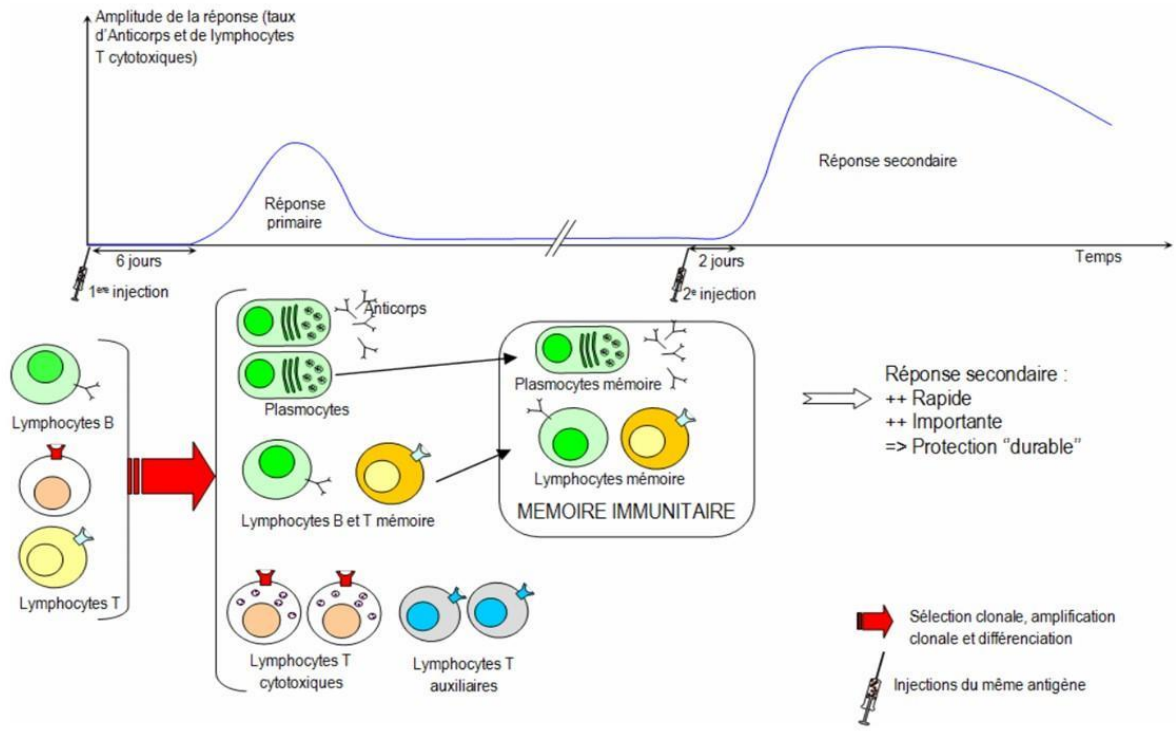


Figure 4 : Mémoire immunitaire : réponse primaire et secondaire et intervention concertée des lymphocytes T et B (Source : Banque de Schémas-Académie de Dijon).

5. Caractéristiques de l'immunogène

La réactivité de l'hôte dépend aussi des propriétés immunogéniques du vaccin. Les vaccins « inertes » uniquement constitué à base de protéine mettent en jeu la mémoire immunologique thymo-dépendante faisant intervenir les lymphocytes T mémoire. Une nouvelle injection déclenchera alors l'ascension des IgG protectrices.

Les antigènes polysaccharidiques induisent une réponse ne faisant intervenir que les lymphocytes B (thymo-indépendante). Elle sera donc moins complète et moins durable et avec un effet limité sur les rappels. L'efficacité de ces vaccins peut être très amoindrie voire nulle chez les très jeunes enfants (moins de 2 ans) (15).

Les vaccins complets induisent des réactions immunitaires de grande diversité car ils stimulent beaucoup plus violemment le système immunitaire. Certaines de ces réactions peuvent être graves.

Les vaccins sous-unités, issus d'une meilleure connaissance de la structure des antigènes et des facteurs de virulence des agents infectieux peuvent avoir une activité stimulatrice plus précise mais souvent moins intense.

Dans les cas où la stimulation du système immunitaire est insuffisante, il sera souvent nécessaire de recourir à des adjuvants (cf. parti 3.3.). L'adjuvant aura alors deux fonctions :
- garder l'antigène à proximité du site d'injection,
- activer les cellules présentatrice d'antigène de manière à favoriser la reconnaissance immune et la production d'interleukines (8).

5.1. Caractéristiques pharmacocinétiques utiles en clinique

La réponse immunitaire dépend donc en grande partie de la composition du vaccin (caractéristiques de l'immunogène, présence ou non d'adjuvant etc…), de sa voie d'introduction, de sa dose ainsi que du nombre d'administrations.

5.1.1. Voie d'introduction

L'administration de l'antigène par la voie sous-cutanée ou intradermique, voies habituelles des vaccins, entraîne une forte réponse immunitaire alors que l'administration par voie intraveineuse, surtout à fortes doses, n'est pas très immunogène, et peut même induire un état de tolérance.

A noter que depuis une vingtaine d'année, l'usage tend à indiquer la voie intramusculaire pour l'administration des vaccins bien que la voie sous-cutanée ou intradermique entraine généralement une meilleure réponse immunitaire. La raison historique

en est l'absence de constitution de nodules sous cutanés comme ce fut le cas pour des vaccins tétravalents à la fin des années 1970. La persistance de certains adjuvants tel l'aluminium semble aussi différente selon que la voie d'injection a est intramusculaire ou sous cutanée.

Certains vaccins, en particulier le vaccin anti-poliomyélite, qui est vivant mais atténué, peuvent s'administrer par voie orale. Il n'est plus indiqué en primo vaccination vu les risques de réactivation (7). D'autres vaccins utilisés dans la prévention des infections des voies aériennes peuvent s'administrer par voie perlinguale ou par voie locale, nasale et bronchique.

5.1.2. Dose d'antigène

D'une manière générale, une très forte dose d'un antigène ou son administration répétée à de très faible dose inhibent la réponse immunitaire (phénomène de tolérance). L'optimisation de dose est faite au cours des phases cliniques I et II mais sans que la mesure des antigènes circulants n'ait de valeur décisionnelle. C'est donc sur des notions de tolérance et de pourcentage de séroconversion que nous choisissons la dose usuelle.

5.1.3. Répétition de l'administration

Les vaccins s'administrent habituellement en deux ou trois fois séparées par un intervalle d'environ 2 à 4 semaines. La première administration dite sensibilisante entraîne une réponse primaire de faible intensité et transitoire. La deuxième et la troisième administration entraînent une réponse dite secondaire beaucoup plus intense et durable que la première. Les administrations postérieures à l'administration sensibilisante font intervenir la mémoire immunitaire, probablement liée à la longue durée de vie de certains lymphocytes T et B.

5.1.4. L'influence des injections successives

Pour un grand nombre de vaccins où l'immunité ne paraît pas assez durable, il est effectué des rappels, le plus souvent 18 mois après la primo vaccination puis tous les 5 ans à l'âge adulte. Pour certains vaccins une seule série d'injection suffit. C'est le cas pour la rubéole ou les oreillons ; en revanche pour la rougeole un rappel à l'âge adulte s'avère nécessaire.

Certains vaccins ne demandent qu'une seule administration. Le rappel se faisant par exemple un an plus tard dans le cas de grippe mais sa composition change d'une année à

l'autre. Elle peut se faire dix ans plus tard dans le cas de fièvre jaune. Un rappel à un an et éventuellement plusieurs années plus tard réactive la réponse immunitaire. Comme la demi-vie des anticorps (en particulier celle des IgG qui est pourtant la plus longue) est de seulement 21 jours, se sont les lymphocytes B stimulés par le vaccin qui continuent à synthétiser des anticorps longtemps après la vaccination. Cette persistance s'explique par la longue durée de vie, de l'ordre de plusieurs années, de certains lymphocytes. Par ailleurs certaines infections qui passent inaperçues lors de diverses épidémies peuvent jouer un rôle dans la réactivation du système immunitaire.

5.2. Situations à risque ou déconseillées

Les contre-indications réelles des vaccinations sont extrêmement limitées ; elles sont explicitées dans l'AMM de chacun des vaccins. Nous pouvons malgré tout retenir que les vaccins vivants sont, d'une façon générale, contre-indiqués en cas d'immunodépression et, le plus souvent, contre-indiqués chez la femme enceinte.

En deuxième lieu, il conviendra de préciser le statut allergique du patient (allergie notamment à certains antibiotiques : Néomycine, streptomycine, kanamycine etc... car certains vaccins en contiennent des traces). En cas de réaction vaccinale exagérée (telle les encéphalites), nous pouvons être amené à prescrire des corticoïdes. Dans ce cas, leur effet immunosuppresseur supprime les effets de la vaccination.

5.3. Précautions d'emploi

La vaccination est une thérapeutique préventive et donc les risques induits doivent être aussi bas que possible. Cette évaluation de la balance bénéfice/risque est évidemment modifiée en cas de brutale épidémie d'une affection nouvelle et dont les prémices paraissent redoutables. Nous pouvons nous rappeler de l'exemple de la pandémie H1N1 (Hémagglutinine 1 Neuraminidase 1) en 2009-2010. Sur des données épidémiologiques fragmentaires venues des premières infections dans l'hémisphère sud l'OMS (Organisation Mondiale de la Santé) recommanda une vaccination de masse le plus tôt possible. L'Europe et l'Amérique du nord firent fabriquer en urgence des dizaines de millions de doses de vaccins. Il fût supposé qu'il fallût 2 injections pour obtenir une immunité satisfaisante. En France les commandes furent massives mais la vaccination fût très minoritaire du fait d'une méfiance voire d'une défiance

générale vis-à-vis des vaccins et d'une faible adhésion de certains professionnels de santé au programme de vaccination (16). Avec le recul nous avons pu constater que quelques centaines de personnes, sans facteur de risque préalable et le plus souvent très jeunes, étaient décédées du fait de l'infection (alors que chaque année quelques milliers de patients âgés ou de santé précaire décèdent de la grippe saisonnière). Il fut aussi découvert a posteriori quelques cas de syndrome de narcolepsie (pathologie grave du système veille/sommeil) imputables au vaccin sans qu'aucun mécanisme ou facteur de risque ne puisse être clairement identifié. Cet épisode donne un aperçu des principaux points qui sont continuellement discutés à propos de la vaccination car ce type de médicament relève d'un usage éminemment politique.

- Intérêt collectif contre risque individuel.

- Rapport coût/efficacité d'une campagne en particulier en cas d'urgence face à un agent infectieux nouveau (17).

- Difficulté à faire passer des informations et décisions d'action préventive qui sont anxiogènes dans les pays démocratiques où l'accès à de nombreuses informations contestataires et non contrôlées peut induire rapidement des forces d'oppositions à des mesures de santé publique urgentes (16).

Plus récemment en France nous avons assisté à de nombreux débats sur l'intérêt et les éventuels risques à faire passer la vaccination des jeunes enfants de 3 à 11 vaccins (devenus obligatoires en 2018) avant l'entrée en collectivité.

5.4. Effets indésirables

Ils sont nombreux dans leur énumération et varié dans leur diversité mais ceux qui concentre l'attention sont les effets indésirables graves (EIG) rares voire très rares (moins de 1/10 000 sujets exposés). Ce sont ces derniers qui amènent des conflits durables entre les autorités de santé et des associations de patients qui ont été victimes de tels EIG. La faible incidence de ces EIG les rend donc quasi indétectable lors des phase clinique I II et III. Elles sont donc souvent détectées après la commercialisation du vaccin lors de la phase IV. Il reste problématique que les autorités sanitaires puissent exiger des preuves de l'origine vaccinale de ces EIG tandis qu'en pharmacovigilance nous n'en sommes généralement qu'à établir des relations comme vraisemblables ou très vraisemblables. Ces relations sont d'autant plus délicates à établir que les réactions immunitaires de chacun sont très variables et donc que

nous entrons dans un effet de type idiosyncrasique selon la classification et non dans un effet de type classique qui implique une relation dose/exposition dépendant.

Parmi les effets indésirables graves certains sont communs entre la maladie et la vaccination. Ainsi le syndrome de Guillain – Barré peut être observé après vaccination contre la grippe saisonnière, en revanche sa fréquence estimée est environ 10 fois plus faible qu'après avoir contracté la maladie de la grippe.

Dans d'autres cas des EIG ont été observés après vaccination alors qu'ils n'existent pas dans le cadre de la maladie naturelle. Cela est le cas pour des épisodes d'inflammation chronique du système nerveux central (IC-SNC) après vaccination contre le virus de l'hépatite B (VHB). Si ces cas sont très rares et parfois mal documentés ils ont néanmoins les apparences d'épisodes de sclérose en plaques (SEP). En l'absence de revaccination les phases de démyélinisation sembleraient pouvoir s'amender mais dans certains cas des séquelles très lourdes persistent après les épisodes d'IC-SNC, toutefois rien ne prouve la responsabilité du vaccin dans cette inflammation (18).

Sachant que chez les sujets adultes et sains un décès totalement inexpliqué peut survenir avec une incidence de 1/1 000 000 de sujets, il est statistiquement inévitable que surviennent un décès de ce type à la suite d'une vaccination. C'est là toute la difficulté de détecter et alerter sur des effets graves et très rares à la suite de la vaccination.

Plus globalement, nous pouvons être rassurant sur les vaccinations usuelles pratiquées en occident car elles sont globalement très sûres. Bien évidemment différents signes bénins sont fréquents (ex. Rhino-conjonctivite après vaccination anti-morbilleuse, fièvre et troubles de l'appétit après les premières vaccinations plurivalentes du nourrisson, douleur voire inflammation au point d'injection, etc…). Les résumés des caractéristiques du produit et notices d'information des patients explicitent bien ces effets indésirables. Il est donc important de les relire et de les expliquer avant de vacciner.

5.5. Surveillance des effets de la vaccination

La surveillance des effets indésirable doit être adaptée à la gravité des signes cliniques ; ainsi en cas de réaction encéphalique, il est urgent d'hospitaliser en neurologie voire en réanimation. Les complications cardio-pulmonaires peuvent aussi être très préoccupantes en particulier chez le sujet fragilisé ou le très jeune nourrisson ce qui doit amener là aussi à hospitaliser sans délai.

Quant à la surveillance des effets bénéfiques, elle est assez complexe. Elle va reposer principalement sur une analyse épidémiologique. Le nombre de cas et donc l'évolution de l'incidence ou de la non-récidive d'une maladie fréquente (ex. grippe saisonnière) sera compilé. Ces données sont moins tangibles quand il s'agit d'affections plus rares ou de gravité variable et plus encore si nous avons à faire à une couverture vaccinale partielle et/ou des expositions au risque restreintes à des populations dites « à risques » mais peu surveillées (comme pour les hépatites B).

Dans certains cas, il sera recommandé de reprendre la vaccination ou faire un rappel selon les résultats d'une sérologie. Chaque vaccin a son profil de durabilité, les campagnes de prévention vaccinale doivent donc s'adapter à ces particularités ainsi qu'aux priorités de santé publique.

6. Conclusion

La vaccination a pour objectif essentiel d'induire une immunité de population. Ainsi les individus sont protégés directement (immunité active) et indirectement (immunité de groupe), l'état immunitaire de la population créant un obstacle à la circulation des agents infectieux. Dans de nombreuses maladies, en particulier virales, c'est le moyen le plus sûr et le plus efficace d'échapper à la maladie alors que celle-ci peut donner des formes très graves (rougeole, poliomyélite, hépatite B, Covid19) voire rapidement mortelles (tétanos, diphtérie, méningites à méningocoques, broncho-pneumopathies à pneumocoques, Haemophilus).

Cependant étant donné la nature préventive du vaccin, la tolérance aux effets secondaire au sein de la population peut être très faible. Nous avons aussi pu voir, avec l'exemple récent des vaccins contre le covid19, une forte appréhension face à des vaccins d'une technologie nouvelle (ARNm) et souvent mal comprise par la population (16). Nous verrons alors dans cette thèse comment des méthodes bioinformatiques peuvent aider à la compréhension et à la conception de certains vaccins afin d'en améliorer soit sa production soit son efficacité.

Enfin, pour ceux qui gardent des doutes voire une opposition à toute vaccination, rappelons-nous les millions de vies épargnées durant le siècle dernier où les grandes maladies contagieuses sévissaient encore (variole aujourd'hui disparue dans le monde, et poliomyélite aujourd'hui disparue en France). Et pour ce qui est des maladies plus récentes, combien de patients auraient aimés bénéficier d'un vaccin anti-VIH, anti-VHC, anti-Ebola ou encore

aujourd'hui quand des virus émergeants réapparaissent ou se révèlent comme c'est le cas actuellement avec la pandémie de SARS-Cov2 qui affecte le monde et nos sociétés depuis 2020.

BIBLIOGR APHIE DU CHAPITRE 1

1. Gross CP, Sepkowitz KA. The myth of the medical breakthrough: smallpox, vaccination, and Jenner reconsidered. Int J Infect Dis IJID Off Publ Int Soc Infect Dis. sept 1998;3(1):54-60.

2. Cohen JM. « Remarkable solutions to impossible problems »: lessons for malaria from the eradication of smallpox. Malar J. 23 sept 2019;18(1):323.

3. Zimmermann P, Curtis N. Factors That Influence the Immune Response to Vaccination. Clin Microbiol Rev. 20 mars 2019;32(2).

4. Minor PD. Live attenuated vaccines: Historical successes and current challenges. Virology. mai 2015;479-480:379-92.

5. Wyatt HV. Polio immunization: benefits and risks. J Fam Pract. sept 1978;7(3):469-74.

6. Trombetta CM, Gianchecchi E, Montomoli E. Influenza vaccines: Evaluation of the safety profile. Hum Vaccines Immunother. 4 mars 2018;14(3):657-70.

7. Bandyopadhyay AS, Garon J, Seib K, Orenstein WA. Polio vaccination: past, present and future. Future Microbiol. 2015;10(5):791-808.

8. Jarząb A, Skowicki M, Witkowska D. [Subunit vaccines--antigens, carriers, conjugation methods and the role of adjuvants]. Postepy Hig Med Doswiadczalnej Online. 27 nov 2013;67:1128-43.

9. Gerlich WH. Medical virology of hepatitis B: how it began and where we are now. Virol J. 20 juill 2013;10:239.

10. McKee SJ, Bergot A-S, Leggatt GR. Recent progress in vaccination against human papillomavirus-mediated cervical cancer. Rev Med Virol. mars 2015;25 Suppl 1:54-71.

11. Li Y-D, Chi W-Y, Su J-H, Ferrall L, Hung C-F, Wu T-C. Coronavirus vaccine development: from SARS and MERS to COVID-19. J Biomed Sci. 20 déc 2020;27:104.

12. Tatsis N, Ertl HCJ. Adenoviruses as vaccine vectors. Mol Ther J Am Soc Gene Ther. oct 2004;10(4):616-29.

13. Zhang C, Maruggi G, Shan H, Li J. Advances in mRNA Vaccines for Infectious Diseases. Front Immunol. 2019;10:594.

14. Wang R, Pan W, Jin L, Huang W, Li Y, Wu D, et al. Human papillomavirus vaccine against cervical cancer: Opportunity and challenge. Cancer Lett. 28 févr 2020;471:88-102.

15. Saso A, Kampmann B. Vaccine responses in newborns. Semin Immunopathol. nov 2017;39(6):627-42.

16. Dubé E, Laberge C, Guay M, Bramadat P, Roy R, Bettinger JA. Vaccine hesitancy. Hum Vaccines Immunother. 1 août 2013;9(8):1763-73.

17. Doherty M, Buchy P, Standaert B, Giaquinto C, Prado-Cohrs D. Vaccine impact: Benefits for human health. Vaccine. 20 déc 2016;34(52):6707-14.

18. Mailand MT, Frederiksen JL. Vaccines and multiple sclerosis: a systematic review. J Neurol. juin 2017;264(6):1035-50.

Chapitre 2 : Modélisation et mécaniques moléculaires

1. Hiérarchisation des structures biologiques : De l'atome à la macromolécule

L'étude exhaustive des phénomènes biologiques nécessite la connaissance fine des structures des différentes molécules impliquées dans ceux-ci. Les principales molécules effectrices de nos cellules sont les protéines. Elles peuvent agir et interagir de concert entre elles, mais aussi avec des plus petits éléments organiques comme, entre autre, des lipides, des neurotransmetteurs, des ions etc. Cependant, ces complexes macromoléculaires sont de taille et de composition très variables. En effet, la taille des structures (poly)peptidiques peut varier de quelques angströms (Å) à plusieurs centaines et peuvent être lié à des structures peptidiques, des lipides (membranaires ou non), des molécules organiques, des ions, etc. Les entités de nature polypeptidiques peuvent se structurer pour transmettre un signal cellulaire ou alors pour empêcher la transmission de ce signal, afin d'aboutir à la régulation positive (ou activatrice) ou négative (ou inhibitrice) d'un processus physiologique au niveau cellulaire. Les éléments impliqués dans l'initialisation et la modulation des processus cellulaires sont appelés effecteurs et sont de nature très variable : ligand endogène (cytokine, …) ou exogène (cofacteur, principe actif, …), modifications post-traductionnelles des protéines (phosphorylation) ou mutation génétique (insertion / délétion ou mutation ponctuelle). L'acteur central reste néanmoins la protéine sur laquelle l'effecteur induit une action. Pour comprendre l'ensemble de ces processus, la description la plus fine et la plus fidèle possible est donc nécessaire. Ainsi la première étape de ce travail réside dans la compréhension des différents niveaux d'organisation des atomes qui forment et structures les protéines.

Le premier niveau d'organisation des protéines est la structure primaire. Elle va déterminer la succession des acides aminés composants la protéine, nous employons alors le terme de séquence primaire (cf. Figure 5). De cette séquence primaire, une organisation locale et souvent régulière entre résidus apparaît et forme des structures particulières. Cette organisation régulière des acides aminés entre eux sont appelés structures secondaires. Nous

pouvons dénombrer 3 principales familles que sont les hélices, les beta-feuillets et les boucles qui eux sont des fragments non-structurés. De l'organisation de ces structures secondaires va survenir le repliement tridimensionnel (structure 3D) de la protéine que nous nommerons structure tertiaire et duquel la fonction protéique est fortement liée. Enfin, une architecture pluriprotéique peut apparaître afin de former des complexes macromoléculaires (stables ou métastables) : nous parlons alors de structure quaternaire. L'ensemble de ces structures est étroitement relié à la nature des interactions stéréo-physico-chimiques entre les éléments, et à la quantité d'énergie associée à ces interactions.

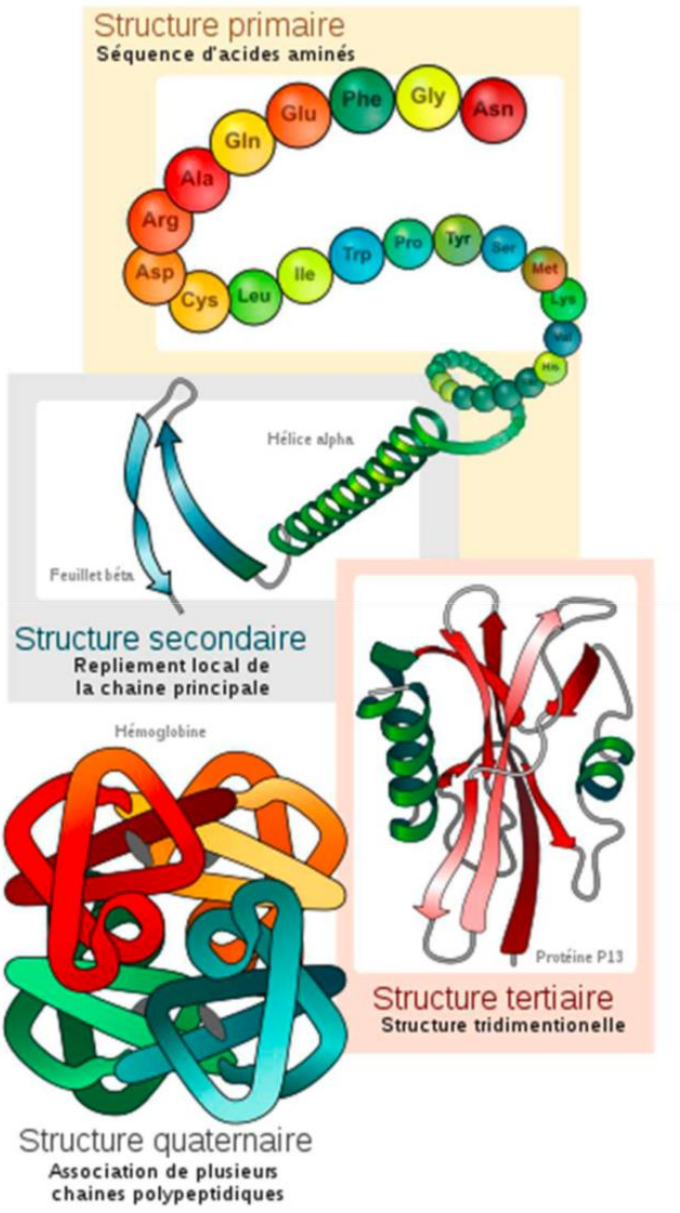


Figure 5 : Les différents niveaux d'organisation des protéines. (Image issue du site http://www.courspharmacie.com/biochimie/structures-des-proteines.html)

La structure primaire des protéines, résultant de la liaison covalente de deux résidus d'acides aminés (liaison peptidique), est associée à une énergie très importante (>100 kcal/mol). Cette liaison n'est donc formée ou rompue qu'à l'aide d'enzymes très spécialisées. L'hydrolyse de telles liaisons est possible mais extrêmement lente, elles constituent les liaisons les plus stables dans les conditions physiologiques. Les autres niveaux structurels sont liés à des liaisons non-covalentes ayant une énergie plus faible, dont la formation et la rupture est en partie possible dans les conditions physiologiques. Ils vont donc apporter une certaine plasticité à la protéine et autoriser les mouvements au sein des complexes : nous parlons alors de dynamique conformationnelle ou de transition macromoléculaire.

Les structures secondaires, formées par des liaisons non-covalentes, restent des structures le plus souvent stables car associées à des gains d'énergie importants. Elles résultent de la formation de liaisons hydrogène entre les atomes de la chaîne principale (composée des atomes communs à tous les acides aminés, la chaîne latérale étant formée à contrario de groupements spécifiques à chaque acide aminé) de résidus plus ou moins éloignés dans la séquence primaire. Dans les structures de type hélice, l'éloignement des résidus conditionne le pas du tour de l'hélice, et donc son type. Si l'énergie d'une de ces liaisons hydrogène isolée peut paraître faible comparée aux liaisons covalentes (entre 0,5 et 3,5 kcal/mol (1,2)), la stabilité des structures secondaires est généralement associée à la présence d'un grand nombre de liaisons hydrogène. Cependant, certaines protéines montrent des variations très importantes de leurs structures secondaires, comme la calmoduline dont l'hélice centrale peut se courber. Il existe aussi des protéines caractérisées par l'absence totale de structures secondaire.

La structure tertiaire des protéines est associée à différentes forces, covalentes (ponts disulfure) ou non (interactions de van der Waals, électrostatiques etc…). La formation des ponts disulfure (une liaison covalente entre les atomes de soufre de la chaîne latérale de deux résidus de cystéine) est associée à une énergie de 70 kcal/mol environ (3). Le gain énergétique, à l'échelle de la protéine ou du peptide, du repliement tridimensionnel associé est cependant plus faible, de l'ordre de 2 à 5 kcal/mol (4,5). Surtout, contrairement aux liaisons peptidiques, les ponts disulfures sont sensibles notamment aux attaques d'agents nucléophiles et leur rupture a généralement pour conséquence une perte de fonction liée à la perte de la structure secondaire de la protéine. Les forces non-covalentes qui participent à la

formation des structures tertiaires et quaternaires peuvent être de deux types : forces de van der Waals ou interactions électrostatiques.

Les forces de van der Waals résultent de l'interaction transitoire entre les nuages électroniques de deux atomes. Généralement de faible énergie (< 1 kcal/mol), elles jouent néanmoins un rôle non-négligeable étant donné le grand nombre d'interactions présentes au sein de complexes de plusieurs milliers d'atomes. Les forces électrostatiques proviennent de l'effet réciproque de deux charges électriques, et sont décrites par la loi de Coulomb. Elles peuvent être attractives (les charges du dipôle sont opposées, ex. : pont salin) ou répulsives (les charges du dipôle sont de même nature), et sont généralement faibles (< 2 kcal/mol). Enfin, des interactions apolaires (entre groupements non chargés) surviennent également au sein des protéines, des effets hydrophobes se mettent alors en place, qui peuvent avoir une contribution considérable dans la stabilisation de la structure 3D. La présence omniprésente de molécules d'eau autour de la protéine (l'environnement naturel majoritaire de la plupart des biomolécules) peut considérablement pénaliser énergétiquement la protéine si des résidus hydrophobes sont positionnés à la surface. Les résidus hydrophobes vont être tournés vers l'intérieur de la protéine afin de limiter la surface en contact avec l'eau. Ils forment ainsi au sein des protéines un cœur hydrophobe présentant de multiples interactions, faibles individuellement (< 0.7 kcal/mol) mais fortes une fois regroupées (>40 kcal/mol) (6).

La résultante de ces forces explique la structure tridimensionnelle des protéines (forme globulaire ou linéaire, nature des structures secondaires) mais également la formation (ou séparation) de complexes macromoléculaires. La notion de « complexe macromoléculaire » est omniprésente au niveau cellulaire puisqu'une cellule est un environnement saturé en protéines, et que chaque protéine est à la fois au centre de son réseau d'interactions protéique et à la marge des réseaux de ses protéines partenaires. Réseaux auxquels il faut rajouter les partenaires de natures diverses (hormones, cytokines, sucres, molécules inorganiques, etc…). Les interactions entre les différents membres d'un réseau ont ainsi la capacité d'influer l'un sur l'autre, et de modifier leurs propriétés respectives pour moduler leur activité et permettre l'acquisition de nouvelles propriétés de la protéine (capacité de fixation de nouveaux ligands en autre). Ce phénomène est essentiellement dynamique car il s'appuie sur la modification permanente de l'environnement à la fois interne (perturbation du réseau d'interactions entre les résidus) et externe (perturbation du réseau d'interactions entre macromolécules) des différents effecteurs. In fine, l'ensemble de ces modifications

dynamiques permet la transmission d'un signal cellulaire, conduisant par exemple à l'activation de la transcription d'un groupe de gènes.

L'aspect dynamique revêt une importance primordiale dans la compréhension de la fonction d'une protéine, car sa dynamique reste étroitement associée à sa structure tridimensionnelle. L'intégration de la dynamique moléculaire est l'élément clé pour explorer les relations entre séquence, structures et fonctions d'une protéine.

2. La caractérisation expérimentale des structures protéiques

La recherche de la plus petite unité du vivant a longtemps constitué centre d'intérêt important pour la science. La découverte des cellules par Robert Hooke en 1665 a constitué un premier pas qui s'est prolongé avec la caractérisation successive des organites intracellulaires (le noyau cellulaire par Robert Brown en 1831, la mitochondrie par Albert von Kölliker en 1857, l'appareil de Golgi par Camillo Golgi en 1883, etc.). La biologie structurale a continué à étudier ce pan de la biologie et produit aujourd'hui de nombreuses structures protéiques de tailles variées (de petits polypeptides à des ribosomes entiers) et à différentes résolutions (de l'ordre de l'angström Å par rayons X à plusieurs dizaines d'angströms par cryo-microscopie électronique (cryo-ME)). Le développement des bases de données a également joué un rôle primordial dans la publication et le partage des structures résolues. Aujourd'hui, plus de 170 000 structures, dont la grande majorité a été résolue par cristallographie aux rayons X, sont accessibles à tous (cf. Figure 6).

2.1. Détermination des structures protéiques par cristallographie aux rayons X

Cette technique expérimentale est basée sur l'interprétation de la diffraction de rayons X à travers un cristal obtenu à partir d'une solution concentrée de la protéine d'intérêt. Cette solution contient, entre autre que la protéine, d'autres produits comme typiquement, un ligand, un inhibiteur ou encore des ions, ainsi que des produits stimulant la solvatation ou l'initiation de la cristallisation. La cristallogenèse est le processus qui permet d'obtenir des cristaux et

constitue une étape cruciale : plus le cristal sera pur et périodique, moins la carte de diffraction obtenue comprendra de bruit issu des impuretés du signal. Elle représente ainsi une étape souvent limitante dans le processus de cristallographie. Les principales conditions qui régulent la qualité du cristal sont la pureté de la solution protéique, la concentration, le pH, la température, la présence d'ions, l'emploi d'additifs et le savoir-faire du cristallographe. Afin de garder les protéines en solution dans leurs conformations physiologiques, la cristallogenèse emploie des processus lents tels que la diffusion en phase vapeur pour générer des cristaux dont la taille est de l'ordre du micron (µm) et qui sont composés malgré tout de molécules de solvant aqueux à hauteur de 20-80 %. Des molécules annexes sont ajoutées en règle générale, afin de permettre la formation de contacts cristallins et de favoriser la stabilité des protéines, améliorant ainsi la qualité de la diffraction.

Au cours de la cristallisation les molécules forment une structure périodique, ou un cristal. Les critères fondamentaux sont donc la pureté du cristal, ainsi que l'agencement plus ou moins ordonné des espèces cristallisées (la périodicité) et sa taille (la taille minimale explorable est 5 x 10 x 30 μm^3). Un cristal est une structure solide, composée d'un arrangement ordonné et périodique des éléments qui la composent, dans toutes les directions de l'espace : cet arrangement constitue la maille du cristal, qui est présent de manière répétée par translation dans le cristal. Lorsque la maille est composée de protéines disposées de manière symétrique, une unité asymétrique est constituée du plus petit espace dont la répétition (par rotation, symétrie ou inversion) permet de reconstituer la maille et donc le cristal.

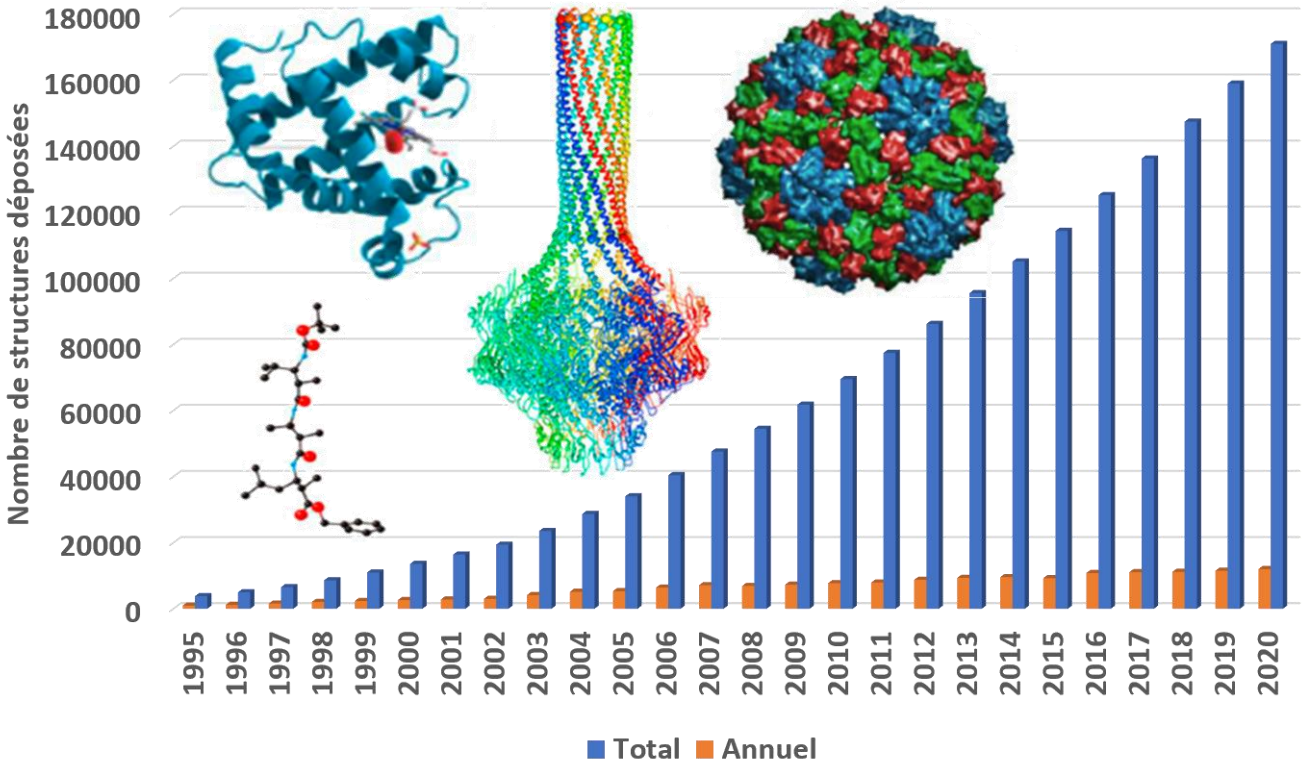


Figure 6 : Evolution du nombre des structures présentes dans la Protein Data Bank et leur complexité croissante. Le nombre total par année et le nombre des structures déposées chaque année sont respectivement montrés en bleu et en orange. Les structures, de simple à complexe, sont illustrées par un peptide, un lysosyme, un fragment de bactériophage et un virus entier.

Les rayons X sont diffractés au contact des nuages électroniques des atomes constituant des motifs répétitifs du cristal. La détection de ces motifs répétitifs par Max von Laue (récompensé par un prix Nobel en 1914) a constitué la base du développement de cette technique qui a abouti en 1957 à la première structure cristallographique (7). Le faisceau de rayons X qui rencontre le cristal provoquant la dispersion du faisceau lumineux dans des directions spécifiques. La répétition périodique des mailles permet la détection de motif périodique par les détecteurs, qui renseignent ainsi la taille de la maille cristalline et génèrent les cartes de diffraction à partir desquelles est déterminée la carte de densité électronique. À partir de cette densité, la position moyenne des atomes du cristal peut être déterminée, ainsi que leurs liaisons chimiques, leur entropie et d'autres informations. La construction et l'affinement du modèle diffractant sont réalisés en incorporant la composition supposée du cristal (séquence primaire de la protéine) et en comparant l'intensité de diffraction calculée avec celle observée.

La cristallographie permet ainsi d'obtenir les coordonnées atomiques des atomes dont le nuage électronique est suffisant pour diffracter les rayons X, ce qui exclut de fait les atomes d'hydrogène, à l'exception des structures à très hautes résolution (<1 Å) (8). La précision et l'exactitude des coordonnées, groupées sous le terme de résolution globale, sont exprimées en angströms (Å) : les atomes séparés par une distance similaire ou inférieure ne seront pas discernables au niveau de la carte de densité électronique, mais la position des atomes sera généralement déduite à partir de la forme de la carte de densité et des connaissances de la séquence primaire (nature du ou des acides aminés impliqués). Cependant, la résolution globale de la protéine n'est pas uniforme, et elle va grandement dépendre de la régularité des conformations de la protéine du cristal. Les régions les plus flexibles ne peuvent être résolues, car certains fragments des protéines dans le cristal affichent des variations de position trop importantes, produisant une carte des densités électroniques indéchiffrable. Ces régions non résolues peuvent consister en un simple groupement de chaînes latérales jusqu'à un domaine entier de la protéine. La taille des complexes macromoléculaires (composés de protéines, ligands et/ou acides nucléiques) étudiés varie ainsi de quelques acides aminés à plusieurs milliers, pour des résolutions inférieures à 3 Å (9).

2.2. Méthodes émergentes de caractérisation expérimentale des structures

Il existe d'autres méthodes expérimentales de résolution de structures protéiques. Parmi celles qui ont permis la résolution du plus grand nombre de structures (en dehors de la cristallographie aux rayons X) nous pouvons citer la résonance magnétique nucléaire (RMN) et la cryo-microscopie électronique (cryo-ME). A titre illustratif sur les 170 968 structures que comptait la Protein Data Bank (PDB) fin-2020, 151 463 sont issues de la cristallographie aux rayons X, 13 182 de la RMN et 6 385 de la cryo-ME.

L'usage de la RMN pour déterminer la structure tridimensionnelle des macromolécules en solution commence en 1985 par l'équipe de Kurt Wüthrich (10) et lui valut un prix Nobel en 2002. C'est à l'époque la deuxième technique permettant de résoudre la structure tridimensionnelle des macromolécules biologiques à l'échelle atomique mais ce n'est pas son seul apport à la biologie structurale. L'une de ses contributions majeures est certainement d'avoir popularisé l'idée que les macromolécules biologiques sont des objets profondément

dynamiques, animés de mouvements à toutes les échelles de temps (de la picoseconde à la minute), et d'avoir donné à la biologie les moyens d'explorer ces mouvements avec une résolution atomique. Cette observation est rendue possible car contrairement à la cristallographie aux rayons X, l'analyse d'un échantillon par RMN se fait en phase liquide. Une des limitations principales de la RMN étant la difficulté à résoudre la structure des protéines de grande taille rendant sont usage limité à des protéines de poids inférieurs à 30 kilodalton (kDA) soit généralement des protéines de moins de 300 acides aminés.

La méthode de cryo-ME correspond à une technique particulière de préparation d'échantillons biologiques utilisée en microscopie électronique en transmission. Elle fut développée vers la fin des années 1980, lorsque l'équipe de Jacques Dubochet proposa d'utiliser de l'éthane liquide pour la préparation des échantillons de protéines, de macromolécules ou de virus. La congélation ultra rapide et à très basse température (inférieure à -185 °C) était, et reste toujours, la clé de cette préparation. En effet, la vitesse de congélation est suffisamment élevée pour éviter que l'eau ne cristallise. Cela favorise alors uniquement la formation de glace à l'état vitreux (glace amorphe) (11). L'échantillon est alors placé dans le vide poussé du microscope électronique puis observé, à -180 °C. Un des avantages de cette méthode est que le complexe macromoléculaire peut rester dans un environnement relativement proche de ses conditions physiologiques tant en termes de pH et que de concentration en sel. Pendant longtemps cette technique n'était pas assez résolutive pour reconstruire la structure tout-atome des protéines avec un fort degré de certitude mais avec les progrès techniques constants qui ont et continuent d'être accomplis ces dernières années, de plus en plus de macromolécules avec une résolution <4,5 Å sont ajoutés chaque année à la PDB par la méthode de cryo-ME. Bien qu'à ce jour, il demeure plus de structure résolut par RMN (13227) que par cryo-ME (6674), nous pouvons noter que le nombre de découvertes annuelle de nouvelles structures par la méthode de cryo-ME augmente exponentiellement ces dernières années tandis que celles issues de la RMN sont en nombre décroissant depuis 2007

(figure 7). La cryo-ME est donc une méthode sur laquelle nous pourrons de plus en plus compter dans le futur.

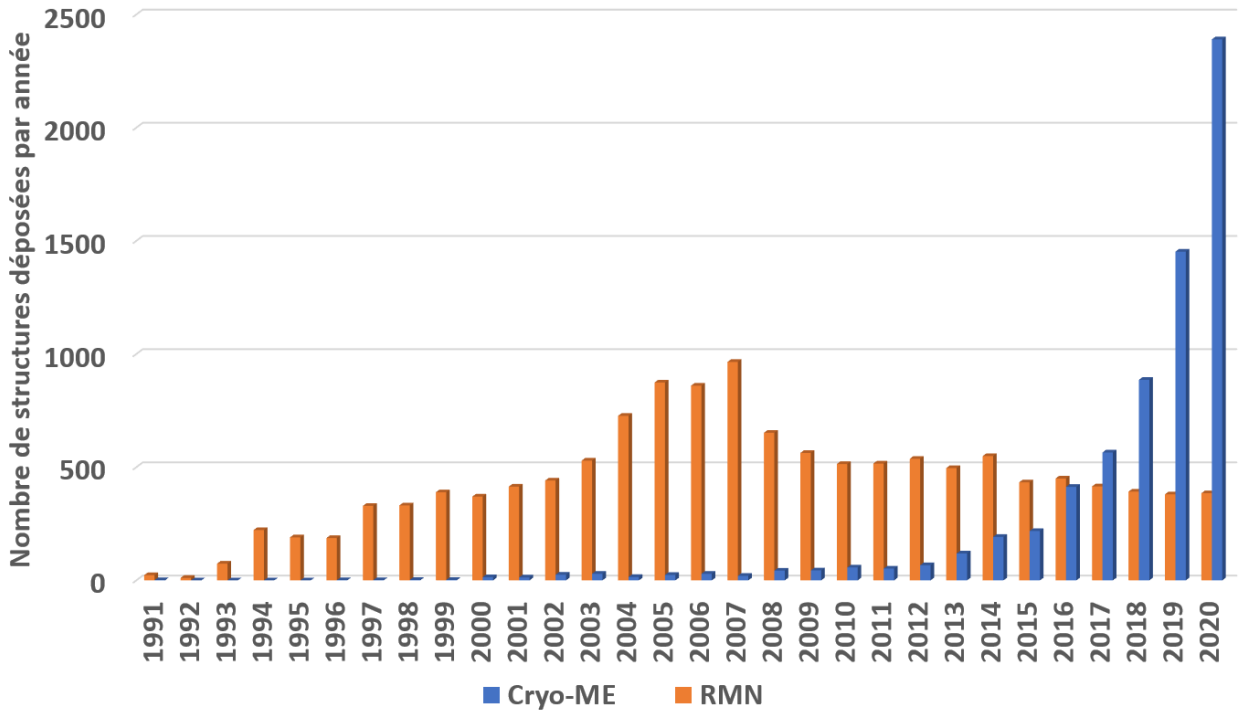


Figure 7 : Evolution du nombre de structures annuellement découvertes en fonction de la méthode utilisée. En bleu le nombre de structure résolus par cryo-ME et en orange par RMN.

2.3. La Protein Data Bank (PDB) : une précieuse source de structures

La PDB (Protein Data Bank, www.rcsb.org) (12) est la base de données mondiale qui regroupe la quasi-totalité des informations que nous possèdons concernant les structures tridimensionnelles de molécules biologiques, quelles que soient leur nature (protéique ou acide nucléique), leur origine (humaine, bactérienne, virale) ou la méthode expérimentale utilisée (cristallographie aux rayons X, RMN, Cryo-ME…). D'autres banques de données coexistent (BMRB : Biological Magnetic Resonance Data Bank, spécialisée dans le recueil de structures résolues par résonance magnétique nucléaire, etc.) mais la PDB constitue actuellement la banque de données de référence par son exhaustivité (plus de 170 000 structures disponibles fin-2020). Cette base de données a standardisé le format des fichiers

décrivant les structures, et permet l'intégration de toutes les informations nécessaires à la description de la structure (type d'atomes, leurs coordonnées, facteurs thermiques) et à son obtention (organisme de la protéine et système d'expression, séquence peptidique de la protéine versus séquence peptidique résolue, présence de ligands, etc…). Enfin, malgré son exhaustivité, de nombreux systèmes macromoléculaires sont présentés plusieurs fois avec des changements plus ou moins importants (résolution améliorée, présence de ligands différents etc…) tandis que de nombreuses structures et arrangements structuraux restent non résolus. Ainsi, si la PDB constitue un formidable point d'appui pour toutes les études de modélisation, la limite principale de celle-ci est qu'elle n'offre qu'un regard limité quant à la « structure exacte » d'une protéine d'intérêt (la grande partie des protéines cristallisées représente des objets modifiés, des séquences partielles et/ou bien modifiées) et à la dynamique des protéines, élément essentiel qui gouverne la fonction des protéines et des systèmes macromoléculaires.

2.4. Analyse, représentation et visualisation des données structurales

En parallèle de la résolution d'un nombre accru de structures protéiques, des outils de visualisation ont été développés afin de permettre la représentation, l'inspection et l'analyse des objets biologiques d'intérêt. Les nombreux types de représentations graphiques permettent de visualiser les structures à différents niveaux, de le superposer pour permettre d'intégrer des analyses systématiques de manière simultanée. Les nombreux logiciels de visualisation (PyMOL (13), Maestro (14), MView (15), Chimera (16), VMD (Visual Molecular Dynamics) (17), Antheprot (18) etc…) jouent donc aujourd'hui un rôle important dans la biologie structurale. Ils permettent une première approche de la structure 3D d'une protéine, des structures secondaires et du positionnement des domaines constituant une protéine. Couplé à l'affichage des séquences et la mise en évidence des résidus clés (selon la littérature et les données cliniques ou biologiques), cette étape de familiarisation avec la protéine constitue généralement la première étape des projets de modélisations. Afin d'étudier ou de comparer le mode d'action de protéines apparentées ou le mode d'activation/d'inhibition de ligands, les logiciels graphiques modernes donnent la possibilité de comparer de multiples structures de protéines similaires ou les interactions des ligands liés à une même cible. Au contraire des fichiers de structures protéiques contenant généralement une information limitée (la séquence et les coordonnées atomiques), les outils de visualisation peuvent générer et

afficher des informations supplémentaires découlant de la structure (potentiels électrostatiques, charges etc…). Un exemple typique est la visualisation des surfaces moléculaires, qui peut être calculée quasi-instantanément avec les moyens informatiques actuels. La représentation des propriétés physico-chimiques sur la surface des molécules, telle que les partenaires la perçoivent in vivo, caractérise les systèmes biologiques et sont porteurs d'une information particulièrement pertinente. La superposition d'une même protéine liée à différents partenaires, ou de plusieurs protéines apparentées, permet ainsi d'apprécier rapidement les variations des sites de liaison lorsque leurs surfaces sont affichées.

La dynamique des complexes biologiques reste complexe à se représenter sans les visualiser. La visualisation des mouvements, générés à la volée grâce à des algorithmes dédiés (notamment via l'analyse des modes normaux) ou en fournissant un fichier comportant les données de dynamique moléculaire, est un élément important car, en plus de la représentation, il peut parfois orienter l'analyse de la dynamique et de ses caractéristiques, notamment dans le cadre de projets très exploratoires pour lesquels peu de données structurales initiales sont disponibles. Le développement de ces projets lié principalement à la partie logicielle de visualisation, va de pair avec une évolution du matériel informatique. L'apparition de nouvelles surfaces de visualisation virtuelle (CAVEs : cave automatical virtual environments, écrans 3D, casque de réalité virtuelle) autorise aujourd'hui la représentation en trois dimensions (3D) de structures (et non plus la projection d'un élément tridimensionnel sur une surface) et une immersion accrue, qui permet de formuler des concepts ou hypothèses grâce à la perception de détails subtiles difficilement détectables ou non-détectables par ailleurs (19,20). Ces environnements restent cependant encore peu diffusés à travers le monde, du fait de leur coût, si bien que leur utilisation ne concerne qu'une minorité des chercheurs. Des modèles structuraux physiques ont également fait leur apparition notamment grâce à l'impression 3D. Reliés à une interface adéquate, ils permettent la manipulation des objets virtuels de manière intuitive et facilitent la communication machine-chercheur ou chercheur-chercheur (21). Enfin, l'interactivité évolue également d'une façon progressive et la manipulation d'objets avec un retour de forces est maintenant possible, permettant d'explorer en temps réel différentes hypothèses comme pour l'amarrage moléculaire (22).

3. Modélisation de la structure tridimensionnelle des protéines

Le repliement des protéines dans l'espace a été l'objet de nombreuses hypothèses et études. Christian Anfinsen, par ses travaux sur la ribonucléase, a montré que la dénaturation de la conformation active d'une protéine, notamment via la rupture de quatre ponts disulfures, engendrait la perte d'activité enzymatique (23,24). Le lien entre la séquence primaire et le repliement spatial des protéines a ainsi été démontré. Quatre-vingt-un pourcents des chaines latérales des résidus non-polaires ont ainsi tendance à se tourner vers l'intérieur de la protéine pour former un cœur hydrophobe (cf. partie 1 de ce chapitre), tandis que cette orientation est préférée par 63% et 54% des chaînes latérales polaires et chargées, respectivement (25). D'autres facteurs ont néanmoins contribué dans le processus de repliement des protéines : l'action du solvant et des protéines chaperonnes, la présence de cofacteurs, le pH, un environnement membranaire. C'est donc aussi l'environnement cellulaire particulier (protéines associées au ribosome, densité en protéines, lipides etc.) qui est responsable du repliement à l'issue de la biosynthèse de la protéine (26), ou de l'absence de structures stables comme dans le cas des protéines ou régions de protéines intrinsèquement désordonnées (27).

Le développement des moyens bio-informatiques et l'accroissement du nombre de structures disponibles (stockées dans les bases de données structurales) ont permis la mise au point de différentes méthodes de prédiction de la structure tridimensionnelle. Plusieurs approches de prédiction coexistent, s'appuyant plus ou moins (voir pas du tout) sur les données structurales disponibles. Ces approches sont devenues d'autant plus pertinentes du fait de la concomitance de 2 phénomènes :

(1) D'une part, l'accroissement exponentiel de la capacité de calcul du matériel informatique au cours du temps conformément à la loi de Moore que nous pouvons simplifier par un doublement de la puissance informatique à coût constant tous les 2 ans (28). Bien que cette loi de Moore, énoncé une première fois en 1965, n'est plus tout à fait exacte aujourd'hui du fait des limites physiques liées à la miniaturisation des composants électroniques (effet tunnel) et à des considérations économiques, l'industrie des composants informatique en a pleinement profité pendant plus de 50ans. L'innovation logiciel et architecturel permet aussi de poursuivre la course à l'efficacité computationnelle (FLOPS/watt).

(2) D'autre part, le coût de la résolution expérimentale de la structure d'une protéine a tendance à augmenter car ces techniques sollicitent du matériel de plus en plus performant et

des équipes de plus en plus grandes. En effet, les structures d'intérêt les plus simples à résoudre ont eu tendance à être résolues en premier (les protéines globulaires) tandis que des protéines d'intérêt plus complexes à synthétiser constituent encore aujourd'hui un défi à relever. Nous pouvons prendre pour exemple la faible représentativité des protéines membranaires dans la PDB, de l'ordre de 1,5%, alors qu'elles représentent près de 25% du total des protéines (29,30).

Après des décennies de développement, de nombreuses méthodes de prédiction sont en cours de construction et/ou disponibles pour les utilisateurs du monde entier (31). Dans cette partie, nous allons donner une brève introduction à plusieurs méthodes de prédiction de structures plus ou moins largement utilisées. Elles peuvent être réparties en deux grandes familles : les prédictions directement issues de modèles structuraux donc s'inspirant de structures préalablement résolues, et celles issues de la modélisation *ab initio*, signifiant « à partir de zéro », basées sur des principes physiques plutôt que (directement) sur des structures existantes. Une dernière méthode sera aussi abordée dans le cas des complexes de deux ou plusieurs protéines, où les structures des protéines sont connues ou peuvent être prédites avec suffisamment de précision : il s'agit de l'amarrage protéine-protéine. Cette dernière méthode peut aussi bien servir à la prédiction de la structure quaternaire d'une protéine, nécessaire à son activité biologique, qu'à l'étude d'une partie de son interactome (32).

3.1. La prédiction de structures basée sur des modèles structuraux

Pour la plupart des protéines cibles, la matrice structurelle souhaitable peut être identifiée à partir de PDB par alignement de séquences ou méthode d'enfilage. Étant donné que les informations conformationnelles du modèle sont beaucoup plus fiables que celles provenant d'ailleurs (en particulier lorsque la protéine cible et le modèle sont hautement homologues), la précision de prédiction de la méthode basée sur le modèle est généralement plus élevée que les autres méthodes, ce qui la rend très populaire dans la pratique.

