Concepts et méthodes d'analyse numérique de la dynamique des cavités au sein des protéines et applications à l'élaboration de stratégies novatrices d'inhibition

Nathan Desdouits

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité : **Sciences de la Vie**
Ecole doctorale : **CdV**

Présentée par

Nathan Desdouits

Pour obtenir le grade de

DOCTEUR de L'UNIVERSITE PIERRE ET MARIE CURIE

Sujet de la thèse :

Concepts et méthodes d'analyse numérique de la dynamique des cavités au sein des protéines et applications à l'élaboration de stratégies novatrices d'inhibition

Soutenue le 29 Mai 2015 devant le jury composé de :

Pr. Catherine VENIEN-BRYAN	Président
Pr. Anne-Claude CAMPROUX	Rapporteur
Pr. Frédéric CAZALS	Rapporteur
Pr. Claudine MAYER	Examinateur
Pr. Michael NILGES	Examinateur, Directeur de thèse
Dr. Arnaud BLONDEL	Examinateur, Directeur de thèse

Les cavités sont le lieu privilégié des interactions d'une protéine avec ses ligands, et sont donc déterminantes pour sa fonction. Cette dernière est aussi influencée par la dynamique de la protéine. Bien que les cavités aient été étudiées depuis les années 70, les travaux sur leur dynamique ne sont apparus que récemment et peu de méthodes sont disponibles malgré leur intérêt, notamment pour le criblage virtuel et la conception de médicaments.

Les cavités d'une protéine définissent un ensemble foisonnant très dynamique et labile. Ainsi, identifier "une" cavité le long d'une trajectoire est ardu car le site d'intérêt peut être sujet à des divisions, fusions, disparitions et apparitions. Je propose donc une méthode pour résoudre cette question et pouvoir exploiter la dynamique des cavités de façon systématique et rationnelle. Cette méthode classifie les cavités selon les groupes d'atomes les entourant, avec les algorithmes et paramètres que j'ai identifiés comme procurant les meilleurs critères qualitatifs et quantitatifs de suivi de ces cavités.

Par ailleurs, pour caractériser les évolutions principales de la géométrie des cavités en relation avec la dynamique de la protéine, j'ai développé une méthode basée sur l'Analyse en Composantes Principales (ACP). Cette méthode peut être utilisée pour sélectionner ou construire des conformations ayant des cavités de géométrie spécifiée. Deux exemples d'applications sont traitées : la sélection de conformations ayant des cavités de géométries diverses et l'étude de l'évolution du réseau de cavités internes de la myoglobine lors de la diffusion du monoxyde de carbone en son sein.

Ces deux méthodes ont été utilisées pour trois projets de criblage virtuel ciblant respectivement le domaine de coupure-ligation de l'ADN-gyrase de *M. tuberculosis*, la subtilisine 1 de l'agent du paludisme *P. vivax* et GLIC, homologue procaryote des récepteurs ionotropes pentamériques humains. Les molécules sélectionnées à l'aide de ces méthodes ont permis d'identifier une molécule active contre la subtilisine 1 ainsi que deux inhibiteurs et deux potentiateurs de GLIC.

Mots clés : *Cavités de protéines, dynamique des cavités, criblage virtuel, conception rationnelle d'inhibiteurs, Analyse en Composantes Principales, identification dynamique des cavités, allostérie*

Cavities are the prime location of the interactions between a protein and its ligands. As an important feature of the protein fold, they are crucial for its functions. The protein function is also influenced by its dynamics. Although cavities have been studied since the seventies, specific studies on their dynamical behaviour only appeared recently. Few methods can tackle this aspect, despite its interest for virtual screening and drug design.

Protein cavities define an extremely labile and dynamic ensemble. Identifying "one" cavity along a trajectory ("tracking" a cavity) is therefore an arduous task, because the underlying site can be subjected to several events of fusions, divisions, apparitions and disparitions. I propose a method to help resolve this question, thus enabling systematic and rational dynamical exploitation of protein cavities. This method classify cavities using the atom groups that flank it, using algorithms and parameters that I identified as giving best qualitative results for cavity tracking.

To characterize the main directions of evolution of cavity geometry, and to relate them with the dynamics of the underlying structure, I developed a method based on Principal Component Analysis (PCA). This method can be used to select or build conformations with given cavity shapes. Two examples of applications have been treated : the selection of conformations with diverse cavity geometries, and the analysis of the myoglobin cavity network evolution during the diffusion of carbon monoxide in it.

These two methods have been used in three projects involving virtual screening, targeting the breakage-reunion domain of *M. tuberculosis* DNA-gyrase, *P. vivax* subtilisin 1 and GLIC, an procaryotic model of human pentameric ligand-gated ion channel. These methods enabled us to identify an inhibitor of subtilisin 1 and four effectors of GLIC (two inhibitors and two potentiators).

Keywords : *Protein cavities, cavities dynamics, virtual screening, rational drug design, Principal Component Analysis, dynamical identification of cavities, allostery*

Remerciements

Tout d'abord, j'aimerais remercier Michael Nilges de m'avoir accepté au sein du laboratoire de Bioinformatique Structurale et d'avoir pris la responsabilité de diriger ma thèse. Tu es pour beaucoup pour l'excellente ambiance régnant au sein du laboratoire, c'était une chance de t'avoir comme encadrant et comme collègue. J'aimerais également chaleureusement remercier Arnaud Blondel pour m'avoir codirigé et encadré pendant le stage de Master et le doctorat. Ta disponibilité, ton enthousiasme et tes conseils m'auront permis de passer 4 années sereines mais productives. Merci également pour ces innombrables et interminables discussions, pas toujours scientifiques mais toujours enrichissantes. Plus qu'un directeur de thèse, tu es et restera un mentor pour moi.

J'aimerais également remercier l'intégralité des membres passés et présent de BiS pour leur enthousiasme et leur bonne humeur au jour le jour. J'ai notamment une pensée toute particulière pour Fabien, Edithe, Damien, Yannick, Guillaume, Mathias, Nathalie, Simeon, Isidro et Silke. J'aimerais remercier Tru, notre ingénieur système, et Renée Communal et Maya Um, secrétaires successives du laboratoire, pour leur aide précieuse.

Ce travail a été partiellement réalisé dans le cadre de collaborations. Sans mes collaborateurs, tout un pan de cette thèse n'aurait pas vu le jour. A Sanofi, je remercie tout particulièrement Kwame Ammaning. A l'Institut Pasteur, je remercie Claudine Mayer, Stéphanie Petrella, Pierre-Jean Corringer, Marc Delarue, Jean-Christophe Barale et tout particulièrement Anaïs Menny, Ludovic Sauguet et Frédéric Poitevin pour leurs aides précieuses, les échanges fructueux que j'aie eu avec eux et les opportunités qu'ils ont pu m'ouvrir.

Un grand merci également à ma famille et mes amis, à Louviers, Rouen, Paris et ailleurs, pour leur bonne humeur et leur soutien, leur aide parfois. J'ai une pensée particulièrement émue pour Aurore, soutien sans faille, correctrice vaillante et cobaye patiente, sans qui j'aurais eu beaucoup plus de difficultés à tenir le rythme de ces quatre années.

Enfin, je souhaiterais remercier Catherine Venien-Bryan, Anne-Claude Camproux, Frédéric Cazals, et Claudine Mayer, pour avoir accepté de faire partie de mon jury de thèse.

Table des matières

Chapitre I

L'analyse fonctionnelle des cavités pour développer de nouvelles stratégies d'identification de molécules actives

1 Les phases de dévelopement d'un médicament

1.1 Historique et enjeux économiques

Avant le développement de la médecine moderne, les médicaments étaient principalement des remèdes à base de plantes ou de minéraux, utilisés pour soulager les symptômes dont souffraient les patients sans apporter nécessairement de solution pour traiter la maladie sous-jacente[1]. Les avancées en chimie au cours du XIX^{e} siècle ont rendu possible l'identification, l'extraction et la purification des molécules actives présentes dans ces remèdes[1, 2, 3]. Les travaux de Pasteur et de ses successeurs ont aidé à identifier les causes des maladies (microbes, virus, parasites, champignons, déréglement de mécanismes biologiques...). Malgré cela, et jusqu'à la moitié du XX^{e} siècle environ, les médicaments étaient découverts principalement par l'étude des substances thérapeutiques naturelles ou simplement par sérendipité[4]. Ainsi, c'est totalement par hasard que Fleming découvrit le premier antibiotique (la péniciline) en 1928[5]. Le processus de découverte des médicaments s'est ensuite progressivement rationnalisé. Le développement des chimiothèques et des méthodes de biologie moléculaire a rendu possible le criblage de plus en plus de composés chimiques sur des cellules vivantes[3]. L'essor de la génétique moléculaire au cours des années 70 a permis d'identifier plus rapidement les protéines à cibler et de les tester directement[2, 3]. Depuis la fin des années 80, le développement de la bioinformatique (génomique comme structurale) a également permis de mieux optimiser les composés (méthodes QSAR), puis de guider leur conception[2]. Au cours des années 90, la découverte de nombreuses molécules actives a été possible grâce à l'utilisation des techniques de criblage à haut débit. Dans le même temps, les conditions

d'entrée des médicaments sur le marché se sont progressivement durcies pour assurer la sécurité des patients.

Ainsi, malgré ces avancées, le coût de développement d'une nouvelle substance chimique (*NCE* : *New chemical entity*) a progressivement augmenté pour dépasser le milliard de dollars au cours des années 2000[6]. Le développement des médicaments candidats est également une entreprise risquée, puisque la plupart des projets sont arrêtés avant la fin des études cliniques, pour des raisons diverses (toxicité, manque d'efficacité, problème de brevet...)[7]. L'industrie pharmaceutique est constituée d'entreprises parmi les plus grosses du monde, ce qui leur permet d'avoir les moyens financiers nécessaires pour pouvoir développer en parallèle plusieurs médicaments candidats. Le milieu académique et les entreprises de plus petites tailles font néanmoins partie intégrante de cette industrie en relayant et apportant l'innovation nécessaire dans les premières phases de la conception d'un médicament. Les preuves de concepts développées dans ces petites entreprises sont généralement revendues aux "big pharma" pour réaliser les phases cliniques.

Le nombre de nouvelles molécules mises sur le marché varie d'année en année, mais est relativement stable depuis les 20 dernières années[8], autour de 30 molécules par an pour le marché américain (voir figure I.1.a). L'augmentation quasi exponentielle du coût de développement d'un médicament[6, 9] reste toutefois problématique (figure I.1.b). Il est donc de plus en plus important de trouver de nouvelles voies de développement tout en essayant de diminuer les risques et les coûts de chaque étape, notamment au tout début d'un projet lorsque l'investissement est encore limité.

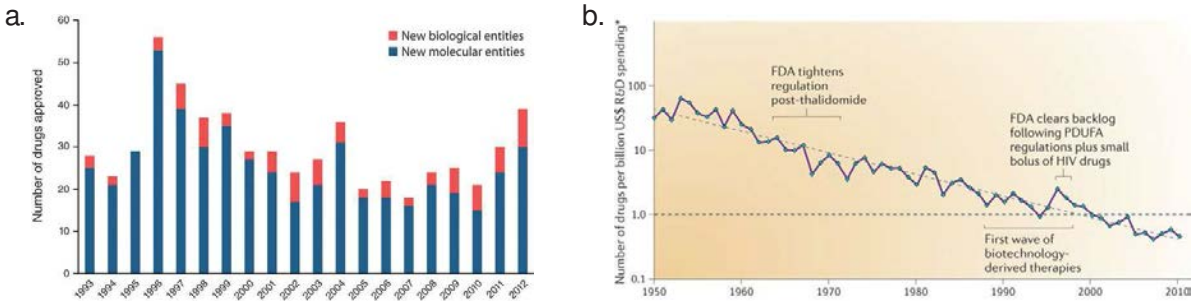


Figure I.1 – **a.** Nombre de nouvelles entités moléculaires approuvées par la FDA (Food and Drug Agency) de 1993 à 2012. Source : Jiang 2013[8]. **b.** Evolution du nombre de molécules approuvées par milliard de dollard, de 1950 à 2010 (échelle logarithmique en ordonnées). Tiré de Scannell *et al.*, 2012[9].

Le développement d'un médicament se déroule en plusieurs phases[10, 11]. La première phase, dite de découverte, est l'étude des mécanismes d'action liés à une maladie, et la détermination de molécules pour bloquer ces mécanismes. Elle se divise en plusieurs étapes : l'étude de la pathologie, la génération de touches (*target-to-hit*), la sélection de têtes de séries (*hit-to-lead*) et l'optimisation des têtes de séries [1][11]. A la fin de cette phase, plusieurs dizaines ou centaines de molécules ont été sélectionnées et testées sur un modèle simple (tests biochimiques ou phénotypiques). Les molécules les plus prometteuses (ayant une grande affinité pour la cible ou une bonne efficacité contre l'infection) sont ensuites conduites en phase préclinique. La phase préclinique consiste à

1. Pour la définition des termes *touches* et *têtes de séries*, voir section 1.3 de ce chapitre

réaliser une première estimation de l'efficacité et de la toxicité des molécules à l'aide de modèles animaux.

Les molécules présentant des propriétés concluantes à l'issue de la phase préclinique passent en phase d'essais cliniques où la molécule est testée pour la première fois sur des volontaires humains[12, 13]. Les essais cliniques sont décomposés en 4 phases distinctes[12]. Durant la phase I, la molécule est administrée à un petit nombre de volontaires sains afin de tester sa toxicité et son évolution cinétique (cette phase peut être traitée différemment pour certains composés anticancéreux). Les molécules passant la phase II (non toxiques et aux propriétés cinétiques appropriées) sont testées sur un petit nombre de patients malades, afin de déterminer la plus petite dose efficace et d'estimer les effets secondaires éventuels à plusieurs doses. L'efficacité du nouveau médicament est évaluée en phase III dans une étude à grande ampleur impliquant de nombreux patients volontaires. La molécule testée est administrée à une partie des patients, tandis que l'autre partie est soumise à un traitement de référence ou à un placebo dans le cas où aucun traitement n'existe. Les essais la de phase III permettent de déterminer avec précision les doses et modes d'administration les plus efficaces et les risques associés au médicament. A la suite de ces essais, l'industriel peut demander une autorisation de mise sur le marché (AMM) aux autorités de sécurité du médicament. Il doit faire la preuve de l'efficacité et de l'absence de dangerosité du médicament[11]. Enfin, la phase IV correspond aux études post-commercialisation sur le long terme.

Le coût de développement varie d'un projet à un autre, mais augmente sensiblement au cours des différentes phases précliniques et cliniques[7] (tableau I, dernière colonne). A chaque étape du projet, le médicament peut être jugé dangereux ou inefficace et le développement arrêté[14] (tableau I, 2^{e} colonne), ce qui signifie généralement une perte conséquente pour l'industriel qui développe le projet[7]. Il est donc crucial de diminuer le taux de perte, et donc de diminuer le risque inhérent à chaque projet. Pour cela, un effort particulier doit être réalisé dès les phases d'étude du mécanisme et précliniques pour éviter de porter plus loin des projets voués à l'échec tout en restant innovant[15]. La multiplication des familles moléculaires est une voie de diminution du risque : on peut plus facilement s'affranchir d'une famille s'avérant posséder un inconvénient rédhibitoire s'il est possible de se rabattre sur une autre famille également prometteuse (tableau I, 4e colonne).

Durant ma thèse, je me suis intéressé au développement rationnel de médicaments basés sur de petite molécules chimiques ciblant une protéine. Je m'inscris exclusivement dans la phase de découverte, au niveau de l'étude du mécanisme et de la découverte des touches *in silico*. Mon but est de déterminer de nouvelles méthodes permettant de mieux tirer partie des modèles du fonctionnement de la protéine cible afin de sélectionner des molécules plus pertinentes pour les phases de criblage et donc de multiplier le nombre de touches à coût constant.

Phase		Tx. de succès	Durée (ans)	# molécules	Coût (M$)
Découverte		50%	4.5	14.6	674
dont	découverte des touches	80%	1.0	24.3	94
	touches à têtes de séries	75%	1.5	19.4	166
	optimisation	85%	2.0	14.6	414
Préclinique		69%	1.0	12.4	150
Phase I		54%	1.5	8.6	273
Phase II		34%	2.5	4.6	319
Phase III		70%	2.5	1.6	314
Enregistrement		91%	1.5	1.1	48
Total		4%	13.5	1.0	1,778

Tableau I – Coût, durée, taux de succès et nombre de molécules moyennes nécessaires pour chaque phase d'un projet de développement de médicament. Source : Paul *et al.*. 2010[7].

1.2 La phase de découverte : définition d'une cible biologique

A la base même d'un projet de développement d'un médicament se trouve la pathologie. Le développement rationnel de médicaments implique de déterminer le mécanisme sous-jacent à la maladie : mécanisme d'infection d'un virus ou d'une bactérie, cible d'une toxine, voies déréglées par un cancer ou une maladie génétique. La première étape d'un projet consiste donc à déterminer ce mode de fonctionnement ainsi qu'un angle d'attaque potentiel pour traiter la pathologie.

1.2.1 Modélisation du mécanisme pathogénique

L'établissement du mécanisme biologique fait intervenir de nombreuses méthodes issues de disciplines diverses. Dans le cadre d'agents infectieux, le séquençage et l'annotation du génome de l'agent pathogène permet d'avoir une première idée du rôle de certaines protéines, mais rend également possible la production des protéines du pathogène par génie génétique. Les études de localisation de l'expression des protéines (micropuce ADN, microscopie confocale, immunofluorescence, ...), la détermination des interactions entre protéines (double hybride, spectrométrie de masse, ...) et les méthodes de *knockout* ou d'extinction de gène permettent d'affiner le modèle de fonctionnement et de déterminer les gènes *essentiels* de la pathologie. Les protéines produites par ces gènes peuvent alors être étudiées en détail, ce qui fournit de nombreuses informations pouvant être utilisées pour la recherche d'effecteurs. Ces études peuvent être très longues et donner lieu à de multiples interprétations ; elles constituent la base de la recherche fondamentale en biologie.

Lorsque la pathologie est bien comprise, il est possible de sélectionner une ou plusieurs protéines jouant un rôle déterminant dans son mécanisme. Ces protéines seront les cibles des campagnes de développement de médicament. Elles peuvent potentiellement provenir de n'importe

quel type, mais on dénombre quelques familles particulièrement importantes[16, 17, 18] :

— les récepteurs couplés aux protéines G (RCPG ou GPCR en anglais) : cette famille de protéines transmembranaires joue le rôle de senseurs, et est impliquée dans de nombreux processus physologiques (régulation du système immunitaire, récepteurs de certains neurotransmetteurs, contrôle du rythme cardiaque, de la pression sanguine, ...). Entre 40% et 60% des médicaments sur le marché ciblent un RCPG[17, 19].

— les kinases : élément central de multiples voies de signalisation et de régulation, elles sont la cible de nombreuses molécules anticancéreuses[16].

— les protéases : souvent utilisées par les procaryotes et les virus pour hydrolyser leurs polyprotéines (ensemble non fonctionnel de protéines synthétisées en un seul passage de la polymérase)[20] ou pour maturer un précurseur vers une nouvelle forme fonctionnelle[21, 22, 23]. Les protéases ont aussi un rôle dans certaines voies de signalisation[24].

— les canaux ioniques : cibles de nombreux anesthésiques, de composés psychotropes et de médicaments contre l'hypertension[25]

1.2.2 Etude structurale de la cible

Au fur et à mesure du développement d'un projet de développement de médicament, il devient de plus en plus important de connaître la structure de la cible au niveau atomique, pour affiner les différents modèles et optimiser efficacement un composé intéressant. Il existe différentes techniques d'obtention d'une structure protéique, mais les plus courantes restent la cristallographie par rayons X et la résonnance magnétique nucléaire (RMN).

La cristallographie par rayons X donne accès de façon précise à une image de la protéine, à condition de pouvoir en former des cristaux. Les cristaux de protéine sont formés en purifiant et en concentrant la protéine cible dans un tampon de composition précise (figure I.2.a). Le cristal est ensuite soumis à un flash de rayons X, ce qui permet de capturer le motif de diffraction du réseau cristallin (figure I.2.b). Le motif de diffraction est ensuite utilisé pour déterminer la densité éléctronique de la protéine, sous réserve de pouvoir déterminer les phases. Les atomes sont ensuites placés dans la densité éléctronique à l'aide de logiciels comme *O* ou *Coot* (figure I.2.b), ce qui permet après plusieurs étapes de rafinement d'obtenir la structure de la protéine.

La résonance magnétique nucléaire (RMN) quant à elle utilise plusieurs propriétés des noyaux atomiques liés à leur spin (moment magnétique nucléaire). Les expériences de RMN utilisent des séquences d'impulsions d'un champ magnétique très intense pour conditionner les spins de certains de ces atomes. La relaxation des spins (retour à l'état initial) est ensuite analysée pour déterminer l'environnement éléctronique des atomes. En croisant les résultats de plusieurs types d'expérience (HSQC[26], COSY[27], NOESY... - figure I.3.b), il est possible d'obtenir des informations sur les distances et les angles entre différentes paires d'atomes de la protéine. Ces informations sont utilisées pour définir des contraintes qui sont utilisées dans des logiciels de simulations (comme ARIA[28] ou CYANA[29]) pour réaliser et affiner des modèles de structure. Les variations entre

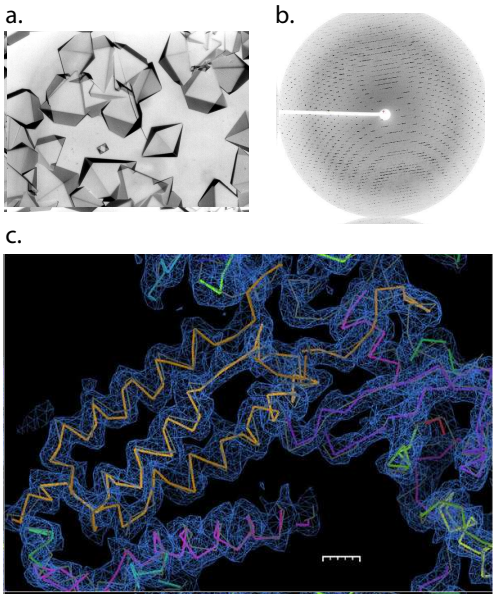


Figure I.2 – Détermination de structure par cristallographie aux rayons X. a. Production de cristaux de protéines. **b.** Exemple de motif de diffraction **c.** Assignement d'une structure dans sa densité électronique

modèles sont le reflet à la fois de la dynamique interne de la protéine et des différentes incertitudes dans les données expérimentales et de la simulation (figure I.3.c).

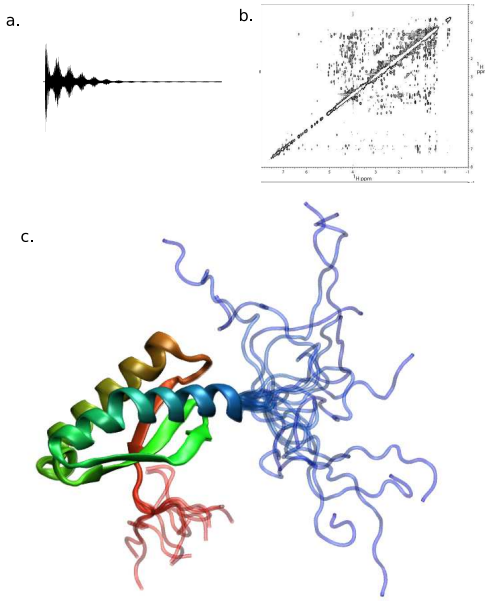


Figure I.3 – Détermination de structure par résonance magnétique nucléaire. a. Le signal mesuré, somme des signaux induits par les atomes. **b.** Exemple de spectre NOESY **c.** Ensemble de structures déterminées avec ARIA ; la partie centrale de la protéine est bien définie, alors que les partie C- et N-terminales sont très variables du fait de l'absence de contraintes sur ces parties.

1.2.3 Détermination d'un site orthostérique ou allostérique

L'obtention de la structure atomique permet de sélectionner précisément le site de liaison préférentiel des futurs ligands. Souvent, on préfèrera cibler le site orthostérique, c'est-à-dire le site de fixation du substrat (dans le cas d'une enzyme) ou d'une molécule inhibitrice déjà connue. Il est alors important d'avoir des indications structurales permettant de déterminer le site avec certitude. Dans l'idéal le site orthostérique est déterminé en co-cristallisant la protéine avec le ligand (substrat ou effecteur), mais la connaissance des acides aminés impliqués dans la fonction (déterminés par RMN ou mutations par exemple) peut suffire.

Il peut être préférable dans certains cas de ne pas viser le site orthostérique, soit parce qu'il n'en existe pas comme pour les protéines n'étant pas des enzymes (protéines de structure par exemple[30]), soit parce que le site orthostérique n'est pas avantageux en tant que cible, pour des raisons de spécificité par exemple (sites ATP notamment[16]). Dans ce cas, il est nécessaire de sélectionner un site dit allostérique, c'est-à-dire un site d'interaction autre que le site orthostérique ayant un lien avec la fonction de la cible. L'identification de sites allostériques peut également être l'occasion de créer de nouvelles opportunités sur des cibles déjà exploitées dans le passé. Il n'est pas trivial de déterminer si une région particulière d'une protéine peut être un bon site allostérique. Le chapitre II, dédié au suivi des cavités le long d'une dynamique, traite notamment de ce point.

1.3 Trouver et optimiser une "touche"

Le site choisi est ensuite criblé afin de déterminer des molécules potentiellement actives : les *touches*. Ces molécules sont raffinées, les meilleures sont définies comme "têtes de séries" et sont alors optimisées. Ces différentes appellations sont fonction de l'affinité, qui doit être la plus importante possible tout en limitant la toxicité et l'insolubilité. La découverte d'une touche provient souvent de criblages hauts débits, de criblages virtuels, ou d'un criblage dit "basé sur le ligand" lorsqu'un composé actif est déjà connu. Ces différentes méthodes nécessitent la mise en place d'un test biologique permettant d'évaluer l'efficacité des molécules sélectionnées.

L'optimisation des touches est le travail du chimiste médicinal. L'objectif est de modifier un composé en ajoutant, supprimant ou modifiant des groupes chimiques, afin de rendre le composé plus efficace, moins toxique et plus longtemps biodisponible. Les outils de relation structure-activité ($QSAR$) permettent de rationaliser cette approche.

1.3.1 Criblage haut-débit

Le criblage permet de déterminer une liste réduite de molécules, ayant une activité et une bonne affinité, à partir d'une banque de molécules chimiques (une chimiothèque) plus ou moins grande. Il existe deux grandes stratégies de criblage. Le criblage haut débit est la solution classique qui consiste à tester l'intégralité d'une chimiothèque de façon automatisée à l'aide de robots[31]. Ce type de criblage est relativement fiable car il ne dépend que de la qualité du test et de la

chimiothèque. Un autre avantage est qu'il ne requiert pas de connaissances sur la structure de la cible, il peut donc être utilisé dès la conception du test biologique. Malheureusement, il est aussi très onéreux, notamment lors de tests de chimiothèques de plusieurs centaines de milliers, voire plusieurs millions de composés.

1.3.2 Criblage virtuel basé sur la structure de la cible

Le criblage virtuel consiste à filtrer les molécules *in silico* afin de rejeter celles qui n'ont a priori que peu de chances d'avoir un effet[32]. Pour cela, on utilise des logiciels d'arrimage moléculaire pour simuler la pose du ligand à l'intérieur du site choisi. Les molécules les plus prometteuses (celles qui sont les plus adaptées au site) sont ensuite commandées et testées de la même façon que pour un criblage haut débit. Cette méthode permet de réduire drastiquement le nombre de molécules testées, donc de diminuer le coût relatif à l'achat ou au stockage des composés, ainsi que le coût des tests. Malheureusement, la méthode est moins fiable, le taux d'erreur (à la fois faux positifs et faux négatifs) restant relativement important ; il est donc possible de passer à côté de molécules actives. Un autre problème est la nécessité absolue de connaître la structure atomique de la cible, ce qui n'est pas forcément chose aisée. Toutefois, pour beaucoup de projets en phase de preuve de concept, pour lequel il n'existe pas de test automatisé ou assez de ressources pour entreprendre le test d'une chimiothèque complète, le criblage virtuel est une étape obligatoire.

1.3.3 Criblage virtuel basé sur la structure du ligand

Lorsque la cible possède un substrat ou une molécule active connue, il est possible de réaliser une recherche de composés dite "basée sur le ligand". Le but est de rechercher des molécules d'une chimiothèque dont la forme et le placement des fonction chimiques ressemblent à la molécule de départ (substrat ou molécule active). Ces recherches ne nécessitent pas d'avoir la connaissance de la structure de la cible, ce qui peut être très avantageux dans certains cas où la structure est complexe à obtenir (protéines membranaires difficilement cristallisables, gros complexes protéiques). Cette problématique ne sera pas étudiée dans ce manuscrit.

1.3.4 Optimisation d'une touche (hit-to-lead)

Une fois un petit nombre de touches déterminées, il est généralement nécessaire de les optimiser. L'optimisation d'un composé consiste à modifier, supprimer ou ajouter des groupements chimiques au composé, pour :

— améliorer l'affinité du composé pour la cible

— améliorer la biodisponibilité du composé en augmentant sa capacité à passer les barrières (intestinales, cellulaires...) et en diminuant les possibilités de métabolisation.

— diminuer la toxicité

Pour cela, il est nécessaire de concevoir des tests de toxicité, de solubilité et de biodisponibilité en plus des tests d'efficacité déjà utilisés lors des phases de recherche. Cette étape d'optimisation

est cruciale, car elle doit permettre de passer d'une touche assez peu efficace et potentiellement dangereuse à un médicament candidat administrable à un patient.

L'optimisation repose sur une très bonne connaissance de la chimie médicinale : certains groupements sont à proscrire car notoirement toxiques ou métabolisables, d'autres permettent d'augmenter ou de diminuer la solubilité d'un composé. Le chimiste médicinal doit jongler avec l'ensemble de ces contraintes tout en tenant compte de la structure du composé et de son interaction avec le site ciblé. Pour cela, il réalise une relation structure-activité (*SAR* pour *Structure-activity relationship*), aidé si besoin par des méthodes statistiques (*QSAR* pour *Quantitative SAR*, *QSAR3D*). Ces méthodes (notamment l'ACP, discutée au chapitre III) sont appliquées sur les résultats de premiers essais d'optimisation pour déterminer les déterminant chimiques de l'activité du composé et guider ainsi les efforts du chimiste. Les méthodes de *QSAR3D* utilisent également les données structurales du site ciblé pour guider l'établissement de cette relation.

1.4 Phases précliniques, cliniques et autorisation de mise sur le marché

La phase préclinique correspond au test des molécules sur des modèles animaux : souris, chien, porc, singe... Ces tests doivent pouvoir déterminer des doses minimales d'efficacité et de toxicité, ainsi que la dose maximale tolérée[12, 10, 11]. Si la toxicité est trop forte ou l'efficacité trop faible, le projet peut être arrêté dès cette étape. Les doses déterminées sur ces modèles animaux permettent d'avoir une première idée des doses utilisables sur l'humain.

La partie purement clinique du projet se décompose en trois phases. Ces phases impliquent un nombre croissant de patients volontaires : de quelques dizaines d'individus en phase I jusqu'à plusieurs milliers en phase III. Ces phases permettent de s'assurer que la molécule est effectivement efficace et non toxique, d'affiner les connaissances sur le métabolisme et les effets secondaires du médicament candidat, et de régler les doses qui seront utilisées pour le lancement du produit sur le marché.

Les tests de phase I sont réalisés sur des patients volontaires en bonne santé, ou au contraire en impasse thérapeutique. Ces tests ont pour objectif d'évaluer les doses ne provoquant pas d'effets indésirables et d'étudier le métabolisme du composé chez l'humain. Cette phase permet également de recenser les effets secondaires associés à l'utilisation du médicament testé. Les médicaments anticancéreux peuvent ne pas passer par une phase I du fait de leur mécanisme particulier.

La phase II consiste à une première étude de l'efficacité du médicament. Ces tests sont effectués sur un nombre modéré (plusieurs centaines) de patients volontaires touchés par la pathologie ciblée. Cette phase se décompose en deux sous-phases, IIa et IIb. La phase IIa permet de vérifier l'efficacité du traitement, tandis que la phase IIb vise à déterminer la dose thérapeutique à utiliser.

Lors de la phase III, le médicament est testé sur un nombre important de patients volontaires, jusqu'à plusieurs milliers voire dizaine de milliers de malades. Il est nécessaire non seulement de prouver que le médicament possède une efficacité, mais que celle-ci soit supérieure à l'efficacité d'un traitement de référence. Si ces tests s'avèrent concluant (efficacité supérieure aux médicaments

préexistants et absence de toxicité aux doses efficaces), le médicament peut être déposé à l'agence de sécurité du médicament pour être avalisé.

L'autorisation de mise sur le marché (AMM) est délivrée par les autorités de régulation des médicaments propres à chaque pays ou région (ANSM en France, EMA en Europe, FDA aux Etats-Unis). L'autorisation n'est délivrée qu'après l'étude du dossier transmis par l'entreprise pharmaceutique à la fin de l'étude clinique. Après la mise sur le marché, le nouveau médicament débute la phase IV des essais cliniques qui correspond au suivi du médicament. Le but est d'amasser des connaissances sur le mécanisme et les effets du médicament, en recensant notamment les effets indésirables rares et les effets du composé sur le très long terme. L'AMM peut ainsi être retirée lors de la découverte d'effets secondaires dangereux ou d'un manque d'efficacité à long terme.

2 La recherche de nouvelles stratégies d'inhibition in silico

Comme nous avons pu le voir dans la section précédente, la recherche de nouveaux médicaments est un défi nécessitant la mobilisation de ressources importantes. Dans un domaine très compétitif, la découverte de nouvelles stratégies d'inhibition de pathogènes par des voies innovantes ouvre de nouvelles opportunités, ce qui peut être un atout majeur. L'amélioration des méthodes de sélection de composés *in silico* permet à la fois d'accélérer la découverte de nouveaux inhibiteurs, de diminuer le coût du criblage des composés et d'améliorer les chances d'identifier des molécules actives. Je me place ici dans un contexte exclusivement basé sur la structure, dans l'optique de réaliser des criblages virtuels les plus efficaces possibles en tirant parti au maximum des données structurales existantes. Pour atteindre cet objectif, je réalise des modèles du fonctionnement des protéines cibles afin de choisir la voie d'action la plus pertinente pour un projet de développement de médicaments, via la détermination de poches allostériques ou la dynamique du site actif. L'utilistation de nouvelles stratégies sur des cibles nouvelles ou déjà connues permet en effet de découvrir et développer des composés pouvant contourner les brevets industriels déjà en place ou évitant des problèmes de toxicité et de biodisponibilité. La meilleure caractérisation des sites d'intérêt permet également d'élargir les possibilités de criblage et de sélectionner des composés qui n'auraient pas pu l'être précédemment.

Dans cette section, j'introduirai les bases théoriques nécessaires à la compréhension des mécanismes d'association entre une protéine et un ligand. Je ferai donc la revue des outils utilisés pour modéliser les associations possibles avec des composés chimiques, mais également pour comprendre le fonctionnement de la cible. Cela inclue donc les logiciels de prédiction d'arrimage moléculaire (*docking* protéine-ligand), la dynamique moléculaire, le calcul de chemins de transition et l'analyse des cavités de la protéine ciblée. Nous soulignerons l'importance de la prise en compte de la dynamique de la cible pour chacun de ces types d'analyse.

2.1 Principes physiques de l'association protéine-ligand

2.1.1 Thermodynamique de l'association protéine-ligand

Deux molécules ne vont s'associer que si elles ont un "intérêt commun" à le faire. Plus précisément, l'enthalpie libre totale du complexe protéine-ligand ($G_{complexe}$) doit être inférieure à la somme des enthalpies libres du récepteur seul (G_{rec}) et du ligand seul (G_{lig}) [2]. On définit la différence d'enthalpie libre d'association ΔG_{assoc} ainsi :

$$\Delta G_{assoc} = G_{complexe} - (G_{rec} + G_{lig})$$

ΔG_{assoc} doit donc être strictement négatif pour que l'interaction entre le récepteur et son ligand soit favorable. De fait, plus ΔG est négatif, plus l'affinité du récepteur pour son ligand est importante.

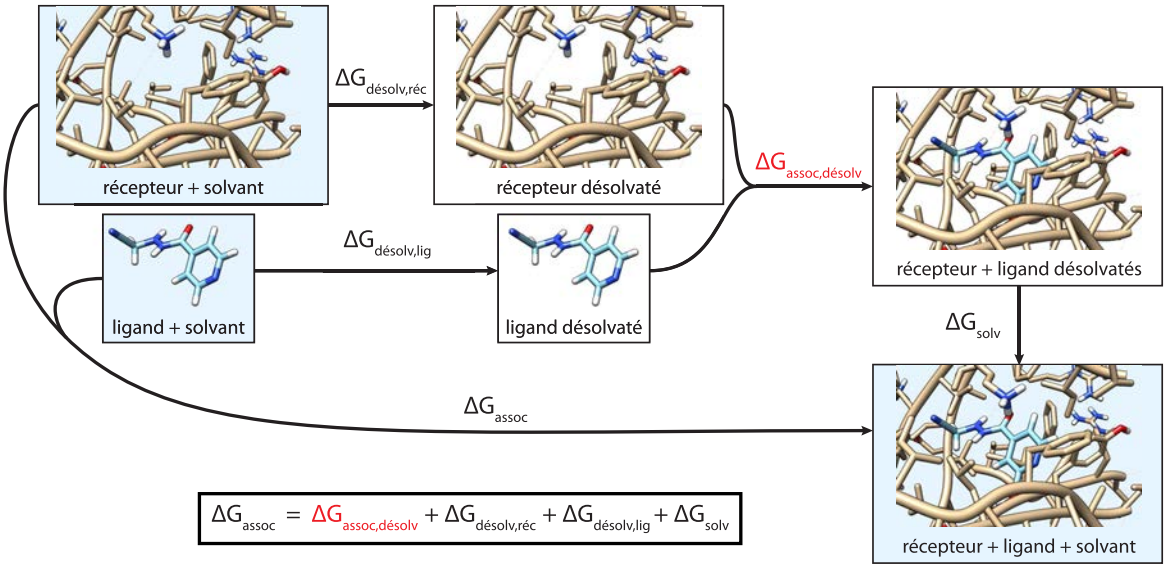


FIGURE I.4 – Prise en compte de l'énergie de solvatation dans l'énergie d'association protéine-ligand.

La prise en compte de l'effet du solvant étant en général complexe à modéliser[33], on décompose l'énergie d'association pour faire émerger les termes d'association en milieu "sec" (figure I.4), plus facilement calculable :

$$\begin{aligned} \Delta G_{assoc} = & \quad \Delta G_{assoc,desolv} + \Delta G_{desolv,rec} + \Delta G_{desolv,lig} + \Delta G_{solv} \\ = & \quad \Delta H_{assoc,desolv} + \Delta H_{desolv,rec} + \Delta H_{desolv,lig} + \Delta H_{solv} \\ & -T\left(\Delta S_{assoc,desolv} + \Delta S_{desolv,rec} + \Delta S_{desolv,lig} + \Delta S_{solv}\right) \end{aligned}$$

Les termes d'entropie (ΔS) sont en général difficiles à estimer[33] et peuvent varier grandement

2. L'association entre une protéine et son ligand se déroule naturellement à pression constante et en phase condensée. On peut donc utiliser l'enthalpie libre G au lieu de l'énergie libre F, la différence entre ces deux grandeurs, PV, étant quasi constante (P est constante et ΔV est généralement très faible).

d'un ligand à l'autre et d'un récepteur à l'autre. De plus, la balance $\Delta H/\Delta S$ est très influencée par le solvant et en particulier par les ions.

On se concentrera donc sur les termes d'enthalpie (ΔH), c'est-à-dire les termes correspondant aux interactions, qui sont appréhendables plus facilement. On peut ainsi décomposer le terme d'enthalpie désolvatée :

$$\Delta H_{assoc,desolv} = H_{rec,lig} + \Delta H_{lig,lig} + \Delta H_{rec,rec}$$

L'enthalpie d'association $H_{rec,lig}$, qui correspond à l'énergie d'interaction entre le récepteur et son ligand, doit donc compenser non seulement les interactions propres au récepteur ($\Delta H_{rec,rec}$) et au ligand ($\Delta H_{lig,lig}$) pouvant être perdues lors de l'association, mais surtout la perte de certaines interactions avec le solvant ($\Delta H_{desolv,rec}$ et $\Delta H_{desolv,lig}$).

Le nombre d'interactions est donc un élément très important de l'association entre une protéine et un ligand. Les concavités à la surface et les trous à l'intérieur de la protéine permettent de créer un grand nombre d'interactions simultanées en permettant la formation de liaisons sur un plus grand nombre de faces du ligand. Cela permet d'augmenter l'enthalpie d'association $H_{rec,lig}$ et donc en principe d'augmenter l'affinité du récepteur pour son ligand. Un autre bénéfice est que plus il y a d'interactions, plus le ligand doit être spécifique du récepteur pour pouvoir se lier. Le ligand ne pourra en effet pas former autant d'interactions ou aura des incompatibilités stériques avec un récepteur légèrement différent. Cette spécificité est un avantage certain dans le cadre de la conception de médicament (toxicité et effets secondaires réduits). Ces trous et concavités dans la protéine seront désignés sous le terme commun de cavité (je définirai plus en détail les cavités dans la section 2.5 de ce chapitre). La forme des cavités explique également la spécificité de l'association protéine-ligand puisque le ligand doit avoir une forme suffisament proche de celle de la cavité pour pouvoir y entrer et réaliser ces interactions qui agissent souvent à courte portée (ex : liaisons hydrogènes). De plus, la complémentarité de forme permet de maximiser les contacts entre ligand et récepteur et donc de tirer profit au maximum des forces de van der Waals attractives. Le détail de la forme géométrique de la cavité peut ainsi être déterminante pour certaines familles de ligand.

2.1.2 Modèles dynamiques de l'association protéine-ligand

Le premier modèle de la reconnaissance entre une protéine et son ligand est le modèle "clé-serrure" (figure I.5.a), introduit par Fischer dès 1894[34]. C'est un modèle statique : le récepteur possède une cavité dont la forme correspond exactement à celle du ligand. Ce modèle s'est avéré inexact pour de nombreuses protéines pour lesquelles les structures apo (libres) et holo (liées) sont nettement différentes. L'ajustement induit (figure I.5.b), correspondant à une transformation du récepteur au cours de la liaison avec le ligand, a ainsi été introduit par Koshland en 1958[35] pour répondre à ce problème. Ce modèle implique une certaine flexibilité du récepteur et introduit donc une notion de dynamique absente du modèle clé-serrure. Un troisième modèle, celui de la sélection conformationnelle (figure I.5.c), postule que la conformation holo du récepteur fait partie

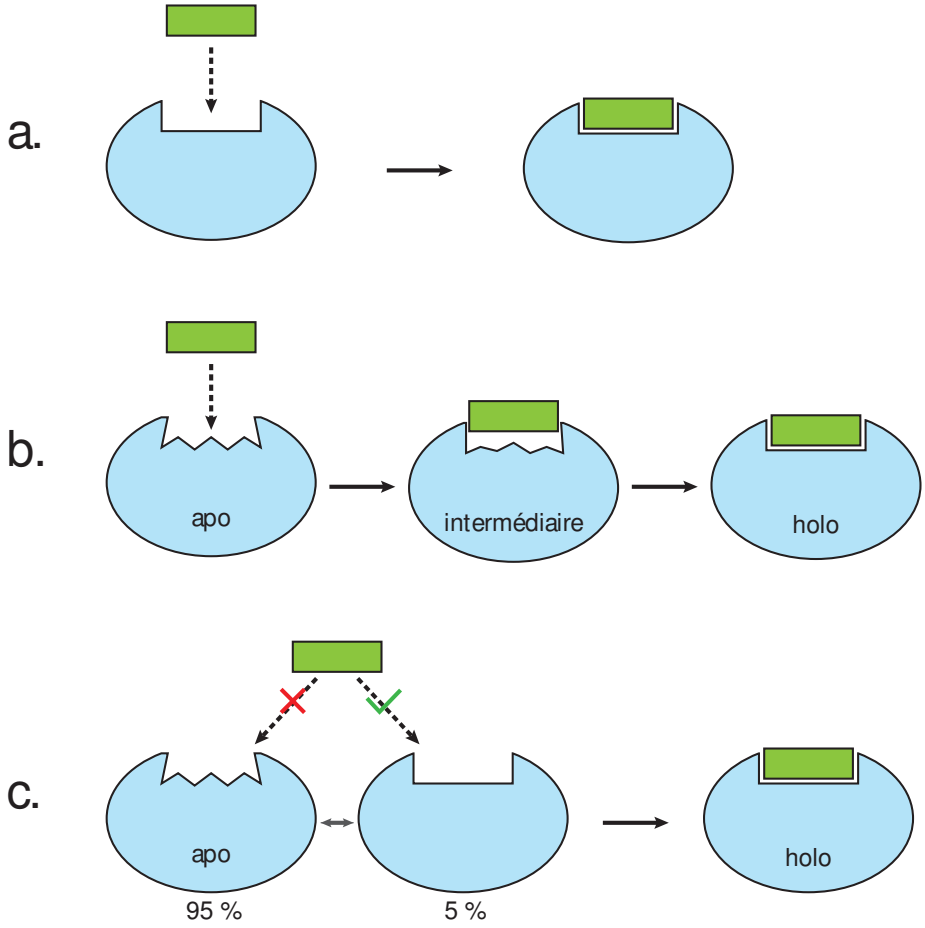


Figure I.5 – Modèles d'association protéine-ligand. a. Modèle clé-serrure. **b.** Modèle d'ajustement induit. La conformation du récepteur change au cours de l'association. **c.** Modèle de sélection conformationnelle. La forme holo (conformation du récepteur lorsqu'il est lié au ligand) est présente de façon très minoritaire lorsque le ligand n'est pas lié. Le ligand ne se fixe que sur cette forme et "reconnaît" donc cette conformation parmi les autres.

de l'ensemble des conformations visitée par le récepteur non lié, mais que cette conformation est simplement beaucoup plus rare que la conformation apo (elle n'est donc pas mesurée). Dans ce modèle, le ligand n'interagit pas avec le récepteur s'il n'est pas dans le bon état et "sélectionne" donc la conformation holo qui est ainsi stabilisée par la liaison avec le ligand. Ce modèle a été introduit petit à petit, suite à diverses observations faites notamment sur les anticorps qui peuvent se lier à différents antigènes grâce à ces changements de conformations[36, 37, 38, 39]. Il a été formalisé pour les interactions protéine-protéine par Kumar *et al.*[40].

Les modèles d'ajustement induit et de sélection conformationnelle ont été longtemps en compétition. Toutefois, il est sans doute probable que la plupart des associations protéines-ligand suivent une combinaison des deux modèles à des degrés divers[41, 42, 43]. Certaines sources privilégient en revanche l'hypothèse d'une association provoquée par la sélection conformationnelle[44], notamment à faible concentration[43]. On peut voir dans ces deux modèles des paradigmes utiles pour l'arrimage moléculaire que j'aborde dans la section suivante.

2.2 L'arrimage moléculaire (*docking*)

L'arrimage moléculaire, ou *docking* en anglais, consiste à prédire la structure d'un complexe à partir des structures de ses composants. En particulier, le docking protéine-ligand permet de prédire la *pose* d'une petite molécule dans le site d'une protéine. Les logiciels de docking associent à chaque pose un score correspondant généralement à une estimation de l'énergie d'association du ligand. Les ligands ayant un bon score (i.e : une faible énergie) sont ainsi plus susceptibles de se lier à la cible que les autres. Le docking consiste donc à trouver, pour un ligand et une protéine cible donnés, la pose du ligand dans la protéine donnant la meilleure énergie. C'est donc un outil très utilisé dans le domaine de la conception de médicament, car il permet de sélectionner parmi un grand nombre de petites molécules celles qui sont les plus à même de se fixer à la protéine cible. Ce problème est très complexe, car il fait intervenir un grand nombre de degrés de libertés : ceux de la protéine, ceux du ligand et ceux des associations possibles entre les deux (trois degrés de translation et trois de degrés de rotation). Afin de simplifier ce problème, les logiciels de docking tentent de limiter voire de supprimer certains de ces degrés de liberté :

— soit en limitant les degrés de libertés translationnels et rotationnels en limitant la zone ciblée à un site spécifique. Cette méthode est généralement utilisée lorsque l'on connait le site à cibler, ce qui est le cas la plupart du temps dans les projets de conception de médicament. Il est aussi possible de discrétiser l'espace des degrés de libertés translationnels et rotationnels exploré.

— soit en réduisant les degrés de liberté de la protéine, en rigidifiant les atomes de la chaîne principale et/ou des chaînes secondaires.

— soit en réduisant les degrés de liberté du ligand, en totalité ou en partie, en décomposant le ligand en fragments rigides reliés par des liaisons pouvant tourner librement ou par incréments d'angle.

— soit par une combinaison de l'un des trois points précédents.

De façon évidente, plus on supprime de degrés de liberté, plus on simplifie le modèle et la complexité du problème mais moins la pose prédite et son score sont plausibles. Actuellement, la plupart des logiciels de docking prennent en compte en partie la flexibilité du ligand. Certains de ces logiciels peuvent également prendre en compte la flexibilité du récepteur en rendant flexibles les chaînes latérales[45, 46, 47]. Ces logiciels peuvent donc modéliser en partie le phénomène d'ajustement induit. Pour modéliser le phénomène de sélection de conformations, il est possible de réaliser un docking sur de multiples conformations, soit en réalisant le docking indépendamment sur plusieurs conformations, soit en utilisant des méthodes dédiées[48, 49, 50, 51].

Les logiciels de docking utilisés pour le criblage virtuel ont vocation à docker des centaines de milliers, voire des millions de composés sur une cible en un temps raisonnable. Afin d'atteindre cette vitesse d'exécution, ces logiciels se limitent donc à l'exploration d'un seul site de la protéine, suppriment ses degrés de liberté (récepteur rigide) et discrétisent ceux du ligand (décomposition en fragments rigides et échantillonnage des angles de torsion). C'est le cas par exemple de DOCK[52,

53, 54], AutoDock[55, 56] et AutoDock Vina[57], FRED[58], Glide[59], FlexX[60, 61] ou GOLD[62, 63].

Il existe un grand nombre de subtilités caractérisant chaque logiciel de docking. La fonction de score en est la différence principale, mais d'autres différences existent : les méthodes de décomposition du ligand et d'échantillonnage des angles de torsion, la pose du ligand, l'élagage de l'arbre des possibilités... Les étapes de criblage virtuel du chapitre IV utilisent les logiciels DOCK 6 et FlexX, qui vont donc être explorés plus en détail.

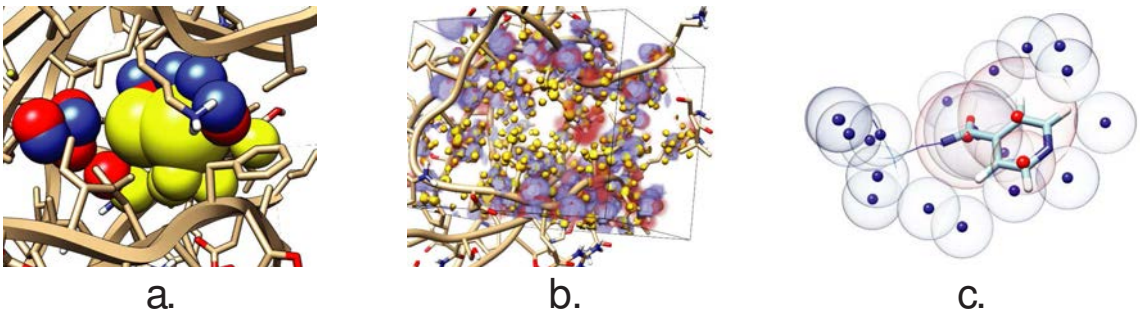


Figure I.6 – Fonctionnement de DOCK a. Sphères remplissant le site visé du récepteur. Les couleurs représentent le type d'interaction de chaque sphère (ici choisi au hasard). **b.** Grille d'interaction générée pour le même récepteur (bleu, rouge : électrostatique, jaune : contact/van der Waals). **c.** Pose de l'ancre : certains atomes de l'ancre (représentée par des bâtons, le reste du ligand est représenté en fil de fer) sont alignés sur les centres des sphères de la couleur complémentaire à l'atome. Les centres des sphères sélectionnées sont représentées en rouge, les autres centres en bleu.

DOCK est l'un des premiers algorithmes de docking. Le logiciel ne gérait initialement que le docking rigide, sans prendre en compte la flexibilité du ligand, mais cette fonctionnalité a depuis été implémentée. Le fonctionnement de DOCK se base sur la superposition des atomes du ligand avec les centres de sphères remplissant le site du récepteur. Ces sphères sont générées par sphgen (figure I.6.a), et la contribution du récepteur à l'interaction est précalculée sous forme de grille (figure I.6.b). DOCK attribue ensuite à chaque sphère une "couleur" correspondant à l'interaction prédominante dans cette région (donneur ou accepteur de liaison hydrogène, interaction hydrophobe, charge positive/négative ou neutre). Le ligand est d'abord décomposé en fragments rigides (figure I.7.a), séparés par des liaisons permettant la rotation. Un des fragments (généralement le plus gros) est désigné comme l'ancre. Ce fragment est placé dans le récepteur de telle sorte que ses atomes se rapprochent du centre des sphères de "couleur" complémentaires à sa nature (donneur/accepteur, hydrophobe/hydrophobe, charges positives/négatives, figure I.6.c). Les autres fragments sont ensuite ajoutés itérativement de la même façon (figure I.7.b et c), l'échantillonnage des angles de torsion du ligand provenant d'une table précalculée. Les poses sont préfiltrées par un score de "bump" qui discrimine les poses présentant des recouvrements trop importants entre les atomes du ligand et ceux du récepteur. Le score de chaque pose intermédiaire est calculé à partir des grilles d'interactions, prend en compte les contacts avec le récepteur et les interactions électrostatiques. Les poses intermédiaires sont filtrées à chaque étape à l'aide d'un algorithme de clustering prenant en compte le score et la diversité géométrique des poses, afin de réduire la combinatoire des poses. Les poses finales sont optimisées à l'aide d'un champ de force (voir sec-

tion 2.3) et peuvent être éventuellement rescorées par une fonction de score plus précise, telle que MMGBSA[64] ou PBSA[65].

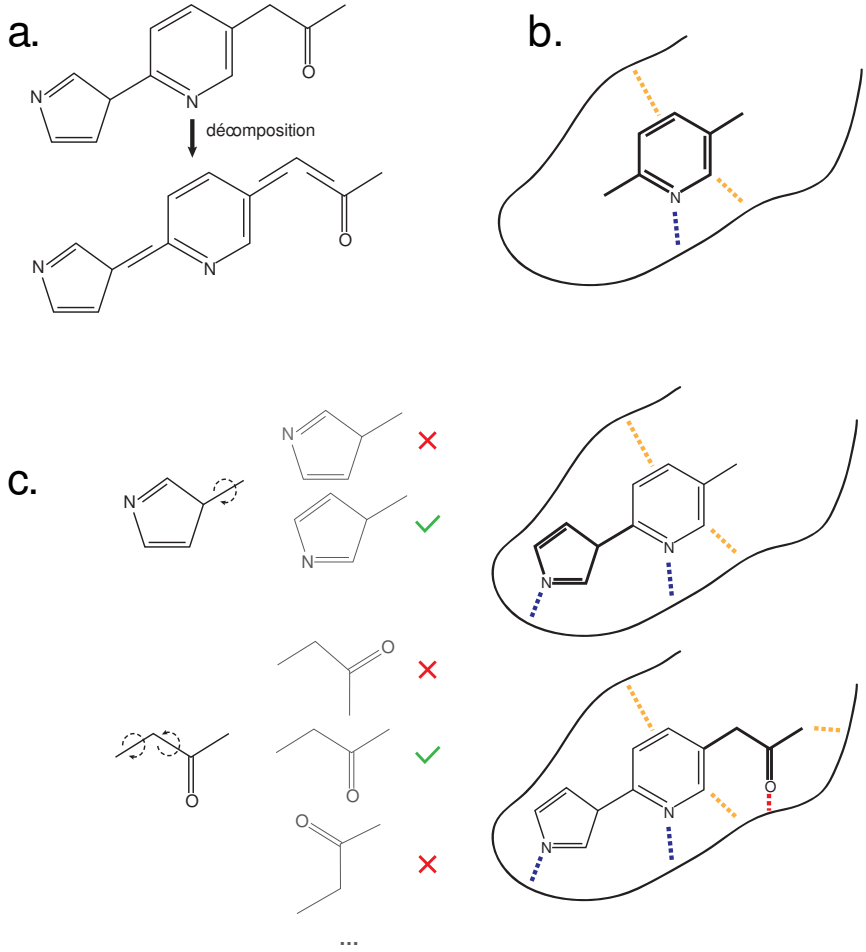


Figure I.7 – Etapes de la pose d'un ligand. a. Décomposition du ligand en fragments rigides. **b.** Placement du premier fragment (ancre) dans le récepteur. **c.** Reconstruction itérative du ligand, fragment par fragment, via l'échantillonnage des angles de torsion.

FlexX est un logiciel plus récent développé au milieu des années 90[60, 61]. FlexX modélise la surface d'interaction des atomes par des surfaces définies par des coupes d'une sphère (figure I.8.a). Le ligand est décomposé de la même façon que DOCK et suit le même schéma de reconstruction itérative du ligand (figure I.7). L'ancre est placée à l'aide d'un algorithme de hachage géométrique qui fait correspondre des triangles reliant les surfaces d'interaction du ligand à des points définis sur les surfaces d'interaction du récepteur (figure I.8.b). Le score est une estimation de l'énergie libre d'association et comprend des termes pour les liaisons hydrogènes, les ponts salins, la surface de contact hydrophobe et la perte de degrés de libertés. Les poses intermédiaires proches sont regroupées par un algorithme de partitionnement hiérarchique, et seules les k poses de meilleur score sont sélectionnées pour l'itération suivante.

Les diverses approximations réalisées lors du calcul de la pose, nécessaires pour la rapidité du calcul, impactent fortement sa précision[66]. La pose et son énergie associée sont donc fortement

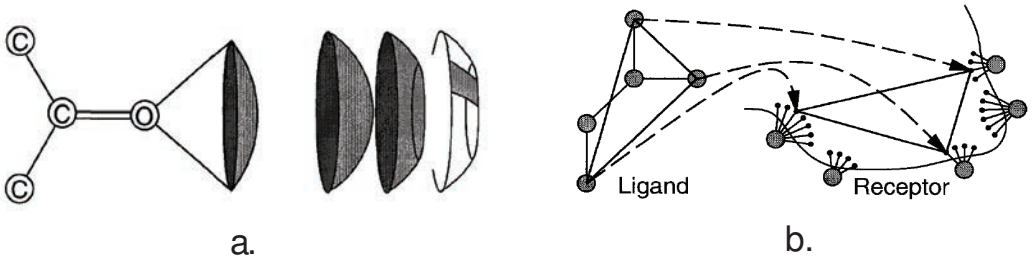


Figure I.8 – Fonctionnement de FlexX a. Surfaces d'interaction utilisées par FlexX pour le placement de l'ancre et des fragments. A gauche : exemple de surface d'interaction pour un groupe carbonyle. A droite, de gauche à droite : surfaces coniques (électrostatique, donneur de liaison hydrogène, cation-pi), coniques tronquées (accepteur de liaison hydrogène), rectangles sphériques (idem). **b.** Placement de l'ancre par *hachage géométrique*. Images tirée de Rarey *et al.* 1996[61].

hypothétiques[33]. Ainsi, les tests d'efficacité de ces algorithmes visent à mesurer le rang d'un ligand connu parmi un nombre de composés "leurres"[67] ou la précision de la pose d'un ligand connu[68]. Un bon algorithme placera le ligand connu parmi les composés de meilleur score et/ou à proximité de la pose connue, mais il n'est absolument pas garanti que ce soit toujours le cas. Dans le cadre d'un criblage virtuel, il est donc nécessaire de sélectionner un grand nombre de composés afin d'augmenter les chances d'obtenir une touche. En outre, les algorithmes de docking n'ont pas vocation à filtrer les composés particulièrement réactifs (et donc peu intéressant car peu spécifiques), toxiques ou insolubles. Il faut donc réaliser des filtres supplémentaires : fonctions chimiques, solubilité, toxicité (ADMET)...

Ni DOCK, ni FlexX ne prennent en compte la flexibilité du récepteur. Certains logiciels de docking laissent toutefois la possibilité de rendre flexibles certaines chaînes latérales, au détriment de la rapidité d'exécution[45, 46]. Pour prendre en compte la flexibilité du site actif, il est également possible d'utiliser plusieurs conformations du récepteur, ce qui revient à considérer le modèle de sélection de conformation. Cette approche, appelée *ensemble docking*, peut utiliser des structures expérimentales ou provenant de la modélisation[49, 50, 51]. En particulier, l'utilisation de conformations générées par dynamique moléculaire a récemment donné de bons résultats[50, 69, 70, 71]. Enfin, on peut noter que de nouvelles méthodes d'échantillonnage de la conformation du site visé ont été récemment développées spécifiquement pour ce type d'applications. Parmi ces méthodes, on peut citer SCARE[72], les techniques de fumigation[73] et de pressurisation[74]. Nous nous intéresserons plus en détail à la sélection de conformations à partir de la géométrie des cavités au chapitre III.

2.3 La dynamique moléculaire

2.3.1 Principe général de la dynamique moléculaire

La Dynamique Moléculaire (DM, molecular dynamics en anglais) est une technique permettant de simuler l'évolution de plusieurs molécules au cours du temps[75]. A partir d'une structure atomique et de vitesses initiales, on peut utiliser le principe fondamental de la dynamique, $m\mathbf{A} =$

F, pour générer des conformations successives dans le temps. Il existe plusieurs composants clés rentrant en compte dans une dynamique moléculaire :

- la *topologie* du système, qui définit l'ensemble des atomes et leurs liaisons
- les *coordonnées* et *vitesses initiales* des atomes du système. Les coordonnées proviennent généralement de structures déterminées expérimentalement (voir section 1.2.2), tandis que les vitesses initiales sont souvent générées automatiquement.
- les *thermostats/barostats* éventuels, qui permettent de contrôler la température et/ou la pression de l'environnement selon l'ensemble thermodynamique choisi.
- le *champ de force*, qui définit l'ensemble des forces appliquées aux atomes du système sous forme de potentiels. Les atomes ont le plus souvent une représentation ponctuelle symbolisant la position du noyau. Le nom "champ de force" dénote aussi les paramètres associés aux différents atomes, liaisons et forces (masse, charge, constantes des potentiels, ...).
- l'*intégrateur*, qui est l'algorithme produisant une nouvelle conformation à partir des coordonnées des conformations et/ou des vitesses précédentes, du champ de force et des éventuels thermostats, barostats et autres forces éventuellement appliquées.

2.3.2 Ensembles thermodynamiques, thermostats et barostats

Une dynamique moléculaire peut être réalisée dans plusieurs ensembles thermodynamiques différents, selon les grandeurs que l'on souhaite garder constantes au cours du temps. Les plus couramment utilisés sont :

- l'ensemble microcanonique (NVE) : la quantité de matière (N), le volume (V) et l'énergie (E) sont constants
- l'ensemble canonique (NVT) : la quantité de matière, le volume et la température (T) sont constants. Cet ensemble nécessite un thermostat pour échanger de l'énergie avec le système et le garder à température constante.
- l'ensemble isotherme-isobare (NPT) : la quantité de matière, la pression (P) et la température sont constantes. Cet ensemble nécessite à la fois un thermostat pour garder la température constante et un barostat pour garder la pression constante.

Le système est défini dans un volume fini (éventuellement variable dans le cas de l'ensemble NPT), généralement choisi comme périodique afin de simuler un système infini et éviter des effets de bords peu réalistes (ondes de pression).

Le principe sous-jacent aux thermostats est de moduler la vitesse des particules du système pour réduire (ou augmenter) la température pour la faire tendre vers la température cible. Cette modulation des vitesses peut être réalisée par une dilatation ou contraction (thermostat de Berendsen[76]), via l'ajout de degrés de libertés supplémentaires (thermostat de Nosé-Hoover[77, 78]), ou par l'ajout de "collisions" aléatoires (thermostat d'Andersen[79] et dynamique de Langevin). Le thermostat de Berendsen ne sera pas traité ici, car il n'est pas réalisé strictement dans l'ensemble canonique et peut engendrer des artefacts non réalistes (effets dit de "solvant

chaud, soluté froid" et du "glaçon volant"). Le thermostat de Nosé-Hoover inclu un degré de liberté supplémentaire au système qui représente le bain thermostaté. Ce bain possède une "masse" Q : plus Q est grande, plus le bain échangera de l'énergie avec le système. Le thermostat de Nosé-Hoover garantit que le système est effectivement simulé dans l'ensemble canonique. Le processus est également strictement déterministe puisqu'il n'emploie pas de nombres aléatoires. Il n'évite cependant pas complétement le phénomène de "solvant chaud, soluté froid".

La dynamique de Langevin applique une approche différente visant à émuler l'effet du solvant sur un soluté via l'ajout de deux types de forces supplémentaires au système. La première est décrite comme une force de friction, qui réduit la vitesse des atomes du système proportionnellement à leur vitesse ($\dot{\mathbf{X}}$) et à la fréquence de collision avec le solvant (γ) : $\mathbf{F}_{\text{friction}} = -\gamma m\dot{\mathbf{X}}$. La seconde est une fluctuation aléatoire simulant des chocs, qui est proportionnelle à la fréquence de collision γ ainsi qu'à la vitesse du solvant, donc à la température T_0 du bain : $\mathbf{F}_{\text{chocs}} = \sqrt{2\gamma k_b T_0 m}\mathbf{R}$, avec $\mathbf{R}$ un processus gaussien de moyenne nulle et de déviation standard égale à 1. L'équation du mouvement devient donc :

$$\begin{aligned} m_i\mathbf{A}_i = m_i\ddot{\mathbf{X}}_i &= \mathbf{F}_i + \mathbf{F}_{\text{friction},i} + \mathbf{F}_{\text{chocs},i} \\ &= \mathbf{F}_i - \gamma m_i\dot{\mathbf{X}}_i + \sqrt{2\gamma k_b T_0 m_i}\mathbf{R}_i \end{aligned}$$

Le "thermostat" de Langevin fait donc tendre progressivement les vitesses des atomes du système vers des vitesses correspondant à la température du bain thermostaté.

Les barostats fonctionnent tous sur un principe commun qui est de faire évoluer le volume du système (i.e : de la boîte) en fonction de la pression interne afin de s'approcher d'une valeur de pression cible. Pour cela, le volume V de la boîte peut être considéré comme un degré de liberté supplémentaire de "masse" Q, à l'instar de la méthode de Nosé-Hoover pour les thermostats. Ce degré de liberté peut être interprété comme un piston de masse Q, s'opposant plus ou moins aux changements de volume ; un piston de petite masse fait osciller le volume rapidement, au contraire d'un piston de grande masse qui oscillera plus lentement. Lorsque Q tend vers l'infini, on retrouve le comportement d'une dynamique moléculaire classique sans changement de volume.

2.3.3 Le champ de force

L'élément le plus important et le plus complexe des programmes de dynamique moléculaire est le champ de force (*force field* en anglais). Il définit en effet l'ensemble des forces appliquées à chaque atome du système. Le champ de force choisi a donc un impact direct et important sur le résultat d'une dynamique moléculaire[80, 81]. Les valeurs des paramètres du champ de force sont déterminées de façon empirique en tentant d'approcher au maximum les propriétés mesurées d'un grand nombre d'espèces chimiques[82, 83, 84]. Le champ de force définit deux grandes catégories de forces : les forces agissant à travers les liaisons covalentes, et les forces agissant à distance.

