Développement et implémentation d'une approche par fragments pour le design d'ARNs modifiés simple brin avec évaluation sur des protéines de liaison à l'ARN et un modèle d'étude la Bêta-Sécrétase 1

Taher Yacoub

université PARIS-SACLAY

Développement et implémentation d'une approche par fragments pour le design d'ARNs modifiés simple brin avec évaluation sur des protéines de liaison à l'ARN et un modèle d'étude la Bêta-Sécrétase 1

Development and implementation of a fragment-based approach to design single-stranded modified RNAs with evaluation on RNA-binding proteins and the Beta-Secretase 1 enzyme

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et Dynamique des systèmes vivants (SDSV)
Spécialité de doctorat : Biochimie et Biologie Structurale
Graduate School : Life Sciences and Health. Référent : Faculté des Sciences d'Orsay

Thèse préparée dans l'unité de recherche **Institute for Integrative Biology of the Cell (I2BC) (Université Paris-Saclay, CEA, CNRS)**, sous la direction de **Fabrice LECLERC**, Chargé de Recherche et le co-encadrement de **Yann PONTY**, Directeur de Recherche (LIX, Institut Polytechnique de Paris, CNRS, INRIA)

Thèse soutenue à Paris-Saclay, le 17 janvier 2024, par

Taher YACOUB

Composition du jury

Membres du jury avec voix délibérative

Raphael GUEROIS Directeur de Recherche, Université Paris-Saclay, CEA	Président
Juliette MARTIN Directrice de Recherche, Université Claude-Bernard Lyon, CNRS	Rapportrice & Examinatrice
Dorian MAZAURIC Chargé de Recherche (HDR), Université Côte d'Azur, INRIA	Rapporteur & Examinateur
Fariza TAHI Professeure des Universités, Université Paris-Saclay	Examinatrice

THESE DE DOCTORAT
NNT : 2024UPASL002

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Développement et implémentation d'une approche par fragments pour le design d'ARNs modifiés simple brin avec évaluation sur des protéines de liaison à l'ARN et un modèle d'étude la Bêta-Sécrétase 1

Mots clés : Aptamères, Approche par Fragments, Alogirthmes de complexité paramétrés, Color-Coding, Maladie d'Alzheimer, Beta-Sécrétase 1

Résumé : Les ARNs sont des biomolécules avec des fonctions différentes ou générées de façon artificielle. Nous pouvons citer les aptamères, de courtes séquences d'ARN (ou ADN). Leur génération comme médicaments est une approche limitée. Des modifications chimiques introduites dans les aptamères permettent d'améliorer leurs propriétés, mais leur diversité chimique est restreinte et constitue une limitation majeure. Nous proposons une nouvelle approche computationnelle pour s'affranchir de ces limitations, et qui modélise ce problème en un problème informatique. Cette nouvelle approche a été évaluée afin d'en tester sa robustesse et sa pertinence. Pour la mettre à l'épreuve, une vraie étude a été réalisée sur une protéine impliquée dans la maladie d'Alzheimer. Une analyse structurale a été réalisée dans l'objectif de générer des aptamères modifiés avec notre nouvelle approche. Les résultats montrent la possibilité de cibler les sites fonctionnels de cette protéine avec des nucléotides modifiés.

Title : Development and implementation of a fragment-based approach to design single-stranded modified RNAs with evaluation on RNA-binding proteins and the Beta-Secretase 1 enzyme

Keywords : Aptamers, Fragment-based design, Parameterized complexity algorithms, Color-Coding, Alzheimer's disease, Beta-Secretase 1

Abstract : RNAs are biomolecules with different functions or generated artificially. We can cite aptamers (short sequences of RNA or DNA). Their generation as drugs is a limited approach. Chemical modifications introduced in the aptamers enable to improve their properties, but their chemical diversity is restricted and is a major limitation. We propose a new computational approach to overcome these limitations, and to model this problem into a mathematical and computer problem. This new approach has been evaluated in order to test its robustness and relevance on seven RNA/protein complexes. A real therapeutic study was carried out on a protein involved in Alzheimer's disease to evaluate this new approach. A structural analysis was carried out with the aim of generating modified aptamers using our new approach. The results show the possibility of targeting the functional sites of this protein with modified nucleotides.

Liste de Publications

[1] Yacoub, T., González-Alemán, R., Leclerc, F. *Color Coding for the Fragment-Based Docking, Design and Equilibrium Statistics of Protein-Binding ssRNAs*. Pre-Print 2023 hal : hal-03816423v2

[2] Yacoub, T., González-Alemán, R., Montero-Cabrera, L., Alvarez-Ginarte, Y., Leclerc, F. *Integrative Modeling to Identify Selectivity Features and Nucleotide Fragments For BACE1 Inhibiton*. Unpublished

Je tiens à adresser mes remerciements à mes superviseurs, Fabrice LECLERC et Yann PONTY, de m'avoir accordé leur confiance pour collaborer sur ce projet de recherche, qui m'a permis de grandir dans l'interdisciplinarité. Ce travail de près de trois années fut un challenge incroyable, avec des discussions scientifiques enrichissantes. Je les remercie également de leur écoute, de leur patience, de leur disponibilité et de leur compréhension.

J'adresse mes remerciements à la Fondation Vaincre Alzheimer, pour avoir financé ce projet de thèse et m'avoir permis de travailler sur un sujet qui est un enjeu majeur de la santé publique. Cela m'a permis de mieux percevoir l'intérêt de faire de la prévention auprès du grand public, notamment en participant au concours "Ma Thèse en 180 secondes".

J'adresse mes remerciements à Claire Toffano-Nioche (membre de l'équipe SSFA), à Sebastian Will (membre de l'équipe AMIBIO) ainsi qu'aux membres de mon comité de suivi (Liliane Mouawad, Pierre Barraud, Dominique Barth) pour leur écoute et leurs conseils.

J'adresse mes remerciements à la totalité des membres de l'équipe SSFA et de l'équipe AMIBIO de m'avoir aidé à m'intégrer au sein d'un environnement interdisciplinaire et de m'avoir perçu comme un membre à part entière.

J'adresse mes remerciements à Isaure Chauvot De Beauchêne pour une collaboration enrichissante et de m'avoir donné l'opportunité de travailler à ses côtés durant la fin de ma thèse.

Je remercie les membres du jury d'avoir porté intérêt à mon sujet de recherche et pour nos discussions profondes.

Enfin, je conserve une pensée pour toutes les personnes de mon entourage qui m'ont accompagné et encouragé durant ces trois années.

Table des matières

1 - Introduction Générale

1.1 . Approche par Fragments

L'approche par Fragments est une méthode qui fut proposée par le laboratoire Abbott en 1996[1]. Cette approche est utilisée à la fois dans le milieu académique et industriel afin de concevoir des molécules puissantes, et pour lesquelles cinq molécules ont été approuvées par la FDA à ce jour[2–6]. En effet, elle consiste à convertir un fragment ou un ensemble de fragments ("fragments-hits") en une molécule thérapeutique ("molecule-lead").

Par analogie à la règle des 5 de Lipinski, un fragment est défini selon la règle des 3 (RO3), bien que non absolue notamment dans le cadre d'un design d'ARN simple brin (**ARNsb** ou **oligonucléotide**). En théorie, un fragment est défini comme ayant un poids moléculaire inférieur à 300Da, un logP inférieur à 3, et établissant au plus 3 liaisons hydrogènes accepteurs et donneurs.

Le faible poids moléculaire des fragments limite leur affinité potentielle, de l'ordre du mM généralement dans le meilleur des cas. Par conséquent, un fragment n'a de valeur intrinsèque que dans le contexte de sa proximité avec d'autres fragments dont la liaison ("linking"), fusion ("merging"), ou extension ("growing") permet d'obtenir un ligand optimisé.

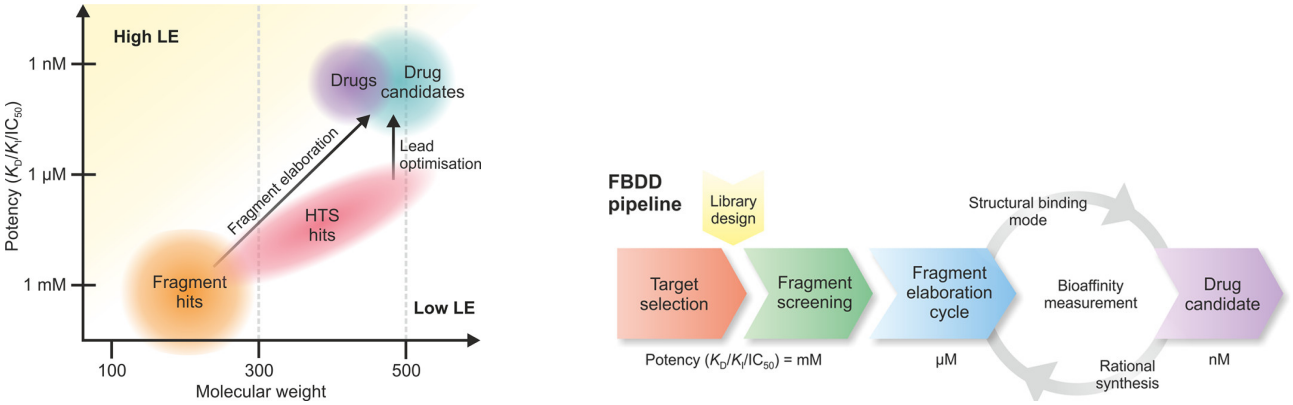
Cette approche par fragments repose sur trois principales étapes[7,8] :

1. la constitution d'une bibliothèque de fragments,
2. le criblage sur une protéine cible et l'identification des fragments-hits,
3. l'optimisation des fragments-hits en molécule-lead

1.1.1 . Les étapes de l'approche par fragments

La constitution de la librairie de fragments est un point de départ important. S'il n'existe aucune recommandation sur le nombre de fragments qu'elle doit contenir, elle doit être composée de fragments qui incluront des groupes fonctionnels qui faciliteront leur optimisation, et éviter de contenir des *scaffold* instables ou toxiques tels que les groupes alkyl ou acyle. De plus, il est suggéré que la bibliothèque contienne des fragments respectant la règle RO3 bien que cela ne soit pas obligatoire comme en attestent certaines études. D'autres filtres peuvent être appliqués comme des propriétés physicochimiques bien particulières (logP, l'aire de surface polaire...). Ainsi, la constitution de la bibliothèque de fragments permet de filtrer les fragments souhaités et non-souhaités qui serviront de bases pour la conception ("design") d'une molécule active.[7,8]

Dès lors que la bibliothèque de fragments est constituée, l'étape suivante consiste à identifier un fragment-hit qui servira de point d'ancrage pour développer un composé actif. Due à leur faible niveau d'affinité, allant de 0.1mM à 10mM ou plus, quelques techniques expérimentales bien précises



(a) Comparaison entre un fragment-hit et une molécule-lead

(b) Pipeline de la conception d'une molécule candidate-médicament selon la stratégie par Fragments

FIGURE 1.1 – Conception de molécules candidate-médicaments selon l'approche par Fragments.[7]

sont utilisées lors de cette étape. Nous pouvons citer le *Thermal Shift Assay* (TSA), la Spectorscopie RMN (Résonance Magnétique Nucléaire), la Spectroscopie de Masse, la Résonance plasmon de surface ou encore la Cristallographie à rayon X (RX). Parallèlement à ces méthodes expérimentales, le Criblage Virtuel est une approche computationnelle complémentaire possédant ses avantages et ses inconvénients[8].

L'étape d'optimisation est la dernière étape après l'identification des fragments-hits. Elle consiste à optimiser un fragment-hit ayant une affinité de l'ordre du mM à une molécule-lead avec une affinité de l'ordre du nM. Cette étape peut être réalisée selon trois stratégies différentes (Figure 1.2).

La première (extension) est le *Fragment-Growing*. À partir d'un fragment-hit interagissant avec une sous-poche du site de liaison, des groupements chimiques sont ajoutés au fragment afin de lui octroyer la possibilité de faire de nouvelles interactions avec la protéine. D'un point de vue général, une bonne stratégie consiste à ajouter des groupements polaires dans un premier temps et, *in-fine*, des groupements hydrophobes afin de s'assurer de la bonne solubilité de la molécule finale. Ainsi, à chaque nouvelle modification, le fragment optimisé sera testé et évalué au travers de différentes métriques telles que la constante de dissociation (K_d) et le *Ligand-Efficiency* (LE)[7].

La deuxième stratégie (fusion) est le *Fragment-Merging*. Elle repose sur l'identification de deux fragments-hits au minimum, chacun composé d'un groupement chimique similaire. La superposition de ces deux fragments, au niveau de ce groupement similaire, permet dès lors de fusionner les deux fragments en un seul.

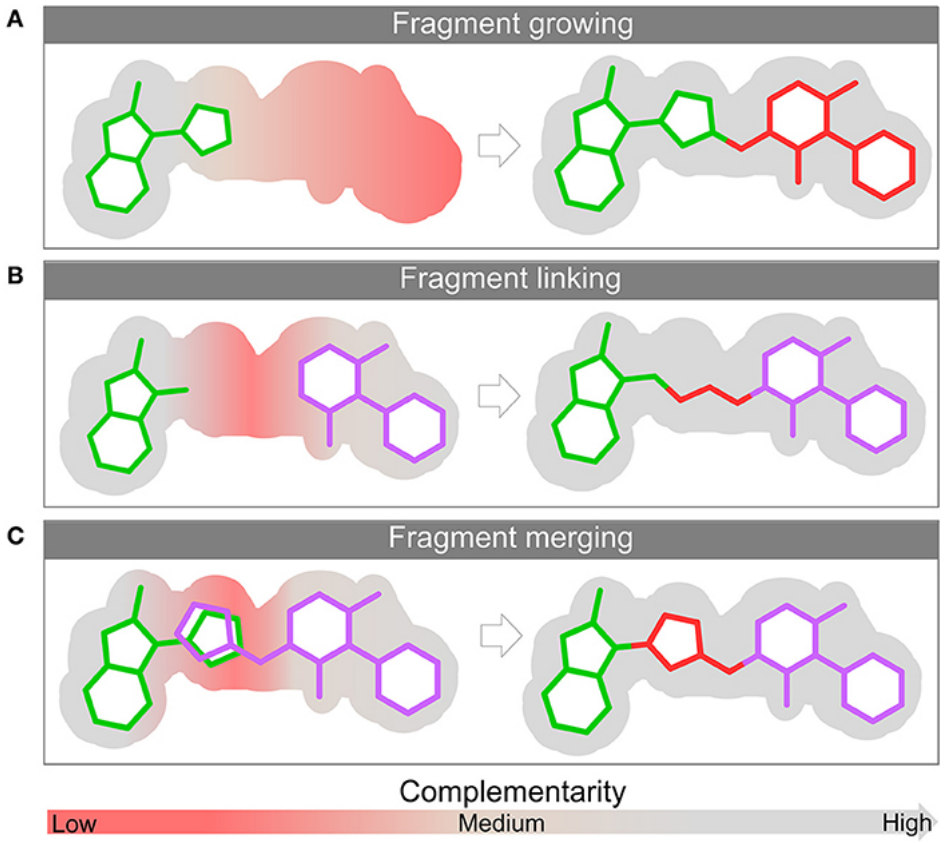


FIGURE 1.2 – Illustration des trois stratégies de l'approche par fragments (Growing, Linking et Merging), pour concevoir une molécule-lead à partir de fragments-hits pour une région de liaison d'intérêt.[10]

La troisième stratégie (liaison) est le *Fragment-Linking*. Elle repose sur l'identification, au minimum, de deux fragments-hits liés à deux sous-poches voisines du site de liaison. L'idée principale consiste à connecter ces fragments entre eux par l'intermédiaire d'un *linker* si nécessaire. Il est supposé, en théorie, que l'énergie libre de la molécule-lead soit meilleure que l'addition des énergies libres associées à chacun des fragments pris séparément, due principalement à la barrière entropique des corps rigides[9].

1.1.2 . Défis des méthodes computationnelles

Le criblage computationnel consiste à prédire une pose ou mode de liaison natif, c'est-à-dire prédire la position, l'orientation et la conformation natives d'un fragment liant une cible biologique (*e.g.* la poche catalytique d'une protéine). L'identification de cette pose native va reposer sur trois facteurs importants :

1. l'échantillonnage conformationnel du fragment ("poses"),

2. l'attribution d'un score ("scoring"),
3. la solvatation.

Algorithmes de recherche. L'échantillonage conformationnel consiste à balayer l'ensemble des modes de liaisons possibles, c'est-à-dire toutes les positions et conformations du ligand au sein de la région cible. On distingue deux méthodes d'algorithmes de recherche :

- la méthode systématique, qui repose sur la modification graduelle de mouvements de translation, rotation et torsion du ligand ; ou par le criblage d'un fragment à différentes positions et rotations dans la région d'intérêt,[11,12]
- la méthode stochastique avec, par exemple Monte Carlo, qui consiste à placer un ligand arbitrairement dans la région d'intérêt, à lui attribuer une valeur score et à générer ensuite une nouvelle configuration.[11]

Cette notion d'échantillonnage est donc très importante pour avoir des prédictions fiables. En effet, un mauvais échantillonnage ou un échantillonage insuffisant conduirait à ne pas prédire les poses natives (et donc réaliste), ce qui peut être un frein important dans le cadre du design de molécules à but thérapeutique entre autres (voir définition 1.1.3).

Fonctions de Scores. Le scoring consiste à attribuer une valeur numérique (un score) à chaque pose, et pour lequel le meilleur score correspondrait idéalement à la pose native[12]. Ce score peut être assimilé à un "pseudo-K_D", et donc en quelque sorte à l'affinité de la pose vis-à-vis de la cible biologique. On distingue différentes catégories de fonctions de score[11], parmi lesquelles :

- La fonction de score basée sur le champ de force. Le champ de force décrit un système par un ensemble d'équations et de paramètres afin de calculer les énergies qui y sont associées. L'énergie potentielle du système est définie comme la somme des énergies de liaisons liées (définies comme la connectivité correspondant aux énergies de liaisons, des angles et angles dièdres) et non-liées (définies au travers l'espace par les interactions de Van der Waals, électrostatiques et hydrogènes)[11].

Solvatation. Une étude de González-Alemán et *al*. a évalué deux paramètres, le "Docking Power" et le "Screening Power" avec le logiciel de criblage virtuel MCSS (voir la section 1.1.4), en présence (modèle `STDW`) ou absence (modèle `SCAL`) de solvatation. Le Docking Power est défini comme la capacité de la fonction de score à identifier le mode de liaison natif du ligand, tandis que le Screening Power est défini comme la capacité de la fonction de score à identifier le vrai ligand à partir d'un *pool* de molécules aléatoires.

En particulier, cette étude montre que le modèle SCAL inclut un biais dans le scoring en favorisant les purines (Adénosine et Guanosine). Le modèle STDW décrit mieux les contributions liées associées aux angles de torsions, et est ainsi meilleur à la fois sur les paramètres de Docking et Screening powers. Ainsi, le modèle STDW permet de mieux discriminer des ligands très similaires, tandis que les poses les mieux scorées reproduisent pour la plupart le mode de liaison natif[13].

Ainsi, ces trois facteurs sont à prendre en considération, et correspondent à des défis des approches computationnelles pour la précision des prédictions.

1.1.3 . Définitions et différences entre le "Design" et le "Docking"

Design. Le "design" consiste à concevoir de nouvelles molécules *de-novo*, généralement à but thérapeutique ou pouvant servir de sondes moléculaires, et interagissant avec une cible biologique (*e.g.* protéines, ARN...). Cette conception repose sur différentes stratégies (Figure 1.3) :

- La première repose sur la **connaissance de la structure 3D** d'une protéine ("récepteur" ou "cible") si elle est connue. On parle alors de la stratégie SBDD ("Structure-Based Drug Discovery"). Cette structure 3D est généralement obtenue par différentes méthodes expérimentales telles que la cristallographie à rayons X (RX) ou la résonance nucléaire magnétique (RMN). Si elle n'est pas connue, il subsiste la possibilité de créer un modèle 3D de cette structure à partir de sa séquence primaire et de la connaissance de la structure 3D de protéines homologues à cette dernière (modélisation par homologie)[14].
- En absence de connaissances 3D du récepteur, une autre stratégie repose sur la **connaissance des caractéristiques de ligands** (charges, hydrophobicité...) interagissant avec ce récepteur afin de créer un **pharmacophore**, se définissant comme la partie moléculaire responsable de l'activité biologique recherchée[14,15]. Cette stratégie est alors nommée LBDD ("Ligand-Based Drug Discovery")

L'approche par fragments peut être appliquée dans le cadre du design, couplée au SBDD, et consiste en quatre différentes étapes[16] que sont :

1. la détermination des positions optimales des fragments dans le site de liaison (*i.e.* modulant l'activité biologique de la protéine cible),
2. la connexion des fragments ensemble pour obtenir une molécule compatible avec la forme de la cavité,
3. l'estimation ou la prédiction de l'affinité selon une fonction de score,
4. la synthèse et évaluation des ligands prédits

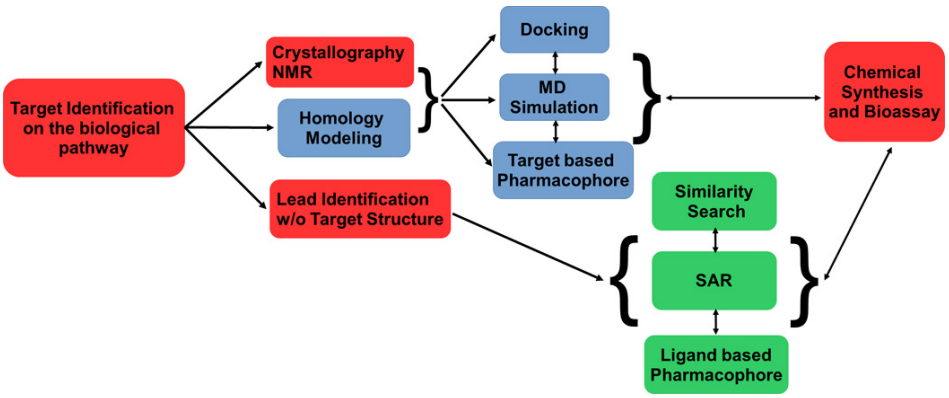


FIGURE 1.3 – Workflow pour la conception de médicaments *de-novo*. Les méthodes expérimentales correspondent aux cases rouges. Les méthodes computationnelles SBDD (reposant sur la connaissance 3D de la structure) correspondent aux cases bleues. Les méthodes computationnelles LBDD (reposant sur la connaissance de propriétés connues de ligands interagissant avec la cible biologique) correspondent aux cases vertes.[14]

Docking. Le criblage ("docking") consiste à modéliser l'interaction d'un complexe protéine/ligand, pour lequel (i) la séquence du ligand ou sa formule chimique est connue et, (ii) la structure 3D de la protéine est connue.
L'objectif est de prédire, classiquement, le mode de liaison du ligand et l'énergie d'interactions associée. En pratique, un **échantillonnage conformationnel** du ligand par rapport au récepteur est réalisé, et chaque conformère est évalué par une **fonction de score**. En général, le conformère de meilleur score prédit par le logiciel de docking est supposé correspondre au mode de liaison natif du ligand à la surface du récepteur.

L'approche par fragments est également applicable dans le cadre du docking. Elle consiste à découper le ligand en fragments, à cribler chacun de ces fragments sur la protéine, et à réassembler les meilleures poses[17]. Dans ce cadre-là, deux explorations sont donc réalisées : (i) une exploration de la conformation globale du ligand par assemblage des fragments, et (ii) une exploration de la conformation locale de chaque fragment[17].

1.1.4 . Approche par Fragments basée sur les méthodes computationnelles pour le design ou le docking d'ARNsb

Le travail de cette thèse a été réalisé dans le cadre du développement et l'application de l'approche par fragments pour modéliser les interactions de complexes Protéine/AN (Acide Nucléique). Ainsi, il est important de rappeler que la définition du fragment, considéré ici, ne correspond pas à la définition classique et théorique de la règle RO3.

Dans le cadre de cette thèse, un **fragment est défini comme un mononucléotide ou un trinucléotide** selon l'approche computationnelle utilisée.

Approche basée sur l'outil NUCLEAR couplé à MCSS pour du design d'ARNsb

L'approche NUCLEAR couplé à MCSS est une approche itérative entre deux outils bien distincts :

- MCSS ("Multi-Copy Simultaneous Search") pour le criblage virtuel de fragments sur une structure 3D pour laquelle les coordonnées sont connues[16],
- NUCLEAR ("NUCLeotide AssembleR") pour l'identification de hotspots à but de design d'ARNsb.[18].

Multi-Copy Simultaneous Search (MCSS). L'outil MCSS adresse la première étape du design, c'est-à-dire le criblage de fragments à la surface de la protéine selon un modèle tout-atome. Il va ainsi prédire les positions énergétiquement favorables des fragments dans une région de liaison déterminée, appelée la **map de fonctionnalité du site de liaison**. Ces fragments énergétiquement favorables peuvent donc être considérés comme des fragments-hits connectables entre eux. Le pipeline de MCSS est décrite dans la figure 1.4.