3.1.1. La modélisation par homologie

La modélisation par homologie (ou modélisation comparative) s'appuie sur les données structurales d'une protéine dont la séquence en acides aminés est proche de la protéine que nous cherchons à modéliser. Cette méthode repose sur l'hypothèse (validée par la majorité des observations) proposant que deux protéines ayant une bonne similitude de séquence partagent avec une grande probabilité une bonne similarité de structure tridimensionnelle. La différence de structure, représentée comme la déviation de la moyenne quadratique des distances interatomiques des atomes de la chaîne principale, a ainsi été corrélée à la fraction d'acides aminés mutés entre deux protéines homologues (33). L'utilisation de ce type de méthodes est liée à l'augmentation des données structurales disponibles (au sein de la PDB notamment, cf. Figure 6), qui ont permis la modélisation par homologie d'un nombre croissant de protéines, et l'amélioration de la précision des modèles générés. Ainsi, Zhang et Skolnick ont montré en 2003 que des modèles comparables aux structures expérimentales à basse résolution peuvent être générés à partir des données de la PDB (comprenant plus de 23 000 structures à cette date), pour des modèles de taille moyenne (moins de 200 résidus) (34).

L'étape initiatrice de la modélisation par homologie est la recherche d'un alignement de séquences homologues. La séquence (ou la fraction de séquence connue) de la protéine à modéliser (la cible) est confrontée aux bases de données structurales à la recherche de séquences proches (les supports) de la cible, en filtrant les résultats positifs en fonction de la disponibilité des structures associées et de la similarité des séquences. Cette étape, généralement effectuée par un algorithme tel que BLAST (Basic Local Alignment Search Tool) (35) ou FASTA (36,37), permet de sélectionner un sous-ensemble de séquences protéiques similaires à celle de la protéine cible et recouvrant tout ou partie de la séquence cible. Différents paramètres (pénalité pour un vide dans la séquence : gap, type de matrice de distance utilisée, etc.) influent sur la qualité de la sélection, et la sélection du ou des supports finaux se base sur les notions d'identité et de similarité des séquences, du recouvrement des séquences, etc. parfois évalués grâce à des scores associés (z-score, E-value). Les informations associées aux structures des supports (résolution de la structure, comparaison des structures secondaires observées pour la protéine support et prédites pour la protéine cible) sont également prises en compte. L'identité des séquences cible-support constitue cependant le critère essentiel : une identité de séquence inférieure à 20% produira des résultats discutables dans le meilleur des cas, alors qu'une identité de séquence supérieure à 50% génèrera des modèles de bonne qualité en règle générale (38). Pour les séquences à

très faible identité, la modélisation par homologie semble être moins adaptée que d'autres méthodes. Alors que plus l'identité de séquence augmente, moins le modèle généré par homologie contiendra d'erreurs. Ces erreurs se localiseront de plus en plus dans les parties très variables/flexibles de la protéine, comme les chaînes latérales et les boucles. La précision des modèles générés est également en partie conditionnée par la qualité de l'alignement entre séquence cible et séquence(s) support(s). L'alignement généré par logiciel est choisi, jugé et corrigé par le chercheur et dépend par conséquent de sa capacité à détecter l'alignement optimal à partir des données à sa disposition (structures secondaires prédites de la protéine cible, structures secondaires de la séquence support, conservation des résidus clés, etc.).

Une fois l'alignement obtenu, des modèles vont être générés à partir de cette information, générant un jeu de coordonnées pour chaque atome lourd (non-hydrogène). Plusieurs méthodologies peuvent être utilisées : l'élaboration de contraintes spatiales à partir de l'alignement, l'assemblage de fragments conservés, ou encore la recherche de courts segments correspondants à la séquence cible. La première méthodologie, l'élaboration de contraintes spatiales à partir de l'alignement, est la plus couramment utilisée dans la modélisation par homologie. Elle consiste à générer un jeu de critères géométriques sous forme de contraintes appliquées aux coordonnées internes de la protéine (distances inter-atomes, angles dièdres). Les contraintes sont ensuite optimisées de manière itérative, ce qui permet également de modéliser les régions désordonnées des protéines comme les boucles (39).

L'évaluation des modèles générés est réalisée par l'estimation d'une énergie, issue de potentiels statistiques ou du calcul des interactions physiques au sein des modèles : plus l'énergie du système est faible, meilleur est le modèle. Les potentiels statistiques s'appuient sur la fréquence d'occurrence des interactions intramoléculaires dérivées de la PDB (ou éventuellement d'une autre base de données), et peuvent ainsi produire des scores détaillés (parfois résidu par résidu) en plus d'un score global. Ces potentiels statistiques présentent le défaut d'être moins fiables pour les types de protéines peu représentées dans la base de données initiale. Un exemple typique est l'évaluation des structures de protéines membranaires qui étaient encore peu nombreuses quelques années auparavant. Les potentiels énergétiques, quant à eux, se basent à la fois sur les théories de la mécanique classique, où les atomes sont considérés comme des sphères interagissant les uns sur les autres, ainsi que sur l'hypothèse que la conformation native des protéines est celle présentant l'énergie potentielle totale la plus faible.

3.1.2. Modélisation de protéines par enfilage

En dehors de la modélisation par homologie, un autre type de méthode peut être employé : La modélisation d'une protéine par enfilage (threading en anglais) ou la modélisation par reconnaissance des repliements (40,41). Cette méthode ne nécessite pas la connaissance de la structure d'une protéine support homologue, mais s'appuie sur la structure des protéines partageant des structures secondaires et certaines propriétés physico-chimiques (exposition au solvant, hydrophobicité, etc…) similaires à celles de la protéine cible. La modélisation par reconnaissance de repliement se base sur le fait qu'il existe un nombre limité de repliements dans la nature (42,43), et que ceux de la plupart des protéines sont déjà représentés dans la PDB. Aussi une protéine ne possédant pas d'homologues de structure connue a de bonnes chances d'adopter un repliement déjà présenté dans les bases de données structurales. Un modèle de la protéine cible est généré (en commençant par la chaîne principale puis la chaîne latérale des résidus) en adoptant le repliement le plus plausible étant donné sa séquence, par enfilement successif des résidus. Parmi les logiciels utilisant ce type de méthode nous pouvons citer RaptorX (44), HHpred (45) ou encore I-TASSEUR (Iterative Threading ASSEmbly Refinement) (46).

3.2. La prédiction *ab initio* de structures de protéines

Enfin, certaines protéines n'ont pas encore de structures résolues suffisamment homologues pour permettre leur modélisation en utilisant l'une des méthodologies décrite précédemment. La modélisation de structures uniquement à partir de la séquence constitue alors la seule alternative. Elle porte le nom de modélisation *de novo*, modélisation *ab initio* ou encore modélisation libre. Le postulat d'Anfinsen, qui affirme que la structure d'une protéine est entièrement encodée dans sa séquence d'acides aminés (24,47), sous-tend l'hypothèse que pour une protéine relativement petite et de forme globulaire, la conformation stable de la protéine est proche du minimum global de sa fonction d'énergie. Ses acides aminés hydrophobes se trouvent au centre de la conformation alors que ceux hydrophiles sont exposés au solvant (l'eau) à la surface. Cependant, toutes les protéines ne correspondent pas à l'ensemble des critères d'Anfinsen comme entre autre les protéines membranaires. Pour autant, même en considérant le postulat d'Anfinsen comme correct dans la majorité des cas, la prédiction du « pliage » d'une protéine reste une tâche complexe car la fonction d'énergie utilisée est empirique et donc entachée d'erreurs et d'approximation. De plus l'exploration de

l'ensemble des conformations possibles d'une protéine possède une grande combinatoire qui augmente exponentiellement avec le nombre de résidus qui la compose. Concrètement, chaque acide aminé a deux degrés de liberté au niveau du squelette peptidique. Même si nous considérons seulement des rotations par pas de 10 degrés pour chaque degré de liberté, nous obtenons un nombre de configurations possibles qui grossit extrêmement vite en fonction du nombre d'acides aminés. C'est le paradoxe de Levinthal (48), explorer séquentiellement toutes les conformations possibles dans l'espoir de trouver celle de plus basse énergie est calculatoirement impossible alors que les protéines adoptent leur conformation native en quelques secondes tout au plus (49,50). Cela suggère que les protéines ne se plient pas de façon aléatoire mais suivent une « trajectoire » menant vers la conformation de plus basse énergie.

Le pliage de protéine *in-silico* consiste en un ensemble de simulations. Différentes approches ont été explorées afin d'y parvenir en des temps raisonnables. Certaines reposent sur la modélisation de fragments de petites tailles (3 à 20 résidus) qui sont ensuite étendus et/ou assemblés (51), tandis que d'autres utilisent des techniques de simulation de Monte Carlo qui explorent l'espace possible formé par les conformations de la protéine cible (52,53). C'est finalement une combinaison de ces deux approches qui est le plus souvent utilisée et qui semblait donner les meilleurs résultats (53).

Une autre approche qui consiste à décrire le processus entier du repliement d'une protéine repose sur la simulation en dynamique moléculaire de la séquence d'acides aminés (54). Le protocole de ces simulations fait souvent intervenir des méthodes de recuit simulé ou d'échange de réplica couplé à la dynamique moléculaire afin d'en accélérer l'échantillonnage conformationnel (55,56). L'usage de champ de force gros grains (GG) consistant à représenter des groupes d'atomes par une unique particule permet encore d'accélérer l'échantillonnage comme avec les champs de force UNRES (UNited RESidue) et Martini (57,58). Ces dernières simulations restent cependant rares pour le moment et ne sont efficaces que pour des protéines de petite taille, étant donné le temps de calcul nécessaire pour simuler un processus qui se déroule *in vivo* généralement sur plusieurs millisecondes et la limitation des champs de force mais cet aspect sera plus généralement abordé dans la partie 4 de ce chapitre.

L'état de l'art concernant la prédiction de structure protéique est évalué tous les deux ans depuis 1994 lors de l'expérience CASP (Critical Assessment of Structure Prediction). Cette expérience a pour objectif de tester les méthodes de prédiction de structure. Des

protéines, dont la structure vient d'être résolue mais pas encore publiée, sont proposées aux prédicteurs. Ceux-ci doivent tenter de prédire, selon la catégorie, la structure *ab initio*, la structure par homologie ou la structure secondaire d'une protéine. Pour comparer la performance des différents algorithmes sur la prédiction des structures 3D, nous utilisons le score GDT (Global Distance Test). Ce score calcul, selon un certain seuil de tolérance, le pourcentage d'acide aminé correctement positionné dans la structure prédite par rapport à la structure résolut expérimentalement. Selon le classement issu de la dernière édition de CASP (CASP14), nous retrouvons parmi les programmes librement accessibles les plus performants, I-TASSER et QUARK (59) lorsque qu'aucune structure homologue n'est disponible.

Cependant, un nouvel acteur a récemment fait son apparition lors du concours de CASP13 qu'est l'équipe DeepMind de Google à travers son algorithme Alphafold. Dès sa 1ère participation au concours (décembre 2018 pour CASP13), l'équipe de DeepMind avec l'algorithme AlphaFold a été le grand gagnant de cette édition en obtenant un GDT moyen de 61,4 pour les protéines résolubles uniquement par méthode *ab initio* tandis que le second groupe atteignait un score GDT moyen de 42,7 (12,59). Cet écart est conséquent sachant que depuis plus de 10 ans les meilleures équipes dépassaient rarement un GDT moyen de 35. A la dernière édition de CASP en décembre 2020 (CASP14) une nouvelle version d'AlphaFold a atteint un GDT moyen de 88 (60), s'approchant alors du seuil de 90 et pour lequel une structure peut être considérée comme résolut à une précision expérimentale. Cet algorithme se caractérise par l'usage intensif de réseau de neurones profond permettant de construire une matrice de distance entre les résidus à partir de la séquence d'acides aminés. Cela permet de construire des fragments de la protéine dont l'énergie globale est ensuite minimisée par des simulations de recuit simulé. Dans la 1ère version d'AlphaFold, il s'agissait de réseau de neurones convolutif, cependant il semblerait que la seconde version de l'algorithme fait intervenir des réseaux de neurones avec mécanisme d'attention. Malheureusement cet outil n'est pas encore librement accessible à la communauté scientifique. De plus, du fait du nombre et de la profondeur des réseaux de neurones que l'algorithme utilise, il est très probable que la puissance de calcul nécessaire pour accomplir une prédiction soit relativement conséquente.

3.3. L'amarrage protéine-protéine

L'ensemble des structures des protéines est loin d'avoir été résolu expérimentalement mais un nombre toujours croissant de structures continue d'abonder la PDB chaque année (cf.

partie 2 du chapitre). En parallèle et en complémentarité de la caractérisation expérimentale une recherche très active des méthodes de prédiction *in-silico* des structures protéiques est menée (cf. partie 3 du chapitre). Cependant ces dernières concernent essentiellement la prédiction des structures tertiaires des protéines (mise à part quelques exceptions s'agissant de la construction par homologie). Pourtant les protéines sont des objets biologiques qui fonctionnement très souvent en association avec d'autre protéines d'où l'importance de la connaissance de la structure quaternaire pour nombreuses d'entre elles. C'est pour répondre à cette problématique que les méthodes d'amarrage protéine-protéine se sont développées et dont le but est de prédire la structure d'un complexe à partir des structures ou modèles des partenaires isolés (cf. figure 8). Les méthodes d'amarrages sont d'autant plus complémentaires avec les méthodes expérimentales, qu'il est généralement plus difficile de déterminer la structure d'un complexe que celle d'une protéine individuelle.

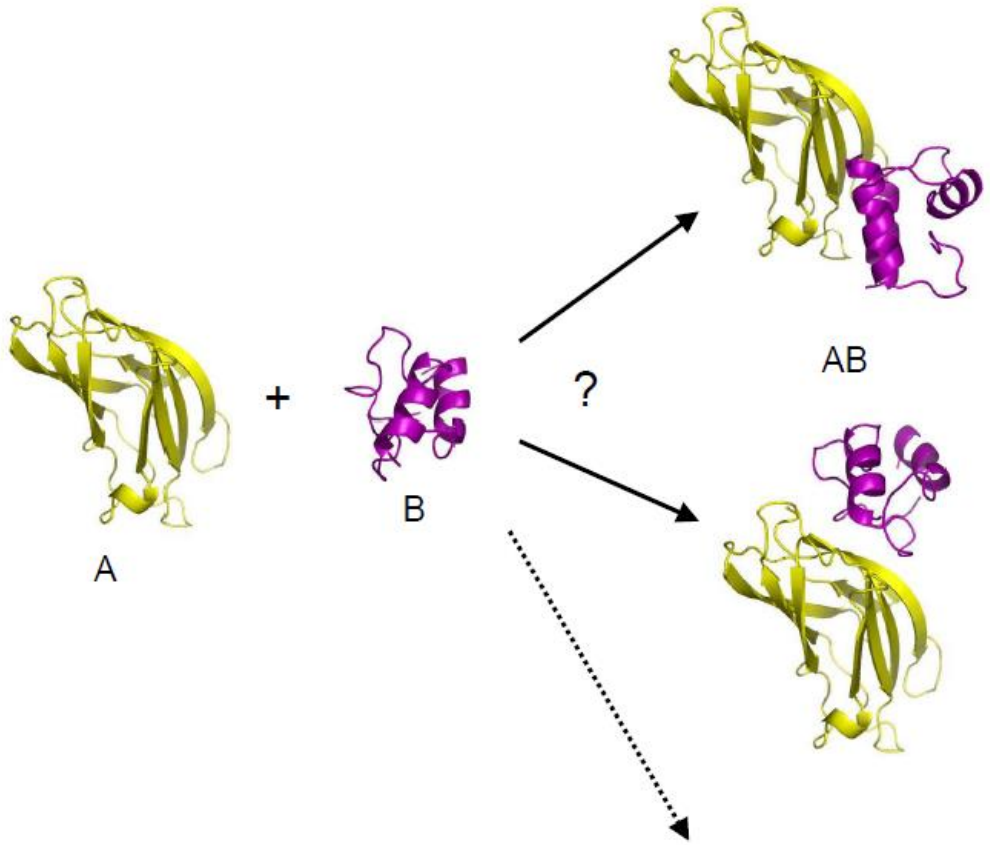


Figure 8 : Illustration du problème de l'amarrage. Comment associer la protéine A et la protéine B ? Plusieurs configurations AB sont obtenues, laquelle est représentative ?

Le problème se divise en deux étapes : d'abord, l'espace est exploré pour obtenir toutes les conformations possibles, et ensuite, ces conformations sont triées dans le but de

classer en premier la conformation la plus proche de la forme native. Avec une approximation de corps rigides, si nous considérons chaque partenaire comme une sphère de 15 Å de rayon à la surface de laquelle les propriétés atomiques sont décrites sur une grille de 1 Å , une recherche systématique présente 10^9 modes distincts d'association (61). La question est ensuite de déterminer, parmi ces modes d'association, lequel est le mode natif.

Pour pouvoir accéder aux changements de conformation et aux mouvements des chaînes latérales et des bases, le modèle doit être de type "soft". Cela signifie que les molécules doivent pouvoir légèrement s'interpénétrer. De plus, les molécules doivent être considérées comme des ensembles de sphères articulées. Ainsi, il est possible de traiter aussi bien les molécules issues de résolution de structures de protéines seules (non-liées) ou complexées (liées).

La première étape qui consiste à générer un ensemble de poses exhaustif repose sur la notion de complémentarité de surface. De nombreux algorithmes ont été développé pour la résolution ce problème (62–66) mais il semblerait que ceux utilisant la transformation de Fourier rapide (en anglais Fast Fourier Transform (FFT)) soient les plus performants et sont maintenant les plus largement utilisés (67–70). Avec les méthodes utilisant la FFT une grille cubique est tracée. A chaque point, il est attribué un poids qui est négatif et important si le point est situé à l'intérieur de la protéine A, nul s'il est à l'extérieur et égale à 1 s'il est proche de la surface ; il est fait de même pour la protéine B. Le produit sera donc important et positif (donc défavorable) si les deux volumes moléculaires s'interpénètrent, et il sera négatif (donc favorable) pour les points qui appartiennent à la surface d'une molécule et au volume de l'autre. Lorsque la molécule A est translatée par rapport à la molécule B, le score peut être rapidement calculé par transformation de Fourier rapide (FFT), si la grille de A est identique à la grille de B. La grille doit donc être redéfinie à chaque nouvelle orientation pour que la recherche soit complète. Cette approche présente de nombreux avantages : les poids peuvent contenir des informations sur les propriétés physico-chimiques de la surface, et la résolution peut être ajustée tout en limitant le nombre de termes de Fourier calculés dans la somme. Avec les différentes méthodes de complémentarité de forme, la problématique de la recherche exhaustive des poses peut être considérée comme raisonnablement résolue. Reste à déterminer parmi les milliers de poses laquelle est la bonne.

La deuxième étape du processus d'amarrage qui consiste à trier les poses obtenues par une fonction de score, reste le point le plus critique à améliorer. Si beaucoup de fonctions de score reposent principalement sur l'usage de fonctions énergétiques (71–75), d'autre

reposent aussi sur les propriétés physico-chimiques des atomes héritées des méthodes d'amarrage des petites molécules. Bien que ces dernières aient été adaptées à l'amarrage protéine-protéine, leurs performances demeurent insuffisantes (76–78). Une approche qui semble donner de meilleurs résultats qu'avec des fonctions de score énergétiques, utilise le principe du regroupement (clustering) (79). Cette méthode consiste à trier les poses en les regroupant par famille/grappe. Une grappe représente un ensemble de poses relativement proches les unes des autres. Dans chaque grappe il n'est retenu qu'une seule pose, celle ayant le plus grand nombre de poses voisines. Les meilleures poses seront celles issus des grappes les plus nombreuses.

Ces deux dernières décennies, stimulées par l'expérience d'amarrage CAPRI (Critical Assessment of PRediction of Interactions), plusieurs méthodes ont vu le jour. Comme pour CASP et la prédiction de structure tertiaire, CAPRI est un test à l'aveugle des algorithmes d'amarrage de macromolécules (80). CAPRI existe depuis 2001 et permet de tenir une synthèse de l'état de l'art dans le domaine de l'amarrage protéine-protéine. Dans cette expérience, chaque équipe doit prédire le mode d'association de deux protéines à partir de leur structure tridimensionnelle. La structure du complexe, résolue expérimentalement, n'est dévoilée aux participants et publiée qu'à l'issue des soumissions. A ce jour l'algorithme automatique donnant les meilleurs résultats pour l'amarrage protéine-protéine est celui de Cluspro utilisant la méthode de triage dite par regroupement (ou par grappe) (81).

4. Etude de la dynamique des systèmes biologiques

4.1. Principes généraux

Tous les processus biologiques reposent sur les propriétés dynamiques des composants cellulaires et les changements structuraux/conformationnels liés à l'exercice d'une fonction biologique donnée prennent place dans des temps variables (cf. Figure 9). L'étude et la compréhension de ces processus nécessitent donc de prendre en compte cette dimension temporelle, sans se limiter à une analyse tridimensionnelle statique. L'augmentation des moyens de calcul (supercalculateurs de très haute performance) a ainsi permis à l'étude de la dynamique moléculaire de se développer par diverses méthodes au cours des années 70 et 80 avant de constituer un champ d'étude bien spécifique, ne se limitant pas à la biologie. La science des matériaux a également beaucoup profité de ces nouvelles méthodes

computationnelles. La dynamique moléculaire permet l'étude de processus prenant place dans des échelles de temps allant de la femtoseconde (fs) à la milliseconde (ms), pour des systèmes biologiques de petite taille tel que les peptides (82)) comme pour des systèmes très complexes et de grande taille tel que des complexes membranaires multiprotéiques en présence de multiples ligands (83) ou de capside virale (84).

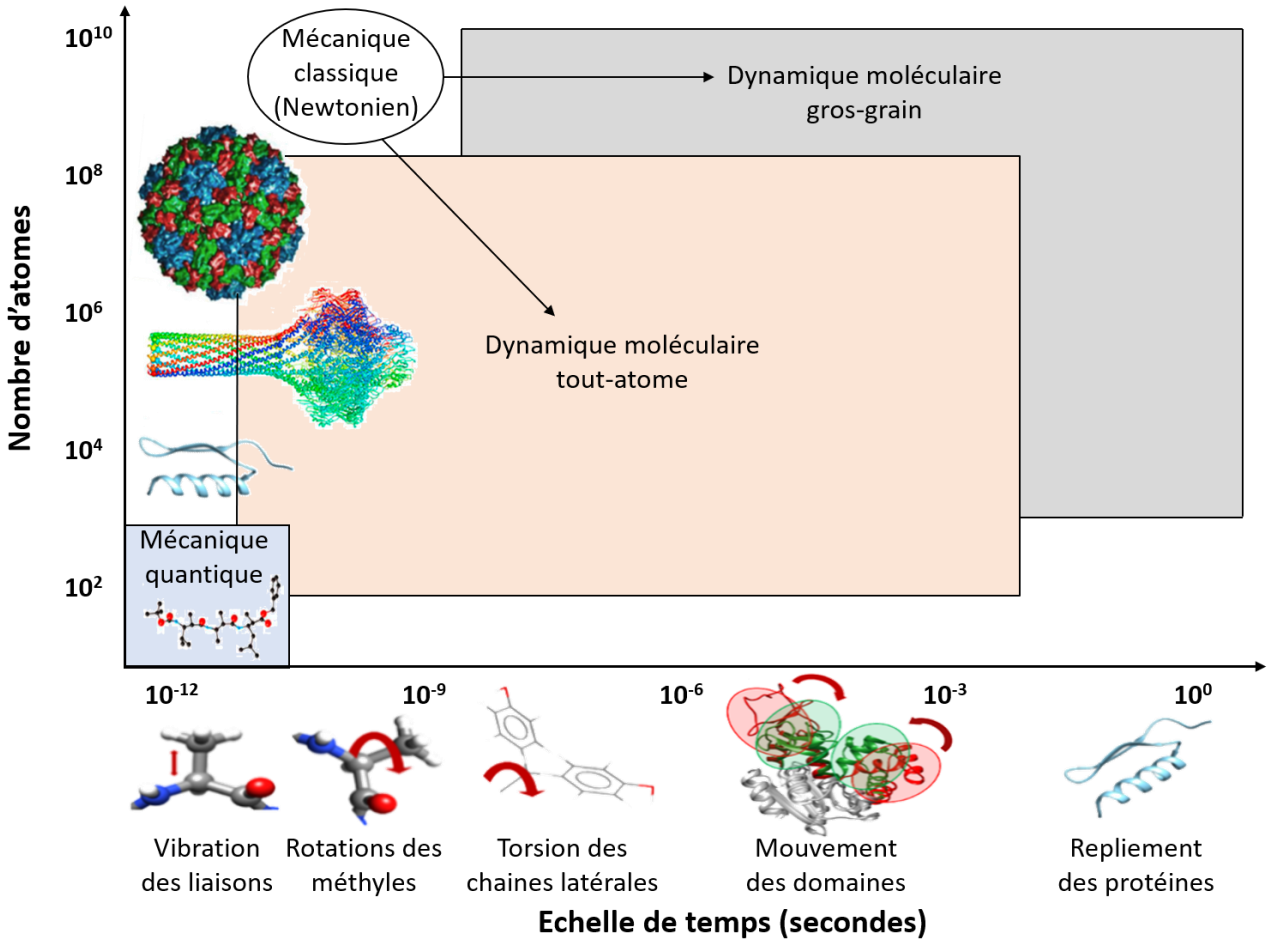


Figure 9 : Echelle de temps explorable en simulation en fonction de la technique utilisé et par rapport à la taille du système étudié en nombre d'atomes. Les différentes échelles de temps permettent d'explorer différentes amplitudes de mouvements.

Différents paramètres sont à prendre en compte lorsque nous souhaitons effectuer une simulation de dynamique moléculaire d'une macromolécule. Le processus étudié, la taille du système ainsi que les ressources en temps de calcul disponibles qui vont déterminer la durée des simulations ainsi que de la méthodologie à utiliser. Trois principales méthodologies existent : la mécanique quantique (QM), la mécanique moléculaire dite classique (MM) et les

simulations gros grains (GG). Ces 3 méthodes peuvent être dérivés en méthodes hybrides tel que la QM/MM et la MM/GG employées en fonction des besoins. Toutes ces méthodes reposent sur un principe commun : les mouvements atomiques sont dirigés par les forces qui s'exercent sur les atomes. Ainsi, un champ de forces décrit les interactions que chaque atome établit avec son environnement, et permet de décrire la physique du système à chaque instant de la simulation. Ces champs de forces ont principalement 2 origines :
(1) les simulations de dynamique moléculaire dite *ab initio* ou quantique utilisant un potentiel énergétique en se basant sur les principes de la mécanique quantique, prenant en compte la structure électronique des atomes du système.
(2) Les simulations de dynamique moléculaire dites « classiques » utilisant un champ de forces « empirique », dont les paramètres sont dérivés des bases de données structurales (PDB) ou de simulations elles-mêmes issues de la mécanique quantique.

Les méthodes classiques présentent comme avantage d'être bien moins coûteuses en temps de calcul que les méthodes quantiques. En contre parti, elles ont le désavantage de ne pouvoir représenter que des changements de topologie (pas de formation/rupture de liaison covalente) ainsi que d'être bien moins précises car ne reposant pas sur la structure électronique des atomes. Le choix de la méthode à utiliser dépendra ainsi des caractéristiques du système que nous cherchons à étudier ainsi que de l'échelle de temps nécessaire à son observation. La modélisation de réactions chimiques ne peut pas être réalisée par des simulations de dynamique dite « classique ». En revanche, pour l'exploration des phénomènes longs (supérieur à la nanoseconde) sur des systèmes comportant plusieurs milliers d'atomes il sera impossible d'utiliser des simulations de dynamique moléculaire par des méthodes de mécaniques quantiques car elles sont trop coûteuses en temps de calcul. Chacune des approches impliquent donc de faire des compromis sur le champ d'étude que nous décidons d'investir.

Afin de répondre aux limitations explicitées ci-dessus (temps de simulations accessibles, coût des calculs, observation de réactions chimiques), de nouvelles approches ont été introduites. Les deux approches les plus notables sont les simulations hybrides QM/MM et les simulations gros grains. Dans la première, une partie du système est traité une méthode de mécanique quantique (QM) puis le reste l'est par méthode de mécanique moléculaire classique (MM). En limitant la description quantique à une petite partie du système, le temps de calcul est réduit considérablement, bien qu'il reste supérieur à une simulation en mécanique classique. En revanche cette méthode hybride nous permet de

profiter des avantages offerts par la mécanique quantique au niveau de la région d'intérêt (tel qu'un site de liaison ou un site actif par exemple). La deuxième approche, que sont les simulations en gros grains, est à l'inverse une approche représentant des groupes d'atomes par une unique particule, un gros grain. Nous perdons donc à nouveau la possibilité de simuler des réactions chimiques en plus de perdre en précision par rapport à la mécanique moléculaire classique. La taille du système (comprendre le nombre de particules constituant le système) étant réduit, cela rend possible des simulations de très larges systèmes pendant sur des échelles de temps qui sont inaccessibles avec les autres approches mais cela se fait au détriment de la description tout-atome des systèmes étudiés. Enfin, pour augmenter les échelles de temps explorable lors d'une simulation, il est possible d'optimiser l'interaction entre le code du logiciel de simulation de dynamique moléculaire et l'architecture du matériel informatique. L'émergence de la technologie CUDA (Compute Unified Device Architecture) en 2007 et son intégration dans les logiciels de simulation a permis l'usage intensif des GPU pour ces dernières offrant ainsi des gains de performance substantiels pour les calculs parallèles (85-87). Du côté des cluster de calcul la machine Anton, dont l'architecture a été spécifiquement conçus pour des simulations de dynamiques moléculaire, peut générer des simulations de l'ordre de la milliseconde (88) en un temps d'exécution inférieur à celui d'un cluster de calcul traditionnel (89).

4.2. Dynamique moléculaire en mécanique moléculaire classique

4.2.1. Les équations du mouvement

La représentation de la mécanique classique en dynamique moléculaire est la forme la plus courante dans le champ de la biologie computationnelle. En effet, la plupart des processus étudiés ne comportent pas d'évènements impliquant un changement de la connectivité des atomes (réactions enzymatiques). Une description à l'échelle atomique est néanmoins nécessaire à l'analyse fine des modifications structurales et dynamiques. La mécanique moléculaire classique est donc le standard en dynamique moléculaire lorsqu'on étudie les protéines. Les simulations de dynamique moléculaire doivent résoudre, pour l'ensemble des N atomes du système, les équations du mouvement de Newton :

$$F_i = m_i \frac{\partial^2 r_i}{\partial t^2}, i = 1 \ldots N$$ Equation 1

où F_i sont les forces s'exerçant sur l'atome i, m_i la masse de l'atome i, $r_i = (x_i, y_i, z_i)$ les coordonnées de l'atome i et t le temps.

Les forces sur les atomes à un instant t sont calculées à partir de la dérivée du potentiel d'énergie potentielle $V(r_1, r_2, \ldots, r_N)$:

$$F_i(t) = -\frac{\partial V(r_1(t), r_2(t), \ldots, r_N(t))}{\partial r_i(t)}$$ Equation 2

Ces équations sont résolues simultanément de manière itérative pour un pas de temps δt, appelé par conséquent « pas d'intégration ». Chaque itération génère un nouveau jeu de coordonnées à l'ensemble des atomes qui, mises bout à bout, constituent une trajectoire de dynamique moléculaire. Les systèmes tendent à s'équilibrer, lorsqu'un système atteint cet état, la simulation est dite « à l'équilibre ».

Le pas d'intégration représente l'intervalle de temps séparant deux évaluations successives de la fonction d'énergie. Sa valeur doit être suffisamment petite pour ne pas discrétiser certaines quantités mesurables. Dans les faits, cela consiste à choisir un pas d'intégration inférieur au mouvement de la plus haute fréquence du système (théorème de Nyquist-Shannon (90)). La fréquence d'un mouvement étant liée à la masse des particules impliquées dans ce mouvement vibratoire, la plus haute fréquence des systèmes biologiques est liée à l'élongation des liaisons covalentes impliquant des atomes d'hydrogène (liaisons C–H, O–H ou N–H). La fréquence vibratoire de ces mouvements est supérieure à 3000 cm^{-1} à 310K, d'où la nécessité d'utiliser un pas d'intégration de l'ordre de la femtoseconde (1 fs = 10^{-15} s).

L'intégration des pas de temps peut se faire par plusieurs méthodes différentes, nous présenterons ici l'algorithme dit « leapfrog », ou « saut de grenouille » (91). Cet algorithme est une méthode de résolution numérique de l'équation différentielle (équation 1), en utilisant les positions atomiques r au temps t et les vitesses atomiques v au temps $-\frac{1}{2}\delta t$. Les positions des atomes ainsi que leurs vecteurs vitesses sont alors calculées grâce aux équations suivantes :

$$r(t+\delta t) = r(t) + \delta t \cdot v(t+\frac{1}{2}\delta t)$$ Equation 3

$$v\left(t+\frac{1}{2}\delta t\right) = v\left(t-\frac{1}{2}\delta t\right) + \frac{\delta t}{m}F(t)$$ Equation 4

Cet algorithme calcule les vitesses des atomes de manière explicite (à l'inverse de l'algorithme de Verlet (92) par exemple), bien qu'à l'instant $t + \delta t$. Les vitesses à l'instant t peuvent être approximées grâce à la relation :

$$v(t) = \frac{1}{2}\left[v\left(t-\frac{1}{2}\delta t\right) + v\left(t+\frac{1}{2}\delta t\right)\right]$$ Equation 5

Afin de démarrer une dynamique moléculaire, un jeu de vitesses initiales doit être fourni afin d'amorcer la dynamique du système. Si aucune donnée ne permet d'obtenir les vitesses initiales, un jeu de vitesses à $t = t_0 - \frac{1}{2}\delta t$ sera généré automatiquement en suivant une distribution aléatoire de Boltzmann à une température donnée :

$$p(v_i) = \sqrt{\frac{m_i}{2\pi k_B T}}\, exp\left(-\frac{m_i v_i^2}{2k_B T}\right)$$ Equation 6

où m_i et v_i la masse et la vitesse de l'atome i, k_B la constante de Boltzmann, k_B = 8.314510.10-3 kj.mol^{-1}.K^{-1} et T la température.

L'énergie cinétique du système E_c est obtenue à partir des vitesses calculées grâce à l'équation 6 et doit respecter l'égalité suivante :

$$E_c = \sum_{i=1}^{N} \frac{1}{2} m_i v_i^2 = \frac{1}{2} N_{DL} k_B T$$ Equation 7

où $N_{DL} = 3N - N_c - N_{com}$, N est le nombre d'atomes du système, N_c est le nombre de contraintes appliquées au système et N_{com} = 3 ou 6 en fonction des mouvements de translation et/ou de rotation du système qui sont retirés du mouvement. Une fois les vitesses initiales connues, l'intégration des formules des équations 3 et 4 est possible.

4.2.2. Champ de forces

La fonction d'énergie utilisée dans l'équation 2 est décrite par le champ de force, qui comprend des termes représentant à la fois les interactions liantes (les liaisons covalentes) et les interactions non liantes (interactions électrostatiques et de van der Waals). Voici les différents composants d'un champ de forces :

$$E_{totale} = E_{liantes} + E_{non-liantes} \qquad \text{Equation 8}$$

Avec $\begin{cases} E_{liantes} = E_{élongation} + E_{angle} + E_{dièdre} \\ E_{non-liantes} = E_{électrostatique} + E_{van\ der\ Waals} \end{cases}$ Equation 9

L'énergie d'élongation correspond à la variation de la longueur des liaisons covalentes, et est représentée par une fonction de potentiel harmonique générique de type :

$$E_{élongation} = \sum k(r_{ij} - r_{ij0})^2 \qquad \text{Equation 10}$$

où k est la constante de force de liaison, r_{ij} est la longueur instantanée de la liaison entre les atomes i et j, et r_{ij0} est la longueur de la liaison de référence. Les variations de l'angle formé par trois atomes sont également représentées par un potentiel harmonique :

$$E_{angle} = \sum k\,(\theta_{ijk} - \theta_{ijk0})^2 \qquad \text{Equation 11}$$

avec k la constante de force angulaire, θ_{ijk} l'angle formé par les atomes i, j et k à l'instant t, et θ_{ijk0} l'angle i-j-k de référence. Les angles dièdres correspondent à la rotation de deux groupements autour d'une liaison et implique donc quatre atomes, i,j,k et l. L'angle dièdre Φ autour de la liaison j-k est l'angle formé par les deux plans i-j-k et j-k-l. L'énergie des angles dièdres est donc décrite par la fonction générique de l'équation 12 :

$$E_{dièdre} = \sum \sum_i k_\phi(1 + \cos(n\phi - \phi_s)) \qquad \text{Equation 12}$$

où k_Φ est la constante de torsion, n est la périodicité de la rotation et Φ_s est l'angle de phase.

Les interactions non-liantes sont également représentées par les fonctions suivantes. Pour chaque paire d'atomes i et j séparés par une distance r, les interactions de van der Waals sont décrites par le potentiel de Lennard-Jones :

$$E_{van\ der\ Waals}(r_{ij}) = 4\varepsilon^0 \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] \qquad \text{Equation 13}$$

où r^0_{ij} est la distance pour laquelle les forces attractives et répulsives s'annulent et ε^0 l'opposé de la valeur du potentiel au point r^0_{ij} . Enfin les interactions électrostatiques entre les charges électriques q_i et q_j des atomes i et j, respectivement, sont données par :

$$E_{électrostatique} = \sum \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{\varepsilon_r r_{ij}} \qquad \text{Equation 14}$$

où ε_0 est la permissivité du vide et ε_r la permissivité du milieu.

Cette forme générique est partagée par de nombreux champ de forces, mais d'autres termes peuvent être ajoutés afin de mieux décrire les propriétés physiques du système, comme notamment les angles dièdres impropres afin de maintenir certains groupes (les cycles aromatiques par exemple) planaires. Les paramètres des champs de forces ont eux aussi deux origines possibles : pour les molécules disposant de larges données (typiquement les acides aminés et les acides nucléiques), les paramètres vont être extraits des bases de données afin de générer des champs de forces dits « empiriques ». Dans le cas contraire (généralement les ligands), les paramètres vont être dérivés de calculs de mécanique quantique.

Le calcul des forces non-liantes, électrostatiques (équation 14) et de van der Waals (équation 13), se base sur la notion de pair d'atomes. L'inconvénient de cette méthode est que le nombre d'interactions augmente exponentiellement avec le nombre d'atome car elle varie en N^2. Pour limiter l'impact du calcul de ces forces, une limite de distances est appliquée et seuls les atomes séparés par une distance inférieure à cette limite seront pris en compte pour le calcul des interactions non-liantes. Cette valeur seuil doit être inférieure à la moitié de la plus petite longueur de la boîte de simulation pour éviter l'interaction de particules avec l'une de leur image périodique. Des approches permettent ensuite de limiter l'effet de la troncation, qui implique dans la forme décrite ci-dessus une discontinuité dans le champ de forces : le potentiel peut être diminué pour atteindre la valeur de zéro au niveau de la valeur seuil (« shift

»), ou un modificateur est appliqué au potentiel à partir d'une distance pour atteindre la valeur de zéro au niveau du seuil (« switch ») (cf. figure 10).

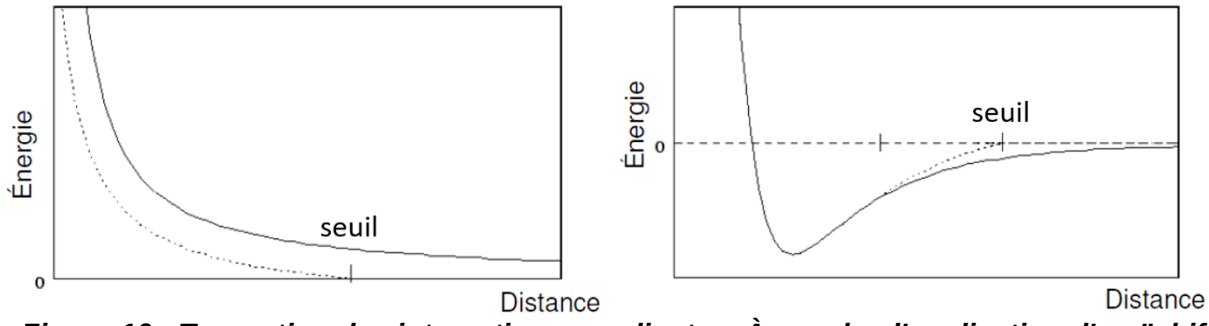


Figure 10 : Troncation des interactions non-liantes. À gauche, l'application d'un "shift" change le potentiel, qui est nul pour la valeur seuil. À droite, le potentiel est décalé à partir d'une limite pour devenir nul à la valeur seuil. Figures issues du site du logiciel NAMD (*NAnoscale Molecular Dynamics) *: http://www.ks.uiuc.edu/Research/namd/2.9/ug/node23.html

Enfin, pour les interactions longue distance, c'est l'algorithme Particle Mesh Ewald (PME) qui est généralement utilisé. Cet algorithme a été appliqué pour la première fois au calcul du potentiel électrostatique en 1993 (93). La sommation de l'énergie électrostatique d'un système et de ses images périodiques est lente à converger. Plus particulièrement, les interactions longue-distance sont les plus lentes à converger ; la sommation d'Ewald propose de changer l'espace de calcul de ces interactions de l'espace réel vers un espace de Fourier, dont la convergence sera plus rapide (94). L'approche PME améliore les performances de la sommation d'Ewald en utilisant une grille et une transformation de Fourier rapide pour le calcul des interactions longue-distance. La complexité du calcul varie selon N•log(N) au lieu de N^2 pour la sommation d'Ewald, d'où un gain de temps considérable.

Le contrôle de la température est un élément important puisque la température d'un système est corrélée à la somme des énergies cinétiques particules qui la composent et donc à la quantité de mouvements observée (cf. équation 15). De même, la pression est reliée au carré de la vitesse des particules via la relation 15 :

$$P = \frac{1}{3}\frac{N}{V}mv^2 \qquad \text{Equation 15}$$

Le maintien de ces deux variables d'état (température et pression) à des niveaux constants au cours de la simulation doivent correspondre aux conditions physiologiques des protéines pour considérer la simulation comme réaliste. Combiné à un nombre constant de

particules, nous obtenons un ensemble isotherme-isobare, NPT (Nombre de particules, Température et Pression constantes). L'utilisation de l'ensemble micro-canonique (NVE, Nombre de particules, Volume et Énergie constants) ou canonique (NVT, Nombre de particules, Volume et Température constants) peut également être envisagé ; cependant les erreurs liées aux approximations (résolution des équations du mouvement, etc.) engendrent des variations de température et de pression au cours du temps. Afin d'éviter ces dérives, la température et à la pression sont couplés à des bains (pool) pour maintenir ces valeurs constantes.

4.3. Les limites de la mécanique moléculaire classique

Les simulations de dynamique moléculaire par mécanique moléculaire classique reposent sur des théories physiques bien définies mais qui présentent néanmoins certaines limites. La confrontation des données obtenues est un moyen de contrôle quant à la validité et la précision des simulations réalisées. Cependant, ces données ne sont pas forcément disponibles, et surtout elles ne sont pas établies aux mêmes échelles de grandeur de temps (heures versus nanosecondes) et de taille (échelle macroscopique versus atomistique).

Les trajectoires de dynamique moléculaire sont toujours analysées sous l'hypothèse ergodique, qui affirme qu'à l'équilibre, la valeur moyenne d'une grandeur calculée de manière statistique est égale à la moyenne d'un très grand nombre de mesures prises dans le temps. En statistique mécanique, cette hypothèse permet de s'affranchir partiellement de la nécessité d'explorer l'ensemble de l'espace conformationnel (l'ensemble des conformations que peut adopter une molécule), si un nombre suffisant de conformations représentatives est généré. Dans la pratique, parcourir l'ensemble de cet espace conformationnel par simulations de dynamique moléculaire n'est pas réalisable, car il est immense. Générer plusieurs trajectoires permet de calculer les valeurs d'intérêt en générant d'autres conformations représentatives de l'ensemble en utilisant des vitesses initiales différentes, les sous-espaces conformationnels explorés dans chaque dynamique diffèrent alors partiellement. Certaines approches computationnelles permettent de générer un plus grand nombre de conformations en orientant/contraignant les déplacements du système.

Parmi celles-ci, la méta-dynamique est utilisée lorsqu'un obstacle énergétique important empêche l'exploration du paysage conformationnel (95). Un potentiel est ajouté au paysage énergétique afin de limiter le retour du système aux points déjà explorés : les puits

énergétiques vont ainsi être comblés progressivement et la barrière énergétique va devenir accessible. Les simulations de dynamique dirigée, à partir d'une structure de départ et d'une structure cible, vont guider linéairement l'évolution du système de l'un à l'autre en appliquant un potentiel harmonique aux atomes (96). Plutôt que de se servir d'une structure cible résolue, il pourra s'agir d'un nuage électronique issu d'une cryo-microscopie (97). Cette technique porte le nom de fitting flexible en dynamique moléculaire (MDFF). Si cette approche permet de franchir rapidement des obstacles énergétiques, elle peut également déformer la structure lors de réarrangements majeurs et ne suit pas nécessairement le parcours de plus faible énergie. Les simulations de dynamiques moléculaires guidées appliquent également des contraintes, mais elles l'appliquent au centre de masse du système et non à des atomes. Ainsi, les déformations de structures sont moindres, mais le système ne suit pas forcément les variations du paysage énergétique. D'autres méthodes ont été développées afin de répondre à ces limites, mais ce champ d'étude reste à ce jour très ouvert (98,99).

L'utilisation de la mécanique moléculaire classique pour décrire le système et intégrer les équations de Newton est suffisamment précise pour la plupart des atomes à température ambiante, mais peut présenter certaines limitations dans le cas de transfert des protons, dont la dynamique est essentiellement dominée par des effets quantiques. Les liaisons hydrogènes sont un exemple de ces effets à prédominance quantique. En pratique, tous les mouvements vibratoires rapides (dont la fréquence est supérieure à 200 cm^{-1}) peuvent potentiellement se comporter d'une manière inappropriée. Appliquer des contraintes sur ce type de mouvements (notamment les mouvements vibratoires impliquant des atomes d'hydrogène) va limiter ces effets, en plus de permettre l'augmentation du pas d'intégration, donc augmenter le temps de simulation pour une quantité de calcul constant. D'autres types d'interactions peuvent être biaisés, il s'agit des interactions non-liantes. Comme décrit dans le paragraphe précédent, ce type de forces est tronqué à partir d'une distance seuil. De plus, les champs de force ne modélisent pas les effets de polarisabilité qui prennent place dans les molécules. Dans la pratique, ces effets ne sont pas (trop) mal traités par les champs de forces, ce qui explique que nous les utilisons toujours.

Les paramètres des champs de forces sont principalement issus des données structurales, par nature fortement hétérogènes et d'une précision variable. Les conditions expérimentales (température notamment) qui ont permis de générer ces données peuvent fortement fluctuer et diffèrent souvent des conditions de simulation. Et ceci notamment pour certaines simulations qui introduisent des amples variations de température (recuit simulé ou

échange de réplica par exemple). Les paramètres du champ de forces peuvent alors se révéler de faible qualité et amener des distorsions de la structure atomique des molécules simulées. Ainsi, si l'utilisation de champ de forces empiriques démontre une robustesse croissante, ces outils restent perfectibles et sont enrichis ou corrigés régulièrement.

Parmi les champs de force les plus utilisés, pour les simulations de macromolécule tel que les protéines et les acides nucléiques, nous pouvons citer les 3 principaux que sont CHARMM (Chemistry at HARvard Macromolecular Mechanics), AMBER (Assisted Model Building with Energy Refinement) et GROMOS (GROningen MOlecular Simulation) (100). Chacun de ces champs de force sont associés à un logiciel permettant d'effectuer les calculs pour les simulations et portant le même nom que le champ de force associé (à ceci près que le champ de force GROMOS est associé au logiciel GROMACS pour GROningen MAchine for Chemical Simulations). Bien que chaque logiciel soit particulièrement optimisé pour le champ de force qui lui est associé, il est possible d'utiliser en pratique chaque champ de force sur chacun des programmes. Le choix de l'utilisation d'un logiciel de simulation ainsi que du champ de force sera déterminé par le système que nous cherchons à étudier, les performances du logiciel et les outils qu'il offre à la disposition de l'utilisateur. Au cours de cette thèse nous avons principalement utilisé le programme AMBER avec son champ de force éponyme car il est parfaitement adapté pour l'étude des systèmes protéiques et que le logiciel offre parmi les meilleures performances de simulation grâce à son accélération des calculs par l'usage de GPU (85,101). Ainsi, des simulations de large complexe et sur des échelles de temps convenable peuvent être accompli par du matériel facilement accessible et pour un coût très raisonnable comparé à l'usage de cluster CPU.

5. Prédiction *in-silico* de d'affinité de liaison protéine-protéine

La prédiction *in-silico* de l'affinité de liaison protéine-protéine est à la croisée des chemins entre la modélisation des structures et la dynamique des complexes protéiques. Elle constitue l'un des défis majeurs de la modélisation moléculaire. L'interaction des protéines est d'une importance fondamentale pour pratiquement tous les processus dans les systèmes vivants. La plupart des fonctions dans une cellule sont médiées par l'assemblage de protéines pour former des dimères ou oligomères transitoires afin d'agir comme des enzymes, des

transporteurs, ou pour stabiliser la forme de la cellule (102,103). De nombreuses interactions entre les protéines d'une cellule sont en principe possibles mais seule une fraction des complexes et assemblages putatifs est en effet formée et possède une pertinence fonctionnelle (104). L'énergie libre d'affinité protéine-protéine détermine la stabilité de l'association et les conditions de formation et dissociation du complexe. Par conséquent, une compréhension complète des processus cellulaires nécessite non seulement une connaissance de toutes les interactions protéines-protéines possibles, mais également un aperçu quantitatif de la structure et de la stabilité des complexes formés (103,105). Ces dernières années, la possibilité de concevoir de nouveaux complexes synthétiques protéine-protéine avec une fonction souhaitée a pris un essor significatif (106,107). L'objectif étant souvent de modifier les protéines naturelles existantes de telle manière que la géométrie et l'affinité d'une interaction protéine-protéine connue soit modifié pour favoriser ou non sa formation. C'est un outil essentiel de la conception de protéine. La prédiction informatique de l'affinité de liaison protéine-protéine nécessite généralement la structure tridimensionnelle du complexe ou au moins un modèle de la structure du complexe. Dans le cas où la structure du complexe est inconnue la prédiction peut servir à déterminer quelles sont les poses pertinentes issu des simulations d'amarrage protéine-protéine. La prédiction informatique de l'affinité de liaison protéine-protéine peut donc aussi servir à la modélisation de la structure tridimensionnelle des complexes non résolue.

L'interaction entre un récepteur et son ligand est un phénomène complexe pouvant mettre en jeu de nombreux effets : un changement dans l'entropie conformationnelle, translationnelle et vibrationnelle des partenaires, un réarrangement du solvant, la modification des interactions électrostatiques et de van der Waals entre les partenaires et avec le solvant ou encore la réorganisation de contre-ions (108). Cette complexité rend l'estimation de l'affinité difficile et couteuse. Pour répondre à ce problème, différentes méthodes ont été développées qui peuvent être regroupé en 3 familles :

- les approches les plus simples et les moins couteuses permettent d'estimer l'affinité d'un grand nombre de complexes. Pour arriver à de telles performances, les termes énergétiques sont simplifiés et souvent couplés à des valeurs statistiques regroupées dans des fonctions de score. Cependant, la simplicité de ces fonctions fait qu'il est souvent difficile de discerner deux ligands ayant une différence d'affinité inférieure à 1,5 kcal/mol (109).
- Pour déterminer précisément une affinité, les méthodes basées sur la mécanique statistique, appelées perturbation d'énergie libre (ou FEP pour Free Energy

Perturbation), peuvent être utilisées. Bien qu'ayant un pouvoir prédictif élevé, ces méthodes sont extrêmement couteuses car elles nécessitent de simuler les états initiaux et finaux de la réaction mais également les intermédiaires reliant ces deux états. Ces méthodes sont aussi appelées méthodes alchimiques. Parmi elles ont peut citer les plus connu que sont l'intégration thermodynamique et l'échantillonnage en parapluie.

- Enfin, entre ces deux groupes de méthodes, un autre groupe donnant des performances intermédiaires existe. Comme pour la FEP, ces méthodes sont basées sur l'échantillonnage des conformations mais uniquement des états initiaux et finaux. Elles sont, de ce fait, bien moins couteuses en temps de calcul que les méthodes FEP mais reste théoriquement plus justes que les fonctions de score car plus rigoureuses méthodologiquement. Cependant, ces méthodes nécessitent une étape de paramétrisation, elles sont donc appelées méthodes semi-empiriques.

Etant donné le coût computationnel des méthodes alchimiques au regard de la taille des systèmes que nous étudions dans cette thèse, nous nous intéresserons ici uniquement aux méthodes semi-empiriques ainsi qu'à celles faisant appel à des fonctions de score.

5.1. Les fonctions de score

Les fonctions de score ont pour la plupart été mises au point et utilisées pour discriminer les poses d'amarrage les plus pertinentes. Certaines donnent aussi une estimation de l'affinité de liaison entre un ligand (L) et son récepteur (R). L'affinité peut être mesurée expérimentalement en déterminant la constante d'association à l'équilibre (Keq) qui représente le rapport des concentrations entre le complexe récepteur-ligand (RL) et les formes libres de R et L lorsque la réaction a atteint l'équilibre. Par ailleurs, la constante Keq est directement reliée à la variation de l'énergie libre de Gibbs (ΔG) qui peut aussi être décrite par les variations des contributions enthalpique (ΔH) et entropique (ΔS) :

$$\Delta G = -RT\ ln\ (Keq) = \Delta H - T\Delta S \qquad \text{Equation 16}$$

où R est la constante universelle des gaz parfaits et T la température.

Dans des conditions de température et de pression constantes et lorsque le système a atteint l'équilibre, une variation négative de ΔG correspond à un processus spontané d'association. L'équation 16 montre que la magnitude de ΔG est déterminée par la constante Keq. Nous pouvons donc considérer que la valeur de ΔG reflète l'affinité de liaison entre un récepteur et un ligand, ou encore la stabilité du complexe RL. L'équation 16 montre également que ΔG résulte des contributions enthalpique et entropique intervenant dans le processus de liaison. Dans des conditions où la pression est constante et la variation de volume du système ne varie pas, la variation d'enthalpie peut être définie comme la variation de l'énergie totale du système faisant suite à l'association d'un ligand à son récepteur. Elle découle d'une balance énergétique résultant de la formation et de la rupture d'un ensemble d'interactions. Une liaison enthalpiquement favorable peut ainsi mettre en jeu une perte de liaisons hydrogènes et d'interactions de van der Waals/électrostatiques entre chacun des solutés et le solvant, perte qui doit être compensée par la formation de nouvelles interactions non covalentes établies entre le récepteur et le ligand. L'énergie interne des solutés intervient également dans la contribution enthalpique puisque les conformations du récepteur et du ligand peuvent être davantage contraintes dans l'état du complexe RL et donc énergétiquement moins favorables qu'elles ne peuvent l'être dans leurs formes libres respectives. L'équation 16 montre qu'un ΔH négatif (favorable) conduira à un processus d'association spontané uniquement si cette variation n'est pas compensée par une variation d'entropie défavorable.