Les forces agissant à travers les liaisons covalentes (*bonded terms*) sont liées à la topologie chimique du système qui définit ces liaisons. Elles font intervenir des atomes liés par une liaison covalente à leurs voisins, et sont définies à partir de potentiels. Ces potentiels sont relativement

rapides à calculer car la liste des atomes affectés est fixe et ne dépend pas de la position des atomes. En outre, le nombre de termes composant ces potentiels croît linéairement avec le nombre d'atomes. Parmi ces potentiels, quatre se retrouvent dans la plupart des logiciels de dynamique moléculaire :

— les forces de liaisons dépendent de la longueur l de la liaison entre deux atomes. Elles sont habituellement définies par un potentiel harmonique centré sur une longueur d'équilibre l_0 (figure I.9, potentiel **1**) :

$$E_{liaison} = \sum_{\text{liaisons}} \frac{k_l}{2}(l - l_0)^2$$

— le potentiel des angles de liaison dépend de l'angle θ entre trois atomes liés consécutivement par des liaisons covalentes. Il est souvent défini par un potentiel harmonique centré sur un angle d'équilibre θ_0 (potentiel **1**) :

$$E_{angle} = \sum_{\text{angles}} \frac{k_a}{2}(\theta - \theta_0)^2$$

— le potentiel des angles de torsion (ou angles dièdres) dépend de l'angle dièdre ω entre quatre atomes liés consécutivement par des liaisons covalentes. Le potentiel utilisé dépend des atomes concernés et peut posséder plusieurs minima. Il est généralement défini par des séries de fonctions cosinus (paramétrisées par les constantes A_n et ϕ_n, potentiel **2**) :

$$E_{torsion} = \sum_{\text{torsions}} \sum_{n} \frac{A_n}{2}(1 + cos(n\omega + \phi_n))$$

— le potentiel des angles de torsion impropres se rapproche des angles de torsion mais est défini sur des ensembles d'atomes pour corriger empiriquement certaintes propriétés de forme, comme la planéité des systèmes conjugués (potentiel **1**).

Il existe également d'autres potentiels de correction, comme le Urey-Bradley[84] (couplage liaison-angles) ou les termes de couplage entre les angles dièdres ϕ et ψ CMAP[80], basés sur la carte de Ramachandran (dans le programme CHARMM).

Les forces agissant à distance (*non-bonded terms*) sont fonction de la distance entre deux atomes. En pratique, elles s'appliquent aux atomes non liés de façon covalente. En principe, ces forces pourraient faire intervenir l'ensemble des $n(n-1)/2$ paires d'atomes d'un système ce qui en fait l'étape de calcul la plus lourde. Il est donc nécessaire d'avoir recours à des approximations permettant d'accélérer substantiellement le temps de calcul. Ces forces sont principalement de deux types :

— les forces de van der Waals et la répulsion interatomique (principe d'exclusion de Pauli) sont modélisées par la formule de Lennard-Jones (potentiel **3**) :

$$E_{vdw} = \sum_{i \neq j} \varepsilon_{ij} \left(\left(\frac{d_{0ij}}{d_{ij}} \right)^{12} - 2 \left(\frac{d_{0ij}}{d_{ij}} \right)^{6} \right)$$

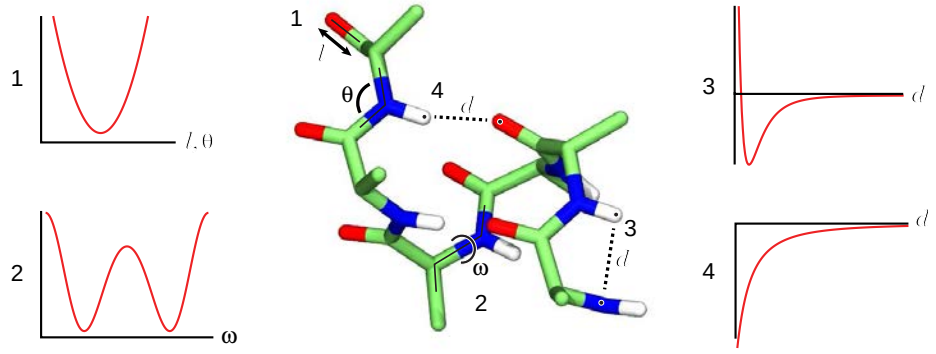


FIGURE I.9 – Potentiels des forces définies lors d'une dynamique moléculaire. 1. Potentiel harmonique (longueurs des liaisons, angles entre deux liaisons, angles de torsion impropres). **2.** Potentiel périodique en série de Fourier (angles de torsion) **3.** Potentiel de Lennard-Jones (forces de van der Waals) **4.** Potentiel de Coulomb (forces éléctrostatiques)

Il est possible d'ignorer les interactions au delà d'un certain seuil ; pour garder la continuité de la fonction d'énergie, on peut par exemple décaler la fonction pour qu'elle devienne nulle au niveau du seuil.

— les forces électrostatiques sont modélisées par un potentiel de Coulomb (potentiel **4**) :

$$E_{electrostatique} = \frac{1}{4\pi\varepsilon_0} \sum_{i \neq j} \frac{q_i q_j}{d_{ij}}$$

Dans le cas de simulations en condition périodique, il est possible d'accélérer les calculs en découpant le potentiel en deux parties selon les distances entre atomes. Les interactions à courte portée sont ainsi calculées normalement par sommation directe en tenant à jour la liste des paires d'atomes proches. Les interactions à longue portée sont elles calculées par sommation d'Ewald[85, 86] dans l'espace réciproque de Fourier, ce qui a l'avantage d'être une méthode rapide et de tenir compte des interactions à longue portée. Une fonction de partage est utilisée pour partitionner de façon continue et dérivable les interactions de courte portée des interactions de longue portée.

On peut également ajouter un terme prenant en compte spécifiquement les liaisons hydrogènes, mais la plupart des champs de forces plus récents ne l'utilise pas, ces interactions étant modélisées en adaptant les deux termes précédents.

L'ensemble de ces termes définit un champ de force *classique*, qui peut être éventuellement modifié ou amélioré pour rendre compte de phénomènes plus complexes, au détriment de la vitesse de calcul : polarisabilité (oscillateur de Drude), réactions chimiques, modèles hybrides entre mécanique classique et quantique ($QMMM$), ...

2.3.4 L'intégrateur et la génération des vitesses intitiales.

L'intégrateur est l'algorithme produisant une nouvelle conformation à partir de la conformation précédente et du champ de force. Il existe plusieurs algorithmes d'intégration, tous basés sur des méthodes de différence finie : méthodes de Verlet[87] et Verlet avec vélocités[88], *leapfrog*[89], méthode de Beeman[90]. Un algorithme très utilisé est la méthode de Verlet avec vélocités[88], qui utilise les coordonnées $\mathbf{X}$, les vitesses $\mathbf{V}$ et l'accélération $\mathbf{A}$ (calculée par le champ de force) de la conformation au temps t pour produire les coordonnées et vitesses de la conformation au temps $t + \Delta t$:

$$\mathbf{X}(t + \Delta t) = \mathbf{X}(t) + \mathbf{V}(t)\Delta t + \frac{1}{2}\mathbf{A}(t)\Delta t^2$$
$$\mathbf{V}(t + \Delta t) = \mathbf{V}(t) + \frac{\mathbf{A}(t) + \mathbf{A}(t + \Delta t)}{2}\Delta t$$

Il présente le double avantage de correspondre à un schéma d'intégration centré, donc précis et robuste, et de donner les vitesses au temps t (contrairement aux méthodes Verlet et leapfrog). Le pas de temps Δt est fixe, et doit être suffisament petit pour simuler correctement des phénomènes oscillatoires rapides et les chocs entre particules à la température du système. Le but est d'éviter d'obtenir des conformations sortant de la zone où l'évolution du potentiel est compatible avec les lois de conservation de l'énergie (atomes qui se chevauchent, liaisons trop longues). En général, Δt est fixé à 1 fs pour une dynamique classique réalisée aux conditions standard de température et de pression, mais peut être augmenté à 2 voire 4 fs en appliquant des algorithmes spécifiques, notamment SHAKE[91], qui fixent la longueur des liaisons impliquant des atomes d'hydrogène.

L'algorithme de Verlet avec vélocités nécessite un jeu de vitesses intiales, puisqu'elles sont utilisées dès la 1re étape. Ces vitesses sont reliées à la température T du système. La composante de la vitesse de chaque atome dans chacune des 3 directions de l'espace, v_{ix}, est donc tirée aléatoirement d'une distribution de Maxwell-Boltzmann :

$$p(v_{ix}) = \sqrt{\frac{m_i}{2\pi k_B T}} exp\left(-\frac{m_i v_{ix}^2}{2k_b T}\right)$$

où $p(v_{ix})$ est la probabilité que l'atome i ait une vitesse v_{ix} dans la direction x.

2.3.5 Le solvant

Le solvant utilisé pour l'immense majorité des simulations de dynamique moléculaire est l'eau, éventuellement associé à quelques ions pour neutraliser le système ou simuler une concentration en sels. Ce solvant peut être simulé de façon implicite, en utilisant une dynamique de Langevin et des constantes diéléctriques variables, ou de façon explicite, en incorporant un grand nombre de molécules d'eau dans le système. Les modèles de solvants implicites permettent de réaliser des simulations beaucoup plus longues, mais beaucoup moins précises qu'avec des solvants explicites.

Il existe plusieurs modèles de solvants implicites, plus ou moins complexes, dont le point commun est de substituer les molécules d'eau par un modèle continu. Ces modèles décomposent

l'énergie de solvatation du soluté[3], ΔG_{solv}, en deux termes représentant l'énergie provenant des interactions électrostatiques (ΔG_{elec}) et l'énergie apolaire (hydrophobe) $\Delta G_{apolaire}$:

$$\Delta G_{solv} = \Delta G_{elec} + \Delta G_{apolaire}$$

Le terme $\Delta G_{apolaire}$ prend en compte les interactions ne faisant pas intervenir les charges, notamment les interactions de van der Waals ainsi que le gain d'énergie lié à la désorganisation du solvant. Ce terme est généralement modélisé à l'aide de la surface accessible au solvant A : $\Delta G_{apolaire} = \sigma A$, où σ est un facteur déterminé empiriquement.

Le terme ΔG_{elec} doit, quant à lui modéliser, les interactions électrostatiques avec le solvant. Un modèle communément considéré comme le modèle "étalon" des solvants implicites est le modèle de Poisson-Boltzman qui donne le potentiel électrostatique d'un ensemble de charges ponctuelles plongées dans un solvant. Le modèle de Poisson-Boltzmann est très complexe à calculer et n'est donc généralement pas utilisé dans les simulations de dynamique moléculaire. Pour cette raison, d'autres modèles ont été élaborés. Le modèle le plus simple est de remplacer la permittivité du vide ε_0 par celle de l'eau, ε_{eau} (on a $\varepsilon_{\text{eau}}/\varepsilon_0 \sim 80$). Ce modèle améliore le comportement des résidus situés à la surface de la protéine, près du solvant, mais détériore la précision au sein de la protéine, dont la permittivité est située entre 2 et 4. Un modèle très utilisé est le modèle de Born généralisé[92], qui ajoute le terme ΔG_{Born} au potentiel électrostatique :

$$\Delta G_{solv} = \Delta G_{Born} = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\sum_{i}^{N}\sum_{j}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i R_j \exp\left(-\frac{r_{ij}^2}{4R_i R_j}\right)}}$$

Dans cette équation, les termes R_i et R_j sont proportionnels au degré d'enfouissement des atomes i et j. Nous avons utilisé un modèle de solvant implicite dérivant du modèle de Born généralisé, ACE[93], pour le calcul de chemin de transition avec *POE* (voir section 2.4.2).

2.3.6 Les logiciels de dynamique moléculaire

Les premières simulations de dynamique moléculaire datent de la fin des années 1950[94]. Les premiers modèles utilisés dans ces simulations faisaient intervenir un champ de force très simple (sphères dures) et ne simulaient que les collisions entre particules. L'application de la dynamique moléculaire aux protéines date du milieu des années 1970, avec les travaux de Levitt et Warshel[95], corécipiendaires avec Martin Karplus du prix Nobel 2013. Ce dernier est à l'origine du développement du programme CHARMM[96], un des premiers logiciels de dynamique moléculaire, toujours mis à jour et utilisé de nos jours. CHARMM utilise les champs de force développés par le groupe de développement de CHARMM, notamment dans le groupe de MacKerell[84]. D'autres moteurs de dynamique moléculaires ont été développés : NAMD[97], AMBER[98] et sa famille de champs de force associée, GROMACS[99] et son champ de force GROMOS, ou bien MMTK[100].

3. On note que l'utilisation de la notation ΔG procède d'un abus de language car elle caractérise normalement un ensemble représentant un état thermodynamique

Une des avancées récentes notable des logiciels de dynamique moléculaire est l'utilisation des cartes graphiques pour accélérer les calculs.

2.3.7 Les dérivés de la Dynamique Moléculaire

Le principe de la dynamique moléculaire a été étendu au cours du temps à l'aide d'un grand nombre de méthodes, en ajoutant ou modifiant des termes au champ de force, en modifiant l'intégrateur, etc. Les dérivés permettant de simuler des intermédiaires entre deux conformations, typiquement active et inactive, sont particulièrement utiles pour l'analyse du fonctionnement d'une protéine. Ils permettent de mieux comprendre les mécanismes moléculaires responsables de l'activation d'une cible et donc de choisir de façon plus efficace un site et une conformation à cibler. Ils peuvent également être utilisés comme première étape d'un calcul de chemin de transition[101] (voir section suivante).

Un des dérivés les plus utilisés est la dynamique moléculaire dirigée (*SMD*, pour *Steered Molecular Dynamics*). Le principe est de tirer un ensemble d'atomes du système vers des coordonnées cibles, par l'ajout d'un potentiel basé sur la distance entre les coordonnées du système et de la cible. Il est ainsi possible de tirer l'ensemble des atomes d'une protéine de sa conformation active à sa conformation inactive pour générer des conformations intermédiaires. La dynamique moléculaire ciblée (*TMD*, *Targeted Molecular Dynamics*) vise à contraindre la structure à se déplacer dans une hypersphère définie par une valeur de RMSD à la structure cible. Cette valeur de RMSD évolue avec le temps de la valeur initiale (RMSD entre la structure de départ et la structure cible) jusqu'à atteindre 0, ce qui assure que la structure finale est identique à la structure cible.

Ces méthodes très efficaces sur des transitions simples se trouvent prises par des barrières infranchissables dans des cas plus complexes. Il est alors nécessaire d'utiliser d'autres outils plus poussés pour calculer des chemins de transition.

2.4 Calcul de chemins de transition

2.4.1 Intérêt, principe et méthodes de calcul de chemins de transition

Lorsqu'il existe des structures de la protéine dans deux états différents, typiquement forme active-forme passive, il est possible de réaliser un calcul de chemin de transition. Un chemin de transition est une collection de conformations intermédiaires représentant un chemin plausible entre deux structures initiales et finales. L'intérêt d'un tel chemin est de modéliser le fonctionnement de la cible de façon très détaillée. Le modèle aide à générer de nouvelles stratégies d'inhibition, notamment en déterminant de nouveaux sites potentiellement allostériques. Les conformations intermédiaires peuvent alors servir de base pour l'étape de criblage virtuel.

Le problème du calcul de chemin de transition est extrêmement complexe. Pour paraphraser Bolhuis[102], il s'agit en quelque sorte de "lancer des cordes au dessus de chaînes de montagnes accidentées, dans le noir". On peut également ajouter que ces montagnes ne sont malheureusement pas situées dans un espace à 3 dimensions, mais dans l'espace des degrés de liberté d'une protéine,

de l'ordre de 10,000 dimensions, ce qui rend la tâche d'autant plus complexe... De fait, il existe plusieurs méthodes, plus ou moins subtiles, pour réaliser de tels chemins. Les plus "brutales" échantillonnent les chemins à l'aide de dynamiques moléculaires standards. Ces méthodes sont généralement utilisées pour calculer le chemin de repliement de petites protéines, comme pour le projet Folding@HOME[103, 104] ou les travaux du groupe de D.E. Shaw[105]. Les chemins de transition calculés par simple dynamique moléculaire ne peuvent être observés que pour des trajectoires de l'ordre de plusieurs millisecondes et nécessitent donc des ressources extrêmement importantes (super-ordinateurs, clusters de plusieurs dizaine de milliers de processeurs, ...).

Pour obtenir des chemins de transition de molécules plus grosses de façon plus efficace, il existe d'autres méthodes utilisant différents types de biais afin d'arriver à un chemin plausible. Il est tout d'abord possible d'utiliser les dérivés de la dynamique moléculaire décrits dans la section précédente, comme la SMD ou la TMD. L'inconvénient principal de ces méthodes est qu'elles introduisent un biais fort et déforment ainsi le paysage énergétique échantillonné. Le chemin calculé est donc très marqué par ce biais et s'éloigne d'un chemin d'énergie libre minimal, ce qui diminue grandement la plausibilité du chemin. D'autres algorithmes modifient cette surface d'énergie libre afin de permettre au système d'échantillonner plus librement son espace des phases. Ces modifications sont faites de telle façon qu'il soit possible de retrouver la surface d'énergie libre initiale (correspondant à l'ensemble canonique). Par exemple, le principe de la métadynamique[106] est de forcer le système à s'écarter de son état initial en ajoutant un terme énergétique défavorable aux endroits déjà visités (figure I.10.a). La méthode de l'*umbrella sampling*[107] permet quant à elle de diminuer les barrières d'énergie libre, facilitant l'échantillonnage de plusieurs bassins. Lorsque ces méthodes sont utilisées pour déterminer un chemin de transition, il est nécessaire de définir des coordonnées réactionnelles de faible dimension, décrivant l'état d'avancement du système le long du chemin. Ce sous-espace des phases décrit par les coordonnées réactionnelles permet de définir une surface d'énergie libre que l'on peut alors moduler pour faciliter les possibilités d'évolution du système afin d'obtenir un échantillonnage correct. Ces algorithmes laissent le système relativement libre et approximent correctement l'énergie libre le long du chemin défini par les coordonnées réactionnelles choisies, mais ils sont biaisés par ce choix.

Un troisième groupe d'algorithme utilise un chemin initial (pouvant être déterminé automatiquement ou par d'autres méthodes), et le raffine de façon itérative afin de diminuer son énergie. Une de ces méthodes, le *string of swarms*[108], consiste à produire une multitude de dynamiques très courtes à partir de chaque conformation du chemin, puis à moyenner les trajectoires résultantes pour déterminer les directions d'évolution diffusive de chaque conformation (figure I.10.b). Ces conformations sont ensuites déplacées dans ces directions moyennes afin d'aligner le vecteur de diffusion moyen et la tangeante au chemin. Ce déplacement est réalisé en gardant une contrainte de distance entre deux conformations consécutives pour éviter de regrouper toutes les conformations dans le bassin de plus basse énergie. Enfin, *CPR* (pour *Conjugate peak refinement*[109]) est une méthode visant à optimiser un chemin sur la surface d'enthalpie du système au lieu de la surface d'énergie libre. La valeur de l'enthalpie d'une conformation est en effet directement accessible via le champ de force car elle ne prend pas en compte l'entropie du système, difficile à estimer. Pour

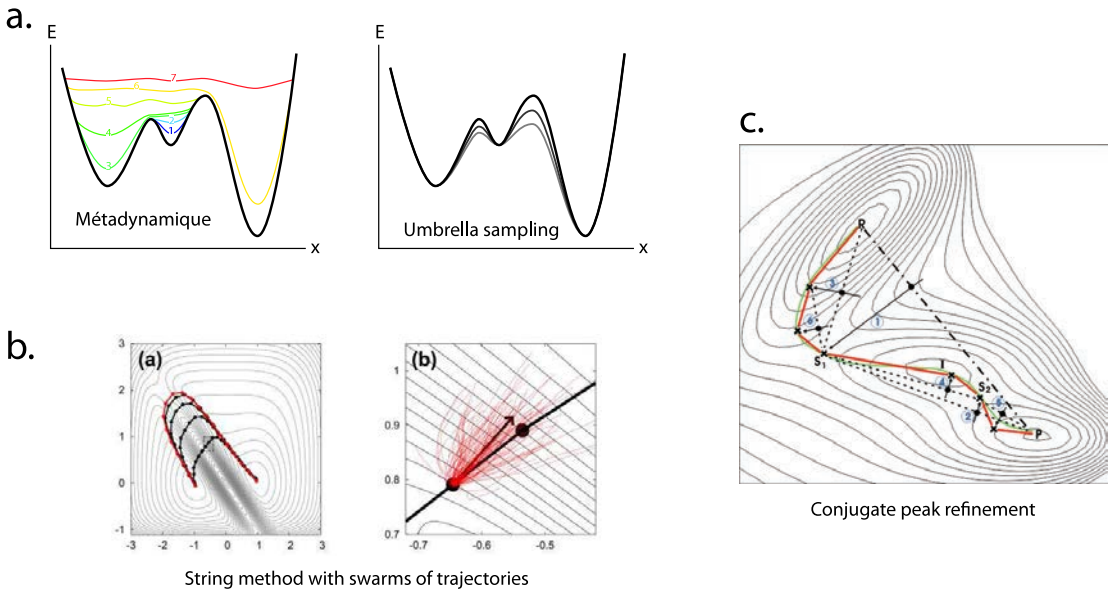


Figure I.10 – Trois familles de méthodes de calcul de chemins de transition. **a.** Méthodes de calcul d'énergie libre (Métadynamique, Umbrella sampling). Métadynamique, à gauche : la surface d'énergie libre définie par la coordonnée réactionnelle (en noir) est progressivement "remplie" (courbes 1 à 7), ce qui oblige le système à évoluer vers d'autres bassins. Umbrella sampling, à droite : la surface d'énergie libre (en noir) est modifiée (aplatie, courbes grises) pour autoriser un échantillonnage plus étendu, mais d'une façon "réversible" (il est possible de retrouver les valeurs réelles d'une quantité à l'aide d'une formule mathématique). **b.** *String method with swarms of trajectories*. (a) Les conformations définissant le chemin sont réparties de façon homogène sur une "corde". (b) Lors d'une itération, les conformations évoluent dans la direction moyenne déterminée par une nuée de trajectoires de dynamique moléculaire non biaisées. **c.** *Conjugate Peak Refinement* (*CPR*) : le chemin (initial : point-trait, intermédiaire après 6 itérations : rouge, réel : vert) est raffiné itérativement en optimisant le maximum énergétique (point noir) dans le sous-espace orthogonal au chemin (flèche) jusqu'à atteindre un minimum (croix). L'optimisation s'arrête lorsque les maxima locaux ne peuvent plus être optimisés : il s'agit de point de selles sur la surface d'enthalpie du système.

calculer le chemin, *CPR* optimise itérativement l'énergie de la conformation d'énergie maximale du chemin. Cette optimisation se fait dans le plan bisecteur des intermédiaires voisins de part et d'autre du chemin (figure I.10.c). *CPR* n'utilise pas de notion de dynamique et n'utilise que les principes de la mécanique moléculaire. Le chemin produit est un chemin adiabatique à 0 K, il n'est donc pas possible d'utiliser de solvant explicite puisque celui-ci gèlerait lors de l'optimisation.

2.4.2 L'approche *POE* (*Path Optimization and Exploration*)

L'approche *POE*, développée au laboratoire par A. Blondel, est une méthode utilisant *CPR* à grande échelle pour calculer le chemin de transition de grands systèmes lors de mouvements de grande amplitude (ex : mouvements de domaines). Le problème principal de *CPR* est sa tendance à réaliser des chemins topologiquement très complexes, dont la plupart des mouvements n'ont pas d'intérêt fonctionnel (figure I.11.a). De plus, *CPR* ayant pour objet de chercher les points de selle le long d'un chemin, il ne se débarasse pas spontanément des avatars topologiques tels que les croisements de chaînes principales ou latérales. Pour contrer ce défaut, *POE* tente de trouver des raccourcis entre des conformations du chemin situées relativement près dans l'espace mais éloignées dans la séquence de conformations du chemin.

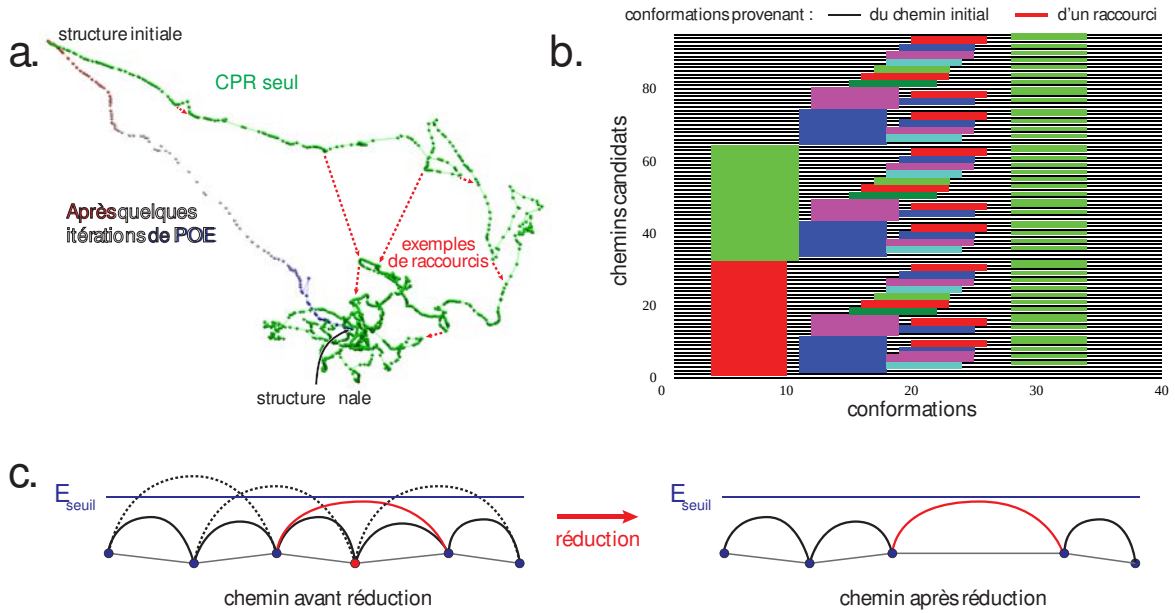


Figure I.11 – Optimisation de chemins de transition à l'aide de *POE*. **a.** Un chemin de transition d'une protéine obtenu avec *CPR* (vert), et le même chemin après optimisation avec *POE* (bleu). Un exemple de raccourci pouvant être entrepris est indiqué avec une flèche rouge. **b.** Combinatoire de la reconstruction du chemin complet à partir des raccourcis calculés avec *CPR*. Les conformations provenant du chemin initial sont indiqués en ligne noire, les séries de conformations remplacées par des raccourcis sont indiqués en segment épais de couleur. **c.** Schéma explicatif de la continuité énergétique du chemin produit par *POE* après réduction. Les conformations sont représentées par des ronds (bleus ou rouges), l'interpolation linéaire entre deux conformations est représentée par un segment gris. L'évolution de l'énergie entre deux conformations est schématisée par une parabole, le seuil d'énergie E_{seuil} par un trait bleu. Lorsque l'énergie entre deux conformations non consécutives ne dépasse pas le seuil d'énergie (parabole rouge), les conformations intermédiaires peuvent être supprimées. Ce n'est pas possible lorsque l'énergie dépasse le seuil (paraboles en pointillés).

Ces racourcis sont modélisés dans un premier temps par des "lignes droites" (des interpolations linéaires) dans l'espace des conformations puis optimisés à l'aide de *CPR*. Les raccourcis qui ont pu être produits et qui apportent une amélioration sont ensuite combinés en plusieurs chemins candidats (figure I.11.b). Ces chemins candidats sont ensuite optimisés une nouvelle fois à l'aide de *CPR*. Le chemin final, qui sera utilisé comme point de départ pour l'itération suivante, est ensuite choisi parmi ces candidats selon des critères d'énergie et de complexité (nombre de conformations, longueur curvilinéaire, barrières de basse énergie). Afin de réduire la complexité du chemin à la suite de cette recombinaison, *POE* sélectionne en priorité les raccourcis diminuant à la fois l'énergie, le nombre de conformations intermédiaires et la distance curvilinéaire entre le départ et l'arrivée. La plupart du temps, pour un chemin composé de n intermédiaires, il est trop coûteux d'optimiser les $(n-1)(n-2)/2$ raccourcis possibles. Il est donc nécessaire de passer par une heuristique pour déterminer les raccourcis ayant le plus de chance d'aboutir à un chemin plus simple et d'énergie plus basse avant de tenter de les calculer.

L'heuristique de sélection des raccourcis que j'ai développée, *tip-ext*, consiste en un programme permettant de calculer le nombre d'intersections de chaînes principales et/ou secondaires entre deux conformations (figure I.12). Les chaînes principales sont modélisées par un segment de droite entre deux C_α consécutifs, tandis que les chaînes latérales sont modélisées par un segment entre

un C_α et un atome situé loin sur la chaîne secondaire (figure I.12.a). Pour détecter si deux chaînes se croisent entre deux conformations i et j, *tip-ext* définit un tétraèdre reliant les extrémités des deux chaînes. Le volume est calculé analytiquement le long de l'interpolation linéaire faisant passer de la conformation i à la conformation j selon un paramètre λ. S'il existe $\lambda_0 \in [0, 1]$ tel que le volume s'annule, et si les segments modélisant les chaînes se croisent pour ce λ_0, alors on considère que ces deux chaînes risquent de se croiser lors du calcul du raccourci de i vers j (figure I.12.b). On évite donc de faire le calcul de transition du raccourci entre les conformations i et j, afin de gagner du temps. Cette heuristique donne de très bon résultats, en prédisant notamment de façon assez précise quand un calcul de raccourci échoue à cause d'énergies trop élevées (du fait d'un croisement entre deux liaisons). Elle permet également de définir des cartes topologiques pour des chemins de transition (voir figure I.12.c).

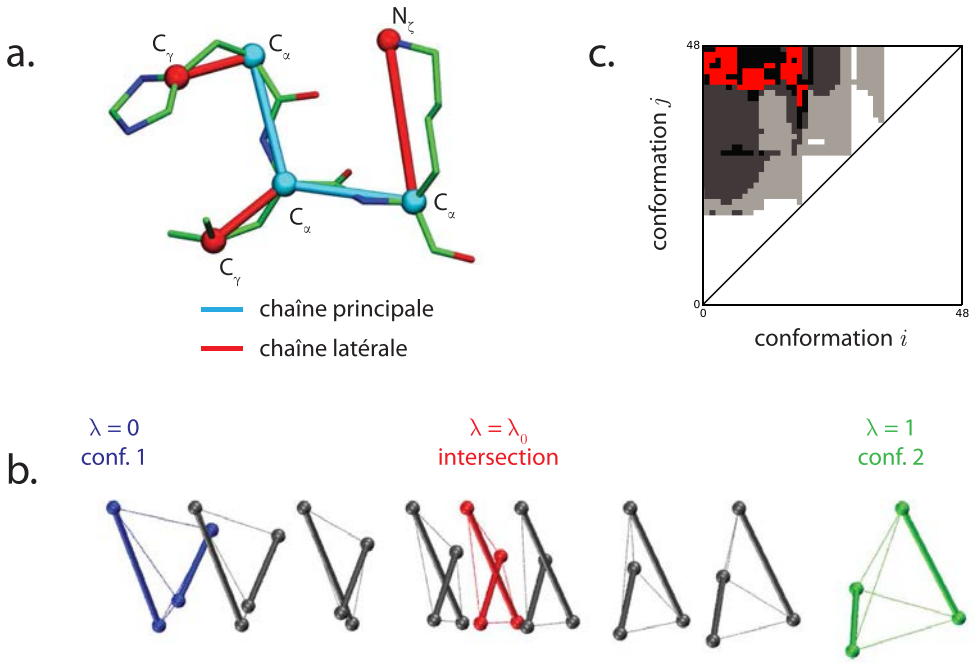


Figure I.12 – *tip-ext* : une heuristique de sélection de raccourcis dans un chemin de transition. **a.** Modélisation des chaînes principales et chaînes latérales : les chaînes principales sont modélisées par le segment entre deux C_α consécutifs, les chaînes latérales par un segment entre le C_α et un atome de la chaîne latérale situé relativement loin dans la chaîne (N_ζ pour les lysines, par exemple). **b.** Intersection de chaînes : la détection d'une intersection de deux chaînes entre deux conformations est calculée à partir de l'interpolation linéaire (de paramètre λ) du tétraèdre défini par les sommets des deux chaînes. **c.** La carte topologique d'un chemin est définie par le nombre d'intersections entre les chaînes de chaque paire de conformations i, j d'un chemin. Sur cette carte, la valeur est normalisée par le nombre I d'intersections entre la structure initiale et finale (noir : I intersections, blanc : aucune intersection, rouge : plus de I intersections).

Outre le calcul de raccourcis dans le chemin, une deuxième simplification est effectuée en réduisant le nombre de conformations intermédiaires, ce qui permet également de diminuer le nombre de raccourcis à calculer lors de l'itération suivante. Les conformations supprimées lors de cette étape de réduction sont choisies afin que l'énergie du chemin entre deux conformations successives ne dépasse pas un seuil prédéfini (figure I.11.c). Les structures linéairement interpolés entre deux conformations consécutives d'un chemin produit par *POE* ont donc une énergie inférieure à un

seuil donné, ce qui garantit que le chemin ne passe pas au travers de zones d'énergie très élevée. La réduction permet de ne garder que les conformations essentielles au chemin. En outre, le seuil d'énergie est abaissé à chaque itération de *POE* afin d'aider à l'optimisation du chemin. *POE* a été utilisé à plusieurs reprises pour le calcul de chemins de transition de protéines, notamment dans le cadre de la recherche d'inhibiteurs : récepteur de l'acide rétinoïque[110], proline racémase[111] et toxine de l'anthrax[112].

2.5 Les cavités au sein des protéines

La fonction des protéines est pratiquement toujours portée par leur forme. C'est en effet la configuration des atomes dans l'espace qui autorise les interactions spécifiques de la protéine avec son substrat ou ses partenaires. Les cavités, c'est-à-dire les concavités situées à l'intérieur ou à la surface des protéines, sont de fait particulièrement importantes pour leur fonctions. En effet, leur forme concave permet à la protéine de réaliser beaucoup d'interactions en même temps, ce qui augmente l'énergie d'interaction entre la protéine et son substrat (voir section 2.1.1).

2.5.1 Intérêt de l'analyse des cavités

Les cavités ont de multiples fonctions au sein des protéines. Tout d'abord, les cavités et autres défauts d'empilement ont un rôle ambigu dans la stabilité de la protéine. Ces vides dans la structure permettent ainsi un équilibre entre stabilité thermodynamique et flexibilité[113].

Les cavités ont souvent un rôle central dans la fonction des protéines. Les sites actifs des enzymes en sont l'exemple évident : siège de la fonction, les cavités enzymatiques ont une forme particulière permettant de fixer spécifiquement leur substrat. A ce titre, les cavités des sites enzymatiques sont régulièrement étudiées en parallèle de leur structure. Le lien entre les cavités et la fonction d'une protéine peut également apparaître de façon plus subtile et moins spécifique. La myoglobine contient ainsi des cavités dans lesquelles de petits ligands (typiquement le dioxygène O_2, mais également les monoxydes de carbone (CO) et d'azote (NO)) peuvent se déplacer[114, 115, 116]. Ces cavités servent aussi de voie de passage vers l'extérieur de la protéine pour ces petites molécules[117, 118, 119, 120, 121, 122], ce qui influe sur leurs constantes cinétiques d'association. Les cavités occupent également une place centrale dans la fonction des protéines de la famille des cytochromes p450 : le site actif, situé en plein milieu de la protéine, n'est atteignable qu'en passant par une multitude de réseau de cavités servant simultanément de voie d'accès et de crible[123, 124, 125].

Enfin, la détection des cavités est une étape centrale dans un grand nombre de logiciels d'arrimage moléculaire et de criblage virtuel. En effet, la complémentarité de forme entre un ligand et la cavité du site d'intérêt est un facteur clé de la fonction de score utilisée dans ces logiciels[53, 126, 127].

2.5.2 Détection des cavités

Il est important de noter qu'il n'existe pas de définition univoque des cavités au sein des protéines. Dans ce manuscrit, je désignerai comme cavité un volume de l'espace défini par un algorithme de détection qui sera spécifié pour chaque application. Ces algorithmes sont conçus pour identifier des creux relativement concaves à l'intérieur ou à la surface de la protéine, pouvant être vides ou remplis d'eau. Chaque cavité est entourée d'atomes de la protéine qui définissent une *poche*. Cette définition des cavités a le mérite de désigner la plupart des sites d'interaction protéine-ligand.

La surface délimitant le volume d'une cavité peut elle aussi être décrite de plusieurs façons. Dans leur article séminal sur la surface moléculaire[128], Lee et Richards décrivent à la fois la surface de van der Waals d'une molécule (l'union des surfaces de van der Waals de tous ses atomes) et la surface accessible au solvant (SAS). Le volume accessible au solvant consiste schématiquement à placer une molécule de solvant (généralement l'eau, modélisable par une sphère de rayon 1,4 Å) à tous les endroits possibles de l'espace où cette molécule ne chevauche aucun atome de la protéine (figure I.13.a, étapes 1 et 2). Connolly a beaucoup développé les méthodes de calcul et de représentation de la surface accessible au solvant[129, 130]. La SAS, utilisé dans l'ensemble des chapitres de ce manuscrit, est sans doute la définition la plus utilisée dans le calcul des cavités et la plus simple d'implémentation. Les diagrammes de Voronoi et de Laguerre peuvent également être utilisés pour décrire la surface moléculaire, en décrivant les plans séparant de façon équidistante les atomes et le solvant[131].

La définition de la surface moléculaire est une première étape pour décrire les cavités et permet de déterminer les cavités non reliées au solvant. Il est toutefois nécessaire de définir un nouveau critère permettant de séparer les cavités à la surface de la protéine du reste du solvant (le *bulk*). Une méthode régulièrement utilisée lorsque l'on considère la surface accessible au solvant consiste à supprimer les positions de ce dernier qui sont également accessibles à une sphère de taille bien supérieure à celle du solvant (de 3 à 10 Å de rayon en général, figure I.13.a, étape 2 et 3). Les α-shapes[132, 133, 134, 135, 136], dérivés de la tesselation de Voronoi, sont définies comme le sous-ensemble des arêtes de la triangulation de Delaunay contenues dans le volume (de van der Waals) de la protéine. Cette définition permet de discriminer précisément les cavités des "dépressions" de la surface de la protéine, grâce à une analyse du "flux" des tétraèdres obtus vers les tétraèdres aigus définis par ces arêtes de Delaunay.

D'autres filtres peuvent également être utilisés afin de réduire le nombre des cavités détectées et favoriser les cavités intéressantes pour la fonction. Le filtre le plus utilisé discrimine les cavités selon leur volume, mais il existe aussi des filtres sur leur concavité, la régularité de leur surface, la composition en acide aminés de la poche...

2.5.3 Implémentations et logiciels de détection des cavités

Il existe de nombreux logiciels de détection des cavités, chacun ayant leur propre implémentation et variation des principes explorés dans la section précédente. Pérot *et al.*[137] ont re-

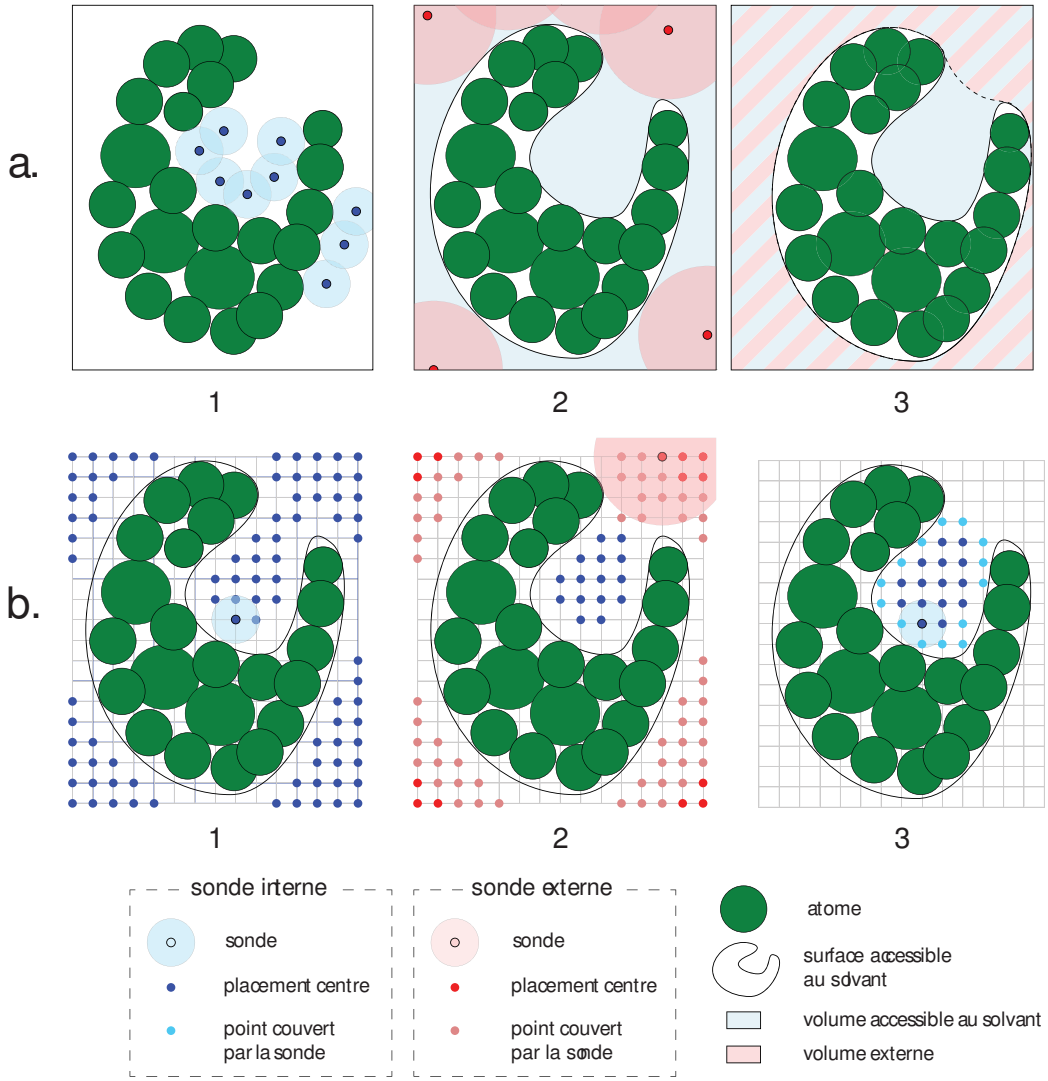


FIGURE I.13 – Le volume accessible au solvant, principe et implémentation sur une grille. a. Définition du volume accessible au solvant et des cavités associées. 1. Le volume accessible au solvant est défini par l'ensemble des positions accessibles à une petite sphère (sonde interne, disques bleus) ne recouvrant pas les atomes (disques verts). 2. Ce volume définit tout l'espace (en bleu), on utilise donc une deuxième sphère de plus grande taille (sonde externe, disque rose) pour exclure le solvant "externe" (*bulk*). 3. Les cavités sont définies par le volume accessible au solvant auquel on soustrait le volume externe (zone uniquement bleue). **b.** Implémentation des cavités type Lee et Richards par discrétisation de l'espace à l'aide d'une grille. 1. Placement des sondes internes sur les points de grille tels qu'aucun recouvrement avec les atomes n'est possible. Les points de grille pouvant acceuillir le centre de la sonde sont indiqués en bleu. 2. Placement des sondes externes, les points recouverts par les sondes externes (en rouge et rose foncé) sont supprimés. 3. Développement des centres des sondes internes sélectionnés (points recouverts en cyan). Les cavités sont définies par l'ensemble des points recouverts par les sondes internes (points bleus et cyans) dont les centres n'ont pas été supprimés par la sonde externe.

censé une grande partie de ces logiciels. La plupart des logiciels de détection des cavités appliquent des filtres pour ne garder que celles ayant le plus de chance de fixer une molécule d'intérêt pour la conception de médicament. Certains logiciels utilisent des principes différents de l'analyse géométrique de la surface de la protéine, comme les données issues de la génomiques (conservation des résidus[138, 139]) ou énergétiques (potentiels[140, 141, 142], sondes physicochimiques[143, 144]). D'autres se spécialisent pour réaliser des tâches spécifiques, comme

la détection de tunnels[145, 146, 147].

Parmis les logiciels généralistes utilisant des méthodes purement géométriques, on retrouve les cavités définies à l'aide de la tesselation de Voronoi ou des alpha-shapes : APROPOS[148], CAST[135] et CASTp[149], les algoritmes dévoloppés au sein du groupe de Deok-Soo Kim[150], fpocket[151] et ses dérivés, SplitPocket[152, 153].

Enfin, les logiciels basés sur une définition proche de la surface accessible au solvant sont les plus nombreux, et présentent une grande variété d'implémentation. Sphgen (utilitaire de détéction de cavité de DOCK[52, 53]), SURFNET[154] et PASS[155] remplissent l'espace avec des sphères. POCKET[156], LigSite[157] et ses dérivés scannent l'espace autour d'un point en suivant des directions définies par un cube autour de ce point. Enfin, TravelDepth[158], LSMS[159] PocketPicker[160], PocketDepth[161], VICE[162] et gHECOM[163], ainsi que `mkgrid`, un logiciel développé au laboratoire, implémentent le principe de la surface accessible au solvant en discrétisant l'espace à l'aide d'une grille régulière. En général, les étapes de la détection suivent le schéma indiqué dans la figure I.13.b. `mkgrid` ajoute une option supplémentaire, permettant d'avoir deux rayons de sondes externes différentes : la sphère de plus petit rayon est utilisée pour placer les centres des sondes externes (points rouges), tandis que la plus grosse sphère est utilisée pour supprimer les points délimitant le volume accessible au solvant. De ce fait, la sonde externe pénètre légèrement à l'intérieur de la protéine, ce qui permet de filtrer un grand nombre de cavités peu intéressantes situées à la surface de la protéine.

2.5.4 L'analyse dynamique des cavités

La dynamique du site de liaison peut avoir une grande importance dans l'association entre une protéine et son ligand (voir section 2.1.1). Les cavités étant le lieu privilégié des interactions protéine-ligand, il semble naturel d'en analyser la dynamique afin d'en tirer des connaissances sur la fonction, ou pour pouvoir améliorer l'étape de criblage virtuel d'un projet de conception de médicament. Pourtant, le lien entre fonction et cavités n'est généralement étudié que dans un cadre statique[164, 165, 166]. Une exception notable provient des globines : la diffusion des ligands au sein de leurs réseaux de cavités est une composante importante de leur fonction, qui a été largement étudiée[118, 119, 121, 122]. L'analyse quantitative de l'évolution de la géométrie des cavités reste toutefois assez peu étudiée, et le plus souvent uniquement à l'aide de descripteurs simplifiés (surface, volume ou position du centre géométrique)[166]. Les logiciels dédiées à la détection et l'analyse des cavités sur de multiples conformations n'ont d'ailleurs commencé à être développés que récemment : Krone *et al.*[167], MDpocket[168] (basé sur fpocket), PPIAnalyzer[169], Provar[170], TRAPP[171] ou KVFinder[172]. Ces développements récents font de l'étude dynamique des cavités un domaine en pleine expansion. A noter toutefois qu'à ma connaissance, il n'existe aucun exemple utilisant précisément la dynamique de la géométrie des cavités pour améliorer la pertinence des étapes de criblage virtuel.

3 Détermination d'inhibiteurs du virus de la dengue : exemple et genèse de l'utilisation de la dynamique des cavités pour la conception de médicament

Ce projet de détermination d'inhibiteurs du virus de la dengue a été lancé avant mon arrivée au laboratoire par Arnaud Blondel courant 2009. Ronan Rocle, à l'époque doctorant au sein du laboratoire, a réalisé la plupart des analyses décrites dans cette section. Au cours de ce projet, il aura utilisé un grand nombre d'analyses préfigurant des méthodes développées au cours de ma thèse. Ce projet est donc à la fois un exemple de projet de conception de médicament réalisé au sein du laboratoire, mais également le point de départ du développement de ces méthodes. J'ai été associé à ce projet après le départ de R. Rocle, mais je n'ai pas pu apporter de contribution technique, le projet étant resté longtemps bloqué au stade des tests biologiques.

3.1 Pathologie, mode d'action du virus et stratégie d'inhibition

3.1.1 Pathologie et implications socioeconomiques

La dengue est une maladie virale tropicale touchant entre 50 et 100 millions de personnes dans le monde[173]. Elle est transmise par la piqure des moustiques tigres (genre *Aedes*). La plupart du temps, l'infection par le virus de la dengue entraîne des symptômes type "état gripal" : forte fièvre, fatigue, céphalées, douleurs musculaires et articulaires. Une autre forme, plus rare et plus grave, existe : la dengue sévère, ou dengue hémorragique. Cette forme est caractérisée par une fièvre hémorragique, avec vomissements et douleurs abdominales. Elle touche environ 500 000 personnes par an, dont 2,5% des cas sont mortels. De façon inquiétante, le réchauffement climatique augmente les zones de viabilité du moustique vecteur de la dengue, ce qui expose environ 40 à 60% de la population mondiale au virus (figure I.14). On dénote par exemple un premier cas autochtone en France métropolitaine en 2010. La dengue est donc un problème de santé mondial au coût humain et économique[174, 175] important.

Le virus de la dengue ($DENV$) existe sous la forme de 4 sérotypes différents (dénotés DEN-1 à DEN-4), de séquences très similaires. Malgré la similarité entre les quatre sérotypes, une première infection par un sérotype ne garantit qu'une imunité temporaire aux autres sérotypes. Plus grave encore, cela augmente à plus long terme les chances de développer une forme hémorragique lors d'une seconde infection par un autre sérotype[176]. Cette particularité crée une difficulté majeure pour le développement d'un vaccin contre la dengue, car un vaccin candidat doit protéger des quatre sérotypes simultanément. A ce jour, il n'existe pas encore de vaccin ni d'antiviral sur le marché, les soins consistant à traiter les symptômes de la maladie, et la prévention à contrôler la

Figure I.14 – Zones touchées par l'épidémie de dengue en 2013 (source : OMS[173]).

population de moustiques. Toutefois, un candidat vaccin est actuellement développé par Sanofi-Pasteur et vient de passer la phase III des essais cliniques[177], avec une efficacité globale estimée à 60,8%, et des efficacités par sérotypes allant entre 42,3% et 77,7%.

3.1.2 Mécanisme d'infection et stratégie d'inhibition

Le virus de la dengue est un virus enveloppé à ARN positif, du genre *Flavivirus*. L'ARN du virus code pour 10 protéines : 3 protéines structurales C (capside), prM/M (membrane) et E (enveloppe), et 7 protéines non structurales (NS1, NS2A, NS2B, NS3, NS4A, NS4B et NS5). Le virus pénètre dans la circulation sanguine lorsque le moustique pique l'hôte. La protéine E, qui recouvre la totalité de l'enveloppe du virus, interagit avec des protéines présentes à la surface des cellules hôtes[178, 179, 180], déclenchant l'endocytose du virus. La diminution du pH dans l'endosome déclenche ensuite un changement de conformation de la protéine E, qui passe d'une forme dimérique "plate"[181] à une forme trimérique en forme de pic, qui vient se ficher dans la membrane cellulaire et la tirer vers la membrane virale, ce qui déclenche la fusion de ces dernières (cf. figure I.15). La protéine E est composée de 3 domaines : le domaine II (représenté en jaune) constitue la "pointe" de la protéine, au bout de laquelle se situe une zone hydrophobe permettant l'ancrage à la membrane dévoilée lors de l'acidification (le peptide de fusion[182, 183]). Le domaine III (bleu) est la partie venant se coller au domaine II lors du changement de conformation, afin de rapprocher les membranes virales et cellulaires. Enfin, le domaine I (rouge) est le domaine central reliant les domaines II et III.

La fusion des membranes permet l'entrée de la capside, et donc du matériel génétique du virus, dans le cytoplasme de la cellule hôte. L'ARN viral est alors traduit en une unique polyprotéine par le ribosome de la cellule hôte, et clivée à l'aide d'une protéase virale pour former l'ensemble des composants du virus. Ces protéines sont ensuite translocalisées vers le réticulum endoplasmique rugueux, où ils sont assemblés pour former les virions. A la sortie du virion par exocytose, le précurseur prM est clivé pour former la protéine M, ce qui induit un changement structural[185]. Le virion est alors arrivé à maturité et peut donc infecter de nouvelles cellules.

Pour ce projet, Arnaud Blondel et Ronan Rocle ont choisi de cibler le processus de fusion pour

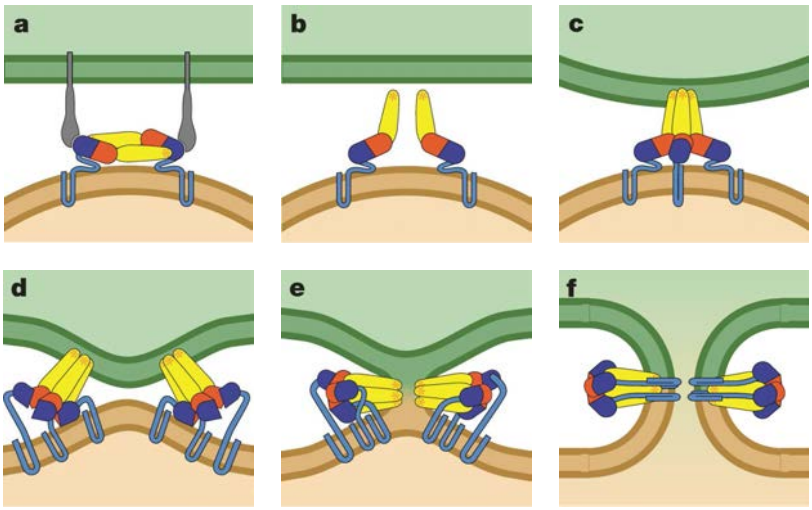


Figure I.15 – Mécanisme de fusion des membranes virales et cellulaires, déclenché par un changement de conformation de la protéine E (représentée en bleu, rouge et jaune ici). a. Forme dimérique *préfusion*. c. Forme trimérique *préfusion*. f. Forme trimérique *postfusion*. Repris de Modis 2004[184].

bloquer l'infection par le virus. L'objectif est ici de cibler la protéine E au moment du passage de sa forme dimérique "préfusion" à sa forme trimérique "postfusion", lors du changement de pH dans l'endocyte. Ce choix se justifie par l'existence des structures préfusion[181] et postfusion[184] (voire figure I.16 pour le détail des structures), ce qui permet de modéliser le mécanisme permettant de passer de l'une à l'autre.

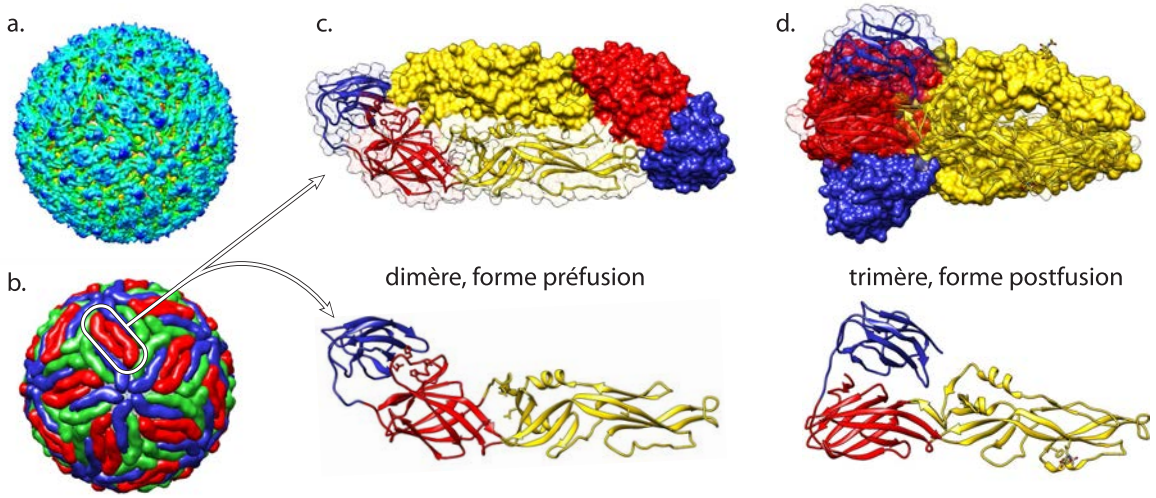


Figure I.16 – Structures de l'enveloppe du virus et de la protéine E. a. Densité électronique du virus mature à 28°C, forme préfusion. **b.** Modèle de l'agencement des protéines d'envelope. L'unité asymétrique est composée de 3 monomères de couleur rouge, bleu et vert, agencés en un icosaèdre de 90 dimères. **c.** et **d.** Structures du dimère (forme préfusion, **c.**) et du trimère (forme postfusion, **d.**) de la protéine E. Domaine I en rouge, domaine II en jaune, domaine III en bleu. En haut : surface du dimère/trimère complet ; la surface d'un des monomère est transparente pour révéler les structures secondaires sous-jacentes. En bas : conformation du monomère au sein du dimère/trimère.

3.1.3 Particularités de la cible

La particularité principale de la protéine E en tant que cible pour un projet de conception de médicament est son absence de site actif : ce n'est pas une enzyme. En effet, la fonction de la protéine E est activée par un simple changement de pH et transmise par un changement de conformation. Il est donc nécessaire de trouver un site "allostérique", pour lequel la liaison d'une petite molécule devrait permettre de bloquer le mouvement de la forme préfusion à la forme postfusion.

3.2 Chemin de transition de la forme préfusion à la forme postfusion

Plusieurs projets de conception de médicaments visent le site β-OG[186, 187, 188], situé au milieu du domaine II de la protéine et déjà exploré au cours du chapitre III. Nous cherchons à déterminer d'autres sites potentiellement intéressants pour bloquer la fonction, dans le but de nous différencier et d'apporter de nouvelles stratégies d'inhibition. Un chemin de transition d'un monomère de la forme préfusion à la forme postfusion a ainsi été réalisé afin de déterminer les sites les plus intéressants pour inhiber ce changement de conformation. Ce chemin a été produit au laboratoire en utilisant l'approche *POE* (voir section 2.4.2). Pour déterminer un chemin initial, deux *SMD* (voir section 2.3.7) partant respectivement des structures pré et postfusion sont produites par paliers, en tirant chacune des trajectoires vers la structure finale du segment précédent de l'autre trajectoire. Les structures cibles sont ainsi mises à jour régulièrement. Ce chemin a ensuite été raffiné par 11 itérations successives de *POE* réalisées en solvant implicite (*ACE*[93]), faisant passer la distance curvilinéaire de la trajectoire de 153Å (chemin initial à partir des SMDs) à 20,4Å, soit un ratio distance curvilinéaire/RMSD passant de 12,4 à 1,65 (RMSD pre-postfusion = 12.3 Å). Le nombre de conformations diminue également, passant de 305 à 58. Le chemin perd en complexité spatiale, comme l'indiquent le chemin visualisé sur les deux premières composantes principales, figure I.15.a, et la diminution de la complexité topologique, figure I.15.b. Ce chemin a été utilisé pour identifier et valider les cavités dont la forme change au cours du changement de conformation. Deux cavités ont été identifiées, au sein desquelles la fixation d'une petite molécule pourrait potentiellement empêcher leur diminution de volume et ainsi bloquer le mouvement (non présenté, projet industriel).

3.3 Dynamique moléculaire et sélection des structures

Afin de réaliser un échantillonnage des conformations, une dynamique moléculaire de 1 ns de la forme dimérique de la protéine E a été produite en solvant explicite. Les trajectoire de chacun des deux monomères ont été extraites, alignées puis concaténées pour former une unique trajectoire de 20 000 conformations. Les cavités ont été calculées sur cette trajectoire concaténée à l'aide de

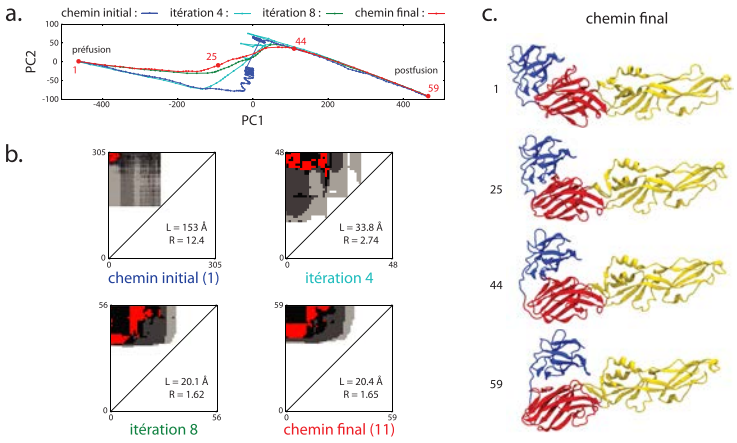


Figure I.17 – Chemin de transition de la forme préfusion à la forme postfusion. a. Projections du chemin initial après SMD (bleu) et du chemin final après 11 itérations de *POE* (rouge) sur les deux premières composantes principales (PC1-2) du chemin initial (abscisse : PC1, ordonnée : PC2). **b.** Evolution de la carte topologique des chemins de transition. Les cartes sont données pour les trajectoires résultant des itérations 1, 4, 8 et 11 de *POE* (de gauche à droite et de bas en haut). Un point i, j de la carte correspond au nombre d'intersections de chaînes principales ou secondaires lors du passage de la conformation i à la conformation j (voir 2.4 pour plus de détail sur le calcul de ce nombre). Cette valeur est codée en niveau de gris, en normalisant par le nombre d'intersections entre les conformations initiales et finales (noir : 1, nombre d'intersections entre conformations initiales et finales ; blanc : 0, pas d'intersection). Lorsque la valeur pour un point i, j est supérieure à la valeur entre les conformations initiales et finales, il est représenté en rouge. **c.** Changement de conformation du monomère au cours du chemin final après la 11^{e} itération de *POE*. De haut en bas : structure initiale, structure 24/59, structure 43/59, structure finale.

`mkgrid`. Les poches sont définies comme l'ensemble des acides aminés situés à moins de 6 Å de chaque cavité.

L'analyse des cavités et de la dynamique structurale a été réalisée par Ronan Rocle au laboratoire, avant la conception des méthodes de suivi et d'analyse des cavités explorées dans les chapitres II et III. Une analyse visuelle des cavités a permis de révéler leur fragmentation dans un grand nombre de conformations. Les cavités ont été isolées à l'aide de boîtes englobantes déterminées visuellement. L'évolution et la fragmentation de ces cavités ont été étudiées en calculant le centre géométrique de chaque cavité (globale ou fragments), et en utilisant là aussi des boîtes englobantes. Résolue dans cette étude grâce à des critères de position et de géométrie, l'identification des cavités s'est toutefois clairement posée.

La dynamique structurale des poches des cavités sélectionnées était analysée en utilisant l'ACP (voir principe général section 2.1). Cela nécessitait de définir une poche qui convienne à l'ensemble des cavités de la trajectoire, fragments ou non. La projection des trajectoires d'une des poches sur ses trois premières composantes principales est représentée figure I.18, en même temps que la variation de son volume (représentée par un code couleur). Cette projection est divisée en plusieurs zones, autant que de conformations différentes utilisées pour le criblage. Les conformations utilisées pour le criblage sont ensuite séléctionnées en prenant les structures les plus proches des centres de ces zones, puis en séléctionnant les conformations de plus petit ou grand volume, et enfin les conformations dont l'énergie éléctrostatique est la plus favorable.

L'utilisation des cavités faite ici préfigure des outils développés aux chapitres II et III. Les

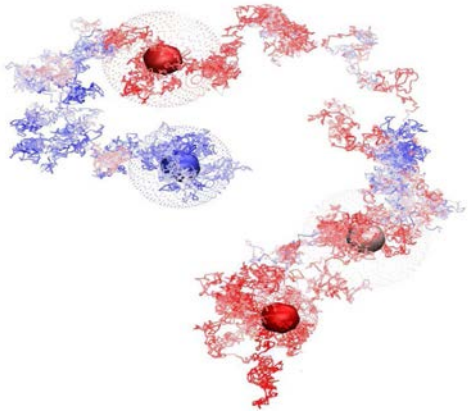


Figure I.18 – Projection de la trajectoire d'une poche sur ses 3 premières composantes principales. Les points correspondant à chaque structure de la poche sont reliés dans l'ordre de la trajectoire. La couleur du trait correspond à la variation du volume de la cavité associée, du bleu (faible volume) au rouge (grand volume). Les structures choisies sont indiquées par des sphères.

méthodes utilisées sont laborieuses et manuelles, et leur automatisation a été un élément central de mon travail. L'utilisation de boîtes englobantes pour extraire les cavités d'intérêt a laissé la place à une méthode plus précise et systématique exposée au chapitre II. De même, bien qu'une classification relativement poussée des cavités ait été réalisée, la variation de la géométrie des cavités n'est pas précisément prise en compte dans la sélection des conformations. Ce point sera traité au chapitre III.

3.4 Criblage et sélection des molécules

Le criblage a été réalisé à l'aide du logiciel FlexX, en utilisant la Chimiothèque Nationale, une métachimiothèque de composés produits dans divers laboratoires français, ainsi que la ChemDiv, une chimiothèque commerciale. A l'époque du criblage, elle comprenait environ 40 000 composés. Le criblage a été réalisé sur 12 conformations séléctionnées à l'étape précédente. Pour chaque conformation, l'état de protonation des histidines présentes a été varié, ce qui représente une quarantaine de criblages différents. Un classement entrelacé des composés a ensuite été utilisé afin de réaliser la séléction sur plusieurs critères à la fois (énergie minimale, énergie moyenne, famille chimique ; voir chapitre IV section 2.4 pour le détail de l'algorithme utilisé). Les 200 molécules de plus haut rang ont ensuite été commandées et envoyées au laboratoire de Félix Rey pour réaliser des tests d'efficacité.

3.5 Tests préliminaires

Une première série de tests a d'abord été envisagée, comprenant des mesures biophysiques (dénaturation thermique, microcalorimétrie) et biochimiques (test de fusion des membranes). Ces tests ont été par la suite jugés trop complexes et trop coûteux en protéines. Ils ont donc été abandonnés au profit de tests d'infection du virus sur des cellules Vero. Une solution de virus et

de composés est mise en contact avec un tapis de cellules en culture, puis rincée. Après un temps d'incubation, les plages de cellules mortes sont comptées : plus le composé est efficace, moins les cellules meurent.

Ces tests d'infections ont été réalisés par Joe Cockburn au sein du laboratoire de Félix Rey. Cette série, réalisée en trois exemplaires sur 80 composés, a permis d'identifier une dizaine de composés potentiellement efficace. Des mesures d'effets-doses pour ces composés sont attendus afin de s'assurer de cette efficacité et de commencer l'étape d'optimisation des composés.

4 Objectifs de la thèse

Le principal objectif de cette thèse est de proposer de nouveaux inhibiteurs ou effecteurs dans différents projets de conception de médicaments, éventuellement selon un mécanisme d'action novateur. Ces effecteurs sont sélectionnés en associant des méthodes d'analyse du fonctionnement d'une protéine, notamment la dynamique des cavités, avec des méthodes de criblage virtuel. L'approfondissement des méthodes d'analyse des cavités est également un point central de ma thèse.

4.1 Proposition d'inhibiteurs potentiels pour différentes cibles

J'ai eu l'opportunité de contribuer à trois projets de conception de médicaments au cours de ma thèse. Le premier projet vise l'ADN-gyrase de *Mycobacterium tuberculosis* et a été entrepris en collaboration avec le groupe de Claudine Mayer à l'Institut Pasteur. Un deuxième projet, en collaboration avec Sanofi, a pour objectif de déterminer de nouvelles molécules ciblant les parasites responsables du paludisme en bloquant la subtilisine de type 1, protéine nécessaire à la sortie du parasite des érythrocytes. Enfin, un dernier projet vise à déterminer des effecteurs d'un récepteur procaryote ionotrope modulé par le proton, GLIC, homologue de divers récepteurs neuronaux humains, dont les récepteurs $GABA_A$ et les récepteurs nicotiniques.

4.2 Développement d'outils d'analyse dynamique des cavités

Afin d'améliorer la pertinence des criblages réalisés au cours de ces différents projets, au delà de l'analyse des informations de la littérature, des données fournies par nos collaborateurs et du fonctionnement des cibles, j'ai utilisé au maximum les informations obtenues à partir de l'analyse des cavités des protéines ciblées. Pour obtenir ces informations, j'ai développé de nouvelles méthodes d'analyse de la dynamique des cavités, et j'ai pu les éprouver sur ces différentes applications.

