Développement d'une méthode pour la détection de cibles secondaires de ligands

Inès Rasolohery

Université Paris Diderot

École Doctorale Bio Sorbonne Paris Cité

THÈSE DE DOCTORAT

Discipline : Bioinformatique

présentée par

Inès RASOLOHERY

Développement d'une méthode pour la détection de cibles secondaires de ligands

---

dirigée par Frédéric GUYON et Gautier MOROY

Soutenue le 22 novembre 2016 devant le jury composé de :

Pr. C. ETCHEBEST	Université Paris Diderot	Présidente
Pr. M. MONTÈS	CNAM	Rapporteur
Pr. B. OFFMANN	Université de Nantes	Rapporteur
Dr. J. CHOMILIER	Université Pierre et Marie Curie	Examinateur
Dr. G. ANDRÉ-LEROUX	INRA	Examinatrice
Dr. F. GUYON	Université Paris Diderot	Directeur
Dr. G. MOROY	Université Paris Diderot	Co-directeur

Université Paris Diderot

École Doctorale Bio Sorbonne Paris Cité

Thèse de doctorat

Discipline : Bioinformatique

présentée par

Inès Rasolohery

Développement d'une méthode pour la détection de cibles secondaires de ligands

---

dirigée par Frédéric Guyon et Gautier Moroy

Soutenue le 22 novembre 2016 devant le jury composé de :

Pr. C. Etchebest	Université Paris Diderot	Présidente
Pr. M. Montès	CNAM	Rapporteur
Pr. B. Offmann	Université de Nantes	Rapporteur
Dr. J. Chomilier	Université Pierre et Marie Curie	Examinateur
Dr. G. André-Leroux	INRA	Examinatrice
Dr. F. Guyon	Université Paris Diderot	Directeur
Dr. G. Moroy	Université Paris Diderot	Co-directeur

Remerciements

"À partir de maintenant, j'commence mon ascension
J'ai plus peur du vide, d'affronter la spirale sans fond"
Casseurs Flowters

Je ne réalisais pas la portée de l'aventure dans laquelle je me lançais lorsque je débutais ma thèse, il y a trois ans de cela. La thèse fut une expérience enrichissante dont je sors plus mûre. Il faut du courage pour affronter l'inconnu, mais l'issue de ce périlleux moment est une élévation personnelle sans prix.

Ma profonde reconnaissance va à Frédéric Guyon et Gautier Moroy pour m'avoir donnée cette opportunité de réaliser mes travaux de thèse sous leur co-direction. Vous avez toujours été disponibles pour moi et m'avez poussée à donner le meilleur de moi-même, j'ai énormément appris à vos côtés. Merci Frédéric pour ta bienveillance et ta patience. Gautier, merci pour nos discussions scientifiques et non-scientifiques.

J'adresse ma sincère gratitude à mes rapporteurs Matthieu Montes et Bernard Offmann pour leurs remarques qui m'ont permise d'améliorer ce manuscrit. Je remercie également Catherine Etchebest, Gwenaëlle André-Leroux et Jacques Chomilier d'avoir accepté d'être membres de mon jury de thèse pour évaluer mon travail.

Je remercie Bruno Villoutreix de m'avoir accueillie au sein de l'unité Molécules Thérapeutiques *in silico* et l'université Paris Diderot pour m'avoir financée.

Je remercie très chaleureusement les membres du MT*i* passés et présents. Tout particulièrement, je remercie ceux avec qui j'ai eu le plaisir de partager des discussions, des déjeuners ou des *after-works*. Vous avez contribué à faire de cette thèse une expérience inoubliable. Mes plus profonds remerciements vont à ceux qui ont partagé mon bureau, pour finalement devenir des amis. Merci à Mélaine et Alexandre pour tous vos conseils et pour nos discussions interminables. Ikram, ta présence et ton éternel sourire m'ont aidée à plusieurs reprises. Enfin, je tiens à remercier Dhoha pour sa bonne humeur quotidienne et toutes ses attentions qui furent de précieuses aides.

Un immense merci plus personnel à tous mes amis qui ont su trouver les mots pour m'encourager dans les moments moins évidents et m'offrir des moments de pause.

Merci Thomas. Ton aide, ta patience et ta bonne humeur furent mes moteurs dans cette dernière année.

Je pense bien sûr aussi aux aînés de la famille et cousins qui ont toujours été là. Ces précieux moments familiaux m'ont permise de me ressourcer pour repartir de plus belle.

Papa, Maman et Frank, vous avez été mon soutien infaillible de chaque instant. Pour tout ce que vous m'avez apportée, pour m'avoir inlassablement suivie et supportée, merci. *Ankehitriny, antenaiko fa afaka mijoro sy mirehareha ianareo nahatafita zanaka.*

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

De nombreux médicaments sont connus pour interagir avec des protéines différentes de la cible initiale, ce qui peut entraîner des effets secondaires. Pour certains médicaments, leurs effets secondaires se sont avérés être efficaces pour traiter des pathologies différentes de celles qui étaient ciblées dans un premier temps. L'utilisation de ces médicaments a par la suite été ré-orientée afin d'exploiter leur action dans un contexte plus judicieux. D'autres effets secondaires ont conduit à la restriction de la prescription du médicament selon le cas, voire à l'arrêt total de la prescription. Une autre possibilité qui est aussi explorée est l'utilisation d'un médicament qui a plusieurs cibles qui sont toutes connues pour jouer un rôle dans une même pathologie.

Ces actions multiples sont principalement dues à des interactions entre le médicament et des protéines différentes et précisément aux similitudes des zones d'interaction des médicaments sur les protéines avec lesquelles ils interagissent. C'est pourquoi nous nous sommes intéressés aux atomes de la protéine qui sont en surface et proches du ligand. Nous avons défini ces atomes comme étant le patch d'interaction pour un ligand donné.

La recherche d'un patch connu pour interagir avec un médicament, sur une ou plusieurs protéines permettrait de détecter de potentielles cibles protéiques pour ce médicament. En effet, des similitudes structurales entre le patch d'intérêt et la surface d'une protéine ciblée indiqueraient une interaction probable du médicament avec cette protéine. Ainsi, l'identification de protéines présentant des patches similaires

à leur surface permettrait soit d'apporter des éléments d'explications pour les effets secondaires mis en évidence, soit de participer à la conception d'un médicament aux interactions secondaires limitées ou choisies.

La comparaison de sites de liaison est une problématique qui a entraîné le développement d'outils qui utilisent différentes approches pour déterminer si deux sites de liaison sont susceptibles de fixer un même ligand. Toutefois, il n'existe pas de programme permettant de rechercher un patch défini par l'utilisateur, pouvant être un site de liaison ou une partie de surface, sur une surface protéique. Les outils actuellement disponibles permettent de détecter si deux sites de liaison sont susceptibles de se lier à un même ligand, et imposent des ligands et des cavités bien définies.

Dans ce manuscrit, je décris le développement d'une méthode appelée PatchSearch permettant de rechercher des similitudes géométriques et physico-chimiques entre un patch de surface issu d'une protéine et la surface d'une autre protéine. Cette recherche de similitudes se base sur une approche algorithmique originale de recherche de quasi-cliques dans un graphe produit, qui prend en compte les caractéristiques propres à la surface protéique.

Ce manuscrit débute par une présentation générale du médicament, de son parcours au sein de l'organisme ainsi que des interactions multiples qui peuvent exister entre le médicament et les protéines présentes dans l'organisme. Je détaille par la suite l'état de l'art des différentes méthodes qui permettent de comparer les sites de liaison entre eux ou de détecter d'éventuelles similitudes locales entre des protéines.

Dans le chapitre suivant, j'explique les méthodes que nous avons développées pour extraire des patches et/ou des surfaces, ainsi que l'approche algorithmique que nous avons implémentée dans PatchSearch pour identifier des similitudes entre un patch donné et une surface.

Nous avons appliqué PatchSearch dans différents contextes qui sont présentés dans les trois chapitres de résultats qui suivent la présentation de la méthode.

Dans le premier chapitre d'application de PatchSearch, je montre l'intérêt de notre méthode par rapport à celle des cliques qui est plus classiquement utilisée lors de la comparaison structurale.

Ensuite, je présente la comparaison des performances de PatchSearch par rapport aux autres outils dédiés à la détection de similitudes entre sites d'interaction, ainsi que la mise en application de PatchSearch sur des cas de médicaments connus pour se lier à plusieurs protéines.

Chapitre 2

Contexte biologique

2.1 Médicaments : généralités et conception

2.1.1 Histoire des composés thérapeutiques

Comment le principe de molécule à usage thérapeutique a-t-il émergé ?
Au cours de l'Égypte ancienne, la médecine comprenait l'étude du psychisme [104] entre les organes et leur pharmacopée se composait de minéraux, de végétaux, et de produits animaux et humains entre autres [13]. L'association d'éléments naturels avec des propriétés curatives commençait à s'établir.
Plus tard dans l'Antiquité, Hippocrate et Galien ont élaboré vers 138 avant J.-C. la médecine selon laquelle la santé humaine correspond aux quatre fluides corporels ou humeurs : le sang, le phlegme, la bile noire et la bile jaune [42]. D'après leur théorie, une maladie serait le fruit du déséquilibre d'une de ces quatre humeurs. Ce n'est qu'au XVI$^{\text{ème}}$ siècle qu'émergea la notion de médicament pour soigner une maladie, sous l'impulsion de Paracelse [136]. Ce dernier soutenait que tous les phénomènes scientifiques pouvaient être expliqués par l'alchimie [170], soit la science de transmutation des métaux. Il appliqua ainsi certains préceptes alchimistes de métallurgie à la biologie humaine : notamment, le médicament doit être adapté à la maladie du patient, et doit être administré à une dose optimale pour permettre son activité curative. Il affirmait que "tout est poison rien n'est poison : c'est la dose qui fait le poison" [43], ce qui peut se traduire par une relation proportionnelle entre la dose de

médicament administrée et l'effet : une importante dose de médicament entraînera un effet accru du médicament. Cette dernière théorie fut contredite ces dernières années , notamment par l'exemple des perturbateurs endocriniens [165] : dans ce cas, : la relation dose-effet peut ne pas être linéaire, de faibles doses peuvent avoir plus d'impact que des doses plus élevées.

Suite aux théories paracelsiennes, l'idée d'allier la chimie et la physiologie permit l'émergence du principe de la médecine dite chimique. Plus tard, au XVII^{ème} siècle, les études de botanique aboutirent à l'extraction de principes actifs. Un principe actif est une substance présentant des propriétés thérapeutiques, c'est-à-dire visant à guérir ou soulager une maladie donnée. Notamment, les premières substances thérapeutiques à être extraites de végétaux furent la morphine (isolée en 1803 à partir de l'opium [178]) et l'acide salicyclique, extrait de l'écorce de saule au XIX^{ème} siècle, qui sera par la suite modifié et commercialisé sous sa forme la plus connue, l'aspirine [51].

En 1967, le code de la santé publique a défini le médicament de la façon suivante : "toute substance ou composition présentée comme possédant des propriétés curatives ou préventives à l'égard des maladies humaines ou animales, ainsi que tout produit pouvant être administré à l'Homme ou à l'animal en vue d'établir un diagnostic médical ou de restaurer, corriger ou modifier leur fonction organique". Intéressons-nous maintenant plus particulièrement à la seconde partie de la définition du médicament, il est administré à l'Homme, entre autres, dans le but "d'établir un diagnostic médical ou de restaurer, corriger ou modifier leur fonction organique".

D'une manière générale, on prend le médicament dans deux cas de figure :

— pour la prévention d'une maladie. La personne qui prend le médicament n'est pas malade, mais se sait exposée. C'est le cas par exemple des traitements antipaludiques qui sont proposés aux personnes se rendant dans des zones connues pour abriter des moustiques porteurs du parasite du paludisme. La personne prémunit son organisme de molécules thérapeutiques pouvant faire face à une éventuelle maladie qui surviendrait.

— pour la guérison d'une maladie. Le patient présente un état pathologique et se voit administré un traitement médicamenteux afin de revenir à un état sain.

À quoi correspond plus précisément un état pathologique ?
Lorsqu'une fonction de l'organisme est anormale, le corps est dans un état qualifié de pathologique. Un tel état équivaut généralement à un dysfonctionnement d'un ou plusieurs organes impliqués dans une même fonction biologique ou non, ou d'un tissu physiologique. Selon l'ampleur de ses conséquences, les symptômes manifestés par l'organisme peuvent être visibles à l'échelle microscopique, de la cellule par exemple, ou à l'échelle macroscopique, dans ce cas les symptômes sont visibles à l'oeil nu.

Ainsi, suite à la manifestation d'un état pathologique de l'organisme, la prise d'un médicament vise à rétablir les concentrations des différents types de molécules chimiques à des valeurs proches des normales ou alors à stimuler le métabolisme ou l'élimination de substances toxiques, afin que l'organisme recouvre son fonctionnement normal. Toutefois, il est nécessaire de réglementer la conception du médicament afin de maximiser les chances du médicament de remplir cette fonction.

2.1.2 Conception

La conception d'un médicament repose sur 3 approches distinctes [148] :

— la conception basée sur le mécanisme : le médicament cible une molécule spécifique impliquée dans le mécanisme suspecté d'être à la cause du dysfonctionnement.

— la conception basée sur la fonction : le médicament a pour but de rétablir la fonction biologique déréglée à son niveau normal. Ceci implique qu'il puisse agir à différents niveaux ou dans plusieurs mécanismes différents.

— la conception basée sur la physiologie du patient : le médicament doit avoir un effet sur les symptômes de la pathologie.

La première voie de conception des médicaments, parmi celles évoquées précédemment, est la plus choisie si les connaissances concernant le mécanisme ciblé ont été approfondies. Les effecteurs moléculaires (en grande majorité des protéines) impliqués dans le mécanisme en question ont été étudiés expérimentalement et la

plupart de leurs interactions ont été déterminées ou sont du moins étudiées. Néanmoins, il est important de modérer cette dernière affirmation : le médicament peut atteindre sa cible mais il peut également avoir une action sur d'autres protéines ou molécules non ciblées. Ce point sera traité dans la prochaine partie 2.3. Il est important de préciser qu'un médicament est efficace s'il parvient à atteindre sa cible dans le corps humain.

En France, la conception du médicament est régie par différentes étapes, dans le but d'optimiser son efficacité.
Les différentes étapes de la conception du médicament sont résumées dans la Figure 2.1 :

— la recherche fondamentale, approfondie par la suite en recherche appliquée : les différentes méthodes de recherches visent à approfondir au mieux les mécanismes que l'on souhaite cibler. Cette phase correspond à la phase de recherche et à l'identification expérimentale des protéines à cibler (Figure 2.1).

  Les approches bioinformatiques permettent d'enrichir les connaissances ou d'émettre des hypothèses sur le fonctionnement des protéines concernées. Si les structures des protéines en question ont été résolues, il est possible d'étudier les différences structurales entre les formes normale et pathologique d'une même protéine, voire de simuler l'association entre la protéine et le médicament. De telles expériences peuvent apporter des éléments de réponse sur les problématiques relatives aux interactions impliquant le médicament et sa cible, ce qui est le but de notre étude. Une fois ces mécanismes élucidés dans la limite des informations biologiques disponibles, la recherche appliquée vise à mettre au point des molécules supposées actives sur les protéines ciblées.

— les essais pré-cliniques : l'efficacité du composé médicamenteux et sa toxicité sont testées sur des animaux.

  C'est également l'occasion d'étudier une première fois le comportement du médicament au sein d'un organisme, même si ce n'est pas l'humain. Les tests de toxicologie permettent de déterminer les doses toxiques du médicament.

— les essais cliniques : le médicament est testé sur l'Homme. Son efficacité, sa toxicité, son métabolisme et la recherche d'éventuels effets secondaires sont

étudiés chez l'humain. Plusieurs phases s'enchaînent lors des essais cliniques :

- la phase I se focalise sur l'étude du devenir du médicament une fois ingéré par l'organisme, jusqu'à son élimination.
- la phase II détermine la dose optimale de médicament à administrer pour que son effet soit atteint et qu'il soit toléré par l'organisme.
- la phase III évalue le rapport entre le bénéfice et la prise de risque tous deux liés au médicament. Elle permet notamment de quantifier les effets secondaires relatifs au médicament, pour enfin aboutir à la conclusion quant à l'apport de bénéfices de ce dernier.

— un dossier d'autorisation de mise sur le marché ou AMM est présenté auprès d'agences d'évaluation du médicament. La qualité, l'efficacité et la toxicité du médicament sont estimées afin de permettre la conception du médicament à des fins commerciales.

— si le médicament obtient l'AMM, il est finalement fabriqué de manière industrielle et peut être administré aux patients.

FIGURE 2.1: **Conception d'un médicament** Schéma représentant les différentes étapes de la conception des médicaments [15].

Notons que lorsque les médicaments sont mis sur le marché, d'autres essais de pharmacovigilance continuent d'avoir lieu afin d'améliorer son efficacité lorsque cela est possible (la phase IV dans la Figure 2.1).

Cet enchaînement d'étapes a été mis en place, entre autres, suite aux effets observés lorsque certains médicaments entraînaient des effets inattendus. Une méthode permettant de prédire en amont des interactions non désirées d'une molécule potentiellement thérapeutique est particulièrement utile et pertinente dans ce contexte.

2.2 Interactions du médicament dans l'organisme

Il est nécessaire d'étudier les différentes transformations que peut subir le médicament avant d'atteindre la protéine ciblée. Je détaillerai donc dans un premier temps les différentes étapes rythmant la progression du médicament dans l'organisme une fois son absorption.

2.2.1 Progression du médicament vers sa cible

Le médicament peut être absorbé entre autres par les voies orales, cutanées, sous-cutanées, respiratoires, rectales ou intraveineuses. La forme d'administration d'un médicament la plus répandue, car la plus pratique et la plus confortable pour le patient est la voie orale. L'étude de la progression du médicament dans l'organisme correspond à déterminer ses propriétés dites ADME (Absorption, Distribution, Métabolisme, Excrétion) : l'absorption par l'organisme, la distribution à l'organe ou au tissu ciblé, le métabolisme et l'excrétion du médicament.

Les propriétés ADME d'un médicament correspondent à son trajet dans le corps humain, une fois qu'il est absorbé par l'organisme : son temps de vie dans le sang, sa distribution vers les cellules ou tissus ciblés, comment il est dégradé ou métabolisé [4]. Dans un premier temps, afin que l'effet pour lequel le médicament a été conçu se réalise, il est essentiel que le principe actif parvienne à sa cible. Dans un second temps, une fois que l'interaction a eu lieu et que l'activité physiologique est revenue à la normale, le principe actif et ses métabolites ne doivent pas persister dans l'organisme et doivent être éliminés. L'optimisation des critères pharmacocinétiques du médicament constitue donc une étape cruciale une fois la molécule médicamenteuse mise au point.

Lors de la conception d'un médicament, il est essentiel de s'assurer qu'il sera absorbé par l'organisme, et ce pour la majorité des patients auxquels il sera administré. Avant d'être distribué au contact de la cellule ou du milieu ciblé *via* la circulation systémique, le médicament peut subir des modifications lors de son passage dans des organes tels que l'estomac, le foie, ou les poumons, c'est ce qu'on appelle un effet de premier passage. Suite à ce potentiel premier passage et à sa résorption

(ou absorption), le médicament se trouve dans la circulation sanguine et peut être distribué auprès de sa cible notamment. Cette quantité de médicament qui atteint la circulation sanguine correspond à la biodisponibilité du médicament.

Tout particulièrement, la forme du médicament doit être optimisée pour permettre son passage dans la circulation sanguine. Afin de passer la membrane des entérocytes (les cellules des parois intestinales), le composé thérapeutique peut soit être pris en charge par un transporteur membranaire (une protéine transmembranaire dont le rôle est d'assurer le passage de molécules du milieu extracellulaire vers le milieu intracellulaire), soit ses caractéristiques physico-chimiques lui permettent de diffuser au travers de la membrane. Par conséquent, dans le cas d'une diffusion à travers la membrane, le médicament doit être avoir une liposolubilité suffisante pour passer au travers des lipides des membranes des entérocytes.

Les caractéristiques du médicament doivent donc être optimisées pour maximiser ses chances d'atteindre sa cible dans l'organisme, en quantité suffisante afin d'être efficace. De plus, elles doivent permettre au médicament de limiter le plus possible les interactions possibles avec des molécules différentes de la cible, ce qui est un cas de figure dans lequel s'inscrit notre étude.

Une fois que le principe actif arrive au contact de sa cible moléculaire, l'interaction entre les deux partenaires s'opère si leur affinité respective est suffisamment forte, permettant ainsi au médicament d'induire l'effet pour lequel il a été conçu. Le plus souvent, le partenaire moléculaire du médicament est une protéine.

D'un point de vue structural, la structure d'un complexe protéine-ligand peut être déterminée soit par diffraction aux rayons X [86] (exemple dans la Figure 2.2), soit par Résonance Magnétique Nucléaire ou RMN [108] ou encore par microscopie électronique à transmission [67]. Une fois que la structure du complexe a été résolue, elle peut être mise à la disposition de la communauté scientifique sous la forme d'un fichier au format ou PDB, qui est mis en ligne sur le site de la Protein Data Bank (PDB) [17].

Dans le cas où le complexe protéine-ligand est inconnu, il est possible de proposer une structure probable du complexe à l'aide d'outils bioinformatiques appropriés, principalement des programmes d'amarrage moléculaire ou *docking* [124] ou

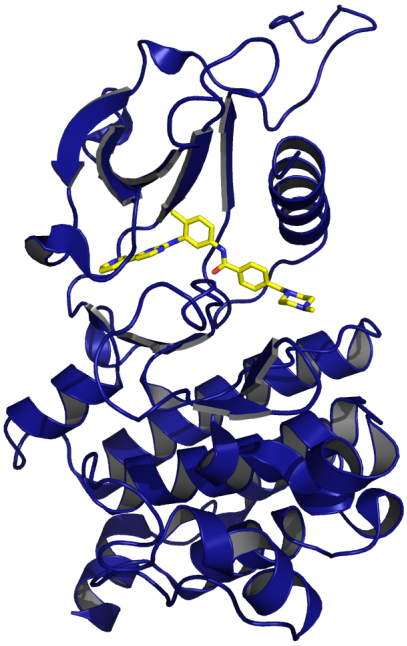


FIGURE 2.2: **Exemple de complexe protéine-ligand.** Est présentée ici la structure cristalline du complexe de la protéine Abl-kinase avec l'imatinib comme ligand (code PDB : 3k5v [188]). La protéine est colorée en bleu et est représentée en rubans. Le ligand en jaune et est représenté en bâtons.

encore de simuler les paramètres de l'interaction à l'aide de modèles associant des descripteurs statistiques à une activité biologique ou modèles QSAR (Quantitative Structure Activity Relationship) [95].

Toutefois, il existe également des méthodes expérimentales pour mettre en évidence l'existence d'interactions entre une protéine et une molécule. Certaines de ces méthodes sont capables d'évaluer quantitativement l'affinité de l'interaction.

— Résonance Plasmonique de Surface (SPR) [149].

— Calorimétrie à Titration Isotherme (ITC) [50].

— Spectroscopie d'absorption [90].

— Mesure de la fluorescence [103].

2.2.2 Le site de liaison : au coeur de l'interaction entre la protéine et le médicament

Les protéines peuvent interagir avec des ligands qui peuvent être de différentes natures, protéines, peptides, acides nucléiques ou petites molécules, à travers leur site de liaison. Ce site correspond à l'ensemble des atomes des résidus de la protéine qui sont en contact avec les atomes du ligand [143, 97]. L'environnement physico-chimique du site de liaison est généralement complémentaire à celle du ligand [10] et met en jeu des interactions hydrophobes, hydrophiles, polaires ou apolaires suivant les propriétés du site de fixation et du ligand.

La forme du site de liaison peut varier : plus en forme de poches ou de cavités [28] ou plus plate, dans le cas des interactions entre deux protéines par exemple [96]. La forme du site de liaison joue un rôle dans la comparaison entre différentes protéines capables d'interagir avec un même ligand. Il est à noter que l'interaction peut être stabilisée à l'aide d'autres molécules tels que des co-facteurs : ce sont des molécules organiques ne faisant pas partie de la protéine qui les aident à assumer leur fonction biologique.

2.3 Interactions multiples

2.3.1 Interactions ciblées et interactions secondaires

En théorie, un médicament est conçu dans le but d'atteindre sa cible et d'effectuer une interaction exclusive avec celle-ci. Mais en plus de cette interaction ciblée, le médicament peut aussi se lier à d'autres protéines de l'organisme. De telles interactions sont qualifiées de secondaires, et aboutissent aux effets qualifiés également de secondaires. Les protéines impliquées dans ces interactions secondaires sont qualifiées de cibles secondaires ou plus communément d'*off-targets*. De plus, il est important de différencier les effets secondaires des effets indésirables. Les effets secondaires peuvent être indésirables mais peuvent également être bénéfiques dans d'autres cas. Comme je l'ai précédemment évoqué en première partie : l'évaluation de l'équilibre entre le bénéfice et le risque encouru de la prise de médicament a lieu au cours des

essais cliniques lors de la conception du médicament. Ces effets secondaires, détectés très tardivement lors de la conception du médicament, sont la cause de nombreux échecs de la mise sur le marché [154].

Certains médicaments ont fait l'objet de procès suite à leur commercialisation à cause de la manifestation d'effets secondaires délétères remarqués chez des patients, le foetus, ou la descendance des patients. Ces effets indésirables sont survenus suite à une ou plusieurs interactions imprévues lors de la conception du médicament.

Un des exemples les plus connus en France, le Médiator © fit l'objet de poursuites judiciaires suite aux effets secondaires qu'il provoquait. Ce médicament fut prescrit de 1976 à 2009. Le composé thérapeutique mis au point par les laboratoires Servier est le benfluorex, et le principe actif est la norfenfluramine. Le Médiator était administré afin de permettre aux personnes atteintes du diabéte de type II. Suite aux effets coupe-faim observés, il fut également prescrit dans le cadre de la perte de poids. [139, 134]. Il a été noté que de nombreux patients qui prenaient ce médicament développaient par la suite des complications cardiaques [132, 22] pouvant engager leur pronostic vital. De tels effets n'étaient pas prévus lors de la conception du Médiator, il est par conséquent question d'effets secondaires indésirables.

2.3.2 Interactions multiples et ré-orientation de médicaments

Dans le cas où les protéines interagissant avec le médicament sont différentes de la protéine ciblée, deux cas de figures se distinguent :

— le ligand n'a pas une cible unique. Sa mise au point a été optimisée pour qu'il agisse sur plusieurs protéines à la fois. C'est ce qu'on appelle des ligands polypharmacologiques [72, 38, 144, 119, 190, 76, 142]. Le ligand peut cibler plusieurs protéines impliquées dans un même mécanisme cellulaire, ou alors plusieurs protéines qui ont une action dans des voies différentes. L'hypothèse de la polypharmacologie est une alternative de la théorie d'après laquelle un médicament interagirait de manière spécifique et exclusive pour sa cible. En effet, les ligands médicamenteux spécifiques pour une protéine unique relèvent plus de l'exception que de la règle [38, 144, 56, 76, 167]. La mise au point de ligands polypharmacologiques a été étudiée pour des pathologies complexes

[21] telles que le cancer [40]. Plusieurs études expérimentales ont montré que cibler plusieurs protéines, avec une affinité moindre, impliquées dans une même voie s'avère plus efficace que de cibler spécifiquement une seule protéine [38, 3]. En effet, en 2005, Csermely P. et Agston V. proposent des modélisations des réseaux d'interaction des voies de signalisations chez la bactérie *E. coli* et *S. cerevisiae*, et démontrent ainsi que dans la plupart des cas, suite à une "attaque" partielle de 3 à 5 cibles dans chacune des voies, les réseaux sont plus impactés que lorsqu'une seule cible est fortement touchée [3]. La probabilité de modifier le fonctionnement de la voie ciblée est plus importante lorsque le ligand est polypharmacologique et interagit donc avec plusieurs protéines de cette voie. Ce phénomène s'explique par l'organisation compensatrice de la machinerie cellulaire : si un seul noeud d'un réseau d'interactions est touché, l'activité manquante sera compensée, tandis qu'il sera plus compliqué de combler l'activité de plusieurs protéines altérées [181].

— le ligand a été mis au point dans le but de cibler une seule protéine. Les interactions avec les autres protéines, ou "*off-targets*" ainsi mises en évidence sont des interactions inattendues.

Un des exemples les plus connus de repositionnement de médicaments est le sildenafil (commercialisé sous le nom de Viagra ©) [20]. Ce médicament était initialement conçu pour traiter les coronaropathies dans les années 1980. Au cours des essais cliniques de phases I et II, le sildenafil ne montrait pas d'effet significatif sur l'angine de poitrine qui est un des symptômes des maladies coronaires. Le principe actif semblait toutefois provoquer un effet secondaire inattendu, l'érection pénienne. En effet, le sildenafil a un effet inhibiteur sur la phosphodiestérase de type 5, qui est un enzyme qui dégrade le guanosine monophosphate cyclique (GMPc). Le GMPc est une molécule impliquée dans le mécanisme de l'érection pénienne [36]. L'administration de sildenafil aux patients atteints de troubles de l'érection permettait ainsi de corriger ces troubles. Le médicament fut par la suite autorisé sur le marché pharmaceutique en 1998 aux Etats-Unis et en 1999 en France. La Figure 2.3 indique les protéines humaines connues pour interagir avec le sildenafil et dont les structures ont été résolues.

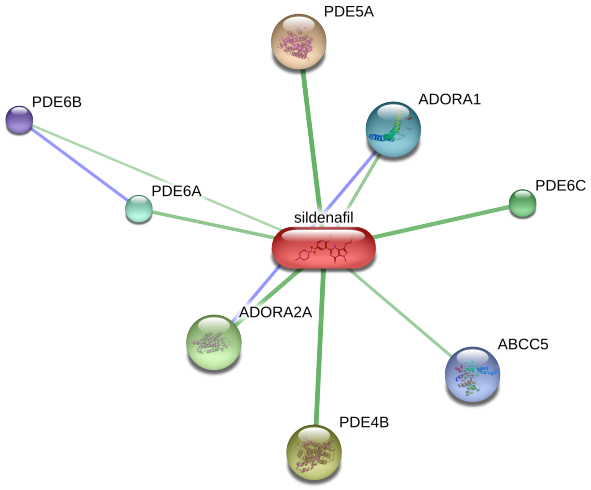


FIGURE 2.3: **Exemple de médicament découvert par sérendipité.** Schéma issu de la base de données STITCH [99, 100]. STITCH répertorie des informations d'interactions connues et prédites entre des composés chimiques et des protéines. Ce schéma représente les différentes interactions réalisées entre le sildenafil et les différentes protéines chez *Homo sapiens*. Plus une interaction est forte, plus le trait la représentant est épais. Les traits bleus représentent des interactions entre deux protéines. Les traits verts représentent des interactions entre le composé chimique et une protéine.

Les effets secondaires sont donc des effets non prévus lors de la conception du médicament. Dans le cas du sildénafil, le principe actif a présenté un effet secondaire qui s'est révélé bénéfique pour traiter des troubles érectiles, ce qui illustre que la notion d'effet indésirable dépend de la condition du patient.

Un autre exemple de médicament qui a été repositionné est le baclofène. Ce médicament a été mis au point en 1974 dans le but de traiter les spasmes musculaires, mais depuis 2014, l'agence nationale de sécurité du médicament ou ANSM française l'a introduit [6] dans le cadre du traitement de la dépendance à l'alcool.

Plus récemment, en août 2016, un autre exemple de repositionnement de médicament a été proposé par Xu M. et ses collègues [182] afin de bloquer la multiplication du virus Zika, essentiellement transmis par le moustique *Aedes aegypti*. Après avoir testé l'efficacité de 6000 molécules thérapeutiques approuvées ou en phase d'essais

cliniques, ils ont mis en évidence deux substances :

— le niclosamide, qui est un médicament inclus dans le traitement du ténia.

— l'emricasan, un composé thérapeutique actuellement en phase d'essai clinique, qui aurait un rôle dans le traitement de la fibrose hépatique.

Ces deux médicaments ont démontré leur efficacité en inhibant la production de particules virales : plus particulièrement, les résultats de l'étude suggèrent que les composés ont une action inhibitrice sur l'infection juste après l'entrée du virus dans la cellule, lors de l'étape de réplication de l'ARN.

Néanmoins, on dénombre également des médicaments présentant des effets secondaires aux conséquences néfastes pour le foetus de la femme enceinte : c'est le cas du thalidomide. A la base, ce médicament fut mis au point pour traiter les troubles du sommeil et les nausées [110]. Il fut commercialisé en 1957, mais en 1961, de nombreux cas de lésions des nerfs et de malformations de bébés nés de femmes ayant pris ce médicament ont été recensés [111]. Des examens poussés sur la procédure de mise sur le marché du médicament ont montré que les tests cliniques et précliniques ont été réalisés sans tenir compte de la possibilité d'effets tératogènes (le médicament augmente le risque de malformation du foetus) chez l'humain [110]. Plus précisément, il a été démontré qu'il avait une action inhibitrice sur la formation de vaisseaux sanguins [39], ce qui conduisait à de telles malformations. Cependant, les études suivantes sur le thalidomide ont montré que ce médicament peut être utilisé pour traiter la lèpre, et qu'il est également efficace dans le traitement des myélomes multiples [141]. Le thalidomide, ainsi que ses dérivés peut être commercialisé, mais ne peut être prescrit aux femmes enceintes. Un autre type de ré-orientation de molécules thérapeutiques est l'approche SOSA ("Sélective Optimization of Side Activities") [173, 172]. La molécule testée semble présenter une activité secondaire d'intérêt : le but de l'approche SOSA est d'optimiser cette activité secondaire, quitte à ce que l'effet primaire recherché soit réduit ou n'ait plus lieu.

2.3.3 Identification de cibles secondaires

Il est actuellement admis qu'une fois absorbées par l'organisme, les molécules thérapeutiques sont fortement susceptibles d'interagir avec plusieurs partenaires, ciblés ou non. Une étude menée sur les réseaux d'interaction par Mestres J. et ses collègues [121] confirme cette affirmation en mettant en évidence qu'en moyenne, un médicament peut interagir avec 6 cibles différentes. L'identification de potentielles cibles secondaires d'un médicament permettrait d'anticiper au mieux les interactions d'un médicament donné avant de l'administrer aux patients. Afin de mieux comprendre et prévoir ces interactions multiples [184, 181, 33, 3], il est nécessaire de rechercher si le site de liaison propre au médicament se trouve également sur d'autres protéines différentes de la cible : les protéines qui présentent un site de liaison similaire à celui de la cible seraient ainsi susceptibles d'effectuer des interactions non désirées avec le médicament. Cette recherche de cibles secondaires nécessite par conséquent de rechercher des similitudes entre cette zone d'interaction protéine-ligand que nous avons définie comme le patch, et les surfaces des autres protéines potentiellement ciblées par la molécule thérapeutique. De nombreux outils ont été développés afin de proposer une comparaison de sites d'interaction, et sont présentés dans le chapitre suivant.

Chapitre 3

État de l'art des différentes méthodes de comparaison de sites de liaison

La zone de la protéine en interaction avec le ligand a des appellations qui diffèrent selon le contexte d'étude :

— le site de liaison : il est constitué des atomes directement impliqués dans la liaison du ligand à la protéine. Le site d'interaction peut être formé d'une poche qui est une région concave, dans laquelle la surface de la protéine forme un creux, ce qui favorise l'interaction avec le ligand. Dans ce contexte, l'intérêt est porté sur la géométrie de la zone d'interaction. L'interaction est favorisée par la complémentarité structurale entre la forme de la poche et celle du ligand.

— le patch : d'une façon générale, c'est un ensemble d'atomes appartenant à la surface de la protéine : ils sont donc exposés au solvant. Dans notre cas, il s'agit plus particulièrement des atomes au voisinage direct du ligand, le patch comprend donc les atomes de la surface impliqués dans l'interaction avec le ligand. Les atomes du patch sont donc plus exposés au solvant par rapport à ceux du site de liaison.

Dans la suite de cette partie, ainsi que du manuscrit, je décris la détermination et la comparaison des zones d'interaction entre une protéine et un ligand, en tant que sites de liaison en général. La comparaison de sites de liaison et de patches permet de déterminer si une protéine est susceptible d'interagir avec un ligand en se basant sur des caractéristiques propres à sa surface.

La différence entre les outils de comparaison de sites de liaison réside dans la façon de déterminer les sites de liaison, ainsi que dans les méthodes de comparaison. Cette partie est dédiée à présenter les différents outils et pour certains d'entre eux, à comprendre les approches mathématiques et algorithmiques utilisées pour la recherche de similitudes.

3.1 Méthodes de détection de sites de liaison

Lorsqu'un site de liaison est étudié, il est nécessaire dans un premier temps de l'extraire de la protéine à laquelle il appartient. Deux cas sont possibles :

— le complexe protéine-ligand a déjà été obtenu. Les résidus impliqués dans l'interaction ont déjà été déterminés de façon expérimentale ou à l'aide du complexe protéine-ligand s'il a été cristallisé.

— le complexe protéine-ligand n'est pas connu. Dans ce cas, il est nécessaire d'utiliser une méthode de prédiction de potentiels sites de liaison sur la structure de la protéine étudiée.

La plupart des méthodes de prédiction de sites de liaison s'appuie sur une représentation géométrique de la protéine. L'approche privilégiée est la recherche de toutes les zones de la structure de la protéine qui auraient une forme de poche, ou de cavité dans laquelle le ligand pourrait éventuellement se fixer. Cette recherche de poches repose donc sur le postulat que le ligand se fixe forcément dans une zone concave de la protéine.

Depuis plus de vingt ans, différentes équipes de recherche se sont attelées à cette tâche de détection de poches de liaison à la surface des protéines.

Levitt D.G. *et al.* [112] sont parmi les premiers à avoir étudié la question en 1992. Leur outil POCKET [112] propose de parcourir les atomes de la protéine dans les différents axes de l'espace à l'aide d'une sphère et ainsi de déterminer si celle-ci se trouve dans une zone de contact avec d'autres atomes de la protéine ou de solvant (non-contact). Ils proposent d'appeler poches les régions dans lesquelles certaines zones de non-contact sont entourées de zones de contact (qu'ils appellent comme

une succession d'événements "PSP", pour "Protein-Solvent-Protein").

Plus tard, en 1997, LIGSITE [68] a été mis au point en reprenant le principe de la recherche d'événement protéine-solvant-protéine. Ils proposent cette fois-ci un espacement de grille plus faible (de 0,75 Å au plus, contre 2,0 Å pour POCKET). Cet outil prédit des cavités protéiques à partir de 2 événements PSP et est indépendant des rotations possibles des poches. Plus récemment, ce logiciel a été mis à jour en prenant en compte la conservation des résidus des poches sous le nom de LIGSITE^{csc}[73].

Une autre méthode de prédiction de cavité protéique a été d'utiliser le principe dit de l'*α-shape* qui consiste à représenter la surface à l'aide des diagrammes de Voronoi et de prédire les poches à l'aide de la triangulation de Delaunay ([113], [138]). Les différents centres des atomes de la surface de la protéine sont des points reliés entre eux, les triangles formés sont étudiés selon leur nature (obtus, si ce sont des triangles des cavités, ou aigu si c'est un triangle dans une cavité).

En 1998, Laskowski R.A. propose l'outil SURFNET [106] qui permet de prédire des cavités en ajustant des sphères entre des atomes donnés de façon à ce qu'elle soit seulement tangente à ces atomes. Le rayon de la sphère est ajusté jusqu'à ce qu'il ne soit pénétré par aucun atome voisin. Si la taille finale de la sphère est supérieure à 1,0 Å, alors la position et le rayon de la sphère sont gardés. L'ensemble des informations relatives aux sphères gardées aboutit à une carte de densités qui permettra par la suite de représenter une cavité.

Nayal M. *et al.* [129] proposent de détecter des poches en faisant rouler deux sphères de rayons différents : 1,4 Å pour obtenir la surface et l'autre de rayon 5,0 Å pour déterminer une enveloppe globale de la protéine. D'après leur définition, une cavité correspond aux atomes de la surface se trouvant à une distance seuil de 2,0 Å sous l'enveloppe globale.

D'autres méthodes de prédiction de sites de liaison les plus connues sont : PASS [23], CastP [19], LigandFit [166], PocketPicker[171] ou encore SiteMap [61] : on parcourt la protéine à l'aide d'une sphère dans plusieurs directions, à chaque nouvel atome de la protéine rencontré par la sphère, son indice d'enfouissement (BI pour

"Buriedness Index") est incrémenté. On étudie les atomes présentant des BI compris entre 16 et 26. Un algorithme de partitionnement ou clustering permet finalement de rassembler les atomes voisins et d'identifier de potentielles poches. Ainsi, la cavité la plus large de la protéine est calculée et est considérée comme un possible patch.

Enfin, d'autres méthodes se basent sur l'énergie d'interaction entre un potentiel ligand et la protéine afin de déterminer l'emplacement d'un possible patch. Q-SiteFinder [107], par exemple, calcule les énergies de van der Waals d'une sonde de méthyle avec la protéine, garde les sondes dont les énergies sont les plus favorables et les rassemble selon leur proximité pour déterminer un potentiel patch. L'outil ODA [47] (pour "Optimal Docking Area") découpe la surface de la protéine en plusieurs parties et calcule leur énergie de desolvatation de docking pour chacune d'entre elles. Celle qui présente l'énergie la plus faible et donc la plus favorable, correspond ainsi à une zone optimale de docking, et par conséquent, à un potentiel patch.

Ces méthodes de recherche de poches s'appuient pour la plupart d'entre elles sur le postulat que la plus grande poche serait potentiellement celle où le ligand pourrait se fixer. Néanmoins, il existe également des poches à l'intérieur de la protéine dans lesquelles le ligand peut être internalisé (par exemple si la protéine se "referme" suite à la fixation du ligand), ou encore il est possible que le patch étudié ne corresponde qu'à une interaction transitoire entre la protéine et le ligand, par conséquent le ligand n'interagirait pas forcément sur une zone concave de la protéine. Ces changements conformationnels n'étant pas toujours faciles à appréhender, nous avons opté délibérément pour une détermination du site de liaison à partir de structures expérimentales de complexes protéine-ligand.

3.2 Méthodes de comparaison des sites de liaison

Une fois que les sites de liaison ont été identifiés, plusieurs méthodes de comparaison sont possibles. Le principe général de la comparaison de sites de liaison s'appuie sur l'appariement entre les deux sites, qui peut reposer uniquement sur les caractéristiques géométriques des sites de liaison, sur les propriétés physico-chimiques, sur les deux, ou encore sur d'autres grandeurs descriptives des sites de liaison ; puis dans

la recherche de similitudes au sein de cet appariement.

3.2.1 Outils utilisant des descripteurs de sites de liaison

En 2007, Kahraman *et al.* [79] développent une méthode représentant chaque site de liaison sous la forme de surface "étoilée", dont les points vérifient les fonctions harmoniques sphériques (principe introduit par Morris R. J. [125]). L'ensemble des normales en ces points par rapport à la surface permet d'obtenir pour chaque surface, un ensemble de coefficients uniques. Le calcul de la similitude entre les poches correspond finalement à la distance euclidienne entre les vecteurs de coefficients de celles-ci.

En 2008, Kupas *et al.* [101] propose de découper les cavités détectées par LIGSITE en sites de liaison circulaires et de résumer leurs caractéristiques physico-chimiques sous la forme de vecteurs de 20 dimensions. Par la suite, un algorithme de Self-organizing maps ou SOM permet d'identifier des vecteurs similaires au sein d'un "paysage" sur les cartes 2D SOM.

Hoffmann *et al.* [71] propose en 2010 un outil Sup-CK qui prend en compte les distributions des masses atomiques des deux nuages de points comparés (qui correspondent aux sites de liaison comparés). D'après leur approche, la distance entre les poches est également la distance entre les distributions des masses du patch_1 et du patch_2. Il propose un deuxième score, Sup-CK_L qui prend en compte les charges partielles atomiques.

3.2.2 Outils utilisant les tables de hachage

Une autre approche permettant la comparaison de sites de liaison entre eux est l'utilisation des tables de hachage. Une table de hachage correspond à un ensemble de valeurs accessibles par des clés qui ont la caractéristique d'être uniques. Dans le cas des comparaisons s'appuyant sur la géométrie, les clés sont des triplets uniques d'atomes et à chaque clé sont associées les distances entre les atomes de ce triplet. Ce type d'approche était utilisé dans un premier temps pour la reconnaissance en vision computationelle [105]. Un même groupe de distances retrouvé entre deux surfaces

comparées est un patch commun. L'algorithme TESS [168] développé en 1997, par exemple, utilise cette méthode.

Pennec X. et *et al.* [137] ont également proposé cette approche en 1998 : ils représentent chaque résidu sous la forme de trièdres (les plans formés par les angles entre les atomes N, C_α et C) et appellent cette représentation une image ; une protéine correspond donc à un ensemble d'images. Ils comparent ainsi deux structures données en calculant, pour chaque paire d'images (une image de la première structure, et une image de la deuxième structure), six poses invariantes issues de paramètres de transformation rigide pour passer d'une image de la poche_1 à la poche_2. Leur table de hachage contient ainsi autant de clés qu'il y a de paires d'images, et à chaque paire d'images, la valeur associée est un vecteur de dimension 6. Leur recherche de similitudes entre les patch_1 et patch_2 s'appuie sur le principe qu'une sous-structure commune présente une transformation similaire entre chaque paire d'images. Suite à un algorithme de partitionnement ou *clustering*, la transformation commune au plus grand groupe de paires est considérée comme étant la plus informative et l'appariement donné par ces paires indique la sous-structure commune.

Cette approche est pertinente à utiliser dans les cas où les caractéristiques géométriques sont fortement conservées. Toutefois, lorsque la similarité entre deux sites de liaison diminue, les performances de cette méthode sont altérées.

3.2.3 Outils effectuant une recherche de clique dans un graphe produit

La comparaison de sites de liaison peut aussi être calculée en appariant les atomes ou les caractéristiques des deux sites de liaison sous la forme d'un graphe produit et en recherchant une clique dans ce graphe. Un graphe contient des noeuds dont certains sont reliés entre eux par des arêtes. Dans un graphe produit, les noeuds représentent des atomes ou des caractéristiques de ceux-ci, tandis que les arêtes représentent des relations de distances conservées entre les atomes des deux sites comparés. Une clique consiste en un sous-graphe du graphe produit dont tous les noeuds sont reliés entre eux. La recherche de la plus grande clique

équivaut donc à identifier un patch commun entre deux sites de liaison comparés et s'effectue généralement avec l'algorithme de Bron-Kerbosch [24] qui est l'algorithme de référence pour identifier les cliques maximales d'un graphe.