L'entropie représente une mesure du désordre qui, dans le contexte d'un système simple composé d'un complexe protéique baignant dans un milieu aqueux, peut être interprété par le nombre d'états que peut adopter chacune des entités moléculaires du système. Les états possibles d'une molécule intègrent par exemple l'ensemble de ses conformations potentielles, ses possibilités de translations et rotations, ou encore les fréquences de vibrations de ses atomes. Lorsqu'un ligand se fixe à un récepteur, le nombre de ses états possibles réduit drastiquement, ce qui entraîne une forte pénalité entropique. Un phénomène important permet cependant de contrebalancer cette dernière : l'effet hydrophobe. Pour saisir son action, considérons une molécule hydrophobe (ou non polaire). Ses propriétés physico-chimiques ne lui permettent pas d'établir de liaisons hydrogènes (ou d'interactions électrostatiques "fortes") avec des molécules d'eau. Son introduction dans un milieu aqueux tend à rompre une partie du réseau dynamique de liaisons hydrogènes que forment les molécules d'eau entre elles. Les molécules d'eau au voisinage direct d'une surface hydrophobe ont toutefois tendance à s'orienter de manière à maximiser leur potentiel à établir des liaisons hydrogènes avec les molécules d'eau environnantes. Cet agencement tend à réduire leur mobilité et entraîne donc une augmentation de l'ordre qui est entropiquement

pénalisante (110,111). Par ailleurs, plusieurs études suggèrent que les molécules d'eau entourant un composé non polaire tendent à former moins de liaisons hydrogènes, conférant également une pénalité enthalpique à son énergie de solvatation (112,113). L'effet hydrophobe désigne le processus amenant des composés non polaires à spontanément s'associer en milieu aqueux. Le caractère spontané de cette réaction résulte, notamment, d'une diminution de la surface nette hydrophobe solvatée à la suite de l'association, cette dernière s'accompagnant d'une libération des molécules d'eau bénéfique sur le plan entropique. Ces dernières peuvent alors à nouveau participer au réseau de liaisons hydrogènes formé par les molécules d'eau environnantes apportant ainsi une contribution enthalpique favorable. La complémentarité stérique entre les régions non polaires contribue également à leur agrégation en favorisant des interactions de van der Waals.

L'effet hydrophobe représente donc un facteur majeur de stabilité des assemblages biologiques. En effet, il est essentiel au repliement des protéines et des acides nucléiques, à la formation des membranes lipidiques, et évidemment aux interactions intermoléculaires. Plusieurs travaux ont par exemple montré que l'affinité d'un ligand envers son récepteur peut être augmentée par l'ajout de groupements hydrophobes (114).

Les fonctions de score sont des modèles mathématiques utilisés pour estimer une énergie d'interaction entre un récepteur et un ligand. Dans une approche d'amarrage réalisée pour prédire leur mode d'interaction, les fonctions de score doivent pouvoir permettre de distinguer, parmi un ensemble varié de poses, celles qui sont pertinentes. Dans l'idéale elles doivent aussi être capables d'identifier quels sont les couples (ligand, récepteur) qui peuvent réellement se lier entre eux et même pouvoir les classer selon leur affinité respective. Le nombre de poses à évaluer au cours de calculs d'amarrage peut varier de plusieurs milliers à plusieurs millions selon la stratégie d'échantillonnage adoptée et le nombre de couple à tester. Par conséquent, les fonctions de score doivent aussi être rapides. L'efficacité de leurs temps de calculs ne peut cependant être obtenue sans approximations ou simplifications du modèle décrivant le processus de liaison, et s'atteint donc au détriment de la précision dans l'évaluation de l'énergie d'interaction. De nombreuses fonctions de score ont été développées dans le but de répondre à ces contraintes d'efficacité et de précisions.

Elles peuvent se classer en quatre grandes catégories :

1) Les fonctions de score basées sur les champs de force, généralement CHARMM ou AMBER. Elles correspondent à un ensemble de paramètres et de fonctions permettant de définir un système et de décrire son paysage d'énergie potentielle. Ces fonctions de score au

formalisme très simple présentent l'avantage d'être relativement peu coûteuses en temps de calculs. Elles restent cependant très approximatives puisqu'elles n'estiment qu'une contribution enthalpique à l'énergie de liaison. Des facteurs importants intervenant dans l'affinité de liaison sont en effet négligés, comme l'effet du solvant ou les composantes entropiques.

2) Les fonctions empiriques. Elles estiment l'affinité de liaison ΔG entre un couple récepteur-ligand sur la base d'un ensemble de descripteurs énergétiques, chacun étant pondérés par un coefficient. Les descripteurs peuvent inclure différents termes énergétiques empruntés à la mécanique moléculaire (énergies de van der Waals, électrostatiques) mais aussi d'autres composantes idéalement non corrélées considérant par exemple l'hydrophobicité et/ou la polarité d'un site de liaison, l'accessibilité au solvant, l'entropie d'un ligand ou encore d'autres propriétés. Les coefficients associés à chacun des termes énergétiques sont déterminés par des analyses de régressions linéaires de manière à optimiser la corrélation entre des énergies d'association estimées par la fonction de score et des affinités de liaison connues pour des complexes dont la structure est résolue expérimentalement.

3) Les fonctions reposant sur des potentiels statistiques. Ces fonctions de score sont basées sur une idée de la mécanique statistique permettant de dériver des potentiels de forces moyennes à partir de distributions de mesures observées entre paires d'atomes. Si des contacts entre une paire d'atomes donnée apparaît plus souvent que dans un état de référence, alors ces contacts auront une énergie d'interaction favorable. Les fonctions de score à potentiel statistique partagent avec les fonctions empiriques le fait d'essayer de capturer implicitement des facteurs intervenant dans le processus de liaison qui sont difficiles à modéliser explicitement.

4) Les fonctions basées sur des techniques d'apprentissage automatique. Elles se distinguent de toutes ces fonctions conventionnelles puisqu'elles dérivent à partir d'un jeu d'entraînement une relation mathématique entre un large ensemble de descripteurs dont la forme n'est plus définie *a priori* et qui ne suit donc plus nécessairement un modèle additif et linéaire. Plusieurs études comparatives ont montré la supériorité des modèles non-linéaires dans leur capacité à prédire l'affinité de liaison (115–117). Cependant ces études comparatives restent cantonnées à l'interaction protéine-ligand (le ligand n'étant pas lui-même une protéine).

Dans cette thèse nous utiliserons en particulier une fonction de score de type empiriques, celle issue des travaux de Vangone et Bonvin. Cette fonction de score est celle qui donne les meilleures performances au regard du coefficients de corrélation (-0.73)

concernant la prédiction de l'affinité protéine-protéine (118). Ce score calcul le ΔG d'un complexe selon la formule suivante :

$$\Delta G_{calc} = 0.09459\,ICs_{charged/charged} + 0.10007\,ICs_{charged\_apolar} - 0.19577\,ICs_{polar/polar} + 0.22671\,ICs_{polar/apolar} - 0.18681\%NIS_{apolar} - 0.13810\%NIS_{charged} + 15.9433. \quad \text{Equation 17}$$

où ICs sont les contacts inter-résidu (Inter-residue Contacts) et %NIS représente le pourcentage des surfaces non interactives (Non-Interacting Surface). Les deux paramètres étant triés et pondérés en fonction de la nature électrostatique des résidus concernés sur chaque chaine du complexe.

5.2. Les méthodes semi-empiriques

Les fonctions de score sont rapides à exécuter mais leurs puissances prédictives peuvent être insuffisante pour discriminer l'énergie d'interaction entre 2 complexes proche. En revanche les méthodes exactes, qui sont supposées être fortement prédictives, sont très couteuses en calcul et donc difficilement applicables à pour un grand nombre de complexes ou des complexes de grandes tailles. C'est pourquoi il a été élaboré des approches moins rigoureuses et moins couteuses que les méthodes exactes car elles ne simulent que les états initiaux et finaux de la réaction. L'énergie libre de liaison, ΔG, est alors déterminée de la manière suivante :

$$\Delta G_{liaison} = [\Delta G_{associé}] - [\Delta G_{dissocié}] \quad \text{Equation 18}$$

où $\Delta G_{associé}$ représente l'énergie libre de la forme associée du complexe et $\Delta G_{dissocié}$ celle de la forme dissociée. La plupart du temps, les énergies libres sont moyennées sur un ensemble de conformations plutôt que calculées sur une conformation unique. Pour cela, des simulations de dynamique moléculaire ou Monte Carlo sont effectuées puis des conformations sont extraites à intervalles réguliers. Cet ensemble conformationnel est symbolisé dans l'équation 18 par les crochets.

Les deux principaux représentant de ces méthodes basées sur cette approximation sont les modèles d'énergies d'interaction linéaire (en anglais LIE pour Linear Interaction Energy) et les modèles de types MM-PBSA (Molecular Mechanics, Poisson–Boltzmann Surface Area). De nombreux variants de ces méthodes ont été développés au cours des dernières années. Dans cette thèse nous ne nous attarderons que sur la méthode de MM-PBSA car elle est, non seulement la méthode semi-empirique la plus utilisé de manière générale, mais aussi celle sur laquelle nous possédons le plus de recul quant à son utilisation pour le calcul de l'interaction protéine-protéine (119).

Initialement développée par Kollman *et al.* (120), elle estime l'énergie libre de liaison à partir de l'énergie libre des partenaires associés et dissociés :

$$\Delta G_{liaison} = [G_{PL}] - [G_P] - [G_L] \qquad \text{Equation 19}$$

L'énergie libre de chaque état est calculée comme suit :

$$G = E_{MM} + G_{solv} - TS \qquad \text{Equation 20}$$

où E_{MM} correspond à l'énergie totale calculée à l'aide de la mécanique moléculaire, Gsolv représente l'énergie libre de solvatation, S est l'entropie conformationnelle et T la température. Comme dans le cas du LIE, les approches de type MM-PBSA sont généralement appliquées à un ensemble de conformations générées par dynamique moléculaire ou Monte Carlo en solvant explicite. Les différents termes énergétiques sont ensuite déterminés lors d'une étape de post-traitement pendant laquelle les molécules d'eau sont retirées et remplacées par un continuum diélectrique. Nous allons maintenant détailler plus précisément ces 3 différents termes énergétiques.

1) Les termes de mécanique moléculaire : le terme E_{MM} est issu de la mécanique moléculaire et décrit les interactions liées (liaisons, angles, dièdres, impropres) et non liées entre les atomes (électrostatique et van der Waals). Elles sont calculées à partir des paramètres du champ de force.

$$E_{MM} = E_{liée} + E_{elec} + E_{vdW} \qquad \text{Equation 21}$$

Les termes d'énergie liée des protéines et du complexe peuvent être de l'ordre de quelques centaines de kcal/mol et introduire une incertitude importante dans l'estimation de l'énergie libre. Afin de limiter cette incertitude, les conformations des états lié et dissocié sont généralement extraites de la trajectoire du complexe (121–123). De ce fait, le récepteur et le ligand présentent exactement la même conformation dans les deux états, les termes $E_{liée}$ s'annulent donc. Cette approche, appelée mono-trajectoire, réduit considérablement le temps de calcul puisqu'il n'est plus nécessaire de simuler le récepteur et le ligand dans leur état dissocié. L'approche mono-trajectoire fait cependant l'hypothèse que la liaison n'entraine pas de changements conformationnels importants.

2) Les termes d'énergie de solvatation : l'énergie de solvatation, G_{solv}, décrit les interactions entre le soluté et le solvant mais également l'écrantage des interactions électrostatiques entre les atomes du soluté. Actuellement, le meilleur moyen de décrire le solvant consiste à le traiter de manière explicite à l'aide de modèles moléculaires représentant les molécules d'eau (par exemples les modèles SPC (Simple Point-Charge), SPC/E, TIP3P (Transferable Intermolecular Potential with 3 Points), TIP4P ou TIP5P). Ces modèles sont cependant trop couteux pour permettre un calcul rapide. Les méthodes de MM-PBSA traitent donc le solvant de manière implicite. L'énergie de solvatation est alors décomposée en deux termes, un terme ${G_{solv}}^{pol}$ décrivant les effets électrostatiques du solvant et un terme ${G_{solv}}^{apol}$ décrivant les interactions apolaires entre le soluté et le solvant.

$$G_{solv} = {G_{solv}}^{pol} + {G_{solv}}^{apol} \qquad \text{Equation 22}$$

Le terme ${G_{solv}}^{pol}$ est généralement traité par un terme de Poisson-Boltzmann (PB) ou de Born généralisé (GB) dans lesquels le soluté forme une cavité de constante diélectrique faible (généralement comprise entre 1 et 8) entourée d'un continuum de constante diélectrique élevée (généralement 80). Le terme ${G_{solv}}^{apol}$ correspond à la formation d'une cavité dans le solvant ainsi qu'aux forces attractives et répulsives de van der Waals entre le soluté et le solvant. Il est généralement proportionnel à la surface accessible au solvant (SASA). Le terme GB est bien moins couteux en temps de calcul tout en donnant une approximation généralement proche du terme PB concernant le pouvoir prédictif (116). Cette problématique du temps de calcul est d'autant plus importante sur les gros systèmes tel que les complexes protéiques, c'est pourquoi la variante MM-GBSA (Molecular Mechanics, Generalized Born Surface Area) sera privilégié dans les travaux de cette thèse.

3) Le terme entropique : le terme TS vise à prendre en compte l'entropie conformationnelle du soluté. Ce terme est généralement calculé par une analyse en modes normaux des conformations issues d'une simulation puis minimisées. Cette approche est cependant très couteuse en temps de calcul car un grand nombre de conformations est généralement nécessaire pour que les valeurs convergent (123). De plus, le caractère prédictif qu'apporte de ce terme n'a été clairement démontré dans aucune revue systématique, il est donc généralement omis lors des calculs d'affinité (124,125). Une mauvaise approximation de ce terme ou son omission dans les calculs de prédiction d'énergie libre d'interaction demeure malgré tout une des principales limites méthodologiques au MM-PB(GB)SA.

6. Conclusion

Dans ce chapitre, nous avons fait un tour d'horizon non exhaustif des méthodes bioinformatiques que nous pouvons utiliser pour l'étude structurale des protéines. Nous nous sommes plus particulièrement attardés sur celles dont nous nous sommes servit au cours des travaux de cette thèse. La modélisation de nouvelles protéines et des interactions qu'elles forment les unes envers les autres revêt d'un intérêt majeur dans le processus de recherche et de développement de l'industrie pharmaceutique. L'étude des antigènes ainsi que de leur interaction avec d'autre protéines (en particulier les anticorps) est une composante cruciale dans la réalisation de candidat vaccin. En ce sens, les méthodes de bioinformatiques structurales, bien que perfectibles, sont en perpétuelles améliorations/évolutions, ceci afin d'être de plus en plus fiables et de mieux compléter/assister les méthodes expérimentales.

Dans cette dynamique, les algorithmes d'apprentissage automatique sont de plus en plus utilisés et démontre parfois une certaine supériorité, tant en terme de performance qu'en terme d'efficacité, face à des méthodes plus classiques. Nous pouvons prendre pour exemple des algorithmes d'apprentissage automatique tel que AlphaFold pour la modélisation ainsi que certaines fonctions de score pour l'interaction protéine-ligand, que nous avons cité respectivement en partie 3.2. et 5.1. de ce chapitre. Ces méthodes d'apprentissages automatiques performantes ont comme point commun d'utiliser un type d'algorithme en particulier qui se nomme apprentissage en profondeur. Ce type d'algorithme n'a pas encore fait l'objet d'une utilisation dans ce champ d'étude qu'est la prédiction de l'énergie affinité

protéine-protéine et qui nous intéresse ici particulièrement. C'est pourquoi il fera l'objet du prochain chapitre introductif ainsi que d'un chapitre d'étude à part entière.

BIBLIOGRAPHIE DU CHAPITRE 2

1. Davis AM, Teague SJ. Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. Angew Chem Int Ed Engl. 15 mars 1999;38(6):736-49.

2. Williams DH, Searle MS, Mackay JP, Gerhard U, Maplestone RA. Toward an estimation of binding constants in aqueous solution: studies of associations of vancomycin group antibiotics. Proc Natl Acad Sci U S A. 15 févr 1993;90(4):1172-8.

3. Creighton TE. Disulphide bonds and protein stability. BioEssays. 1 févr 1988;8(2-3):57-63.

4. Thornton JM. Disulphide bridges in globular proteins. J Mol Biol. 15 sept 1981;151(2):261-87.

5. Katz BA, Kossiakoff A. The crystallographically determined structures of atypical strained disulfides engineered into subtilisin. J Biol Chem. 25 nov 1986;261(33):15480-5.

6. Pace CN, Fu H, Fryar KL, Landua J, Trevino SR, Shirley BA, et al. Contribution of hydrophobic interactions to protein stability. J Mol Biol. 6 mai 2011;408(3):514-28.

7. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature. 8 mars 1958;181(4610):662-6.

8. Schmidt A, Teeter M, Weckert E, Lamzin VS. Crystal structure of small protein crambin at 0.48 Å resolution. Acta Crystallograph Sect F Struct Biol Cryst Commun. 1 avr 2011;67(Pt 4):424-8.

9. Polikanov YS, Steitz TA, Innis CA. A proton wire to couple aminoacyl-tRNA accommodation and peptide-bond formation on the ribosome. Nat Struct Mol Biol. sept 2014;21(9):787-93.

10. Williamson MP, Havel TF, Wüthrich K. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. J Mol Biol. 20 mars 1985;182(2):295-315.

11. Dubochet J, Adrian M, Chang JJ, Homo JC, Lepault J, McDowall AW, et al. Cryo-electron microscopy of vitrified specimens. Q Rev Biophys. mai 1988;21(2):129-228.

12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 1 janv 2000;28(1):235-42.

13. DeLano W. The PyMOL Molecular Graphics System, v1.8. (Schrödinger, LLC).

14. Schrödinger Release 2020-2: Maestro. (Schrödinger, LLC, New York, NY, 2020).

15. Viani L. MView: A tool for visualization and analysis of molecular properties.

16. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. oct 2004;25(13):1605-12.

17. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph. févr 1996;14(1):33-8, 27-8.

18. Geourjon C, Deléage G, Roux B. ANTHEPROT: an interactive graphics software for analyzing protein structures from sequences. J Mol Graph. sept 1991;9(3):188-90, 167.

19. Beuming T, Kniazeff J, Bergmann ML, Shi L, Gracia L, Raniszewska K, et al. The binding sites for cocaine and dopamine in the dopamine transporter overlap. Nat Neurosci. juill 2008;11(7):780-9.

20. Lewis D. The CAVE artists. Nat Med. mars 2014;20(3):228-30.

21. McIntosh-Smith null, Tchertanov null, Nerukh null, Stone null, Baaden null, Hayward null, et al. Computing power revolution and new algorithms: GP-GPUs, clouds and more: general discussion. Faraday Discuss. 2014;169:379-401.

22. Levieux G, Tiger G, Mader S, Zagury J-F, Natkin S, Montes M. Udock, the interactive docking entertainment system. Faraday Discuss. 2014;169:425-41.

23. Sela M, White FH, Anfinsen CB. Reductive cleavage of disulfide bridges in ribonuclease. Science. 12 avr 1957;125(3250):691-2.

24. Anfinsen CB. Principles that govern the folding of protein chains. Science. 20 juill 1973;181(4096):223-30.

25. Pace CN, Shirley BA, McNutt M, Gajiwala K. Forces contributing to the conformational stability of proteins. FASEB J Off Publ Fed Am Soc Exp Biol. janv 1996;10(1):75-83.

26. Van den Berg B, Wain R, Dobson CM, Ellis RJ. Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell. EMBO J. 1 août 2000;19(15):3870-5.

27. Huber R, Bennett WS. Functional significance of flexibility in proteins. Biopolymers. janv 1983;22(1):261-79.

28. Schaller RR. Moore's law: past, present and future. IEEE Spectr. juin 1997;34(6):52-9.

29. Loll PJ. Membrane protein structural biology: the high throughput challenge. J Struct Biol. avr 2003;142(1):144-53.

30. Rawlings AE. Membrane proteins: always an insoluble problem? Biochem Soc Trans. 15 2016;44(3):790-5.

31. Lafita A, Bliven S, Kryshtafovych A, Bertoni M, Monastyrskyy B, Duarte JM, et al. Assessment of protein assembly prediction in CASP12. Proteins. 2018;86 Suppl 1:247-56.

32. Vakser IA. Protein-Protein Docking: From Interaction to Interactome. Biophys J. 21 oct 2014;107(8):1785-93.

33. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. avr 1986;5(4):823-6.

34. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A. 25 janv 2005;102(4):1029-34.

35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 5 oct 1990;215(3):403-10.

36. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 22 mars 1985;227(4693):1435-41.

37. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. avr 1988;85(8):2444-8.

38. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 5 oct 2001;294(5540):93-6.

39. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci Publ Protein Soc. sept 2000;9(9):1753-73.

40. Lathrop RH. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng. sept 1994;7(9):1059-68.

41. Westhead DR, Collura VP, Eldridge MD, Firth MA, Li J, Murray CW. Protein fold recognition by threading: comparison of algorithms and analysis of results. Protein Eng. déc 1995;8(12):1197-204.

42. Chothia C. Proteins. One thousand families for the molecular biologist. Nature. 18 juin 1992;357(6379):543-4.

43. Leonov H, Mitchell JSB, Arkin IT. Monte Carlo estimation of the number of possible protein folds: effects of sampling bias and folds distributions. Proteins. 15 mai 2003;51(3):352-9.

44. Peng J, Xu J. RaptorX: exploiting structure information for protein alignment by statistical inference. Proteins. 2011;79 Suppl 10:161-71.

45. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 1 juill 2005;33(Web Server issue):W244-248.

46. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods. janv 2015;12(1):7-8.

47. Englander SW, Mayne L. The nature of protein folding pathways. Proc Natl Acad Sci U S A. 11 nov 2014;111(45):15873-80.

48. Karplus M. The Levinthal paradox: yesterday and today. Fold Des. 1 juin 1997;2:S69-75.

49. Lipman EA, Schuler B, Bakajin O, Eaton WA. Single-molecule measurement of protein folding kinetics. Science. 29 août 2003;301(5637):1233-5.

50. Schuler B, Eaton WA. Protein folding studied by single-molecule FRET. Curr Opin Struct Biol. févr 2008;18(1):16-26.

51. Floudas CA. Computational methods in protein structure prediction. Biotechnol Bioeng. 1 juin 2007;97(2):207-13.

52. Bonneau R, Baker D. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct. 2001;30:173-89.

53. Deng H, Jia Y, Zhang Y. Protein structure prediction. Int J Mod Phys B [En ligne]. 20 juill 2018 [cité le 25 juill 2020];32(18). Disponible: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6407873/

54. Breda A, Santos DS, Basso LA, de Souza ON. Ab initio 3-D structure prediction of an artificially designed three-alpha-helix bundle via all-atom molecular dynamics simulations. Genet Mol Res GMR. 5 oct 2007;6(4):901-10.

55. Laughton CA. A study of simulated annealing protocols for use with molecular dynamics in protein structure prediction. Protein Eng Des Sel. Oxford Academic; 1 févr 1994;7(2):235-41.

56. Zhou R. Replica exchange molecular dynamics method for protein folding simulation. Methods Mol Biol Clifton NJ. 2007;350:205-23.

57. Liwo A, Ołdziej S, Kaźmierkiewicz R, Groth M, Czaplewski C. Design of a knowledge-based force field for off-lattice simulations of protein structure. Acta Biochim Pol. 1997;44(3):527-47.

58. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink S-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. J Chem Theory Comput. mai 2008;4(5):819-34.

59. [En ligne]. Groups Analysis: zscores - CASP13; [cité le 25 juill 2020]. Disponible: https://www.predictioncenter.org/casp13/zscores_final.cgi?model_type=first&gr_type=server_only

60. [En ligne]. Groups Analysis: zscores - CASP14; [cité le 15 janv 2021]. Disponible: https://www.predictioncenter.org/casp14/zscores_final.cgi?model_type=first&gr_type=server_only

61. Connolly ML. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. Biopolymers. juill 1986;25(7):1229-47.

62. Fischer D, Norel R, Wolfson H, Nussinov R. Surface motifs by a computer vision technique: searches, detection, and implications for protein-ligand recognition. Proteins. juill 1993;16(3):278-92.

63. Lee RH, Rose GD. Molecular recognition. I. Automatic identification of topographic surface features. Biopolymers. août 1985;24(8):1613-27.

64. Cherfils J, Duquerroy S, Janin J. Protein-protein recognition analyzed by docking simulation. Proteins. 1991;11(4):271-80.

65. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol. 14 juin 1976;104(1):59-107.

66. Wang H. Grid-search molecular accessible surface algorithm for solving the protein docking problem. J Comput Chem. 1991;12(6):746-50.

67. Jiang F, Kim SH. « Soft docking »: matching of molecular surface cubes. J Mol Biol. 5 mai 1991;219(1):79-102.

68. Carter P, Lesk VI, Islam SA, Sternberg MJE. Protein-protein docking using 3D-Dock in rounds 3, 4, and 5 of CAPRI. Proteins. 1 août 2005;60(2):281-8.

69. Wiehe K, Pierce B, Mintseris J, Tong WW, Anderson R, Chen R, et al. ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. Proteins. 1 août 2005;60(2):207-13.

70. Smith GR, Sternberg MJE, CAPRI blind trial. Evaluation of the 3D-Dock protein docking suite in rounds 1 and 2 of the CAPRI blind trial. Proteins. 1 juill 2003;52(1):74-9.

71. Heifetz A, Katchalski-Katzir E, Eisenstein M. Electrostatics in protein-protein docking. Protein Sci Publ Protein Soc. mars 2002;11(3):571-87.

72. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins. 1 juin 2002;47(4):409-43.

73. Fernández-Recio J, Abagyan R, Totrov M. Improving CAPRI predictions: optimized desolvation for rigid-body docking. Proteins. 1 août 2005;60(2):308-13.

74. Comeau SR, Vajda S, Camacho CJ. Performance of the first protein docking server ClusPro in CAPRI rounds 3-5. Proteins. 1 août 2005;60(2):239-44.

75. Chen R, Weng Z. A novel shape complementarity scoring function for protein-protein docking. Proteins. 15 mai 2003;51(3):397-408.

76. Daily MD, Masica D, Sivasubramanian A, Somarouthu S, Gray JJ. CAPRI rounds 3-5 reveal promising successes and future challenges for RosettaDock. Proteins. 1 août 2005;60(2):181-6.

77. Zhang C, Liu S, Zhu Q, Zhou Y. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J Med Chem. 7 avr 2005;48(7):2325-35.

78. Zhang C, Liu S, Zhou Y. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci Publ Protein Soc. févr 2004;13(2):391-9.

79. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. Nat Protoc. févr 2017;12(2):255-78.

80. Janin J. Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst. déc 2010;6(12):2351-62.

81. Lensink MF, Nadzirin N, Velankar S, Wodak SJ. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. Proteins Struct Funct Bioinforma. 2020;88(8):916-38.

82. Díaz N, Suárez D. Extensive simulations of the full-length matrix metalloproteinase-2 enzyme in a prereactive complex with a collagen triple-helical peptide. Biochemistry. 10 févr 2015;54(5):1243-58.

83. Koldsø H, Shorthouse D, Hélie J, Sansom MSP. Lipid clustering correlates with membrane curvature as revealed by molecular simulations of complex lipid bilayers. PLoS Comput Biol. oct 2014;10(10):e1003911.

84. Zhao G, Perilla JR, Yufenyuy EL, Meng X, Chen B, Ning J, et al. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature. 30 mai 2013;497(7451):643-6.

85. Götz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J Chem Theory Comput. American Chemical Society; 8 mai 2012;8(5):1542-55.

86. Rodrigues CI, Hardy DJ, Stone JE, Schulten K, Hwu W-MW. GPU acceleration of cutoff pair potentials for molecular modeling applications. Dans: Proceedings of the 5th conference on Computing frontiers [En ligne]. New York, NY, USA : Association for Computing Machinery; 2008. p. 273-82. (CF '08).

87. Harvey MJ, Giupponi G, Fabritiis GD. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. J Chem Theory Comput. American Chemical Society; 9 juin 2009;5(6):1632-9.

88. Dror RO, Young C, Shaw DE. Anton, A Special-Purpose Molecular Simulation Machine. Dans: Padua D, rédacteur. Encyclopedia of Parallel Computing [En ligne]. Boston, MA : Springer US; 2011 [cité le 23 juill 2020]. p. 60-71. Disponible: https://doi.org/10.1007/978-0-387-09766-4_199

89. Piana S, Lindorff-Larsen K, Shaw DE. Atomic-level description of ubiquitin folding. Proc Natl Acad Sci U S A. 9 avr 2013;110(15):5915-20.

90. Nyquist H. Certain Topics in Telegraph Transmission Theory. Trans Am Inst Electr Eng. avr 1928;47(2):617-44.

91. Hockney RW, Goel SP, Eastwood JW. Quiet high-resolution computer models of a plasma. J Comput Phys. 1 févr 1974;14(2):148-58.

92. Swope WC, Andersen HC, Berens PH, Wilson KR. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. J Chem Phys. American Institute of Physics; 1 janv 1982;76(1):637-49.

93. Darden T, York D, Pedersen L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J Chem Phys. American Institute of Physics; 15 juin 1993;98(12):10089-92.

94. Ewald PP. Die Berechnung optischer und elektrostatischer Gitterpotentiale. Ann Phys. 1921;369(3):253-87.

95. Laio A, Parrinello M. Escaping free-energy minima. Proc Natl Acad Sci U S A. 1 oct 2002;99(20):12562-6.

96. Schlitter J, Engels M, Krüger P. Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. J Mol Graph. juin 1994;12(2):84-9.

97. McGreevy R, Teo I, Singharoy A, Schulten K. Advances in the molecular dynamics flexible fitting method for cryo-EM modeling. Methods San Diego Calif. 01 2016;100:50-60.

98. Grubmüller H, Heymann B, Tavan P. Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. Science. 16 févr 1996;271(5251):997-9.

99. Isralewitz B, Izrailev S, Schulten K. Binding pathway of retinal to bacterio-opsin: a prediction by molecular dynamics simulations. Biophys J. déc 1997;73(6):2972-9.

100. Ponder JW, Case DA. Force fields for protein simulations. Adv Protein Chem. 2003;66:27-85.

101. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J Chem Theory Comput. 10 sept 2013;9(9):3878-88.

102. Nooren IMA, Thornton JM. Diversity of protein–protein interactions. EMBO J. John Wiley & Sons, Ltd; 15 juill 2003;22(14):3486-92.

103. Marsh JA, Teichmann SA. Structure, dynamics, assembly, and evolution of protein complexes. Annu Rev Biochem. 2015;84:551-75.

104. Johnson GT, Autin L, Al-Alusi M, Goodsell DS, Sanner MF, Olson AJ. cellPACK: a virtual mesoscope to model and visualize structural systems biology. Nat Methods. janv 2015;12(1):85-91.

105. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. Nature Publishing Group; janv 2013;10(1):47-53.

106. King NP, Bale JB, Sheffler W, McNamara DE, Gonen S, Gonen T, et al. Accurate design of co-assembling multi-component protein nanomaterials. Nature. Nature Publishing Group; juin 2014;510(7503):103-8.

107. Schreiber G, Fleishman SJ. Computational design of protein–protein interactions. Curr Opin Struct Biol. 1 déc 2013;23(6):903-10.

108. Simonson T. The Physical Basis of Ligand Binding. Dans: 2015. p. 3-43.

109. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov. Taylor & Francis; 4 mai 2015;10(5):449-61.

110. Ball P. Water as an active constituent in cell biology. Chem Rev. janv 2008;108(1):74-108.

111. Southall NT, Dill KA, Haymet ADJ. A View of the Hydrophobic Effect. J Phys Chem B. American Chemical Society; 1 janv 2002;106(3):521-33.

112. Ji N, Ostroverkhov V, Tian CS, Shen YR. Characterization of vibrational resonances of water-vapor interfaces by phase-sensitive sum-frequency spectroscopy. Phys Rev Lett. 7 mars 2008;100(9):096102.

113. Richmond G. Structure and bonding of molecules at aqueous surfaces. Annu Rev Phys Chem. 2001;52:357-89.

114. Houk KN, Leach AG, Kim SP, Zhang X. Binding affinities of host-guest, protein-ligand, and protein-transition-state complexes. Angew Chem Int Ed Engl. 20 oct 2003;42(40):4872-97.

115. Ashtawy HM, Mahapatra NR. A comparative assessment of ranking accuracies of conventional and machine-learning-based scoring functions for protein-ligand binding affinity prediction. IEEE/ACM Trans Comput Biol Bioinform. oct 2012;9(5):1301-13.

116. Wang C, Zhang Y. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. J Comput Chem. 30 janv 2017;38(3):169-77.

117. Wójcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep. 25 avr 2017;7:46710.

118. Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein–protein complexes. Levitt M, rédacteur. eLife. eLife Sciences Publications, Ltd; 20 juill 2015;4:e07454.

119. Chen F, Liu H, Sun H, Pan P, Li Y, Li D, et al. Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein-protein binding free energies and re-rank binding poses generated by protein-protein docking. Phys Chem Chem Phys PCCP. 10 août 2016;18(32):22129-39.

120. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res. déc 2000;33(12):889-97.

121. Gohlke H, Case DA. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. J Comput Chem. 30 janv 2004;25(2):238-50.

122. Lepsík M, Kríz Z, Havlas Z. Efficiency of a second-generation HIV-1 protease inhibitor studied by molecular dynamics and absolute binding free energy calculations. Proteins. 1 nov 2004;57(2):279-93.

123. Genheden S, Ryde U. Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies. Proteins. mai 2012;80(5):1326-42.

124. Foloppe N, Hubbard R. Towards predictive ligand design with free energy based computational methods? Curr Med Chem. 2006;13(29):3583-608.

125. Rastelli G, Del Rio A, Degliesposti G, Sgobba M. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. J Comput Chem. mars 2010;31(4):797-810.

Chapitre 3 : L'apprentissage automatique par les réseaux de neurones artificiels

L'apprentissage automatique fait partie de ce que nous appelons communément l'intelligence artificielle (IA). L'IA est un terme générique qui se définit comme l'ensemble des théories et des techniques mises en œuvre afin de réaliser des machines et algorithmes capables de simuler l'intelligence. Mais bien que le concept d'IA englobe celui de l'apprentissage automatique, c'est bel et bien par ce dernier que l'IA a pu enfin commencer à dépasser les capacités humaines dans certains domaines.

Si au cours des deux dernières décennies, nous avons pu constater un développement fulgurant des algorithmes d'apprentissage automatique, l'un d'eux s'est particulièrement distingué ces dernières années par des succès retentissants. Il s'agit de l'apprentissage en profondeur par réseaux de neurones artificiels (Deep Learning).

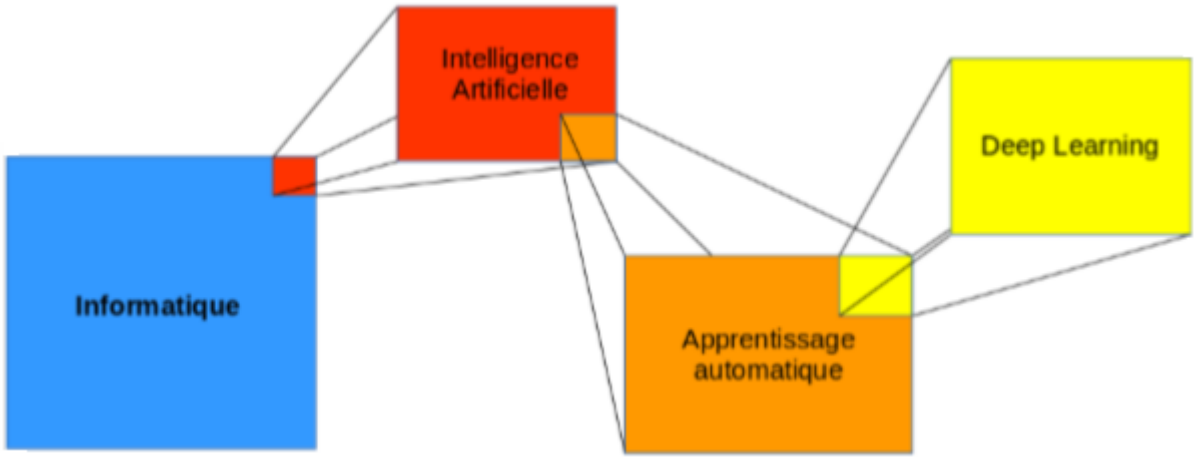


Figure 11 : Schéma résumant la place du deep learning dans le monde de l'informatique.

L'un des premiers succès retentissants de l'apprentissage en profondeur par réseaux de neurones artificiels fut le concours de reconnaissance et de classification d'images

ImageNet de 2012 (1). Ce concours met en compétition des équipes de recherche du monde entier chaque année pour classifier automatiquement des images selon 1000 catégories différentes. En 2012, l'équipe qui l'a emporté l'a fait avec un réseau de neurones dit profond. Ce fut une première pour cette compétition. En effet, la grande complexité de la tâche semblait jusque-là trop importante pour pouvoir utiliser des réseaux de neurones. Cependant, là où le taux d'erreur de classification d'images des autres méthodes ne descendait pas sous les 25%, celle du réseau de neurones de l'équipe Krizhevsky descendit à 16% (2). Toutes les meilleures équipes se sont alors mises à l'utilisation de cette méthode d'apprentissage automatique faisant passer le taux d'erreur de 14% en 2013 à moins de 5% en 2017. En octobre 2015, une nouvelle étape importante est atteinte lorsque le programme « AlphaGo » de Google gagne contre un des meilleurs joueurs au jeu de Go (3). C'est la première fois qu'un programme de go bat un joueur professionnel dans un match avec des parties sur un goban de taille normale (19×19) et sans handicap. Le jeu de Go est un jeu de plateau considéré comme le plus dur du monde (4).

Aujourd'hui l'apprentissage automatique par réseaux de neurones artificiels est omniprésent dans le secteur des technologies, de l'information et de la communication. Il s'applique aussi bien dans le cadre des véhicules autonomes (5) que de celui de la reconnaissance faciale, des algorithmes de recherche ou du ciblage publicitaire. Mais son champ d'application est aussi présent dans tous les domaines des sciences tels que la physique, la bio-informatique (6) et la médecine (7–10). Comme divers exemples à ces 3 disciplines nous pouvons citer :

- La recherche sur les particules élémentaires exotiques (11).
- La prédiction du repliement des protéines avec l'algorithme AlphaFold (12).
- L'aide au diagnostic médical comme avec reconnaissance automatique des cancers en imagerie médicale (13).

Dans ce chapitre nous développerons dans une première partie les bases de l'apprentissage automatique en citant les principaux algorithmes utilisés. Puis dans une deuxième partie nous nous étendrons plus particulièrement sur un de ces algorithmes qu'est l'usage des réseaux de neurones artificiels.

1. L’apprentissage automatique

1.1. Généralités

L'apprentissage automatique est un champ d'étude de l'IA qui se fonde sur des approches mathématiques et statistiques pour donner aux algorithmes la capacité d’«apprendre» à partir de données. Cette notion d’apprentissage signifie que l’algorithme est capable d'améliorer ses performances à résoudre des tâches sans être explicitement programmés pour chacune d’elles. Là où les algorithmes « classiques » ne font que suivre une suite de règles écrites par un humain, un algorithme d’apprentissage automatique est capable de s’adapter à une expérience. Il va en quelque sorte extraire des règles de cette base d’expérience et donc modifier ses propres paramètres au cours de l’apprentissage. C’est par cet ajustement automatique des paramètres que les algorithmes d’apprentissage automatique se distinguent des autres algorithmes ou les paramètres sont fixés à l’avance par l’humain. Pour que l’algorithme puisse apprendre, il faut un moyen de calculer son erreur et de la corriger pour diminuer l’erreur de la prédiction.

Les taches que résolvent les algorithmes d’apprentissage automatique dépendent de la nature des données en sortie. Nous parlerons de régression si les données de sorties sont des valeurs numériques et de classification s’il s’agit de catégoriser les données d’entrée.

1.2. Les étapes d'un projet d'apprentissage automatique

Un projet d’apprentissage automatique ne se réduit pas à un ensemble d'algorithmes. Il doit suivre une succession d'étapes, plus ou moins chronologique, que nous pouvons réduire à 6.

1) Elle commence par l'acquisition des données. Un algorithme d’apprentissage automatique se nourrit des données qu’on lui donne en entrée, cette étape est donc cruciale. La réussite du projet passera donc par la récolte en quantité suffisante des données pertinentes.

2) Vient ensuite la préparation et le nettoyage de la base de données. Très souvent, les données recueillies doivent être retouchées avant utilisation. En effet, certains attributs sont inutiles, d'autre doivent être modifiés afin d'être interprétable par l'algorithme, et enfin certains éléments sont tout simplement inutilisables car leurs données sont incomplètes. Il existe plusieurs techniques pour préparer les données telles que la vérification manuelle des données, la transformation automatique des données ou encore la normalisation.

3) La création du modèle : un modèle correspond à l'algorithme d'apprentissage qui a été mis en œuvre ainsi que le choix des hyperparamètres qui lui ont été définis. Les hyperparamètres sont l'ensemble des paramètres qui ne sont pas affectés au cours de l'entrainement de l'algorithme d'apprentissage comme, entre autre, la taille et nombre de couches cachées, le pas d'apprentissage, la taille des lots/mini-lots ou le nombre d'époques/cycles d'apprentissage. Toutes ces notions seront abordées plus loin dans ce chapitre.

4) La mise en apprentissage du modèle. Cette étape peut aussi s'appeler « phase d'entrainement ». Une base de données d'entrainement va permettre l'ajustement des paramètres de l'algorithme afin de minimiser l'erreur de la prédiction. Le calcul de l'erreur de la prédiction est effectué sur une deuxième base de données que nous appelons base de données de validation. Cette dernière permet de surveiller que l'algorithme ne sur-apprend pas. Cela implique qu'il n'apprend pas par cœur les exemples (problème de sur-apprentissage). Le sur-apprentissage sera abordé plus précisément dans la deuxième partie de ce chapitre.

5) L'évaluation de l'algorithme. Une fois l'algorithme d'apprentissage automatique entraîné, il est évalué sur un troisième ensemble de données, souvent nommée base de données de test, afin de comparer la performance de l'algorithme avec celle des autres algorithmes correspondant à l'état de l'art.

6) Le déploiement de l'algorithme. Le modèle est alors mis en production pour faire des prédictions sur des données dont nous ignorons les résultats attendus. Durant cette phase, les paramètres de l'algorithme sont alors fixes, comme pour un algorithme classique.

Certains systèmes peuvent poursuivre leur apprentissage malgré leur mise en production pour peu qu'ils aient un moyen d'obtenir un retour sur la qualité des résultats produits. L'apprentissage est alors continu. Mais dans la plupart des projets ce sont les phases

3 et 4 qui sont plusieurs fois répétées afin de tester plusieurs algorithmes d'apprentissage différents et plusieurs combinaisons d'hyperparamètres.

1.3. L'importance de la base de données

Comme le dit Jean-Claude Heudin dans *Comprendre le Deep Learning : Une introduction aux réseaux de neurones :* « Ce n'est pas forcément celui qui a le meilleur algorithme qui gagne, c'est celui qui a le plus de données ». Cette citation illustre bien pourquoi la donnée est le « pétrole » du monde numérique. En revanche, elle ne s'attarde pas sur l'importance que peut avoir le travail humain sur la pertinence et le traitement de cette donnée. En effet toutes règles générales qui sortiraient d'une base de données biaisée seraient, elles aussi, tout autant biaisées.

La qualité de l'apprentissage de l'algorithme et donc de l'analyse des données dépend alors en grande partie de la compétence de l'opérateur pour préparer l'analyse. Elle dépend aussi de la complexité du modèle, s'il est spécifique ou généraliste, et de son adéquation/adaptation à traiter le sujet. Quoi qu'il en soit, la qualité de l'apprentissage dépendra de facteurs initiaux contraignants, principalement liées à la base de données. Et la qualité de la base de données dépendra entre autres :

- Du nombre d'exemples. Plus les exemples sont rares et plus un apprentissage efficace est difficile. Mais plus les exemples sont nombreux et plus le besoin de mémoire informatique est élevé. L'apprentissage en sera par conséquent plus long.
- Du nombre et de la qualité des attributs décrivant ces exemples. L'éloignement entre deux exemples numériques (prix, taille, poids, intensité lumineuse, etc…) est facile à établir tandis que celle entre deux attributs catégoriels (beauté, utilité…) est plus subjective et délicate. Ici aussi un nombre trop faible d'attributs peut limiter la capacité d'apprentissage de l'algorithme. Mais un nombre plus élevé va engendrer un besoin de mémoire et de calcul informatique plus élevé. De plus, si un trop grand nombre d'attributs non pertinents sont pris en compte, cela peut limiter la vitesse d'apprentissage ainsi qu'augmenter le risque de sur-apprentissage.
- Du pourcentage de données renseignées par rapport aux quantités de données manquantes. Cela jouera sur le biais de représentativité des données, il peut donc

être pertinent de réduire la quantité de données d'une base de donnée afin de la rendre plus équilibré.

- De la qualité des données. Le nombre et la répartition des valeurs douteuses (erreurs potentielles, valeurs aberrantes…) aura une influence sur la qualité de la généralisation de l'algorithme et donc sur l'efficacité de l'apprentissage.

Le travail à fournir sur la collecte des données brutes est donc primordiale. Mais un travail tout aussi primordiale doit être effectué sur la transformation des données pour qu'elles soient utilisables en tant que données d'entrée interprétables par l'algorithme. Par exemple les algorithmes ont très souvent besoin de données d'entrée de taille fixe.

1.4. Différentes catégories d'apprentissage

Les algorithmes d'apprentissage peuvent se trier en différentes catégories selon la manière/méthode dont nous décidons de les faire apprendre. La méthode utilisée sera principalement dépendante du problème qui doit être résolu ainsi que des données disponibles pour résoudre le problème posé. Ces catégories ne sont pas hermétiques entre elles. Certaines méthodes peuvent être des hybrides de plusieurs catégorie et il n'est pas rare qu'un algorithme face appel à plusieurs méthode différentes pour arriver à ces fins. Bien que nous puissons dénombrer un grand nombre de méthode d'apprentissage, voici la liste des cinq méthodes les plus couramment utilisées :

1.4.1. Apprentissage supervisé

L'apprentissage supervisé est une méthode d'apprentissage automatique consistant à apprendre une fonction de prédiction à partir d'exemples étiquetés. Nous distinguons ici les problèmes de régression des problèmes de classification. Ainsi, nous considèrons que les problèmes de prédiction d'une variable quantitative sont des problèmes de régression tandis que les problèmes de prédiction d'une variable qualitative sont des problèmes de classification. Les exemples étiquetés peuvent alors constituer une base d'apprentissage, et la fonction de prédiction apprise peut-être appelée « hypothèse ». Nous supposons alors que cette base d'apprentissage est représentative d'une population d'échantillons plus large. Le but des méthodes d'apprentissage supervisé est donc de correctement généraliser les règles qui sous-

tendent l'hypothèse. Cela doit aboutir à une fonction qui fasse des prédictions correctes sur des données non présentes dans la base de données d'apprentissage.

1.4.2. Apprentissage non supervisé

Contrairement à la l'apprentissage supervisé, l'apprentissage non supervisé s'utilise quand le système ou l'opérateur ne dispose que d'exemples ne possédant pas d'étiquettes. Aucun expert n'est alors requis. L'algorithme doit découvrir/extraire par lui-même la structure plus ou moins cachée des données. Le partitionnement de données (en anglais data clustering) est un type de problème pouvant faire appel à des algorithmes d'apprentissage non supervisé. L'algorithme doit ici, cibler les données selon leurs attributs disponibles afin de les classer en groupes homogènes d'exemples. Une fonction de similarité est généralement calculée selon une notion de distance entre paires d'exemples. C'est ensuite à l'opérateur d'associer/déduire ou créer du sens pour chaque groupe (14). Par exemple, un épidémiologiste veut dans un ensemble assez large de victimes de cancer du foie tenter de faire émerger des hypothèses explicatives. L'algorithme doit différencier différents groupes d'individus, que l'épidémiologiste chercherait ensuite à associer à divers facteurs explicatifs : origines géographique, génétique, habitudes ou pratiques de consommation, expositions à divers agents potentiellement ou effectivement toxiques afin d'en extraire des facteurs de risque ou de protection.

1.4.3. Apprentissage semi-supervisé

L'apprentissage semi-supervisé peut être effectué de manière probabiliste ou non. Il vise à faire apparaître une distribution sous-jacente des exemples dans leur espace de description. Il est mis en œuvre quand des données manquent d'étiquettes ou quand seulement une partie des données sont étiquetés. Le modèle peut alors utiliser des exemples étiquetés concomitamment à des exemples non étiquetés mais pouvant néanmoins renseigner sur le problème posé. A ce titre, cette méthode peut être considéré comme une méthode hybride entre l'apprentissage supervisé et non supervisé. L'intérêt de ce type d'apprentissage provient du fait que l'étiquetage de données nécessite souvent l'intervention d'un opérateur humain. Et lorsque les jeux de données sont très grands, l'étiquetage des données peut s'avérer être une opération fastidieuse. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique car il permet d'adapter le modèle

à la structure du problème (15). Ce mode d'apprentissage a par exemple été utilisé en médecine pour faire de l'aide au diagnostic ou pour choisir les moyens les moins onéreux de tests de diagnostic.

1.4.4. Apprentissage par renforcement

Dans l'apprentissage par renforcement, l'algorithme apprend un comportement à partir d'observations. L'action de l'algorithme sur l'environnement produit une valeur de retour, souvent assimilé à une récompense, qui guide l'algorithme d'apprentissage. Cette valeur de récompense peut être positive ou négative. L'algorithme cherche, au travers d'expériences itérées, un comportement décisionnel favorable dans le sens ou son but est de maximiser la somme des récompenses positives au cours du temps. C'est une des principales méthode d'apprentissage qui fut utilisé dans le cas de l'algorithme AlphaGo qui apprenait à jouer au jeu de Go en jouant des parties contre lui-même (16).

1.4.5. Apprentissage par transfert

Ce dernier type d'apprentissage peut être vu comme la capacité d'un algorithme à reconnaître et appliquer ses « connaissances » et « compétences » apprises à partir de tâches antérieures sur de nouvelles tâches ou domaines. Souvent, cette opération n'est possible que lorsque les problèmes à résoudre partageant des similitudes. Toute la difficulté consiste à identifier les similitudes entre la ou les tâche(s) source(s) et la ou les tâche(s) cible(s), pour ensuite pouvoir transférer la connaissance de l'une à l'autre (17). La contrainte principale étant que les données d'entrées d'un système à l'autre doivent être de même nature ou du moins facilement interchangeable pour être interprétable d'un domaine à l'autre. L'apprentissage par transfert peut aussi s'opérer dans le but de renforcer ou accélérer l'apprentissage d'un algorithme en le fusionnant avec d'autre algorithme ayant déjà acquis de l'expérience sur une tâche similaire. Ce type de transfert est relativement bien utilisé pour les réseaux de neurones.

1.5. Différents types d’algorithmes

Il existe une grande diversité d’algorithmes d’apprentissage automatique, le but ici n’est pas d’être exhaustif mais d’énumérer les plus connus et les plus couramment utilisés.

- Les machines à vecteur de support (SVM) très efficaces dans les problèmes de classification ;
- La méthode des k plus proches voisins (k-NN) pour un apprentissage supervisé ;
- Les arbres de décision, méthodes à l'origine des forêts d’arbres aléatoires et par extension également du boosting ;
- Les méthodes statistiques comme le modèle de mixture gaussienne (GMM) ;
- La régression logistique ;
- L'analyse discriminante linéaire (ADL) ;
- Les algorithmes génétiques et la programmation génétique.
- Les réseaux de neurones, dont les méthodes d'apprentissage en profondeur (deep learning en anglais) peuvent s’appliquer à tous les types d’apprentissage (supervisé, non-supervisé etc…). Ils se sont imposés comme la méthode de référence pour résoudre la plupart des problèmes qui s’appliquent à l’apprentissage automatique dès lors que nous possèdons une base de données suffisante. C’est donc celle-ci qui sera développé dans cette thèse.

2. Les réseaux de neurones artificiels

Un réseau de neurones artificiels est l’association, en un graphe plus ou moins complexe, d’objets élémentaires que sont les neurones artificiels. Ces neurones artificiels sont aussi appelés neurones formels. Les réseaux de neurones artificiels se distinguent entre eux par l’organisation de leur graphe. Cela implique le nombre de neurones, la manière dont ils sont connectés les uns des autres et la présence ou non de boucles dans le réseau. Comme son nom l’indique, le neurone artificiel tire l’origine de sa conception à la compréhension du neurone biologique. Le neurone biologique est une cellule eucaryote se caractérisant par la présence :

• des synapses (connexions avec les autres neurones),

• des dendrites (entrées du neurone),

• des axones (les sorties du neurone),

• et d’un noyau qui active les sorties en fonction des stimulations en entrée.

Un neurone artificiel peut se résumer en une simple représentation mathématique et informatique de ce qu’est un neurone biologique. Le neurone artificiel possède une ou plusieurs entrées ainsi qu’une ou plusieurs sorties qui correspondent respectivement aux dendrites et aux axones. Bien que les réseaux de neurones artificiels soient d’inspiration biologique, il aura fallu plusieurs décennies de recherche fondamentale et appliquée pour passer de la conception mathématique d’un neurone artificiel à des réseaux de neurones complexes et performants tels qu’on les connait aujourd’hui.

2.1. Du neurone artificiel à l’apprentissage profond

2.1.1. Le neurone artificiel

On entend parler pour la première fois du concept de neurone artificiel en 1943 lorsque Warren McCulloch et Walter Pitts publient leur modèle mathématique et informatique du neurone biologique : le neurone formel. Son fonctionnement est simple, le neurone active sa sortie (sortie active = 1) si ses entrées dépassent un certain seuil (18). A ce stade le neurone formel est directement inspiré du neurone biologique mais est plus une invention conceptuelle qu’un réel outil applicable à la résolution de problème complexe.