4.2.1 Suivi des cavités

Lors d'un projet impliquant une étape de criblage virtuel, il est nécessaire de choisir un site particulier, donc une cavité particulière. Il est donc nécessaire de pouvoir identifier les différentes

cavités apparaissant au cours de la dynamique, et de pouvoir extraire les propriétés de chacune : variation du volume, de la surface, des résidus composant la poche, de la forme. Cette identification n'est pas un problème trivial, car les cavités au sein des protéines sont un ensemble à la définition subjective et très labile : elles disparaissent, apparaissent, se divisent, fusionnent entre elles au cours du temps. Le chapitre II traite donc du suivi des cavités d'une protéine sur une multitude de conformations. Le suivi des cavités permet ainsi de réaliser des analyses systématiques des cavités d'une protéine, facilitant et rationalisant le choix des cavités à cibler et leurs conformations.

4.2.2 Analyse de la dynamique des cavités et sélection de conformations pour le criblage

Une fois le site choisi, il faut choisir une ou plusieurs conformations des résidus composant le site pour pouvoir démarrer le criblage virtuel. La plupart du temps, les conformations choisies proviennent de structures obtenues expérimentalement. Des criblages sur des conformations issues de modélisation moléculaire ont été entrepris avec succès[49, 50, 51], généralement en utilisant la dynamique moléculaire[50, 69, 70, 71], ou bien des techniques d'échantillonnage utilisant des contraintes sur la structure[72, 73, 74]. Malheureusement, ces techniques ne font que peu référence aux cavités sous-jacentes bien qu'elles constituent un élément important de l'association protéine-ligand en posant une contrainte forte sur la forme du ligand. Lorsque c'est le cas, ces techniques utilisent des indicateurs ne prenant pas en compte les détails géométriques des cavités : volume, surface, position du centre géométrique... Afin d'améliorer la qualité du criblage, j'ai donc développé un ensemble de méthodes permettant de choisir ou de construire des conformations suivant l'évolution des cavités choisies. J'ai choisi d'analyser la dynamique des cavités en utilisant l'analyse par composantes principales (ACP), et l'ensemble du chapitre III est donc dédié à l'étude de l'application de cette méthode sur les cavités.

4.2.3 Développement d'un logiciel pour l'automatisation des analyses des cavités au sein des protéines

L'ensemble de ces méthodes (suivi des cavités, ACP sur les cavités) ont été implémentées dans un module Python destiné à faciliter la manipulation des cavités aussi bien sous forme statique que dynamique. Le logiciel sera disponible à l'adresse http ://TODO.fr/. Les fonctionnalités et l'architecture du programme ainsi qu'une courte introduction de son fonctionnement sont données en annexe, section 1.

Chapitre II

Détection et suivi des cavités au sein des protéines

1 Les cavités au sein des protéines : un ensemble labile et non univoque, difficile à suivre

1.1 Suivre les cavités au cours du temps : intérêt et exemples d'applications

L'étude de la dynamique des cavités est un domaine assez récent et en pleine expansion comme indiqué au chapitre précédent. Lors d'un projet impliquant un criblage virtuel ou pour analyser la fonction d'une protéine, il peut être intéressant de suivre l'évolution de certaines propriétés d'une cavité en particulier. Ces propriétés peuvent être le volume de la cavité ou sa surface, ou encore des propriétés associées aux résidus entourant la cavité : hydrophobicité, conservation... Il peut également être intéressant de suivre l'évolution de la géométrie de la cavité à l'aide de méthodes comme l'analyse en composantes principales (voir chapitre III pour une étude détaillée de ce point). Pour extraire ces propriétés d'une cavité particulière, il est toutefois nécessaire d'identifier la cavité parmi les autres : c'est le but du suivi des cavités.

Les applications du suivi des cavités sont multiples. Lors de l'étude de la fonction d'une protéine, cette analyse permet de repérer des cavités temporaires, n'apparaissant pas dans la structure cristallographique, et d'estimer leur impact éventuel sur la fonction. On peut également étudier la stabilité de la forme d'une cavité et en tirer des indications sur son rôle fonctionnel. Il est aussi possible de détecter des phénomènes de "respiration" des réseaux de cavités en étudiant les corrélations ou anticorrélations entre différentes cavités. Ce type d'étude peut avoir son importance dans des familles de protéines comme les globines[118, 119, 121, 122] ou les cytochromes p450[123, 124, 125]. Lors de projet impliquant un criblage virtuel, le suivi des cavités peut per-

mettre la détermination de sites allostériques. Identifier une cavité d'intérêt et en extraire les propriétés permet également d'aider à la séléction des conformations à utiliser lors d'un criblage. Ces propriétés peuvent en effet jouer sur la qualité des composés séléctionnés :

— l'utilisation de conformations pour lesquelles la cavité étudiée a des volumes variés permet de sélectionner des composés de structures chimiques plus diverses

— l'étude de la conservation des résidus de la poche permet de sélectionner des conformations ne présentant que des résidus conservés, afin de séléctionner des composés n'interagissant pas avec des résidus susceptibles de muter et diminuer ainsi le risque d'apparition de résistances.

— l'étude de l'hydrophobicité de la poche permet d'éviter d'orienter la sélection vers des composés peu solubles

Il est donc particulièrement intéressant de pouvoir identifier une ou plusieurs cavités et de les suivre au cours du temps. Malheureusement, ce n'est pas un problème simple du fait de la grande variabilité des cavités. De fait, il n'a à ce jour et à ma connaissance quasiment jamais été traité précisément dans des études scientifiques (à l'exception notable des travaux d'Eyrisch et Helms[189, 166]).

1.2 Problématique du suivi des cavités

1.2.1 Une grande variété de définitions, sans consensus

L'étude de la dynamique d'une structure est une tâche relativement aisée théoriquement, puisqu'il suffit de suivre l'évolution des coordonnées des atomes au cours du temps. Cette définition de la structure par ses coordonnées atomiques est univoque, ce qui n'est pas le cas des cavités. Nous avons pu voir dans le chapitre précédent que la définition même des cavités est sujette à diverses interprétations. L'étude de la dynamique des cavités se heurte donc à un premier problème, qui est le choix de leur représentation. On peut définir quatre différences principales entre ces interprétations des cavités :

— le choix de la surface interne, près de la protéine présente le moins de diversité : SAS, tesselation de Voronoi...

— le choix de la surface externe, séparant la cavité du solvant (*bulk*), peut être très variable : grande sphère exclusive, distance aux atomes, flux de tétraèdres, notion de profondeur...

— le choix de la représentation numérique des cavités entraîne également des différences de forme (sphères, points de grilles, tétraèdres...)

— enfin, les filtres raffinant la surface et supprimant les cavités jugées "peu intéressantes" sont légion ; chaque logiciel possède sa propre spécificité à ce niveau

L'immense variété des paramètres jouant sur la présence et la forme des cavités fait qu'une cavité détectée par un logiciel n'aura pas la même forme voire sera absente en utilisant un autre logiciel.

1.2.2 La variabilité des cavités au cours du temps

Chacun de ces paramètres agit comme un seuil, une limite plus ou moins artificielle sur un espace continu, l'extérieur de la protéine. Ces seuils sont problématiques lorsque l'on veut comparer les cavités provenant de deux logiciels différents, comme on vient de le voir. Malheureusement, ils sont tout aussi problématiques lorsque l'on veut comparer des cavités de deux conformations différentes d'une même protéine. En effet, de petites variations dans les coordonnées peuvent potentiellement avoir de gros effets sur les cavités détectées. Par exemple, une petite cavité peut diminuer légèrement de volume et passer en dessous du seuil de volume minimal, ce qui entraîne sa disparition (figure II.1.a). De façon moins extrême, un goulot d'étranglement dans une cavité peut se réduire suffisamment pour diviser la cavité en deux (figure II.1.b). Inversement, deux cavités pouvant apparaître bien distinctes selon différents critères (géométrique, fonctionnel...) peuvent fusionner, perdant ainsi leur identité. Les cavités sont donc un ensemble en pratique très labile, sujet au cours d'une dynamique à de très nombreux événements d'apparition, de disparition, de division et de fusion. De plus, la "respiration" des protéines fait que des cavités peuvent se former à n'importe quel endroit en leur sein[190]. Le suivi des cavités se pose alors comme un problème en pratique complexe et difficile à traiter, mettant à mal la conception que l'on en a sur une structure statique.

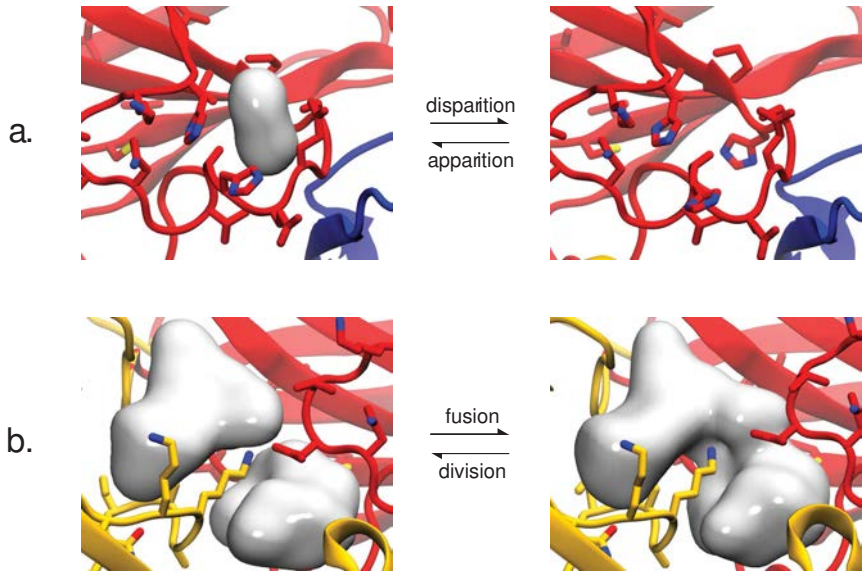


Figure II.1 – Exemples d'événements liés à l'utilisation de seuils dans la définition des cavités. a. Evènement de disparition/apparition de cavité lorsque la cavité passe en dessous du seuil de volume. **b.** Evènement de fusion/division lié à la définition de la surface de la cavité. L'écart entre les conformations de gauche et de droite est de 10ps et le changement structural sous-jacent est très limité.

1.3 Esquisses de solutions

1.3.1 Suivi d'une cavité unique lors d'une dynamique

Ces événements d'apparition/disparition et de fusion/division sont problématiques lorsqu'il s'agit de suivre l'évolution des cavités au cours du temps. En effet, comment traiter une cavité lorsqu'elle se divise ? Doit-on alors suivre une seule des deux cavités résultantes, et si oui laquelle ? Si on décide de suivre les deux cavités, comment traiter les cas de fusion sans risquer de s'éloigner de la cavité d'origine ? Comment suivre la cavité lorsqu'elle disparaît pour réapparaître un instant plus tard en s'étant déplacée ? Dans ce genre de situation, on risque rapidement soit de ne suivre l'évolution que d'une petite partie d'une cavité, soit d'étendre progressivement l'étude à un grand nombre de cavités différentes.

Il existe plusieurs solutions pour contourner ce problème lorsqu'on ne s'intéresse qu'à "une seule" cavité. La solution la plus simple est de limiter la détection de cavité à la région de l'espace autour de la cavité d'intérêt, typiquement à l'aide d'une forme géométrique comme un cube ou un cylindre[167, 191]. Cette méthode peut manquer de précision lorsque la cavité d'intérêt dépasse le cadre ou quand d'autres cavités rentrent dans le cadre. Elle n'est en outre pas du tout adaptée à l'étude de systèmes soumis à de larges déplacements de sous-domaines. Une solution pour éviter ces problèmes est de spécifier un groupe de résidus clés (résidus catalytiques, en contact avec un ligand...) et de filtrer les cavités pour ne garder que celles étant suffisament proches de ces résidus.

1.3.2 Suivi de l'ensemble des cavités d'une protéine

Etendre les solutions du suivi d'une cavité unique au suivi de l'ensemble des cavités d'une protéine n'est pas immédiat. La solution du découpage de l'espace n'est plus applicable, puisque l'espace considéré est l'ensemble de la protéine. De même, il est très laborieux de définir des résidus clés pour l'ensemble des cavités de la protéine, ce qui demande une très bonne connaissance du système et de ses cavités.

Je propose une méthode qui consiste à réaliser une correspondance entre chaque cavité et les résidus les entourant. J'appelerai les résidus en contact avec la cavité l'*empreinte* de la cavité sur la structure. Les empreintes des cavités de l'ensemble des conformations d'une dynamique peuvent être regroupées par similarité en utilisant un algorithme de partitionnement (voir annexe 3). Chaque groupe déterminé par l'algorithme définit alors une cavité unique tout au long de la trajectoire. Suivre une cavité consiste à ne considérer que les cavités appartenant à son groupe. Eyrisch et Helms avaient proposé une solution similaire pour résoudre ce problème[189, 166], passée malheureusement relativement inaperçue, et que nous avons redécouverte indépendamment, Arnaud Blondel et moi-même. A noter qu'Ashford *et al.*[170] ont également développé l'idée d'utiliser les empreintes des cavités dans leur logiciel Provar sans toutefois l'appliquer au suivi des cavités.

Ce chapitre est dédié au développement, à l'amélioration et la généralisation de cette méthode. L'approche que j'ai développée initialement consiste à utiliser un algorithme de partitionnement

hiérarchique en considérant des empreintes booléennes (résidus situés à moins d'une certaine distance seuil des cavités) comparées par la distance de Jaccard. L'approche utilisée par Eyrisch et Helms est similaire, la différence principale provenant de l'ajout d'une contrainte empêchant l'existence de plus d'une unique cavité de chaque groupe par conformation. J'ai décidé pour ma part de ne pas utiliser cette contrainte, considérant qu'il est acceptable d'avoir des cavités pouvant se diviser en plusieurs morceaux au cours d'une dynamique. Différents algorithmes de partitionnements et différentes empreintes seront testés pour tenter de définir la méthode la plus pertinente et la plus robuste pour le suivi de l'ensemble des cavités. J'utiliserai les cavités détectées par `mkgrid` dans ce chapitre. La méthode est toutefois généralisable à n'importe quelle méthode de détection grâce au système de correspondance cavité-empreinte.

2 Méthodes et controles

2.1 Protéines étudiées, dynamique moléculaire et détection des cavités

Nous avons utilisé quatre systèmes pour tester l'efficacité des algorithmes de suivi des cavités :

— la myoglobine de cachalot

— le dimère de la protéine d'enveloppe du virus de la dengue

— le facteur œdemateux de la toxine d'anthrax (EF)

— la tyrosine kinase Abl (ABL1)

Ces quatre systèmes ont été choisis car ils possèdent tous plusieurs sites intéressantes d'un point de vue fonctionnel. Ces différents sites sont explicités section 2.7.

Les paramètres de dynamique moléculaire utilisés pour chacun de ces quatre systèmes sont indiqués en annexe, section 2. Pour EF, deux trajectoires provenant de deux dynamiques réalisées respectivement en présence et en absence de la calmoduline ont été concaténées puis échantillonnées pour former une trajectoire de 2000 conformations. Les autres dynamiques ont été échantillonnées afin d'obtenir des trajectoires de 1000 conformations chacune. Les cavités ont été détectées pour chaque conformation de ces trajectoires à l'aide de `mkgrid`, en utilisant une sonde interne de 1.4 Å de rayon, une petite sonde externe de 5.2 Å de rayon (placement des centres) et une grande sonde externe de 8 Å de rayon (suppression des points de grille externes).

2.2 Principe général de l'algorithme du suivi des cavités

Les cavités sont détectées pour chaque conformation d'une trajectoire. Chaque cavité de chaque conformation possède un identifiant unique déterminé par la méthode de détection (`mkgrid`) ne permettant pas de suivre la cavité au cours du temps. J'appellerai ces cavités non suivies les cavités *instantanées*. Les empreintes de chaque cavité instantanée sont calculées puis regroupées à l'aide

d'un algorithme de partitionnement. Les groupes ainsi formés définissent les cavités suivies au cours du temps, ce sont les cavités *transverses*. Lorsqu'une cavité instantanée semble provenir de la fusion de plusieurs cavités transverses, elle peut être découpée en plusieurs lobes. Les empreintes de ces lobes sont ensuites utilisées pour assigner un numéro de cavité transverse à chaque lobe en utilisant le résultat de l'algorithme de partitionnement précédemment réalisé. Un schéma explicatif du processus est donné en figure II.2.

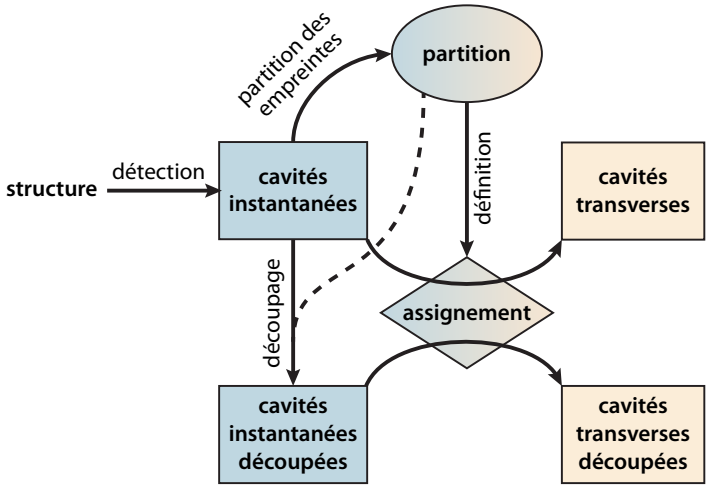


Figure II.2 – Schéma de l'algorithme de suivi des cavités.

2.3 Définition des empreintes structurelles des cavités

Une définition assez naturelle consiste à utiliser les atomes situés à moins de 5 Å d'une cavité pour définir son empreinte, définition d'ailleurs utilisée par Eyrisch et Helms. Cette définition très précise produit des empreintes de taille $3n_{atome}$, qui peuvent prendre beaucoup de temps à être traitées par les algorithmes de partitionnement. Pour réduire la taille de ces empreintes, je propose de définir des *groupes structuraux* comme des sous-ensembles d'atomes de la protéine. Les empreintes sont alors décrites non plus sur les atomes mais sur les groupes structuraux. Typiquement, ces groupes peuvent être les résidus (définissant des empreintes de taille n_{res}), mais aussi les chaînes principales et chaînes latérales des résidus pris séparément (taille $2n_{res} - n_{GLY}$) ou tout simplement les atomes.

Ces empreintes sont purement booléennes, basées sur un seuil de distance. Il est possible de définir d'autres empreintes en se basant sur la valeur de distance minimale entre un groupe structural et la cavité. Je définis donc trois types d'empreintes différentes (voir figure II.3) :

— les empreintes de *distance* fp^{dist} considèrent la distance euclidienne minimale entre la cavité c et le groupe structural g :

$$fp_c^{dist}(g) = \min_{i \in c, j \in g} d(i,j)$$

avec $d(i,j)$ la distance euclidienne entre le point de grille i et l'atome j. Ces empreintes ne seront pas utilisées dans ce chapitre, car elles mettent l'accent sur les groupes structuraux

éloignés de la cavité considérée, ce qui conduit à des assignements peu satisfaisants, notamment pour les systèmes présentant de larges mouvements (données non précisées ici). Elles sont toutefois utiles pour définir les deux autres types d'empreintes.

— les empreintes booléennes ou empreintes de *contact* sont définies à partir des empreintes de distance à l'aide d'un seuil de distance $d_{contact}$:

$$\begin{aligned} fp_c^{contact}(g) &= \left(\min_{i \in c, j \in g} d(i,j) < d_{contact} \right) \\ &= \left(fp_c^{dist}(g) < d_{contact} \right) \end{aligned}$$

— les empreintes de *contact-distance* dérivent des empreintes de distance mais tombent à 0 au delà d'une distance seuil. Elles conservent donc le contexte local des empreintes de contact tout en limitant les effets de bords engendrés par le saut de valeur de 0 à 1 des empreintes de contact. Elles peuvent être définies à partir des empreintes de distance à l'aide d'un seuil de distance $d_{contact}$:

$$\begin{aligned} fp_c^{contdist}(g) &= max(0, d_{contact} - \min_{i \in c, j \in g} d(i,j)) \\ &= max(0, d_{contact} - fp_c^{dist}(g)) \end{aligned}$$

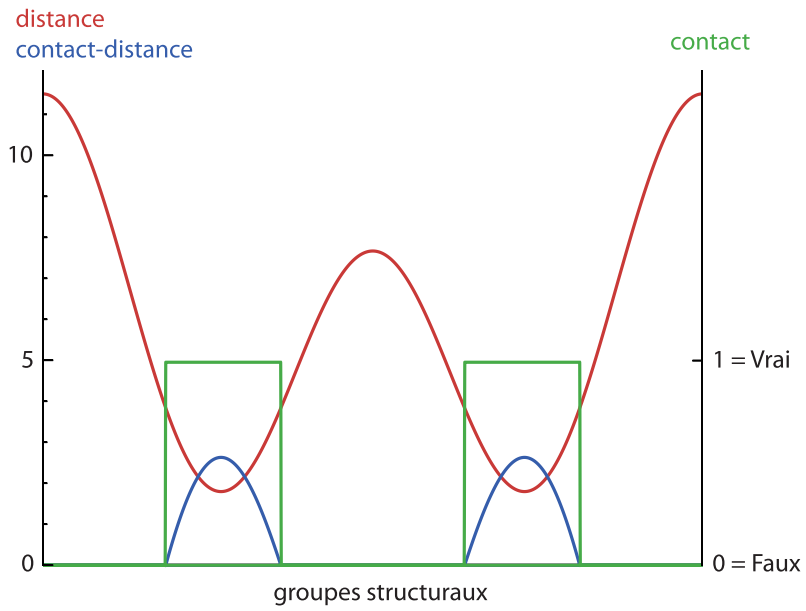


FIGURE II.3 – Forme des trois types d'empreintes. En abscisse, la liste des groupes structuraux. En ordonnée, la valeur prise par l'empreinte pour chaque groupe. En rouge, empreinte de distance ; en bleu, empreinte de contact-distance ; en vert, empreinte de contact (valeurs booléennes).

2.4 Partitionnement des empreintes

2.4.1 Les différentes distances envisagées

Le partitionnement d'un ensemble de points en haute dimension (ici les empreintes) demande la plupart du temps d'avoir une mesure de distance entre les points. Les distances utilisées peuvent être adaptées à différent types d'empreintes : réelles (distance, contact-distance) ou booléennes (contact). Dans ce chapitre, je considèrerai quatres distances différentes :

— la distance euclidienne (empreintes réelles) :

$$d(a,b) = \sqrt{\sum_{i}^{n} (a_i - b_i)^2}$$

— la distance cosinus (empreintes réelles) :

$$d^{cos}(a,b) = \frac{a \cdot b}{\|a\|\|b\|}$$

— la distance de Jaccard (empreintes booléennes) :

$$d^{rec}(a,b) = \frac{|a \cap b|}{|a \cup b|}$$

— la distance utilisée par Eyrisch et Helms (EH, empreintes booléennes) :

$$d^{EH}(a,b) = \frac{|a \cap b|}{\min(|a|,|b|)}$$

Je calcule ensuite la distribution des distances entre empreintes pour chacune des distances considérées (figure II.4). On remarque que pour la plupart des systèmes et des distances, la distribution est plus ou moins bimodale. Il existe un premier ensemble de faibles distances, et un ensemble plus grand de grandes distances. L'ensemble de faibles distances correspond aux distances entre les empreintes de cavités proches dans l'espace, il comprend donc les distances entre empreintes de mêmes cavités transverses. Il est donc possible de définir une distance seuil d_{seuil} située entre ces deux modes afin de sélectionner préférentiellement les paires d'empreintes susceptibles de correspondre à la même cavité transverse. J'ai donc choisi d_{seuil} comme la distance correspondant au premier minimum local de la distribution après avoir passé le premier maximum local (correspondant au premier mode). Cette distance sera utilisée dans les algorithmes de partitionnement afin d'éviter de définir manuellement le nombre de partitions (donc le nombre de cavités transverses) voulues. Elle est représentée par un point sur la courbe dans la figure II.4.

Cette définition peut être problématique. En effet, les distributions des distances euclidiennes, Jaccard et EH des empreintes d'ABL1 ne sont pas bimodales, la distance seuil ne peut donc pas être définie pour ces distributions. Par ailleurs, les modes correspondant aux faibles distances apparaissent plus nettement pour la distance cosinus. La distance cosinus appliquée aux empreintes

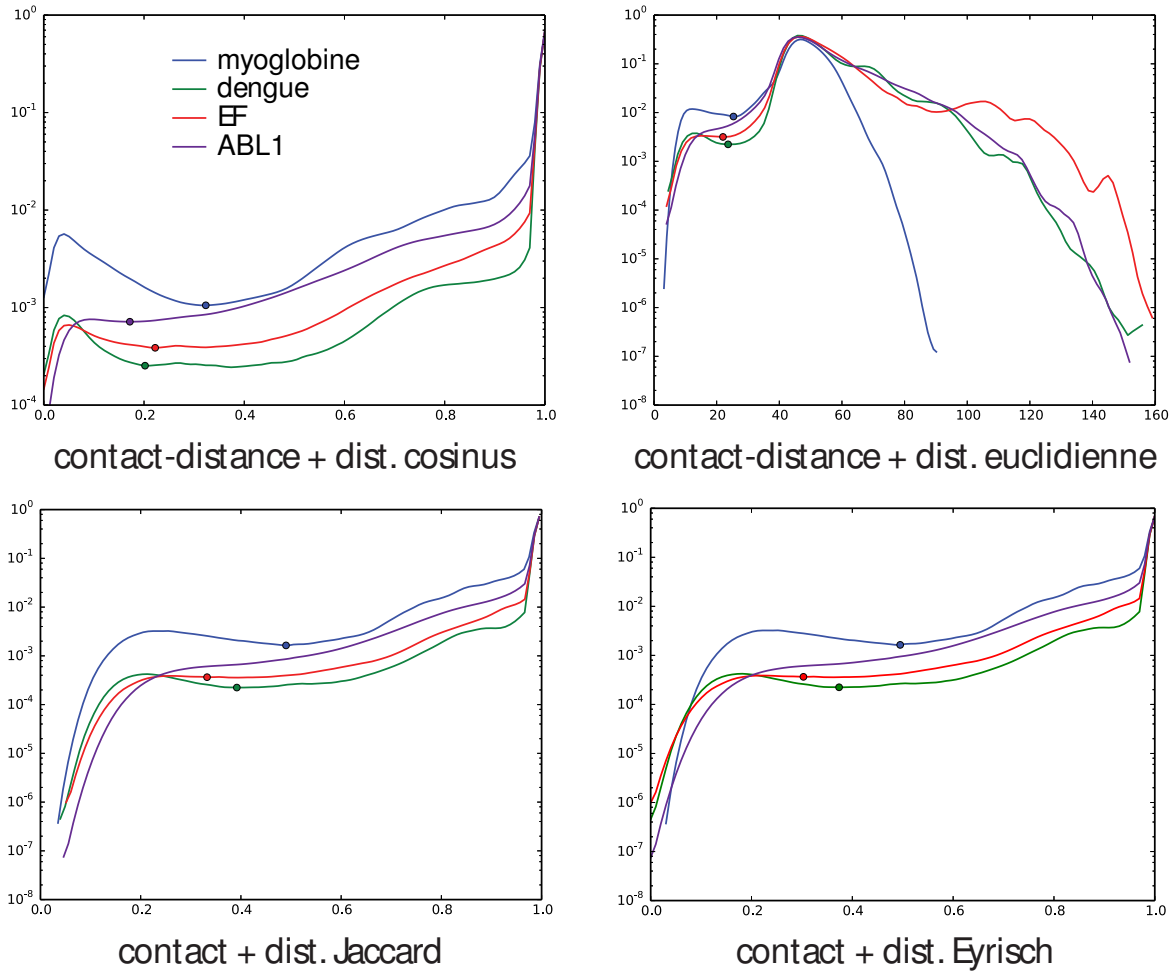


Figure II.4 – Distribution des distances entre toutes les paires d'empreintes pour la myoglobine (courbe bleue), la protéine E du virus de la dengue (vert), EF (rouge) et ABL1 (cyan). L'échelle en ordonnée est logarithmique ; la distribution est normalisée et lissée par une fenêtre glissante de Hamming de largeur 0.05. De gauche à droite et de haut en bas : distance cosinus sur empreintes de contact-distance ; distance euclidienne sur empreintes de contact-distance ; distance Jaccard sur empreintes de contact ; distance utilisée par Eyrisch et Helms sur empreintes de contact.

de contact-distance semble donc plus robuste. Elle sera donc utilisée en priorité, sauf pour comparer les résultats du suivi des cavités avec ceux provenant de la méthode d'Eyrisch et Helms. Le seuil de distance utilisé par Eyrisch et Helms est fixé à 0.85, comme indiqué dans leur article. J'utiliserai également des seuils de distance fixes respectivement choisis à 0.15, 0.3, 0.5, 0.7, 0.8, 0.85 et 0.9.

2.4.2 Les différents algorithmes de partitionnement : avantages et inconvénients

Une fois la matrice de distances et la distance seuil établies, on peut réaliser le partitionnement et l'assignement des empreintes. Ces deux étapes sont séparées : le partitionnement permet de définir des groupes d'empreintes similaires, tandis que l'assignement utilise le résultat du partitionnement pour donner un identifiant à chaque empreinte. Cette décomposition a deux avantages, étant donné que le partitionnement est l'étape la plus coûteuse en temps de calcul :

— il est possible de réaliser le partitionnement sur un sous-ensemble des empreintes puis de réaliser l'assignement après coup sur l'ensemble des empreintes

— il est possible de réaliser l'assignement des lobes des cavités sans avoir à refaire le partitionnement après l'étape de découpe de ceux-ci

De nombreux algorithmes de partitionnement ont été décrits dans la littérature. Rui *et al.* en ont réalisé une revue relativement récente et complète[192]. Chaque classe d'algorithmes est plus ou moins adaptée aux différents problèmes de partitionnement existant. Pour la méthode de suivi des cavités, il est nécessaire de considérer des algorithmes pouvant réaliser le partitionnement de données assez nombreuses (n_{inst}, soit jusqu'à 72507 pour EF) à partir d'une matrice de distances (et non des données d'entrée). La topologie des données du problème est également spécifique, puisque comme vu dans la section précédente, je recherche des groupes d'empreintes proches les unes des autres mais pas forcément très séparés des autres groupes. Le nombre de partitions doit être également déterminé à partir des données, préférentiellement à partir d'un critère de distance (voir section précédente). L'algorithme se doit enfin d'être suffisamment rapide pour traiter ces données en un temps raisonnable (quelques heures tout au plus), et éventuellement d'être suffisamment simple pour être programmé aisément ou disponible dans une librairie dédiée facile d'accès. A partir de ces critères, j'ai identifié quatre algorithmes de partitionnement : DBSCAN, le partitionnement spectral de graphe et les partitionnements hiérarchiques de type *UPGMA* et *complet*[1]. Ces algorithmes sont décrits en détail en annexe 3, mais une courte description de leur fonctionnement ainsi que de leurs avantages et leurs inconvénients pour le partitionnement des empreintes de cavités sont exposés dans le tableau I. A noter que l'utilisation de cartes auto-organisatrices partitionnées hiérarchiquement a été considérée un temps (voir annexes, section 3.3), mais abandonnée du fait de l'influence du choix de la taille de la carte sur le résultat du partitionnement (et notamment du nombre de partitions) et de la difficulté de relier cette taille à un paramètre du système.

2.5 Assignement des empreintes non utilisées durant l'étape de partitionnement

Pour assigner des identifiants aux empreintes à partir des résultats du partitionnement, nous avons utilisé deux méthodes. La méthode la plus rapide et la plus générale consiste à calculer l'empreinte moyenne de chaque partition, puis d'assigner à chaque empreinte le numéro de partition de l'empreinte moyenne la plus proche. L'autre méthode consiste à assigner à chaque empreinte le numéro de partition de l'empreinte ayant servi au partitionnement la plus proche. Dans les deux cas, les empreintes ayant servi au partitionnement gardent le numéro de partition à laquelle elles sont associées. La deuxième méthode peut également être légèrement modifiée pour la rendre plus cohérente vis-à-vis de l'algorithme de partitionnement spectral de graphe (voir annexes, dernier paragraphe de la section 3.5 p.162).

1. Le partitionnement hiérarchique *simple* n'a pas été retenu du fait de sa susceptibilité au phénomène de "chaînage" : une unique partition qui grossit au fur et à mesures que des points lui sont ajoutés.

Algorithme	Description	Avantages et Inconvénients
Partitionnement hiérarchique	Initialement, chaque empreinte définit son propre groupe. Les groupes sont fusionnés deux à deux pour former un nouveau groupe selon un critère de distance : moyenne (resp. maximum) des distances entre les points de chaque classe pour UPGMA (resp. complet). L'algorithme se termine lorsque tous les groupes ont été fusionnés en un groupe unique regroupant tous les points. L'arbre ainsi défini est coupé à un certain niveau défini à partir de d_{seuil} pour déterminer les partitions.	+ Possibilité de changer d_{seuil} sans avoir à refaire le partitionnement + Possibilité de donner un nombre de partitions voulu - Les partitions ne sont pas découpées dans les zones de faible densité *a priori* - Pour l'algorithme "complet", le seuil de distance choisi automatiquement est moins pertinent du fait de la définition de distance entre classes
DBSCAN	Un graphe de distance est défini en reliant chaque paire d'empreintes situées à moins de d_{seuil} l'une de l'autre. Les points de cœurs sont définis comme les empreintes ayant au moins *minPts* voisins. Les points de cœurs voisins partagent le même numéro de partition. Les points directement liés à un point de cœur héritent de son numéro de partition. Les points non numérotés sont regroupés dans une partition "bruit".	+ Partitionnement dans les zones de faible densité +/- Présence d'une partition "bruit" un peu fourre-tout mais utile pour la détection des cavités fusionnées - Impossibilité de donner un nombre de partitions voulu - Assignement un peu moins cohérent des empreintes non utilisées pour le partitionnement - Nécessité de choisir le paramètre *minPts*
Partitionnement spectral de graphe	Un graphe de distance est défini en reliant chaque paire d'empreintes situées à moins de d_{seuil} l'une de l'autre. La valeur de chaque arête (i, j) vaut $(d_{seuil} - d_{ij})/d_{seuil}$. La matrice de transition du graphe est calculée et diagonalisée. Les empreintes appartiennent au numéro du vecteur propre pour laquelle la projection du vecteur de transition de l'empreinte est maximale.	+ Partitionnement dans les zones de faible densité + Possibilité de donner un nombre de partition voulu + Assignement cohérent des empreintes non utilisées pour le partitionnement - Relativement lent et coûteux en mémoire (calcul des distances + diagonalisation)

Tableau I – Courte description, avantages et inconvénients des algorithmes de partitionnement utilisés dans ce chapitre.

Eyrisch et Helms utilisent un partitionnement hiérarchique de type complet, avec comme contrainte l'impossibilité pour deux cavités d'une même conformation de faire partie d'une même

partition. Pour tester cette méthode, cette contrainte sera modélisée en donnant la valeur 1 (distance maximale qu'il est possible d'atteindre avec la distance EH) à l'ensemble des distances entre deux empreintes tirées d'une même conformation. L'utilisation des méthodes d'assignement rend complexe la mise en place de cette contrainte, que ce soit pour l'assignement des empreintes de lobes de cavités fusionnés ou pour celui des empreintes non utilisées lors de l'échantillonnement. En effet, dans ces deux cas les empreintes n'apparaissent pas dans la matrice de distances d'origine, et il est donc nécessaire de prendre en compte le résultat de l'assignement au cas par cas, conformation par conformation, ce qui est très coûteux en temps. L'algorithme d'Eyrisch et Helms ne sera donc testé qu'en absence d'échantillonnage et de découpage des cavités fusionnées.

2.6 Traitement des cavités fusionnées

Parfois, plusieurs cavités fusionnent pour n'en former qu'une de très grande taille. Il peut être souhaitable de diviser ces cavités en plusieurs morceaux afin de retrouver les lobes correspondant aux cavités transverses pertinentes. Idéalement, le découpage devra se faire dans des zones géométriques favorables, les goulots d'étranglement (voir figure II.6).

2.6.1 Détection

La détection des cavités fusionnées peut se faire de plusieurs façons. La méthode la plus simple consiste à diviser systématiquement les cavités dont la taille dépasse un volume seuil donné. L'inconvénient de cette méthode est qu'elle peut diviser de grosses cavités pourtant cohérentes et pertinentes d'un point de vue fonctionnel. L'idéal est d'inférer à partir du résultat de l'algorithme de partitionnement. Une méthode générale consiste à comparer les empreintes moyennes de chaque partition avec une "somme" d'empreintes moyennes (en réalité, la valeur maximale des deux empreintes pour chaque groupe structural ; voir figure II.5.a). Si une somme d'empreinte est plus proche d'une empreinte unique que de l'un de ses composants, alors cette empreinte unique est constituée de deux partitions plus petites qui ont fusionnées (figure II.5.b et c). Il faut donc diviser toutes les cavités de cette partition. Pour DBSCAN, il peut également être intéressant de diviser les empreintes tombant dans la partition "bruit".

2.6.2 Division des cavités au niveau des goulots d'étranglement

Une fois qu'une cavité est annotée comme étant fusionnée, la division se déroule comme suit :

1. On définit une sonde, de taille superieure à la sonde interne utilisée pour la détection (figure II.6.1).
2. Les zones de la cavité dans lesquelles il est possible de placer le centre de la sonde sans qu'elle ne dépasse de la cavité sont calculées (figure II.6.2).
3. S'il n'existe qu'une seule zone connexe, la cavité ne peut pas être découpée. On augmente donc la taille de la sonde et on recommence à l'étape 2.

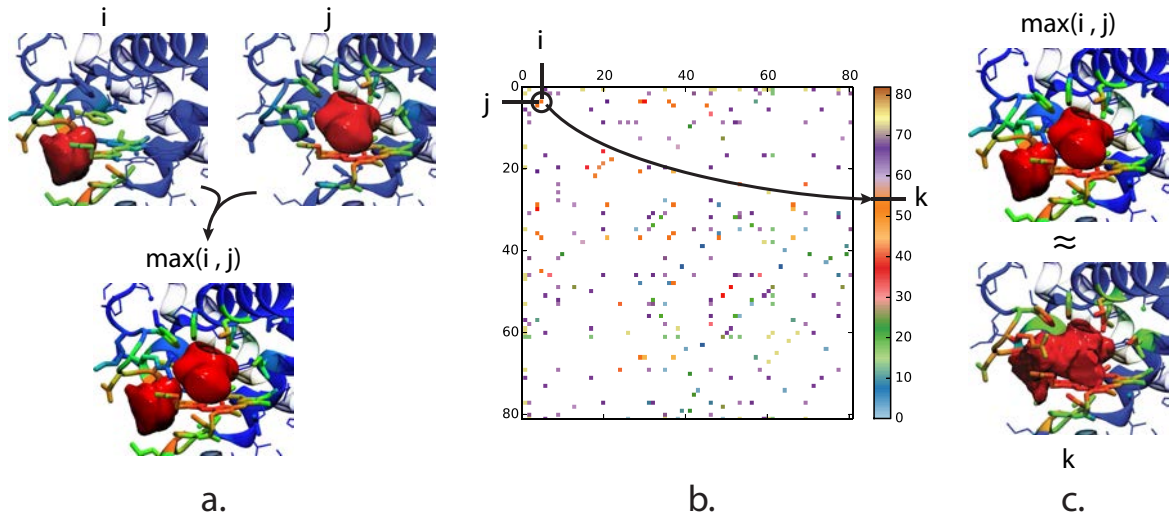


Figure II.5 – Détection des cavités fusionnées. a. Les empreintes moyennes (i et j) peuvent être combinées pour former une empreinte "fusionnée" (max(i, j)). Les empreintes moyennes de i et j sont indiquées par un code couleur sur la structure (bleu : atomes éloignés, rouge : atomes très proches). Les cavités moyennes de i et de j sont représentées sur leurs empreintes respectives et sont superposées sur l'empreinte combinée. **b.** L'ensemble des combinaisons possibles de fusions entre i et j est comparé avec les empreintes moyennes d'origine. Chaque point du graphe (i, j) coloré indique que la combinaison formée par les empreintes moyennes i et j est plus proche d'une empreinte moyenne k que de i ou de j. Le code couleur du point indique le numéro de la partition k la plus proche. **c.** L'empreinte moyenne de k est très proche de la combinaison de i et j. Les cavités de la partition k seront donc découpées, car elles ont été identifiées comme des cavités fusionnées.

4. S'il existe plusieurs zones non connexes, on étend ces zones au reste de la cavité en utilisant l'algorithme de *ligne de partage des eaux* (*watershed* en anglais) pour définir les sous-cavités qui composent la cavité d'origine (figure II.6.3-5).
5. L'empreinte de chaque sous-cavité est calculée et est utilisée pour redéfinir l'affectation de chacune de ces sous-cavités.
6. Si après division et affectation, il reste une sous-cavité pouvant être considérée comme une cavité fusionnée, on recommence l'étape de division sur cette cavité.

Dans tous les cas, on s'arrête si la taille de la sonde dépasse un seuil donné. Dans ce chapitre le rayon d'origine de la sonde est de 1.6 Å, le pas d'augmentation du rayon est de 0.2 Å et le rayon maximal est de 3 Å.

2.7 Mesure de la qualité du suivi des cavités

Elaborer une métrique de qualité du suivi des cavités d'une protéine est difficile entre autres car un bon assignement des cavités peut être subjectif. J'essayerai d'objectiver au maximum la mesure tout en vérifiant graphiquement que ce que l'on obtient correspond bien à nos attentes. Il existe plusieurs possibilités pour mesurer la qualité de l'affectation de l'algorithme de suivi, le plus simple et le moins rigoureux restant la mesure "à l'œil". Cette solution à l'avantage de coller avec une définition intuitive des cavités et reste donc une validation tout à fait pertinente et souvent

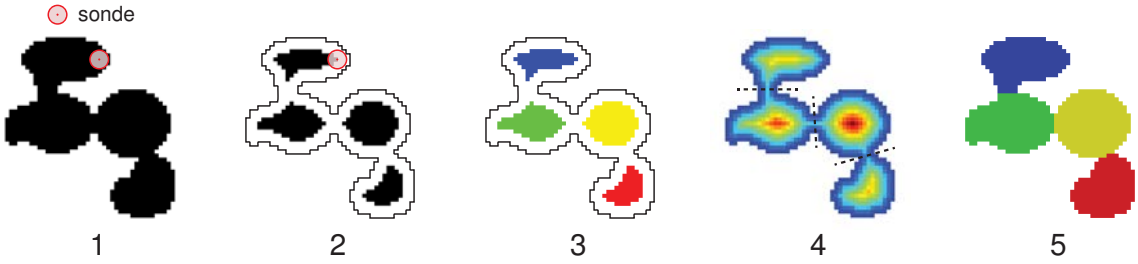


Figure II.6 – Découpage géométrique des cavités fusionnées. L'exemple est artificiel et représenté en 2 dimensions. **1.** Une sonde est choisie de taille plus grande que la sonde de détection. **2.** Les zones accessibles au centre de la sonde sont calculées. **3.** Chaque élément connexe se voit attribué un numéro différent (représenté ici par une couleur). **4.** La distance à l'extérieur (représentée en gradient de couleur du bleu vers le rouge) est utilisée pour définir des goulots d'étranglement. **5.** L'algorithme *watershed* étend chaque zone jusqu'à atteindre un goulot.

incontournable. Il est malheureusement quasiment impossible d'identifier tous les assignements insatisfaisants sur des milliers de conformations.

Pour essayer de s'affranchir des a priori, il est possible de suivre les cavités d'une protéine pour laquelle plusieurs sites ont été bien décrits dans la littérature. On peut alors vérifier que le suivi est cohérent et stable pour les cavités de ce site, en déterminant si les identifiants affectés aux cavités proches d'un ensemble de résidus clés (la poche) a tendance à prendre une valeur unique. Nous suivrons cette approche en suivant les cavités des quatre systèmes et en sélectionnant la cavité dont l'empreinte est la plus proche d'une définition de poche tirée de la littérature :

— pour la myoglobine, les atomes situés à moins de 5 Å d'un des quatre atomes de xénon[114] définissent la poche pour ce site de fixation du xénon (dénotés **Xe1** à **Xe4** selon la nomenclature de Tilton *et al.*)

— pour la protéine E du virus de la dengue, je définis trois poches décrites précédemment dans la littérature et utilisées dans des projets d'identification d'inhibiteurs : le site β-**OG**[181, 30] et les sites dits **1** et **2**[193, 194].

— Deux sites ont été utilisés pour la toxine de l'anthrax (EF) : le site **catalytique**[195] et le site **SABC**, utilisé pour identifier un inhibiteur allostérique de la toxine[112]

— pour ABL1, deux sites ont été utilisés : le site de fixation de l'**imatinib**[196], ligand compétitif de l'ATP, et le site de fixation allostérique de **GNF-2**[197].

Une vue d'ensemble de ces sites est représentée figure II.7.

La cavité la plus proche du centre géométrique des résidus séléctionnés est considérée. Si aucune cavité ne se trouve à moins de 5 Å du centre, on considère qu'il n'y a pas de cavité pour cette conformation. Deux mesures sont alors effectuées pour chaque site :

— la première mesure, $I_{quali,1}$, est calculée comme le nombre d'apparition de la cavité transverse observée le plus souvent, divisé par le nombre total de fois qu'une cavité appraît dans le site. Une mesure $I_{quali,1}$ élevée indique que l'affectation de la cavité du site est univoque, et donc stable.

— la seconde mesure, $I_{quali,div}$, est calculée comme la fréquence pour laquelle la cavité transverse n'est pas divisée (numéro assigné une unique fois pour une même conformation). Une

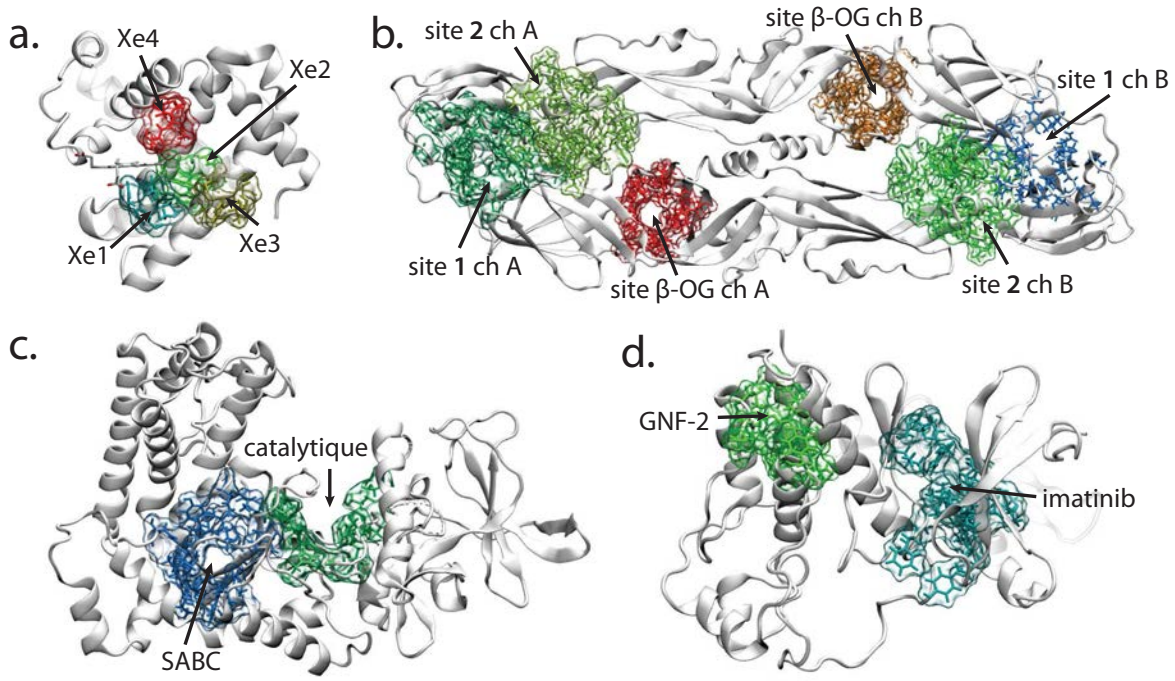


Figure II.7 – Les sites utilisés dans ce chapitre. a. Les sites de la myoglobine : Xe1 à Xe4. **b.** Les sites de la protéine E du virus de la dengue : sites **1**, **2** et β-OG des chaînes A et B. **c.** Les sites de EF : site actif (catalytique) et site allostérique (SABC). **d.** Les sites d'ABL1 : site de fixation de l'imatinib et de GNF-2.

mesure $I_{quali,div}$ basse indique que le site tend à être réparti fréquemment sur deux cavités ou plus, ce qui est insatisfaisant (une cavité transverse coupée en deux la plupart du temps devrait être préférentiellement décomposée en deux cavités transverses).

Je définis également des bornes sur ces mesures pour considérer le suivi de la cavité du site comme une réussite ou un échec. Pour considérer un suivi de cavités comme réussi, on suppose que $I_{quali,1}$ et $I_{quali,div}$ doivent être tous deux supérieurs à 0.75. Ces bornes ont été choisies pour donner une certaine sélectivité tout en qualifiant un nombre suffisant de protocoles de suivi pour pouvoir bien caractériser les différentes approches.

3 Résultats

L'objectif de cette section est de déterminer la meilleure méthode de suivi des cavités au cours d'une dynamique. Pour cela, les mesures décrites section 2.7 seront calculées sur les résultats de suivi des cavités de chacune des quatre dynamiques décrites section 2.1, en faisant varier les paramètres suivants (les notations en italiques seront utilisées dans le reste de ce chapitre par soucis de concision) :

— le groupe structural : *atomes*, chaînes principales/latérales (*CPCL*), *résidus*

— l'échantillonnage des cavités utilisées pour le partitionnement : *1/1* (toutes les cavités pour toutes les conformations) ou *1/10* (les cavités d'une conformation sur dix)

— la distance seuil utilisée : calculée en fonction de la distribution des distances (*auto*) ou fixe (*0.15*, *0.30*, *0.5*, *0.7*, *0.8*, *0.85*, *0.9*)

— l'algorithme de partitionnement : *DBSCAN*, *UPGMA*, Hiérarchique-complet (*complet*), Partitionnement spectral de graphe (*Spectral*), et lorsque c'est possible, Eyrisch et Helms (*EH*)

— la méthode d'assignement : par empreinte moyenne (*moyenne*) ou par cavité la plus proche (*minimum*)

— le découpage des cavités fusionnées : *avec* et *sans*

Cela représente l'analyse de 576 trajectoires de suivi de cavités pour chaque protéine, soit 2304 trajectoires en tout.

3.1 Nombre de cavités instantanées et limitations mémoires

Pour chacune des protéines considérées, le nombre des cavités instantanées sur l'ensemble de la trajectoire dépasse les dizaines de milliers (tableau II, colonne 4). Ce nombre important est problématique lors de l'étape de partitionnement, puisque celle-ci nécessite de calculer la matrice de distance pair-à-pair, de taille $n_{inst}(n_{inst}-1)/2$ et qui peut prendre énormément de place en mémoire (tableau II, colonne 5). Le partitionnement de telles matrices peut ainsi prendre plusieurs heures voire plusieurs jours (à supposer que la mémoire vive disponible soit suffisante pour mener à bien le calcul). Cela rend donc relativement peu pratique le calcul de cavité sans échantillonnage pour de grosses protéines. Au vu du grand nombre d'analyses prévues pour chaque trajectoire (voir section précédente), je me limiterai à l'étude de la myoglobine et d'ABL1 pour l'échantillonnage *1/1*.

Protéine	# atomes	# résidus	# cavités instannées	mémoire utilisée pour la matrice de distance (Go)
Myoglobine	2534	164	11 863	0.52
Protéine E (*dengue*)	12258	788	58 376	12.7
EF	9942	465	72 507	19.6
ABL1	4326	268	27 308	2.77

Tableau II – Nombre de résidus, d'atomes et de cavités instantanées pour chacune des trajectoires utilisées dans ce chapitre

3.2 Comparaison des algorithmes de partitionnement sans découpage des cavités fusionnées

L'objectif de cette section est de déterminer les paramètres optimaux (seuil de distance et méthode d'assignement) pour chacun des algorithmes testés, en fonction de la protéine considérée.

Pour cela, les mesures présentées section 2.7 ont été réalisées sur l'ensemble des trajectoires de suivi des cavités en faisant varier chacun des paramètres individuellement. On ne considèrera dans cette section que les suivis de cavités sans découpage des cavités fusionnées.

En tout, sur les 7394 suivis des sites sélectionnés, 1393 valident les critères de stabilité définis en section 2.7 ($I_{quali,1} > 0.75$ et $I_{quali,div} > 0.75$), soit 18.9%. Ce pourcentage assez bas démontre la nécessité de choisir les bons algorithmes et paramètres associés afin de réaliser un suivi convenable. Le taux de réussite des suivis validant les critères de stabilité en fonction des sites étudiés et des algorithmes utilisés est indiqué figure II.8. On remarque tout d'abord que les sites étudiés dans ce chapitre sont plus ou moins faciles à suivre. Ainsi, les sites Xe2 et Xe4, le site catalytique d'EF et le site GNF-2 d'ABL1 paraissent particulièrement difficiles à suivre. On note, dans la figure III.13 du chapitre suivant, que les poches Xe2 et Xe4 sont en fait relativement ambigues. De même, il a déjà été relevé que la poche catalytique d'EF a une forme très particulière, en forme de crevasse incurvée et des voies d'entrée multiples[198].

Un autre point notable est l'efficacité de l'algorithme de partitionnement hiérarchique complet, qui semble moins dépendre des paramètres que les autres algorithmes et qui produit ainsi plus souvent des suivis corrects. DBSCAN semble peu adapté au problème de suivi des cavités : seules 6 cavités ont été correctement suivies, provenant d'une combinaison très restreinte de paramètres. Enfin, chaque algorithme possède ses propres forces et faiblesses selon les sites suivis. On peut ainsi noter que le seul algorithme capable de suivre Xe4 correctement est le partitionnement spectral de graphe et non le partitionnement hiérarchique complet, pourtant plus constant pour les autres sites.

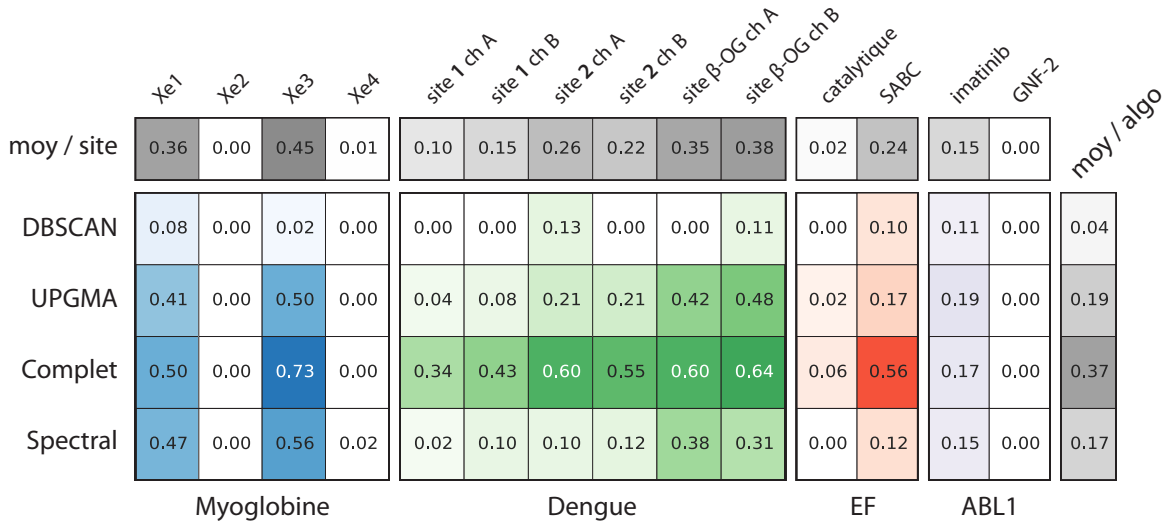


FIGURE II.8 – Fraction du nombre de combinaison de paramètres produisant des suivis de cavités "réussis" (stables) sur le nombre de combinaisons testées pour chaque site sélectionné et chaque algorithme. La ligne du haut (dégradé de gris) correspond à la moyenne des valeurs de l'ensemble des algorithmes pour chacun des sites étudiés. La dernière colonne à droite (en dégradé de gris) correspond à la moyenne des valeurs de l'ensemble des sites étudiés pour chacun des algorithmes testés.

La figure II.9 permet d'avoir une vision plus détaillée des paramètres efficaces pour chaque algorithme de partitionnement. Sans surprise, la valeur du seuil de distance a un grand impact

sur le résultat du suivi des cavités. DBSCAN est plus efficace pour de petites valeurs de seuil, mais n'atteint tout de même pas l'efficacité des autres algorithmes de partitionnement. Ceci peut s'expliquer par le fait que DBSCAN fusionne les partitions à partir du moment où la distance entre deux empreintes est inférieure à la distance seuil, ce qui n'est pas le cas des autres algorithmes. Le seuil de distance automatique semble inadapté pour la plupart des cas, il est donc préférable de favoriser des seuils plus grands, notamment pour les algorithmes de partitionnement hiérarchique. Une autre conclusion est la supériorité de la méthode d'assignement *minimum* par rapport à la méthode *moyenne*.

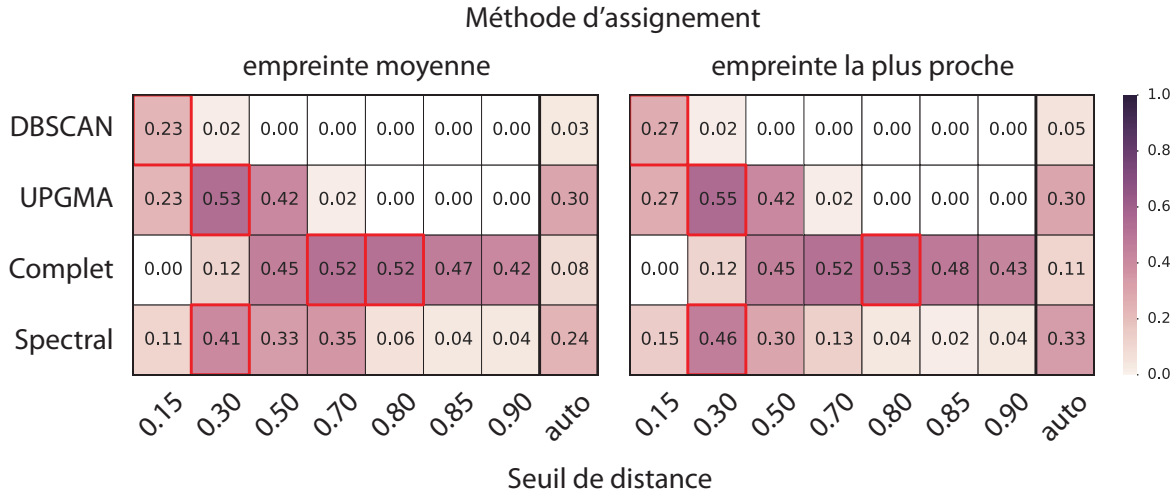


Figure II.9 – Fraction du nombre de combinaisons de paramètres produisant des suivis de cavités "réussis" (stables) sur le nombre de combinaisons testées pour chacun des algorithmes, méthodes d'assignement et seuil de distance utilisés. Les rectangles rouges indiquent la valeur maximale pour un algorithme et une méthode d'assignement donnés.

A partir de cette étude, on peut conclure que les paramètres de partitionnement adaptés pour chaque algorithme sont :

— pour le partitionnement hiérarchique complet, un seuil élevé autour de 0.80 et l'assignement par empreinte la plus proche

— pour UPGMA, un seuil moyen autour de 0.30 et l'assignement par empreinte la plus proche

— pour le partitionnement spectral de graphe, un seuil moyen autour de 0.30 et l'assignement par empreinte la plus proche

— pour DBSCAN, un seuil faible (< 0.15) et l'assignement par empreinte la plus proche

3.3 Comparaison avec la méthode d'Eyrisch et Helms

Comme indiqué section 2.5, la méthode d'Eyrisch et Helms est difficilement applicable lors de l'utilisation des méthodes d'assignement. Cela limite *de facto* la comparaison aux cas ne faisant pas appel à un échantillonnage ni au découpage des cavités fusionnées, donc à l'analyse de la myoglobine et d'ABL1 avec un échantillonnage *1/1*. A noter que pour l'échantillonnage *1/1*

d'ABL1, le partitionnement spectral de graphe n'a pas pu aboutir (problème mémoire lors de la diagonalisation). Il n'y a donc pas de résultat pour cet algorithme.

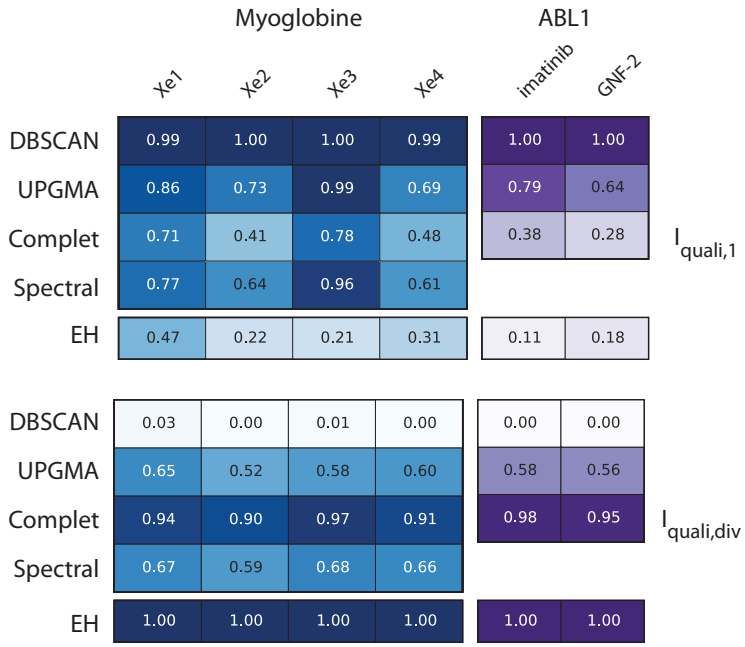


Figure II.10 – Comparaison des valeurs moyennes de $I_{quali,1}$ et $I_{quali,div}$ de chacun des algorithmes de partitionnement et de la méthode d'Eyrisch et Helms (EH) pour chaque site de la myoglobine et d'ABL1. En haut : valeurs moyennes de $I_{quali,1}$, en bas : valeurs moyennes de $I_{quali,div}$.

Sur les 18 suivis des sites séléctionnés avec la méthode d'Eyrisch et Helms (6 sites, 3 groupes structuraux), aucun n'a débouché sur un suivi considéré comme réussi. En cause, la contrainte interdisant aux cavités d'une même structure d'appartenir à la même partition à des conséquences extrêmement négatives sur l'indice $I_{quali,1}$, dont les valeurs vont de 0.1 à 0.59 (figure II.10). Cette contrainte implique aussi mécaniquement que les suivis réalisés par la méthode d'Eyrisch et Helms ont un indice $I_{quali,div}$ parfait (égal à 1), puisqu'aucune division de cavité n'a lieu.

3.4 Effet de l'échantillonnage

Comme vu dans la section 3.1, échantillonner les cavités avant de réaliser le partitionnement peut s'avérer nécessaire pour des questions de mémoire et de temps de calcul. Intuitivement, l'échantillonnage devrait dégrader la qualité du suivi des cavités, puisque le partitionnement ne prend pas en compte une grande partie des cavités et peut donc passer à côté de certaines cavités transverses peu fréquentes. Pour mesurer cette éventuelle dégradation, j'ai comparé l'évolution de la fraction de suivis réussis pour chaque site lors du passage de l'échantillonnage *1/1* à l'échantillonnage *1/10* (figure II.11).

On remarque que l'effet de la réduction par 10 du nombre de cavités traitées n'a que rarement des effets négatifs sur le suivi (suivi de Xe1 avec UPGMA et hiérarchique complet), voire même améliore les chances de réaliser un suivi pertinent (DBSCAN, partitionnement spectral de graphe,

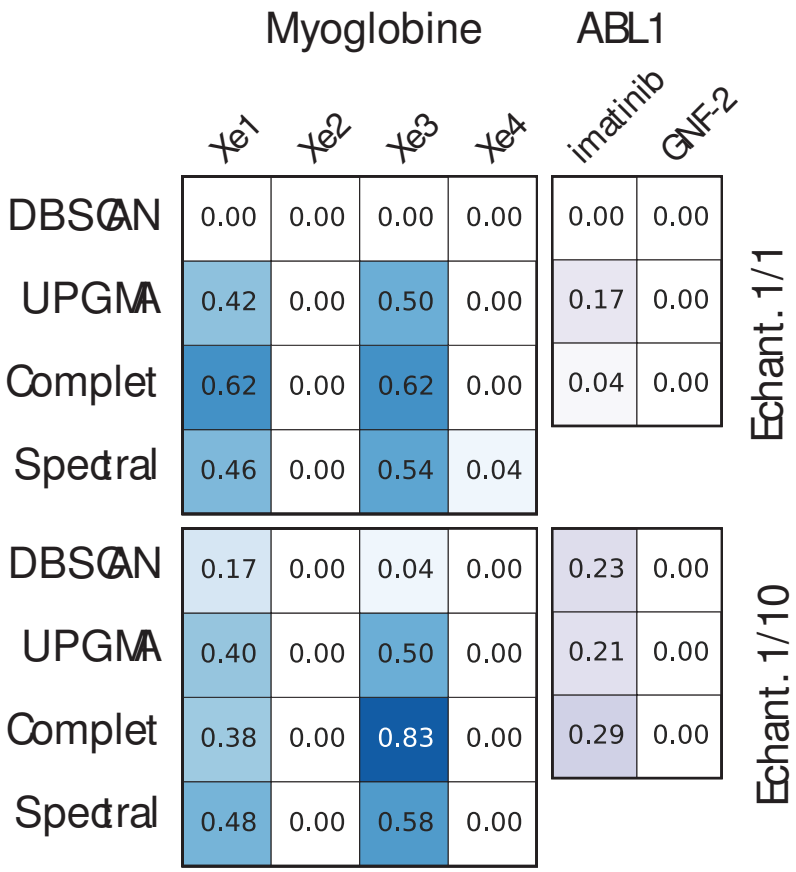


Figure II.11 – Fraction du nombre de combinaisons de paramètres produisant des suivis de cavités "réussis" (stables) sur le nombre de combinaisons testées en fonction de l'échantillonnage pour chacun des algorithmes utilisés et chacun des sites étudiés. En haut : échantillonnage *1/1*, en bas : échantillonnage *1/10*.

Xe3 avec hiérarchique complet). Il semble donc particulièrement intéressant de limiter le nombre de cavités à traiter, non seulement pour limiter l'utilisation de la mémoire et le temps de calcul, mais bien également pour limiter le nombre de cavités "intermédiaires" pouvant faire le lien entre deux partitions.

3.5 Comparaison des résultats pour chacun des groupes structuraux

La motivation première derrière la définition de nouveaux groupes structuraux en plus des groupes type *atomes* provient surtout du gain de temps lors du calcul des empreintes et des distances entre empreintes. On peut supposer que l'utilisation de groupes moins précis tels que *CPCL* et *résidus* devrait engendrer des suivis de cavités moins satisfaisants. La figure II.12 indique la fraction de suivis réussis en fonction des groupes structuraux utilisés. De façon surprenante, l'utilisation de groupes *atomes* ne permet pas forcément d'obtenir des suivis plus efficaces. En effet, pour la protéine E du virus de la dengue et ABL1, un certain nombre de sites sont plus

efficacement suivis en utilisant les groupes *CPCL* ou *résidus*. Il est donc pour ce type de cas doublement intéressant d'utiliser des groupes plus généraux. A noter tout de même que pour la myoglobine, il semble toujours plus intéressant d'utiliser un groupement par atomes. Cela peut s'expliquer par la distance très courte entre chacune des cavités Xenon qui peuvent donc être bordées par des atomes différents de mêmes résidus.