En 2000, Kinoshita *et al.* [85] propose l'outil eF-site qui subdivise le patch en triangles et associe à chacun d'entre eux son potentiel électrostatique, son hydropathie, ainsi que le vecteur normal.
Les noeuds du graphe produit correspondent ainsi à l'appariement entre deux triangles de propriétés similaires qui sont quantifiées par un score qui est relatif au potentiel électrostatique et à l'hydropathie de chaque triangle.
Une arête relie deux noeuds lorsque les vecteurs normaux aux triangles sont compatibles, ce qui dépend d'un deuxième score qui est calculé en fonction de la distance, des angles formés et de l'angle dièdre.

En 2003, SuMO [77] propose de présenter les atomes des sites de liaison sous la forme de groupements chimiques et d'effectuer une triangulation sur ces groupements chimiques. Cet outil représente l'appariement entre les deux sites de liaison sous la forme d'un graphe produit dans lequel les noeuds sont des paires de triangles aux propriétés similaires (longueur d'arêtes, densité, orientation et groupements chimiques), et les arêtes représentent des relations d'adjacence entre les triangles comparés. Des sous-graphes avec une forte densité d'arêtes sont équivalents à des paires de triangles appariées, donc à des paires de groupements chimiques identiques et finalement à des sites de liaison similaires.

La même année, Schmitt S. *et al.* utilisent également la recherche de cliques dans leur méthode CavBase [151]. Dans leur graphe produit, chaque noeud correspond à un appariement entre une paire d'atomes de chaque patch aux propriétés physico-chimiques similaires, tandis que deux noeuds sont connectés par une arête si les atomes correspondant dans chacun des sites de liaison sont distants d'au plus 2,0 Å.

Enfin, Najmanovich *et al.* [128] proposent en 2008 le programme IsoCleft dont la comparaison de sites de liaison repose également sur la théorie des graphes, notamment sur la recherche de la clique maximale. Le graphe produit est construit en prenant en compte les caractéristiques physico-chimiques ainsi que les distances

inter-atomiques comme présenté précédemment. La clique maximale est identifiée en deux étapes : dans un premier temps, IsoCleft détecte la clique maximale uniquement parmi le graphe de correspondances effectuées entre les atomes de C_α. L'outil utilise ensuite sa matrice de transformation ainsi que ses vecteurs de translation afin de pouvoir superposer tous les atomes des deux sites de liaison. La deuxième étape consiste à rechercher une fois de plus la clique maximale mais cette fois-ci sur les noeuds du graphe produit qui sont issus d'appariements impliquant tous les atomes sauf les hydrogènes et les C_α.

Plus récemment, l'outil ProBiS a été proposé en 2010 par Konc J. *et al* [89]. Leur algorithme repose également sur la recherche d'une clique maximum dans un graphe produit (les noeuds et arêtes sont respectivement construits selon une similitude de propriétés physico-chimiques et des distances internes aux protéines similaires). La clique maximum du graphe produit est la clique qui contient le plus grand nombre de noeuds possible. Afin d'accélérer la recherche de la clique maximum, ProBiS découpe le graphe produit en plusieurs sous-graphes et recherche la clique maximum dans ces sous-graphes. Finalement, les cliques maximales qui partagent au moins cinq noeuds en commun sont fusionnées : l'appariement final est l'ensemble des atomes présentant des similitudes physico-chimiques et structurales entre les deux surfaces comparées.

3.2.4 Autres méthodes de recherche de sites de liaison similaires

L'outil FINDSITE [26] mis au point par Brylinski M. *et al.* proposent d'exploiter la conservation de structures entre différentes protéines connues : ce programme aligne plusieurs complexes protéine-ligand sur la protéine requête qui ont un taux d'identité de séquence inférieur ou égal à 35% par rapport à la séquence de la protéine requête et la significativité de l'alignement entre deux structures données est quantifiée par un Z-score qui tient compte de la différence d'énergie de cet alignement par rapport au meilleur alignement. Les structures des ligands pris en compte sont celles dont les alignements ont un Z-score supérieur ou égal à 4. un Z-score

≥ 4. Les centres de masses des ligands dans ces complexes sont rassemblés selon leur proximité spatiale. Au final, le centre de chaque groupe de centres de masse correspond à un potentiel site de liaison.

La méthode Patch-Surfer est proposée dans le même but en 2011 par Sael L. *et al.* en 2011 [146, 145] : cet outil divise la poche de liaison en patches de surface, les représente à l'aide des descripteurs 3D Zernike et les compare en utilisant un algorithme pondéré bipartite.

Brylinski M. a également développé la méthode eMatchSite [25] très récemment en 2014, qui représente chaque patch sous la forme d'une table de hachage dont les clés sont les sites de liaison et les valeurs associées sont les distances caractéristiques entre les C-α, les C-β et les centres géométriques des chaînes latérales des résidus regroupés selon leurs caractéristiques physico-chimiques. Le score de similarité correspond au score PocketMatch [185].

Le Tableau 3.1 récapitule les différentes méthodes précédemment détaillées.

3.3 Conclusions et approches choisies

Bien que leur topologie s'y prête, les sites de liaisons ne sont pas formés de poches bien définies. Par exemple, Kahraman A. et ses collègues [79] utilisent les poches prédites par SURFNET mais les considèrent comme trop larges pour être ainsi exploitées. Par conséquent, ils proposent de les réduire en utilisant trois types de modélisations :

— un modèle de cavités qui ne garde que les résidus conservés au cours de l'évolution, c'est le modèle "Conserved".

— un modèle de cavités basé sur les résidus connus pour interagir avec le ligand ou qui pourraient être impliqués dans des liaisons hydrogènes, c'est le modèle "Interact".

— un modèle de cavités qui s'appuie sur les atomes de la protéine en contact avec le ligand, c'est le modèle "Ligand".

Lorsque son outil compare les cavités de chacun de ces modèles entre elles , les résultats obtenus selon le troisième modèle, ou modèle "Ligand", sont meilleurs par

Tableau 3.1: **Outils de comparaison de sites de liaison** Liste des approches *in silico* développées permettant de comparer des sites de liaison protéiques citées dans ce chapitre.

Méthodes	**Année**	**Principe**
TESS [168]	1997	tables de hachage géométriques
Pennec *et al.* [137]	1998	tables de hachage géométriques
eF-site [85]	2000	clique dans un graphe produit
CavBase [151]	2002	clique dans un graphe produit
SuMO [77]	2003	clique dans un graphe produit
Kahraman *et al.* [79]	2007	descripteurs de sites de liaison
Kupas *et al.* [101]	2008	descripteurs physico-chimiques de sites de liaison
IsoCleft [128]	2008	clique dans un graphe produit
ProBiS [89]	2010	clique dans un graphe produit
Sup-CK [71]	2010	descripteurs de masses atomiques
Patch-Surfer [146, 145]	2011	descripteurs 3D Zernike
eMatchSite [25]	2014	table de hachage

rapport à ceux obtenus avec les deux autres (les modèles "Conserved" et "Interact"). Ainsi, les différentes techniques de détermination de poches de liaison indiquent une incertitude sur la détermination des zones d'interaction de la protéine avec le ligand. Cette même difficulté est rencontrée par Sael L. et Kihara D. [146, 145] qui déterminent une poche de liaison en extrayant les atomes de la protéine se trouvant dans les rayons passant par le centre de la poche (détectée par LIGSITE), ou par le centre de masse du ligand lorsqu'il est présent. Cette méthode s'appuie donc soit sur la topologie de la surface des protéines, soit sur le centre de masse du ligand et donc ne permet pas de prendre en compte tous les atomes qui sont au contact du ligand.

D'autres part, les différents outils de prédiction de sites de liaison précédemment

présentés indiquent de potentiels sites de liaison pour de petites molécules, il existe également des outils pour prédire des sites de liaison de peptides par exemple, comme PEP-Site Finder [147]. Il n'existe cependant pas d'outils qui traitent les sites de liaison de différentes natures, petites molécules comme peptides. Ces deux raisons nous ont motivés à opter pour l'étude de la similarité des patches qui sont des structures que l'on peut extraire à partir de la surface de la protéine et du ligand de façon automatisée, qui sont connues expérimentalement pour interagir avec le ligand, . De plus, notre outil est innovant car indépendant de la nature du ligand : nous pouvons ainsi exploiter des patches de peptides, de petites molécules ou encore d'acides nucléiques, ce qui n'a pas été mis au point à ce jour.

Les différentes approches développées au cours des dernières décennies et listées ici témoignent que la recherche de sites de liaison similaires a déjà été longuement étudiée. ProBiS est une des seules méthodes qui propose de comparer un site de liaison d'une protéine donnée à la surface d'une autre protéine. La contrainte de cette méthode est qu'elle nécessite la comparaison des surfaces complètes des deux protéines, ce qui implique des calculs plus longs qu'une recherche d'un patch, qui est une sous-partie d'une surface, sur toute une surface. L'utilisation de cette méthode est également problématique lors de la comparaison de sites de liaison se trouvant à l'interface de domaines protéiques.

Les différentes méthodes précédemment exposées sont limitées sur les points suivants :

— elles ne permettent pas de rechercher si, pour un patch de surface donné, il existe un patch similaire sur une autre protéine.

— elles ne prennent pas explicitement en compte les déformations structurales.

— elles ne permettent pas de rechercher des sites de liaison indépendamment de la nature du ligand.

C'est dans ce contexte que s'inscrit la mise au point de PatchSearch qui s'affranchit de la nature du ligand et prend explicitement en compte la flexibilité de la protéine et est transposable à d'autres protéines tout en assurant des temps de calculs raisonnables.

Le but du travail décrit dans ce manuscrit est de détecter pour un ligand donné, de potentielles cibles. Pour atteindre cet objectif, nous avons choisi de rechercher des similitudes entre un patch extrait à partir du ligand d'intérêt et de rechercher s'il existe un patch similaire dans les surfaces des protéines ciblées. Nous proposons ainsi un outil, PatchSearch, qui recherche des similitudes entre un patch correspondant à un site de liaison ou déterminé par l'utilisateur, et la surface d'une protéine ciblée. Notre méthode s'appuie sur la recherche d'un sous-graphe du graphe produit qui permet de retrouver à la fois un coeur conservé de façon rigide entre le patch et la surface, et les parties moins conservées du patch.

Chapitre 4

Recherche de patches similaires

Le but de notre méthode, PatchSearch, est d'identifier à la surface d'une protéine, un patch similaire à celui donné en requête. Le ligand pourrait être de différentes natures : une petite molécule, un acide nucléique ou encore un peptide. Dans cette étude, nous nous sommes particulièrement concentrés sur les complexes de protéine avec de petites molécules.

Le début de ce chapitre est consacré au processus d'extraction de surfaces et de patches à partir des fichiers PDB. Ensuite, je détaillerai le fonctionnement de l'algorithme de PatchSearch qui procède à la comparaison entre le patch et la surface, dans le but de mettre en évidence d'éventuelles similitudes. Enfin, j'exposerai les outils d'évaluation des patches ainsi identifiés.

4.1 Processus d'extraction des patches et des surfaces protéiques

Dans un premier temps, le patch est extrait à partir d'un complexe protéine-ligand, c'est le patch qualifié de "requête", que l'on souhaite identifier sur la surface d'une protéine cible. Le patch étant lui-même une partie de surface, les étapes d'extraction des atomes de la surface sont donc identiques pour les procédés d'extraction du patch et celui d'extraction de la surface.

Seule la vérification de la proximité des atomes par rapport au ligand est une

étape propre à l'extraction des patches. Aussi, je présenterai d'abord l'extraction de surface, puis je préciserai le déroulement de l'extraction du patch.

L'étape préliminaire à effectuer est le téléchargement du fichier PDB concerné à partir du site de la PDB [17]. Ce format de fichier offre une représentation standardisée des données de structures macromoléculaires qui ont été déterminées soit par diffraction aux rayons X, soit par RMN.

Les différentes étapes de la préparation de la surface d'un fichier PDB sont :

1. visualisation de la structure à l'aide de PyMol [153],
2. ajout des atomes manquants et l'optimisation de la structure,
3. calcul de l'accessibilité relative des atomes de la protéine,
4. extraction du patch ou de la surface,
5. typage des atomes, pour finalement aboutir au fichier final au format PDB.

4.1.1 Visualisation de la structure de la protéine étudiée

Dans un premier temps, nous procédons à un examen visuel de la protéine dont on extrait la surface. Dans la plupart des cas, la surface est extraite d'une seule chaîne protéique ou d'une seule sous-unité. En effet, les sous-unités protéiques étant généralement similaires, nous considérons que si un patch est identifié sur la surface d'un monomère d'une protéine donnée, il sera également potentiellement reconnu sur les autres sous-unités. L'exploitation des données et la recherche de similitudes sur une seule chaîne permet ainsi de réduire le temps de calcul lorsque les appariements sont effectués entre le patch requête et une surface cible. Il est donc nécessaire soit après une visualisation de la protéine, soit en récupérant les identifiants des chaînes de façon automatique, de spécifier la chaîne protéique dont on souhaite extraire la surface. Nous extrayons donc les coordonnées cartésiennes des atomes de la chaîne d'intérêt dans un fichier PDB à part.

La visualisation des structures PDB est effectuée à l'aide du logiciel PyMol.

4.1.2 Ajout des atomes manquants et optimisation de la structure de la protéine

Nous appliquons par la suite l'outil PROPKA [159, 133] (version 3.0) à la structure de la chaîne protéique extraite. Nous utilisons la version installable avec les paramètres suivants : le pH est à 7, le champ de force utilisé est CHARMM, et le format du fichier de sortie est le format PDB de CHARMM. Le but principal de ce programme est de calculer les états de protonation des groupements titrables ainsi que de combler, pour un nombre restreint de résidus, des positions d'atomes manquantes. Dans le cadre de notre étude, nous avons utilisé PROPKA afin d'avoir des résidus complets. Plus précisément, PROPKA détermine dans un premier temps le taux d'atomes lourds manquants par rapport à toute la structure et permet de déterminer des positions atomiques manquantes dans le cas où il y a au plus 10% des atomes lourds manquants. En effet, dans une structure de protéine résolue par diffraction aux rayons X, il peut manquer les coordonnées cartésiennes de certains résidus soit car la protéine a été clivée, soit car ces résidus sont trop mobiles et la densité électronique n'a pu être déterminée autour de ces résidus. La structure de la protéine peut différer selon le pH du milieu dans lequel la protéine a été cristallisée. Lors de l'optimisation des structures protéiques à l'aide de PDB2PQR, nous avons toujours supposé que la structure avait été obtenue à pH neutre, soit 7. Nous supposons ainsi que nous évaluons la reconnaissance des patches recherchés dans ces conditions de pH.

Les fichiers de sorties sont générés par PDB2PQR [44]. Les coordonnées des atomes d'hydrogènes ont également été ajoutées dans les structures ne comportant pas d'hydrogènes. Toutefois, le positionnement de ces atomes manque souvent d'exactitude et leur présence permet une reconnaissance moins spécifique des patches recherchés. Suite à ces constatations, nous avons décidé de retirer les atomes d'hydrogènes des structures issues de PDB2PQR.

4.1.3 Extraction des atomes exposés à la surface de la protéine

Une définition de la surface d'une macromolécule est nécessaire afin de mieux comprendre le calcul de la surface protéique effectué par les logiciels dédiés. En 1980, Kodak *et al.* [176] proposent de définir la surface accessible au solvant comme la surface sur laquelle une molécule d'eau peut être placée en faisant des interactions de van der Waals avec cet atome sans qu'il n'y ait d'encombrement stérique avec les autres atomes de la protéine. Cette définition s'apparente au principe de la détermination des poches protéiques par SURFNET [106]. Nous avions besoin d'un outil téléchargeable afin de pouvoir extraire automatiquement des surfaces sur de larges jeux de données, nous ne pouvions donc pas avoir recours aux serveurs.

Nous avons porté notre choix sur Naccess [74] qui est un outil dédié à la détermination d'accessibilité au solvant au niveau de chaque résidu, et plus particulièrement au niveau de chaque atome. Ce logiciel calcule l'accessibilité au solvant d'une macromolécule donnée grâce à l'algorithme développé par Lee et Richards [109] qui utilise une sphère représentant une molécule d'eau de rayon 1,4 Å, "roulant" sur les atomes de la protéine. Le trajet du centroïde de la sphère permet d'obtenir la surface accessible au solvant de chaque atome de la protéine.

Dans un premier temps, nous avons recherché des seuils d'accessibilité au solvant évoqués dans la littérature. Nous avions arbitrairement fixé le seuil d'ASA relative à 40% (qui était celui déterminé par Trellet *et al.* en 2013 [163] pour déterminer un résidu comme faisant partie de la surface. Nous avons par la suite utilisé des seuils plus faibles trouvés dans la littérature : Zhou *et al.* [189] propose un seuil de 10% de l'aire nominale maximale du résidu, en utilisant le logiciel DSSP, tandis que Yan *et al.* [183] propose le seuil de 25% de l'aire nominale maximale.

Nous avons décidé par la suite d'affiner le seuil d'ASA pour chaque atome. En effet, nous avons souhaité extraire uniquement les atomes constituant la surface, plutôt que les résidus entiers, afin d'avoir une meilleure précision lors de la recherche de similitudes entre le patch donné et la surface ciblée. Les atomes des patches et des surfaces doivent être extraits selon un même seuil d'accessibilité relatif. Toutefois, les

fichiers de sortie de Naccess renseignent sur les accessibilités absolues au solvant de chaque atome et sur l'accessibilité relative au solvant de chaque résidu de la protéine (calculée à partir de valeurs d'accessibilité standard implémentées dans un fichier annexe du logiciel). De toute évidence, les accessibilités absolues des atomes ne nous permettaient pas d'extraire les atomes de façon uniforme selon une unique valeur seuil. Nous avons donc créé une librairie de 20 tripeptides ala-X-ala (X étant le résidu considéré) sur laquelle nous avons calculé les accessibilités absolues maximales des atomes. Nous avions donc pour chaque atome de la protéine son accessibilité dans la protéine et celle dans un résidu, qui correspond donc à une accessibilité maximale en théorie. L'accessibilité relative de chaque atome pour chaque type d'acide aminé est calculée en utilisant la formule 4.1.

$$\text{Accessibilité}_{\text{relative}} = \frac{\text{Accessibilité}_{\text{absolue}} \text{ dans la protéine} \times 100}{\text{Accessibilité}_{\text{absolue}} \text{ dans le résidu}} \tag{4.1}$$

Par la suite, les applications successives de PatchSearch aux différents jeux de données et les améliorations apportées aux programmes d'extraction de surfaces et de patches nous ont amené à diminuer ce seuil d'accessibilité au solvant relative à une valeur de 1% qui nous permet d'obtenir des surfaces et des patches contenant suffisamment d'atomes pour la comparaison entre un patch et une surface.

Nous avons fait des essais notamment avec des patches de certaines ligands complexés avec des protéines issues de la base de données MTLD [33]. Un exemple de ligand choisi est le sunitnib, pour lequel on a calculé la distance moyenne mesurée entre le patch identifié dans une surface donnée et le patch déjà connu dans cette surface lorsque l'on identifie des patches de ce ligand sur des protéines dont les structures en complexe avec ce ligand sont connues. Les complexes issus de la MTLD [33] contenant ce ligand sont 2Y7J, 3TI1 et 3MIY. A partir de chaque complexe, nous avons extrait les patches et les surfaces successivement aux seuils de 1%, 2,5%, 5% et 7,5% d'ASA. Pour chacun des seuils, nous avons recherché chacun des 3 patches sur les surfaces des deux autres protéines connues pour lier le sunitinib et à chaque patch identifié, nous avons mesuré la distance, la Dc, entre le patch identifié et le patch connu dans la surface (je reviendrai sur cette distance plus loin dans ce chapitre).

Le but est d'identifier un patch de sunitinib qui soit le plus proche (en terme de distance euclidienne) possible de celui qui est déjà connu dans la surface ciblée, comme présenté dans le Tableau 4.1.

Seuil d'ASA	**Dc minimale (Å)**	**Dc maximale (Å)**
1%	0,84	3,28
2,5%	1,22	3,68
5%	1,37	10,11
7,5%	1,12	10,70

Tableau 4.1: **Tableau récapitulatif des différents seuils d'ASA testés pour l'extraction des patches et surfaces.** Les patches et les surfaces pour certains complexes de la MTLD [33] ont été extraits suivants les différents seuils d'ASA présentés dans ce tableau. Les distances, Dc, mesurées entre les patches identifiés sur les surfaces des protéines liant le sunitinib et les patches connus sur ces surfaces nous ont guidé pour déterminer quel seuil d'ASA relatif garder.

Comme on peut le voir dans le Tableau 4.1, certains patches extraits avec les seuils de 5% et de 7,5% sont identifiés avec des Dc maximales élevées, qui, comme je l'aborderai plus tard dans cette partie, témoignent d'un patch non retrouvé car éloigné du patch à identifier. Nous avons ainsi décidé d'extraire les atomes des surfaces et des patches avec le seuil le plus fin de 1% afin d'obtenir des surfaces qui contiennent l'intégralité des atomes exposés.

Le taux d'accessibilité relative seuil qui a été gardé pour la suite est de 1% : cela signifie que tous les atomes présentant une accessibilité relative supérieure ou égale à 1% ont été gardés dans le fichier de la surface.

Selon leur implication dans les interactions avec de potentiels ligands et la forme de la surface protéique, certains atomes de carbones ont été traités à part.

Traitement des cycles des résidus aromatiques

Les résidus aromatiques sont connus pour leurs interactions de *stacking* grâce à la géométrie particulière de leurs cycles aromatiques. Toutefois, dans le cas où les carbones de deux cycles aromatiques sont appariés, ces $(6 \times (6 - 1))/2$ soient 15 distances conservées augmenteraient le score de similitude artificiellement alors que ce sont des distances conservées dans les cycles aromatiques de tous les résidus aromatiques (soient le tryptophane, la phénylalanine et la tyrosine) : les appariements entre les atomes de résidus aromatiques seraient donc fortement privilégiés, au détriment des appariements entre d'autres types d'atomes. Par conséquent, nous avons choisi de représenter les cycles aromatiques des résidus aromatiques sous le forme d'un seul atome : leur centre de gravité. Les coordonnées du centre de gravité ou centroïde d'un cycle aromatiques sont calculées comme le milieu de deux atomes opposés, en l'occurence les carbones CD1 et CE2. Le cas du tryptophane est particulier : il comprend un cycle benzénique et un cycle pyrrole (les deux formant un groupement indole). Le centroïde du cycle benzénique est calculé comme étant le milieu entre les atomes CD1 et CD2. Le centroïde du cycle pyrrole est calculé avec les atomes CZ2 et CE3. L'azote du cycle pyrrole est gardé, il peut en effet être impliqué dans une éventuelle interaction polaire avec un ligand.

Les nouveaux atomes ainsi créés sont appelés CG4 et les centroïdes des cycles pyrroles des tryptophanes sont nommés CG5. Les autres atomes appartenant aux chaînes latérales des résidus aromatiques sont traités comme les autres atomes des autres résidus.

Nous gardons le centroïde d'un cycle dans la surface protéique lorsqu'un des atomes de son cycle a une accessibilité relative supérieure à 1%.

Ajout des carbones α

La forme générale de la surface protéique dépend principalement de l'agencement des carbones α. Nous avons donc jugé important de les garder dans la surface protéique extraite. Les carbones α de chaque résidu présentant au moins un atome exposé à la surface sont par conséquent gardés dans le fichier de surface extraite.

Au terme de cette étape, nous obtenons ainsi un fichier PDB contenant les atomes

de la surface protéique. Cette étape ainsi que les précédentes décrites plus haut sont communes à l'extraction de surface et de patch, elles ne seront donc pas décrites dans la section suivante qui est propre à l'extraction du patch.

4.1.4 Extraction du patch

Dans le cas de l'extraction d'un patch, la visualisation de la structure PDB est d'autant plus importante car cela permet de déterminer si :

— le ligand interagit avec une ou plusieurs chaînes protéiques. Si le ligand se trouve entre deux chaînes (Figure 4.1), il est nécessaire d'extraire le patch entre deux chaînes et donc de garder les deux chaînes concernées.

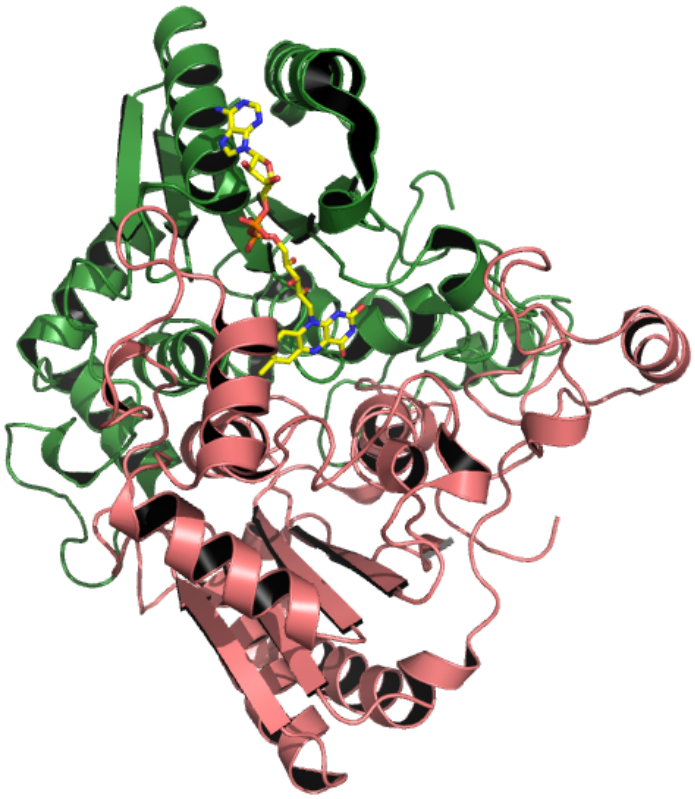


FIGURE 4.1: **Exemple de ligand entre deux chaînes.** Exemple du complexe 1h69 [46] : le FAD (en jaune) se trouve entre les chaînes A (en vert) et C (en saumon) de la NADPH-quinone oxydoréductase. Le ligand est représenté sous forme de bâtons et la protéine en rubans.

— plusieurs molécules du même ligand sont dans le fichier : cela correspond généralement à un excès de molécules du ligand lors de la cristallisation du complexe protéine-ligand, ou d'un enzyme qui présente deux sites de liaison (Figure 4.2). Il est donc essentiel de visualiser les différentes molécules de ligand au contact de la protéine pour choisir celle que l'on souhaite garder pour extraire le patch. Dans ces cas-là, nous gardons l'exemplaire du ligand qui partage le plus d'interactions avec la protéine.

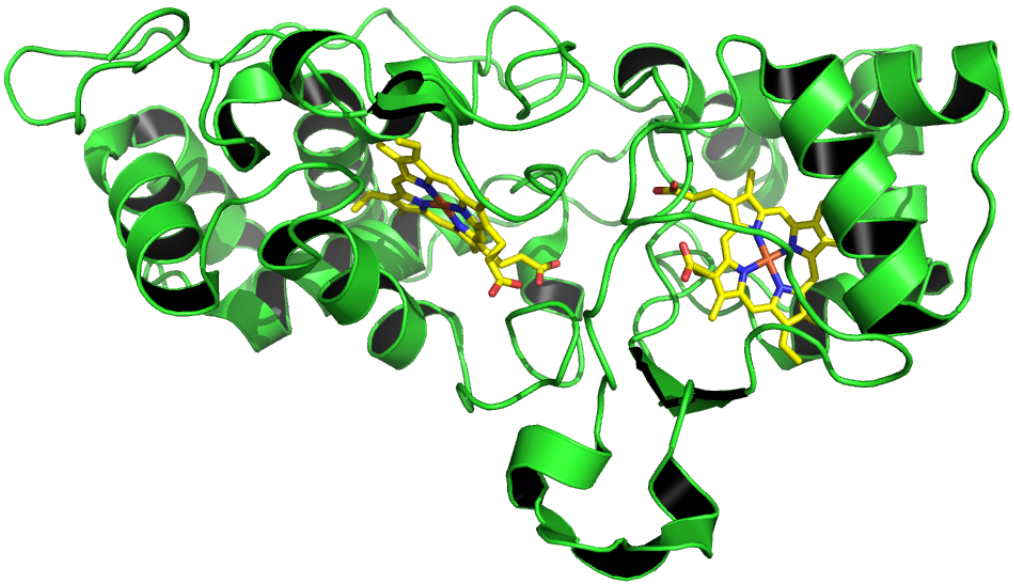


FIGURE 4.2: **Exemple de structure avec plusieurs molécules du même ligand.** Exemple du complexe 1iqc [156] : 2 molécules d'hème (en jaune) se lient sur une même chaîne issue de la di-hème peroxydase.

Deux informations supplémentaires par rapport aux informations à fournir pour extraire la surface (l'identifiant de la structure PDB et la chaîne protéique) sont requises :

— le nom de résidu du ligand à partir duquel on souhaite extraire le patch d'extraction. Tous les ligands ont un nom de 3 lettres dans la colonne assignée au nom des résidus de la protéine (par exemple "GLC" pour le glucose).

— le nom de la chaîne à laquelle appartient le ligand.

La surface protéique est extraite comme décrit dans la section précédente.

Les coordonnées atomiques du ligand sont isolées dans un fichier PDB. Chaque atome de la surface protéique se trouvant dans un rayon de 5 Å autour du ligand est ajouté au fichier de patch. Nous avons fixé ce seuil de distance car, après analyse visuelle des patches, ce seuil nous permet de garder les atomes du patch qui sont impliqués dans l'interaction avec le ligand.

Ce seuil de distance de 5 Å est en accord avec les distances d'extraction de sites de liaison utilisés par les autres méthodes de comparaison de sites de liaison [71, 79, 89] (entre 4 et 5,3 Å). A la fin de cette étape, nous avons un fichier au format PDB contenant les coordonnées atomiques du patch extrait.

Lors de la mise en place de la méthodologie d'extraction des patches, nous avons été confrontés à des cas de figures différents de la méthode classique pour laquelle un seul ligand interagit avec une seule chaîne, comme précédemment décrit.

Extraction sur deux chaînes

Nous avons dû traiter le cas des complexes dans lesquels les ligands se trouvent entre deux chaînes. En effet, les protéines résultent le plus souvent de l'assemblage de chaînes polypeptidiques. Le ligand peut ainsi se trouver entre plusieurs chaînes de la protéine. Les deux chaînes polypeptidiques à garder sont déterminées en visualisant le complexe : il peut s'agir des chaînes présentant le repliement le plus favorable à une interaction avec le ligand. Les deux chaînes sont traitées comme une seule et même suite de résidus dès l'étape d'optimisation de la structure.

Présence de plusieurs ligands

Enfin, nous avons également dû prendre en compte les cas où plusieurs conformations du même ligand interagissent avec une même protéine (comme montré dans la Figure 4.3 par exemple).

Si le facteur d'occupation d'une des deux conformations est supérieur à 0,5, dans ce cas, nous pouvons opter pour cette conformation de ligand (et ne pas tenir compte de la seconde conformation). Par contre, dans le cas où elles sont toutes les deux le même facteur d'occupation, la visualisation de chacune des conformations permettra de déterminer laquelle effectue le plus d'interactions spécifiques avec la

protéine. Dans l'exemple de la Figure 4.3, nous avons gardé la conformation A dans laquelle le ligand est plus orienté vers la transthérétine favorisant ainsi les interactions avec la protéine.

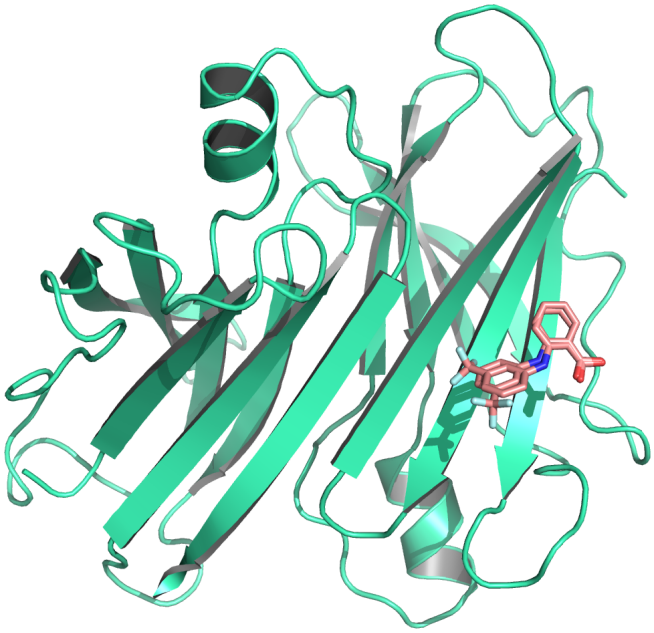


FIGURE 4.3: **Complexe contenant deux conformations du ligand.** Deux conformations pour l'acide flufénamique ou FLF sont en interaction la transthyrétine (Id PDB : 1bm7).

4.1.5 Typage des atomes selon leur nature

Une fois le patch ou la surface extraite, nous obtenons un fichier au format PDB. L'étape de typage qui va suivre est commune aux procédés d'extraction de patches et de surfaces.

La recherche de patches similaires repose précisément sur une géométrie conservée et un ensemble d'atomes de même type que le patch donné en requête.

En effet, si deux protéines peuvent fixer un même ligand, elles présentent donc toutes deux des patches similaires, notamment par leur environnement physico-chimique constitué d'atomes capables d'effectuer des interactions du même type. Nous cherchons donc à exploiter deux types d'informations à partir des patches et des surfaces :

— la configuration spatiale des atomes, indiquée par leurs coordonnées cartésiennes.

— l'implication des atomes dans des interactions connues pour être plus spécifiques, telles que les liaisons hydrogènes et le *stacking*, ou dans le squelette protéique. Ces informations sont obtenues à l'aide du type des atomes donné en dernière colonne dans les fichiers PDB : C, O, N et S.

Néanmoins, ce typage proposé par la PDB n'est pas clair quant à l'implication des carbones dans un cycle aromatique, dans la chaîne latérale ou encore dans le squelette polypeptidique. Par conséquent, nous avons typé ces carbones :

— CA : pour carbone aromatique. Les centroïdes des cycles aromatiques (nommés CG4 et CG5) sont ainsi typés CA.

— Ca : pour carbone α. Les carbones du squelette polypeptidique et porteurs des chaînes latérales sont typés Ca.

— C : pour tous les autres carbones, des chaînes latérales.

Les autres types d'atomes N, O et S proposés par dans les fichiers de la PDB sont gardés.

Ce typage nous permettra par la suite de rechercher précisément des similitudes :

— sur la base des atomes formant le patch : oxygènes, azotes, soufres et centroïdes aromatiques.

— d'assurer une reconnaissance de la conservation de la forme générale du patch requête avec la prise en compte des carbones α.

Ce traitement est appliqué sur les fichiers issus des extractions de patch et de surface.

Au terme de cette étape, nous avons des fichiers PDB dont les structures peuvent être visualisées à l'aide de PyMol (Figures 4.4 et 4.5) et exploitables par PatchSearch.

L'ensemble des étapes permettant de générer un fichier de patch ou un fichier de surface est récapitulé dans la Figure 4.6

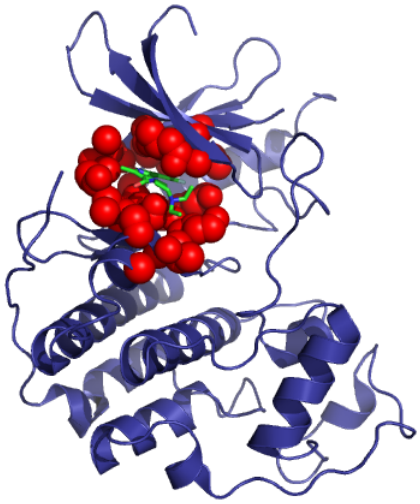


FIGURE 4.4: **Exemple de patch.** Est représenté le patch extrait du complexe 3ti1 (en rouge).

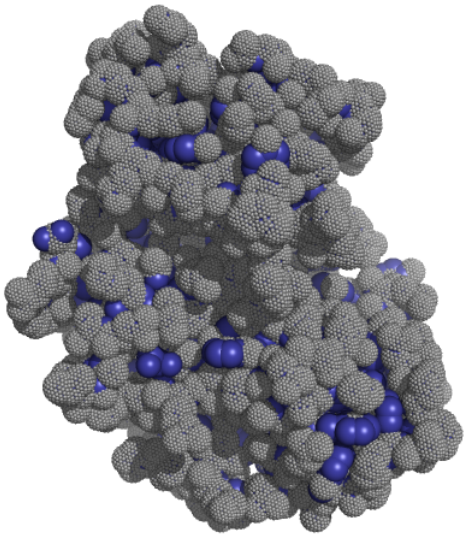


FIGURE 4.5: **Cas de deux ligands sur un même chaîne.** Est représentée la surface extraite sur 3ti1 (en gris), la protéine est en bleue en-dessous.

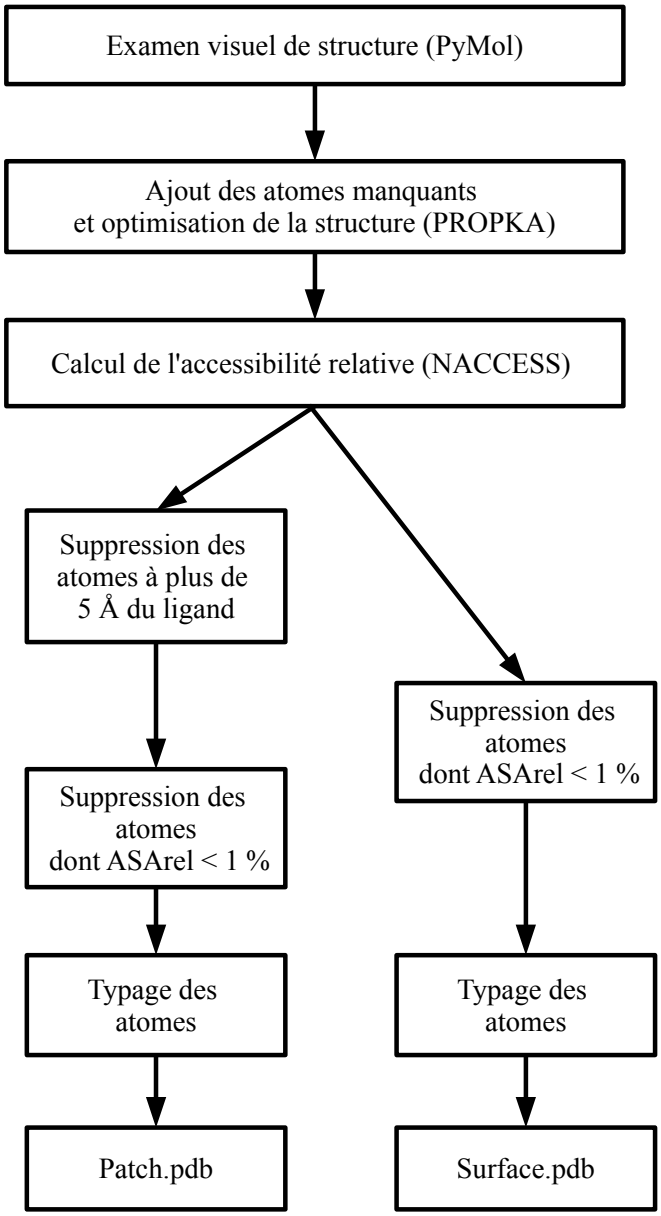


FIGURE 4.6: **Schéma des étapes permettant de générer un patch ou une surface à partir d'un fichier PDB.**

4.2 Algorithme pour comparer un patch à une surface donnée

Le but de PatchSearch est d'identifier, s'il existe, un patch similaire au patch requête sur une surface donnée. PatchSearch peut être utilisé dans deux cas de comparaisons :

— un fichier $\texttt{patch}_{\texttt{A}}\texttt{.pdb}$ (ou le patch requête) et un autre fichier $\texttt{patch}_{\texttt{B}}\texttt{.pdb}$. On souhaite identifier si ces deux patches sont susceptibles d'interagir avec un même ligand.

— un fichier $\texttt{patch}_{\texttt{A}}\texttt{.pdb}$ et un fichier $\texttt{surface}_{\texttt{C}}\texttt{.pdb}$. On souhaite rechercher s'il existe un patch similaire au patch_{A} sur la $\text{surface}_{\text{C}}$ et, par conséquent, si la protéine pourrait interagir avec un même ligand.

La recherche de similitudes dans les deux cas s'effectue de la même façon, seul le calcul du score final diffère. J'exposerai précisément la recherche d'un patch dans une surface donnée grâce à la recherche de la meilleure clique dans le graphe produit, puis je détaillerai également l'algorithme d'extension de la clique trouvée en quasi-clique.

4.2.1 Construction du graphe de correspondances

La recherche de patches similaires nécessite de déterminer s'il existe des configurations spatiales similaires entre les atomes du patch_{A} et ceux de la $\text{surface}_{\text{C}}$, tout en tenant compte des propriétés physico-chimiques des atomes.

Principe du graphe de correspondance

L'algorithme de PatchSearch s'appuie sur la théorie des graphes. Un graphe (Figure 4.7) correspond à un ensemble de sommets, dont certains peuvent être reliés par des arêtes. Les graphes que nous traitons ici sont des graphes simples non orientés, c'est-à-dire que les arêtes vont dans les deux sens.

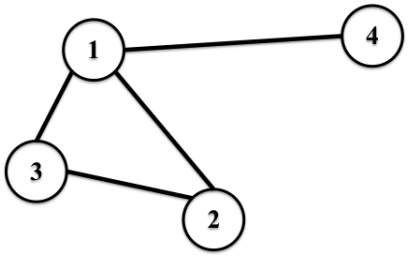


FIGURE 4.7: **Exemple de graphe simple non orienté.** Exemple de graphe composé de 4 sommets et de 4 arêtes.

L'algorithme utilisé par PatchSearch construit un graphe de correspondance ou graphe produit de la façon suivante :

— un noeud représente une correspondance entre deux atomes de même type. Par exemple, un S du patch_A apparié avec un S de la surface_C, c'est le noeud_1 dans le graphe de correspondance. Il en est de même pour le CA du patch_A apparié avec un CA de la surface_C qui correspondent au noeud_2 du graphe produit.

  Nous vérifions que les atomes appariés entre eux appartiennent à des résidus de nature proche en calculant le score de substitution issu de la matrice de substitution BLOSUM62 (pour "BLOcks Substitution Matrix"), [69] du résidu auquel appartient l'atome du patch_A par le résidu auquel appartient l'atome de la surface_C. Si ce score est supérieur ou égal à zéro, les deux atomes sont appariés. Ainsi, des atomes appartenant à des résidus de propriétés physio-chimiques similaires sont appariés :

  — chargé avec chargé

  — aromatique avec aromatique

  — apolaire avec apolaire

— une arête entre deux noeuds représente des distances conservées entre les atomes appariés dans le patch_A et dans la surface_C : si la distance entre le S du patch_A et le CA du patch_A (d_1, voir Figure 4.8) et celle entre le S de la surface_C et le CA de la surface_C ($\text{d}_{1'}$, voir Figure 4.8) sont conservées à un seuil de distance près, alors une arête relie le noeud_1 et le noeud_2. Ce seuil

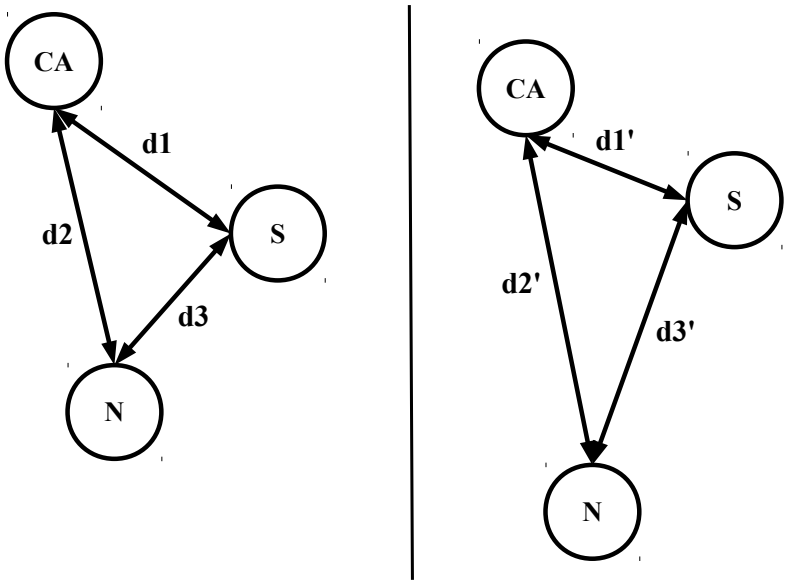


FIGURE 4.8: **Atomes composant le patch$_{\mathbf{A}}$ (à gauche) et la surface$_{\mathbf{C}}$ (à droite).** Les atomes sont représentés dans les cercles et nommés par leur types. Les distances sont annotées "d".

de distance, *mindist*, correspond à la distance minimale entre deux atomes du patch$_{\mathrm{A}}$. Cette distance est très faible, elle varie en général entre 1,1 et 1,2 Å. Il était nécessaire d'imposer un seuil de distance minimal entre les couples d'atomes appariés afin d'éviter les cas d'appariements surjectifs (Figure 4.11) : un atome du patch apparié avec deux atomes de la surface (dit aussi appariement *one-to-many*) et l'inverse (appariement *many-to-one*).

Une représentation du graphe de correspondances entre les atomes du patch$_{\mathrm{A}}$ et ceux de la surface$_{\mathrm{C}}$ lorsque ces appariements sont possibles est proposée dans la Figure 4.9. Un patch similaire au patch$_{\mathrm{A}}$ retrouvé dans la surface$_{\mathrm{C}}$ est un patch qui répond aux deux conditions suivantes :

— ses atomes ont les mêmes propriétés physico-chimiques et appartiennent à des résidus de même nature que ceux du patch$_{\mathrm{A}}$.

— ses distances inter-atomiques sont conservées par rapport à celles du patch$_{\mathrm{A}}$ et sont représentées par les arêtes.