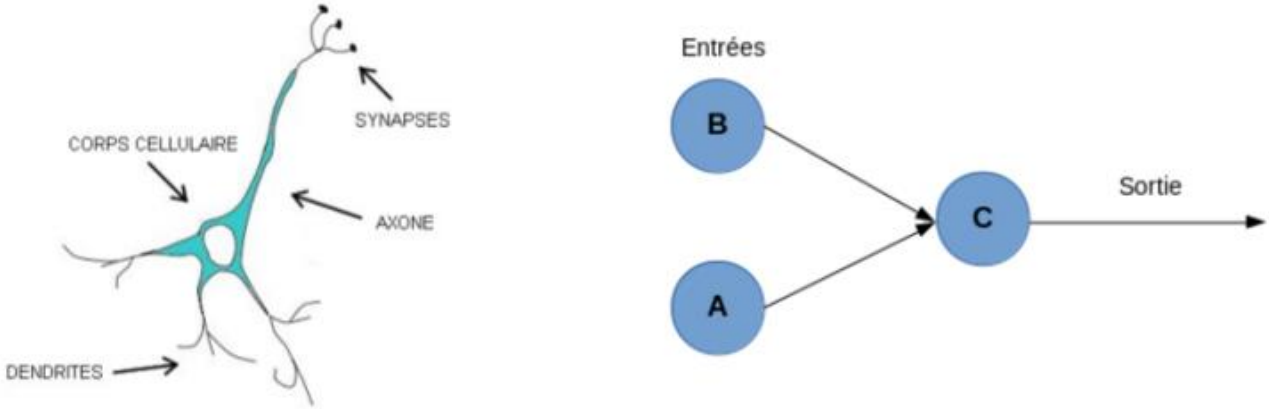


Figure 12 : À gauche le schéma d'un neurone biologique et à droite le schéma du neurone formel de 1943.

2.1.2. Le perceptron

Ce n'est qu'en 1957 que Frank Rosenblatt met au point le perceptron. Le perceptron peut être vu comme le type de réseau de neurones le plus simple. C'est un classifieur linéaire (19). L'apprentissage d'un perceptron consiste à trouver les poids (ou coefficients) du neurone qui permettent de renvoyer une valeur souhaitée et donc la bonne classe. Si le problème est linéairement séparable, un théorème assure que la règle du perceptron permettra toujours de trouver une séparation entre deux classes.

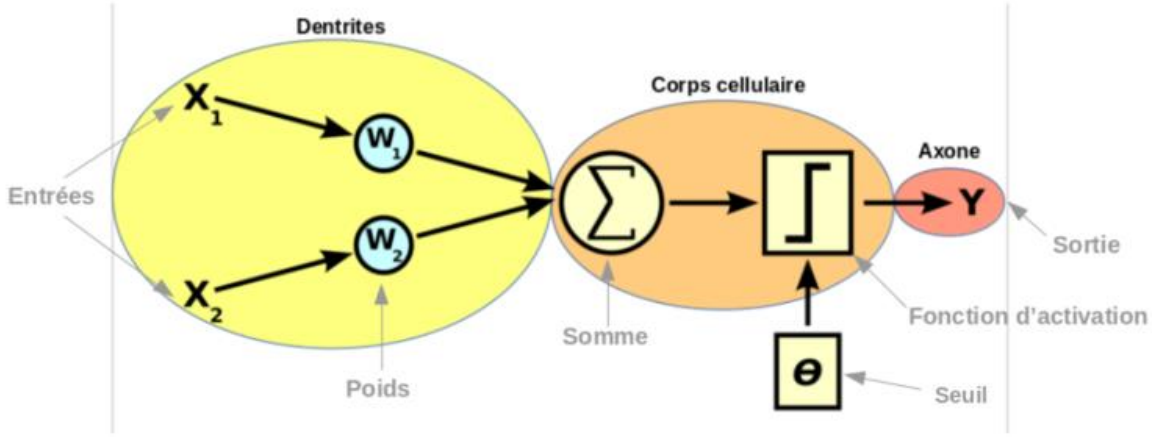


Figure 13 : Schéma représentatif d'un perceptron et comparaison avec les structures d'un neurone biologique.

Le perceptron sur le schéma ci-dessus fonctionne de la façon suivante : X_1 et X_2 sont les deux entrées (cela peut être comparé aux signaux qu'un neurone biologique reçoit par le biais des synapses d'un autre neurone). W_1 et W_2 sont des poids qui vont respectivement pondérer X_1 et X_2 (poids synaptiques). Ensuite, le symbole Σ indique une somme, il est donc effectué la somme des deux entrées X_1 et X_2 pondérées par W_1 et W_2. Le résultat de cette somme est la valeur d'entrée qui aura pour rôle de fonction d'activation (dernier cadre). La fonction d'activation est une fonction mathématique appliquée à un signal en sortie d'un neurone artificiel. Le terme de "fonction d'activation" vient de l'équivalent biologique "potentiel d'activation", seuil de stimulation qui, une fois atteint, entraîne une transmission du signale au neurone. C'est cette fonction qui détermine la sortie Y. Le symbole θ représente le seuil d'activation. Lorsque la valeur en entrée de la fonction d'activation est supérieure à ce seuil le neurone est actif (la sortie est égale à 1), lorsque cette valeur est inférieure à ce seuil il est non-actif (la sortie vaut alors 0 ou -1). Quand la valeur en entrée de la fonction d'activation est aux alentours du seuil, cette phase est appelée la phase de transition.

Historiquement, les premiers perceptrons utilisaient la fonction de seuil. Cette dernière permettait d'obtenir une prédiction binaire de type "oui" ou "non". Néanmoins, par son gradient nul en tous points, elle ne permet pas d'entraîner le réseau par l'algorithme de descente de gradient que nous verrons dans la partie 2.2.1. Il a donc fallu la remplacer par des fonctions dont les gradients sont non nuls. Les fonctions Sigmoïde et Tanh (Tangente hyperbolique) ont un comportement semblable à la fonction de seuil, mais avec un gradient non nul et une transition plus lisse (figure 14). La principale différence entre ces deux fonctions étant que la fonction Sigmoïde ramène les valeurs d'entrées entre 0 et 1, alors que la fonction Tanh les ramène entre −1 et 1. Bien que ces dernières aient été très utilisées, en 2011 l'introduction d'une nouvelle fonction nommée ReLU (Rectified Linear Unit) (20) les a quasiment toutes remplacées car elle permet un entraînement plus rapide des réseaux de neurones .

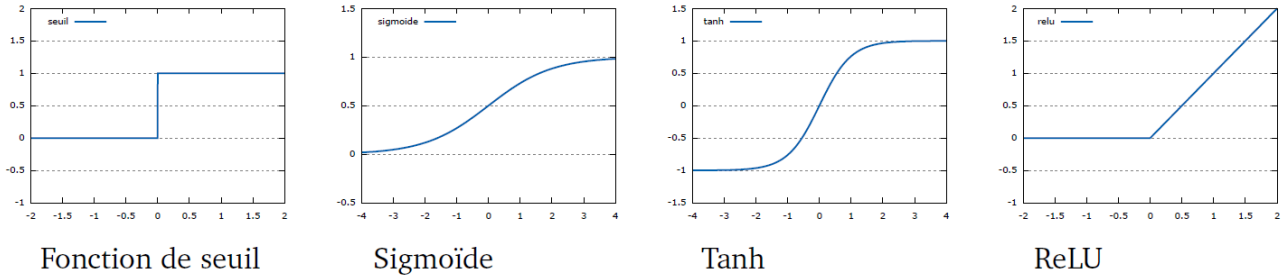


Figure 14 : Les fonctions d'activations les plus connues. De gauche à droite : la fonction de seuil, Sigmoïde, Tanh et ReLU.

2.1.3. Le perceptron multicouche

Les limites techniques de l'algorithme du perceptron ont vite été atteintes. En effet, un perceptron à une seule couche ne peut séparer les classes que si elles sont séparables de façon linéaire. Dans le cas d'un classifieur linéaire, les données d'entraînement doivent être classifiées en catégories correspondantes de sorte que si des classifications sont appliquées à deux catégories, toutes les données d'entraînement doivent être rangées dans ces deux catégories. Il ne peut y avoir que deux catégories de classification, il s'agit donc d'un classifieur binaire. En 1969 Marvin Lee Minsky et Seymour Papert publièrent un ouvrage qui porta un coup dur à la communauté scientifique gravitant autour des réseaux de neurones. En effet, leur livre mettait en exergue les limitations théoriques du perceptron, et plus généralement des classifieurs linéaires, notamment l'impossibilité de traiter des problèmes non linéaires ou de connexité. Ces limitations qui ne s'appliquaient qu'au modèle de perceptron monocouche

furent implicitement étendu à tous modèles de réseaux de neurones artificiels (21). Ce champ de recherche semblait alors dans une impasse. La recherche sur les réseaux de neurones artificiels perdit donc une grande partie de son attrait et par conséquence perdit aussi une grande partie des financements publics et privés.

C’est en 1986 que toutes ces limites tombèrent grâce à une nouvelle génération de réseaux de neurones, capables de traiter avec succès des problèmes de classifications non linéaires. Il s’agissait du perceptron multicouche. Le concept repose sur le fait qu’en combinant plusieurs couches de neurones, chacune capable de réaliser des opérations simples, il est possible de résoudre des problèmes complexes. Pour ce faire, les neurones artificiels d'une couche sont reliés à la totalité des neurones artificiels des couches adjacentes. Ces liaisons sont soumises à un coefficient qui altère l'effet de l'information sur le neurone de destination (22). Ainsi, le poids de chacune de ces liaisons est l'élément clef du fonctionnement du réseau. L’apprentissage d'un perceptron multicouche passe donc par la détermination des meilleurs coefficients/poids, applicables à chacune des connexions inter-neuronales, pour résoudre un problème. Il a été montré qu’un réseau de neurones multicouches peut théoriquement approximer n’importe quelle fonction si ce dernier possède un nombre de couches et de neurones suffisamment élevé. Le nombre d’entrées d’un perceptron multicouches est déterminé par le type des entrées et le nombre de sorties par la tâche étudiée. Les couches au centre du réseau sont appelées couches cachées, car elles permettent d’apprendre des représentations intermédiaires propres au réseau dont il est difficile d’avoir connaissance.

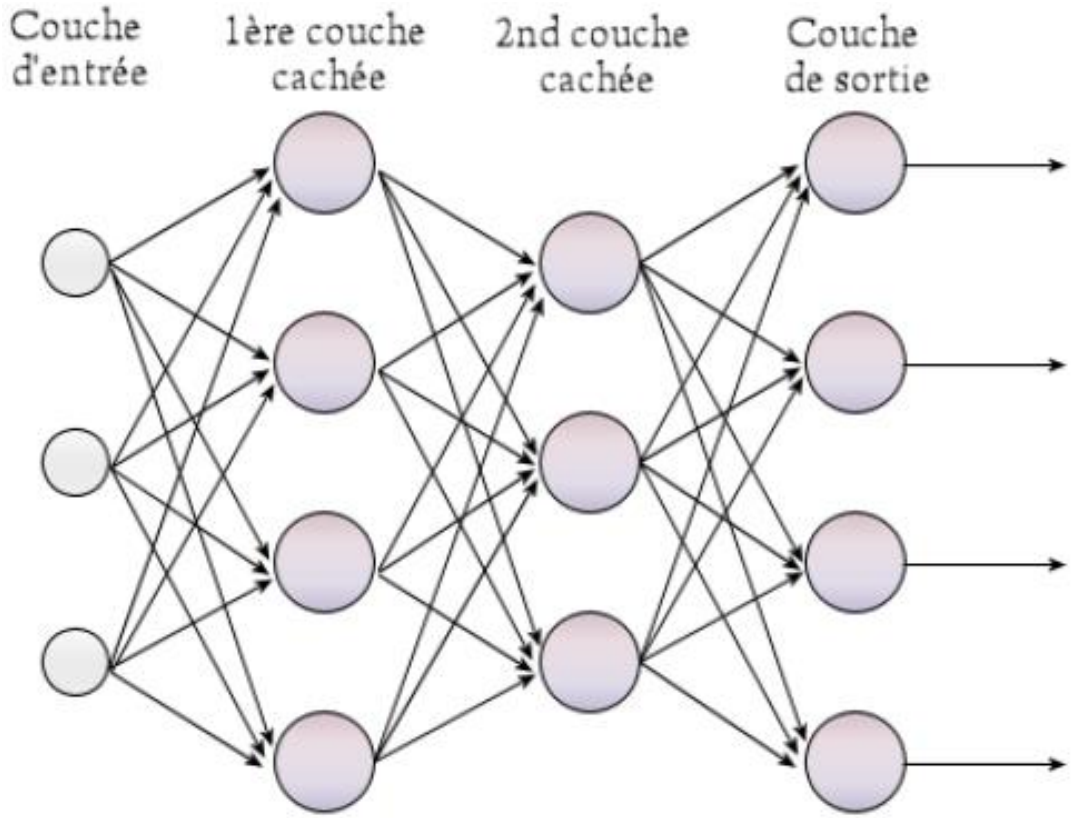


Figure 15 : Schéma d'un perceptron multicouche unidirectionnel. Chaque perceptron est connecté à l'ensemble des perceptrons de la couche suivante et précédente.

Le fait d'ajouter des couches à un perceptron rend l'apprentissage de celui-ci plus difficile car cela consiste à trouver les valeurs des poids de l'ensemble des connexions constituant le réseau (ce qui peut potentiellement faire un nombre de poids très important à trouver). C'est grâce à un algorithme appelé rétro-propagation du gradient que ces difficultés purent être surmontées (23).

Cependant, il s'est rapidement révélé que, pour résoudre des problèmes plus complexes, il était nécessaire de multiplier les couches dans les réseaux de neurones. Cette complexification des réseaux de neurones entraîne deux difficultés.

- Plus le réseau est profond (autrement dit, plus il possède un grand nombre de couches) et plus il faut de puissance de calcul pour pouvoir faire le faire apprendre et l'utiliser.

- Mais aussi, plus le réseau est profond et moins l'algorithme d'apprentissage (la rétro-propagation du gradient) fonctionne correctement. C'est ce que bous appelons la perte du gradient.

Ces deux difficultés que sont les limitations matérielles de l'époque et la notion de perte du gradient, ont limité l'utilisation des réseaux de neurones pendant près de deux décennies.

2.1.4. Apprentissage profond et réseau de neurones actuel

L'apprentissage profond ou apprentissage en profondeur (en anglais : deep learning) consiste à multiplier le nombre de couches cachées de neurones artificiels. Usuellement il est considéré qu'au-delà de 2 couches cachées nous avons affaire à de l'apprentissage en profondeur. Le concept d'apprentissage profond se matérialise à partir des années 2010 grâce à la convergence de plusieurs facteurs :

- L'arrivée à maturité des bases théoriques et techniques vues précédemment, dont en particulier les neurones artificiels multicouches.
- L'apparition de machine dont la puissance de calcul permet de traiter une quantité massive de données. Nous pouvons souligner en particulier, l'émergence de l'utilisation des cartes graphiques dans les calculs parallèles. Cette émergence a été démocratisée par la technologie CUDA. C'est cette même technologie qui permet de faire tourner des simulations de dynamique moléculaire sur carte graphique. L'usage de CUDA pour les réseaux de neurones artificiels fut popularisé par l'équipe de Krizhevsky en 2012 avec le réseau AlexNet lors de la compétition ImageNet (2).
- Des bases de données suffisamment grandes qui ont pu émerger grâce au développement des technologies de l'information. Car l'un des points limitants lors de tout apprentissage aujourd'hui reste la grande quantité d'exemples nécessaires à une généralisation correcte des règles par le réseau.

Aujourd'hui, il existe de nombreuses architectures neuronales allant au-delà du « simple » réseau de neurones multicouches. La dernière partie de ce chapitre présente l'architecture des réseaux utilisée dans cette thèse : les réseaux de neurones convolutifs (RNC) (que nous verrons souvent dans la littérature abrégé par CNN pour Convolutional

Neural Network) (24). Mais il existe de nombreux autres types d'architectures non abordées dans cette thèse telles que les réseaux de neurones récurrents, bidirectionnels, auto-organisés, auto-encodeurs, avec mécanisme d'attention, les réseaux de croyances profondes (Deep Belief Network) (25) ou encore les machines de Boltzmann restreintes (Restricted Boltzmann Machine) (26).

2.2. L'apprentissage d'un réseau de neurones

2.2.1. Rétro-propagation et descente de gradient

Entraîner un réseau de neurones consiste mettre à jour des poids du réseau jusqu'à ce que l'erreur de prédiction soit minime. Pour cela, il est définit une fonction de perte (Loss function en anglais) qui permet de calculer l'erreur faite par rapport aux valeurs cibles (27). Nous utilisons ensuite un algorithme de descente de gradient qui ajustera les poids du réseau afin de diminuer l'erreur de prédiction (28). Cependant pour que l'algorithme de descente de gradient fonctionne, il a besoin de connaître le gradient de l'erreur pour chacun des paramètres du réseau. Ce calcul du gradient s'effectue grâce à l'algorithme de rétro-propagation (back-propagation en anglais) développé en 1986 par Rumelhart (23). Il s'occupe de calculer le gradient de l'erreur pour chaque paramètre du réseau, de la dernière couche vers la première. Pour cela il fonction en deux temps :

1) calculer l'erreur en comparant la sortie avec le résultat attendu,

2) propager l'erreur de couche en couche vers l'arrière.

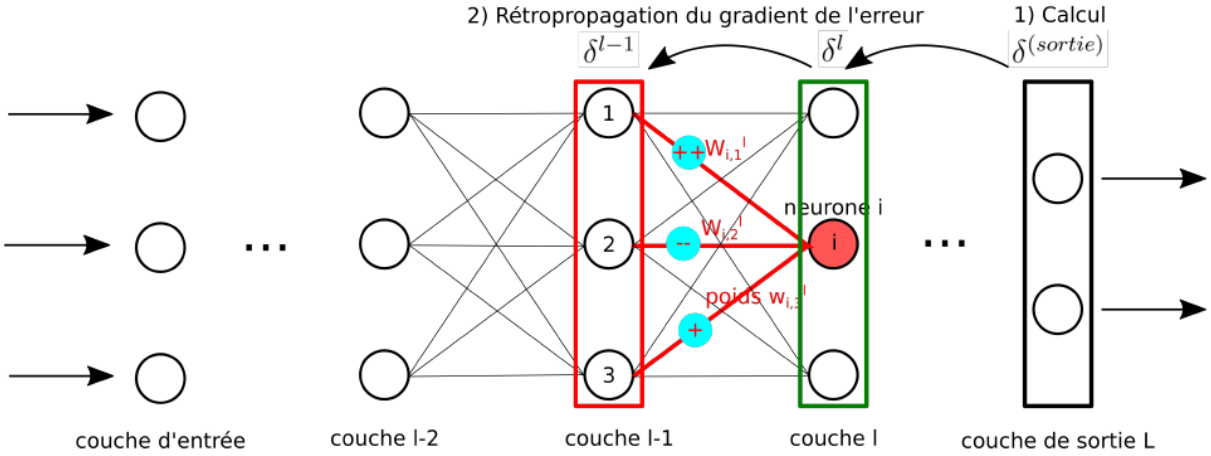
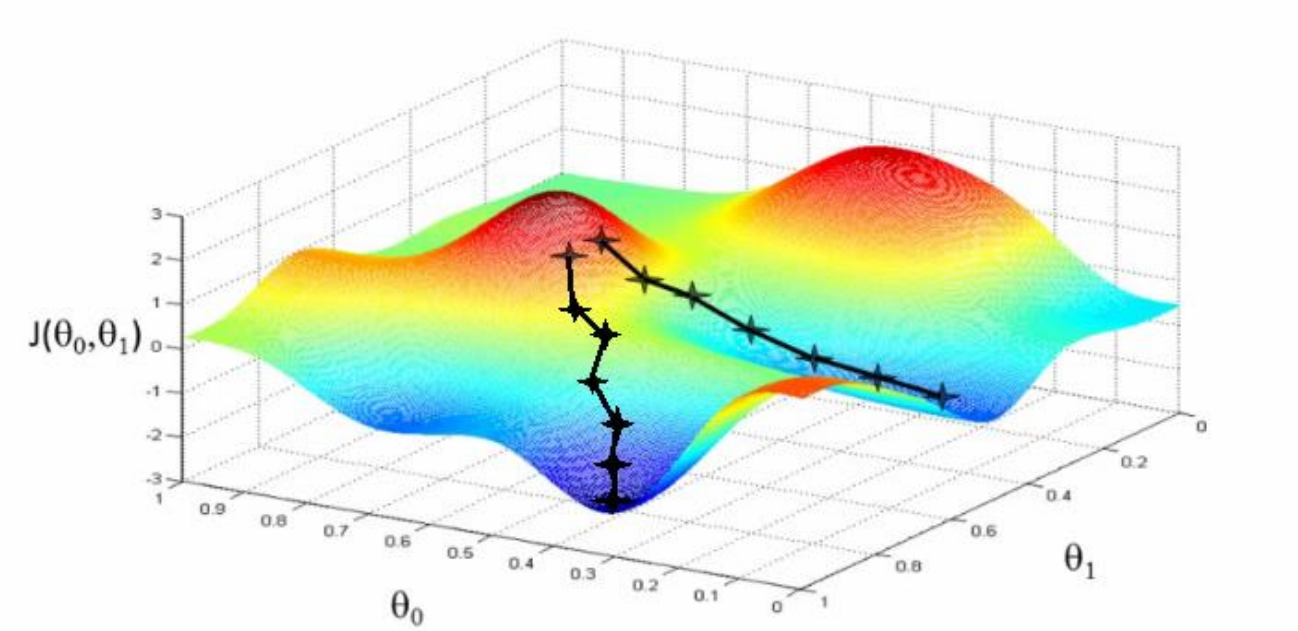


Figure 16 : Schéma de la rétro-propagation du gradient de l'erreur depuis la couche de sortie jusqu'aux poids des connexions synaptiques du neurone « i » de la couche « l » (vert) issues des 3 neurones de la couche précédente « l - 1 » (en rouge).

L’algorithme de descente du gradient utilise donc les résultats de l’algorithme de rétro-propagation pour mettre à jour les poids. Ce double processus est itératif et



procède par améliorations successives. Le but étant de s’approcher du minimum global, donc d’une combinaison de poids où l’erreur de prédiction est la plus faible. Au cours d’une descente de gradient, les erreurs sont corrigées en fonction du degré d’importance des neurones à participer à l’erreur finale. Les poids synaptiques qui contribuent à engendrer une erreur trop importante dans le résultat finale sont modifiés de manière plus significative que les poids qui engendrent une erreur marginale. Les poids sont ainsi soit diminués, soit augmentés, et le modèle est mis à jour à chaque cycle. Le gradient de l’erreur donnant le sens de la pente, les poids sont modifiés de tel sorte qu’ils descendent cette pente afin de diminuer l’erreur. Un hyperparamètre, appelé taux d’apprentissage, détermine la longueur du pas à faire dans la direction du gradient. La taille du pas détermine donc l’ampleur de la correction. Une fois les paramètres (poids) mis à jour, le nouveau gradient est calculé (par retro-propagation) puis l’algorithme se dirige à nouveau dans la direction opposée de le pente (par descente de gradient). Nous minimisons ainsi l’erreur en avançant pas à pas dans la direction de la pente (itération après itération).

Figure 17 : Illustration de la descente de gradient sur une fonction de potentiel J arbitraire qui dépend de deux variables, θ_0 et θ_1. La descente se fait des sommets

(rouge) vers les creux (bleu). Le pas d'apprentissage est représenté par une droite entre deux croix.

2.2.2. Optimisation des calculs du gradient et optimiseur de la descente de gradient

Il existe un paramètre qui joue sur la vitesse d'apprentissage ainsi que sur la capacité qu'a le réseau à apprendre. Ce paramètre est le pas d'apprentissage, il détermine l'amplitude de la variabilité des poids entre deux mises à jour. Un faible pas d'apprentissage provoque des faibles variations de poids ce qui rend l'apprentissage plus lent qu'avec un grand pas d'apprentissage. Un faible pas d'apprentissage peu garantir une certaine stabilité dans la progression de l'apprentissage du réseau mais cela peut aussi empêcher le réseau d'atteindre les poids optimums en bloquant le processus de descente de gradient dans des minimum locaux.

Il existe trois variantes connues pour calculer le gradient de l'erreur qui diffèrent en fonction du nombre de données utilisées :

- La première, appelée descente de gradient par lot (batch en anglais) utilise toutes les données de la base d'entraînement pour calculer l'erreur globale. Ainsi, le gradient donne la direction exacte de la pente permettant de minimiser l'erreur globale. Si cette dernière présente l'avantage de calculer le gradient exacte (et donc la pente exacte) pour minimiser l'erreur sur l'ensemble des données d'entraînement, elle est très lourde à mettre en place. En effet, cela implique que, pour chaque pas, il est nécessaire d'évaluer l'ensemble des données d'entraînement afin d'obtenir l'erreur de chacun des exemples. Or si la base de données contient énormément de données cela peut être très long voir irréalisable. Or plus la taille du lot (donc la base de données) est grande, et plus la mémoire informatique nécessaire pour traiter les calculs est grande. Tant et si bien qu'il ne sera presque jamais possible d'utiliser cette méthode.
- Une deuxième variante consiste donc à ne calculer le gradient que sur une donnée à la fois. Cette variante est appelée descente de gradient stochastique (SGD = Stochastic Gradient Descent), car les exemples d'entrainement sont choisis de manière aléatoire. Elle consiste à mettre à jour les poids après chaque exemple ce qui permet de ne pas saturer la mémoire informatique. Le défaut étant que le gradient est calculé à chaque étape et ne correspond donc pas au véritable

gradient global. Néanmoins, minimiser les erreurs des exemples un-à-un permet une convergence moyenne vers un minimum local. Il est possible qu'un pas soit fait dans la mauvaise direction, mais la moyenne des pas effectués ira dans la bonne direction. En effet, un exemple d'entraînement compliqué (ou faux) dans la base de données n'aura une influence sur le gradient que lorsque ce dernier sera considéré, ce qui entraînera un pas dans une mauvaise direction, mais qui sera corrigé par les exemples suivants. Cela a pour effet de permettre une meilleure généralisation en limitant l'effet des données parasites mais demande beaucoup de calcul à cause du grand nombre d'itérations.

- Néanmoins, même si la direction globale des gradients permet de se diriger vers un optimum, les gradients individuels de chacun des exemples présentent de grosses variations, entraînant souvent l'algorithme à changer de directions. Par conséquent, une troisième variante de l'algorithme de descente de gradient présente un juste milieu entre la descente de gradient par lot et la SGD. L'idée est de calculer le gradient pour un sous-ensemble de données, tirées aléatoirement dans la base de données. Ce sous-ensemble est appelé mini-lot en conséquence l'algorithme correspondant est appelé descente de gradient par mini-lot. L'utilisation d'un mini-lot pour calculer l'erreur permet d'obtenir un gradient plus proche du véritable gradient global tout en gardant l'avantage de ne pas évaluer l'ensemble des données pour chaque pas de l'algorithme. La taille du mini-lot sera donc un équilibre à trouver entre occupation de mémoire informatique et vitesse de calcul.

La descente de gradient par mini-lot est aujourd'hui devenue la norme pour l'entraînement des réseaux de neurones tant et si bien que par abus de langage nous entendrons souvent parler de SGD même lorsqu'il s'agit de descente de gradient par mini-lot. Cependant, elle garde le défaut (moins prononcé) de faire des pas dans la mauvaise direction, ralentissant ainsi l'entraînement des réseaux. Pour pallier ce problème, plusieurs améliorations ont été développées, permettant à l'algorithme de converger plus vite vers un minimum local. Ces algorithmes, appelés optimiseurs, consistent à estimer la direction optimale du gradient basée sur le gradient obtenu à chaque cycle avec l'erreur faite sur le mini-lot.

Une première amélioration possible qui permet d'estimer une direction globale est d'ajouter une inertie dans le gradient (momentum en anglais). Cette technique consiste à

combiner le gradient obtenu à l'étape t avec l'estimation obtenue à l'étape t − 1. L'inertie ajoutée au calcul du gradient permet d'éviter les changements de directions trop importants et donc d'éviter aux paramètres de trop zigzaguer (29). Cette inertie fonctionne à l'image d'une balle qui roule le long d'une pente et qui prend de la vitesse lors de la descente. Le pas d'apprentissage augmente lorsque les gradients successifs pointent dans la même direction. Il diminue lorsque les gradients ont des directions différentes.

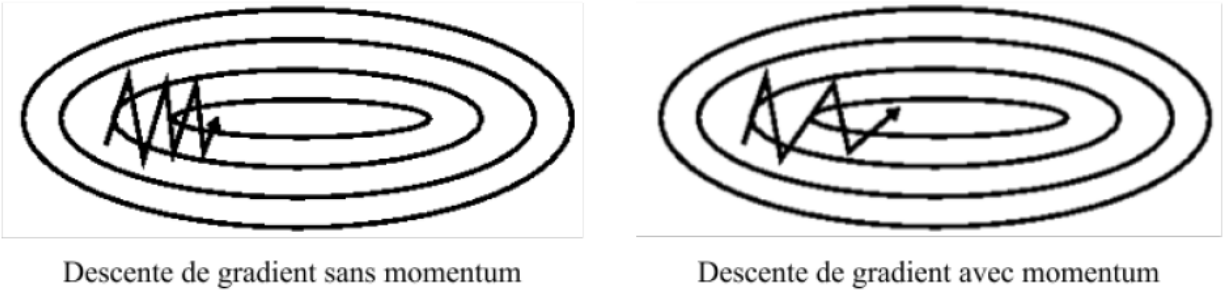
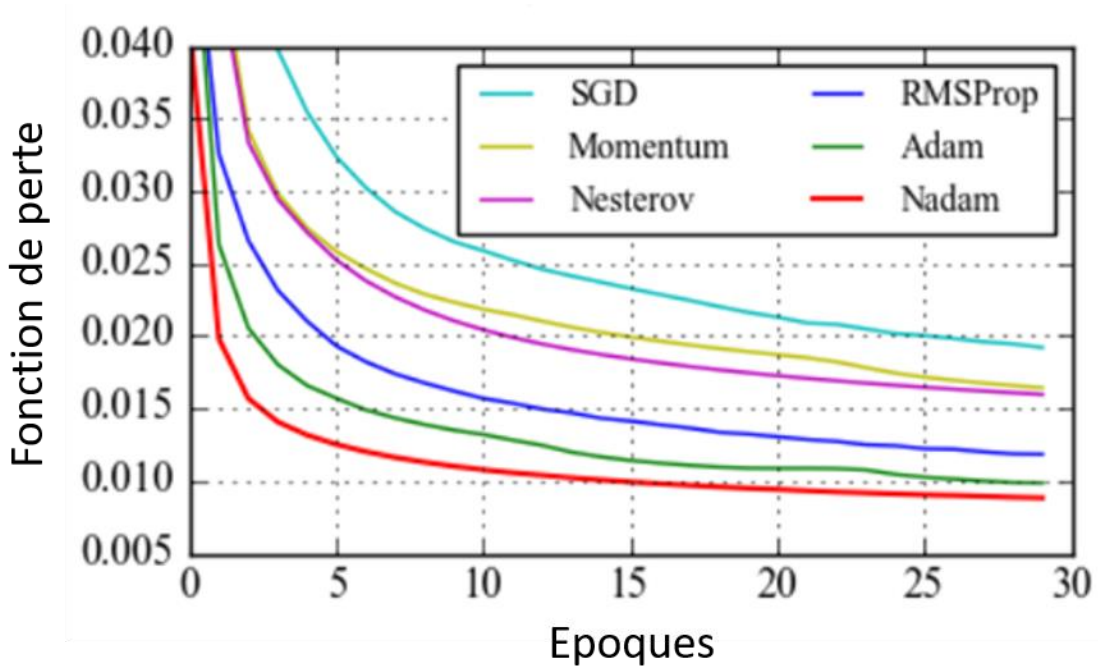


Figure 18 : Schéma de la descente de gradient avec momentum (droite) et sans momentum (gauche). Sans momentum, la descente de gradient progresse lentement car il oscille beaucoup d'une pente à l'autre. Avec momentum, la convergence s'accélère car les oscillations sont amorties. Tirée du papier Srivastava et al** *(35))***

Basés sur cette idée d'inertie de nombreux autres optimiseurs présentent des différences plus ou moins importantes. Nous pouvons par exemple présenter l'optimiseur NAG (Nesterov accelerated gradient) qui calcule le gradient après avoir avancé d'un pas dans la direction précédente (30), AdaGrad (Adaptive Gradient), qui adapte le pas d'apprentissage pour chacun des paramètres indépendamment (31), RMSProp (Root Mean Square



Propagation) qui est une amélioration d'AdaGrad et qui permet d'éviter des changements trop brutaux dans les pas d'apprentissage des paramètres (32), Adam (Adaptive moment estimation) qui peut être vu comme une combinaison entre RMSProp et de Adagrad, ou encore Nadam (Nesterov-accelerated Adaptive Moment Estimation) qui essaie d'améliorer la stabilité d'Adam en le combinant avec NAG (33). La figure 19 présente une comparaison de la performance de différents optimiseurs.

Figure 19 : Comparaisons de la performance de différents optimiseurs sur un réseau de neurone convolutif autoencoder (tirée de la publication de la méthode Nadam** (33)**).

Le problème avec tous ces optimiseurs est qu'il est souvent difficile de les comparer car ils ont des comportements différents en fonction des données utilisées. Par conséquent, il n'existe pas de règles précises sur l'utilisation d'un optimiseur plutôt qu'un autre. Néanmoins, il est possible d'en privilégier certains par rapport à l'objectif recherché. Par exemple, si l'objectif est de tester ou d'optimiser une nouvelle architecture, il peut être intéressant d'utiliser un optimiseur ayant peu d'hyperparamètres. Ainsi une bonne optimisation de l'optimiseur sera plus facile et il sera plus aisé de vérifier le fonctionnement de l'architecture testée. Au contraire, si l'objectif est d'entraîner une architecture ayant déjà fait ses preuves, il peut être intéressant d'utiliser un optimiseur plus complexe permettant ainsi une finesse plus précise dans les hyperparamètres et donc une optimisation plus précise. Malgré le grand nombre d'hyperparamètre que possède des optimiseurs récents comme l'Adam et le Nadam leurs performances dépassent souvent celui des optimiseurs plus anciens sans pour autant avoir à modifier leurs hyperparamètres.

Le processus d'apprentissage est donc un processus itératif dans lequel se succèdent plusieurs cycles de rétro-propagation du gradient de l'erreur puis de descente de gradient. Une itération est l'application d'un cycle (rétro-propagation du gradient de l'erreur puis de descente de gradient) sur un exemple/lot/mini-lot. Lorsque l'ensemble des exemples/lot/mini-lot de la base de données d'apprentissage sont passés par un cycle, nous parlons alors d'une époque (epoch en anglais). Ce nombre d'époques est un autre hyperparamètre à fixer. Un grand nombre d'époques présente l'avantage de laisser le système apprendre plus longtemps et de potentiellement s'améliorer davantage. En revanche cela entraîne une augmentation du temps de calcul. Une méthode alternative consiste à fixer une sorte de « patience » au système. Si le réseau ne s'est pas amélioré d'un certain seuil durant un nombre d'époques prédéfini par la patience, il s'arrêtera. En revanche, s'il trouve une amélioration significative, il réinitialise son compteur de patience. Malgré le fait qu'un réseau continue à apprendre, un nombre d'époques trop grand peut occasionner un phénomène de sur-apprentissage. La

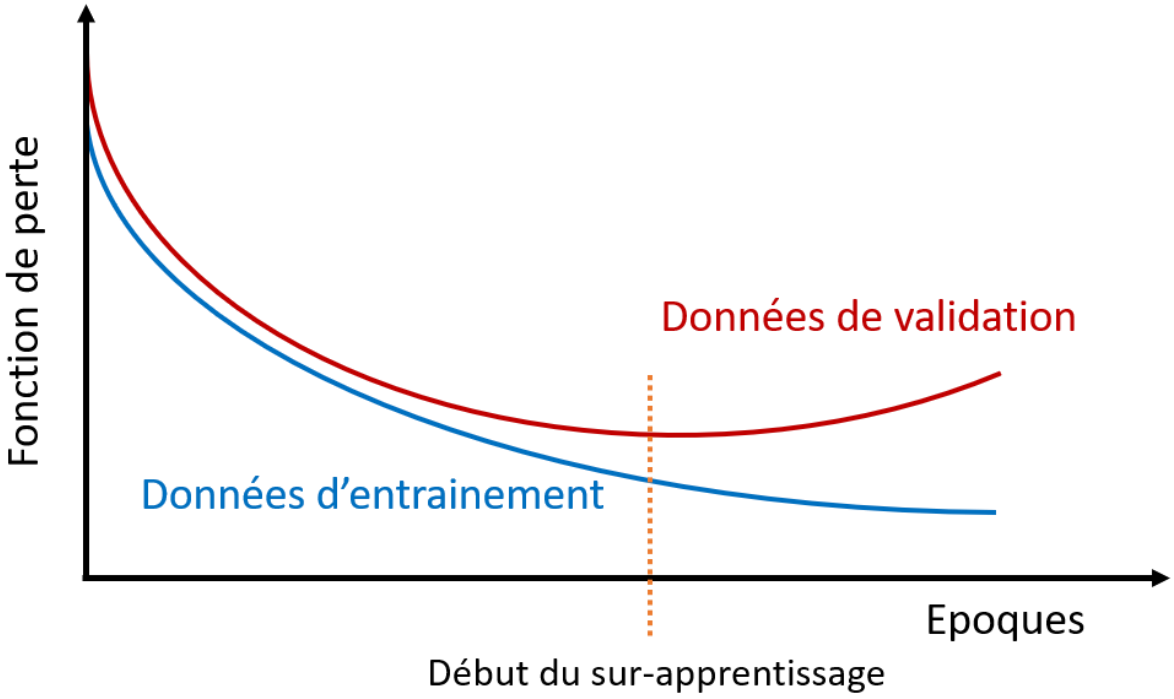
disposition d'une base de données de validation pendant l'apprentissage est donc requise. En effet, cette étape de validation pendant l'apprentissage est nécessaire pour procéder à une auto-évaluation de la performance du réseau et surveiller l'apparition ou non du phénomène de sur-apprentissage. La calibration des paramètres qui sera finalement conservée sera celle ayant obtenue le meilleur résultat sur la base de données de validation au cours de l'apprentissage.

En conclusion, de nombreux hyperparamètres sont à tester et à optimiser dans un réseau de neurones afin d'en améliorer son apprentissage : optimiseur, taille de mini-lot, taille et nombre de couches cachées, nombre d'époques, etc... Cette optimisation des hyperparamètres nécessite d'exécuter un grand nombre d'expériences, ceci afin de tester les différentes combinaisons possibles. Cependant, la quantité de combinaison possible entraîne rapidement une explosion combinatoire qui rend toute recherche exhaustive irréaliste en termes de temps d'exécution. Cela fait partie des difficultés liées à la manipulation des réseaux de neurones.

2.2.3. Le sur-apprentissage

Comme mentionné dans la section précédente, lorsqu'on entraîne un réseau de neurones nous séparons systématiquement les données d'entraînement des données de validation de sorte qu'il n'y a pas de chevauchement entre les bases de données. Cette séparation a pour objectif de simuler une situation réelle, cela implique d'une part d'avoir des données qui sont étiquetées pour l'entraînement et d'autre part des données dont les étiquettes sont inconnues du réseau pour la validation. De ce fait, il existera toujours une différence statistique entre les données d'entraînement et les données de validation. Idéalement cette différence doit être réduit au maximum, pour cela les bases de données doivent être le plus homogène possible entre elles. Néanmoins, la différence de répartition des exemples entre les données étiquetées disponible et l'ensemble des cas qui existe dans la réalité peut entraîner un problème de sur-apprentissage. En effet, le modèle entraîné peut se sur-spécialiser sur les données d'entraînement et ne plus être capable de généraliser pour inférer correctement sur les données de validation. Il apprend par cœur les exemples. Cela a pour conséquence direct de provoquer une décorrélation entre l'erreur (Loss en anglais) que commet l'algorithme sur la base de données d'apprentissage et l'erreur qu'il commet sur la base de donnée de validation. Typiquement l'erreur diminue d'époque en époque et de manière similaire entre les 2 bases de données mais lorsque le phénomène de sur-

apprentissage apparait, l'erreur sur la base de données de validation se met à augmenter (figure 20). Cela témoigne que l'algorithme commence à créer des règles trop spécifiques à la base de donnée d'apprentissage pour qu'elles se généralisent à d'autre cas et en particulier



les exemples de la base de données de validation.

Figure 20 : Illustration du phénomène de sur-apprentissage. Lorsque l'apprentissage dure trop longtemps l'augmentation de la fidélité aux données d'apprentissage (courbe bleu) détériore les capacités de généralisation de l'algorithme qui se traduit par l'augmentation de l'erreur sur les données de validation (courbe rouge).

Pour éviter ce genre de problème qui est très présent en apprentissage automatique ainsi que dans l'apprentissage en profondeur, il existe différentes solutions :

- arrêter l'entrainement avant une trop forte divergence d'erreur entre les données d'entrainement et les données de validation. Cela peut avoir pour conséquence de limiter la performance du réseau.

- Utiliser certaines couches spécifiques telles que le sous-échantillonnage ou la normalisation du lot (batch normalisation) (34).

- Faire de l'augmentation de données en créant des nouvelles données à partir des données d'entrée (déformer les données, leur ajouter du bruit, inverser les images horizontalement ou leur appliquer des rotations aléatoires etc…).

- Dégrader l'influence des poids (weight decay) en ajoutant un terme de régularisation dans leur fonction qui empêche les poids d'avoir chacun une trop grande valeur absolue. Cette méthode consiste à pénaliser (dissuader d'utiliser) des valeurs importantes pour les poids. Cela force le réseau à ne pas se focaliser sur un caractère spécifique.

- Décimer aléatoirement les poids à l'entraînement pour forcer de la redondance dans les paramètres. Cette méthode s'appelle le dropout. Le principe du dropout est simple. En phase d'entraînement, à chaque évaluation d'un nouvel exemple d'entrée, des neurones ont une certaine probabilité d'être désactivés (35) tel qu'illustré dans la figure ci-dessous.

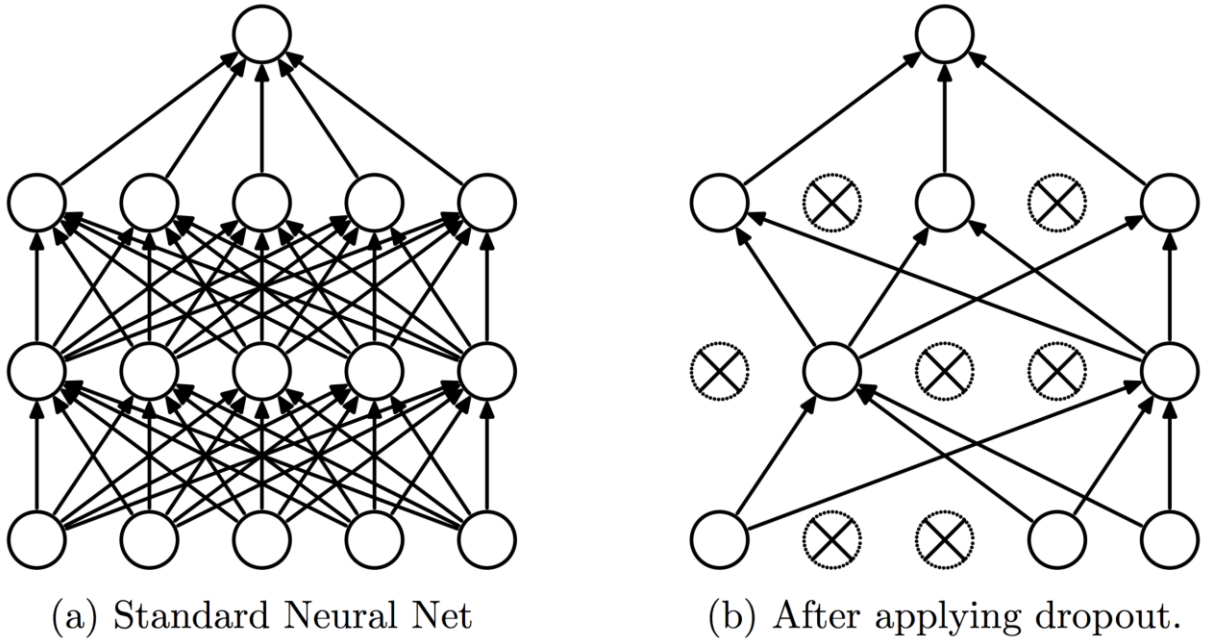


Figure 21 : Illustration de l'opérateur de Dropout (tirée du papier Srivastava et al** (35)**). À droite le réseau entièrement connecté. À gauche, des neurones ont été aléatoirement désactivés.

2.3. Les réseaux de neurones convolutifs

Dans le cadre de cette thèse, nous nous sommes intéressés au calcul de l'énergie d'interactions que les protéines peuvent avoir entre elles lorsqu'elles forment un complexe à partir de leurs informations structurales. Un des moyens de traiter ce problème en utilisant les réseaux de neurones est d'utiliser directement les images des structures 3D de ces complexes. Les réseaux de neurones convolutifs (RNC) sont particulièrement efficaces pour traiter les informations issues d'images. Ils semblent donc tout indiqués pour la résolution de ce type de problème et cela grâce à la présence d'une ou plusieurs couches de convolution caractéristiques de ce type réseau de neurones.

Le RNC est inspiré par le cortex visuel des vertébrés. Début 1962, des travaux ont montré que le cortex visuel des mammifères contient des arrangements complexes de cellules, responsables de la détection de la lumière dans les sous-régions du champ visuel. Ces arrangements qui ont la particularité de se chevaucher sont appelés champs réceptifs (36). Deux types de cellules de base sont identifiés. Les cellules simples, qui répondent à des pics caractéristiques (grand contraste, forte intensité) à l'intérieur de leur champ récepteur. Et les cellules complexes, qui ont des champs récepteurs plus grands et sont localement invariantes à la position exacte du motif. Ces cellules agissent comme des filtres locaux sur l'espace d'entrée. C'est inspiré par ces découvertes que Le Cun *et al.* développe en 1990 le réseau neuronal convolutif « LeNet » dédié spécifiquement à la classification d'images de chiffres manuscrits. Ces images ont la particularité de ne nécessiter qu'un prétraitement minimal des données (24). Contrairement à la plupart des travaux qui se faisaient jusque-alors, ce réseau reçoit directement des données à deux dimensions (2D), à savoir des images, plutôt que des données à une dimension (1D). Cela met en jeu la capacité de ces nouveaux réseaux à traiter de grandes quantités d'information de bas niveau. Il n'est donc plus nécéssaire de convertir lourdement la donnée brute via des fonctions mathématiques finement choisies qui auraient fait appel à un savoir-faire ou une expertise humaine exigeante. Ainsi, bien choisir le type d'architecture de réseau neuronal selon la tâche de prédiction à effectuer évite d'avoir à effectuer un prétraitement des données trop important et qui nécessiterait lourd travail de pré-traitement. Le réseau AlexNet dont nous avons parlé en 1er partie et qui fit grande impression lors de la compétition ImageNet de 2012 (2) est directement inspiré du réseau LeNet de Le Cun *et al*.

2.3.1. La convolution

Si nous voulons utiliser des images en entrée d'un réseau de perceptrons multicouches, il est nécessaire de créer une valeur d'entrée par pixel (mot-valise créé en contractant phonétiquement « picture element ») de l'image. Nous parlerons de voxel (mot-valise créé à l'image de pixel mais en contractant « volume element ») dans le cadre d'image en 3 dimensions. Or le nombre de voxels, dans une image 3D, peut être très grand (allant de 1000 pour de très petites images de taille 10 × 10 × 10 à 1000000 pour des images de taille 100 × 100 × 100). Cela implique que la première couche du réseau utilise de nombreux perceptrons, ce qui entraîne une augmentation importante du nombre de paramètres du réseau. De plus, en faisant cela nous n'exploitons pas les informations spatiales des voxels

(quel voxel se situe où dans l'image) et surtout le réseau ne possède pas les informations de localisation (il ne sait pas quels sont les voisins d'un voxel).

La prise en compte du voisinage d'un pixel/voxel peut se faire grâce à un opérateur mathématique, appelé opérateur de convolution. Elle consiste à déplacer un filtre en le faisant glisser sur une image et à réaliser une convolution (un produit matriciel) de ce filtre avec l'image sous-jacente. Ce filtre est une matrice carrée de poids et ces poids seront déterminés pendant l'apprentissage par la descente de gradient.

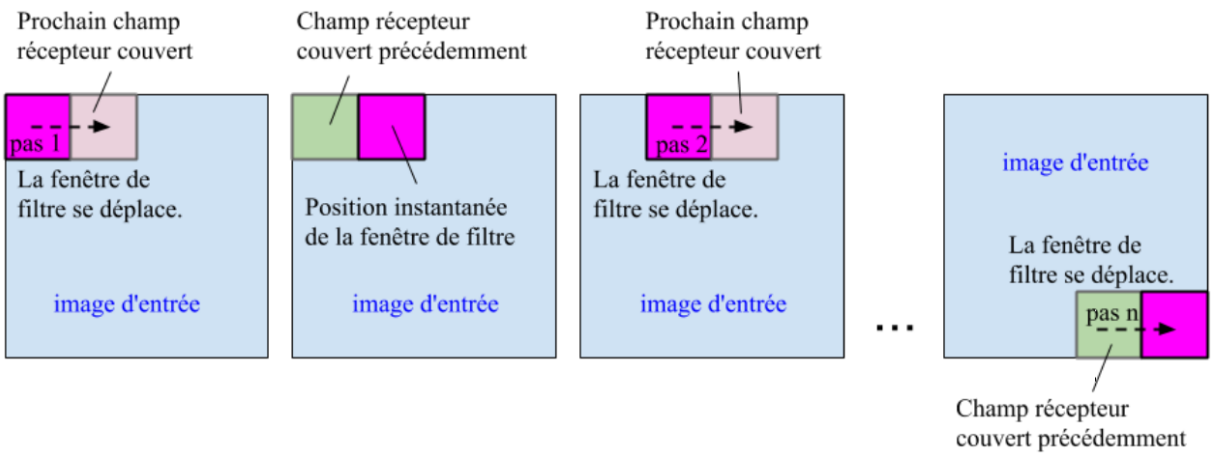


Figure 22 : Schéma du glissement de la fenêtre de filtre sur l'image d'entrée.

L'objectif de la convolution est de faire ressortir des caractéristiques (features) dans les images d'entrée. Dans ce contexte, le concept de caractéristique est assimilé au filtre. Un exemple bien connu de convolution est celui du filtre de Sobel utilisé en traitement d'image pour la détection de contours.

Figure 23 : Illustration du filtre de Sobel sur une image de test (Lenna), image servant fréquemment de référence pour les algorithmes de traitement d'image.

Dans chaque couche convolutive, chaque filtre est répliqué sur tout le champ visuel. Ces unités répliquées partagent la même paramétrisation, cela signifie que les poids sont identiques d'un champ récepteur à l'autre et forment finalement une carte de caractéristiques (feature map). Cela signifie que tous les neurones d'une même couche convolutive répondent aux mêmes caractéristiques. Cette réplication permet ainsi de détecter les caractéristiques quelle que soit leur position dans le champ visuel. Une information visuelle sera alors traitée de la même façon, quelle que soit sa localisation dans l'espace. Cette propriété porte le nom d'invariance par translation et c'est une caractéristique fondamentale des RNC. Un autre avantage de la couche de convolution est qu'une image peut être décomposée en plusieurs filtres (donc en plusieurs convolutions) à chaque couche créant ainsi plusieurs cartes de caractéristiques pour une même image pouvant faire ressortir autant de propriétés différentes.

2.3.2. Le sous-échantillonnage

Dans un réseau de neurones convolutif, les couches de convolutions sont alternées avec des couches de sous-échantillonnage (appelé pooling en anglais). Le sous-échantillonnage réduit la taille spatiale d'une image intermédiaire, réduisant ainsi la quantité de paramètres et donc de calcul à effectuer. Ce processus permet aussi de réduire l'espace mémoire ainsi que le nombre de neurones à l'entrée des parties entièrement connectées. Son rôle principal est de concentrer l'information tout en conservant les plus pertinentes (37–39).

Généralement le sous-échantillonnage a pour conséquence d'accélérer la vitesse d'apprentissage. Il est donc courant d'insérer périodiquement une couche de sous échantillonnage entre deux couches de convolution successives au sein d'une architecture de RNC.

L'opération de sous-échantillonnage consiste à choisir un représentant d'une zone spatiale en fonction d'un critère prédéfini. L'image d'entrée est découpée en une série de fenêtres de n pixels de côté ne se chevauchant pas (sous-échantillonnage). Le signal en sortie de fenêtres est défini en fonction des valeurs prises par les différents pixels de la fenêtre. Bien que les différents paramètres soient personnalisables, dans une très grande majorité des cas, la taille de la fenêtre de sous-échantillonnage ainsi que les pas sont de 2 × 2, ce qui permet de réduire le nombre de paramètre par 4 (réduction de moitié de la hauteur et de la largeur). Plusieurs stratégies d'agrégation des valeurs sont possibles. Dans la figure ci-dessous nous donnons trois exemples d'agrégation des valeurs d'entrées :

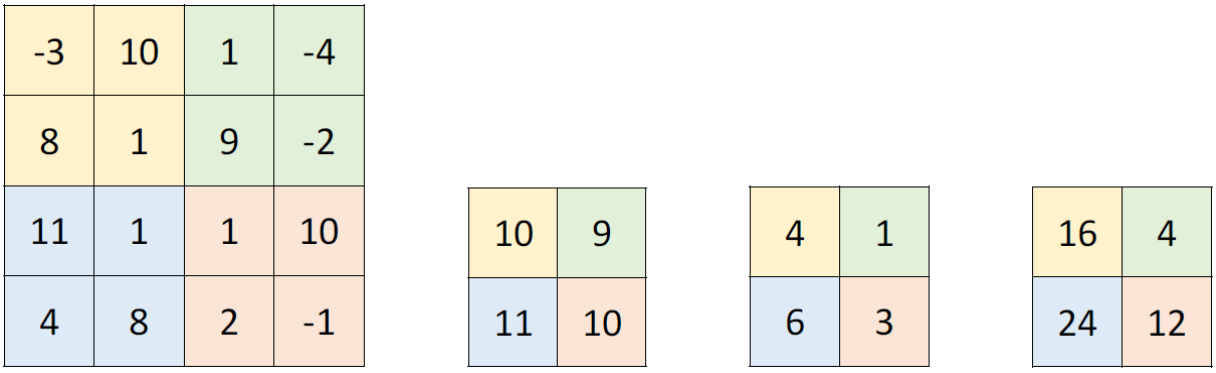


Figure 24 : Exemples de sous-échantillonnages. De gauche à droite : l'entrée, les résultats avec un opérateur de sous-échantillonnage par valeur maximale (maximum pooling), de sous-échantillonnage par valeur moyenne (average pooling) et de sous-échantillonnage par somme (sum pooling).

- Un sous-échantillonnage par valeur maximale, appelé maximum pooling, consiste à récupérer la valeur maximale dans la fenêtre d'observation.
- Un sous-échantillonnage par valeur moyenne, appelé average pooling, consiste à calculer la moyenne des valeurs dans la fenêtre d'observation.
- Un sous-échantillonnage par somme des valeurs, appelé sum pooling, consiste à calculer la somme des valeurs dans la fenêtre d'observation.

De nombreuses autres stratégies sont possibles tel que le sous-échantillonnage stochastique (stochastic pooling en anglais) (40). Néanmoins, le sous-échantillonnage le plus couramment utilisé aujourd'hui est celui par valeur maximale avec une fenêtre de taille 2 × 2 et un pas de 2 × 2. Une des conséquences directes du sous-échantillonnage est la perte d'information. Cette perte d'information peut être considéré comme minime dans le cadre d'un sous échantillonnage par valeur maximale car nous gardons le terme le plus activé par rapport au filtre précédent. De plus, cette dégradation de la source permet de limiter le sur-apprentissage en supprimant les informations non pertinentes.