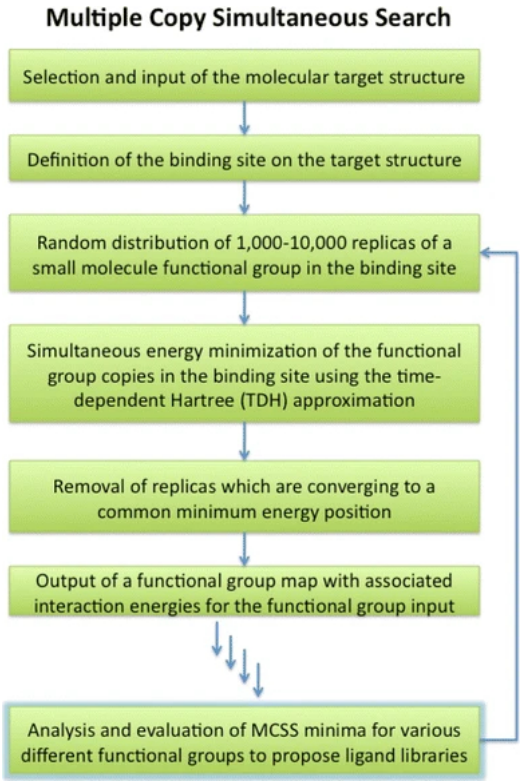


FIGURE 1.4 – Les différentes étapes de l'outil MCSS[16]

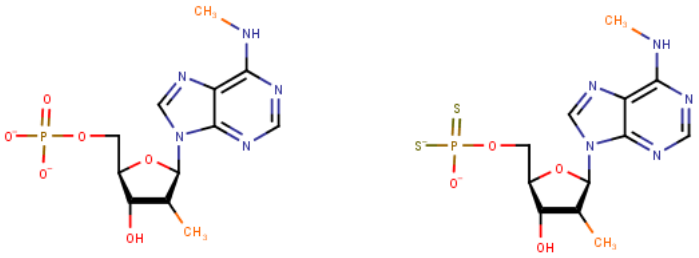


FIGURE 1.5 – Un exemple d'un nucléotide (code MMA) présent dans la bibliothèque : MMA est un dérivé de l'adénosine et présente une modification sur le ribose à la position O2' (groupement hydroxyle substitué par un méthyl), sur la base nucléique (ajout d'un groupement méthyl), en considérant à la fois le groupe phosphorothioate (PS_2) et le groupe phosphate (PO)

La première étape consiste à la construction d'une bibliothèque composée de fragments de mononucléotides. À l'occurrence, cette bibliothèque est composée de **444 groupes au total**. On dénombre :

- 5 nucléotides standards que sont : Adénosine (A), Thymidine (T)/Uradine (U), Guanosine (G), Cytosine (C)
- 106 modifications sur la base nucléique et/ou le ribose (18 modifications pour l'adénine, 29 modifications pour la guanine, 18 modifications pour la cytosine et 45 modifications pour la thymine/uracile)
- Sur ces 111 groupes, il est considéré à la fois les groupes avec une fonction phosphate ou phosphorothioate (111×2) (voir figure 1.5)
- Sur ces 222 groupes, il est considéré à la fois les configurations ANTI et SYN (définitions en section 1.4.1) (222×2).

La deuxième étape consiste à distribuer aléatoirement 1.000 ou 10.000 copies (réplicas) de chaque groupe de la librairie sur la région de liaison. Il est important de noter que **ces réplicas ne se voient pas les uns des autres**, mais voient le champ de force de la protéine. Les positions de ces différentes copies sont minimisées par un gradient conjuguant de minimisation. Tous les réplicas convergeant vers le même minima énergétique (selon un seuil de RMSD inférieur à 0.2Å) sont éliminés afin de conserver qu'une seule copie de celle-ci. Enfin, une cartographie de fonctionnalité, associée à chaque groupe, est fournie dans un fichier de sortie ("output") associée à une fonction score[16].

Cette fonction de score est définie par des contributions électrostatiques (E_{el}^{inter}) et de Van der Waals (E_{vdw}^{inter}), avec en plus une pénalité correspondant à la déviation de la conformation du fragment par rapport à son énergie minimum ($E_{conf}^{fragment}$)[13] :

$$\Delta E_{MCSS}^{binding} = \Delta E_{conf}^{fragment} + \Delta E_{vdw}^{inter} + \Delta E_{el}^{inter} \tag{1.1}$$

La contribution de van der Waals et la contribution électrostatique au score sont définies par les équations 1.2 et 1.3[13] :

$$E_{vdw} = \sum_{excl(i,j)=1} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) sw(r_{ij}^2, r_{on}^2, r_{off}^2) \tag{1.2}$$

$$E_{el} = \sum_{excl(i,j)=1}^{\epsilon=3} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}^2} sw(r_{ij}^2, r_{on}^2, r_{off}^2) \tag{1.3}$$

NUCLeotide AssembleR (NUCLEAR) pour le design. NUCLEAR est un prologiciel Python qui prédit des chaînes ARNsb à but de **design** thérapeutique.[18]
NUCLEAR prend pour données d'entrée (i) les coordonnées 3D du récepteur, et (ii) les coordonnées 3D des poses générées par MCSS. Un pré-clustering de poses peut être réalisé afin d'éliminer toutes les poses similaires selon un seuil RMSD, généralement fixé à 0.5Å.
Deux matrices sont ensuite générées :

- une **matrice de connectivité (connexion)** indiquant quelles poses sont connectables selon un critère de distance entre les atomes O3′ du nucléotide n et C5′ de la pose voisine $n+1$,
- **une matrice de clashes** (stériques) indiquant que deux poses ne peuvent être connectées.

Enfin, une recherche de séquences est réalisée pour générer des séquences d'ARNsb de taille k, avec k défini ici comme la longueur de la séquence de l'ARNsb cible (voir figure 1.6) .

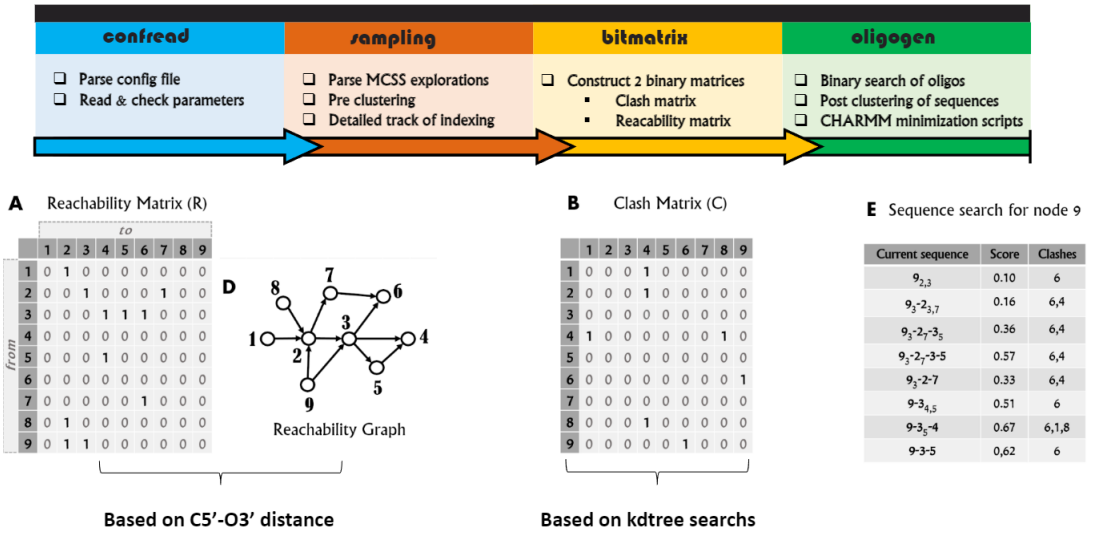


FIGURE 1.6 – Les différentes étapes de l'outil NUCLEAR pour prédire des ARNsb à la surface d'une protéine à partir du jeu de données MCSS[18]

NUCLEAR pour l'analyse de spots. De plus, NUCLEAR possède une extension pour analyser le résultat du criblage de fragments obtenu avec MCSS (voir figure 1.7), afin d'identifier ce qui

s'apparente aux *hotspots*, *warm-spots* et *specific-spots*[18]. La procédure est décrite comme suit :

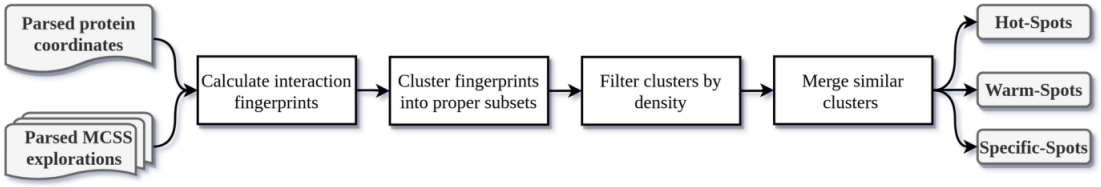


FIGURE 1.7 – Les différentes étapes de l'outil NUCLEAR pour rechercher et identifier des spots à la surface d'une protéine à partir du jeu de données fourni par MCSS[18]

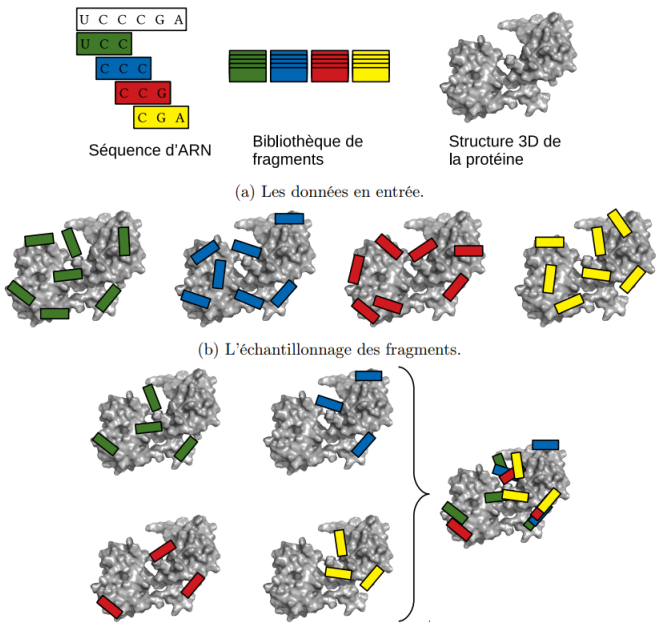
1. Calculer les empreintes ("fingerprints"), définies comme un ensemble de contacts (selon un seuil de distance généralement compris entre 3.0Å et 3.5Å) entre un réplica MCSS pour un fragment donné et la protéine.
2. Réaliser un clustering de contacts (fingerprints) à un niveau de résolution du résidu (*i.e.*, est-ce que le résidu de la protéine est impliqué dans une interaction avec un réplica ?).
3. Filtrer les clusters selon le critère de densité, défini comme le nombre de réplicas divisé par le nombre unique de résidus appartenant à ce cluster.
4. Fusionner les clusters ayant une certaine valeur de recouvrement par rapport à un seuil de l'index de Tanimoto pré-défini par l'utilisateur. En output, différents fichiers TSV sont générés avec les informations suivantes : le numéro d'identification du spot, le nombre de poses et de groupes chimiques constituant ce spot, la densité et les résidus de la protéine constituant ce spot.

Approche basée sur l'outil ssRNATTRACT pour du docking d'ARNsb

La méthode ssRNATTRACT est un protocole itératif pour modéliser le mode de liaison d'un ARNsb à la surface d'une protéine. Elle a été notamment développée pour réaliser du **docking** d'ARNsb sur une protéine (*i.e.* reproduire le mode de liaison natif d'un ARNsb expérimental à la surface d'une protéine).
Trois éléments sont principalement donnés en entrée : i) les coordonnées 3D de la protéine cible résolue par RMN ou par RX, ii) une bibliothèque constituée de fragments de trinucléotides et iii) la séquence de l'ARNsb recherchée.

Le protocole repose sur trois étapes (Figure 1.8) :

- la construction de bibliothèques de fragments trinucléotidiques,
- le docking de ces fragments à la surface de la protéine,



(a) Les données en entrée.

(b) L'échantillonnage des fragments.

FIGURE 1.8 – Approche illustrée de ssRNATTRACT[19]

- l'assemblage des fragments selon un critère de recouvrement de fragments chevauchants.

Création de la bibliothèque avec ProtNAff Le choix, ici, qu'un fragment soit un trinucléotide au lieu d'un mononucléotide comme dans le cas de MCSS, résulte de la stratégie du fragment-merging de ssRNATTRACT. En effet, le choix des trinucléotides facilite la connexion des fragments et limite les conformations potentiellement défavorables en termes de torsions entre fragments reconnectés, car les conformations des trinucléotides correspondent à des conformations canoniques rencontrées dans les structures d'ARN.

Ainsi, un effort particulier est la construction de librairies de conformations représentatives pour couvrir idéalement l'espace conformationnel avec les 4 nucléotides standards dont l'ensemble des combinaisons n'est pas complètement représenté dans les structures expérimentales disponibles pour les régions simple-brin. La bibliothèque de fragments va donc contenir l'ensemble des séquences et conformères possibles associés à chaque séquence afin d'augmenter les chances de prédire correctement une pose native. La génération automatisée de cette bibliothèque de fragments est réalisée par l'outil ProtNAff à partir de la base de données PDB[19].

Il est important de noter ici que le modèle utilisé pour modéliser les fragments et la protéine sont du modèle gros-grain afin de lisser la surface énergétique du complexe ARNsb-protéine. Autrement dit, 3 à 4 atomes lourds en fonction de la nature du nucléotide sont remplacés par un pseudo-atome, donnant respectivement un total de 7 et 8 pseudo-atomes pour les purines et les pyrimidines et tolérant ainsi une imprécision sur le positionnement atomique (Figure 1.9).

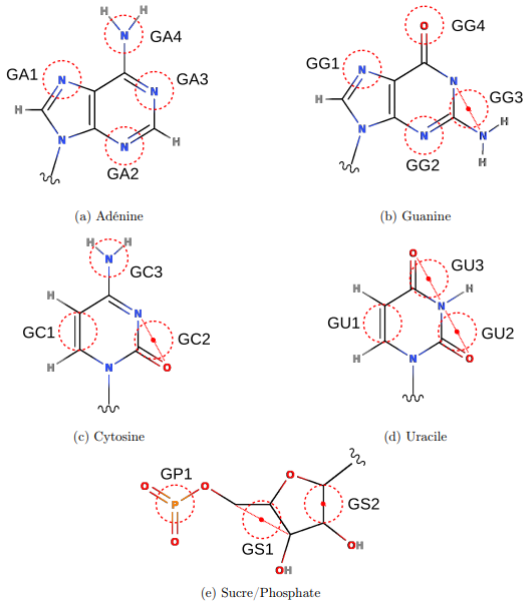


FIGURE 1.9 – Illustration du modèle gros-grain utilisé par ssRNATTRACT[17]

Criblage de fragments avec ATTRACT. Le criblage de fragments repose sur la séquence recherchée et sur le découpage de ces derniers en trinucléotides, afin d'amarrer uniquement les fragments trinucléotidiques correspondant. Ainsi, une séquence de n nucléotides correspond à une séquence de n-2 trinucléotides (*e.g.*, AUGGU → AUG-UGG-GGU). Ce docking repose sur un docking rigide avec un échantillonnage réalisé selon deux stratégies au choix nommées **Randsearch** (placement aléatoire de fragments) ou **Systsearch** (placement de fragments à des positions régulières).
Chaque pose est ensuite minimisée et classée selon la fonction énergétique d'ATTRACT : les poses redondantes à un seuil de 0.2Å sont éliminées, tandis que les 10^6 poses de meilleurs scores (tout conformère confondu) sont conservées[17]. La fonction énergétique, ici, correspond à la somme des énergies d'interactions individuelles de toutes les paires de pseudo-atomes, calculée selon le potentiel de type Lennard-Jones avec une composante électrostatique implicite (Figure 1.10) :

Fait important, dû au modèle gros-grain utilisé lors du criblage, les clashes entre les fragments et la protéine sont tolérés contrairement à l'approche MCSS où les clashes entre les fragments et la protéine ne sont pas tolérés.

Assemblage des fragments. L'assemblage des fragments repose sur la création d'un graphe dirigé de connectivité et pour lequel une définition sera donnée à la section 1.3. Il est considéré que deux fragments sont connectables (*i.e.* fusionnable selon l'approche Fragment-Merging) si leurs séquences

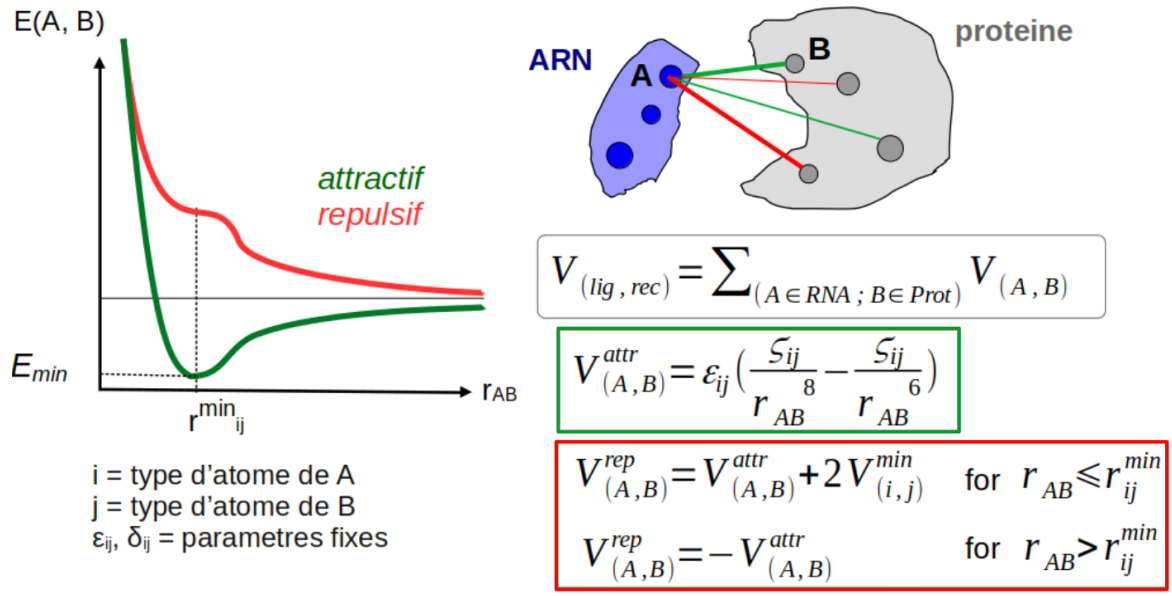


FIGURE 1.10 – Fonctions de scores utilisées dans ssRNATTRACT. (Figure issue du Manuscrit HDR de Isaure Chauvot de Beauchêne)

sont compatibles avec la séquence recherchée, et si les pseudo-atomes de leurs nucléotides communs ont une distance en dessous d'un seuil d'*overlap* fixé, typiquement de 2Å. La conversion d'une chaine de trinucléotides en une chaine mononucléotidique se fait par une moyenne des coordonnées des atomes chevauchants.
De plus, l'assemblage peut être restreint par la notion **d'ancres** si existante, c'est-à-dire par la présence de poses conservées[20] ; ce qui permet ainsi de contraindre la direction et la géométrie lors des connexions, de réduire la combinatoire et d'augmenter les chances de prédire la solution expérimentale native.

ssRNATTRACT a été évaluée sur 9 complexes (RRM (Motif de reconnaissance de l'ARN) : `1B7F, 1CVJ, 2MGZ, 2YH1, 3NNH, 4BS2, 4N0T` ; PUF (Domaine Pumilio) : `3BX3, 5BZV`). Des solutions reproduisant le mode de liaison natif ont été retrouvées à haute résolution ($\leqslant$ 2.0Å) pour 4 de ces protéines dans le Top10, et pour 5 de ces protéines dans le Top100, conduisant ainsi à une bonne prédiction du mode d'interaction des ARNsb[20].

Cette méthodologie démontre donc une bonne efficacité dans la prédiction du mode d'interaction d'une solution native, et donc pour du **docking d'ARNsb**. Dans une perspective de design utilisant des nucléotides modifiés, la construction de libraries trinucléotidiques intégrant une liste substantielle de modifications serait plus problématique.

	MCSS/NUCLEAR	**ssRNATTRACT**
Modèle	Tout-Atome	Gros-Grain
Nature du Fragment	Mono-nucléotides	Tri-nucléotides
Application	Design	Docking
Clashes Autorisés	Aucun	Entre les poses et le récepteur dû au modèle gros-grain
Connectivité	Critère de distance entre le O3′ de la pose n et le C5′ de la pose $n + 1$	Chevauchement basée sur la RMSD entre les nucléotides 2 et 3 de la pose n et les nucléotides 1 et 2 de la pose $n + 1$

TABLE 1.1 – Différences entre les approches basées sur MCSS/NUCLEAR et ssRNATTRACT appliquées dans le cadre de cette thèse

1.2 . Notions d'algorithmique fondamentale

En informatique, la notion de **complexité** d'un algorithme vise à expliciter la relation entre la taille (d'encodage) n de la donnée en entrée, aussi appelée **instance**, et son temps d'exécution. Caractériser la complexité d'un problème consiste à appréhender la difficulté du problème, et de déterminer à l'avance le temps de calcul et la mémoire utilisée par un algorithme pour résoudre ce problème. En effet, si la résolution d'un problème bioinformatique nécessite une quantité de mémoire importante ou un temps de calcul non-raisonnable, alors nous pouvons soit nous demander s'il est pertinent de créer un algorithme pour résoudre un tel problème, soit y réfléchir pour trouver une solution. Ainsi, la caractérisation de la complexité du problème permet d'avoir une vue élargie du problème associé, et guide la recherche d'un algorithme adapté.

1.2.1 . Complexité classique

Un modèle classique restreint cette analyse au cas le pire, et s'intéresse au plus long temps d'exécution observé pour une instance de taille donnée. Il s'agit alors de trouver une fonction $f(n)$ qui représente un bon ordre de grandeur pour le temps d'exécution $c(n)$ le plus long sur une instance n, ce pour toute valeur de n. Pour être plus précis, on introduit les notations suivantes :

- Algorithme en $\mathcal{O}(f(n))$: Le temps d'exécution réel $c(n)$ pour l'instance la plus chronophage est borné par :
$$\lim_{n\to+\infty} \frac{c(n)}{f(n)} < \infty;$$
- Algorithme en $\Theta(f(n))$: Le temps d'exécution réel $c(n)$ pour l'instance la plus chronophage est, à une constante près, tel que :
$$0 < \lim_{n\to+\infty} \frac{c(n)}{f(n)} < \infty.$$

La complexité d'un algorithme est dite polynomiale si, pour un entier i, elle est en $\mathcal{O}(n^i)$, c'est-à-dire que son exécution nécessite au pire n^i opérations élémentaires. Si, par ailleurs, sa complexité est en $\Theta(n^i)$, alors le temps de calcul nécessaire à son exécution sur un processeur simple croit alors de façon proportionnelle à n^i.

Classes de problèmes (P vs NP). La plupart des problèmes algorithmique, et en tout cas l'intégralité de ceux considérés au sein de cette thèse, appartiennent à deux classes :

- La classe P, pour "Déterministe Polynomiale", comprend des problèmes admettant des algorithmes de complexité polynomiale sur une machine déterministe (*e.g.* ordinateur domestique) ;
- La classe NP, pour "Non Déterministe Polynomiale", comprend les problèmes pour lesquels il

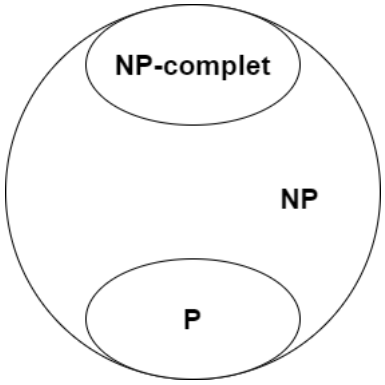


FIGURE 1.11 – Schéma d'illustration des différentes classes de complexité

est possible de vérifier la validité d'une solution en temps polynomial.

Intuitivement, la classe NP regroupe des problèmes qui pourraient être résolus en temps polynomial si la phase d'exploration (explosion combinatoire) des solutions potentielles pouvait être distribuée sur un nombre infini de machines, et qu'un nombre exponentiel de possibles pouvait ainsi être répartis en temps linéaire, puis vérifiés en temps polynomial.

En particulier, tout problème dans P est aussi dans NP, et nous avons donc $\mathsf{P} \subseteq \mathsf{NP}$. En revanche, nous ne savons pas, à l'heure actuelle, s'il existe des problèmes propres à NP, c'est-à-dire $\mathsf{P} \neq \mathsf{NP}$.