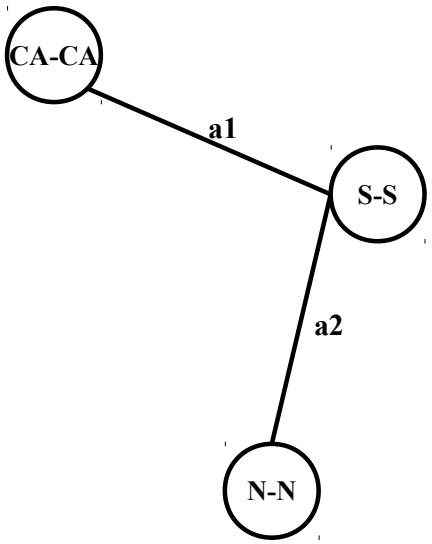


FIGURE 4.9: **Construction du graphe de correspondances entre les atomes du patch$_{\mathbf{A}}$ et ceux de la surface$_{\mathbf{C}}$.** Les cercles représentent des noeuds du graphe de correspondance, construits à partir des appariements possibles entre les atomes du patch$_{\mathrm{A}}$ et ceux de la surface$_{\mathrm{C}}$. Les traits reliant les cercles représentent les arêtes du graphe. Elles sont établies lorsque les distances entre les couples d'atomes appariés sont conservées dans le patch$_{\mathrm{A}}$ et dans la surface$_{\mathrm{C}}$.

Motif d'atomes appariés dans le graphe de correspondance

Il nous a paru plus pertinent de rechercher la conservation de types d'atomes et de distances entre le patch et la surface pour les atomes connus pour jouer un rôle dans l'interaction, plutôt que de prendre directement en compte l'ensemble des atomes. Par conséquent, nous avons choisi de restreindre les noeuds de ce graphe aux atomes typés accepteurs (O) de liaisons hydrogènes, donneurs (N) de liaisons hydrogènes, aromatiques (CA) et aux C_α (Ca). Dans la suite de ce manuscrit, j'emploi le terme de "motif" pour désigner l'ensemble des atomes précédemment cités. Il ne s'agit ni de motifs de séquences ni de motifs structuraux. Ces atomes sont en règle générale directement impliqués dans des interactions dites spécifiques entre le patch et le ligand, ou dans le cas des C_α, influent directement sur la forme du patch. Nous avons donc considéré les atomes précédemment cités comme essentiels à retrouver dans des configurations identiques. Les carbones qui n'entrent pas dans les catégories précédentes (typés C) ne sont pas pris en compte dans un premier temps car nous les considérons comme non impliqués préférentiellement dans des interactions spécifiques.

Injectivité des appariements établis

Nous avons aussi vérifié que chaque noeud du graphe de correspondance couple un seul atome du patch_A avec un seul atome de la surface_C. Il ne peut pas y avoir plusieurs atomes de la surface_C qui s'apparient à un même atome du patch. Chaque atome de la surface_C peut être apparié à un ou zéro atome du patch_A, c'est une fonction injective (Figure 4.10), qui apparie les deux groupes d'atomes entre eux.

Si cet appariement avait été surjectif (Figure 4.11), plusieurs atomes de la surface_C auraient été appariés à un même atome du patch_A et le nombre de similitudes entre les deux surfaces comparées aurait été artificiellement augmenté, sans refléter le degré de similarité entre le patch_A et la surface_C.

Ce premier graphe de correspondance contient donc autant de noeuds que d'appariements injectifs possibles entre les atomes du motif CA, Ca, O, N et S. Les noeuds sont reliés par des arêtes lorsque les distances entre les atomes ainsi appariés

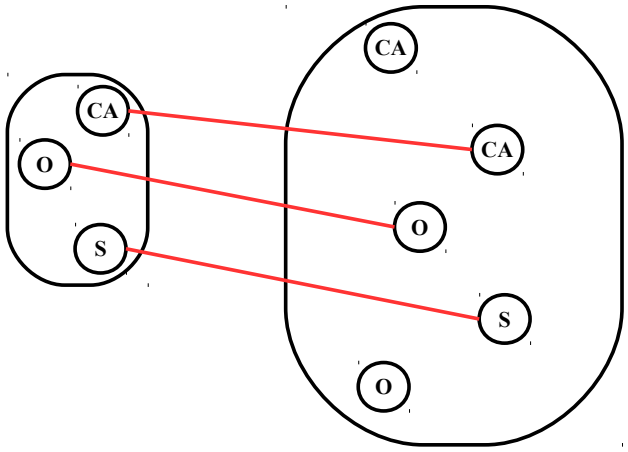


FIGURE 4.10: **Appariement injectif entre un patch et une surface.** Les atomes du patch_A sont représentés à gauche et ceux de la surface_C sont représentés à droite. Chaque trait rouge représente un appariement entre deux atomes.

sont conservées.

4.2.2 Recherche de la meilleure clique

Clique dans le graphe de correspondance

Une fois que la construction du graphe de correspondance est achevée comme dans la Figure 4.12, nous cherchons à garder un sous-graphe particulier : un sous-graphe dans lequel tous les noeuds sont reliés à tous les autres noeuds : c'est un graphe complet, appelé aussi clique, colorée en rouge dans la Figure 4.13.

Si l'on revient au but de la comparaison du patch_A et de la surface_C, nous souhaitons identifier un sous-ensemble d'atomes (qui peut également correspondre dans le meilleur des cas à l'intégralité des atomes du patch_A) de la surface_C présentant des caractéristiques physico-chimiques de même type et des distances inter-atomiques conservées par rapport au patch_A.

Les couples d'atomes appariés dans les noeuds d'une clique déterminée dans le graphe de correspondance précédemment construit :

— correspondent à un environnement physico-chimique et une forme générale du patch et de la surface très conservés

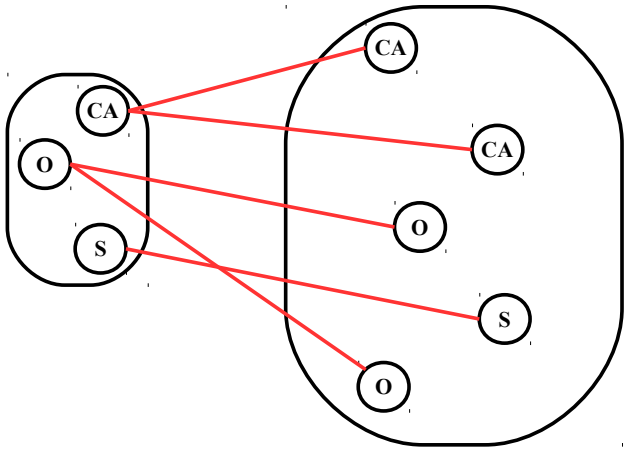


FIGURE 4.11: **Appariement surjectif entre un patch et une surface.** Les atomes du patch_A sont représentés à gauche et ceux de la surface_C sont représentés à droite. Chaque trait rouge représente un appariement entre deux atomes.

— sont dans des configurations similaires

Finalement, trouver une clique dans ce graphe de correspondance correspond à trouver un patch fortement similaire au patch_A dans la surface_C.

Recherche des meilleures cliques

Les cliques maximales du graphe de correspondance sont calculées en utilisant l'algorithme de Bron-Kerbosch [24]. Une clique maximale est une clique dans laquelle il n'est plus possible de rajouter de noeud, sous peine d'obtenir un sous-graphe n'étant plus une clique.

Une fois toutes les cliques calculées, nous avons éliminé celles qui étaient de tailles strictement inférieures à 4 : c'est-à-dire toutes les cliques qui comptent moins de 4 noeuds. Par exemple, la clique présentée dans la Figure 4.14 est gardée, c'est une clique de taille 6 dont tous les noeuds sont reliés à 5 autres noeuds, ils sont donc de degré 5 (le degré d'un sommet étant le nombre d'arêtes adjacentes à ce sommet). Comme nous l'avons noté précédemment, les cliques sont un cas particulier de graphes dans lesquels chaque sommet est relié à tous les autres sommets : par conséquent, tous les sommets ont le même nombre de voisins.

Parmi les cliques calculées, nous avons gardé celles présentant le meilleur BC-

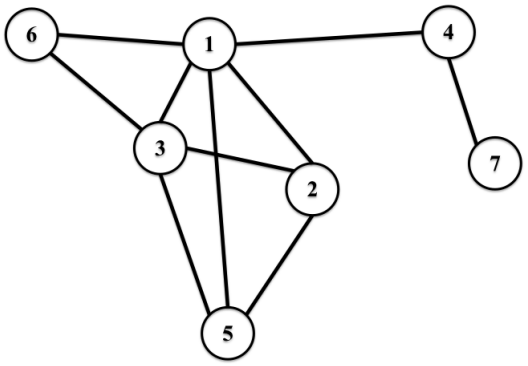


FIGURE 4.12: **Graphe de correspondance final.**

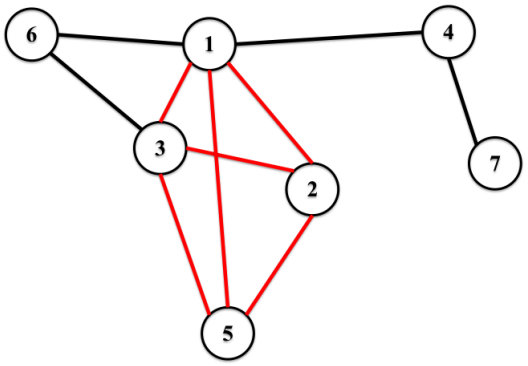


FIGURE 4.13: **Clique dans un graphe de correspondance** Les arêtes de la clique sont colorées en rouge.

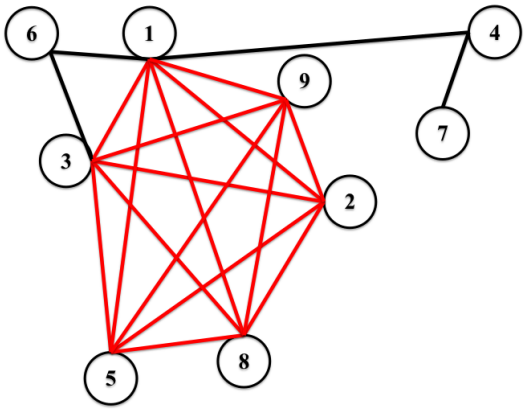


FIGURE 4.14: **Clique de taille 6.** Les arêtes de la clique sont colorés en rouge.

score [59], score basé sur un noyau de Binet-Cauchy qui reflète la similitude entre deux structures et est indépendant de la rotation d'une structure par rapport à l'autre par exemple. Nous avons choisi de garder ce score par rapport à d'autres scores de mesure de similitude structurale connus (tels que le TM-score ou le RMSD) car il est :

— indépendant de la taille des structures comparées.

— insensible aux larges déformations.

— très sensible aux petites différences de structures locales.

Lors du calcul du BC-score entre deux ensembles de points, les coordonnées de tous les tétraèdres possibles sont calculées pour chaque ensemble de points. Par la suite, à chaque ensemble de points correspond le vecteur des volumes signés des tétraèdres entre ses points. Le BC-score est la corrélation entre les deux vecteurs de volumes ainsi obtenus. Le BC-score varie entre -1, lorsque les structures comparées sont les images miroir l'une de l'autre (ces cliques sont supprimées) et 1 lorsque les structures sont identiques. Calculer le BC-score sur les patches appariés est une étape importante : deux patches en miroir l'un de l'autre ont des atomes de nature similaire et des distances internes conservées, mais n'ont pas la même forme. Nous supprimons donc toutes les cliques présentant des BC-scores négatifs.

En terme de structure, nous gardons donc les groupes d'atomes qui ont été appariés sur la base de leur type et des propriétés physico-chimiques des résidus auxquels ils appartiennent, et qui ont au moins 6 distances conservées entre eux. De

plus, le BC-score supérieur à zéro assure qu'il n'y a pas de symétrie entre le patch requête et le patch similaire identifié à présent.

4.2.3 Enrichissement des meilleures cliques en quasi-cliques

Principe de la construction d'une quasi-clique

Le but est d'identifier dans la surface un patch similaire à celui donné en requête aux distances inter-atomiques moins conservées que le coeur rigide. La configuration des atomes de ce patch ainsi identifiée doit rester la plus proche possible de celle du patch donné en requête : c'est la raison pour laquelle nous avons fixé une distance *maxdist* de 3 Å.

Afin de pouvoir choisir le plus judicieusement possible une clique sur laquelle travailler par la suite, nous avons fusionné les cliques qui partageaient au moins 4 noeuds en commun. Nous travaillons ensuite sur le plus grand sous-graphe obtenu au terme de ces fusions, qui est une première quasi-clique [2]..

Nous avons considéré nécessaire de prendre en compte une autre caractéristique des patches protéiques : leur flexibilité. L'algorithme de PatchSearch est ainsi adapté à la géométrie des protéines et à une de leur propriété intrinsèque qu'est la flexibilité.

Les noeuds de la quasi-clique sont rajoutés s'ils répondent aux conditions suivantes :

— chaque noeud est relié à au moins 4 noeuds de la clique initiale (voir la Figure 4.15)

— les appariements de tous les types d'atomes sont pris en compte. En plus des appariements réalisés précédemment avec les atomes CA avec CA, Ca avec Ca, N avec N et O avec N, nous travaillons avec les noeuds qui sont des appariements de C avec des C.

— les noeuds sont cette fois-ci reliés entre eux selon le seuil *maxdist* qui est fixé à 3 Å. Cela signifie qu'on accepte de créer un lien entre deux couples d'atomes appariés si les couples d'atomes concernés sont espacés au plus de 3 Å dans le patch et dans la surface.

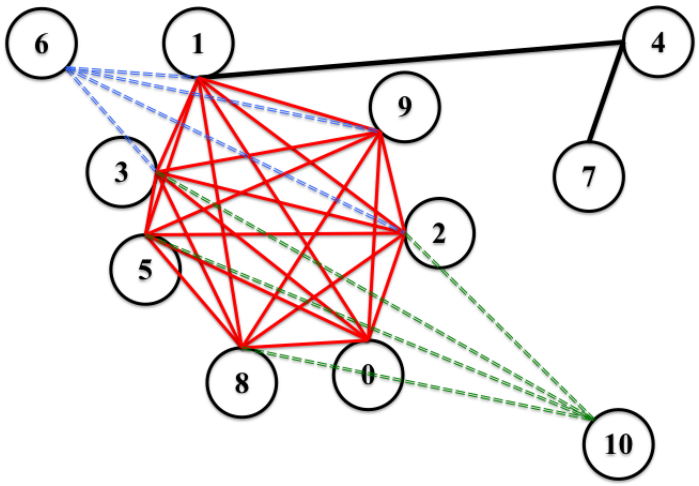


FIGURE 4.15: **Exemple de quasi-clique.** Enrichissement d'une clique en quasi-clique. Les noeuds de la clique sont reliés par des arêtes rouges, tandis que les noeuds ajoutés à la clique sont reliés à au moins 4 noeuds de la clique par des arêtes en pointillées bleues (pour le noeud 6) et vertes (pour le noeud 10).

— les conditions d'établissement d'une arête entre deux noeuds sont identiques à celles appliqués lors de la construction de la clique : comme pour la clique, une arête est rajoutée dans la graphe de correspondance uniquement si elle maintient cette relation univoque d'appariement d'un atome du patch avec un atome de la surface.

Les différentes étapes des appariements et les contraintes de conservation de distances indiquent que le patch trouvé à l'aide de la quasi-clique est similaire à 1,5 Å près (dû au *mindist*) au patch requête au niveau de la partie rigide, et au plus, à 3 Å (dû au *maxdist*) près au niveau des parties moins rigides.

Conditions d'ajout des noeuds dans la quasi-clique

Afin que l'agencement des atomes du patch trouvé soit le plus similaire à celui du patch requête, nous cherchons à diminuer le plus possible la déviation entre les atomes des deux patches. Ainsi, lorsque nous établissons un lien entre deux noeuds, il est important de tendre à diminuer le plus possible la déviation entre les atomes appariés dans le patch et dans la surface. Aussi, dans le cas où un noeud est un candidat pour la quasi-clique soit il répond aux conditions énoncées précédemment,

soit il reste à vérifier que les atomes appariés répondent bien à la règle qui implique qu'un atome du patch ne peut être associé qu'à un seul atome de la surface (voir les appariements injectifs dans la Figure 4.10).

Dans le cas où l'on souhaite ajouter un noeud qui apparie un atome déjà apparié avec un autre atome, nous calculons les déviations maximales dans les deux cas : avec le noeud initial, et avec le nouveau noeud à la place du noeud de base. Nous décidons de garder le noeud qui nous permet d'obtenir une déviation maximale qui soit la plus proche possible de 0 Å.

Nous gardons les quasi-cliques les mieux enrichies, tant en termes de nombres d'atomes et qu'en termes de BC-scores et de RMSD les plus favorables pour obtenir une différence structurale la plus faible possible. Ces deux scores de conservation structurale sont utilisés afin de s'assurer que les atomes du patch requête et ceux du patch trouvé présentent des structures proches. Nous les utilisons comme outils d'aide à la détermination des meilleures quasi-cliques, mais ils ne sont pas pris en compte lors du calcul du score final de PatchSearch.

Quasi-clique finale et patch identifié

Le patch finalement reconnu sur la surface_C comme étant similaire au patch_A présente les caractéristiques suivantes :

— au moins 4 de ses atomes (C_α, centroïde aromatique, S, O, N) sont dans une configuration très similaire à ceux du patch_A (à 1,5 Å près, le *mindist*) avec lesquels ils sont appariés. Cet ensemble d'atomes retrouvés constitue une partie très conservée du patch, tant au niveau des distances qu'au niveau des propriétés physico-chimiques des résidus auxquels ils appartiennent.

— les autres atomes du patch reconnu ont été appariés avec les atomes du patch_A sur les mêmes critères de conservation des propriétés physico-chimiques, mais leurs distances inter-atomiques sont moins conservées (à 3 Å près, le *maxdist*).

Les étapes rythmant l'algorithme de PatchSearch pour reconnaître un patch_A sur une surface_C sont récapitulées dans la Figure 4.16.

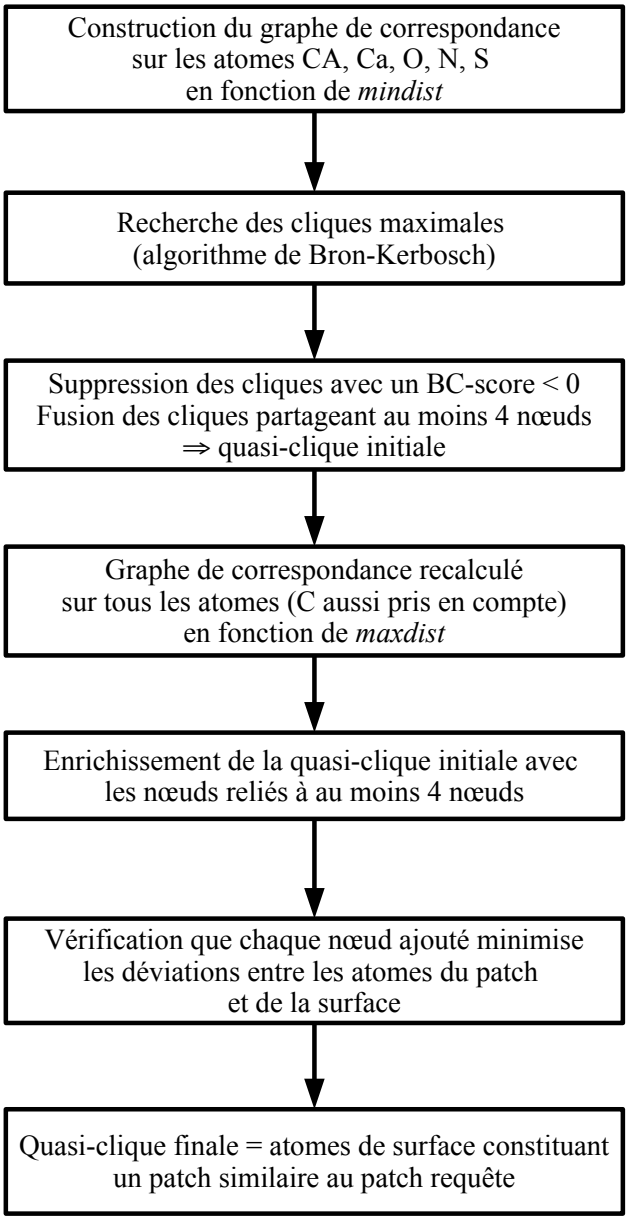


FIGURE 4.16: **Schéma récapitulatif des différentes étapes opérées par PatchSearch pour identifier un patch similaire à un patch requête dans une surface donnée.**

4.2.4 Score du patch trouvé et sorties de PatchSearch

Une fois que PatchSearch a identifié un patch similaire à une requête sur une surface, il est nécessaire de quantifier les appariements ainsi effectués sous la forme d'un score.

Le score du patch finalement trouvé s'exprime en fonction de la taille de l'alignement (soit le nombre d'atomes appariés entre le patch_A et la surface_C) et de la taille du patch requête. Le calcul du score est détaillé dans la Formule 4.2 : $\text{N}_\text{alignement}$ est le nombre d'atomes de la surface_C qui ont été appariés avec ceux du patch_A à l'aide de la quasi-clique. $\text{N}_\text{requête}$ est le nombre d'atomes constituant le patch_A..

$$score = \frac{N_\text{alignement}}{N_\text{requête}} \tag{4.2}$$

Le calcul du score diffère lorsque PatchSearch recherche des similitudes entre deux patches et est présenté dans la Formule 4.3 : N_cible est le nombre d'atomes constituant le patch_B sur lequel on a recherché un patch similaire au patch_A.

$$score = \frac{N_\text{alignement}}{N_\text{requête} + N_\text{cible} - N_\text{alignement}} \tag{4.3}$$

Au cours de la première étape de recherche de clique dans le graphe de correspondance initial, si aucune clique (donc aucune configuration de 4 atomes similaire au patch_A) n'a pu être trouvée dans la surface_C, le score retourné est de 0. Au contraire, si tous les atomes du patch ont pu être appariés à un ensemble d'atomes, on a $\text{N}_\text{alignement} = \text{N}_\text{requête}$ et on obtient alors un score de 1. Plus il y a d'atomes qui ont pu être appariés entre le patch_A et la surface_C, plus le score est proche de 1.

Nous nous sommes basés sur le jeu de données mis en place par Gunasekaran et ses collègues [58] pour définir un score seuil à partir duquel on déterminera qu'un patch contenant suffisamment d'atomes a été détecté dans la surface d'une protéine donnée. Nous nous sommes appuyés sur la distribution des scores obtenus en utilisant l'approche des quasi-cliques lorsque l'on recherchait un patch extrait d'une protéine *holo* dans la surface de la même protéine dans la forme *holo* (voir Figure 4.17).

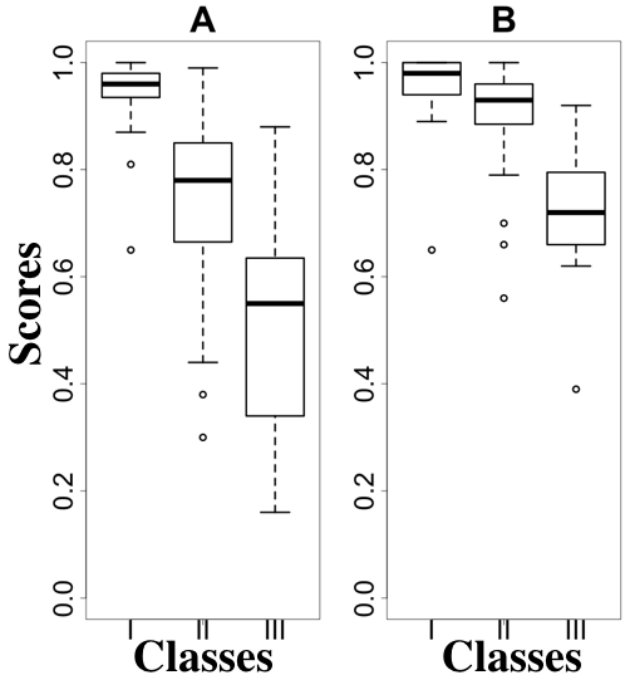


FIGURE 4.17: **Détermination d'un score seuil de PatchSearch : comparaison des distributions des scores obtenus à l'aide des quasi-cliques et des cliques sur le jeu de données de Gunasekaran.** Distributions des scores obtenus par PatchSearch lorsqu'on recherche le patch_{holo} sur la surface_{apo} correspondante, en utilisant les cliques (**A**) et les quasi-cliques (**B**). Les différentes classes I, II et III de protéines ont été définies en calculant le RMSD (tout atomes) entre le patch_{holo} et le patch_{apo} de chaque protéine.

Comme montré dans la Figure 4.17, le score minimal pour lequel PatchSearch permet d'identifier le patch d'une protéine *holo* dans la surface de la protéine *apo* correspondante est de 0,39 (je détaille ce cas dans le chapitre 5). Nous avons ainsi déterminé 0,4 comme étant le score seuil pour identifier un patch dans une structure donnée.

Comme le montre la Figure 4.18 obtenue à partir des scores et des patches issus de ce même jeu de données, toutes les tailles de patches ne sont pas représentées de façon homogène. Ce jeu compte majoritairement des patches de 20 à 60 atomes. La droite de régression ne permet pas d'établir une relation linéaire entre les scores et les tailles de patches. La corrélation entre ces deux variables est de -0,26, ce qui est donc très faible. Nous ne pouvons donc pas établir avec certitude que le score

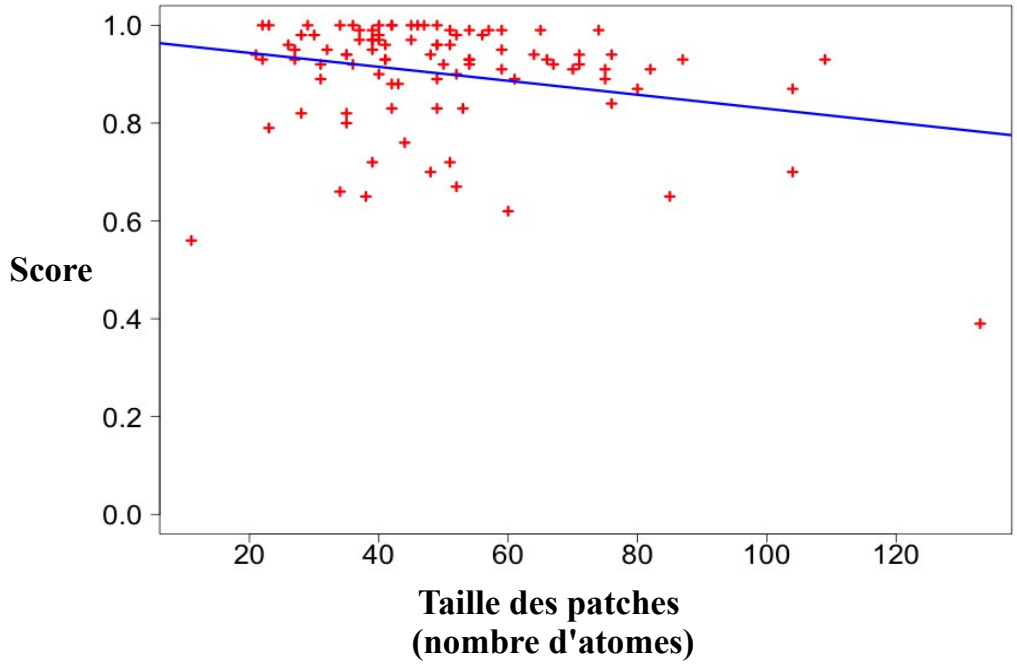


FIGURE 4.18: **Variation des scores de PatchSearch obtenus en fonction de la taille des patches donnés en requêtes.** Les scores ont été obtenus par PatchSearch lorsqu'on recherche le patch_{holo} sur la surface_{apo} correspondante. En bleue, est représentée la droite de régression linéaire dont les coefficients ont été calculés en fonction des scores et des tailles des patches.

varie de façon corrélée avec la taille des patches. Cette figure indique éventuellement que, dans ce jeu de données, les patches qui comptent plus de 80 atomes sont plus difficiles à identifier. Afin d'analyser de façon plus concluante les variations de score en fonction de la taille des patches, il serait nécessaire d'avoir un jeu de patches dont les tailles auraient une distribution plus homogène.

4.3 Illustration du fonctionnement de l'algorithme

Afin d'illustrer le fonctionnement de l'algorithme de PatchSearch, nous allons traiter l'exemple de l'identification d'un patch spécifique à un médicament, le sunitinib. Le patch choisi est extrait du complexe identifié dans la PDB par 2y7j. Dans ce fichier pdb, la phosphorylase kinase-γ 2 est complexée au sunitinib. Le but de

notre exemple est de tester s'il existe un patch similaire à celui de 2y7j sur la surface de la protéine 3ng5 qui est impliquée dans le transport de la thyroxine.

4.3.1 Extraction du patch requête et de la surface ciblée

Avant de pouvoir extraire le patch de 2y7j et la surface de 3ng5, il est nécessaire dans un premier temps de les télécharger à partir du site de la PDB.

Une fois que les fichiers sont téléchargés, le complexe 2y7j et la protéine 3ng5 doivent être visualisés à l'aide de PyMol. Un tel examen permet de déterminer que le ligand sunitinib porte le nom de résidu "B49", et se trouve en un seul exemplaire lorsqu'il est au contact de la chaîne A de la phosphorylase kinase-γ 2. Par défaut, nous extrayons les surfaces à partir de la première chaîne du fichier PDB. Lors de l'extraction des patches, la visualisation moléculaire permet de déterminer si la disposition du ligand lui permet d'effectuer plus d'interactions avec une chaîne plutôt qu'avec une autre, dans les cas de ligands entre deux chaînes. Nous extraiyons donc le patch de la chaîne A de la protéine A, à partir du ligand B49 qui appartient lui aussi à la chaîne A dans le fichier pdb. Nous visualisons également 3ng5 afin de déterminer de quelle chaîne protéique nous allons extraire la surface : nous choisissons la chaîne B. En effet, dans le cas de cette protéine, elle présente des patches connus pour d'autres ligands : la chaîne A est complexée avec deux molécules d'épigallocatéchine gallate (ou EGCG) et une molécule de glycérol, tandis que la chaîne B est complexée à seulement une molécule d'EGCG. Lorsque cela est possible, j'extrais la surface de la chaîne protéique qui est complexée avec le moins de molécules possible. Lors de l'extraction de surface, le choix de la chaîne protéique se fait de manière arbitraire et est automatisable.

Les programmes d'extraction du patch et de la surface requièrent d'avoir les logiciels NACCESS et PROPKA (version 3.0) installés. De plus, afin de calculer l'accessibilité relative de chaque atome pour chaque résidu, il est également nécessaire d'avoir une librairie de résidus pdb, un répertoire contenant un fichier PDB par résidu. Les langages de programmation qui doivent être installés pour faire fonctionner les programmes sont Python, R et C++. Les fonctions implémentées en C++ permettent de traiter rapidement les informations structurales afin d'obtenir

des graphes. Les graphes, noeuds et arêtes ainsi générés sont traités par la suite à l'aide des fonctions R de la librairie igraph. Le package igraph de R est nécessaire afin de rechercher les cliques dans des graphes. Les fonctions implémentées en C++ appelées dans le programme PatchSearch sont :

— graph.edgelist() : cette fonction permet de créer un graphe à partir d'une liste d'arêtes.

— maximal.cliques() : cette fonction renvoie la liste des cliques maximales.

La librairie Rcpp permet d'intégrer les fonctions implémentées en C++ au programme PatchSearch programmé en R Une fois que ces pré-requis sont installés sur la machine, il est possible de faire fonctionner les programmes d'extraction de patch et de surface, ainsi que le programme PatchSearch.

Dans un premier temps, le patch d'interaction est extrait de 2y7j (Figure 4.19). La surface de la transthyrétine est aussi extraite du fichier 3ng5 (Figure 4.20).

4.3.2 Établissement des correspondances du graphe produit

Nous allons maintenant détailler les différentes étapes de l'algorithme de PatchSearch qui s'opèrent lorsque le programme est appelé pour identifier un patch requête sur une surface ciblée. Dans le cadre de l'exemple utilisé dans ce chapitre, nous recherchons le patch extrait de 2y7j sur la surface de 3ng5.

Le fichier $2y7j_{patch}$ compte 63 atomes, tous types d'atomes confondus, tandis que $3ng5_{surface}$ en contient 541.

Dans un premier temps, tous les atomes du groupe d'atomes lourds initial (O,N,CA,Ca,S) du $2y7j_{patch}$, soient 37 atomes, sont appariés avec ceux de $3ng5_{surface}$ qui appartiennent au même groupe d'atomes, soient 315 atomes. Chacun des 37 atomes de $2y7j_{patch}$ sont appariés avec les atomes de $3ng5_{surface}$ lorsque les correspondances sont possibles.

Au total, le graphe de correspondance compte 1251 noeuds. Par exemple, l'atome numéro 1 du $2y7j_{patch}$, le O de la valine 29, est apparié, entre autres, avec les atomes suivants de $3ng5_{surface}$:

— l'atome 38 : O de l'alanine 19.

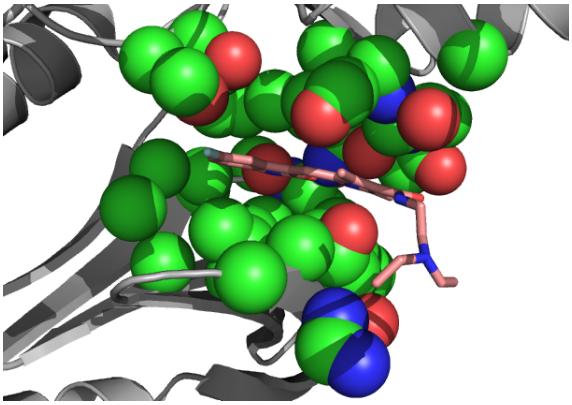


FIGURE 4.19: **Patch extrait du complexe 2y7j.** Les atomes sont représentés sous la forme de sphères. Les atomes de carbone sont colorés en vert, les oxygènes en bleu et les azotes en rouge. La phosphorylase kinase-γ 2 est représentée en mode rubans.

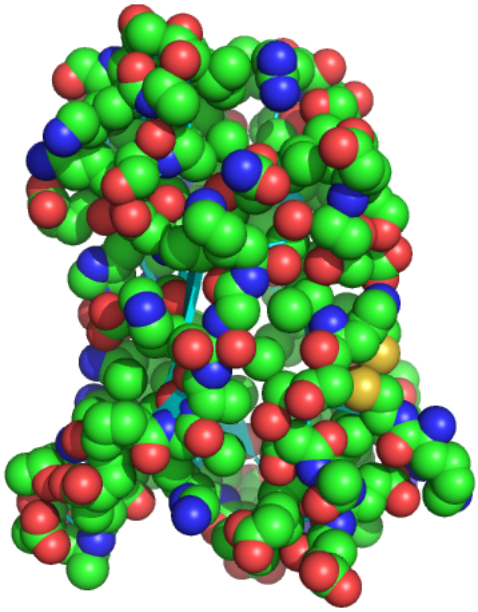


FIGURE 4.20: **Surface extraite de la protéine 3ng5.** La surface a été extraite de la chaîne B de la transthyrétine du complexe 3ng5.

— l'atome 41 : O de la valine 20.

— l'atome 65 : O de l'isoleucine 26.

— l'atome 87 : O de la valine 32.

D'après les scores de substitution de la matrice BLOSUM62, cet atome d'oxygène de $2y7j_{patch}$ est apparié avec des atomes d'oxygènes de $3ng5_{surface}$, et les résidus auxquels appartiennent ces atomes ont des scores de substitution supérieurs ou égaux à 0 : 0 pour VAL→ ALA, 4 pour VAL → VAL et 3 pour VAL → ILE.

4.3.3 Établissement d'arêtes entre les noeuds du graphe produit

L'étape suivante consiste à relier deux noeuds par une arête lorsque les atomes appariés correspondants sont distants au plus du *mindist* calculé comme la distance minimale des matrices de distances du patch et de celle de la surface, soit 1,22 Å. Seuls les couples d'atomes dont les distances internes (soit à la fois dans le patch et dans la surface) sont similaires à 1,22 Å près sont gardés pour construire le graphe de correspondance et sont reliés par des arêtes. En tout, le graphe de correspondance contient 53513 couples de noeuds reliés par des arêtes.

4.3.4 Recherche des meilleures cliques

Une fois que le graphe de correspondance est ainsi construit, nous utilisons des fonctions de la librairie "igraph" afin de rechercher les cliques dans ce graphe.

On impose que la clique maximale doit être de taille minimale 4. En effet, afin de localiser précisément un point dans l'espace, il est nécessaire de connaître la distance de ce point à au moins 4 autres points de l'espace. Dans notre exemple, la clique maximale est de taille 9, elle correspond donc au critère précédemment énoncé. Il existe donc 9 atomes dans $2y7j_{patch}$ (Figure 4.21) et 9 atomes dans $3ng5_{surface}$ (Figure 4.22) qui sont de types identiques, qui appartiennent à des résidus aux propriétés physico-chimiques similaires, et se trouvant dans des configurations très proches, à 1,22 Å près.

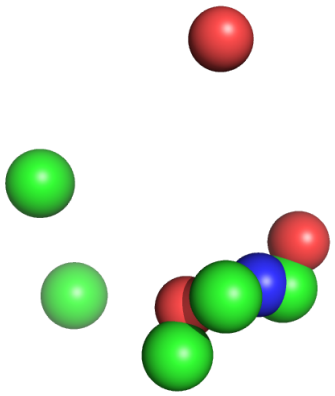


FIGURE 4.21: **Clique trouvée : atomes du patch de $\mathbf{2y7j_{patch}}$.** Atomes de $2y7j_{patch}$ appariés avec des atomes de $3ng5_{surface}$, après la construction de la clique.

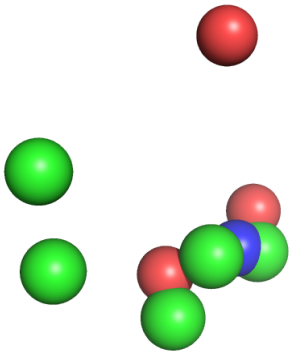


FIGURE 4.22: **Clique trouvée : atomes du patch de $\mathbf{3ng5_{surface}}$.** Atomes de $3ng5_{surface}$ appariés avec des atomes de $2y7j_{patch}$ après la construction de la clique.

Le BC-score calculé entre les atomes de $2y7j_{patch}$ appariés avec ceux de $3ng5_{surface}$ est de 0,86, ce qui signifie que les configurations de ces deux ensembles d'atomes sont très similaires et ne sont pas en miroir l'une de l'autre.

4.3.5 Enrichissement des meilleures cliques en quasi-cliques

Nous cherchons maintenant à déterminer s'il existe des parties similaires entre le patch et la surface comparés, mais plus flexibles que les atomes de la clique précédemment déterminée.

Une première recherche dans le graphe de correspondance déjà établi permet de rajouter à la clique tous les noeuds qui partagent au moins 4 arêtes avec les noeuds de la clique. Nous obtenons ainsi une quasi-clique dont tous les noeuds sont au moins

de degré 4 : tous les noeuds sont reliés à au moins 4 autres noeuds de ce sous-graphe.

Puis, dans un deuxième temps, nous recalculons les noeuds du graphe de correspondance, mais cette fois-ci en prenant en compte tous les atomes, et en acceptant un score de substitution de la matrice BLOSUM62 négatif. Le nouveau graphe de correspondance compte 9624 noeuds. Un nombre d'appariements aussi important s'explique par la composition de la surface d'une protéine qui est très riche en carbones qui ne sont ni des C_α, ni des carbones de cycles de résidus aromatiques (voir Tableau 4.2), et également par les conditions d'appariements qui sont moins strictes que précédemment pour la construction de la clique.

Fichier	**N**	**S**	**O**	**CA**	**Ca**	**C**
$2y7j_{patch}$	6	0	10	6	20	26
$3ng5_{surface}$	68	2	118	120	107	226

Tableau 4.2: **Composition en types d'atomes de $2y7j_{patch}$ et de $3ng5_{surface}$.**

De nouveau, la quasi-clique doit être enrichie selon le même principe que précédemment : tous les noeuds qui sont reliés à au moins 4 noeuds de la clique peuvent être potentiellement rajoutés. On calcule donc les arêtes qui relient les nouveaux noeuds candidats à la quasi-clique cette fois-ci avec un paramètre d'écart de distance (*maxdist*), qui permet de modifier la flexibilité avec laquelle on souhaite retrouver la configuration d'atomes aux distances similaires. Pour chaque nouveau noeud qui n'appartient pas à la quasi-clique précédemment calculée, on calcule l'écart de distance entre les atomes appariés par ce noeud et entre les atomes appariés pour les autres noeuds de la quasi-clique. Si cet écart de distance est inférieur à 3 Å (la valeur à laquelle nous avons fixé *maxdist*), le noeud est un candidat potentiel pour la quasi-clique et est relié aux autres noeuds du graphe de correspondance.

Au final, le graphe de correspondance compte 697 couples de noeuds reliés entre eux, et autant d'arêtes.

4.3.6 Finalisation des meilleures quasi-cliques

Enfin, la dernière étape pour finaliser les quasi-cliques identifiées est l'ajout de noeuds qui permettent un appariement avec des écarts de distance minimaux. Pour chaque noeud qui est compatible avec la quasi-clique, il est nécessaire de vérifier s'il permet un appariement d'atomes de 3ng5 plus proche de la configuration des atomes de $2y7j_{patch}$. Si le RMSD calculé entre les atomes du patch appariés avec ceux de la surface est inférieur ou égal à celui de la quasi-clique, alors le noeud est rajouté dans la quasi-clique. Au final, on obtient la meilleure quasi-clique qui compte 18 atomes présentée dans les Figures 4.23 et 4.24, qui présente un BC-score de 0,79.

Le BC-score est plus faible qu'avec la clique, ce qui correspond à des distances moins conservées. En effet les 18 atomes de $2y7j_{patch}$ sont appariés avec les 18 atomes de $3ng5_{surface}$ avec un écart-entre les paires d'atomes d'au maximum 3 Å (*maxdist*).

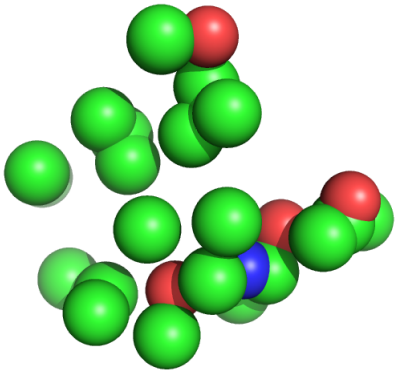


FIGURE 4.23: **Appariement final pour $2y7j_{patch}$.** Atomes de $2y7j_{patch}$ appariés avec des atomes de $3ng5_{surface}$ après l'enrichissement de la quasi-clique.

Le score attribué à cet alignement s'exprime en fonction de la taille de l'alignement par rapport à la taille de $2y7j_{patch}$. Cependant, comme les cycles des résidus aromatiques sont représentés par leur centroïde, il faut prendre en compte cette "simplification" et multiplier le nombre de centroïdes (CA) par 6 à la fois parmi les atomes de l'alignement, et parmi ceux de $2y7j_{patch}$. Le score obtenu est calculé en fonction de $N_{alignement}$ et $N_{2y7jpatch}$ en utilisant les totaux du Tableau 4.3, et vaut 0,34.

Nous venons d'étudier le protocole d'extraction de patches et de surfaces, puis de détailler les étapes les étapes détaillées qui permettent à PatchSearch d'identifier

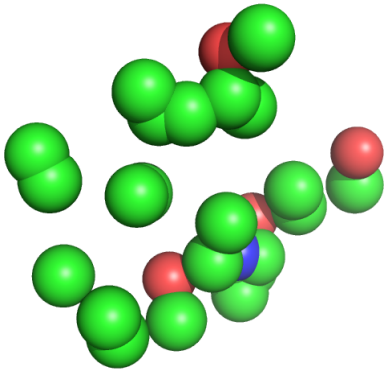


FIGURE 4.24: **Appariement final pour $\mathbf{3ng5_{surface}}$.** Atomes de $3ng5_{surface}$ appariés avec des atomes de $2y7j_{patch}$.

Fichier	**N**	**S**	**O**	**CA**	**Ca**	**C**	**Total**
$2y7j_{patch}$	6	0	10	6×6	20	26	68
alignement	1	0	3	1×6	8	5	23

Tableau 4.3: **Composition en types d'atomes des surfaces comparées.**

un potentiel patch similaire au patch requête sur une surface donnée.

Il était important d'évaluer les performances de PatchSearch sur des protéines qui présentaient des patches connus afin de valider notre approche. Nous aborderons dans la partie suivante les mesures de l'efficacité de PatchSearch pour reconnaître un patch sur une protéine connue pour présenter ce patch.

4.4 Évaluation de la reconnaissance d'un patch

Le but de notre démarche est d'appliquer PatchSearch à la recherche de potentielles cibles secondaires ou *off-targets* pour un composé thérapeutique, dans le cas de possibles interactions multiples. Le patch de ce composé serait recherché sur :

— les surfaces de protéines connues pour être souvent impliquées dans des effets secondaires.

— les surfaces de protéines impliquées dans une voie ciblée par le médicament.

— ou encore sur des protéines appartenant à une toute autre voie, dans un contexte de repositionnement de médicament.

Néanmoins, lors de la mise en place d'un tel outil, il a fallu vérifier que les patches identifiés sur des *off-targets* (ou autres protéines) correspondent effectivement à ceux connus pour être impliqués dans l'interaction avec le ligand. A l'aide des jeux de données présentés dans les chapitres suivants, nous avons ajusté les différents paramètres de préparation des patches et surfaces et de recherche de similitudes en appliquant PatchSearch à des protéines déjà connues pour présenter un patch similaire au patch requête.

De plus, de nombreux outils se sont déjà attelés à la comparaison de sites de liaison, et la plupart d'entre eux utilisent l'AUC (*Area Under the Curve*) qui permet d'évaluer la spécificité et la sensibilité de la reconnaissance d'un groupe de patches.

Dans un premier temps, j'expliquerai le calcul de l'AUC, puis je présenterai notre évaluation de la reconnaissance d'un patch sur une surface.

4.4.1 Calcul d'une AUC

La plupart des outils de comparaison de patches cherchent à comparer un patch_1 à un autre patch_2 connu pour lier un ligand, et à renvoyer un score indiquant si le patch_1 pourrait lier le même ligand que le patch_2. Dans le meilleur des cas, tous les patches liant un même ligand seraient retrouvés avec des scores forts proches de 1, et au contraire, la comparaison de deux patches liant des ligands différents aboutirait à un score faible, proche de 0.

La performance d'un outil de reconnaissance de patch se traduit par sa capacité à :

— identifier avec un score élevé deux patches connus pour interagir avec le même ligand : c'est la sensibilité de l'outil.

— discriminer deux patches qui lient des molécules différentes avec un score très faible : c'est la spécificité de la comparaison de patches.