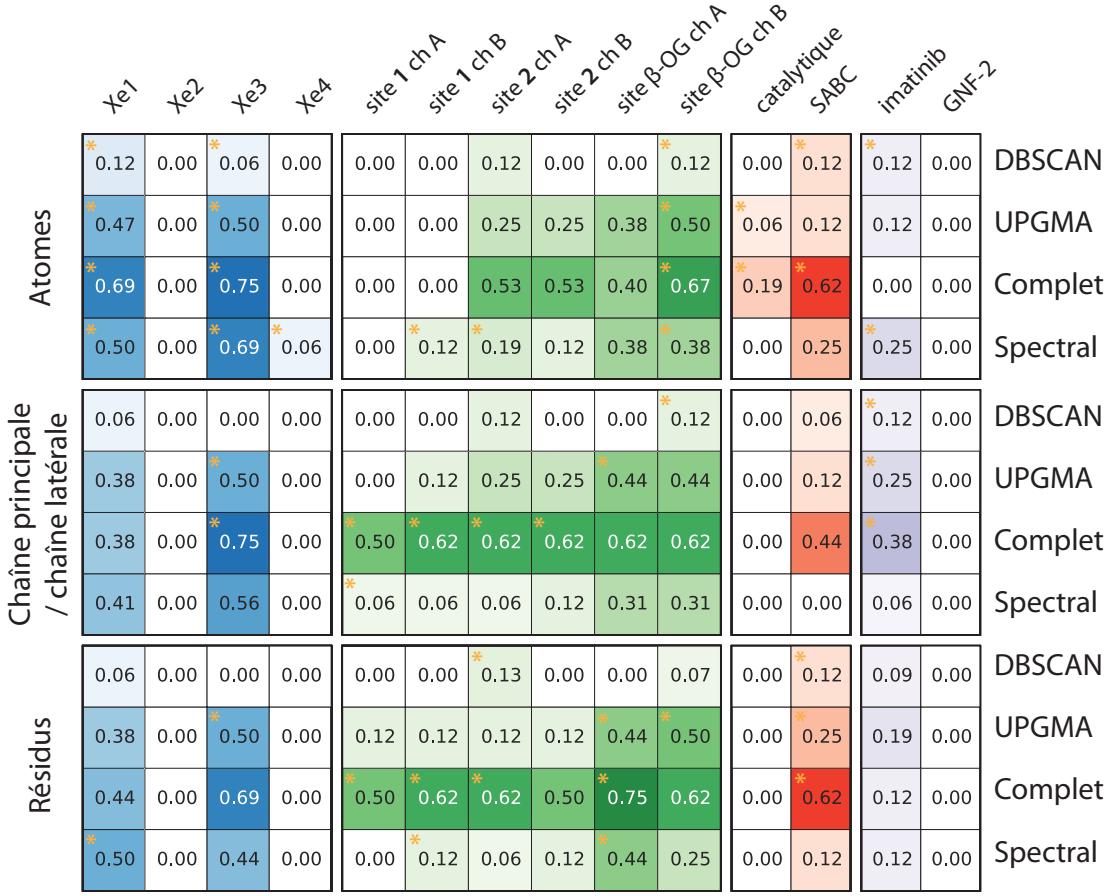


FIGURE II.12 – Fraction du nombre de combinaisons de paramètres produisant des suivis de cavités "réussis" (stables) sur le nombre de combinaisons testées en fonction du groupe structural choisi pour chacun des algorithmes utilisés et chacun des sites étudiés. De haut en bas : groupes *atomes*, groupes chaîne principale/chaîne latérale (*CPCL*), groupes *résidus*. Les valeurs les plus hautes pour chaque paire algorithme-site sont annotées par des étoiles orangées.

3.6 Effet du découpage des cavités fusionnées

Le découpage des cavités fusionnées a été mis en place afin d'améliorer le suivi pour des cas un peu complexes. L'étape étant optionnelle, l'efficacité du découpage doit être mesurée dans le meilleur des cas sur l'ensemble des suivis observés. La figure II.13 indique la plus grande évolution relative de l'indice $I_{quali,1}$ lorsque l'on applique le découpage des cavités, pour chaque algorithme et chaque site étudiés. L'effet du découpage semble relativement modéré pour la plupart des sites et des algorithmes. Le découpage peut toutefois être très intéressant pour certains sites en

particulier, comme Xe1, Xe3, les sites **1** et **2** de la protéine E, et le site de fixation de l'imatinib. On note également que l'effet du découpage peut être très positif lors de l'utilisation des algorithmes *complet* et *spectral*.

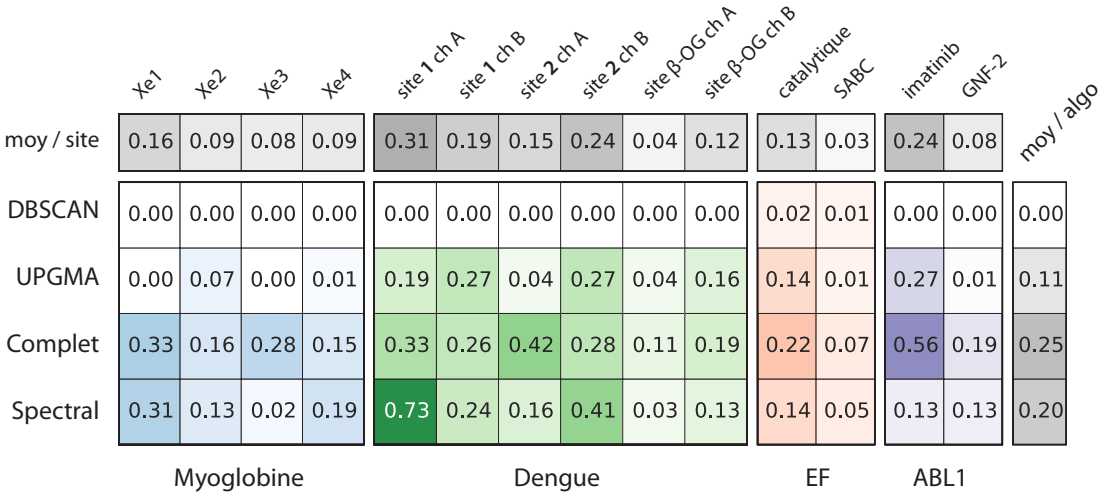


Figure II.13 – Ecart maximum relatif (pourcentage d'augmentation) observé de l'indice $I_{quali,1}$ pour chacun des algorithmes de partitionnement et chacun des sites étudiés. La ligne du haut et la colonne la plus à droite correspondent respectivement à la moyenne des valeurs du tableau pour chaque site et pour chaque algorithme de partitionnement.

3.7 Combinaisons optimales de paramètres

Le tableau III donne pour chaque algorithme les paramètres permettant de suivre de façon pertinente le plus de sites possibles (le paramètre d'échantillonnage n'est pas indiqué car les suivis avec un échantillonnage *1/1* n'ont pas tous été produits). Les algorithmes de partitionnement hiérarchiques (complet et UPGMA) permettent de suivre le plus de sites à la fois. A noter la grande variété des paramètres, notamment pour les groupes.

La figure II.14 donne les valeurs moyennes de $I_{quali,1}$ et $I_{quali,div}$ sur l'ensemble des sites pour chacun de ces paramètres optimaux. Il apparait nettement sur ce graphe que l'algorithme UPGMA donne les meillers valeurs pour chacun des deux indices.

4 Discussion

Dans ce chapitre, j'ai posé la question du suivi des cavités au cours d'une dynamique, établi des bases pour l'aborder et proposé une méthode général pour sa résolution. Un grand nombre de paramètres portant sur les différentes étapes de la méthode ont été testés sur quatre protéines. Quatorze sites précédemment décrits dans la littérature ont été utilisés pour mesurer l'influence de ces paramètres sur la qualité du suivi des cavités à l'aide de deux critères de qualité, $I_{quali,1}$ et $I_{quali,div}$. Lorsque les paramètres sont bien choisis, cette méthode permet de suivre efficacement

ID	Partitionnement	groupes	assign.	seuil	découpage	# sites correct. suivis
1	DBSCAN	atomes	moy.	0.15	oui	6
2		-	-	-	non	-
3		-	min.	-	oui	-
4		-	-	-	non	-
5	UPGMA	résidus	moy.	0.30	oui	10
6		-	-	-	non	-
7		-	min.	-	-	-
8	Complet	CPCL	min.	0.50	oui	10
9	Spectral	atomes	moy.	0.70	oui	8
10		-	-	-	non	-

Tableau III – Paramètres permettant de suivre le maximum de sites correctement pour chaque algorithme.

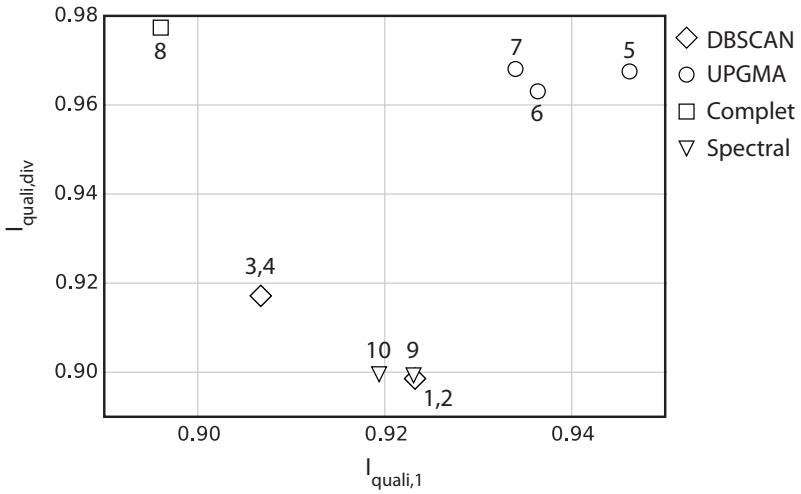


Figure II.14 – Valeurs de $I_{quali,1}$ (abscisse) et $I_{quali,div}$ (ordonnée) pour chacun des 10 jeux de paramètres optimaux. Les numéros correspondent à ceux de la colonne ID du tableau III.

la plupart de ces sites. A partir de ces résultats, j'ai pu déterminer la liste de ces paramètres pour chacun des algorithmes. Ce travail ouvre la voie à l'étude dynamique systématique de l'ensemble des cavités d'une protéine. L'analyse individuelle des propriétés de chacune des cavités transverses devrait permettre d'en dégager les caractéristiques ce qui permet d'exploiter en parallèle les cavités les plus prometteuses pour la recherche de molécules effectrices. Le suivi des cavités est donc un outil précieux pour l'étude fonctionnelle des protéines ainsi que pour l'établissement de stratégies novatrices de criblages virtuels, notamment dans le cadre de projets de conception de médicaments. J'ai ainsi pu l'utiliser dans deux projets d'identification d'inhibiteur, ciblant l'ADN-gyrase de *M. tuberculosis* et la subtilisine 1 de *P. vivax* et développés au chapitre IV.

4.1 Revue des paramètres et guide du suivi des cavités

D'après la figure II.8 et le tableau III, les algorithmes de partitionnement hiérarchique donnent les résultats les plus constants. Il faut toutefois noter que d'autres méthodes peuvent être plus adaptées au suivi de certains sites problématiques. Le partitionnement spectral de graphe notamment est le seul algorithme ayant permis de suivre correctement le site Xe4 de la myoglobine. Cet algorithme tend à réaliser ses meilleures performances en utilisant le groupement par *atomes*, tandis que UPGMA donne de meilleurs résultats avec le groupement par *résidus*. Cela explique peut-être ce résultat, le groupement par *atomes* permettant une analyse plus détaillée du réseau intriqué des cavités de la myoglobine. Le suivi utilisant DBSCAN ainsi que la méthode d'Eyrisch et Helms produisent des résultats rarement convainquants et diamétralement opposés dans leurs défauts : DBSCAN a tendance à regrouper beaucoup de cavités dans une même partition, tandis que la méthode d'Eyrisch et Helms assigne plusieurs numéros de cavités transverses à chaque site. A ce titre, je recommande d'éviter l'utilisation de la méthode d'Eyrisch et Helms à moins d'avoir un besoin impérieux de ne pas assigner le même numéro de cavité transverses à deux cavités de la même conformation.

La valeur de la distance seuil optimale dépend des algorithmes de partitionnement utilisés dans ce chapitre (figure II.9 et tableau III). La distance seuil automatique décrite section 2.4.1 n'est finalement jamais la distance optimale pour réaliser le suivi. Il serait utile de pouvoir développer une méthode automatique adaptée à chaque algorithme, car les seuils fixes ont dû être choisis arbitrairement et ne s'adaptent donc pas aux données d'entrée. On pourrait pas exemple tenter de comparer la distribution de l'ensemble des distances avec celle de l'ensemble des distances entre cavités de la même conformation pour guider le choix d'une autre distance seuil automatique plus adaptée.

Les notions de groupes structuraux et d'échantillonnage des conformations, introduites en premier lieu pour diminuer le temps de calcul et l'empreinte mémoire des empreintes, ont finalement un effet plus subtil qu'imaginé (figure II.12 et figure II.11). On retrouve le groupement par résidu trois fois dans la liste des paramètres optimaux, et le groupement CPCL une fois. Réduire la complexité de la représentation des empreintes semble ainsi bénéfique dans de nombreux cas. Cependant, le cas des cavités intriquées de la myoglobine et notamment de la poche Xe4 montre qu'une description plus fine pourrait dans certains cas être nécessaire.

Les deux méthodes d'assignement explorés dans ce chapitre donnent des résultats satisfaisants. La méthode d'assignement utilisant l'empreinte la plus proche (*minimum*) donne des résultats légérement meilleurs en moyenne que la méthode utilisant l'empreinte moyenne la plus proche (*moyen*), comme le montre la figure II.9. Toutefois, la présence majoritaire de l'assignement *moyen* dans l'ensemble des paramètres optimaux (tableau III) ainsi que le gain en temps et en mémoire qu'il apporte pousse à favoriser son utilisation.

Le découpage des cavités impacte en moyenne assez peu la qualité du suivi (données non communiquées). Cette étape peut cependant être très utile pour améliorer sensiblement le suivi dans certains cas difficiles, comme démontré section 3.6. Toutefois, le coût en mémoire et en temps

de calcul (proportionnel au nombre des cavités) de l'étape de découpage rend son utilisation plus difficile, et je conseille donc de ne l'utiliser que si les cavités produites paraissent le demander.

Au final, lors de l'étude de la dynamique des cavités d'une nouvelle protéine, je conseille l'utilisation d'UPGMA sur les cavités extraites d'une fraction des conformations (ici une centaine), en utilisant un seuil de distance de 0.3, le groupement structural par *résidus*, l'assignement des empreintes non partitionnées vers le numéro de l'empreinte moyenne la plus proche (assignement *moyen*) et éventuellement le découpage des cavités s'il subsiste des cas difficiles.

4.2 Les limites de l'étude et de la méthode

Bien que l'étude du suivi des cavités présentée dans ce chapitre traite d'un grand nombre de paramètres, il subsiste des zones améliorables pouvant expliquer certains problèmes, ce qui est susceptible d'altérer les conclusions finales.

Une faiblesse de cette étude est le traitement des distances seuils. En effet, DBSCAN semble demander des distances seuils très basses, qui n'ont pas été échantillonnées aussi finement que les seuils plus élevés, finalement peu utiles pour cet algorithme. Il est possible que ce défaut d'échantillonnage des seuils bas soit la cause des résultats très médiocres des suivis réalisés à l'aide de DBSCAN. De manière générale, il devrait être intéressant d'élargir l'étude à un échantillonnage plus dense des distances seuils, éventuellement en "zoomant" sur les zones plus intéressantes de chaque algorithme (< 0.15 pour DBSCAN, $0.2 - 0.5$ pour UPGMA et spectral, $0.5 - 0.85$ pour complet).

Une autre zone d'amélioration de cette étude provient du fait qu'elle repose sur la définition provenant de la littérature d'une série de sites. Il n'est pas garanti que ces sites soient pleinement représentatifs de l'ensemble des sites intéressants de chacune de ces protéines, ou même des sites intéressants des protéines en règle générale. L'absence de suivi correct du site Xe2 de la myoglobine et du site de fixation de GNF-2 d'ABL1 peuvent également provenir d'une faiblesse dans la méthode d'établissement de la cavité de référence. En effet, cette sélection est basée conjointement sur la définition des résidus bordant le site, pouvant être sujette à caution notamment dans le cas du site de fixation de GNF-2, et d'un critère de distance consensuel mais arbitraire (5 Å) pour définir cette cavité de référence. Les cavités extraites par cette méthode peuvent donc être biaisées, ce qui fausse les résultats. D'autre part, les seuils arbitraires appliqués à $I_{quali,1}$ et $I_{quali,div}$, bien que raisonnables, peuvent être discutés. Leur utilisation vient principalement de la difficulté de traiter des données à partir d'un couple d'indices. Par exemple, l'utilisation de DBSCAN conjointement à un seuil de distance élevé donne des valeurs de $I_{quali,1}$ parfaites mais un indice de qualité $I_{quali,div}$ nul. Il est donc difficile de priorétiser ces indices. La création de mesures quantitatives "objectives" est une approche a priori prometteuse que j'ai initialement tenté d'entreprendre. Malheureusement, il est complexe de définir de telles mesures qui soient à la fois pertinentes, facilement interprétables et faciles à manipuler. A ce titre, l'utilisation dans cette étude des indices $I_{quali,1}$ et $I_{quali,div}$ et de leurs seuils associés, bien que non idéale, a le mérite d'être facilement interprétable (succès/échec du suivi) et aisément manipulables (fraction

de suivis réussis).

Enfin, la méthode elle-même possède ses limites intrinsèques. La plus importante est particulièrement bien retranscrite par l'équilibre subtil entre les indices $I_{quali,1}$ et $I_{quali,div}$ et la nécessité de définir des seuils arbitraires sur ces deux indices pour définir le "succès" du suivi des cavités. La méthode jongle en effet en permanence entre deux forces : le partitionnement des cavités similaires et l'identification de plusieurs cavités instantanées d'une même conformation à une même cavité transverse. Cet équilibre est arbitré principalement par la valeur de distance seuil, dont la valeur optimale dépend fortement de l'algorithme de partitionnement utilisé. En outre, les cavités ne peuvent pas toutes être traitées à l'identique et certaines cavités fonctionnelles intéressantes peuvent passer au travers des mailles du filet du suivi des cavités. C'est le cas notamment des cavités des sites Xe2 et Xe4 de la myoglobine, du site catalytique d'EF et du site de fixation de GNF-2 d'ABL1, qui sont particulièrement complexes à suivre. Le cas des sites Xe4 de la myoglobine et du site catalytique d'EF sont d'ailleurs particulièrement intéressants, puisque les algorithmes les plus robustes (partitionnement hiérarchiques avec les paramètres optimaux) n'arrivent pas à les suivre, tandis que d'autres algorithmes (*spectral*) et d'autres combinaisons de paramètres peuvent être utilisés pour les suivre. La diversité des cas de figure et le caractère intrinséquement subjectif de ce qu'est une bonne cavité transverse sont donc deux limites fortes à l'obtention d'une méthode de suivi des cavités "parfaite".

Malgré ces limitations, l'existence de combinaisons de paramètres permettant le suivi de 10 des 14 sites étudiés semble indiquer que cette méthode est relativement robuste, et qu'elle peut donc être utilisée dans la majorité des études dynamiques de cavités de protéines. J'estime donc que malgré ses limites, cette méthode est arrivée à un bon degré de maturité. Elle devrait pouvoir être exploitée dans de nombreux cas requiérant la détection et l'analyse en parallèle de la dynamique d'une multitude de cavités.

4.3 Perspectives

Comme vu dans la section précédente, l'étude présentée dans ce manuscrit présente quelques limites qu'il sera bon de corriger afin de renforcer le socle scientifique de l'utilisation de la méthode de suivi des cavités et de ses paramètres. Une étude plus poussée du choix de la distance seuil et la définition d'une distance seuil automatique plus adaptée sont deux des axes les plus importants à développer afin d'identifier les paramètres réellement optimaux de cette méthode. La définition de mesures quantitatives et objectives du suivi des cavités est un point crucial et l'affinement des critères permettrait également de renforcer les conclusions de l'étude. Ce dernier point reste toutefois délicat. L'application de la méthode de suivi des cavités dans plusieurs projets impliquant l'étude systématique et dynamique des cavités devrait également permettre d'identifier les points faibles et points forts de chacun des paramètres sur des cas pratiques. Cette méthode est en effet à un point de maturité suffisant pour envisager son utilisation massive, seule façon d'identifier empiriquement les combinaisons les plus efficaces en général et en fonction des problèmes rencontrés, ainsi que leur impact dans les différentes applications envisageables, comme la conception

d'effecteurs ou l'analyse de la dynamique fonctionnelle des protéines.

Chapitre III

Utilisation de l'ACP sur les cavités de protéines

1 La dynamique de la géométrie des cavités, un aspect encore peu exploité

Comme vu précédemment au chapitre I, les cavités ont été principalement utilisées pour expliquer des processus biologiques et dans le cadre d'applications type arrimage moléculaire ou criblage virtuel. Dans les deux cas, malgré l'importance des cavités sur la dynamique des systèmes, ce sont principalement des données statiques et/ou qualitatives qui sont utilisées. La plupart des études portant sur la fonction des cavités utilisent une seule ou un petit nombre de structures cristallographiques. Une exception majeure est la myoglobine qui a fait l'objet de nombreuses études dynamiques sur ses cavités internes[118, 119, 121, 122], l'objectif étant d'expliquer la dynamique de ses ligands (dioxygène et monoxyde de carbone notamment). De même, dans le cadre de projets de développements de médicament, la plupart des logiciels de criblage virtuel ou d'arrimage moléculaire utilisent des structures uniques. Les méthodes prenant en compte la flexibilité du site visé uilisent souvent une multitude de structures mais ne considèrent pas la géométrie de la cavité (méthodes de docking sur un ensemble de structures[199], docking avec bibliothèques de rotamères[48], moyenne des énergies d'interaction[200]). Ceux qui le font de manière directe appliquent généralement des contraintes ne tenant pas compte les données dynamiques, comme SCARE[72] ou les procédés dits de fumigation[73] et de pressurisation[74].

La plupart des méthodes de détection des cavités ne sont pas adaptées à une étude dynamique et les logiciels spécialisés dans ce type d'études sont relativement récents (voir chapitre I). Une partie de ces logiciels ne font que détecter les cavités[167, 169, 170, 171, 172]. De plus, la plupart d'entre eux n'utilisent pas la géométrie complète de chaque cavité, mais se focalisent sur des descripteurs plus simples, comme le volume, la surface ou la position du centre géométrique[167, 168, 169, 170, 171, 172]. Il parait donc intéressant et relativement novateur de pouvoir analyser en détail l'évolution de la géométrie des cavités, d'abord pour mieux en connaître les propriétés, mais également pour guider la sélection, l'échantillonnage ou la création de structures pour les

projets de conception de médicaments impliquant des étapes de criblage virtuel.

L'ACP est une méthode simple permettant de visualiser les directions d'évolution d'un système, mais aussi de compresser efficacement des données avec une perte minimale de variance. L'application de l'ACP sur les cavités permet ainsi de bénéficier de ces deux avantages. La détermination de directions d'évolutions de plus grande variance permet de repérer les cavités ou les lobes dont les apparitions sont corrélées. Il est ensuite possible d'utiliser ces informations pour corréler ces variations avec la fonction d'une protéine. Cela peut être précieux pour la conception de médicaments. Nous avons choisi une représentation des cavités sous forme de grille booléenne, il est donc nécessaire d'étudier le comportement de l'ACP sur de tels objets afin de définir un cadre d'utilisation cohérent. A noter que l'ACP sur des cavités représentées par des grilles a déjà été proposée indépendamment par Craig *et al.* avec PocketAnalyzer$^{\mathrm{PCA}}$[201]. Notre formalisation basée sur l'espace des pas de temps nous a permis une analyse poussée du résultat de l'ACP, ainsi que l'étude des corrélations entre mouvements des cavités et des atomes qui n'avait pas été entreprise dans cet article. Nous nous proposons donc de présenter nos apports dans ce chapitre.

2 L'ACP sur les cavités : définition et outils d'analyse

2.1 Principe de l'Analyse en Composantes Principales et application aux cavités

Par définition, l'Analyse en Composantes Principales (ACP, ou PCA en anglais) consiste à calculer puis à diagonaliser la matrice de covariance, V, d'une liste de descripteurs, D. Les descripteurs correspondants aux coordonnées atomiques, D^{atomes}, sont simplement les coordonnées de chaque atome (soit $3 * n^{atomes}$) pour chacune des s conformations de la trajectoire. D^{atomes} est donc une matrice de taille $3n^{atomes} \times s$. Pour les descripteurs de géométrie des cavités, j'introduis la matrice D^{cav}, composée de l'état de chaque point de grille pour chacune des s conformations. Pour simplifier les calculs on ne considère que les points de grilles apparaissant au moins une fois dans la trajectoire (soit n^{cav} point de grilles, le *domaine de définition*). L'état d'un point de grille vaut 1 si le point se trouve dans une cavité dans cette structure, et 0 sinon. La matrice D^{cav} est donc de taille $n^{cav} \times s$. L'ACP utilise la matrice des descripteurs centrés, M, définie pour chaque conformation $i \in [1, n^{cav}]$ comme $M_i = D_i - C$, avec $C = \langle D_i \rangle$ le vecteur de coordonnées moyennes. Il faut noter que cette étape de recentrage fait passer la description des cavités d'une représentation booléenne à une représentation réelle, entre 0 et 1. La matrice de covariance V est ensuite calculée classiquement dans l'*espace des coordonnées* : $V_{desc} = 1/s \cdot M \cdot M^T$. La diagonalisation de V produit une liste de vecteurs propres v_i, qui définissent les directions d'évolutions des composantes principales CP_i, et leurs valeurs propres associées λ_i (triées de telle sorte que $\lambda_1 > ... > \lambda_n \geq 0$), qui donnent à la fois la variance expliquée par chaque v_i et l'amplitude de

CP_i. Les composantes principales résultantes appartiennent à l'espace des descripteurs et peuvent donc être visualisées directement.

L'ACP peut aussi être réalisée dans l'*espace des pas de temps* : $V_{step} = 1/s \cdot M^T \cdot M$, de taille $s \times s$. De façon très opportune, bien que V_{desc} et V_{step} n'aient pas la même taille, leurs valeurs propres non nulles sont identiques, et leurs vecteurs propres v_j sont aisément reliés (voir en annexe, section 4.1 pour les preuves) :

$$v_{desc,j} = \sqrt{s\lambda_j} \cdot v_{step,j} \cdot M \tag{III.1}$$

Ainsi, les composantes principales d'un espace peuvent être calculées rapidement à partir des composantes principales de l'autre. Pour des questions de performance, il est généralement préférable de réaliser les calculs dans l'espace des pas de temps, puisque $s < n^{cav}$ dans la plupart des cas. Le calcul dans cet espace permet aussi de comparer les composantes principales des structures et des cavités, les matrices ayant les mêmes dimensions et faisant références aux mêmes entités : les différents pas de la trajectoire.

2.2 Reconstructions de cavités et de structures via les composantes principales

Il est assez courant, lorsque l'on réalise une ACP sur des coordonnées atomiques, de reconstruire une trajectoire variant autour de la structure moyenne le long d'une composante principale i :

$$T_i(\mu) = C + \mu * v_{desc,i} \tag{III.2}$$

Cette technique permet de visualiser simplement l'évolution des coordonnées en la représentant par un mouvement continu. En adaptant ce principe aux cavités, il est ainsi possible de visualiser leur évolution à partir des composantes principales. La principale différence est le passage d'une grille aux valeurs réelles produites par la formule $G^r = C^{cav} + \lambda * v^{cav}_{desc,i}$ à des cavités booléenes : $G^b_g = (G^r_g > c)$ pour tout point de grille g. Un seuil de coupure c de 0.5 s'est montré satisfaisant et a été utilisé pour le reste de ce chapitre. Les conséquences du passage des valeurs réelles à des valeurs booléennes sont explorées page 76.

Comme vu précédemment (section 2.1), lorsque l'ACP est réalisée dans l'espace des conformations, il est possible de comparer les vecteurs propres d'une trajectoire atomique à ceux des cavités qui lui sont associées. Dans cet espace, il est donc possible d'utiliser les vecteurs propres calculées sur les cavités pour reconstruire des structures. Pour cela on utilisera l'équation III.1, dans laquelle le vecteur propre de cavités dans l'espace des conformations v^{cav}_{step} est utilisé pour construire un vecteur propre de coordonnées atomiques v^{atomes}_{desc} en passant par la matrice des descripteurs atomiques centrés M^{atomes} :

$$v^{atomes}_{desc,j} = \sqrt{s\lambda_j} \cdot v^{cav}_{step,j} \cdot M^{atomes} \tag{III.3}$$

Finalement, l'équation III.2 permet de réaliser la reconstruction proprement dite.

2.3 Indices de qualité et contrôle des outils d'analyse des composantes principales

Les différents outils d'analyse des composantes principales des cavités dans les différents espaces reposent sur des passages entre formalismes discrets et continus. Nous avons donc cherché à mettre en place des métriques et des cas tests de base pour en évaluer la pertinence et la qualité.

2.3.1 Cavités modèles

Afin de mieux comprendre les propriétés de l'ACP sur les grilles (discrètes) de booléens, deux "cavités" modèles ont été créées.

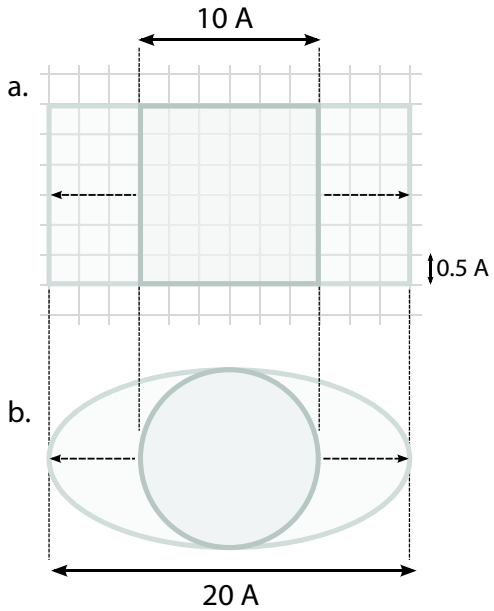


Figure III.1 – Cavités modèles utilisées dans ce chapitre. 1. Cube étiré selon l'axe des abscisses de 10 Å à 20 Å. 2. Sphère étirée selon l'axe des abscisses de 10 Å à 20 Å.

La première cavité (figure III.1.1) a été générée sur une grille de maille 0.5 Å en étirant un cube de 10 Å de côté en un pavé de 20x10x10 Å, par pas de 0.05 Å des deux côtés du cube. La même procédure a été appliquée sur une sphère de rayon 5 Å pour générer la seconde trajectoire (figure III.1.2). Ces cavités sont donc les représentations sur une grille d'objets déformés linéairement au cours du temps dans une unique direction.

2.3.2 Mesures de similarité

Il peut être nécessaire de comparer deux grilles afin de déterminer si celles-ci sont proches l'une de l'autre. Deux indices de similarité seront utilisés pour cela : S et S^r. Dans le cas où les

grilles décrivent des objets booléens (typiquement, lorsque l'on compare des cavités à des pas i et j différents), j'utilise l'indice de Jaccard S :

$$S_{ij} = \frac{G_i^b \cap G_j^b}{G_i^b \cup G_j^b}$$

Pour comparer des grilles à valeurs réelles (pour traiter des comparaisons de composantes principales ou de cavités moyennes par exemple), il est nécessaire d'utiliser un autre indice de similarité adapté à ce type de valeurs. J'utilise pour cela le produit scalaire normalisé des deux grilles, S^r :

$$S_{ij}^r = \frac{1}{\|G_i^b\|\|G_j^b\|} \sum_{l=1}^{n^{cav}} G_{i,l}^b G_{j,l}^b$$

2.4 Systèmes étudiés dans ce chapitre et détection des cavités

En plus des cavités modèles, des trajectoires provenant de dynamique moléculaire de quatres protéines différentes ont été utilisées pour étudier la dynamique des cavités sur des systèmes réels. Ces quatres protéines sont le lysozyme d'œuf de poule (PDB : 2LYZ), la myoglobine du cachalot lié au monoxyde de carbone (provenant des tests de CHARMM), la toxine d'anthrax EF en complexe avec la calmoduline (dénoté EF-CaM, PDB : 1K90) et un dimère de la protéine d'envelope du virus de la Dengue (E ou DenV E).

Trois trajectoires de 120 ns ont été produites pour le lysozyme et la myoglobine, en solvant explicite. Pour des questions de temps de calcul, seule une trajectoire de 10 ns a été produite pour la protéine E. La trajectoire de 15 ns du complexe EF-CaM utilisée dans ce chapitre provient quant à elle d'un précédent article (Laine *et al.*, 2008[112]). Les détails de production des trajectoires sont données en annexe (section 2), et un résumé des trajectoires utilisées est donné table I.

Système	Nbre atomes	Longueur
Protéine E	12258	10ns
Complexe EF-CaM	9942	15ns
Myoglobine (MbCO)	2534	120ns ×3
Lysozyme	1960	120ns ×3

Tableau I – Longueur et nombre d'atomes des trajectoires utilisées dans ce chapitre

Du fait de la détection sur une grille positionnée de façon absolue, il est absolument crucial de bien aligner les structures avant de détecter les cavités. Dans ce chapitre, les trajectoires sont donc toutes alignées sur leur structure cristallographique correspondante. Les cavités ont été détectées au moyen de deux logiciels, `mkgrid` et `gHECOM` (voir section 2.5.3), tous deux produisant des cavités sur une grille de maille 0.5Å, en utilisant une sonde interne de 1.4Å. Il subsiste beaucoup plus de points de grille dans les sorties de `gHECOM` que dans celles de `mkgrid`, du fait de l'absence de filtres dans `gHECOM`. Il a donc été nécessaire de choisir des tailles de sondes externes différentes, soit 10Å

de rayon pour `mkgrid` et 3Å pour `gHECOM`, afin de limiter la taille des fichiers de sortie de `gHECOM`. La figure III.2 donne un aperçu des protéines et de leurs cavités.

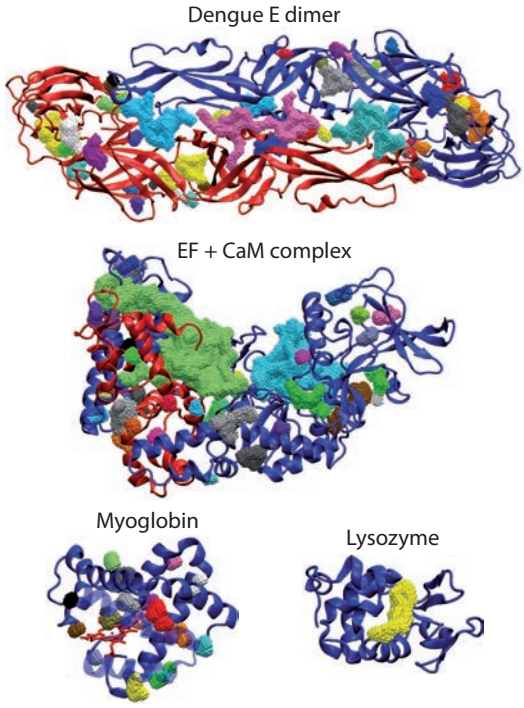


Figure III.2 – Les quatre protéines utilisées dans ce chapitre, ainsi que leurs cavités calculées avec `mkgrid`.

3 Résultats

3.1 Volume et dynamique des cavités pour `mkgrid` et `gHECOM`

Au vu du volume des domaines de définition des trajectoires de cavités (tableau II), on remarque que celles-ci remplissent, au cours de la trajectoire, une fraction très importante du volume englobant la protéine (ce volume est défini par les points de grilles non accessibles par une sonde de rayon 10Å). Ce volume est comparable au volume moyen des protéines. Les cavités sont donc très mobiles et se retrouvent à beaucoup d'endroits dans la protéine.

Les deux logiciels de calcul de cavité `mkgrid` et `gHECOM` utilisés dans ce chapitre présentent des différences importantes de volumes. Ces différences sont expliquées par le fait que `mkgrid` élargit la sonde externe au moment de supprimer les points de grille et empiète donc plus sur les cavités proche du solvant. `mkgrid` applique en outre un filtre de concavité. `mkgrid` produit donc des cavités dont la somme des volumes est généralement beaucoup plus petite que les cavités produites par `gHECOM` (tableau II).

Système	Durée traj.	RMSD (Å)	enveloppe (Å^3)	Vol. du domaine de définition (Å^3) mkgrid	(%)	gHECOM	(%)	vol. atom. moyen (Å^3)
Dengue	10 ns	1.78	239,459	36,314	(15.2)	128,016	(53.5)	127,688
EF-CaM	10 ns	1.80	201,669	37,555	(18.7)	102,954	(51.1)	101,773
Lysozyme1	100 ns	1.91	47,833	9,992	(20.9)	20,486	(42.9)	20,197
Lysozyme2	100 ns	1.42	42,716	7,621	(17.8)	17,405	(40.7)	20,070
Lysozyme3	100 ns	1.46	45,079	8,022	(17.8)	18,324	(40.6)	20,021
Lysozyme4	10 ns	1.17	37,260	4,379	(11.8)	13,847	(37.2)	20,072
Myoglobine1	100 ns	1.31	49,308	5,447	(11.0)	27,066	(54.9)	26,113
Myoglobine2	100 ns	1.20	48,636	5,177	(10.6)	31,265	(64.3)	26,045
Myoglobine3	100 ns	1.33	49,047	5,245	(10.7)	32,254	(65.8)	26,092
Myoglobine4	10 ns	1.12	45,234	3,436	(7.6)	28,342	(62.7)	26,050

Tableau II – Volume de l'enveloppe et du domaine de définition des cavités. La valeur donnée dans la colonne RMSD correspond à la valeur moyenne du RMSD sur les derniers 2/3 des trajectoires atomiques. Le volume est donné en Å^3 pour les cavités de `mkgrid` et `gHECOM`. La fraction du volume du domaine de définition sur le volume de l'enveloppe est donné entre parenthèse, en pourcentages. Le volume atomique est calculé comme le volume inaccessible à une sphère de 1.4 Å de rayon. Ce volume est moyenné sur l'ensemble des conformations.

3.2 Autocorellation temporelle des trajectoires de cavités

En utilisant la mesure de similarité définie précédemment (section 2.3.2), on peut comparer l'échelle temporelle d'évolution des cavités. La figure III.3 montre la tendance des cavités à diverger au cours du temps, à la manière d'une marche aléatoire. On retrouve ce type de profil d'auto-correlation pour les trajectoires atomiques, typique d'un comportement diffusif. En augmentant l'échantillonnage des conformations de la protéine, on peut donc améliorer l'échantillonnage des cavités.

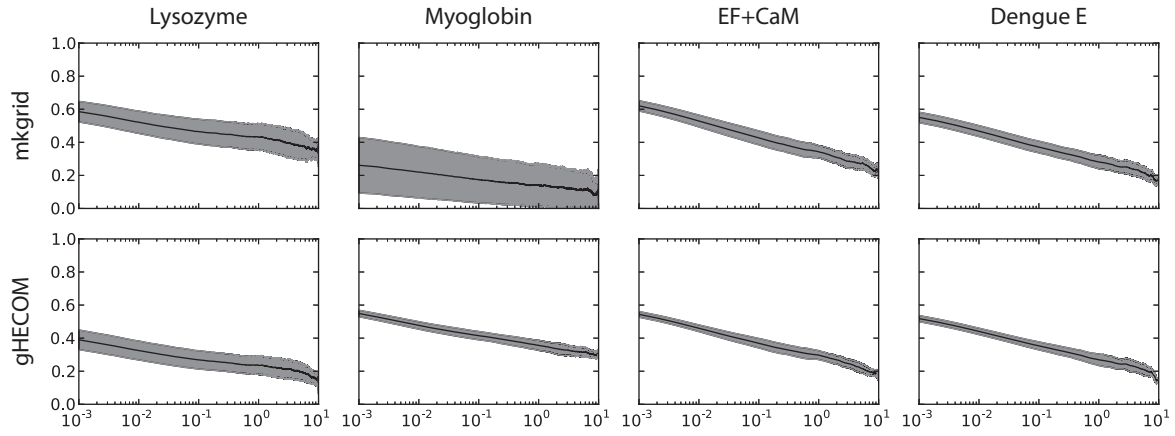


Figure III.3 – Similarité moyenne entre cavités en fonction de l'écart de temps (ligne noire). Les cavités sont calculées avec `mkgrid` (1er rangée) et `gHECOM` (2^{e} rangée) pour les quatre systèmes étudiés, d'après les trajectoires de 10ns. L'échelle de l'abscisse est logarithmique et l'échelle de temps est exprimée en ns. La déviation standard de la similarité est indiquée par la surface grisée.

3.3 Propriétés générales des composantes principales des cavités

L'ACP est classiquement utilisée sur des séries de nombres réels (images, nuages de points, trajectoire atomiques...). Il est donc nécessaire de bien comprendre les implications de l'application de l'ACP à une grille de points booléens. Pour cela, plusieurs outils ont été conçus pour analyser la géométrie et la composition des composantes principales.

3.3.1 Premier exemple

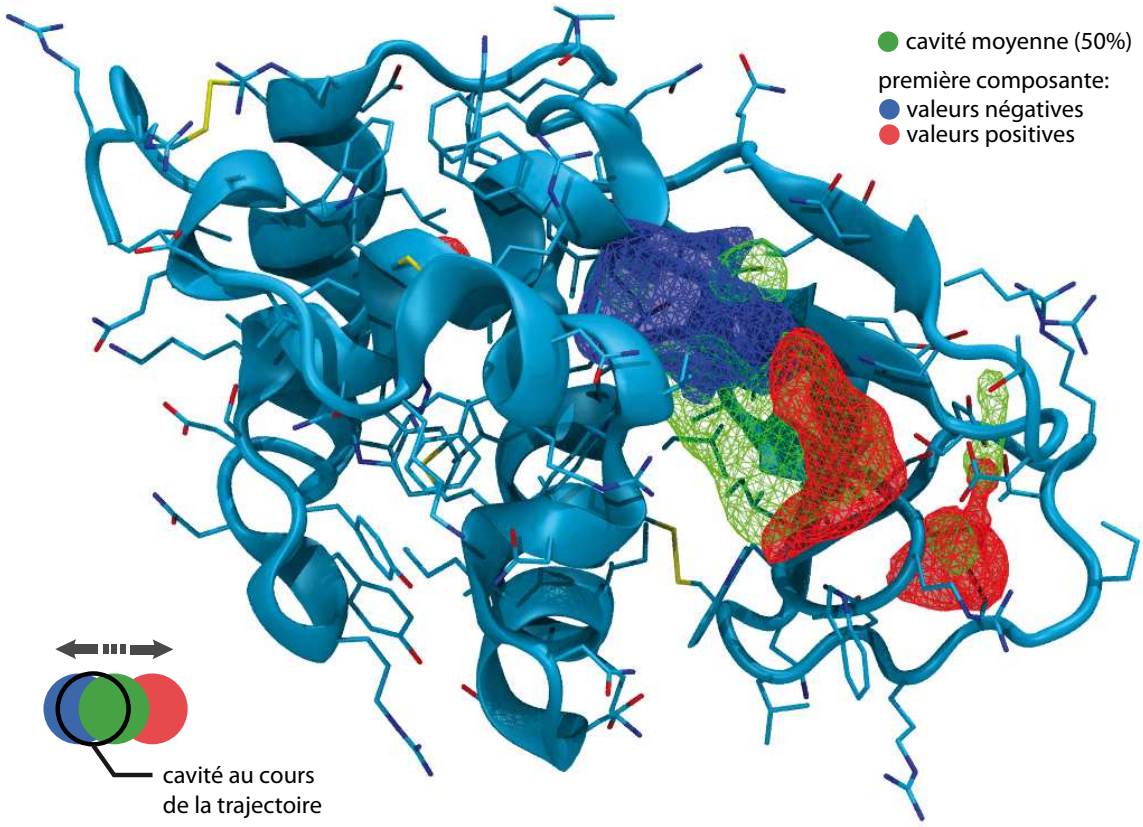


Figure III.4 – Première composante principale de cavité d'une trajectoire de 100ns du lysozyme. Les cavités ont été produite par `gHECOM`. La grille verte représente l'isosurface à 50% de la cavité moyenne C. La première composante des cavités est représentée par une grille bleue (resp. rouge) au niveau d'une isosurface de valeur négative (resp. positive).

La figure III.4 montre la réprésentation d'une composante principale de cavité (ici, la 1re CP d'une trajectoire de lysozyme de 100ns). Les zones bleues et rouges sont de signes opposés, et sont donc anticorrélées (à noter que le signe d'une composante principale est arbitraire, seul le changement de signe est important). Les zones de même signe sont corrélées positivement. Ainsi, lors de la trajectoire, une cavité va avoir tendance à évoluer entre les zones bleues et rouges, alternativement.

3.3.2 Spectre

Le spectre des valeurs propres indique la répartition de la variance dans la trajectoire. Plus les premières valeurs propres sont grandes, plus la trajectoire est dominée par un petit nombre de grandes directions d'évolution. Les dernières valeurs propres correspondent à de petits mouvements locaux que l'on peut voir comme un bruit de fond. Plus ces valeurs sont importantes, plus ce bruit l'est aussi. On remarque en premier lieu que le spectre des valeurs propres d'une trajectoire de cavité présente un profil plus diffus que celui d'une trajectoire atomique (figure III.5). Les premières composantes principales des cavités sont donc moins informatives que celles des trajectoires atomiques, on peut donc dire que les trajectoires de cavités sont plus "bruitées" que les trajectoires atomiques. Les valeurs propres des cavités produites par `gHECOM` sont elles aussi plus diffuses que celles produites par `mkgrid`.

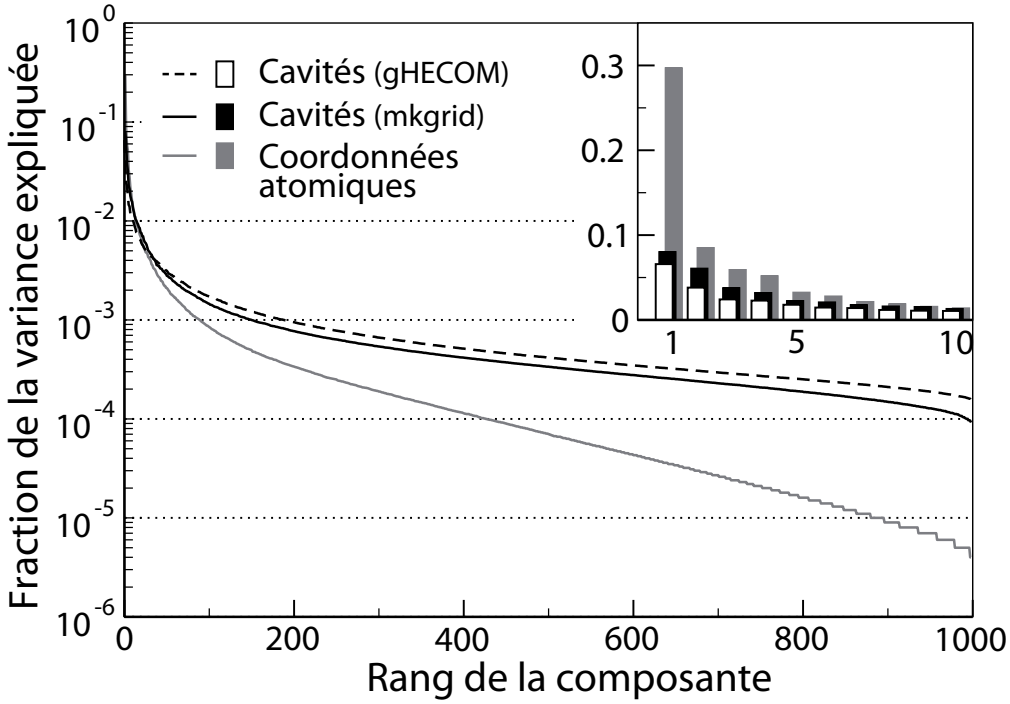


Figure III.5 – Le spectre des valeurs propres des trajectoires atomiques et des cavités `mkgrid` et `gHECOM` correspondantes, pour une trajectoire de 100ns du lysozyme. Les valeurs données sont normalisée pour correspondre à la fraction de variance expliquée par chaque composante. L'abscisse du graphe principal est logarithmique, et celle du graphe inclu linéaire.

3.3.3 Autocorrelation spatiale des composantes principales de trajectoire de cavités

Afin de déterminer l'échelle caractéristique des variations de chaque composante principale dans l'espace des cavités, j'ai conçu un nouvel indice, nommé ARA (pour *average radial autocorrelation*, aucorrelation radiale moyenne) :

$$ARA(r) = \frac{1}{r^2} \int_{v \in S_r} \mathrm{d}A \int_{\mathbf{x} \in \mathbb{R}^3} \mathrm{d}\mathbf{x} CP(\mathbf{x}) CP(\mathbf{x} - \mathbf{v})$$

où dA est un petit élément de surface de la sphère S_r de rayon r, situé en $\mathbf{v}$. A noter, pour des raisons de performance, l'intégrale correspondant à l'autocorrelation est calculée par Transformée de Fourier Rapide (FFT).

Afin de résumer cette fonction en une seul valeur, j'utiliserai également son intégrale (notée ARAI) :

$$ARAI = \int_0^{\frac{d}{2}} ARA(r)\mathrm{d}r$$

où d est la taille du côté le plus petit de la grille.

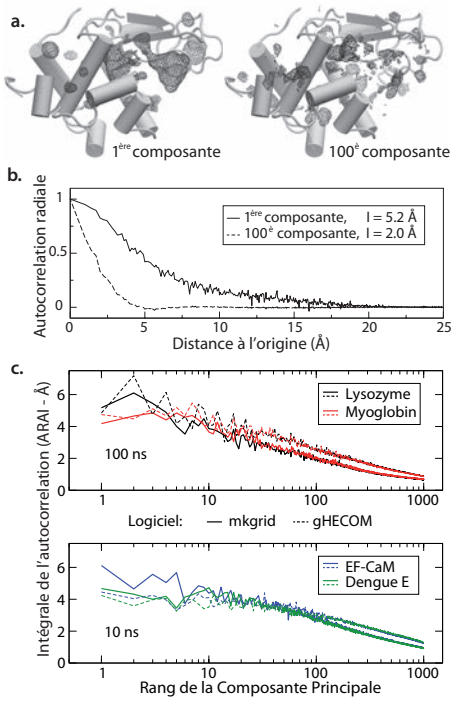


Figure III.6 – Autocorrelation spatiale des composantes de cavités. **a.** Isosurfaces de la 1^{re} (à gauche) et de la 100^{me} (à droite) composante principale des cavités du lysozyme (représentation en grille). La structure du lysozyme est également indiquée. **b.** Variation de l'ARA pour les deux composantes représentées en **a**. Les valeurs correspondantes d'ARAI pour chaque courbe sont indiquées en légende (I). **c.** ARAI des composantes des trajectoires de 100ns du lysozyme (en noir) et de la myoglobine (en rouge), et des trajectoires de 10ns du complexe EF-CaM (en bleu) et de la protéine E de la Dengue (en vert) pour des trajectoires de cavités produites par `mkgrid` (lignes continues) et `gHECOM` (pointillés). L'axe du rang des composantes principales (abscisse) est en échelle logarithmique.

La figure III.6.c montre une décroissance de l'autocorrelation quand on va vers les composantes principales de plus haut rang (valeur propre plus faible). Une valeur d'ARAI élevée correspond à une autocorrelation plus élevée, donc à une échelle d'évolution spatiale plus grande, plus *lisse* (figure III.6.b). Les premières composantes principales correspondent donc à des variations sur de grandes distances et à des mouvements globaux des cavités, alors que les composantes principales de plus faible rang fournissent des informations très locales, proches d'un "bruit de fond" (figure III.6.a). On peut noter également que la courbe ARAI montre d'importantes variations entre les différentes composantes principales de faible rang. Ces différents bruits peuvent provenir de deux sources. Premièrement, les trajectoires de cavités peuvent être intrinsèquement très vola-

tiles, avec des variations locales de petite ou grande amplitude. Deuxièmement, l'application même de l'ACP sur une grille de booléens peut provoquer des artefacts visibles dans les composantes principales. Ce dernier point est traité dans la section suivante.

3.3.4 Conséquences de l'application de l'ACP sur une grille de booléens

En utilisant les cavités modèles, il est possible de déterminer les conséquences de l'application de l'ACP sur une grille de booléens. En effet, les objets sous-jacents aux cavités modèles (cube, sphère) sont étirés linéairement sur un seul axe, ce qui correspond à une unique direction d'évolution, suggérant donc une unique composante principale. Le passage de ces objets sous forme de grilles implique la perte de la linéarité de cette évolution, qui devient une évolution par paliers (un saut de 0 à 1 à chaque fois que le bord de l'objet passe un point de grille).

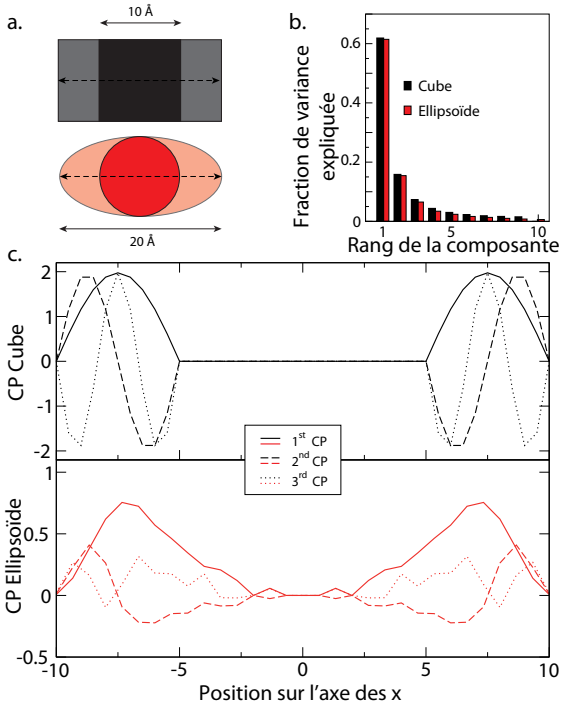


Figure III.7 – Effet de la discrétisation de l'espace sur les composantes principales. a. Vue schématique des cavités modèles utilisées (cube en noir, ellipsoïde en rouge ; les zones étirées sont respectivement colorées en gris et rose). **b.** Contribution des modes à la variance globale pour le cube (barres noires) et la sphère (barres rouges). Ne sont indiquées que 10 des 79 valeurs propres non-nulles pour la sphère. **c.** Profil des trois premières composantes principales, moyennées le long de l'axe des abscisses, pour le cube étiré (en noir) et la sphère étirée (en rouge).

La figure III.7.b montre que le spectre des valeurs propres des cavités modèles n'est pas constitué que d'une composante, mais bien d'une multitude. Ainsi, la contribution de la première valeur propre au mouvement global d'étirement est de 61.9% pour le cube et 61.9% pour la sphère. De même, il existe respectivement 9 et 79 valeurs propres non nulles pour ces deux systèmes (la différence étant due à la représentation plus subtile des ellipsoïdes sur une grille). On observe donc une "fuite" dans les composantes principales de plus haut rang. La répartition dans l'espace des

trois premières composantes principales (figure III.7.c) est proche de fonctions siunusoïdes, au lieu des fonctions linéaires attendues pour un tel étirement réalisé à vitesse constante. Ainsi, l'ACP sur les cavités modèles révèle et quantifie les limites de l'application de cette méthode sur une représentation discrétisée de l'espace. Cette conclusion permet de mettre en perspective certaines interprétations des composantes principales de cavités, notamment dans leur comparaison avec les composantes principales des trajectoires atomiques, plus classiques.

3.3.5 Effet de la troncature lors de la reconstruction de trajectoire des cavités

La troncature en booléen nécessaire à la reconstruction de trajectoire de cavités induit une perte de linéarité, comme indiqué section 2.2. De plus, lorsque s est plus petit que n^{cav} (ce qui est généralement le cas), la base formée par les s vecteurs propres non nuls est incomplète et n'engendre qu'un sous espace de l'espace des cavités, de dimension n^{cav}. Ces deux observations indiquent que la troncation vers les valeurs booléennes d'une combinaison linéaire des vecteurs propres v_i (telle qu'engendrée par la reconstruction, équation III.2) doit génériquement "fuire" en dehors du sous espace des vecteurs propres, de dimension s.

Pour évaluer l'étendue de cette "fuite", je définis un déplacement cible, comme le déplacement par rapport à la cavité moyenne de la reconstruction, avant la troncature : $\Delta^* = G^r - C^{cav}$. Je compare ce déplacement cible avec le déplacement effectif, après troncature : $\Delta^{\mathit{eff}} = G^b - C^{cav} = (G^r \geq c) - C^{cav}$.

Ces comparaisons sont effectuées en utilisant 4 indices, compris entre 0 à 1, et dénommés I_{self}, I_{comp}, I_{sub} and I_{other}. I_{self} est la fraction du déplacement effectif allant dans la direction idéale :

$$I_{self} = \frac{\Delta^{\mathit{eff}} \cdot \Delta^*}{\|\Delta^{\mathit{eff}}\| \, \|\Delta^*\|}$$

Ici, $\|X\|$ dénote la norme du vecteur X dans l'espace complet à n^{cav} dimensions. Plus I_{self} est grand, plus la cavité tronquée est proche de la cavité cible, et moins la "fuite" est importante. I_{sub} est la fraction du déplacement effectif située dans le sous-espace défini par la trajectoire des cavités d'origine (le sous espace à s dimensions). On le définit par produit scalaire avec les s vecteurs propres v_i :

$$I_{sub} = \left(\sum_{i=1}^{s} \left(\frac{\Delta^{\mathit{eff}}}{\|\Delta^{\mathit{eff}}\|} \cdot v_i \right)^2 \right)^{1/2}$$

I_{comp} est la fraction du déplacement efficace se situant dans l'hyperplan orthogonal à Δ^*, dans le sous espace de la trajectoire d'origine à s dimensions :

$$I_{comp} = \left(I_{sub}^2 - I_{self}^2 \right)^{1/2}$$

Pour terminer, I_{other} est la fraction du déplacement effectif orthogonal à Δ^* dans l'espace complet des cavités.

$$I_{other} = \left(1 - I_{self}^2 \right)^{1/2}$$

Les courbes de I_{self} et I_{comp} pour les cavités modèles (figure III.8.a) montrent une chute nette de la similarité aux alentours de la cavité moyenne. La présence d'un trou lorsque les cavités approchent C est attendue, puisque le déplacement idéal Δ^* dans la cavité moyenne est nul ; les indices sont de fait non définis lorsque $G^r = C$. Le fait que ce creux ait une largeur non nulle est par contre un effet direct du passage des valeurs réelles aux valeurs booléennes. Cela s'explique par le fait que le passage aux booléens, $\Delta^{eff} - \Delta^*$, est a priori indépendant de Δ^*, et est donc d'autant plus ressenti lorsque $\|\Delta^*\|$ est petit. On retrouve ce phénomène pour les courbes de I_{self}, I_{comp} et I_{sub} pour les reconstructions de cavités (figure III.8.b).

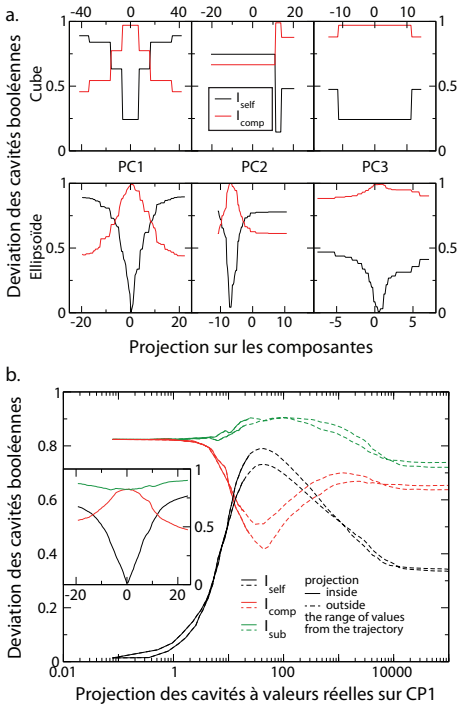


Figure III.8 – Reconstructions des cavités, effets du passage aux booléens et de l'extrapolation. a. Effet du passage aux booléens sur les reconstructions le long des trois premières composantes (CP1-3) des cavités modèles (cube au dessus, ellipsoïde en dessous). Les courbes de I_{self} et I_{comp} sont respectivement en noir et rouge. A noter que pour l'ensemble des conformations, I_{sub} vaut 1 et I_{other} est égal à I_{comp} ; ces deux indices ne sont donc pas représentés dans cette figure. L'abscisse spécifie la valeur des projections des cavités sur les 3 premières composantes *avant* le passage aux booléens. **b.** Reconstructions réalisées le long de la première composante principale (CP1) des cavités `mkgrid` d'une trajectoire de 100ns du lysozyme. I_{self}, I_{comp} et I_{sub} sont indiqués en noir, rouge et vert respectivement, en un trait continu pour les projections comprises dans l'intervalle des valeurs de la trajectoire d'origine, et en pointillés au delà. L'abscisse du graphe principal est en échelle logarithmique, et la valeur de la projection est donnée en valeur absolue (d'où la courbe doublée indiquant à la fois les parties positives et négatives de la courbe). L'abscisse du graphe en encadré est en échelle linéaire.

Une autre information révélée par cette analyse (figure III.8) est que les cavités reconstruites extrapolées, c'est-à-dire les cavités $T(\mu)$ pour lesquelles $T(\mu)\cdot\Delta^* > max_i(D_i^{cav}\cdot\Delta^*)$ peuvent avoir des valeurs de I_{self} très grandes. Ces cavités ressemblent donc beaucoup à la composante principale utilisée pour la reconstruction, et ce même au delà de l'espace couvert par l'échantillonnage support de l'analyse.

3.4 Compression des descripteurs de cavités et utilisation pour la sélection de cavités représentatives

L'ACP peut être vue comme une rotation de l'espace de départ, un changement de repère qui maximise la variance dans chacune des directions définies par les composantes principales. La projection d'une cavité sur un vecteur propre donne donc une coordonnée dans ce nouvel espace. Cette coordonnée correspond en outre à la ressemblance de cette cavité avec la composante principale étudiée. En suivant l'évolution des valeurs des projections sur quelques vecteurs propres à la fois, on peut suivre l'évolution de la forme des cavités au cours de la trajectoire.Cela permet de repérer les conformations ayant des géométries de cavités similaires, ou au contraire de détecter des changements soudains de forme.

Il faut noter que quelque soit le nombre N de vecteurs propres utilisés pour réduire la dimension du repère, la variance expliquée par ces N premiers vecteurs propres est optimale. En projetant dans ce repère on obtient un nouveau vecteur de taille N, qui est ainsi la description la plus fiable des cavités en n'utilisant que N descripteurs car elle maximise la quantité d'information inclue pour ces N dimensions sans déformer l'espace. L'ACP peut donc être utilisée pour compresser l'espace de départ, souvent très grand (le nombre n^{cav} de points de grille pouvant dépasser la centaine de milliers pour des protéines de grande taille) en un espace beaucoup plus petit, en perdant un minimum d'information. Ces descripteurs compressés peuvent ensuite être traités par des algorithmes qu'il serait trop coûteux d'appliquer aux descripteurs initiaux.

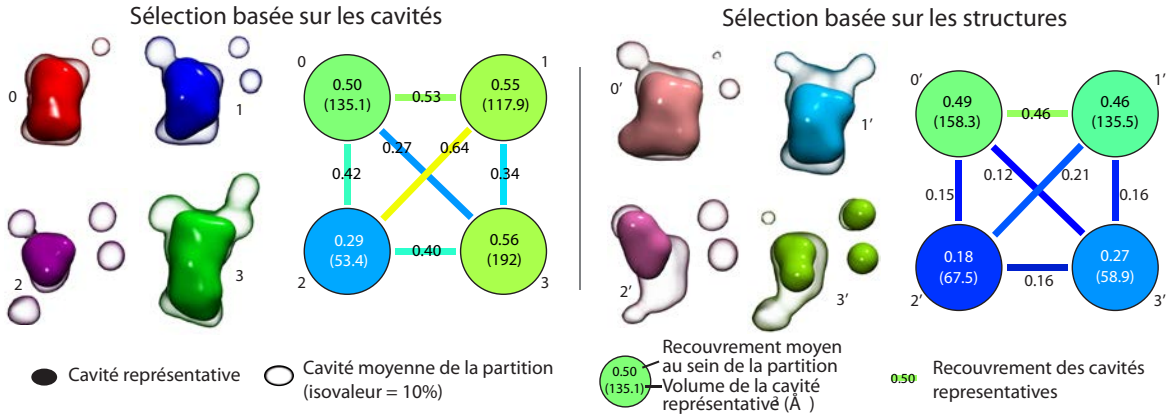


Figure III.9 – Evaluation de la diversité des cavités sélectionnées par leur forme de cavité ou leur structure. Les cavités représentatives (opaques) et les cavités moyennes (isosurface à 10%) des 4 groupes de cavités définies par l'algorithme des *k-moyennes* (à gauche : partitionnement basé sur les cavités, à droite : basé sur les structures) de la poche β-OG de la protéine E de la Dengue. Les cavités ont été produites avec `gHECOM`. Les diagrames donnent la mesure de la moyenne de la similarité des cavités de chaque groupe avec leur centroïde (similarité intra, dans les disques) ainsi que la similarité entre les différents centroïdes (similarité inter, lignes) pour les deux types de partitionnement (basé sur les cavités ou les structures). Le code couleur va du bleu (0 : pas de similarité) au rouge (1 : identité) et correspond à la valeur indiqué sur chaque symbole.

Un exemple de cette façon de procéder est donné figure III.9 en utilisant un algorithme de partitionnement sur des descripteurs compressés pour déterminer des cavités représentatives de classes diverses. La cavité utilisée dans cet exemple est la cavité β-OG, cible de nombreux projets

d'antiviraux[181, 30, 194]. L'ACP sur la trajectoire de cavité de β-OG permet ici de diminuer la dimension du descripteur de cavités de 18 131 à 100, tout en capturant 74% de la variance totale. L'algorithme de *k-moyennes* est ensuite utilisé pour partitionner les descripteurs en quatre groupes. La même procédure a aussi été appliquée en utilisant les 100 premières composantes principales de la trajectoire atomique de la poche entourant β-OG (capturant ici 91% de la variance totale de la poche). Pour cet exemple, les cavités des conformations sélectionnées en utilisant les coordonnées atomiques sont légèrement plus diversifiées que celles sélectionnées en utilisant les cavités (les similitudes des cavités moyennes sont plus faibles). Par contre, les cavités sélectionnées à l'aide de l'ACP sur les cavités montrent une similitude moyenne bien plus importante, et sont donc plus représentatives que les cavités sélectionnées à l'aide de la PCA sur les structures. Les cavités sélectionnées à l'aide de l'ACP sur les cavités sont en moyenne plus volumineuses (125±57 Å^3) que les cavités sélectionnées par l'ACP sur les structures (105±49 Å^3, voir également le tableau III).

# cavité	Volume de la cavité représentative (^{3})	
	selection par les cavités	sélection par les structures
0	135.13	158.25
1	117.88	135.5
2	58.38	67.5
3	192	58.88

Tableau III – Volume des cavités representatives de la poche β-OG, selectionnées par la géométrie des cavités ou des structures

3.5 Correlation entre l'evolution des structures et de leurs cavités associées

La possibilité de réaliser l'ACP dans l'espace des conformations ouvre une voie intéressante : la comparaison des composantes principales des trajectoires atomiques et des cavités. En effet, lorsque l'ACP est réalisée dans cet espace, les composantes principales sont exprimées en terme de combinaisons linéaires d'indices des pas de trajectoire, et non de descripteurs. Dans ce cas, les composantes principales peuvent être facilement comparées, puisqu'elles font références aux mêmes objets, les indices de la trajectoire. Pour réaliser cette comparaison, j'utilise le RMSIP$_n$ (Root Mean Square Inner Product, la racine de la moyenne du produit scalaire au carré)[202, 118] :

$$RMSIP_n = \left(\sum_{k=1}^{n} \frac{1}{n} \sum_{l=1}^{n} \left(v_l^{cav} \cdot v_k^{coord} \right)^2 \right)^{1/2} \tag{III.4}$$

La valeur $n = 10$ généralement utilisée dans la littérature est également utilisée ici. Plus la valeur est proche de 1, plus les espaces définis par les n premiers vecteurs propres de chaque ensemble sont proches, donc plus les variations correspondantes sont semblables.