De la décision à l'optimisation. Pour des raisons techniques, ces classes de complexités concernent des problèmes formulés comme des problèmes de décision, c'est-à-dire pour lesquelles la réponse est réduite à "Oui" ou "Non". Par exemple, le problème de décider si un entier naturel appartient à une liste non triée est un problème de décision, qui peut être résolu en temps $\mathcal{O}(n)$ (linéaire) par une simple boucle sur les éléments de la liste, et appartient donc à la classe P.

En revanche, les problèmes d'optimisation peuvent toujours être ramenés, au moins en théorie, à des problèmes de décision, en ramenant une question type :

$\mathcal{P}$: Quel est le coût minimal d'une solution respectant des contraintes C ?

à une formulation typique des problèmes de décision :

$\mathcal{P}'$: Existe-t-il une solution de coût supérieur à K respectant des contraintes C ?

S'il existe un algorithme polynomial pour le problème de décision $\mathcal{P}'$, et sous des hypothèses raisonnables sur le coût maximal (*e.g.* borné par $n!$), alors la recherche dichotomique d'une valeur optimale $K^\star$ (au-delà de laquelle la réponse à $\mathcal{P}'$ devient "Non") résulte en un algorithme polynomial pour $\mathcal{P}$.

Les problèmes NP-complets. Parmi les problèmes NP, un sous-ensemble important est constitué par les problèmes NP-complet, ceux dont la résolution en temps polynomial permettrait une résolution en temps polynomial de tous les problèmes de NP. Un exemple historique de problème NP-complet est le problème de Satisfaisabilité (3SAT) qui consiste à trouver des affectations (Vrai ou Faux) à des variables logiques (x_1, x_2 ...) telles qu'un ensemble de formules conjonctives, chacune impliquant trois variables (*e.g.* $x_1 \vee \overline{x_2} \vee x_3$), soient toutes simultanément interprétées à vrai. Autrement dit, si on considère une liste de membres où chaque membre fait part d'une liste d'exigence : existe-t-il, oui ou non, une solution pour satisfaire au moins une des exigences de chacun des membres ?

Il n'existe pas à ce jour d'algorithme capable de résoudre 3-SAT en temps polynomial, et le problème a même été démontré NP-complet. Néanmoins, si une solution est proposée, alors il est possible de vérifier sa validité en temps polynomial : le problème 3-SAT est donc aussi dans NP. Le problème est, en un sens, exemplaire au sein de la classe NP, et c'est cette exemplarité que capture la notion de problème NP-complet.

Concrètement, un problème est NP complet s'il est à la fois :

- NP-difficile, c'est-à-dire au moins aussi dur que tous les problèmes de NP. En pratique, montrer la NP-difficulté d'un problème nécessite de montrer que sa résolution en temps polynomial implique l'existence d'un algorithme polynomial pour (au moins) un problème NP-complet de référence. L'existence d'un tel algorithme impliquerait alors une résolution en temps polynomial pour tous les problèmes NP, puis pour tous les problèmes de NP ;
- Dans NP, c'est-à-dire qu'il est possible de tester la validité d'une solution proposée en temps polynomial.

Notons que la théorie de Cook-Levin[21,22] montre que tout problème NP peut être ramené au problème SAT : si on trouve un algorithme capable de résoudre le problème SAT en temps polynomial, alors tout problème NP peut être résolu en temps polynomial.

Sous l'hypothèse, généralement admise en algorithmique, que $\mathsf{P} \neq \mathsf{NP}$, alors un problème prouvé NP-complet ne peut pas admettre d'algorithme en temps polynomial dans le cas au pire.

1.2.2 . Algorithmique et complexité paramétrée

Pour des problèmes motivés par la pratique, le statut de problème NP-complet représente souvent un point de branchement, où l'algorithmicien doit typiquement choisir entre limiter ses ambitions (résolution inexacte, approchée ou heuristique) et assumer un coût dépendant du problème, et exponentiel dans le pire cas (algorithmes faiblement exponentiels, programmation mathématique ...). Au sein de cette deuxième catégorie, le paradigme de la **complexité paramétrée** fournit une approche, pratique

et élégante, pour une résolution exacte des problèmes NP-complets. Elle repose sur le constat que, pour de nombreuses applications concrètes, les instances considérées sont plus contraintes que ne le laisse supposer la formulation du problème algorithmique. En particulier, certaines caractéristiques des instances, appelées **paramètres**, peuvent être contraintes à des valeurs beaucoup plus faibles que celles induites par le cas au pire. Il suffit donc d'assumer une explosion combinatoire de la complexité sur le(s) paramètre(s), et de viser une complexité polynomiale en n, afin d'obtenir un algorithme efficace sur les données réelles, associés à des faibles valeur de(s) paramètre(s).

C'est souvent le cas en bioinformatique, où de nombreux problèmes peuvent être formalisés comme des problèmes sur les graphes. La plupart de ces problèmes sont difficiles dans le cas au pire, mais les cas difficiles coincident alors souvent avec des instances complexes (*e.g.* denses, c'est-à-dire $|E| \in \Theta(n^2)$), alors que les instances concrètes ressemblent davantages à des arbres. Pour cette raison, de nombreuses méthodes en génomique comparative[23] et bioinformatique structurale[24] se basent sur le paramètre de **largeur arborescente**, prenant de faibles valeurs sur des graphes quasi-arborescents, et obtiennent des algorithmes polynomiaux quand celle-ci est bornée.

À un niveau théorique, l'approche de complexité paramétrée consiste à concevoir des algorithmes dont la complexité s'exprime en fonction de la taille n de l'instance I, mais aussi un paramètre $k = f(I)$ de cette instance tel que, si k est borné par une constante, alors le problème peut être résolu efficacement en temps polynomial. On distingue au moins deux niveaux de complexité :

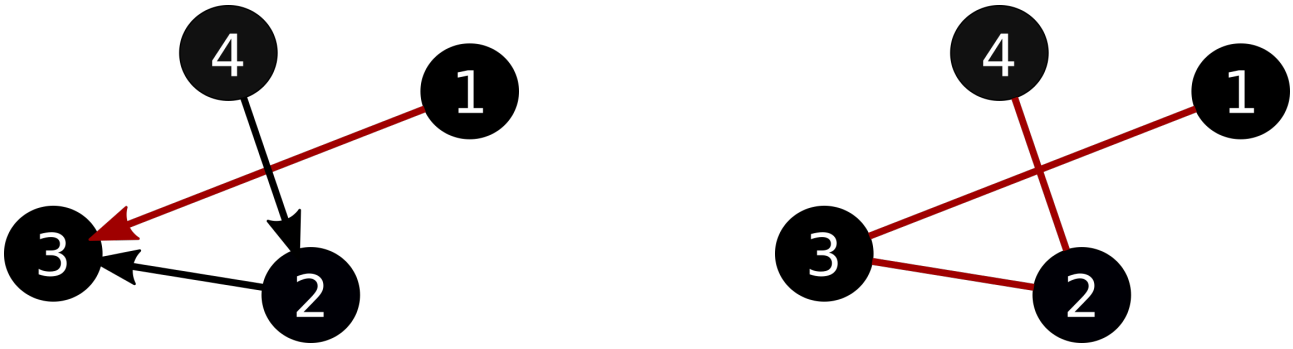
- les algorithmes **FPT** (Fixed Parameter Tractable) possèdent une complexité de la forme $\mathcal{O}(f(k) \times P(n))$, où P est un polynome de degré indépendant de k ;
- les algorithmes **XP** (slicewise polynomial), dont la complexité est en $\mathcal{O}(n^{f(k)})$.

L'algorithmique paramétrée possède aussi des limites, matérialisées par la W **hierarchie**. L'appartenance d'un problème/paramètre à $W[0]$ implique l'existence d'un algorithme FPT. En revanche, l'existence d'un algorithme FPT pour un problème/paramètre difficile pour la classe $W[1]$ (ou $W[2]$ ...) rendrait fausse l'**hypothèse de temps exponentiel (ETH)**, une hypothèse classique en algorithmique. Celle-ci prétend que le problème 3-SAT ne peut être résolu en temps $2^{\mathcal{O}(n)}$. Elle implique en particulier que $P \neq NP$, et qu'il n'existe pas d'algorithme pour le problème qui soit à la fois exact et s'exécute en temps sousexponentiel sur n.

1.3 . Algorithmique des Graphes

Un graphe constitue une représentation abstraite de connexions ou de relations d'un ensemble d'objets pour modéliser, ou abstraire, un objet du monde réel. La notion de graphe se prête à la définition de problèmes algorithmiques en des termes qui ne nécessitent plus de tenir compte de la réalité

sous-jacente à sa définition.



(a) Graphe dirigé (b) Graphe non-dirigé

FIGURE 1.12 – Exemple d'un graphe dirigé et d'un graphe non-dirigé. Un exemple de chemin/chaine, avec pour point de départ le noeud 1, est illustré en rouge, pour chaque cas

Tout d'abord, définissons intuitivement ce qu'est un graphe. Un graphe est un modèle de connectivité constitué d'un ensemble de sommets et d'arêtes reliant ces sommets. Nous distinguons deux types de graphes : (i) les graphes dirigés ; et (ii) les graphes non-dirigés (Figure 1.12) et pour lesquels nous proposons les définitions formelles suivantes :

Definition 1.3.1 (Graphe dirigé): Un graphe dirigé $G = (V, E)$ est un graphe constitué de V noeuds/sommets, et de $E \subseteq V \times V$ arcs dirigés, *i.e.* des paires ordonnées $(v_i, v_j) \in V \times V$ de sommets exprimant la possibilité d'une transition $v_i \rightarrow v_j$.

Certaines relations d'adjacence étant naturellement symétriques, une version non-dirigée des graphes est aussi parfois considérée, et définie comme suit.

Definition 1.3.2 (Graphe non-dirigé): Un graphe non-dirigé $G = (V, E)$ est un graphe constitué de V noeuds/sommets, et de $E \subseteq V^2$ arêtes, *i.e.* de paires non-ordonnées $\{v_i, v_j\}$ de sommets, connectant v_i à v_j sans notion de direction.

Chemins et chaînes. Un graphe peut contenir différents motifs, dont l'exemple le plus simple est celui des **chemins et chaines** (Figure 1.12). De façon informelle, un chemin (ou une chaine) peut être vu comme une suite de sommets empruntés successivement. L'emploi de l'un ou l'autre des termes dépend de la nature du graphe, respectivement dirigé ou non-dirigé. De même, nous proposons les définitions formelles suivantes :

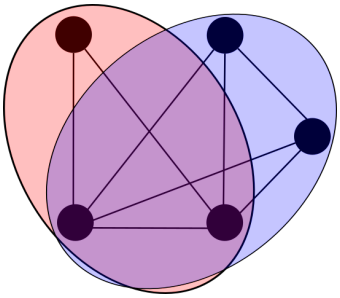


FIGURE 1.13 – Exemple des cliques dans un graphe $G(V, E)$

Definition 1.3.3 (Chemin): Étant donné un graphe dirigé $G = (V, E)$, on appelle **chemin** une séquence de sommets $(v_1, \ldots, v_k)$ telle que, pour tout $i \in [1, k-1]$, $(v_i, v_{i+1}) \in E$ est un arc dans G.

Une **chaine** est définie de façon similaire pour les graphes non-dirigés, la seule différence ayant trait au caractère non-ordonné des arêtes $\{v_i, v_{i+1}\} \in E$.

Le concept de chemin est central au problème du **voyageur de commerce**, un problème classique en algorithmique des graphes : Etant donné des distances inter-noeuds, il consiste à trouver le chemin le plus court permettant de passer exactement une fois par chacun des noeuds du graphe. Cette formulation peut alors être adaptée à un réseau routier reliant différentes villes. On considérera alors un graphe noté $G = (V, E)$, où V est le nombre de noeuds du graphe (ici associés aux villes), et E le nombre d'arêtes (ici associés aux chemins). Une solution au problème du voyageur de commerce fournit alors une tournée véhiculaire permettant de parcourir, à coût minimal, un ensemble de destinations sans jamais revenir sur ses pas.

Il n'existe à l'heure actuelle aucun algorithme permettant de le résoudre de façon garantie exacte et efficace. Pire, ce problème fait partie d'une famille de problèmes dits NP-complets qui, sous certaines hypothèses, n'admettent aucun algorithme efficace[25] (cf section précédente pour plus de détails).

Cliques. Les sous-ensembles de noeuds fortement connectés représentent un autre motif d'intérêt au sein des graphes, et sont à l'origine de problèmes algorithmiques difficiles et intéressants. Plus précisément, une **clique** représente un sous-ensemble, ou sous-graphe, d'un graphe donné comme illustré dans la Figure 1.13, dont la définition formelle est la suivante :

Definition 1.3.4 (Clique): Une clique d'un graphe $G = (V, E)$ est un sous-ensemble $V' \subseteq V$ de sommets tous connectés deux-à-deux.

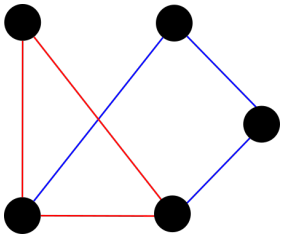


FIGURE 1.14 – Schéma d'un chemin auto-évitant et non-auto-évitant. Soit le graphe $G(V, E)$. Deux chemins sont représentés : un chemin auto-évitant (en bleu) passant par un noeud unique, et un chemin non-auto-évitant (en rouge) repassant deux fois par le même noeud

Le problème consistant à trouver une clique maximale dans un graphe est NP est NP complet, et W[1]-difficile au regard de la complexité paramétrée.

Ensembles stables (indépendants). Étant donné un graphe $G(V, E)$, un **ensemble stable** est défini comme un sous-ensemble de noeuds V pour lesquels les noeuds ne sont pas adjacents entre eux. On dit aussi que cet ensemble est **indépendant**.

Definition 1.3.5 (Ensemble stable/indépendant): Étant donné un graphe $G = (V, E)$, un ensemble de sommets $S \subseteq V$ est dit **stable** (alt. **indépendant**) si, pour tout $u \neq v \in S^2$, $(u, v) \notin E$.

Un problème classique d'optimisation est le problème du **stable maximum** ("Maximum Independent Set") qui est défini comme la recherche du stable de cardinalité maximum. Il est NP-complet, et W[1] difficile pour $k := |V|$ du point de vue de la complexité paramétrée, y compris sur les instances géométriques (*e.g.* graphes issus d'intersections de segments/disques)[25,26]

1.3.1 . Chemins auto-évitants (k-chemins)

Étant donné un graphe $G(V, E)$, un chemin est dit **auto-évitant** de taille k, aussi appelé un k**-chemin** s'il ne visite pas deux fois un même noeud du graphe (voir figure 1.14).

Definition 1.3.6 (k-chemin): Étant donné un graphe $G(V, E)$, un k-chemin (ou chemin auto-évitant) est un chemin qui passe par une suite de sommets $v_0, \ldots, v_k$ une seule et unique fois.

L'existence, dans un graphe donné, d'un k-chemin est un problème combinatoire de référence, qui peut être formalisé comme suit.

Problème d'existence d'un k-chemin
Entrée : Graphe dirigé $G = (V, E)$; Entier $k \in \mathbb{N}^+$.
Sortie : Vrai s'il existe un k-chemin, *i.e.* une séquence de noeuds $c := (v_1, \ldots, v_k) \in V^k$ tels que :

- Les noeuds consécutifs de c sont connectés : $\forall i \in [1, k-1], (v_i, v_{i+1}) \in E$;
- Les noeuds sont deux-à-deux distincts : $\forall i \neq j, v_i \neq v_j, \forall i \neq j$;

Faux sinon.

Quand $k = n$, un k-chemin passe une et une seule fois par l'ensemble des noeuds de V, et on appelle alors un tel chemin **chemin Hamiltonien** du graphe G. La recherche d'un tel chemin est un problème complexe à résoudre exactement en temps raisonnable, hormis pour des valeurs limitées de n. En particulier, une recherche *force brute* d'un n-chemin (Hamiltonien) dans le contexte d'une (quasi)-clique pourrait nécessiter la considération d'un nombre de chemins en $\mathcal{O}(n!)$, c'est à dire de l'ordre de $(n/e)^n$. Ainsi, un graphe contenant 71 noeuds contiendrait 5.10^{99} k-chemins pour $k = n$. Ainsi, rechercher un k-chemin revient à une complexité NP-complet. Une façon de résoudre exactement ce problème est d'utiliser une approche de complexité paramétrée FPT en utilisant une technique nommée le **color-coding**.

Résolution paramétrée par Color coding

La technique du **color-coding** a été introduite par Alon, Yuster et Zwick[27]. Elle a été utilisée dans le contexte de la Bio-informatique pour rechercher et compter les occurrences de motifs dans les réseaux biologiques[28–30].

La principale clé du color-coding est d'associer une **coloration** $\kappa : V \to [1, k]$ à un graphe donné en entrée $G = (V, E)$, et remplace la recherche (difficile) pour un chemin (ou motif) de longueur k (k-chemin) avec une recherche (facile) d'un chemin colorié, utilisant chacune des k couleurs exactement une seule fois. Les chemins coloriés peuvent être optimisés et comptés en un temps linéaire sur $n + |E|$, et uniquement exponentiel sur k. Il est clair que tout chemin colorié est également un k-chemin puisqu'il utilise des sommets distincts. Ainsi, pour une seule coloration, l'ensemble des chemins coloriés est uniquement un sous-ensemble de k-chemins. Cependant, pour une coloration κ donnée, plusieurs k-chemins existant x dans G peuvent ne pas être coloriés dans une coloration uniforme aléatoire, un évènement avec probabilité $\mathbb{P}(p \text{ non-colorié}) = 1 - \frac{k!}{k^k}$.
Pour améliorer les chances de trouver un chemin k, la méthode itère la recherche de chemins coloriés dans un ensemble de α colorations aléatoires uniformément distribuées de G. En supposant des colorations uniformes $\kappa_1, \ldots, \kappa_\alpha$, la probabilité d'atteindre un k-chemin après α coloriages est exactement

de

$$\mathbb{P}(\text{p colorié pour plusieurs colorations} \mid \alpha) = 1 - \left(1 - \frac{k!}{k^k}\right)^{\alpha}. \tag{1.4}$$

Par ailleurs, le nombre attendu de colorations aléatoires nécessaires avant de trouver x (c'est-à-dire x colorié par rapport à au moins une des colorations) est donnée par $k^k/k! \in \mathcal{O}(\sqrt{k}\, e^k)$ en utilisant la formule de Stirling.

On peut aussi utiliser la **dérandomisation** pour dériver de cette approche en un algorithme efficace, déterministe et exact. Pour cela, on doit construire une famille de colorations qui, pris ensemble, représente tous les k-chemins du graphe de départ. Naor et *al.*[31] proposent une construction explicite pour une telle famille, consistant en $e^k k^{\mathcal{O}(\log k)} \log n$ colorations qui couvrent toutes les occurrences possibles de k-chemins. L'itération de la recherche de chemins coloriés optimaux sur la famille permet alors d'obtenir un algorithme exact en temps global $\mathcal{O}((2e)^k k^{\mathcal{O}(\log k)}(n + |E|) \log n)$, utilisant $\mathcal{O}(2^k n)$ mémoire.

1.3.2 . Application de la théorie des graphes aux approches par fragments

Nous pouvons retrouver des exemples pour lesquels la théorie des graphes a été appliquée aux approches par fragments, par exemple les travaux de Sarfaraz et *al.* ou les travaux de De Beauchene et *al.*[20]. Ces travaux ont tous deux porté sur du Docking de molécules chimiques et d'ARNsb respectivement, et pour lesquels la méthodologie est différente. Le docking par fragments est défini dans ces deux études comme (i) la construction d'une bibliothèque de fragments issue de la fragmentation de ligands (au travers par exemple de l'algorithme "BRICS"), et (ii) le criblage virtuelle de cette bibliothèque sur la région d'intérêt de la biomolécule (par exemple, la poche de liaison d'une protéine) (Section 1.1.3 pour une définition plus complète).

Docking de molécules chimiques. Sarfaraz et *al.* ont proposé un "benchmark" sur des complexes protéines-ligands avec différentes familles protéiques, telles que par exemple les enzymes Cytochromes P450, les récepteurs couplés aux protéines G (GPCR) ou les Bromodomaines[32]. Leur méthode peut être résumée comme suit.

Tout d'abord, des ligands sont fragmentés en différents fragments rigides avec l'algorithme BRICS ("Breaking Retrosynthetically Interesting Chemical Substructures"), de manière rétro-synthétique afin d'éviter que deux fragments aient des sous-structures similaires. On définit une sous-structure comme des parties ou fonctions chimiques du ligand (phényl, amide, ester...). Ainsi, un fragment peut correspondre à des cycles aromatiques ou non-aromatiques, ou à des chaînes (au sens chimique du terme) de type "C-C", "C-OH", "C=O", ou encore "C-C-C=O" par exemple. Cela constitue une

bibliothèque de fragments, qui est criblée dans la région de liaison de la protéine cible.

Par la suite, un graphe de connectivité $G(V, E)$ est créé selon le raisonnement suivant (Figure 1.15). Deux ensembles, Q et T sont définis et correspondent respectivement aux poses de fragments criblés (une pose est définie comme un fragment avec une orientation et une position) et aux sous-structures chimiques du ligand. Le graphe auxiliaire (graphe de compatibilité) est défini comme la compatibilité des éléments dans les ensembles Q et T, pour lequel les noeuds correspondent à la combinaison de deux éléments associée avec le même type chimique :

$$Vertex(v_{ik}) = (q_i, , , t_k), q_i \in Q, t_k \in T, \text{ and } \Phi(q_i) = \Phi(t_k) \tag{1.5}$$

avec la fonction Φ décrivant le type chimique de la sous-structure, et les arêtes comme la connexion/liaison chimique entre deux noeuds. Ainsi, comme illustré à la Figure 1.15, un noeud correspond à une sous-structure similaire entre un fragment et une sous-structure du ligand.

Par la suite, des cliques maximales du graphe de compatibilité G sont recherchées. Une clique maximale est définie ici comme une clique qui inclut le maximum d'arêtes sans violer la connexion complète requise. Par exemple, dans la Figure 1.15, une clique maximale est de taille 3 car trois poses peuvent être alignées aux trois sous-structures du ligand. Enfin, chaque clique maximale est transformée en une structure 3D.

Ainsi, cette application montre un exemple de la théorie des graphes basée sur l'approche par Fragments, selon la stratégie Fragment-Linking consistant à connecter des fragments entre eux pour obtenir une molécule finale.

Docking d'ARNsb. Sur un principe similaire, De Beauchêne et *al.* ont proposé un benchmark pour du docking d'ARNsb sur des protéines. Une description générale de leur méthodologie est décrite en Section 1.1.4. Dans cette partie, nous re-décrirons très succinctement certaines parties du protocole, et nous nous focaliserons sur la définition du graphe de connectivité $G(V, E)$.

Comme dans l'exemple précédent, une bibliothèque de fragments est construite à partir de la défragmentation d'ARNsb issue de complexes ARN/Protéine de la base de données PDB. Rappelons qu'un fragment est défini dans cette étude comme un trinculéotide. Une sous-bibliothèque est ensuite criblée sur la région d'intérêt de la protéine, et pour laquelle la sous-bibliothèque contient uniquement des fragments respectant la séquence de l'ARNsb recherchée. Ainsi, par exemple, pour un ARN de séquence cible AUGCU, les fragments appartenant à la sous-bibliothèque sont AUG, UGC et GCU.

L'ensemble des fragments criblés est ensuite transformé en un graphe de compatibilité $G(V, E)$, dans lequel un noeud correspond à un fragment, et une arête comme la connexion de deux fragments se chevauchant par l'intermédiaire des deux derniers nucléotides du fragment n et des deux premiers nucléotides du fragment $n + 1$ (Figure 1.16).

À partir de ce graphe, des chaînes de taille k sont prédites et sont ensuite converties en 3D afin de prédire le mode de liaison natif associé à la séquence d'ARN cible.

Ainsi, cette deuxième application montre un exemple de la théorie des graphes basée sur l'approche par Fragments, selon la stratégie Fragment-Merging, appliquée cette fois-ci sur les acides nucléiques.

D'autres applications. Bien que nous ayons fait un zoom de la théorie des graphes appliquée aux approches par Fragments, d'autres applications peuvent être retrouvées dans les domaines de la Chémo-informatique et de la Bio-informatique. Notamment, nous pouvons citer le problème "Trouver MCS (Maximum Common Substructures)", qui est un problème NP-complet et important dans des projets de découverte de médicaments.
Le problème MCS consiste à montrer des relations de similarité moléculaire par des caractéristiques sous-structurelles communes de deux graphes chimiques ou plus[33]. Il existe deux sous-problèmes MCS : (i) Trouver le plus grand fragment unique entre deux composés alignés (MCS connecté), (ii) Trouver l'ensemble des fragments qui maximise le nombre d'atomes ou de liaisons (MCS déconnecté) (Figure 1.17). Plusieurs algorithmes ont été proposés afin de résoudre ce problème de façon exacte ou approximative[33].