Le but est d'évaluer si le score de l'outil permet de discriminer correctement les vrais positifs des vrais négatifs et, par la même occasion, minimise donc le nombre de faux positifs et de faux négatifs. L'ensemble de ces scores est trié dans l'ordre décroissant. Pour chaque seuil de score (le premier score, puis le deuxième score, et

ainsi de suite jusqu'aux n-ième score), on évalue si ce score seuil permet d'identifier le plus de positifs possibles et le moins de négatifs possible. Cette évaluation revient à comptabiliser les cas positifs trouvés au-delà du seuil de score, ainsi que les négatifs, qui n'auraient pas dû être retrouvés si ce seuil de score était pertinent. Pour chaque comparaison au-delà du seuil de score, on est face à 2 cas différents :

— un $\text{patch}_{\text{ligandA}}$ comparé sur un $\text{patch}_{\text{ligandA}}$: c'est un vrai positif car les deux patches sont connus pour interagir avec le même ligand A.

— un $\text{patch}_{\text{ligandA}}$ comparé sur un $\text{patch}_{\text{ligandB}}$: c'est un faux positif car les deux patches sont connus pour interagir avec des ligands différents (ligand A et ligand B), PatchSearch devrait donc identifier le moins de similitudes possible entre les deux patches.

Le nombre de comparaisons positives (n_{pos}) attendues est connu, de même que le nombre de comparaisons négatives ($n_{nég}$). Ainsi, pour chaque score seuil, on peut calculer la sensibilité de ce score qui est le taux de positifs qu'on a su retrouver à l'aide de ce score, et 1 - la spécificité, qui est le taux de négatifs qu'on a retrouvés.

La Figure 4.25 montre le calcul de l'AUC pour chaque score seuil.

Le taux de vrais positifs (TVP) est détaillé dans la Formule 4.4 : $n_{\text{positifs trouvés}}$ est le nombre de comparaison de vrais positifs trouvés au-dessus du score seuil. $n_{\text{positifs total}}$ est le nombre de comparaison de vrais positifs total. Le taux de faux positifs (TFP) est présenté dans la Formule 4.5 : $n_{\text{négatifs trouvés}}$ est le nombre de comparaison de faux positifs trouvés au-dessus du score seuil. $n_{\text{négatifs total}}$ est le nombre de comparaison de vrais négatifs total. Ces taux permettent ainsi de calculer l'AUC à ce score précis. De manière graphique, la représentation pour chaque valeur de score seuil du TVP en fonction du TFP est la courbe ROC [64], ou *Receiver Operating Characteristic*. L'AUC totale correspond à l'aire sous cette courbe ROC. Pour chaque valeur de score seuil, l'AUC_i est donc l'aire sous la courbe entre le score seuil et le score précédent : cette aire est calculée à l'aide de la formule des trapèzes (Formule 4.6, i et $i-1$ sont les indices des scores seuil à partir duquel on calcule l'AUC). Finalement, la valeur total de l'AUC correspond à l'aire sous la courbe, soit à la somme des aires des différents trapèzes précédemment calculés (Formule 4.7).

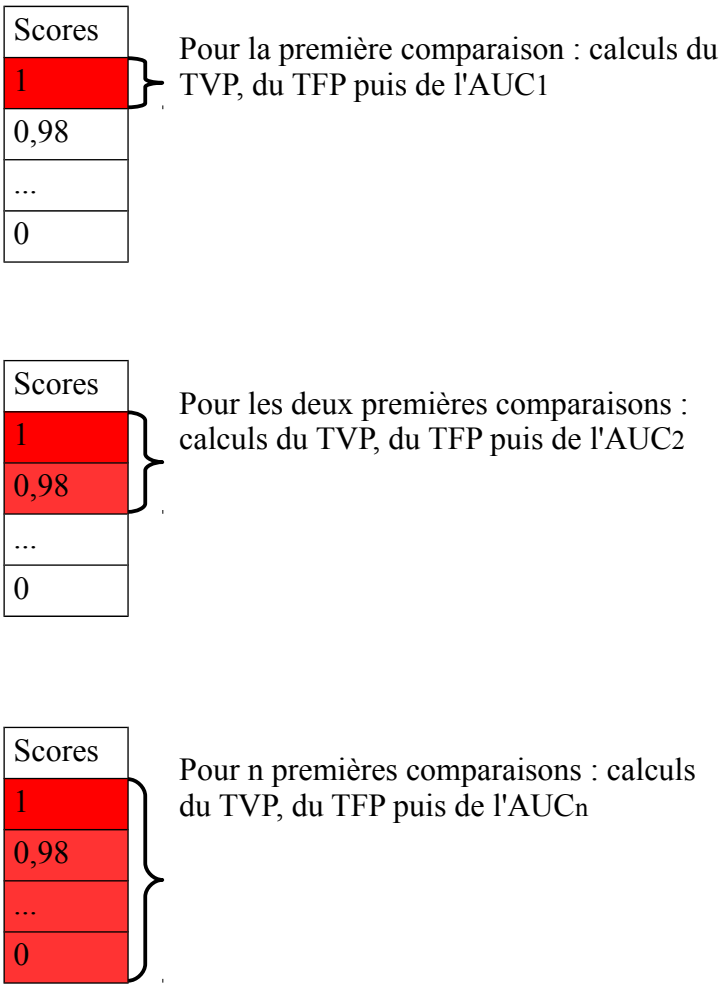


FIGURE 4.25: **Schéma détaillant la méthode de calcul d'AUC.** Les scores des différentes comparaisons effectuées sont triés par ordre décroissant. Chaque valeur de score est considérée comme un seuil pour lequel on calcule l'AUC : d'abord pour la première comparaison, puis pour les deux premières, pour les n-1 premières comparaisons, enfin pour les n comparaisons.

$$TVP = \frac{n_{\text{positifs}}\text{trouvés}}{n_{\text{positifs}}total} \tag{4.4}$$

$$TFP = \frac{n_{\text{négatifs}}\text{trouvés}}{n_{\text{négatifs}}\text{total}} \tag{4.5}$$

$$AUC_i = (TFP_i - TFP_{i-1}) \times \frac{TVP_i + TVP_{i-1}}{2} \tag{4.6}$$

$$AUC_{\text{totale}} = \sum_{i=1}^{n} AUC_i \tag{4.7}$$

La valeur de l'$\text{AUC}_{\text{totale}}$ renseigne sur la performance générale d'un outil :

— un outil optimal permet de discriminer parfaitement les VP des VN et l'AUC calculée vaut 1.

— un outil moins performant ne permet pas de discriminer efficacement les VP des VN et l'AUC calculée vaut 0,5. Ses scores correspondent à ceux qui auraient été obtenus à l'aide d'un outil aléatoire.

4.4.2 Identification d'un patch similaire sur une surface donnée

Lors du développement de PatchSearch, il a été nécessaire d'évaluer la capacité de notre outil à identifier un patch sur une surface pour laquelle ce patch a déjà été déterminé. Par exemple, si deux protéines (protéine_A et protéine_B) sont expérimentalement connues pour lier un même ligand, il est donc possible d'extraire les patches correspondants (patch_A et patch_B), et de rechercher le patch_A sur la surface_B en vérifiant que le patch trouvé corresponde bien au patch_B que nous connaissons déjà.

Pour vérifier la précision du patch reconnu, nous avons utilisé une distance indiquant si la localisation du patch identifié est correcte, la *Dc* (pour Distance entre centroïdes). La distance est calculée selon la Formule 4.8 : $\text{Dist}_{i,j}$ est calculée entre les points i et j. x_i, y_i, et z_i sont les coordonnées cartésiennes du point i, tandis que x_j, y_j, et z_j sont les coordonnées du point j. Cette distance correspond à la

distance entre le centre de gravité du patch connu (sur la surface_B) et le centre de gravité du patch identifié (le patch_B appartenant à la surface_B). Le patch identifié et le patch déjà connu sont tous deux dans la même surface, la *Dc* peut donc se calculer directement à partir des coordonnées des atomes des deux patches.

Les coordonnées atomiques des différents centres de gravité sont calculées en utilisant la formule 4.9 : G_x, G_y et G_z sont les coordonnées cartésiennes du centre de gravité G d'un nuage de n points.

$$Dist_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{4.8}$$

$$G_x = \frac{\sum_{i=1}^{n} x_i}{n} \quad G_y = \frac{\sum_{i=1}^{n} y_i}{n} \quad G_z = \frac{\sum_{i=1}^{n} z_i}{n} \tag{4.9}$$

Les figures 4.26 et 4.27 présentent des patches de formes différentes avec leur centres de gravité respectifs en rouge ainsi calculés.

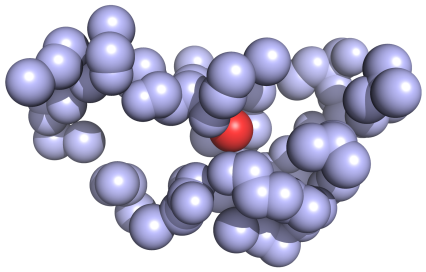


FIGURE 4.26: **Exemple de centre de gravité d'un patch enfoui** Est représenté ici le patch extrait du complexe 3hec et son centre de gravité. Le centre de gravité est coloré en rouge, les atomes du patch sont colorés en bleu.

La recherche d'un patch contre la surface de la protéine dont il a été extrait (par exemple lorsque l'on recherche le patch_A sur la surface_A) donne toujours une *Dc* de 0 Å : le patch requête est intégralement retrouvé, par conséquent, son centre de gravité correspond à celui du patch identifié.

Pour la recherche d'un patch extrait d'une protéine différente, plus la valeur de la *Dc* est faible, plus le recouvrement entre le patch identifié et le patch connu est fort, ce qui indique que PatchSearch identifie un patch déjà existant sur une protéine différente de la protéine dont il a été extrait. Une Dc faible signifie qu'on a identifié

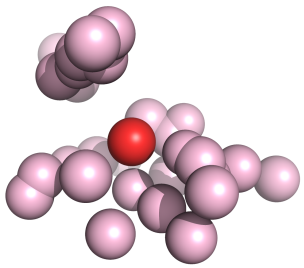


FIGURE 4.27: **Exemple de centre de gravité d'un patch plus exposé.** Est représenté ici le patch extrait du complexe 2a3a et son centre de gravité. Le centre de gravité est coloré en rouge, les atomes du patch sont colorés en rose.

un patch qui partage un nombre significatif d'atomes en commun avec le patch déjà connu dans la surface ciblée. Le critère du BC-score qui permet de ne garder que des appariements entre des groupes d'atomes aux dispositions similaires assure que le patch identifié est dans une orientation correcte.

Dans l'exemple de la Figure 4.28, le patch de l'imatinib extrait du complexe 1t46 est recherché sur la surface de 3hec. Le patch identifié contient un nombre important d'atomes du patch connu de 3hec : la Dc calculée vaut 0,48Å , PatchSearch a donc localisé efficacement le patch de 3hec.

Par contre, dans l'exemple de la Figure 4.29, le patch de la théophylline extrait du complexe 4eoh est recherché sur la surface de 2a3a. Le patch trouvé (en sphères roses) est fortement éloigné du patch connu de 2a3a : la Dc calculée vaut 29,30Å.

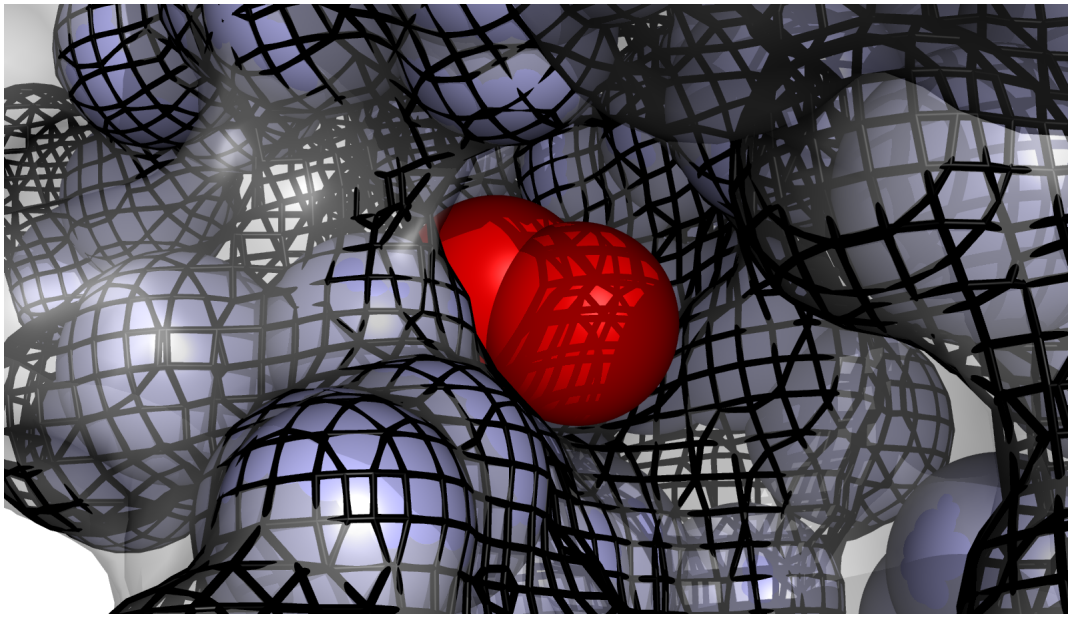


FIGURE 4.28: **Exemple où la *Dc* est inférieure à 1 Å.** Le patch trouvé est représenté en grillage noir et le patch connu de 3hec est représenté sous le forme de sphères bleues. Les centres de gravité des deux patches sont colorés en rouge.

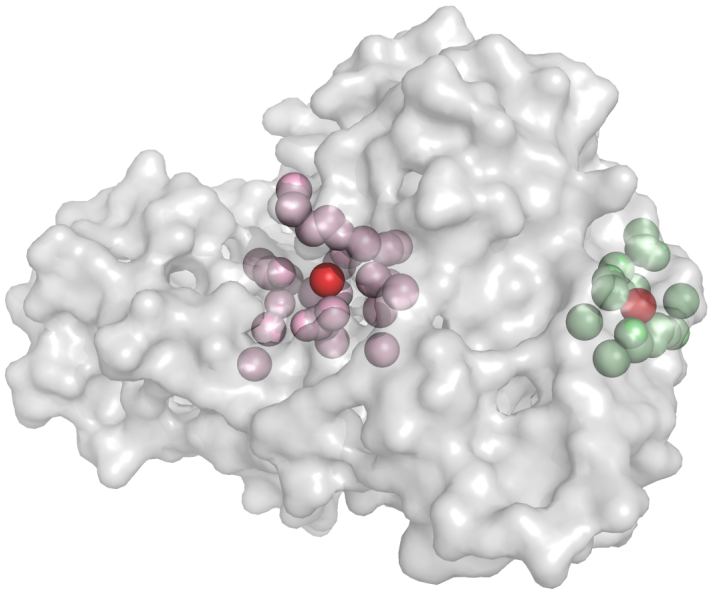


FIGURE 4.29: **Exemple de *Dc* importante.** Le patch trouvé est en sphères roses et le patch connu est en sphères vertes. Les centres de gravité des deux patches sont colorés en rouge.

4.5 Discussion et conclusions

4.5.1 Mise au point des différents outils présentés

La mise en place des différents outils présentés dans ce chapitre résulte de travaux d'optimisation des méthodes utilisées et de leurs paramètres. En effet, nous nous sommes d'abord appuyés sur des paramètres issus de recherches bibliographiques. Les premiers patches étaient extraits sur des protéines dont les résidus impliqués dans les interactions avaient été expérimentalement déterminés. Nous avons mis au point par la suite l'extraction des patches sur des complexes de façon plus automatisée (le nom du ligand ainsi que sa chaîne doivent être spécifiés) afin de pouvoir l'appliquer sur de larges jeux de données. Nous avons finalement abouti aux valeurs seuils telles que l'$ASA_{relative}$ à 1%, la distance du ligand à 5Å. Ces paramètres ont été fixés suite aux scores obtenus au cours des essais effectués sur les protéines présentées plus loin dans la section 7 qui s'apparentent le plus aux cas d'application de PatchSearch qui sont la recherche de cibles multiples connues pour interagir avec un même ligand.

Dans un premier temps, la comparaison entre un patch et une surface était basée sur différents algorithmes qui avaient pour but de retrouver des sous-graphes riches en arêtes dans un graphe produit, ce qui équivaut à des sous-parties fortement conservées entre le patch et la surface. Une de nos priorités a été en effet d'optimiser le taux de distances conservées entre les deux surfaces comparées. Nous avons par la suite modifié cette méthode en prenant en compte la notion de k-cores, qui sont des sous-graphes dans lesquels tous les noeuds sont reliés à au moins k voisins [155, 14]. Dans un 3-core par exemple, tous les noeuds sont reliés à au moins 3 autres noeuds du graphe. Cette méthode, bien que rapide, n'est pas efficace sur notre problème de reconnaissance d'un patch sur une surface. En effet, elle procède à tous les appariements possibles et ne vérifie pas qu'on fait bien un appariement unique entre un atome du patch et un atome de la surface. Par conséquent, il est possible d'obtenir un patch dit "similaire" qui compte plus d'atomes que le patch trouvé, et surtout qui n'est plus représentatif du nombre de distances inter-atomiques conservées dans les appariements du graphe produit.

4.5.2 Conclusions

Ce chapitre détaille le fonctionnement d'outils applicables à de larges jeux de données :

— les programmes qui permettent d'extraire soit un patch, soit une surface, à partir d'un fichier PDB. Ces programmes prennent en argument des fichiers contenant les chemins des fichiers PDB à traiter ainsi que les noms de ligands et de chaînes.

— PatchSearch, un outil qui identifie les similitudes entre un patch et une surface. Les résultats de PatchSearch tels que les meilleurs patches identifiés et les scores associés peuvent être directement exploités par la suite, et permettre à la fois une analyse individuelle de chaque cas testé, et globale lorsque PatchSearch est appliqué à un large jeu de données.

Toutefois, rappelons qu'une visualisation du patch étudié est primordiale avant de le rechercher sur une ou plusieurs cibles, afin de vérifier qu'il ait bien été extrait autour du ligand d'intérêt. Les paramètres par défaut de ces programmes sont modifiables selon les besoins de l'utilisateur, ainsi que la conservation géométrique avec laquelle il souhaite identifier des patches similaires à la requête. Si l'utilisateur souhaite étudier un patch contenant les atomes d'une protéine dans un rayon plus étendu ou plus restreint autour du ligand, il est possible de modifier le seuil de distance autour duquel on souhaite extraire le ligand. Il en est de même pour le seuil d'$\mathrm{ASA}_{\mathrm{relative}}$ qui peut varier selon que l'on souhaite prendre en compte ou non l'exposition des atomes au solvant. De plus, l'utilisateur peut contrôler la flexibilité avec laquelle il souhaite identifier un patch similaire à l'aide du paramètre *maxdist* de PatchSearch : une valeur de *maxdist* élevée permet d'identifier des atomes appariés aux distances moins conservées.

Enfin, nous avons aussi implémenté des outils pour évaluer les patches identifiés sur des cibles, dans le cas où celles-ci sont connues pour présenter le patch étudié. Toutefois, l'optimisation des paramètres de PatchSearch a toujours été effectuée sur des cibles dites "vraies positives" ce qui nous permet de quantifier les vraies positives et les faux négatives lors des différents tests. Nous n'avions pas de cibles

“vrais négatives”, soient des protéines connues pour ne pas lier certains ligands sur lesquels nous pourrions améliorer la détermination du score seuil de PatchSearch notamment. Pour cela, il faudrait disposer de données quantitatives d’affinité très faibles de protéines mises au contact de ligands.

Chapitre 5

Reconnaissance d'un patch sur une protéine flexible

Ce premier chapitre de résultats est consacré à l'analyse de l'efficacité de PatchSearch pour identifier un patch dans une conformation différente. PatchSearch recherche d'abord une clique dans un graphe produit, ce qui est une étape commune aux autres outils de comparaison de structures qui utilisent la théorie des graphes. Toutefois, notre programme procède par la suite à l'extension des meilleures cliques identifiées en quasi-cliques. A notre connaissance, cette méthode est originale comparée aux autres outils d'identification de similitudes entre structures.

Aussi, cette approche soulève plusieurs interrogations : les quasi-cliques ainsi identifiées permettent-elles d'obtenir des résultats du même ordre, voire meilleurs que les cliques ? La description de l'extension des cliques en quasi-cliques indique que, théoriquement, PatchSearch devrait reconnaître des atomes appariés dont les écarts de distances maximum sont fixés, par défaut, à 3 Å.

5.1 Flexibilité structurale des protéines

Les protéines sont des macromolécules sujettes aux modifications structurales, [63]. L'environnement réactionnel des protéines enzymatiques joue également un rôle dans ces changements de conformations intrinsèques. En effet, comme il est évoqué en introduction, la nature même d'une protéine induit un comportement

dynamique caractéristique. Cette flexibilité peut se manifester par des mouvements de grande amplitude impliquant des mouvements de la chaîne principale ou par des mouvements locaux, essentiellement des mouvements des chaînes latérales.

Par conséquent, lorsque nous étudions un patch issu d'une structure où l'enzyme est déjà liée à son substrat, le patch équivalent dans la protéine seule peut présenter des résidus voire des structures secondaires dans des orientations ou des configurations différentes. Cette théorie avait été énoncée par Koshland en 1958 notamment sous le nom "d'ajustement induit" [93] : la liaison du substrat à l'enzyme entraîne des changements réversibles au niveau des caractéristiques géométriques de la structure enzymatique, ce qui s'explique par l'orientation que doivent adopter certains résidus impliqués directement dans la catalyse afin que l'enzyme puisse exercer son activité. D'autres travaux avaient également été menés sur une éventuelle capacité des enzymes à se souvenir de leurs différents états de transition, les qualifiant ainsi d'enzymes mnémoniques [140, 174].

5.2 Description d'un jeu de protéines flexibles

Nous souhaitions appliquer PatchSearch sur des protéines flexibles afin de proposer des éléments de réponses aux problématiques précédemment posées. Le jeu de données sur lequel nous nous sommes appuyés a été mis en place par Gunasekaran et ses collègues en 2007 ([58]), et présente la particularité d'avoir deux catégories de protéines :

— un groupe de 97 complexes protéine-ligand (les protéines "*holo*").

— un deuxième groupe qui correspond aux mêmes protéines, mais en absence de ligand (les protéines "*apo*").

Les structures des protéines *holo*/*apo* ont été cristallisées séparément, et sont chacune téléchargée dans des fichiers pdb distincts.

5.2.1 Classes selon Gunasekaran

Les protéines ont été distribuées en 3 classes selon la flexibilité entre le patch de la forme *holo* et celui de la forme *apo* correspondante. Plus précisément, le déplacement des Cα entre les structures *holo* et *apo* de chaque couple, précédemment superposées à l'aide d'un algorithme d'extension combinatoire (CE) d'alignement structural [157, 57], a été calculé afin de quantifier les changements conformationnels.

Trois groupes de couples de structures *holo*/*apo* sont rassemblés en classes selon le RMSD calculé entre les formes *holo* et *holo* :

— la classe I : les protéines dont le RMSD des Cα des formes *holo* et *apo* est inférieur à 0,5 Å (40 protéines).

— la classe II : les protéines dont le RMSD des Cα des formes *holo* et *apo* est supérieur ou égal à 0,5 Å et inférieur ou égal à 2,0 Å (35 protéines).

— la classe III : les protéines dont le RMSD des Cα des formes *holo* et *apo* est supérieur à 2,0 Å (22 protéines).

La liste des structures des complexes et des formes *apo* est présentée dans le Tableau A.1 (Annexe A).

5.2.2 Mise en place de nouvelles classes

Toutefois, la mesure effectuée par Gunasekaran tient compte uniquement des mouvements importants du squelette polypeptidique mais ne reflète pas les modifications locales, et donc plus fines, de la conformation du site de liaison. En effet, des mouvements importants de chaînes latérales influencent l'environnement polaire ou hydrophobe du patch, dû à un enfouissement de certains atomes impliqués dans des interactions prépondérantes avec le ligand par exemple. Par conséquent, nous avons calculé les RMSD entre les atomes du patch dans la conformation *holo* et les mêmes atomes dans le patch équivalent dans la conformation *apo*.

Le protocole suivant a été appliqué pour chacun des 97 couples :

— extraction de la surface de la forme *apo*,

— chaque atome du patch$_{apo}$ a été extrait en comparaison aux atomes du patch$_{holo}$ en recherchant s'il répondait aux critères suivants :

- il appartient à la surface$_{apo}$,

- il appartient à la même chaîne, au même numéro de résidu et porte le même nom d'atome que l'atome correspondant du patch$_{holo}$.

— les deux structures sont superposées et leur RMSD est calculé sur tous les atomes en utilisant la commande `rms` de PyMol.

Nous obtenons ainsi différents types de données : des patches-*holo* extraits à partir des protéines en forme *holo*, des surfaces-*apo* extraites à partir des protéines en forme *apo* et des patches-*apo* extraits des surfaces-*apo* par correspondance avec les patches-*holo*.

Nous avons ainsi défini nos classes calculées en fonction du RMSD de tous les atomes du patch$_{holo}$ et du patch$_{apo}$ dans le Tableau A.2 (Annexe A).

Nous avons environ un quart des protéines de la classe I$_{\text{Gunasekaran}}$ qui se retrouvent essentiellement dans la classe II (14 structures), et 2 sont assignées dans la classe III. Ces deux structures sont assignées dans la classe III notamment pour chacun des cas à cause de mouvements d'un résidu en particulier.

Pour le cas du couple 1arm/1yme (Figure 5.1), la fonction hydroxyle de la tyrosine 248 de la carboxypeptidase présente un RMSD de 10,7 Å entre la forme *holo* 1arm, et la forme *apo* 1yme. Le RMSD moyen entre les patches des deux formes est de 2,12 Å, ce qui explique que cette protéine soit en classe III.

Le couple 1byq/1yer (Figure 5.2), le RMSD entre les C-α des patches des deux formes est de 2,7 Å et le RMSD entre les deux patches est de 2,09 Å.

Finalement, le calcul du RMSD entre les patches des formes *holo* et *apo* a permis de classer les protéines selon des modifications structurales plus précises, et nous permet ainsi d'appliquer PatchSearch sur des patches aux changements conformationnels connus et quantifiés.

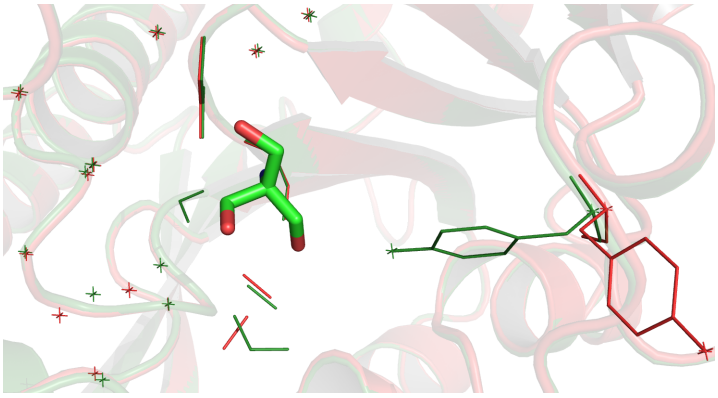


FIGURE 5.1: **Superposition des patches issus des formes *holo* et *apo* de la carboxypeptidase.** En vert, le patch extrait de la forme *holo* de la carboxypeptidase (Id PDB : 1arm). En rouge, la patch extrait de sa forme *apo* (Id PDB : 1yme). Le ligand est représenté en mode bâtons.

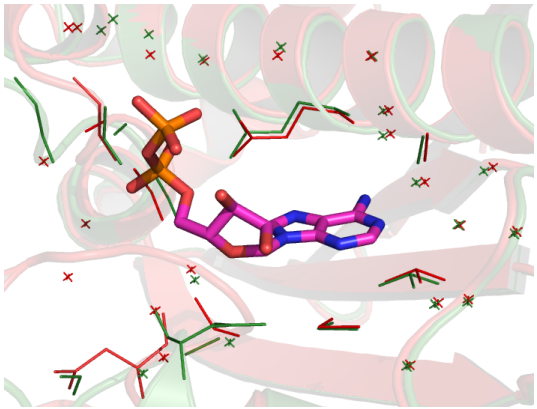


FIGURE 5.2: **Superposition des patches extraits de la forme *holo* du domaine N-terminal de la protéine de choc thermique 90.** En vert, le patch extrait de sa forme *holo* (Id PDB : 1byq). En rouge, le patch extrait de sa forme *apo* (Id PDB : 1yer). Le ligand est représenté en bâton.

5.3 Intérêt d'utiliser des quasi-cliques pour reconnaître des patches

5.3.1 Expériences réalisées et présentation des résultats

Nous avons utilisé PatchSearch, tantôt avec l'approche de cliques, tantôt avec l'approche de quasi-cliques, pour rechercher les patches$_{holo}$ sur les surfaces$_{apo}$. Nous avons comparé les résultats des cliques calculées sur tous les atomes et les quasi-cliques calculées initialement sur un motif d'atomes donné puis étendues avec tous les atomes.

Les résultats des scores obtenus à l'aide des deux approches sont représentés sous forme de boîtes à moustaches. Cette représentation permet de visualiser la distribution des scores, par rapport à la médiane, qui est la ligne épaisse dans la boîte à moustaches. La boîte se divise en deux rectangles, un en-dessous et et un autre au-dessus de la médiane :

— le rectangle inférieur correspond au deuxième quartile : il indique les limites inférieures et supérieures du deuxième quart des scores.

— le rectangle supérieur correspond au troisième quartile : il indique les limites inférieures et supérieures du troisième quart des scores.

Le premier quartile est délimité de l'extrémité basse de l'arête partant du rectangle inférieur au début de ce premier rectangle et contient le premier quart des scores, tandis que le quatrième et dernier quartile, qui contient le dernier quart des données s'étend de la fin du deuxième rectangle à l'extrémité haute de l'arête partant de celui-ci. Les données ne rentrant pas dans la majorité de la distribution ou "*outliers*" sont représentées en dehors des boîtes et segments sous la forme de cercles.

5.3.2 Comparaison des résultats obtenus avec les cliques par rapport à ceux des quasi-cliques

Analyses générales

Nous avons comparé les distributions des scores obtenus avec les cliques et avec les quasi-cliques en regroupant successivement les protéines selon les classes définies par Gunasekaran (Figure 5.3), puis selon les classes pour lesquelles nous avons calculé le RMSD entre les patches$_{holo}$ et patches$_{apo}$ (Figure 5.4).

Le but de cette démarche est de déterminer l'intérêt d'utiliser la recherche de quasi-cliques pour détecter des similitudes entre un patch et une surface plutôt que de rechercher des cliques.

Nous remarquons que les scores médians obtenus avec les classes I avec les cliques et les quasi-cliques sont similaires pour les deux types de regroupements (les classes déterminées par Gunasekaran et celles que nous avons calculées) : 0,94 et 0,96 pour la classe de Gunasekaran et 0,96 et 0,98 pour notre classe.

Nous tirons une conclusion similaire pour les classes II de Gunasekaran et la nôtre. Lors des comparaisons avec les protéines des classes définies par Gunasekaran, les cliques ont un score médian de 0,82 et les quasi-cliques ont un score médian de 0,93. Les classes que nous avons définies nous permettent d'obtenir des scores médians similaires : 0,78 avec les cliques et 0,93 avec les quasi-cliques.

Les deux façons de regrouper les protéines présentent des différences pour les protéines de la classe III. Les protéines de la classe III de Gunasekaran ont des scores médians élevés, que ce soit en utilisant l'approche des cliques (score médian : 0,63) ou l'approche des quasi-cliques (score médian : 0,87). Les protéines de notre classe III ont des scores médians plus faibles : 0,55 en utilisant les cliques et 0,72 avec les quasi-cliques.

Ces observations sont cohérentes avec nos classes et celles définies par Gunasekaran : bien que délestée d'un quart de ses données initiales, la classe I comprend toujours les protéines qui présentent le moins de modifications structurales entre les formes *holo* et *apo*, par conséquent les scores sont élevés quelque soit le type de regroupement utilisé. La classe II comprend certaines protéines des deux autres

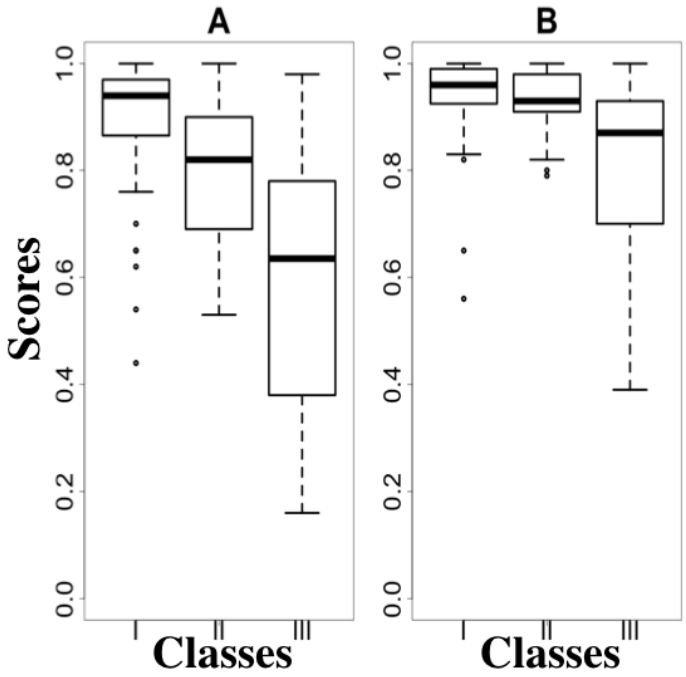


FIGURE 5.3: **Comparaison des distributions des scores obtenus à l'aide des quasi-cliques et des cliques sur le jeu de données de Gunasekaran avec les classes de l'article.** Distributions des scores obtenus lorsqu'on recherche le patch_{holo} sur la surface_{apo} correspondante, en utilisant les cliques (A) et les quasi-cliques (B). Les classes I, II et III sont déterminées par Gunasekaran.

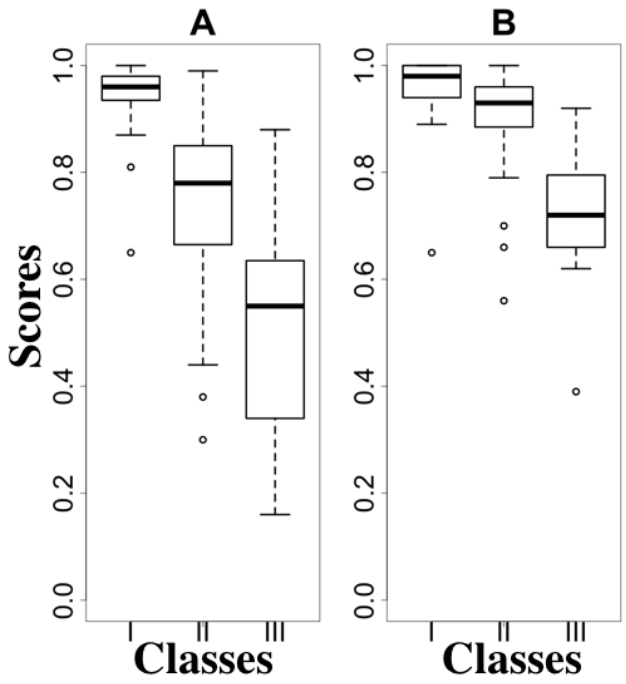


FIGURE 5.4: **Comparaison des distributions des scores obtenus à l'aide des quasi-cliques et des cliques sur le jeu de données de Gunasekaran avec nos classes.** Distributions des scores obtenus lorsqu'on recherche le patch_{holo} sur la surface_{apo} correspondante, en utilisant les cliques (**A**) et les quasi-cliques (**B**). Les différentes classes de protéines ont été définies en calculant le RMSD entre le patch_{holo} et le patch_{apo} de chaque protéine.

classes, donc finalement, ses scores s'équilibrent : elle comprend les protéines qui étaient à la limite supérieure de la classe I de Gunasekaran et celles qui étaient à la limite inférieure de la classe III de Gunasekaran. Finalement, la classe III est celle qui subit le plus de changements et ne compte maintenant que les protéines dont les changements structuraux de tous les atomes du patch sont pris en compte, soient 11 protéines au total, au lieu de 22 selon la classe de Gunasekaran. Les scores propres à cette classe sont par conséquent plus faibles : les patches étant beaucoup plus flexibles, PatchSearch identifie moins d'atomes pour les protéines de cette classe. Un exemple type illustre ce propos, plus loin dans ce chapitre.

Par la suite, nous exploiterons exclusivement les résultats selon le regroupement que nous avons mis au point.

Analyses des cas extrêmes

Les scores des cliques/quasi-cliques sont fortement similaires pour la classe I, on remarque qu'il existe des comparaisons pour lesquelles on obtient un score de 1 peu importe l'approche algorithmique employée. Pour la classe II, le meilleur score obtenu avec les cliques est de 0,99 pour le couple 1bk9/1psj. Enfin, pour les protéines de la classe III, on obtient au maximum un score de 0,88 en recherchant un patch similaire à l'aide des cliques avec le couple 1arm/1yme, tandis que les quasi-cliques permettent d'obtenir un score d'alignement de 0,92 pour le même couple de protéines.

Dans la classe II, le score du couple 5cna/1enq qui a des scores de 0,30 avec les cliques et 0,55 avec les quasi-cliques s'explique par les différences d'atomes présents dans la surface de la protéine entre les formes *holo* et *apo*. PatchSearch apparie le patch de 5cna qui compte 34 atomes avec 24 atomes de $1enq_{surface}$. Le patch *apo* de 1enq ne contient que 25 atomes. Les résidus impliqués dans le patch de 5cna ont subi des changements structuraux importants (Figure 5.5), impliquant ainsi que certains d'entre eux présentent des atomes qui ne sont pas présents dans la surface de la forme apo 1enq, et logiquement, ils ne sont pas présents non plus dans le patch de 1enq. C'est le cas notamment des résidus :

— ASP_{208} qui présente un RMSD de 6,0 Å entre fonctions carboxylates de

l'ASP_{208} de 5cna et 1enq (Figure 5.6),

— les fonctions alcool du résidu TYR_{12} présentent un RMSD de 9,5 Å entre 5cna et 1enq (Figure 5.7).

Le RMSD global mesuré entre les deux formes est de 1,93 Å, ce qui correspond à une moyenne entre les RMSD entre les résidus précédemment évoqués et les résidus qui sont proches dans les formes *holo* et *apo* (Figure 5.5).

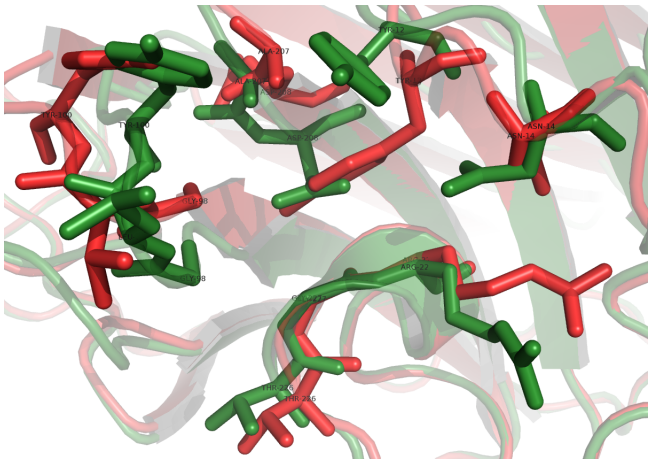


FIGURE 5.5: **Superposition entre les patches extraits des formes *holo* et *apo* de la concanavaline A.** En vert, le patch extrait de la forme *holo* de la concanavaline A (Id PDB : 5cna). En rouge, lle patch extrait de la forme *apo* de la concanavaline A (Id PDB : 1enq).

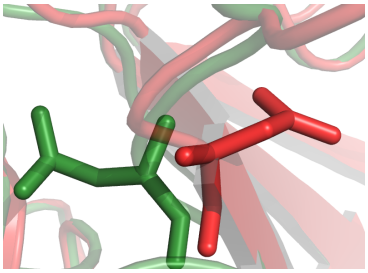


FIGURE 5.6: Comparaison des ASP_{208} extraits des $5cna_{patch}$ (en vert) et $1enq_{patch}$ (en rouge).

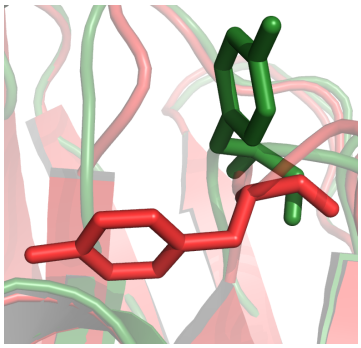


FIGURE 5.7: Comparaison des TYR_{12} extraits des $5cna_{patch}$ (en vert) et $1enq_{patch}$ (en rouge).

Un cas extrême de la classe III est celui de 1ake qui présente le score le plus bas en clique, 0,16 et également en quasi-cliques, 0,39. Le RMSD entre les patches des deux formes est de 6,08 Å. Comme pour le cas précédent, nous avons mesuré les distances entre certains résidus impliqués dans les patches de deux conformations :

— la distance entre les fonctions amines des ARG_{36} est de 6,1 Å (Figure 5.8a).
— celle entre les CA des TYR_{133} est de 15,1Å (Figure 5.8b).
— celle entre les amines des ARG_{156} est de 10,9 Å (Figure 5.8c).
— celle entre les carboxyles des ASN_{138} est de 13,1Å (Figure 5.8b).

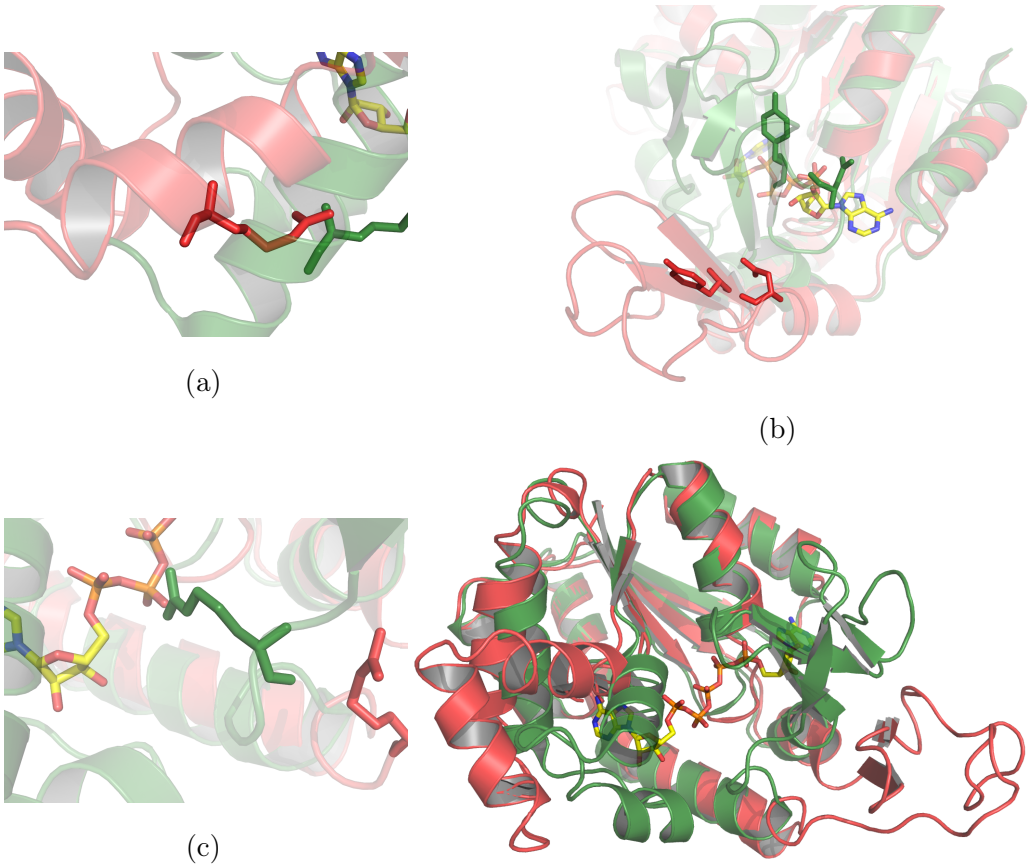
Ces distances importantes sont dues à des mouvements importants de structures secondaires impliquées dans le site de liaison au ligand (Figure 5.8d).

Enfin, le pire des cas de la classe III avec l'approche des cliques est le couple 1hii/1hsi. Seulement 32 atomes du patch de 1hii sont alignés sur les 104 en utilisant les cliques, ce qui aboutit à un score de 0,31. Par contre, lorsque l'on recherche des similitudes avec les quasi-cliques, on peut apparier 46 atomes, ce qui aboutit à un score de PatchSearch de 0,7.

Les quasi-cliques permettent d'apparier quasiment autant d'atomes que les cliques dans les cas où les structures du patch requête et de la surface sont fortement similaires, soit lorsque le RMSD est inférieur ou égal à 2,0 Å. Ce résultat est cohérent avec la définition de la clique et d'une quasi-clique résultant d'appariements d'atomes aux distances internes quasiment toutes conservées, à 2,0 Å près : les deux méthodes aboutissent quasiment au même résultat.

Dans les cas où les distances internes du patch sont moins conservées d'une forme *holo* à une forme *apo*, soient les couples de protéines pour lesquels le RMSD est supérieur à 2,0 Å, ce qui équivaut à notre classe III, l'apport des quasi-cliques est clairement visible : PatchSearch est plus performant que les cliques seules, ce qui se reflète par un appariement d'un plus grand nombre d'atomes.

Cette conclusion montre que l'utilisation de PatchSearch dans la recherche d'*off-target* est pertinente : en effet, les couples de la classe III sont représentatifs des déformations des sites de liaison dans les *off-targets*.



(a) (b) (c) (d)

FIGURE 5.8: **Exemples de résidus impliqués dans les patches issus de la forme *holo* (Id PDB : 1ake) et de la forme *apo* (Id PDB : 4ake) de l'adénylate kinase. Le RMSD (calculé sur tous les atomes) entre les patches de 1ake et 4ake est de 6,08 Å. (a)** : Exemple des ARG_{36} de $1ake_{patch}$ et $4ake_{patch}$. **(b)** : Exemple des TYR_{133} et ASN_{138} de $1ake_{patch}$ et $4ake_{patch}$. **(c)** : Exemple des ARG_{156} de $1ake_{patch}$ et $4ake_{patch}$. **(d)** : Superposition du complexe 1ake et de sa forme *apo* 4ake.

5.4 Vérification des patches appariés

Nous avons également évalué l'identification du patch *apo* précédemment déterminé, lorsque l'on recherche le patch *holo* sur la surface$_{apo}$ de chaque couple à l'aide de la mesure de la Dc. La distribution des Dc obtenues est présentée dans la Figure 5.9 et une focalisation sur les effectifs de cas aux Dc élevées est proposée dans l'histogramme de la Figure 5.10.