3. Conclusion

Dans ce chapitre nous avons vu ce qu'est un algorithme d'apprentissage automatique. Nous nous sommes ensuite concentrés sur un type d'algorithme d'apprentissage ayant fait ses preuves dans de nombreux domaines scientifiques et industriels ces dernière années, l'apprentissage profond via les réseaux de neurones. Après avoir détaillé les différents éléments de construction d'un réseau de neurones ainsi que les algorithmes utilisés pour les entraîner, nous avons pu constater que le choix d'une architecture ainsi que de l'optimiseur et de ces hyperparamètres demande une bonne compréhension de leur fonctionnement. Nous nous sommes ensuite concentrés sur une architecture de réseau de neurones spécialisé dans l'interprétation d'image, les réseaux de neurones convolutifs. Cette architecture semble bien adaptée à l'utilisation des structures 3D des complexes protéiques dans la prédiction de leur affinité étant donné le succès qu'elle a pu avoir dans la prédiction de l'affinité de complexe protéine-ligand (41). Parmi les barrières à l'utilisation des réseaux de neurones demeure le besoin d'une grande quantité d'information, nécessaire au processus d'apprentissage, ainsi que les besoins intensifs en capacité de calcul informatique. Un des objectifs de cette thèse sera alors d'étudier la viabilité de cette approche dans cette problématique spécifique qu'est la prédiction par méthode bioinformatique de l'énergie d'interaction d'un complexe protéique.

BIBLIOGRAPHIE DU CHAPITRE 3

1. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 1 déc 2015;115(3):211‑52.

2. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, éditeurs. Advances in Neural Information Processing Systems 25 [Internet]. Curran Associates, Inc.; 2012. p. 1097–1105. Disponible sur: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

3. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature. janv 2016;529(7587):484‑9.

4. The Mystery of Go, the Ancient Game That Computers Still Can't Win [Internet]. WIRED. [cité 27 juill 2020]. Disponible sur: https://www.wired.com/2014/05/the-world-of-computer-go/

5. Cireşan D, Meier U, Masci J, Schmidhuber J. Multi-column deep neural network for traffic sign classification. Neural Netw. août 2012;32:333‑8.

6. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 29 juill 2016;bbw068.

7. Machine Learning in Medicine. N Engl J Med. 27 juin 2019;380(26):2588‑90.

8. DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, et al. A review of machine learning in obesity: Machine learning in obesity research. Obes Rev. mai 2018;19(5):668‑85.

9. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. déc 2017;42:60‑88.

10. Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Comput Biol Med. juill 2018;98:126‑46.

11. Baldi P, Sadowski P, Whiteson D. Searching for exotic particles in high-energy physics with deep learning. Nat Commun [Internet]. sept 2014;5(1). Disponible sur: http://www.nature.com/articles/ncomms5308

12. AlQuraishi M. AlphaFold at CASP13. Valencia A, éditeur. Bioinformatics. 1 nov 2019;35(22):4862‑5.

13. Codella N, Cai J, Abedini M, Garnavi R, Halpern A, Smith JR. Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images. In: Machine Learning in Medical Imaging [Internet]. Springer, Cham; 2015 [cité 27 juill 2020]. p. 118‑26. Disponible sur: https://link.springer.com/chapter/10.1007/978-3-319-24888-2_15

14. Figueiredo MAT, Jain AK. Unsupervised learning of finite mixture models. IEEE Trans Pattern Anal Mach Intell. mars 2002;24(3):381‑96.

15. Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In ACM Press; 1998. p. 92‑100. Disponible sur: http://portal.acm.org/citation.cfm?doid=279943.279962

16. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature. oct 2017;550(7676):354‑9.

17. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A Survey on Deep Transfer Learning. In: Artificial Neural Networks and Machine Learning – ICANN 2018 [Internet]. Springer, Cham; 2018 p. 270‑9. Disponible sur: https://link.springer.com/chapter/10.1007/978-3-030-01424-7_27

18. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. déc 1943;5(4):115‑33.

19. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386‑408.

20. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. In: PMLR [Internet].2011. p. 315‑23. Disponible sur: http://proceedings.mlr.press/v15/glorot11a.html

21. Press TM. Perceptrons | The MIT Press [Internet]. [cité 29 juill 2020]. Disponible sur: https://mitpress.mit.edu/books/perceptrons

22. Learning Internal Representations by Error Propagation - MIT Press books [Internet]. [cité 29 juill 2020]. Disponible sur: https://ieeexplore.ieee.org/document/6302929

23. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. oct 1986;323(6088):533‑6.

24. LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, et al. Handwritten Digit Recognition with a Back-Propagation Network. In: Touretzky DS, éditeur. Advances in Neural Information Processing Systems 2 [Internet]. Morgan-Kaufmann; 1990. p. 396–404. Disponible sur: http://papers.nips.cc/paper/293-handwritten-digit-recognition-with-a-back-propagation-network.pdf

25. Hinton GE, Osindero S, Teh Y-W. A Fast Learning Algorithm for Deep Belief Nets. Neural Comput. juill 2006;18(7):1527‑54.

26. Salakhutdinov R, Hinton G. Deep Boltzmann Machines. In: PMLR [Internet]. 2009 [cité 29 juill 2020]. p. 448‑55. Disponible sur: http://proceedings.mlr.press/v5/salakhutdinov09a.html

27. Stemmer G, Steidl S, Nöth E, Niemann H, Batliner A. Comparison and Combination of Confidence Measures. In: Text, Speech and Dialogue [Internet]. Springer, Berlin, Heidelberg; 2002 [cité 29 juill 2020]. p. 181‑8. Disponible sur: https://link.springer.com/chapter/10.1007/3-540-46154-X_25

28. Kelley HJ. Gradient Theory of Optimal Flight Paths. ARS J. oct 1960;30(10):947‑54.

29. Qian N. On the momentum term in gradient descent learning algorithms. Neural Netw. janv 1999;12(1):145‑51.

30. Nesterov Y. A method for unconstrained convex minimization problem with the rate of convergence o(1/k^2) [Internet]. 1983 [cité 31 juill 2020]. Disponible sur: https://www.semanticscholar.org/paper/A-method-for-unconstrained-convex-minimization-with-Nesterov/ed910d96802212c9e45d956adaa27d915f5d7469

31. DuchiJohn, HazanElad, SingerYoram. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. J Mach Learn Res [Internet]. 1 juill 2011 [cité 31 juill 2020]; Disponible sur: https://dl.acm.org/doi/abs/10.5555/1953048.2021068

32. Dauphin YN, de Vries H, Bengio Y. Equilibrated adaptive learning rates for non-convex optimization. arXiv.org [Internet]. 15 févr 2015 [cité 31 juill 2020]; Disponible sur: https://arxiv.org/abs/1502.04390v2

33. Dozat T. Incorporating Nesterov Momentum into Adam. 18 févr 2016 [cité 31 juill 2020]; Disponible sur: https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ

34. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ArXiv150203167 Cs [Internet]. 2 mars 2015; Disponible sur: http://arxiv.org/abs/1502.03167

35. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15(56):1929‑58.

36. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. janv 1962;160:106‑54.

37. Cireşan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J. High-Performance Neural Networks for Visual Object Classification. ArXiv11020183 Cs [Internet]. 1 févr 2011; Disponible sur: http://arxiv.org/abs/1102.0183

38. Gradient-based learning applied to document recognition - IEEE Journals & Magazine [Internet]. [cité 3 août 2020]. Disponible sur: https://ieeexplore.ieee.org/document/726791

39. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images; 2009 [cité 3 août 2020]. Disponible sur: https://scholar.google.com/scholar_lookup?title=Learning%20multiple%20layers%20of%20features%20from%20tiny%20images&author=A.%20Krizhevsky&publication_year=2009

40. Zeiler MD, Fergus R. Stochastic Pooling for Regularization of Deep Convolutional Neural Networks. ArXiv13013557 Cs Stat [Internet]. 15 janv 2013; Disponible sur: http://arxiv.org/abs/1301.3557

41. Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. J Chem Inf Model. 26 2018;58(2):287-96.

PARTIE 2 : TRAVAIL DE THESE

Chapitre 4 : Etude bio-informatique de la pseudo particule virale de l'hépatite B

L'hépatite B, l'une des infections hépatiques les plus courantes au monde, est causée par le virus de l'hépatite B (VHB). Via les cellules infectées ce virus génère, entre autres, des particules non pathogènes avec des structures de surface antigénique similaires à celles trouvées dans le virus complet (la particule de Dane). Ces particules non pathogènes qu'on appellera ici particules sous virale de l'hépatite B (PSVB) sont utilisées sous une forme recombinante pour produire des vaccins efficaces. La structure atomique des PSVB est actuellement non résolue, et les seules données structurelles existantes pour la particule entière ont été obtenues par microscopie électronique avec une résolution maximale de 12 Å. Comme pour de nombreuses protéines auto-assembleuses, la PSVB est un biosystème complexe. Cette complexité résulte de nombreuses sources d'hétérogénéité et les outils d'analyse bio-immuno-chimique traditionnels sont souvent limités dans leur capacité à décrire pleinement la surface moléculaire de la particule. Pour la particule de vaccin contre l'hépatite B (PSVB), aucune donnée de résolution atomique n'est disponible à ce jour. Dans cette étude, nous avons utilisé les principaux éléments bien connus de la structure de PSVB pour reconstituer et modéliser l'assemblage complet de la particule à un niveau moléculaire (assemblage de protéines, formation et maturation des particules). Des modèles atomiques de PSVB ont été construits sur la base d'une revue exhaustive des données expérimentales, d'une analyse des séquences d'acides aminés, d'une modélisation par enfilage itératif et d'approches de dynamique moléculaire.

1. Problématiques

Le virus de l'hépatite B (VHB - famille des Hepadnaviridiae) provoque l'hépatite B et les maladies hépatiques qui en découlent. La prévention contre le VHB a progressé au cours des 30 dernières années grâce à la vaccination (1). Néanmoins, le VHB demeure une menace constante dans le domaine de la santé publique ; (887 000 décès dans le monde en 2015) (2) et l'amélioration d'un vaccin à usage intensif est un défi perpétuel. Le vaccin anti-VHB est un vaccin recombinant à base de particules de type sous virale (PSVB) et est la forme antigénique

la plus utilisée. Il est composé d'une seule protéine, la protéine antigénique de surface de l'hépatite B (HBsAg) et de phospholipides (PLs). La structure et l'antigénicité de l'HBsAg ont été largement décrites par un panel d'approches et de résultats moléculaires, structuraux et immuno-biochimiques que nous détaillerons tout au long de ce chapitre (3–50).

La protéine HBsAg peut adopter différentes dispositions de surface dans les virus de l'hépatite B. Sous forme de protéine membranaire dans la particule de Dane ou sous forme de particule sous virale tubulaire et sphérique (22–28). La conformation de cette dernière est générée au travers de longs bioprocédés en plusieurs étapes, dans et hors de cellules de levure (29). Dans tous les cas, une étape de maturation de la particule a été décrite après leur extraction de levure ou leur sécrétion cellulaire, qui peut durer entre un minimum de 60-120 heures (11) à un maximum de 23 jours (19), selon les paramètres utilisés dans le procédé. À 37 °C et en présence d'isothiocyanate de potassium (KSCN) (11,12), l'augmentation de la taille des particules de 10 à 20% observée pendant la maturation de la PSVB (25,26) était plus rapide et reflète une réorganisation moléculaire spontanée en dépendante de l'environnement intra et extracellulaire.

La teneur en phospholipides des PSVB (le rapport entre protéine HBsAg / di-oléoyl-phosphatidylcholine (DOPC) est d'environ 60/40%) (3–8), et leurs interactions détaillées avec leur environnement ont été en partie décrites (9-13). Les données expérimentales publiées dépendent largement du temps d'observation de la synthèse et de la maturation de la PSVB. Ces résultats seront discutés dans le cadre de notre modèle, qui nous a permis de reconstituer l'assemblage chronologique et moléculaire de la PSVB. À cet égard, de nombreuses données expérimentales ont été analysées afin d'en extraire des données convergentes. Il faut garder à l'esprit que beaucoup de ces données sont issues de différentes PSVB (des sous-types «ad/yw/r» (14-16)), produites dans différents organismes de mammifère ou de levure (17,18) produites par différents bio-processus et contrôlées par différentes bio-analyses (19,50). Ces données ont été utilisées pour modéliser et reconstituer la dynamique d'assemblage et la chronologie de la maturation de la PSVB. L'intégration de ces données expérimentales observées chez toutes les PSVB à différents stades de formation et de maturation implique :

- la taille des particules (20-24)
- la variation de la taille pendant la maturation (25,26)
- la densité des particules (20), la composition et le rapport protéine/PL (3,4,6)
- la surface des protubérances antigéniques (22-25)
- et le nombre de protéines HBsAg par particule bien que cette dernière donnée reste controversée (23,24).

De nombreuses données structurelles sont disponibles à ce jour : données de cryo-ME (23,24), microscopie à force atomique (MFA) (22) et données par diffusion des rayons X (27). Toutes tendent à considérer les protubérances antigéniques comme des sous-unités condensées d'oligomères de protéine HBsAg en interaction avec les PLs environnants. Les PSVB de mammifère (24) ou de levure (23), la cryo-ME et les reconstructions d'images 3D à 12 Å ont montré le même type d'échafaudage protéine/PL pour un total de 24 protubérances par particule avec une morphologie des protubérances et de surface similaire. Des sous-populations de particules de petite et de grande taille (20,6 ± 0,8 nm ; 22,5 ± 0,6 nm) de la même masse ont été observées, toutes avec une symétrie octaédrique. L'évolution moléculaire de la petite à la grande taille de la particule est modélisée dans cette étude et s'explique par un changement de conformation entre deux unités moléculaires asymétriques. Cette modélisation, qui sera détaillée plus en avant, pourrait expliquer l'augmentation de la taille et l'évolution conformationnelle des protubérances pendant la maturation qui est nécessaire à l'antigénicité (22–25).

La plasticité/flexibilité de la surface antigénique de HBsAg pendant la maturation des particules a été décrite comme résultant principalement des ponts disulfures de 14 cystéines fortement conservés de la séquence HBsAg (30) et de leurs réseaux de commutation pendant la genèse et la maturation des PSVB (31,32). Parmi les 14 cystéines présentes dans chaque monomère de HBsAg, 10 sont considérées comme des acides aminés clés au cours de la formation des particules (29,31,35–48) grâce à un réarrangement dynamique intra et inter protéique des ponts disulfures (38–42). Malheureusement, le réseau de liaison disulfure n'est pas clairement établi. Par exemple, certains résidus cystéine se sont révélés faire partie des ponts disulfures inter ou intramoléculaires, tout comme ils peuvent avoir été identifiés comme étant sous forme libre en fonction du stade de maturation à laquelle l'observation a été réalisée (29,38–42). De plus, la synthèse des HBsAg et son organisation progressive sous forme d'oligomère dépendent d'un environnement phospholipidique membranaire spécifique dans lequel l'organisation des PLs n'est pas bien décrite. Malgré cela, les vaccins recombinant à base de PSVB obtenus à partir de différentes souches de levure et de différents processus de maturation suggèrent que les conformations épitopiques finales convergent vers un même équilibre indépendamment de son environnement (17). Ces épitopes sont reconnus par des anticorps monoclonaux spécifiques et dépendent du niveau de maturation et d'antigénicité des particules. Ainsi, les anticorps monoclonaux peuvent fournir un outil utile pour déterminer quelles combinaisons de ponts disulfures peuvent correspondre aux résultats expérimentaux. Ceci est particulièrement utile pour les épitopes discontinus (49), pour lesquels différents

segments de la séquence de HBsAg doivent être situés dans une surface inférieure à 1000 $Å^2$ (surface moyenne de l'interaction entre un épitope et son paratope). Cette méthode peut également aider à clarifier les transformations intra et intermoléculaires se produisant au cours de la maturation des particules et a été utilisée pour mettre à l'épreuve nos modèles de PSVB.

2. Construction des modèles

Dans cette étude, nous avons compilé toutes les données disponibles pour proposer un modèle atomique de la PSVB, afin de mieux comprendre son organisation et sa conformation. Ce travail a pour but d'aider à améliorer la conception et la formulation du vaccin contre le VHB. Les résultats expérimentaux et les structures moléculaires de surface associées décrits dans la littérature ont été combinés pour développer un modèle de la protéine HBsAg par une méthode de modélisation *ab initio* d'enfilage itératif. Ce modèle a ensuite été utilisé pour reconstituer deux modèles de particules d'HBsAg. La dynamique moléculaire de l'ensemble des modèles de particules a été réalisée pour mieux comprendre 1) la différence de taille des deux modèles de la particule, 2) les domaines distincts, l'orientation et l'oligomérisation des monomères HBsAg et 3) l'organisation des protéines HBsAg avec les PLs.

La particule HBsAg est composée d'un nombre déterminé de la protéine HBsAg. La séquence d'HBsAg est longue de 226 résidus. Compte tenu des profils de densité de microscopie électronique des PSVB publiés pour les petites et les grosses particules (24), nous pouvons observer que les deux particules présentent une configuration de «cube adouci», qui correspond à une symétrie octaédrique. Un cube adouci est composé de 60 arêtes, 6 carrés et 32 triangles équilatéraux. Comme modélisé dans une publication de Mulder *et al* (23), les protéines HBsAg caractérisées par des profils électroniques occupent 24 triangles équilatéraux, qui forment 12 paires de triangles reliées le long d'un bord. Ces paires de triangles à deux pointes antigéniques sont les unités élémentaires asymétriques, composées de 4, 6 ou 8 copies des protéines (il n'y a pas de consensus dans la littérature à ce sujet (23,24)), conduisant à un nombre total de 48, 72 ou 96 copies de la protéine HBsAg par particule. Les dernières surfaces de 6 carrés et 8 triangles équilatéraux, qui ne sont pas observables par microscopie électronique, sont remplies par des PLs.

La controverse sur le nombre de protéines HBsAg est due à l'absence de modèle atomique PSVB. À ce jour, il n'y a pas de structure expérimentale de la protéine HBsAg en

raison de l'impossibilité actuelle de sa cristallisation. La modélisation d'homologie n'est pas fiable, car aucun modèle pertinent ne peut être trouvé dans la PDB (la similitude maximale n'excède pas 10%). Cependant, une prédiction de structure secondaire via le logiciel Antheprot (51) a été réalisée. La composition des résidus révèle un faible nombre de résidus chargés (en particulier les résidus chargés négativement). Plusieurs méthodes de prédiction de structure secondaire ont été testées, aboutissant à la prédiction d'une structure hydrophobe en hélice alpha pour 3 segments correspondant aux résidus T5-N40, H60-I100, I150-I226, avec une séquence proche des structures de glissière à leucine pour les résidus L77 à L98.

2.1. Fabrication du monomère HBsAg

Un modèle 3D a été généré à l'aide de la méthode Iterative Threading ASSEmbly Refinement (I-TASSER) (52). Pour cela nous avons utilisé la séquence de la protéine HBsAg (référence UniprotKB P03141-3, sérotype adw2 et isoforme S) utilisée dans la fabrication des vaccins recombinants contre l'hépatite B. Un réseau de ponts disulfures intramoléculaires connu comme essentiel à la structuration de la protéine a été utilisé comme contraintes (C107-C138, C137-C149 et C139-C147) (38–40). Quatre modèles de protéines ont été générés et classés selon un score de confiance, le C-score (53). La qualité structurale des protéines modélisées est lui évalué par le TM-Score. Le C-score est calculé en fonction de la pertinence des alignements de séquence, de la taille des séquences qui sont enfilées et des paramètres de convergence des simulations d'assemblage de la structure. Il est typiquement de l'ordre de [-5,5], où un C-score de valeur plus élevée signifie un modèle avec une confiance élevée et vice-versa. Le TM-score est à l'image du RMSD (Root Mean Square Deviation), un score qui permet de mesurer la similitude structurelle entre deux structures et qui est généralement utilisé pour mesurer la précision de la modélisation d'une structure lorsque la structure native est connue. Dans le cas où la structure native n'est pas connue, les TM-score des modèles sont calculés par rapport aux structures natives ayant servi à l'enfilage. Le TM-score est borné entre [0,1], un TM-score > 0,5 indique un modèle ayant une topologie très proche des structures natives et un TM-score < 0,17 signifie une similitude trop éloignée pour être pertinente. Le modèle ayant obtenu le meilleur C-score a obtenu 4,1 (confiance élevée). Son TM-score est de 0,28 (ce qui implique une certaine modification du modèle par rapport aux structures d'origine) avec une précision moyenne pour le modèle HBsAg estimé à environ 4Å. Le modèle sélectionné est montré sur la figure 25 et est en accord avec les méthodes de prédiction de structure secondaire SOPMA intégré dans le logiciel Antheprot (51). Ce modèle a été optimisé énergétiquement à l'aide du logiciel Molecular Operating Environment (MOE

2016) en utilisant le champ de force AMBER10:EHT dont les paramètres par défaut sont optimisés pour les protéines.

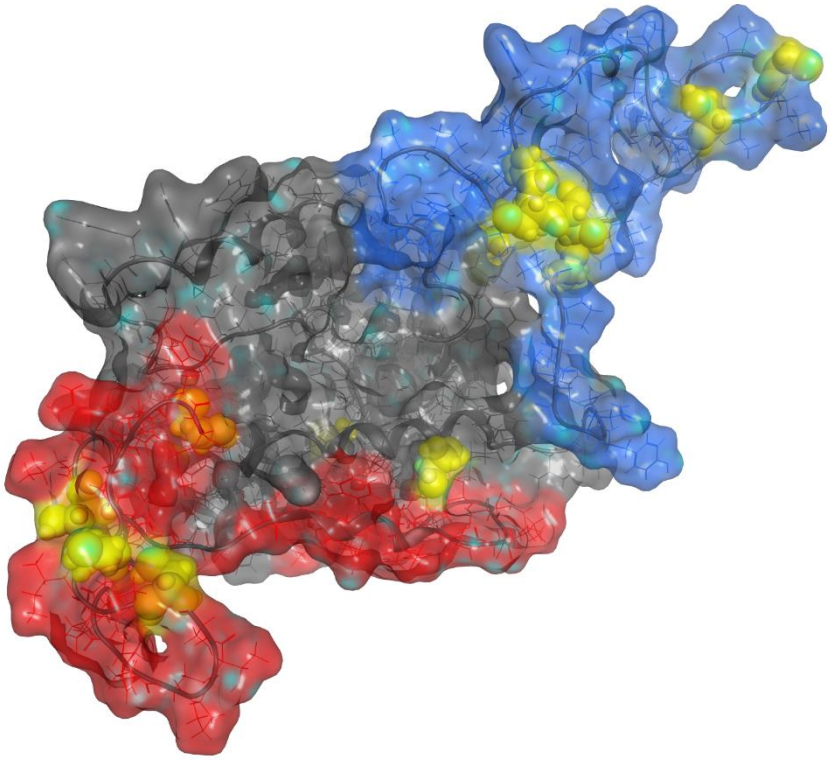


Figure 25: Modèle de la protéine monomère HBsAg obtenue par modélisation avec I-TASSER. Le modèle est affiché sous forme de ruban superposé à tous les atomes de protéines. La séquence des boucles antigéniques (BA) (D99-F180), très flexible, est exposée au solvant et colorée en bleu. Le noyau en hélice α, qui forme le domaine central rigide, est représenté en gris. La séquence des boucles capsidiques (BC) (R23-C90), qui est internalisée dans la particule, est colorée en rouge. Les cystéines de chaque domaine sont représentées par des sphères jaunes.

2.2. Fabrication des oligomères

Le modèle optimisé a été utilisé avec l'algorithme d'amarrage protéine-protéine ClusPro pour construire un modèle de dimère. Le serveur ClusPro (54) a été utilisé pour toutes les simulations d'amarrage protéine-protéine. Les simulations d'amarrage protéine-protéine ont été mises en place sans contraintes de proximité. La simulation a créé plusieurs modèles. Afin de valider le meilleur modèle, nous avons utilisé les données d'accessibilité des séquences des BA et des BC en vérifiant leur localisation interne ou externe sur la particule. Le modèle de dimère sélectionné est en accord avec l'exposition des BA et des BC et possède le score énergétique le plus faible -630 kJ/mol (haute stabilité) en plus d'être le mieux classé

par le score de regroupement. Dans ce modèle, la séquence BA (résidus 99-160) et la séquence BC (qui interagit avec la capside dans le cadre de la particule de Dane) (résidus 23-90) sont orientées vers les côtés opposés, ce qui semble cohérent avec le présupposé des orientations intérieure pour les BC) et extérieure pour les BA au sein la PSVB. En utilisant la cartographie expérimentale des épitopes (49), nous pouvons observer que plusieurs anticorps ont pu reconnaître d'autres séquences qui ne font pas partie des BA telles que les séquences 158-175, 186-207 et 208-226. Cela suggère une orientation vers l'extérieur pour ces résidus aussi. Les autres modèles de dimères issus de l'amarrage protéine-protéine avaient des résidus, identifiés comme des épitopes, qui sont internalisés, ou qui présentent de grandes surfaces hydrophobes exposées au solvant. Ces modèles-là ont donc été rejetés.

Le modèle de dimère HBsAg sélectionné a été préparé et minimisé avec le logiciel MOE. En utilisant le même protocole ce dimère a été soumis à une nouvelle simulation d'amarrage. Ces simulations aboutissent à la formation de tétramères qui ont été eux aussi sélectionnés en fonction de leur score d'amarrage et de leur cohérence expérimentale. Dans le modèle tétramérique sélectionné, les quatre protéines HBsAg ont leurs cystéines des séquences des BC situées dans la même région spatiale, ce qui permet la création de ponts disulfures intermoléculaires, comme observé expérimentalement (38). Ce complexe a été étendu à une forme multimère par l'ajout de nouveaux dimères au tétramère. Des modèles de tétramères, hexamères et octamères ont été construits en utilisant cette approche. Pour chaque modèle, un angle de rotation de 8° entre les dimères a pu induire une courbure sur la face interne de la particule. Les meilleurs modèles dimère, tétramère, hexamère et octamère ont été ajustés aux profils électroniques obtenu par cryo-ME à l'aide du logiciel Chimera en utilisant l'option « Fit » du module UCSF Chimera Map. Seule la forme tétramérique correspond aux critères des données expérimentales et donne le meilleur score de corrélation du recouvrement du nuage électronique (0,63 pour la cryo-ME EMD(Electron Microscopy Data)-1158 et 0,66 pour EMD-1159). Les formes hexamèriques et octamériques ne pouvaient pas être convenablement contenues dans le volume du nuage électronique des deux cryo-ME. Le

modèle conservé comme étant l'unité constitutive de base de la particule est donc celui d'une poutre tétramérique tel qu'illustré dans la figure 26.

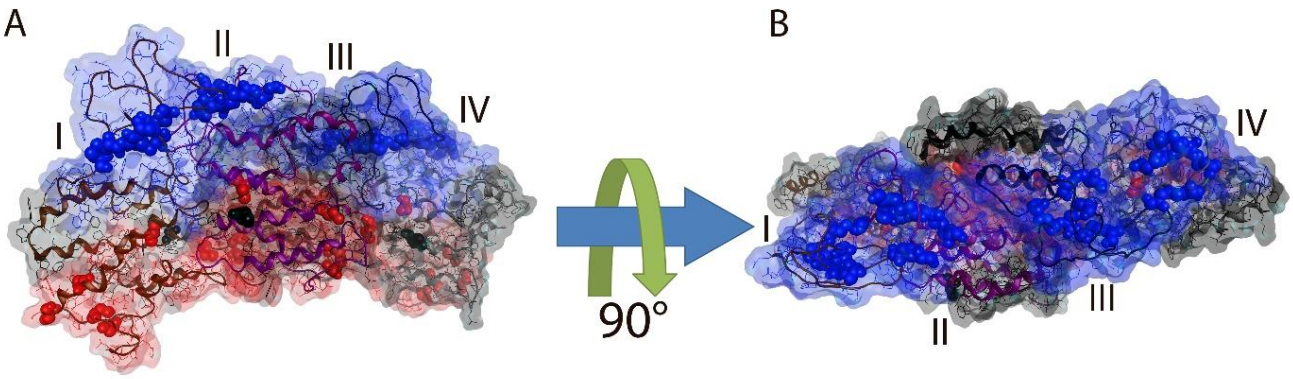


Figure 26: Poutre tétramérique HBsAg (2 unités asymétriques dimériques). Le modèle du tétramère (un dimère de dimères) est affiché avec les monomères I, II, (premier dimère) et III, IV (deuxième dimère). Les surfaces sont colorées en bleu pour les BA (99-178) et en rouge pour les BC (23-90). Le reste de la séquence est coloré en gris. La figure B est une vue qui correspondrait à une vue de l'extérieur de la particule (vue du dessus) et la figure A est tournée de 90 ° le long de l'axe y.

2.3. Fabrication des particules

2.3.1. Création et ajustement de l'échafaudage

C'est sur la base du modèle tétramérique réalisé dans la partie précédente, qu'ont été conçus les échafaudages protéiques des deux modèles atomiques des particules HBsAg complètes (PSVB). Les cryo-ME des grandes (EMD-1159) et des petites particules (EMD-1158) observées par Gilbert *et al.* (24), ont été utilisés comme « calque » pour positionner les tétramères (les poutres) et ainsi construire l'échafaudage protéique de la PSVB. Pour cela, le tétramère a été dupliqué, à partir de sa meilleure disposition dans le nuage électronique, selon une symétrique octaédrique permettant de reproduire l'architecture du cube adouci. Cette opération a été effectué à l'aide du module « Map » de Chimera. Pour la forme petite et la forme grande de la particule, 12 modèles de tétramère ont ainsi été placés et ajustés aux nuages de cryo-ME portant le nombre total de monomères d'HBsAg dans une particule à 48. Le recouvrement des protéines par le nuage électronique étant effectué sur des structures rigides, structures elles-mêmes issu d'un modèle créé *ab initio*, elles peuvent comporter de nombreuses erreurs ou approximations. C'est pourquoi il a été effectué une simulation de dynamique moléculaire à ajustement flexible (en anglais Molécular Dynamique Flexible Fitting (MDFF)) afin d'ajuster au mieux la structure de l'ensemble des 48 monomères d'HBsAg aux 2

nuages électroniques EMD-1158 (pour la forme petite soit ≃19 nm) et EMD-1159 (pour la forme grande soit ≃22 nm).

Les simulations MDFF ont été réalisées à l'aide du logiciel NAnoscale Molecular Dynamics (NAMD) 2.12. Le champ de force CHARMM36 a été utilisé pour tous les calculs. L'échafaudage protéique des particules a été simulé sous vide et la constante diélectrique a été réglée à 80. La température a été réglée à 300 K en utilisant la dynamique de Langevin avec une constante d'amortissement de 5 ps^{-1}. Le gradient de force de la grille (un paramètre de la simulation, propre à la MDFF, qui contrôle l'équilibre entre le terme de l'énergie potentielle dérivée du nuage électronique issu de la cryo-ME et le champ de force standard de la dynamique moléculaire) a été fixé à 0,3. Des contraintes de structures ont été appliquées à l'aide d'un potentielle de force harmonique afin de préserver la chiralité, la conformation des liaisons peptidiques et les éléments de structure secondaire. Les simulations de MDFF ont été effectuées jusqu'à ce que la convergence du RMSD de la structure soit atteinte ce qui représente des simulations de l'ordre de la dizaine de nanoseconde que çe soit pour la forme petite ou pour la forme grande. Les coefficients de corrélation du recouvrement des nuages électroniques par les protéines sont passés de 0,63 à 0,85 pour l'EMD-1158, et de 0,66 à 0,93 pour l'EMD-1159. La qualité de l'ajustement des nuages électroniques a confirmé l'hypothèse que les poutres moléculaires observées sont composées de tétramères construits comme un dimère de dimères.

2.3.2. Formation complète de la PSVB

La PSVB n'est pas une particule uniquement constituée de protéines c'est pourquoi afin de la modéliser de manière plus réaliste il faut encore y ajouter les phospholipides. C'est pourquoi des molécules de phospholipides organisées sous forme de bicouches de 1,2-dioléoyl-sn-glycéro-3-phosphocholine (DOPC) ont été ajoutées à l'échafaudage protéique afin de remplir les surfaces vides restantes. Pour cela, un modèle carré et plan de 100x100Å de bicouche DOPC pré-équilibré a été généré avec le constructeur de membrane du programme VMD. Ensuite, le modèle de membrane a été ajusté dans l'espace vacant en fonction des surfaces hydrophobes des protéines environnantes conformément à la disposition que nous donne le serveur web PPM (Positioning of Proteins in Membranes) qui est spécialisé dans la reconnaissance de la surface membranaire des protéines. Enfin, toutes les molécules de DOPC trop proches d'une protéine HBsAg (moins de 1Å de distance entre 2 atomes) ont été éliminées. Ce processus a été effectué pour les 6 carrés et les 8 triangles équilatéraux du

modèle à de la grande particule, et uniquement pour les 6 carrés de la petite particule, qui ne possède pas de triangles équilatéraux. Les quantités de PL ajoutés correspondent d'elles-mêmes au rapport protéine/lipide expérimentalement connu (le rapport HBsAg/DOPC étant d'environ 60/40%) (3–8). Un total de 847 molécules de DOPC a été ajouté pour la grande particule tandis que c'est 782 molécules de DOPC qui ont été ajoutées pour la petite particule.

2.3.3. Dynamique moléculaire des systèmes complets de PSVB

Afin d'équilibrer le système dans son ensemble, les modèles de la petite et de la grande particule avec leurs PLs ont été préparés aux vues d'une dynamique moléculaire sans contrainte. A cet effet les modèles ont été insérés dans une boîte de solvant d'eau de type TIP3P avec une marge de 12 Å de côté autour des protéines à laquelle des ions Na+ et Cl- ont été ajoutés pour une concentration de 0,15 Molaire afin de neutraliser le système. Le système comptabilise alors respectivement un total de plus de 1,2 millions d'atomes pour la grande particule et plus de 900 000 atomes pour la petite particule. Les systèmes ont alors été minimisés par 5000 cycles de descente la plus raide (steepest descent) puis 5000 cycles de gradient conjugué. Un préchauffage de 5 picosecondes faisant passer la température de 0 à 100K est réalisé. Ensuite un chauffage progressif est réalisé sur 50 picosecondes pour faire passer la température du système de 100K à 300K. Durant ces 2 étapes de chauffage une force de contrainte fixe les PLs afin de maintenir leurs cohésions au cours de la montée en température.

Pour chaque système, une simulation de dynamique moléculaire a été réalisée pendant plus de 100 ns afin d'équilibrer le système et d'en observer sa stabilité. Les calculs des trajectoires ont été réalisés par le programme AMBER16 avec les champs de force ff14SB et lipid14. Les conditions des boîtes périodiques ont été réalisées dans un ensemble NPT (constante Number Pressure and Temperature) signifiant à température et pression constantes. La température des simulations a été réglée à 300 K à l'aide du thermostat Langevin, une pression de 1 bar a été utilisée avec un pas d'intégration de 2 femtosecondes. Les termes électrostatiques à longue portée sont pris en charge par l'algorithme Particle-Mesh Ewald (PME) et le seuil des termes non liés tel que les interactions de Van Der Waals a été réglé à 10 Å. Le RSMD des structures évolue lentement pour atteindre un plateau aux alentours des 70 ns (figure 27). Cette lente vitesse de progression du RSMD est lié à la taille des systèmes étudiés en plus des grands réarrangements qui se produisent quand des

structures ne sont pas issues d'une résolution expérimentale. Nous pouvons aussi noter l'existence d'un écart significatif entre les structures initiales et celles à l'équilibre (compris entre 5 et 6 Å). Cet écart peut s'expliquer par 3 principaux facteurs :

- les structures initiales sont chacune issu d'une dynamique moléculaire sous contrainte, la MDFF.
- Les lipides ajoutés exercent des contraintes sur les protéines qui ne peuvent être ajustées par la boite périodique puisque les lipides sont tous inclus dans le système clos que forme la particule. Dans un système classique, une bicouche lipidique s'équilibre en interagissant avec elle-même au travers des 4 faces du plan qu'elle forme vis-à-vis de la boite périodique.
- Comme pour les lipides, une partie des molécules d'eau sont pris au piège quand elles se situent à l'intérieur de la sphère que forme la particule. Sphère rendue étanche par la présence des lipides. Ceci provoque une légère réorganisation du volume de la particule en particulier durant les premières nanosecondes.

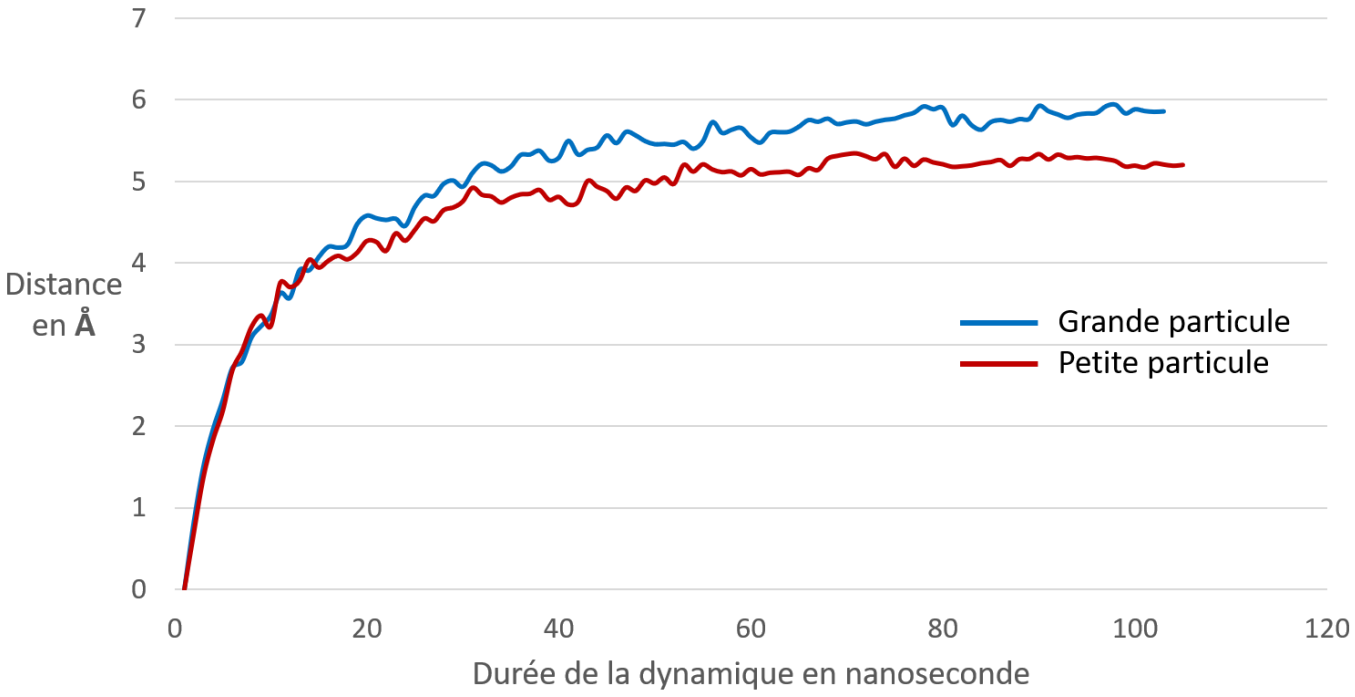


Figure 27 : RMSD de la particule grande (bleu) et petite (rouge) au cours de la dynamique moléculaire. L'écart moyen des distances des carbones α des structures échantillonnées au cours de la dynamique sont comparées à la structure initiale.

Pour les deux simulations, les ponts disulfures ont été enlevés (5,20) et convertis en groupements thiols, afin de vérifier la stabilité atomique de la particule. Comme le montre la figure 28, les deux modèles de PSVB sont en accord avec les données de cryo-ME expérimentales en termes d'organisation spatiale, avec un recouvrement de l'ordre de 85 et 93%, respectivement pour la petite et la grande particule.

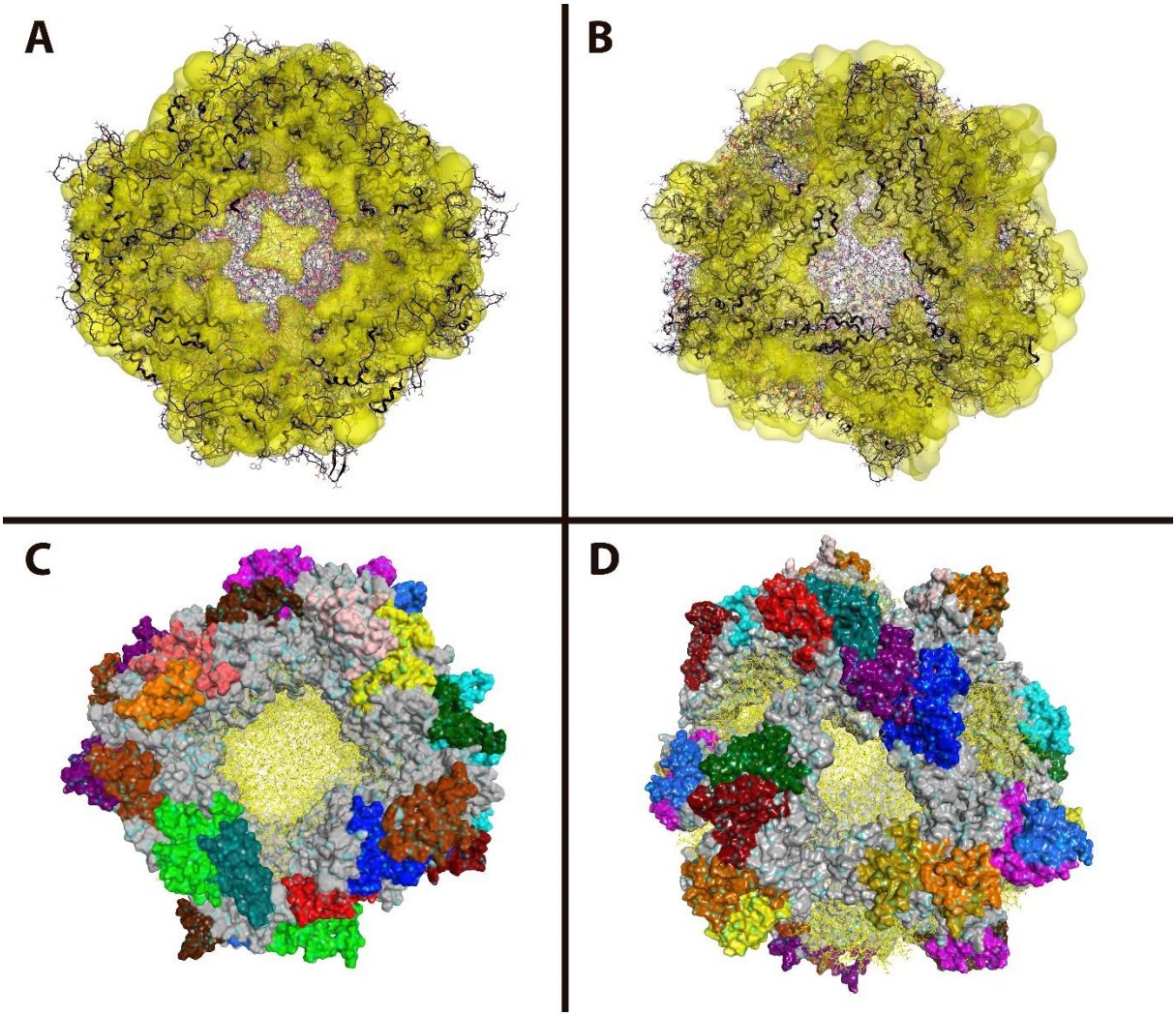


Figure 28 : Modèles atomiques de la petite (gauche: A / C) et grande (droite: B / D) particule. Panneau supérieur (A / B): pour chaque particule (petite et grande), les atomes sont superposés à la densité cryo-ME (petite EMD-1158 / grande EMD-1159) avec une surface jaune transparente. Le petit et le grand modèle sont affichés avec des rubans noirs et des lignes pour les protéines et en ligne grise pour les phospholipides DOPC. En raison de leurs flexibilités, les boucles antigéniques débordent parfois du nuage de densité électronique. Cette flexibilité peut expliquer pourquoi BA ne sont pas bien résolus à la surface de la particule. Panneau inférieur (C / D) : La surface accessible au solvant a été calculée pour les protéines et colorée en gris, à l'exception des 48 BA (99-178) qui sont colorées par chaîne (soit une couleur par monomère). Les molécules de DOPC sont affichées sous forme de lignes jaunes.

Les deux particules sont stables pendant toute la simulation dynamique moléculaire et sont restées pseudo-sphériques. Les molécules de DOPC subissent une optimisation de leurs positions et établissent des contacts avec les protéines HBsAg adjacentes. L'analyse des différents termes énergétiques du champ de force révèle que la dynamique atteint un pré-équilibre au bout de 2 ns. Après les simulations, la taille mesurée des particules est de 198 ± 16 Å et 225 ± 10 Å, respectivement pour la petite et la grande particule. Comme décrit d'abord par Gilbert *et al* (24), puis par Mulder *et al* (23) et indirectement par Short *et al* (28) (22 nm de diamètre pour les microtubes de HBsAg), les structures expérimentales des petites et grandes particules (respectivement 18-20 nm et 22 nm) correspondent aux deux sous-populations observées respectivement avant et après maturation (22–25). Par conséquent, nous proposons que les particules immatures correspondent à la sous-population des petites particules, tandis que les particules immunologiquement matures correspondent à la sous-population des grandes particules. Pendant la maturation des particules dans les levures, des protubérances sont observées à la surface des particules, ce qui correspond à nos modèles. Les protubérances obtenues dans la grosse particule ont une hauteur de 2,5 ± 0,5 nm au-dessus de la surface lipidique (4,0 ± 0,2 nm d'épaisseur), avec un diamètre à mi-hauteur de 3,9 ± 0,2 nm et une distance inter-protubérance de 7,9 ± 0,4 nm.

La transformation entre les deux états peut s'expliquer par un changement dans le réseau interne des ponts disulfures. Ce décalage peut être décrit comme un mouvement de « diaphragme » issu d’une réorganisation spatiale des tétramères les uns par rapport aux autres, comme illustré sur la figure 29. Ce mécanisme de glissement peut expliquer toute la réorganisation de la particule. Il convient de noter que dans la configuration immature, seuls les surfaces carrées de PL sont observées, tandis que des surfaces de PL triangulaires et carrés se trouvent dans les particules matures et forment *in fine* un cube adouci.

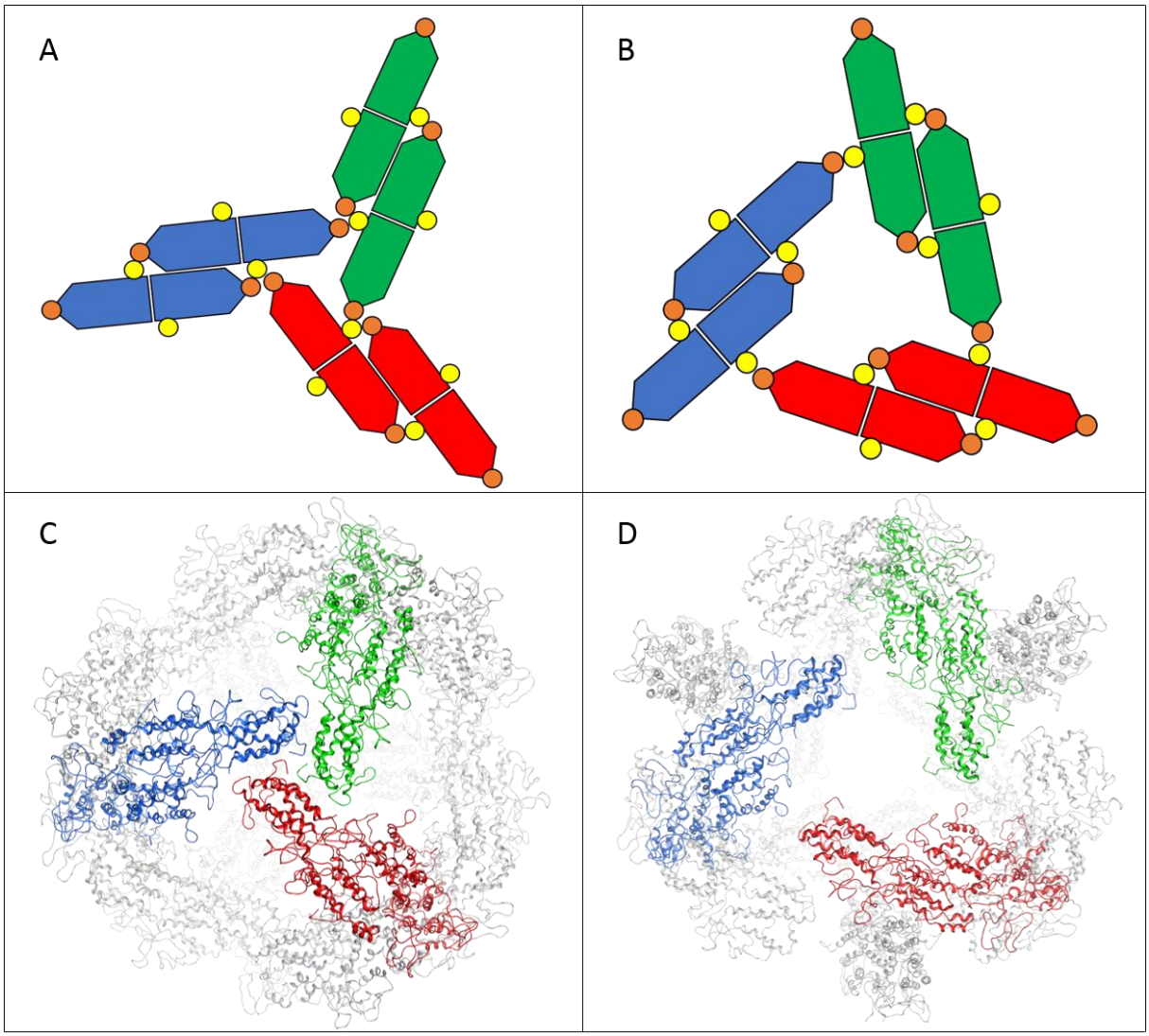


Figure 29 : modèle de diaphragme moléculaire fermé/ouvert. Panneau A / C : particule de conformation petite dite « fermée ». Panneau B / D : grande particule dite de conformation « ouverte ». Pour chaque particule, l'image inférieure (C / D) affiche trois tétramères formant le diaphragme avec une représentation en ruban. Les figures supérieures (A / B) sont des vues schématiques de ces tétramères. Chaque tétramère est formé par l'association de deux dimères adjacents. Les boules oranges représentent un amas composé des cystéines C65, C69, C48 et C76. Les boules jaunes représentent les cystéines C221. Dans la conformation fermée, les tétramères sont liés les uns aux autres par des ponts disulfures établies entre C76 (en orange) du premier tétramère et C221 (en jaune) du deuxième tétramère. Dans la particule mature, ces ponts disulfures s'établissent aussi entre C76 (en orange) du premier tétramère et C221 (en jaune) du deuxième tétramère mais faisant partie de l'autre dimère. Le passage à la conformation ouverte est produit par rotation/translation, brisant les liaisons disulfures pour en créer de nouvelles entre les billes oranges (C76) et jaunes (C221).

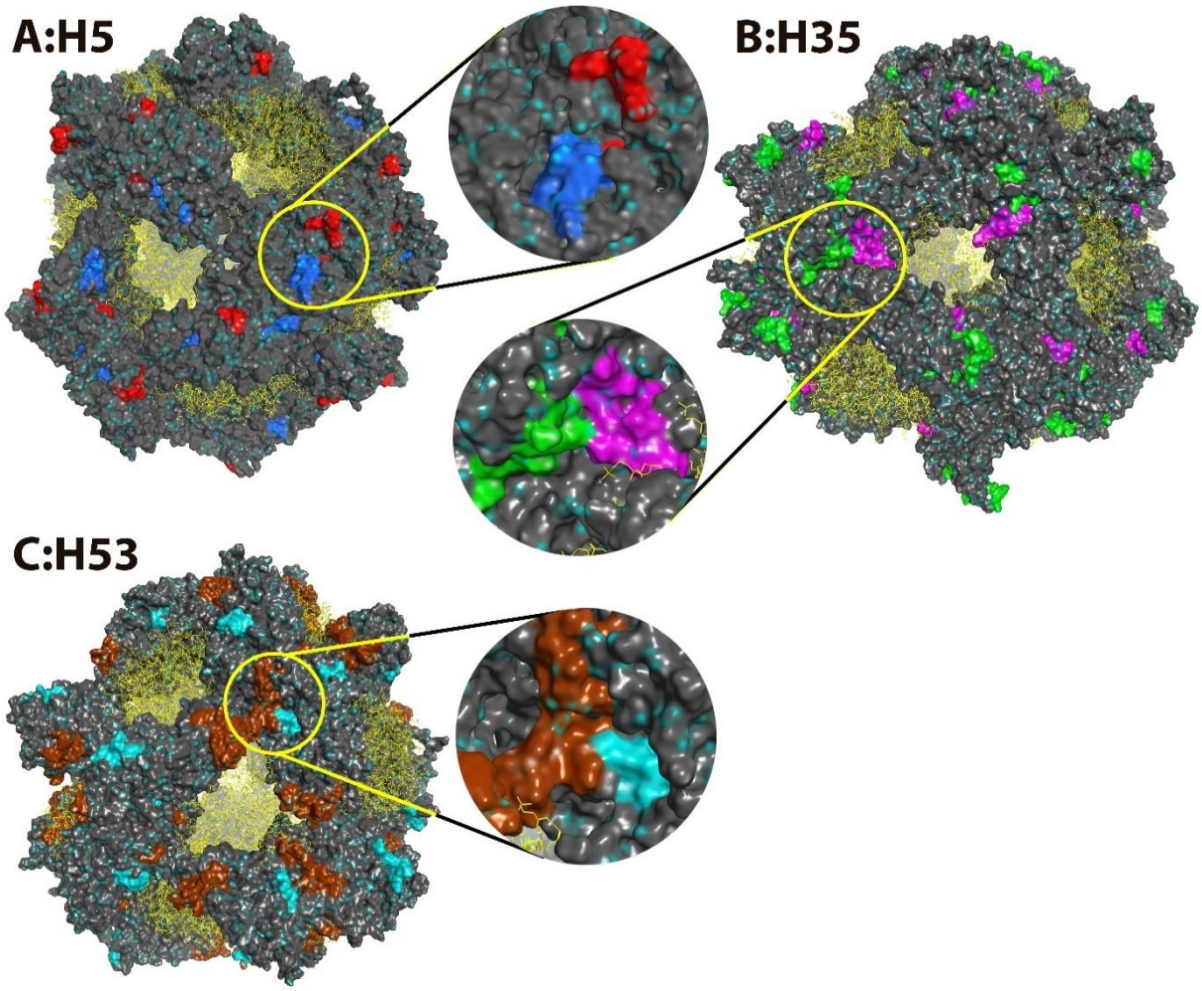


Figure 30 : Localisation d'épitopes discontinus. Représentation en surface accessible au solvant du modèle de la particule ouvert. Les surfaces protéiques sont affichées en gris, à l'exception des séquences impliquées dans les épitopes discontinus (couleurs). Les phospholipides sont affichés en lignes jaunes. Les trois épitopes correspondant aux anticorps monoclonaux H5 (A), H35 (B) et H53 (C) de la surface PSVB sont affichés en couleurs : les couleurs des résidus discontinus sont pour H5 (rouge: 101-106 / bleu: 158-167); H35 (vert: 121-130 / rose: 166-175) et H53 (bleu clair: 186-207, marron: 112-117).