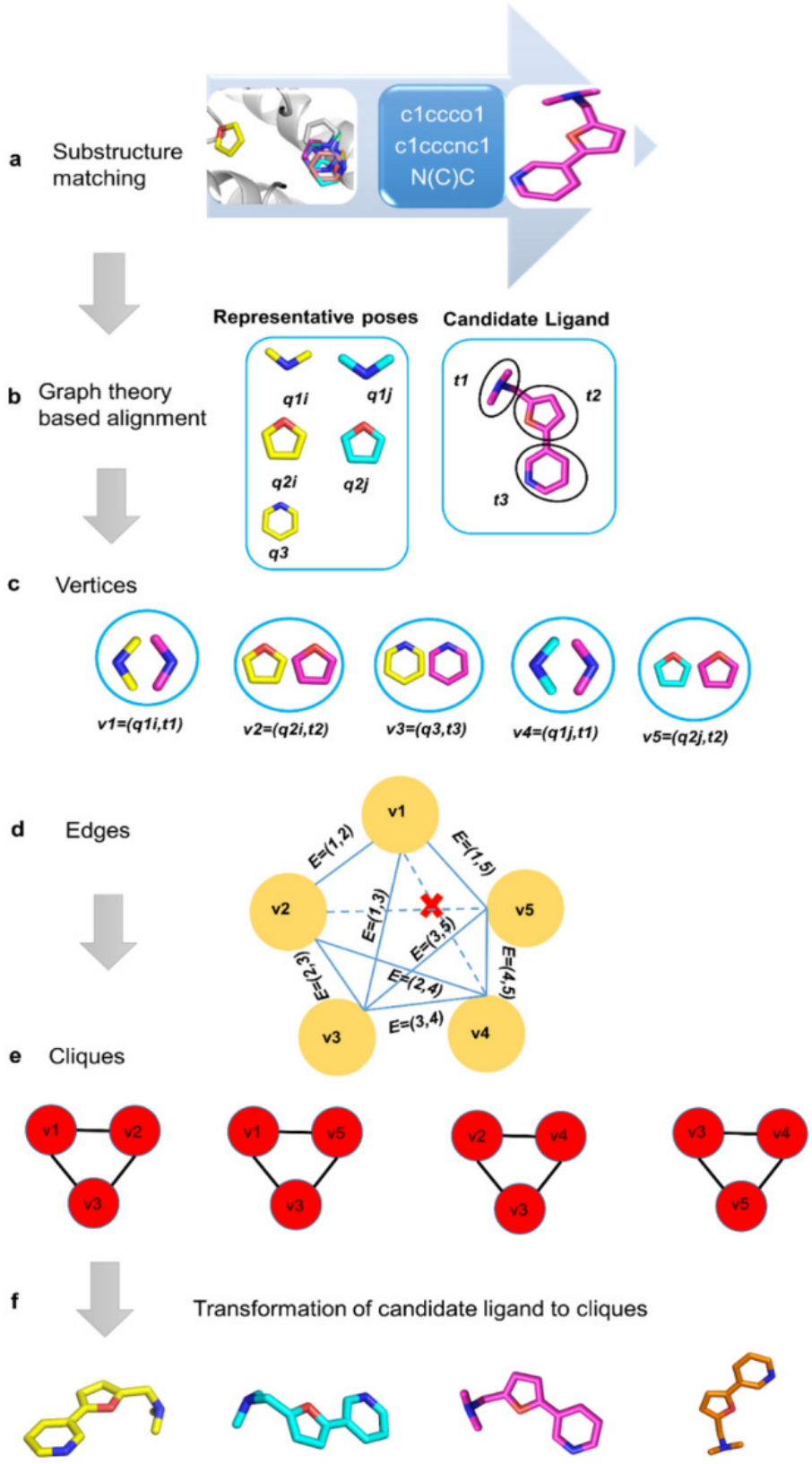


FIGURE 1.15 – Représentation du protocole de la création du graphe de connectivité à la reconstruction de molécules en 3D[32]

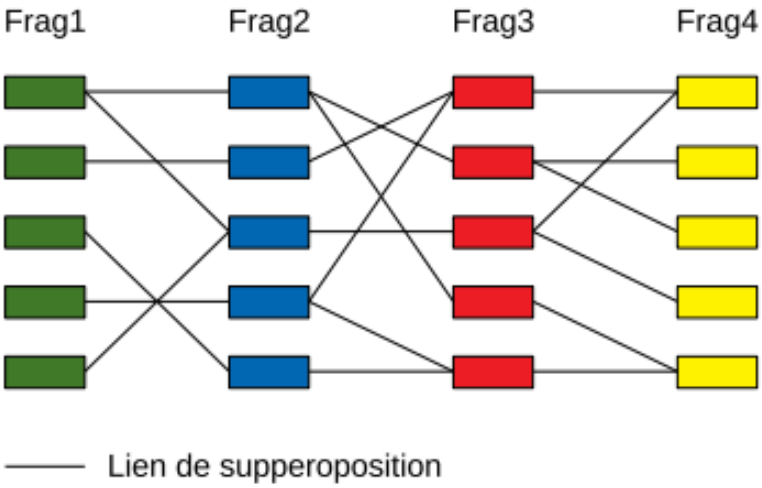


FIGURE 1.16 – Représentation du graphe de poses pour lequel chaque arête (lien de superposition) constitue une connexion possible entre deux fragments successives basée sur la RMSD entre leur chevauchement[17].

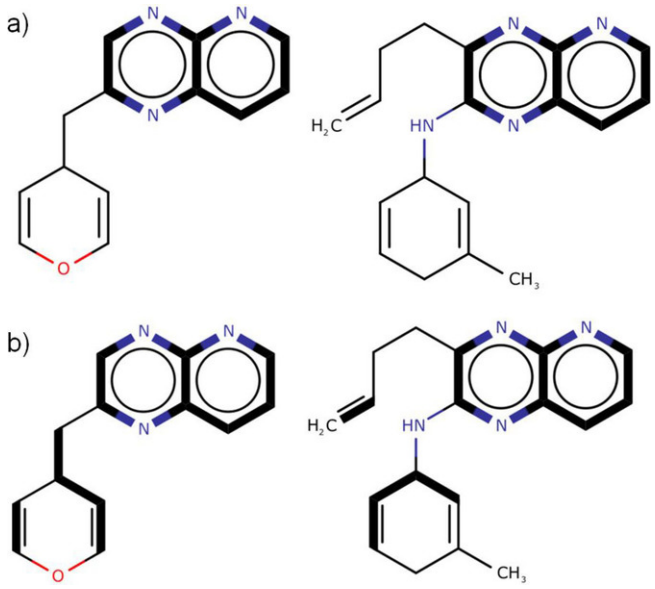
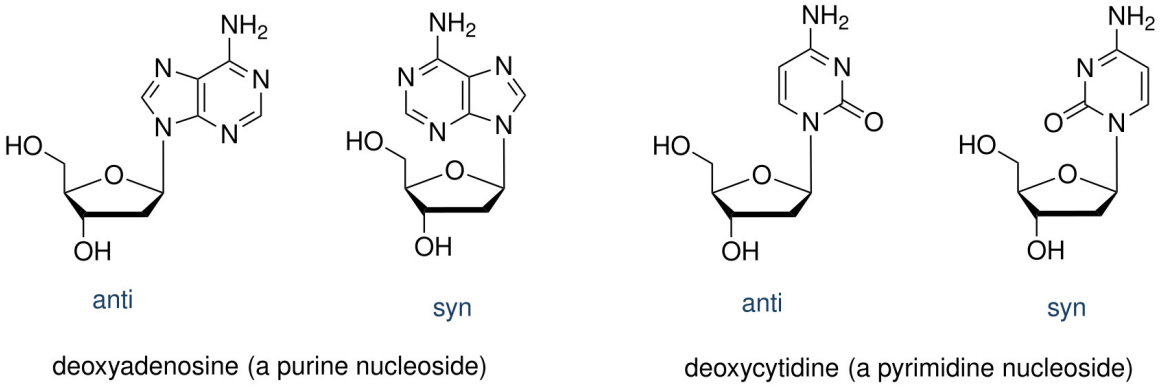


FIGURE 1.17 – Illustration des problèmes MCS, avec (A) le problème MCS déconnecté et (B) le problème MCS connecté. Le problème MCS connecté consiste à trouver le plus grand fragment commun entre deux molécules alignées. En l'occurrence, ce fragment correspond à la sous-structure marquée en gras. Le problème MCS déconnecté consiste à maximiser le nombre de liaisons ou d'atomes communs entre deux molécules alignées.

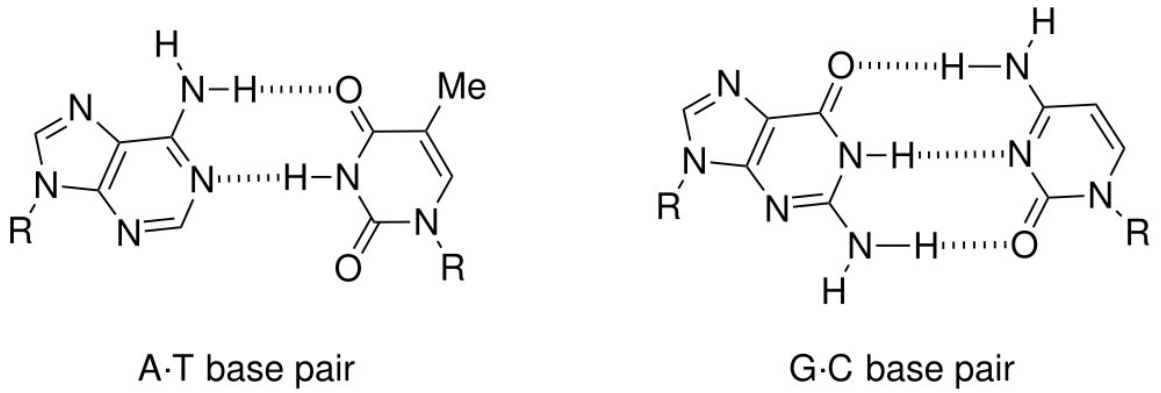
1.4 . Généralités sur les oligonucléotides

Les oligonucléotides (ONs) peuvent interférer avec différentes biomolécules afin de moduler des voies cellulaires facilement ou difficilement "druggable" (on entend ici la capacité d'une petite molécule à moduler efficacement l'activité d'une protéine *in-vivo*). En l'occurrence, les ONs peuvent être conçus afin de cibler de l'ARN/ADN ou toute cible n'appartenant pas à la classe des acides nucléiques[34]. Nous proposons de voir brièvement en revue les différentes familles d'ONs et de se focaliser principalement sur l'une d'elles que sont les **aptamères**.

1.4.1 . Quelques définitions



(a) Illustration des conformations ANTI et SYN



(b) Illustration de l'appariement des bases par la face Watson-Crick

FIGURE 1.18 – Illustrations tirées de https://atdbio.com/nucleic-acids-book/Nucleic-acid-structure

Les ONs sont constitués de nucléotides. Ces nucléotides sont composés de trois entités principales : un acide phosphorique, un ribose (pentose) et une base nucléique. Il subsiste 5 bases nucléiques, séparées en deux sous-familles : les pyrimidines d'une part constituées de l'Adénine (A) et la Guanine (G) ; les purines d'autre part constituées de la Cytosine (C), la Thymine (T) caractéristique de l'ADN,

et l'Uracile (U) caractéristique de l'ARN.

On distingue deux conformations de la base nucléique, qui sont la conformation ANTI (la base nucléique est orientée "à l'extérieur" par rapport au ribose), et la conformation SYN (la base nucléique est orientée "à l'intérieur" par rapport au ribose), comme illustrées à la figure 1.18a.

Deux faces sont caractérisées sur la base nucléique : la face Hoogsteen, et la face Watson-Crick impliquée principalement dans les appariements avec d'autres bases nucléiques. Ces appariements sont définis comme des liaisons de faibles énergies (liaisons hydrogènes) entre les bases nucléiques de A et G, ou entre C et T/U, ce qui permet par exemple à l'ADN de former une hélice double brin ou à un ARNsb de se structurer (figure 1.18b).

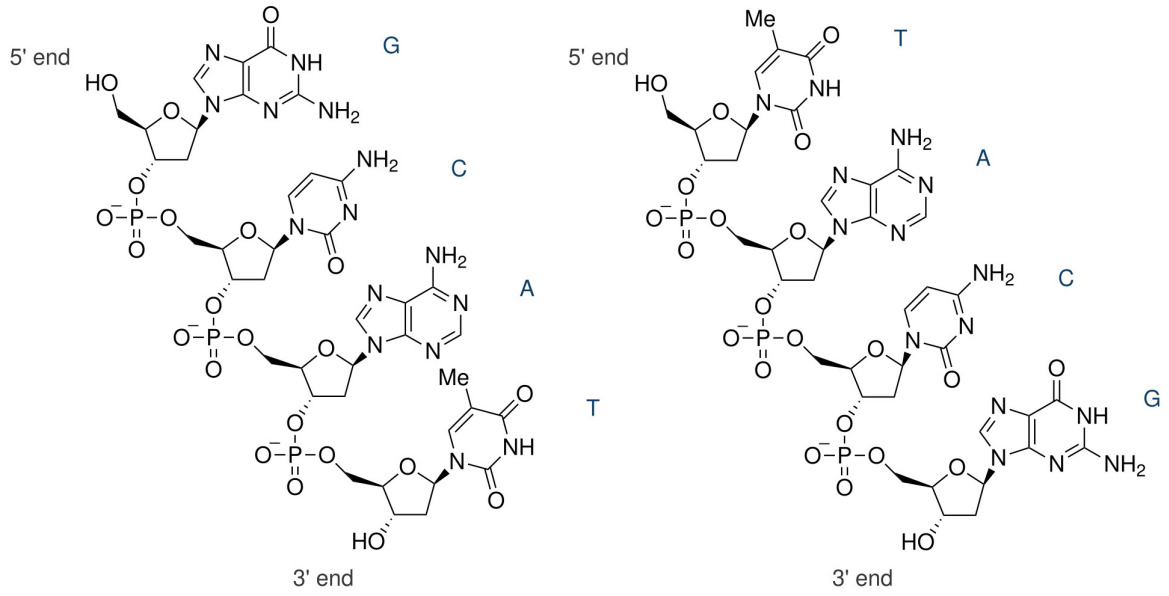


FIGURE 1.19 – Illustration d'un oligonucléotide selon la polarité 5′ → 3′ (https://atdbio.com/nucleic-acids-book/Nucleic-acid-structure)

Les ONs sont définis comme étant de courtes chaines d'ARN ou d'ADN (en général d'une dizaine de nucléotides), pour laquelle sa polarité est 5′ → 3′. Autrement dit, le nucléotide $n+1$ se lie à la position 3′ du nucléotide n par l'intermédiaire de son acide phosphorique situé à l'extrémité 5′ (figure 1.19). Au niveau du vocabulaire, un oligonucléotide composé de n nucléotides est appelé un **n-mer**.

1.4.2 . Brève revue des différentes familles d'ONs

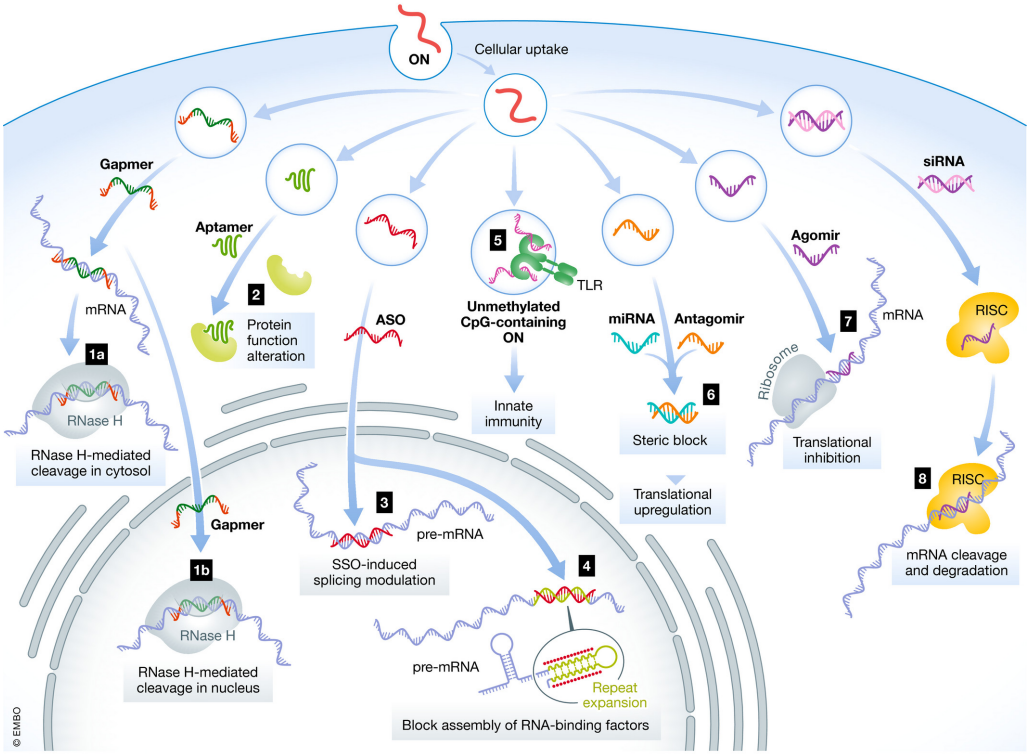


FIGURE 1.20 – Illustration de différentes familles d'ONs[35]

Les ONs ciblant les acides nucléiques

Oligonucléotides Antisens (AONs). Les AONs correspondent à une classe thérapeutique d'ONs ciblant des ARNs, et interagissent par un appariement avec la face Watson-Crick. Parmi les AONs, on peut citer :

- les "Gapmers" qui vont s'hybrider à l'ARN messager (ARNm). Cette hybridation induit la dégradation de l'ARNm par la Ribonucléase H, et donc empêcher le processus de traduction de l'ARNm en une protéine. Parmi les médicaments étant un gapmer, on peut citer le Mipomersen pour le traitement de patients atteint d'hypercholestérolémie familiale.[34]
- Les AONs pour le "splice switching", liant un pre-ARNm dans le noyau cellulaire afin de moduler le processus d'épissage. Parmi les traitements approuvés, on peut citer le Nusinersen indiqué pour l'atrophie musculaire spinal. En effet, cette maladie est causée par le dysfonctionnement d'une protéine SMN due à un épissage défectueux. Le traitement permet de restaurer un ARNm correct codant pour la protéine SMN fonctionnelle[34].

MicroRNAs (miRNAs). Les miRNAs sont des ARNs non-codants d'une vingtaine de nucléotides s'hybridant sur la région 3' non-traduite de l'ARNm. Leur but consiste à éteindre l'expression

d'un gène par la méthylation directe de l'ADN, inhiber la traduction d'un ARNm ou conduire à sa dégradation[34].

Petits ARNs interférants (siRNAs). Les siRNAs sont des molécules d'ARN double brins de 20 à 25 paires de bases, qui causent la dégradation de l'ARN cible[34]. Une fois rentré dans la cellule, cet ARN va former un complexe multiprotéique appelé RISC, où il sera déroulé en un ARN simple brin, afin de rentrer en complémentarité avec la portion de l'ARNm cible. Cette complémentarité va induire le clivage et la dégradation de l'ARNm, empêchant ainsi la traduction de cette dernière en une protéine[34].

Les ONs ciblant des biomolécules autres que les acides nucléiques

Aptamères. Les aptamères sont des ONs simples brins qui peuvent se replier en une structure en trois dimensions. Ils sont composés généralement d'environ 30 à 60 nucléotides (A, T/U, G, C), bien que des séquences plus courtes puissent être envisagées. Ils peuvent interagir avec des protéines, des carbohydrates ou encore des lipoglycans.
Parmi les aptamères qui sont sur le marché, on peut citer le Pegaptanib, un inhibiteur du VEGF et indiqué pour la dégénérescence maculaire liée à l'âge (DMA). On peut également citer des aptamères ou des ligands dérivés de nucléotides interagissant avec des protéines ne liant pas naturellement d'acides nucléiques, telle que la Beta-Sécrétase 1[36–38] qui fait l'objet d'une étude particulière dans cette thèse.

1.4.3 . Les aptamères et leurs dérivés

La méthode classique SELEX pour la conception d'aptamères

L'approche SELEX ("Systematic Evolution of Ligands by EXponential enrichment") est une méthode expérimentale conventionnelle pour la sélection et la synthèse d'aptamères à haute affinité contre des cibles biologiques, et pour laquelle la méthode est décrite par la figure 1.21. Différentes étapes itératives sont réalisées durant ce processus :

1. Contre-sélection : À partir de bibliothèques aléatoires d'ONs et en présence de la cible biologique immobilisée sur une matrice, les ONs ne se liant pas spécifiquement à cette dernière sont éliminés,
2. Un pool d'ONs est ensuite incubé avec la cible à une certaine concentration : les ONs liés à la cible sont collectés au travers d'une élution
3. Les ONs sélectionnés sont soumis à une amplification par PCR au travers d'une transcription inverse

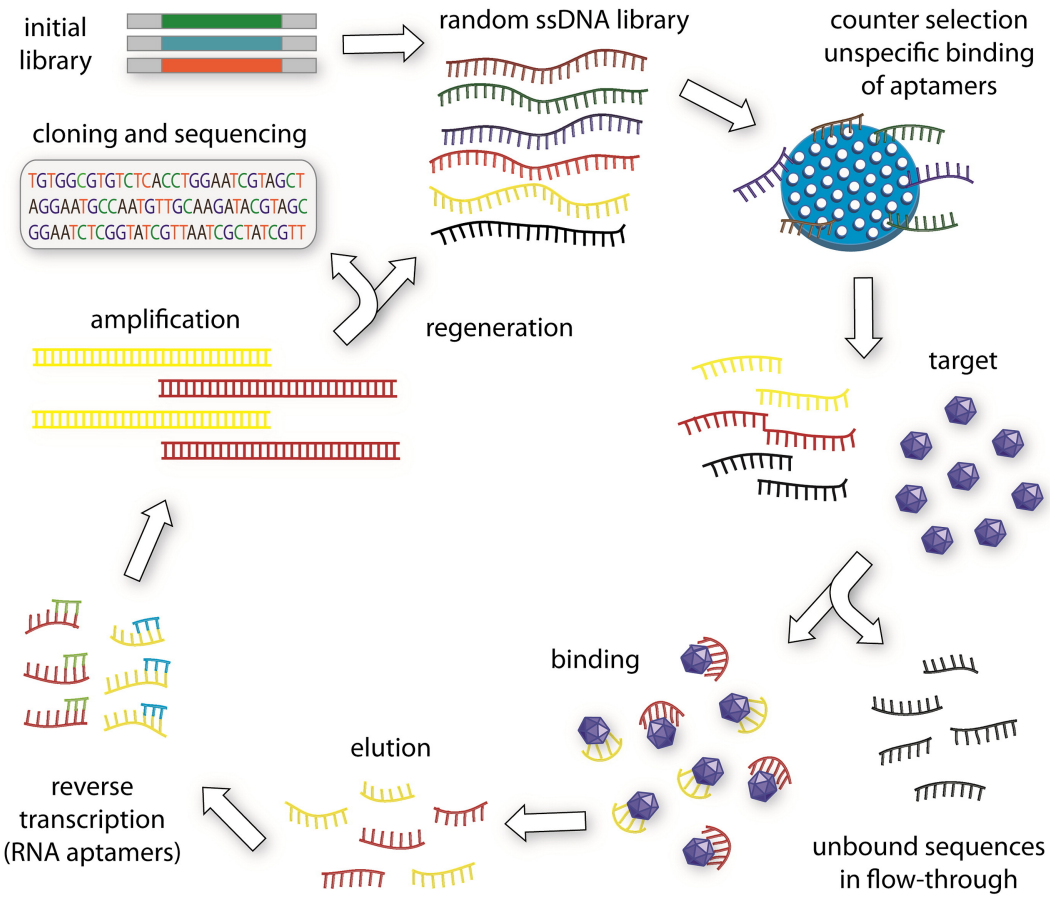


FIGURE 1.21 – Workflow de l'approche SELEX[39]

4. Un nouveau cycle commence avec les ONs retenus par le précédent cycle.

Cette approche possède néanmoins quelques inconvénients. Parmi ceux-ci, la génération des aptamères peut prendre quelques semaines à un mois afin d'obtenir des aptamères hautement spécifiques à la cible biologique, au travers d'une vingtaine de cycles effectuée. De plus, cette approche peut ne pas être robuste ou reproductible d'une expérience à l'autre.