La plupart des patches sont trouvés avec une Dc inférieure à 4 Å, soient 89 des couples de protéines. Parmi les 8 autres cas où les patches identifiés ne semblent pas co-localisés avec les patches *apo*, nous étudierons plus précisément deux cas :

— le cas que nous avons évoqué plus tôt, 1ake/4ake pour lequel le Dc est de 6,24Å.

— le cas extrême, 1aj7/2rcs pour lequel la Dc est de 9,78 Å.

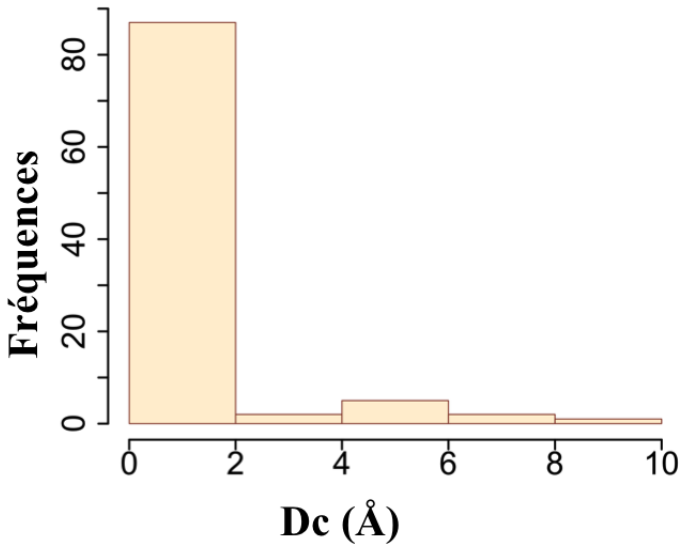


FIGURE 5.9: **Distribution des Dc obtenues en appliquant PatchSearch sur les 97 couples de protéines du jeu de données de Gunasekaran.**

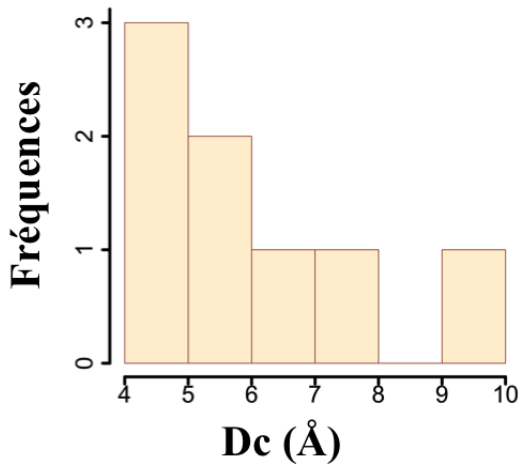


FIGURE 5.10: **Distribution des Dc extrêmes obtenues.**

5.4.1 Mouvement de grande amplitude entre les formes *holo* et *apo*

Le cas 1ake/4ake a déjà été évoqué dans la section 5.3. Nous avions vu que le patch identifié ne comprenait pas la boucle qui effectue un mouvement de repli sur la protéine pour entourer le ligand. Les différences de patches impliqués par le mouvement de cette boucle influent déjà sur les patches de 1ake et de 4ake qui n'ont que la moitié des atomes proches comme on peut le voir dans la superposition des deux protéines dans la Figure 5.11.

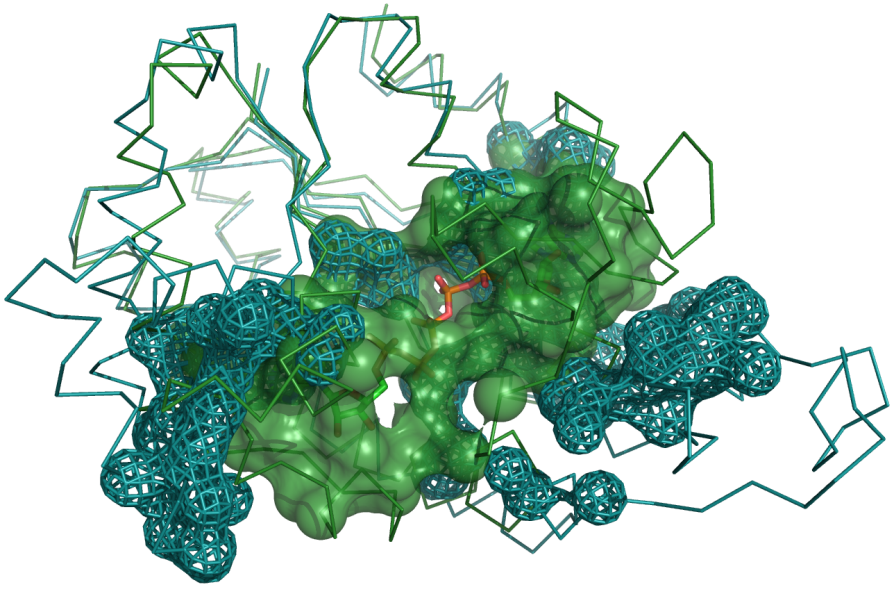


FIGURE 5.11: **Superposition des patches extraits des complexes 1ake et 4ake.** Les chaînes polypeptidiques des protéines 1ake et 4ake sont colorées respectivement en vert et en bleu. La protéine du complexe 1ake est colorée en vert, son patch est représenté en surface. La protéine du complexe 4ake est colorée en bleu et son patch est en grillage bleu.

Le patch a été identifié sur 4ake avec le paramètre *maxdist* de 3 Å (Figure 5.12). Mais lorsque nous augmentons le paramètre *maxdist* à 6,0 Å, le patch retrouvé est plus large et son centre de gravité se rapproche donc de celui du patch de 1ake (Figure 5.13). Néanmoins, il serait difficile d'identifier un patch sur 4ake dont la Dc serait inférieure à 4 Å par exemple : les distances entre les atomes de la boucle et ceux du coeur du patch sont trop importantes dans la forme *apo*, ce qui aboutit à

un RMSD entre les deux formes de 6,07 Å.

5.4.2 Mouvements de chaînes entre les structures *apo* et *holo*

Le couple 1aj7/2rcs a la Dc la plus importante de notre jeu de données, soit 9,78 Å. La visualisation de la forme *holo* et de la forme *apo* (Figure 5.14) permet de constater que les deux structures semblent identiques au niveau de la chaîne polypeptidique.

Les patches *holo* et *apo* sont, quant à eux, de formes tout à fait différente (voir la Figure 5.15) : ces différences de formes sont dues à l'absence de presque la moitié des atomes correspondant au patch de 1aj7 dans la surface de 2rcs : le patch de 1aj7 contient 54 atomes, tandis que son correspondant dans 2rcs n'en compte que 28. Ce patch d'interaction contient des atomes des chaînes L et H, le patch de 1aj7 contient 19 atomes de la chaîne H, tandis que celui de 2rcs en a seulement 9. Ces différences significatives d'effectifs d'atomes pour la chaîne H dans les différents patches indiquent que bien que les modifications structurales entre les formes *holo* et *apo* soient modérées (leur RMSD vaut 1,68 Å), elles ont suffi à induire l'enfouissement de presque la moitié des atomes de la forme *holo* sous la surface de la protéine. Finalement, la Dc élevée calculée correspond à la distance entre le patch de 2rcs trop incomplet pour être pris comme référence, et le patch identifié sur la surface de 2rcs, qui correspond à un patch de forme tout à fait similaire à celle du patch de 1aj7 (Figure 5.16).

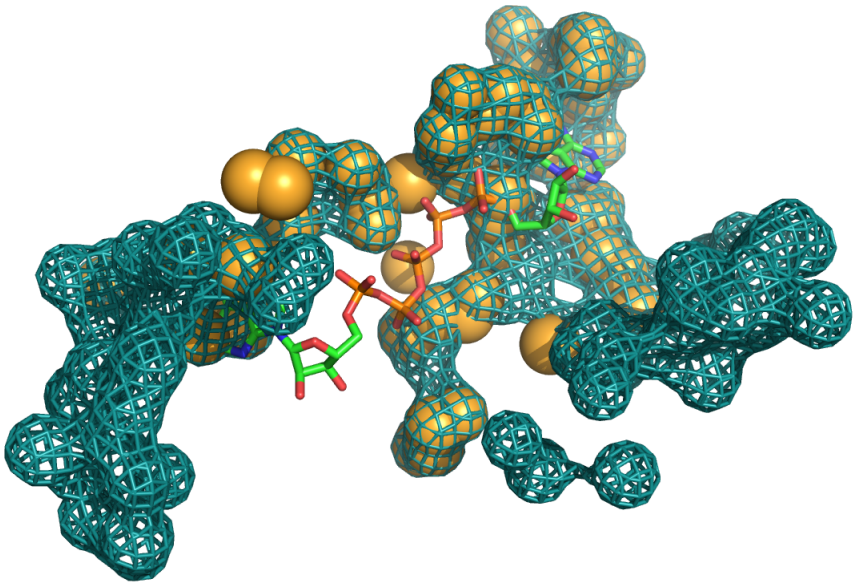


FIGURE 5.12: **Superposition du patch identifié sur 4ake et patch connu de 4ake, maxdist = 3 Å.** Le patch identifié par PatchSearch sur 4ake est coloré en orange et est représenté sous forme de sphères. Le patch connu de 4ake est coloré en bleu et est représenté sous la forme d'un grillage.

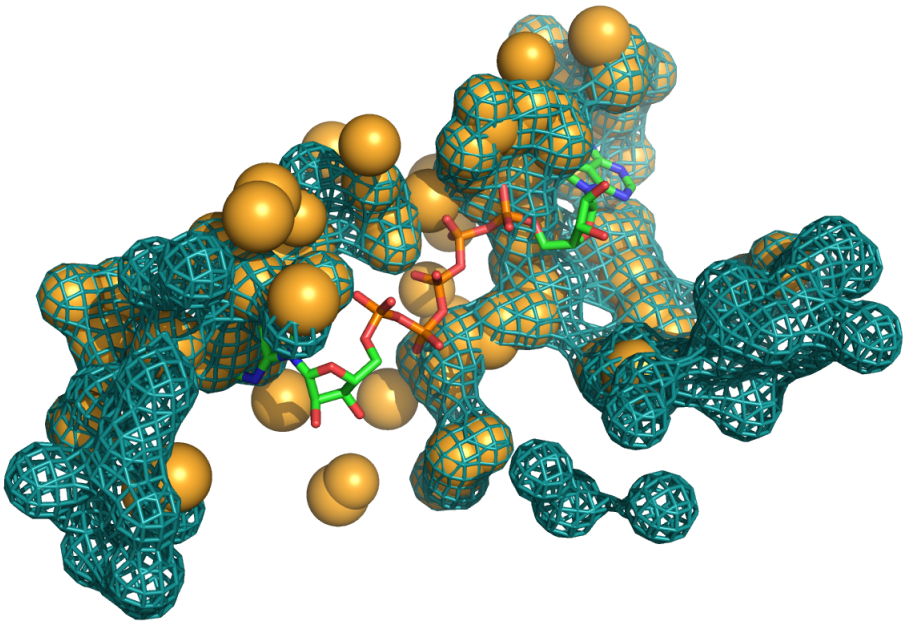


FIGURE 5.13: **Superposition du patch identifié sur 4ake et du patch connu de 4ake, maxdist = 6 Å.** Le patch identifié par PatchSearch sur 4ake est coloré en orange et est représenté sous forme de sphères. Le patch connu de 4ake est coloré en bleu et est représenté sous la forme d'un grillage.

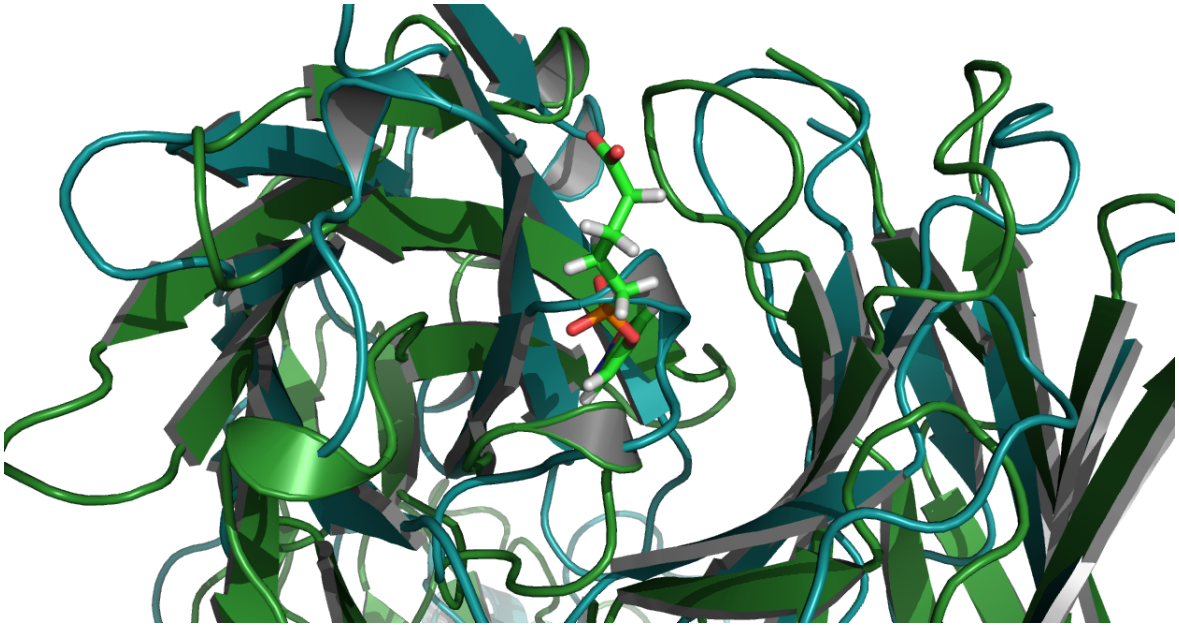


FIGURE 5.14: **Superposition des formes *holo* et *apo* de l'hommunoglobuline 48g7.** Superposition des formes *holo* (Id PDB : 1aj7) et *apo* (Id PDB : 2rcs) de l'immunoglobuline 48g7. La protéine de 1aj7 est colorée en vert et est représentée en mode rubans. La protéine de 2rcs est colorée en bleue et est représentée sous forme de rubans.

5.5 Discussion et conclusions

Ce chapitre expose l'application de PatchSearch dans des cas concrets de patches conformationnellement modifiés. Nous montrons ainsi que la reconnaissance d'un patch, dans une protéine de conformation différente, basée sur la méthodes des cliques est efficace si les structures sont très similaires : au maximum, les écarts entre deux atomes appariés ne peuvent varier qu'entre 1,5 à 2,0 Å. Pour ces cas précis, la recherche de quasi-cliques améliore faiblement les résultats. Nous avons été en présence de ces cas dans la classe I, voire dans la classe II aussi. Les résultats obtenus par PatchSearch sur les protéines de la classe III mettent en évidence les performances de PatchSearch pour :

— apparier la majorité des patches avec des scores significatifs et

— localiser des patches de façon précise par rapport au patch connu de la protéine.

La recherche de patches sur des surfaces ayant subi des déformations structurales

de plus forte amplitude s'apparente plus aux couples de la classe III qui reflète :

— des grands mouvements de chaînes latérales. Zavodszky M.I. et Kuhn L.A. [187] ont étudié les mouvements des chaînes latérales des résidus impliqués dans les sites de liaison et ont conclu que de petits angles de rotations sont environ 20% plus fréquents dans les sites de liaisons que dans le reste de la surface protéique, et plus précisément que les modifications d'angles dièdres dans la chaîne latérale est caractéristique des sites de liaison, afin d'optimiser l'interaction avec le ligand.

— de mouvements de structures secondaires de grande amplitude.

Les résultats de ce chapitre suggèrent que PatchSearch est adapté pour la reconnaissance des patches dont la conformation structurale a beaucoup évolué entre les formes *apo* et *holo* : dans le cas où les atomes du patch sont exposés dans les deux formes d'une même protéine, PatchSearch parvient à identifier le patch dans les différentes conformations d'une même protéine. Le problème de l'exposition des atomes au solvant pourrait se poser par exemple dans l'étude de patches issus de canaux : l'extraction des surfaces doit prendre en compte l'enfouissement des atomes du canal qui sont au contact de la membrane.

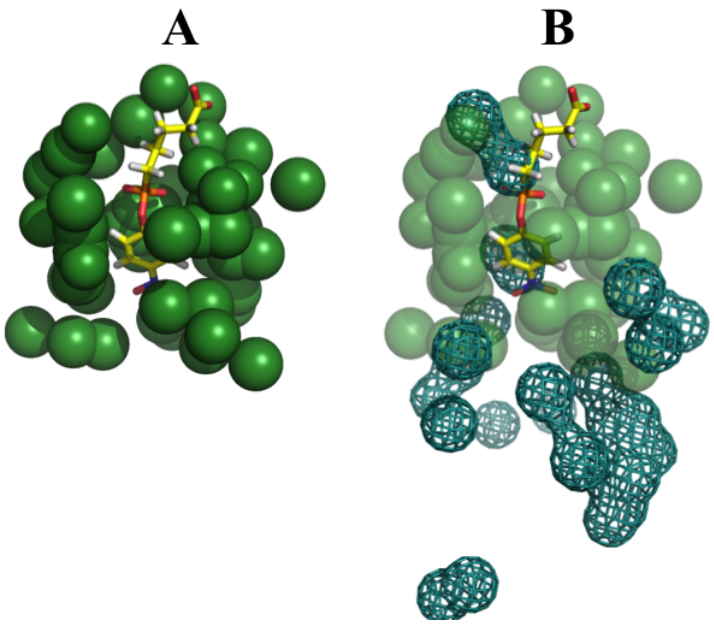


FIGURE 5.15: **Superposition des patches issus de 1aj7 et 2rcs. A** : patch extrait de 1aj7. Le patch de 1aj7 est coloré en vert et est représenté sous forme de sphères. Le ligand est représenté en bâtons. **B** : Superposition des patches de 1aj7 et 2rcs. Le patch de 2rcs est coloré en bleu et est représenté en grillage.

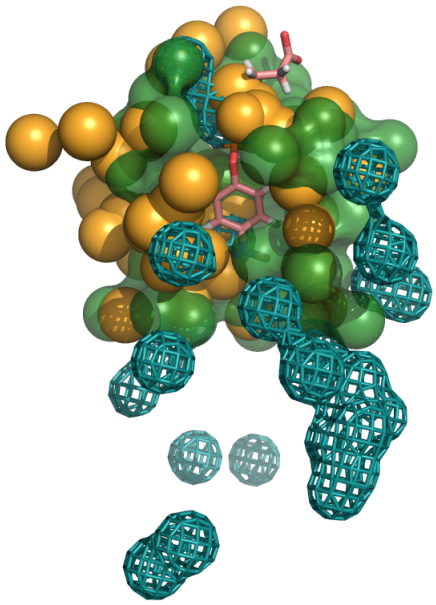


FIGURE 5.16: **Superposition des patches de 1aj7, de 2rcs et du patch identifié sur 2rcs.** Le patch de 1aj7 est coloré en vert et est représenté sous forme de sphères. Le ligand est représenté en bâtons. Le patch de 2rcs est coloré en bleu et est représenté en grillage. Le patch identifié par PatchSearch sur la surface de 2rcs est coloré en orange et est représenté sous la forme de sphères.

Chapitre 6

Recherche de patches sur des protéines différentes

Les résultats présentés dans le chapitre précédent montrent que PatchSearch reconnait un patch sur des surfaces d'une même protéine dans des conformations différentes. Toutefois, nous avons aussi souhaité étudier si PatchSearch parvenait à reconnaître des similitudes entre deux patches issus de protéines différentes, mais interagissant avec un même ligand. Cette démarche correspond ainsi à l'identification de similitudes entre différentes surfaces protéiques, et s'inscrit dans notre volonté de détecter d'éventuelles *off-targets* pour un ligand donné. De plus, elle s'inscrit dans la recherche ambitieuse de patches "signatures" qui pourrait permettre à terme d'identifier des ligands de protéines encore orphelines.

Plusieurs outils de comparaison de patches partagent ce même but, ce qui entraîna la conception de jeux de données protéiques communément employés dans ce domaine, dits également “benchmarks”. Nous avons donc entrepris d'appliquer PatchSearch sur ces benchmarks afin de comparer ses performances aux autres outils et d'évaluer sa capacité à détecter des similitudes entre un patch donné et les surfaces des différentes cibles du même ligand.

6.1 Description des complexes protéine-ligand utilisés

6.1.1 Jeu de référence mis en place par Kahraman

Le premier groupe de complexes protéines-ligand sur lequel nous nous basons a été mis au point en 2007 par Kahraman et ses collègues [79] dans le but de comparer les patches de protéines non-homologues entre eux, ainsi que la forme des ligands entre eux. Ce groupe de complexes a été choisi selon les critères suivants :

— deux protéines interagissant avec un même ligand doivent appartenir à deux super-familles d'homologie différentes : c'est le niveau le plus élevé de classification CATH [135] qui recense les structures PDB expérimentalement résolues et les classifie. En d'autres termes, il n'existe pas de relation d'homologie entre les domaines protéiques.

— pour chaque protéine impliquée dans un processus enzymatique, une vérification a été effectuée concernant l'implication effective de son ligand dans une réaction enzymatique (à l'aide des numéros EC [12], EC pour "Enzyme Commission", ce numéro permet de classer les enzymes d'après les réactions qu'elles catalysent).

— pour les protéines non-enzymatiques, une recherche a été effectuée dans la base de données protéiques Uniprot [9] afin de s'assurer que le ligand est effectivement connu pour effectuer une interaction avec la protéine d'intérêt.

Au total, ce jeu de données, que nous appellerons par la suite KD (pour "Kahraman Dataset") comptabilise 100 patches issus de 100 complexes (Tableau Annexe B.1), subdivisés en 9 groupes de ligands (Figure 6.1) à partir desquels les patches ont été extraits : AMP, ATP, FAD, NAD, Stéroïdes, FMN, Glucose (GLC), Hème et Phosphate (PO_4). Notons qu'initialement, le complexe 1qla était dans le groupe des protéines liant l'hème, mais ce complexe était obsolète par la sortie du complexe 2bs2 dont la résolution est meilleure.

Le jeu de données mis en place par Kahraman est devenu un benchmark classique de référence dans le domaine de la comparaison de patches [71, 18, 25, 45, 85, 146,

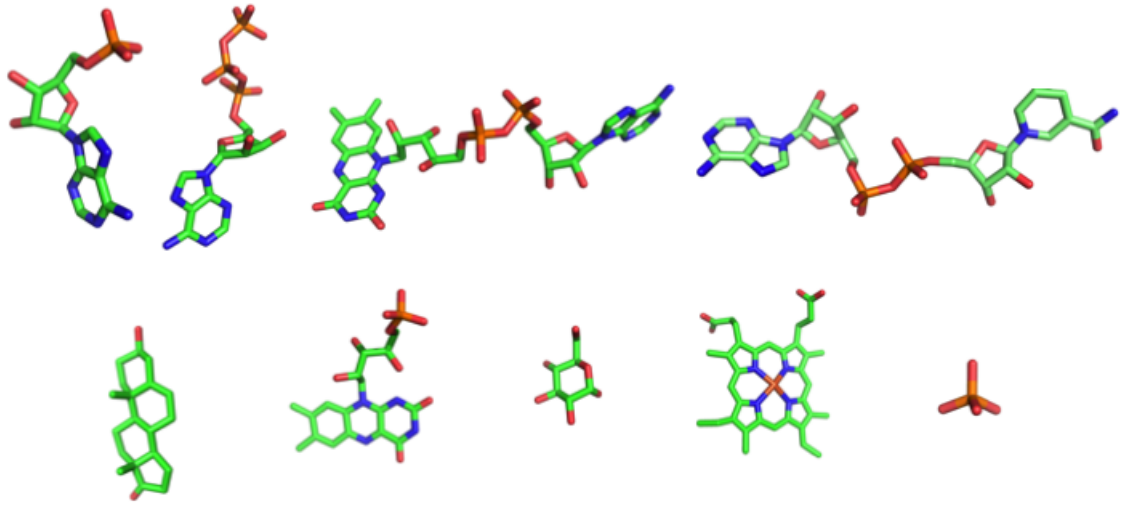


FIGURE 6.1: **Structures des ligands du benchmark de Kahraman.** De gauche à droite : AMP, ATP, FAD, NAD, Stéroïdes, FMN, GLC, Hème et PO_4.

145].

6.1.2 Second jeu de référence utilisé : "Homogeneous dataset"

A la suite de ces observations, Hoffmann et ses collaborateurs ont mis en place un autre jeu de données que nous nommerons HD pour "Homogeneous Dataset" constitué de 99 patches, distribués en 10 groupes. Dans HD, les ligands (présentés dans la Figure 6.2) sont de tailles (en nombre d'atomes) similaires (Tableau Annexe B.2). Le jeu de données de base est constitué de 100 patches. Le complexe 1k87 a été remplacé par la structure 4o8a. Toutefois, cette nouvelle structure ne contient pas de ligand. Par conséquent, nous avons exploité 99 complexes au lieu des 100 complexes du jeu initial.

En plus des patches extraits par Hoffmann, nous avons extrait les patches et les surfaces des complexes de KD et HD en utilisant le protocole d'extraction expliqué dans la section 4.

6.1.3 Préparation des patches issus de HD

Lors du développement de leur méthode sup-CK [71], Hoffmann et ses collaborateurs ont extrait leur propres patches (correspondant aux atomes se trouvant entre

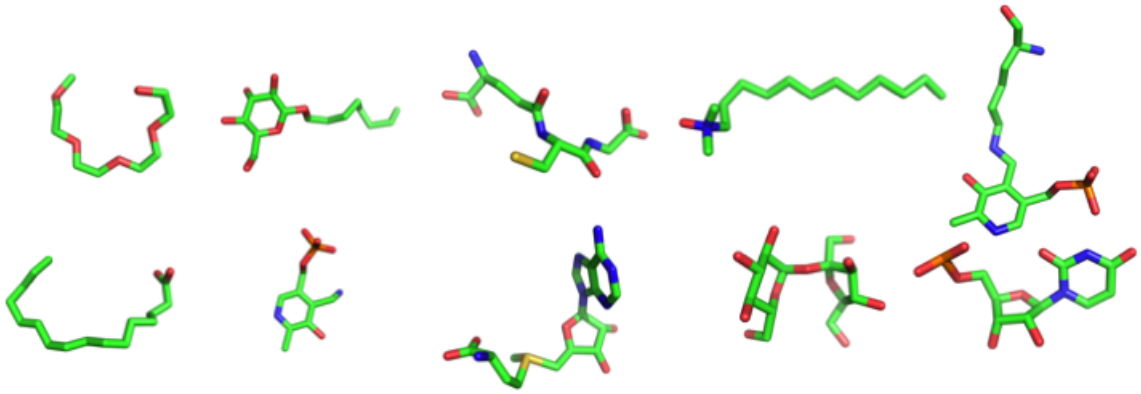


FIGURE 6.2: **Structures des ligands du "Homogeneous Dataset".** De gauche à droite : 1PE, BOG, GSH, LDA, LLP, PLM, PMP, SAM, SUC, U5P.

5,3 Å et 6,0 Å du ligand) et les ont mis en ligne. Nous les avons téléchargés à partir de l'adresse `http://cbio.mines-paristech.fr/paris/paris.html`.

Nous avons au final 4 jeux de données :

— K_{nous} et HD_{nous} : les patches et surfaces du KD et HD extraits selon notre protocole.

— KD_{orig} et HD_{orig} : les patches de KD et HD extraits par Hoffmann que nous avons récupérés.

Afin de pouvoir utiliser PatchSearch sur les patches d'Hoffmann, nous les avons traités de la façon suivante :

1. les centroïdes des résidus aromatiques sont calculés sur les résidus entiers.
2. pour chaque résidu aromatique : si au moins un de ses carbones des cycles aromatiques se trouve dans un rayon de 5,3 Å du ligand, les carbones aromatiques de ce résidu sont remplacés par son centroïde.

Nous avons également effectué une vérification afin d'extraire nos patches de la même façon que ceux de KD_{orig} et HD_{orig} :

— les patches doivent être extraits sur les mêmes chaînes protéiques, et ce, même si ceux de KD_{orig} et HD_{orig} ont été extraits sur des chaînes différentes de celles explicitées dans les articles de référence. Les protéines concernées par ces changements de chaînes sont :

- pour KD : 1gn8 (chaîne A → B), 1e8g (A → B), 1qpa (A → B), 1fbt (A → B), 1ho5 (A → B), 1l7m (B → A), 1rdq (E → I).
- pour HD : 1b4w (D → A), 1bjw (B → A), 1cmc (B → A), 1dug (B → A), 1dxr (M → H), 1y10 (B → A), 2hd0 (E → A).

— si plusieurs molécules de ligand sont présentes, le patch doit être extrait sur la même molécule.

- pour KD : 1iqc (patch extrait sur HEM_{401}), 1ejd (PO_4-$_{2422}$), 1gyp (PO_4-$_{359}$), 1h6l (PO_4-$_{501}$), 1e9g (PO_4-$_{3002}$).
- pour HD : 1aij (LDA_{312}), 1ar1 (LDA_{272}), 1dxr (LDA_{1259}), 1fx8 (BOG_{473}), 1g8i ($1PE_{951}$), 1jjo (SUC_{2380}), 1kmo (LDA_{742}), 1thq (LDA_{200}), 1xkw (LDA_{2001}), 2c37 ($U5P_{403}$), 2hdo (BOG_{218}).

— si les protéines sont multimériques et que le ligand se trouve entre plusieurs chaînes, les chaînes traitées doivent être identiques à celles dont sont extraits les patches de KD_{orig} et HD_{orig} :

- pour KD : 1o9t (A, B), 1tid (A, B), 1h69 (A,C), 1jqi (A, B), 1f5v (A, B), 1d7c (A, B), 1qax (A, B), 1euc (A, B), 1nf5 (A, B).
- pour HD : 1a0t (P, R), 1aia (A, B), 1b4w (A, D, C), 1bwo (A, B), 1c51 (A, C), 1dbt (A, B), 1fgx (A, B), 1iug (A, B), 1iyh (A, C), 1jg8 (A, D), 1o57 (C, D), 1pq2 (A, B), 1r4w (A, B), 1umx (H, M), 1x8 (B, C), 2czv (A, C, D), 2iu8 (A, B), 2p4b (A, B).

6.2 Recherche de similitudes entre deux patches d'un même ligand

Comme précédemment décrit (section 6.1.3), nous avons préparé les patches extraits par Hoffmann. Les AUC résultant des recherches de similitudes entres les différents patches sont présentées dans le Tableau 6.1.

Nous constatons que les AUC_{totale} sont significativement inférieures aux $AUC_{ligands}$ dans les tests effectués sur le KD. Cette différence est due aux écarts

Tableau 6.1: **Résultats des différentes comparaisons de patches de KD et HD.**

Comparaisons	AUC_{totale}	$AUC_{ligands}$
$\mathbf{K_{nous}}$ **vs** $\mathbf{K_{nous}}$	0,67	0,73
$\mathbf{K_{orig}}$ **vs** $\mathbf{K_{orig}}$	0,78	0,73
$\mathbf{HD_{nous}}$ **vs** $\mathbf{HD_{nous}}$	0,67	0,67
$\mathbf{HD_{orig}}$ **vs** $\mathbf{HD_{orig}}$	0,74	0,67

D'une part, l'AUC_{totale} correspond à la somme des AUC obtenues pour tous les patches d'un même ligand. D'autre part, l'$AUC_{ligands}$ est l'AUC obtenue pour tous les patches d'un ligand divisée par le nombre de patches de ce ligand dans le jeu de données d'étude.

d'effectifs de patches entre les différentes classes de ligands pour le KD. En effet, certaines classes contiennent un nombre important de patches :

— le phosphate (PO_4) qui compte une vingtaine de patches,

— l'hème qui en compte 16,

— d'autres groupes ne présentent que très peu de patches comme le glucose et les stéroïdes qui comptent chacun 5 patches.

Les AUC moyennes du GLC (0,81) et des stéroïdes (0,78) sont très élevées (Tableau 6.2) et comptent autant que les AUC moyennes des ligands PO_4 (0,61) et hème (0,77). Ces différences d'effectifs entraînent finalement une AUC moyenne par ligands biaisée qui n'est pas représentative de la diversité d'effectifs entre les différentes classes de patches et augmentent donc artificiellement l'$AUC_{ligands}$. Nous ne garderons donc que la valeur d'AUC_{totale} pour comparer nos résultats à ceux des autres programmes de comparaison de patches. Ce problème d'AUC biaisée ne se pose pas pour le HD : il y a autant d'effectifs pour toutes les classes de ligands (10 patches), mis à part le groupe de 9 patches de 1PE.

Par la suite, nous comparons les résultats obtenus sur les patches K_{orig} et HD_{orig} en AUC_{totale}

Tableau 6.2: **AUC obtenues pour retrouver les patches des ligands de Knous.**

Ligands	**AUC**
AMP	0,74
ATP	0,64
PO_4	0,61
GLC	0,81
FAD	0,74
HEM	0,77
FMN	0,65
AND	0,96
NAD	0,64

6.3 Comparaison des performances de PatchSearch aux outils existants

Les jeux de données KD et HD sont utilisés par d'autres outils de comparaison de patches afin d'évaluer leurs performances [71, 18, 25, 45, 85, 146, 145]. Nous avons souhaité comparer les performances de PatchSearch, en terme d'identification de similitudes entre deux patches, à celles des autres programmes qui y sont dédiés.

Le Tableau 6.3 permet de conclure que PatchSearch obtient des résultat similaires à ceux des autres outils de comparaison de patches.

Tableau 6.3: **AUC obtenues par les différents outils de comparaison de patches sur les jeux de données KD et HD.**

	PatchSearch	H. S.	sup-CK	MultiBind	PSIM	Patch-Surfer	IsoCleft
KD	0,78	0,64	0,86	0,71	0,79	0,81	0,62
HD	0,74	NR	0,71	NR	NR	0,84	NR

Dans la première ligne sont présentés les différents outils dont les performances sont comparées. Les lignes suivantes contiennent successivement les performances des outils de comparaison de sites de liaison sur le KD puis sur le HD. NR signifie "Non Renseignée". H. S. représente les harmonies sphériques développées par Kahraman.

Kahraman et ses collègues ont déterminé l'influence des caractéristiques "taille" et "forme" des patches sur les AUC obtenues :

— lorsque seule la forme des patches est prise en compte, l'AUC obtenue est de 0,64.

— lorsque seule la taille des patches est utilisée, ils obtiennent une AUC de 0,73.

— lorsque la forme et la taille des patches sont prises en compte, leur AUC s'élève à 0,77.

L'ajout du paramètre "taille du patch" est un critère permettant d'augmenter les AUC pour leur outil comme pour PatchSearch, qui dans ces conditions, a une AUC de 0,82. Il en est de même pour Patch-Surfer : lorsque les caractéristiques du patch sont prises en compte sans la taille, l'outil obtient une AUC de 0,81 et 0,84 lorsque la taille est prise en compte.

L'outil sup-CK [71] obtient les meilleures AUC présentées ici. Néanmoins, les paramètres de PatchSearch sont constants quelque soient les patches comparés, contrairement à sup-CK qui optimise son paramètre σ (qui mesure la sensibilité de la mesure de similarité en fonction du déplacement des atomes) pour chaque comparaison suite à une double validation croisée. Selon l'approche utilisée qui tient compte des charges partielles ou seulement des caractéristiques géométriques, ce

paramètre varie entre 1 et 4. Toutefois, les auteurs de cet article lui attribuent les avantages suivants :

— la superposition de deux patches ne nécessite aucune similitude de structure ou de séquence.

— sup-CK ne superpose pas chaque atome de façon individuelle, mais superpose globalement des groupes d'atomes.

Nous avons également souhaité nous comparer à l'outil eMatchSite développé plus récemment en 2014, par Brylinski [26]. Cette méthode permet d'aligner les structures des sites de liaison protéique indépendamment de leur séquence en acides-aminés. Cet outil obtient une AUC de 0,69 sur le KD et 0,92 calculée sur le HD. Cependant, ces résultats de classifications ont été obtenus sur des sous-jeux de données restreints : 53 patches issus du KD et 51 patches issus du HD. Les comparaisons effectuées n'étant spécifiées ni dans l'article correspondant aux expériences réalisées, ni sur le site répertoriant les données sur lesquels l'auteur s'est basé, nous ne pouvons pas comparer ses AUC obtenus par rapport aux nôtres et à celles des autres outils qui ont été calculées sur l'intégralité des comparaisons.

Finalement, PatchSearch obtient des résultats similaires à ceux obtenus par les autres outils de comparaison de patches. Cette première indication positive sur la comparaison de patches est un préalable pour l'application de PatchSearch, mais ne correspond pas au but premier de la conception de PatchSearch qui est l'identification d'un patch sur la surface d'une autre protéine.

6.4 Reconnaissance de patches sur les surfaces des protéines de KD et HD

Nous avons également évalué la capacité de PatchSearch à identifier les différents types de patches sur les surfaces correspondantes : nous avons donc recherché chacun des 100 patches sur les 100 surfaces.

L'AUC_{ligand} est calculée en prenant en compte uniquement les comparaisons de patches et de surfaces d'un ligand donné. Les AUC_{ligand} obtenues pour le KD et

Tableau 6.4: **AUC obtenues pour retrouver les patches des ligands de KD.**

Ligands	**AUC**
AMP	0,68
ATP	0,61
PO_4	0,61
GLC	0,81
FAD	0,67
HEM	0,75
FMN	0,53
AND	0,78
NAD	0,63

Tableau 6.5: **AUC obtenues pour retrouver les patches des ligands de HD.**

Ligands	**AUC**
PMP	0,76
SUC	0,53
LDA	0,56
BOG	0,63
PLM	0,54
SAM	0,66
U5P	0,64
GSH	0,68
1PE	0,50

le HD sont respectivement présentées dans les Tableaux 6.4 et 6.5. Ces valeurs d'AUC nous indiquent que PatchSearch reconnaît les patches de KD et HD avec une sensibilité et une spécificité proche d'un détecteur aléatoire : la plupart des AUC sont proches de 0,5.

Néanmoins, pour les protéines considérées de HD et KD, il est difficile de discuter de la spécificité pour reconnaître un patch sur une surface donnée. Il faudrait en effet considérer comme faux positif un patch d'ATP qui a été identifié sur une surface liant l'AMP par exemple, au-delà d'un score seuil. Dans cet exemple, affirmer qu'il s'agit d'un faux positif peut être faux pour deux raisons :

— la protéine peut présenter des patches pour deux ligands différents. Par exemple, dans le complexe 1cqx (Figure 6.3), la flavohémoglobine est composée d'un domaine de la famille des globines liant l'hème et d'un deuxième domaine à activité oxydoréductase qui lie le FAD. Cette protéine doit donc être considérée comme vrai positif lorsque chacun des patches d'hème et de FAD sont identifiés à sa surface.

— si l'on souhaitait vraiment définir une protéine comme partenaire négatif d'un

ligand donné, il serait nécessaire d'étudier expérimentalement l'interaction de la protéine en présence du ligand.

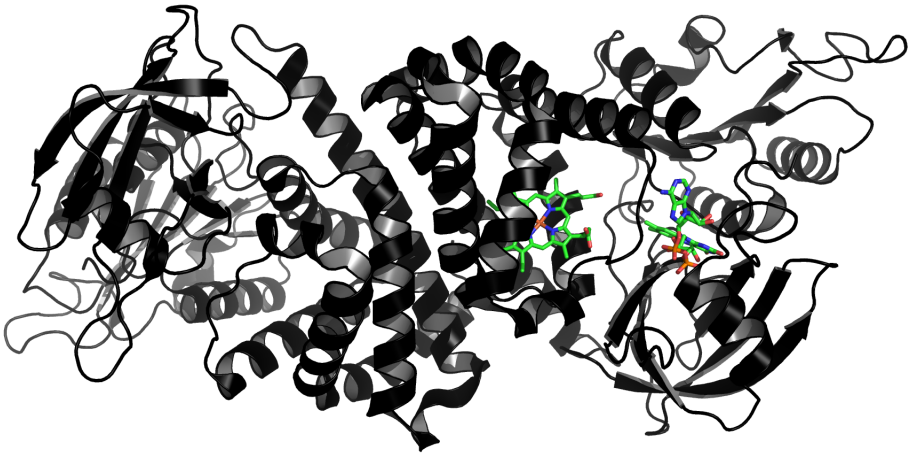


FIGURE 6.3: **Exemple de la flavohémoglobine qui présente des patches pour des ligands différents.** La flavohémoglobine lie à la fois l'hème et le FAD (Id PDB : 1cqx).

D'autre part, les ligands du HD n'ont pas été vérifiés pour leur spécificité d'interaction avec leurs protéines partenaires, contrairement aux ligands du KD. Le pentaéthylène glycol (PEG) par exemple est un polymère utilisé pour cristalliser les protéines, par conséquent, sa présence au voisinage d'une protéine et certainement pas spécifique. Il en est de même pour le beta-octylglucoside (BOG) et l'oxyde de lauryl-diméthylamine (LDA) qui sont des tensioactifs utilisés comme détergents pour solubiliser les protéines membranaires. Par exemple, la cytochrome oxydase C est complexée avec plusieurs molécules de LDA (Figure 6.4) dans le complexe 1ar1 : cette enzyme a été extraite d'une membrane, et le LDA a permis de la solubiliser. Il en est de même pour le complexe 1i78 (Figure 6.5) qui contient une protase extraite d'une membrane externe, solubilisée à l'aide de molécules de BOG.

Cette expérience met ainsi en évidence les difficultés éprouvées pour discuter de la spécificité d'un patch identifié sur la surface d'une protéine donnée. Le but premier de cette démarche était de mesurer la spécificité et la sensibilité de PatchSearch pour identifier les patches des ligands des KD et HD. Toutefois, l'existence de "faux positifs" montrent une des limites à la détection de cibles secondaires pour un ligand donné. De plus, comme je l'ai détaillé plus tôt, il est essentiel de disposer

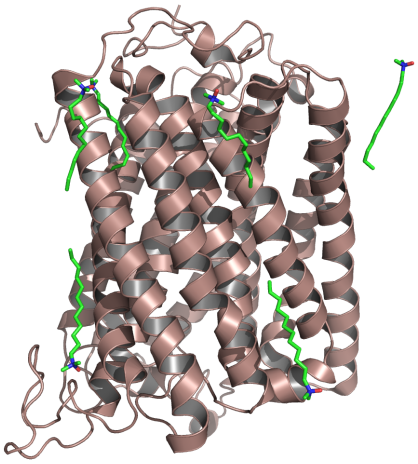


FIGURE 6.4: **Exemple de la cytochrome oxydase C complexée avec plusieurs molécules de ligand.** La cytochrome oxydase C est complexée à six molécules de LDA (Id PDB : 1ar1).

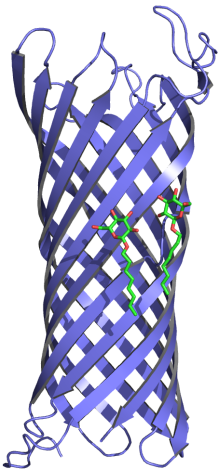


FIGURE 6.5: **Exemple d'une protéase complexée avec plusieurs molécules de ligand.** Protéase complexée à deux molécules de BOG (Id PDB : 1i78).

d'interactions avec une forte affinité et par conséquent des ligands choisis dans ce but afin de pouvoir discuter de l'efficacité de notre outil.

6.5 Discussion et conclusions

Finalement, le jeu de données proposé par Kahraman ne paraît pas être le benchmark adapté pour étudier la capacité d'un outil de recherche de patches sur une surface donnée. En effet, il présente certains aspects qui sont discutables :

— certaines protéines sont cristallisées avec plusieurs ligands différents, ce qui ne nous permet pas d'émettre un avis sur la spécificité de notre reconnaissance de patch sur une surface donnée,

— tous les ligands ne sont pas représentés à effectifs similaires, comme nous en avons discuté dans la section 6.4. Il n'est par conséquent pas possible d'affirmer qu'en moyenne PatchSearch reconnaît mieux les patches de GLC par rapport à ceux de PO_4, par exemple.

— certains des ligands choisis partagent de fortes similitudes structurales (Tableau 6.1) :

- l'AMP, l'ATP, le FAD et le NAD présentent tous trois groupements communs : une adénine, couplée à un ribose, lui-même lié à un phosphate (un pour l'AMP, deux pour le NAD et le FAD, trois pour l'ATP),
- il en est de même pour les ligands FMN et FAD qui sont tous les deux des dérivés de la riboflavine et par conséquent, partagent ce groupement en commun. Ces groupements communs peuvent amener les patches à présenter également des similitudes.

Ainsi, la notion de faux-positifs lorsqu'on compare un patch d'AMP et un autre d'ATP avec un fort score n'est pas forcément aussi tranchée que lorsque l'on compare des patches extraits de ligands plus différents (comme un patch de GLC avec un patch d'hème par exemple).

Une classification des ligands au préalable selon les différents groupements qu'ils présentent s'avère nécessaire. Certains outils de comparaison ont tenu compte de ces fortes similitude entre ligands : par exemple, lorsque le programme IsoCleft est testé sur le KD en 2008 [128], ils comptabilisent un fort score de similitude trouvé entre deux patches d'AMP, ATP, FAD ou NAD comme positif, et de même pour les patches de FMN ou FAD. L'étude de Brylinski avance le même argument de similitude de ligands et montre qu'en groupant les patches d'AMP, ATP et NAD, l'AUC d'eMatchSite sur le KD augmente à 0,79 (sans ce regroupement, elle vaut 0,69).

Le benchmark HD a l'avantage de comparer des classes de patches d'effectifs similaires. Cependant, les ligands choisis ne sont pas tous les ligands physiologiques impliqués dans une interaction spécifique avec les protéines co-cristallisées.

Ces différentes raisons nous ont poussé à n'examiner que les scores obtenus par PatchSearch et à ne pas calculer les Dc des patches trouvés sur les surfaces ciblées, les interactions des ligands avec les protéines n'étant pas forcément toutes spécifiques.

Les expériences présentées tout au long de ce chapitre mettent en évidence que PatchSearch est aussi performant que les autres outils existants pour reconnaître deux patches d'un même ligand. Le but principal du développement de cet outil

reste néanmoins d'identifier des similitudes entre un patch donné et la surface d'une protéine cible. Afin d'évaluer la sensibilité et la spécificité de la reconnaissance de patches sur des cibles secondaires, il serait nécessaire de disposer d'un jeu de données contenant des complexes protéine-ligand ayant une forte affinité mais également des cas de protéines connues pour ne pas lier les ligands du jeu de données ou au mieux avec une très faible affinité. Ainsi, un patch reconnu sur une protéine avec laquelle le ligand concerné a une affinité quasiment nulle pourrait être considéré comme faux positif à plus juste titre. Dans le cas où on ne dispose pas de ces informations expérimentales, il serait intéressant de disposer d'une banque de complexes protéine-ligand dont les énergies d'interaction ont été calculées.