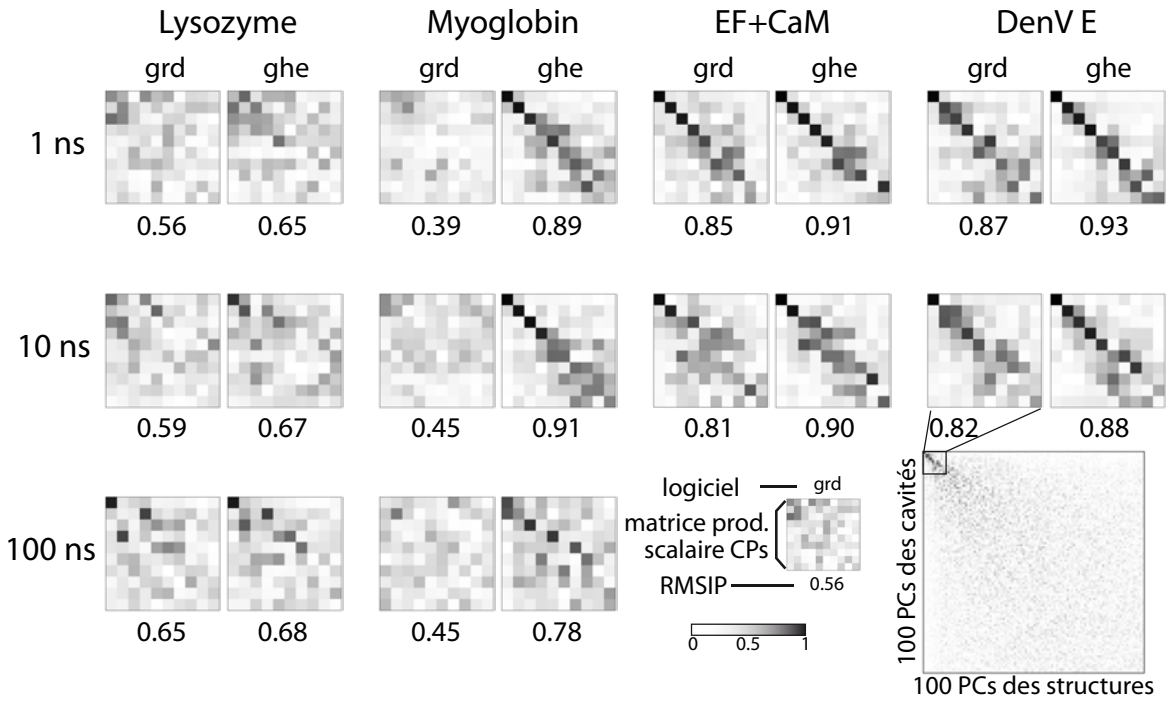


Figure III.10 – Similarités des vecteurs propres des cavités et des structures. Les matrices de produit scalaire des 10 premières composantes principales des cavités et des structures sont indiquées ici pour les quatre systèmes étudiés. Les trajectoires de 1ns, 10ns et 100ns (pour le lysozyme et la myoglobine) ont été analysées avec `mkgrid` (grd) et `gHECOM` (ghe). La valeur absolue est donnée en niveau de gris, du blanc (0, orthogonalité) au noir (1, identité). Les valeurs de RMSIP_{10} sont données en dessous de chaque matrice.

La figure III.10 montre les différentes matrices de produit scalaire des 10 premiers vecteurs propres de structures et de cavités, ainsi que le RMSIP_{10} correspondant. Les valeurs de RMSIP_{10} sont élevées, variant généralement entre 0.5 et 0.9. On peut également noter que les valeurs de RMSIP_{10} sont systématiquement plus élevées pour les cavités `gHECOM` que pour les cavités `mkgrid`. Cela peut s'expliquer par la présence des filtres dans `mkgrid` qui peuvent supprimer des cavités de surface de façon plus abrupte que le seul critère de volume de `gHECOM`. Le bruit associé fait donc diminuer les scores de RMSIP_{10} de `mkgrid`. De façon marquante, les valeurs de RMSIP_{10} varient très fortement en fonction des systèmes. Une tendance associant une plus grande valeur de RMSIP_{10} aux gros systèmes apparaît, mais il est difficile d'en tirer des conclusions au vu du faible nombre de cas traités. On peut noter aussi la faible influence de la longueur de la trajectoire sur la similarité des composantes structure-cavité.

Ces valeurs de corrélation montrent que les limites soulevées dans les sections 3.3.2 et 3.3.4 sont finalement peu pénalisantes dans les cas pratiques. Cette corrélation suggère également que l'ACP est un outil efficace permettant de relier dans les deux sens les positions atomiques à la géométrie des cavités.

3.6 Construction de structures ciblant des géométries de cavités

Comme vu section 2.2, il est possible de reconstruire des structures en utilisant les composantes principales des cavités. On peut alors calculer les cavités de ces nouvelles structures, et les comparer aux cavités de la trajectoire d'origine. La figure III.11 montre qualitativement que les cavités des structures reconstruites semblent plus proches de la forme décrite par la composante principale que les cavités d'origine. Cela est effectivement démontré figure III.12.a pour la trajectoire de 10ns de la Dengue : les cavités provenant des structures reconstruites ont une valeur de I_{self} systématiquement plus élevée que les cavités d'origine. On remarque qu'en outre, tant que les projections des cavités restent dans l'intervalle de valeurs de la trajectoire d'origine, ces cavités de structures reconstruites sont globalement comparables aux cavités reconstruites décrites en section 3.3.5 (page 80). En revanche, les cavités des structures reconstruites extrapolées (en dehors du cadre de la trajectoire d'origine) divergent de plus en plus de la composante principale de départ. Leur valeur de I_{sub} plonge elle aussi, ce qui indique que des cavités apparaissent à des endroits non explorés dans la trajectoire d'origine. En somme, la structure devient trop déformée pour reproduire les cavités de la trajectoire d'origine.

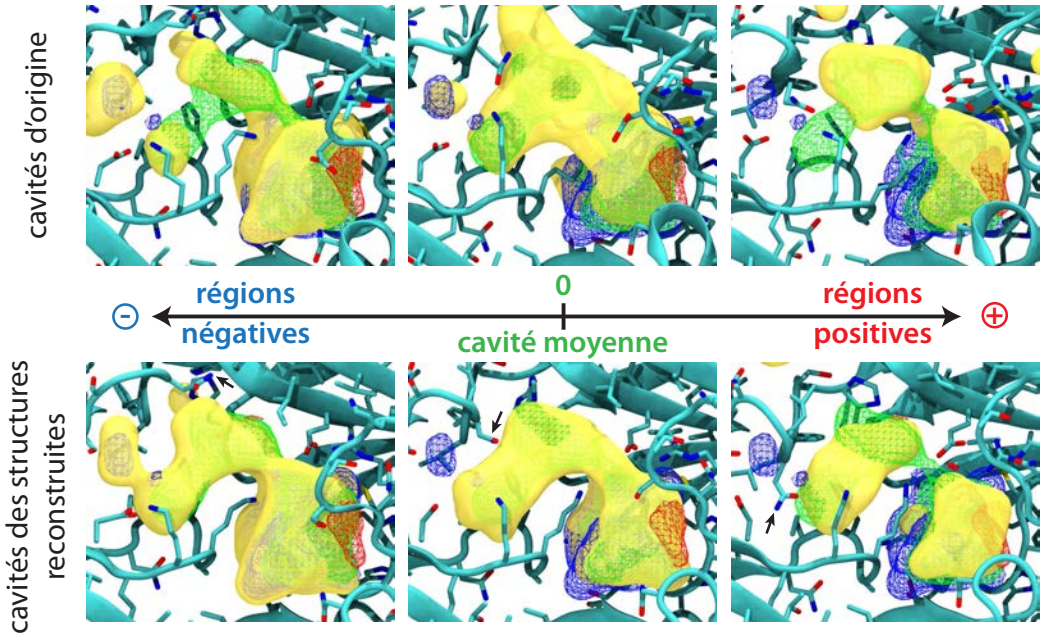


Figure III.11 – Comparaison des cavités d'origine et des cavités des structures reconstruites de la dynamique de 10ns de la protéine E de la Dengue (cavités `mkgrid`). Les cavités représentées sont celles ayant la valeur I_{self} la plus élevée (en jaune ; à gauche : projection négative, à droite : projection positive, au milieu : projection la plus proche de 0). En haut : cavités d'origine de la trajectoire, en bas : cavités des structures reconstruites. Les valeurs négatives (resp. positives) de la première composante principale sont représentées par un grillage bleu (resp. rouge). La cavité moyenne est représentée par un grillage vert.

Ces informations sont résumées pour chaque système par la valeur maximale de I_{self} qu'il est possible de tirer des trajectoires de cavités d'origine, reconstruites ou provenant de structures reconstruites (figure III.12.b). On retrouve les mêmes conclusions qu'au paragraphe précédent pour toutes les trajectoires de la protéine E de la Dengue et pour le complexe EF-CaM. En revanche, les

cavités provenant de structures reconstruites semblent moins proches de la composante principale de reconstruction que les meilleures cavités d'origine pour les trajectoires du lysozyme et de la myoglobine. Leur score d'I_{self} reste cependant très bon, ce qui peut être intéressant pour augmenter l'échantillonnage des cavités le long d'une direction d'évolution.

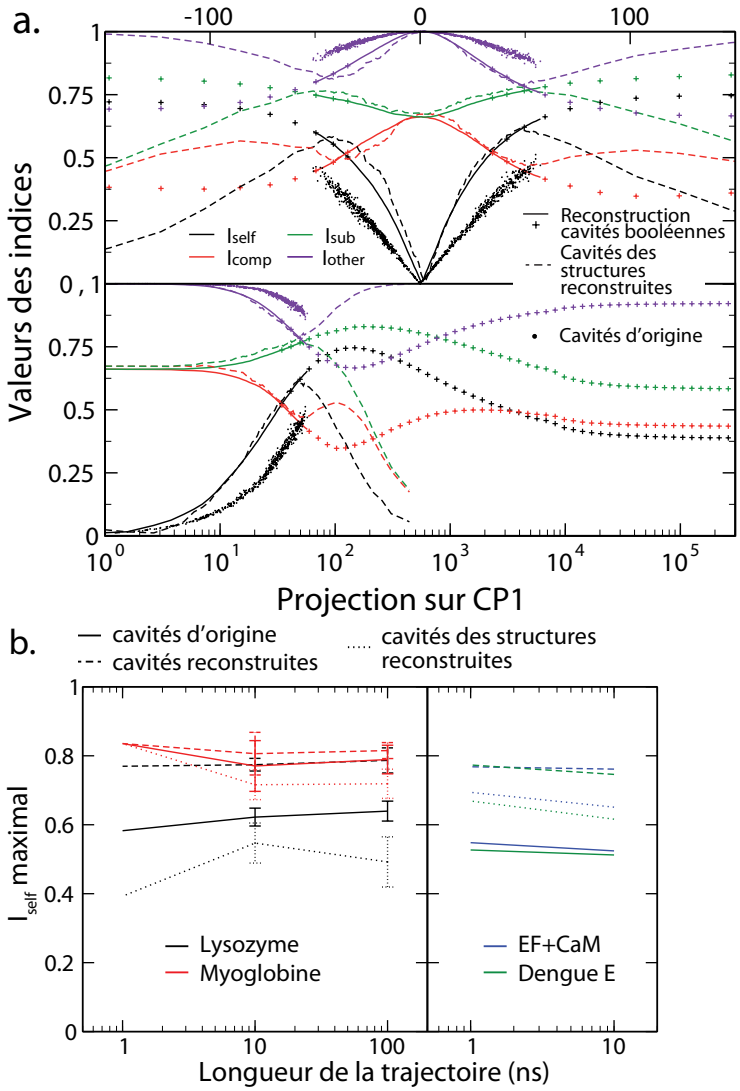


Figure III.12 – Reconstruction de structures à l'aide de composantes principales de cavités. a. Similarité de la première composante principale de la trajectoire de cavités `mkgrid` de 10ns de la protéine E avec différents types de cavités. Les cavités d'origine (nuage de points), les reconstructions de cavités le long de la première composante de cavités (ligne continue, et symboles + au delà de l'extension maximale de la trajectoire d'origine) et les cavités des structures reconstruites le long de cette composante (pointillés) sont analysées. Les structures ont été reconstruites jusqu'à 15 fois la déviation standard des projections de la trajectoire. Les indices I_{self} (en noir), I_{comp} (en rouge), I_{sub} (en vert), I_{other} (en violet) sont représentés. L'abscisse est linéaire pour le graphe supérieur, logarithmique pour le graphe inférieur (seules les valeurs positives des projections sont indiquées ici). **a.** Les valeurs maximales de I_{self} des cavités d'origine (ligne continue), des cavités reconstruites (traits interrompus) et des cavités provenant des structures reconstruites (pointillés) sont représentées pour les différentes échelles de temps des systèmes étudiés (lysozyme en noir, myoglobine en rouge, complexe EF-CaM en bleu et protéine E en vert).

3.7 Utiliser l'ACP sur les cavités pour étudier la "respiration" de la myoglobine

L'ACP permet de décomposer un mouvement en plusieurs directions globales d'évolution. La possibilité de relier les évolutions des structures et des cavités en passant par l'espace des pas de temps peut être exploitée dans une analyse fonctionnelle, pour tirer des informations biologiques pertinentes d'une trajectoire de cavités. Pour illustrer ce point, j'ai analysé l'importance de l'évolution des cavités dans le phénomène de "respiration" de la myoglobine permettant le mouvement de petits ligands (dioxygène, monoxyde de carbone, monoxyde d'azote) au sein de son réseau de cavités internes. La myoglobine comporte en effet une multitude de cavités très proches les unes des autres et en général séparées du solvant extérieur. Ces cavités comprennent les quatre cavités Xenon (elles ont été observées pour la première fois par crystallographie par rayons X en présence de Xenon à haute pression), Xe1 à Xe4, la poche distale, située au dessus de l'hème et dans laquelle se loge le ligand lorsqu'il est attaché à l'hème, ainsi que quelques cavités observées plus tardivement, notamment les cavités "fantômes" Ph1 et Ph2 (non observées ici). De nombreuses études sur la myoglobine[114, 115, 116, 117, 118, 119, 120, 121, 122] ont permis de déterminer les différentes voies de passage et de sortie des ligands au sein de ce réseau de cavité, ainsi que les résidus "portes" associés au passage d'une cavité à l'autre ou à la sortie du ligand vers le solvant.

Les trois trajectoires de la myoglobine utilisées dans ce chapitre comprennent une molécule de monoxyde de carbone (CO), initialement située dans une poche au dessus de l'hème, mais détachée et donc libre d'évoluer dans l'espace. Au cours de ces trois trajectoires, le CO visite une multitude de cavités :

— dans la première trajectoire, le CO visite la poche distale (DP), Xe4, puis retourne dans DP et y reste jusqu'à la fin des 120ns

— dans la seconde trajectoire, le CO visite la poche distale (DP), puis oscille entre les poches Xe4, Xe2 et Xe1 avant d'entrer dans Xe3 vers la fin de la trajectoire

— dans la troisème trajectoire, le CO visite la poche distale (DP), une petite poche au dessus de DP, puis sort après 8ns par une sortie située entre les résidus F46, H48, L49, M55, D60 et L61.

Ces trois trajectoires ont été concaténées pour former une trajectoire globale de 248ns (la trajectoire 3 étant tronquée après la sortie du CO). L'ensemble des positions sucessives du CO a ensuite été partitionnée à l'aide de l'algorithme DBSCAN (voir page 162) selon les différents sites de liaisons (DP et Xe1 à Xe4, plus une partition "exception"). Ces positions sont représentées figure III.13, et la composition des partitions est détaillée dans le tableau IV.

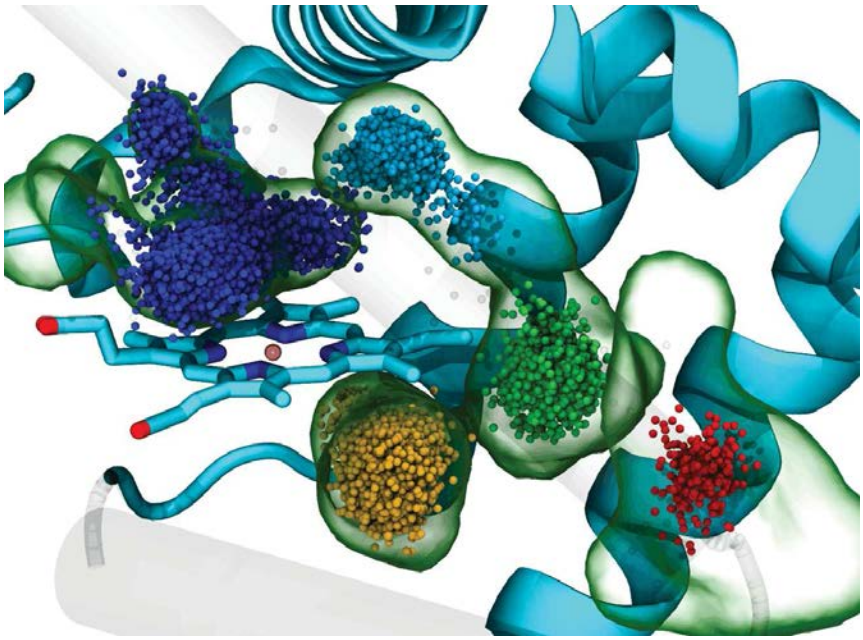


Figure III.13 – Définition des sites de liaison du CO dans la myoglobine. Positions de l'atome d'oxygène du monoxyde de carbone au cours des 248ns de la trajectoire concaténée (24 800 positions). Le code couleur correspond à la définition du site de liaison par DBSCAN : DP en bleu, Xe4 en cyan, Xe2 en vert, Xe1 en jaune et Xe3 en rouge. L'isosurface à 3% de la cavité moyenne est indiquée en contour vert.

Site	DP	Xe4	Xe2	Xe1	Xe3	"exceptions"
# conf.	12224	1789	1559	8954	231	43

Tableau IV – Nombre de conformations pour lesquelles CO est lié pour chaque site., définis par l'algorithme DBSCAN.

Les cavités utilisées sont celles produites par `mkgrid` pour la trajectoire de 248ns définie ci-dessus. Les cavités ont été filtrées pour ne garder que les cavités internes, celles ayant au moins un point de grille à moins de 0.5 Å d'une des 24,800 positions du CO. Pour chaque site de liaison du CO, on définit un jeu de cavités internes moyen calculé sur les pas de temps pour lesquels le CO s'y trouve. La projection de ces cinq jeux de cavités moyennes sur les vecteurs propres des cavités de la trajectoire de 248ns (figure VI.6 en annexe) montre une ressemblance très forte avec la première composante principale, ainsi qu'une ressemblance de la cavité moyenne de Xe2, relativement centrale, avec la 3^{e} composante.

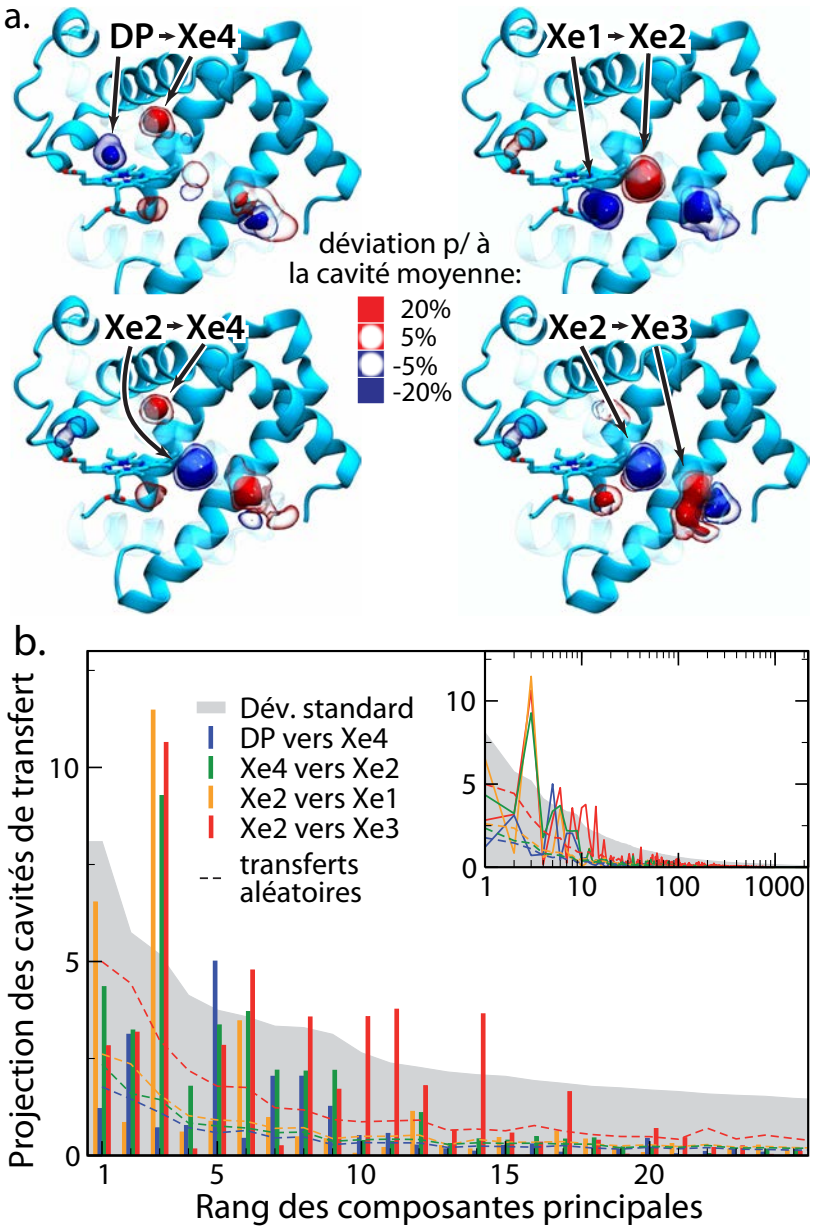


Figure III.14 – Localisation et amplitude des cavités de transfert. a. Cavités de transfert des paires de sites de liaison du CO adjacentes. De haut en bas et de gauche à droite : Xe4 → DP, Xe4 → Xe2, Xe2 → Xe1, et Xe3 → Xe2. Les isosurfaces positives des cavités de transfert (transfert positif net de volume par rapport à la cavité moyenne) sont représentées en rouge, et les isosurfaces positives en bleu (surface opage : 20%, contour : 5%). **b.** Valeurs absolues des projections des cavités de transfert sur les composantes principales des cavités, en bleu pour le transfert DP → Xe4, en vert pour Xe4 → Xe2, en jaune pour Xe2 → Xe1 et en rouge pour Xe2 → Xe3. La valeur propre de chaque composante (correspondant à la déviation standard des projections de l'ensemble des cavités sur cette composante) est représentée en gris. Les pointillés représentent les valeurs absolues pour l'hypothèse de transfert aléatoire (voir annexe 4.4).

Il existe 4 couples de sites de liaisons pour lesquels on observe des transferts d'un site à l'autre entre deux conformations successives : DP → Xe4, Xe4 → Xe2, Xe2 → Xe1 et Xe2 → Xe3. On peut donc définir des différences de jeux de cavités moyennes entre deux sites de liaisons, en faisant simplement la différence de l'une par rapport à l'autre. J'appelle ces différences de jeux de

cavités moyennes des *jeux de cavités de transfert*. Elles représentent le changement de forme des cavités internes lorsque le CO passe d'un site à un autre. Ces jeux de cavités de transfert sont représentés dans la figure III.14.a, et révèlent des variations notables de volume de certaines cavités lorsque le CO change de site. Globalement, le volume des sites liant le CO semble diminuer (resp. augmenter) lorsque le CO part du site (resp. arrive dans le site). On observe en outre plusieurs effets de changement de localisation de certaines cavités, en particulier pour le site de fixation Xe3 qui semble bouger en fonction de la localisation du CO. La projection des cavités de transfert sur les composantes principales (figure III.14.b) indique une forte ressemblance des cavités de transfert Xe4 → Xe2, Xe2 → Xe1 et Xe2 → Xe3 avec la 3^{e} composante principale. De même, la cavité de transfert Xe2 → Xe1 est très proche de la 1re composante principale, indiquant que ce transfert est une composante majeure de l'évolution des cavités. On observe également que les projections de toutes les cavités de transfert sur les 10-15 premières composantes principales ont des amplitudes significatives par rapport à l'hypothèse nulle de transfert aléatoire (voir annexe 4.4 pour une définition précise). Cela implique que ces transferts spécifiques sont particulièrement importants pour la dynamique globale des cavités de la myoglobine. Ce résultat est renforcé par la part des cavités de transfert dans la variance globale des cavités (tableau V)

Site 1	Site 2	Norme de la cavité de transfert	fraction du RMS de la déviation totale (%)
DP	Xe4	7.1 (3.1)	35.0
Xe4	Xe2	12.8 (3.9)	63.5
Xe2	Xe1	13.9 (4.6)	69.1
Xe2	Xe3	15.5 (8.9)	76.7

Tableau V – Amplitude du mouvement décrit par les cavités de transfert. La norme des cavités de transfert est indiquée - voir annexe, section 4.2 pour l'interprétation de ce nombre. La colonne de droite donne la fraction de cette norme sur la racine de la déviation carrée moyenne de la trajectoire des cavités. La norme des cavités de transfert aléatoire (voir annexe, section 4.4) est indiquée entre parenthèse.

3.8 Evolution locale contre évolution globale

L'ensemble des résultats présentés jusqu'ici, à l'exception de la sélection de cavités consensus, a été réalisé sur l'ensemble des cavités de chaque système. Il n'était pas clair a priori que ces types de résultats s'appliquent à l'analyse d'une cavité particulière, notamment pour la comparaison des composantes principales des trajectoires de structures et de cavités. Pour voir si ce résultat à l'échelle globale était transposable à l'échelle d'un site, j'ai extrait à l'aide du suivi des cavités (décrit au chapitre précédent) trois cavités isolées : la cavité β-OG de la Dengue (voir section 3.4), une cavité dite *inter-domaine* entre les deux monomères de la protéine E de la Dengue (déjà observée figure III.11), et la cavité *enzymatique* de EF. Les poches de chacune de ces cavités ont été définies comme l'ensemble des résidus s'approchant à moins de 5Å dans au moins une conformation au cours de la trajectoire. Deux trajectoires de cavités ont été considérées pour

chacune de ces poches :

— une trajectoire de cavités "non réalignées", correspondant aux cavités déjà calculées sur l'ensemble du système et pour lequel l'alignement structural avait été fait sur la structure complète

— une trajectoire de cavités "réalignées", pour lesquelles les structures ont été réalignées en minimisant le RMSD de la poche seule, et les cavités recalculées sur ces nouvelles trajectoires.

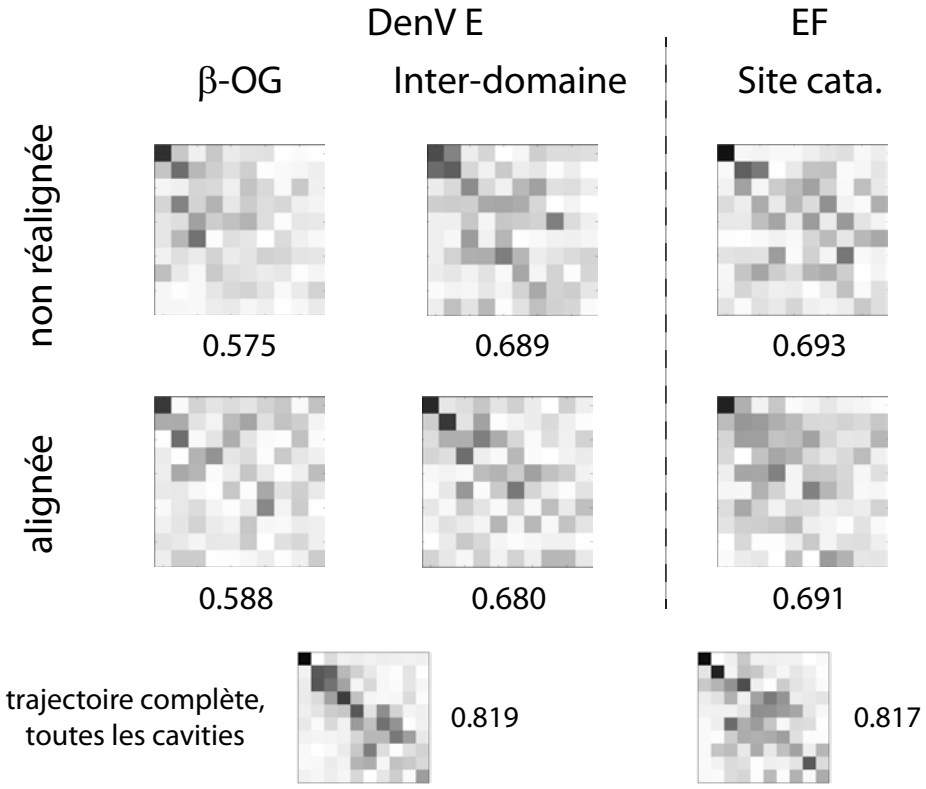


Figure III.15 – Comparaison des composantes principales de cavités isolées, et effet du réalignement. Matrices de produit scalaire (taille 10×10) et valeurs de RMSIP correspondantes pour les composantes principales de trois cavités isolées (abscisse) et des poches correspondantes (ordonnée). La rangée notée "alignée" correspond aux cavités calculées sur les structures réalignées sur la poche. La rangée "non réalignée" correspond aux cavités calculées sur les structures alignées sur le système complet, comme pour la figure III.10. Les matrices des produits scalaires et les valeurs de RMSIP pour le système complet (toutes les cavités et la structure complète) sont rappelées pour comparaison.

La figure III.15 présente les résultats de l'analyse des matrices de produits scalaires et de RMSIP déjà entreprise dans la section 3.5. On observe que les résultats précédents se transposent assez bien à l'analyse de cavités uniques, malgré une baisse du RMSIP, et ce quelque soit la nature de l'alignement (système complet ou poche seule). La valeur de RMSIP pour une cavité unique est par ailleurs relativement proche des valeurs observées pour le lysozyme et la myoglobine, ce qui renforce la tendance observée précédemment qui semble indiquer que le RMSIP augmente avec la taille du système étudié.

4 Discussion

Dans ce chapitre, j'ai analysé l'évolution géométrique des cavités des trajectoires de dynamique de quatre protéines. La première découverte de cette étude est l'extrême variabilité des cavités, qui évoluent de façon très importante au cours d'une DM, jusqu'à couvrir une grande partie du volume de la protéine (tableau II). A ce titre, l'ACP permet de décomposer cette évolution chaotique en modes d'évolution facilement interprétables (figure III.4). En réalisant le calcul dans l'espace des pas de temps, il est aussi possible de comparer les composantes principales des coordonnées atomiques et des cavités, ce qui m'a permis de découvrir une corrélation forte entre les deux (figure III.10). Cette similitude dans l'évolution suggère que les cavités et les coordonnées atomiques peuvent être utilisées en parallèle à des fins de suivi, de manipulation et de sélection de structures. Cela m'a permis d'identifier des conformations ayant des formes de cavités spécifiques, décrites par les premières composantes principales (figures III.11 et III.12), et même de construire de nouvelles structures ayant des formes de cavités proches de la forme cible (figure III.12).

En plus d'aider à la sélection de cavités de formes spécifiques, l'ACP est un outil puissant pour faciliter l'utilisation de la dynamique des cavités, en permettant la compression de leurs descripteurs tout en conservant un maximum d'informations. En appliquant cette compression j'ai pu utiliser des algorithmes de partitionnement sur les cavités, ce qui a rendu possible la sélection de conformations représentatives des différents groupes de géométries de cavités trouvées dans une dynamique (figure III.9). L'application de l'ACP aux cavités peut également compléter l'analyse de la fonction d'une protéine, comme en témoigne la découverte des relations fortes entre la dynamique des cavités internes de la myoglobine et la diffusion du CO en son sein (figure III.14 et tableau V).

La PCA sur les cavités apparait donc comme un outil puissant, autant pour améliorer la pertinence des étapes de criblage virtuel dans les projets de conception de médicament, que pour analyser la fonction de protéines sous un angle nouveau.

4.1 Les limites et conditions d'utilisation de la méthode

L'application de l'ACP sur les cavités présente quelques limitations, liées à la nature des descripteurs de cavités. Tout d'abord, les cavités étant définies sur des grilles dont la position est fixe dans l'espace, leur analyse (et particulièrement l'ACP) peut être très sensible à l'alignement du système. Des conformations mal alignées ou des domaines très mobiles peuvent ainsi produire de larges variances peu pertinentes, ce qui impacte directement la géométrie et les contributions des premières composantes principales des cavités. Toutefois, il est intéressant de noter que cette limite potentielle dans l'utilisation des cavités n'a pas vraiment été rencontrée lors des travaux présentés dans ce chapitre. Les cavités locales ont même pu être analysées sans avoir à refaire un alignement et une détection de cavités, comme le montrent les résultats de la section 3.8.

La représentation des cavités sous forme de grilles, centrale dans cette approche, peut également poser des problèmes de gestion de mémoire. En effet, pour des trajectoires de grosses

protéines de plusieurs dizaines de milliers de conformations, la réprésentation des cavités peut mobiliser plusieurs gigaoctets de mémoire, ce qui demande des machines relativement puissantes (selon les standards actuels).

Enfin, l'ACP est formalisée dans un espace de valeurs réelles, ce qui pose problème lors de son application sur des grilles booléennes discrétisant l'espace comme c'est le cas pour les cavités. Cela provoque des "fuites", diminuant la contribution des composantes de haut rang et augmentant celle des autres composantes. Malgré cela, la forte corrélation entre les composantes des cavités et des structures et la pertinence des reconstructions de conformations à l'aide de composantes de cavités indiquent que cette limitation théorique n'a que peu d'impact sur les applications pratiques présentées dans ce chapitre.

4.2 L'ACP sur les cavités comme méthode d'analyse de la fonction des protéines

L'évolution des cavités au cours du temps est manifestement très dynamique. Il est notamment fréquent d'observer des cavités de toutes tailles apparaitre ou disparaitre. Derrière ce comportement en apparence aléatoire se cachent des tendances et des variations globales qui n'ont pu être identifiées qu'à l'aide d'une méthode du type de l'ACP. Cette méthode ouvre donc de nouvelles possibilités d'inspection de la dynamique d'une protéine, via la description de l'évolution de ses cavités. De plus, l'utilisation des calculs dans l'espace des pas de temps permet de révéler le lien fort entre la géométrie des cavités et l'état fonctionnel de la protéine. Ainsi, et malgré le fait que les cavités peuvent apparaître comme une représentation déformée, floutée et donc peu informative de la structure d'une protéine, j'ai pu faire émerger des liens forts entre les modes d'évolution des cavités internes de la myoglobine et la diffusion du CO de site en site. Cela renforce la pertinence de ce nouvel outil, venant compléter l'analyse traditionnelle de la dynamique des protéines dans les nombreux cas où la dynamique des cavités peut potentiellement jouer un rôle important dans la fonction. La famille des cytochromes p450 est un exemple pour lequel une telle analyse pourrait aider à mieux comprendre les voies et les mécanismes liés aux entrées et sorties des ligands et des molécules d'eau du site actif vers le solvant[123, 124, 125]. De même, ce type d'analyse appliquée à la famille des perméases pourrait donner de nouvelles informations sur les mécanismes de transport et de reconnaissance[203, 204, 205]. Enfin, il peut être intéressant d'utiliser l'ACP sur les cavités pour étudier les relations entre les différentes cavités d'un système afin de détecter de potentiels sites allostériques, notamment dans le cas de protéines dont les mouvements sont directement impliqués dans la fonction[112].

4.3 Les liens forts entre structure et cavités devraient ouvrir la voie à de nouvelles opportunités pour la conception rationelle de médicaments

La comparaison des composantes principales des structures et des cavités m'a permis de dévoiler la corrélation forte entre leurs évolutions. En ajoutant la possibilité d'interchanger les composantes principales de structures et les cavités dans l'espace des pas de temps, j'ai pu montrer qu'il est possible de construire de nouvelles conformations dont les cavités sont très proches d'une géométrie donnée. Ce point encourageant doit être nuancé par le fait que de telles structures, résultant d'une reconstruction linéaire autour d'une structure moyenne, sont généralement assez déformées. Elles doivent donc être vérifiées et éventuellement corrigées avant d'être effectivement utilisées dans un projet de criblage virtuel.

Je pense que la méthode de construction de structures présentée ici a de bonnes chances d'être un atout dans la conception de médicaments, au vu des exemples encourageants d'utilisations de structures construites réalisées récemment (pressurisation [74], fumigation [73], SCARE[72]...). L'utilisation de l'analyse de la dynamique des cavités a l'avantage d'utiliser des informations extraites de conformations déjà échantillonnées afin de sélectionner les conformations ou de les construire si besoin. Elle pointe vers les zones flexibles des récepteurs, et permet également de construire des conformations ayant des particularités intéressantes pour le criblage, tout en s'appuyant sur la dynamique réelle des cavités et non sur des contraintes introduisant des biais. Elle permet ainsi de sélectionner ou de construire des conformations ayant des cavités de volumes et de géométries diverses, pour lesquelles de petites cavités sont fusionnées en une cavité consensus, ou pointant vers des acides aminés spécifiques. De fait, cette procédure semble prometteuse pour améliorer la pertinence et la diversité des composés sélectionnés dans un projet de conception de médicaments.

5 Conclusions

L'analyse en composantes principales (ACP) appliquée aux cavités, malgré quelques limites techniques, est un outil puissant et robuste, qui devrait ouvrir de nouvelles opportunités de visualisation et d'exploration de la dynamique des cavités. Cette technique permet de décomposer et de classifier l'évolution dynamique de la géométrie des cavités au sein des protéines, ce qui est d'autant plus utile au vu du comportement très volatile et tumultueux des cavités. Elle permet de dévoiler et d'analyser des mécanismes fonctionnels subtiles impliquant les cavités, comme le prouve mon étude sur le rôle de l'évolution des cavités internes dans la diffusion du CO au sein de la myoglobine. Je pense que cet outil peut également aider à améliorer la pertinence des étapes de criblage virtuel au cours d'un projet de conception rationnelle de médicaments. Cet outil permet en effet la sélection et la reconstruction de conformations basées sur la connaissance de la

dynamique et de la flexibilité des cavités, ce qui devrait permettre de "dérisquer" les projets de conception de médicaments en augmentant la diversité des composés actifs.

Chapitre IV

Application des méthodes ciblant les cavités à des projets de conception rationnelle de médicaments

1 L'ADN gyrase de la tuberculose

1.1 La tuberculose

1.1.1 Pathologie

La tuberculose est une maladie due à un bacille, *Mycobacterium tuberculosis* (*Mtb*), caractérisée par de la fièvre, des frissons, une fatigue générale et une perte d'appétit. Dans la forme pulmonaire de la maladie (90% des cas), on observe des douleurs toraciques ainsi qu'une toux récurrente, parfois accompagnée par une hémoptysie ("tousser du sang")[206]. Les principaux facteurs augmentant le risque d'infection par *Mtb* sont la surpopulation et la malnutrition, ce qui en fait un marqueur du niveau de pauvreté. La tuberculose touche plus de 9 millions de personnes par an ; en 2013, 1,5 millions de personnes en sont mortes[206]. Cela fait de la tuberculose la 2^{e} cause de décès par maladie infectieuse, juste après le sida[207]. La lutte contre la tuberculose est donc un enjeu majeur de santé mondial et de nombreux plans ont été développés afin de développer de nouvelles thérapies et d'améliorer l'accès aux soins dans les zones de faibles revenus.

La tuberculose est traitée par des cocktails d'antibiotiques. Le traitement standard fait intervenir la streptomycine, l'isoniazide, la rifampicine, la pyrazinamide, et l'ethambutol[206]. Lorsque la souche traitée est résistante - on parle de souche MDR (Multidrug-resitant) ou XDR (extensively drug-resistant), une deuxième série d'antibiotiques peut être utilisée : des aminoglycosides, des fluoroquinolones, des thioamides... La résistance des souches de *Mtb* aux antibiotiques de première ligne est un problème de santé majeur : on estime à environ 500 000 cas par an le nombre d'infections par une souche multirésistante (MDR) dont 9% sont même ultrarésistantes (XDR)[206]. Il est donc très important de déterminer de nouvelles familles de composés efficaces contre la tuberculose.

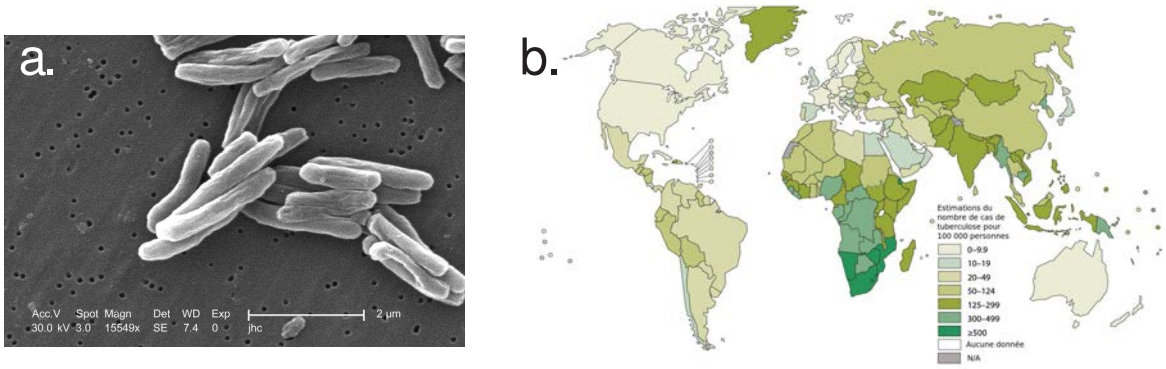


Figure IV.1 – La tuberculose. a. Bacilles de la tuberculose observés au microscope électronique, grossissement 15549x (source : Center for Disease Control). **b.** Prévalence des cas de tuberculose dans le monde en 2013 (source : OMS[206]).

1.1.2 Stratégie d'inhibition et particularité de la cible

Pour enrouler son génome, *M. tuberculosis* utilise une protéine appelée l'*ADN gyrase*. Il s'agit d'une topoisomérase de type II, qui hydrolise l'ATP pour cliver un segment d'ADN et passer un autre segment à travers l'espace créé, avant de resouder le brin clivé[208]. L'ADN gyrase est la seule topoisomérase permettant de surenrouler l'ADN ; c'est aussi la seule topoisomérase de *M. tuberculosis* qui ne possède pas de topoisomérase de type IV. Elle est ainsi nécessaire à la survie des bactéries, et est donc ciblée par un certain nombre d'antibiotiques[209], comme les fluoroquinolones (ciprofloxacine) ou les aminocoumarines (novobiocine).

La structure complète de l'ADN gyrase n'est pas connue avec précision, mais les structures de certains de ses domaines chez différentes espèces sont connues, notamment le domaine de coupure-ligation[210, 211, 212]. Ce domaine est central dans le modèle de fonctionnement à 2 portes[213, 214] (figure IV.2) :

1. un segment d'ADN (segment G) est fixé et clivé en haut du domaine de coupure-ligation
2. le domaine s'ouvre dans sa partie supérieure (la porte ADN)
3. l'autre segment d'ADN (segment T) passe dans l'espace vide central du domaine
4. la porte ADN se referme et le segment G est reformé par ligation
5. le domaine s'ouvre dans sa partie inférieure (porte C) et laisse passer le segment T
6. la porte C se referme

Bloquer une de ces étapes permettrait d'empêcher les fonctions de surenroulement et de déroulement au niveau de la fourche de réplication, toutes deux nécessaires à la survie de *Mtb*. Il serait même possible de stabiliser le complexe ADN-gyrase permettant de rendre l'ADN "inutilisable". L'équipe de Claudine Mayer a pu résoudre la structure du domaine coupure-ligation de la gyrase de *Mtb* dont les deux portes sont fermées (structure *fermée*). Les structures des formes *porte ADN ouverte* et *porte C ouverte* proviennent de modèle par homologie réalisé par le groupe de C. Mayer à partir des structures du domaine coupure-ligation d'un homologue de la gyrase, la topoisomérase IV[215, 210].

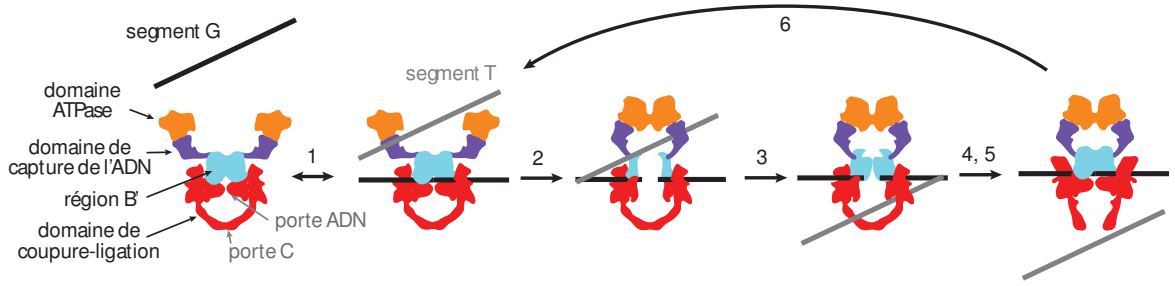


Figure IV.2 – Modèle de fonctionnement de l'ADN gyrase. Les étapes reprennent la numérotation donnée en page 98. Schémas repris de Champoux 2001[208].

Le but de ce projet est de déterminer un modèle à l'échelle atomique du mécanisme du domaine catalytique à l'aide de l'approche *POE* (voir section 2.4.2). Ce modèle, couplé à l'analyse dynamique des cavités (voires chapitres II et III), devrait servir à terme pour définir un site cible afin de déterminer des inhibiteurs potentiels de l'ADN gyrase. Les particularités de ce projet sont la distance assez élevée entre les structures initiales et finales (RMSD : 17.4 Å et 7.99 Å respectivement), la production de deux chemins pour modéliser un unique mécanisme et l'utilisation de structures issues de la modélisation par homologie.

1.2 Modèle du mécanisme du domaine catalytique

1.2.1 Calcul du chemin de transition

J'ai tout d'abord réalisé le chemin de transition du domaine de coupure-ligation (domaine rouge sur la figure IV.2). On considèrera ici qu'il n'y a pas d'hystérésis prononcée et que les étapes d'ouvertures (étapes 1-2 et 5-6) et de fermetures (étapes 3-4 et 7) sont plus ou moins symétriques l'une de l'autre dans le temps. On peut donc calculer deux chemins de transition du domaine de coupure ligation pour modéliser le mécanisme :

— le chemin entre la forme *fermée* et la forme *porte ADN ouverte*, qui modélise les étapes 1 à 4 du mécanisme (chemin **1**)

— le chemin entre la forme *fermée* et la forme *porte C ouverte*, qui modélise les étapes 5 à 7 du mécanisme (chemin **2**)

La forme *fermée* a été minimisée en plaçant des contraintes l'empêchant de trop s'éloigner de la structure cristallographique. Les structures basées sur l'homologie ont été légèrement retravaillées pour corriger certains problèmes structuraux. Les monomères des structures ouvertes ont été alignés sur les monomères correspondants de la structure *fermée*. Les dimères produits ont ensuite été minimisées en contraignant les angles de Ramachandran et la distance à la structure *fermée*. Une étape de recuit simulé a également été réalisée en solvant implicite (ACE) pour relacher au maximum les structures ouvertes. Une contrainte additionnelle a été ajoutée à la structure *porte C ouverte* pour élargir l'espacement au niveau de la porte, initialement trop petit pour laisser passer un segment d'ADN.

Le chemin de transition a été calculé à l'aide de l'approche *POE*, en utilisant un chemin initial linéaire (interpolation linéaire entre la structure *fermée* et l'une des structures ouvertes) et le modèle d'éléctrostatique ACE[93]. Seules deux itérations de *POE* ont été nécessaires pour réaliser chacun des chemins et les raffiner à un seuil d'énergie satisfaisant. Le tableau I résume les caractéristiques des chemins finaux.

Chemin	RMSD (Å)	Long curv. (Å)	Ratio	# structures intermédiaires
fermée → *porte ADN ouverte* (**1**)	17.4	22.9	1.32	28
fermée → *porte C ouverte* (**2**)	7.99	9.15	1.15	13

Tableau I – Caractéristiques des chemins de transition produits dans cette section. La colonne RMSD correspond au RMSD entre la structure initiale (structure *fermée*) et la structure finale (une des deux structures ouvertes). La deuxième colonne correspond à la longueur curvilinéaire le long du chemin de transition (somme des RMSD des structures consécutives). La troisième colonne correspond au ratio longueur curvilinéaire / RMSD. Plus cette valeur est élevée, plus le chemin est complexe. La dernière colonne correspond au nombre de structures intermédiaires entre les structures initiales et finales (incluses).

1.2.2 Suivi des cavités

Les cavités ont été calculées à l'aide de `mkgrid` et suivies en utilisant les algorithmes développés au chapitre II. Les empreintes utilisées sont booléennes, comparées par la distance de Jaccard. Le partitionnement utilisé lors du suivi est un partitionnement hiérarchique de type UPGMA. Les cavités fusionnées n'ont pas été découpées.

Une série de cavités ont été identifiées (figure IV.3), toutes situées entre le "coude" formé par les deux superhélices ("coiled-coil") et la porte ADN. Lors de la transition, ces cavités voient leur volume augmenter. Par ailleurs, seules trois d'entre elles existent dans la strucure *fermée*. Ces cavités semblent particulièrement intéressantes d'un point de vue mécanistique ; en effet, la fixation d'un ligand au niveau du coude rendrait impossible le mouvement d'ouverture de la porte C.

On notera également la tendance à l'augmentation de volume des cavités entre la structure *fermée* et les structures ouvertes. Cette augmentation est due à une accumulation de défauts d'empilement (*packing defects*) qui est un indicateur d'un problème dans la modélisation des structures[216]. La décision a donc été prise de retravailler les modèles par homologie pour obtenir des structures plus compactes.

1.3 Conclusions et perspectives

L'ADN gyrase de *Mycobacterium tuberculosis* est une cible intéressante pour le développement d'un médicament contre la tuberculose. Le mécanisme de fonctionnement du domaine coupure-ligation est modélisable à l'aide de deux chemins de transition, qui ont été réalisés à l'aide de l'approche *POE*. Les chemins se sont avérés très faciles à optimiser, avec des ratios longueur curvilinéaire sur RMSD de 1.32 et 1.15, ce qui est très faible (en comparaison, le chemin de

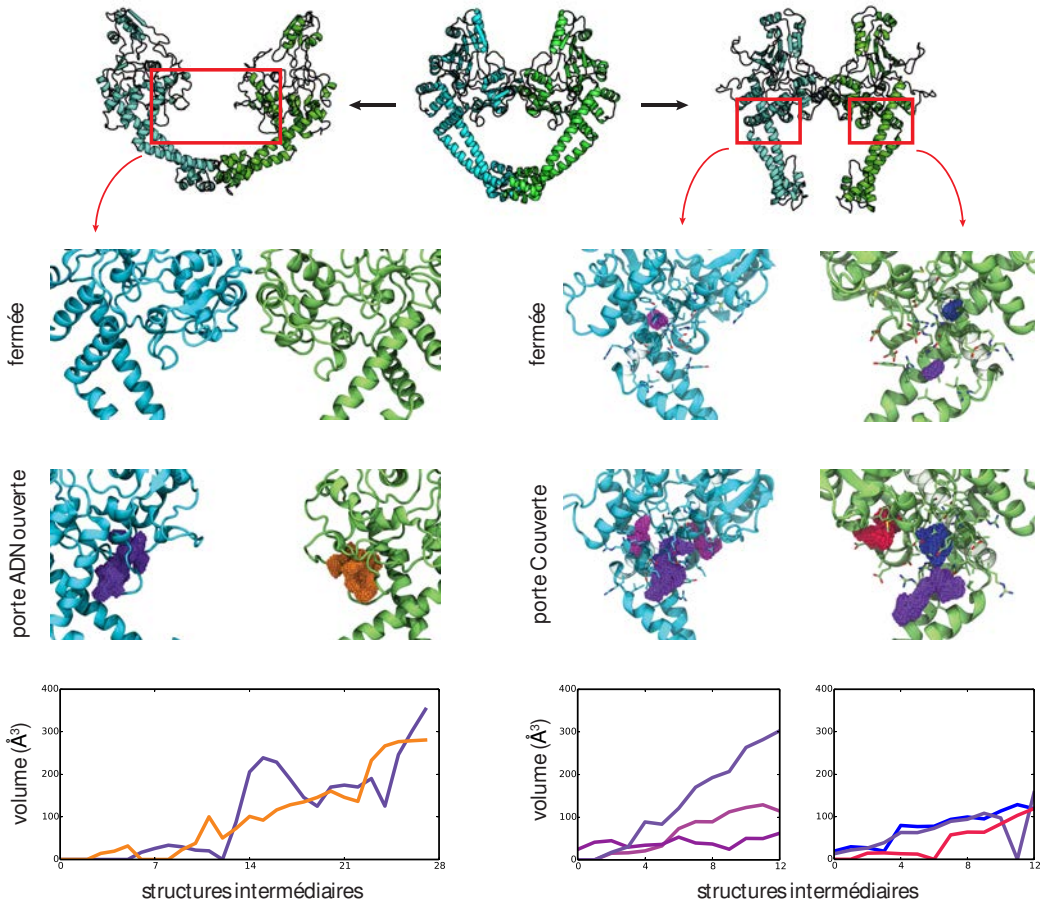


FIGURE IV.3 – Evolution des cavités intéressantes de la gyrase lors des transitions. A gauche : évolution de deux cavités lors de l'ouverte de la porte ADN (en haut : pas de cavité dans la structure *fermée*, au milieu : cavités présentes dans la structure totalement ouverte, en bas : volumes des cavités au cours de la transition). A droite : idem pour l'ouverture de la porte C. La zone zoomée est indiquée par des cadres rouges sur les structures ouvertes.

transition de la toxine de l'anthrax est de 3.2). Le suivi des cavités a permis d'identifier des sites potentiellement intéressants, dans lesquelles une petite molécule pourrait bloquer le mouvement.

Pour l'instant malheureusement, l'absence de structure cristallographique pour les formes ouvertes oblige à utiliser des structures issues de la modélisation. Ces structures ont des cavités de grande taille indiquant de probables lacunes dans la modélisation de la structure, ce qui rend le criblage risqué. Il est donc préférable d'affiner les modèles par homologie existants en attendant d'obtenir de nouvelles structures.

2 La subtilisine I des agents du paludisme

2.1 Contexte et mécanisme d'infection de la malaria

2.1.1 Pathologie et implications socio-économiques

La malaria, ou paludisme, est une maladie caractérisée par des accès cycliques de fièvre, accompagnées par un état de fatigue, une anémie, des pertes d'appétit, vertiges, céphalées, ... Les formes graves, comme le neuro-paludisme ou l'anémie sévère, peuvent être fatales. La maladie, due à une famille de parasites appartenant au genre *Plasmodium* (*P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale* et *P. knowlesi*), est transmise lors des piqûres par les moustiques femelles du genre *Anophèle* (figure IV.4.a). La maladie est localisée dans les régions tropicales d'Afrique, d'Asie et d'Amérique du Sud, dans les zones où le moustique - le vecteur - et des individus affectés - le réservoir - sont présents (figure IV.4.b). Malgré les importants efforts récents de contrôle, près de la moitié de la population mondiale demeure exposée au paludisme et l'on recense chaque année plus de 200 millions de cas cliniques pour plus de 600 000 décès[217].

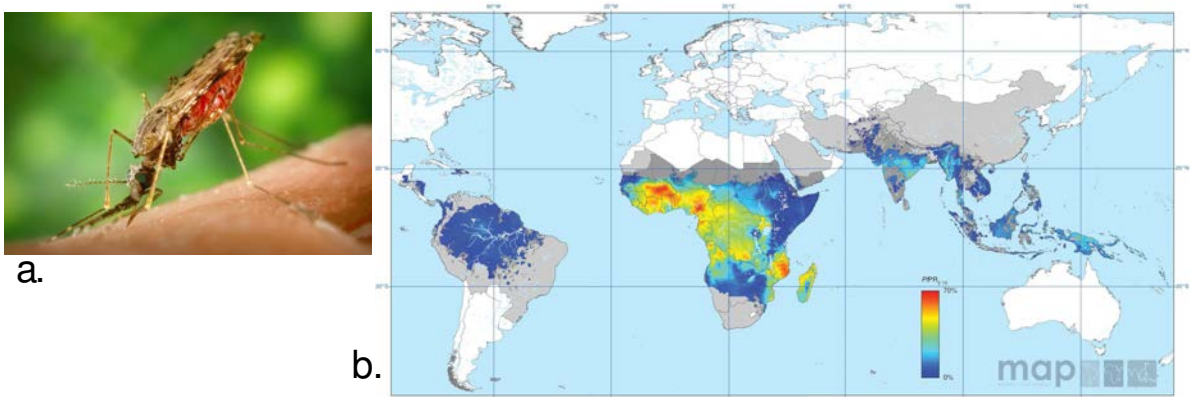


Figure IV.4 – a. Moustique anophèle (source : Center for Disease Control and Prevention[218]) **b.** Prévalence de *P. falciparum* chez les 2-10 ans dans le monde en 2010 (source : Gething *et al.* 2011[219])

Comme la plupart des agents infectieux, *Plasmodium* a la capacité de développer rapidement des résistances aux médicaments utilisés. Ainsi, bien qu'il existe plusieurs médicaments originellement efficaces contre la malaria, plusieurs d'entre eux sont déconseillés par l'OMS du fait des résistances[217]. Cette apparition de résistances est la raison pour laquelle seule les traitements basés sur l'artémisinine (*ACT*, Artemisinin-based Combined based Therapy) sont recommandés par l'OMS, malgré l'existence de plusieurs autres médicaments antipaludéens (chloroquine, quinine, sulfadoxine-pyriméthamine, méfloquine, amodiaquine, doxycycline, artémisinine). Des financements importants, issus en particulier du "Fond Mondial contre le SIDA, la tuberculose et le Paludisme" (émanation du G7) ou d'organisations caritatives, comme la Fondation "Bill et Melinda Gates", ainsi qu'une maîtrise du coût du traitement (moins de 1$ par personne) permettent aux ACT d'être aujourd'hui utilisés en première intention dans 79 des 88 pays où le paludisme est endémique[217]. Cependant, la sélection de parasites résistants à l'artémisinine et

ses dérivés[220, 221, 222, 223, 224] relance la nécessité de trouver de nouvelles molécules antipaludéennes.

2.1.2 Mécanisme d'infection

Le mécanisme d'infection des parasites du genre *Plasmodium* est commun à toutes les espèces, nonobstant quelques subtilités. Le cycle de vie du parasite se décompose en deux étapes majeures, chez l'insecte et chez l'humain (figure IV.5). Le moustique est infecté lorsqu'il se nourrit du sang d'une personne infectée par *Plasmodium*. Les gamétocytes mâles et femelles du parasite fusionnent pour former un zygote, lequel après la méiose se différencie en ookinète, qui après avoir traversé la paroie intestinale du moustique se divise pour former de nombreux sporozoïtes. Ces sporozoïtes rejoignent les glandes salivaires du moustique d'où ils seront injectés lors d'un nouveau repas de sang infectant ensuite l'Homme.

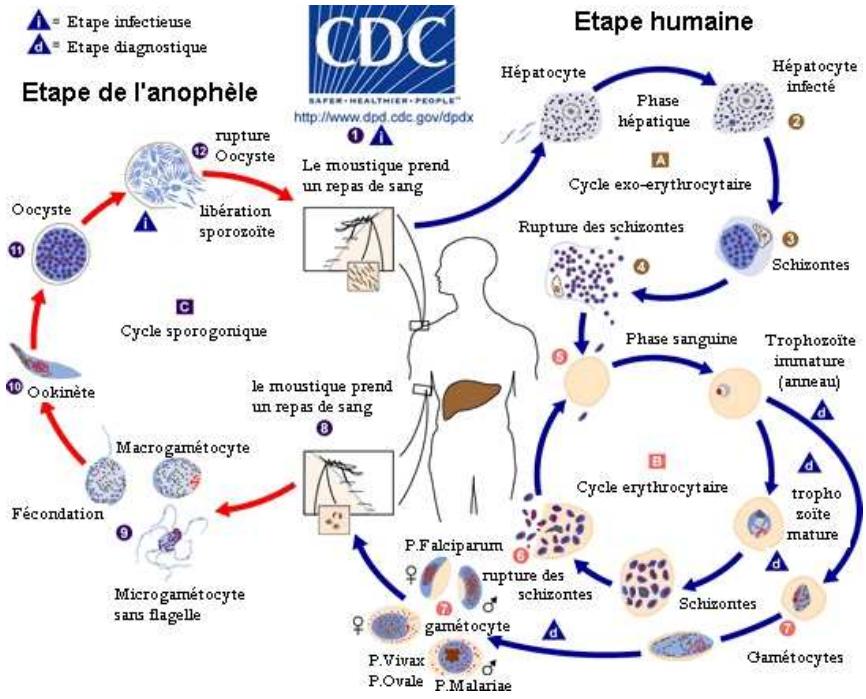


Figure IV.5 – Cycle de vie de *Plasmodium* (source : CDC[218])

Déposés dans la peau, les sporozoïtes gagnent le foie via la circulation sanguine et envahissent des hépatocytes dans lesquels ils se différencient en schizontes, formant des milliers de mérozoïtes hépatiques. Durant cette phase, l'infection est cliniquement asymptomatique. La rupture des membranes de la vacuole parasitophore (dans laquelle se développe les parasites) et de l'hepatocyte hôte libère de grandes quantités de mérozoïtes, qui vont infecter les érythrocytes (globules rouges). Les parasites maturent et se divisent au sein des érythrocytes, qui finissent par éclater après 48 à 72h selon les espèces (hémolyse). Cette rupture libère de nouveaux mérozoïtes dans le sang, qui infectent alors d'autres globules rouges. Cette étape de libération des mérozoïtes est responsable des accès de fièvre caractéristiques de la maladie. Certains parasites intra-érythrocytaires

se différencient en une forme sexuée, les gamétocytes, prêts à infecter un nouveau moustique.

Trois protéines proches de la famille des subtilisines bactériennes, SUB1, SUB2 et SUB3, sont retrouvées dans l'ensemble des espèces du genre *Plasmodium*. Ces protéases jouent un rôle crucial dans la rupture de la membrane erythrocytaire et le mécanisme d'invasion du parasite[225]. En particulier, SUB1 (subtilisine-like 1) est responsable du clivage de certaines protéines du parasite, notamment les protéases de la famille SERA[21, 23] et les protéines de surface des mérozoïtes, MSP1, MSP6 et MSP7[22], agissant au stade de l'entrée *Plasmodium* dans les globules rouges. La fonction de ces protéines est mal connue, mais il semblerait qu'elles soient responsables de la rupture des membranes[226]. SUB1 est essentielle au cycle de vie sanguin du parasite[23] et s'est avérée être une cible potentiellement intéressante pour la conception d'un médicament antipaludéen[21, 22, 225].

Les structures de SUB1 de *P. vivax* (PvSUB1)[227] et de *P. falciparum* (PfSUB1)[228] ont été résolues en 2014. PvSUB1 est composée d'un domaine catalytique globulaire de 334 résidus et d'une prorégion (située en amont) de 164 résidus. SUB1 est auto-inhibée par sa prorégion, et est activée par un pic de concentration de calcium qui permet de cliver cette prorégion pour activer le domaine catalytique.

2.1.3 Particularités de la cible

Nous ciblons ici le site actif de SUB1, tout en se laissant la possibilité de cribler un site allostérique éventuel. SUB1 est une cible assez complexe d'un point de vue biologique car exprimée à l'intérieur du parasite. La molécule devra passer deux membranes avant d'accéder à sa cible : la membrane érythrocytaire et la membrane du parasite. En outre, SUB1 est une protéase, son substrat naturel est donc un peptide. Ce type de cible a un site actif longiligne favorisant la sélection de molécules flexibles, ayant plus difficilement de hautes affinités. Le site catalytique de SUB1 est globalement assez polaire ce qui est bon pour la spécificité mais moins pour l'affinité. Il y a également une poche hydrophobe risquant de favoriser la sélection de composés peu solubles[229].

2.2 Dynamique moléculaire et détection des cavités

L'unité asymétrique de la structure cristallographique de *Pv*SUB1 comporte deux copies de la protéine (dénotées A et B), comprenant chacune le domaine catalytique et sa prorégion. Deux dynamiques moléculaires de 10 ns ont donc été réalisées (10 000 conformations sauvegardées, paramètres donnés en annexe section 2), une pour chaque copie (trajectoires 1A et 1B). Deux dynamiques supplémentaires ont été réalisées en ne gardant que le domaine catalytique de chacune des copies et en utilisant les mêmes paramètres (trajectoires 2A et 2B). La protéine des trajectoires 1A et 1B a été extraite et les trajectoires ont été concaténées pour former la trajectoire **1**, de même pour les trajectoires 2A et 2B (trajectoire **2**) Le domaine catalytique de la trajectoire **1** a ensuite été concaténé à celui de la trajectoire **2** pour former la trajectoire **2c**. Afin de faciliter l'étude dynamique des cavités (notamment le suivi), seule une conformation sur deux a été gardée dans cette trajectoire concaténée, pour former une trajectoire de 20 000 conformations.

Les cavités ont été détectées à l'aide de `mkgrid` sur les trajectoires **1** et **2c**, puis suivies à l'aide de la méthode développée au chapitre II (empreintes booléennes et partitionnement hiérarchique, pas de division des cavités fusionnées).

2.3 Sélection des cavités et conformations d'intérêt

Le suivi des cavités nous a permis d'étudier en détail l'évolution de chacune des cavités du système selon des critères de volume, de conservation des résidus et de changement de leur forme géométrique. A partir de cette analyse, nous avons déterminé 3 cavités transverses potentiellement intéressantes pour le criblage virtuel (figure IV.6).

1. la cavité catalytique (trajectoire **2c**), pour développer un inhibiteur compétitif
2. une cavité dite "enfouie" (trajectoire **2c**), dont les résidus sont conservés chez *Plasmodium*
3. une cavité dite "côté" (trajectoire **1**) située à la surface de la protéine entière et pouvant impacter son activation

La cavité "côté" a été rapidement abandonnée en raison de son exposition, en effet la forme entière de la protéine (domaine catalytique + prorégion) ne se trouve que durant l'étape de développement du parasite précédent la rupture des érythrocytes et est protégée par une troisième membrane. La cavité "enfouie" a été considérée mais n'a finalement pas été retenue, car de volume trop petit.

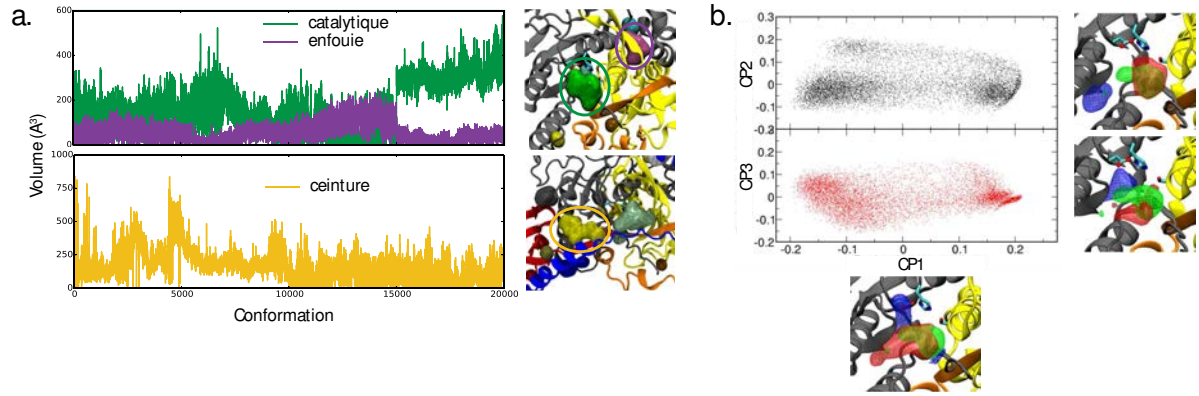


Figure IV.6 – Evolution du volume des trois cavités présélectionnées et de la forme géométrique de la cavité catalytique.

La sélection des conformations du site catalytique pour l'étape de criblage virtuel a été réalisée selon le même schéma qu'expliqué dans le chapitre III, section 3.4. Les résidus situés à moins de 5 Å de la cavité catalytique dans au moins 95% des conformations définissent le site (la poche). L'algorithme des k-moyennes a été utilisé sur la projection de la trajectoire de la cavité catalytique sur ses 100 premiers vecteurs propres pour définir 15 partitions de l'espace géométrique des cavités (figure IV.7). Pour chaque partition, la structure d'énergie électrostatique la plus faible est choisie comme représentante. Ces structures et la définition du site ont été envoyées au groupe Structural Design and Informatics à Sanofi pour réaliser le criblage virtuel.