Comparaison avec les immunothérapies

En biochimie, les aptamères sont généralement considérés comme des alternatives aux immunothérapies. Parmi les différences, on peut noter spécifiquement :

- la **stabilité**. Les anticorps monoclonaux sont sensibles à la température, faisant que leur conservation doit être réalisée à basse température pour ne pas être irréversiblement dénaturés. Contrairement à ces derniers, les aptamères sont stables à température ambiante, et leur dé-

naturation (c'est-à-dire le passage d'une structure repliée à une structure non-repliée) est réversible.

- l'**immunogénécité**. En général, un traitement par immunothérapie déclenchera une réaction du système immunitaire, réaction limitée avec l'administration d'aptamères.
- le **mode d'adminsitration**. Un traitement basé sur une immunothérapie est délivré par voie intraveineuse (IV), tandis que les aptamères peuvent être administrés par inhalation, perfusion ou administration locale. C'est le cas du Pegaptanib par injections intravitréennes.

Les aptamères possèdent néanmoins quelques limitations. Parmi celles-ci, on peut citer les **effets hors-cibles** ou une **dégradation rapide** par les endo- ou exonucléases. Les effets hors-cibles se caractérisent comme la capacité d'une molécule à interagir avec des biomolécules autres que la cible biologique. Une façon de limiter ces problèmes consiste à appliquer des modifications chimiques.

Les pseudo-aptamères, une classe d'aptamères avec des modifications chimiques

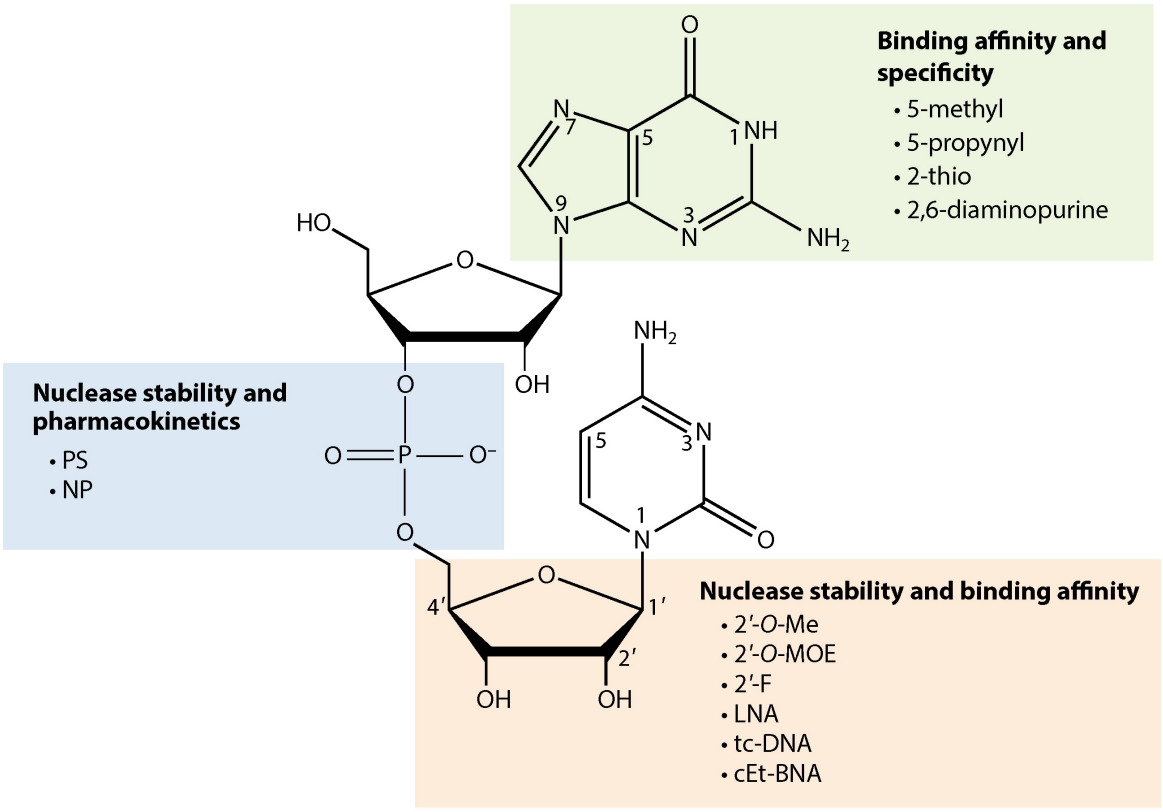
Les modifications chimiques peuvent être réalisées sur au moins une des entités qui constitue un nucléotide, à savoir la base nucléique, l'acide phosphorique, le ribose ou le squelette phosphodiester ("backbone") comprenant à la fois l'acide phosphorique et le ribose.

Liaison Phosphodiester (Acide Phosphorique). Classiquement, une modification sur la liaison phosphodiester consiste à substituer le groupe phosphate PO_4^{2-} (PO) par un phosphorothioate (ou thiophosphate) PSO_3^{2-} (PS) ou $PS_2O_2^{2-}$ (PS_2). Cette modification confère à l'ONs une résistance vis-à-vis de la dégradation par les nucléases. En général, cette modification est appliquée sur quelques nucléotides de l'aptamère, car des toxicités liées à celle-ci ont pu être rapportées[34,42]

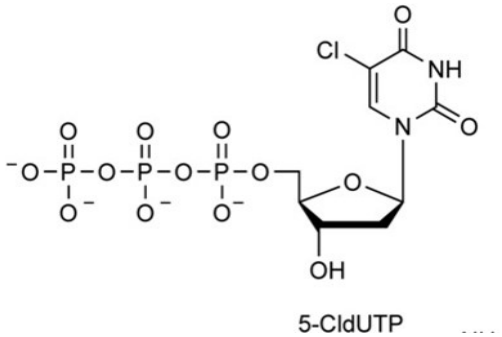
Ribose. Tout comme la modification sur le groupe PO, la modification sur le ribose a pour but de réduire la sensibilité des aptamères vis-à-vis des nucléases. Cette modification est appliquée à la position 2' du ribose. Parmi les modifications classiques, on peut citer le 2'-O-Méthyl, ou encore le 2'O-Méthylethyl qui peut aussi jouer un rôle sur la modulation de l'affinité. Une autre des modifications avec des propriétés similaires est le 2'-fluoro[34].

Base Nucléique. Les modifications au niveau de la base nucléique ont pour but d'améliorer l'affinité des aptamères vis-à-vis de la cible biologique, ce qui réduirait les effets hors-cibles. Parmi les modifications, on peut citer des modifications à la position 5' des pyrimidines, dont par exemple le 5-chloro-UTP (figure 1.22b), un nucléotide modifié incorporé dans la conception d'aptamères ciblant l'enzyme Beta-Sécrétase 1 afin de moduler son activité[40]. Un autre exemple sont les SOMAmers, qui

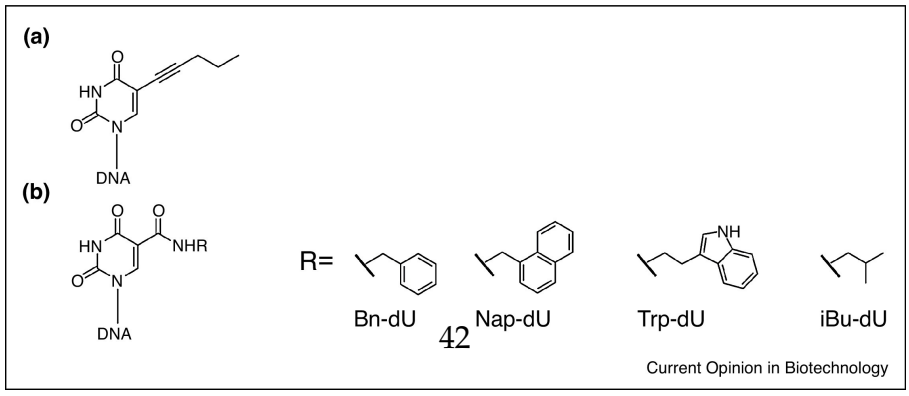
présentent des modifications de nature cyclique (figure 1.22c)[41,43]. Ces modifications peuvent aussi avoir un intérêt dans le design d'ONs simple brin, en perturbant les appariements Watson-Crick et en les empêchant ainsi de se replier en structure.



(a) Exemples classiques de modifications chimiques réalisées sur les nucléotides[34]



(b) Exemple d'une modification chimique sur l'Uracile à la position 5[40]



(c) Exemple de modifications pour le design de SOMAmers[41]

FIGURE 1.22 – Exemple de modifications chimiques sur les nucléotides

1.5 . Un modèle d'étude sur la Beta-Sécrétase 1

1.5.1 . Généralités

La Beta-Sécrétase 1 (BACE1) est une enzyme impliquée dans la maladie d'Alzheimer (MA), une démence neurodégénérative progressive, dont le premier cas a été décrit par Alois Alzheimer en 1907. Deux formes de la maladie ont été identifiées en fonction de l'âge d'apparition[44,45] :

- la forme familiale ou héréditaire ("FAD") qui compte pour 1 à 5% de tous les cas d'Alzheimer et qui est associée à une apparition précoce et,
- la forme sporadique ("SAD") qui est la plus majoritaire associée à une apparition tardive.

D'un point de vue biologique, la MA est décrite par deux principales caractéristiques conduisant à la mort neuronale : (i) les **plaques amyloïdes** constituées de peptites β-amyloïdes (Aβ) aggregés entre les neurones[46], et (ii) les **enchevêtrements neurofibrilaires** constitués d'une accumulation de protéines Tau hyperphosphorylés dans les neurones.
L'accumulation des plaques amyloides perturbe la communication neuronale au niveau des synapses, tandis que l'accumulation de la protéine Tau provoque un blocage du transport de nutriements et de molécules essentielles à la survie et fonctionnement des neurones[47].

Face aux causes multifactorielles de la maladie, différentes hypothèses ont été proposées pour expliquer le processus physiopathologique de la maladie[45]. Nous décrirons ici succinctement qu'une seule parmi ces hypothèses, à savoir **l'hypothèse de la cascade amyloide**.

1.5.2 . Hypothèse de la cascade amyloïde

L'hypothèse de la cascade amyloïde a été proposée par Hardy et Higgins en 1992. Ils ont suggéré que les fragments Aβ initient la cascade pathologique de la maladie d'Alzheimer, conduiraient à l'enchevêtrement neurofibrilaires et à la mort des cellules neuronales[48]. Bien que l'hypothèse ait pu faire l'objet de discussion, plusieurs preuves tendent à démontrer sa validité.

Réaction catalytique responsable des peptides Aβ

Avant de mentionner ces différentes preuves, il est indispensable de décrire la réaction catalytique responsable de la formation des peptides Aβ afin de prendre connaissance de tous les acteurs impliqués dans la voie amyloïdogénique.

Les peptides Aβ proviennent de la catalyse de la **protéine précurseure de l'amyloïde (APP)**, une glycoprotéine membranaire de type I. Plusieurs isoformes de l'APP existent, avec un nombre de résidus compris entre 695 (APP695, la forme la plus abondante exprimée par les neurones cérébraux)

et 770 (APP770, principalement exprimée au niveau des plaquettes et cellules périphériques). Cette protéine est impliquée comme régulateur de la formation et la réparation synaptique, et le transport neuronal antérograde[49].

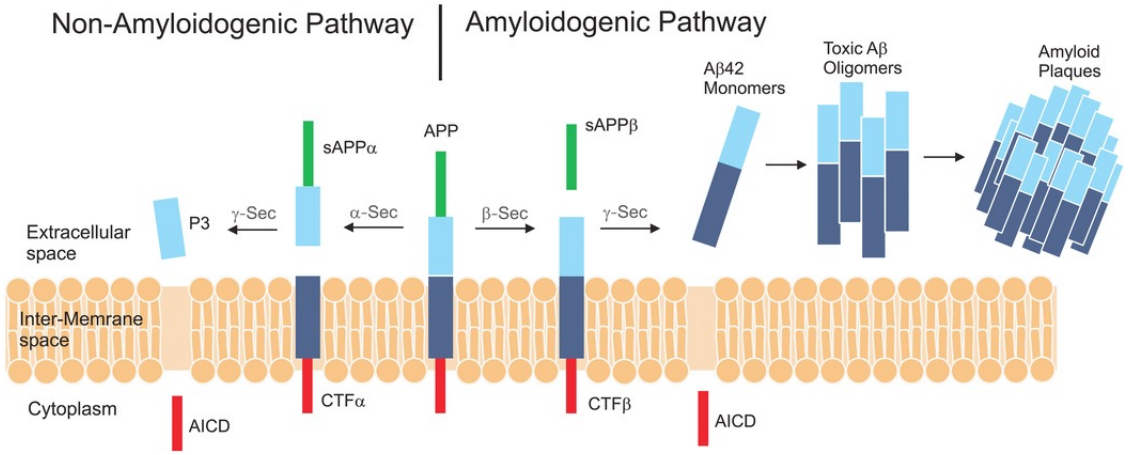


FIGURE 1.23 – Schéma des voies amyloidogénique impliquée dans la formation de plaques amyloides, et non-amyloidogénique[50]

Deux principales voies catalytiques sont impliquées dans la catalyse de cette glycoprotéine humaine (Figure 1.23) : (i) la voie **non-amyloidogénique** pouvant conduire à la formation de fragments aux propriétés neuroprotectrices[51] et (ii) la voie **amyloidogénique** conduisant à la formation des peptides Aβ.

Ces deux voies impliquent l'action de deux réactions catalytiques successives, avec une différence notable pour la première réaction entre les deux voies. En effet, la voie non-amyloïdogénique implique en premier l'action de l'enzyme **α-sécrétase**, tandis que la voie amyloïdogénique implique l'action de l'enzyme **β-sécrétase (BACE1)** conduisant à une libération différente de produits. La réaction catalysée par l'α-sécrétase conduit à la formation du fragment sAPPα et du fragment C83, correspondant à l'extrémité C-terminal de l'APP constitué de 83 acides aminés. La β-sécrétase, quant à elle, peut cliver à deux niveaux :

- entre Tyr606 et Glu607 conduisant à la libération du fragment C89
- entre Met596 et Asp597 conduisant à la libération du fragment C99.

La deuxième réaction catalytique implique l'action du complexe enzymatique **γ-sécrétase** sur les fragments C83, C89 et C99, conduisant à la formation de peptitdes P3 et d'Aβ respectivement. Il est, ici, intéressant de noter que le fragment C99 peut être clivé à différents sites par la γ-sécrétase conduisant à la production de différentes isoformes Aβ. Les deux principales isoformes sont l'isoforme Aβ40, la

plus majoritaire, et l'isoforme Aβ42 qui est cytotoxique à des niveaux d'expression pathologique[49,51].

Des facteurs de risques génétiques liés à la voie amyloïdogénique

Des mutations autosomales dominantes du gène de l'APP ont été décrites par différentes études comme étant une cause génétique de la maladie d'Alzheimer[52]. Parmi elles, des mutations ont montré une augmentation de la production des peptides Aβ, comme la mutation Swedish[53] et la mutation A673V[54]. À l'inverse, les porteurs de la mutation A673T de l'APP sont moins à risque d'être atteint du déclin cognitif dû à une diminution du clivage de l'APP par BACE1[55]. Par ailleurs, en 1990, Yoshikai et *al*. ont montré que les personnes atteintes du syndrome de Down, ayant une extra-copie de ce chromosome, développaient une FAD précoce[56].

Parmi les autres facteurs de risques génétiques, nous pouvons aussi citer les gènes PSEN1 et PSEN2, localisés respectivement sur les chromosomes 19 et 1. Les deux protéines résultantes PSEN1 et PSEN2 participent à la formation du site catalytique de la γ-sécrétase[57]. Plus de 100 mutations de PSEN1 et 11 mutations de PSEN2 ont été identifiées comme autosomale dominante provoquant une MA précoce[44]. Par exemple, il a été montré que la mutation L166P sur PSEN1 peut causer une réduction de l'activité de la γ-sécrétase, et donc conduire à une décroissance de proportion de l'isoforme $A\beta_{1-42}$. À l'inverse, la mutation G384A sur PSEN1 conduit à une augmentation de l'isoforme $A\beta_{1-42}$ causant l'AD précoce[58,59].

Ces résultats mettent donc en avant **le lien entre la voie amyloïdogénique du clivage de l'APP et l'apparition de maladie d'Alzheimer précoce**.

Un autre facteur de risque est lié au gène APOE présent sur le chromosome 19. Ce gène est responsable de la formation d'une lipoprotéine de 299 acides aminés. Trois allèles APOE ont été identifiés : $\epsilon 2$, $\epsilon 3$ et $\epsilon 4$ conduisant aux isoformes lipoprotéiques ApoE2, ApoE3 et ApoE4 respectivement. Principalement exprimé dans le foie et le cerveau, ApoE joue un rôle important dans le métabolisme lipidique. Des études ont montré que ApoE4 peut augmenter l'agrégation des peptides Aβ et perturber leur clairance[60–62]. ApoE4 a donc été identifié comme un risque majeur de la FAD contrairement à ApoE2 et ApoE3.

1.5.3 . Des immunothérapies vis-à-vis des plaques amyloïdes

Parmi les traitements contre la maladie d'Alzheimer, nous pouvons citer les traitements symptomatiques et les traitements **"disease-modifiers"**.

Les traitements symptomatiques consistent à accompagner le patient malade dans son bien-être et

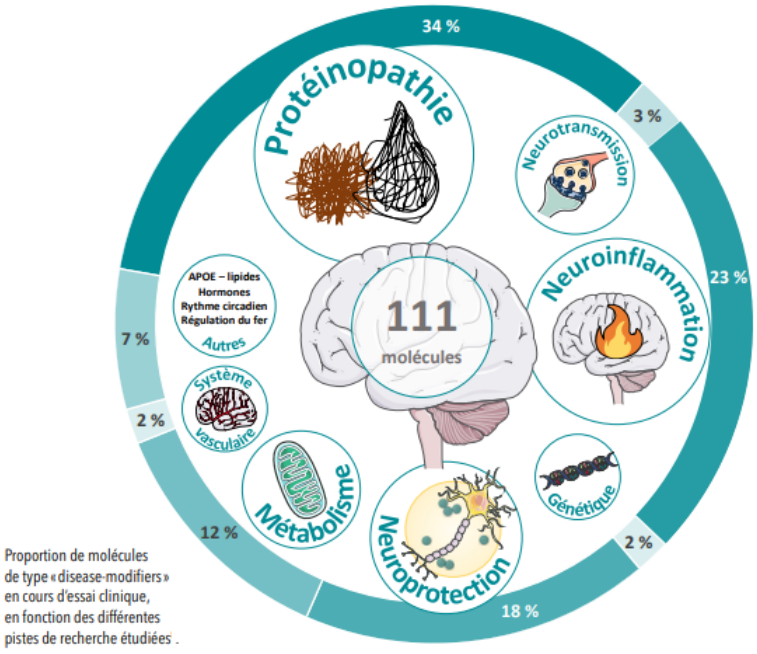


FIGURE 1.24 – Proportions de molécules en phase clinique en fonction de la classe thérapeutique[63]

sa qualité de vie face aux symptômes causés par la maladie (les troubles de langage, difficultés à s'organiser ou à s'orienter, pertes de mémoire...). Néanmoins, ces traitements ne ralentissent pas la progression de la maladie.

Ce n'est que très récemment que trois traitements affectant la progression de la maladie ont vu le jour. Ces traitements consistent en des immunothérapies ciblant les dépôts amyloides dans le cerveau.
Le premier qui fut approuvé est l'Aducanumab[64], mais fut l'objet de discussions scientifiques sur sa réelle efficacité à ralentir le déclin cognitif[65].
Le deuxième est le Lecanemab[66], définitivement approuvé en juin 2023 sur le marché américain, qui aurait montré ralentir le déclin cognitif de 27% chez les patients Alzheimer.
Enfin, le dernier est le Donanemab[67], qui aurait montré un ralentissement cognitif de 35%. Ce dernier est encore en phase III, en attente d'une approbation par les autorités de santé.

Ainsi, l'ensemble de ces preuves biologiques et thérapeutiques met en avant la considération à porter vis-à-vis du rôle amyloïdes comme un des responsables de l'apparition précoce de la maladie d'Alzheimer, et comme une cible de choix pour ralentir le traitement cognitif.

Néanmoins, rappelons que plusieurs molécules de type "diseaser-modifiers" sont en étude clinique, avec différentes cibles biologiques (Figure 1.24)

1.5.4 . La Beta-Sécrétase 1 : un modèle d'étude

Caractéristiques cellulaires de BACE1

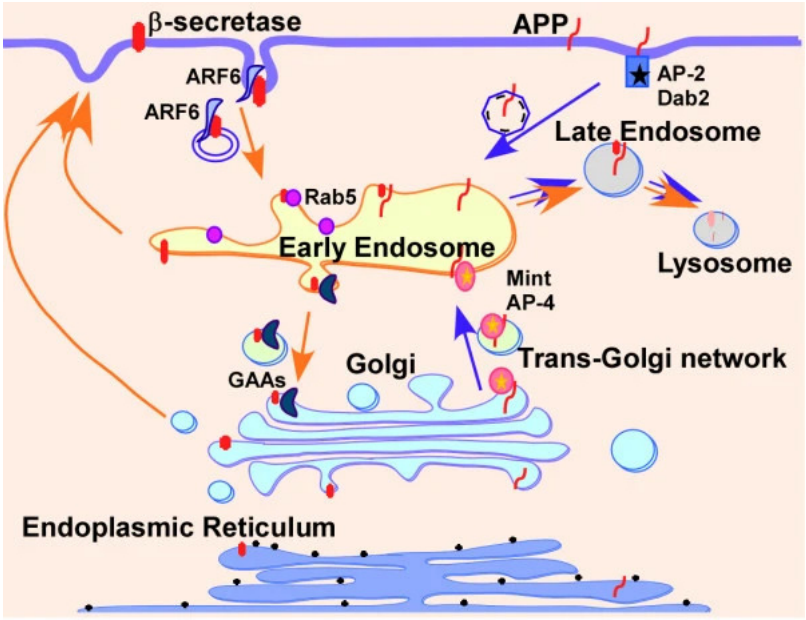


FIGURE 1.25 – Trafic cellulaire de l'APP et de BACE1[68]

BACE1 est une enzyme transmembranaire de type I, composée de 501 acides aminés. Elle appartient à la famille des protéases aspartiques. **Hautement exprimée dans les neurones cérébraux**, BACE1 est synthétisée sous sa forme zymogène avec la présence d'un prodomaine dans le réticulum endoplasmique. L'enzyme, prise en charge par des protéines de transports (AP-1, Arf-1 et Arf-4), est exportée vers le Golgi où il subit un certain nombre de modifications post-traductionnelles (acétylation, glycosylation, phopsphorylation, ubiquitination, SUMOylation). Il peut être noté que les modifications post-traductionnelles peuvent impacter l'activité amyloidogénique de BACE1, ainsi que sa localisation[68]. Suite à cela, BACE1 est (i) soit dirigée vers la membrane plasmique pour être internalisée dans l'endosome, (ii) soit dirigée directement dans l'endosome. Son **activité étant optimale à pH acide**, son action catalytique a principalement lieu dans les compartiments acides intracellulaires tels que **l'endosome** ou le *trans*-Golgi (TGN). En effet, il a été montré que l'APP et BACE1 sont co-localisées dans le TGN et l'endosome précoce au cours de leur trafic cellulaire (Figure 1.25), et que ces deux compartiments cellulaires sont les principaux sites de production des peptides Aβ[69].

Il est important de noter que BACE1 peut catalyser une trentaine de substrats, et qu'elle peut donc être impliquée dans d'autres voies physiologiques (Figure 1.26)[71,72]. En effet, différentes études ont montré que l'extinction ("knock-out") du gène *BACE1* chez la souris peut induire une hypomyélination

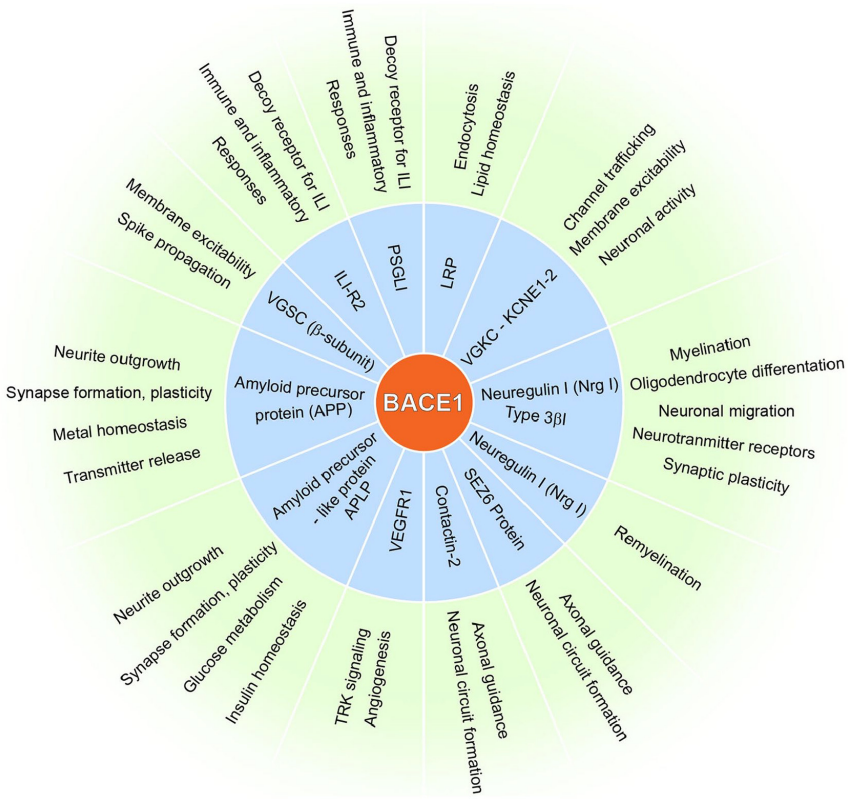


FIGURE 1.26 – Quelques substrats catalysés par BACE1[70]

(substrat : NRG1), des troubles de la mémoire, une altération de la neurogénèse et de l'astrogénèse (substrat : Jag1), des anormalités rétiniennes (substrat : VEGFR1), ou encore avoir un rôle dans la fonction synaptique[72]. Par conséquent, une inhibition complète de l'activité BACE1, dans le cadre d'un traitement préventif de la maladie d'Alzheimer, conduira inévitablement à des effets indésirables lourds. Par contre, une inhibition partielle (dose-dépendante) de BACE1 n'interfère pas avec la transmission synaptique[73,74].