Chapitre 7

Identification de patches spécifiques aux ligands polypharmacologiques

Ce dernier chapitre de résultats expose l'application de PatchSearch dans la reconnaissance de patches de ligands interagissant avec des cibles multiples. Ces ligands polypharmacologiques sont utilisés dans différentes pathologies et/ou ciblent plusieurs protéines impliquées dans un même processus. Á ce propos, la revue de l'état de l'art axée sur la conception de ligands à cibles multiples réalisée par L. Costantino et D. Barlocco en 2012 [37] rappelle ainsi la nécessité de cibler plusieurs noeuds des cascades de signalisation cellulaires afin de contourner les systèmes de secours développés par la cellule et d'optimiser les chances d'observer un effet phénotypique du traitement. Je présenterai également l'application de ProBiS sur les mêmes protéines. Cet outil permet d'identifier s'il existe d'éventuels sites de liaison similaires entre deux protéines données en se basant sur un alignement structural local de surface. , afin de comparer les performances des deux outils pour identifier des patches de ligands polypharmacologiques sur leur *off-targets*.

7.1 Ligands étudiés et cibles multiples

En 2015, la "Multiple Target Ligand Database" (MTLD) [33] a mis à la disposition de la communauté scientifique 1732 ligands dont les structures de complexes avec leurs protéine-cibles sont répertoriées, ce qui aboutit à un total de 12 759 struc-

tures pdb. Cette base de structures est accessible à l'adresse `www.mtdcadd.com`. Les critères appliqués aux ligands et aux cibles protéiques sont :

— les structures obtenues sont toutes issues de diffractions aux rayons X et sont toutes des protéines.

— la résolution est inférieure à 3,0 Å, ce qui implique une qualité acceptable des structures étudiées ainsi que des informations exploitables sur les interactions entre les protéines et les ligands.

— les ligands comptent tous plus de 8 atomes lourds.

— les protéines ciblées par un même ligand partagent une identité de séquence inférieure à 35%.

Lors de la constitution de cette base de données, les auteurs ont principalement étudié la structure des ligands et celle des sites de liaison entre les ligands et leur cible. Leur but était de répertorier des structures de complexes protéine-ligand afin de pouvoir étudier la polypharmacologie des ligands de ces complexes. L'intérêt est préférentiellement porté sur la nature des ligands plutôt que sur celle des protéines ciblées. Par conséquent, les structures des cibles n'ont pas été filtrées ou analysées par rapport à leur similitudes structurales.

Cette base de données contient au total 222 médicaments approuvés, parmi lesquels nous en avons choisi 3, impliqués dans des traitements de cancer, dont les structures sont présentées dans la Figure 7.1 :

- l'imatinib, vendu sous le nom de Gleevec ©. Cette molécule est utilisée pour cibler différents types de cancers, notamment la leucémie myéloïde chronique, ou encore les tumeurs de l'estomac ou du pancréas.
- le sunitinib commercialisé sous le nom de Sutent ©. Ce médicament est utilisé dans le traitement du cancer du rein, et de l'adénome gastrique (dans le cas où la tumeur est résistante au traitement à l'imatinib).
- le sorafenib, commercialisé sous le nom de Nexavar ©, qui traite également le cancer du rein, du foie, ou de la thyroïde.

Une majorité des protéines ciblées par ces médicaments sont des protéines de type kinases (Tableau 7.1) : leur localisation cellulaire peut varier selon leur type

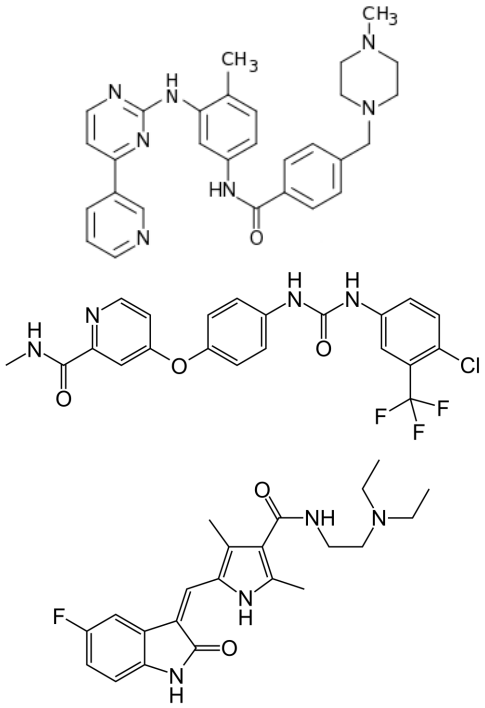


FIGURE 7.1: **Structures des ligands polypharmacologiques choisis dans la MTLD**. De haut en bas : imatinib, sorafenib et sunitinib.

(membrane, cytoplasme ou noyau), et selon les voies de signalisation dans lesquelles elles sont impliquées. Ces protéines font partie des cibles préférentielles lors de la conception de médicaments : elles interviennent dans des fonctions essentielles de la cellule et certaines de leurs mutations peuvent entraîner des pathologies tels que certains cancers [41].

Une cinquième cible était proposée pour l'imatinib : la ribosyldihydronicotinamide deshydrogénase (id PDB : 3fw1), mais nous avons choisi de ne pas étudier son patch car le ligand (en orange dans la Figure 7.2) n'interagit pas directement avec la protéine, mais par l'intermédiaire d'un groupement flavine-adénine dinucléotide (en bleu).

Ligands	**Id PDB**	**Fonctions des protéines ciblées**
Sunitnib	3g0e	facteur de croissance à activité tyrosine kinase
	2y7j	sérine/thréonine phosphorylase kinase
	3ti1	sérine/thréonine kinase dépendante des cyclines
	3miy	tyrosine kinase de lymphocyte T
Sorafenib	1uwh	sérine/thréonine kinase du proto-oncogène B-raf
	4asd	récepteur de VEGF à activité tyrosine kinase 2
	3rgf	kinase dépendante des cyclines 8
	3gcs	"Mitogen-Activated Protein" (MAP) kinase 4
Imatinib	1t46	tyrosine-kinase Kit
	1xbb	spleen tyrosine-kinase (syk)
	3hec	MAP kinase 14
	3k5v	tyrosine kinase du proto-oncogène ABL_1

Tableau 7.1: **Liste des complexes étudiés issus de la MTLD.**
Pour chaque complexe, sont présentés le nom du ligand ainsi que la protéine en interaction.

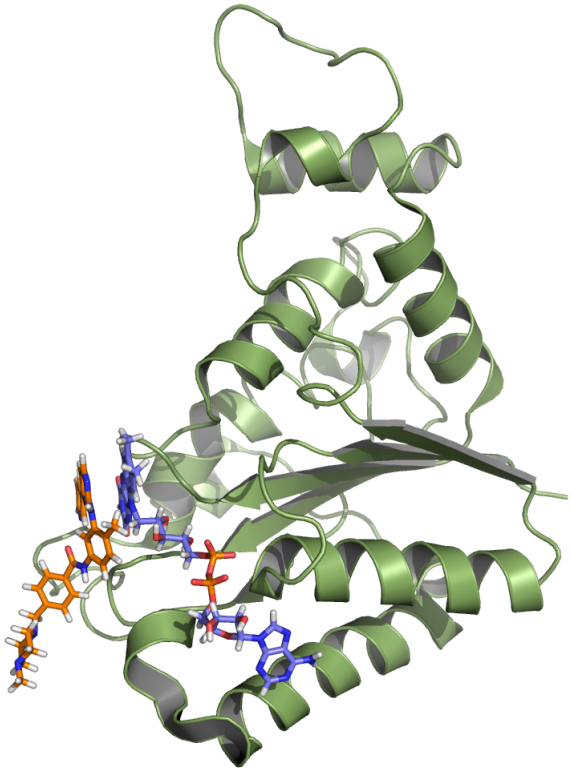


FIGURE 7.2: **Cible proposée dans la MTLD non utilisée.** La ribosyldihydronicotinamide deshydrogénase interagit avec l'imatinib par l'intermédiaire du FAD (Id PDB :3fw1).

7.2 Application de PatchSearch sur des patches extraits de complexes de la MTLD

7.2.1 Recherche des patches de ligands polypharmacologiques

Nous avons extrait les patches et les surfaces de chacun des complexes listés dans le Tableau 7.1, et pour chaque ligand, nous avons recherché chacun de ses patches sur les surfaces des trois autres protéines cibles proposées dans la MTLD. Nous avons réalisé les mêmes expériences à l'aide du serveur de ProBiS [89, 88] : cet outil représente les surfaces des protéines comparées sous la forme de graphes, dans lesquels les noeuds sont des appariements de groupements fonctionnels similaires entre les deux surfaces et les arêtes sont des relations de distances conservées entre les groupements appariés. Les graphes de chaque surface sont chacun divisés en sous-graphes, puis chaque sous-graphe d'une protéine est comparé aux autres sous-graphes de l'autre protéine, et un graphe produit est construit lorsque la matrice des différences de distances est compatible avec un seuil de similarité. La clique maximum de chaque sous-graphe est recherchée, et le regroupement des cliques maximum partageant au moins 5 noeuds permet d'obtenir un morceau de surface avec de fortes similitudes structurales entre les deux protéines comparées.

La matrice de scores présentée dans le Tableau 7.2 montre les différents scores obtenus par PatchSearch lorsque l'on recherche les 4 patches issus de chacun des 4 complexes de la MTLD connus pour contenir du sunitinib dans les 4 surfaces des *off-targets* de la MTLD connues pour lier ce ligand. Le score d'un patch identifié dans la surface dont il est extrait, par exemple du patch de 2y7j recherché dans la surface de 2y7j, vaut toujours 1 : la surface ciblée contient exactement les mêmes atomes que ceux du patch dans une disposition identique. On remarque que les scores obtenus lors de la recherche du patch de 2y7j dans la surface de 3ti1, soit 0,82, et celui de la recherche du patch de 3ti1 dans la surface de 2y7j, soit 0,67, diffèrent. Cette non-symétrie de la matrice s'explique par la formule du calcul du score de PatchSearch : en effet, ce score dépend de la taille (en terme de nombre d'atomes) du patch requête

et de celle du patch identifié. Les deux patches ont des tailles différentes : 63 atomes pour 2y7j et 69 atomes pour 3ti1. De plus, on n'apparie pas les mêmes atomes (et les mêmes nombres d'atomes) dans les deux comparaisons :

— pour la recherche du patch de 2y7j dans la surface de 3ti1 : on apparie 51 atomes.

— pour la recherche du patch de 3ti1 dans la surface de 2y7j : on apparie 48 atomes.

Finalement, la non-symétrie des scores obtenus par PatchSearch s'explique par les différences de compositions des patches requêtes et par les différences d'atomes appariés qui en découlent.

	2y7j patch	**3ti1** patch	**3miy** patch	**3g0e** patch
2y7j surface	1	0,67	0,80	0,59
3ti1 surface	0,82	1	0,77	0,61
3miy surface	0,84	0,71	1	0,71
3g0e surface	0,71	0,66	0,70	1

Tableau 7.2: **Matrice des scores obtenus lors de la recherche de patches du sunitinib dans les surfaces des *off-targets* connues pour ce ligand dans la MTLD.**
Le score de PatchSearch correspond au ratio du nombre d'atomes du patch identifié dans la surface ciblée par le nombre d'atomes du patch entré en requête. Ce score varie donc entre 0 et 1.

Les résultats des patches ainsi identifiés à l'aide PatchSearch et ProBiS sont présentés dans le Tableau 7.3. Les résultats de ProBiS sont les protéines ou "hits" qui présentent un z-score est supérieur ou égal à 1 : ce score est conseillé par défaut par les auteurs [89]. Les patches des ligands à cibles multiples sont spécifiques de ces

ligands, nous avons donc pu mesurer pour chaque patch trouvé, la Dc correspondante ou la distance entre les centroïdes du patch identifié dans la surface ciblée et celui du patch déjà connu dans la même surface.

7.2.2 Analyses des patches identifiés et comparaison des résultats de PatchSearch et ProBiS

Les patches de sorafenib sont reconnus avec des scores élevés pour les deux outils, et une Dc moyenne très faible pour PatchSearch : les 4 patches sont reconnus en moyenne à moins de 2 Å du patch connu sur la protéine cible.

Les performances pour identifier les patches d'imatinib sont similaires pour les patches de 1t46 et 1xbb : dans les deux cas, ProBiS et PatchSearch ne parviennent pas à identifier le patch d'imatinib sur la cible 1xbb. Par contre ProBiS reconnaît le patch d'imatinib de 3hec sur 1xbb.

Les différents patches d'imatinib sont recherchés sur 1xbb et identifiés avec des scores moyens faibles : lorsque le patch requête est 1t46, le score moyen est de 0,49, 0,46 avec le patch requête de 3k5v et 0,45 avec le patch de 3hec. Lors de ces trois expériences, les Dc moyennes mesurées sont importantes et témoignent que seule une partie du patch a correctement été identifiée : les Dc moyennes mesurées varient de 4,47 Å à 5,37 Å. Nous avons ainsi poussé l'analyse du cas particulier du patch de 1xbb, notamment à l'aide de l'examen visuel de l'imatinib dans la structure de 1xbb : le ligand se trouve dans une configuration différente par rapport à celle retrouvée dans les autres complexes (Figure 7.4) : l'imatinib est dans une cis-conformation, le patch a donc une forme moins étendue que dans les 3 autres complexes (Figure 7.5).

Le patch de 1xbb compte 62 atomes tandis que celui de 3hec en contient 88, celui de 3k5v 87 atomes et celui de 1t46 en a 96. Toutefois, lorsque l'on recherche le patch de 1xbb à partir de celui de 1t46 par exemple (Figure 7.4), on remarque qu'il y a un recouvrement partiel entre le patch identifié (en sphères bleues) et celui de 1xbb (en grillage rose). Pour ce cas précis, PatchSearch parvient tout de même à identifier

Ligand	**Patch requête**	**"Hits" de ProBiS**	**Z-score de ProBiS**	**Score moyen de PatchSearch**	**Dc (Å)**
Sorafenib	1uwh	3	3.21	0,60 ± 0,17	1.48
	4asd	3	2,92	0,72 ± 0,14	0,71
	3rgf	3	3,05	0,67 ± 0,08	1,15
	3gcs	3	3,01	0,57 ± 0,07	1,87
Imatinib	1t46	2	3,41	0,71 ± 0,19	2,4
	3k5v	2	3,42	0,65 ± 0,17	2,39
	1xbb	2	3,34	0,70 ± 0,06	6,22
	3hec	3	2,85	0,66 ± 0,19	1,98
Sunitinib	3g0e	0	-	0,64 ± 0,06	1,23
	2y7j	3	2,93	0,79 ± 0,07	1,23
	3ti1	0	-	0,68 ± 0,03	0,77
	3miy	1	2,92	0,76± 0,05	1,49

Tableau 7.3: **Comparaison des scores moyens obtenus avec PatchSearch et ProBiS pour identifier des patches de ligands de la MTLD sur des surfaces d'*off-targets* connues.**

Pour chaque ligand, chacun des 4 patches a été recherché dans les surfaces des 3 autres *off-targets* connues issues de la MTLD. Un "hit" de ProBiS correspond à une protéine présentant un patch de surface identifié comme similaire à celui de la protéine requête. Dans ce tableau, sont présentés les scores moyens de PatchSearch ainsi que les z-scores moyens des hits de ProBiS.

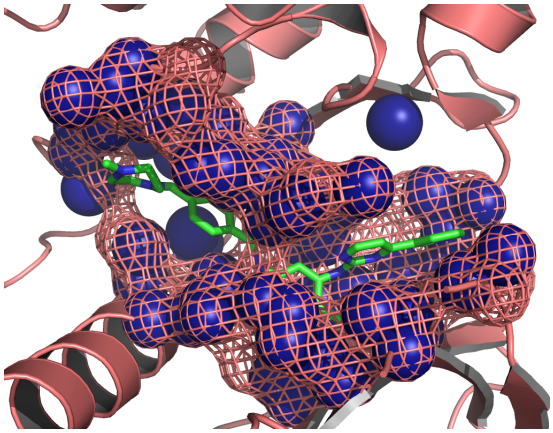


FIGURE 7.3: **Identification d'un patch d'imatinib sur 3hec.** Superposition entre le patch de 3hec et le patch d'imatinib identifié par PatchSearch (patch requête : 1t46).

une partie du patch de 1xbb. Néanmoins, PatchSearch reconnaît les autres patches d'imatinib avec une localisation correcte sur les surfaces des protéine cibles comme l'illustre la Figure 7.3 avec l'exemple du patch de 3hec (en grillage saumon) correctement identifié lorsqu'on utilise celui de 1t46 (le patch identifié est en sphères bleues).

Enfin, ProBiS identifie les patches de sunitinib lorsque les patches requêtes sont 2y7j (ProBiS renvoie les 3 cibles) et 3miy (ProBiS n'identifie ce patch que sur 2y7j). Lorsque les patches de sunitnib de 3g0e et de 3ti1 sont utilisés comme requêtes, ProBiS ne parvient pas à les identifier sur les protéines cibles. D'autre part, les patches de sunitnib sont tous retrouvés par PatchSearch avec des scores moyens élevés (tous supérieurs à 0,6) et des Dc moyennes inférieures à 2,0 Å. Les patches ainsi identifiés sont donc localisés de façon correcte par rapport aux patches de sunitnib déjà connus sur les cibles, comme l'illustre la Figure 7.6, la distance entre le centroïde du patch de 2y7j (le patch est représenté en grillage rouge) et celui du patch identifié (en sphères vertes) à l'aide du patch requête 3ti1 est de 0,9 Å.

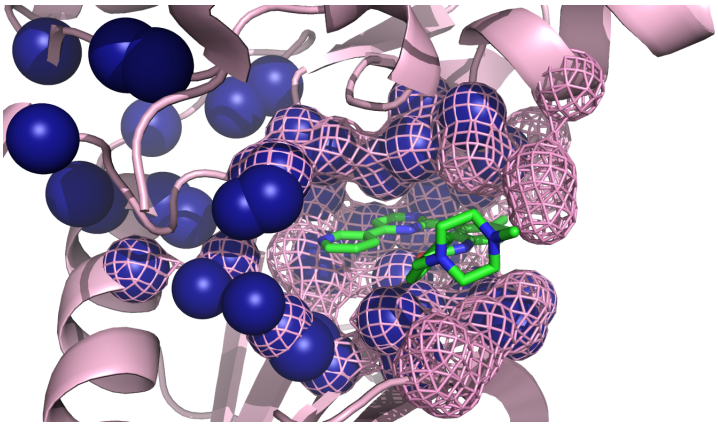


FIGURE 7.4: **Identification d'un patch d'imatinib sur 1xbb.** Superposition entre le patch de 1xbb en grillage rose et le patch d'imatinib identifié par PatchSearch en sphères bleues (patch requête : 1t46).

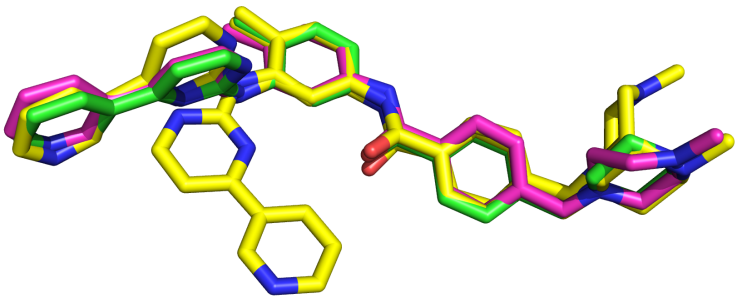


FIGURE 7.5: **Superposition des différentes conformations *cis* et *trans* de l'imatinib.** En jaune : l'imatinib en conformation *cis*. En magenta et vert : l'imatinib en conformation *trans*.

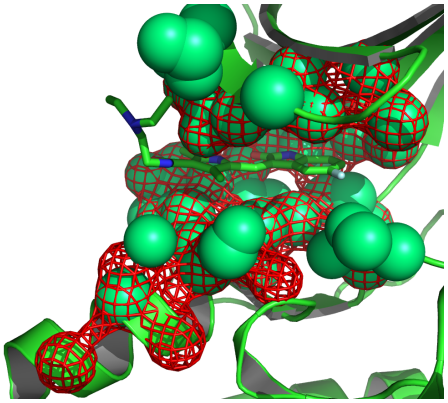


FIGURE 7.6: **Identification du patch de sunitinib de 3ti1 sur 2y7j.** Superposition entre le patch de 2y7j en grillage rouge et le patch de sunitnib identifié par PatchSearch en sphères vertes (patch requête : 3ti1).

7.3 Discussion et conclusion

Performances de PatchSearch sur les complexes de la MTLD

PatchSearch parvient à reconnaître les patches de ligands polypharmacologiques sur les surfaces des différentes cibles connues :

— avec des scores d'alignement élevés, en moyenne, tous les scores sont supérieurs à 0,6.

— les patches ainsi identifiés sur les surfaces des cibles sont correctement localisés par rapport aux patches connus.

Le cas du patch de l'imatinib dans le complexe 1xbb montre que bien qu'un groupement de l'imatinib soit orienté dans un angle différent de la conformation communément connue, une partie du patch peut être retrouvée et il existe un recouvrement non négligeable entre le patch ainsi identifié grâce aux autres patches requêtes (1t46, 3k5v et 3hec) et celui qui est connu sur 1xbb.

De plus, la méthode de PatchSearch donnait des résultats soit équivalents à ProBiS dans le cas de la recherche de patches de sorafenib, soit meilleurs que ceux de ProBiS, ce qui fut le cas pour les patches de sunitinib notamment. Dans le cas du sorefenib, les deux outils identifient les trois cibles de ce ligand à chaque expérience. Pour PatchSearch, les patches sont identifiés avec des scores moyens supérieurs à

0,5. Ils sont correctement localisés, les Dc moyennes mesurées sont inférieures à 2,0Å. Dans le cas du sunitinib, les cibles de ce médicament ne sont pas identifiées lorsqu'on utilise les patches requêtes 3g0e et 3ti1, tandis que PatchSearch parvient à les identifier avec des scores moyens supérieurs à 0,6 et à les localiser correctement avec des Dc moyennes inférieures à 2Å.

Analyse des patches de kinases

Les kinases jouent un rôle important dans le fonctionnement de la cellule et leur intervention a été démontrée dans le cadre de pathologies telles que les tumeurs [130, 114]. Les résidus essentiels impliqués dans leur mécanisme d'action ont été mis en évidence expérimentalement [186, 91] et sont :

— la paire de résidus lysine et glutamate (K et E) qui établissent un contact polaire caractéristique. La lysine permet de lier les groupements α et β phosphates de l'ATP.

— le motif aspartate - phénylalanine - glycine (D-F-G) qui forme une triade conservée parmi toutes les kinases. Le résidu aspartate joue un rôle essentiel dans la catalyse de la réaction.

— le motif histidine/tyrosine - arginine - aspartate (H/T-R-D) est également conservé au sein des kinases. L'histidine ou la tyrosine interagit avec l'aspartate et la phénylalanine du motif DFG.

Ces résidus jouent ainsi un rôle structural et fonctionnel dans l'action de la kinase sur son substrat.

Les médicaments présentés dans ce chapitre (sorefenib, imatinib et sunitinib) sont des inhibiteurs des kinases ciblées et interagissent au même endroit que le substrat. Nous avons donc vérifié la présence de ces résidus dans les patches extraits des différentes kinases étudiées à partir des 12 complexes protéine-ligand étudiés.

Nous avons identifié la paire K-E dans chacun des patches issus des complexes kinases-sorafenib (1uwh, 4asd, 3rgf et 3gcs) ainsi que dans les complexes kinases-imatinib (1t46, 3k5v, 1xbb et 3hec).

Analyse des patches des kinases de la MTLD

Les patches issus des complexes contenant le sunitinib (3g0e, 2y7j, 3ti1 et 3miy) ne présentent aucun des motifs précédemment présentés :

— le patch issu du complexe 3g0e [52] ne présente que l'aspartate 810 qui a le rôle dans la réaction de catalyse. Dans ce complexe, le sunitinib se lie à la conformation désactivée et autoinhibée de la kinase. L'intérêt de cette étude réside dans l'importance de cibler des kinases dans des conformations diversifiées.

— le complexe 2y7j n'a pas de publication associée.

— dans le complexe 3ti1 [118], l'interaction du sunitinib avec la kinase s'effectue à travers des interactions polaires (avec le squelette carboné du Glu81 et de la Leu83) et des interactions de van der Waals (avec la Phe80 et Ile10).

— dans le complexe 3miy [102], l'hélice C-terminale de la kinase est dans une conformation dite "in" qui correspond à l'état actif (au contraire de la conformation "out" qui est assimilée à l'état inactif).

Les patches issus des complexes contenant le sorafenib ne contiennent pas de résidus de la triade HRD. Seul le patch issu du complexe 1uwh contient les résidus du motif DGF et la paire K-E, comme montré dans la Figure 7.7.

Le patch du complexe 3hec, montré dans la Figure 7.8 est le seul patch parmi les 12 patches extraits de kinases qui présente tous les résidus précédemment cités comme essentiels pour l'interaction entre la kinase et l'ATP (la paire K-E, la triade DGF et la triade HRD).

Étude de la présence des résidus-clé dans les patches identifiés par PatchSearch

Nous avons donc recherché si PatchSearch parvenait systématiquement à identifier les résidus-clés présents dans les patches des complexes 1t46, 3k5v et 3hec. Les résultats de cette expérience sont présentés dans le Tableau 7.4.

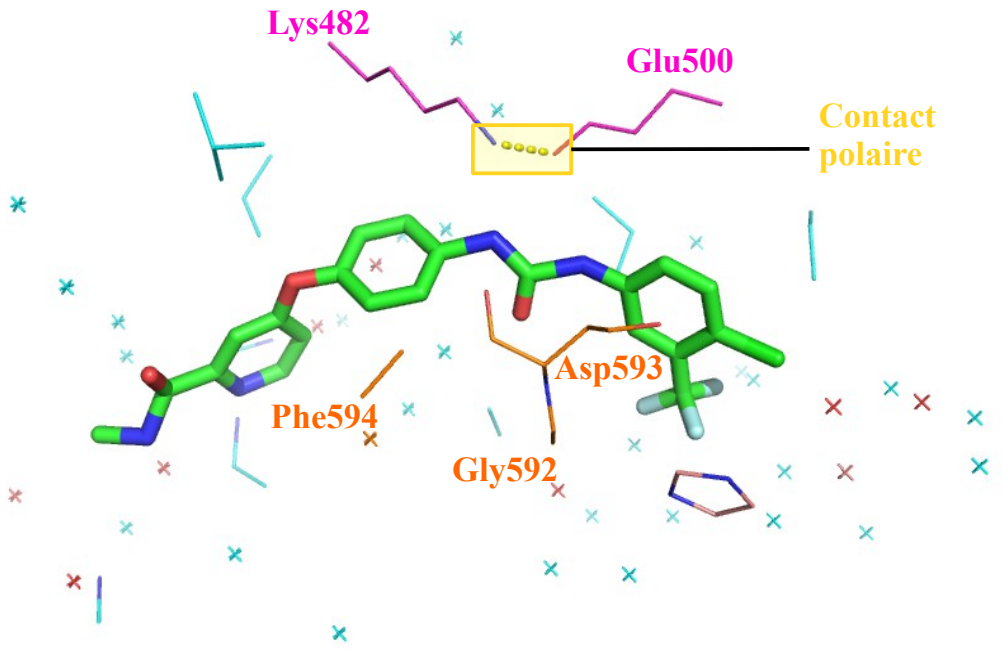


FIGURE 7.7: **Patch issu du complexe 1uwh.** Les résidus de la paire K-E sont colorés en magenta. Le contact polaire établi entre eux est coloré et entouré en jaune. Les résidus de la triade DGF sont colorés en orange. Les autres atomes du patch sont colorés en cyan. Le sorafenib est représenté en bâtons et est coloré en vert.

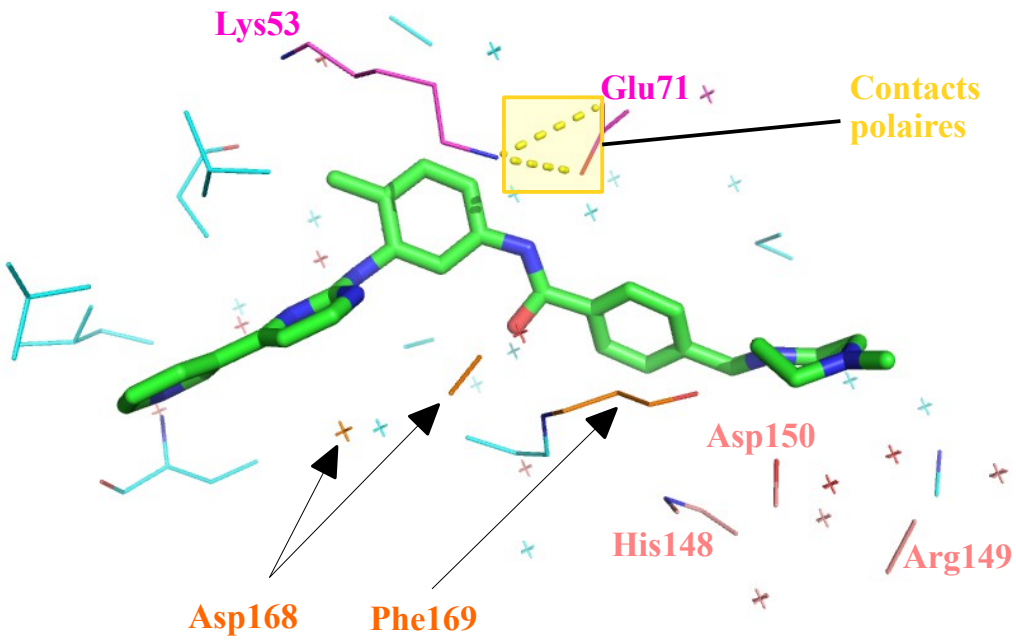


FIGURE 7.8: **Patch issu du complexe 3hec.** Les résidus de la paire K-E sont colorés en magenta. Le contact polaire établi entre eux est coloré et entouré en jaune. Les résidus de la triade DGF sont colorés en orange. Les résidus de la triade HRD sont colorés en rose saumon. Les autres atomes du patch sont colorés en cyan. L'imatinib est représenté en bâtons et est coloré en vert.

	1t46 patch	**3k5v** patch	**3hec** patch
1t46 surface	6/6	6/6	6/6
3k5v surface	6/6	6/6	6/6
3hec surface	6/7	7/7	7/7

Tableau 7.4: **Nombre de résidus-clés identifiés par PatchSearch lorsque l'on recherche des patches de l'imatinib sur les cibles connues.**
La ligne du haut présente les différents patches donnés en entrée de chaque expérience tandis que la première colonne contient les surfaces des cibles sur lesquelles chacun des patches a été recherché. Pour chaque recherche de patch, le nombre de résidus-clés identifiés et le nombre de résidus-clés connus pour le patch connu de la cible sont notés.

Ces résultats mettent en évidence que PatchSearch retrouve tous les résidus-clés dans 5 expériences (si on ne tient pas compte des recherches d'un patch contre sa propre surface) et dans l'expérience restante (la recherche du patch 1t46 sur la surface 3hec), la quasi-totalité des résidus-clés du patch de 3hec sont identifiés (soient 6 résidus sur les 7 connus).

Dans le cas des kinases se liant à l'imatinib, les patches ainsi identifiés par PatchSearch prennent en compte les résidus connus pour jouer un rôle structural et fonctionnel dans l'activité de ces kinases.

Discussion

V.J. Haupt et ses collaborateurs ont étudié l'influence de la flexibilité du ligand sur le nombre de cibles différentes que ce ligand peut avoir [66]. D'après leur étude, de tels cas sont minoritaires : sur les 164 médicaments qui avaient plus d'une cible connue, 18% ont 2 conformères et 14% en ont 3. Cette caractéristique d'un ligand donné à pouvoir interagir avec des cibles secondaires est sa promiscuité. Ils ont par la suite comparé les différentes corrélations entre certaines caractéristiques de l'interaction protéine-médicament et la capacité d'un médicament à interagir avec plusieurs cibles différentes. Plus le coefficient de corrélation (r) est proche de 1, plus les deux variables sont liées positivement. Toutefois, lorsque r est proche de 0, cela

signifie que les variables sont faiblement voire pas du tout liées. Leurs études ont mis en évidence :

— qu'il existe une forte corrélation (r=0,81) entre la similarité de patch et la promiscuité du ligand.

— que la similarité structurale entre les protéines cibles est aussi liée aux interactions multiples (r=0,76).

— et que la flexibilité du ligand est faiblement corrélée aux différentes interactions qu'il peut avoir (r=0,20).

Ces corrélations sont en accord avec notre démarche et nos conclusions obtenues au fil des expériences menées durant mon travail de thèse. Notre outil parvient à identifier des patches similaires propres à un même ligand avec des scores élevés et à une localisation correcte, ceci correspond à la promiscuité du ligand qui est fortement corrélée avec la similarité des patches de ce ligand. Ensuite, nos observations sur le patch de 1xbb indiquent que le ligand peut être flexible, il peut tout de même être reconnu par sa cible et le patch est identifié en partie sur les autres cibles connues.

Les médicaments dont nous avons étudié la détection de cibles dans ce chapitre sont impliqués dans des traitements contre certains types de cancers et ont tous entre 4 et 5 cibles différentes de structures connues. Ces chiffres sont cohérents avec ceux issus de l'étude de Hase T. et ses collègues en 2009 [65] : en moyenne, un médicament impliqué dans une maladie différente du cancer est connu pour avoir 4,24 cibles, tandis qu'un médicament impliqué dans un traitement anti-cancer compte en moyenne 7,82 cibles connues. Ainsi, l'identification d'un patch sur les différentes protéines susceptibles d'interagir avec le ligand d'intérêt permettrait de mieux comprendre les interactions secondaires d'un ligand donné, l'utilisation de PatchSearch s'inscrit tout à fait dans cette démarche.

Chapitre 8

Conclusions générales et Perspectives

Conclusions

L'analyse de la spécificité d'interaction d'un ligand pour une protéine donnée a nécessité la réflexion et le développement de différents outils présentés dans ce manuscrit :

— un programme permettant d'extraire un patch d'un complexe protéine-ligand donné et un programme d'extraction de surface protéique.

— un programme qui détecte des similitudes entre un patch et une surface, permettant ainsi de déterminer s'il existe un patch similaire au patch requête dans la surface.

Les outils ainsi développés peuvent ainsi être appliqués à de larges jeux de données.

Nous avons appliqué nos outils dans différents cas :

— la reconnaissance d'un patch sur la surface de la même protéine qui aurait subi des déformations structurales.

— l'identification de similarités entre des patches d'un même ligand.

— la détection localisée d'un patch issu d'un médicament sur les différentes protéines connues pour interagir avec ce médicament.

Les résultats des études décrites dans ce manuscrit soulignent l'importance de la visualisation des complexes protéine-ligand, avant et après extraction du patch.

L'emplacement du ligand conditionne fortement la manière dont le patch sera extrait par la suite : par exemple, un ligand qui interagit avec la protéine par l'intermédiaire d'une autre molécule ne permet pas d'extraire de patch spécifique à l'interaction précise du ligand avec la protéine. Un autre aspect nécessaire à prendre en compte concerne l'interaction d'un ligand entre deux chaînes d'une protéine.

PatchSearch correspond à l'implémentation d'une méthodologie originale de quasi-cliques, qui permet de détecter si une surface protéique présente des similitudes avec un patch donné ou pas, et par conséquent, si cette protéine pourrait être une cible pour le ligand dont est extrait le patch recherché. Les résultats de la comparaison entre l'approche plus classiquement utilisée des cliques, et celle des quasi-cliques, ont montré que PatchSearch identifie correctement un patch, ou du moins une partie significative. De toute évidence, la méthode des cliques est fortement sensible aux modifications structurales de la surface et permet préférentiellement de retrouver des patches fortement conservés. Les atomes exposés au solvant étant ceux dont la mobilité est la plus élevée par rapport au coeur de la protéine, il est essentiel de mettre en place un outil qui prenne en compte cette flexibilité.

La comparaison des résultats de PatchSearch à ceux des autres outils mettent en évidence

— que PatchSearch obtient des résultats similaires aux autres outils de comparaison de patches.

— l'importance de la mise en place de jeux de données contenant des complexes protéine-ligand dont la spécificité des interactions a été vérifiée au cours d'études cinétiques par exemple, et dont les ligands sont significativement différents entre eux (sans groupement commun par exemple).

— que PatchSearch identifie des patches de ligands polypharmacologiques non reconnus par ProBiS notamment.

PatchSearch s'avère finalement être un outil adapté à la recherche de patches propre à un ligand sur des protéines, dans le but d'identifier des patches similaires. L'utilisation d'un tel outil s'inscrit dans la démarche de recherche de cibles secondaires pour un ligand donné.

Perspectives

PatchSearch pourrait être appliqué à la reconnaissance d'un patch sur les différentes conformations d'une protéine obtenues au cours d'une dynamique moléculaire : ces résultats pourraient permettre de régler les paramètres de PatchSearch pour prendre en compte la flexibilité à la fois des chaînes latérales et des structures secondaires. Il aurait également été intéressant d'appliquer PatchSearch sur plus d'exemples de protéines pour lesquelles le patch entre les formes libres et liées subit un changement de grande amplitude.

Dans ce travail nous avons précisément étudié des interactions protéine-ligand pour mettre au point l'outil PatchSearch, mais il serait intéressant de l'appliquer dans la recherche des patches issus de complexes protéine-peptide ou protéine-acide nucléique.

Il n'est pas à négliger le rôle de certaines molécules d'eau se trouvant au coeur de l'interaction entre la protéine et son ligand. Les molécules d'eau peuvent être impliquées dans des réseaux d'interactions polaires entre la protéine et son partenaire. Dans le cas où le rôle de certaines molécules d'eau aurait été avéré, il serait intéressant de prendre en compte les atomes de ces molécules d'eau comme faisant partie du patch d'interaction.

Dans le cas où PatchSearch a détecté de potentielles cibles secondaires avec des scores élevés, il serait intéressant d'utiliser ces protéines lors de simulations de *docking*. De cette façon, nous pourrions vérifier que le patch identifié par PatchSearch correspondrait également à un patch auquel se lierait une pose du ligand dont l'énergie serait la plus favorable.

De la même façon qu'il est possible d'avoir des protéines assignées comme "vraies positives" lorsqu'un patch est retrouvé dans leur surface, il serait intéressant de tester PatchSearch sur des protéines dite "vraies négatives", c'est-à-dire des protéines qui sont connues pour n'avoir aucune interaction significative avec ligand donné. En effet, dans les jeux de données de référence, nous avons remarqué à plusieurs reprises que la notion de "faux positif" d'une protéine par rapport à son interaction avec un ligand varie selon les cas, notamment pour les protéines qui présentent des patches

pour plusieurs ligands différents. L'application de PatchSearch sur ces protéines permettrait de déterminer un score seuil en-dessous duquel le patch identifié n'est pas considéré comme significatif.

Enfin, comme nous l'avons soulevé dans la partie 7.3, il existe des protéines pour lesquelles l'implication de certains résidus-clés dans la fonction protéique ou dont le rôle structural essentiel dans la structure du site de liaison a été mis en évidence de façon expérimentale. Il serait intéressant d'appliquer la recherche de patches ne contenant que ces résidus-clés dans des surfaces de protéines pour lesquelles aucun ligand n'est connu. L'identification d'un patch similaire dans les surfaces de protéines dites "orphelines" pourrait ainsi donner des pistes pour l'étude *in silico* et *in vitro* de potentielles interactions entre le ligand du patch requête et la protéine que l'on cible.

Annexe A

Données issues du jeu de Gunasekaran

Classe I	154l (153l) ; 1a0t (1a0s) ; 1a8u (1a7u) ; 1afa (1afd) ; 1agw (1fgk) ; 1aha (1ahc) ; 1aqm (1aqh) ; 1arm (1yme) ; 1awb (2hhm) ; 1b5d (1b49) ; 1bk9 (1psj) ; 1bm7 (1bmz) ; 1br6 (1rtc) ; 1bso (2blg) ; 1bxq (3app) ; 1byq (1yer) ; 1com (2chs) ; 1dcp (1dco) ; 1did (1xla) ; 1dil (2sil) ; 1dmy (1dmx) ; 1duc (1dun) ; 1dud (1dup) ; 1eus (1eur) ; 1gmp (1gmq) ; 1hor (1dea) ; 1hvq (2hvm) ; 1icm (1ifb) ; 1kev (1ped) ; 1log (1loe) ; 1nft (1tfa) ; 1pnf (1png) ; 1pnl (1pnk) ; 1rca (1aqp) ; 1tal (2alp) ; 1vps (1vpn) ; 1xzb (1xza) ; 2enb (1ena) ; 4tim (1ag1) ; 5enl (3enl)
Classe II	1ai2 (3icd) ; 1fut (1fus) ; 1gd1 (2gd1) ; 1igb (1amp) ; 1qpq (1qpo) ; 1ra1 (5dfr) ; 1xva (1bhj) ; 1a26 (2paw) ; 1adl (1alb) ; 1aj0 (1ajz) ; 1alw (1alv) ; 1aq7 (2ptn) ; 1beh (1bd9) ; 1ben (1trz) ; 1ddt (1sgk) ; 1epb (1epa) ; 1fga (2fgf) ; 1fkl (1fkk) ; 1gca (1gcg) ; 1jdc (1jda) ; 1jef (135l) ; 1jul (1igs) ; 1kel (1kem) ; 1ltt (1lts) ; 1mjl (1mjk) ; 1mzm (1mzl) ; 1nhk (2nck) ; 1pgn (2pgd) ; 1ptr (1ptq) ; 1ses (1sry) ; 1ubw (1ubv) ; 2izf (2izd) ; 3pca (2pcd) ; 6cha (4cha) ; 2gal (1bkz)
Classe III	1a9p (1a9o) ; 1aj7 (2rcs) ; 1ake (4ake) ; 1anf (1omp) ; 1brp (1brq) ; 1byb (1bya) ; 1cen (1ceo) ; 1ebg (1ebh) ; 1hex (1xaa) ; 1hii (1hsi) ; 1hnl (1lz4) ; 1lca (4tms) ; 1mpj (3ins) ; 1o0r (1fgx) ; 1swd (1swa) ; 1vrt (1rtj) ; 2bgu (2bgt) ; 2cht (2chs) ; 3gal (1bkz) ; 4fua (1fua) ; 5cna (1enq) ; 7tim (1ypi)

Tableau A.1: **Liste des structures PDB composant le jeu de données de Gunasekaran.** Les structures ont été regroupées en trois classes, selon les variations structurales des C-α des sites de liaison entre les formes *holo* et *apo*. Les identifiants des complexes sont suivis des identifiants des structures *apo* entre parenthèses.

Classe I	1a0t ; 1a8u ; 1afa ; 1agw ; 1aha ; 1aqm ; 1awb ; 1b5d ; 1com ; 1dcp ; 1did ; 1dil ; 1dmy ; 1dud ; 1eus ; 1gca ; 1hor ; 1hvq ; 1jul ; 1kel ; 1log ; 1pnl ; 1ptr ; 1tal ; 1vps ; 2izf ; 3gal ; 3pca ; 4tim ; 5enl ; 6cha ; 154l
Classe II	1a9p ; 1a26 ; 1adl ; 1ai2 ; 1aj0 ; 1aj7 ; 1alw ; 1aq7 ; 1ben ; 1bk9 ; 1bm7 ; 1br6 ; 1brp ; 1bso ; 1bxq ; 1cen ; 1ddt ; 1duc ; 1epb ; 1fga ; 1fkl ; 1fut ; 1gd1 ; 1gmp ; 1hex ; 1icm ; 1igb ; 1jdc ; 1jef ; 1kev ; 1lca ; 1ltt ; 1mjl ; 1mpj ; 1mzm ; 1nft ; 1nhk ; 1pgn ; 1pnf ; 1qpq ; 1ra1 ; 1rca ; 1ses ; 1swd ; 1ubw ; 1xva ; 1xzb ; 2bgu ; 2cht ; 2enb ; 2gal ; 4fua ; 5cna
Classe III	1ake ; 1anf ; 1arm ; 1byb ; 1byq ; 1ebg ; 1hii ; 1hnl ; 1o0r ; 1vrt ; 7tim

Tableau A.2: **Nouvelles classes des structures du jeu de données de Gunasekaran.** Liste des structures du jeu de données mis en place par Gunasekaran, classées selon les RMSD entre tous les atomes des patches des formes *holo* et *apo*.

Annexe B

Reconnaissance de patches sur des protéines différentes

Ligands	AMP	ATP	FAD	FMN	Glucose
Id PDB	12as, 1amu, 1c0A, 1ct9, 1jp4, 1kht, 1qb8, 1tb7, 8gpb.	1a0i, 1a49, 1ayl, 1b8a, 1dv2, 1dy3, 1e2q, 1e8x, 1esq, 1gn8, 1kvk, 1o9t, 1rdq, 1tid.	1cqx, 1e8g, 1evi, 1h69, 1hsk, 1jqi, 1jr8, 1k87, 1pox, 3grs.	1dnl, 1f5v, 1ja1, 1mvl, 1p4c, 1p4m.	1bdg, 1cq1, 1k1w, 1nf5, 2gbp.

Ligands	Hème	NAD	PO_4	Stéroïdes
Id PDB	1d0c, 1d7c, 1dk0, 1eqg, 1ew0, 1gwe, 1iqc, 1naz, 1np4, 1po5, 1pp9, 1qhu, 2bs2, 1qpa, 1sox, 2cpo.	1ej2, 1hex, 1ib0, 1jq5, 1mew, 1mi3, 1o04, 1og3, 1qax, 1rlz, 1s7g, 1t2d, 1tox, 2a5f, 2npx.	1a6q, 1b8o, 1brw, 1cqj, 1d1q, 1dak, 1e9g, 1ejd, 1euc, 1ew2, 1fbt, 1gyp, 1h6l, 1ho5, 1l5w, 1l7m, 1lby, 1lyv, 1qf5, 1tco.	1e3r, 1fds, 1j99, 1lhu, 1qkt.

Tableau B.1: **Liste des complexes du benchmark de Kahraman.** Liste des 100 complexes protéine-ligand dont sont extraits les 100 patches du KD. AMP : Adénosine Monophosphate. ATP : Adénosine Triphosphate. FAD : Flavine Adénine Dinucléotide. FMN : Flavine Mononucléotide. NAD : Nicotinamide Adénine Dinucléotide. PO_4 : Phosphate.