Une visualisation de l'épitope discontinu sur la surface du modèle est présentée sur la figure 30 (vue de surface tétramérique partielle). Cette figure est basée sur une cartographie expérimentale réalisée avec divers anticorps monoclonaux reconnaissant des épitopes

discontinus (49). Les trois épitopes correspondant aux anticorps monoclonaux H5 (figure 30A), H35 (figure 30B) et H53 (figure 30C) de la surface de la PSVB ont été analysés afin de vérifier la proximité finale des séquences discontinues. Les séquences discontinues reconnues par les anticorps monoclonaux sont les résidus 101-106 et 158-167 pour H5, 121-130 et 166-175 pour H35 et enfin 186-207 et 112-117 pour H53. Toutes les surfaces reconnues sont inférieures à 1000 $Å^2$, ce qui est une taille appropriée pour un site de reconnaissance antigénique. Dans tous les cas, les deux séquences sont contiguës et accessibles pour les interactions intermoléculaires, ce qui suggère que les différentes séquences des épitopes discontinus sont correctement positionnées à la surface de la particule. La proximité des surfaces et leur accessibilité n'ont pas été définies comme des contraintes dans le modèle lors du processus de construction des particules.

3. Discussion

3.1. Deux types de particules de diamètre différent

L'assemblage des particules suit ces différentes étapes : synthèse de la protéine HBsAg, dimérisation, tétramérisation puis formation de l'échafaudage oligomère de PSVB (forme fermée), avant la maturation antigénique en extracellulaire (forme ouverte). La reconstitution de cette séquence chronologique permet d'expliquer les écarts dans les données expérimentales publiées, liés à différents moments d'observation. La maturation antigénique des particules est un processus lent (20 jours à 23°C), qui implique une réorganisation de la surface et des événements biochimiques peu fréquents. A 37°C et en présence d'un oxydant, le temps de maturation est réduit à 90 heures. L'antigénicité finale obtenue avec ces particules s'est avérée suffisante pour une utilisation vaccinale (11). La structuration de la particule pendant sa maturation repose sur la dynamique des ponts disulfures intra et interprotéique, combinée à d'autres facteurs tels que la formation de jonctions lipides-protéines et des interactions particules-solvants.

Des études d'anticorps ont montré que le processus de maturation comprend une modification structurelle de la séquence des boucles antigéniques (BA) exposée (11,12) par le biais de mécanismes moléculaires qui ne sont pas encore complètement compris. Dans les particules issues de la levure, les particules immatures sont caractérisées par une seule population de petite taille (25). Les particules du sérotype adw2 produites par Hansenula polymorpha ont un diamètre moyen d'environ 18 nm avant maturation, et 22-23 nm après

maturation (22). Une augmentation de taille supérieure à 10% a été observée au cours du processus de maturation avec un autre sous-type (ayw), dans lequel les diamètres des particules se sont avérés être de 17 et 22 nm (20). Par rapport à l'infection *in situ* par le VHB, dans laquelle les particules sont sécrétées en continu à partir de cellules infectées, deux sous-populations de particules coexistantes ont été détectées dans des échantillons de sang allant d'une taille de 18-20 nm à 22 nm (24). De même, les différences de taille entre les PSVB de levure et d'origine mammifère soutiennent l'hypothèse que dans les deux systèmes d'expression, 18-20 nm correspond aux particules immatures et 22-23 nm à l'état mature. Les données obtenues à l'aide de nos modèles semblent confirmer l'hypothèse de l'existence d'un processus de maturation aboutissant à l'agrandissement d'une même particule (en termes de composition) au cours du temps, au détriment de l'hypothèse selon laquelle il existerait un vaste ensemble hétérogène de structure concernant la forme sphérique.

3.2. Validation de l'organisation globale des particules

Les résultats des simulations dynamiques moléculaires (100 ns) pour les deux modèles de particules chevauchent de plus de 85% avec les nuages électroniques de cryo-ME des petites et grandes particules (24). Des protubérances apparaissent clairement chez les formes ouvertes des particules comme nous le montrent d'autres données expérimentales (23). La détection expérimentale de ces protubérances a été difficile, probablement à cause de la flexibilité des BA qui sont moins rigides que le cœur de la particule. Alors que l'échafaudage protéique de surface est bien caractérisé comme une structure pseudo-capside rigide et stable (23,24), l'organisation exacte des protéines à la surface des particules reste encore hypothétique, le nombre de monomères d'HBsAg par particule variant entre 48 et 96 (22–28). Différentes configurations du modèle de la PSVB (dimère, trimère, tétramère de HBsAg par unité asymétrique) ont été étudiées, et divers paramètres tels que le côté exposé des séquences BC/BA, l'orientation de la surface polaire et du noyau apolaire des hélice, ainsi que l'encombrement stérique du monomère HBsAg ont été prises en compte.

L'adéquation entre les deux modèles finaux et les nuages électroniques de cryo-ME expérimentaux ont été affinés par dynamique moléculaire à ajustement flexible (MDFF). Ces simulations ont ainsi pu confirmer que l'organisation de l'unité asymétrique est composée de dimères, les autres configurations étant trop encombrantes. Les autres types d'organisations de HBsAg n'ont pas été retenus selon plusieurs facteurs. D'une part un encombrement stérique non réaliste. Au-delà de 48 protéines d'HBsAg par particule, il est impossible de faire

tenir les protéines dans le nuage électronique. D'autre part les BA et BC n'étaient pas correctement exposées (à l'intérieur de la particule pour les BC et à l'extérieur de la particule pour les BA).

3.3. Protrusions des boucles antigéniques (BA)

Des caractéristiques, telles que l'hétérogénéité des surfaces et des BA ont été analysées afin de comprendre l'état des transformations entre le stade immature et le stade mature (immunogène). Des protrusions correspondant aux BA étaient apparentes à la fois à l'état mature et à l'état immature mais avec une évolution progressive de la morphologie au cours de la maturation (25). Ces résultats sont similaires qu'il s'agisse de particules issues de la levure ou de particules issues de cellules de mammifères (22-24). Un changement accéléré de la morphologie de la surface a été observé à la fois au cours d'un traitement redox et au cours d'un traitement thermique (12,25). Au cours de la maturation, les modifications du réseau de ponts disulfures au niveau des BA sont observées, avant de passer à une structure à pointes plus rigide. Après l'extraction des particules à partir de cellules de levure ou après la sécrétion de cellules de mammifères, l'environnement de la particule passe d'un système de membrane intracellulaire dédié à la synthèse et à l'oligomérisation des protéines HBsAg (première particule immature stable (5,20)) à un milieu aqueux extracellulaire dans lequel les particules « mûrissent ». Dans les deux environnements, les mécanismes moléculaires sont distincts et la maturation spontanée est probablement due à une diminution de l'enthalpie libre du système dans le nouvel environnement.

Une étude analysant le temps de maturation des particules indique une augmentation progressive du nombre de ponts disulfures au cours de ce processus (5). Ce temps de maturation se retrouve amplement diminué par un traitement redox (réduction puis oxydation), d'élévation du pH ou de la chaleur (5,12,25). Cela signifie que les 8 cystéines présentes à la surface de la séquence des BA peuvent être réduites avant maturation, sans destruction de la particule, et que les réseaux de ponts disulfures inter et intramoléculaires se remanient de préférence dans le sens de la formation d'une particule mature. Ce phénomène est en accord avec le fait que nos modèles de PSVB (en particulier la forme immature) sont restés stables au cours de la dynamique malgré l'absence de cystines. Qui plus est, le nombre de groupements thiols détectés dans les particules issues de cellules de levure ou de mammifère dépend du stade auquel le processus de maturation est analysé. Il a été démontré qu'un à trois groupements thiols par protéine peuvent être observés avant la maturation, mais plus

aucune n'est détectée une fois que la particule est mature (38–41). Cela indique que le réseau des ponts disulfures implique l'ensemble des 8 cystéines des BA.

À l'exception de la cystéine C107 isolée, les sept autres cystéines sont regroupées en 3 zones, ce qui peut augmenter leur capacité à former des ponts disulfures : une séquence de 3 cystéines contiguës (C137-C138-C139), et 2 autres séquences de 2 cystéines (C121-XX-C124) et (C147-X-C149). La proximité et la densité des cystéines (toutes étant situées dans une zone inférieure à 2500 $Å^3$), combinées à la flexibilité des BA, favorise le réarrangement des cystéines dans de nouveaux réseaux cystines. En raison de ces changements dans les réseaux intra et intermoléculaires des cystines, les surfaces des BA peuvent adopter une conformation différente conduisant à une amélioration de leur antigénicité au cours du processus de maturation. Cependant au vu de la combinatoire des réseaux cystines possibles au niveau des BA et du manque d'information concernant les paires de cystéines impliquées, il nous a été impossible de déterminer le réseau de ponts disulfures caractéristique de la conformation immunogène des BA.

3.4. Phospholipides (LPs) et antigénicité :

Le rôle des PLs est essentiel pour la formation, la maturation et l'antigénicité des particules. Leur remplacement a été évalué expérimentalement et l'antigénicité des particules a été améliorée à l'aide de PLs sélectionnés (7). Néanmoins, la structure et le rôle des PLs ne sont pas encore clairement compris. En utilisant des données RMN, il a été démontré que de l'eau était présente à l'intérieur des particules et qu'elle y restait enfermé au cours de la maturation (6,48). Ceci indique que la structure de la surface des particules est un réseau de jonction protéo-lipidique étanche. Cette imperméabilité des particules est en accord avec nos modèles qui se retrouvent étanches une fois les phospholipides placés et équilibrés et au fait qu'une plus grande quantité de molécules d'eau est contenue dans la forme ouverte de la particule que dans la forme fermée. Ceci peut s'expliquer par le fait que la particule, gagnant en taille (et donc en volume) au cours de la maturation, va avoir tendance à se remplir de molécules d'eau plutôt qu'à s'en vider.

Il a été remarqué une diffusion limitée des lipides au sein de la particule (8), ceci n'est pas typique d'une bicouche lipidique membranaire. Ce phénomène peut très bien s'expliquer à travers nos modèles par le cloisonnement des PLs, sous forme de petits îlots indépendants, par les poutres tétramériques. Cette bicouche de PL discontinue occupe les surfaces vides (pas de densité électronique par cryo-ME) dans l'échafaudage de protéine HBsAg et génère

une particule hermétique et rigide avec des molécules de solvant piégées. Il a aussi pu être montré l'existence de PL piégé à l'intérieur des particules (48). Comme ces lipides internes sont dans un environnement aqueux, ils peuvent être organisés en liposomes, micelles et nanoparticules ou ils peuvent être liés à la surface interne des PLs de la particule. Au cours de la transformation de la particule immature à la forme mature (de la petite (≃18 nm) à la grande particule (≃22 nm)), le diamètre et le volume des particules augmenteront sans aucune modification de leur poids moléculaire. Ainsi, ces lipides pourraient agir comme un réservoir de PL qui participe à la maturation des PSVB. L'apparition des ilôts lipidiques triangulaires dans la forme mature modifie l'éloignement des BA entre les tétramères par rapport à la forme immature. Cela étant, ce changement d'environnement stérique des BA joue probablement un rôle dans la conformation de ces dernières et semble nécessaire à leur bonne antigénicité. Les PLs ayant été ajoutés uniquement à la surface des particules, la légère différence du ratio protéine/lipide dans nos modèles entre la forme immature et mature (respectivement 782 et 847 molécules de DOPC) peut s'expliquer par une présence des PLs nécessairement supérieure à la surface de la particule mature étant donné la plus grande surface de cette dernière par rapport à la forme immature.

4. Conclusion

Cette étude présente les premiers modèles atomiques de particules sous virale de l'hépatite B (PSVB) utilisées massivement pour la vaccination contre le virus de l'hépatite B. Créé à l'aide de plusieurs techniques de modélisation puis affiné par plusieurs types de dynamiques moléculaires, notre modèle peut être utilisé comme base pour expliquer des observations hétérogènes et contradictoires publiées antérieurement. Nos deux modèles sont en bon accord avec les profils électroniques de cryo-ME publiés et la caractérisation de surface MFA (22-24). Un recouvrement atomique de 90% des nuages électroniques à une résolution de 12 Å (24) a été obtenue avec nos modèles et caractérise deux sous-populations de particules, l'une de petite taille et l'autre de grande taille. Grâce à ces modèles, le nombre de monomères HBsAg par particule, leurs orientation, organisation, oligomérisation et interactions avec les PLs ont pu être décrites. La variation de taille entre les PSVB immatures et matures a également été observée en laboratoire chez Sanofi Pasteur pendant le processus de maturation des particules du vaccin anti-VHB et a été corrélée avec une augmentation de l'antigénicité et de la réactivité croisée des anticorps pour les PSVB issues de la levure ainsi que celle d'origine de mammifère. Nos résultats contribuent à mieux comprendre l'évolution chronologique des complexes protéines-lipides lors de l'assemblage des particules. Les modèles de particules HBsAg ont aidé à reconstituer une séquence chronologique et moléculaire étape par étape de la formation de PSVB, tant au niveau de l'assemblage que de la maturation des particules. Ces modèles permettent de mieux comprendre les changements moléculaires qui se produisent pendant la maturation des particules et peuvent être utilisés pour optimiser la formulation du vaccin contre les particules HBsAg à l'avenir. Il est malgré tout important de garder à l'esprit que « tous les modèles sont faux, mais certains sont utiles » (George Box), à ce titre nous espérons que nos modèles demeureront plus utiles que faux.

Ce travail a fait l'objet d'une publication scientifique dans un journal international à comité de lecture :

Laurent Berthier, Olivier Brass, Gilbert Deleage, Raphaël Terreux. Construction of atomic models of full hepatitis B vaccine particles at different stages of maturation. J Mol Graph Model. 2020 Jul;98:107610. doi: 10.1016/j.jmgm.2020.107610

BIBLIOGRAPHIE DU CHAPITRE 4

1. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 310 Diseases and Injuries, 1990-2015: A Systematic Analysis for the Global Burden of Disease Study 2015. *Lancet Lond. Engl.* **2016**, *388* (10053), 1545–1602.

2. WHO. Hepatitis B Fact Sheet. (2018).

3. Gavilanes, F.; Gonzalez-Ros, J. M.; Peterson, D. L. Structure of Hepatitis B Surface Antigen. Characterization of the Lipid Components and Their Association with the Viral Proteins. *J. Biol. Chem.* **1982**, *257* (13), 7770–7777.

4. Gavilanes, F.; Gomez-Gutierrez, J.; Aracil, M.; Gonzalez-Ros, J. M.; Ferragut, J. A.; Guerrero, E.; Peterson, D. L. Hepatitis B Surface Antigen. Role of Lipids in Maintaining the Structural and Antigenic Properties of Protein Components. *Biochem. J.* **1990**, *265* (3), 857–864.

5. Wampler, D. E.; Lehman, E. D.; Boger, J.; McAleer, W. J.; Scolnick, E. M. Multiple Chemical Forms of Hepatitis B Surface Antigen Produced in Yeast. *Proc. Natl. Acad. Sci. U. S. A.* **1985**, *82* (20), 6830–6834.

6. Greiner, V. J.; Egelé, C.; Oncul, S.; Ronzon, F.; Manin, C.; Klymchenko, A.; Mély, Y. Characterization of the Lipid and Protein Organization in HBsAg Viral Particles by Steady-State and Time-Resolved Fluorescence Spectroscopy. *Biochimie* **2010**, *92* (8), 994–1002.

7. Gómez-Gutiérrez, J.; Rodríguez-Crespo, I.; Peterson, D. L.; Gavilanes, F. Reconstitution of Hepatitis B Surface Antigen Proteins into Phospholipid Vesicles. *Biochim. Biophys. Acta* **1994**, *1192* (1), 45–52.

8. Satoh, O.; Umeda, M.; Imai, H.; Tunoo, H.; Inoue, K. Lipid Composition of Hepatitis B Virus Surface Antigen Particles and the Particle-Producing Human Hepatoma Cell Lines. *J. Lipid Res.* **1990**, *31* (7), 1293–1300.

9. Jezek, J.; Chen, D.; Watson, L.; Crawford, J.; Perkins, S.; Tyagi, A.; Jones-Braun, L. A Heat-Stable Hepatitis B Vaccine Formulation. *Hum. Vaccin.* **2009**, *5* (8), 529–535.

10. Braun, L. J.; Jezek, J.; Peterson, S.; Tyagi, A.; Perkins, S.; Sylvester, D.; Guy, M.; Lal, M.; Priddy, S.; Plzak, H.; et al. Characterization of a Thermostable Hepatitis B Vaccine Formulation. *Vaccine* **2009**, *27* (34), 4609–4614.

11. Zhao, Q.; Towne, V.; Brown, M.; Wang, Y.; Abraham, D.; Oswald, C. B.; Gimenez, J. A.; Washabaugh, M. W.; Kennedy, R.; Sitrin, R. D. In-Depth Process Understanding of RECOMBIVAX HB® Maturation and Potential Epitope Improvements with Redox Treatment: Multifaceted Biochemical and Immunochemical Characterization. *Vaccine* **2011**, *29* (45), 7936–7941.

12. Zhao, Q.; Wang, Y.; Abraham, D.; Towne, V.; Kennedy, R.; Sitrin, R. D. Real Time Monitoring of Antigenicity Development of HBsAg Virus-like Particles (VLPs) during Heat- and Redox-Treatment. *Biochem. Biophys. Res. Commun.* **2011**, *408* (3), 447–453.

13. Böttcher, B.; Tsuji, N.; Takahashi, H.; Dyson, M. R.; Zhao, S.; Crowther, R. A.; Murray, K. Peptides That Block Hepatitis B Virus Assembly: Analysis by Cryomicroscopy, Mutagenesis and Transfection. *EMBO J.* **1998**, *17* (23), 6839–6845.

14. Guerrero, E.; Swenson, P. D.; Hu, P. S.; Peterson, D. L. The Antigenic Structure of HBsAg: Study of the d/y Subtype Determinant by Chemical Modification and Site Directed Mutagenesis. *Mol. Immunol.* **1990**, *27* (5), 435–441.

15. Ben-Porath, E.; Wands, J. R.; Marciniak, R. A.; Wong, M. A.; Hornstein, L.; Ryder, R.; Canlas, M.; Lingao, A.; Isselbacher, K. J. Structural Analysis of Hepatitis B Surface Antigen by Monoclonal Antibodies. *J. Clin. Invest.* **1985**, *76* (4), 1338–1347.

16. Valenzuela, P.; Gray, P.; Quiroga, M.; Zaldivar, J.; Goodman, H. M.; Rutter, W. J. Nucleotide Sequence of the Gene Coding for the Major Protein of Hepatitis B Virus Surface Antigen. *Nature* **1979**, *280* (5725), 815–819.

17. Diminsky, D.; Schirmbeck, R.; Reimann, J.; Barenholz, Y. Comparison between Hepatitis B Surface Antigen (HBsAg) Particles Derived from Mammalian Cells (CHO) and Yeast Cells (Hansenula Polymorpha): Composition, Structure and Immunogenicity. *Vaccine* **1997**, *15* (6–7), 637–647.

18. Yamaguchi, M.; Sugahara, K.; Shiosaki, K.; Mizokami, H.; Takeo, K. Fine Structure of Hepatitis B Virus Surface Antigen Produced by Recombinant Yeast: Comparison with HBsAg of Human Origin. *FEMS Microbiol. Lett.* **1998**, *165* (2), 363–367.

19. McAller, W. J.; Wasmuth, E. H. Hepatitis B Surface Antigen Vaccine Production. US4088748A, 1976.

20. Valenzuela, P.; Medina, A.; Rutter, W. J.; Ammerer, G.; Hall, B. D. Synthesis and Assembly of Hepatitis B Virus Surface Antigen Particles in Yeast. *Nature* **1982**, *298* (5872), 347–350.

21. Heermann, K. H.; Goldmann, U.; Schwartz, W.; Seyffarth, T.; Baumgarten, H.; Gerlich, W. H. Large Surface Proteins of Hepatitis B Virus Containing the Pre-s Sequence. *J. Virol.* **1984**, *52* (2), 396–402.

22. Milhiet, P.-E.; Dosset, P.; Godefroy, C.; Le Grimellec, C.; Guigner, J.-M.; Larquet, E.; Ronzon, F.; Manin, C. Nanoscale Topography of Hepatitis B Antigen Particles by Atomic Force Microscopy. *Biochimie* **2011**, *93* (2), 254–259.

23. Mulder, A. M.; Carragher, B.; Towne, V.; Meng, Y.; Wang, Y.; Dieter, L.; Potter, C. S.; Washabaugh, M. W.; Sitrin, R. D.; Zhao, Q. Toolbox for Non-Intrusive Structural and Functional Analysis of Recombinant VLP Based Vaccines: A Case Study with Hepatitis B Vaccine. *PloS One* **2012**, *7* (4), e33235.

24. Gilbert, R. J. C.; Beales, L.; Blond, D.; Simon, M. N.; Lin, B. Y.; Chisari, F. V.; Stuart, D. I.; Rowlands, D. J. Hepatitis B Small Surface Antigen Particles Are Octahedral. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (41), 14783–14788.

25. Zhao, Q.; Wang, Y.; Freed, D.; Fu, T.-M.; Gimenez, J. A.; Sitrin, R. D.; Washabaugh, M. W. Maturation of Recombinant Hepatitis B Virus Surface Antigen Particles. *Hum. Vaccin.* **2006**, *2* (4), 174–180.

26. Greiner, V. J.; Manin, C.; Larquet, E.; Ikhelef, N.; Gréco, F.; Naville, S.; Milhiet, P.-E.; Ronzon, F.; Klymchenko, A.; Mély, Y. Characterization of the Structural Modifications Accompanying the Loss of HBsAg Particle Immunogenicity. *Vaccine* **2014**, *32* (9), 1049–1054.

27. Aggerbeck, L. P.; Peterson, D. L. Electron Microscopic and Solution X-Ray Scattering Observations on the Structure of Hepatitis B Surface Antigen. *Virology* **1985**, *141* (1), 155–161.

28. Short, J. M.; Chen, S.; Roseman, A. M.; Butler, P. J. G.; Crowther, R. A. Structure of Hepatitis B Surface Antigen from Subviral Tubes Determined by Electron Cryomicroscopy. *J. Mol. Biol.* **2009**, *390* (1), 135–141.

29. Huovila, A. P.; Eder, A. M.; Fuller, S. D. Hepatitis B Surface Antigen Assembles in a Post-ER, Pre-Golgi Compartment. *J. Cell Biol.* **1992**, *118* (6), 1305–1320.

30. Norder, H.; Hammas, B.; Löfdahl, S.; Couroucé, A. M.; Magnius, L. O. Comparison of the Amino Acid Sequences of Nine Different Serotypes of Hepatitis B Surface Antigen and Genomic Classification of the Corresponding Hepatitis B Virus Strains. *J. Gen. Virol.* **1992**, *73 (Pt 5)*, 1201–1208.

31. Salisse, J.; Sureau, C. A Function Essential to Viral Entry Underlies the Hepatitis B Virus “a” Determinant. *J. Virol.* **2009**, *83* (18), 9321–9328.

32. Tleugabulova, D. Size-Exclusion Chromatographic Study of the Reduction of Recombinant Hepatitis B Surface Antigen. *J. Chromatogr. B. Biomed. Sci. App.* **1998**, *713* (2), 401–407.

33. Kreutz, C. Molecular, Immunological and Clinical Properties of Mutated Hepatitis B Viruses. *J. Cell. Mol. Med.* **2002**, *6* (1), 113–143.

34. Prange, R.; Nagel, R.; Streeck, R. E. Deletions in the Hepatitis B Virus Small Envelope Protein: Effect on Assembly and Secretion of Surface Antigen Particles. *J. Virol.* **1992**, *66* (10), 5832–5841.

35. Eble, B. E.; Lingappa, V. R.; Ganem, D. The N-Terminal (Pre-S2) Domain of a Hepatitis B Virus Surface Glycoprotein Is Translocated across Membranes by Downstream Signal Sequences. *J. Virol.* **1990**, *64* (3), 1414–1419.

36. Eble, B. E.; Lingappa, V. R.; Ganem, D. Hepatitis B Surface Antigen: An Unusual Secreted Protein Initially Synthesized as a Transmembrane Polypeptide. *Mol. Cell. Biol.* **1986**, *6* (5), 1454–1463.

37. Eble, B. E.; MacRae, D. R.; Lingappa, V. R.; Ganem, D. Multiple Topogenic Sequences Determine the Transmembrane Orientation of the Hepatitis B Surface Antigen. *Mol. Cell. Biol.* **1987**, *7* (10), 3591–3601.

38. Mangold, C. M.; Unckell, F.; Werr, M.; Streeck, R. E. Analysis of Intermolecular Disulfide Bonds and Free Sulfhydryl Groups in Hepatitis B Surface Antigen Particles. *Arch. Virol.* **1997**, *142* (11), 2257–2267.

39. Mangold, C. M.; Unckell, F.; Werr, M.; Streeck, R. E. Secretion and Antigenicity of Hepatitis B Virus Small Envelope Proteins Lacking Cysteines in the Major Antigenic Region. *Virology* **1995**, *211* (2), 535–543.

40. Mangold, C. M.; Streeck, R. E. Mutational Analysis of the Cysteine Residues in the Hepatitis B Virus Small Envelope Protein. *J. Virol.* **1993**, *67* (8), 4588–4597.

41. Guerrero, E.; Peterson, D. L.; Franco, F. G. G. *Model for the Protein Arrangement in HBsAg Particles Based on Physical and Chemical Studies*; 1988.

42. Stirk, H. J.; Thornton, J. M.; Howard, C. R. A Topological Model for Hepatitis B Surface Antigen. *Intervirology* **1992**, *33* (3), 148–158.

43. Prange, R.; Mangold, C. M.; Hilfrich, R.; Streeck, R. E. Mutational Analysis of HBsAg Assembly. *Intervirology* **1995**, *38* (1–2), 16–23.

44. Prange, R.; Streeck, R. E. Novel Transmembrane Topology of the Hepatitis B Virus Envelope Proteins. *EMBO J.* **1995**, *14* (2), 247–256.

45. Wounderlich, G.; Bruss, V. Characterization of Early Hepatitis B Virus Surface Protein Oligomers. *Arch. Virol.* **1996**, *141* (7), 1191–1205.

46. Bruss, V.; Ganem, D. Mutational Analysis of Hepatitis B Surface Antigen Particle Assembly and Secretion. *J. Virol.* **1991**, *65* (7), 3813–3820.

47. Patzer, E. J.; Nakamura, G. R.; Simonsen, C. C.; Levinson, A. D.; Brands, R. Intracellular Assembly and Packaging of Hepatitis B Surface Antigen Particles Occur in the Endoplasmic Reticulum. *J. Virol.* **1986**, *58* (3), 884–892.

48. Grélard, A.; Guichard, P.; Bonnafous, P.; Marco, S.; Lambert, O.; Manin, C.; Ronzon, F.; Dufourc, E. J. Hepatitis B Subvirus Particles Display Both a Fluid Bilayer Membrane and a Strong Resistance to Freeze Drying: A Study by Solid-State NMR, Light Scattering, and Cryo-Electron Microscopy/Tomography. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **2013**, *27* (10), 4316–4326.

49. Chen, Y. C.; Delbrook, K.; Dealwis, C.; Mimms, L.; Mushahwar, I. K.; Mandecki, W. Discontinuous Epitopes of Hepatitis B Surface Antigen Derived from a Filamentous Phage Peptide Library. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93* (5), 1997–2001.

50. Rutter, W. J.; Valenzuela, P. D. T.; Hall, B. D.; Ammerer, G. Synthesis of Human Virus Antigens by Yeast. US6544757B1, April 8, 2003.

51. Deléage, G.; Combet, C.; Blanchet, C.; Geourjon, C. ANTHEPROT: An Integrated Protein Sequence Analysis Software with Client/Server Capabilities. *Comput. Biol. Med.* **2001**, *31* (4), 259–267.

52. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12* (1), 7–8.

53. Zhang, Y.; Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins* **2004**, *57* (4), 702–710.

54. Kozakov, D.; Hall, D. R.; Xia, B.; Porter, K. A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S. The ClusPro Web Server for Protein-Protein Docking. *Nat. Protoc.* **2017**, *12* (2), 255–278.

55. Trabuco, L. G.; Villa, E.; Mitra, K.; Frank, J.; Schulten, K. Flexible Fitting of Atomic Structures into Electron Microscopy Maps Using Molecular Dynamics. *Struct. Lond. Engl. 1993* **2008**, *16* (5), 673–683.

Chapitre 5 : Etude in-silico de l'affinité protéine-protéine

Durant l'étude précédente, qui consistait en la construction d'un modèle de pseudo particule virale de l'hépatite B, il n'a pas été mesuré d'énergies d'interactions quant à l'assemblage des protéines. Seuls les scores d'amarrage moléculaire ainsi que la cohérence avec le nuage électronique ont été pris en compte pour la validité de la modélisation.

Il est communément admis que l'énergie libre la plus basse d'une protéine est celle représentative de la conformation fonctionnelle de cette dernière. Cependant, compte tenu des approximations inhérentes aux modèles de la mécanique moléculaire classique, (approximations nécessaires du fait de la taille des macromolécules que sont les protéines) le meilleur programme d'amarrage protéine-protéine (ClusPro) réalise le classement de ses poses en combinant un score d'interaction avec un score de regroupement (clustering). De plus ce score d'interaction n'est pas donné sous forme d'énergie libre, il ne peut donc pas servir à quantifier l'affinité qu'ont deux protéines l'une envers l'autre. Cela ne nous permet pas non plus de comparer l'énergie d'interaction de différents complexes entre eux. Cette limitation inclut aussi l'effet des mutations ponctuelles au sein d'un complexe ce qui limite cet outil au classement des poses d'amarrage. C'est pourquoi il est nécessaire d'explorer d'autres méthodes de calcul d'affinité protéine-protéine pour faire de la conception de protéine (protein design) et par prolongement de l'aide à la conception de vaccin par méthode de bioinformatique.

1. Etude de cas dipeptide-anticorps

1.1. Problématiques

La grippe saisonnière est une infection virale aiguë provoquée par un virus grippal. Il existe 3 types de grippe saisonnière – A, B et C. Les virus grippaux de type A se subdivisent en sous-types en fonction des différentes sortes et associations de protéines de surface du virus. Parmi les nombreux sous-types des virus grippaux A, les sous-types A (H1N1) et A

(H3N2) circulent actuellement chez l'homme. Le virus grippal circulant A (H1N1) est aussi écrit A (H1N1) pdm09 puisqu'il a été à l'origine de la pandémie de 2009 et a ensuite remplacé le virus A (H1N1) de la grippe saisonnière qui circulait avant 2009. Les seuls les virus grippaux qui ont été à l'origine de pandémies sont ceux de type A (1).

Les virus de la grippe B en circulation peuvent être divisés en 2 principaux groupes, ou lignées, appelés les lignées B/Yamagata et B/Victoria. Les virus de la grippe B ne sont pas classés en sous-types. Le virus grippal de type C quant à lui n'est que très rarement détecté et ne cause généralement que des infections bénignes, ses répercussions sur la santé publique ne sont par conséquent que de moindre importance (2).

Les vaccins contre la grippe saisonnière se concentrent donc souvent sur le type A. Pour ce qui est du choix des souches virales, il est effectué par l'Organisation mondiale de la santé (OMS) qui fait un suivi des épidémies mondiales de grippe tout au long de l'année. Elle choisit ensuite les souches les plus à risque d'émerger et de causer une épidémie (1,2). Cela peut parfois permettre la production de vaccin très efficace lorsque le virus choisi pour la fabrication du vaccin est identique aux réelles souches émergentes (3). Il peut parfois arriver que les souches qui causent les épidémies soient différentes de celles attendues, ce qui n'amène qu'une protection mineure pour les gens vaccinés, car les anticorps produits grâce au vaccin ne réagissent que faiblement au virus circulant. Ce phénomène se produit lorsque le virus mute de manière importante ou que la souche a été mal choisie. Ce fut le cas en France lors de l'hiver 2015-2016 ou une inadéquation entre la souche B/Yamagata contenue dans le vaccin et la souche circulante B/Victoria entraina une forte diminution de l'efficacité de ce dernier. Le vaccin avait alors protégé moins de 10% des vaccinés selon Skowronski *et al* (4).

Un point important en faveur de la vaccination saisonnière est la réémergence d'anciennes souches virales très contagieuses et/ou hautement pathogènes de grippe. En effet, il a été montré une grande similitude entre l'hémagglutinine du virus de la grippe A (H1N1) de 2009 et celui de la grippe espagnole de 1918 qui a fait des millions de morts à travers le monde (5).

Dans les années 2010, les vaccins antigrippaux ne sont encore que relativement efficaces (44% de succès en moyenne) en raison de l'évolution antigénique rapide du virus et en raison de contraintes de fabrication conduisant souvent à une non-concordance des vaccins et des souches dominantes du moment (6). L'idée d'un vaccin universel est évoquée depuis plusieurs années, ou plus précisément l'idée d'un vaccin « largement protecteur ». Selon une modélisation récente (2019) un vaccin universel efficace à 75% en moyenne

réduirait fortement l'impact épidémiologique de la grippe mondiale, et grâce à une chute de l'incidence et des hospitalisations, ferait économiser 3,5 milliards de dollars/an de frais médicaux rien qu'aux USA. Les frais médicaux directement imputables à la grippe saisonnière fait considéré le développement d'un tel vaccin comme d'une priorité scientifique élevé par le NIAID (6).

Dans cette perspective il est nécessaire d'avoir à disposition des tests permettant de passer au crible les anticorps circulants ayant pour cible l'hémagglutinine et la neuraminidase (les 2 protéines de surface du virus de la grippe) et qui pourraient être un point d'entrée à une sorte d'« épitome » grippal. Une preuve de concept a été réalisée par la société Pepscan montrant qu'il était possible de réaliser une cartographie des épitopes conformationnels en se basant sur les résultats de Corti D. *et al* (7). Dans cette publication, un anticorps monoclonal qui neutralise potentiellement toutes les souches de grippes de type A a été isolé ainsi que sa position sur une l'hémagglutinine de sous-type H3.

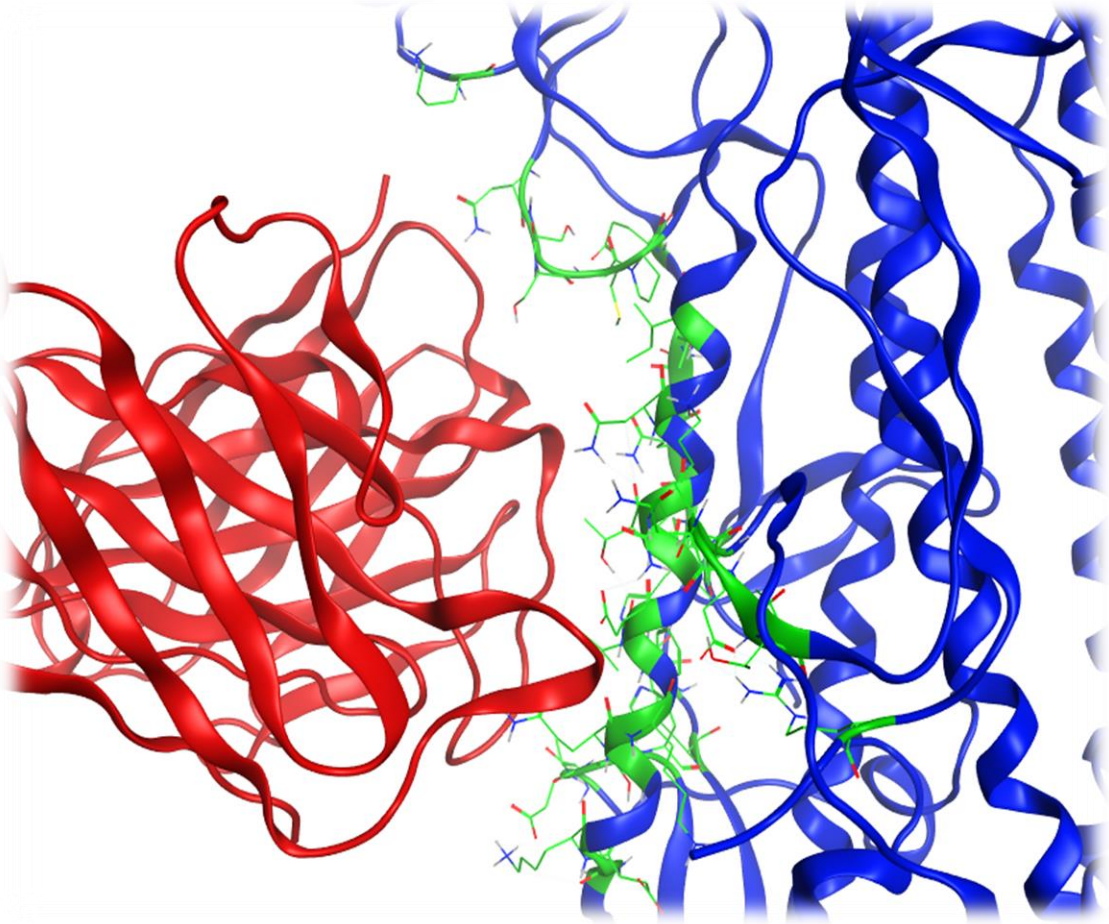


***Figure 31** : représentation en ruban du complexe 3ZTJ avec l'anticorps FI6 en rouge, l'hémagglutinine H3 en bleu et en vert l'épitope. Les résidus de l'épitope sont représentés sous forme de bâtons.*

Pepscan a donc recréé la partie de l'hémagglutinine qui contient le déterminant antigénique principal pour l'anticorps FI6 et qui est conformé en hélice alpha. Pour s'assurer de la stabilité conformationnelle de cet antigène sous la forme d'un peptide de 26 acides aminés, la séquence a été modifiée afin de former une superhélice. Une superhélice, ou coiled-coil, est un motif structural de protéines pouvant faire intervenir deux à sept hélices alpha enroulées ensemble les unes autour des autres. Les dimères et les trimères de coiled-coil sont les structures les plus fréquentes (8). Un grand nombre de protéines présentant ce type de structures sont impliquées dans des fonctions biologiques importantes, comme par exemple, les facteurs de transcription impliqués dans l'expression génétique. Les structures en superhélices contiennent généralement des acides aminés hydrophobes. Nous y retrouvons en grande majorité l'isoleucine, la leucine ou la valine. Lorsqu'une séquence de coiled-coil est enroulée en une hélice α, ces résidus hydrophobes s'alignent automatiquement en une bande formant elle-même une hélice gauche autour de l'hélice alpha. Il en résulte une structure

amphiphile qui, dans le cytoplasme d'une cellule, tend à se dimériser de telle sorte que les deux hélices alpha s'enroulent l'une autour de l'autre par la mise en contact de leurs résidus hydrophobes respectifs. Il existe malgré tout quelques superhélices droites qui ont aussi puent être observées dans la nature et dans des protéines synthétiques (9).

Sur cette hélice de 26 acides aminés ont donc été effectuées 8 mutations en leucine et isoleucine à intervalle de 3 et 2 acides aminés afin de créer une superhélice qui se dimérise et se stabilise spontanément. Ces peptides coiled-coil sont synthétisés dans des puits à peptides dans lesquels chaque puits exprime une séquence chevauchante de la précédente. L'anticorps FI6 est alors introduit dans ces puits et une mesure de la densité optique (OD) permet de déterminer quels coiled-coil fixent l'anticorps (figure 32).

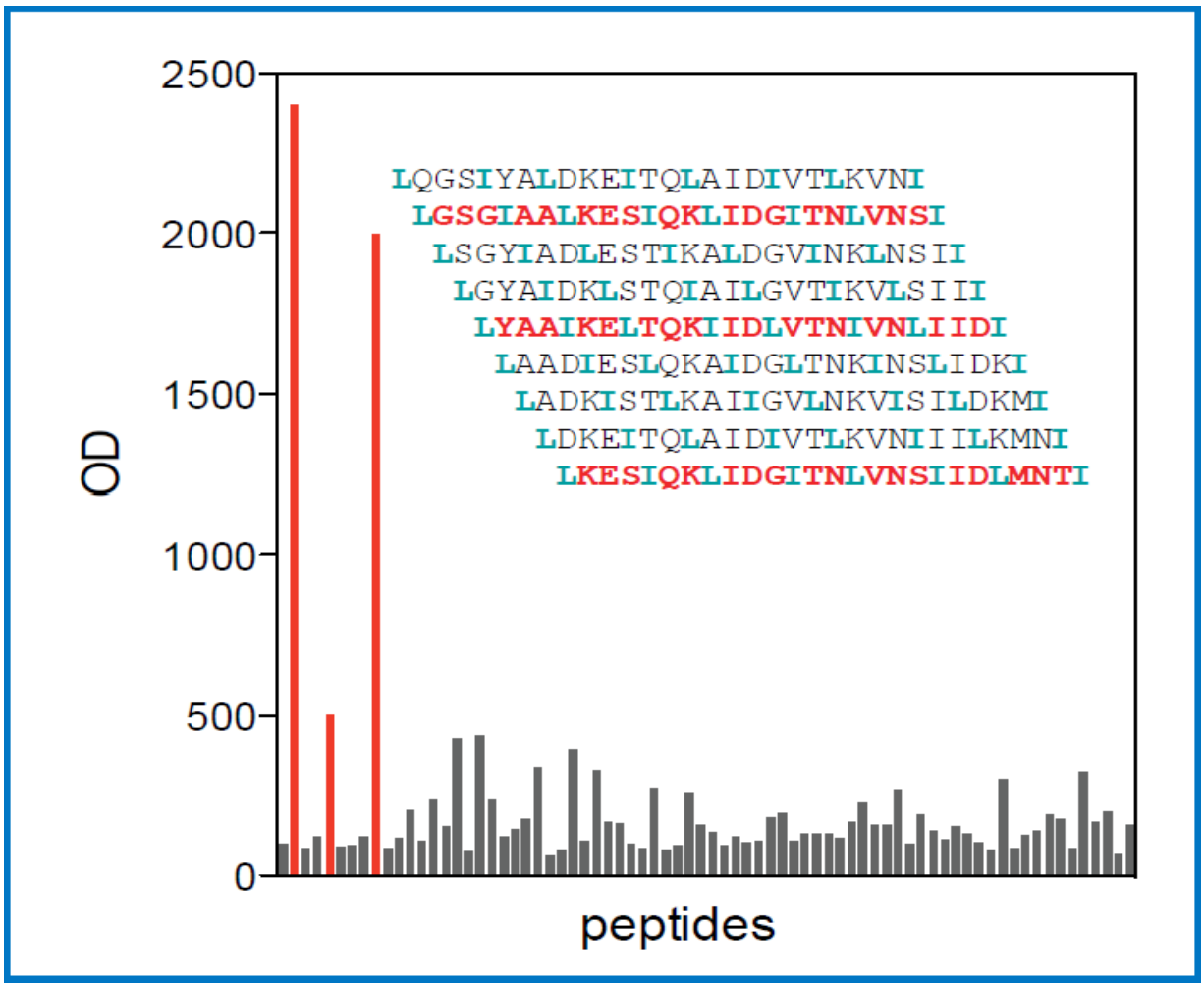


Figure 32 : Résultats des densités optiques en fonction des séquences glissantes des coiled-coil (seul les 9 premières séquences sont affichées). Les résidus mutés par rapport à la séquence d'origine sont en bleu (leucines et isoleucines). En rouge sont mis en évidence les 3 séquences interagissant le mieux avec l'anticorps.

Le but sera de savoir si les techniques de bio-informatiques permettent de corroborer ces résultats et donc d'arriver à une bonne corrélation entre les résultats obtenus expérimentalement en densité optique et ceux issus des modèles *in-silico* en termes d'énergie libre d'interaction.

1.2. Fabrication des modèles

L'ensemble des modèles ont été construits suivant un protocole en 4 étapes :

1) Construction des hélices alpha à partir des séquences d'acides aminés avec le programme de modélisation *ab initio* QUARK (10).

2) Docking des hélices alpha ainsi formées par le programme FRODOCK (Fast Rotational DOCKing) afin de créer les coiled-coil d'homodimères (11).

3) Docking des coiled-coil avec l'anticorps FI6 (référence PDB : 3ZTJ) avec le programme ClusPro (12) qui possède un algorithme optimisé pour les amarrages faisant intervenir les anticorps (13).

4) Calcul de l'énergie d'interaction par différentes méthodes bioinformatiques : PRODIGY (PROtein binDIng enerGY prediction) (14), MM-GBSA (15), en dynamique moléculaire ou seulement à partir des poses minimisées.

1.2.1. Construction des hélices alpha

La construction des peptides de 26 acides aminés ne relève pas d'un défi à proprement parler étant donné qu'ils sont tous issus d'une structure résolue d'hémagglutinine et faisant partie d'une région conformée en hélice alpha. Cependant, il existe plusieurs paramètres à prendre en compte qui peuvent affecter la conformation en hélice alpha de ce peptide. D'une part, en absence de son environnement complet, la conformation du peptide peut ne pas être stable. Et d'autre part, les mutations occasionnées pour la formation du coiled-coil peuvent modifier la structure secondaire du peptide. Pour ces raisons, nous avons privilégié l'usage d'un outil de modélisation *ab initio* tel que QUARK à la création de modèles par homologie.

En effet la construction d'un modèle par homologie ne prend pas en compte la perte de l'environnement protéique du peptide. De plus, une structure issue d'une construction par homologie ne voit sa conformation globale que peu réarrangé par rapport à la structure dont

elle est issue. Il peut donc être nécessaire d'effectuer des méthodes d'explorations conformationnelles, tel que la dynamique moléculaire, pour observer l'effet qu'impliquent l'ensemble des mutations (entre la séquence homologue et la séquence cible) sur la structure. Ce processus alourdi grandement les temps de calculs nécessaires à la création et à la validation des modèles.

L'usage d'une méthode de modélisation *ab initio* tel que QUARK permet de s'affranchir des problèmes mentionnés précédemment. L'algorithme de QUARK tente donc de résoudre la structure des protéines juste à partir de la séquence (16). La séquence soumise à l'algorithme est d'abord divisée en fragments de 1 à 20 résidus où plusieurs structures de fragments sont récupérées à chaque position à partir de structures expérimentales indépendantes. Les modèles de structures entières (26 acides aminés dans notre cas) sont ensuite assemblés à partir des fragments en utilisant des simulations de Monte Carlo par échange de réplica guidées par un champs de force spécialement conçu pour le programme. L'usage des simulations de Monte Carlo par échange de réplica permet de rendre compte de la flexibilité de la protéine et donc de confirmer si, oui ou non, la conformation en hélice est la plus représentative de la conformation native. Cette méthode a aussi pour avantage de reconstruire « naïvement » le peptide, là ou une construction par homologie partirait d'une structure biaisée car représentative d'une conformation dans un environnement protéique inexistant (figure 33). QUARK a été choisi face à d'autres programme de modélisation *ab initio* car il est à ce jour le plus performant parmi ceux disponibles (17).

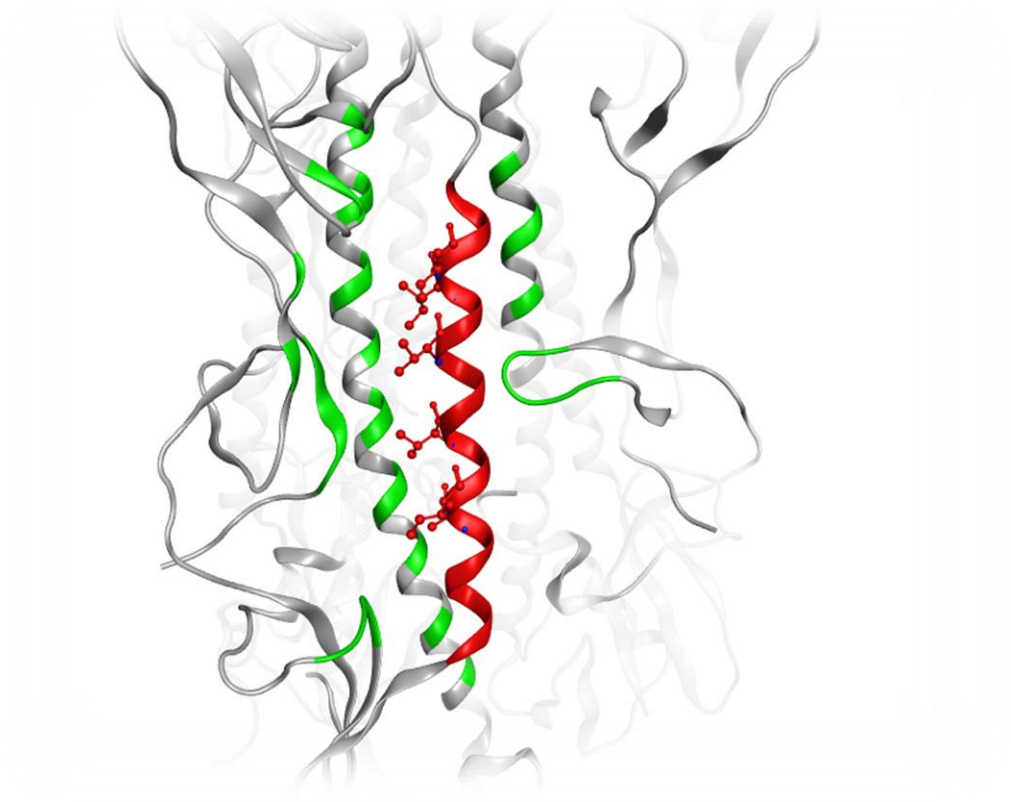


Figure 33 : Vue du peptide au sein de son environnement. La structure de l'hémagglutinine est représentée en ruban : en rouge le peptide antigène d'intérêt et en vert les résidus qui interagissent avec lui. Sont représentés sous forme de balles et de bâtons les atomes (hors hydrogène) des résidus hydrophobes susceptibles de se replier sur eux même dans un environnement aqueux.

Le programme nous propose alors 5 modèles les plus représentatifs issus des simulations de Monte Carlo par échange de réplica. Pour chaque peptide le modèle avec la plus faible énergie libre de Gibbs est conservé. Tous les modèles sont conformés en une unique hélice alpha cependant pour certaines séquences des propositions de modèles en hélices alpha avec un coude figurent parmi les autres propositions ce qui suggère une instabilité du peptide. Cette instabilité a pu être confirmée par la dynamique moléculaire en solvant décrit à l'aide du programme AMBER.

A cet effet, le modèle a été inséré dans une boîte de solvant d'eau de type TIP3P avec une marge de 12 Å de côté autour des protéines à laquelle des ions Na^+ et Cl^- ont été ajoutés pour une concentration de 0,15 Molaire afin de neutraliser le système. Le système a alors été minimisé par 5000 cycles de descente la plus raide (steepest descent) puis 5000 cycles de gradient conjugué. Un préchauffage de 5 picosecondes faisant passer la température de 0 à

100K est réalisé. Ensuite un chauffage progressif est réalisé sur 50 picosecondes pour faire passer la température du système de 100K à 300K. A la suite du chauffage une simulation de dynamique moléculaire a été réalisée pendant 100 ns afin d'équilibrer le système et d'en observer sa stabilité. Les calculs des trajectoires ont été réalisés par le programme AMBER16 avec les champs de force ff14SB. Les conditions de la boîte périodique ont été réalisées dans un ensemble NPT. La température des simulations a été réglée à 300 K à l'aide du thermostat Langevin, une pression de 1 bar a été utilisée avec un pas d'intégration de 2 femtoseconde. Les termes électrostatiques à longue portée sont pris en charge par l'algorithme Particle-Mesh Ewald (PME) et le seuil des termes non liés tel que les interactions de van der Waals a été réglé à 10 Å.

Au cours de cette dynamique nous voyons clairement l'hélice se déformer jusqu'à créer un coude au milieu de la séquence aux alentours des 20 ns. Coude qui se forme du fait de l'action du solvant vis-à-vis des régions hydrophobes de l'hélice (figure 34). Bien que la forme coudée du peptide ne se stabilise pas entièrement au cours de la dynamique, elle ne retrouve à aucun moment sa structure initiale en simple hélice.

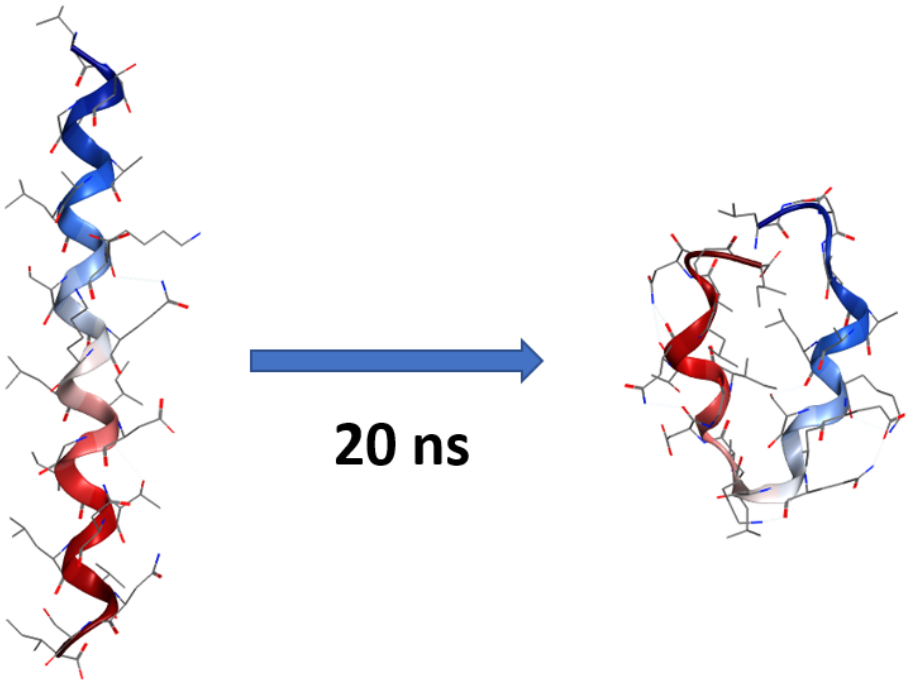


Figure 34 : Evolution du peptide antigène au cours de la dynamique moléculaire. Représentation en ruban avec dégradé de couleur du bleu au rouge, respectivement de la partie N-terminale à la partie C-terminale. Les atomes sont représentés sous forme de bâtons (hors hydrogène).

Les résultats de la dynamique moléculaire confirment donc la pertinence de l'approche de Pepscan consistant à dimériser les hélices afin de leurs permettre de conserver leurs conformations épitopiques.