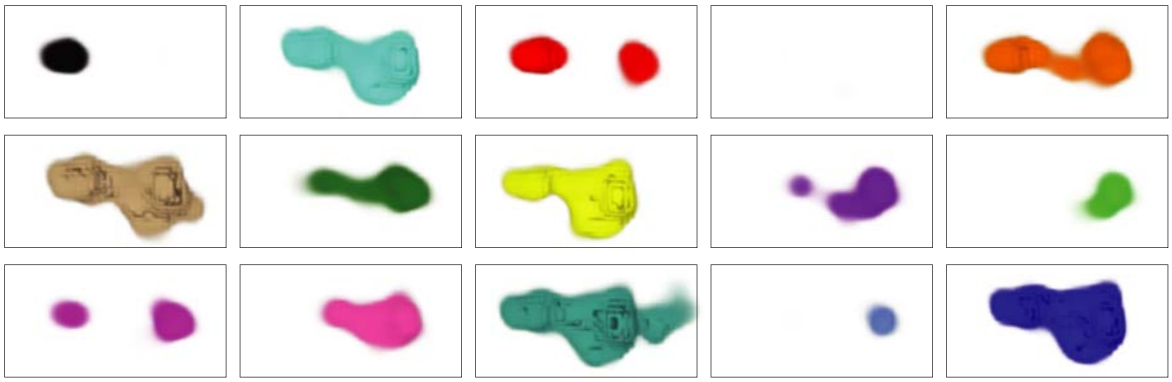


Figure IV.7 – Distibution moyenne des cavités (cavité moyenne) pour chacune des 15 partitions de l'espace des cavités. La partition correspondant à la cavité moyenne semblant vide comporte une multitude de cavités de volume très faible situés à différents endroits.

2.4 Criblage virtuel et test des composés

Le criblage virtuel a été réalisé entièrement à Sanofi à partir d'une chimiothèque propriétaire. Il s'est déroulé en deux temps. La première sélection de composés comportait plusieurs milliers de molécules sélectionnées à partir des résultats de plusieurs criblages. Seuls quelques pourcents des composés sélectionnés provenaient du criblage d'une partie des structures que j'ai modélisées et fournies à notre partenaire. Les tests de cette première sélection, assez stringeants (faible concentration de composés testés), n'ont révélé que très peu de composés actifs. Ces composés ont par la suite été rejetés en raison d'autres propriétés peu satisfaisantes. On a pu toutefois remarquer qu'une des molécules les plus intéressantes était issue du criblage virtuel des structures que nous avons apportées.

Une deuxième sélection a été réalisée à partir du criblage virtuel du reste des structures provenant de nos modèles (7 structures). Les résultats de ce criblage (score des composés pour chaque structure) ainsi qu'un partitionnement chimique des composés (19 partitions) nous ont été transmis sous forme anonymisée afin de réaliser la sélection des composés à tester. La sélection a été réalisée à l'aide d'un algorithme de tri entrelacé (voir section 3.6), utilisant deux classements de composés fondés sur le score :

— un classement *diversité* qui range les composés selon le score le plus favorable parmi les poses sur les 7 structures

— un classement *consensus* qui range principalement les composés selon le nombre de structures pour lesquelles le docking a réussi, puis secondairement sur la moyenne des scores

Les composés sont sélectionnés alternativement parmi ces deux listes en ignorant un composé si sa partition chimique est déjà surreprésentée. Les 1000 composés de meilleur rang selon cette méthode ont été choisis pour être testés. Les tests sont toujours en cours, les résultats préliminaires semblant un peu plus concluants que précédemment.

2.5 Conclusion et perspectives

Ce projet aura été l'occasion pour moi de faire l'expérience d'une collaboration avec une équipe de Sanofi, acteur majeur de l'industrie pharmaceutique. J'ai pu y découvrir un environnement complètement tourné vers la production de résultats, très efficace mais peu enclin à développer de nouvelles méthodes selon le mode académique. J'ai donc appliqué les méthodes développées au cours de ma thèse et décrites dans les chapitres II et III pour la sélection des composés inhibiteurs de SUB1, protéase essentielle dans le cycle de vie et d'infection des parasites responsables du paludisme. Les résultats du premier criblage n'ont malheureusement pas été très concluants. Un point positif est qu'une des rares molécules actives provenait du criblage virtuel sur une des structures que j'ai apportées. Les résultats du second criblage virtuel basé sur la structure n'étant pas encore disponibles pour le moment, nous ne pouvons pas encore tirer de conclusion sur l'apport de l'utilisation des données de cavité pour le criblage virtuel dans ce projet.

3 GLIC, récepteur ionotrope pentamérique analogue des récepteurs $GABA_A$ humains

3.1 Mecanisme et intérêt médical

3.1.1 Fonctionnement des récepteurs ionotropes et intérêt pharmaceutique

Les récepteurs ionotropes pentamériques humains sont des protéines localisées au niveau de la membrane des synapses. Ces récepteurs lient des neurotransmetteurs, ce qui ouvre leur canal entre le milieu extérieur et le cytoplasme du neurone. L'ouverture de ce canal permet le passage sélectif d'un type d'ions ce qui génère un courant électrique. Ainsi, ce sont ces récepteurs qui convertissent le signal chimique délivré par les neurotransmetteurs en signal électrique utilisé par les neurones pour transmettre une information. Les cinq sous-unités formant chaque récepteur ont des séquences variées, et peuvent s'assembler en homo ou hétéropentamères. Les récepteurs ionotropes sont en général spécifiques d'un seul type d'ion et d'une seule famille de neurotransmetteurs. Les principales familles de récepteurs ionotropes pentamériques sont :

— les récepteurs nicotiniques de l'acetylcholine (nAChRs), se liant comme son nom l'indique à la fois à l'acetylcholine et à la nicotine (sélectifs des cations Na^+, K^+ et parfois Ca^{2+}).

— les récepteurs $GABA_A$, se liant au neurotransmetteur GABA (sélectifs de Cl^-)

— le récepteur sérotoninergique 5-HT_3, liant la sérotonine (Na^+, K^+, Ca^{2+})

— les récepteurs GlyR, liant la glycine et d'autres petits acides aminés (Cl^-)

Les fonctions de ces récepteurs sont multiples et très variées, et dépendent largement de la composition des sous-unités. Ces récepteurs régulent un grand nombre de phénomènes : mouvements musculaires, mémoire, anxiété, apprentissage, nausée, addiction... Ainsi, les récepteurs

$GABA_A$ sont responsables de la plupart des phénomènes inhibiteurs du système nerveux central, tandis que certains récepteurs nicotiniques de l'acetylcholine sont responsables de certaines sensations de "récompense", mais aussi de la transmission des mouvements musculaires volontaires. Ce sont donc des cibles de choix pour l'industrie pharmaceutique. Malheureusement, les récepteurs ionotropes pentamériques restent des cibles difficiles à étudier de par leur nature et leur localisation (difficulté de production, molécule transmembranaire difficile à cristalliser, mode de fonctionnement complexe et subtile, effets de "haut niveau" sur le comportement et la santé complexes à modéliser).

3.1.2 GLIC : structure et mouvements

Les travaux de Aravind et collab.[230] ont permis de déterminer un homologue bactérien du récepteur $GABA_A$ humain, nommé GLIC (Gloeobacter Ligand-gated Ion Channel). Cet homologue a la particularité d'être activé par les protons, donc à pH acide. Plus récemment, Corringer et collab. ont déterminé une structure ouverte[231] et localement ouverte de GLIC en pH acide[232] (figure IV.8, gauche), puis une structure fermée à pH neutre[233] (figure IV.8, droite). L'ensemble de ces structures permettent d'établir des hypothèses sur le fonctionnement de GLIC et plus particulèrement sur les mouvements du pentamère lors de l'ouverture du canal. La fermeture du canal s'accompagne ainsi de deux mouvements généraux de l'ensemble de la protéine : un mouvement de torsion[234, 233] et un mouvement dit d'*éclosion* du domaine extracellulaire[233]. Cette éclosion s'accompagne d'ailleurs d'une ouverture du domaine extracellulaire au niveau de la jointure avec le domaine transmembranaire (figure IV.9.a).

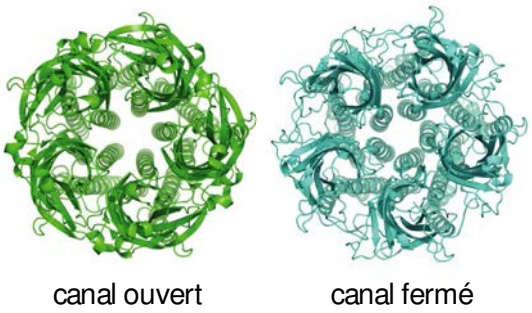


Figure IV.8 – Vue de dessus des structures ouvertes (gauche) et fermées (droite) de GLIC.

3.1.3 Stratégie de recherche d'effecteurs

L'ouverture de cette partie du domaine extracellulaire lors de la fermeture du canal fait apparaitre un réseau de plusieurs cavités à l'intérieur de cette zone (figure IV.9.b) La stabilisation de ce réseau de cavités devrait donc stabiliser la forme fermée du récepteur et empécher son ouverture. Une molécule se liant aux résidus constituant ces cavités pourrait ainsi être inhibitrice de GLIC. L'objectif de ce projet est donc d'obtenir de nouveaux effecteurs de GLIC, et en particulier des inhibiteurs, afin d'affiner la compréhension de son mécanisme de fonctionnement.

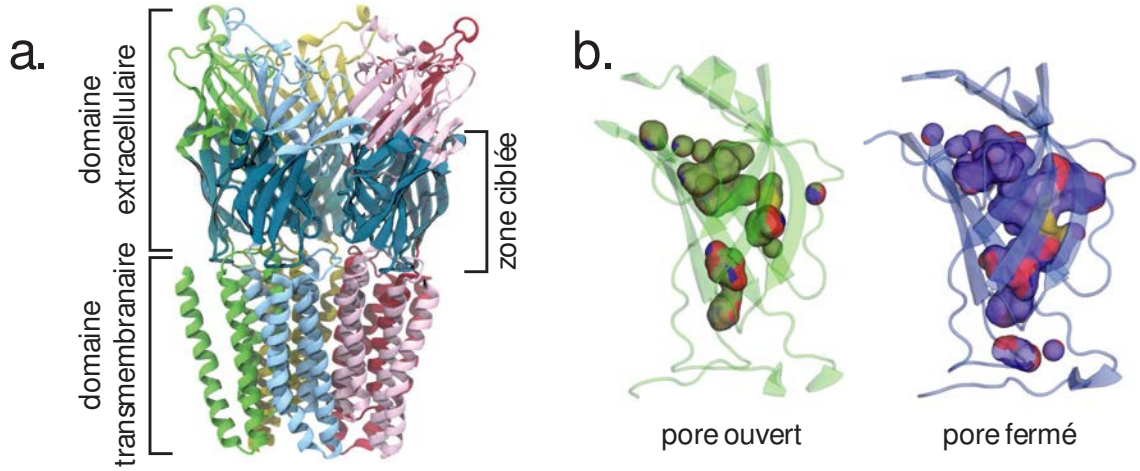


Figure IV.9 – Zone de GLIC ciblée dans ce chapitre. **a.** La zone ciblée, en bleu, est située à la base du domaine extracellulaire, juste au dessus du domaine transmembranaire. **b.** Le volume des cavités de la forme fermée (à droite) augmente par rapport à la forme ouverte (à gauche).

Pour aborder ce projet, nous avons mis en place les étapes et fait les choix suivants. Pour répondre à l'absence de densité électronique des chaînes latérales, nous avons effectué un échantillonnage de ces mêmes chaînes par dynamique moléculaire. Nous avons détecté et analysé la dynamique des cavités du domaine afin de les trier et de sélectionner des conformations pour le criblage virtuel à partir de différents critères (énergie de la structure, volume, diversité). Le criblage est réalisé en utilisant la ZINC[235], une collection de plusieurs chimiothèques commerciales recensant plus de 10 millions de composés a priori disponibles à la commande. Ce choix se justifie par la possibilité de commander rapidement de nombreux composés et par la grande diversité chimique que l'on peut attendre d'une chimiothèque de cette taille.

Du fait des difficultés et incertitudes importantes de ce projet (voir section suivante), les étapes d'affinement suivant l'identification des premières touches sont faites en se concentrant sur les données concrètes provenant des tests. Je me suis donc concentré pour cette étape sur une recherche de composés similaires aux composés actifs (*ligand-based*).

3.1.4 Particularités de la cible

La cible peut être considérée comme une cible risquée et difficile à cribler en raison de la faible résolution (4.35Å) - la définition de la densité électronique est toutefois relativement bonne pour une si faible résolution du fait de la possibilité de moyenner les 4 copies du pentamère de l'unité asymétrique. L'absence de densité électronique pour la majorité des chaînes latérales de la zone ciblée demande une étape supplémentaire de modélisation de leur positionnement, ce qui affaiblie la pertinence du criblage, la position des atomes devenant hypothétique.

Un deuxième facteur de risque provient des tests biologiques qu'il n'est possible d'effectuer qu'en petit nombre. Il s'agit de mesures d'électrophysiologie, c'est-à-dire de la différence de potentiel entre l'intérieur et l'extérieur d'une cellule, en l'occurence un œuf de xénope. Ces mesures ne peuvent pas être automatisées, et doivent donc être réalisées manuellement, œuf par œuf, ce

qui demande beaucoup de temps. Il n'est donc possible de tester qu'environ 30 à 50 composés par série de tests, loin des milliers de composés habituellement testés pour un projet de criblage. Ce faible nombre de composés testés diminue les chances de trouver un composé actif parmi ceux qui sont sélectionnés suite au criblage virtuel.

Dernièrement, la taille de la chimiothèque utilisée (plus de 10 millions de composés) implique d'effectuer un travail de partitionnement chimique en amont afin de limiter la durée du criblage virtuel tout en gardant l'essentiel de la diversité chimique d'origine.

3.2 Echantillonnage des chaînes latérales

Les chaînes latérales ne figurant pas dans la densité électronique ont été reconstruites par estimation d'après les diverses structures de GLIC par L. Sauguet. Une étape d'échantillonnage sous contrainte des chaînes latérales a été réalisée avec CHARMM pour simuler leur flexibilité et obtenir une certaine diversité des cavités. Tous les résidus n'étant pas voisin d'une des cavités ciblées dans les monomère cristallographique (distance minimale >5Å) sont fixés totalement, et donc exclus des calculs d'énergie. Les atomes des chaînes principales des résidus non fixés sont maintenus en place par des contraintes élastiques sur leurs positions d'origine. Enfin, les atomes des chaînes latérales proche des cavités sont laissés libre, à l'exception de certains résidus pour lesquelles les données structurales convergent vers une conformation fixe : K151 (doit pointer vers le solvant), Q124 (structure la boucle Cys), E26, R192 et W160. Les résidus dont les chaînes latérales sont libres sont donc numérotés 24, 27, 30, 37, 39, 114, 126, 128, 155-158, 162, 165, 188 et 190. L'énergie d'une structure est définie comme la somme des énergies des contraintes, de l'énergie de la structure non fixée et de la moitié de l'énergie d'intéraction entre atomes fixés et non fixés.

L'échantillonnage des chaînes latérales se déroule ensuite en trois étapes :

1. les structures cristallographiques de chaque monomère sont minimisées (ce qui donne 20 structures, 5 pour chacun des 4 pentamères de l'unité asymétrique).
2. une DM de 100 000 pas à 3000 K est produite sur chacune des structures minimisées. Pour chaque trajectoire ainsi produite, on minimise une structure sur 1000, ce qui donne 2000 structures supplémentaires.
3. les 10 structures de plus basse énergie de chaque trajectoire sont ensuites utilisées comme point de départ pour un deuxième tour de dynamiques moléculaires produites avec les mêmes paramètres, puis une structure sur 100 est sélectionnée et minimisée.

Ainsi, au total, 22 020 structures ont été produites au cours de l'échantillonnage (figure IV.10).

3.3 Etude dynamique des cavités et sélection des conformations

Les cavités ont été détectées sur les 22 020 structures avec `mkgrid` (rayon de la sonde interne : 1.1Å). Les cavités présentes à l'ouverture du réseau de cavité, au niveau des résidus E35, K33

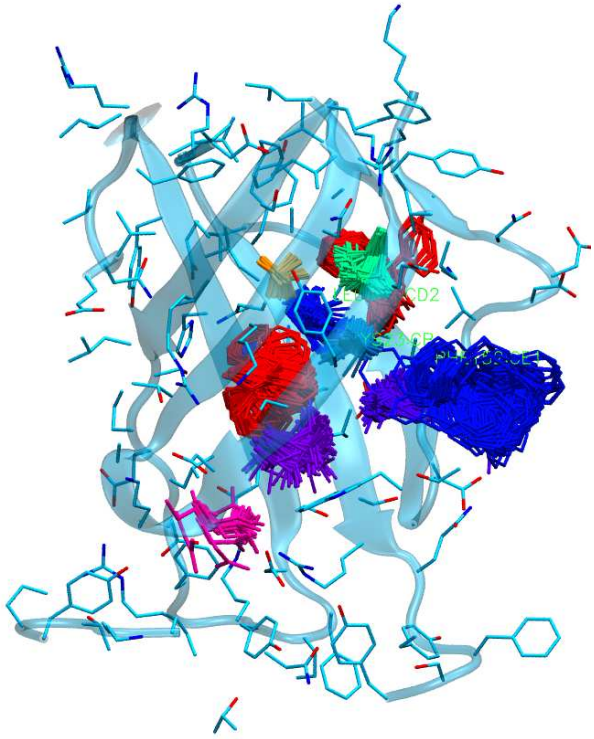


Figure IV.10 – Echantillonnage des chaînes latérales. Ne sont montrées ici que les chaînes latérales pointées directement vers les cavités.

et K248, ont été filtrées car jugées peu intéressantes (pas assez enfouies). Les cavités ont ensuite été analysées par ACP (Analyse en Composantes Principales, voire chapitre III). La projection des cavités sur les 3 premiers vecteurs propres révèle quatre directions d'évolution majeures (figure IV.11.a). Ces quatre directions correspondent à quatre zones géographiques à l'intérieur du domaine ciblé : enfouie en haut du réseau (zone 1), en dessous de PHE37 près de la sortie (2), au dessus de PHE37 (3) et sur le côté, près de ILE162 (4). Lorsque la projection des cavités d'une conformation est située loin de l'origine dans une de ces directions, c'est qu'il existe une cavité de gros volume dans la zone géographique correspondante. Les vecteurs définissant ces directions sont déterminés manuellement, en faisant la différence entre les cavités situées le plus loin possible dans chaque direction et la cavité moyenne. Pour chaque direction, on sélectionne ensuite 100 conformations de projection maximale sur ce vecteur. La conformation représentative choisie pour le criblage est finalement la conformation d'énergie électrostatique minimale parmi ces 100 conformations. Nous avons voulu conserver une structure cristallographique, ce qui nous a poussé à remplacer un de ces quatre modèles par la structure la plus adaptée provenant d'un monomère de la structure cristallographique. J'ai donc choisi la structure correspondant à la chaîne N pour remplacer la cavité sur le côté, car c'est la structure pour laquelle la cavité est la plus volumineuse et la plus archétypale parmi les quatre zones géographiques et les 20 monomères cristallographiques. On obtient ainsi 4 conformations, présentées figure IV.11.b, qui sont représentatives d'une certaine direction d'évolution des cavités, tout en étant les plus favorables possibles d'un point de

vue énergétique.

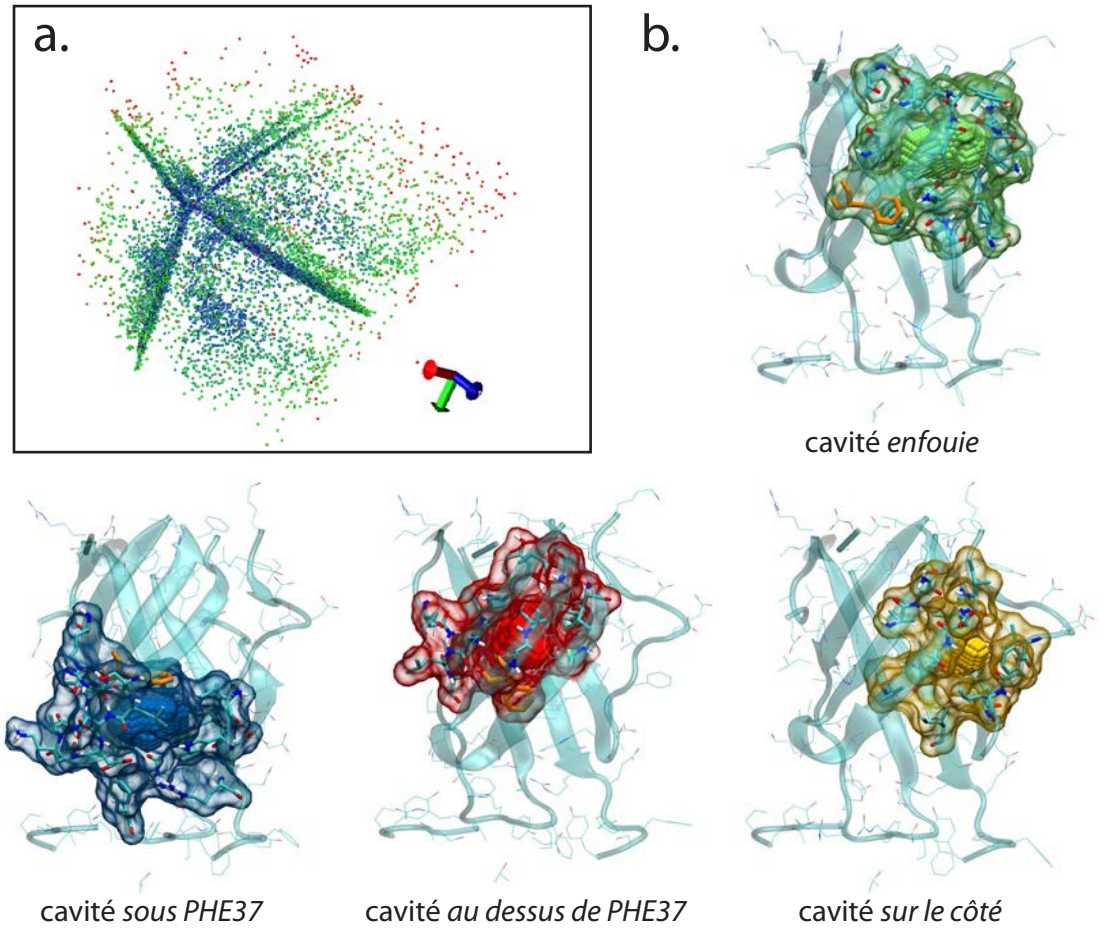


Figure IV.11 – Sélection des conformations utilisées pour le criblage. **a.** Projection de chaque conformation sur les trois premiers vecteurs propres de cavité. La projection en 2D des trois axes est indiqué en bas à gauche. Le code couleur indique le volume de l'ensemble des cavités de chaque structure, en variant du bleu (0 Å^3) au rouge (285 Å^3). **b.** Les quatre conformations représentatives et leurs cavités. Les cavités, les résidus alentours et leur surface accessible au solvant sont représentés en vert pour la cavité enfouie, en bleu pour la cavité sous PHE37, en rouge pour cavité au dessus de PHE37 et en jaune pour la cavité sur le côté. Le résidu PHE37 est représenté en orange.

Pour définir la poche (les résidus entourant la cavité) de chaque représentant, les cavités ont été redétectées en utilisant une sonde interne de 1.1Å, afin de détecter les éventuels canaux très fins situés aux alentours des cavités. Les résidus situés à moins de 5Å sont considérés comme faisant partie de la poche.

3.4 Mise en place d'un échantillonnage de la ZINC

La ZINC[235] est une "méta-chimiothèque" : elle agrège une multitude de chimiothèques commerciales dans un format unifié et donne des informations permettant de commander les composés. La version de la ZINC utilisé dans ce chapitre date de décembre 2013. J'utilise plus précisément le sous-ensemble *instock*, comprenant les composés *a priori* disponibles directement à la commande,

soit 10 125 419 composés. Le criblage complet de la ZINC sur les quatre conformations retenues sur 256 cœurs prendrait plusieurs mois ; pour accélérer le procédé, j'ai réalisé un échantillonnage représentatif de la ZINC, permettant de faire le criblage en deux étapes :

1. les représentants de chaque groupe déterminé par le partitionnement de la ZINC sont criblés sur chaque structure

2. les groupes pour lesquels les représentant semblent intéressants (pose correcte et énergie de docking suffisament basse) sont sélectionnés, et l'ensemble des composés appartenant à ces groupes sont criblés sur les structures correspondantes

De cette manière, le criblage est effectué sur les composés les plus pertinents a priori, en évitant de gaspiller des ressources sur des composés qui n'ont que peu de chance d'être actifs.

3.4.1 Partitionnement de la ZINC à l'aide d'une carte auto-organisatrice

Afin de partitionner une chimiothèque, il est nécessaire de définir une mesure de similarité (ou de distance) entre deux composés. Pour cela, les composés sont traduits en vecteurs de fonctions chimiques, les empreintes de Morgan, très utilisées dans la littérature et proches des empreintes ECFP4 et FCFP4. La sous-structure définie par l'environnement atomique de chaque atome à une et deux liaisons de distance est traduite en un *bit*, une valeur numérique variant de 0 à $2^{32}-1$ (figure IV.12.a). Le compte des bits de tous les atomes définit l'empreinte de Morgan d'un composé. L'empreinte peut être utilisée telle quelle ou compressée par modulo en un vecteur booléen de taille fixe (figure IV.12.b), ici 2048 bits. A noter que cette compression génère des "conflits" (un bit compressé peut représenter une multitude de sous-structures différentes) et perd le décompte des sous-structures. La fonction de similarité utilisée pour comparer deux empreintes non compressées est une extension de la similarité de Jaccard adaptée aux vecteurs de valeurs entières :

$$S(A,B) = \frac{\sum_i min(A_i, B_i)}{\sum_i max(A_i, B_i)}$$

Les empreintes compressées sont elles comparées à l'aide d'une extension de la similarité de Jaccard aux valeurs réelles entre 0 et 1, afin de mesurer la distance au centroïde des nœuds auxquels ils sont affectés.

$$S^{compress}(A,B) = \frac{A \cdot B}{A + B - A \cdot B}$$

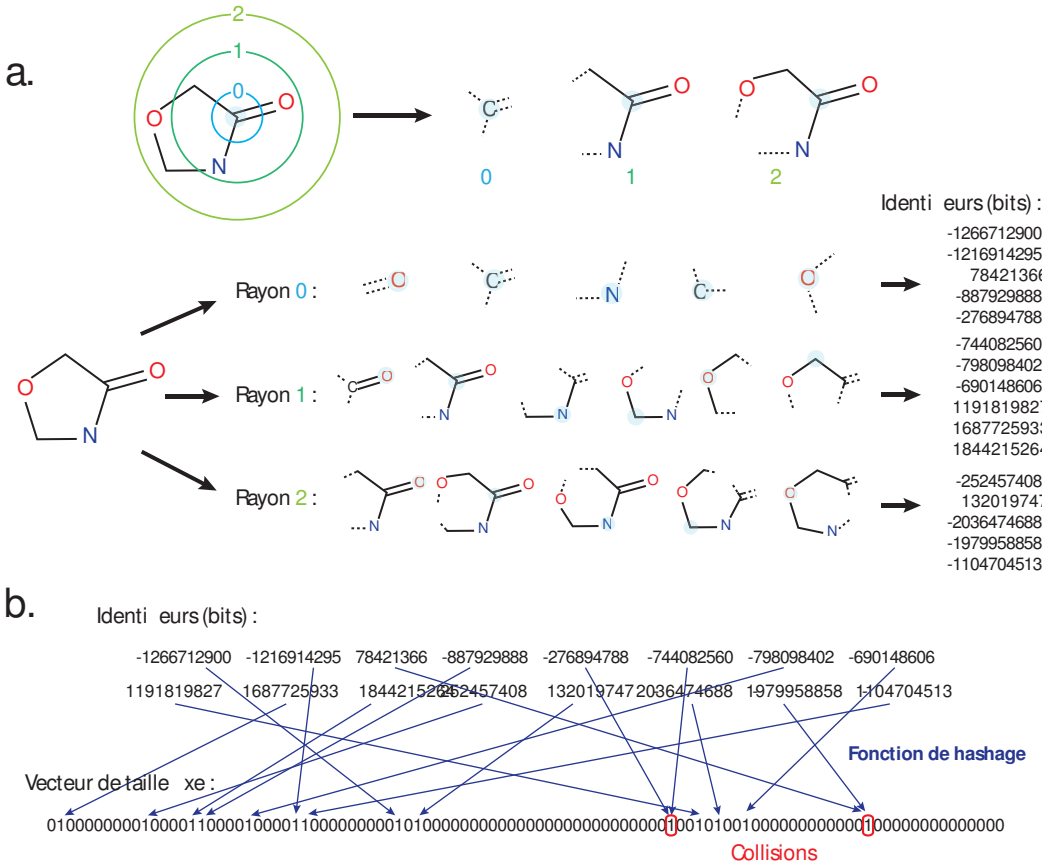


Figure IV.12 – Description des composés par empreintes de Morgan. a. La molécule est décomposée en sous-structures centrées sur chaque atome. Les sous-structures sont associées à différentes valeurs entières (les *bits*). **b.** L'empreinte peut être compressée en un vecteur booléen "hashé".

La première étape de l'échantillonnage de la ZINC consiste à réaliser son partitionnement à l'aide d'une carte auto-organisatrice (l'algorithme est décrit section 3.3). Cet algorithme est en effet très bien adapté aux jeux de donnés massifs de grande dimension, car sa complexité est avantageuse ($O(n \times m \times d)$ avec n le nombre de vecteurs d'entrée, d leur taille, et m le nombre de neurones composant la carte) tout en prenant en compte la densité des données d'entrée. L'ensemble des empreintes compressées des composés de la ZINC est passé deux fois durant la phase d'apprentissage, ce qui génère une carte de 2500 neurones dont 2150 sont effectivement associés à des composés.

3.4.2 Sous-partitionnement de la ZINC et définition des composés représentants

Chaque composé est associé au neurone le plus proche. On appelera également neurone l'ensemble des composés associés à un neurone particulier. La deuxième étape du partitionnement de la ZINC consiste à utiliser l'algorithme des *k-medoïdes* (cf. annexes 3.1) pour partitionner chacun de ces neurones en groupes de composés en utilisant cette fois-ci les empreintes non compressées. Le nombre de groupe pour chaque neurone est proportionnel à la variance de l'ensemble des distances entre les composés associés au neurone : plus un neurone regroupe des composés différents,

plus il sera subdivisés en groupes. Le médoïde de chaque groupe devient alors le représentant de ce groupe. Cette étape de partition permet de définir plus de 700 000 groupes et autant de représentants. Ces représentants sont utilisés pour la première étape du criblage.

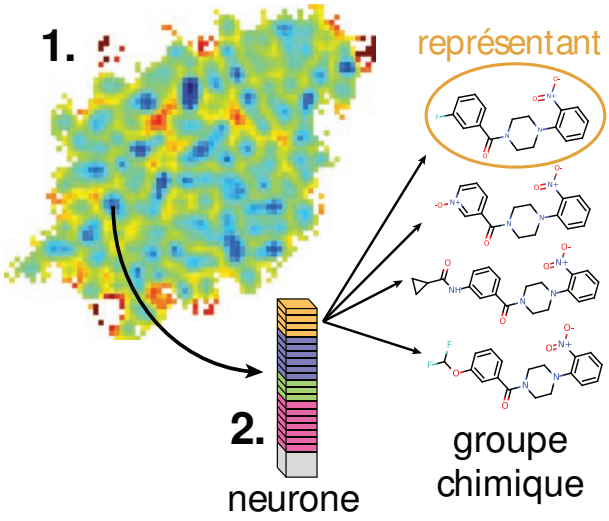


Figure IV.13 – Partitionnement et échantillonnage de la ZINC. La carte SOM est composée d'une collection de neurones (étape 1) entrainés sur les composés de la ZINC. Les composés associés à chaque neurones sont partitionnés à l'aide de l'algorithme des k-médoïdes (étape 2). Le médoïde de chaque partition est considéré comme son composé représentant pour la 1re étape du criblage.

3.5 Criblages virtuels et présélection des composés

Les deux étapes de criblage ont été réalisées à l'aide des logiciels FlexX [60] et DOCK [52, 53] (voir section 2.2). Les poses des ligands provenant de DOCK sont rescorés en utilisant le score MMGBSA. Seule la pose de plus basse énergie est considérée. La première étape du criblage consiste à docker l'ensemble des représentants déterminés lors du sous-partitionnement de la ZINC. L'enfouissement de chaque ligand est déterminé comme suit :

— chaque atome du ligand est discrétisé en considérant les huits sommets du voxel de la grille utilisée pour détecter les cavités

— si l'un de ces sommets correspond à un point de la cavité, l'atome est considéré comme correctement placé à l'intérieur de la cavité

— la fraction du nombre d'atomes considérés comme correctement placés sur le nombre d'atomes du ligand définit le taux d'enfouissement e du ligand ($0 \leq e \leq 1$)

Les diagrammes représentant l'énergie du docking en fonction du taux d'enfouissement de chaque ligand sont donnés dans la figure IV.14. Il est possible de tirer plusieurs conclusions de ce diagramme. Tout d'abord, les poses de FlexX ont plus tendance à sortir du cadre de la cavité d'origine que celles de DOCK. Une des raisons est la réduction des rayons de van der Waals des atomes réalisée par FlexX, permettant de modéliser de façon très crue la flexibilité du récepteur. Cela permet au ligand de se glisser dans des interstices normalement trop petits et donc non compris

dans les cavités telles que calculées par `mkgrid`, donc d'être classé par FlexX alors qu'il est rejeté par DOCK. On remarque également que la cavité *enfouie*, la plus volumineuse des cavités sélectionnées, permet de docker un plus grand nombre de ligand avec une énergie et une pose bien enfouie. On retrouve ce comportement à moindre échelle pour la cavité *au dessus de PHE37*. A l'inverse, les cavité *sous PHE37* et *sur le côté* sont beaucoup plus petites et ne permettent aux logiciels de ne docker correctement que les composés les plus petits. Le cas de la cavité *sur le côté* est particulièrement sensible puisque très peu de poses de bonne énergie et de bon enfouissement sont trouvées. Enfin, on peut relever que les composés ayant pu être dockés à la fois par DOCK et par FlexX ont un meilleur enfouissement et une meilleure énergie.

La sélection des représentants pour la deuxième étape du criblage se fait à l'aide de critères sur l'énergie de docking et le taux d'enfouissement (pointillés colorés dans la figure IV.14). Les valeurs des critères d'énergie et de taux d'enfouissement sont choisies afin d'équilibrer le nombre de composés issus de DOCK et FlexX et de chacune des cavités. Ces valeurs sont données en annexe, section 5.2.

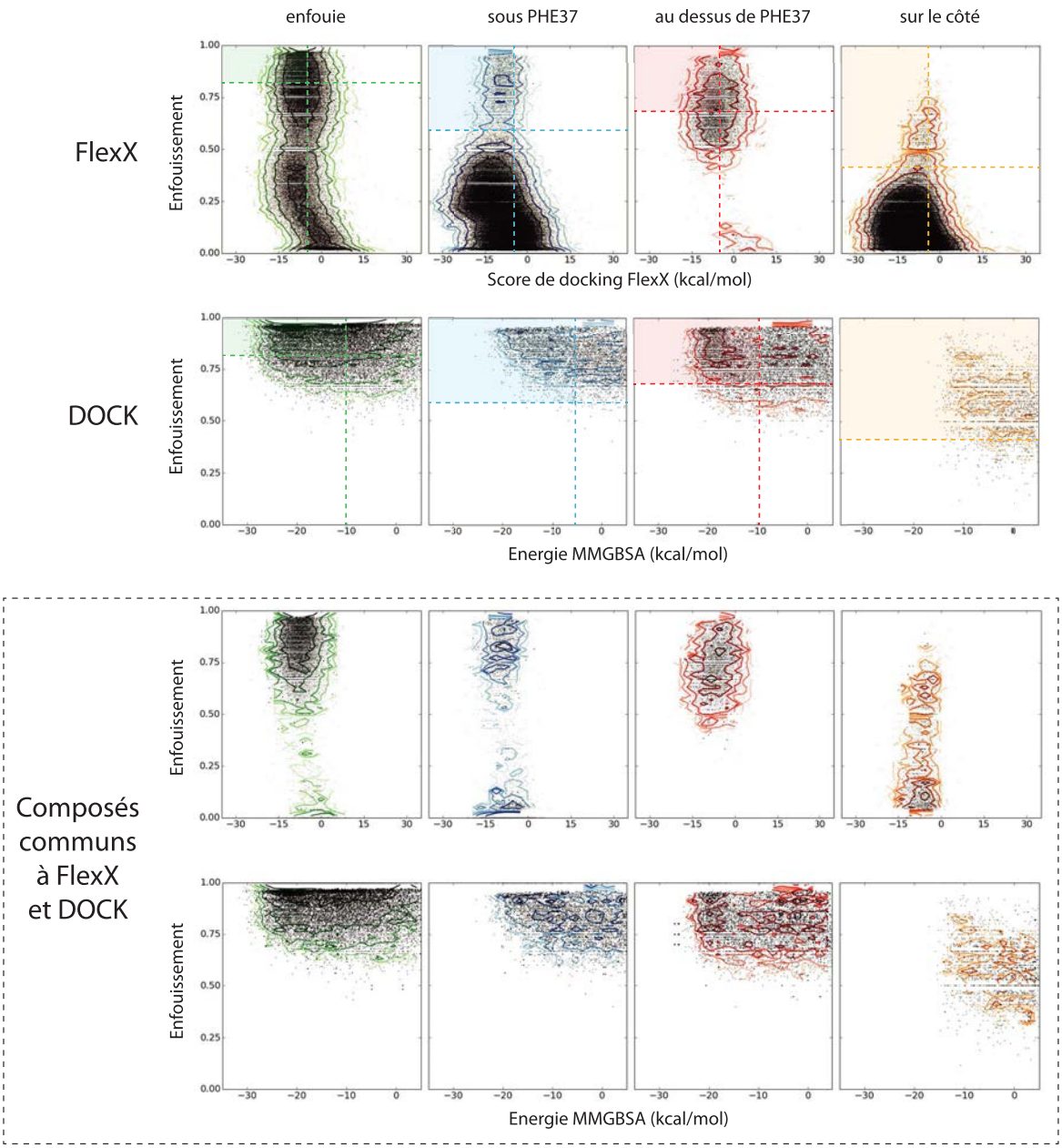


Figure IV.14 – Diagramme score-taux d'enfouissement des poses des représentants après criblage à l'aide de FlexX et DOCK. De gauche à droite : docking sur la cavité enfouie (vert), la cavité sous PHE37 (bleu), la cavité au dessus de PHE37 (rouge) et la cavité sur le côté (jaune orangé). De haut en bas : docking réalisé avec FlexX, docking réalisé avec DOCK, composés effectivement dockés par FlexX et DOCK (en haut : diagrammes des poses de FlexX, en bas : diagramme des poses de DOCK). En abscisse, l'énergie de docking des poses (score FlexX ou énergie MMGBSA) ; en ordonnée, la fraction d'atomes effectivement contenus dans la cavité détectée précédemennt. Les lignes de niveau sont logarithmiques et donnent un aperçu de la "densité" de poses dans le graphe. Les traits pointillés verticaux et horizontaux correspondent aux critères (en énergie et taux d'enfouissement) de sélection des composés. Les composés situés dans le rectangle coloré seront sélectionnés pour l'étape suivante du criblage.

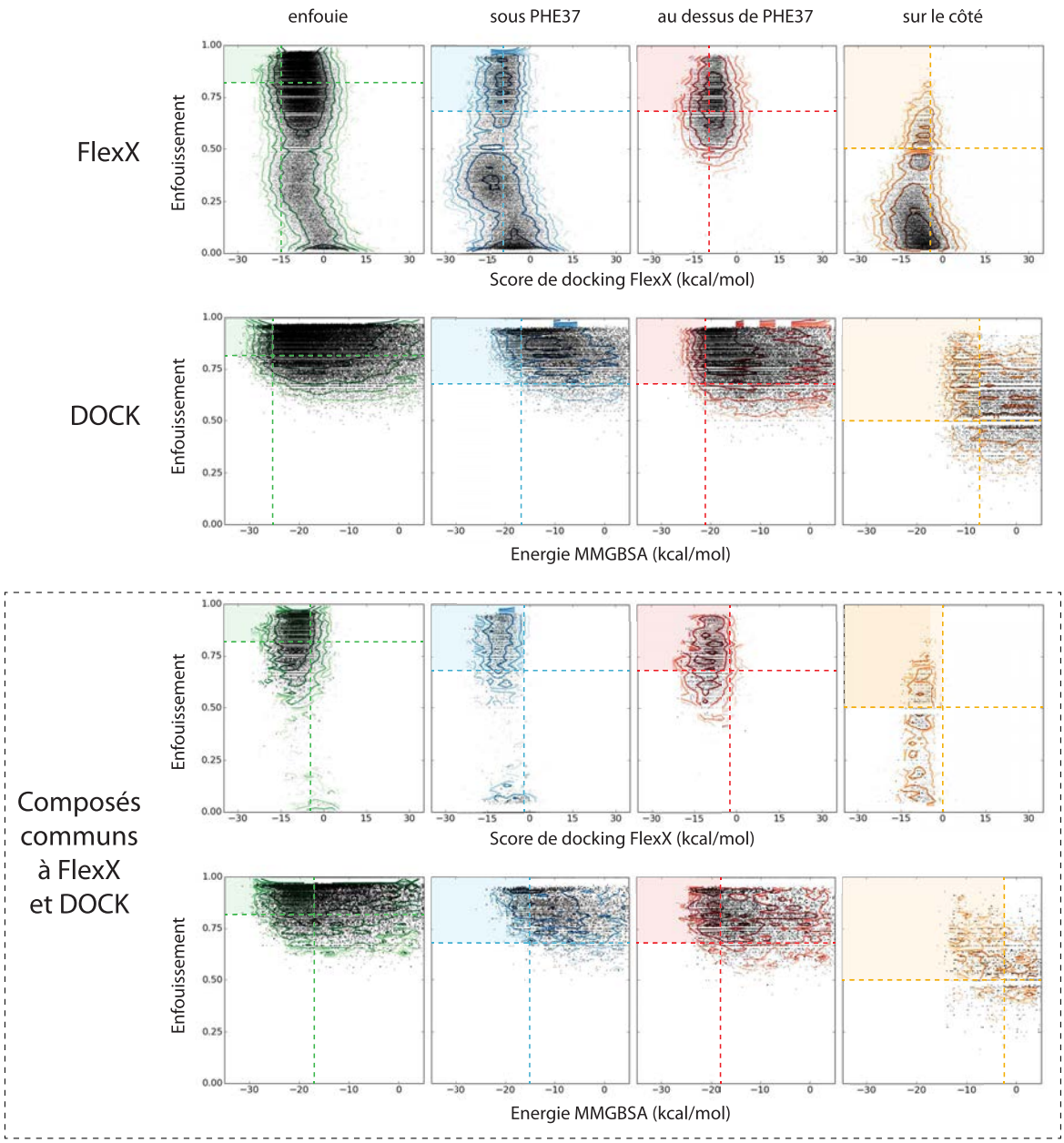


Figure IV.15 – Diagramme score-enfouissement des composés criblés lors du 2e criblage. Voir figure IV.14 pour la description des éléments de cette figure. A noter que dans le cadre "composé communs à FlexX et DOCK" les composés doivent vérifier à la fois les critères pour FlexX et pour DOCK pour être sélectionnés (ils doivent être situés dans les deux carrés à la fois).

Pour chaque cavité, l'ensemble des composés appartenant aux familles des représentants sélectionnés (soit plus de 700 000 composés) sont criblés à l'aide de FlexX et DOCK comme pour le premier criblage. Le diagramme score-enfouissement pour ce criblage est donné figure IV.15. Comparativement au criblage précédent, nous avons remarqué que les composés sont bien plus concentrés dans les zones "intéressantes" (basse énergie-enfouissement élevé). Cette observation justifie *a posteriori* l'utilisation d'un criblage en deux temps, car les composés sélectionnés via les représentants sont effectivement plus susceptibles de docker correctement. Les critères de sélection des composés sont ici plus drastiques afin de réduire le nombre de composés à traiter. Les composés ayant pu être dockés à la fois par DOCK et par FlexX sont sélectionnés en priorité sur des critères d'énergie plus relachés. Ils doivent cependant obéir à l'ensemble des quatre critères pour être sélectionnés de cette façon. A noter que la méthode de classement des composés utilisée ici (voir section suivante) implique que les critères de sélection sur les composés issus uniquement de DOCK ou de FlexX n'ont pas vraiment d'importance. Les composés sélectionnés lors de cette deuxième étape doivent finalement être classés pour définir la liste des composés à commander et à tester.

3.6 Classement des composés présélectionnés

Le classement des composés doit vérifier plusieurs points afin de conserver le plus de chances d'identifier des touches :

1. Les composés doivent avoir la meilleur affinité possible pour le récepteur.
2. Les cavités d'origines de la sélection des composés devraient être représentées relativement uniformément. Le but ici est de limiter le risque d'un biais défavorable issu de la modélisation (poche mal modélisée ou n'ayant pas d'impact sur la fonction).
3. Les composés sélectionnés devraient appartenir à des familles chimiques diverses. En effet, deux composés d'une même famille ont plus de chance d'avoir une affinité apparentée pour une même cible (à quelques ordres de grandeur près toutefois). Sélectionner des composés issus de familles différentes permet donc de multiplier les chances de trouver une famille chimique efficace sur ces poches nouvelles.

Le meilleur indicateur de l'affinité à notre disposition est le score donné par le logiciel de docking. Les scores de DOCK et de FlexX sont différents et ne peuvent donc pas être comparés. De plus, un composé ayant un bon score DOCK n'a ni plus ni moins de chance d'être un bon ligand qu'un composé ayant un bon score FlexX - les deux logiciels doivent donc être traités *a priori* sur un pied d'égalité. On peut aussi considérer qu'un composé docké par deux logiciels a plus de chance d'avoir une bonne affinité pour la cible. Afin de traiter le point n°2, il est également nécessaire de traiter les scores des composés de façon indépendante pour chaque cavité. Je considère donc trois classements de composés pour chacune des quatre cavités, pour un total de 12 listes de composés :

1. Les composés issus de la sélection sur DOCK rangés par score croissant (liste *dock*).

2. Les composés issus de la sélection sur FlexX rangés par score croissant (liste *flexx*).
3. Les composés issus de la sélection sur DOCK et FlexX en même temps, rangés par $min(rang_{DOCK}, rang_{FlexX})$ croissant (liste *dock-flexx*).

Enfin, je traiterai le point n°3 en réalisant le partitionnement des composés en 6 familles chimiques à l'aide de l'algorithme des k-médoïdes, de la même façon qu'expliquée dans la section 3.4.2.

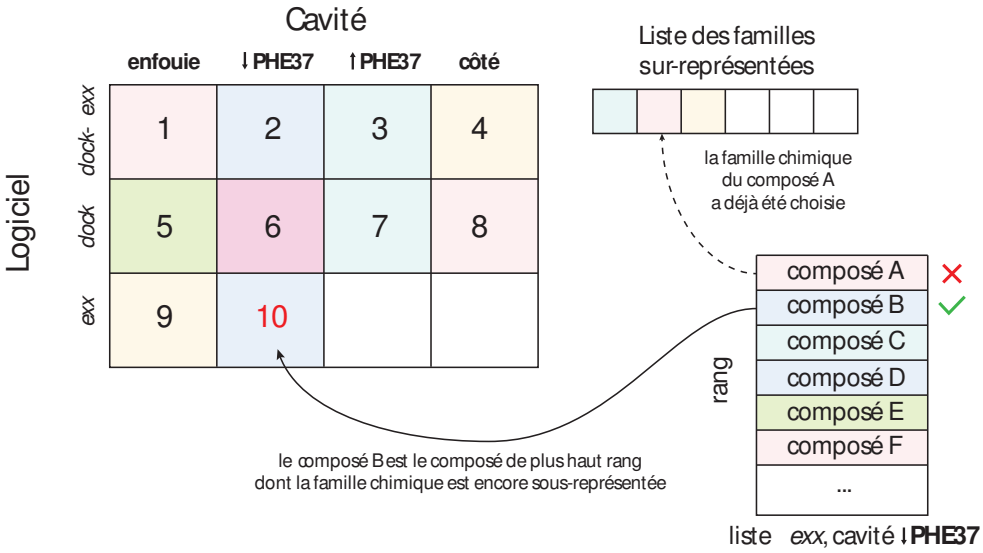


Figure IV.16 – Schéma de l'algorithme de classement entrelacé utilisé pour réaliser le classement global des composés. Le tableau indique le rang global des composés sélectionnés. Le composé en court de sélection est le numéro 10, qui doit provenir de la liste *flexx* de la cavité *sous PHE37*. Les familles chimiques des composés sont notées par des couleurs. Lorsque la famille chimique d'un composé est sur-représentée il ne peut pas être sélectionné, et on essaye alors les suivants jusqu'à l'identification d'un composé d'une famille sous-représentée.

J'ai choisi d'utiliser un algorithme de classement entrelacé afin de combiner les listes en un classement global des composés. Pour cela, les composés sont sélectionnés un par un à partir des 12 listes, à la façon d'un pilulier (voir figure IV.16). Un composé est sélectionné lorsqu'il est en tête de la liste courante, sauf dans le cas où sa famille chimique est déjà strictement surreprésentée. L'algorithme est répété en boucle jusqu'à épuisement de l'ensemble des listes.

3.7 Tests des composés

Les tests des composés sont réalisés par mesure éléctrophysiologique sur des œufs de xénope (voir figure IV.17.a). Les œufs sont placés dans un flux d'une solution tampon à pH neutre, puis à pH acide. Le proton étant un agoniste de GLIC, le passage en pH acide induit l'ouverture des canaux, ce qui crée un courant entre les deux éléctrodes. Ce courant est mesuré tout au long de l'expérience. Le composé est ajouté au flux de tampon et on mesure l'évolution du courant entre le cytoplasme et le milieu extérieur de l'œuf (figure IV.17.b). Ces tests étant particulièrement chronophages, il était nécessaire de ne sélectionner qu'une cinquantaine de composés à tester. Certains composés de la liste n'ont pu être commandés pour des raisons pratiques (stocks épuisés ou composé onéreux). D'autre part, nous avons tenu à ne pas multiplier les fournisseurs, et à éviter de

sélectionner des composés possédant des fonctions chimiques titrables risquant d'interférer avec le fonctionnement pH-dépendant du récepteur. Les meilleurs remplaçants (même catégorie/famille) ont alors été choisis manuellement. Au final, 35 composés ont été reçus et 26 effectivement testés (les composés reçus mais non testés n'étant pas solubles dans le tampon).

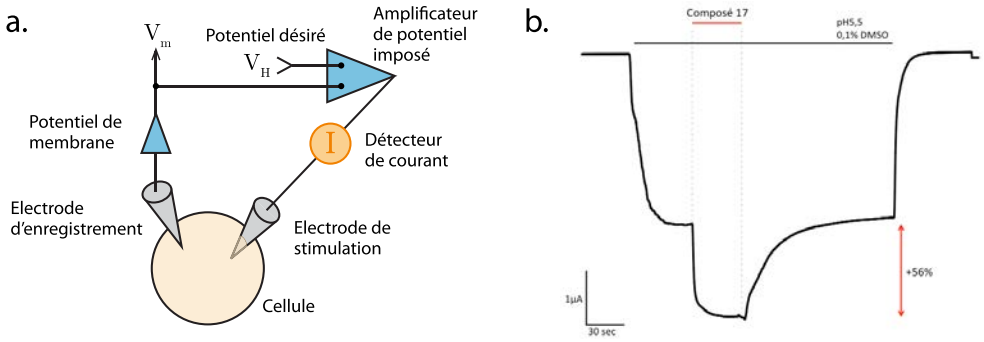


Figure IV.17 – Principe du test d'efficacité des composés : l'éléctrophysiologie. a. Des œufs de xénopes utilisés ont été génétiquement modifiés pour exprimer GLIC à leur surface. Le principe de la mesure est de mesurer le courant entre le cytoplasme d'un œuf et le milieu extérieur induit par la différence de potentiel entre deux éléctrodes. Lorsque les canaux sont fermés, les éléctrons ne peuvent pas passer, le courant est donc quasi nul. Lorsque les canaux sont ouverts, les éléctrons passent par les canaux, ce qui induit un courant. La mesure du courant permet donc de déterminer dans quel état se trouve les canaux. **b.** Exemple de mesure d'un composé potentiateur. On se place en pH acide pour favoriser l'ouverture du canal. Lorsque l'on ajoute un composé potentiateur, les canaux s'ouvrent plus ou plus souvent et le courant augmente (zone encadrée par des pointillés). A l'inverse, lorsque l'on ajoute un composé inhibiteur, les canaux se ferment et le courant diminue. On peut ainsi mesurer directement l'efficacité d'un potentiateur ou d'un inhibiteur en mesurant la variation du courant au moment de l'ajout d'un composé (double flèche rouge).

Les tests ont été réalisés au laboratoire de Pierre-Jean Corringer par Anaïs Menny à l'Institut Pasteur. parmi les molécules testées, 4 se sont avérées avoir une efficacité tangible, soit en tant que potentiateurs (tableau II, molécules A, B), soit en tant qu'inhibiteur (molécule C, D). Parmi les autres molécules, certaines ont un effet faible, d'autres aucun effet mesurable (voir annexe 5.3). Les quatre composés les plus actifs sont toutefois d'une efficacité limitée, les concentrations utilisées lors des tests étant de l'ordre de 30 à 100 µM.

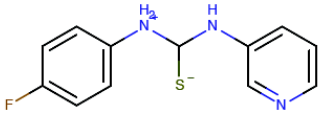
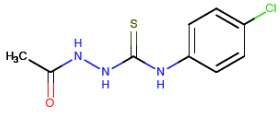
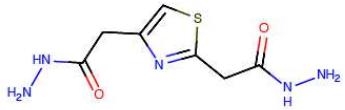
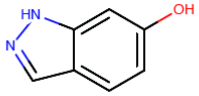
ID	Composé	Cavité ciblée	Concentration testée	Effet sur la fonction	n
A		1	340µM 34µM	+35,8% +17%	5 2
B		1	500µM 50µM	+22,1% < +2%	3 1+
C		1	1mM 100µM	-16,4% -5.4%	3 1+
D		2	1mM	-16%	1+

Tableau II – Efficacité des composés actifs. La colonne n correspond au nombre de répétition de l'expérience. Les numéros de cavités ciblées correspondent respectivement à : 1. cavité *enfouie*, 2. cavité *sous PHE37*.

3.8 Sélection de composés similaires aux molécules actives

Afin d'obtenir des composés plus efficaces, j'ai réalisé une dernière étape de criblage sur la ZINC, basée entièrement sur la similarité chimique avec les molécules les plus actives. Les composés de la ZINC sont ordonnés en fonction de leur similarité avec chacun des quatre composés actifs (voir section 3.4.1). Un filtre sur la structure chimique est également appliqué afin de limiter la recherche aux molécules qui ne différent que par un ou deux groupes. Ces filtres utilisent les SMARTS (`http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`) et des critères de nombre d'atomes pour limiter la sélection aux molécules les plus pertinentes. Les molécules finales ont été choisies par Pierre-Jean Corringer et Anaïs Menny afin de séléctionner en priorité des composés ayant des propriétés intéressantes pour d'éventuelles études cristallographiques (dispersion anomale). La liste de ces molécules est donnée en annexe, figure VI.9. Ces composés ont été commandés et devraient être testés peu de temps après l'écriture de ce manuscrit.

3.9 Conclusions et perspectives

Ce projet consistait à déterminer des effecteurs de GLIC ciblant un sous-domaine du domaine extracellulaire localisé juste au dessus du domaine transmembranaire. Ce sous-domaine voit son réseau de cavité interne grossir lors de la fermeture du canal du fait du phénomène global "d'éclosion" du domaine extracellulaire. Ce projet comportait de nombreuses incertitudes et risques (des

chaînes latérales manquantes mais surtout un très faible nombre de composés que l'on pouvait tester), rendant *a priori* assez hypothétique l'identification d'effecteurs, même à haute concentration. L'utilisation de l'analyse dynamique de ce réseau de cavités et d'un criblage virtuel en deux étapes basé sur une hiérarchisation de la ZINC nous a pourtant permis d'identifier quatre effecteurs, deux potentiateurs et deux inhibiteurs, sur seulement 26 composés testés. La nature même de ces effecteurs est toutefois intriguante : les sites visés, présents uniquement dans la structure fermée de GLIC, auraient du favoriser largement la découverte d'inhibiteurs. La sélection d'analogues des quatre composés effecteurs portant des fonctions chimiques intéressantes pour les études cristallographiques devrait nous permettre, si ceux-ci sont suffisament efficaces, d'obtenir de nouvelles connaissances sur le mécanisme d'action de GLIC et de ses effecteurs.

Chapitre V

Conclusions

Durant ma thèse, j'ai développé un certain nombre de concepts permettant de formaliser, appréhender et analyser la dynamique des cavités, dans un but fondamental, mais aussi pour aider à la découvertes de composés effecteurs de différentes protéines. J'ai eu l'occasion de les appliquer à des projets de découvertes, ce qui m'a permis de les adapteur aux cas pratiques. J'ai en particulier développé deux méthodes novatrices permettant d'étudier précisément, de façon systématique et relativement facilement la dynamique des cavités au sein des protéines.

J'ai ainsi développé les méthodes de suivi des cavités, développant des concepts introduits par ailleurs par Eyrisch et Helms en 2007[189, 166]. En identifiant les cavités non pas en fonction de leur position spatiale, mais par les éléments structuraux situés à proximité, il est en effet possible de s'affranchir des problèmes dûs aux déplacement rendant l'étude des cavités complexe. Les empreintes que forment ces éléments autour des cavités peuvent alors être facilement regroupées à l'aide d'algorithmes de partitionnement. Divers types d'empreintes, distances et algorithmes de partitionnement ont été testés et validés à l'aide d'indicateurs qualitatifs et quantitatifs. Les résultats de ces tests indiquent que les algorithmes et paramètres que j'ai avancé sont plus pertinentes et plus flexibles que la méthode proposée par Eyrisch et Helms. Ceci est notamment dû au fait que j'ai souhaité dès le début avoir une approche permettant de s'accomoder d'événements de fusion et de divisions éventuelles des cavités suivies. Une étape supplémentaire optionnelle a été ajoutée, la détection et le découpage des cavités que l'on peut considérer comme fusionnées. Cette étape permet d'améliorer dans certains cas les performances du suivi des cavités. Malgré ces améliorations notables, il existe toujours des situations difficiles pour l'ensemble des algorithmes (petites cavités mobiles, réseaux de cavités, crevasses très sensibles aux effets de bords). Cette étude approfondie m'a tout de même permi de définir un ensemble de paramètres favorables, et de différencier l'impact de ces différents paramètres sur le suivi des cavités. Ces méthodes ont été mises en pratique dans deux des trois projets de détermination d'inhibiteurs. Le suivi des cavités de la gyrase a ainsi révélé quelques cavités dont le volume varie sensiblement au cours de la transition fonctionnelle, et qui pourraient donc être de bons candidats pour des sites allostériques. Le suivi des cavités de la subtilisine 1 de *P. vivax* a permis d'extraire efficacement la cavité enzymatique et d'identifier une autre cavité allostérique potentiellement intéressante pour le développement d'un médicament antipaludéen.

La deuxième méthode d'analyse de la dynamique des cavités que j'ai développée pendant la thèse est la formulation de l'analyse en composantes principales pour les cavités. Cette méthode classique d'analyse des structures de protéines s'applique bien sur les cavités décrites par des grilles, malgré quelques limitations qui se sont finalement avérées peu génantes. Les composantes principales des structures et des cavités de quatre systèmes se sont notamment révélées assez fortement corrélées, ce qui permet de construite des conformations dont les cavités ont des géométries proches de celles décrites par les composantes principales des cavités. L'application de l'ACP sur les cavités de la myoglobine a révélé la correlation très marquée entre la position de la molécule de CO au sein du réseau de cavité et la forme des cavités formant ce réseau. Cette méthode s'avère ainsi très utile pour analyser la fonction de certaines protéines. L'ACP sur les cavités facilite grandement la sélection de conformations selon la forme de leurs cavités. Explorée en tant qu'exemple dans le chapitre III, ce deuxième type d'application a par ailleurs été utilisé dans le projet de criblage virtuel de la subtilisine 1 de *P. vivax* en collaboration avec Sanofi. Enfin, l'ACP sur les cavités nous a permis de classifier les cavités d'un sous-domaine du domaine extracellulaire de GLIC en quatre familles et d'en sélectionner quatre formes archétypales pour les utiliser dans un criblage virtuel.

Le développement du module python PyCav (voir annexes, section 1), implémentant l'ensemble de ces méthodes, devrait faciliter leur utilisation et la manipulation de trajectoires de cavités en règle générale. Ce module de 16000 lignes est largement testé et documenté et est facilement extensible grâce à son architecture orientée objet. Malgré son utilité, la difficulté d'accès que représente un module Python pour un non-spécialiste est susceptible de limiter son utilisation par la communauté des biologistes structuraux. Il semble donc pertinent d'orienter les futures voies de développement vers la création d'une interface graphique, ce qui pourrait toutefois nécessiter un travail important. Une solution intermédiaire serait le développement d'un module pour l'un des logiciels majeurs de visualisation de structures moléculaires. Les logiciels VMD, Chimera et PyMol sont de bons candidats, notamment au vu de l'intégration de Python dans leurs environnements respectifs.

Durant ma thèse, j'aurai participé à 3 projets de recherche de composés effecteurs. Le premier projet ciblait la gyrase de *M. tuberculosis*. J'ai pu produire un chemin de transition et réaliser la détection et le suivi des cavités. Ce projet à révélé plusieurs cavités potentiellement intéressantes en vue de l'étape de criblage virtuel. Malheureusement, le projet a été suspendu avant cette phase car les modèles par homologie de certaines structures n'étaient pas suffisament aboutis.

Le deuxième projet ciblait la subtilisine 1 de *P. vivax*. Ce projet était particulièrement intéressant du fait de la collaboration avec Sanofi qui m'a permis de travailler en proche collaboration avec l'industrie pharmaceutique. J'ai pu observer une efficacité et une soif de résultats marquante, avec parfois une tendance à éviter la prise de risque et de l'innovation. A ce titre, je pense que les collaborations entre l'industrie et l'académie ne peuvent qu'être fructueuses, les laboratoires académiques pouvant combler ce manque de prise de risque, comme nous avons pu le faire en apportant notre expérience sur l'étude des cavités. Les campagnes de tests des composés issus du criblage virtuel n'ont malheureusement pas été très concluantes jusqu'ici. Les raisons de ces

résultats en demi-teinte peuvent venir des défis inhérents à la cible, ou plus prosaïquement des conditions stringeantes appliquée aux tests des composés, visant à identifier des touches de hautes affinités dans un projet comportant d'autres pistes de détermination d'inhibiteurs à des stades différents. Une autre campagne de tests a par ailleurs été lancée et les résultats devraient venir à moyen terme.

Le troisième projet consistait à déterminer des composés ayant un effet sur GLIC, qu'il soit potentiateur ou inhibiteur, en se basant sur une structure déterminée par Sauguet *et al.* au laboratoire de Marc Delarue. La résolution assez basse de la structure cristallographique de cette protéine difficile, couplée à la complexité du test des composés sur œufs de xénopes imposant une limite drastique au nombre de molécules à tester, faisait que ce projet était à haut risque. Au cours de celui-ci, j'ai pu développer mes propres méthodes de chémoinformatique en réalisant la partition en deux étapes de la ZINC, permettant un accès hiérarchique aux composés par des représentants. Les résultats intermédiaires de ce projet sont très encourageants : sur les 26 composés testés, 4 semblent avoir un effet notable sur la fonction de GLIC (2 potentiateurs et 2 inhibiteurs) et une dizaine ont un faible effet. Ces résultats doivent toutefois être modulés par la concentration élevée des composés testés imposée par leur très faible nombre.

Ces trois projets très différents ont apporté des conclusions variées : nécessité de modèles structuraux affinés pour la gyrase, résultats en demi-teinte (mais encore en cours) pour la subtilisine 1, et résultats très encourageants pour GLIC. De fait, il parait prématuré de donner une conclusion définitive sur l'impact de l'utilisation des données dynamiques de cavités pour un projet de conception de médicaments. L'échelle de temps de ce type de projet (plusieurs mois entre le criblage virtuel et l'exploitation des tests des composés) ainsi que les ressources mises en jeu ralentissent la réalisation d'une étude scientifique pour ce type de développement méthodologiques. Ce type d'étude devrait en effet se baser sur la comparaison de résultats de criblages réalisé avec et sans l'utilisation de ces méthodes. De tels travaux ont toutefois été en partie réalisés dans des travaux antérieurs[111], mais aussi dans une certaine mesure dans le projet GLIC, où une structure cristallographique a été délibérement choisie pour l'étape de criblage (mais en utilisant nos méthodes). A ce titre, un prolongement important de ces travaux de thèses serait d'évaluer l'impact de l'utilisation de la dynamique des cavités sur les performances de criblages virtuels. Il est possible d'envisager d'utiliser des jeux de données conditionnés et faciles d'utilisation comme la DUD-E[67]. Je n'aurais pas réalisé ce type d'étude pendant la thèse pour des questions de temps et de priorités. J'ai en effet préféré me concentrer sur la conception de ces méthodes et sur l'aspect plus concret de la découverte de nouveaux inhibiteurs proposé par les projets de conception de médicaments auxquels j'ai pu participer. Il était en effet très important de confronter les méthodes en développement à des cas concrets afin de faire resortir des difficultés innatendues et ainsi d'en renforcer la pertience. Ainsi, la mise à l'épreuve de ces nouvelles méthodes sur des cas tests et sur des cas concrets sont complémentaires. Leur utilisation à grande échelle est indispensable et constitue une sorte de démonstration par empirisme qui peut s'étaler sur une longue période.

BIBLIOGRAPHIE

1. Ng Rick. Drugs : from discovery to approval. Hoboken, New Jersey : , John Wiley & Sons, ; , 2nd edition, 2011.
2. Drews Jürgen. Drug discovery : A historical perspective. Science 2000 ;**287**(5460) :1960–1964.
3. Pina AnaSofia, Hussain Abid, Roque AnaCecíliaA. An historical overview of drug discovery. In *Ligand-Macromolecular Interactions in Drug Discovery,* Roque Ana Cecília A., editorvolume 572 of *Methods in Molecular Biology* 3–12. Humana Press 2010.
4. Li Jie Jack. History of Drug Discovery, chapter 1, 1–42. John Wiley & Sons, Inc. 2013.
5. Fleming Alexander. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of b. influenzae. British Journal of Experimental Pathology, June, 1929 ;**3** :226–236.
6. DiMasi Joseph A, Hansen Ronald W, Grabowski Henry G. The price of innovation : new estimates of drug development costs. Journal of Health Economics 2003 ;**22**(2) :151–185.
7. Paul Steven M, Mytelka Daniel S, Dunwiddie Christopher T, Persinger Charles C, Munos Bernard H, Lindborg Stacy R, Schacht Aaron L. How to improve r&d productivity : the pharmaceutical industry's grand challenge. Nat Rev Drug Discov, Mar, 2010 ;**9**(3) :203–214.
8. Jiang Kevin. Near-record number of approvals signals drug development shift. Nature Medicine, Feb, 2013 ;**19**(2) :114.
9. Scannell Jack W, Blanckley Alex, Boldon Helen, Warrington Brian. Diagnosing the decline in pharmaceutical r&d efficiency. Nat Rev Drug Discov, Mar, 2012 ;**11**(3) :191–200.
10. Goodnow Jr Robert A. Hit and lead identification : Integrated technology-based approaches. Drug Discovery Today : Technologies 2006 ;**3**(4) :367–375.
11. Rosier Jan A., Martens Mark A., Thomas Josse R. Drug Life Cycle 1–11. John Wiley & Sons, Ltd 2014.
12. ICH E8. General considerations for clinical trials. In *London, UK : International Conference on Harmonisation*, 1997.
13. Rosier Jan A., Martens Mark A., Thomas Josse R. Drug development : General aspects. In *Global New Drug Development*, 23–90. John Wiley & Sons, Ltd 2014.
14. Kennedy Tony. Managing the drug discovery/development interface. Drug Discovery Today 1997 ;**2**(10) :436–444.
15. DiMasi Joseph A. The value of improving the productivity of the drug development process. PharmacoEconomics 2002 ;**20**(3) :1–10.
16. Cohen Philip. Protein kinases—the major drug targets of the twenty-first century ? Nature reviews Drug discovery 2002 ;**1**(4) :309–315.
17. Filmore David. It'sa gpcr world. Modern drug discovery 2004 ;**7**(11) :24–28.
18. Overington John P, Al-Lazikani Bissan, Hopkins Andrew L. How many drug targets are there ? Nature reviews Drug discovery 2006 ;**5**(12) :993–996.
19. Leifert Wayne R, editor. G Protein-Coupled Receptors in Drug Discovery. Berlin, Germany : Humana Press ; 1st edition 2009.
20. Debouck C. The hiv-1 protease as a therapeutic target for aids. AIDS Res Hum Retroviruses, Feb, 1992 ;**8**(2) :153–164.
21. Yeoh Sharon, O'Donnell Rebecca A, Koussis Konstantinos, Dluzewski Anton R, Ansell Keith H, Osborne Simon A, Hackett Fiona, Withers-Martinez Chrislaine, Mitchell Graham H, Bannister Lawrence H, Bryans Justin S, Kettleborough Catherine A, Blackman Michael J. Subcellular discharge of a serine protease mediates release of invasive malaria parasites from host erythrocytes. Cell, Dec, 2007 ;**131**(6) :1072–1083.