Ligands	PMP	SUC	LLP	LDA	BOG
Id PDB	1a0g, 1aia, 1fg7, 1kta, 1mdo, 1uu1, 1zc9, 2c81, 2cjg, 2e7u.	1l0g, 1m98, 1pt2, 1tj4, 1uc2, 1w2t, 1ylj, 1jgi, 1jj0, 1a0t.	1a8i, 1ax4, 1bjw, 1bw0, 1cl1, 1cs1, 1d7k, 1iug, 1j04, 1jg8.	1aij, 1ar1, 1c8u, 1dxr, 1f7s, 1kmo, 1ojd, 1thq, 1umx, 1xkw.	1aua, 1b4w, 1fx8, 1i78, 1k8q, 2czv, 2hd0, 2p4b, 2z73, 3b6h.
Ligands	SAM	GSH	U5P	1PE	PLM
Id PDB	1cmc, 1eiz, 1hmy, 1i9g, 1msk, 1nt2, 1nw3, 1p91, 1qzz, 1r30.	1dug, 1eem, 1fw1, 1iyh, 1jlv, 1r4w, 1y1a, 2fls, 2imd, 2pbj.	1dbt, 1fgx, 1g8o, 1i5e, 1wlj, 2b56, 2bln, 2bmu, 2c37, 2j4j.	1g8i, 1o57, 1q0r, 1s7g, 1y10, 1zx8, 2byn, 2haw, 2idb.	1b56, 1eh5, 1m66, 1mgp, 1o6u, 1pq2, 1sz7, 2fik, 2iu8, 2nwl.

Tableau B.2: **Liste des complexes du benchmark homogène.** Liste des 99 complexes protéine-ligand dont sont extraits les 99 patches du HD. PMP : pyridoxine-5'-phosphate. SUC : sucrose. LLP : acide hexanoïque 2-amino-6-[[3-hydroxy-2-méthyl-5-(phosphonooxyméthyl)pyridine-4-yl]méthylideneamino]. LDA : lauryl diméthylamine-n-oxyde. BOG : béta octyl glucoside. SAM : S-adénosylméthionine. GSH : glutathion. U5P : uridine-5'-monophosphate. 1PE : pentaéthylène glycol. PLM : acide palmitique.

Annexe C

Article en révision

PatchSearch: a fast computational method for off-target detection

Inès Rasolohery, Gautier Moroy, and Frédéric Guyon*

Université Paris Diderot, INSERM UMRS 973, Molécules Thérapeutiques in Silico, Paris, France

E-mail: frederic.guyon@univ-paris-diderot.fr

Abstract

Many therapeutical molecules are known to bind several proteins, which can be different of the initially targeted one. Such unexpected interactions with proteins called off-targets can lead to adverse effects. Potential off-target identification is important to predict and avoid drug side effects and it can also be used to discover new targets for existing drugs. We propose a new program named PatchSearch, which implements a local non-sequential searching for similar binding sites on protein surfaces with a controlled amount of flexibility. It is based on detection of quasi-clique in product graphs representing all the possible matchings between the two compared structures. This method has been assessed on three protein-ligand benchmarks and on three molecules used to treat some cancer diseases and known to lead to adverse effects. The experiments conducted in this study show that the PatchSearch method could be useful in the early identification of drug adverse effects or could contribute to the repurposing of existing drugs. The program and the benchmarks presented in this paper are available as an R package at `https://github.com/MTiPatchSearch`.

1 Introduction

During drug design process, numerous promising molecules are not translated into effective drugs. Poor pharmacokinetics, lack of efficacy and toxic or adverse effects are the cause of around 90% of failures (*1*). Those side effects are often due to the interaction of the molecule with non-targeted partners, referred to as off-targets, such as enzymes, receptors or ion channels.

This ability of a molecule to interact with several proteins is also called polypharmacology. Based on drug-protein interaction data from annotated chemical libraries (*2*), (*3*), it has been shown that over 50% of drugs interact with more than five proteins and only 15% of them target only one single protein. Roughly, one drug is able to interact with on average six proteins (*4*).

The identification of off-targets in early-stage of drug design is therefore important to decrease the failure rate observed in clinical trials and to speed-up the drug discovery process. Moreover, the knowledge of the potential critical off-targets would lead to rationalize the drug design by chemical modification of compound in order to increase its specificity for the target protein. The identification of ligands able to bind different proteins has also an application in drug repositioning or repurposing. This approach consists in using existing drugs for new medical applications. Since the approved drugs have well-known pharmacokinetic properties and toxicity profiles, drug repositioning enables to drastically shorten drug development timelines and significantly lower the risks of failure, thus considerably reducing costs.

Based on the large amount of protein structures currently available in the Protein Data Bank, many computational methods have been proposed to detect structural similarities among proteins at different scale levels.

These methods essentially differ from the global structural alignment of proteins as they do not consider sequential order of the amino acids. This implies the use of different and more complex algorithms.

Approaches can be classified based on how they represent the structures and surfaces and

on the type of algorithm used to score and identify similarities. In a general way, they are divided into two main classes: alignment free and alignment based methods.

Alignment free methods do not try to find matching of atoms or residues and perform an overall comparison of global properties and characteristics such as shape, surface descriptors and physicochemical residue properties. These characteristics are atom coordinates combined to atom types compared with a convolution kernel (Sup-CK (*5*)), spherical harmonics ((*6*),(*7*)), three dimensional Zernike descriptors (Patch-Surfer (*8*, *9*)), internal distance distributions (PocketMatch (*10*)).

On the other hand, alignment based methods compute alignments between atoms or residues of compared binding sites, SOIPPA (*11*),eMatchSite (*12*), (*13*).

These alignments are different of the classical sequence alignment because binding site atoms are not always in the same sequence order in the two compared structures. Sequence order-independent alignments of residues or atoms are in general far more difficult to compute than alignment free comparisons. However, these methods allow identification of atom or residue sets involved in the binding with the ligand. In particular, geometric hashing can provide matching between atoms or features (*14*). In the geometric hashing, query and target surfaces are represented by atom triplet coordinates and a hash table permits to efficiently retrieve all similar triplets. The largest set of consistent similar triplets gives the best structural alignment. The following methods are based on geometric hashing: TESS (*15*), SitesBase (*16*), SiteEngine (and I2I-SiteEngine)(*17*), MultiBind (*18*), (*19*), TIPSA (*20*),SuMo (*13*) uses an heuristic to compare graph of surface chemical features,PCalign (*21*). Another classical approach to compute matching or atom superimposition is based on graph theory. By searching for clique in product graphs, it allows detecting all similar spatial arrangements of atoms in two different structures (*22*). The Bron-Kerbosh is the most efficient algorithm to this purpose. The first methods developing this strategy have been proposed by (*23*) (*24*),(*25*), (*24*) and more recently, clique algorithms are used in CavBase (*26*), eF-site (*27*). Nodes of a product or correspondance graph are pairs of atoms

one from the binding site and the second one from the targeted protein surface. Edges associate such pairs of atoms whenever the distances between the atoms in the binding site and in the protein surface match within a given tolerance. Methods based on the graph theory first compute the product graph as previously described, and then search for cliques in this graph. A clique is a subgraph in which all nodes are connected to all nodes. Therefore, it represents a set of atoms, which are in same relative positions in both surfaces.

Yet, this clique strategy, though accurate, presents some drawbacks: comparisons of proteins counting a few thousand atoms lead to some very large product graphs, and hence to large amount of running time. Indeed, algorithms for computing cliques such as the Bron-Kerbosch algorithm, run in exponential time for general graphs.

Moreover, a product graph clique implies that all atom inner distances are similar between the compared protein surfaces and does not take into account flexibility of binding sites. This strict definition of a clique makes it sensitive to small structural changes induced by protein dynamics or due to inaccuracy in protein structures solved by experimental methods. Furthermore, cliques can be very numerous and overlapping, and further computations are required to select the best ones.

Most of the above approaches compare binding sites by performing a global alignment restricted to binding sites. But few of them implement a local non sequence ordered alignment providing a localisation of a similar binding site on the target surface. ProBis(*28*) searches for similar surface patches, by non sequential local structural alignment of entire protein surface against a structure database using a clique detection technique.

To address these issues, we propose a program named PatchSearch based on a quasi-clique detection. A quasi-clique is a dense subgraph of the correspondence graph. In PatchSearch, a quasi-clique is constructed from a small core clique, which is enriched with new nodes highly connected to it. By searching for quasi-cliques, distance similarity constraints between the binding sites are relaxed, consequently, it is possible to retrieve similar binding sites with

some different inner distances.

Based on this quasi-clique searching, the PatchSearch algorithm is able to compare two binding sites at the atomic level or to perform a local non-sequential alignment of a binding site with a whole protein surface. This method aims to predict whether a ligand could interact with off-target proteins, which may lead to some hypothesis regarding its specificity for its target.

First, we assess this method on two benchmarks frequently used for binding sites comparison. Second, we test PatchSearch on a dataset of proteins in different conformational states. Finally, we study three approved drugs able to bind to multiple proteins.

2 Methods

2.1 Surface and Patch extraction

Protein structures are first processed using PROPKA (*29*, *30*) in order to rebuild the missing atoms and to optimize their positions. Then, solvent accessibility was calculated using NACCESS (*31*). We defined an atom as being solvent accessible if its relative accessibility is over 1%. The solvent accessible atoms and C_α belonging to a residue that has at least one of its atoms exposed to the solvent are kept for the structural similarities calculations. The atoms that are not solvent accessible are removed. Moreover, the aromatic ring atoms are replaced by their centroid (phenylalanine or tyrosine: 1 centroid, tryptophan: 2 centroids). In this study, a patch is constituted by the solvent accessible atoms, which have a distance smaller than 5 Åto the ligand.

2.2 Correspondence graphs

Patches and protein surfaces are represented by sets of atom coordinates combined with their physicochemical properties. Each atom of the patch and the surface is assigned a type corresponding to a physicochemical property or to a specific atom: N, O, S, C, Ca (for C_α).

Moreover aromatic rings are replaced by one or two pseudo-atoms typed A.

In PatchSearch, a correspondence graph (or product graph) is constructed as follows: nodes are pairs of atoms, a patch atom and a target atom, with identical type. We call such a pair a correspondence. A correspondence graph node represents a possible matching between atoms of the two compared structures. Let us denote nodes (i, i') representing a correspondence between patch atom i and a target atom i' where i and i' belong to the same atom type. An edge connects nodes (i, i') and (j, j') if

$$|\text{dist}(i, j) - \text{dist}(i', j')| \leq \Delta d$$

where Δd is a given distance tolerance criterion and $dist(i, j)$ is the euclidean distance between atoms i and j. An edge represents two patch atoms and two surface atoms, which are separated by the same distance up to a distance difference given by Δd.

A clique is a subgraph with all nodes connected to each other. Therefore, a clique represents a set of consistent correspondences between the two compared structures and gives subset of atoms in the same relative position in both structures.

We add the constraint that no two edges share a common vertex, equivalently an edge cannot connect two correspondences sharing a common atom, for instance (i, i') and (i, j') or (i, i') and (j, i'). In PatchSearch, this condition is fulfilled by choosing a threshold Δd, which is less than the minimum distance between patch atoms or target atoms of the same type.

$$\Delta d \leq \min_{i,j \text{ such that } type(i)=type(j)} dist(i, j)$$

With this threshold, for two pairs (i, i') and (i, j') we have

$$|\text{dist}(i, j) - \text{dist}(i', j')| = \text{dist}(i', j') > \Delta d$$

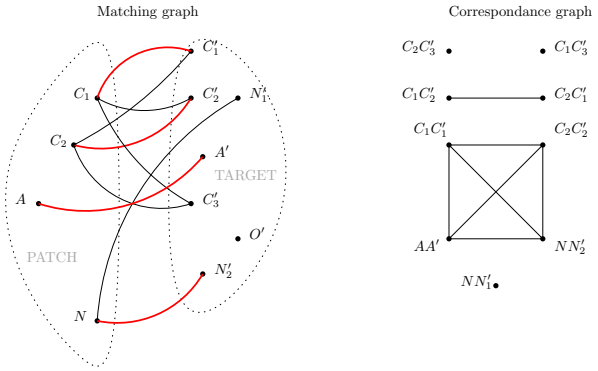


Figure 1: Construction of a correspondence graph. Correspondence or matching between atoms C_1C_1' is linked to correspondence C_2C_2' as distance between C_1 and C_2 is equivalent to distance between C_1' and C_2'. However, no edge exists between correspondences C_1C_1' and C_1C_2' as they share a common atom C_1

and such pairs cannot be linked by an edge.

With the no common atom condition, a clique provides a bijective mapping between patch and target surface and represents a possible structural alignment

2.3 Clique searching

Cliques may be efficiently retrieved using a variant of the Bron-Kerbosch algorithm (*32*). This algorithm has a worse-case running time of $0(dn3^{d/3})$ where n is the number of vertices in the correspondence graph and d is the degeneracy of the graph. Hence the running time required to retrieve all the maximal cliques depends on n, which is the number of pairs of atoms and on d, which depends on the number of pairs of atom in the same relative positions. Given two proteins structures with n_1 and n_2 atoms, the number of vertices can reach $n1 \times n2$. In proteins, physical constraints and structures of amino-acids imply that numerous inter-atomic distances are similar in both protein structures. Actually, atom types “C” and “Ca” may show regular spatial arrangements. For this reason, aromatic rings have been reduced to one or two pseudo-centers. These equivalent distances generate a great number of edges in the product graph with a high number of non significative small cliques. The number of spurious cliques can be controlled by selecting the types of atoms

involved into the construction of the graph, and by the distance deviation tolerance Δd. A stringent distance tolerance condition implies by definition less edges, whereas a higher deviation tolerance permits to retrieve more edges, consequently identifying more flexible spatial similarities. First, to reduce the size of the correspondence graph and the number of spurious cliques, a graph is constructed with a small distance tolerance criterion and a chosen restricted subset of atoms (S, N, A and O atoms). Maximal cliques corresponding to a sufficient number of atoms and with a positive determinant criterion are selected. Clique gives a subset of matching atoms in patch and surface with approximatively conserved inter-atomic distances. This does not imply the same spatial configuration. One of the two structural elements can be the mirror image of the other. Indeed, a structure and its mirror-symmetric image have identical inter-atomic distances. Such a mirror symmetry can be detected by a simple matrix determinant calculation. We have previously presented a new structural score based on this matrix determinant named Binet-Cauchy score (*33*). This score is used to remove cliques due to mirror images of the query and to assess the quality of the structural similarity given by the obtained cliques.

In more detail, we detect false matches due to chirality with the following criteria:

$$\text{mirror}(X, Y) \ = \text{sign}\det(X^\top Y) \tag{1}$$

This criteria is negative for mirror image conformations. More precisely, the mirror conformation of a structure can be exactly superimposed to $-X$. More generally, it can easily be proved that

$$\text{mirror}(X, Y) \leq 0 \Longleftrightarrow \text{rmsd}(-X, Y) \leq \text{rmsd}(X, Y) \tag{2}$$

When $\det(X^\top Y)$ is negative then the second structure Y is closer to the mirror conformation of X than to X itself.

2.4 Quasi-clique construction

PatchSearch proceeds in two steps: first, a graph is constructed with a small distance tolerance criterion (less than 1 Å) and with the chosen restricted subset of atoms. All maximum cliques are then found with the Bron-Kerbosh algorithm. Based on a structural similarity score (*33*), the best clique is selected. This clique gives the rigid core of the patch.

This clique gives an initial matching, which is then improved by considering links formed by all type of atoms neighboring the atoms given by the initial clique. The clique is enriched with selected links, which increases the structural similarity score of the query patch and the surface. More precisely, a link is selected and added to the quasi-clique if it is connected to at least four nodes in the clique. Indeed, four distances are necessary to guaranty a consistent matching between patch and surface.

Additionally, the selected link must preserve the bijectivity of the mapping. In case it shares an atom with one or two links of the clique, it induces a many-to-one or one-to-many mapping. This property of the algorithm permits to avoid another time-consuming step based for example on the Hungarian algorithm to identify a one-to-one matching.In this case, it is included into the quasi-clique if its insertion and the removal of the initial link implies a lower maximal deviation.

$$\text{MaxDev} = \max_{i \sim i', j \sim j'} |\text{dist}(i,j) - \text{dist}(i',j')| \tag{3}$$

where $i \sim i'$ means that atom i in patch matches atom i' in target surface.

The entire process is repeated iteratively until no further improvement in the maximal deviation can be obtained. The algorithm stops whenever no link fulfilling the two above conditions can be found.

This heuristic constructs quasi-cliques defining 3D similarities with more flexibility. The initial cliques represent preserved rigid cores between structures while the extended quasi-cliques yield matchings with a higher tolerance and represent the moving parts of the protein

surface.

At the end of the algorithm, depending on the use of the program, two different scores are computed: the first one assesses the similarity between two patches and is equivalent to the Tanimoto index:

$$s = \frac{N_{match}}{N_1 + N_2 - N_{match}}$$

where N_1 and N_2 denote the number of atoms in two patches and N_{match} the number of matched atoms The second one quantifies the local similarity of a patch and target surface:

$$s = \frac{N_{match}}{N_1}$$

Another criteria is used to assess the localization of the recognized patch on a protein surface with the true patch when it is known. The Dc criteria is the euclidean distance between the centroids of the known patch and the retrieved patch.

$$Dc(p_1, p_2) = \|c(p_1) - c(p_2)\|$$

where p_1 is the patch bound to the studied ligand, p_2 is the patch bound to the same ligand on an off-target protein and $c(p_1)$ and $c(p_2)$ are the centroids of the atoms of the two patches.

2.5 Datasets

Kahraman Dataset Kahraman and coworkers have built a dataset to study the shape variations in protein binding pockets (*34*). This dataset is composed by 100 protein binding sites; each one interacts with one of nine ligand types: adenosine monophosphate (AMP), adenosine-triphosphate (ATP), flavin-adenine dinucleotide (FAD), flavin mononucleotide (FMN), D-glucose (GLC), Heme, nicotinamide-adenine-dinucleotide (NAD), PO4 and Steroids.

A shape matching method, developed by the authors, was applied on the dataset to

compare the shapes of protein binding pockets to the shapes of their ligands. Using a shape descriptor, the authors are able to classify correctly the binding pockets interacting with rigid ligands. However, the variety of conformations observed in flexible ligands complicates the classification based on the shape alone. Therefore, their analyses suggest that the molecular recognition cannot be only explained by shape complementarities. Additional physicochemical properties, such as electrostatic potential or hydrophobicity, seem to be required to obtain a suitable classification of binding pockets. Nowadays, this dataset, referred as Kahraman dataset, is frequently used as a benchmark to assess the performance of several methods devoted to the comparison and the recognition of protein binding sites (*5*, *8*, *20*, *35*).

Homogeneous dataset We considered a second dataset, called Homogeneous Dataset (HD), made up by Hoffmann and coworkers (*5*), in which the sizes of the binding sites are closer than those in the Kahraman dataset. This benchmark is composed by 100 pockets extracted from non-redundant proteins, binding to ligands of similar sizes. This dataset is divided into ten groups of ten pockets binding one of ten ligand types:4âĂŹ-deoxy-4âĂŹ-aminopyridoxal-5âĂŹ-phosphate, sucrose, 2-lysine(3-hydroxy-2-methyl-5-phosphonooxymethyl-pyridin-4-ylmethane), lauryl dimethylamine-N-oxide, b-octylglucoside, palmitic acid, S-adenosylmethionine uridine-5âĂŹ-monophosphate, glutathione and pentaethylene glycol.

Gunasekaran dataset It is widely established that conformational changes in the binding sites can occur upon ligand binding. It means that various ligands can induce motions of flexible parts in a binding site. In order to prove the ability of PatchSearch to take into account the flexibility of binding sites, we tried to recognize the same binding site on the same protein in different conformations. For this purpose, we applied PatchSearch to the Gunasekaran dataset, which comprised 97 proteins in both holo (in complex with ligand) and apo (unbound) forms (*36*). Based on the structural changes between holo and apo forms, this benchmark was divided in three classes: class I (no conformational change: Cα displacement $\leq$ 0.5 Å), class II (moderate conformational changes, 0.5 Å$\leq$ Cα displacement

$\leq$ 2.0 Å) and class III (large conformational changes: Cα displacement $\geq$ 2.0 Å).

MTLD: multiple target ligand database This database presented in (*37*) comprises different structures of proteins, which have been solved in interaction with a same ligand. On the whole, MTLD is composed of 1,732 polypharmacological ligands, among them 222 are approved drugs, able to bind to 44,996 sites. Interestingly, the MTLD statistical analysis shows that 795 ligands can bind with two different proteins, 740 ligands interact with from three to ten proteins and 197 ligands bind to over ten proteins. We have chosen three approved drugs used in cancer treatments referenced by MTLD: Sutent, Imatinib and Sorafenib. PatchSearch is employed as a tool to predict and study their polypharmacological effects.

3 Results and discussion

3.1 Binding site similarity recognition

Although we aim to identify off-target proteins, we first assess the PatchSearch scoring when used as a patch similarity score. The Kahraman dataset (*34*) has been used for the assessment of several methods devoted to the identification of similar binding sites (*38*), this is the reason why we choose this dataset to compare our results to those obtained by some of these methods: spherical harmonics (SH) (*34*), convolution kernel (Sup-CK) (*5*), geometric hashing (MultiBind) (*18*), molecular surface shape alignment (PSIM) (*39*) and 3D Zernike descriptor (PatchSurfer) (*8*, *9*).

The assessment of each methods was based on a ROC analysis and the AUC (area under curve) criteria. For each ligand, we have computed all their score of similarity with all pockets and rank these scores in decreasing order. The fraction of positive hits above a given rank and the fraction of negative hits below this rank, represents a point of the ROC curve.

In a first step, we have evaluated PatchSearch performances using clique and quasi-clique approaches. The summary of the results are given in Table 1. For each approaches, we performed all-against-all patches comparisons. Whereas AUC value of the pure clique approach, 0.52, is slightly above the random ranking, the quasi-clique approach gives significantly good results with corresponding AUC value of 0.78. The relatively poor results of pure clique approach is likely due to a more stringent constraint about interatomic distances in the structures of the compared binding sites. The structural distortions caused by experimental conditions, conformational changes upon the ligand binding or by differences in the binding mode of flexible ligands explain that quasi-clique is the most relevant approach for PatchSearch. Nonetheless, it was pointed out that, in the Kahraman dataset, the binding sites have various sizes depending on each ligand. Therefore, the volume or the size of the pocket has been shown to enhance the performances of the different methods (Table 1). The PatchSearch program yields the same observations with a pocket size descriptor: AUC values are increased from 0.52 to 0.60 and from 0.78 to 0.82 for clique and quasi-clique approaches respectively. The best results were obtained by sup-CK, with AUC values of 0.86 and 0.89 with a volume descriptor. Whereas the PatchSearch assessment is realized with constant parameters, the sup-CK score has been optimized independently for each ligand.

To avoid the problem of the size effect of the binding pocket in the Kahraman dataset, the Homogeneous dataset was built to contain binding sites with similar pocket volumes. On this dataset, quasi-clique approach also performs slightly better than the pure clique one, with AUC values of 0.77 and 0.74 respectively. As expected, the performances of all methods are slightly weaker than in Kahraman dataset. However, except for sup-CK, the performances are not significantly deteriorated between both datasets. Whatever the method used, the pocket recognition is nearly identical when the size or the volume of the binding pocket is combined to the initial score.

3.1.1 Time assessment

The overall running time necessary to compute all-against-all patch similarity is 2083 seconds on a PC GNU/Linux Intel Xeon CPU at 2.40GHz with one core used. The average running time of a single patch comparison is 0.37 second. If a standard clique approach is used under the same conditions, the overall time is 11500 seconds with an average time of 2.07 seconds. The average time ratio between the two approaches is over 9. These durations measure the whole processing time of the method including parsing of the PDB files, the clique detection based on a reduced set of atoms and the quasi-clique construction. The clique approach performs the clique detection given all the atoms of the patch and the target surfaces whereas in the PatchSearch algorithm, the clique detection is performed on a subset of selected atoms.

The running time assessment to perform the search of a patch onto a target surface gives very close results. This evaluation is performed on the Gunasekaran dataset. The average time is almost proportional to the number of atoms of the target surface (not shown) and the average time ratio between the clique and quasi-clique approaches is equal to 8.16.

Table 1: AUC for different methods applied to Kahraman dataset and Homogenous dataset. Values are taken from (a) Hoffman et al. (*5*), (b) Spitzer et al. **??** and (c) Sael et al. (*8*, *9*). NA: not available.

Methods	Kahraman dataset		Homogeneous dataset	
	Shape	Shape + size	Shape	Shape + size
PatchSearch: Clique	0.52	0.60	0.67	0.74
PatchSearch: Quasi-Clique	0.78	0.82	0.74	0.77
Spherical Harmonics (a)	0.64	0.77	NA	NA
sup-CK (a)	0.86	0.89	0.71	0.72
MultiBind (a)	0.71	NA	0.69	NA
PSIM (b)	0.79	NA	0.76	NA
PatchSurfer (c)	0.81	0.84	NA	NA

Another reliable method, called eMatchSite has been developed to carry out alignments of ligand binding sites (*12*). It can recognize efficiently similar binding pockets in proteins

with different global structures. It is based on machine learning techniques that predict pocket alignments with protein-ligand complex templates and then estimate the probability of binding a same ligand from many sequential, structural, geometry and physicochemical properties of known binding sites.

This method yields AUC values of 0.69 and 0.92 for Kahraman dataset and Homogeneous dataset respectively. However, these results are not directly comparable to ours, since the author has selected subsets of the both datasets.

Even if Kahraman dataset and Homogeneous dataset are often employed to compare methods for the assessment of the similarities between binding sites they are not really appropriate for the drug design field. In these dataset, the ligands do not show pharmacological relevance: they are biological cofactors or bind non-specifically to proteins. Furthermore, in Kahraman dataset, some ligands are closely related, for instance AMP and ATP share a same moiety, such as the adenine moiety, which is contained in AMP, ATP and NAD. Overall, on both datasets, the quasi-clique approach of PatchSearch provides better accurately than those obtained with pure clique approach and exhibits equivalent performances to the other methods specifically designed for binding sites comparisons. However, it is important to note that PatchSearch is developed for the recognition of similar binding sites onto protein surfaces.

3.1.2 Quasi-clique and clique approaches comparison on the Gunasekaran dataset

As previously mentioned, two (or more) patches interacting with a same ligand do not have exactly conserved inner distances, depending on the protein environment or state.

Consequently, we extracted patches from holo protein structures (proteins bound to a ligand), and we tested patch recognition on the apo state of the same protein (free protein).

We first assessed the patch recognition efficiency on this dataset. Patches of the class I are recognized with the highest specificity and sensitivity, the AUC value is 1 using cliques or quasi-cliques. The results are slightly better using quasi-cliques rather than cliques for class

II (AUC value of 0.98 with cliques and 0.99 with quasi-cliques) and class III (respectively 0.95 and 0.96).

We next evaluate the ability of PatchSearch to localize the patches extracted from proteins in holo form onto the corresponding apo form surface using clique and quasi-clique approaches. The patch recognition is measured by the PatchSearch score and by the fraction of correctly retrieved residues.

In this experiment, for all classes, we obtain a mean fraction of correctly retrieved residues close to 1 indicating that besides some very difficult cases, the PatchSearch score gives the fraction of correctly retrieved atoms and therefore that PatchSearch succeeds in precisely localizing the expected binding site onto the apo form. As shown in the Figure 2, clique score values tend to drastically decrease when holo and apo forms have gone through larger conformational changes (class II and class III proteins). We notice that using cliques does not permit to align the whole patches, especially for class III patches (maximal score using cliques: 0.98 for 3GAL/1BKZ). Scores obtained on class I using cliques or quasi-cliques are similar (clique median score: 0.94 , quasi-clique median score: 0.96) indicating that both methods managed to precisely retrieve the corresponding patches onto the apo surface (Figure 2). An example of identification of holo patches from the class I is shown in the figure 3: 1agw and 1fgk are both crystal structures of the tyrosine kinase domain of fibroblast growth factor receptor 1. 1agw is the form bound to SU4984 inhibitor, whereas 1fgk is the apo form. The RMSD between binding site of both forms is 0.32 Å. Both holo and apo forms are very close, which explains the high overlap between the patch found on the apo form 1fgk using the 1agw patch as a query.

In class II, median score is 0.82 and 0.93 for cliques and quasi-cliques respectively. In class III, the proteins have undergone larger conformational changes and cliques are significatively less efficient than in class I and II: median score for cliques is 0.63, whereas median score for quasi-cliques is 0.87. For example, 1hii and 1hsi are both crystal structures of the HIV-2 protease and belong to the class III proteins (figure 4). 1hii is the form bound to CGP 53820

pseudosymmetric inhibitor, whereas 1hsi is the apo form. The RMSD between binding site of both forms is 2.92 Å, nevertheless, PatchSearch is able to accurately identify the 1hii patch on 1hsi with a high score (0.7) and the proportion of correct retrieved residues is 0.93. For all the proteins of the dataset, the quasi-clique approach allows to identify larger parts of the patch, especially the flexible ones. The difficult cases correspond to protein movements implying distances between patches over 3 Å. The Δd threshold parameter is set to 3 Å, and PatchSearch can only retrieve subpart of the patch on the corresponding apo form. In all cases, this part of the patch has been subject to a lesser deformation and is precisely localized on the apo surface. For example, the score value for the recognition of the 1ake patch on 4ake surface is quite low (0.39). The RMSD between patch in the holo and apo forms is high (6.1 Å), corresponding to a movement of a loop. For this particular case, we choose to increase the Δd parameter to 6 Å, which allows us to recognize the patch with a better score (0.69) and the rate of correct identified residues is acceptable (0.59) given the high distortions between holo and apo forms.

3.2 Recognition of patches binding some polypharmacological ligands

PatchSearch was applied in relevant pharmacological cases, *i.e.* approved drugs known to bind to multiple proteins. We choose three drugs designed for cancer therapy from the MTLD dataset (*37*): imatinib, sutent and sorafenib. Imatinib has been approved by the Food and Drug Administration for the treatment of chronic myelogenous leukemias and gastrointestinal stromal tumors, sutent for renal cell carcinomas, gastrointestinal stromal tumors and pancreatic neuroendocrine tumors and sorafenib for renal cell carcinomas, hepatocellular carcinomas and thyroid cancers. For each drugs, four structures of drug-protein complexes are available in the PDB (Table 2). We attempted to retrieve the drug binding site for a given protein in the 3 other protein surfaces with PatchSearch. The same experiment was performed with ProBis and compared with PatchSearch results. For sorafenib, both Patch-

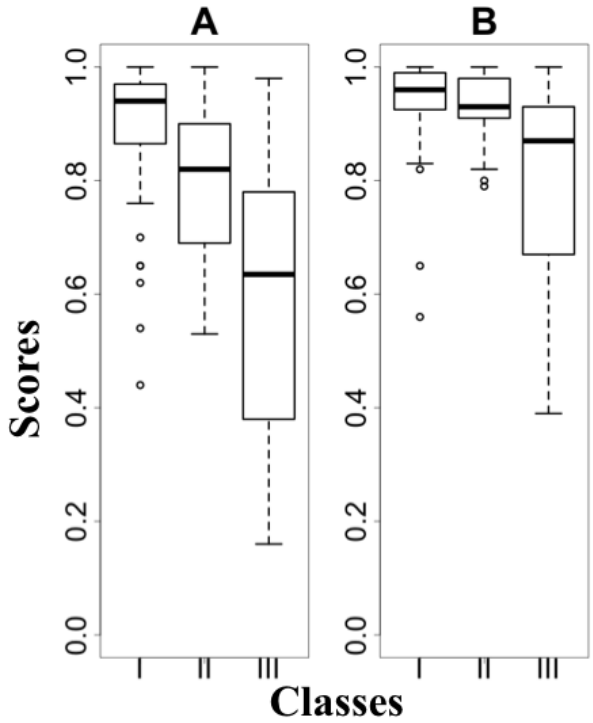


Figure 2: **Comparison of clique (A) and quasi-clique (B) scores for each class of the proteins from the Gunasekaran dataset.**

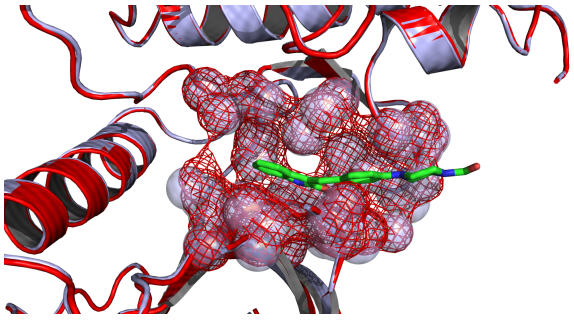


Figure 3: **Example of holo patch found on apo form on class I proteins.** 1agw patch atoms are represented in red mesh and atoms of the patch found on 1fgk are represented in blue spheres.

Search and ProBis succeed in recognizing the three others structures. For PatchSearch, the binding sites are found accurately on the surfaces ($Dc \leq 2.0$ Å), with an average score of 0.64. Most of the imatinib patches are correctly found by both tools, except for 1xbb. When analyzing imatinib patches in 1xbb, 1t46, 3hec and 3k5v structures, we noticed that the spleen tyrosine kinase in 1xbb adopts an activated kinase conformation (which has not been

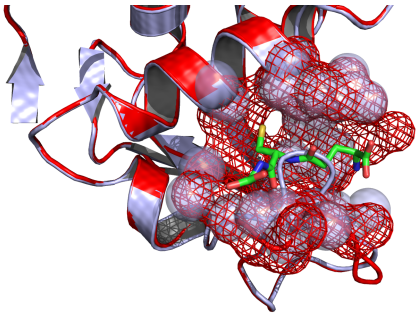


Figure 4: **Example of holo patch found on apo form on class III proteins.** 1hii is colored in blue whereas 1hsi is colored in red. Both proteins are represented as cartoon. 1hii patch atoms are represented in red mesh and atoms of the patch found on 1hsi are represented in blue spheres.

seen in the other structures). Moreover, in 1xbb, the ligand is in its cis-conformation, which is more folded and compact: its interactions with the kinase are more limited. Those two different conformations could explain that imatinib patches are less efficiently recognized on 1xbb surface. ProBiS does not find any patch on the three off-targets for the sutent patches 3g0e and 3ti1, whereas PatchSearch identifies them with scores higher than 0.6 (respectively 0.64 and 0.68), and low Dc (1.23 Åand 0.77 Å). The figure 6) shows the superposition of the 2y7j patch and the patch found with PatchSearch on 2y7j using 3ti1 patch as a query: sutent patch is found on the human phosphorylase kinase with a Dc of 0.9 Åand the score is 0.67. Another example of superposition of found sutent patch and the initial known patch is shown in the Figure 7 : sutent patch is found with a higher score (0.71) and a very low Dc, 0.77 Å. For example, when searching for the 3ti1 patch (Figure 5) on the other surfaces (2y7j, 3miy and 3g0e), we identified them with good PatchSearch scores (respectively 0.67, 0.71 and 0.66) and some very low Dc (0.9 Å, 0.77 Åand 0.64 Å). Patches found on 2y7j (Figure 6) and on 3miy (Figure 7) overlap their own sutent patches, thus showing a good localization of the patches found by PatchSearch. For most of these cases (except for 1xbb), PatchSearch succeeds in detecting similar binding sites in different proteins and retrieve the correct localizations onto the whole protein surfaces. ProBiS is able to obtain equivalent results for sorafenib and imatinib, but, on the contrary of PatchSearch, ProBiS is unable to

detect correctly all the similar binding sites, which interact with sutent.

Table 2: **Comparison of PatchSearch and ProBiS performances on three approved drugs used in cancer therapy.**

Ligand	Query patch	ProBiS hits	ProBiS z-score	PatchSearch score	Dc (Å)
Sorafenib	1uwh	3	3.21	0.60	1.48
	4asd	3	2.92	0.72	0.71
	3rgf	3	3.05	0.67	1.15
	3gcs	3	3.01	0.57	1.87
Imatinib	1t46	2	3.41	0.71	2.4
	3k5v	2	3.42	0.65	2.39
	1xbb	2	3.34	0.7	6.22
	3hec	3	2.85	0.66	1.98
Sutent	3g0e	0	no hit	0.64	1.23
	2y7j	3	2.93	0.79	1.23
	3ti1	0	no hit	0.68	0.77
	3miy	1	2.92	0.76	1.49

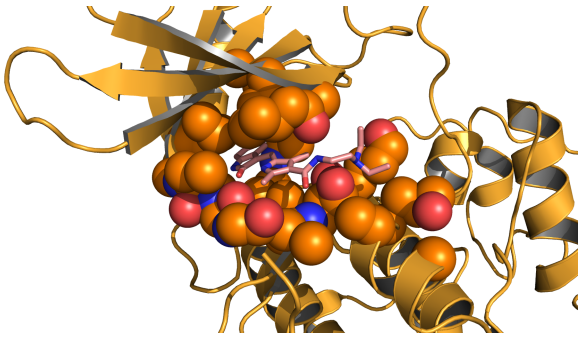


Figure 5: **Patch extract from the complex of the cyclin-dependent kinase 2 bound with B49 (PDB id: 3ti1).** Patch atoms are represented with orange spheres, the ligand (B49) with sticks and the protein with a cartoon structure.

4 Conclusions

PatchSearch is a tool designed to identify off-targets protein in a structural databank whenever the structure of a complex of a drug bound to its target is known. For this purpose,

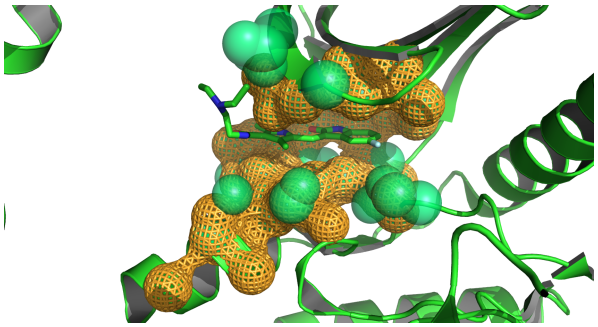


Figure 6: **Example of sutent patch found on 2y7j.** Patch atoms of 2y7j are represented as green spheres, whereas atoms found using PatchSearch with 3ti1 query patch on 2y7j are in orange mesh.

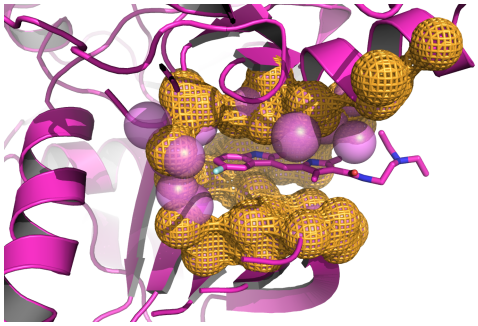


Figure 7: **Example of sutent patch found on 3miy.** Patch atoms of 3miy are represented as magenta spheres, whereas atoms found using PatchSearch with 3ti1 query patch on 3miy are in orange mesh.

it is based on an efficient and original quasi-clique detection approach to recognize specific patches. It not only computes a similarity score between two binding sites, but also it is able do accurately align patch locally on a whole off-target surface.

Moreover, even if the patch is derived from an experimental binding site in the results presented here, the patch can be also manually specified by the user through the selection of relevant residues, for instance protein-protein interaction hotspots.

We showed that PatchSearch can recognize patch with a controlled amount of flexibility. This has been assessed on a benchmark consisting of holo and corresponding apo protein conformations. In conclusion, this novel tool could be useful for detecting unwanted interactions of drugs with non-targeted proteins, in order to prevent some potential side effects. Furthermore, by facilitating the uncovering of unexpected off-targets, PatchSearch could

contribute to the repurposing of existing drugs.

5 Competing interests

The authors declare that they have no competing interests.

6 Author's contributions

Algorithms was developped by (FG). (IR), (GM) and (FG) designed the experiments, analysed the results and wrote the manuscript. All authors read and approved the final manuscript.

7 Acknowledgements

The authors would like to thank the doctoral school Bio Sorbonne Paris Cité and Université Paris Diderot for the recurring funds.

References

1. van de Waterbeemd, H., and Gifford, E. (2003) ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug. Discov. 2*, 192–204.

2. Wishart, D., Knox, C., Guo, A., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res. 36*, D901–906.

3. Bento, A., Gaulton, A., Hersey, A., Bellis, L., Chambers, J., Davies, M., Krüger, F., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., and Overington, J. (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res. 42*, D1083–1090.

4. Mestres, J., Gregori-Puigjané, E., Valverde, S., and Solé, R. (2009) The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst. 5*, 1051–1057.

5. Hoffmann, B., Zaslavskiy, M., J.Ph.Vert„ and Stoven, V. (2010) A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics 11*, 99.

6. Ritchie, D. W., and G. J. L. Kemp, G. J. L. (1999) Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J. Comp. Chem. 20*, 383–395.

7. Morris, R., Najmanovich, R., Kahraman, A., and Thornton, J. (2005) Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics 21*, 2347–2355.

8. Sael, L., and Kihara, D. (2010) Binding Ligand Prediction for Proteins Using Partial Matching of Local Surface Patches. *Int. J. Mol. Sci. 11*, 5009–5026.

9. Sael, L., and Kihara, D. (2012) Detecting Local Ligand-Binding Site Similarity in Non-Homologous Proteins by Surface Patch Comparison. *Proteins 80*, 1177–1195.

10. Yeturu, K., and Chandra, N. (2008) PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinformatics 9*, 543.

11. Xie, L., and Bourne, P. E. (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments. *Proc. Natl. Acad. Sci. USA 105*, 5441–5446.

12. Brylinski, M. (2014) eMatchSite: Sequence Order-Independent Structure Alignments of Ligand Binding Pockets in Protein Models. *PLoS Comput. Biol. 10*, e1003829.

13. Jambon, M., Imberty, A., Deleage, G., and Geourjon, C. (2003) A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins 52*, 137–145.

14. Nussinov, R., and HJ, H. W. (1991) Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl. Acad. Sci. 88*, 10495–10499.

15. Wallace, A. C., Borkakoti, N., and Thornton, J. M. (1997) TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Science 6*, 2308–2323.

16. Gold, N. D., and Jackson, R. M. (2006) Fold Independent Structural Comparisons of Protein-Ligand Binding Sites for Exploring Functional Relationships. *J. Mol. Biol. 355*, 1112–1124.

17. Shulman-Peleg, A., Nussinov, R., and Wolfson, H. (2004) Recognition of Functional Sites in Protein Structures. *J. Mol. Biol. 339*, 607–633.

18. Shatsky, M., Shulman-Peleg, A., Nussinov, R., and Wolfson, H. (2006) The Multiple Common Point Set Problem and its Application to Molecule Binding Pattern Detection. *J. Comp. Biol. 13*, 407–428.

19. Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H. (2008) MultiBind and MAPPIS: Web servers for multiple alignment of protein 3D binding sites and their interactions. *Nucleic Acids Research 36*, W260–264.

20. Ellingson, L., and Zhang, J. (2012) Protein Surface Matching by Combining Local and Global Geometric Information. *PLOS One 7*, e40540.

21. Cheng, S., Zhang, Y., and Brooks, C. (2015) PCalign: a method to quantify physicochemical similarity of protein-protein interfaces. *BMC Bioinformatics 16*, 33.

22. Ullmann, J. (1976) An Algorithm for Subgraph Isomorphism. *J. Assoc. Comput. Mach. 23*, 31–42.

23. Bron, C., and Kerbosch, J. (1973) Algorithm 457: Finding All Cliques of an Undirected Graph. *Comm. ACM 26*, 48–50.

24. Gardiner, E., Artymiuk, P., and Willett., P. (1997) Clique-detection algorithms for matching three-dimensional molecular structures. *Journal of Molecular Graphics and Modelling 15*, 245–253.

25. Grindley, H., Artymiuk, P., Rice, D., and Willett, P. (1993) Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol. 229*, 707–721.

26. Schmitt, S., Kuhn, D., and Klebe, G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol. 323*, 387–406.

27. Kinoshita, K., Furui, J., and Nakamura, H. (2002) Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics 2*, 9–22.

28. Konc, J., and Janezic, D. (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics 26*, 1160–1168.

29. Søndergaard, C. R., Olsson, M. H., Rostkowski, M., and Jensen, J. H. (2011) Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. *Journal of Chemical Theory and Computation 7*, 2284–2295.

30. Olsson, M. H., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. (2011) PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions. *Journal of Chemical Theory and Computation 7*, 525–537.

31. Hubbard, S., and Thornton, J. *'NACCESS' computer program*; 1993.

32. Eppstein, D., Löffler, M., and Strash, D. Listing All Maximal Cliques in Sparse Graphs in Near-optimal Time. Proc. 21st International Symposium on Algorithms and Computation. 2010; pp 403–414.

33. Guyon, F., and Tufféry, P. (2014) Fast protein fragment similarity scoring using a Binet-Cauchy kernel. *Bioinformatics 30*, 784–791.

34. Kahraman, A., Morris, R., Laskowski, R., and Thornton, J. (2007) Shape variation in protein binding pockets and their ligands. *J. Mol. Biol. 368*, 283–301.

35. Bertolazzi, P., Guerra, C., and Liuzzi, G. (2010) A global optimization algorithm for protein surface alignment. *BMC Bioinformatics 11*, 488.

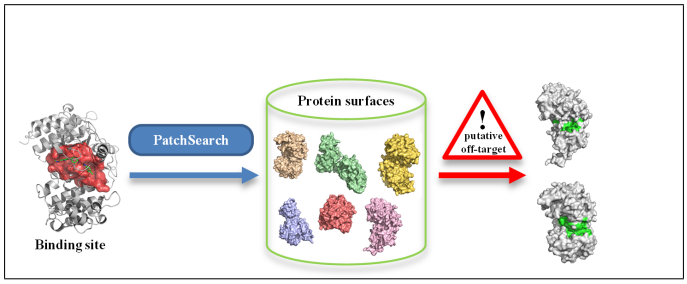
36. Gunasekaran, K., and Nussinov, R. (2007) How Different are Structurally Flexible and Rigid Binding Sites? Sequence and Structural Features Discriminating Proteins that Do and Do not Undergo Conformational Change upon Ligand Binding. *J. Mol. Biol. 365*, 257–273.

37. Chen, C., He, Y., Wu, J., and Zhou, J. (2015) Creation of a free, Internet-accessible database: the Multiple Target Ligand Database. *J. Cheminform. 7*, 14.

38. Jalencas, X., and Mestres, J. (2013) Identification of Similar Binding Sites to Detect Distant Polypharmacology. *Mol. Inf. 32*, 976–990.

39. Spitzer, R., Cleves, A., and Jain, A. (2011) Surface-Based Protein Binding Pocket Similarity. *Proteins: Structure, Function, and Bioinformatics 79*, 2746–2763.