De plus, dans le cerveau, les cellules microglia jouent un rôle important en assurant une barrière protectrice pour les neurones[75], notamment en éliminant par phagocytose des agrégats Aβ[76]. Une étude très récente montre que l'inhibition de BACE1 dans les microglia augmente leur potentiel d'élimination d'agrégats Aβ[77,78]. Par une inhibition dose-dépendante de BACE1[79] et/ou ciblée dans les microglia[77,78], on peut donc s'attendre à des effets bénéfiques pour le traitement de la MA[80].

Par ailleurs, des études thérapeutiques ont montré des effets hors-cibles avec une enzyme homologue à BACE1, appelée la Beta-Sécrétase 2 (BACE2), pour laquelle des molécules ont interagi avec ce dernier de manière non-sélective[79]. Bien qu'elle soit moins exprimée dans le cerveau, et principalement

exprimée dans le pancréas, des études ont mis en avant que BACE2 puisse avoir un effet préventif ou protecteur contre la génération de fragments Aβ. Par conséquent, une inhibition de BACE2 peut entraîner un effet néfaste de la maladie. Ainsi, malgré les échecs répétés d'efficacité d'inhibiteurs de BACE1 dans les essais cliniques, BACE1 demeure une cible thérapeutique d'intérêt.

Caractéristiques structurales de BACE1 et BACE2

Comme mentionné précédemment, BACE1 est une protéine transmembranaire, pour laquelle sa partie catalytique est de forme globulaire, et est composée de deux lobes, respectivement le lobe N-terminal et le lobe C-terminal. Comme toutes les enzymes de la famille à laquelle elle appartient, BACE1 possède deux motifs aspartiques (D93, D228) formant son site catalytique, orientés dans la partie luminale, et impliqués dans la protéolyse du substrat. Ce site actif est aussi composé d'un "flap", une boucle responsable de l'accessibilité ou non du site actif[81,82] (voir figure 1.27). En effet, un inhibiteur placé dans le site actif peut former des liaisons hydrogènes avec le flap, induisant un repliement de ce dernier vers le site actif, ce qui conduit au passage d'une conformation "ouverte" à "fermée". Un fait intéressant est que la conformation du flap peut être influencée en fonction du pH : elle adoptera une conformation ouverte à pH acide (ph = 4 à 5), tandis qu'elle adoptera une conformation inactive à pH neutre (pH = 6 à 7)[81].

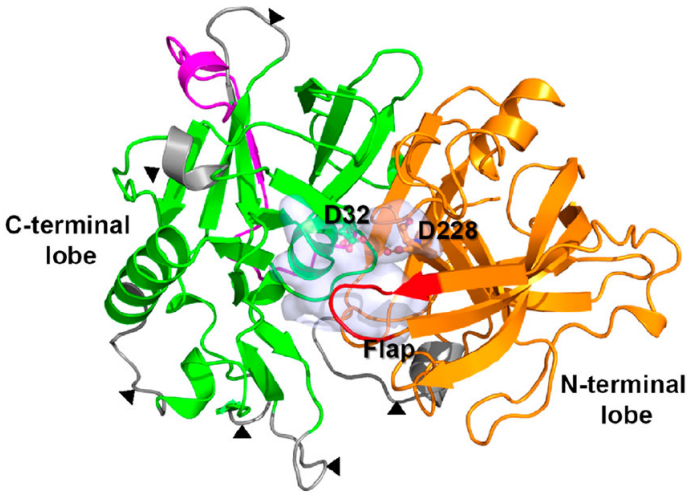


FIGURE 1.27 – Structure 3D de BACE1[83]

Des études ont pu définir l'existence d'une deuxième région capable de moduler l'activité de BACE1, en ayant une action à distance sur la conformation du site actif : on parle alors de site allostérique ou **d'exosite** situé au lobe C-terminal. Contrairement au site actif bien caractérisé par une poche, l'exosite est une surface avec l'absence d'une poche caractéristique. Il a été montré que l'exosite peut

faire l'objet d'une interaction avec un anticorps monoclonal ou d'une interaction avec un peptide[84–86]

Par ailleurs, le manque de spécificité des composés à but thérapeutique entre BACE1 et BACE2 s'explique par la forte homologie entre les deux enzymes, notamment au niveau du site actif qui fut jusqu'à présent la seule région cible de la grande majorité des composés. En effet, BACE1 et BACE2 partage près de 70% de similarité entre leur séquence, dont près de **80% de similarité au niveau du site actif**. De plus, une comparaison structurale entre les deux enzymes montre une **forte similitude au niveau de leur repliement global**, avec une RMSD ("Root Mean Square Deviation") de 1.46Å[87] (Figure 1.28). De ce fait, les composés ciblant le site actif de BACE1 ont de fortes chances d'interagir avec le site actif de BACE2. Néanmoins, des caractéristiques clés de spécificité au niveau du site actif ont été déterminées, comme nous le verrons à la section 1.5.4.

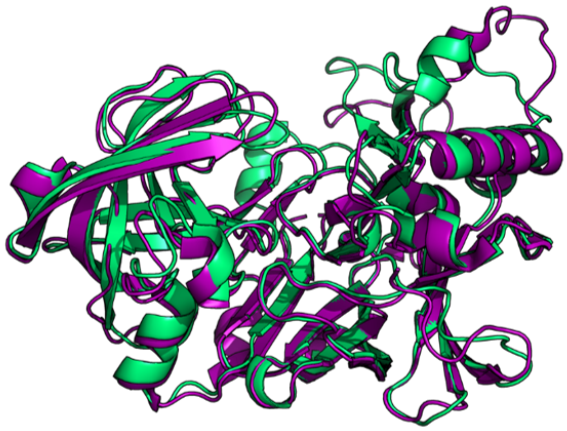
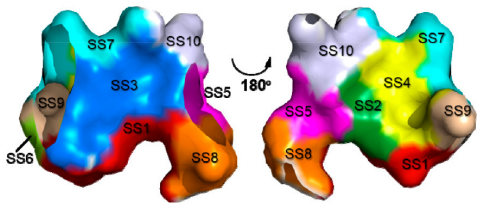


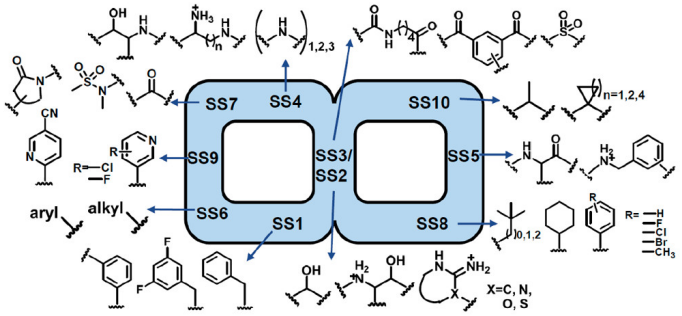
FIGURE 1.28 – Alignement structurale entre BACE1 (en vert) et BACE2 (magenta)

L'intérêt de l'exosite se justifie à la fois par cette haute conservation du site actif entre BACE1 et BACE2, mais également par le fait que cette **région est moins conservée**. En effet, une comparaison de la séquence de l'exosite entre ces deux enzymes montre une similarité à hauteur de 60%. Cet exosite est défini notamment par 3 boucles nommées C (254-257), D (270-274) et F (309-320)[84].

Bien qu'une ancienne nomenclature définissait le site actif en 8 sous-sites[88], Hu et *al.*[83] ont défini une nouvelle nomenclature pour caractériser le site actif en 10 sous-sites à partir de 354 structures cristallines de BACE1 en complexe avec des inhibiteurs. Ces 10 sous-sites reposent sur un modèle d'interactions résidus-ligands, et ont identifié les différentes sous-structures interagissant favorablement avec chacun d'eux (voir figures 1.29 et 1.30).



(a) Représentation 3D des différents sous-sites caractérisant le site actif de BACE1



(b) Schéma d'une représentation en 8 des sous-sites avec les sous-structures favorables à une interaction

FIGURE 1.29 – Représentation des sous-sites du site catalytique de BACE1[83]

Sous-sites	Résidus
SS1	L30, Y71, F108, W115, I118
SS2	D32, D228
SS3	T72, Q73, K107
SS4	G230, T231
SS5	G34, Y198
SS6	I110
SS7	T232, N233, R235, S325
SS8	S35, V69, P70, W76, I126, R128
SS9	G11, G13, Y14, S229, A335
SS10	K224, I226, T329, V332

FIGURE 1.30 – Composition des sous-sites constituant le site actif de BACE1[83]. Les résidus soulignés représentent les résidus non-conservés et substitués chez BACE2.

Caractéristiques clés de la spécificité entre BACE1 et BACE2

Différentes études ont pointé des facteurs clés de la spécificité entre BACE1 et BACE2[14,87,89], en lien avec la conformation intrinsèque du site actif, ou de la substitution de résidus de sa séquence.

Résidus groupes-spécifiques. Le site actif entre BACE1 et BACE2 est relativement bien préservé. Néanmoins, quelques différences notables peuvent être relevées. La première est que le site actif de BACE1 possède un nombre plus élevé d'acides aminés chargés que celui de BACE2 : les interactions électrostatiques pourraient donc être privilégiées lors du design d'inhibiteurs anti-BACE1. De plus, en reprenant la nomenclature des sous-sites de Hu et *al.*, sept résidus ne sont pas identiques. On peut ainsi noter (selon la numérotation de BACE1) :

- la substitution de K107 en N107 dans le SS3 : la chaine latérale de l'asparagine est plus courte que celle de la lysine, et peut établir potentiellement plus de liaison hydrogène.
- la substitution de I110 en L110 dans le SS6 : la chaine latérale hydrophobe de l'isoleucine est plus longue que celle de la leucine. Cette mutation peut impacter la taille de la cavité, qui serait plus petite chez BACE2 que BACE1[89]
- la substitution de N233 en L233 dans le SS7 : la polarité peut être affectée, où l'asparagine prône à établir des interactions électrostatiques et polaires avec le ligand que la leucine.[89]
- la substitution des résidus P70, I126 et R128 en K70, L126 et K128 respectivement dans le SS8 : la substitution de la proline en lysine peut affecter la flexibilité du flap due à la perte de la propriété cyclique apportée par la proline[87]
- la substitution de T329 en N329 dans le SS10

Une dynamique de la poche de liaison différente. Une étude de dynamique moléculaire (MD) a été réalisée par Hernandez-Rodriguez et *al.*[89]. Cette MD a été réalisée sur une conformation de BACE1 (code PDB : `2QP8`) et une conformation de BACE2 (code PDB : `2EWY`), avec CHARMM27, pour une durée totale de 50ns ; avec une sauvegarde de "snapshots" toutes les 5ns afin d'évaluer précisément les changements conformationnels. Leurs résultats ont montré une différence du volume de la poche catalytique durant la MD :

- Au début de la MD, le volume de la poche catalytique entre BACE1 et BACE2 est similaire (environ 350^3)
- Durant la simulation, le volume de la cavité de BACE1 augmentait tandis que le volume de la cavité de BACE2 diminuait
- À la fin de la simulation, le volume de la cavité de BACE1 était de $455Å^3$ contre $335Å^3$ pour BACE2, soit une différence de $120Å^3$.

Cette différence du volume de la poche peut ainsi provoquer un mode de liaison et une affinité différente entre BACE1 et BACE2, induisant l'inaccessibilité de certains sous-sites ou certains résidus du site actif.

1.6 . Objectif de la thèse

Comme il a été mentionné précédemment, la conception d'aptamères avec des modifications chimiques peut être réalisée par voie expérimentale, soit à posteriori de la génération d'aptamères par la méthode SELEX en procédant à des optimisations chimiques, soit par la génération des SOMAmers.

À ce jour, hormis l'outil NUCLEAR. il n'existe aucune approche computationnelle permettant de prédire des aptamères modifiés pour une cible thérapeutique donnée. Or, l'utilisation d'une telle approche peut apporter une aide non-négligeable. En effet, un nombre plus important de solutions (incorporant un nombre plus important de modifications chimiques) peut être étudié par voie computationnelle avec un coût limité en comparaison de la synthèse chimique. Dans la suite du document, nous appelons ces aptamères modifiés, conçus par approche computationnelle des ***in-silico*-mers**.

L'objectif initié avec cette thèse est donc double :

- Concevoir une nouvelle approche algorithmique pour le design d'ONs simple brin à partir de la connaissance 3D de la protéine cible et à partir d'une bibliothèque de fragments criblée sur cette dernière ;
- Démontrer l'intérêt de cette nouvelle approche au travers de différentes preuves de concepts incluant 7 protéines de liaison à l'ARN, et de l'appliquer sur une cible biologique d'intérêt (BACE1) impliquée dans la maladie d'Alzheimer au travers d'un pipeline *in-silico*.

Cette application sur BACE1 (et sur la protéine homologue BACE2) nécessite en amont de réaliser différentes analyses afin de mieux comprendre les caractéristiques jouant un rôle dans la spécificité. C'est pourquoi le travail de cette thèse sur BACE1 et BACE2 s'est focalisé sur des analyses de spécificité, afin de préparer en perspective une application concrète de design d'*in-silico*-mers avec notre méthodologie.

2 - Développement et Implémentation d'une nouvelle approche basée sur les Fragments et le Color-Coding : Docking, Design et Analyse Statistiques à l'Équilibre d'ARNsb liant des protéines

2.1 . Rappels

L'approche par Fragments est une approche généralement appliquée pour le design de composés chimiques et organiques. Comme mentionnée dans la Section 1.1, les fragments sont de petits composés d'une vingtaine d'atomes lourds, et pour lequel le principe consiste à les cribler sur un récepteur, à sélectionner ceux liant spécifiquement la cible biologique afin de créer une molécule-lead *in-fine*. Par ailleurs, cette approche a été aussi utilisée afin de prédire des complexes d'ARNsb de séquences connues avec des Protéines de liaison à l'ARN (RBP)[20,90,91] (Section 1.1.4). La prédiction des telles interactions reposent sur une librairie de fragments nucléotidiques qui doit être la plus diverse possible. Une telle diversité, incluant par exemple des nucléotides modifiés, est cruciale pour le design de molécules thérapeutiques afin d'atteindre l'activité désirée.

Par exemple, l'outil MCSS (Section 1.1.4) permet de cribler une librairie de mononucléotides à la surface de la cible biologique, avec une orientation et position connues. Cependant, l'assemblage de nucléotides successives en un ONs optimal, soit de séquence connue (docking d'ARNsb) soit de séquence inconnue (design d'ARNsb), ne peut être raisonnablement réalisé au travers de la méthode force-brute due à une explosion combinatoire. En effet, le succès de l'approche par Fragments dépend essentiellement d'une densité suffisante de poses qui, à son tour, a un impact sur le nombre de candidats. Le problème est aggravé par la prise en compte des nucléotides modifiés, augmentant ainsi la base d'une croissance exponentielle.

Dans ce travail de thèse, nous proposons de revisiter le docking et le design (Définitions 1.1.3) basés sur l'approche par Fragments au travers du prisme du Color-Coding (Section 1.3.1), qui permet de capturer la notion de l'auto-évitement pour le design d'ARNsb.
Nous montrons à la section 2.2 comment adapter cette technique et comment l'optimiser par une décomposition en cliques afin d'obtenir des algorithmes probabiliste ou exact pour le docking basé sur l'approche par Fragments au travers de la minimisation d'énergie.
La section 2.2.3 décrit comment réaliser un design en relaxant la contrainte de séquence avec une séquence de nucléotides donnée. Le "framework" est encore plus dur pour produire des statistiques à l'équilibre thermodynamique, pour l'étude de caractéristiques ("features") d'intérêts et utiles pour guider le design.

Enfin, nous proposons d'appliquer cette approche sur 7 complexes ARNsb/protéines, afin de discuter des points forts et des limitations de cette approche. Nous verrons en 4.1.1 les différentes extensions à envisager dans le futur.

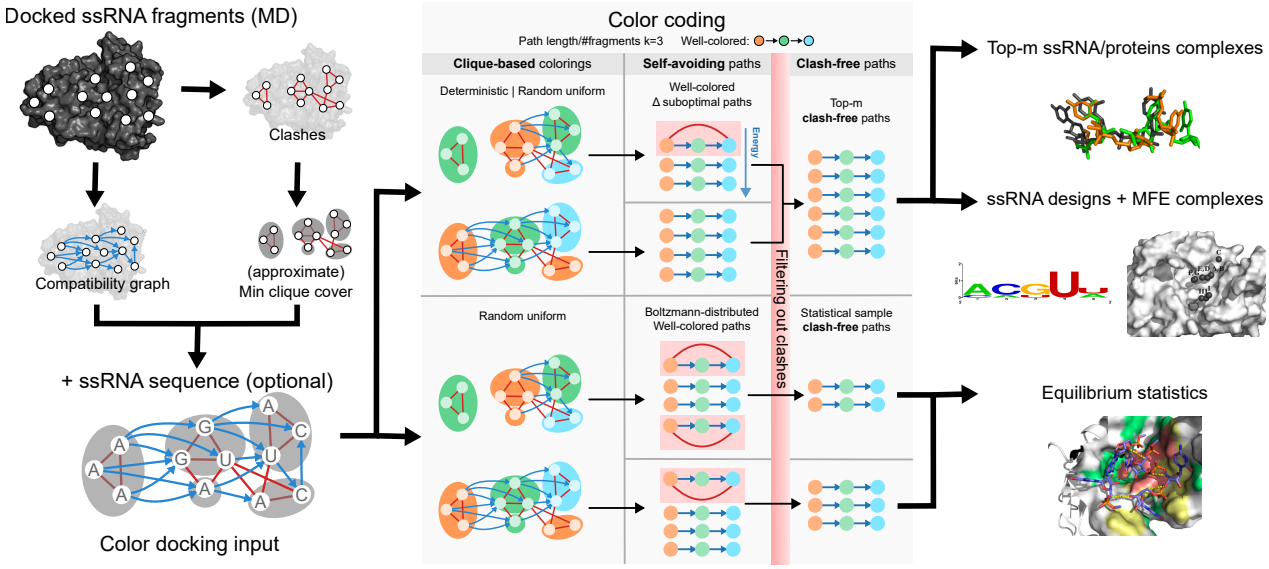


FIGURE 2.1 – Workflow Général : À partir d'un graphe de fragments criblés, pour lequel la matrice de connectivité a été évaluée par un outil externe, notre méthode considère des colorations – aléatoires ou déterministes – à partir desquelles des solutions peuvent être obtenues pour des problèmes computationnelles difficiles.

2.2 . Méthodologie et algorithmique

Posons les hypothèses explicites suivantes. Premièrement, notre docking est supposé être rigide au niveau protéique, ce qui indique que les fragments ARNsb peuvent être criblés individuellement sur la protéine sans perte excessive de précision. Deuxièmement, nous supposons que la longueur ou la composition en nucléotides de l'ARNsb, fournie en données d'entrée ("inputs"), interdit l'adoption de structure secondaires. Troisièmement, le système ARNsb/protéine peut maximiser la probabilité de la configuration jointe. Sous ces trois hypothèses, le docking et le design des ARNsb basés sur les fragments interagissant avec une protéine peuvent tous deux être reformulés comme un problème de graphe.

Définitions et notations. Nous désignons un **fragment** f comme un nucléotide $r(f)$ dans une séquence r d'ARN associée à une conformation 3D de référence. Une **pose** x représente un fragment criblé à la surface d'une protéine, et est définie par la position 3D de ses atomes relative à la protéine. Une paire ordonnée de poses se dit **compatible** si leurs occupations spatiales n'induisent pas de clashes géométriques insolubles, et permettent la connection séquentielle de deux fragments en un ARN plus long. La compatibilité est une relation orientée (associée avec la polarité de l'ARN), dont l'évaluation constitue un problème en soi et fait l'objet d'outils spécialisés pour traiter les ARNsb

courts.

À travers un criblage systématique de fragments à la surface d'une protéine cible, suivi d'une évaluation de connectivité de poses résultantes, on obtient un **graphe de compatibilité de poses**. Il est défini comme un graphe dirigé, c'est-à-dire une paire $G = (V, E)$ où V est un ensemble de poses, et tout arête dirigée $(v, v') \in E$ implique la possibilité de connecter v et v', Dans la suite, on note par $n := |V|$ le nombre de poses dans le graphe. Tout chemin du graphe de connectivité peut être associé à une conformation conjointe protéine/ARNsb, appelée **complexe** par la suite.

On associe ensuite une notion d'**énergie libre** $\Delta G(x)$ à tout complexe $x = (v_1, \ldots, v_k)$, définie comme :

$$\Delta G(x) = \sum_{i=1}^{k} \delta(v_i) + \sum_{i=1}^{k-1} \delta'(v_i, v_{i+1})$$

où $\delta : V \to \mathbb{R} \cup \{+\infty\}$ et $\delta' : V \times V \to \mathbb{R} \cup \{+\infty\}$ sont des termes spécifiques à une procédure de criblage de fragments, qui capturent les contributions individuelles de fragments individuels et connectés par pairs respectivement.

Cependant, quelques paires de fragments peuvent être en **clashes**, occupant des régions géométriques superposées ou trop proximales dans l'espace 3D, conduisant certains chemins du graphe de connectivité à ne pas toujours représenter des candidats prometteurs. Des cas triviaux d'un tel conflit se produisent au sein de complexes qui réutilisent deux fois la même pose. Au-delà de ces cas simples, les clashes peuvent être modélisés en utilisant une **fonction de clashes** $C : V \times V \to \{\mathsf{Vrai}, \mathsf{Faux}\}$. Un chemin x de longueur k est **auto-évitant**, aussi appelé un k**-chemin**, si ses noeuds sont des paires distinctes. Un chemin est **sans-clashes** si et seulement si ses noeuds sont des paires non-clashantes (*i.e.* $\forall 1 \leqslant i < j \leqslant k, C(x_i, x_j) = \mathsf{Faux}$). Notons que l'évitement des clashes induit un auto-évitement tant que, pour chaque pose v, on a $C(v, v') = \mathsf{Vrai}$.

Énoncé du problème et aspects de la complexité. Supposons être à l'équilibre thermodynamique, le complexe le plus stable/probable, pour une séquence donnée de nucléotides r de longueur k, est celui dont l'énergie libre est minimale appelée **MFE** ("Minimum Free-Energy"). De plus, l'assemblage de fragments doit être restreint aux complexes compatibles avec la séquence. Le calcul d'un tel complexe peut être reformulé comme suit :

Problème MFEDOCK
Input : Graphe de connectivité de poses $G = (V, E)$; Fonction de clashes $C : V \times V \to \{\text{Vrai}, \text{Faux}\}$; Fonction d'Energie ΔG ; Séquence de résidus $r = r_1, r_2 \dots r_k$.
Output : Complexe $x^\star = (v_1^\star, v_2^\star, \dots, v_k^\star)$ minimisant l'énergie libre

$$x^\star = \underset{\substack{x=(v_1,\dots,v_k) \text{ tel que} \\ v_i \neq v_j \forall i \neq j, \quad \leftarrow \text{auto-évitement} \\ C(v_i, v_{i+1}) = \text{Faux}, \forall i \quad \leftarrow \text{sans-clashes} \\ \text{et } r(v_i) = r_i, \forall i \quad \leftarrow \text{compatibilité séquence nucléotides}}}{\operatorname{argmin}} \Delta G(x)$$

La complexité de calculs, pour un graphe général, une séquence triviale (homopolymère, $k := |V|$, $C(\cdot,\cdot) = \text{Faux}$) et une fonction d'énergie à valeur unitaire ($(\delta(\cdot) = -1, \delta'(\cdot,\cdot) = 0)$), MFEDOCK résout le problème de décision de l'existence d'un CHEMIN HAMILTONIEN dans G, impliquant la NP-difficulté du problème. Le problème reste robuste, insoluble même quand il est restreint à une sous-classe de graphes d'entrées qui peuvent être tirés sur une surface protéique, tel que des graphes grilles[92]. En outre, pour un graphe de compatibilité complet et une énergie à valeur unitaire, la résolution de MFEDOCK répond à l'existence d'un ensemble k de noeuds non-clashants dans le graphe $(V, \{v, v' \mid C(v, v') = \text{Vrai}\})$, ainsi résolvant le problème MAX INDEPENDENT SET (MIS). Cependant, le problème MIS n'est pas uniquement NP-difficile, mais aussi insoluble (W[1] difficile pour k) du point de vue de la complexité paramétrée sur les instances géométriques (*e.g.* graphes issus d'intersections de segments/disques)[93] Ces résultats indiquent une dureté informatique et robuste du problème, motivant l'exploration d'alternatives et d'heuristiques.