22. Koussis Konstantinos, Withers-Martinez Chrislaine, Yeoh Sharon, Child Matthew, Hackett Fiona, Knuepfer Ellen, Juliano Luiz, Woehlbier Ute, Bujard Hermann, Blackman Michael J. A multifunctional serine protease primes the malaria parasite for red blood cell invasion. EMBO J, Mar, 2009 ;**28**(6) :725–735.
23. Tawk Lina, Lacroix Céline, Gueirard Pascale, Kent Robyn, Gorgette Olivier, Thiberge Sabine, Mercereau-Puijalon Odile, Ménard Robert, Barale Jean-Christophe. A key role for plasmodium subtilisin-like sub1 protease in egress of malaria parasites from host hepatocytes. J Biol Chem, Nov, 2013 ;**288**(46) :33336–33346.
24. Turk Boris. Targeting proteases : successes, failures and future prospects. Nat Rev Drug Discov, September, 2006 ;**5**(9) :785–799.
25. Kaczorowski Gregory J, McManus Owen B, Priest Birgit T, Garcia Maria L. Ion channels as drug targets : the next gpcrs. J Gen Physiol, May, 2008 ;**131**(5) :399–405.
26. Bodenhausen Geoffrey, Ruben David J. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chemical Physics Letters 1980 ;**69**(1) :185–189.
27. Aue W. P., Bartholdi E., Ernst R. R. Two-dimensional spectroscopy. application to nuclear magnetic resonance. The Journal of Chemical Physics 1976 ;**64**(5) :2229–2246.
28. Rieping Wolfgang, Habeck Michael, Bardiaux Benjamin, Bernard Aymeric, Malliavin Thérèse E., Nilges Michael. ARIA2 : Automated noe assignment and data integration in nmr structure calculation. Bioinformatics 2007 ;**23**(3) :381–382.
29. Güntert Peter. Automated nmr structure calculation with cyana. In *Protein NMR Techniques,* Downing A.Kristina, editorvolume 278 of *Methods in Molecular Biology* 353–378. Humana Press 2004.
30. Poh Mee Kian, Yip Andy, Zhang Summer, Priestle John P, Ma Ngai Ling, Smit Jolanda M, Wilschut Jan, Shi Pei-Yong, Wenk Markus R, Schul Wouter. A small molecule fusion inhibitor of dengue virus. Antiviral Res, Dec, 2009 ;**84**(3) :260–266.
31. Mayr Lorenz M, Bojanic Dejan. Novel trends in high-throughput screening. Curr Opin Pharmacol, Oct, 2009 ;**9**(5) :580–588.
32. Shoichet Brian K. Virtual screening of chemical libraries. Nature, Dec, 2004 ;**432**(7019) :862–865.
33. Gilson Michael K, Zhou Huan-Xiang. Calculation of protein-ligand binding affinities. Annu Rev Biophys Biomol Struct 2007 ;**36** :21–42.
34. Fischer Emil. Einfluss der configuration auf die wirkung der enzyme. Berichte der deutschen chemischen Gesellschaft 1894 ;**27**(3) :2985–2993.
35. Koshland D. E. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci U S A, Feb, 1958 ;**44**(2) :98–104.
36. Lancet D., Pecht I. Kinetic evidence for hapten-induced conformational transition in immunoglobin MOPC 460. Proc Natl Acad Sci U S A, Oct, 1976 ;**73**(10) :3549–3553.
37. Foote J., Milstein C. Conformational isomerism and the diversity of antibodies. Proc Natl Acad Sci U S A, Oct, 1994 ;**91**(22) :10370–10374.
38. Leder L., Berger C., Bornhauser S., Wendt H., Ackermann F., Jelesarov I., Bosshard H. R. Spectroscopic, calorimetric, and kinetic demonstration of conformational adaptation in peptide-antibody recognition. Biochemistry, Dec, 1995 ;**34**(50) :16509–16518.
39. Berger C., Weber-Bornhauser S., Eggenberger J., Hanes J., Plückthun A., Bosshard H. R. Antigen recognition by conformational selection. FEBS Lett, Apr, 1999 ;**450**(1-2) :149–153.
40. Kumar S., Ma B., Tsai C. J., Sinha N., Nussinov R. Folding and binding cascades : dynamic landscapes and population shifts. Protein Sci, Jan, 2000 ;**9**(1) :10–19.
41. Grünberg Raik, Leckner Johan, Nilges Michael. Complementarity of structure ensembles in protein-protein binding. Structure, Dec, 2004 ;**12**(12) :2125–2136.
42. Hammes Gordon G, Chang Yu-Chu, Oas Terrence G. Conformational selection or induced fit : a flux description of reaction mechanism. Proceedings of the National Academy of Sciences 2009 ;**106**(33) :13737–13741.

43. Silva Daniel-Adriano, Bowman Gregory R., Sosa-Peinado Alejandro, Huang Xuhui. A role for both conformational selection and induced fit in ligand binding by the lao protein. PLoS Comput Biol, 05, 2011 ;**7**(5) :e1002054.

44. Changeux Jean-Pierre, Edelstein Stuart. Conformational selection or induced fit ? 50 years of debate resolved. F1000 Biol Rep 2011 ;**3** :19.

45. Zavodszky Maria I, Sanschagrin Paul C, Korde Rajesh S, Kuhn Leslie A. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. J Comput Aided Mol Des, Dec, 2002 ;**16**(12) :883–902.

46. McMartin C., Bohacek R. S. Qxp : powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des, Jul, 1997 ;**11**(4) :333–344.

47. Cavasotto Claudio N., Abagyan Ruben A. Protein flexibility in ligand docking and virtual screening to protein kinases. Journal of Molecular Biology 2004 ;**337**(1) :209–225.

48. Claussen H., Buning C., Rarey M., Lengauer T. Flexe : efficient molecular docking considering protein structure variations. J Mol Biol, Apr, 2001 ;**308**(2) :377–395.

49. Barril Xavier, Morley S. David. Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. J Med Chem, Jun, 2005 ;**48**(13) :4432–4443.

50. Totrov Maxim, Abagyan Ruben. Flexible ligand docking to multiple receptor conformations : a practical alternative. Curr Opin Struct Biol, Apr, 2008 ;**18**(2) :178–184.

51. Bottegoni Giovanni, Rocchia Walter, Rueda Manuel, Abagyan Ruben, Cavalli Andrea. Systematic exploitation of multiple receptor conformations for virtual ligand screening. PLoS One 2011 ;**6**(5) :e18845.

52. Kuntz I. D., Blaney J. M., Oatley S. J., Langridge R., Ferrin T. E. A geometric approach to macromolecule-ligand interactions. J Mol Biol, Oct, 1982 ;**161**(2) :269–288.

53. DesJarlais R. L., Sheridan R. P., Seibel G. L., Dixon J. S., Kuntz I. D., Venkataraghavan R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. J Med Chem, Apr, 1988 ;**31**(4) :722–729.

54. Ewing Todd J.A., Makino Shingo, Skillman A. Geoffrey, Kuntz Irwin D. DOCK 4.0 : Search strategies for automated molecular docking of flexible molecule databases. Journal of Computer-Aided Molecular Design 2001 ;**15**(5) :411–428.

55. Goodsell D. S., Olson A. J. Automated docking of substrates to proteins by simulated annealing. Proteins 1990 ;**8**(3) :195–202.

56. Morris Garrett M., Goodsell David S., Halliday Robert S., Huey Ruth, Hart William E., Belew Richard K., Olson Arthur J. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 1998 ;**19**(14) :1639–1662.

57. Trott Oleg, Olson Arthur J. Autodock vina : Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry 2010 ;**31**(2) :455–461.

58. McGann Mark R, Almond Harold R, Nicholls Anthony, Grant J. Andrew, Brown Frank K. Gaussian docking functions. Biopolymers, Jan, 2003 ;**68**(1) :76–90.

59. Friesner Richard A., Banks Jay L., Murphy Robert B., Halgren Thomas A., Klicic Jasna J., Mainz Daniel T., Repasky Matthew P., Knoll Eric H., Shelley Mee, Perry Jason K., Shaw David E., Francis Perry, Shenkin Peter S. Glide : A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. Journal of Medicinal Chemistry 2004 ;**47**(7) :1739–1749. PMID : 15027865.

60. Kramer Bernd, Rarey Matthias, Lengauer Thomas. Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking. Proteins : Structure, Function, and Bioinformatics 1999 ;**37**(2) :228–241.

61. Rarey Matthias, Wefing Stephan, Lengauer Thomas. Placement of medium-sized molecular fragments into active sites of proteins. Journal of Computer-Aided Molecular Design 1996 ;**10**(1) :41–54.

62. Jones G., Willett P., Glen R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J Mol Biol, Jan, 1995 ;**245**(1) :43–53.

63. Jones G., Willett P., Glen R. C., Leach A. R., Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol, Apr, 1997 ;**267**(3) :727–748.

64. Srinivasan Jayashree, Cheatham Thomas E., Cieplak Piotr, Kollman Peter A., Case David A. Continuum solvent studies of the stability of dna, rna, and phosphoramidate−dna helices. Journal of the American Chemical Society 1998 ;**120**(37) :9401–9409.

65. Grant J. Andrew, Pickup Barry T., Nicholls Anthony. A smooth permittivity function for poisson–boltzmann solvation methods. Journal of Computational Chemistry 2001 ;**22**(6) :608–640.

66. Moitessier N., Englebienne P., Lee D., Lawandi J., Corbeil C. R. Towards the development of universal, fast and highly accurate docking/scoring methods : a long way to go. Br J Pharmacol, Mar, 2008 ;**153 Suppl 1** :S7–26.

67. Mysinger Michael M., Carchia Michael, Irwin John. J., Shoichet Brian K. Directory of useful decoys, enhanced (DUD-E) : Better ligands and decoys for better benchmarking. Journal of Medicinal Chemistry 2012 ;**55**(14) :6582–6594. PMID : 22716043.

68. Kellenberger Esther, Rodrigo Jordi, Muller Pascal, Rognan Didier. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 2004 ;**57**(2) :225–242.

69. Nichols Sara E, Baron Riccardo, McCammon J. Andrew. On the use of molecular dynamics receptor conformations for virtual screening. Methods Mol Biol 2012 ;**819** :93–103.

70. Wassman Christopher D., Baronio Roberta, Demir Özlem, Wallentine Brad D., Chen Chiung-Kuang, Hall Linda V., Salehi Faezeh, Lin Da-Wei, Chung Benjamin P., Wesley Hatfield G., Richard Chamberlin A., Luecke Hartmut, Lathrop Richard H., Kaiser Peter, Amaro Rommie E. Computational identification of a transiently open l1/s3 pocket for reactivation of mutant p53. Nat Commun, January, 2013 ;**4** :1407.

71. Tarcsay Ákos, Paragi Gábor, Vass Márton, Jójárt Balázs, Bogár Ferenc, Keserű György M. The impact of molecular dynamics sampling on the performance of virtual screening against gpcrs. Journal of Chemical Information and Modeling 2013 ;**53**(11) :2990–2999. PMID : 24116387.

72. Bottegoni Giovanni, Kufareva Irina, Totrov Maxim, Abagyan Ruben. A new method for ligand docking to flexible receptors by dual alanine scanning and refinement (SCARE). Journal of Computer-Aided Molecular Design 2008 ;**22**(5) :311–325.

73. Abagyan Ruben, Kufareva Irina. The flexible pocketome engine for structural chemogenomics. In *Chemogenomics,* Jacoby Edgar, editorvolume 575 of *Methods in Molecular Biology* 249–279. Humana Press 2009.

74. Withers Ian M., Mazanetz Michael P., Wang Hao, Fischer Peter M., Laughton Charles A. Active site pressurization : A new tool for structure-guided drug design and other studies of protein flexibility. Journal of Chemical Information and Modeling 2008 ;**48**(7) :1448–1454. PMID : 18553961.

75. Leach Andrew R. Molecular Modelling : principles and applications. Harrlow , UK : Prentice Hall ; 2nd edition 2001.

76. Berendsen Herman JC, Postma J Pl M, van Gunsteren Wilfred F, DiNola ARHJ, Haak JR. Molecular dynamics with coupling to an external bath. The Journal of chemical physics 1984 ;**81**(8) :3684–3690.

77. Nosé Shūichi. A molecular dynamics method for simulations in the canonical ensemble. Molecular Physics 1984 ;**52**(2) :255–268.

78. Hoover William G. Canonical dynamics : Equilibrium phase-space distributions. Phys. Rev. A, Mar, 1985 ;**31** :1695–1697.

79. Andersen Hans C. Molecular dynamics simulations at constant pressure and/or temperature. The Journal of Chemical Physics 1980 ;**72**(4) :2384–2393.

80. Best Robert B., Zhu Xiao, Shim Jihyun, Lopes Pedro E. M., Mittal Jeetain, Feig Michael, MacKerell Alexander D. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles. Journal of Chemical Theory and Computation 2012 ;**8**(9) :3257–3273. PMID : 23341755.

81. Best Robert B., Mittal Jeetain, Feig Michael, Jr. Alexander D. MacKerell. Inclusion of many-body effects in the additive CHARMM protein CMAP potential results in enhanced cooperativity of α-helix and β-hairpin formation. Biophysical Journal 2012 ;**103**(5) :1045–1051.

82. Weiner Scott J., Kollman Peter A., Case David A., Singh U. Chandra, Ghio Caterina, Alagona Guliano, Profeta Salvatore, Weiner Paul. A new force field for molecular mechanical simulation of nucleic acids and proteins. Journal of the American Chemical Society 1984 ;**106**(3) :765–784.

83. Cornell Wendy D., Cieplak Piotr, Bayly Christopher I., Gould Ian R., Merz Kenneth M., Ferguson David M., Spellmeyer David C., Fox Thomas, Caldwell James W., Kollman Peter A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society 1995 ;**117**(19) :5179–5197.

84. MacKerell Jr Alex D, Bashford D, Bellott M, Dunbrack Jr Roland Leslie, Evanseck JD, Field MJ, Fischer S, Gao J al, Guo H, Ha S a et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. The Journal of Physical Chemistry B 1998 ;**102**(18) :3586–3616.

85. Ewald P. P. Die berechnung optischer und elektrostatischer gitterpotentiale. Annalen der Physik 1921 ;**369**(3) :253–287.

86. Essmann Ulrich, Perera Lalith, Berkowitz Max L., Darden Tom, Lee Hsing, Pedersen Lee G. A smooth particle mesh ewald method. The Journal of Chemical Physics 1995 ;**103**(19) :8577–8593.

87. Verlet Loup. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. Phys. Rev., Jul, 1967 ;**159** :98–103.

88. Swope William C, Andersen Hans C, Berens Peter H, Wilson Kent R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules : Application to small water clusters. The Journal of Chemical Physics 1982 ;**76**(1) :637–649.

89. Hockney RW. The potential calculation and some applications. Methods in Computational Physics 1970 ;**9** :136–211.

90. Beeman D. Some multistep methods for use in molecular dynamics calculations. Journal of Computational Physics 1976 ;**20**(2) :130–139.

91. Ryckaert Jean P., Ciccotti Giovanni, Berendsen Herman J. Numerical integration of the cartesian equations of motion of a system with constraints : molecular dynamics of n-alkanes. Journal of Computational Physics, March, 1977 ;**23**(3) :327–341.

92. Constanciel Raymond, Contreras Renato. Self consistent field theory of solvent effects representation by continuum models : Introduction of desolvation contribution. Theoretica chimica acta 1984 ;**65**(1) :1–11.

93. Schaefer Michael, Karplus Martin. A comprehensive analytical treatment of continuum electrostatics. The Journal of Physical Chemistry 1996 ;**100**(5) :1578–1599.

94. Alder BJ, Wainwright TEf. Phase transition for a hard sphere system. The Journal of Chemical Physics 1957 ;**27**(5) :1208.

95. Levitt Michael, Warshel Arieh. Computer simulation of protein folding. Nature 1975 ;**253**(5494) :694–698.

96. Brooks B. R., Brooks C. L., Mackerell A. D., Nilsson L., Petrella R. J., Roux B., Won Y., Archontis G., Bartels C., Boresch S., Caflisch A., Caves L., Cui Q., Dinner A. R., Feig M., Fischer S., Gao J., Hodoscek M., Im W., Kuczera K., Lazaridis T., Ma J., Ovchinnikov V., Paci E., Pastor R. W., Post C. B., Pu J. Z., Schaefer M., Tidor B., Venable R. M., Woodcock H. L., Wu X., Yang W., York D. M., Karplus M. CHARMM : The biomolecular simulation program. Journal of Computational Chemistry 2009 ;**30**(10) :1545–1614.

97. Phillips James C., Braun Rosemary, Wang Wei, Gumbart James, Tajkhorshid Emad, Villa Elizabeth, Chipot Christophe, Skeel Robert D., Kalé Laxmikant, Schulten Klaus. Scalable molecular dynamics with NAMD. Journal of Computational Chemistry 2005 ;**26**(16) :1781–1802.

98. Case David A., Cheatham Thomas E., Darden Tom, Gohlke Holger, Luo Ray, Merz Kenneth M., Onufriev Alexey, Simmerling Carlos, Wang Bing, Woods Robert J. The amber biomolecular simulation programs. Journal of Computational Chemistry 2005 ;**26**(16) :1668–1688.

99. Van Der Spoel David, Lindahl Erik, Hess Berk, Groenhof Gerrit, Mark Alan E, Berendsen Herman JC. GROMACS : fast, flexible, and free. Journal of computational chemistry 2005 ;**26**(16) :1701–1718.

100. Hinsen Konrad. The molecular modeling toolkit : a new approach to molecular simulations. Journal of Computational Chemistry 2000 ;**21**(2) :79–85.

101. Huang He, Ozkirimli Elif, Post Carol Beth. A comparison of three perturbation molecular dynamics methods for modeling conformational transitions. J Chem Theory Comput, Apr, 2009 ;**5**(5) :1301–1314.

102. Bolhuis Peter G., Chandler David, Dellago Christoph, Geissler Phillip L. Transition path sampling : Throwing ropes over rough mountain passes, in the dark. Annual Review of Physical Chemistry 2002 ;**53**(1) :291–318.

103. Shirts M., Pande V. S. Screen savers of the world unite ! Science, Dec, 2000 ;**290**(5498) :1903–1904.

104. Zagrovic Bojan, Snow Christopher D, Shirts Michael R, Pande Vijay S. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. Journal of molecular biology 2002 ;**323**(5) :927–937.

105. Lindorff-Larsen Kresten, Piana Stefano, Dror Ron O, Shaw David E. How fast-folding proteins fold. Science 2011 ;**334**(6055) :517–520.

106. Laio Alessandro, Parrinello Michele. Escaping free-energy minima. Proceedings of the National Academy of Sciences 2002 ;**99**(20) :12562–12566.

107. Torrie G.M., Valleau J.P. Nonphysical sampling distributions in monte carlo free-energy estimation : Umbrella sampling. Journal of Computational Physics 1977 ;**23**(2) :187–199.

108. Pan Albert C., Sezer Deniz, Roux Benoit. Finding transition pathways using the string method with swarms of trajectories. The Journal of Physical Chemistry B 2008 ;**112**(11) :3432–3440. PMID : 18290641.

109. Fischer Stefan, Karplus Martin. Conjugate peak refinement : an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. Chemical Physics Letters 1992 ;**194**(3) :252–261.

110. Blondel Arnaud, Renaud Jean-Paul, Fischer Stefan, Moras Dino, Karplus Martin. Retinoic acid receptor : a simulation analysis of retinoic acid binding and the resulting conformational changes1. Journal of Molecular Biology 1999 ;**291**(1) :101–115.

111. Berneman Armand, Montout Lory, Goyard Sophie, Chamond Nathalie, Cosson Alain, d'Archivio Simon, Gouault Nicolas, Uriac Philippe, Blondel Arnaud, Minoprio Paola. Combined approaches for drug design points the way to novel proline racemase inhibitor candidates to fight chagas' disease. PLoS ONE, 04, 2013 ;**8**(4) :e60955.

112. Laine Elodie, Goncalves Christophe, Karst Johanna C., Lesnard Aurélien, Rault Sylvain, Tang Wei-Jen, Malliavin Thérèse E., Ladant Daniel, Blondel Arnaud. Use of allostery to identify inhibitors of calmodulin-induced activation of bacillus anthracis edema factor. Proceedings of the National Academy of Sciences 2010 ;**107**(25) :11277–11282.

113. Eriksson A. Elisabeth, Baase Walter A., Zhang Xue-Jun, Heinz Dirk W., Blaber Michael, Baldwin Enoch P., Matthews Brian W. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. Science 1992 ;**255**(5041) :178–183.

114. Tilton Robert F., Kuntz Irwin D., Petsko Gregory A. Cavities in proteins : structure of a metmyoglobin xenon complex solved to 1.9 .ang. Biochemistry 1984 ;**23**(13) :2849–2857. PMID : 6466620.

115. Elber R., Karplus M. Enhanced sampling in molecular dynamics : use of the time-dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. Journal of the American Chemical Society 1990 ;**112**(25) :9161–9175.

116. Elber Ron. Ligand diffusion in globins : simulations versus experiment. Current Opinion in Structural Biology 2010 ;**20**(2) :162–167. <ce :title>Theory and simulation / Macromolecular assemblages</ce :title>.

117. Brunori Maurizio, Gibson Quentin H. Cavities and packing defects in the structural dynamics of myoglobin. EMBO reports 2001 ;**2**(8) :674–679.

118. Bossa Cecilia, Amadei Andrea, Daidone Isabella, Anselmi Massimiliano, Vallone Beatrice, Brunori Maurizio, Nola Alfredo Di. Molecular dynamics simulation of sperm whale myoglobin : Effects of mutations and trapped CO on the structure and dynamics of cavities. Biophysical Journal 2005 ;**89**(1) :465–474.

119. Ruscio Jory Z., Kumar Deept, Shukla Maulik, Prisant Michael G., Murali T. M., Onufriev Alexey V. Atomic level computational identification of ligand migration pathways between solvent and binding site in myoglobin. Proceedings of the National Academy of Sciences 2008 ;**105**(27) :9204–9209.

120. Tomita Ayana, Sato Tokushi, Ichiyanagi Kouhei, Nozawa Shunsuke, Ichikawa Hirohiko, Chollet Matthieu, Kawai Fumihiro, Park Sam-Yong, Tsuduki Takayuki, Yamato Takahisa, Koshihara Shin-ya, Adachi Shin-ichi. Visualizing breathing motion of internal cavities in concert with ligand migration in myoglobin. Proceedings of the National Academy of Sciences 2009 ;**106**(8) :2612–2616.

121. Scorciapino Mariano Andrea, Robertazzi Arturo, Casu Mariano, Ruggerone Paolo, Ceccarelli Matteo. Breathing motions of a respiratory protein revealed by molecular dynamics simulations. Journal of the American Chemical Society 2009 ;**131**(33) :11825–11832. PMID : 19653680.

122. Gabba Matteo, Abbruzzetti Stefania, Spyrakis Francesca, Forti Flavio, Bruno Stefano, Mozzarelli Andrea, Luque F. Javier, Viappiani Cristiano, Cozzini Pietro, Nardini Marco, Germani Francesca, Bolognesi Martino, Moens Luc, Dewilde Sylvia. CO rebinding kinetics and molecular dynamics simulations highlight dynamic regulation of internal cavities in human cytoglobin. PLoS ONE, 01, 2013 ;**8**(1) :e49770.

123. Wade Rebecca C, Winn Peter J, Schlichting Ilme, Sudarko. A survey of active site access channels in cytochromes P450. Journal of Inorganic Biochemistry 2004 ;**98**(7) :1175–1182. Advances in the Inorganic Biochemistry of Cytochrome P450, Nitric Oxide Synthase and Related Systems.

124. Miao Yinglong, Baudry Jerome. Active-site hydration and water diffusion in cytochrome p450cam : A highly dynamic process. Biophysical Journal 2011 ;**101**(6) :1493–1503.

125. Benkaidali Lydia, André François, Maouche Boubekeur, Siregar Pridi, Benyettou Mohamed, Maurel François, Petitjean Michel. Computing cavities, channels, pores and pockets in proteins from non-spherical ligands models. Bioinformatics 2014 ;**30**(6) :792–800.

126. Chen Rong, Weng Zhiping. A novel shape complementarity scoring function for protein-protein docking. Proteins : Structure, Function, and Bioinformatics 2003 ;**51**(3) :397–408.

127. Venkatachalam C.M., Jiang X., Oldfield T., Waldman M. Ligandfit : a novel method for the shape-directed rapid docking of ligands to protein active sites. Journal of Molecular Graphics and Modelling 2003 ;**21**(4) :289–307.

128. Lee B., Richards F.M. The interpretation of protein structures : Estimation of static accessibility. Journal of Molecular Biology 1971 ;**55**(3) :379–400.

129. Connolly Michael L. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983 ;**221**(4612) :709–713.

130. Connolly Michael L. The molecular surface package. Journal of Molecular Graphics 1993 ;**11**(2) :139–141.

131. Poupon Anne. Voronoi and voronoi-related tessellations in studies of protein structure and interaction. Current opinion in structural biology 2004 ;**14**(2) :233–241.

132. Akkiraju Nataraj, Edelsbrunner Herbert. Triangulating the surface of a molecule. Discrete Applied Mathematics 1996 ;**71**(1–3) :5–22.

133. Edelsbrunner Herbert, Facello Michael, Liang Jie. On the definition and the construction of pockets in macromolecules. Discrete Applied Mathematics 1998 ;**88**(1–3) :83–102. Computational Molecular Biology {DAM} - {CMB} Series.

134. Liang Jie, Edelsbrunner Herbert, Fu Ping, Sudhakar Pamidighantam V, Subramaniam Shankar. Analytical shape computation of macromolecules : I. molecular area and volume through alpha shape. Proteins Structure Function and Genetics 1998 ;**33**(1) :1–17.

135. Liang Jie, Woodward Clare, Edelsbrunner Herbert. Anatomy of protein pockets and cavities : Measurement of binding site geometry and implications for ligand design. Protein Science 1998 ;**7**(9) :1884–1897.

136. Cazals Frédéric, Proust Flavien, Bahadur Ranjit P., Janin Joël. Revisiting the voronoi description of protein–protein interfaces. Protein Science 2006 ;**15**(9) :2082–2092.

137. Pérot Stéphanie, Sperandio Olivier, Miteva Maria A., Camproux Anne-Claude, Villoutreix Bruno O. Druggable pockets and binding site centric chemical space : a paradigm shift in drug discovery. Drug Discovery Today 2010 ;**15**(15–16) :656–667.

138. Huang Bingding, Schroeder Michael. LIGSITE^{csc} : predicting ligand binding sites using the connolly surface and degree of conservation. BMC Structural Biology 2006 ;**6**(1) :19.

139. Glaser Fabian, Morris Richard J., Najmanovich Rafael J., Laskowski Roman A., Thornton Janet M. A method for localizing ligand binding pockets in protein structures. Proteins, February, 2006 ;**62**(2) :479–488.

140. An Jianghong, Totrov Maxim, Abagyan Ruben. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol Cell Proteomics, Jun, 2005 ;**4**(6) :752–761.

141. Halgren Tom. New method for fast and accurate binding-site identification and analysis. Chem Biol Drug Des, Feb, 2007 ;**69**(2) :146–148.

142. Ghersi Dario, Sanchez Roberto. Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. Proteins, Feb, 2009 ;**74**(2) :417–424.

143. Laurie Alasdair T R, Jackson Richard M. Q-sitefinder : an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics, May, 2005 ;**21**(9) :1908–1916.

144. Harris Rodney, Olson Arthur J, Goodsell David S. Automated prediction of ligand-binding sites in proteins. Proteins, Mar, 2008 ;**70**(4) :1506–1517.

145. Smart O. S., Goodfellow J. M., Wallace B. A. The pore dimensions of gramicidin a. Biophys J, Dec, 1993 ;**65**(6) :2455–2460.

146. Petrek Martin, Otyepka Michal, Banas Pavel, Kosinova Pavlina, Koca Jaroslav, Damborsky Jiri. CAVER : a new tool to explore routes from protein clefts, pockets and cavities. BMC Bioinformatics 2006 ;**7**(1) :316.

147. Petřek Martin, Košinová Pavlína, Koča Jaroslav, Otyepka Michal. MOLE : A voronoi diagram-based explorer of molecular channels, pores, and tunnels. Structure 2007 ;**15**(11) :1357–1363.

148. Peters Klaus P., Fauck Jana, Frömmel Cornelius. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. Journal of Molecular Biology 1996 ;**256**(1) :201–213.

149. Binkowski T. Andrew, Naghibzadeh Shapor, Liang Jie. CASTp : Computed atlas of surface topography of proteins. Nucleic Acids Research 2003 ;**31**(13) :3352–3355.

150. Kim Deok-Soo, Cho Youngsong, Kim Donguk, Kim Sangsoo, Bhak Jonghwa, Lee Sung-Hoon. Euclidean voronoi diagrams of 3d spheres and applications to protein structure analysis. Japan Journal of Industrial and Applied Mathematics 2005 ;**22**(2) :251–265.

151. Le Guilloux Vincent, Schmidtke Peter, Tuffery Pierre. Fpocket : An open source platform for ligand pocket detection. BMC Bioinformatics 2009 ;**10**(1) :168.

152. Tseng Yan Yuan, Li Wen-Hsiung. Identification of protein functional surfaces by the concept of a split pocket. Proteins : Structure, Function, and Bioinformatics 2009 ;**76**(4) :959–976.

153. Tseng Yan Yuan, Dupree Craig, Chen Z. Jeffrey, Li Wen-Hsiung. SplitPocket : identification of protein functional surfaces and characterization of their spatial patterns. Nucleic Acids Research 2009 ;**37**(suppl 2) :W384–W389.

154. Laskowski Roman A. SURFNET : A program for visualizing molecular surfaces, cavities, and intermolecular interactions. Journal of Molecular Graphics 1995 ;**13**(5) :323–330.

155. Brady, G.Patrick Jr, Stouten PieterF.W. Fast prediction and visualization of protein binding pockets with PASS. Journal of Computer-Aided Molecular Design 2000 ;**14**(4) :383–401.

156. Levitt David G., Banaszak Leonard J. POCKET : A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. Journal of Molecular Graphics 1992 ;**10**(4) :229–234.

157. Hendlich M., Rippmann F., Barnickel G. LIGSITE : automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model, Dec, 1997 ;**15**(6) :359–63, 389.

158. Coleman Ryan G., Sharp Kim A. Travel depth, a new shape descriptor for macromolecules : Application to ligand binding. Journal of Molecular Biology 2006 ;**362**(3) :441–458.

159. Can Tolga, Chen Chao-I, Wang Yuan-Fang. Efficient molecular surface generation using level-set methods. Journal of Molecular Graphics and Modelling 2006 ;**25**(4) :442–454.

160. Weisel Martin, Proschak Ewgenij, Schneider Gisbert. PocketPicker : analysis of ligand binding-sites with shape descriptors. Chemistry Central Journal 2007 ;**1**(1) :1–17.

161. Kalidas Yeturu, Chandra Nagasuma. PocketDepth : A new depth based algorithm for identification of ligand binding sites in proteins. Journal of Structural Biology 2008 ;**161**(1) :31–42.

162. Tripathi Ashutosh, Kellogg Glen E. A novel and efficient tool for locating and characterizing protein cavities and binding sites. Proteins : Structure, Function, and Bioinformatics 2010 ;**78**(4) :825–842.

163. Kawabata Takeshi. Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins : Structure, Function, and Bioinformatics 2010 ;**78**(5) :1195–1211.

164. Hubbard Simon J., Argos Patrick. Cavities and packing at protein interfaces. Protein Science 1994 ;**3**(12) :2194–2206.

165. Sonavane Shrihari, Chakrabarti Pinak. Cavities and atomic packing in protein structures and interfaces. PLoS Comput Biol, 09, 2008 ;**4**(9) :e1000188.

166. Eyrisch Susanne, Helms Volkhard. What induces pocket openings on protein surface patches involved in protein–protein interactions ? Journal of Computer-Aided Molecular Design 2009 ;**23**(2) :73–86.

167. Krone M., Falk M., Rehm S., Pleiss J., Ertl T. Interactive exploration of protein cavities. Computer Graphics Forum 2011 ;**30**(3) :673–682.

168. Schmidtke Peter, Bidon-Chanal Axel, Luque F. Javier, Barril Xavier. MDpocket : open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics 2011 ;**27**(23) :3276–3285.

169. Metz Alexander, Pfleger Christopher, Kopitz Hannes, Pfeiffer-Marek Stefania, Baringhaus Karl-Heinz, Gohlke Holger. Hot spots and transient pockets : Predicting the determinants of small-molecule binding to a protein–protein interface. Journal of Chemical Information and Modeling 2012 ;**52**(1) :120–133.

170. Ashford Paul, Moss David, Alex Alexander, Yeap Siew, Povia Alice, Nobeli Irene, Williams Mark. Visualisation of variable binding pockets on protein surfaces by probabilistic analysis of related structure sets. BMC Bioinformatics 2012 ;**13**(1) :39.

171. Kokh Daria B., Richter Stefan, Henrich Stefan, Czodrowski Paul, Rippmann Friedrich, Wade Rebecca C. TRAPP : A tool for analysis of transient binding pockets in proteins. Journal of Chemical Information and Modeling 2013 ;**53**(5) :1235–1252.

172. Oliveira Saulo, Ferraz Felipe, Honorato Rodrigo, Xavier-Neto Jose, Sobreira Tiago, de Oliveira Paulo. KVFinder : steered identification of protein cavities as a pymol plugin. BMC Bioinformatics 2014 ;**15**(1) :197.

173. World Health Organization. Global strategy for dengue prevention and control, 2012–2020 (`http://www.who.int/denguecontrol/9789241504034/en/`), August, 2012.

174. Meltzer M. I., Rigau-Pérez J. G., Clark G. G., Reiter P., Gubler D. J. Using disability-adjusted life years to assess the economic impact of dengue in puerto rico : 1984-1994. Am J Trop Med Hyg, Aug, 1998 ;**59**(2) :265–271.

175. Canyon Deon V. Historical analysis of the economic cost of dengue in australia. J Vector Borne Dis, Sep, 2008 ;**45**(3) :245–248.

176. Whitehead Stephen S, Blaney Joseph E, Durbin Anna P, Murphy Brian R. Prospects for a dengue virus vaccine. Nat Rev Microbiol, Jul, 2007 ;**5**(7) :518–528.

177. Villar Luis, Dayan Gustavo Horacio, Arredondo-García José Luis, Rivera Doris Maribel, Cunha Rivaldo, Deseda Carmen, Reynales Humberto, Costa Maria Selma, Morales-Ramírez Javier Osvaldo,

Carrasquilla Gabriel, Rey Luis Carlos, Dietze Reynaldo, Luz Kleber, Rivas Enrique, Miranda Montoya Maria Consuelo, Cortés Supelano Margarita, Zambrano Betzana, Langevin Edith, Boaz Mark, Tornieporth Nadia, Saville Melanie, Noriega Fernando. Efficacy of a tetravalent dengue vaccine in children in latin america. New England Journal of Medicine 2015 ;**372**(2) :113–123. PMID : 25365753.

178. Tassaneetrithep Boonrat, Burgess Timothy H, Granelli-Piperno Angela, Trumpfheller Christine, Finke Jennifer, Sun Wellington, Eller Michael A, Pattanapanyasat Kovit, Sarasombath Suttipant, Birx Deborah L, Steinman Ralph M, Schlesinger Sarah, Marovich Mary A. DC-SIGN (cd209) mediates dengue virus infection of human dendritic cells. J Exp Med, Apr, 2003 ;**197**(7) :823–829.
179. Seema, Jain S. K. Molecular mechanism of pathogenesis of dengue virus : Entry and fusion with target cell. Indian J Clin Biochem, Jul, 2005 ;**20**(2) :92–103.
180. Sessions October M, Barrows Nicholas J, Souza-Neto Jayme A, Robinson Timothy J, Hershey Christine L, Rodgers Mary A, Ramirez Jose L, Dimopoulos George, Yang Priscilla L, Pearson James L, Garcia-Blanco Mariano A. Discovery of insect and human dengue virus host factors. Nature, Apr, 2009 ;**458**(7241) :1047–1050.
181. Modis Yorgo, Ogata Steven, Clements David, Harrison Stephen C. A ligand-binding pocket in the dengue virus envelope glycoprotein. Proceedings of the National Academy of Sciences 2003 ;**100**(12) :6986–6991.
182. Bressanelli Stéphane, Stiasny Karin, Allison Steven L, Stura Enrico A, Duquerroy Stéphane, Lescar Julien, Heinz Franz X, Rey Félix A. Structure of a flavivirus envelope glycoprotein in its low-ph-induced membrane fusion conformation. The EMBO Journal 2004 ;**23**(4) :728–738.
183. Kielian Margaret. Class {II} virus membrane fusion proteins. Virology 2006 ;**344**(1) :38–47. Virology 50th Anniversary Issue.
184. Modis Yorgo, Ogata Steven, Clements David, Harrison Stephen C. Structure of the dengue virus envelope protein after membrane fusion. Nature 2004 ;**427**(6972) :313–319.
185. Mukhopadhyay Suchetana, Kuhn Richard J., Rossmann Michael G. A structural perspective of the flavivirus life cycle. Nat Rev Micro, January, 2005 ;**3**(1) :13–22.
186. Li Ze, Khaliq Mansoora, Zhou Zhigang, Post Carol Beth, Kuhn Richard J., Cushman Mark. Design, synthesis, and biological evaluation of antiviral agents targeting flavivirus envelope proteins. Journal of Medicinal Chemistry 2008 ;**51**(15) :4660–4671. PMID : 18610998.
187. Zhou Zhigang, Khaliq Mansoora, Suk Jae-Eun, Patkar Chinmay, Li Long, Kuhn Richard J., Post Carol Beth. Antiviral compounds discovered by virtual screening of small−molecule libraries against dengue virus e protein. ACS Chemical Biology 2008 ;**3**(12) :765–775. PMID : 19053243.
188. Wilder-Smith Annelies, Ooi Eng-Eong, Vasudevan SubhashG., Gubler DuaneJ. Update on dengue : Epidemiology, virus evolution, antiviral drugs, and vaccine development. Current Infectious Disease Reports 2010 ;**12**(3) :157–164.
189. Eyrisch Susanne, Helms Volkhard. Transient pockets on protein surfaces involved in protein-protein interaction. Journal of Medicinal Chemistry 2007 ;**50**(15) :3457–3464.
190. Desdouits Nathan, Nilges Michael, Blondel Arnaud. Principal component analysis reveals correlation of cavities evolution and functional motions in proteins. Journal of Molecular Graphics and Modelling 2015 ;**55**(0) :13–24.
191. Laurent Benoist, Chavent Matthieu, Cragnolini Tristan, Dahl Anna Caroline E, Pasquali Samuela, Derreumaux Philippe, Sansom Mark S P, Baaden Marc. Epock : rapid analysis of protein pocket dynamics. Bioinformatics, Dec, 2014 ;**e** :In Press.
192. Xu Rui, Wunsch, D. II. Survey of clustering algorithms. Neural Networks, IEEE Transactions on, May, 2005 ;**16**(3) :645–678.
193. Yennamalli Ragothaman, Subbarao Naidu, Kampmann Thorsten, McGeary RossP., Young PaulR., Kobe Bostjan. Identification of novel target sites and an inhibitor of the dengue virus e protein. Journal of Computer-Aided Molecular Design 2009 ;**23**(6) :333–341.
194. Fuzo Carlos A, Degrève Léo. New pockets in dengue virus 2 surface identified by molecular dynamics simulation. J Mol Model, Mar, 2013 ;**19**(3) :1369–1377.

195. Drum Chester L., Yan Shui-Zhong, Bard Joel, Shen Yue-Quan, Lu Dan, Soelaiman Sandriyana, Grabarek Zenon, Bohm Andrew, Tang Wei-Jen. Structural basis for the activation of anthrax adenylyl cyclase exotoxin by calmodulin. Nature, January, 2002 ;**415**(6870) :396–402.

196. Schindler T., Bornmann W., Pellicena P., Miller W. T., Clarkson B., Kuriyan J. Structural mechanism for sti-571 inhibition of abelson tyrosine kinase. Science, Sep, 2000 ;**289**(5486) :1938–1942.

197. Zhang Jianming, Adrián Francisco J, Jahnke Wolfgang, Cowan-Jacob Sandra W, Li Allen G, Iacob Roxana E, Sim Taebo, Powers John, Dierks Christine, Sun Fangxian, Guo Gui-Rong, Ding Qiang, Okram Barun, Choi Yongmun, Wojciechowski Amy, Deng Xianming, Liu Guoxun, Fendrich Gabriele, Strauss André, Vajpai Navratna, Grzesiek Stephan, Tuntland Tove, Liu Yi, Bursulaya Badry, Azam Mohammad, Manley Paul W, Engen John R, Daley George Q, Warmuth Markus, Gray Nathanael S. Targeting bcr-abl by combining allosteric with atp-binding-site inhibitors. Nature, Jan, 2010 ;**463**(7280) :501–506.

198. Martínez Leandro, Malliavin Thérèse E., Blondel Arnaud. Mechanism of reactant and product dissociation from the anthrax edema factor : A locally enhanced sampling and steered molecular dynamics study. Proteins : Structure, Function, and Bioinformatics 2011 ;**79**(5) :1649–1661.

199. Amaro Rommie E, Li Wilfred W. Emerging methods for ensemble-based virtual screening. Curr Top Med Chem 2010 ;**10**(1) :3–13.

200. Knegtel R. M., Kuntz I. D., Oshiro C. M. Molecular docking to ensembles of protein structures. J Mol Biol, Feb, 1997 ;**266**(2) :424–440.

201. Craig Ian R, Pfleger Christopher, Gohlke Holger, Essex Jonathan W, Spiegel Katrin. Pocket-space maps to identify novel binding-site conformations in proteins. J Chem Inf Model, Oct, 2011 ;**51**(10) :2666–2679.

202. Amadei Andrea, Ceruso Marc A., Di Nola Alfredo. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. Proteins : Structure, Function, and Bioinformatics 1999 ;**36**(4) :419–424.

203. Abramson Jeff, Smirnova Irina, Kasho Vladimir, Verner Gillian, Kaback H. Ronald, Iwata So. Structure and mechanism of the lactose permease of escherichia coli. Science 2003 ;**301**(5633) :610–615.

204. Smirnova Irina, Kasho Vladimir, Sugihara Junichi, Kaback H. Ronald. Opening the periplasmic cavity in lactose permease is the limiting step for sugar binding. Proceedings of the National Academy of Sciences 2011 ;**108**(37) :15147–15151.

205. Chaptal Vincent, Kwon Seunghyug, Sawaya Michael R., Guan Lan, Kaback H. Ronald, Abramson Jeff. Crystal structure of lactose permease in complex with an affinity inactivator yields unique insight into sugar recognition. Proceedings of the National Academy of Sciences 2011 ;**108**(23) :9361–9366.

206. World Health Organization. Global tuberculosis report 2014 (`http://www.who.int/tb/publications/global_report/en/`) 2014.

207. World Health Organization. Who tuberculosis facts sheet (`http://www.who.int/mediacentre/factsheets/fs104/en/`).

208. Champoux J. J. Dna topoisomerases : structure, function, and mechanism. Annu Rev Biochem 2001 ;**70** :369–413.

209. Collin Frédéric, Karkare Shantanu, Maxwell Anthony. Exploiting bacterial DNA gyrase as a drug target : current state and perspectives. Appl Microbiol Biotechnol, Nov, 2011 ;**92**(3) :479–497.

210. Dong Ken C, Berger James M. Structural basis for gate-dna recognition and bending by type iia topoisomerases. Nature, Dec, 2007 ;**450**(7173) :1201–1205.

211. Edwards Marcus J, Flatman Ruth H, Mitchenall Lesley A, Stevenson Clare E M, Le Tung B K, Clarke Thomas A, McKay Adam R, Fiedler Hans-Peter, Buttner Mark J, Lawson David M, Maxwell Anthony. A crystal structure of the bifunctional antibiotic simocyclinone d8, bound to dna gyrase. Science, Dec, 2009 ;**326**(5958) :1415–1418.

212. Piton Jérémie, Petrella Stéphanie, Delarue Marc, André-Leroux Gwénaëlle, Jarlier Vincent, Aubry Alexandra, Mayer Claudine. Structural insights into the quinolone resistance mechanism of mycobacterium tuberculosis DNA gyrase. PLoS One 2010 ;**5**(8) :e12245.

213. Roca Joaquim, Wang James C. The capture of a DNA double helix by an ATP-dependent protein clamp : A key step in DNA transport by type II DNA topoisomerases. Cell 1992 ;**71**(5) :833–840.

214. Roca J., Wang J. C. Dna transport by a type ii dna topoisomerase : evidence in favor of a two-gate mechanism. Cell, May, 1994 ;**77**(4) :609–616.

215. Corbett Kevin D, Schoeffler Allyn J, Thomsen Nathan D, Berger James M. The structural basis for substrate specificity in dna topoisomerase iv. J Mol Biol, Aug, 2005 ;**351**(3) :545–561.

216. Sheffler Will, Baker David. RosettaHoles : Rapid assessment of protein core packing for structure prediction, refinement, design, and validation. Protein Science 2009 ;**18**(1) :229–239.

217. World Health Organization. Global malaria report 2014 (`http://www.who.int/malaria/publications/world_malaria` 2014.

218. Center for Disease Control and Prevention. Center for disease control and prevention website (`http://www.cdc.gov/`).

219. Gething Peter, Patil Anand, Smith David, Guerra Carlos, Elyazar Iqbal, Johnston Geoffrey, Tatem Andrew, Hay Simon. A new world malaria map : Plasmodium falciparum endemicity in 2010. Malaria Journal 2011 ;**10**(1) :378.

220. Noedl Harald, Se Youry, Schaecher Kurt, Smith Bryan L., Socheat Duong, Fukuda Mark M. Evidence of artemisinin-resistant malaria in western cambodia. New England Journal of Medicine 2008 ;**359**(24) :2619–2620. PMID : 19064625.

221. Dondorp Arjen M., Nosten François, Yi Poravuth, Das Debashish, Phyo Aung Phae, Tarning Joel, Lwin Khin Maung, Ariey Frederic, Hanpithakpong Warunee, Lee Sue J., Ringwald Pascal, Silamut Kamolrat, Imwong Mallika, Chotivanich Kesinee, Lim Pharath, Herdman Trent, An Sen Sam, Yeung Shunmay, Singhasivanon Pratap, Day Nicholas P.J., Lindegardh Niklas, Socheat Duong, White Nicholas J. Artemisinin resistance in plasmodium falciparum malaria. New England Journal of Medicine 2009 ;**361**(5) :455–467. PMID : 19641202.

222. O'Brien Connor, Henrich Philipp P, Passi Neha, Fidock David A. Recent clinical and molecular insights into emerging artemisinin resistance in plasmodium falciparum. Curr Opin Infect Dis, Dec, 2011 ;**24**(6) :570–577.

223. Phyo Aung Pyae, Nkhoma Standwell, Stepniewska Kasia, Ashley Elizabeth A, Nair Shalini, McGready Rose, ler Moo Carit, Al-Saai Salma, Dondorp Arjen M, Lwin Khin Maung, Singhasivanon Pratap, Day Nicholas P J, White Nicholas J, Anderson Tim J C, Nosten François. Emergence of artemisinin-resistant malaria on the western border of thailand : a longitudinal study. Lancet, May, 2012 ;**379**(9830) :1960–1966.

224. Ashley Elizabeth A., Dhorda Mehul, Fairhurst Rick M., Amaratunga Chanaki, Lim Parath, Suon Seila, Sreng Sokunthea, Anderson Jennifer M., Mao Sivanna, Sam Baramey, Sopha Chantha, Chuor Char Meng, Nguon Chea, Sovannaroth Siv, Pukrittayakamee Sasithon, Jittamala Podjanee, Chotivanich Kesinee, Chutasmit Kitipumi, Suchatsoonthorn Chaiyaporn, Runcharoen Ratchadaporn, Hien Tran Tinh, Thuy-Nhien Nguyen Thanh, Thanh Ngo Viet, Phu Nguyen Hoan, Htut Ye, Han Kay-Thwe, Aye Kyin Hla, Mokuolu Olugbenga A., Olaosebikan Rasaq R., Folaranmi Olaleke O., Mayxay Mayfong, Khanthavong Maniphone, Hongvanthong Bouasy, Newton Paul N., Onyamboko Marie A., Fanello Caterina I., Tshefu Antoinette K., Mishra Neelima, Valecha Neena, Phyo Aung Pyae, Nosten Francois, Yi Poravuth, Tripura Rupam, Borrmann Steffen, Bashraheil Mahfudh, Peshu Judy, Faiz M. Abul, Ghose Aniruddha, Hossain M. Amir, Samad Rasheda, Rahman M. Ridwanur, Hasan M. Mahtabuddin, Islam Akhterul, Miotto Olivo, Amato Roberto, MacInnis Bronwyn, Stalker Jim, Kwiatkowski Dominic P., Bozdech Zbynek, Jeeyapant Atthanee, Cheah Phaik Yeong, Sakulthaew Tharisara, Chalk Jeremy, Intharabut Benjamas, Silamut Kamolrat, Lee Sue J., Vihokhern Benchawan, Kunasol Chanon, Imwong Mallika, Tarning Joel, Taylor Walter J., Yeung Shunmay, Woodrow Charles J., Flegg Jennifer A., Das Debashish, Smith Jeffery, Venkatesan Meera, Plowe Christopher V., Stepniewska Kasia, Guerin Philippe J., Dondorp Arjen M., Day Nicholas P., White Nicholas J. Spread of artemisinin resistance in plasmodium falciparum malaria. New England Journal of Medicine 2014 ;**371**(5) :411–423. PMID : 25075834.

225. Withers-Martinez Chrislaine, Suarez Catherine, Fulle Simone, Kher Samir, Penzo Maria, Ebejer Jean-Paul, Koussis Kostas, Hackett Fiona, Jirgensons Aigars, Finn Paul, Blackman Michael J. Plasmodium subtilisin-like protease 1 (SUB1) : Insights into the active-site structure, specificity and

function of a pan-malaria drug target. International Journal for Parasitology 2012 ;**42**(6) :597–612. Molecular Approaches to Malaria 2012 (MAM 2012).

226. Blackman Michael J. Malarial proteases and host cell egress : an 'emerging' cascade. Cell Microbiol, Oct, 2008 ;**10**(10) :1925–1934.

227. Giganti David, Bouillon Anthony, Tawk Lina, Robert Fabienne, Martinez Mariano, Crublet Elodie, Weber Patrick, Girard-Blanc Christine, Petres Stéphane, Haouz Ahmed, Hernandez Jean-François, Mercereau-Puijalon Odile, Alzari Pedro M, Barale Jean-Christophe. A novel plasmodium-specific prodomain fold regulates the malaria drug target sub1 subtilase. Nat Commun 2014 ;**5** :4833.

228. Withers-Martinez Chrislaine, Strath Malcolm, Hackett Fiona, Haire Lesley F, Howell Steven A, Walker Philip A, Christodoulou Evangelos, Evangelos Christodoulou, Dodson Guy G, Blackman Michael J. The malaria parasite egress protease SUB1 is a calcium-dependent redox switch subtilisin. Nat Commun 2014 ;**5** :3726.

229. Fulle Simone, Withers-Martinez Chrislaine, Blackman Michael J, Morris Garrett M, Finn Paul W. Molecular determinants of binding to the plasmodium subtilisin-like protease 1. J Chem Inf Model, Mar, 2013 ;**53**(3) :573–583.

230. Tasneem Asba, Iyer Lakshminarayan M, Jakobsson Eric, Aravind L. Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal cys-loop ion channels. Genome Biol 2005 ;**6**(1) :R4.

231. Bocquet Nicolas, Nury Hugues, Baaden Marc, Poupon Chantal Le, Changeux Jean-Pierre, Delarue Marc, Corringer Pierre-Jean. X-ray structure of a pentameric ligand-gated ion channel in an apparently open conformation. Nature, Jan, 2009 ;**457**(7225) :111–114.

232. Prevost Marie S, Sauguet Ludovic, Nury Hugues, Renterghem Catherine Van, Huon Christèle, Poitevin Frederic, Baaden Marc, Delarue Marc, Corringer Pierre-Jean. A locally closed conformation of a bacterial pentameric proton-gated ion channel. Nat Struct Mol Biol, Jun, 2012 ;**19**(6) :642–649.

233. Sauguet Ludovic, Shahsavar Azadeh, Poitevin Frédéric, Huon Christèle, Menny Anaïs, Àkos Nemecz, Haouz Ahmed, Changeux Jean-Pierre, Corringer Pierre-Jean, Delarue Marc. Crystal structures of a pentameric ligand-gated ion channel provide a mechanism for activation. Proc Natl Acad Sci U S A, Jan, 2014 ;**111**(3) :966–971.

234. Taly Antoine, Delarue Marc, Grutter Thomas, Nilges Michael, Novère Nicolas Le, Corringer Pierre-Jean, Changeux Jean-Pierre. Normal mode analysis suggests a quaternary twist model for the nicotinic receptor gating mechanism. Biophys J, Jun, 2005 ;**88**(6) :3954–3965.

235. Irwin John J., Shoichet Brian K. ZINC - a free database of commercially available compounds for virtual screening. Journal of Chemical Information and Modeling 2005 ;**45**(1) :177–182.

236. Kohonen Teuvo. Self-organized formation of topologically correct feature maps. Biological Cybernetics 1982 ;**43**(1) :59–69.

237. Ester Martin, Kriegel Hans-Peter, Sander Jörg, Xu Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*volume 96 226–231 1996.

238. Chung Fan RK. Spectral graph theoryvolume 92. New York, NY, USA : American Mathematical Soc. ; 1997.

239. Shi J., Malik J. Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, Aug, 2000 ;**22**(8) :888–905.

240. von Luxburg Ulrike. A tutorial on spectral clustering. Statistics and Computing 2007 ;**17**(4) :395–416.

241. Word J. M., Lovell S. C., Richardson J. S., Richardson D. C. Asparagine and glutamine : using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol, Jan, 1999 ;**285**(4) :1735–1747.

Chapitre VI

Annexes

1 PyCav, un module python d'aide à la manipulation des cavités et structures protéiques

PyCav est un module Python dont l'objectif est de faciliter la manipulation simultanée des trajectoires et des cavités qui leur sont associées. Il est écrit en Python à l'aide de librairies de calcul scientifique (NumPy, SciPy, scikit-learn et scikit-image). Ce module est orienté objet : le code est regroupé sous forme de classes, représentant chacune un type d'objet bien spécifique. Les classes principales sont :

— la classe `Structure` représente les coordonnées et les propriétés atomiques d'une structure simple (statique)

— la classe `Trajectory` représente une trajectoire structurale, c'est-à-dire un ensemble de conformations d'une même protéine

— la classe `Cavity` représente une cavité de façon géométrique, dans l'espace. Il s'agit d'une classe abstraite, c'est-à-dire qu'elle sert de base à d'autres classes implémentant des cavités de différentes façons.

— la classe `GridCavity` implémente la classe `Cavity` pour représenter les cavités définies sur des grilles, telles que celles détectées par `mkgrid` ou gHECOM.

— la classe `Cavities` représente une trajectoire de cavités. Il est possible que la représentation ne soit pas purement géométrique, le but étant d'avoir les meilleurs performances possibles (mémoire ou CPU). Ici aussi, il s'agit d'une classe abstraite.

— la classe `GridCavities` implémente `Cavities`, représentant les trajectoires de cavités définies à l'aide de grilles.

— la classe `System` fait le lien entre les trajectoires de structure et de cavités d'une même protéine. C'est notamment dans cette classe qu'est implémenté l'algorithme de suivi des cavités défini au chapitre II.

1.1 Les fonctionnalités du module

1.1.1 Description des classes principales

La classe `Structure`. La classe `Structure` représente une structure moléculaire. Elle contient une liste de l'ensemble des atomes composant la structure, auxquels sont associés des coordonnées (x, y, z) et des propriétés (nom de l'atome, du résidu parent, numéro d'atome et de résidu, etc). La liste des propriétés reconnues par la classe `Structure` est donnée au tableau I. Il est possible d'accéder à ces propriétés et de les modifier à l'aide de la notation par parenthèse carrée :

```
>>> structure = pycav.Structure("dengue.pdb")
>>> print structure["chain"]
[ A A A ... B B B ]
```

Nom du champ	Description	Note
index	Numérotation interne des atomes (commence par 0)	1
serial	Numérotation des atomes comme indiquée dans le fichier	
hetatom	Booléen répondant à la question : l'atome doit-il être considéré comme un hétéroatome ?	
name	Noms des atomes (ex : CA, N, OT1, ...)	2
resname	Noms de résidu des atomes	2
resid	Numéro de résidu des atomes	
chain	Identifiant de chaîne (une lettre)	
occupancy	Degré d'occupation des atomes	
beta	Facteur beta des atomes (souvent utilisé pour associer une valeur réelle à chaque atome)	
segname	Identifiant de segment des atomes ("segid")	2
realname	Noms des atomes, non modifiés (espaces compris)	
realresname	Noms des résidus, non modifiés (espaces compris)	
realsegname	Identifiant de segment, non modifié (espaces compris)	
user	Champ vide modifiable par l'utilisateur (valeur par défaut : 0)	
properties	Valeur numérique permettant d'enregister certaines propriétés des atomes (chaîne principale, eau, ions...)	1
residue	Numéro interne du résidu, unique pour chaque résidu	1
coords	Les coordonnées atomiques (x, y, z)	

Tableau I – Description des propriétés reconnues par la classe `Structure`. Notes : 1. Le champ est automatiquement généré par PyCav à partir des autres données. 2. Ce champ est généré à partir du champ *real** correspondant.

La classe `Structure` permet également de définir des groupes structuraux tels que définis dans le chapitre II. En plus des méthodes communes à l'ensemble des classes principales (voirs sections suivantes), il est possible d'extraire la séquence (méthode `sequence`) ou de réaliser une carte de densité éléctronique sur un modèle gaussien à partir de la structure (méthode `to_volume`).

La classe `Trajectory`. La classe `Trajectory` représente un ensemble de conformations (une trajectoire) d'une structure moléculaire. Cette classe est composée d'une `Structure` pour contenir les

propriétés atomiques de la structure, et un tableau de taille $n_{conf} \times n_{atomes} \times 3$ contenant les coordonnées de la structure pour chaque conformation. La classe `Trajectory` possède une méthode `pca` permettant de calculer l'ACP de la trajectoire (voir chapitre III). Deux méthodes y sont associées : `build_component` permet de transcrire une composante principale en un objet `Structure` et `linear_stretch` permet de construire un nouvel objet `Trajectory` représentant l'évolution de la structure le long d'une composante principale.

Les classes `Cavity` et `GridCavity`. La classe `Cavity` représente une cavité de façon géométrique. Il s'agit d'une classe abstraite qui doit être implémentée par des classes qui en héritent, comme `GridCavity`. Ces classes doivent implémenter les méthodes `volume` et `surface` permettant de calculer le volume (en $Å^3$) et la surface (en $Å^2$) de la cavité représentée.

La classe `GridCavity` implémente la classe `Cavity` et décrit les cavités d'une unique conformation à l'aide d'une grille en 3 dimensions. Il est possible d'enregistrer plusieurs champs de valeur dans cette grille, accessibles via un nom de champ. Les champs *beta* et *resid* sont les plus courants ; ils correspondent respectivement à une valeur de champ scalaire (une probabilité de cavité par exemple) et à une liste d'identifiants de cavités. La classe `GridCavity` implémente également `score_points`, qui permet de récupérer les valeurs de la grille à chaque point d'une liste donnée en entrée.

Les classes `Cavities` et `GridCavities`. Cette classe abstraite définit un squelette commun. Les classes héritant de `Cavities` doivent ainsi implémenter les méthodes `within`, prenant en entrée une liste de points et retournant la distance des cavités les plus proches, et la méthode `split` permettant de diviser une cavité à l'aide d'une sonde dont le rayon est donné en entrée. Tout comme pour `Cavity`, les classes implémentant `Cavities` doivent fournir des méthodes pour calculer le volume (méthode `volume`) et la surface (méthode `surface`) des cavités pour chaque conformation.

L'implémentation `GridCavities` représente les trajectoires de cavité définies sur des grilles. La grille entière n'est pas gardée en mémoire. Seuls les points apparaissant au moins une fois durant la trajectoire sont effectivement pris en compte. Un objet `GridCavities` comporte donc un ensemble de coordonnées (attribut `points`) et une liste de taille $n_{conf} \times n_{points}$ associant un numéro de cavité à chacun de ces points pour chaque conformation (0 lorsqu'il n'y a pas de cavité pour la conformation courante). La classe `GridCavities` définit plusieurs méthodes supplémentaires :

— `to_trajectory` permet de traduire la trajectoire de cavités en une trajectoire spécifiquement construite à des fins de visualisation (VMD)

— `mean` calcule la cavité moyenne et retourne un objet `GridCavity`

— `pca` calcule l'ACP de la trajectoire des cavités

— `build_component` permet de transcrire une composante principale en un objet `GridCavity`

— `linear_component` permet de représenter l'évolution des cavités le long d'une composante principale (produit un objet `GridCavities`)

— `project` permet de projeter les cavités sur une composante principale ou un vecteur choisi par l'utilisateur

La classe `System`. La classe `System` groupe un objet `Trajectory` et un objet `Cavities` afin de faire le lien entre une structure et ses cavités associées. Cette classe est principalement utilisée pour réaliser le suivi des cavités décrit au chapitre II à l'aide de la méthode `track`. L'algorithme de suivi s'appuie sur des classes permettant de réaliser l'étape de regroupement. Ces classes se situent dans `pycav.tracking_utils` et dérivent de la classe `TrackClustering`. Actuellement, les algorithmes implémentés par ces classes incluent :

— le regroupement hiérarchique (classe `HierarchicalTrackClustering`)

— le regroupement par densité (classe `DBSCANTrackClustering`)

— le regroupement spectral de graphe (classe `SpectralGraphTrackClustering`)

Il est également possible de réaliser l'étape de regroupement par un autre moyen et d'utiliser la classe `PresetTrackClustering` pour réaliser le suivi à partir du résultat. Une fois le suivi des cavités réalisé, il est possible d'utiliser la méthode `follow_properties` pour suivre certaines propriétés de chaque cavités (hydrophobicité, conservation...).

1.1.2 Méthodes d'entrée/sortie

PyCav dispose d'un système d'entrée/sortie extensible et modulaire, basé sur des classes. Deux classes sont les parents de l'ensemble des classes d'entrée et sortie. Ce sont les classes `Reader` (permettant de lire des fichiers) et `Writer` (permettant d'écrire dans des fichiers). Il existe ensuite deux types de classes héritant directement de `Reader` (resp. `Writer`) :

— Les classes de la forme `[X]Reader` (resp. `[X]Writer`) lisent (resp. écrivent) des données sous format [X]

— Les classes de la forme `Base[Y]Reader` (resp. `Base[Y]Writer`) transforment les données pour être lisible par la classe [Y] (resp. transforment les données issues de la classe [Y])

Une troisième "couche" de classes hérite de ces deux types de classes à la fois. Elles combinent donc leurs capacités et permettent de lire des fichiers au format [X] pour les rendre lisible par la classe [Y]. Ces classes sont nommées `[XY]Reader` (resp. `[XY]Writer`).

Par exemple, en prenant [X] = `PDB` et [Y] = `Trajectory`, on peut former les classes suivantes :

— `PDBReader` permet de lire les fichiers au format PDB

— `BaseTrajectoryReader` permet de formatter les données pour être lisible par la classe `Trajectory`

— `PDBTrajectoryReader` hérite de `PDBReader` et `BaseTrajectoryReader` et permet donc de lire les fichiers au format PDB et de formatter les données pour être lisible par la classe `Trajectory`

Lorsque l'on appelle la commande `Trajectory("myoglobine.pdb")`, c'est donc la classe `PDBTrajectoryReader` qui est utilisée en interne pour lire le fichier `myoglobine.pdb`.

A noter que les classes `Reader` et `Writer` s'appuient sur la classe `Frames`, permettant de réaliser une sélection de conformations flexible et uniformisée sur l'ensemble des classes d'entrée/sortie :

```
# lire toutes les conformations du fichier myoglobin.dcd
>>> Trajectory("myoglobin.dcd")
# lire les conformations 1, 3, 5, 7 et 11 du fichier myoglobin.dcd
# (la numerotation suit la convention de Python, qui commence a 0)
>>> Trajectory("myoglobin.dcd", frames=[1, 3, 5, 7, 11])
# lire les conformations impaires du fichier myoglobin.dcd jusqu'a la fin du fichier
# (par convention la premiere structure est de rang impair et de numero 0)
>>> Trajectory("myoglobin.dcd", step=2)
# lire les conformations paires du fichier myoglobin.dcd jusqu'a la conformation 10
>>> Trajectory("myoglobin.dcd", start=1, step=2, stop=10)
```

Pour écrire un objet dans un fichier, il suffit d'utiliser la méthode `write`. Le format d'écriture est automatiquement inféré à partir de l'extension du fichier, mais il est possible de le spécifier à l'aide du mot clé `format`. Un dictionnaire permet de faire l'association entre une extension et la classe de lecture/écriture à utiliser.

```
# ecrit la trajectoire dans le fichier myoglobin_new.dcd (format DCD CHARMM)
>>> traj.write("myoglobin_new.dcd")
# ecrit la trajectoire dans le fichier myoglobin_new au format PDB
>>> traj.write("myoglobin_new", format="pdb")
```

1.1.3 Extraction

Chacune des cinq classes principales (`Structure`, `Trajectory`, `Cavity`, `Cavities` et `System`) possède une méthode `extract` permettant de créer un nouvel object contenant une sous-partie de l'objet d'origine. `extract` prend en argument un *string* décrivant la sélection à effectuer. La syntaxe de cette sélection est inspirée de celle de VMD. Il est ainsi possible de combiner plusieurs propriétés à l'aide des mots clés `and`, `or`, `not`, `same as` ou `within`. Cette méthode permet également d'extraire des conformations à l'aide de la classe `Frames` pour les classes décrivant plusieurs conformations (`Trajectory`, `Cavities` et `System`). Cela permet d'utiliser la même syntaxe que lors de la lecture d'un fichier :

```
traj.extract(frames=[1, 2, 5])
```

La méthode `extract` de `Structure` permet de créer un nouvel objet Structure en sélectionnant certains atomes à partir de leurs propriétés :

```
>>> structure = pycav.Structure("myoglobin.pdb")
>>> print structure
Structure(2534 atoms, 154 residues, grouping is 'atoms')
>>> calphas1to12 = structure.extract("name CA and resid 1 to 12")
>>> print calphas1to12
Structure(12 atoms, 12 residues, grouping is 'atoms')
```

La méthode `extract` de `Trajectory` utilise la même syntaxe (et rajoute la possibilié de sélectionner des conformations comme vu précédemment).