1.2.2. Construction de l'homodimère coiled-coil

Une fois que nous avons réalisé les modèles des peptides de 26 acides aminés nous pouvons créer le coiled-coil par amarrage d'homodimères. Pour ce faire nous avons utilisé le programme FRODOCK car il offre de très bonnes performances en termes de vitesse d'exécution. De plus ce type d'amarrage est relativement simple à effectuer et à vérifier manuellement au vu de la nature des protéines qui lui sont soumises. En effet, il s'agit principalement d'interactions hydrophobe et symétrique étant donné que le complexe forme un homodimère. De plus les poses d'amarrage obtenues avec FRODOCK ont été comparées à celles issues de l'algorithme ClusPro qui est la référence en matière d'amarrage protéine-protéine selon le concours CAPRI (Critical Assessment of PRedicted Interactions) et donne des résultats comparables dans cette situation.

Dans la perspective d'un travail qui doit être automatisable et éventuellement reproduit sur des milliers de séquences, la vitesse d'exécution est un facteur important à prendre en compte et FRODOCK est 10 à 15 fois plus rapide que Cluspro pour le rendu des résultats.

Une dynamique moléculaire a été effectuée sur le coiled-coil de la séquence qui se conformait en coude lorsqu'il était en monomère. La dynamique moléculaire a été réalisée selon le même protocole et avec les mêmes paramètres que pour la dynamique moléculaire de la partie précédente ou le peptide était seul. Les résultats de la dynamique montrent une stabilisation de l'hélice qui demeure dans sa conformation initiale tout au long de la dynamique avec un RMSD autour des 2 Å (figure 35 et 36).

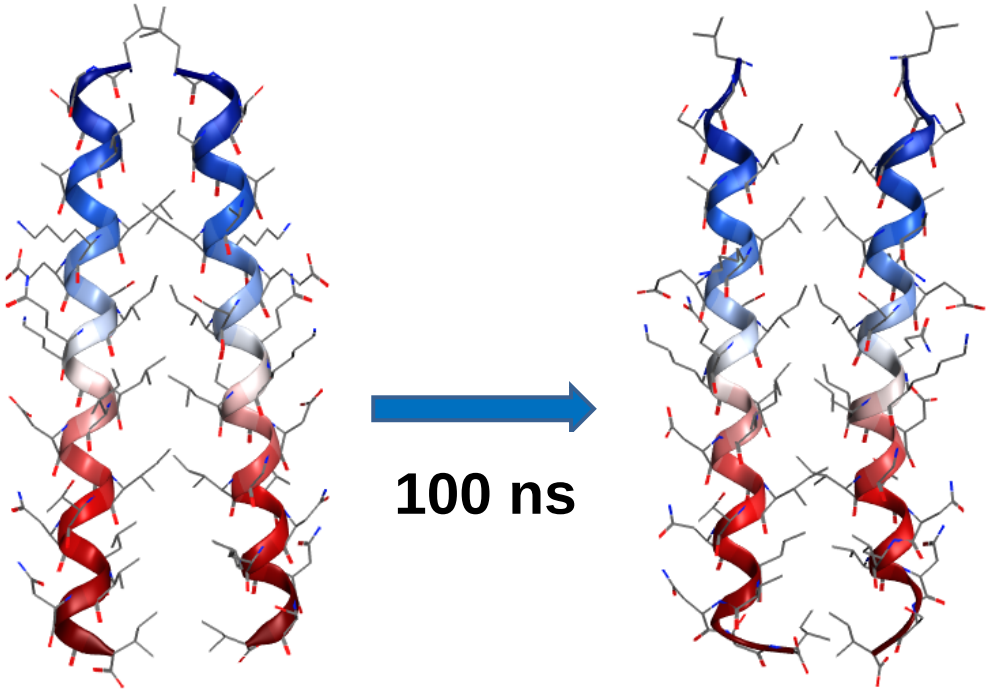


Figure 35 : Evolution du coiled-coil au cours d'une dynamique de 100 ns. Représentation en ruban avec dégradé de couleur du bleu au rouge, respectivement de la partie N-terminale à la partie C-terminale. Les atomes sont représentés sous forme de bâtons (hors hydrogène).

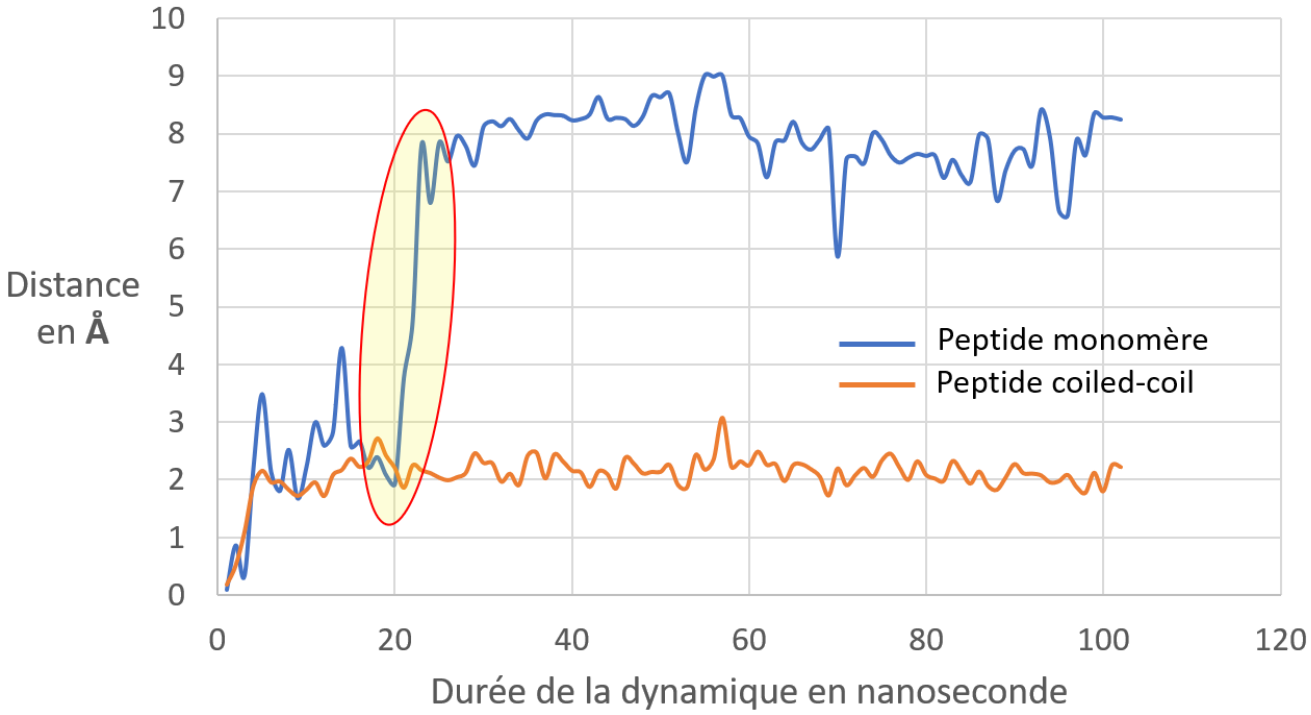


Figure 36 : RMSD du peptide sous forme monomère (bleu) et coiled-coil (orange) au cours de la dynamique moléculaire. Le changement de conformation du monomère (bleu) à partir de 20 ns de dynamique est représenté par le bond du RMSD (ellipse rouge).

1.2.3. Construction du complexe FI6/Coiled-coil

A présent, nous possédons les 2 modèles nécessaires pour étudier l'interaction entre le coiled-coil et l'anticorps. Pour cela nous devons encore former le complexe FI6/coiled-coil. C'est avec l'algorithme Cluspro que l'amarrage est effectué parce que c'est celui qui donne les meilleurs résultats au concours CAPRI et qu'il dispose d'un mode spécifique pour l'amarrage entre un anticorps et un antigène (13). En effet dans les anticorps, la liaison se produit généralement dans des zones que nous appellons régions déterminant la complémentarité (en anglais Complementarity determining regions (CDR)). L'analyse des complexes anticorps-antigène protéique a révélé une asymétrie inhérente au sein de ces interfaces. Plus précisément, les résidus de phénylalanine, de tryptophane et de tyrosine peuplent fortement le paratope de l'anticorps mais pas l'épitope de l'antigène. En concentrant l'exploration dans cette région et en modifiant le potentiel d'interaction de ces résidus, le mode d'amarrage spécifique aux anticorps augmente de 50 à 150% les performances de prédiction d'amarrage (13).

Les résultats sont triés en pondérant 2 scores. L'un issu d'un score d'interaction classique et l'autre avec un score de regroupement (clustering). L'ordre dans lesquels sont trié les poses dépend de la population du regroupement. Plus un groupe rassemble un grand nombre de poses dans un RSMD < 9 Å et plus il sera haut dans le classement. Dans notre cas, et malgré cette méthode de tri, les meilleures poses sont très proches les unes des autres (ce qui signifie que plusieurs groupes de poses se superposent) (figure 37A) et donnes des paires d'interactions entre les résidus très proches de ceux issus de la structure cristallisée (figure 37B).

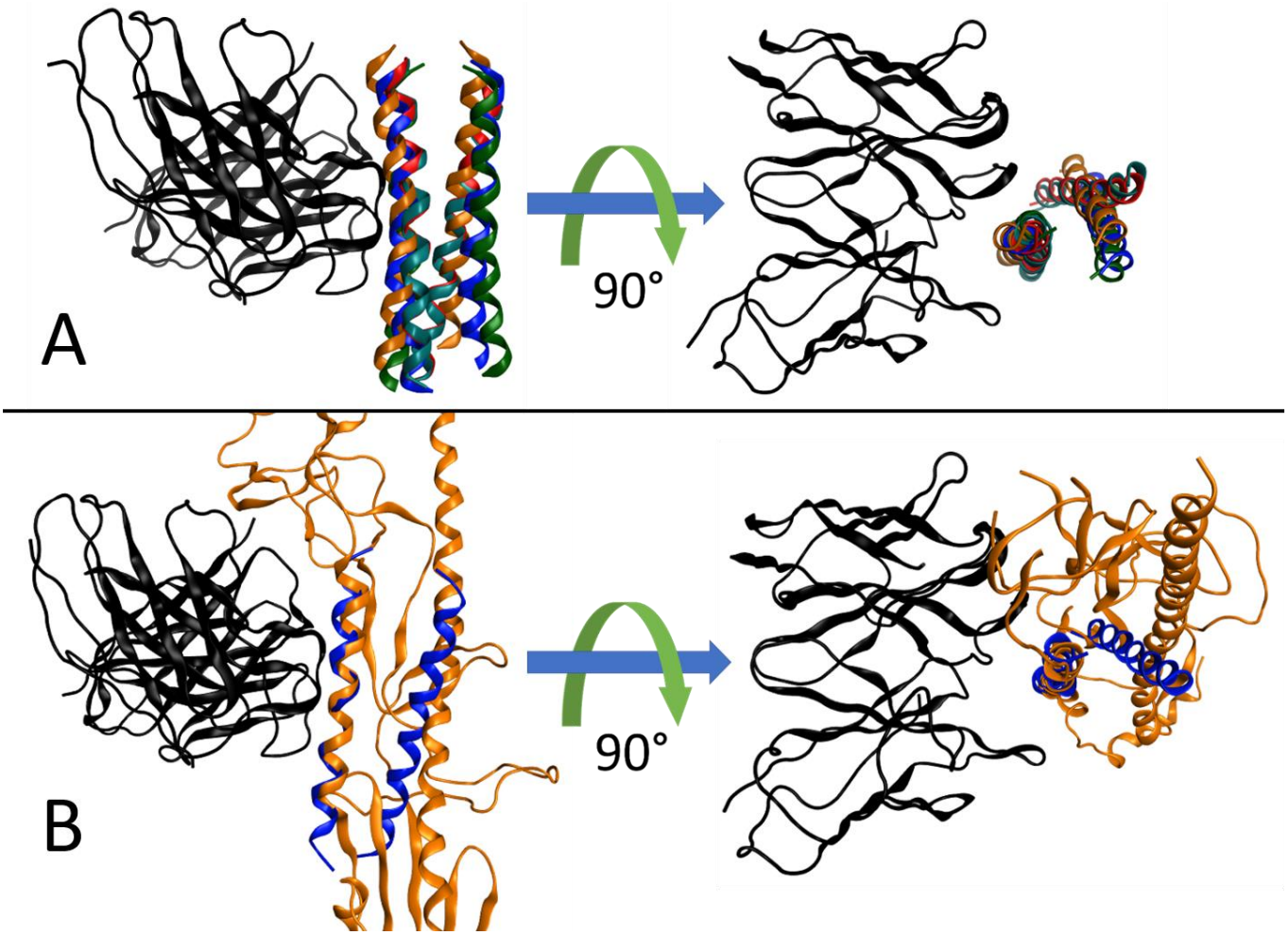


Figure 37 : Représentation de l'interaction de l'anticorps FI6 avec les coiled-coil et comparaison avec le complexe 3ZTJ. Représentation des protéines en ruban. Panneau supérieur (A) : superposition des 5 meilleures poses de coiled-coil (couleurs) vis-à-vis de l'anticorps FI6 (noir) avec ClusPro. Panneau inférieur (B) : superposition de la pose 1 du coiled-coil (bleu) avec l'hémagglutinine (orange) en interaction avec l'anticorps FI6 (noir).

1.2.4. Sélection des poses, minimisation et dynamiques moléculaires

Pour l'ensemble des 74 modèles nous avons un RSMD des 5 premières poses d'amarrage entre elles < 2 Å. Cela signifie qu'elles sont toutes très proches les unes des autres et qu'il existe une sorte de consensus concernant la zone géographique sur laquelle l'anticorps FI6 fixe les coiled-coil, peu importe les séquences. C'est pourquoi les 5 premières poses de chaque modèle ont été utilisées pour le calcul de l'affinité avec le programme PRODIGY.

Dans le protocole de ClusPro, chacune des 10 premières poses voit ses chaines latérales minimisées sur 300 cycles en utilisant uniquement le terme de Van Der Waals du champ de force CHARMM. Cette minimisation supprime les chevauchements atomiques mais ne donne généralement que de très petits changements conformationnels et ne permet pas d'atteindre les minimums locaux énergétiques. Les 5 poses ont donc aussi été reminimisées en utilisant une minimisation plus poussée de 5000 cycles sur l'ensemble des termes du champ de force AMBER et appliquée à l'ensemble de la structure (chaine latérale et squelette) pour ensuite repasser par un calcul d'affinité.

Afin de comparer les 2 protocoles précédents contre la méthode de référence nous réaliserons une dynamique moléculaire de 40 ns pour les 20 premières séquences de coiled-coil en partant à chaque fois du modèle de la meilleure pose de Cluspro afin de réaliser un calcul d'affinité par MM-GBSA. En effet cette méthode nécessite un échantillonnage par dynamique moléculaire surtout lorsque la structure des complexes n'a pas été caractérisée de manière expérimentale (15).

1.3. Résultats

La méthode la plus couramment utilisée pour calculer l'affinité entre 2 protéines est le MM-GBSA. Cette méthode a fait l'objet d'une revue dans laquelle les scores d'affinité obtenus par MM-GBSA sont comparés avec les valeurs d'affinité expérimentales de complexes protéiques dont la structure est résolue. Dans cette étude, le coefficient de corrélation de Pearson est calculé entre l'affinité expérimentale et l'affinité calculée *in-silico* et ce coefficient est de -0.65 (15).

Nous avons donc effectué les calculs d'affinité en utilisant cette méthode en calculant l'énergie d'interaction sur les 10 dernières nanosecondes d'une dynamique de 40 ns et cela

sur les 20 premières séquences de coiled-coil. Pour ce faire nous avons pris comme structures initiales de la dynamique moléculaire la 1ère pose considérée par Cluspro. Les scores d'affinité calculés *in-silico* par MM-GBSA donnent alors un coefficient de corrélation de Person avec les valeurs expérimentales de densité optique de -0.68. Le détail des résultats est condensé dans la figure 38.

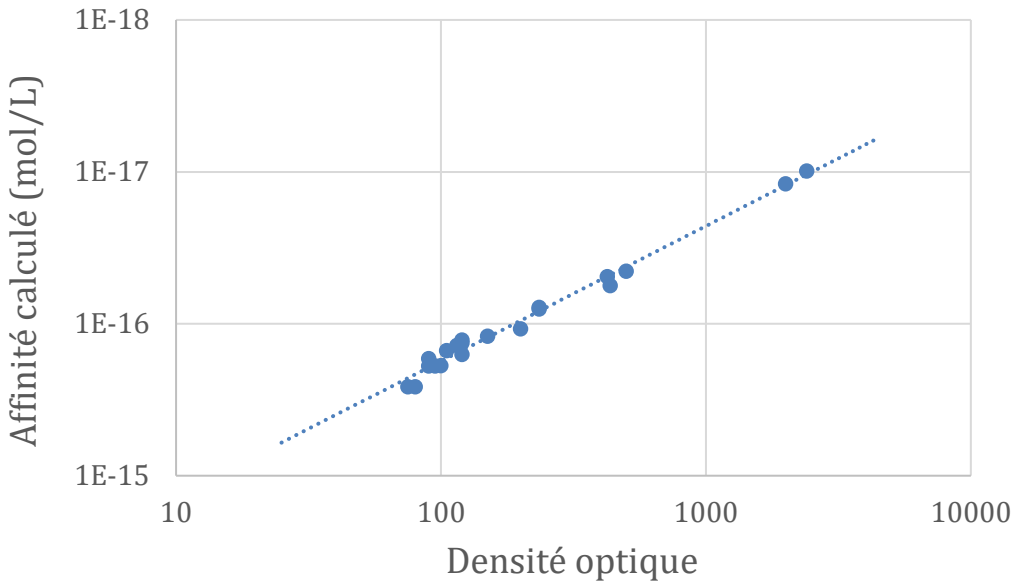


Figure 38 : Droite de régression entre la densité optique et l'énergie d'affinité calculée par MM-GBSA pour les 20 premières séquences.

Le problème de cette méthode est que son temps d'exécution, en particulier à cause de la dynamique moléculaire, est très coûteuse en temps de calcul. Elle n'est donc pas adaptée pour un criblage virtuel sur plusieurs centaines/milliers de séquence de peptides. Cependant, le but de ce 1er protocole a plus pour objectif de nous donner une référence en termes de performance que nous pourrons alors comparer avec les résultats des 2 autres protocoles bien moins expansifs en temps de calcul.

Nous avons alors testé le fait de calculer l'affinité des poses directement à la sortie des résultats d'amarrage, sans autre minimisation que celle effectuée par ClusPro. Le score d'affinité retenu est obtenu en utilisant les 5 meilleures poses et en faisant la moyenne des scores que donne PRODIGY pour chacune de ces poses. Les résultats sont résumés dans la figure 39 et donnent un coefficient de corrélation de Pearson de -0.64 pour les 20 premières séquences et de -0.54 pour l'ensemble des 74 séquences.

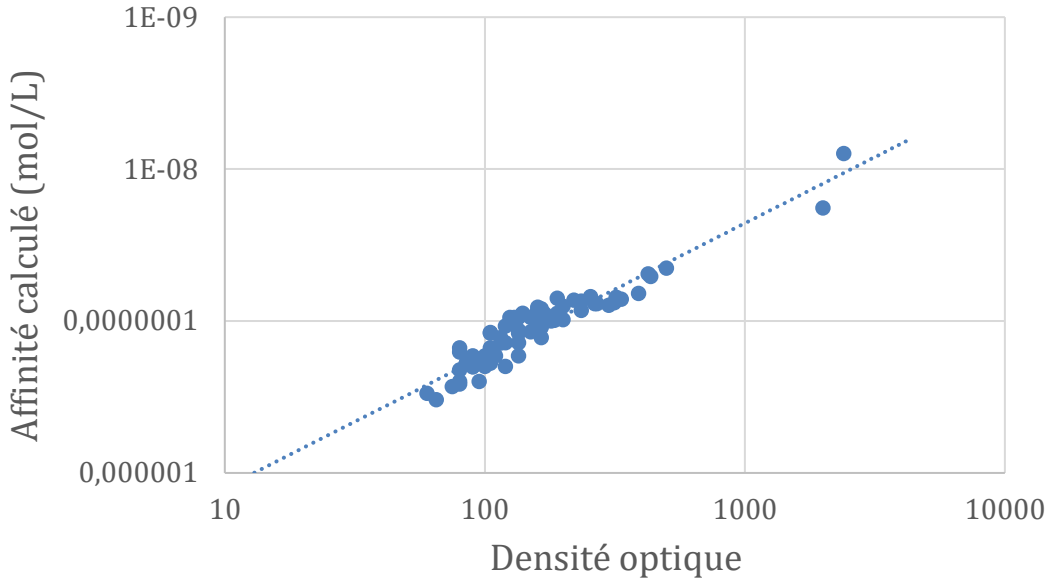


Figure 39 : Droite de régression entre la densité optique et l'énergie d'affinité calculée par PRODIGY pour les 74 séquences sans minimisation supplémentaire.

Le même protocole de calcul d'affinité que le précédent a été cette fois-ci appliqué mais avec des structures qui ont été minimisées avec beaucoup plus d'étapes. Ces étapes de minimisation doivent permettre un réarrangement plus conséquent de la conformation des complexes et devraient donc, en théorie, améliorer la précision des scores d'affinité fourni par PRODIGY. Les résultats sont résumés dans la figure 40 et donnent un coefficient de corrélation de Person de -0.70 pour les 20 premières séquences et de -0.61 pour l'ensemble des 74 séquences.

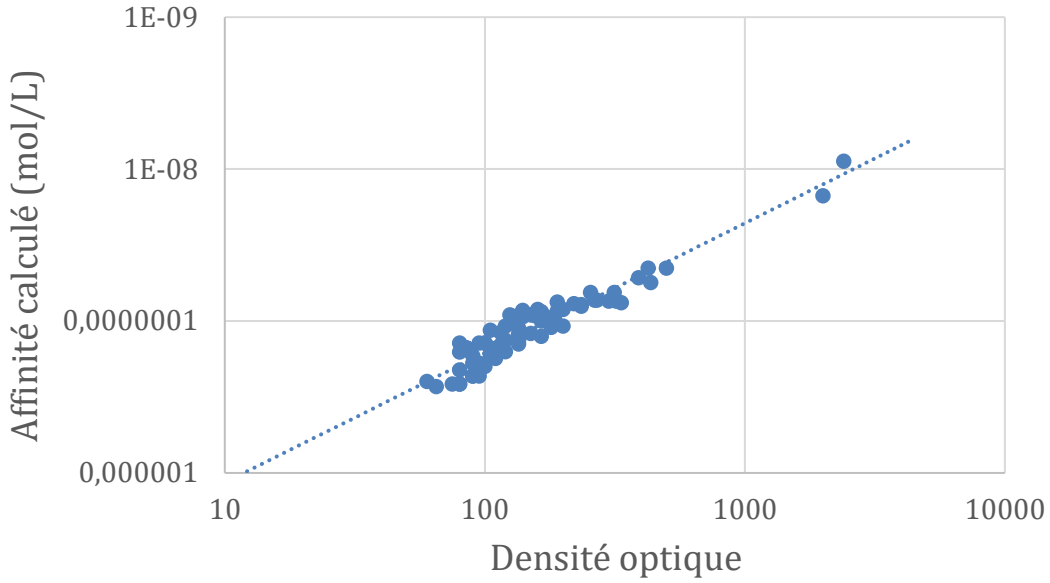


Figure 40 : Droite de régression entre la densité optique et l'énergie d'affinité calculée par PRODIGY pour les 74 séquences avec minimisation supplémentaire.

1.4. Discussion

Nous pouvons constater une bonne corrélation entre les résultats expérimentaux de densité optique et ceux calculés par des moyens de bio-informatique. En effet même en utilisant une méthode simple et rapide qu'est le retraitement des structures par une minimisation, nous constatons un coefficient de corrélation proche de ceux observés dans la littérature par une méthode de référence et bien plus lourd en temps de calcul. Nous pouvons constater que pour les résultats qui correspondent aux structures exposant bel et bien l'épitope d'intérêt (figure 32 : séquences rouge), l'affinité calculée permet une très bonne hiérarchisation vis-à-vis des données expérimentales. Le coefficient de corrélation se met cependant à souffrir lorsque nous tentons de faire correspondre des résultats expérimentaux d'interaction non spécifique (figure 32 : OD < 250) avec les résultats *in-silico*.

Il faut malgré tout nuancer ces résultats par le fait qu'il s'agit d'un cas simple de peptide conformé en hélice alpha, que la structure de l'anticorps est résolue expérimentalement et que les résultats de densité optique n'ont pas une précision suffisante pour être considérés comme purement quantitatifs . Cependant, il semblerait que pour des peptides conformés, un criblage bio-informatique de calcul d'énergie d'interaction peptide-anticorps soit envisageable, sous réserve de posséder la structure de l'anticorps. Bien que ce type d'étude *in-silico* ne puisse se substituer à des études expérimentales, elle peut permettre de rationaliser les séquences

peptidiques d'intérêt et de maîtriser les coûts d'un criblage par puce à peptide par un pré-criblage *in-silico*, en particulier dans le cadre de la création d'un épitome grippal.

2. Conclusion

Dans ce chapitre nous avons pu explorer et tester les forces et les faiblesses de différentes méthodes de calculs d'affinités. Aucune n'est parfaite, d'autant plus que leur efficacité est tributaire de la qualité des modèles sur lesquels ces méthodes sont appliquées. La dynamique moléculaire qui est sensée, en partie, pallier le problème de la précision des modèles ne suffit pas systématiquement à le résoudre. Le manque d'échantillonnage et le blocage d'une conformation dans un puit énergétique peuvent expliquer certaines de ces limites. Ces dernières limites peuvent s'expliquer par le coût computationnel inhérent à cette méthode mais d'autres limites peuvent s'expliquer par une trop grande simplification faite par les équations de la mécanique moléculaire classique. Le fait de négliger les termes entropiques des modifications conformationnelles constitue elle aussi une simplification supplémentaire.

D'autres méthodes qui effectuent un échantillonnage plus local et plus exhaustif, mais aussi plus artificiel comme ceux utilisé dans la conception computationnelle de protéine, peuvent servir d'alternatives efficaces s'agissant de mutation ponctuelle. Mais ces dernières méthodes restent imparfaites car elles sont, elles aussi, tributaires des limites de la mécanique moléculaire classique. Ces méthodes ne sont d'ailleurs pas adaptées pour comparer l'affinité de différents complexes entre eux ni pour étudier les changements conformationnels des protéines. Elles ne permettent donc pas de distinguer les interactions protéines-protéines qui auraient lieu à l'état physiologique de celles qui sont trop improbable pour être représentative.

Bien que la performance des prédictions de notre étude corrobore avec les performances publiées des outils utilisés (14,15), ces performances demeurent inférieures à celles issues des prédictions d'affinité protéine-ligand dont les coefficients de correlation tournent autour de 0.82 (28). Il sera donc entrepris dans le dernier chapitre de cette thèse d'élaborer un nouvel algorithme de calcul d'affinité qui soit suffisamment précis pour analyser l'effet que peuvent avoir des mutations ponctuelles sur l'interaction protéine-protéine et en même temps suffisamment englobant pour comparer l'affinité de différents complexes entre eux.

BIBLIOGRAPHIE DU CHAPITRE 5

1. Peteranderl C, Herold S, Schmoldt C. Human Influenza Virus Infections. Semin Respir Crit Care Med. août 2016;37(4):487-500.

2. Krammer F, Smith GJD, Fouchier RAM, Peiris M, Kedzierska K, Doherty PC, et al. Influenza. Nat Rev Dis Primer [En ligne]. 2018 [cité le 28 août 2020];4(1). Disponible: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7097467/

3. Couch RB. Seasonal Inactivated Influenza Virus Vaccines. Vaccine. 12 sept 2008;26(Suppl 4):D5-9.

4. Skowronski DM, Chambers C, Sabaiduc S, De Serres G, Winter A-L, Dickinson JA, et al. A Perfect Storm: Impact of Genomic Variation and Serial Vaccination on Low Influenza Vaccine Effectiveness During the 2014-2015 Season. Clin Infect Dis Off Publ Infect Dis Soc Am. 01 2016;63(1):21-32.

5. Nabel GJ, Wei C-J, LeΔGerwood JE. Vaccinate for the next H2N2 pandemic now. Nature. Nature Publishing Group; mars 2011;471(7337):157-8.

6. Sah P, Alfaro-Murillo JA, Fitzpatrick MC, Neuzil KM, Meyers LA, Singer BH, et al. Future epidemiological and economic impacts of universal influenza vaccines. Proc Natl Acad Sci U S A. 08 2019;116(41):20786-92.

7. Corti D, Voss J, Gamblin SJ, Codoni G, Macagno A, Jarrossay D, et al. A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. Science. 12 août 2011;333(6044):850-6.

8. Grigoryan G, Keating AE. Structural specificity in coiled-coil interactions. Curr Opin Struct Biol. août 2008;18(4):477-83.

9. Hanukoglu I, Ezra L. Proteopedia entry: coiled-coil structure of keratins. Biochem Mol Biol Educ Bimon Publ Int Union Biochem Mol Biol. févr 2014;42(1):93-4.

10. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowleΔGe-based force field. Proteins. juill 2012;80(7):1715-35.

11. Ramírez-Aportela E, López-Blanco JR, Chacón P. FRODOCK 2.0: fast protein-protein docking server. Bioinforma Oxf Engl. 01 2016;32(15):2386-8.

12. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. Nat Protoc. févr 2017;12(2):255-78.

13. Brenke R, Hall DR, Chuang G-Y, Comeau SR, Bohnuud T, Beglov D, et al. Application of asymmetric statistical potentials to antibody–protein docking. Bioinformatics. Oxford Academic; 15 oct 2012;28(20):2608-14.

14. Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein–protein complexes. Levitt M, rédacteur. eLife. eLife Sciences Publications, Ltd; 20 juill 2015;4:e07454.

15. Chen F, Liu H, Sun H, Pan P, Li Y, Li D, et al. Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein-protein binding free energies and re-rank binding poses generated by protein-protein docking. Phys Chem Chem Phys PCCP. 10 août 2016;18(32):22129-39.

16. Xu D, Zhang Y. Toward optimal fragment generations for ab initio protein structure assembly. Proteins. févr 2013;81(2):229-39.

17. [En ligne]. Groups Analysis: zscores - CASP14; [cité le 15 janv 2021]. Disponible: https://www.predictioncenter.org/casp14/zscores_final.cgi?model_type=first&gr_type=server_only

18. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet Lond Engl. 15 déc 2012;380(9859):2095-128.

19. Attia J, Hatala R, Cook DJ, Wong JG. The rational clinical examination. Does this adult patient have acute meningitis? JAMA. 14 juill 1999;282(2):175-81.

20. Detecting meningococcal meningitis epidemics in highly-endemic African countries. Releve Epidemiol Hebd. 22 sept 2000;75(38):306-9.

21. Oordt-Speets AM, Bolijn R, van Hoorn RC, Bhavsar A, Kyaw MH. Global etiology of bacterial meningitis: A systematic review and meta-analysis. PLoS ONE [En ligne]. 11 juin 2018 [cité le 4 sept 2020];13(6). Disponible: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5995389/

22. Rouaud P, Perrocheau A, Taha MK, Sesboué C, Forgues AM, Parent du Châtelet I, et al. Prolonged outbreak of B meningococcal disease in the Seine-Maritime department, France, January 2003 to June 2005. Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull. juill 2006;11(7):178-81.

23. Shin SH, Kim KS. Treatment of bacterial meningitis: an update. Expert Opin Pharmacother. oct 2012;13(15):2189-206.

24. Lakos A, Torzsa P, Ferenci T. [Bexsero, a novel vaccine against meningococcus]. Orv Hetil. 14 févr 2016;157(7):242-6.

25. Shirley M, Taha M-K. MenB-FHbp Meningococcal Group B Vaccine (Trumenba®): A Review in Active Immunization in Individuals Aged ≥ 10 Years. Drugs. févr 2018;78(2):257-68.

26. Malito E, Lo Surdo P, Veggi D, Santini L, Stefek H, Brunelli B, et al. Neisseria meningitidis factor H binding protein bound to monoclonal antibody JAR5: implications for antibody synergy. Biochem J. 15 déc 2016;473(24):4699-713.

27. Rossi R, Konar M, Beernink PT. Meningococcal Factor H Binding Protein Vaccine Antigens with Increased Thermal Stability and Decreased Binding of Human Factor H. Infect Immun. 2016;84(6):1735-42.

28. Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. J Chem Inf Model. 26 2018;58(2):287-96.

Chapitre 6 : Score d'interaction protéine-protéine à l'aide de l'apprentissage profond

1. Problématiques

Comme montré dans les chapitres précédents les scores d'affinités font face à de nombreuses limites tant en termes de précision qu'au niveau de leur corroboration avec les valeurs expérimentales. Nous avons à faire à des scores calibrés pour une fonction (Score d'amarrage, affinité globale/relative, affinité locale). Quand bien même la fonction d'un score est dévolue à une forte corrélation avec les constantes d'affinité expérimentale (exemple : PRODIGY (1), MM-GBSA (2)), les prédictions concernant les interactions protéines-protéines demeurent bien moins performantes que ce qui est réalisé pour les interactions protéines-ligands (3,4).

C'est dans le but d'apporter une contribution à cet édifice que nous nous sommes lancés dans l'exploration d'une méthode inédite dans le domaine de la prédiction de l'interaction protéine-protéine. C'est par l'approche de l'apprentissage automatique et plus particulièrement par la voie de l'apprentissage en profondeur par des réseaux de neurones convolutifs que nous avons entrepris l'élaboration d'un algorithme de prédiction de l'énergie d'affinité des complexes protéiques. Cette approche a déjà été validée pour la prédiction d'affinité des interactions protéines-ligands (3–5). En extrapolant le fait que la nature des interactions protéines-protéines soient sensiblement identiques à celle des interactions protéines-ligands, cette méthode devrait permettre de se rapprocher de la performance qu'ont les programmes de prédiction d'affinité protéines-ligands pour la prédiction d'affinité protéines-protéines. Nous nommerons cette approche le DLPPS (Deep Learning Protein-Protein Scoring), un outil pour la prédiction de score d'interaction protéine-protéine à l'aide de réseaux de neurones artificiels par apprentissage profond.

2. Elaboration de la base de données

L'une des limites à l'apprentissage automatique et en particulier lorsque nous touchons à l'apprentissage en profondeur est la quantité de données nécessaires à un bon apprentissage. En effet, afin que le réseau puisse généraliser des « règles » qui s'appliquent à l'ensembles des cas, il faut lui apporter un grand nombre d'exemples. Ce qui a été rendu possible par la création d'une base de données regroupant les énergies d'interactions expérimentales de tous les complexes protéiques dont les structures ont elles aussi été résolues expérimentalement. Cette base de données est la PDBbind (6) version 2018.

Cette base de données a pour but de fournir une collection complète des données d'affinité mesurées expérimentalement pour tous les types de complexes biomoléculaires déposés dans la Protein Data Bank (PDB) donc ayant une structure tridimensionnelle résolue. Créée en 2004 ce n'est qu'en 2009 qu'elle intègre les complexes protéiques pour en contenir dans sa version 2018 près de 2416. Cependant, il existe de nombreuses approximations dans les informations concernant les données expérimentales, tant au niveau des données d'affinité qu'au niveau des structures. Il a donc été réalisé une succession de tris :

- Sur les données des affinités expérimentales ; n'ont été conservées que celles possédant des constantes de dissociation (Kd) et ayant des marges d'erreurs inférieures à 50%. Ce qui réduit la base de données initiale à 1947 complexes.
- Sur les fichiers PDB un tri a été effectué pour retirer tous les complexes ayant des structures incomplètes (en N-ter, C-ter ou entre les 2) uniquement si les parties manquantes sont à une distance de plus de 15 Å de distance de l'interface. Ce qui réduit à nouveau la base de données à 1683 complexes.
- Un autre critère de validation des fichiers de PDB est effectué sur les multi-mères pour lesquelles le dimère d'interaction n'est pas clairement défini ainsi qu'à ceux à qui il manque une partie substantielle de la structure même si la distance des bouts manquants est supérieure à 15 Å de distance de l'interface. Ce qui réduit finalement la base de données à 1248 complexes.

Ces 1248 complexes sont ensuite subdivisés en 3 bases de données. Une de 1044 complexes, elle servira à l'apprentissage du réseau à proprement parler. Une autre de 108 complexes, elle servira de base de données de validation. La base de données de validation servira à vérifier que le réseau de neurones n'est pas en train de surapprendre la base de données d'apprentissage. Le sur-apprentissage signifie que le réseau se met à apprendre par

cœur les exemples plutôt que d'en généraliser des règles applicables à l'ensemble des cas. Enfin, une base de données de test de 79 complexes servira de comparaison avec les autres algorithmes de calculs d'affinité, en particulier PRODIGY qui fait office de référence en matière de score d'affinité protéine-protéine (7).

Pour la constitution de la base de données de test (79 complexes) nous avons simplement récupéré celle utilisé par Vangone *et al.* qui a servi dans cette même publication pour juger et comparer la performance de la fonction score de PRODIGY avec d'autres algorithmes (1). Cette base de données a été créé pour servir de référence à l'étude des interactions protéine-protéine par Kastritis *et al.* (8) et est elle-même construit à partir de la base de données de Hwang *et al.* (9) dont le but est de servir de référence à l'étude de l'amarrage protéine-protéine. Elle sera donc la base de données de référence.

Pour la base de données de validation les fichiers de PDB ont été sélectionnés aléatoirement. Nous avons ensuite vérifié que la base de données de validation partage une distribution proche de la base de données d'entrainement concernant différents critères. Cela dans le but que le coefficient de corrélation des 2 base de données soit suffisamment proche l'une de l'autre pour détecter un éventuel découplage qui surviendrait entre les 2 au cours de l'apprentissage. Ce découplage signifierait le début d'un sur-apprentissage. Les critères que nous avons retenu pour la distribution des données sont :

- L'énergie d'affinité des complexes : l'énergie d'affinité moyen de la base de donnée de validation est de -9,9 kcal.mol^{-1} pour un écart type de 2,67. Pour la base de donnée d'entrainement l'énergie d'affinité moyen est de -10.01 kcal.mol^{-1} pour un écart type de 2,72.
- La date de publication des structures : les 2 bases de données partagent une distribution identique de date de publication des structures allant de 1996 à 2018. Les structures les plus récentes étant les plus nombreuses.
- La résolution des structures : pour la base de donnée d'entrainement la moyenne des résolutions est de 2,46 Å pour un écart type de 0.66 et pour la base de données de validation la moyenne est à 2,51 pour un écart type de 0,75.

Nous pouvons noter que notre base de données est bien plus faible que celle utilisée par l'équipe de Jiménez *et al.* (3) dont s'inspire notre approche ; 1248 complexes pour notre base de données tandis que Jiménez *et al.* en rassemble 4057. Cette différence s'explique par le fait que la base de données PDBbind rassemble beaucoup plus de complexes protéine-ligand que de complexes protéine-protéine. La version 2016 de la PDBbind qui a servi pour

l'élaboration de Kdeep (3) contenait 13 308 complexes protéine-ligand contre 2 416 complexes protéine-protéine pour la version 2018 de la PDBbind utilisée pour notre algorithme DLPPS. Nous verrons dans la partie résultats que cela n'a pas posé de problème en terme de sur-apprentissage mais cela peut limiter la performance de l'apprentissage et donc la puissance de la prédiction.

3. Stratégies d'approches

L'usage de réseau de neurones convolutif se justifie par le fait que ce type de réseau excelle dans le traitement des motifs géométriques. Ils se sont fait connaître en particulier dans le domaine de la reconnaissance d'image en deux dimensions (10) mais leur utilisation peut tout aussi bien s'appliquer pour des objets en trois dimensions. Si les images en 2D sont décomposées en pixels, les images en 3D le sont en voxels (figure 41).

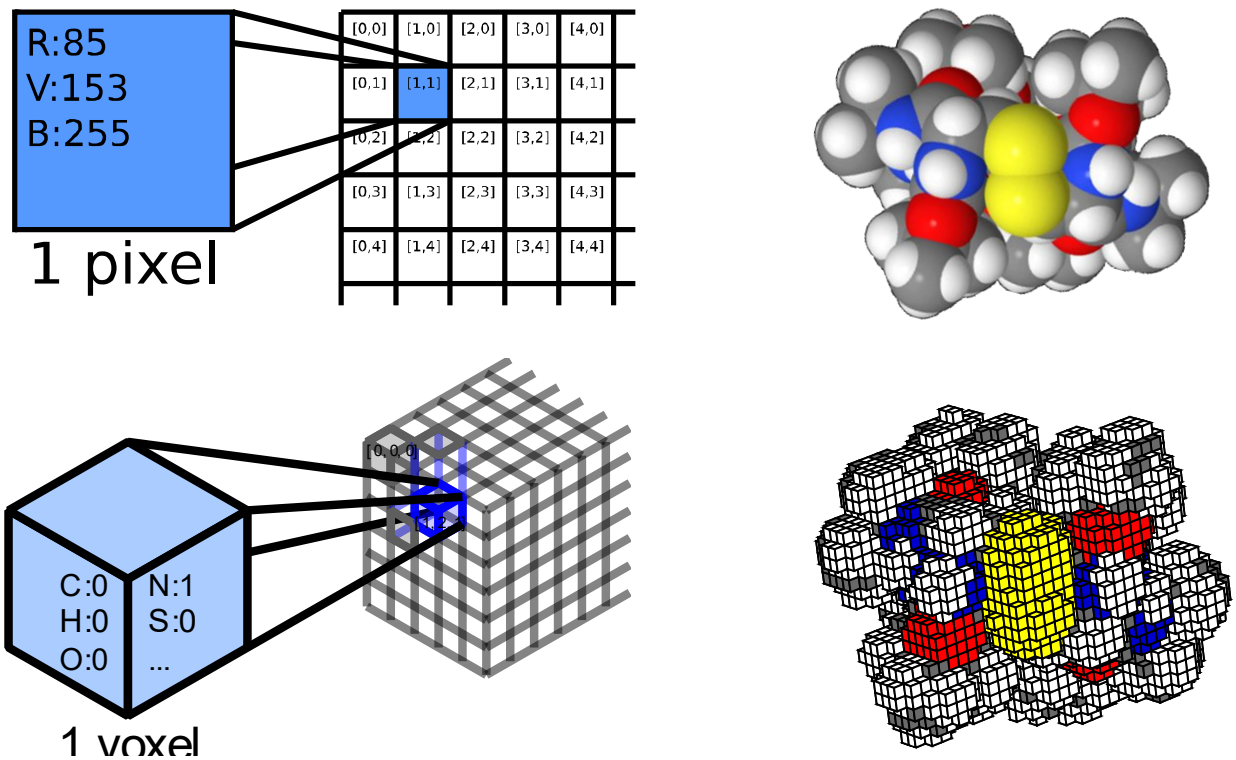


Figure 41 : Comparaison pixel/voxel (droite) et conversion des coordonnées atomiques d'une molécules en voxels (gauche).

Différentes résolutions et différents filtres de voxelisation peuvent être appliqués pour la création des voxels. Il faut bien distinguer les filtres appliqués sur les voxels (que nous appellerons filtre de voxelisation) qui expose donc des caractéristiques explicite et arbitraire, des filtres de convolutions qui sont des cartes de caractéristiques que le réseau créera et modifiera de lui-même. Le tout étant de choisir les filtres de voxels les plus appropriés pour éviter d'alourdir le réseau en termes de temps d'exécution et en mémoire de RAM. En effet si une trop grande résolution et un trop grand nombre de filtres sont appliqués, cela limitera la profondeur du réseau ce qui limitera potentiellement ses performances prédictives. Cela pourra aussi limiter sa vitesse d'apprentissage. C'est pourquoi dans un premier temps nous utiliserons une résolution relativement faible (comparé au réseau de Jiménez *et al.* (3)) afin d'avant tout valider l'approche de l'apprentissage par réseau de neurones. Cette étape permettra aussi de commencer à évaluer différentes architectures de réseau pour voir lesquelles semblent le mieux fonctionner et pour tester différents hyperparamètres.

Le processus de création d'un réseau de neurones optimal ne répond pas à une méthode prédéfinie. Il s'agit plus d'un tâtonnement empirique qui comprend de nombreux allers et retours entre les modifications des paramètres du réseau et des tests d'apprentissages. C'est pourquoi nous nous sommes inspirés, pour point de départ, de la même architecture de réseau utilisée par Jiménez *et al.* (3) qui utilise une même approche de voxelisation des structures (figure 42).

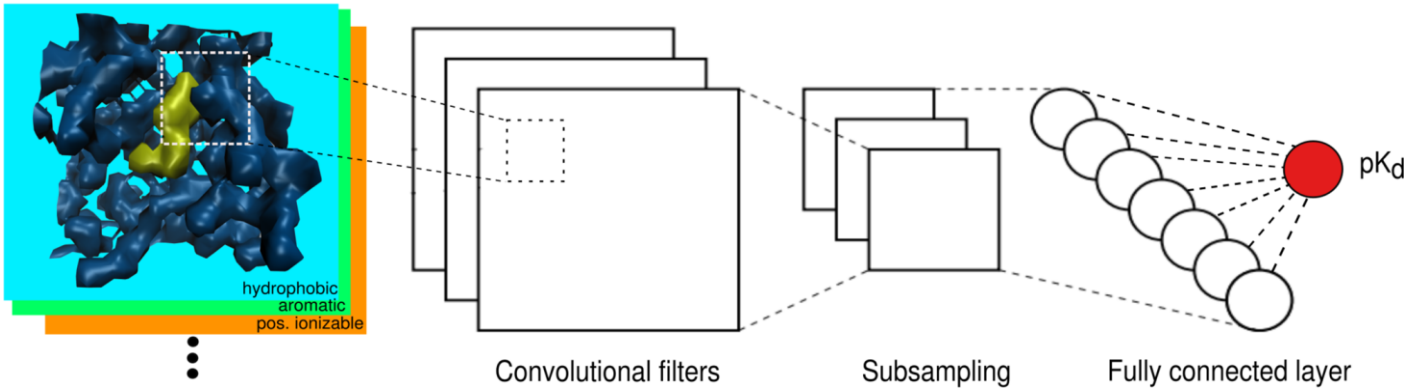


Figure 42 : Schéma simplifié du réseau de neurones convolutif utilisé par l'algorithme Kdeep. De gauche à droite : Les données d'entrée sont des filtres de voxelisation, traités ensuite par des filtres de convolution suivis par du sous-échantillonnage et finissant par une couche de neurones entièrement connectés qui délivrent une prédiction. Extrait de Jiménez et al. (3).

3.1. Annotation des atomes

Les complexes protéine-protéine sont lus à partir de fichiers PDB qui ont été précédemment prétraités pour uniformiser le format. Les atomes d'hydrogène sont rajoutés automatiquement puis l'ensemble des atomes sont annotés à l'aide d'OpenBabel en vue de leur traitement pour la voxelisation :

1. Hydrophobique : carbone lié avec exactement 4 autres atomes, uniquement de carbone ou d'hydrogène et hydrogène lié à ces carbones.
2. Aromatique (critère d'OpenBabel)
3. Accepteur de liaison hydrogène (critère d'OpenBabel)
4. Donneur de liaison hydrogène (critère d'OpenBabel)
5. La chaine protéique à laquelle il appartient
6. La charge (selon Gasteiger-Marsili)
7. Ionisable positivement (charge > 0)
8. Ionisable négativement (charge < 0)
9. Le numéro atomique
10. L'atome type selon le système *PATTY* (Programmable ATom TYping systems (11))
11. Le rayon de Van der Wall (selon les rayons de Van der Wall décrit par Alvarez, S. (12))
12. Les coordonnées spatiales (PDB)
13. Interaction : si l'atome se trouve à moins de 12 Å d'un atome issu de l'autre chaine

3.2. Voxelisation et réalisation de 2 sous-base de données

Les représentations voxelisées des protéines sont créées pour chaque complexe. Avant la création d'une représentation voxelisée de chaque complexe protéique, nous commençons par effectuer 5 rotations aléatoires à chacun de ces derniers. Les complexes sont ensuite centrés au niveau de leur zone d'interaction (grâce à l'annotation « interaction » des atomes, effectuée dans la partie précédente) dans une boîte de 100×100×100 Å. La zone d'interaction doit être entièrement incluse dans cette boîte sinon l'opération est répétée jusqu'à

obtenir 5 orientations différentes du complexe. S'il s'avère que la surface d'interaction est trop étendue pour tenir dans la boîte, le complexe est exclu.

Vient ensuite la voxelisation à proprement parler avec une résolution de un Å. Une résolution de 0.5 Å demanderait beaucoup trop de temps de calcul et de mémoire. Pour calculer les voxels, nous utilisons une grille de 100×100×100 nœuds, centrée en (0,0,0) et avec un pas de 1 Å, chaque nœud correspondant au centre d'un voxel. Une représentation voxelisée de chaque atome est calculée dans une grille de 100x100x100 par une fonction pseudo-gaussienne selon la méthode proposée par Ragoza *et al*. (5). En effet, la fonction s'approche d'une fonction gaussienne (c'est une fonction gaussienne jusqu'au rayon de van der Waals), mais elle est modifiée de telle sorte que la valeur soit nulle à partir de 1.5 fois le rayon de van der Wall.

Une fois que les représentations de tous les atomes du complexe est effectuée, des filtres de voxelisation sont alors créés en combinant ces 2 représentations :

- L'une en fonction de l'appartenance de l'atome à l'une des 2 chaînes.
- L'autre en fonction de l'état d'une des annotations (hydrophobique ou pas, accepteur de liaison d'hydrogène ou pas etc…).

Pour la plupart des filtres de voxelisation, la combinaison des atomes est réalisée en gardant le voxel ayant la valeur maximale. Pour le filtre de charge un poids est appliqué à chaque voxel des atomes avant de les combiner (multiplication par la charge). La combinaison est additive. Dans le cadre de cette étude et dans la perspective de futures expérimentations, nous avons réalisé des voxels ayant pour filtres de voxelisation :

1. Chaine A, chaine B : tous les atomes de chaque chaîne
2. Hydrogène
3. Carbone
4. Azote
5. Oxygène
6. Soufre
7. Hydrophobique
8. Aromatique
9. Accepteur de liaison hydrogène
10. Donneur de liaison hydrogène
11. Charge
12. Interaction

Pour finir, nous avons réalisé deux sous-bases de données voxelisées distinctes en fonction de la manière dont est intégré le traitement de la chaîne lors de la combinaison des filtres :

- Une que nous appelons « séparer » : les filtres sont alors réalisés en doublons (sauf pour le filtre « chaîne »), une combinaison de filtres pour chaque chaîne.
- Une autre que nous appelons « simple » : les filtres sont alors communs aux 2 chaînes sauf pour le filtre chaîne qui est distinct pour chaque chaîne. Ayant moins de filtres, elle sera plus légère et donc plus rapide à traiter/manipuler mais le risque étant que l'information soit moins accessible pour le réseau de neurones ou que l'information soit mal interprétée.

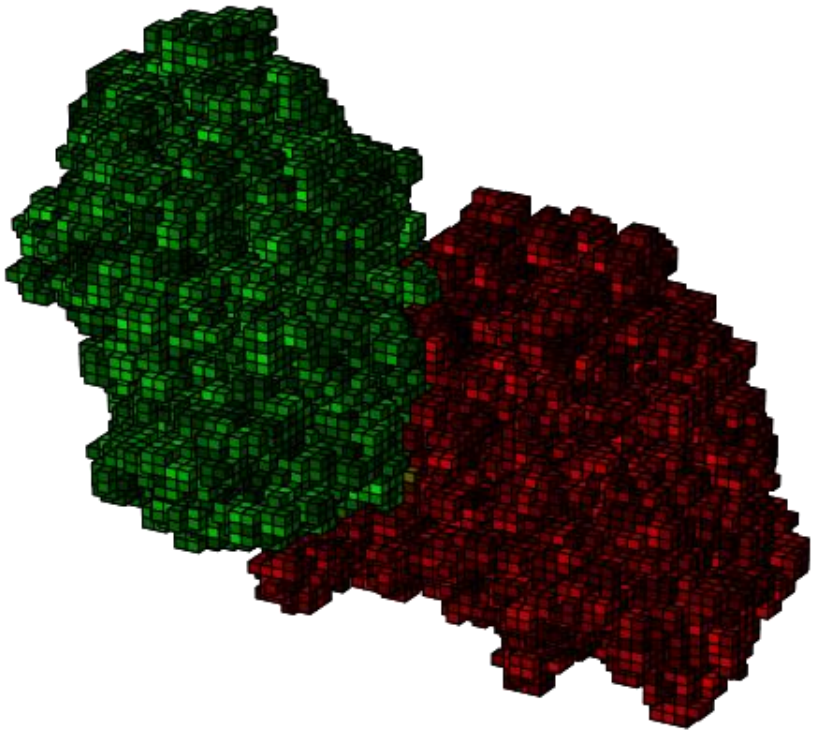


Figure 43 : Représentation du voxel d'un complexe issu de la sous-base de données « simple ». Le filtre actif « chaîne » représente l'encombrement stérique de la chaîne A et B respectivement de couleur verte et rouge. Les intensités de couleurs sont déterminées par la fonction pseudo-gaussienne.

Etant donné le faible nombre de complexes à traiter, nous procédons à une augmentation de données à l'aide de rotations comme il est courant de faire avec les images. Nous avons précédemment réalisé 5 rotations aléatoires différentes avant la voxelisation. Pour chaque rotation aléatoire voxelisée, il existe 24 rotations à 90° réalisables simplement en

modifiant l'ordre et le sens des axes, ce qui donne un total de 120 rotations différentes par complexe pour le traitement des voxels par les réseaux de neurones.

En s'inspirant de Jiménez *et al*. (3), nous utiliserons les filtres :

- chaînes (encombrement stérique),
- charge,
- hydrophobique,
- aromatique,
- accepteur de liaison hydrogène,
- donneur de liaison hydrogène.

Contrairement à Jiménez *et al*. les filtres « ionisable positivement » et « ionisable négativement » ont été remplacé par le filtre « charge ». De plus nous avons ajouté le filtre « Interaction ». Le filtre « métallique » n'a pas lieu d'être dans le cadre des interactions uniquement protéines-protéines contrairement à l'article de Jiménez *et al.* qui concernait les interactions protéine-ligand et pour lesquelles l'influence de certains ions métalliques peut avoir de fortes conséquences quant à l'énergie d'interaction.

Finalement, pour une même structure PDB nous avons 2 représentations :

- Une dite "séparée" ou chaque chaîne empaquette les filtres indépendamment de l'autre : 100×100×100 voxels avec une résolution de 1 Å sur 14 filtres (Chaîne 0, Hydrophobique, Aromatique, Accepteur d'hydrogène, Donneur d'hydrogène, Charge et Interaction ; Chaîne 1, Hydrophobique, Aromatique, Accepteur d'hydrogène, Donneur d'hydrogène, Charge et Interaction). Soit un total de 14 000 000 de paramètres en entrée du réseau et par complexe.
- Une dite "simple" ou les 2 chaînes sont incluses dans l'ensemble des filtres: 100×100×100 voxels avec une résolution de 1 Å sur 8 filtres (Chaîne 0, Chaîne 1, Hydrophobique, Aromatique, Accepteur d'hydrogène, Donneur d'hydrogène, Charge, Interaction). Soit un total de 8 000 000 de paramètres en entrée du réseau et par complexe.