2.2.1 . Garantir l'auto-évitement via une technique de color coding

Compte tenu de l'état de complexité du MFEDOCK, nous abordons initialement une version restreinte du problème qui considère uniquement des clashes résultant à partir de la réutilisation de certaines poses. En d'autres mots, nous optimisons l'énergie d'optimisation sur les chemins auto-évitants, ce qui vaut à fixer $C(\cdot,\cdot) = \text{Faux}$. Dans cette configuration, le problème algorithmique reste NP-difficile, mais le simplifie à un problème *Fixed-Parameter Tractable* (FPT), pratiquement solvable, pour les chemins de longueur k utilisant la technique du color-coding[27] (Section 1.3.1).

Les chemins bien coloriés, une alternative frugale en mémoire. Pour contourner cette exigence de mémoire substantielle, nous considérons une variante du color-coding basée sur les **chemins bien-coloriés**. Un chemin bien-colorié est un k-chemin dont les couleurs dans κ ne sont pas uniquement distincts, mais apparaissent dans un ordre spécifique, supposées être $1 \to 2 \to \cdots \to k$ sans perte de généralité. Par souci de simplicité, nous disons qu'une coloration κ **atteint** (resp. **rate**) un k-chemin x quand x est bien colorié (resp. non bien-colorié). Pour toute coloration donnée, l'opti-

mal/MFE k-chemin est obtenu en temps $\mathcal{O}(k.(n+|E|))$ (*e.g.* utilisant un algorithme de programmation dynamique). Celui-ci commence par calculer la MFE bien-coloriée pour une coloration donnée κ via la récurrence suivante :

$$\mathrm{mfe}_\kappa = \min_{\substack{v \in V \\ \text{tel que } \kappa(v)=1}} \mathrm{mfe}_\kappa[v, 1] \tag{2.1}$$

$$\mathrm{mfe}_\kappa[v, m] = \begin{cases} E(v) & \text{Si } m=k \\ \min\limits_{\substack{(v,v') \in E \text{ tel que} \\ \kappa(v')=m+1}} \delta(v) + \delta'(v, v') + \mathrm{mfe}_\kappa[v', m+1] & \text{Sinon} \end{cases} \tag{2.2}$$

Une fois la matrice mfe calculée en temps $\mathcal{O}(k.(n + |E|))$, une remontée (backtrack) peut être réalisée pour retrouver en temps $\mathcal{O}(k.n)$ une séquence de noeuds, distincts car bien coloriés, atteignant l'énergie minimale mfe_κ.

> Theorème 1: Pour une coloration κ donnée, le chemin bien colorié d'énergie minimale est calculable en temps $\mathcal{O}(k \times (n + |E|))$.

Démonstration. Prouvons d'abord la validité de l'équation de programmation dynamique ci-dessus. Nous n'explicitons pas l'opération de backtrack, qui peut être trivialement adaptée de celui des sous-optimaux, présenté en équation (2.5).

Plus particulièrement, nous considérons la propriété que, une fois la récurrence calculée, on trouve dans $\mathrm{mfe}_\kappa[v, m]$ l'énergie minimale d'un chemin bien colorié au regard de κ, commençant par le noeud v de couleur m, et ayant comme longueur $k' := k - m + 1$. Pour cela, on raisonne par récurrence, en remarquant que, pour $k' = 1$, le seul chemin bien colorié est réduit à v, et son énergie est alors $E(v)$ comme calculé par l'équation (2.2), en notant que $k' = 1$ implique $m = k$.

On pose alors l'hypothèse de récurrence, et on suppose que la propriété est vérifiée pour tout $k' \leqslant i$. Considérons alors les chemins de taille $k' := i + 1$, commençant par un noeuds prescrit v bien colorié ($\kappa(v) = m$). Au sein de ces chemins, le noeud v est suivi d'un noeud v', voisin de v ($(v, v') \in E$), lui-même bien colorié ($\kappa(v) = m + 1$) et suivi d'un suffixe de taille $k' = i$ débutant par v'. L'énergie minimale d'un tel chemin inclut la contribution propre $\delta(v)$ de v, l'énergie de liaison $\delta(v, v')$ de v et v', et l'énergie minimale du suffixe de taille i. Cette énergie est égale à $\mathrm{mfe}_\kappa[v', m + 1]$ par application de l'hypothèse de récurrence.

L'énergie d'un chemin bien colorié de taille i débutant par deux noeuds v et v' est donc égale à

$$\delta(v) + \delta(v, v') + \mathrm{mfe}_\kappa[v', m + 1].$$

En minimisant sur tous les voisins v' de v, on obtient exactement le calcul proposé par l'équation (2.2), ce qui démontre la correction de la formule pour $k' = i + 1$, et permet de conclure sur la validité pour tout k' de la récurrence pour $\mathrm{mfe}_\kappa[v, m]$, et donc de mfe_κ en minimisant l'énergie sur les alternatives disponibles pour le premier noeud.

En terme de complexité, le nombre de valeurs $\mathrm{mfe}_\kappa[v, m]$ à calculer est en $\mathcal{O}(n \times k)$. Le calcul d'une partie droite nécessite au pire n itérations de la minimisation, ce qui induirait une complexité en $\mathcal{O}(n^2 \times k)$. Cependant, on peut raffiner l'analyse en remarquant que, pour un m donné, deux valeurs de v nécessitent une minimisation sur des ensemble disjoints d'arêtes $(v, v') \in E$. Il s'ensuit que, en sommant sur tous les noeuds v possibles, le nombre total d'itérations des boucles de minimisation peut être borné par $|E|$ au lieu de n^2, d'où la complexité annoncée. □

Le fait d'être bien colorié ne contraint que les couleurs attribuées aux k noeuds de x, ce qui ne laisse qu'une seule possibilité parmi k^k colorations possibles, de sorte que la probabilité qu'une coloration uniforme aléatoire atteigne x est simplement $\mathbb{P}(x \text{ bien colorié}) = 1/k^k$. En itérant sur α colorations indépendantes tirées aléatoirement $\kappa_1, \ldots, \kappa_\alpha$, la probabilité qu'un chemin k soit manqué par tous les coloriages est alors

$$\mathbb{P}(x \text{ raté par } \alpha \text{ colorations aléatoires}) = \left(1 - \frac{1}{k^k}\right)^\alpha. \tag{2.3}$$

Cette propriété est valable pour tout chemin k, y compris le chemin MFE $x^\star$. En conséquence, pour toute tolérance cible $\varepsilon \in (0, 1)$, il suffit de fixer

$$\alpha := \left\lceil \frac{\log \varepsilon}{\log\left(1 - \frac{1}{k^k}\right)} \right\rceil \in \Theta(k^k \log \frac{1}{\varepsilon})$$

et nous obtenons le résultat suivant.

> Proposition 2: En absence de clashes ($C(-) = \mathsf{Faux}$), le problème MFEdock admet un algorithme probabiliste qui renvoie la solution optimale avec probabilité $1 - \varepsilon$, en temps FPT sur k et logarithmique sur ε.

Plus précisément, une itération sur le nombre requis de colorations induit une complexité temps en $\mathcal{O}(k^{k+2}\,(n + |E|)\,\log \frac{1}{\varepsilon})$, et une mémoire linéaire sur k et n.

La dérandomisation peut être aussi utilisée dans le contexte de chemins bien-coloriés. Ici, la construction de Alon et al.[94], couplée avec les premiers résultats de Schmidt et Siegel[95], fournit une famille

Data : Complexe de longueur k, Graphe de connectivité de poses $G = (V, E)$, Tolérance ε
Result : Complexe optimal pour MFEDOCK avec probabilité $1 - \varepsilon$

$M \leftarrow \left\lceil \frac{\log \varepsilon}{\log\left(1-\frac{1}{k^k}\right)} \right\rceil$;
$(mfe, mfe_{complex}) \leftarrow (+\infty, \varnothing)$;
for $i \in [1, M]$ **do**
    $\kappa \leftarrow$ `RandomUniformColor(V,k)`;
    $mfe_\kappa \leftarrow$ `ComputeMFE` (κ,k,G);
    if $mfe_\kappa < mfe$ **then**
        $mfe_{complex} \leftarrow$ `BacktrackMFE(`κ`,k,G,`mfe_κ`)`;
    end
end
return $mfe_{complex}$;

Algorithme 1 : Algorithme probabiliste pour MFEDOCK sans considération des clashes ($C(-) = \text{Faux}$)

de $k^{\mathcal{O}(k)} \log n$ colorations qui atteint chaque k-chemin, impliquant ainsi un algorithme déterministe exact pour le MFEDOCK sans contrainte de clashes.

Proposition 3: En absence de clashes ($C(-) = \text{Faux}$), le problème MFEDOCK peut être résolu en temps FPT sur k.

Cette complexité est maintenant en temps $\mathcal{O}(k^{\mathcal{O}(k)} n. \log n)$, marginalement plus élevée que pour les chemins coloriés, tout en utilisant une mémoire très réduite, linéaire sur le nombre de poses.

Rejet à partir de sous-optimaux pour produire des complexes MFE sans-clashes. De façon à récupérer les chemins MFE sans-clashes, et ainsi proposer une solution pour le MFEDOCK, nous proposons de l'extraire de la liste des sous-optimaux Δ **(auto-évitants)**, définie comme ayant une distance d'énergie d'au plus Δ à partir de la MFE auto-évitante. La liste des Δ-sous-optimaux peut être produite par l'utilisation d'une version adaptée du schéma de Waterman/Byers[96]. Une fois la matrice mfe calculée comme ci-dessus en temps $\mathcal{O}(k.(n + |E|))$, la liste exhaustive de Δ-sous-optimaux

Data : Complexe de longueur k, Graphe de connectivité de poses $G = (V, E)$, Tolérance ε
Result : Complexe sans clash optimal pour MFEdock avec probabilité $1 - \varepsilon$

$M \leftarrow \left\lceil \frac{\log \varepsilon}{\log\left(1-\frac{1}{k^k}\right)} \right\rceil$;
$(\text{mfe}, \text{mfe}_{\text{complex}}) = (+\infty, \varnothing)$;
subopts $\leftarrow \varnothing$;
$\Delta \leftarrow 0$;
step $\leftarrow 1$;
while $\Delta \leqslant$ *Emax* − *Emfe* + *step* **do**
  for $i \in [1, M]$ **do**
    $\kappa \leftarrow$`RandomUniformColor(V,k);`
    $\text{Emfe}_\kappa \leftarrow$ `ComputeMFE(`κ`,k,G);`
    $(\text{subopts}_\kappa, \text{Emax}) \leftarrow$ `ComputeSubopts(`κ`,k,G);`
    if $\Delta \geqslant$ *Emax* − *Emfe* **then**
      subopts $\leftarrow$ `BacktrackSubopts(`κ`,k,G,`$subopts_\kappa$`);`
    end
  end
  $\Delta \leftarrow \Delta$+step
end
return subopts;

Algorithme 2 : Algorithme probabiliste pour MFEdock retournant le complexe sans clashes d'énergie-libre minimale (MFE)

peut être obtenue en utilisant un *backtrack* modifié :

$$\text{subopts}_\kappa(\Delta) \to \bigcup_{\substack{v\in V \text{ s.t} \\ r(v)=r_1}} \text{subopts}_\kappa(v, 1; \Delta') \tag{2.4}$$
$$[\text{si } \Delta' := \Delta - (\text{mfe}_\kappa[v, 1] - \text{mfe}_\kappa) \geqslant 0]$$

$$\text{subopts}_\kappa(v, m; \Delta) \to \begin{cases} \{v\} \ [\text{si } m = k] \\ \bigcup_{\substack{(v,v')\in E \text{ s.t} \\ \kappa(v')=m+1 \\ \text{et } r(v')=r_{m+1}}} \{v\} \otimes \text{subopts}_\kappa(v', m+1; \Delta') \\ \qquad [\text{si } m < k \text{ et } \Delta' \geqslant 0] \end{cases} \tag{2.5}$$

Theorème 4: La procédure de backtrack décrite ci-dessus renvoie l'intégralité des chemins bien coloriés ayant une énergie dans $[\text{mfe}_\kappa, \text{mfe}_\kappa + \Delta]$.

Un tel backtrack s'exécute essentiellement en temps et mémoire en $\Theta(k\,D)$, où D est le nombre total de Δ-sous-optimaux, et est attendu croître exponentiellement avec Δ.

Un algorithme exact, en temps exponentiel dans le pire cas, peut être obtenu pour calculer la MFE sans-clashes. Il commence par calculer la MFE auto-évitante globale $E^\star_{SA}$, utilisant une famille dérandomisée $\boldsymbol{\kappa} := (\kappa_i)_i$ de colorations. Il itère ensuite plusieurs fois sur l'ensemble de la famille en utilisant une valeur croissante Δ jusqu'à

$$\Delta \geqslant \Delta_{\max} := E^-_{\text{sans-clashes}} - E^\star_{SA}, \tag{2.6}$$

où $E^-_{\text{sans-clashes}}$ dénote la structure sans-clashes d'énergie minimale, observée jusqu'à présent au sein des Δ-sous-optimaux sur $\boldsymbol{\kappa}$. Une fois l'inégalité ci-dessus satisfaite, l'algorithme peut alors simplement retourner le complexe MFE sans-clashes parmi les sous-optimaux engendrés à distance au plus $\Delta_{\max}$, c'est-à-dire la structure $S^\star$ avec une énergie $E^-_{\text{sans-clashes}}$.

Proposition 5: Une fois une valeur Δ atteinte telle que la condition (2.6) est satisfaite, alors la structure sans-clashes d'énergie la faible rencontrée représente une solution pour MFEdock.

Démonstration. En effet, pour tout complexe sans-clashes $S' \neq S^\star$, si S' est trouvé dans la liste combinée de Δ-sous-optimaux, alors il a une énergie plus élevée que $E^-_{\text{sans-clashes}}$ par définition. Si S' n'est pas listée comme un $\Delta_{\max}$-sous-optimal pour $\boldsymbol{\kappa}$, alors pour toute coloration κ qui atteint S', on a $\text{mfe}_\kappa + \Delta_{\max} \leqslant \Delta G(S')$. Comme $E^\star_{SA} \leqslant \text{mfe}_\kappa$, on conclut avec

$$\Delta G(S^\star) = E^-_{\text{sans-clashes}} \leqslant E^\star_{SA} + \Delta_{\max} \leqslant \text{mfe}_\kappa + \Delta_{\max} \leqslant \Delta G(S').$$

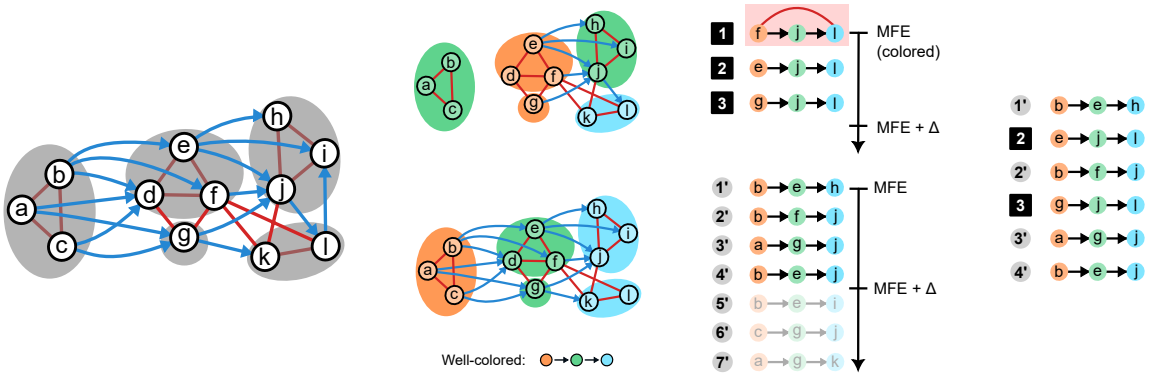


FIGURE 2.2 – **Exemple du color-coding basé sur les cliques monochromes.** À partir de l'instance MFEDOCK (A), incluant les arcs de compatibilités (bleues) et les arêtes de clashes (rouge), une couverture de cliques est calculée de façon heuristique (gris). Une famille de coloration est ensuite générée (B ; aléatoirement ou de façon déterministe), et un algorithme de programmation dynamique permet de construire une liste de k-chemins (auto-évitants) (C). Parmi ces chemins, seuls f → j → l présentent des clashes (boîte rouge) et sont filtrés pour obtenir une liste fusionnée de Δ-suboptimaux sans-clashes. Notamment, les deux colorations ci-dessus sont suffisantes pour atteinte tous les chemins sans-clashes. De plus, j → l → i, un k-path valide qui présente deux noeuds en clashes, ne peut être bien colorié dans la couverture de cliques actuelle.

De plus, l'énergie d'une telle alternative S' est alors supérieure à celle de $S^\star$, ce qui nous permet de conclure que notre algorithme est correct. □

Bien évidemment, les performances pratiques de l'algorithme vont dépendre, de façon critique, de la valeur $\Delta_{\max}$ atteinte, c'est-à-dire de la différence d'énergie entre la MFE auto-évitante et les complexes sans-clashes. Pour atténuer ce problème, nous introduisons dans la section 2.2.2 à venir une optimisation basée sur les cliques, que nous illustrons en Figure 2.2.

2.2.2 . Des cliques et des clashes : Réduire la densité des clashes grâce aux cliques monochromes

Durant la phase de docking initiale, les fragments individuels se regroupent généralement autour de hotspots à la surface de la protéine. D'une part, une telle accumulation est bénéfique pour la résolution des poses sélectionnées car, dans une certaine mesure, elle permet de simuler la flexibilité. D'autre part, une haute densité locale peut résulter à une explosion combinatoire de clashes, réduisant drastiquement la densité des complexes sans-clashes, tout en entravant les performances de notre algorithme. Nous préférons donc concentrer nos efforts sur un sous-ensemble de chemins auto-

évitants présentant une bonne densité de complexes sans-clashes.

Cliques de noeuds clashants peuvent être définies sur une seule couleur

Du fait de l'origine profondément géométrique de la notion de clashes, il est fréquent que les poses en situation de clashes deux à deux s'aggrègent sous la forme de cliques.

Definition 2.2.1 (clique clashante): Une clique clashante $\mathcal{C}$ est définie comme un ensemble de paires de poses clashant tous les uns avec les autres.

Nous remarquons alors trivialement que les noeuds d'une **clique clashante** $\mathcal{C}$, c'est-à-dire un ensemble de paires de poses en clashes, ne peuvent pas apparaitre plus d'une fois au sein d'un complexe sans-clashes.

De tels clashes peuvent être évités en limitant la génération aléatoire à des colorations **monochromes** par rapport à $\mathcal{C}$, c'est-à-dire utilisant une unique couleur pour tous les noeuds de la clique. Cette restriction ne réduit pas l'espace de recherche, constitué des k-chemins sans-clashes. En effet, tout k-chemin sans-clashes, comportant un noeud $v_c \in \mathcal{C}$ et atteint par une coloration arbitraire κ, est aussi atteint par une coloration clique-monochrome κ' construite telle que

$$\kappa' : v \to \begin{cases} \kappa(v_c) & \text{si } v \in \mathcal{C} \\ \kappa(v) & \text{sinon.} \end{cases}$$

Par ailleurs, tout k-chemin qui emprunterait deux noeuds ou plus à $\mathcal{C}$ présenterait alors un clashe. En interdisant l'exploration de ce type de chemins, la restriction des coloration aléatoires aux cliques monochrome augmente la densité de complexes sans-clashes au sein de l'espace de recherche.

Enfin, dans une perspective de dérandomisation, cette restriction autorise des chemins sans-clashes complets à être atteints par une petite famille de colorations, car les cliques entières peuvent alors être traitées comme des noeuds individuels. En appliquant cet argument de façon répétée, cette observation, et cette stratégie globale, se généralise à une collection de cliques disjointes dans le graphe de clashes.

Proposition 6: Etant donné une collection de cliques disjointes $\{C_i\}_i$, une restriction à des colorations monochromes sur chacun des C_i conserve la correction des algorithmes, probabilistes et exacts, basés sur le color coding.

En revanche, deux cliques $\mathcal{C}$ et $\mathcal{C}'$ qui se chevaucheraient ne peuvent pas être forcées à être simultanément monochromes. En effet, en raison de son chevauchement avec $\mathcal{C}'$, la couleur de $\mathcal{C}$ se propagerait à cette dernière, ce qui conduirait à traiter $\mathcal{C} \cup \mathcal{C}'$ comme une unique clique. Ceci amènerait alors à ignorer tous les k-chemins sans-clashes empruntant une paire de noeuds $v, v' \in \mathcal{C} \times \mathcal{C}'$, alors que v et v' pourraient être mutuellement compatibles ($C(v, v') = \text{Faux}$).

Afin de minimiser les temps d'exécution, et maximiser la densité de chemins sans-clashes dans l'espace d'exécution, nous prétraitons donc au préalable les clashes de façon à contraindre un maximum des cliques. Concrètement, nous décomposant G et C sous la forme d'une **couverture par cliques**, une partition de noeuds de V en un ensemble de cliques $\mathfrak{C} = \{\mathcal{C}_i\}$, tout en essayant de maximiser la proportion de clashes couverte par une clique $\mathcal{C}_i$. Bien qu'il ne soit pas strictement équivalent, ce problème est lié au problème de COUVERTURE DE CLIQUES MINIMALES, et très probablement difficile, d'où la mise en oeuvre d'une approche heuristique.

Une solution pragmatique pour décomposer les clashes en cliques non-chevauchantes

Pour résoudre de façon pragmatique ce problème, nous avons implémenté une **heuristique gloutonne de la couverture de la clique minimale** qui initialise la couverture $\mathfrak{C} := \varnothing$ et, à chaque itération, débute par le noeud v^+ ayant le noeud de degré maximal dans le reste du graphe. On initialise une clique $\mathcal{C} := \{v^+\}$ et une liste de voisins communs $\mathcal{N} := \text{voisins}(v^+)$. Ainsi, jusqu'à $\mathcal{N} = \varnothing$, il alterne :

1. Choisir un noeud $v \in \mathcal{N}$ de degré maximal dans $\mathcal{N}$, et est ajouté à $\mathcal{C}$;
2. Mettre à jour la liste des voisins communs par $\mathcal{N} := \mathcal{N} \cap \text{voisins}(v)$;

La clique $\mathcal{C}$ est ensuite ajoutée à la couverture $\mathfrak{C}$, retirée du graphe de cliques pour de futures itérations (choix de v^+, construction de $\mathcal{C}$...) jusqu'à que tous les noeuds aient été retirés du graphe de clashes, et ajoutés comme partie d'une clique à $\mathfrak{C}$. Bien que cette heuristique ne fournisse pas de garanties formelles quant à ses résultats, nous avons constaté qu'elle s'exécutait correctement pour nos instances typiques (voir section 2.2.5).

2.2.3 . Design rationnel d'ARNsb comme une relaxation du docking

La conception rationnelle dans le contexte du docking basée sur l'approche par fragments exige habituellement que l'aptamère conçu satisfasse deux propriétés : le **Design Positif** exige que le ligand ait une affinité optimale ou une faible énergie libre d'interaction, vis-à-vis de la protéine cible ou de la région de liaison cible ; le **Design Négatif** contraint le ligand à se lier spécifiquement à une région donnée de la protéine. Il est intéressant de noter que ces critères sont, au moins, partiellement pris en compte par une simple relaxation du problème MFEDOCK.

La modification requise consiste simplement à spécifier partiellement (*e.g.* code IUPAC), ou même à ignorer complètement (*e.g.* masque poly-N) la séquence de l'ARNsb sans complexité supplémentaire. Dans cette configuration, résoudre le problème MFEDOCK fournit un complexe MFE, pour lequel à la fois la séquence ARNsb $r(x_1).r(x_2)\cdots r(x_k)$ et sa conformation MFE peuvent être dérivées. Plus précisément, nous pouvons montrer que :

1. Aucune alternative de séquences a une meilleure affinité que $r^\star$ envers la protéine (design positif) ;
2. Le site de liaison induit à la surface de la protéine par $x^\star$ est la cible la plus problabe pour $r^\star$ (design négatif).