Graphical TOC Entry



Bibliographie

[1] Christopher D Aakre, Julien Herrou, Tuyen N Phung, Barrett S Perchuk, Sean Crosson et Michael T Laub. « Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates ». In : *Cell* 163.3 (2015), p. 594–606.

[2] James Abello, Mauricio GC Resende et Sandra Sudarsky. « Massive quasi-clique detection ». In : *Latin American Symposium on Theoretical Informatics*. 2002, p. 598–612.

[3] Vilmos Ágoston, Péter Csermely et Sándor Pongor. « Multiple weak hits confuse complex systems : a transcriptional regulatory network as an example ». In : *Physical Review E* 71.5 (2005), p. 051909.

[4] J Ahokas et O Pelkonen. « Pharmacokinetics : How Does The Body Handle Drugs ». In : *Pharmacology, Encyclopedia of Life Support Systems* (2007).

[5] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers et David J Lipman. « Basic local alignment search tool ». In : *Journal of molecular biology* 215.3 (1990), p. 403–410.

[6] O Ameisen et R De Beaurepaire. « Suppression de la dépendance à l'alcool et de la consommation d'alcool par le baclofène à haute dose : un essai en ouvert ». In : *Annales Médico-psychologiques, revue psychiatrique*. T. 168. 2. Elsevier. 2010, p. 159–162.

[7] Jianghong An, Maxim Totrov et Ruben Abagyan. « Pocketome via comprehensive identification and classification of ligand binding envelopes ». In : *Molecular & Cellular Proteomics* 4.6 (2005), p. 752–761.

[8] Stefano Angaran, Mary Ellen Bock, Claudio Garutti et Concettina Guerra. « MolLoc : a web tool for the local structural alignment of molecular surfaces ». In : *Nucleic acids research* (2009), p. 1–6.

[9] Rolf Apweiler, Amos Bairoch, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane et al. « UniProt : the universal protein knowledgebase ». In : *Nucleic acids research* 32.suppl 1 (2004), p. D115–D119.

[10] Mauricio Arenas-Salinas, Samuel Ortega-Salazar, Fernando Gonzales-Nilo, Ehmke Pohl, David S Holmes et Raquel Quatrini. « AFAL : a web service for profiling amino acids surrounding ligands in proteins ». In : *Journal of computer-aided molecular design* 28.11 (2014), p. 1069–1076.

[11] Ted T Ashburn et Karl B Thor. « Drug repositioning : identifying and developing new uses for existing drugs ». In : *Nat Rev Drug Discov* 3.8 (2004), p. 673–683. issn : 1474-1776.

[12] Amos Bairoch. « The ENZYME database in 2000 ». In : *Nucleic acids research* 28.1 (2000), p. 304–305.

[13] Thierry Bardinet. *Les papyrus médicaux de l'Egypte pharaonique*. 2002.

[14] Vladimir Batagelj et Matjaz Zaversnik. « An O (m) algorithm for cores decomposition of networks ». In : *arXiv preprint cs/0310049* (2003).

[15] BD4Cancer. *Les phases des essais cliniques*.

[16] Andreas Bender, Josef Scheiber, Meir Glick, John W Davies, Kamal Azzaoui, Jacques Hamon, Laszlo Urban, Steven Whitebread et Jeremy L Jenkins. « Analysis of pharmacology data and the prediction of adverse

drug reactions and off-target effects from chemical structure ». In : *ChemMedChem* 2.6 (2007), p. 861–873.

[17] Frances C Bernstein, Thomas F Koetzle, Graheme JB Williams, Edgar F Meyer, Michael D Brice, John R Rodgers, Olga Kennard, Takehiko Shimanouchi et Mitsuo Tasumi. « The protein data bank ». In : *European Journal of Biochemistry* 80.2 (1977), p. 319–324.

[18] Paola Bertolazzi, Concettina Guerra et Giampaolo Liuzzi. « A global optimization algorithm for protein surface alignment ». In : *BMC bioinformatics* 11.488 (2010).

[19] T Andrew Binkowski, Shapor Naghibzadeh et Jie Liang. « CASTp : computed atlas of surface topography of proteins ». In : *Nucleic acids research* 31.13 (2003), p. 3352–3355.

[20] Mitradev Boolell, MJ Allen, SA Ballard, S Gepi-Attee, GJ Muirhead, AM Naylor, IH Osterloh et C Gingell. « Sildenafil : an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. » In : *International journal of impotence research* 8.2 (1996), p. 47–52.

[21] Aislyn DW Boran et Ravi Iyengar. « Systems approaches to polypharmacology and drug discovery ». In : *Current opinion in drug discovery & development* 13.3 (2010), p. 297–309.

[22] K Boutet, Irène Frachon, Yannick Jobic, Christophe Gut-Gobert, Christophe Leroyer, Dominique Carlhant-Kowalski, Olivier Sitbon, Gérald Simonneau et Marc Humbert. « Fenfluramine-like cardiovascular side-effects of benfluorex ». In : *European Respiratory Journal* 33.3 (2009), p. 684–688.

[23] G Patrick Brady Jr et Pieter FW Stouten. « Fast prediction and visualization of protein binding pockets with PASS ». In : *Journal of computer-aided molecular design* 14.4 (2000), p. 383–401.

[24] Coen Bron et Joep Kerbosch. « Algorithm 457 : finding all cliques of an undirected graph ». In : *Communications of the ACM* 16.9 (1973), p. 575–577.

[25] Michal Brylinski. « eMatchSite : Sequence Order-Independent Structure Alignments of Ligand Binding Pockets in Protein Models ». In : *PLoS Comput Biol* 10.9 (2014), e1003829.

[26] Michal Brylinski et Jeffrey Skolnick. « A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation ». In : *Proceedings of the National Academy of sciences* 105.1 (2008), p. 129–134.

[27] Ana M Calvo, Richard A Wilson, Jin Woo Bok et Nancy P Keller. « Relationship between secondary metabolism and fungal development ». In : *Microbiology and Molecular Biology Reviews* 66.3 (2002), p. 447–459.

[28] Marco Cammisa, Antonella Correra, Giuseppina Andreotti et Maria Vittoria Cubellis. « Identification and analysis of conserved pockets on protein surfaces ». In : *BMC bioinformatics* 14.7 (2013), p. 1.

[29] Stephen J Campbell, Nicola D Gold, Richard M Jackson et David R Westhead. « Ligand binding : functional site location, similarity and docking ». In : *Current opinion in structural biology* 13.3 (2003), p. 389–395.

[30] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen et Peer Bork. « Drug target identification using side-effect similarity ». In : *Science* 321.5886 (2008), p. 263–266.

[31] Luisa Castagnoli, Anna Costantini, Claudia Dall'Armi, Stefania Gonfloni, Luisa Montecchi-Palazzi, Simona Panni, Serena Paoluzi, Elena Santonico et Gianni Cesareni. « Selectivity and promiscuity in the interaction network mediated by protein recognition modules ». In : *FEBS letters* 567.1 (2004), p. 74–79.

[32] Gianni Cesareni, Simona Panni, Giuliano Nardelli et Luisa Castagnoli. « Can we infer peptide recognition specificity mediated by SH3 domains ? » In : *FEBS letters* 513.1 (2002), p. 38–44.

[33] Chao Chen, Yang He, Jianhui Wu et Jinming Zhou. « Creation of a free, Internet-accessible database : the Multiple Target Ligand Database. » In : *J. Cheminformatics* (2015), p. 7–14.

[34] Philip Cohen. « Protein kinases—the major drug targets of the twenty-first century ? » In : *Nature reviews Drug discovery* 1.4 (2002), p. 309–315.

[35] Shelley D Copley. « An evolutionary biochemist's perspective on promiscuity ». In : *Trends in biochemical sciences* 40.2 (2015), p. 72–78.

[36] JD Corbin. « Mechanisms of action of PDE5 inhibition in erectile dysfunction ». In : *International journal of impotence research* 16 (2004), S4–S7.

[37] L Costantino et D Barlocco. « Designed multiple ligands : basic research vs clinical outcomes ». In : *Current medicinal chemistry* 19.20 (2012), p. 3353–3387.

[38] Péter Csermely, Vilmos Agoston et Sandor Pongor. « The efficiency of multi-target drugs : the network approach might help drug design ». In : *Trends in pharmacological sciences* 26.4 (2005), p. 178–182.

[39] ROBERT J D'Amato, Michael S Loughnan, Evelyn Flynn et Judah Folkman. « Thalidomide is an inhibitor of angiogenesis ». In : *Proceedings of the National Academy of Sciences* 91.9 (1994), p. 4082–4085.

[40] Arvin C Dar, Tirtha K Das, Kevan M Shokat et Ross L Cagan. « Chemical genetic discovery of targets and anti-targets for cancer polypharmacology ». In : *Nature* 486.7401 (2012), p. 80–84.

[41] Helen Davies, Graham R Bignell, Charles Cox, Philip Stephens, Sarah Edkins, Sheila Clegg, Jon Teague, Hayley Woffendin, Mathew J Garnett, William Bottomley et al. « Mutations of the BRAF gene in human cancer ». In : *Nature* 417.6892 (2002), p. 949–954.

[42] Armelle Debru. *Galen on Pharmacology : Philosophy, History, and Medicine : Proceedings of the Vth International Galen Colloquium, Lille, 16-18 March 1995*. T. 16. Brill, 1997.

[43] WB Deichmann, D Henschler, B Holmstedt et G Keil. « What is there that is not poison ? A study of the Third Defense by Paracelsus ». In : *Archives of Toxicology* 58.4 (1986), p. 207–213.

[44] Todd J Dolinsky, Jens E Nielsen, J Andrew McCammon et Nathan A Baker. « PDB2PQR : an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations ». In : *Nucleic acids research* 32.suppl 2 (2004), W665–W667.

[45] Leif Ellingson et Jinfeng Zhang. « Protein surface matching by combining local and global geometric information ». In : *PloS one* 7.7 (2012), e40540.

[46] Margarita Faig, Mario A Bianchet, Shannon Winski, Robert Hargreaves, Christopher J Moody, Anna R Hudnott, David Ross et L Mario Amzel. « Structure-based development of anticancer drugs : complexes of NAD (P) H : quinone oxidoreductase 1 with chemotherapeutic quinones ». In : *Structure* 9.8 (2001), p. 659–667.

[47] Juan Fernandez-Recio, Max Totrov, Constantin Skorodumov et Ruben Abagyan. « Optimal docking area : a new method for predicting protein–protein interaction sites ». In : *PROTEINS : Structure, Function, and bioinformatics* 58.1 (2005), p. 134–143.

[48] R Fischer. « The lock and key hypothesis of enzymatic action ». In : *Nature* 50 (1894), p. 272–280.

[49] Jonathan B Fitzgerald, Birgit Schoeberl, Ulrik B Nielsen et Peter K Sorger. « Systems biology and combination therapy in the quest for clinical efficacy ». In : *Nature chemical biology* 2.9 (2006), p. 458–466.

[50] Ernesto Freire, Obdulio L Mayorga et Martin Straume. « Isothermal titration calorimetry ». In : *Analytical chemistry* 62.18 (1990), 950A–959A.

[51] Valentin Fuster et Joseph M Sweeny. « Aspirin a historical and contemporary therapeutic overview ». In : *Circulation* 123.7 (2011), p. 768–778.

[52] Ketan S Gajiwala, Joe C Wu, James Christensen, Gayatri D Deshmukh, Wade Diehl, Jonathan P DiNitto, Jessie M English, Michael J Greig, You-Ai He, Suzanne L Jacques et al. « KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients ». In : *Proceedings of the National Academy of Sciences* 106.5 (2009), p. 1542–1547.

[53] Petia Zvezdanova Gatzeva-topalova, Lisa Rosa Warner, Arthur Pardi et Marcelo Carlos. « NIH Public Access ». In : 18.11 (2011), p. 1492–1501. issn : 1878-5832. arXiv : NIHMS150003.

[54] Tim Geppert, Benjamin Hoy, Silja Wessler et Gisbert Schneider. « Context-based identification of protein-protein interfaces and "hot-spot" residues ». In : *Chemistry & biology* 18.3 (2011), p. 344–353.

[55] Michael Giese, Markus Albrecht et Kari Rissanen. « Experimental investigation of anion-π interactions - applications and biochemical relevance. » In : *Chemical communications (Cambridge, England)* 52 (2015), p. 1778–1795. issn : 1364-548X. doi : 10.1039/c5cc09072e. url : http://pubs.rsc.org/en/content/articlehtml/2016/cc/c5cc09072e.

[56] Silvia Giordano et Alessio Petrelli. « From single-to multi-target drugs in cancer therapy : when aspecificity becomes an advantage ». In : *Current medicinal chemistry* 15.5 (2008), p. 422–432.

[57] Chittibabu Guda, Sifang Lu, Eric D Scheeff, Philip E Bourne et Ilya N Shindyalov. « CE-MC : a multiple protein structure alignment server ». In : *Nucleic acids research* 32.suppl 2 (2004), W100–W103.

[58] Kannan Gunasekaran et Ruth Nussinov. « How different are structurally flexible and rigid binding sites ? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding ». In : *Journal of molecular biology* 365.1 (2007), p. 257–273.

[59] Frédéric Guyon et Pierre Tufféry. « Fast protein fragment similarity scoring using a binet–cauchy kernel ». In : *Bioinformatics* 30.6 (2014), p. 784–791.

[60] Thomas A Halgren. « Identifying and characterizing binding sites and assessing druggability ». In : *Journal of chemical information and modeling* 49.2 (2009), p. 377–389.

[61] Tom Halgren. « New Method for Fast and Accurate Binding-site Identification and Analysis ». In : *Chemical biology & drug design* 69.2 (2007), p. 146–148.

[62] LD Hall. « Nuclear magnetic resonance ». In : *Advances in carbohydrate chemistry* 19 (1964), p. 51–93.

[63] Gordon G Hammes. « Multiple conformational changes in enzyme catalysis ». In : *Biochemistry* 41.26 (2002), p. 8221–8228.

[64] James A Hanley et Barbara J McNeil. « The meaning and use of the area under a receiver operating characteristic (ROC) curve. » In : *Radiology* 143.1 (1982), p. 29–36.

[65] Takeshi Hase, Hiroshi Tanaka, Yasuhiro Suzuki, So Nakagawa et Hiroaki Kitano. « Structure of protein interaction networks and their implications on drug design ». In : *PLoS Comput Biol* 5.10 (2009), e1000550.

[66] V Joachim Haupt, Simone Daminelli et Michael Schroeder. « Drug promiscuity in PDB : protein binding site similarity is key ». In : *PLoS one* 8.6 (2013), e65894.

[67] M Arif Hayat et al. *Principles and techniques of scanning electron microscopy. Biological applications. Volume 1.* Van Nostrand Reinhold Company., 1974.

[68] Manfred Hendlich, Friedrich Rippmann et Gerhard Barnickel. « LIGSITE : automatic and efficient detection of potential small molecule-binding sites in proteins ». In : *Journal of Molecular Graphics and Modelling* 15.6 (1997), p. 359–363.

[69] Steven Henikoff et Jorja G Henikoff. « Amino acid substitution matrices from protein blocks ». In : *Proceedings of the National Academy of Sciences* 89.22 (1992), p. 10915–10919.

[70] Marylens HERNANDEZ, Dario GHERSI et Roberto SANCHEZ. « SITEHOUND-web : a server for ligand binding site identification in protein structures ». In : *Nucleic Acids Res* 37.Web Server issue (juil. 2009), W413–6. DOI : 10.1093/nar/gkp281.

[71] Brice HOFFMANN, Mikhail ZASLAVSKIY, Jean-Philippe VERT et Véronique STOVEN. « A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D : application to ligand prediction ». In : *BMC bioinformatics* 11.99 (2010).

[72] Andrew L. HOPKINS, Jonathan S. MASON et John P. OVERINGTON. « Can we rationally design promiscuous drugs ? » In : *Current Opinion in Structural Biology* 16.1 (2006), p. 127–136. ISSN : 0959440X.

[73] Bingding HUANG et Michael SCHROEDER. « LIGSITE csc : predicting ligand binding sites using the Connolly surface and degree of conservation ». In : *BMC structural biology* 6.1 (2006), p. 1.

[74] Simon J HUBBARD et Janet M THORNTON. « Naccess ». In : *Computer Program, Department of Biochemistry and Molecular Biology, University College London* 2.1 (1993).

[75] W-C HWANG, A ZHANG et M RAMANATHAN. « Identification of information flow-modulating drug targets : a novel bridging paradigm for drug discovery ». In : *Clinical Pharmacology & Therapeutics* 84.5 (2008), p. 563–572.

[76] Xavier JALENCAS et Jordi MESTRES. « On the origins of drug polypharmacology ». In : *MedChemComm* 4.1 (2013), p. 80–87.

[77] Martin JAMBON, Anne IMBERTY, Gilbert DELÉAGE et Christophe GEOURJON. « A new bioinformatic approach to detect common 3D sites in protein structures ». In : *Proteins : Structure, Function, and Bioinformatics* 52.2 (2003), p. 137–145.

[78] Wolfgang KABSCH et Christian SANDER. « Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features ». In : *Biopolymers* 22.12 (1983), p. 2577–2637.

[79] Abdullah KAHRAMAN, Richard J MORRIS, Roman A LASKOWSKI et Janet M THORNTON. « Shape variation in protein binding pockets and their ligands ». In : *Journal of molecular biology* 368.1 (2007), p. 283–301.

[80] Shumei KATO, Stacy L MOULDER, Naoto T UENO, Jennifer J WHELER, Funda MERIC-BERNSTAM, Razelle KURZROCK et Filip JANKU. « Challenges and perspective of drug repurposing strategies in early phase clinical trials ». In : *Oncoscience* 2.6 (2015), p. 576–580.

[81] Michael J KEISER, Vincent SETOLA, John J IRWIN, Christian LAGGNER, Atheir I ABBAS, Sandra J HUFEISEN, Niels H JENSEN, Michael B KUIJER, Roberto C MATOS, Thuy B TRAN et al. « Predicting new molecular targets for known drugs ». In : *Nature* 462.7270 (2009), p. 175–181.

[82] Curtis T KEITH, Alexis A BORISY et Brent R STOCKWELL. « Multicomponent therapeutics for networked systems ». In : *Nature reviews Drug discovery* 4.1 (2005), p. 71–78.

[83] Nancy P KELLER et Thomas M HOHN. « Metabolic pathway gene clusters in filamentous fungi ». In : *Fungal Genetics and Biology* 21.1 (1997), p. 17–29.

[84] ES KEMPNER. « Movable lobes and flexible loops in proteins structural deformations that control biochemical activity ». In : *FEBS letters* 326.1-3 (1993), p. 4–10.

[85] Kengo KINOSHITA, Jun'ichi FURUI et Haruki NAKAMURA. « Identification of protein functions from a molecular surface database, eF-site ». In : *Journal of structural and functional genomics* 2.1 (2002), p. 9–22.

[86] Harold Philip KLUG, Leroy Elbert ALEXANDER et al. *X-ray diffraction procedures*. T. 2. Wiley New York, 1954.

[87] Zachary A KNIGHT, Henry LIN et Kevan M SHOKAT. « Targeting the cancer kinome through polypharmacology ». In : *Nature Reviews Cancer* 10.2 (2010), p. 130–137.

[88] Janez Konc, Tomo Česnik, Joanna Trykowska Konc, Matej Penca et Dušanka Janežič. « ProBiS-Database : precalculated binding site similarities and local pairwise alignments of PDB structures ». In : *Journal of chemical information and modeling* 52.2 (2012), p. 604–612.

[89] Janez Konc et Dušanka Janežič. « ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment ». In : *Bioinformatics* 26.9 (2010), p. 1160–1168. issn : 13674803.

[90] DC Koningsberger et Roelof Prins. « X-ray absorption : principles, applications, techniques of EXAFS, SEXAFS, and XANES ». In : (1988).

[91] Alexandr P Kornev, Nina M Haste, Susan S Taylor et Lynn F Ten Eyck. « Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism ». In : *Proceedings of the National Academy of Sciences* 103.47 (2006), p. 17783–17788.

[92] Daniel E Koshland. « The key–lock theory and the induced fit theory ». In : *Angewandte Chemie International Edition in English* 33.23-24 (1995), p. 2375–2378.

[93] DE Koshland. « Application of a theory of enzyme specificity to protein synthesis ». In : *Proceedings of the National Academy of Sciences* 44.2 (1958), p. 98–104.

[94] Timo Krotzky, Thomas Rickmeyer, Thomas Fober et Gerhard Klebe. « Extraction of protein binding pockets in close neighborhood of bound ligands makes comparisons simple due to inherent shape similarity ». In : *Journal of chemical information and modeling* 54.11 (2014), p. 3229–3237.

[95] Hugo Kubinyi. *3D QSAR in drug design : volume 1 : theory methods and applications*. T. 1. Springer Science & Business Media, 1993.

[96] Mélaine A Kuenemann, Olivier Sperandio, Céline M Labbé, David Lagorce, Maria A Miteva et Bruno O Villoutreix. « In silico design of low molecular weight protein–protein interaction inhibitors : overall concept

and recent advances ». In : *Progress in biophysics and molecular biology* 119.1 (2015), p. 20–32.

[97] Irina KUFAREVA, Andrey V ILATOVSKIY et Ruben ABAGYAN. « Pocketome : an encyclopedia of small-molecule binding sites in 4D ». In : *Nucleic acids research* 40.D1 (2012), p. D535–D540.

[98] Michael KUHN, Monica CAMPILLOS, Ivica LETUNIC, Lars Juhl JENSEN et Peer BORK. « A side effect resource to capture phenotypic effects of drugs ». In : *Molecular systems biology* 6.1 (2010).

[99] Michael KUHN, Christian von MERING, Monica CAMPILLOS, Lars Juhl JENSEN et Peer BORK. « STITCH : interaction networks of chemicals and proteins ». In : *Nucleic acids research* 36.suppl 1 (2008), p. D684–D688.

[100] Michael KUHN, Damian SZKLARCZYK, Sune PLETSCHER-FRANKILD, Thomas H BLICHER, Christian von MERING, Lars J JENSEN et Peer BORK. « STITCH 4 : integration of protein–chemical interactions with user data ». In : *Nucleic acids research* (2013).

[101] Katrin KUPAS, Alfred ULTSCH et Gerhard KLEBE. « Large scale analysis of protein-binding cavities using self-organizing maps and wavelet-based surface patches to describe functional properties, selectivity discrimination, and putative cross-reactivity ». In : *Proteins : Structure, Function, and Bioinformatics* 71.3 (2008), p. 1288–1306.

[102] Alan K KUTACH, Armando G VILLASEÑOR, Diana LAM, Charles BELUNIS, Cheryl JANSON, Stephen LOK, Li-Na HONG, Chao-Min LIU, Jerome DEVAL, Thomas J NOVAK et al. « Crystal Structures of IL-2-inducible T cell Kinase Complexed with Inhibitors : Insights into Rational Drug Design and Activity Regulation ». In : *Chemical biology & drug design* 76.2 (2010), p. 154–163.

[103] Joseph R LAKOWICZ. *Principles of fluorescence spectroscopy*. Springer Science & Business Media, 2013.

[104] B Lalanne et S Griffon. « Papyrus Ebers : nouvelle transcription, translittération, traduction ». In : *Pessac : Ed. Association égyptologique de Gironde* (2003).

[105] Yehezkel Lamdan et Haim J Wolfson. « Geometric hashing : A general and efficient model-based recognition scheme ». In : (1988).

[106] Roman A Laskowski. « SURFNET : a program for visualizing molecular surfaces, cavities, and intermolecular interactions ». In : *Journal of molecular graphics* 13.5 (1995), p. 323–330.

[107] Alasdair TR Laurie et Richard M Jackson. « Q-SiteFinder : an energy-based method for the prediction of protein–ligand binding sites ». In : *Bioinformatics* 21.9 (2005), p. 1908–1916.

[108] Paul C Lauterbur. « Image formation by induced local interactions : examples employing nuclear magnetic resonance ». In : (1973).

[109] Byungkook Lee et Frederic M Richards. « The interpretation of protein structures : estimation of static accessibility ». In : *Journal of molecular biology* 55.3 (1971), 379–IN4.

[110] Jean-Jacques Lefrère et Patrick Berche. « Les bébés du thalidomide ». In : *La Presse Médicale* 40.3 (2011), p. 301–308.

[111] Wolfgang Lenz, Rudolf A Pfeiffer, Wilhelm Kosenow et DJ Hayman. « Thalidomide and congenital abnormalities ». In : *The Lancet* 279.7219 (1962), p. 45–46.

[112] David G Levitt et Leonard J Banaszak. « POCKET : A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids ». In : *Journal of molecular graphics* 10.4 (1992), p. 229–234.

[113] Jie Liang, Clare Woodward et Herbert Edelsbrunner. « Anatomy of protein pockets and cavities : measurement of binding site geometry and implications for ligand design ». In : *Protein science* 7.9 (1998), p. 1884–1897.

[114] David W Litchfield. « Protein kinase CK2 : structure, regulation and role in cellular decisions of life and death ». In : *Biochemical Journal* 369.1 (2003), p. 1–15.

[115] Eugen Lounkine, Michael J Keiser, Steven Whitebread, Dmitri Mikhailov, Jeremy Jenkins, Paul Lavan, Eckhard Weber, Allison K Doak, Serge Côté, Brian K Shoichet et Laszlo Urban. « NIH Public Access ». In : 486.7403 (2012), p. 361–367.

[116] Xiao Hua Ma, Zhe Shi, Chunyan Tan, Yuyang Jiang, Mei Lin Go, Boon Chuan Low et Yu Zong Chen. « In-Silico approaches to multi-target drug discovery ». In : *Pharmaceutical research* 27.5 (2010), p. 739–749.

[117] Lora Mak, Scott Grandison et Richard J Morris. « An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison ». In : *Journal of Molecular Graphics and Modelling* 26.7 (2008), p. 1035–1045.

[118] Mathew P Martin, Riazul Alam, Stephane Betzi, Donna J Ingles, Jin-Yi Zhu et Ernst Schönbrunn. « A Novel Approach to the Discovery of Small-Molecule Ligands of CDK2 ». In : *Chembiochem* 13.14 (2012), p. 2128–2136.

[119] Simon K Mencher et Long G Wang. « Promiscuous drugs compared to selective drugs (promiscuity can be a virtue) ». In : *BMC Pharmacology and Toxicology* 5.1 (2005), p. 3.

[120] Jordi Mestres et Elisabet Gregori-Puigjané. « Conciliating binding efficiency and polypharmacology ». In : *Trends in pharmacological sciences* 30.9 (2009), p. 470–474.

[121] Jordi Mestres, Elisabet Gregori-Puigjané, Sergi Valverde et Ricard V Solé. « The topology of drug–target interaction networks : implicit dependence on drug properties and target families ». In : *Molecular BioSystems* 5.9 (2009), p. 1051–1057.

[122] Mark J Millan. « Multi-target strategies for the improved treatment of depressive states : Conceptual foundations and neuronal substrates, drug discovery and therapeutic application. » In : *Pharmacology & therapeutics* 110.2 (2006), p. 135–370.

[123] Gautier Moroy, Elyette Martin, Annick Dejaegere et Roland H Stote. « Molecular Basis for Bcl-2 Homology 3 Domain Recognition in the Bcl-2 Protein Family IDENTIFICATION OF CONSERVED HOT SPOT INTERACTIONS ». In : *Journal of Biological Chemistry* 284.26 (2009), p. 17499–17511.

[124] Garrett M Morris, David S Goodsell, Robert S Halliday, Ruth Huey, William E Hart, Richard K Belew, Arthur J Olson et al. « Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function ». In : *Journal of computational chemistry* 19.14 (1998), p. 1639–1662.

[125] Richard J Morris. « An evaluation of spherical designs for molecular-like surfaces ». In : *Journal of Molecular Graphics and Modelling* 24.5 (2006), p. 356–361.

[126] S Mouly, V Delcey, M Diemer et J-F Bergmann. « Évaluation de l'efficacité d'un médicament : de la découverte à la mise sur le marché ». In : *Journal Français d'Ophtalmologie* 31.1 (2008), p. 75–79.

[127] Rolf Mülhaupt. « Hermann Staudinger and the origin of macromolecular chemistry ». In : *Angewandte Chemie International Edition* 43.9 (2004), p. 1054–1063.

[128] Rafael Najmanovich, Natalja Kurbatova et Janet Thornton. « Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites ». In : *Bioinformatics* 24.16 (2008), p. i105–i111.

[129] Murad Nayal et Barry Honig. « On the nature of cavities on protein surfaces : application to the identification of drug-binding sites ». In : *Proteins : Structure, Function, and Bioinformatics* 63.4 (2006), p. 892–906.

[130] Yasutomi Nishizuka. « The role of protein kinase C in cell surface signal transduction and tumour promotion ». In : *Nature* 308.5961 (1984), p. 693–698.

[131] Britta Nisius, Fan Sha et Holger Gohlke. « Structure-based computational analysis of protein binding sites for function and druggability prediction ». In : *Journal of biotechnology* 159.3 (2012), p. 123–134.

[132] P Noize, M Sauer, P Bruneval, M Moreau, A Pathak, H Bagheri et JL Montastruc. « Valvular heart disease in a patient taking benfluorex ». In : *Fundamental & clinical pharmacology* 20.6 (2006), p. 577–578.

[133] Mats HM Olsson, Chresten R Søndergaard, Michal Rostkowski et Jan H Jensen. « PROPKA3 : consistent treatment of internal and surface residues in empirical p K a predictions ». In : *Journal of Chemical Theory and Computation* 7.2 (2011), p. 525–537.

[134] K Opitz. « Tolerance and cross tolerance to the anorexigenic effect of appetite suppressants in rats. » In : *International journal of obesity* 2.1 (1977), p. 59–68.

[135] Christine A Orengo, AD Michie, S Jones, David T Jones, MB Swindells et Janet M Thornton. « CATH–a hierarchic classification of protein domain structures ». In : *Structure* 5.8 (1997), p. 1093–1109.

[136] Walter Pagel. *Paracelsus : an introduction to philosophical medicine in the era of the Renaissance*. Karger Medical et Scientific Publishers, 1982.

[137] Xavier Pennec et Nicholas Ayache. « A geometric algorithm to find small but highly similar 3D substructures in proteins. » In : *Bioinformatics* 14.6 (1998), p. 516–522.

[138] Klaus P Peters, Jana Fauck et Cornelius Frömmel. « The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria ». In : *Journal of molecular biology* 256.1 (1996), p. 201–213.

[139] P HAYDN Pritchard, MARIANA Bowley, SUSAN L Burditt, J Cooling, HP Glenny, N Lawson, RG Sturton et DN Brindley. « The effects of acute ethanol feeding and of chronic benfluorex administration on the activities of some enzymes of glycerolipid synthesis in rat liver and adipose tissue ». In : *Biochemical Journal* 166.3 (1977), p. 639–642.

[140] BR Rabin. « Co-operative effects in enzyme catalysis : a possible kinetic model based on substrate-induced conformation isomerization. » In : *Biochemical Journal* 102.2 (1967), p. 22C.

[141] SV Rajkumar et TE Witzig. « A review of angiogenesis and antiangiogenic therapy with thalidomide in multiple myeloma ». In : *Cancer treatment reviews* 26.5 (2000), p. 351–362.

[142] A Srinivas Reddy et Shuxing Zhang. « Polypharmacology : drug discovery for the future ». In : *Expert review of clinical pharmacology* 6.1 (2013), p. 41–47.

[143] Dagmar Ringe. « What makes a binding site a binding site ? » In : *Current opinion in structural biology* 5.6 (1995), p. 825–829.

[144] Bryan L Roth, Douglas J Sheffler et Wesley K Kroeze. « Magic shotguns versus magic bullets : selectively non-selective drugs for mood disorders and schizophrenia ». In : *Nature Reviews Drug Discovery* 3.4 (2004), p. 353–359.

[145] Lee Sael et Daisuke Kihara. « Binding ligand prediction for proteins using partial matching of local surface patches ». In : *International Journal of Molecular Sciences* 11.12 (2010), p. 5009–5026.

[146] Lee Sael, David La, Bin Li, Raif Rustamov et Daisuke Kihara. « Rapid comparison of properties on protein surface ». In : *Proteins : Structure, function, and bioinformatics* 73.1 (2008), p. 1–10.

[147] Adrien Saladin, Julien Rey, Pierre Thévenet, Martin Zacharias, Gautier Moroy et Pierre Tufféry. « PEP-SiteFinder : a tool for the blind identification of peptide binding sites on protein surfaces ». In : *Nucleic acids research* 42.W1 (2014), W221–W226.

[148] Frank Sams-Dodd. « Drug discovery : selecting the optimal approach ». In : *Drug discovery today* 11.9 (2006), p. 465–472.

[149] Andrey K Sarychev, Gennady Shvets et Vladimir M Shalaev. « Magnetic plasmon resonance ». In : *Physical Review E* 73.3 (2006), p. 036609.

[150] Thomas Schindler, William Bornmann, Patricia Pellicena, W Todd Miller, Bayard Clarkson et John Kuriyan. « Structural mechanism for STI-571 inhibition of abelson tyrosine kinase ». In : *Science* 289.5486 (2000), p. 1938–1942.

[151] Stefan Schmitt, Daniel Kuhn et Gerhard Klebe. « A new method to detect related function among proteins independent of sequence and fold homology ». In : *Journal of molecular biology* 323.2 (2002), p. 387–406.

[152] Gideon Schreiber et Amy E Keating. « Protein binding specificity versus promiscuity ». In : *Current opinion in structural biology* 21.1 (2011), p. 50–61.

[153] Schrödinger, LLC. « The PyMOL Molecular Graphics System, Version 1.8 ». Nov. 2015.

[154] D Schuster, C Laggner et T Langer. « Why drugs fail-a study on side effects in new chemical entities ». In : *Current pharmaceutical design* 11.27 (2005), p. 3545–3559.

[155] Stephen B Seidman. « Network structure and minimum degree ». In : *Social networks* 5.3 (1983), p. 269–287.

[156] Hideaki Shimizu, David J Schuller, William N Lanzilotta, M Sundaramoorthy, David M Arciero, Alan B Hooper et Thomas L Poulos. « Crystal structure of Nitrosomonas europaea cytochrome c peroxidase and the structural basis for ligand switching in bacterial di-heme peroxidases ». In : *Biochemistry* 40.45 (2001), p. 13483–13490.

[157] Ilya N Shindyalov et Philip E Bourne. « Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. » In : *Protein engineering* 11.9 (1998), p. 739–747.

[158] Lydia Siragusa, Simon Cross, Massimo Baroni, Laura Goracci et Gabriele Cruciani. « BioGPS : navigating biological space to predict polypharmacology, off-targeting, and selectivity ». In : *Proteins* 83.3 (2015), p. 517–32.

[159] Chresten R Sondergaard, Mats HM Olsson, Michal Rostkowski et Jan H Jensen. « Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values ». In : *Journal of Chemical Theory and Computation* 7.7 (2011), p. 2284–2295.

[160] Russell Spitzer, Ann E Cleves et Ajay N Jain. « Surface-based protein binding pocket similarity ». In : *Proteins* 79.9 (2011), p. 2746–63.

[161] Russell Spitzer, Ann E Cleves, Rocco Varela et Ajay N Jain. « Protein function annotation by local binding site surface similarity ». In : *Proteins : Structure, Function, and Bioinformatics* 82.4 (2014), p. 679–694.

[162] Nicholas P Tatonetti, Tianyun Liu et Russ B Altman. « Predicting drug side-effects by chemical systems biology ». In : *Genome Biol* 10.9 (2009), p. 238.

[163] Mikael Trellet, Adrien SJ Melquiond et Alexandre MJJ Bonvin. « A unified conformational selection and induced fit approach to protein-peptide docking ». In : *PloS one* 8.3 (2013), e58769.

[164] Charalampos Tsourakakis, Francesco Bonchi, Aristides Gionis, Francesco Gullo et Maria Tsiarli. « Denser than the densest subgraph : extracting optimal quasi-cliques with quality guarantees ». In : *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, p. 104–112.

[165] Laura N Vandenberg, Theo Colborn, Tyrone B Hayes, Jerrold J Heindel, David R Jacobs Jr, Duk-Hee Lee, Toshi Shioda, Ana M Soto, Frederick S vom Saal, Wade V Welshons et al. « Hormones and endocrine-disrupting chemicals : low-dose effects and nonmonotonic dose responses ». In : *Endocrine reviews* 33.3 (2012), p. 378–455.

[166] Cherayathumadom M Venkatachalam, Xiaohui Jiang, Tom Oldfield et Marvin Waldman. « LigandFit : a novel method for the shape-directed rapid docking of ligands to protein active sites ». In : *Journal of Molecular Graphics and Modelling* 21.4 (2003), p. 289–307.

[167] Ingo Vogt et Jordi Mestres. « Drug-target networks ». In : *Molecular Informatics* 29.1-2 (2010), p. 10–14.

[168] Andrew C Wallace, Neera Borkakoti et Janet M Thornton. « TESS : a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites ». In : *Protein science* 6.11 (1997), p. 2308–2323.

[169] Caihua Wang, Juan Liu, Fei Luo, Zixing Deng et Qian-Nan Hu. « Predicting target-ligand interactions using protein ligand-binding site and ligand substructures ». In : *BMC systems biology* 9.1 (2015), p. 1.

[170] Charles Webster. « Alchemical and Paracelsian medicine ». In : *Cambridge Monographs on the History of Medicine* (1979), p. 301–334.

[171] Martin Weisel, Ewgenij Proschak, Gisbert Schneider et al. « PocketPicker : analysis of ligand binding-sites with shape descriptors ». In : *Chem Cent J* 1.7 (2007), p. 1–17.

[172] Camille G Wermuth. « Selective optimization of side activities : the SOSA approach ». In : *Drug discovery today* 11.3 (2006), p. 160–164.

[173] Camille G Wermuth. « The SOSA approach : an alternative to high-throughput screening ». In : *Medicinal chemistry research* 10.7-8 (2001), p. 431–439.

[174] Edward Whitehead. « The regulation of enzyme activity and allosteric transition ». In : *Progress in biophysics and molecular biology* 21 (1970), p. 321–397.

[175] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang et Jennifer Woolsey.

« DrugBank : a comprehensive resource for in silico drug discovery and exploration ». In : *Nucleic acids research* 34.suppl 1 (2006), p. D668–D672.

[176] Shoshana J Wodak et Joël Janin. « Analytical approximation to the accessible surface area of proteins ». In : *Proceedings of the National Academy of Sciences* 77.4 (1980), p. 1736–1740.

[177] Alastair JJ Wood, William E Evans et Howard L McLeod. « Pharmacogenomics—drug disposition, drug targets, and side effects ». In : *New England Journal of Medicine* 348.6 (2003), p. 538–549.

[178] CY Wu et JJ Wittick. « Separation of five major alkaloids in gum opium and quantitation of morphine, codeine, and thebaine by isocratic reverse phase high performance liquid chromatography ». In : *Analytical chemistry* 49.3 (1977), p. 359–363.

[179] Jinyang Xi et Xin Xu. « Understanding the anion–π interactions with tetraoxacalix[2]arene[2]triazine ». In : *Phys. Chem. Chem. Phys.* 18.9 (2016), p. 6913–6924. issn : 1463-9076.

[180] Lei Xie, Li Xie et Philip E Bourne. « Structure-based systems biology for analyzing off-target binding ». In : *Current opinion in structural biology* 21.2 (2011), p. 189–199.

[181] Li Xie, Thomas Evangelidis, Lei Xie et Philip E Bourne. « Drug discovery using chemical systems biology : weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir ». In : *PLoS Comput Biol* 7.4 (2011), e1002037.

[182] Miao Xu, Emily M Lee, Zhexing Wen, Yichen Cheng, Wei-Kai Huang, Xuyu Qian, Julia TCW, Jennifer Kouznetsova, Sarah C Ogden, Christy Hammack, Fadi Jacob, Ha Nam Nguyen, Misha Itkin, Catherine Hanna, Paul Shinn, Chase Allen, Samuel G Michael, Anton Simeonov, Wenwei Huang, Kimberly M Christian, Alison Goate, Kristen J Brennand, Ruili Huang, Menghang Xia, Guo-li Ming, Wei Zheng, Hongjun Song et Hengli Tang. « Identification of small-molecule inhibitors of Zika

virus infection and induced neural cell death via a drug repurposing screen ». In : *Nat Med* advance online publication (2016).

[183] Changhui YAN, Vasant HONAVAR et Drena DOBBS. « Identification of interface residues in protease-inhibitor and antigen-antibody complexes : a support vector machine approach ». In : *Neural computing & applications* 13.2 (2004), p. 123–129.

[184] Kun YANG, Hongjun BAI, Qi OUYANG, Luhua LAI et Chao TANG. « Finding multiple target optimal intervention in disease-related molecular network ». In : *Molecular Systems Biology* 4.1 (2008), p. 228.

[185] Kalidas YETURU et Nagasuma CHANDRA. « PocketMatch : a new algorithm to compare binding sites in protein structures ». In : *BMC bioinformatics* 9.1 (2008), p. 1.

[186] Tracy A YOUNG, Benedicte DELAGOUTTE, James A ENDRIZZI, Arnold M FALICK et Tom ALBER. « Structure of Mycobacterium tuberculosis PknB supports a universal activation mechanism for Ser/Thr protein kinases ». In : *Nature Structural & Molecular Biology* 10.3 (2003), p. 168–174.

[187] Maria I ZAVODSZKY et Leslie A KUHN. « Side-chain flexibility in protein–ligand binding : The minimal rotation hypothesis ». In : *Protein Science* 14.4 (2005), p. 1104–1114.

[188] Jianming ZHANG, Francisco J ADRIÁN, Wolfgang JAHNKE, Sandra W COWAN-JACOB, Allen G LI, Roxana E IACOB, Taebo SIM, John POWERS, Christine DIERKS, Fangxian SUN et al. « Targeting Bcr–Abl by combining allosteric with ATP-binding-site inhibitors ». In : *Nature* 463.7280 (2010), p. 501–506.

[189] Huan-Xiang ZHOU et Yibing SHAN. « Prediction of protein interaction sites from sequence profile and residue neighbor list ». In : *Proteins : Structure, Function, and Bioinformatics* 44.3 (2001), p. 336–343.

[190] Grant R Zimmermann, Joseph Lehar et Curtis T Keith. « Multi-target therapeutics : when the whole is greater than the sum of the parts ». In : *Drug discovery today* 12.1 (2007), p. 34–42.

Liste des abréviations

1PE : pentaéthylène glycol

ADME : Absorption Distribution Métabolisme Excrétion

AMM : Autorisation de Mise sur le Marché

AMP : Adénosine Monophosphate

ANSM : Agence Nationale de Sécurité du Médicament

ARN : Acide Désoxyribonucléique

ARG : arginine

ASA : Accessible Solvent Area

ASN : asparagine

ATP : Adénosine Triphosphate

AUC : Area Under the Curve

BLOSUM : BLOcks Substitution Matrix

BOG : béta octyl glucoside

Dc : Distance entre centroïdes

EC : Enzyme Commission

FAD : Flavine Adénine Dinucléotide

FMN : Flavine Mononucléotide

GLC : Glucose

GSH : glutathion

H.S. : Harmonies Sphériques

HD : Hoffmann Dataset

HD$_{\mathbf{nous}}$: patches et surfaces extraits des complexes du Hoffmann dataset selon

notre protocole

HD_{orig} : patches et surfaces extraits des complexes du Hoffmann dataset par Hoffmann et ses collègues

KD : Kahraman Dataset

K_{nous} : patches et surfaces extraits des complexes du Kahraman dataset selon notre protocole

K_{orig} : patches et surfaces extraits des complexes du Kahraman dataset par Hoffmann et ses collègues

LDA : lauryl diméthylamine-n-oxyde

LLP : acide hexanoïque 2-amino-6-[[3-hydroxy-2-méthyl-5-(phosphonooxyméthyl)pyridine-4-yl]méthylideneamino]

MAP : Mitogen Activated Protein

MTLD : Multiple Target Ligand Database

NAD : Nicotinamide Adénine Dinucléotide

NR : Non Renseignée

PDB : Protein Data Bank

PO_4 : Phosphate

PLM : acide palmitique

PMP : pyridoxine-5'-phosphate

ROC : Receiver Operating Curve

RMN : Résonance Magnétique Nucléaire

RMSD : Root Mean Squared Deviation

SAM : S-adénosylméthionine

SOM : Self Organizing Maps

SUC : sucrose

syk : spleen tyrosine-kinase

TFP : Taux de Faux Positifs

TVP : Taux de Vrais Positifs

TYR : tyrosine

U5P : uridine-5'-monophosphate

VEGF : Vascular Endothelial Growth Factor

VN : Vrai Négatif

VP : Vrai Positif

Résumé

La détection de potentielles cibles secondaires ou *off-targets* d'un ligand donné requiert la détermination de son site d'interaction et la recherche de sites d'interaction similaires sur d'autres protéines. Dans le but de mener à bien cette étude, nous avons développé PatchSearch : cet outil compare un patch requête, correspondant à un site d'interaction, avec la surface d'une cible potentielle. L'algorithme employé s'appuie sur une méthode originale de recherche de quasi-cliques dans un graphe produit : cette approche identifie des groupes d'atomes du patch appariés avec ceux de la surface ciblée avec des propriétés physico-chimiques conservées et dans des configurations proches. Nous montrons que PatchSearch trouve des patches qui correspondent à ceux qui sont connus sur les surfaces ciblées. De plus, les résultats de l'application de PatchSearch sur des protéines flexibles indiquent que l'approche des quasi-cliques permet de retrouver à la fois les parties rigides et flexibles des patches, contrairement à la recherche classique de cliques. Enfin, les performances de PatchSearch sont équivalentes à celles des autres outils de comparaison de sites de liaison. Nous avons également appliqué PatchSearch sur des *off-targets* de médicaments impliqués dans le traitement de cancers. Nos expériences suggèrent l'utilisation de PatchSearch dans la recherche des éventuelles *off-targets* d'un médicament.

Abstract

Detection of putative off-targets for a ligand requires to search for some similar binding sites onto other proteins surface. In order to achieve this goal, we developed a tool named PatchSearch. This program compares a query patch, which contains the binding site, with the surface a potentially targeted protein. PatchSearch's algorithm is based on an original method searching for some quasi-cliques in a graph product, which identifies some atoms both in the patch and in the surface with conserved physicochemical properties and in similar configurations. We show that PatchSearch efficiently finds known patches on protein surfaces. Moreover, application of PatchSearch on flexible proteins shows that, unlike the classic cliques approach, quasi-cliques method allows to find both rigid and flexible parts of the patches. PatchSearch gets similar results compared to the other binding site comparison tools. We also applied PatchSearch to find patches binding polypharmacological drugs involved in cancer treatment, in order to identify them on known off targets. Our experiments suggest to employ PatchSearch in off-targets detection process.