Pour donner une idée plus concrète, les 14 000 000 de paramètres d'entrée sont en virgule flottante simple précision. Sachant qu'un paramètre en virgule flottante simple précision correspond à un nombre de 32 bits soit 4 octets par paramètre, cela fait un total de 56 mégaoctets par complexe (pour une seule rotation). Ce volume est comparable à une image

2D RVB (Rouge, vert, bleu soit 3 filtres) d'une résolution de 4320×4320 pixels. Il est important de rester dans des dimensions raisonnables pour que les temps de calcul demeurent réalistes. Par exemple passer d'une résolution de voxel de 1 à 0,5 Å multiplie par 8 le nombre de paramètres et donc multiplie tout autant la mémoire et le temps de calcul nécessaire pour traiter un complexe.

3.3. L'architecture du réseau de neurones artificiels

Tout comme Jiménez *et al*, nous avons adapté l'architecture de *SqueezeNet* (Iandola *et al.* (13)) pour réaliser la prédiction de l'énergie d'interaction à partir d'une représentation voxelisée des molécules. SqueezeNet est un réseau réalisé dans le cadre de la classification des images. Son principal avantage est sa légèreté. Sa principale innovation vient de ce que les auteurs ont appelé le *Fire Module* qui offre une solution efficace de diminution du nombre de paramètres tout en préservant la capacité d'apprentissage du réseau.

Un *Fire Module* est composé d'une partie *Squeeze*, composée d'une couche de convolution 1×1×1 et qui réduit le nombre de filtres pour condenser l'information. Suivie d'une partie *Expand*, composée de 2 couches de convolution parallèles (prenant toutes les deux la sortie de la couche de Squeeze) et dont les sorties sont ensuite concaténées.

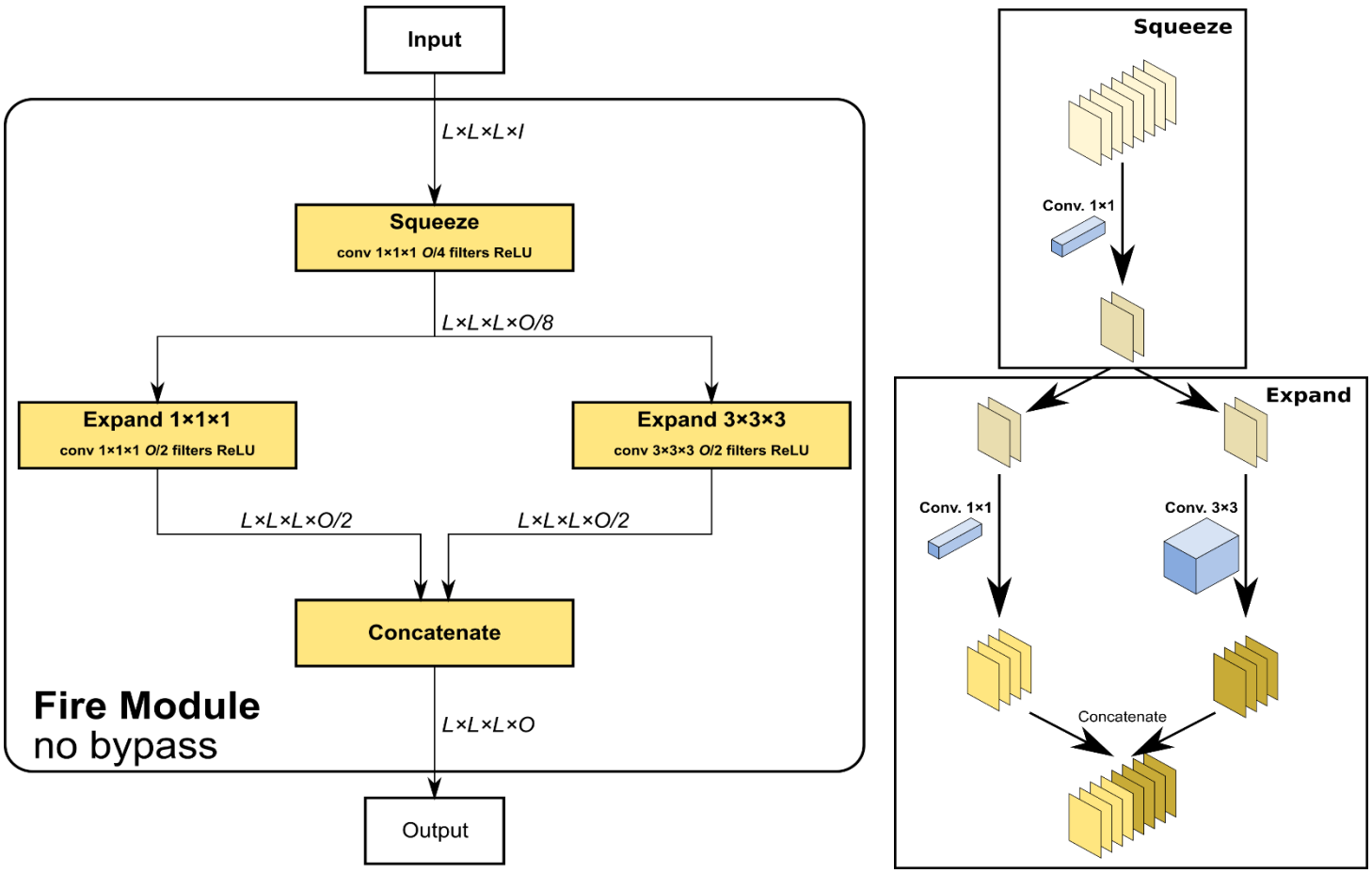


Figure 44 : Schéma explicatif d'un Fire Module. A gauche la représentation exhaustive du Fire Module tel qu'utilisé dans nos réseaux de neurones convolutifs traitant des structures 3D. A droite une version simplifiée du Fire Module s'appliquant à des données en 2D.

Cette façon de faire permet de réduire l'utilisation de la convolution 3×3×3 ainsi que le nombre de filtres en entrée et sortie de ceux-ci. En effet, une telle convolution demande plus de paramètres et de temps de calcul qu'une convolution 1×1×1, le nombre de filtres d'entrée et de sortie multipliant d'autant plus ces quantités. Cependant les convolutions 3×3×3 sont nécessaires pour identifier des motifs dans l'espace.

Les auteurs de *SqueezeNet* ont aussi proposé d'ajouter des *bypass*. Un *bypass* consiste à créer un pont dans le réseau en réalisant une somme de valeurs paires à paires entre la sortie d'une couche et celle d'une couche l'ayant précédé. C'est une solution fréquemment utilisé dans les réseaux de neurones profonds pour limiter le problème de la « disparition du gradients » qui survient quand le nombre de couches dans le réseau est très élevé. Dans notre modèle de réseau le *bypass* est réalisé entre l'entrée et la sortie du *Fire Module*. Cependant, cela nécessite que les deux sorties aient les même dimensions (même

taille et même nombre de filtres). Lorsque le nombre de filtres varie entre les deux sorties, il est possible d'utiliser une convolution 1x1x1 pour ajuster le nombre de filtres. C'est ce qui a été appelé un *bypass* complexe (figure 45).

Trois différentes architectures de réseau ont alors été proposées selon le type de *bypass* utilisé:

- Sans aucun *bypass*.
- *Bypass* simple : seuls les *Fire Modules* sans modification du nombre de filtres possèdent un bypass reliant l'entrée à la sortie.
- *Bypass* complexe : en plus des *Fire Modules* sans augmentation du nombre de filtre, un *bypass* est réalisé sur les *Fire Modules* avec augmentation du nombre de filtres en réalisant une convolution 1×1×1 pour augmenter le nombre de filtres de la couche précédant le *Fire Module*.

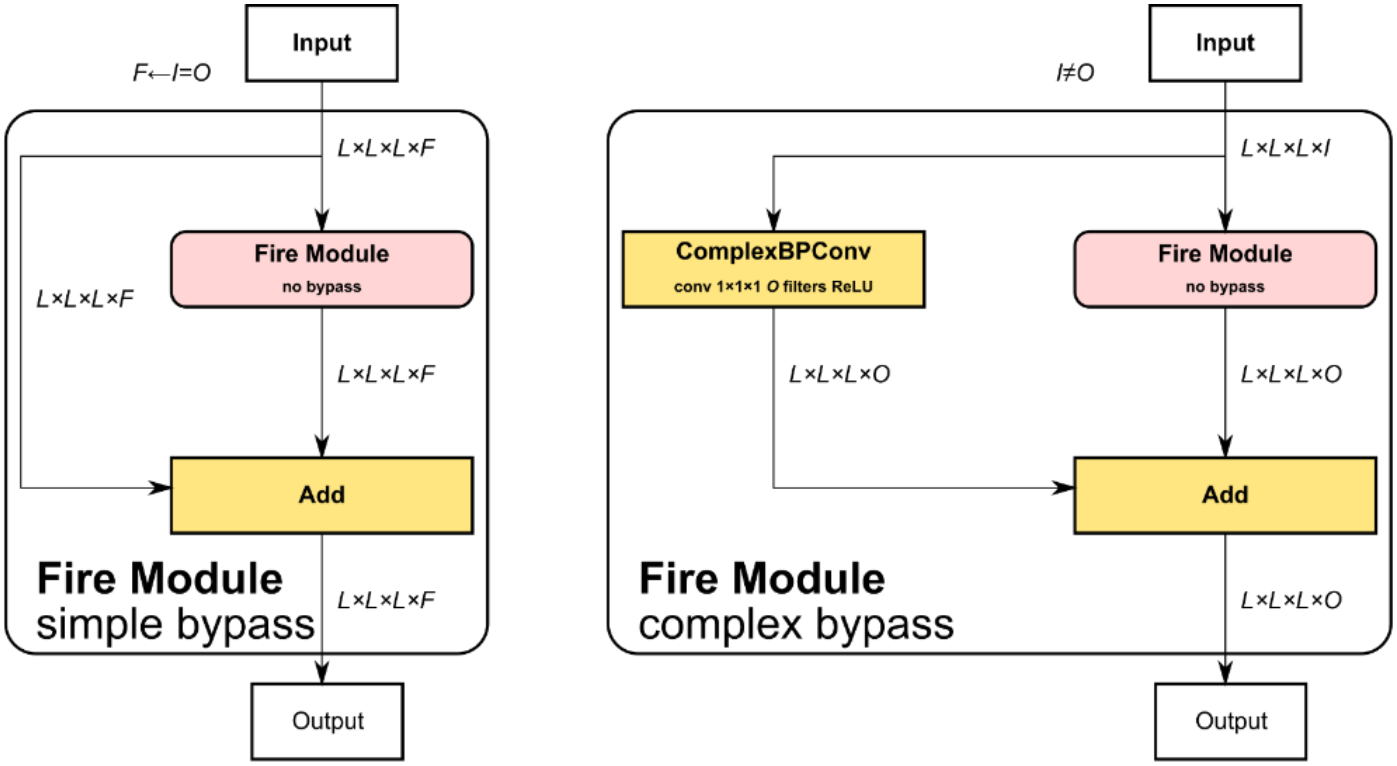


Figure 45 : Schéma explicatif d'un bypass simple et complexe. A gauche : un bypass simple dans lequel sont directement sommés la couche d'entrée du Fire Module avec sa couche de sorite. A droite : un bypass complexe pour lequel une couche de convolution a été appliquée à la couche d'entrée afin d'harmoniser son nombre de filtres à la couche de sortie du Fire Module pour permettre sa sommation.

Tous les 2 *Fire Modules*, le nombre de filtres est augmenté d'une valeur constante. Tous les 4 *Fire Modules*, un sous-échantillonnage (pooling) est réalisé (juste après une augmentation du nombre de filtres).

Le réseau est composé d'une répétition d'un assemblage que nous avons appelé « *SqueezeNet Module* ». Un *SqueezeNet Module* est composé de 4 *Fire Modules* et d'une fonction de sous-échantillonnage (pooling) entre le 3ème et le 4ème *Fire Module*. Durant le 1er et le 3ème *Fire Module*, le nombre de filtres est augmenté d'une valeur constante.

Nous avons donc un *SqueezeNet Module* composé ainsi (voir figure 46) :

1. *Fire Module* + augmentation du nombre de filtres
2. *Fire Module*
3. *Fire Module* + augmentation du nombre de filtres
4. *Pooling*
5. *Fire Module*

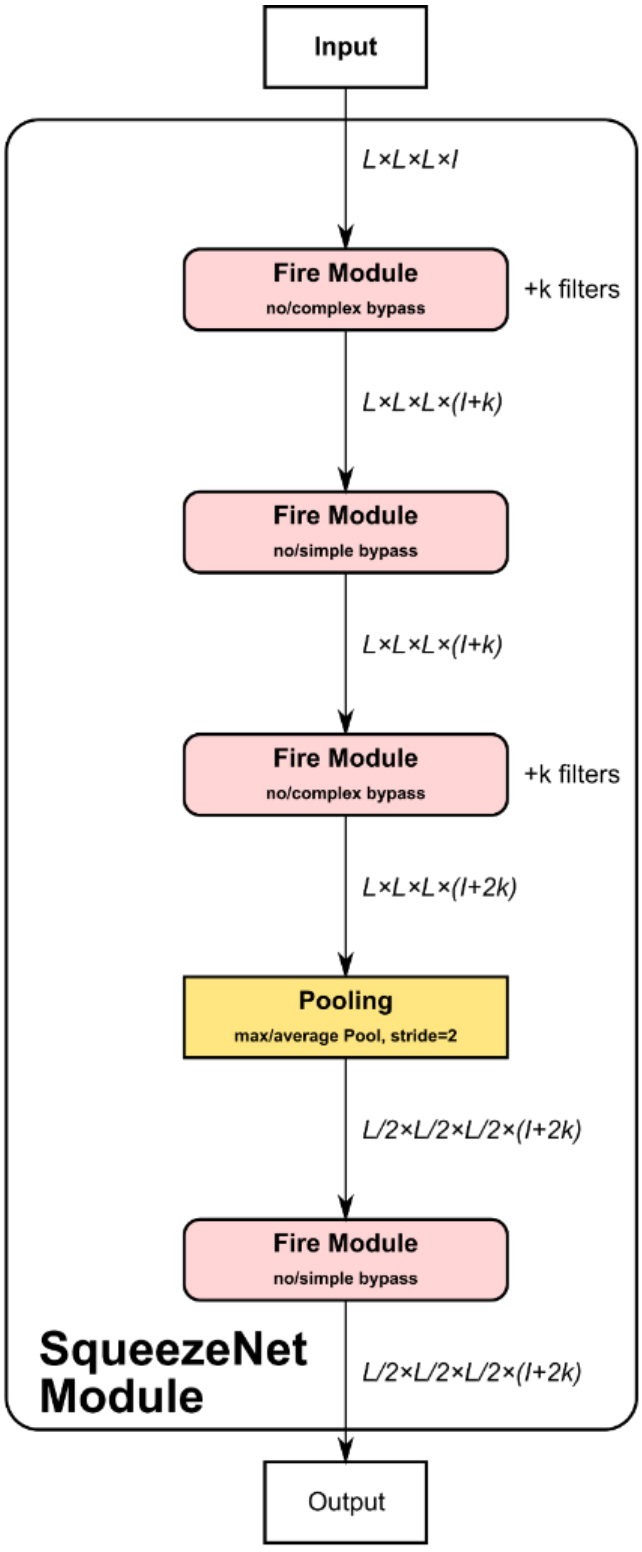


Figure 46 : Schéma explicatif d'un SqueezeNet Module.

3.4. Notre réseau *SqueezeNet*

Nous appliquons une première convolution de 5×5×5 avec un pas de 2 (sous-échantillonnage intelligent) comme Jiménez et al.. S'ensuit 4 blocs de *SqueezeNet Module* (choix purement arbitraire, c'est un paramètre à faire varier). Enfin, le réseau se termine par une convolution réduisant à un seul filtre avant de finir par une couche dense sans activation (figure 47).

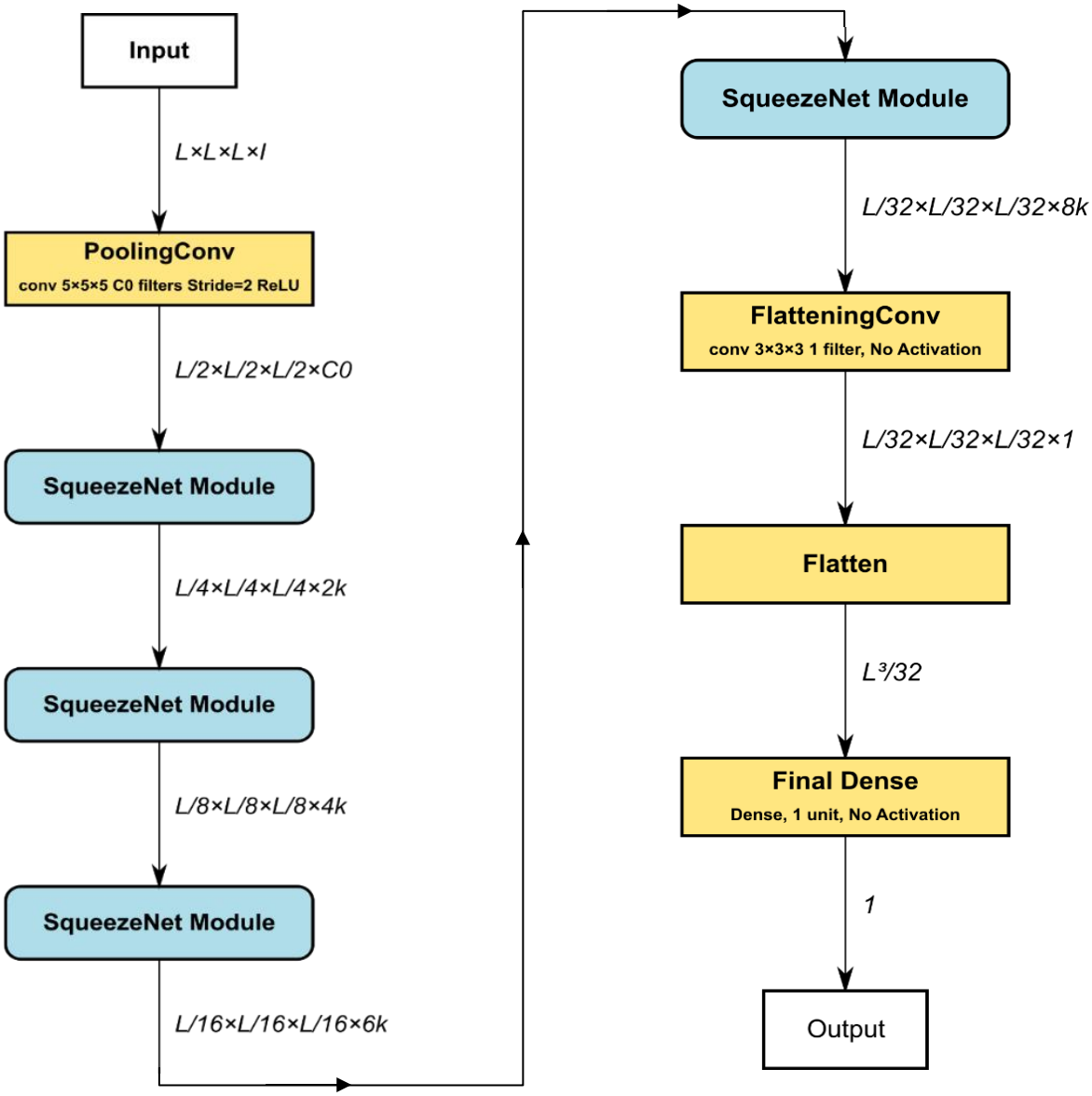


Figure 47 : Schéma explicatif de l'architecture générique de notre réseau de neurones convolutif.

Nous avons fait varier :

- L’entrée : selon si nous utilisons les voxels avec les chaînes « simple » ou « séparer ».
- La fonction de sous-échantillonnage : par valeur moyenne ou par valeur maximale.
- Le Bypass : aucun, simple ou complexe.
- La largeur du réseau (nombre de filtres de convolution par couche) : Normale : qui correspond à une première couche de convolution avec 32 filtres puis augmentation à 64 filtres tous les 1 *Fire Module* sur 2. Large : première couche de convolution avec 96 filtres puis augmentation de 128 filtres tous les 1 *Fire Module* sur 2. Ces derniers ont très rarement eu le temps de finir (à cause d’un nombre trop élevé de paramètres), ils ont donc été mis de côté.
- Le taux d'apprentissage : celui par défaut est de 0.001 (taux par défaut dans Keras, le reste des hyperparamètres est laissé par défaut), un taux d’apprentissage à 0.00001 a aussi été testé.
- Les époques : correspondent au passage d'une rotation (parmi 120) de chaque complexe. L'ordre des complexes et la rotation est aléatoire. Cependant, il est garanti que chaque complexe passe exactement une fois par époque et que les 120 rotations possibles soient toutes passées au bout de 120 époques.

La fonction d’activation utilisée est systématiquement la fonction ReLu (14) comme dans SqueezeNet. Nous avons utilisé l'optimiseur Nadam (15). Le calcul de la fonction de perte est effectué par le calcul de l’erreur quadratique moyenne entre les valeurs expérimentales et les valeurs prédites. Le calcul du gradient de l’erreur est effectué par des mini-lots de 4 complexes. Des lots plus grands sont incompatibles avec les capacités matérielles des GPU (Graphics Processing Unit) disponibles (saturation de la mémoire) et des lots plus petits diminuent la vitesse d’apprentissage. Les réseaux ont tourné pendant 360 époques (l'intégralité des rotations est passé 3 fois). Une prédiction est réalisée sur l'intégralité de la base de données de test (les 79 complexes avec leurs 120 rotations chacun) en fin d'apprentissage afin de comparer la performance des réseaux.

Pour les calculs d’apprentissage nous avons utilisé les serveurs de calculs de l'IN2P3 (Institut national de physique nucléaire et de physique des particules) comprenant des machines de 8 CPU et 2 GPU K80 de 24GB. Le langage informatique dans lequel les réseaux ont été codés est Python. En particulier grâce à la bibliothèque Keras qui permet d’interagir facilement avec les algorithmes d’apprentissage automatique, dans notre cas Tensorflow.

4. Résultats

Nous avons entrainé près d'une centaine de réseaux de neurones avec différentes combinaisons d'hyperparamètres et de bases de données (chaîne séparée et chaîne simple). La performance de l'algorithme est évaluée en calculant le coefficient de corrélation de Pearson sur la base de données de validation et de test. Après avoir déterminé un nombre maximum de *SqueezeNet Modules* (profondeur du réseau) et de filtres convolutif (largeur du réseau) que nous pouvions traiter avec le matériel disponible, nous avons fait varier les autres hyperparamètres. En l'occurrence le réseau ayant obtenu les meilleures performances d'apprentissage est composé de 4 *SqueezeNet Modules* avec une première convolution de 32 filtres convolutif pour chaque filtre de voxel (chaînes, charge, hydrophobique, aromatique, accepteur de liaison hydrogène et donneur de liaison hydrogène). Il en est ressorti plusieurs faits remarquables aboutissant à un réseau de neurones plus performant tel que :

- L'augmentation du nombre de filtres de convolution.
- L'augmentation du nombre de *SqueezeNet Modules*.
- La diminution du taux d'apprentissage (un taux de 0.00001 est plus efficace que celui par défaut de 0.001).
- L'usage de la base de données voxelisée par chaîne simple.
- L'usage d'un sous échantillonnage par valeur moyenne (average pooling) est plus efficace qu'un sous échantillonnage par valeur maximale (maximum pooling).
- L'usage d'un *Bypass* complexe donne de meilleurs résultats que celui d'un *Bypass* simple qui lui-même donne de meilleurs résultats que sans *Bypass*.

Etant donnée la grande diversité des résolutions des complexes structurales (de 4 Å à <1 Å) nous avons testé l'influence d'une minimisation des structures. Le but de cette minimisation est de permettre de limiter quelques artefacts concernant les distances entre certains atomes. En particulier les aberrations de distance telles que les collisions d'encombrement stériques. Mais la minimisation a aussi pour but d'homogénéiser les distances entre les atomes de l'interface pour s'émanciper des différences qui pouvaient être liées aux différentes résolutions des structures entre elles. Les résultats ont montré un meilleur coefficient de corrélation pour les réseaux qui apprenaient sur la base de données minimisées (0.49 contre 0.53 sur la base de données de validation) ce qui confirme la pertinence de cette approche. Nous avons alors mis en production le réseau de neurones le plus performant pour

le tester sur la base de données de test et pour pouvoir le comparer avec les autres algorithmes. Le réseau nous donne cette fois-ci un coefficient de corrélation de 0.44. Les explications de cette diminution seront discutées dans la partie suivante. Ce score de corrélation nous permet de comparer notre algorithme avec d'autres déjà existants. Pour cela nous nous sommes servis d'un agrégateur de fonctions scores fourni par le serveur web CCharPPI (Computational Characterisation of Protein-Protein Interactions) (16). Cette comparaison est synthétisée dans la figure 48.

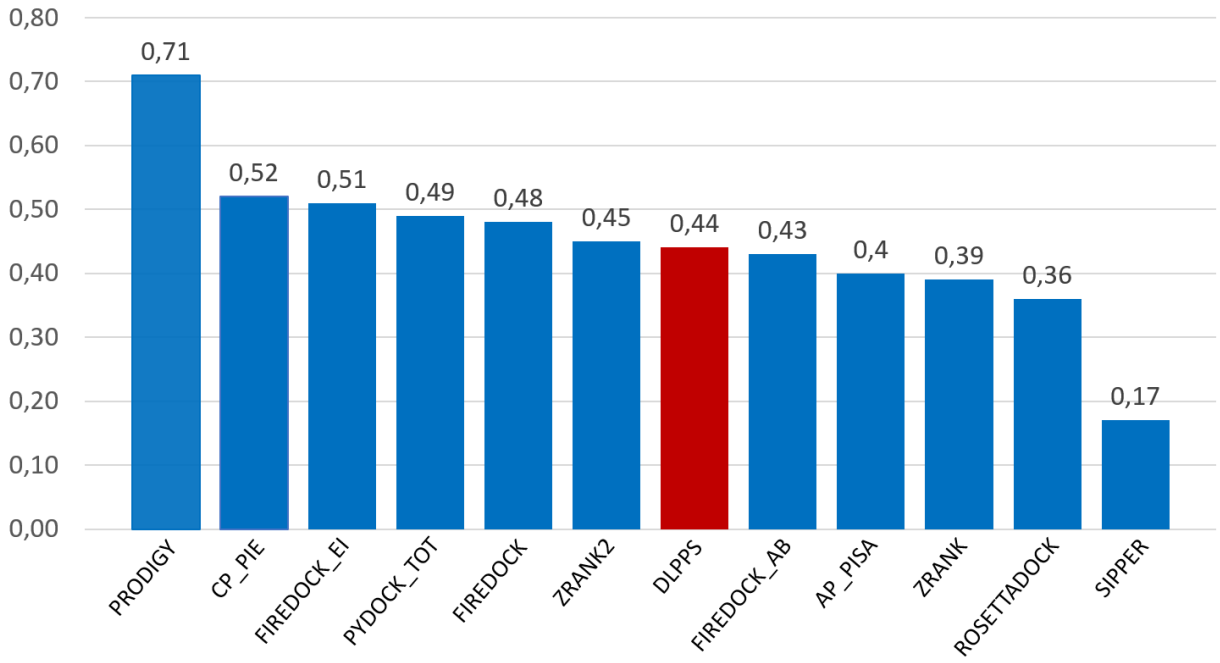


Figure 48 : Coefficient de corrélation de Pearson entre les données expérimentales d'énergie libre d'association des complexes et les prédictions de différents algorithmes. Notre algorithme DLPPS est en rouge.

Comme nous pouvons le constater PRODIGY domine les autres algorithmes, tandis que DLPPS se situe dans la moyenne des performances du reste des algorithmes. Pour comparer plus précisément nos résultats avec l'algorithme de référence, PRODIGY, nous avons voulu voir l'influence que pouvait avoir la minimisation des structures de la base de données de test sur ce dernier. Nous avons pu constater une diminution du coefficient de corrélation passant de 0.71 pour la base de données de test non minimisée, à 0.64 pour celle minimisée. Cette diminution nous a fait soupçonner un potentiel biais de sur-apprentissage dans la conception de l'algorithme PRODIGY. Pour vérifier cette hypothèse, nous avons alors testé la performance de PRODIGY sur notre base de données de validation. Le coefficient de

corrélation tombe alors à 0.27 pour la base de données non minimisée et 0.28 pour celle minimisée. Il n'y a presque pas d'écart entre les structures minimisées et non minimisées ce qui suggère que la minimisation n'a que peu d'influence sur la performance de l'algorithme. L'écart que nous constatons alors entre les performances de PRODIGY sur la base de données de test minimisée et non minimisée est donc probablement lié à ce sur-apprentissage. Nous avons donc testé les autres algorithmes pour voir s'ils souffraient aussi d'une forme de sur-apprentissage. Les données synthétisées dans la figure 49 permettent de constater un phénomène de sur-apprentissage chez tous les algorithmes mis à part le nôtre, DLPPS, ainsi que SIPPER (Scoring by Intermolecular Pairwise Propensities of Exposed Residues).

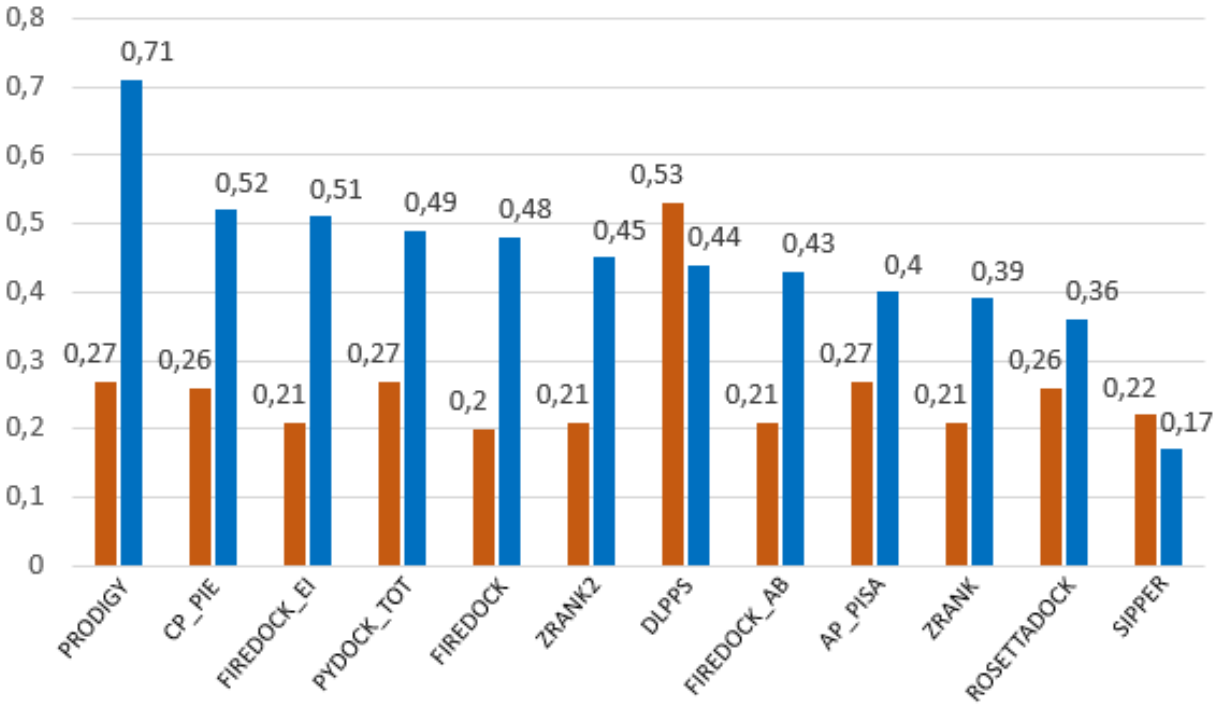


Figure 49 : Comparaison des coefficients de corrélation entre la base de données de test (bleu) et la base de données de validation (rouge) des différents algorithmes.

En analysant plus en détail les résultats de prédiction de notre algorithme DLPPS, nous pouvons constater que la plupart des grands écarts d'énergie libre de liaison (ΔΔG > ±0.5 kcal.mol^{-1}) entre les valeurs expérimentales et les valeurs prédites sont la cause de seulement 10 complexes sur les 79 de la base de données de test (figure 50). Nous ne pouvons pas connaître avec certitude les raisons de ces écarts de prédiction. Généralement, cela se produit dans les algorithmes d'apprentissage automatique lorsque la base de données n'est pas suffisamment fournie en exemples. Nous pouvons supposer que les structures concernées possèdent des particularités en termes de type d'interaction/d'architecture qui sont peu représentés dans la base de données d'entrainement.

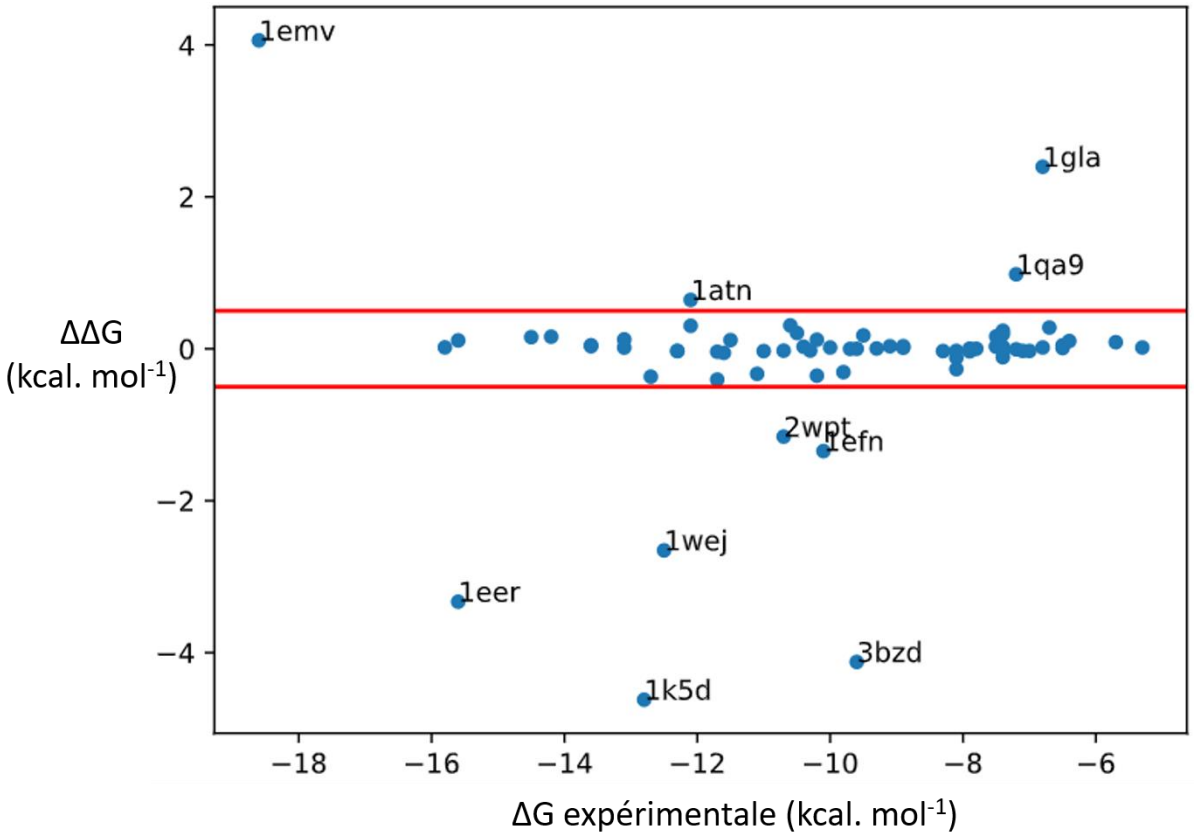


Figure 50 : Ecart entre l'énergie libre de liaison expérimentale et l'énergie libre de liaison prédit par DLPPS. Les lignes rouges délimitent un écart > ±0.5 kcal.mol^{-1}.

Nous avons noté qu'un seul complexe de type anticorps-antigène, sur les 10 que comprend la base de données de test, ne fait partie des complexes ayant un écart d'énergie libre de liaison prédit supérieur à ±0.5 kcal.mol^{-1} avec l'énergie libre de liaison observée expérimentalement. Etant donné que ce type d'interaction est très spécifique et que les anticorps présentent une structure globale similaire entre elles, il nous a semblé pertinent d'entrainer des réseaux de neurones sur une base de données spécifique. Cela a pour

conséquence de diminuer drastiquement la taille des bases de données nouvellement créées mais elles seront aussi beaucoup plus homogènes et spécifiques que les précédentes. Le but est d'améliorer les performances du réseau sur une problématique spécifique en lui permettant de se concentrer sur un type d'interaction en particulier sans être biaisé par les autres. De plus, les interactions anticorps-antigène sont d'une importance particulière dans le cadre de l'immunité et dans la problématique des vaccins.

Il a donc été créé trois bases de données anticorps-antigène. Une de 179 complexes pour l'entrainement du réseau, la seconde de 35 complexes servant de base de données de validation une troisième de 30 complexes pour tester l'algorithme et le comparer. Pour limiter le sur-apprentissage nous avons multiplié par 5 le nombre de rotations par structure par rapport à la base de données totale. Nous nous sommes servis de l'architecture du réseau de neurones qui donnait les meilleurs résultats à la différence que nous avons réduit le nombre de *SqueezeNet Modules* à 2 pour alléger le réseau. Nous obtenons cette fois-ci un coefficient de corrélation de 0.50 entre les énergies libres de liaison expérimentales et celles prédites par notre algorithme. Les résultats comparatifs sont synthétisés dans la figure 51.

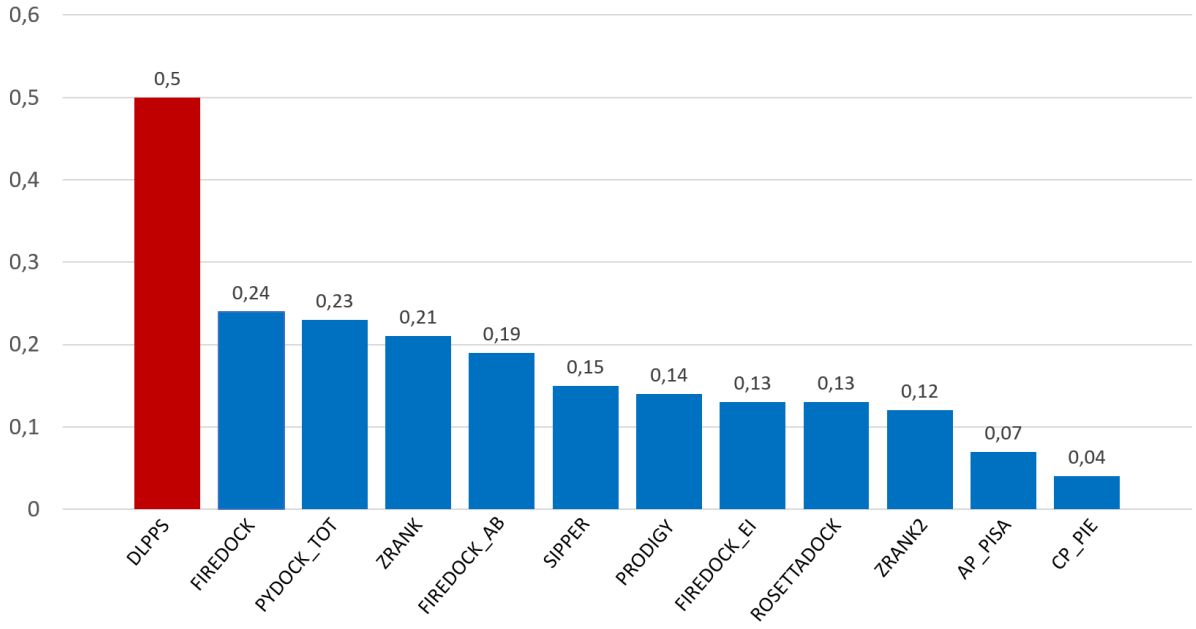


Figure 51 : Coefficient de corrélation de Pearson entre les données expérimentales d'énergie libre d'association des complexes antigène-anticorps et les prédictions de différents algorithmes. Notre algorithme DLPPS est en rouge.

5. Discussions

L'ensemble des réseaux donnait un coefficient de corrélation de Pearson compris entre 0.3 et 0.5 sur les données de validations. En dehors de toute considération de performance, nous avons pu constater que les différents réseaux de neurones donnaient des résultats souvent proches les uns des autres pour une même structure. Cela signifie que les caractéristiques extraites par chaque réseau sont sensiblement identiques mais que leurs différences se jouent plus du côté de la précision et de la vitesse d'apprentissage. Le meilleur réseau a atteint un coefficient de 0.53 pour la base de données de validation mais pour la base de données de test ce coefficient est tombé à 0.44. Cet écart peut s'expliquer par l'inhomogénéité de cette troisième base de données vis-à-vis des bases de données d'entrainement et de validation. D'une part, la base de données de test est plus petite que les 2 autres (79 complexes pour la base de données de test, 108 pour celle de validation et 1044 pour celle d'entrainement). D'autre part elle n'a pas été créée pour répondre à ce critère d'homogénéité puisqu'elle était initialement réalisée pour servir de base de données de référence pour les algorithmes d'amarrages protéine-protéine (8,9). La contrainte première de cette base de données est donc d'avoir accès à des complexes dont les structures de chaque partenaire sont également disponibles sous forme libre ce qui restreint le nombre de complexes disponibles et empêche une sélection aléatoire de ces derniers.

Il est très probable que la surface d'interaction seule ne soit tout simplement pas suffisante pour déterminer avec précision l'énergie d'interaction d'un complexe et cela pour plusieurs raisons.

- Une première raison d'ordre thermodynamique. Chercher à connaître la différence d'entropie entre la forme liée d'un complexe et la forme libre des partenaires sans connaître la structure des formes libres aboutit indubitablement à des approximations.
- Une seconde raison (qui peut être liée à la première) est liée au rôle que peuvent jouer les surfaces non interactives des protéines dans l'interaction protéine-protéine comme cela est décrit dans l'article de Kastritis *et al* (17). L'effet des surfaces non interactives sur l'affinité de liaison peut être expliqué par des contributions électrostatiques à longue portée mais aussi par des interactions surface-solvant.

Un des points positifs de cette étude est que grâce à l'augmentation des données effectuée par les rotations, nous n'avons pas eu à constater le phénomène de sur-apprentissage, ce qui constituait une première crainte face aux quantités de données dont nous disposions. C'est, en partie, ce qui nous a motivé pour créer une nouvelle base de données plus spécifique d'un type de complexe : les complexes de type anticorps-antigène. Cette sous-base de données, bien plus petite que la première, a permis d'atteindre un coefficient de corrélation 0.50 ce qui est légèrement meilleur que pour la base de données mélangeant tous les types de complexes. Cela laisse présager un certain potentiel d'amélioration de l'algorithme au fur et à mesure que les bases de données s'étofferont.

L'un des principaux problèmes de notre approche est que son coût de mise en œuvre, en ce qui concerne la puissance informatique, est très élevé et nécessite du matériel spécifique. Le poids des voxels limite le nombre d'époques ainsi que la taille du réseau de neurones que nous pouvons nous permettre de créer tant en termes de profondeur (son nombre de couches) qu'en termes de largeur (son nombre de filtres). C'est pourquoi nous n'avons pas la prétention d'avoir obtenu le meilleur de ce qui pouvait être fait en matière de réseaux de neurones convolutifs pour traiter cette problématique.

Il est intéressant de mettre en perspective les performances de notre algorithme, DLPPS, avec le fait que cette méthode d'apprentissage automatique permet de contrôler le processus de sur-apprentissage et donc d'éviter de tomber dans ce piège. Ceci n'est pas le cas des méthodes empiriques pour lesquelles le processus de recherche et d'optimisation des paramètres est directement ajusté pour maximiser le coefficient de corrélation sur une unique base de données. Cela aboutit intrinsèquement à une forme de sur-apprentissage comme nous le montre, par exemple, la perte de performance de l'algorithme PRODIGY sur notre base de données de validation (0.27 de coefficient de corrélation pour la base de données de validation contre 0.71 pour celle de test). Il est très probable que si presque tous les autres algorithmes souffrent de sur-apprentissage, cela vient du fait que leurs paramètres ont été calibrés sur la base d'une unique base de données. Face à ce constat, il n'est pas improbable que la performance réelle de notre algorithme de prédiction soit proche voire meilleure que celle des autres algorithmes que nous avons testés.

6. Conclusions et perspectives

Dans ce dernier chapitre nous avons décrit l'élaboration et la mise en œuvre d'une nouvelle approche pour la prédiction de l'énergie d'interaction protéine-protéine. Nous pouvons constater que cette approche est pertinente et peut donner des résultats proches voir meilleurs que ceux obtenus par les algorithmes préexistants faisant référence en la matière. Toutefois, cette approche reste largement perfectible et nous n'avons pas la prétention d'en avoir tiré le meilleur potentiel, ce qui laisse une certaine marge de progression pour l'avenir. Nous pouvons lister un certain nombre d'améliorations propres à l'utilisation des réseaux de neurones convolutifs et qui pourrait augmenter sa capacité d'apprentissage :

- L'usage de nouvelles architectures de réseau tel que le *Fire Module*. L'apprentissage automatique par réseau de neurones est un domaine de recherche en expansion extrêmement rapide. A ce titre, de nouvelle innovation dans le domaine des architectures de réseau, peuvent rendre obsolètes les anciennes architectures d'une année à l'autre.
- L'usage de voxels plus petits (par exemple avec un pas de 0.5 Å, comme pour Kdeep (3)) afin d'améliorer la résolution des représentations des structures que nous donnons à notre réseau.
- L'augmentation de la taille de la boîte afin de prendre en compte l'ensemble de la structure des complexes. Cela permettrait au réseau de prendre en compte d'autre propriétés agissant sur l'énergie d'interaction protéine-protéine que celles concernant la surface d'interaction (17).

Concernant les deux derniers points, leur exploration peut nécessiter une augmentation drastique des capacités de calcul. En effet, doubler la taille de la boîte ou doubler la résolution des voxels multiplie par 8 la quantité de données à traiter et combiner ces deux paramètres multiplierait par 64 le poids en calcul par rapport à la configuration actuelle de notre algorithme. Cependant, avec l'évolution des capacités de calcul des appareils informatiques et l'optimisation de ces appareils en vue de leur utilisation pour les réseaux de neurones, ces barrières peuvent être levées plus rapidement qu'il n'y paraît.

Indépendamment du type de réseau de neurones que nous utilisons (convolutif ou non) nous pourrions augmenter certains hyperparamètres tel que :

- Le nombre d'époques jusqu'à atteindre le phénomène de sur-apprentissage.
- La largeur (nombre de filtres par couche) ainsi que la profondeur (nombre de couches) du réseau afin d'augmenter la capacité d'abstraction du réseau.

Ces paramètres aussi impliquent une plus grande demande en capacité de calcul mais contrairement aux deux précédents, leur augmentation augmente « seulement » d'un facteur linéaire la charge de calcul.

Enfin plus généralement, l'augmentation des bases de données structurales associée à l'augmentation de résultats expérimentaux concernant l'énergie d'interaction protéine-protéine, permettra d'étoffer d'année en année la base de données PDBbind. Cela permettra d'améliorer l'exhaustivité des familles de protéines alimentant les algorithmes d'apprentissage, augmentant d'autant plus leur capacité d'apprentissage. Concernant les interactions protéine-protéine, PDBbind est passé de 1053 structure en 2009 à 2541 en 2019.

Un domaine de recherche émergeant dans le domaine des réseaux de neurones consiste en l'analyse et l'interprétation des raisons ayant abouti à un résultat. Ce genre d'approche est non seulement utile pour optimiser plus intelligemment et rapidement un réseau de neurones mais aussi pour extraire de l'information supplémentaire des résultats de prédiction. Parmi les exploitations que nous pourrions extraire de ce genre de donnée, nous pouvons citer comme exemple l'identification des résidus jouant un rôle crucial dans une interaction (hotspot en anglais). Ce genre d'information est primordial dans le cadre de la conception rationnelle de protéines et pourrait être directement extrait des résultats de ce type d'algorithme.

Nous pouvons donc constater qu'il existe une multitude de facteurs qui permettront d'améliorer les performances et les fonctionnalités de l'approche de DLPPS dans le futur. Il se peut malgré tout que l'étude de l'interface seule soit intrinsèquement insuffisante pour prédire avec précision l'affinité de liaison protéine-protéine. L'avenir de la prédiction *in-silico* de l'énergie d'affinité protéine-protéine se fera probablement en combinant plusieurs approches complémentaires entre elles. Nous espérons néanmoins avoir démontré dans ce dernier chapitre que l'apprentissage automatique par réseaux de neurones convolutifs peut être une approche pertinente et prometteuse dans la réalisation de cette tâche.

BIBLIOGRAPHIE DU CHAPITRE 6

1. Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein–protein complexes. Levitt M, rédacteur. eLife. eLife Sciences Publications, Ltd; 20 juill 2015;4:e07454.

2. Chen F, Liu H, Sun H, Pan P, Li Y, Li D, et al. Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein-protein binding free energies and re-rank binding poses generated by protein-protein docking. Phys Chem Chem Phys PCCP. 10 août 2016;18(32):22129-39.

3. Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. J Chem Inf Model. 26 2018;58(2):287-96.

4. Zheng L, Fan J, Mu Y. OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction. ACS Omega. 16 sept 2019;4(14):15956-65.

5. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-Ligand Scoring with Convolutional Neural Networks. J Chem Inf Model. 24 avr 2017;57(4):942-57.

6. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem. 3 juin 2004;47(12):2977-80.

7. Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. Bioinformatics. Oxford Academic; 1 déc 2016;32(23):3676-8.

8. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AMJJ, et al. A structure-based benchmark for protein–protein binding affinity. Protein Sci. 2011;20(3):482-91.

9. H H, T V, J J, Z W. Protein-protein docking benchmark version 4.0. Proteins. 1 nov 2010;78(15):3111-4.

10. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Dans: Pereira F, Burges CJC, Bottou L, Weinberger KQ, rédacteurs. Advances in Neural Information Processing Systems 25 [En ligne]. Curran Associates, Inc.; 2012 [cité le 18 sept 2020]. p. 1097–1105. Disponible: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

11. Bush BL, Sheridan RP. PATTY: A programmable atom type and language for automatic classification of atoms in molecular databases. J Chem Inf Comput Sci. American Chemical Society; 1 sept 1993;33(5):756-62.

12. Alvarez S. A cartography of the van der Waals territories. Dalton Trans. The Royal Society of Chemistry; 28 mai 2013;42(24):8617-36.

13. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size.

ArXiv160207360 Cs [En ligne]. 4 nov 2016 [cité le 29 sept 2020]; Disponible: http://arxiv.org/abs/1602.07360

14. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. Dans: 2011 [cité le 29 sept 2020]. p. 315-23. Disponible: https://hal.archives-ouvertes.fr/hal-00752497

15. Dozat T. Incorporating Nesterov Momentum into Adam. 18 févr 2016 [cité le 29 sept 2020]; Disponible: https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ

16. Moal IH, Jiménez-García B, Fernández-Recio J. CCharPPI web server: computational characterization of protein–protein interactions from structure. Bioinformatics. Oxford Academic; 1 janv 2015;31(1):123-5.

17. Kastritis PL, Rodrigues JPGLM, Folkers GE, Boelens R, Bonvin AMJJ. Proteins Feel More Than They See: Fine-Tuning of Binding Affinity by Properties of the Non-Interacting Surface. J Mol Biol. 15 juill 2014;426(14):2632-52.

CONCLUSION GENERALE

Au cours des travaux de cette thèse, nous avons pu constater tout l'intérêt des méthodes de bio-informatiques structurales pour répondre à des problématiques de conception et d'optimisation d'épitopes vaccinales. Certaines de ces méthodes ont servi à modéliser des structures n'ayant pas encore été résolu expérimentalement (comme pour la PSVB). D'autres méthodes, comme le calcul de l'affinité d'interaction protéine-protéine, nous ont permis de cartographier virtuellement des épitopes de la grippe vis-à-vis d'un paratope d'intérêt. Les résultats obtenus sur ce dernier sujet offraient, malgré de bons résultats, des performances inférieurs à ceux que l'on obtient lorsque nous étudions les interactions protéine-ligand. C'est pourquoi nous avons cherché à élaborer notre propre méthode de calcul de l'affinité d'interaction protéine-protéine pour tenter de surperformer les méthodes préexistantes.

Nous avons récemment pu voir les excellentes performances que peuvent donner l'usage des méthodes d'apprentissage automatique en profondeur sur des questions fondamentales du champ de la bio-informatique structurale. Parmi ces exemples marquants, nous pouvons citer la prédiction de structure monomérique avec l'algorithme AlphaFold mais aussi la prédiction de l'interaction protéine-ligand avec Kdeep. Fort de ce constat, nous nous sommes inspirés de cette dernière approche pour élaborer notre propre algorithme de prédiction basé sur des réseaux de neurones convolutifs mais cette fois-ci adapté à l'interaction protéine-protéine.

Les premiers résultats obtenus peuvent sembler médiocres en comparaison des autres algorithmes ayant la même fonction. Cependant en élargissant la base de données permettant de faire le comparatif, nous nous sommes aperçus que contrairement à notre algorithme, les autres souffraient très probablement d'un biais de surajustement. Finalement notre approche, bien que perfectible, semble prometteuse. Nous l'avons alors appliquée à un cas d'interaction spécifique, primordiale dans la problématique vaccinale, celui de la prédiction de l'affinité des complexes épitope/paratope. Sur ce sujet, nous pouvons confirmer la supériorité de notre algorithme en comparaison des autres.

Finalement, la performance de notre algorithme tourne autour d'un coefficient de corrélation r^2=0,5 tout type de complexe protéique confondu. Cela signifie qu'il y a encore une grande marge de progression à faire concernant ce type de prédiction. En revanche, comme nous l'avons expliqué à la fin du chapitre précédent, notre approche est perfectible et peut être

améliorée par de nombreux axes tant au niveau de l'usage matériel qu'au niveau de l'optimisation de l'architecture du réseau. Les algorithmes d'apprentissage en profondeur les plus récents ainsi que les plus performants tel qu'AlphaFold ne se contentent ni d'utiliser un seul réseau de neurones ni d'utiliser uniquement cette dernière approche pour résoudre un problème. A ce titre, une des problématiques récurrentes dans le calcul de l'énergie d'affinité de l'interaction protéine-protéine consiste à évaluer avec précision le terme entropique du système. Nous pouvons imaginer une méthode combinant notre approche, qui se concentre sur l'interaction de l'interface du complexe protéique, avec une approche permettant de calculer efficacement le terme entropique à partir d'échantillonnage conformationnel de la structure.

Quoi qu'il en soit la réponse à une question si complexe se fera probablement en combinant plusieurs méthodologies déférentes et sera toujours sujet à l'approximation. L'approximation et l'incertitude des prédictions issues des méthodes de bio-informatiques sont et seront toujours une problématique complexe à appréhender, le seul juge de paix restant, en définitive, la méthode expérimentale.