La méthode `extract` de `GridCavity` permet de séléctionner des zones géométriques dans l'espace (plan, sphère, cube), ainsi que de réaliser des filtres sur les valeurs d'un champ :

```
>>> cavity = pycav.GridCavity("myoglobin.ghe")
# selection des points de grille situes dans une sphere de rayon 5A
# situee aux coordonnes (1., 2., 3.).
>>> cavity_5A = cavity.extract("sphere 5. 1. 2. 3.")
# selection des points de grille situes 3A au dessus d'un plan
# de normale (1., 1., 1.) passant par l'origine.
>>> cavity_plane = cavity.extract("plane 3 over 1. 1. 1.")
# selection des points de grille dont la valeur pour le champ beta est positive
>>> cavity_betapos = cavity.extract("beta > 0.")
```

La méthode `extract` de `GridCavities` permet également de sélectionner des zones géométriques, mais aussi des cavités entières (à partir de leur label ou du mot clé `same`) et leurs alentours (à l'aide du mot clé `within`) :

```
>>> cavities = pycav.GridCavities("myoglobin.ghe")
# selection des cavites touchant une boite de taille 3x3x3
# commencant aux coordonnes (1., 2., 3.).
>>> cavities_touchbox = cavities.extract("same id as box 1. 2. 3. 3. 3. 3.")
```

Il est aussi possible de réaliser un filtre sur le volume ou la surface :

```
# selection des cavites dont le volume est compris entre 100A3 et 400A3
>>> cavities_vol = cavities.extract("volume > 100 and volume < 400")
```

La méthode `extract` de `System` inclue les mêmes mots clés que pour les classes `Trajectory` et `GridCavities`, mais inclue des mots clés permettant d'utiliser les trajectoires de structure et de cavités en même temps :

```
>>> system = System(traj, cavities)
# selection des cavites situees a moins de 5A des residus 5 a 10
>>> system_extract = system.extract("cavids within 5 of resid 5 to 10")
```

1.2 Exemples d'utilisation

1.2.1 Commandes de base

Lecture d'une trajectoire et détection des cavités. La première étape dans l'analyse des cavités est de créer un objet `Trajectory` a partir d'une structure (`myoglobin.pdb`) et d'une trajectoire (`"myoglobin.dcd"`) :

```
>>> import pycav
>>> structure = pycav.Structure("myoglobin.pdb")
>>> trajectoire = pycav.Trajectory("myoglobin.dcd", structure=structure)
```

Il est possible d'avoir un résumé très succinct du contenu des objets à l'aide de la commande `print` :

```
>>> print structure
Structure(2534 atoms, 154 residues, grouping is 'atom')
>>> print trajectoire
Trajectory(10 frames, 2534 atoms, 154 residues, grouping is 'atom')
```

Il est également possible de lire les deux fichiers en même temps avec `Trajectory`. L'objet `Structure` correspondant est alors automatiquement créé, on peut y accéder par `trajectoire.structure` :

```
>>> trajectoire = pycav.Trajectory("myoglobin.dcd", structure="myoglobin.pdb")
>>> print trajectoire.structure
Structure(2534 atoms, 154 residues, grouping is 'atom')
```

On pourra remarquer que le groupe structural par défaut est l'atome (`grouping is 'atom'`). Il peut être changé à l'aide de la commande `set_group` de l'objet `Structure` (il existe pour le moment 3 groupes structuraux : `atoms`, backbone-sidechain) et `residues` :

```
>>> trajectoire.structure.set_group('residues')
>>> print trajectoire.structure
Structure(2534 atoms, 154 residues, grouping is 'residues')
```

La détection des cavités est réalisée à l'aide du paquet `detection`. Les fonctionnalités de ce paquet ne peuvent être utilisées que si le module `pymkgrid` est installé. Ce module se trouve à l'adresse suivante : `http://TODO.org`. Pour détecter les cavités d'une simple structure, on utilisera la méthode `from_structure` :

```
>>> cavite = pycav.detection.from_structure(structure)
>>> print cavite
GridCavity(origin is at (-16., -10.5, -7.5), grid size is 0.5, size is 55x46x30, 1 field
(resid))
```

Pour détecter les cavités d'une trajectoire, on utilisera la méthode `from_trajectory` :

```
>>> cavites = pycav.detection.from_trajectory(trajectoire)
>>> print cavites
GridCavities(10 frames, 49510 grid points, grid size is 0.5)
```

Ecriture des objets dans des fichiers. Il est possible d'écrire les objets dans des fichiers pour pouvoir les relire plus tard. Cette fonction permet également de convertir les fichiers d'un format à un autre.

```
# sauvegarde de la structure au format .cor (coordonnees CHARMM) :
>>> structure.write("myoglobin.cor")
# sauvegarde de la cavite au format .mrc (format courant de densite electronique) :
>>> cavite.write("myoglobin.mrc")
```

Il n'existe pas pour le moment de format universel pour les trajectoire de cavités. Le format utilisé par PyCav se base sur le format `npz` de NumPy :

```
# sauvegarde de la trajectoire de cavites au format .npz :
>>> cavites.write("myoglobin.npz")
# relecture des cavites :
>>> print pycav.GridCavities("myoglobin.npz")
GridCavities(10 frames, 49510 grid points, grid size is 0.5)
```

Extraction Comme indiqué dans la section précédente, il est possible d'extraire des sous-parties des objets créés à l'aide de la commande `extract` :

```
# extraction des Calphas situes a moins de 5A de l'heme
>>> structca = structure.extract("name CA within 5 of resname HEM")
>>> print structca
Structure(44 atoms, 44 residues, grouping is 'residues')
# extraction de la cavite 5
# (notez que la taille de la grille ne change pas, mais que le contenu a ete modifie).
>>> cavite7 = cavite.extract("resid 7")
>>> print cavite7
GridCavity(origin is at (-16., -10.5, -7.5), grid size is 0.5, size is 55x46x30, 1 field
(resid))
```

Il est aussi possible d'extraire des conformations :

```
# extraction des conformations 1 et 2 + extraction des Calphas
>>> trajca =  trajectoire.extract("name CA", stop=2)
>>> print trajca
Trajectory(2 frames, 153 atoms, 153 residues, grouping is 'residues')
```

Volume et surface Le calcul du volume des cavités est réalisé à l'aide des méthodes `volume` et `surface` pour `GridCavity`, et `volumes` et `surfaces` pour `GridCavities` Elles retournent des valeurs uniques (en $Å^3$ et $Å^2$) pour `GridCavity` et un tableau de valeur (une pour chaque conformation) pour `GridCavities`.

```
>>> print cavite.volume(), cavite.surface()
175.75 364.29205544622272
>>> print cavites.volumes()
array([ 796.875, 878.125, 773.75 , 773.75 , 642.125, 655.875, 608.75 , 574.5 , 615.75 ,
793.25 ])
```

Pour calculer le volume ou la surface d'une cavité en particulier, il est nécessaire d'extraire un objet ne contenant que cette cavité, puis de calculer son volume :

```
>>> print cavite.extract("resid 7").volume()
31.25
```

1.2.2 Visualisation de trajectoires de cavités dans VMD

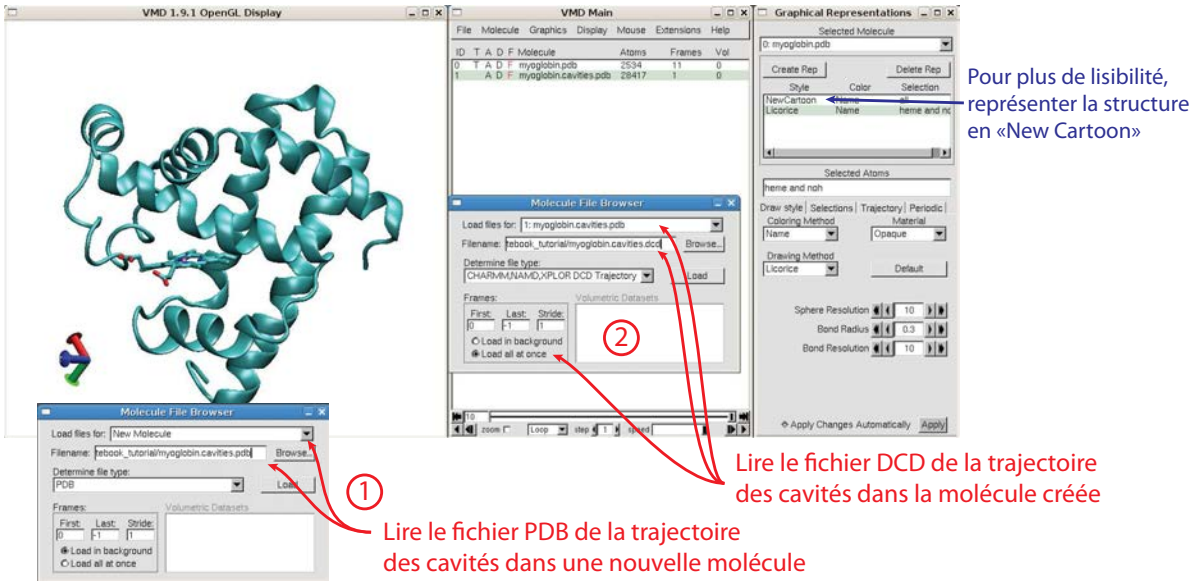
Le format `npz` utilisé par PyCav pour lire des trajectoires de cavités n'est lisible par aucun logiciel de visualisation moléculaire. Il est donc nécessaire de passer par une représentation de trajectoire structurale afin d'écrire dans un format lisible par ces logiciels. La méthode `to_trajectory` permet de créer un objet `Trajectory` intermédiaire permettant d'écrire les fichiers au format PDB et DCD nécessaires à la visualisation :

```
# creation de l'objet Trajectory intermediaire
>>> visutraj = cavites.to_trajectory()
# ecriture du fichier structure
>>> visutraj.structure.write("myoglobin.cavities.pdb")
# ecriture du fichier trajectoire
>>> visutraj.write("myoglobin.cavities.dcd")
```

Afin de visualiser cette trajectoire, nous utiliserons le logiciel VMD (Visual Molecular Dynamics) qui permet nativement de visualiser des trajectoires de structure :

```
# ouverture de vmd en chargeant directement la structure
# et la trajectoire structurale de la myoglobine
vmd myoglobin.pdb myoglobin.dcd
```

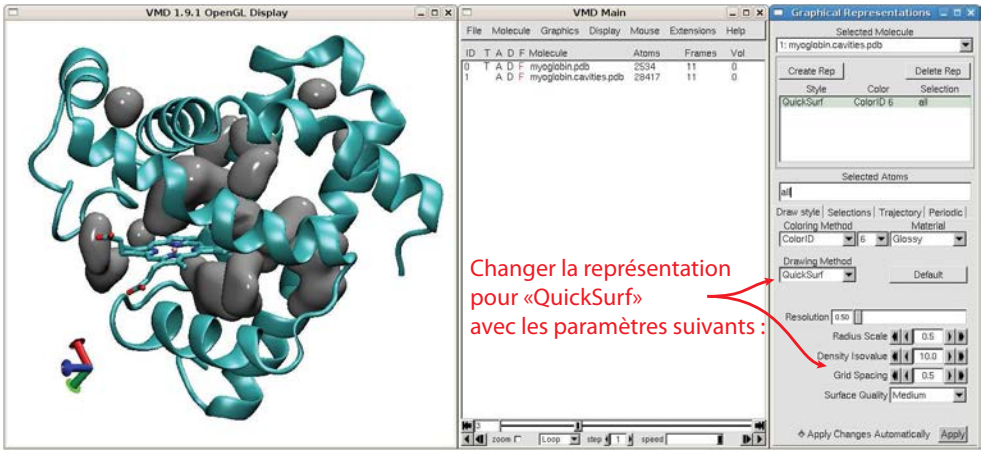
Cela ouvre les fenêtres principales de VMD (fenêtre d'affichage, fenêtre principale et éventuellement représentations graphiques). Pour charger la trajectoire de cavités, il faut d'abord charger le fichier `myoglobin.cavities.pdb` puis charger le fichier `myoglobin.cavities.dcd` à l'intérieur de la molécule créée :



Une fois la trajectoire de cavités chargée, il est possible que la fenêtre d'affichage de VMD n'affiche plus rien. La raison est que VMD zoome automatiquement sur la structure venant d'être chargée, or tous les points de grille du fichier structure des cavités se trouvent au même endroit (à l'origine du repère). Pour pouvoir voir les cavités, il faut donc :

1. changer de conformation courante (touche `+` avec le focus sur la fenêtre d'affichage ou faire glisser l'ascenceur horizontal de la fenêtre principale)
2. dézoomer (molette de la souris)

Pour bien visualiser les cavités, je recommande la représentation graphique `QuickSurf` (paramètres : Radius Scale à 0.5, Density Isovalue à 10.0, Grid Spacing à 0.5) :



Il est aussi possible d'utiliser la représentation `VdW` avec un paramètre Sphere Scale à 0.3 et un Material assez diffusif (AOChalky par exemple).

Les cavités indiquées ici ne sont pas différenciées (elles ont toutes la même couleur), car le suivi des cavités n'a pas encore été réalisé.

1.2.3 Suivi des cavités

Pour suivre les cavités (voir chapitre II), il faut créer un objet `System` associant une trajectoire structurale (objet `Trajectory`) à sa trajectoire de cavités correspondante (objet `GridCavity`) :

```
# creation de l'objet System
>>> system = pycav.System(trajectoire, cavites)
>>> print system
System(10 frames, no footprint, not clustered)
```

Pour calculer les empreintes des cavités, on utilisera la méthode `get_footprints` (les paramètres par défaut réalisent des empreintes de type *contact-distance* avec un seuil de 5 Å, voir chapitre II section 2.3 pour les différentes formes d'empreintes) :

```
# calcul des empreintes
>>> system.get_footprints()
```

Il est alors possible de réaliser la partition de ces empreintes à l'aide de la méthode `cluster_footprints`. Par défaut, cette méthode réalise une partition spectrale d'un graphe de contact à partir d'une distance cosinus (voir chapitre II section 2.4.1). Le nombre de cavités est automatiquement déterminé par cette méthode. Cette étape de partition permet de définir la fonction de passage des cavités instantanées aux cavités transverses, nécessaire pour l'étape d'assignement.

```
# partition des empreintes
>>> system.cluster_footprints()
```

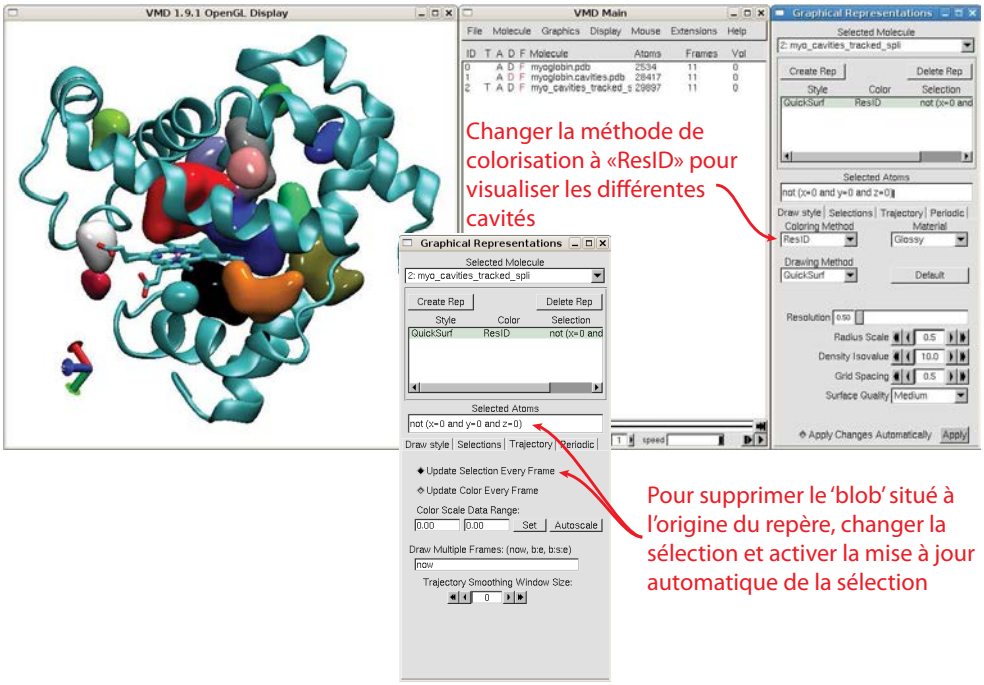
La dernière étape consiste donc à assigner l'identifiant de cavité transverse de chaque empreinte aux cavités correspondantes à l'aide de la méthode `set_cavities_clustering`.

```
# affectation des cavites
>>> system.set_cavities_clustering()
```

On peut enfin écrire les cavités dans les fichiers PDB et DCD comme dans la sous-section précédente :

```
# creation de l'objet Trajectory intermediaire
>>> visutraj_suivi = cavites.to_trajectory()
# ecriture du fichier structure
>>> visutraj_suivi.structure.write("myoglobin.cavities.suivies.pdb")
# ecriture du fichier trajectoire
>>> visutraj_suivi.write("myoglobin.cavities.suivies.dcd")
```

Le chargement de la trajectoire des cavités se fait de la même façon que précédemment. Pour visualiser l'identifiant des cavités par une couleur, il faut sélectionner la méthode de coloration `ResID` :



Les paramètres par défaut de `set_cavities_clustering` ne permettent pas de réaliser la détection des cavités fusionnées et leur division. Pour l'activer, il faut spécifier le mot clé *split_ outliers*.

```
# affectation des cavites + detection et division des cavites fusionnees
>>> system.set_cavities_clustering(split_outliers=True)
```

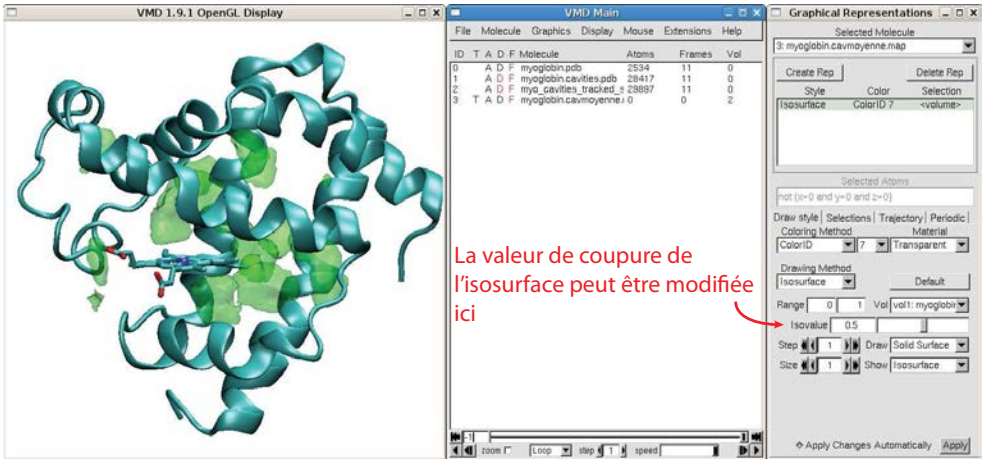
Attention : la division des cavités fusionnées entraîne une modification de l'objet `GridCavities` associé à l'objet `System`. Cette étape est donc irreversible et il faut donc relire ou recalculer les cavités pour retrouver l'objet de départ.

1.2.4 Cavité moyenne et analyse en composantes principales des cavités

La méthode `mean` de la classe `GridCavities` retourne un objet `GridCavity` représentant la "cavité moyenne", c'est-à-dire l'occupation moyenne de chaque point de grille au cours de la trajectoire :

```
# calcul de la cavite moyenne :
>>> cavmoyenne = cavites.mean()
# ecriture de la cavite moyenne en fichier .map (lisible par VMD) et .mrc (lisible par Chimera) :
>>> cavmoyenne.write("myoglobin.cavmoyenne.map")
>>> cavmoyenne.write("myoglobin.cavmoyenne.mrc")
```

Le fichier `.map` est lisible par VMD, qui permet de visualiser des isosurfaces à l'aide de la représentation `Isosurface` :



La méthode `pca` de la classe `GridCavities` permet de réaliser une ACP sur les cavités. Elle retourne trois tableaux :

— un tableau de taille n_{comp} contenant les valeurs propres de l'ACP, correspondant à la variance des cavités le long de chaque composante ($n_{comp} = \min(n_{points}, s) - 1$)

— un tableau de taille $n_{comp} \times n_{points}$ contenant les vecteurs propres de l'ACP dans l'espace des points de grilles

— un tableau de taille $n_{comp} \times s$ contenant les vecteurs propres de l'ACP dans l'espace des pas de temps

```
# ACP sur les cavites :
>>> cavites.pca()
```

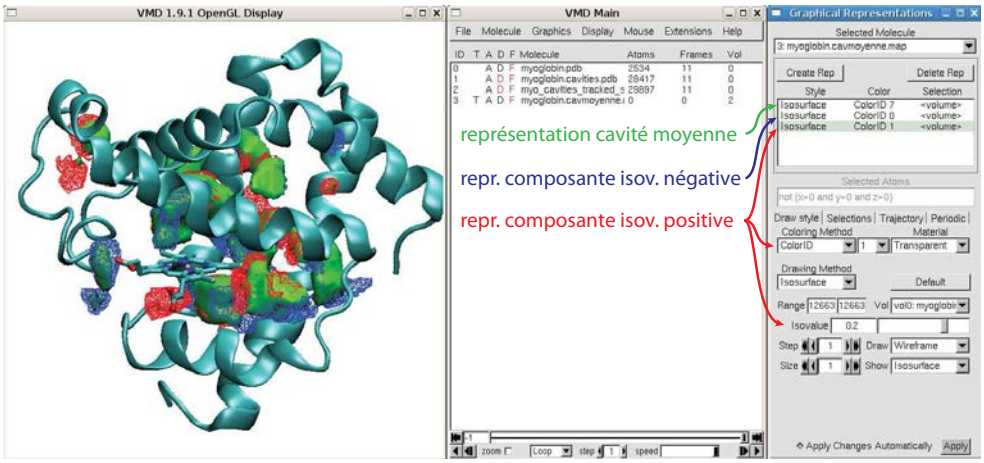
Une fois la méthode `pca` appelée, ces tableaux sont accessibles par les attributs `variances`, `components_points` et `components_step` de l'objet `GridCavities`.

La méthode `build_component` permet de créer un objet `GridCavity` représentant la composante principale choisie dans l'espace des cavités :

```
# representation de la premiere composante principale :
>>> composante1 = cavites.build_component(0)
# ecriture de la composante
>>> composante1.write("myoglobin.composante1.map")
>>> composante1.write("myoglobin.composante1.mrc")
```

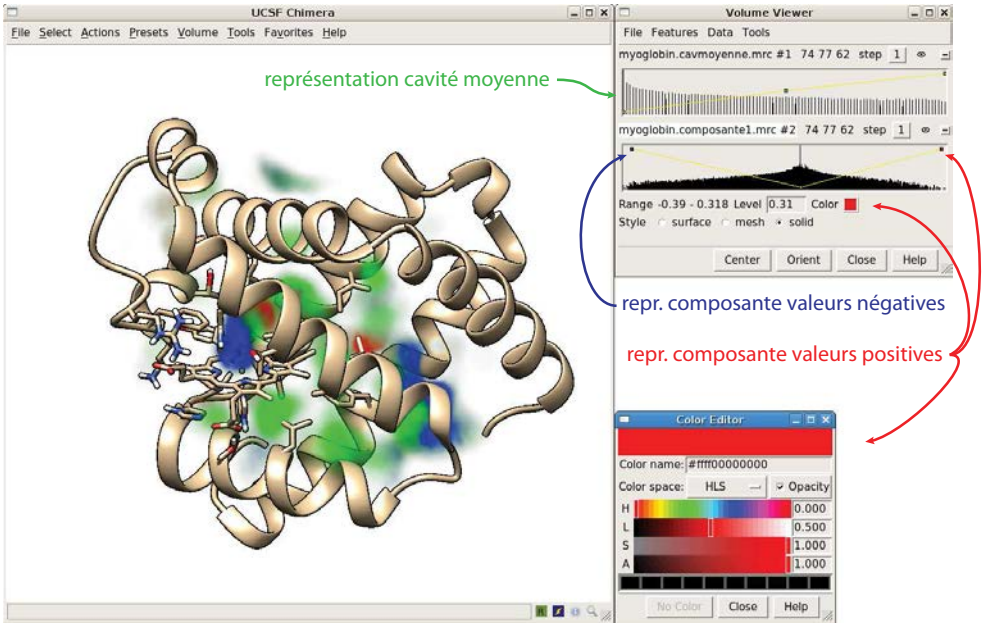
Les composantes principales de cavités sont constituées de zones positives et négatives. Le signe n'a pas d'interprétation physique mais des zones de signes identiques (resp. différents) sont

corrélés (resp. anticorrélés). Pour pouvoir visualiser les zones des deux signes simultanément dans VMD, il est nécessaire de créer deux représentations `Isosurface` avec des isovaleurs différentes :



Chimera propose des options de visualisation de données volumétriques bien plus poussées que VMD, permettant une visualisation en "nuage" plus représentative de la distribution qu'une isosurface. On peut lancer Chimera en spécifiant directement le fichier de structure et les fichiers volumétriques (au format `.mrc`) :

```
# ouverture de chimera en chargeant directement la structure
# et les donnees volumetriques :
chimera myoglobin.pdb myoglobin.cavmoyenne.mrc myoglobin.composante1.mrc
```



1.3 Perspectives

A court terme, le développement de PyCav se dirigera vers l'implémentation des classes `SphereCavity` et `SphereCavities` représentant des cavités sous formes de sphères. L'implémentation

d'autres formats de structure (psf, parm7), de trajectoire (amber, gromacs) et de cavités (notamment basés sur des sphères comme DOCK ou fpocket) est également envisagée. Un changement d'implémentation de l'algorithme de division des cavités fusionnées est également nécessaire pour éviter de modifier l'objet `GridCavities` initial. A moyen terme, il pourrait être très fructueux de développer un module pour VMD ou Chimera pour faciliter l'utilisation du module et permettre une visualisation des cavités plus facile et rapide. A plus long terme, il est possible d'envisager le développement d'un moteur graphique permettant de représenter efficacement la structure et les cavités sans avoir à passer par un logiciel externe. Cela permettrait également d'utiliser au maximum les possibilités de la classe `System`, notamment la sélection conjointe des structures et des cavités.

2 Paramètres de simulation des dynamiques moléculaires

Trois trajectoires de 120 ns de la myoglobine et du lysozyme ont été calculées en prenant des graines aléatoires différentes. Les conformations ont été calculées à l'aide du programme NAMD[97] avec le champ de force CHARMM22 et sauvegardées toutes les 10 ps.

Pour le lysozyme (entrée PDB : 2LYZ), la protéine a été solvatée dans une boîte de $70.4 \times 52.3 \times 49.0$ Å contenant 5192 molécules d'eau de type TIP3P, en plus des 101 molécules d'eau provenant de la structure cristallographique. Les charges positives de la protéine ont été neutralisées à l'aide de 8 ions chlorures. Les atomes de haute énergie (> 10 kcal/mol) ont été minimisés par 100 étapes de descente de gradient. Le solvant a ensuite été chauffé à 300 K et équilibré à l'aide d'une dynamique de Langevin (constante de couplage de $100\,\mathrm{ps}^{-1}$) de 10 ps, en maintenant les atomes de la protéine fixés. Le système a été équilibré à 310 K par une dynamique de Langevin (couplage à $0.1\,\mathrm{ps}^{-1}$). L'algorithme de Particle Mesh Ewald (PME) a été utilisé lors de cette équilibration pour calculer les forces électrostatiques à longue distance, en utilisant une distance seuil de 12 Å. Le pas de temps choisi est de 1 fs. Les liaisons impliquant des atomes d'hydrogène ont été contraintes à l'aide de l'algorithme SHAKE[91]. Les trois simulations de production ont été réalisées avec les mêmes paramètres que pour l'équilibration.

Pour les trajectoires de la myoglobine, le fichier de test "mbco4958" de CHARMM a été utilisé. Ce fichier comporte la myoglobine présolvatée dans une boîte cubique de 55.5 Å de côté contenant 4958 molécules d'eau, ainsi qu'une molécule de monoxyde de carbone (CO) localisée dans la poche distale au dessus de l'hème. L'histidine proximale 93 est liée au fer de l'hème tandis que la molécule de CO est laissée libre. L'équilibration de 1 ns ainsi que les trois simulations de 120 ns ont été produites à l'aide du même protocole que pour les simulations du lysozyme décrites ci-dessus.

J'ai également produit une trajectoire de 10 ns de la protéine d'enveloppe du virus de la dengue (entrée PDB : 1OKE). Le système est solvaté dans une boîte de $180 \times 90 \times 85$ contenant 39 852 molécules d'eau. Le système a été équilibré et simulé de la même façon que pour le lysozyme, à

l'exception de la température (300K) et de la constante de couplage de Langevin ($1\,ps^{-1}$).

Les deux trajectoires de 10 ns de la subtilisine 1 de *P. vivax* (entrée PDB : 4TR2) ont été calculées à l'aide du programme NAMD[97]. Le système a été solvaté dans une boîte périodique de $75 \times 65 \times 62$ Å contenant 8129 molécules d'eau ainsi qu'un ion sodium. Les paramètres utilisés pour l'équilibration et la production de cette trajectoire sont identiques à ceux utilisés pour le lysozyme et la myoglobine.

Une trajectoire de 200 ns a été calculée pour la proteine ABL1 à partir de la structure provenant de la DUD-E[67] avec une version GPU de CHARMM[96], en utilisant la decomposition en domaine (DOMDEC). Le système a été solvaté dans une boîte périodique de taille $76 \times 62 \times 52$ Å contenant 6385 molécules d'eau, 18 ions chlorures et 27 ions sodiums. Les paramètres, champ de force et protocoles de conditionnement sont similaires à ceux des simulations précédentes, en utilisant cependant un pas d'integration de 2 fs, et une température de 298 K. Les structures ont été sauvegardées toutes les 20 ps (10 000 structures).

Enfin, j'ai réutilisé les 10 dernières nanosecondes d'une trajectoire de 15 ns du complexe EF-calmoduline lié à deux calciums déjà produite pour une publication antérieure[112].

Pour réaliser les analyses, les molécules d'eau, les ions et autres ligands éventuels (exceptés le CO et l'hème pour la myoglobine) ont été retirés de l'ensemble des trajectoires, et celles-ci ont été alignées par la méthode des moindres carrés sur les atomes lourds de la structure cristallographique (excepté le CO pour la myoglobine). La molécule de CO a été retirée lors du calcul des cavités de la myoglobine.

3 Algorithmes de partitionnement utilisés pendant la thèse

Plusieurs algorithmes de partitionnement ont été utilisés pendant cette thèse, notamment pour le partitionnement des empreintes de cavités au chapitre II, la définition de cavités consensus et des sites de liaison du CO dans la myoglobine au chapitre III, et enfin pour le partitionnement de la ZINC au chapitre IV. La haute dimensionnalité des espaces considérés ainsi que le très grand nombre de points à classer ont nécessité la recherche d'algorithmes pertinents en terme de résultats, d'empreinte mémoire et de temps de calcul. Les principes de ces algorithmes sont présentés ici. Dans cette section, on notera D la matrice des n descripteurs (également appelés "points") de dimension d. Le but, en règle général, sera de déterminer une partition de l'espace en k groupes, selon différents critères, k pouvant être choisi par l'utilisateur ou déterminé de façon automatique en fonction des données.

3.1 Algorithme des *k-moyennes* et *k-médoïdes*

L'algorithme des k-moyennes est largement utilisé dans de multiples domaines faisant intervenir un besoin de classification. Cet algorithme a l'avantage d'être très performant en terme de vitesse de calcul. Le principe de l'algorithme est d'avancer itérativement vers une partition de l'espace en k groupes pour lesquels l'inertie intracluster est minimale. Il s'agit d'un algorithme "glouton" qui trouve un minimum local, n'ayant aucune garantie d'être le minimnum global.

En règle général, l'algorithme démarre en choisissant k points aléatoires parmi les n possibles pour définir le centre des k partitions. L'algorithme des k-moyennes consiste alors à réaliser itérativement les étapes suivantes :

1. Chacun des n points est placé dans la partition dont le centre lui est le plus proche.
2. La position des centres des k partitions est calculée (centre géométrique)

La convergence est atteinte quand les partitions n'ont pas changé entre deux itérations de l'algorithme (figure VI.1).

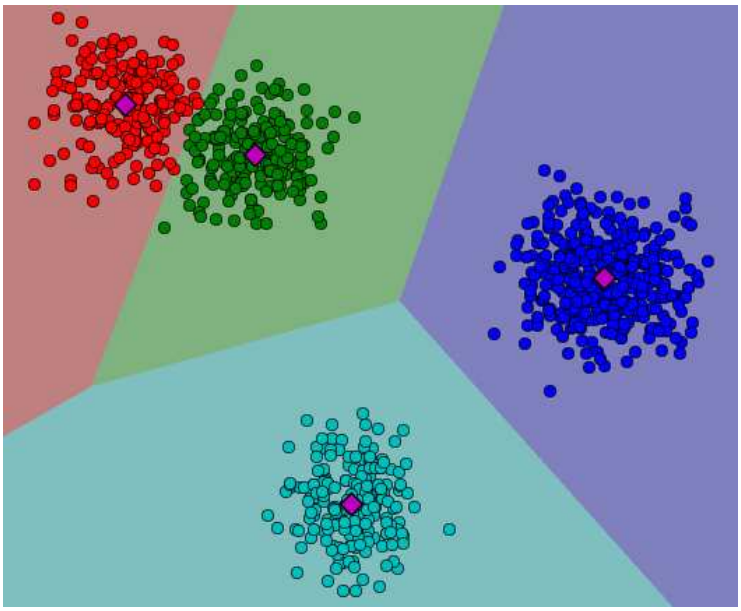


Figure VI.1 – Résultat d'un partitionnement des k-moyennes ($k = 4$) d'un ensemble de points en deux dimensions. Les centres géométriques de chaque partition sont indiqués par un losange violet. Les cellules de Voronoi de chaque partition sont indiquées par le fond coloré.

Une fois les k centres initiaux déterminés, l'algorithme converge relativement rapidement. L'algorithme des k-moyennes est malheureusement très sensible à la sélection initiale, et il est donc nécessaire de lancer l'algorithme un grand nombre de fois en faisant varier cette sélection initiale et en gardant la solution dont l'inertie est la plus faible.

L'algorithme des k-moyennes est simple d'implémentation et extrêmement rapide d'exécution. Il est adapté à la partition rapide de données massives en grandes dimensions. Sa variante des k-médoïdes est idéale pour déterminer k éléments consensus dans un groupe. Malheureusement,

la solution trouvée par les k-moyennes n'est jamais garantie d'être un minimum global. Le partitionnement effectué est également purement géométrique (cellules de Voronoi) et ne prend donc pas du tout en compte la topologie des données d'entrées.

Variante des k-médoïdes. Une variante de l'algorithme des k-moyennes permet de s'affranchir de l'utilisation de la distance euclidienne. Pour cela, au lieu de calculer un centroïde lors de l'étape 2, c'est le médoïde de la partition (l'élément le plus proche des autres éléments en moyenne) qui est sélectionné.

3.2 Regroupement hiérarchique

Les algorithmes de regroupement hiérarchique sont une famille d'algorithmes regroupant des classes d'éléments deux par deux de façon itérative pour former un arbre hiérarchique. L'algorithme démarre en définissant une classe différente pour chacun des éléments du jeu d'entrée, et la fusion de deux classes forme une nouvelle classe utilisable à l'itération suivante (figure VI.2.a). Les étapes de groupement peuvent donc se poursuivre jusqu'à atteindre une classe unique contenant tous les éléments. Le nombre de classe diminuant de 1 à chaque étape, l'algorithme prend n itérations avant d'atteindre la classe unique globale. Pour obtenir une classification en k classes, il faut donc réaliser $n - k$ itérations de l'algorithme.

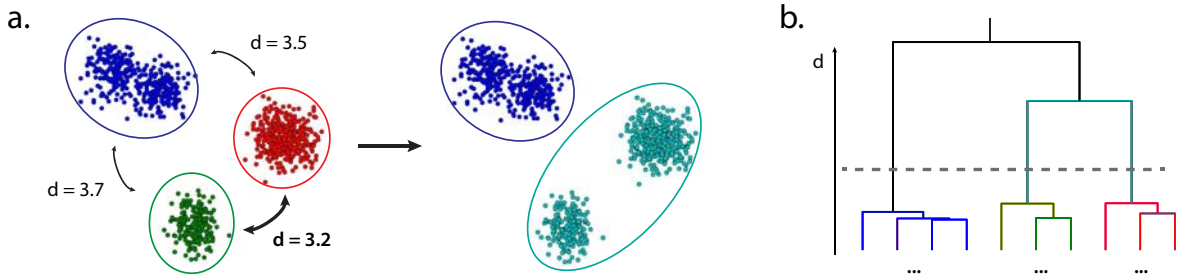


Figure VI.2 – Partitionnement hiérarchique d'un ensemble de points en deux dimensions. a. Etape de groupement de deux classes. Les classes bleu, rouge et verte obtenues à l'itération précédente sont comparées. La paire de classes dont la distance est minimale est la paire rouge-vert, ces deux classes sont donc fusionnées pour former la classe cyan. **b.** Arbre du partitionnement hiérarchique. Les classes sont fusionnées deux à deux jusqu'à atteindre la classe unique (en noir, en haut de l'arbre). La découpe de l'arbre (pointillés gris) permet de définir la partition voulue (ici : bleu, rouge et vert).

Le choix des deux classes à fusionner se fait sur un critère de distance (ou dissimilarité) : à chaque itération, l'algorithme sélectionne les deux classes les plus proches l'une de l'autre. L'utilisateur peut choisir n'importe quelle mesure de distance entre les éléments. Il existe toutefois cinq méthodes principales permettant de combiner ces distances pour comparer deux classes A et B :

1. la liaison *simple* ou *minimale* considère la distance minimale entre les éléments des deux classes : $d(A, B) = \min_{a \in A, b \in B} d(a, b)$
2. la liaison *complète* ou *maximale* considère la distance maximale entre les éléments des deux classes : $d(A, B) = \max_{a \in A, b \in B} d(a, b)$

3. la liaison moyenne ou UPGMA (Unweighted Pair Group Method with Arithmetic Mean) considère la distance moyenne entre les éléments des deux classes : $d(A,B) = \frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b)$ (avec $|A|$ le cardinal de A)
4. la liaison par centroïdes ou UPGMC (Unweighted Pair Group Method with Centroids) considère la distance entre les centroïdes c_A et c_B des deux classes : $d(A,B) = \|c_A - c_B\|$ (uniquement pour la distance euclidienne)
5. la liaison de Ward maximise l'inertie entre deux classes : $d(A,B) = \frac{|A||B|}{|A|+|B|}\|c_A - c_B\|^2$ (uniquement pour la distance euclidienne)

Le partitionnement est réalisé en "coupant" l'arbre ainsi créé (figure VI.2.b). Le nombre de branches au niveau de la coupe correspond ainsi au nombre de partitions. Il existe deux façons de réaliser la coupe : en sélectionnant le nombre de partitions désirées, ou en sélectionnant un niveau de similarité minimal au sein des clusters et en réalisant la coupe à ce niveau.

3.3 Cartes auto-organisatrices (SOM) et SOM émergeantes

3.3.1 Principe général et algorithme d'entraînement

Les cartes auto-organisatrices ont été développées par Kohonen[236]. Le principe est de déformer un maillage de basse dimension (la carte) afin de la faire correspondre à la topologie des données d'entrée. Les vecteurs sont regroupés par proximité : deux vecteurs similaires sont associés à des zones de la carte proche. Les cartes auto-organisatrices sont souvent en deux ou trois dimensions pour pouvoir visualiser facilement cette topologie.

La matrice des descripteurs D est formée de l'ensemble des n vecteurs d'entrée de taille d : $v_1, ..., v_n$. On définit la carte 2D C de dimension $X \times Y \times d$. C est un tenseur de dimension 3, mais il s'agit bien d'un maillage 2D de dimension $X \times Y$ composé de vecteurs de dimension d appelés *neurones*. Les neurones sont initialisés aléatoirement mais sont compris dans l'hyper-boîte englobante de D parallèle aux axes. L'algorithme se déroule alors comme suit pendant T itérations :

1. un vecteur $v_{\sigma(t)}$ est choisi aléatoirement parmi les n vecteurs d'entrée
2. le neurone le plus proche de $v_{\sigma(t)}$ est déterminé (on l'appelle le *BMU*, pour *Best Matching Unit*). On suppose ici que le BMU est situé à la position (x, y) de la carte.
3. la carte est modifiée autour du BMU pour ressembler à $v_{\sigma(t)}$. Cette modification est modulée par une gaussienne centrée en (x, y) :

$$C_{ij}^{t+1} = C_{ij}^{t} + (v_{\sigma(t)} - C_{ij}^{t}) \cdot \alpha_t exp\left(-\frac{(i-x)^2}{\beta_t^2}\right) exp\left(-\frac{(j-y)^2}{\beta_t^2}\right)$$

où t est l'itération courante, et α_t et β_t sont des paramètres d'apprentissage fonction de t (en général des exponentielles décroissantes). A noter qu'il existe une variante dans laquelle la carte est périodique (torique) : le neurone $(1, 1)$ est alors voisin des neurones $(X, 1)$ et

$(1, Y)$. Il suffit pour cela de modifier la gaussienne à l'aide de modulos pour prendre en compte cette pérodicité.

Afin de visualiser la topologie de la carte, on peut calculer sa U-matrix U, matrice de taille $X \times Y$ définie pour chaque neurone comme la moyenne des distances entre le neurone et ses voisins (figure VI.3.a) :

$$U_{ij} = \frac{1}{|V_{ij}|} \sum_{(x,y) \in V_{ij}} d(C_{ij}, C_{xy})$$

où V_{ij} est la liste des neurones voisins de (i, j) et $d(a, b)$ la mesure de distance entre a et b.

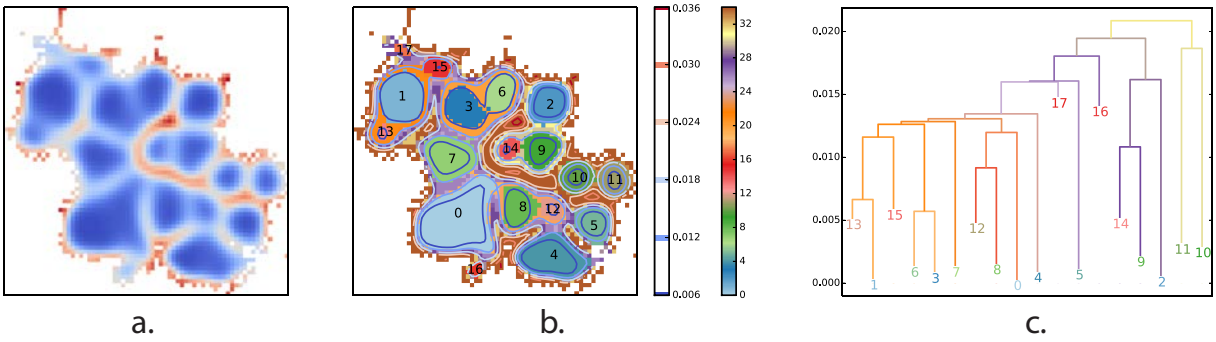


Figure VI.3 – Partitionnement de la U-matrix d'un ensemble d'empreintes de cavités (myoglobine, 100 conformations). **a.** U-matrix torique déployée par remplissage. Le dégradé de couleur va du bleu (valeur de U-matrix faible, neurones très proches) au rouge (valeur de U-matrix élevée, neurones très distants). **b.** Partitionnement hiérarchique de la U-matrix. Chaque neurone est colorié par la valeur du bassin le plus bas. Les feuilles (bassins contenant un minimum local) sont numérotées de 0 à 18. Les couleurs du dégradé au delà de 18 correspondent aux bassins fusionnés. **c.** Dendrogramme correspondant au partitionnement. Les numéros des feuilles sont indiqués à l'emplacement du niveau du minimum de leur bassin. La couleur des numéros des feuilles et des branches de l'arbre correspond aux couleurs de la carte **b**.

3.3.2 Partitionnement

Dans l'article original de Kohonen, chaque neurone définit une partition et les vecteurs d'entrée sont associés à la partition du neurone le plus proche. Lorsque la taille de la carte est suffisament grande (de l'ordre de plusieurs milliers de neurones), des régions peuvent apparaître sur la U-matrix (voir carte page 115). On parle alors de SOMs émergents. Ces zones peuvent être partitionnées à l'aide notamment d'algorithmes de traitement d'images (*waterflooding*) ou de la théorie de Morse.

J'ai ainsi déterminé un algorithme permettant de réaliser le partitionnement hiérarchique d'une carte auto-organisatrice basé sur le principe de *waterflooding*. L'algorithme consiste à faire monter progressivement un niveau d'"eau" λ de la valeur minimale de la carte à sa valeur maximale. A chaque étape, les bassins définis par les zones "immergées" et connexes de la carte définissent des partitions. Lorsqu'un nouveau minimum local est immergé pour la première fois, il définit une nouvelle partition. Lorsque deux bassins se rejoignent au niveau λ, ils sont fusionnés et définissent une nouvelle partition. On garde alors en mémoire cet événement de fusion : le nouveau numéro commun aux deux partitions, les numéros des partitions fusionnées et le niveau λ à partir duquel

les partitions sont fusionnées sont sauvegardés. L'ensemble des états de la carte est également gardé en mémoire. Tous les bassins ont fusionné lorsque le niveau a atteint la valeur maximale de la U-matrix (figure VI.3.b). Il est alors possible de reconstruire un arbre hiérarchique des événements de fusion en fonction du niveau d'eau (figure VI.3.c). Un neurone appartient alors à la partition correspondant au premier bassin dont il a fait partie (figure VI.3.b).

3.4 Partitionnement basé sur la densité spatiale (DBSCAN)

Le partitionnement basé sur la densité spatiale (DBSCAN) a été développé par Ester et al en 1996[237]. Comme son nom l'indique, cet algorithme tente de faire émerger des partitions basées sur la densité de points dans l'espace. Cette densité est évaluée à l'aide d'un critère de distance ε permettant de définir un graphe de contacts (figure VI.4.1). DBSCAN définit ensuite l'ensemble de règles suivant :

— un point p est un point de cœur s'il est connecté à au moins *minPts* points (figure VI.4.2). Les points p_i connectés au point de cœur p sont dit *directement joignables* par p.

— un point q est joignable par un point de cœur p s'il existe une chaîne de points de cœurs $p_1, ..., p_n$ telle que p_1 soit directement joignable par p, q soit directement joignable par p_n et p_{i+1} soit directement joignable par p_i.

— un point de cœur p et l'ensemble des points qui lui sont joignables font partie de la même partition (figure VI.4.3).

— un point b qui n'est joignable par aucun autre point est considéré comme un point de "bruit" et ne fait partie d'aucune partition.

L'ensemble de ces règles permet de définir les différentes partitions (figure VI.4.4).

3.5 Partitionnement spectral de graphe

Le partitionnement spectral de graphe est une méthode de partitionnement apparue dans les années 70 et popularisée au cours des années 90[238, 239, 240]. Dans mon implémentation, je calcule les distances d_{ij} entre chaque paire de points i, j. Je définis ensuite un graphe G dont la matrice d'adjacence A est définie pour chaque paire de point i, j par :

$$A_{ij} = max(0, d_{seuil} - d_{ij})/d_{seuil}$$

La matrice de probabilité des transitions entre chaque sommet du graphe est définie par $P = D^{-1}A$, où D est une matrice diagonale d'éléments diagonaux $D_{ii} = \sum_j A_{ij}$. Cette matrice est ensuite diagonalisée : $\Delta = TPT^{-1}$. Les vecteurs propres T^{pos} de valeurs propres non nulles sont triés par ordre décroissant de valeur propre. Ils définissent une partition du graphe telle que la fréquence de transition d'une partie du graphe au reste du graphe est minimale. Pour définir à quelle partition du graphe appartient un vecteur x, je calcule $A_i^x = max(0, d_{seuil} - d(i, x))$, puis

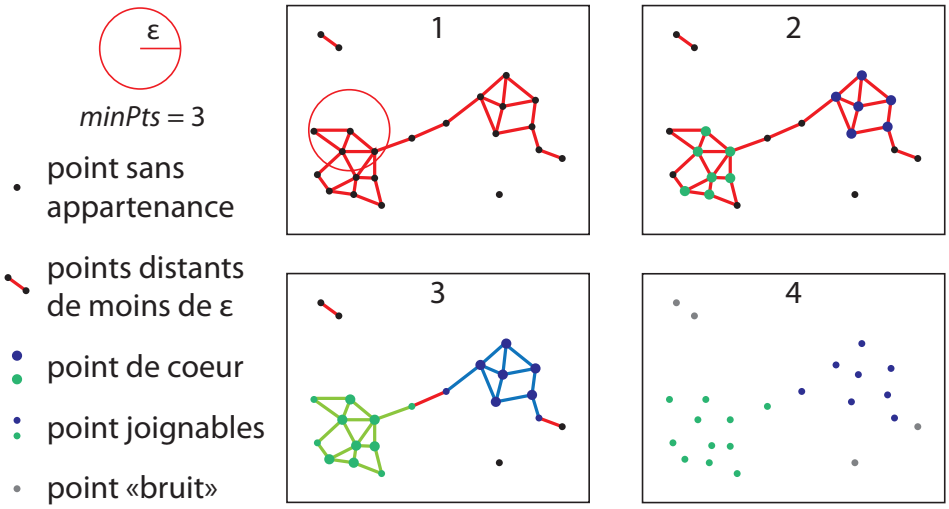


Figure VI.4 – Partitionnement basé sur la densité spatiale. 1. Définition du graphe de contact à partir de la distance ε. Deux points reliés par un trait rouge sont en contact. **2.** Les points de cœur sont définis comme les points en contact avec au moins *minPts* autres points (gros points). Les points de cœur en contact les uns avec les autres font partie de la même partition. **3.** Les points directement en contact avec un ou plusieurs points de cœur font partie de la même partition que les points de cœur correspondant. **4.** La partition de l'ensemble des points inclue les différentes partitions déterminées par les points de cœur (bleu, vert) et une partition "bruit" (gris).

$P^x = D^{-1}A^x$. Le numéro du vecteur propre dont la valeur absolue de la projection sur P^x est maximale est la partition d'appartenance de x : $\text{argmax}_i|P^x T_i^{pos}|$ (figure VI.5.b).

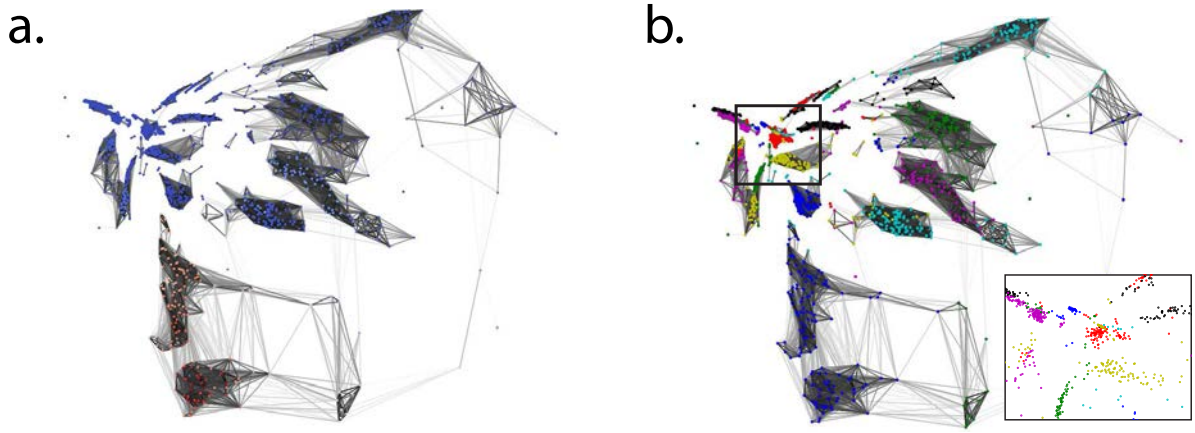


Figure VI.5 – Partitionnement spectral de graphe d'un ensemble d'empreintes de cavités (myoglobine, 100 conformations). Les coordonnées des points correspondent à la projection des empreintes en deux dimensions réalisée à l'aide de l'ACP. Les arêtes du graphe ne sont indiquées que pour les valeurs de A supérieure à 0.5. Leur couleur en niveau de gris correspond à la valeur de A pour cette arête. **a.** Premier vecteur propre du graphe. La couleur des points correspond à la valeur de ce point dans le premier vecteur propre dans une échelle allant du bleu au rouge. **b.** Résultat du partitionnement. La couleur de chaque points correspond au numéro de la partition à laquelle il appartient. Les cycle des couleurs pouvant se répéter, il est possible que deux partitions aient la même couleur. L'encadré en bas à droite correspond à un zoom de la partie entourée, représentée sans les arêtes.

4 Matériel, méthodes et figures supplémentaires pour le chapitre III

4.1 Correspondance des vecteurs et valeurs propres des espaces des descripteurs et des pas de temps

Soit M la matrice des descripteurs centrés ($M = D - C$, M est de taille $n \times s$). On peut définir les deux matrices de corrélation suivantes : $V_{step} = 1/s \cdot M^T M$ et $V_{desc} = 1/s \cdot MM^T$ Soit v_j un vecteur propre de V_{step} de valeur propre λ_j. On peut donc écrire :

$$\begin{aligned} V_{step} \cdot v_j &= \lambda_j v_j \\ M^T \cdot M \cdot v_j &= s\lambda_j v_j \\ M \cdot (M^T \cdot M) \cdot v_j &= s\lambda_j \cdot M \cdot v_j \\ (M \cdot M^T) \cdot (M \cdot v_j) &= s\lambda_j \cdot (M \cdot v_j) \\ V_{desc} \cdot (M \cdot v_j) &= \lambda_j \cdot (M \cdot v_j) \end{aligned}$$

D'où on peut conclure que $M \cdot v_j$ est colinéaire à un vecteur propre de V_{desc}, que l'on appellera N_j. N_j étant normé, on a $N_j = \frac{M \cdot v_j}{\|M \cdot v_j\|}$, et donc :

$$\begin{aligned} \|M \cdot v_j\| &= ((M \cdot v_j)^T \cdot (M \cdot v_j))^{1/2} \\ &= (v_j^T \cdot (M^T \cdot M \cdot v_j))^{1/2} \\ &= (s\lambda_j \cdot v_j^T \cdot v_j)^{1/2} \\ &= (s\lambda_j)^{1/2} \end{aligned}$$

D'où $N_j = (s\lambda_j)^{-1/2} \cdot M \cdot v_j$.

4.2 Interprétation des valeurs de projection sur des composantes principales

La dimension des valeurs de projection est composite : elle contient des valeurs positives et négatives, il est donc impossible de donner un "volume déplacé" équivalent. Il est toutefois possible de comparer ces valeurs à celles d'un vecteur normalisé de déplacement uniforme. Un vecteur dont toutes les valeurs (des points de grilles) vaut 1 correspond à un déplacement de n^{cav} points de grilles. La norme de ce vecteur est $(n^{cav})^{1/2}$, le vecteur normalisé correspondant comprend donc n^{cav} points de grilles ayant tous la valeur $\frac{1}{(n^{cav})^{1/2}}$. Le déplacement total de ce vecteur normalisé

est donc de $(n^{cav})^{1/2}$ points de grilles, soit un volume de $v_n = (n^{cav})^{1/2} \times V_{voxel}$ en Å^3. Ici, nous avons $V_{voxel} = 0.125$ Å^3 et $n^{cav} = 22,880$, le volume de déplacement par "unité de projection" est donc de $v_n \approx 18.9$ Å^3.

4.3 Projection des cavités moyennes de sites de liaison sur les composantes principales de cavités

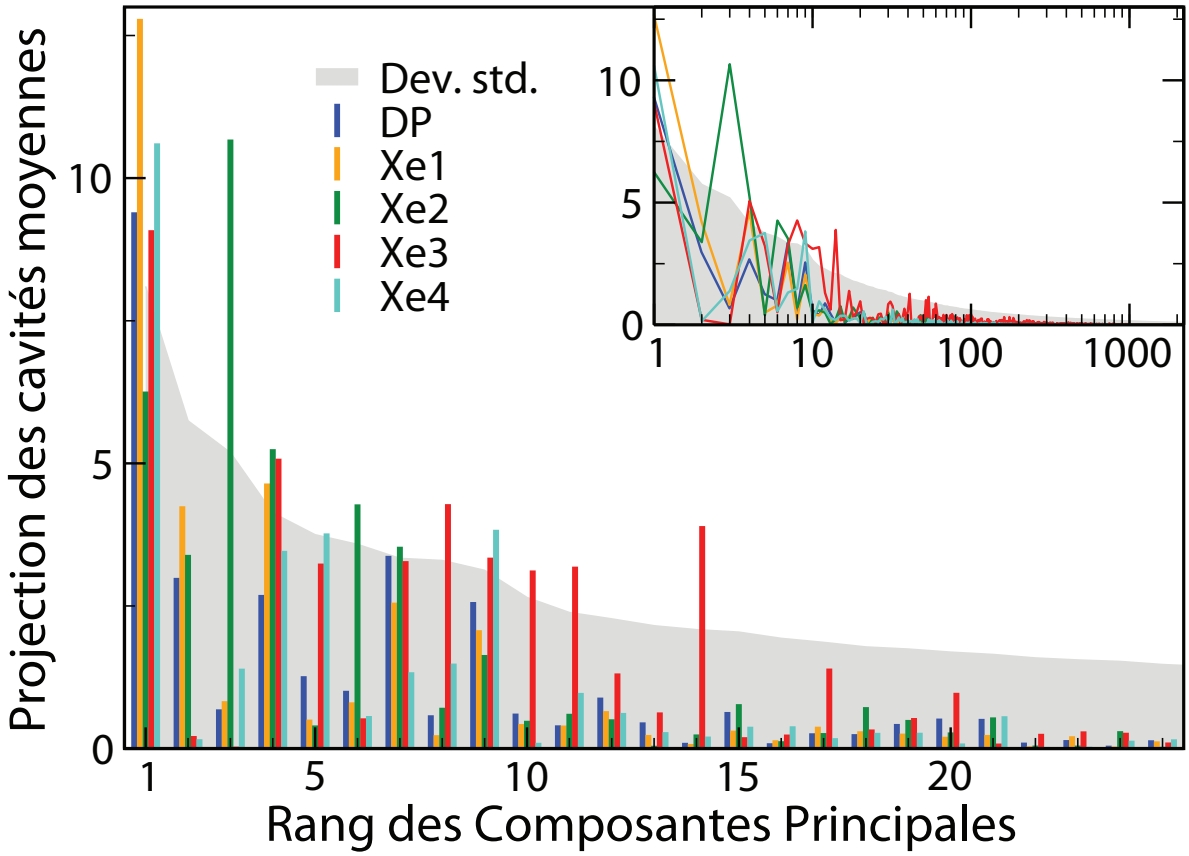


Figure VI.6 – Projection des cavités moyennes par site de liaison sur les composantes principales de cavités. L'abscisse du graphe principal est linéaire, celui du graphe inclu est logarithmique. La projection de la cavité moyenne de la poche distale (DP) est représentée en bleu, Xe1 en jaune, Xe2 en vert, Xe3 en rouge et Xe4 en cyan. La racine de la moyenne des carrés (RMS) des projections des cavités sur chaque composante principale est indiquée en gris.

4.4 Définition des cavités de transfert aléatoires

Je définis $CO(c)$ comme la fonction qui donne le site de liaison i du CO pour la conformation c. La probabilité p_i que le CO soit lié au site i est donc $p_i = 1/n \sum_c \delta_{CO(c),i}$, avec n le nombre de conformations de la trajectoire et δ la fonction identité ($\delta_{i,j} = 1$ si $i = j$, 0 sinon). Je définis ainsi la fréquence de transfert entre deux sites i et j, $f_{i \to j}$, comme suit :

$$f_{i\to j} = p(CO(c+1) = j \mid CO(c) = i) = \frac{\sum_c \delta_{CO(c),i}\delta_{CO(c+1),j}}{\sum_c \delta_{CO(c),i}}$$

Ces fréquences de transfert sont ensuites utilisées pour simuler des processus aléatoires, débutant sur un site i_0 avec une probabilité p_{i_0} et ayant une probabilité de passer d'un site i à un site j égal à $f_{i\to j}$. Au total, 200 processus aléatoires sont générés, tous de taille égale à la trajectoire concaténée de la myoglobine liée au CO. Ces processus aléatoires sont utilisés pour générer des cavités de transfert aléatoires mais obéissant aux fréquences de transfert du CO déterminés par la trajectoire initiale. Elles définissent une hypothèse nulle, à savoir que si les valeurs obtenues avec les cavités de transfert issues de la véritable trajectoire sont sensiblement différentes de celles obtenues avec les cavités de transfert aléatoires, on sait que le véritable transfert du CO d'un site à un autre a une influence notable sur les cavités internes de la myoglobine.

5 Matériel, méthodes et figures supplémentaires pour le chapitre IV

5.1 Dengue : Paramètres de simulation de la dynamique du dimère de la protéine d'enveloppe

La structure cristallographique du dimère (PDB : 1OKE) est utilisée comme structure initiale. Le système est solvaté dans une boîte d'eau de dimension $180 \times 90 \times 80$. Des ions Na^+ et Cl^- sont rajoutés pour neutraliser le système et obtenir une concentration en sel de 150 mM. La protonation des histidines est ajustée à l'aide du logiciel Reduce[241] pour un pH de 7, et vérifiée graphiquement. Le système est minimisé une première fois par descente de gradient en fixant la position des atomes de la protéine. Le champ de force CHARMM22 est utilisé, en utilisant un cutoff à 12 Å et une fonction de switch de 8 à 12 Å. Le système est simulé à 300K à l'aide d'une dynamique de Langevin de coefficient 1 ps^{-1} et un pas de temps de 1 fs. Une première dynamique d'équilibration de 10000 pas (10 ps) est produite, avant de réaliser la dynamique de production de 1 millions de pas (1 ns). Les structures ont été sauvegardées tous les 100 pas (10000 structures au final).

5.2 GLIC : Valeurs des critères de sélection des composés

Les valeurs des critères d'énergie utilisés pour la sélection des représentants sont indiqués pour chaque cavité et chaque logiciel de docking dans le tableau II. Pour être sélectionné, un composé doit satisfaire au moins l'un de ces trois points :

— avoir une énergie FlexX inférieure à la valeur donnée dans la deuxième colonne

— avoir une énergie DOCK (MMGBSA) inférieure à la valeur donnée dans la troisième colonne

— dans le cadre du deuxième criblage, avoir une énergie FlexX inférieure à la valeur donnée dans la quatrième colonne ET une énergie DOCK inférieure à la valeur donnée dans la cinquième colonne

En plus de ces points, la pose du composé doit avoir un taux d'enfouissement e plus élevé que les valeurs données dans la dernière colonne.

	Premier criblage			Deuxième criblage				
						Composé commun		
Cavité	FlexX	DOCK	e	FlexX	DOCK	FlexX	DOCK	e
Enfouie	-15	-26	0.8	-15	-26	-8	-18	0.8
Sous PHE37	-10	-18	0.6	-10	-18	-3	-16	0.7
Au dessus de PHE37	-10	-22	0.7	-10	-22	-3	-18	0.7
Sur le côté	5	-8	0.4	5	-8	0	-3	0.5

Tableau II – Valeurs des critères d'énergie (kcal/mol) et de taux d'enfouissement e utilisés pour la sélection des composés représentants (colonnes 2-4) et des composés finaux (colonnes 5-9).

5.3 GLIC : Composés d'effet faible ou nul sur la fonction de GLIC

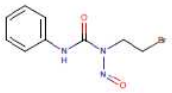
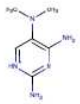
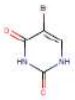
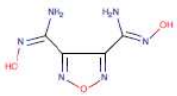
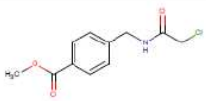
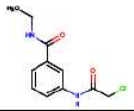
ID	Composé	Cavité ciblée	Concentration testée	Effet sur la fonction	n
NSC79650 **69**		1	100µM 10µM	-8,5% -9%	2+ 1+
NSC38190 **14**		2	500µM 100µM	-10,3% <-2%	2 2
NSC679482 **94**		2	500µM 100µM	-8,1% <-2%	2 1
NSC51035 **42**		2	100µM	-4,4%	1+
NSC19940 **98**		2	500µM 100µM	-4,5% <-2%	2 1
NSC87235 **64**		4	500µM 100µM	<-5% <-2%	2 3
EN300-79721 **36**		4	1mM	-6%	1
STK536828 **107**		2	1mM	-5,3%	1
EN300-01667 **81**		1	1mM	-4,7%	1+
EN300-23207 **21**		1	**1mM**	**-4%**	**1**

Figure VI.7 – Composés testés ayant un faible effet sur la fonction de GLIC.

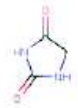
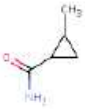
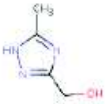
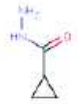
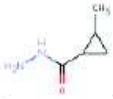
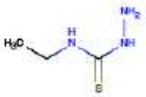
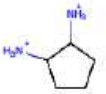
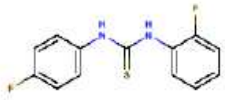
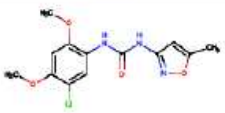
ID	Composé	Cavité ciblée	Concentration testée	Effet sur la fonction	n
EN300-18068 **100**		4	1mM	0	1
EN300-87124 **101**		4	1mM	0	1
EN300-52989 **103**		4	1mM	0	1
EN300-108826 **104**		2	1mM	0	1
EN300-03670 **105**		4	1mM	0	2
EN300-72015 **72**		4	1mM	0	2
STK801323 **88**		4	1mM	<-2%	1
EN300-98951 **60**		4	1mM	<-2%	2
670324 **66**		2	1mM	<-5%	1
NSC760 **39**		3	50µM	0	2
STK130206 **5**		1	340µM	0	1
NSC62921 **4**		4	500µM	<-2%	1
PNU-120,596			10µM	0	2

Figure VI.8 – Composés testés n'ayant pas d'effet sur la fonction de GLIC.

5.4 GLIC : Analogues séléctionnés lors de la deuxième séléction

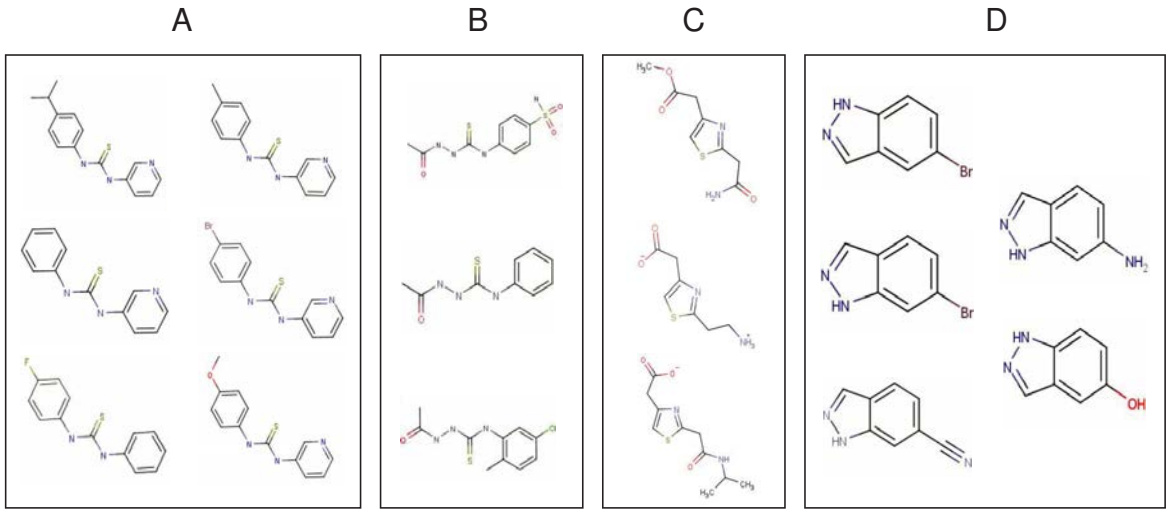


Figure VI.9 – Composés séléctionnés lors de la deuxième étape de séléction des composés (*analoging*). Les classes de chacune des molécules (A, B, C, D) correspondent à l'identifiant de la molécule analogue d'origine (voir chapitre IV, section 3.8)