Notons que cette approche ne permet pas de cibler spécifiquement un site de liaison ou une poche, car la localisation du meilleur complexe est induit par le critère MFE. Elle permet néanmoins produire une variété de séquences qui sont à la fois stables et spécifiques à divers sites de la cible en générant des sous-optimaux et en conservant que la première occurrence de chaque séquence (*i.e.* associée à leur complexe MFE).

2.2.4 . Statistiques à l'équilibre thermodynamique

Bien que ce soit un problème important et difficile à résoudre sur le plan computationnel, le docking par minimisation d'énergie est limité par le fait qu'il se concentre sur la conformation MFE. En effet, à l'équilibre thermodynamique, la probabilité d'un complexe sans-clashes x suit une distribution de Boltzmann :

$$\mathbb{P}(X = x \mid r) = \frac{e^{-\beta.\Delta G(x)}}{\mathcal{Z}_r} \text{ où } \mathcal{Z}_r = \sum_{\substack{x' \text{ sans-clashes} \\ \text{and comp. with } r}} e^{-\beta.\Delta G(x')}$$

est la fonction de partition pour une séquence de nucléotides r, $\mu = RT$ avec R la constante de Boltzmann et T la température absolue. Car le nombre de complexes valides croît typiquement (au moins) de manière exponentielle avec k, la probabilité du complexe MFE devient excessivement petit dans un système de grande taille. À titre d'exemple extrême, pour un graphe de clique donné en entrée, le nombre de complexes croît dans $\Theta(n^k)$ quand $k \ll n$, et même $n! \asymp (n/e)^n$ quand $k = n$, ce qui réduit à la probabilité d'un seul complexe.

Statistiques de Boltzmann

Ceci motive de calculer des statistiques à l'équilibre thermodynamique, c'est-à-dire des propriétés attendues du système sous une distribution de Boltzmann. De telles propriétés sont mesurées par un ensemble de **fonctions caractéristiques ("features")** à valeurs réelles $\{f_1, f_2, \ldots\}$, chacune faisant correspondre un complexe valide à une valeur numérique dans $\mathbb{R}$. Les features peuvent représen-

ter n'importe quelle quantité pertinente (énergie libre, % d'occupation d'un site d'interaction ...), à condition qu'elles puissent être calculées efficacement à partir d'un complexe entièrement spécifié.

L'**espérance du feature** f est définie comme :

$$\mathbb{E}(f(X) \mid r) = \sum_{\substack{x \text{ sans clash} \\ \text{et comp. avec } r}} f(x) \times \mathbb{P}(X = x \mid r).$$

La probabilité d'occurrence d'une propriété peut être aussi calculée comme l' espérance d'un feature binaire (évaluée à 1 si propriété vérifiée et 0 sinon). Par exemple, en définissant $f_c(r) := 1$ ou 0 selon la présence/absence d'un contact avec un résidu A ciblé, l'espérance se simplifie alors en

$$\mathbb{E}(f_c(X) \mid r) = \sum_{x} f_c(x) \times \mathbb{P}(x \mid r) = \sum_{x \text{ s.t. } f_c(x)=1} \mathbb{P}(x \mid r) = \mathbb{P}\left(f_c(X) = 1 \mid r\right).$$

Les moments supérieurs des distributions peuvent finalement être calculés à partir des espérances $f, f^2, f^3 \ldots$ permettant l'accès à des caractéristiques plus fines de la distribution, telles que sa variance/écart-type, son coefficient d'acuité (kurtosis), son asymétrie... ou même des corrélations entre features multiples. D'un point de vue de la complexité, le calcul de la fonction de partition est manifestement plus difficile que le problème d'optimisation, déjà difficile. De plus, telles que définies pour l'optimisation, les familles de colorations utilisées pour la dérandomisation introduiraient typiquement un biais pour les estimations ultérieures, et ne peuvent donc pas être utilisées en l'état.

Estimateurs statistiques à partir des chemins bien-coloriés

Pour contourner ces obstacles, nous adoptons une approche qui estime l'espérance basée sur une séquence de colorations aléatoires $\kappa_1, \kappa_2 \ldots$. Pour cela, nous introduisons l'espérance restreinte aux couleurs d'un feature f étant donné une coloration κ telle que :

$$\mathbb{E}(f(X) \mid r, \kappa) = \sum_{\substack{x \text{ sans-clashes,} \\ \text{comp. avec } r \\ \text{bien col. par } \kappa}} f(x)\, \mathbb{P}(x \mid r, \kappa) = \sum_{\substack{x \text{ sans-clashes,} \\ \text{comp. avec } r \\ \text{bien col. par } \kappa}} f(x)\, \frac{e^{-\beta \Delta G(x)}}{\mathcal{Z}_{r,\kappa}},$$

où

$$\mathcal{Z}_{r,\kappa} = \sum_{\substack{x' \text{ sans-clashes,} \\ \text{comp. avec } r \\ \text{bien col. par } \kappa}} e^{-\beta \Delta G(x')}.$$

Fonction de partition bien coloriée et échantillonage statistique. Pour estimer cette quantité, nous introduisons tout d'abord un schéma de programmation dynamique pour calculer la fonction de partition, restreinte à la coloration et incluant la contribution de chemins en situation de clash :

$$\mathcal{Z}_\kappa = \sum_{\substack{v \in V \\ \text{tel que } r(v)=r_1}} \mathcal{Z}_{v,1} \tag{2.7}$$

$$\mathcal{Z}_{v,m} = \begin{cases} e^{-\beta E(v)} & \text{Si } m = k \\ \displaystyle\sum_{\substack{(v,v') \in E \text{ tel que} \\ \kappa(v')=m+1 \\ e t r(v')=r_{m+1}}} e^{-\beta(E(v)+E'(v,v'))} \times \mathcal{Z}_{v',m+1} & \text{Sinon} \end{cases} \tag{2.8}$$

À partir de la matrice $\mathcal{Z}_{v,m}$, précalculée par programmation dynamique en $\mathcal{O}(nk)$, un backtrack stochastique consiste alors à choisir de façon répétée un noeud avec probabilité proportionnelle à sa contribution, jusqu'à que la condition $m = k$ soit satisfaite, comme décrit dans l'Algorithme 3. Le complexe aléatoire ainsi produit suit alors une distribution de Boltzmann sur les k-chemins bien coloriés.

La valeur moyenne de f sur un ensemble de complexes aléatoires Boltzmann-distribués, filtré pour ne conserver que les chemins sans-clashes, constitue alors un estimateur naif, non-biaisé pour $\mathbb{E}(f(X) \mid r, \kappa)$.

Estimer la fonction de partition sans clashes globale. Une approche similaire peut être utilisée pour estimer la fonction de partition sans clashes. Considérons une séquence de colorations aléatoire indépendentes et uniformes $(\kappa_1, \ldots, \kappa_M)$. Pour chaque coloration κ, on peut estimer naivement la probabilité $\mathbb{P}(x \text{ sans clashes} \mid \kappa)$ qu'une structure, bien coloriée au regard de κ, soit sans clashes. Pour ce faire, on engendre une séquence $\mathbf{x} = (x_1, \ldots)$ de complexes aléatoires bien coloriés, et on renvoie la simple proportion

$$\hat{p}(\mathbf{x}) = \frac{\{x \in \mathbf{x} \mid x \text{ sans clashes}\}}{|\mathbf{x}|}.$$

Un estimateur non-biaisé pour la fonction de partition est alors obtenu via

$$\hat{z}((\kappa_1, \ldots, \kappa_M)) = \frac{k^{n-k} \times \sum_{i=1}^{M} \mathcal{Z}_{\kappa_i} \times \mathbb{P}(x \text{ sans clashes} \mid \kappa)}{M}.$$

Estimer les statistiques sans clashes. Nous en déduisons un estimateur final pour les statistiques à l'équilibre en l'absence de clashes. On définit

$$\widehat{g}(\kappa \mid r) = \frac{k^k}{\mathcal{Z}} \mathcal{Z}_\kappa \times \mathbb{E}(f(X) \mid r, \kappa)$$

Data : Longueur k, Graphe de connectivité de poses $G = (V, E)$, Coloration κ, Matrices de programmation dynamique $\mathcal{Z}_\kappa$ et $\mathcal{Z}_{v,m}$.
Result : Complexe bien colorié aléatoire, Boltzmann distribué, de taille k
Function `StochBacktrack`(v, m) **:**
  if $m = k$ **then**
    return [v];
  else
    $r \leftarrow$ `RandNum()` $\times \mathcal{Z}_{v,m}$; ▷ `Nombre aléatoire uniforme dans` $[0, \mathcal{Z}_{v,m}[$
    for $(v, v') \in E$ *tel que* $\kappa(v') = m + 1$ *et* $r(v') = r_{m+1}$ **do**
      $r \leftarrow r - e^{-\beta(E(v)+E'(v,v'))} \times \mathcal{Z}_{v',m+1}$;
      if $r < 0$ **then**
        return [v]+`StochBacktrack` (v',m+1);
      end
    end
  end
  return ;
$r \leftarrow$ `RandNum()` $\times \mathcal{Z}_\kappa$; ▷ `Nombre aléatoire uniforme dans` $[0, \mathcal{Z}_\kappa[$
for $v \in V$ *tel que* $r(v) = r_1$ **do**
  $r \leftarrow r - \mathcal{Z}_{v,1}$;
  if $r < 0$ **then**
    return `StochBacktrack` (v,1);
  end
end

Algorithme 3 : Backtrack stochastique pour les chemins bien coloriés

et

$$\widehat{f}(\kappa_1, \ldots, \kappa_M \mid r) = \sum_{i=1}^{M} \frac{\widehat{g}(\kappa \mid r)}{M}$$

Theorème 7: L'estimateur $\widehat{f}$ est sans biais, *i.e.* pour une séquence de colorations aléatoires indépendantes $(\kappa_1, \ldots, \kappa_M)$, on a :

$$\mathbb{E}(\widehat{f}(\kappa_1, \ldots, \kappa_M \mid r)) = \mathbb{E}(f(X) \mid r), \tag{2.9}$$

Démonstration. Commençons par raisonner sur l'espérance de $\hat{g}$ pour une unique coloration uniformément distribuée. On a

$$\begin{aligned}
\mathbb{E}(\widehat{g}(\kappa \mid r)) &= \sum_{i=1}^{k^n} \frac{1}{k^n} \times \frac{k^k}{\mathcal{Z}} \times \mathcal{Z}_{\kappa_i} \times \mathbb{E}(f(X) \mid r, \kappa_i) \\
&= \sum_{i=1}^{k^n} \frac{1}{k^{n-k} \times \mathcal{Z}} \times \mathcal{Z}_{\kappa_i} \times \sum_{x \in X_i} \frac{e^{-\beta.\Delta G(x)}}{\mathcal{Z}_{\kappa_i}} \times f(x) \\
&= \sum_{i=1}^{k^n} \frac{1}{k^{n-k} \times \mathcal{Z}} \times \sum_{x \in X_i} e^{-\beta.\Delta G(x)} \times f(x)
\end{aligned}$$

Or, sur l'ensemble des k^n colorations possibles, chacun des chemins apparait exactement k^{n-k} fois, et contribue toujours la même valeur $e^{-\beta.\Delta G(x)} \times f(x)$. On peut donc échanger les sommes de façon à obtenir :

$$\mathbb{E}(\widehat{g}(\kappa \mid r)) = \sum_{x \in X} \frac{e^{-\beta.\Delta G(x)}}{\mathcal{Z}} \times f(x) = \mathbb{E}(f(X) \mid r).$$

On remarque ensuite que l'espérance de l'estimateur $\widehat{f}$, évaluée sur M variables aléatoires indépendantes, est réductible à une somme sur des termes liés à l'espérance de $\widehat{g}$. En effet :

$$\mathbb{E}(\widehat{f}(\kappa_1, \ldots, \kappa_M \mid r)) = \mathbb{E}\left(\sum_{i=1}^{M} \frac{\widehat{g}(\kappa_i \mid r)}{M}\right) = \frac{\mathbb{E}\left(\sum_{i=1}^{M} \widehat{g}(\kappa_i \mid r)\right)}{M}$$

Les termes de la somme sont de simples applications de fonction à des variables aléatoires indépendantes. Les valeurs résultant de leur évaluation sont donc elles-mêmes des variables aléatoires indépendantes, et l'espérance de la somme est donc égale à la somme des espérances. On a donc :

$$\mathbb{E}(\widehat{f}(\kappa_1, \ldots, \kappa_M \mid r)) = \sum_{i=1}^{M} \frac{\mathbb{E}\left(\widehat{g}(\kappa_i \mid r)\right)}{M} = \mathbb{E}(f(X) \mid r)$$

□

2.2.5 . Résultats

Implémentation. Nous avons implémenté notre algorithme pour le MFEDOCK (optimisation ; subopts +/- contrainte de séquence), et l'estimateur statistique dans le logiciel `ColorDocking`, une collection de scripts `Python` à l'interface avec du code `C`, librement téléchargeable avec un jeu de données et des informations supplémentaires pour reproduire des expériences (`https://gitlab.inria.fr/amibio/colordocking`). Toutes les expériences ont été réalisées sur le cluster PBS avec un noyau Linux, avec entre 20GB et 50GB de mémoire réservée, chaque calcul ayant été réalisé sur un seul CPU.

Jeu de données. Nous avons sélectionné 7 complexes protéine/ARNsb pour valider notre méthode. Parmi eux, six complexes sont des complexes RRM-ARN (RRM ; `1B7F`, `1CVJ`, `2MGZ`, `2YH1`, `3NNH` et `4BS2`) et le reste étant un domaine Pumilio (PUF ; `3BX3`). Ceci coincide avec le benchmark sélectionné par De Beauchene et *al.*[20], avec deux structures en moins : `4N0T` qui interagit nativement avec un ARN double-brin ; et `5BZV`, un autre PUF que nous aurions vu comme redondant avec `3BX3`. Pour fournir une configuration réaliste au docking, les protéines ont été préparées et minimisées en utilisant le champ de force `CHARMM36` en absence du ligand ARNsb et de solvant ; ce qui résulte à une altération potentielle du site de liaison à l'ARN et notamment à une position spécifique de la chaine d'ARN.

Pour chacune de nos cibles, nous avons utilisé l'approche `MCSS`[97] pour générer une distribution de 10 000 poses. Ces poses ou fragments sont composés de l'Adenine (A), de la Cytosine (C), de la Guanine (G) et de l'Uracile (U). Seules les 4 000 poses (2 000 conformations ANTI et 2 000 conformations SYN) ont été utilisées pour cette étude. À partir de là, le package Python `NUCLEAR`[98] a été utilisé pour regrouper les poses similaires selon un seuil de 0.5Å RMSD, et pour générer les matrices (connectivité, clashes, scores) en utilisant une valeur de 4.5Å comme valeur maximale pour la distance O3'-C5'. La distance de contact maximale entre la protéine et le fragment a été définie à 3.5Å. Tous les atomes des acides aminés et des nucléotides ont été considérés, excepté un "patch terminal" qui a été omis pour améliorer la connectivité. Toutes les exécutions avec `NUCLEAR` ont été restreintes à la génération des différentes matrices (`run_type = partial`).

Un graphe dirigé (+ la matrice de clashes) a ensuite été généré et donné à notre implémentation du `ColorDocking`, en utilisant des collections de colorations uniformes et aléatoires au niveau des cliques. Nous définissons la tolérance à $\varepsilon = 0.01$, c'est-à-dire pour lequel la MFE sans-clashes est prédite avec probabilité $p = 99\%$. Des sous-optimaux ont été produits, sur la base d'un seuil de sous-optimalité $\Delta = 10$ kcal.mol^{-1}.

Complex	k	#Poses	Target Seq.	α	#Paths	#SA (cliques)	#sans-clashes
1B7F	5	2 171	UUUUU	14 388	5.22×10^8	1.49×10^8	6.98×10^6
1CVJ	5	2 031	AAAAA	14 388	9.44×10^7	2.24×10^7	1.46×10^6
2MGZ	5	4 329	GGUGU	14 388	1.84×10^7	4.89×10^6	3.20×10^5
2YH1	5	2 064	UUUUU	14 388	1.57×10^8	2.97×10^7	1.22×10^6
3BX3	5	7 464	UAUAU	14 388	1.42×10^8	5.39×10^7	5.21×10^6
3NNH	5	4 606	UUUUG	14 388	5.58×10^7	1.85×10^7	3.64×10^6
4BS2	5	8 150	GAAUG	14 388	2.37×10^7	8.65×10^6	1.03×10^6
1B7F	7	2 171	GUUUUUU	3 792 553	-	-	-
1CVJ	8	2 031	AAAAAAAA	77 261 93	-	-	-
1CVJ	8	5 785	AAAAAAAA	77 261 93	-	-	-
1CVJ	8	26 570	AAAAAAAA	77 261 93	-	-	-
2MGZ	7	4 329	UGGUGUG	3 792 553	-	-	-
3BX3	8	7 464	UGUAUAUA	77 261 93	-	-	-
3NNH	6	4 606	UUUUGU	214 856	-	-	-

TABLE 2.1 – Résumé de notre benchmark et analyses de cardinalité des chemins

Analyse de stabilité

La longueur de l'ARN $k = 5$ permet l'exécution de notre algorithme en temps modeste (environ 12 secondes par exécution). De tels temps de calcul autorisent une comparaison des résultats obtenus sur des exécutions successives, de manière à évaluer l'impact de la génération aléatoire des colorations sur la stabilité des prédictions. Ainsi, pour les 7 complexes et pour une configuration de $k = 5$, nous avons réalisé 100 expériences indépendantes. En addition, nous avons utilisé une approche brute-force pour calculer à la fois les complexes MFE auto-évitante et sans-clashes. Comme attendu, nous retrouvons systématiquement le complexe MFE sans-clashes dans nos expériences, à savoir entre 97/100 et 100/100 des expériences sur toutes les cibles. Un tel comportement est attendu, dû à notre choix de $\varepsilon = 0.01$, impliquant 1% de chance de rater la MFE, mais qui pourrait (au moins, en théorie) avoir été affecté par la configuration du paramètre Δ à une valeur fixe. La configuration du paramètre $\varepsilon = 0.63$ a réduit les temps de calcul par un facteur 10, au coût de performances dégradées, avec la MFE sans-clashes trouvée uniquement entre 25/100 et 47/100 des expériences.

Dans une seconde analyse, nous avons recherché si le top 100 des complexes sans-clashes présente une fort chevauchement au travers d'exécutions indépendantes de l'algorithme. Nous avons filtré l'output

pour produire 100 complexes sans-clashes de faible énergie pour une expérience. Nous avons itéré cette expérience 100 fois ; et nous avons constaté que le chevauchement moyen par paire entre deux séries était de 98%, avec des variations très limitées.

Impact de la couverture des cliques monochromes

Complex	#Poses	#Cliques	Max clique size	%Clique edges	Temps moy. (sec) +cliques	Temps moy. (sec) -cliques	$\lvert MFE_{\text{sans-clashes}} - MFE_{SA} \rvert$ +cliques	$\lvert MFE_{\text{sans-clashes}} - MFE_{SA} \rvert$ -cliques
1B7F	2 171	49	298	54.56	10.63	16.37	9.15	11.45
1CVJ	2 031	40	344	60.38	6.75	30.00	1.37	1.37
2MGZ	4 329	63	584	57.64	2.40	5.40	8.45	18.74
2YH1	2 064	49	312	66.40	8.60	15.53	5.2	9.6
3BX3	7 464	70	889	49.29	16.71	18.02	5.3	10.97
3NNH	4 606	55	629	57.62	3.61	12.73	2.52	7.25
4BS2	8 150	98	625	53.79	6.19	12.28	2.07	8.68
1CVJ	5 785	44	1 109	71.50	-	-		

TABLE 2.2 – Propriétés et impact de la couverture de cliques sur les temps d'exécution : valeurs observés pour k = 5, moyennés sur 100 itérations ((std dev ≈ 1).)

k	α	Temps moy. (sec) +cliques	Temps moy. (sec) -cliques
5	14 388	6.75	30.00
6	214 856	56	514
8	3 792 553	2.10^3	1.10^5

TABLE 2.3 – Impact de la couverture de cliques sur les temps d'exécution en fonction de la longueur k sur le système 1CVJ, avec 2 031 poses et 40 cliques

Nous avons ensuite observé l'effet de la coloration, basée sur les cliques, sur la densité des chemins sans-clashes, le temps de calcul et la distance énergétique entre la MFE auto-évitante et la MFE sans-clashes. Pour étudier ces différents points, nous avons aussi utilisé l'approche force-brute pour calculer le nombre de chemins auto-évitants et sans-clashes, sans contrainte.

Nous reportons dans le Tableau 2.2 les couvertures de cliques retournées par notre heuristique

gloutonne. Alors que le nombre de poses est de l'ordre du milliers, le nombre de cliques clashantes varie entre 40 et 100, avec la plus large clique représentant une proportion importante de l'ensemble des sommets (10% à 20%). De plus, une large proportion (50% à 70%) des arêtes clashantes sont internes à la clique. De tels clashes ne peuvent plus se reproduire si on se limite à une coloration basée sur les cliques, ce qui réduit considérablement la probabilité qu'un k-chemin présente un clashe.

Comme nous l'avons vu dans le Tableau 2.1, le nombre de chemins est généralement réduit de 75% quand les cliques monochromes auto-évitantes sont assurées ; ce qui résulte à une réduction du temps de calcul d'un facteur 1.5 à 4. De plus, comme le montre le Tableau 2.2, la distance énergétique entre la MFE auto-évitante et sans-clashes peut être grandement réduite (*e.g.* -10 kcal.mol^{-1} pour `2MGZ`) quand les cliques sont utilisées pour réduire l'espace de recherche.

Dans l'ensemble, la prise en compte des cliques monochromes représente un apport très positif : elle améliore considérablement les temps d'exécution et purifie les chemins auto-évitants pour augmenter la densité des chemins sans-clashes.

Docking par minimisation de l'énergie sous différentes définitions de fragments

En plus de l'ensemble de poses de mononucléotides obtenu avec `MCSS`, nous avons démontré la polyvalence de notre approche en utilisant l'ensemble de poses de trinucléotides (Section 1.1.4)[20]. Pour `1CVJ`, nous avons utilisé le logiciel `ATTRACT`[99] pour générer 1 000 000 de poses non-redondantes (0.2Å pour le seuil de RMSD) en représentation gros-grain. Nous avons construit la matrice de connectivité, en utilisant un critère seuil de 1.8Å pour le chevauchement des nucléotidiques entre poses consécutives. Nous avons ensuite créé un graphe dirigé de poses connectées, en utilisant le package `ssRNATTRACT`[20]. Nous avons utilisé une nouvelle librairie de fragments de trinucléotides d'ARN extraite à partir de la PDB avec le logiciel `ProtNAff`[19].

À partir des 1 000 000 de poses initiales, 5 327 peuvent être assemblées pour obtenir des chaines de taille 5. Ces poses ont été conservées dans le graphe de connectivité final. Nous avons ensuite construit une matrice de clashes de ces poses, en utilisant un critère de distance de 1.5Å entre les atomes lourds en clashes (en excluant les nucléotides chevauchants de poses connectées). De plus, nous avons considéré qu'une paire de poses est incompatible si elle ne peut être connectée uniquement qu'à la même position dans une chaine de taille 5, puisqu'elles ne peuvent être ensemble dans la même chaine. La matrice complète d'incompatibilité (paires de poses en clashes ou incompatibles) a été utilisée pour définir les cliques de poses incompatibles.

L'instance MFEdock résultante n'a besoin d'être exécutée que pour $k = 6$, car chaque fragment

représente ici un trinucléotide, et l'assemblage de 6 fragments est suffisant pour atteindre une solution composée de 8 nucléotides. Cela permet d'exécuter notre algorithme en seulement 34 secondes. En contraste, le temps d'exécution requis avec le jeu de données MCSS, impliquant de fixer le paramètre $k = 8$, est de 2×10^3 (soit approximativement 33 minutes).

Au-delà de la capacité à démontrer que le `ColorDocking` est capable de supporter différentes définitions de fragments, nous n'avons pas analysé la qualité des fragments produits (*e.g* la RMSD par rapport au complexe natif), car notre but n'a pas été de comparer les différents champs de force ou les différentes définitions de fragments.

Design basé sur MFEdock

Pour illustrer la capacité du MFEdock à aborder le design, nous avons considéré une étude de design récemment publiée par Perzanowska et *al.*[100], où un oligonucléotide ciblant une protéine de liaison poly-A (`code PDB: 1CVJ`) a été proposé. Cette étude a inclus des nucléotides modifiés, nous avons considéré une liste étendue de nucléotides : deux sans modifications (`A` et `G`), et cinq avec modifications : l'adenosine et la guanosine avec un phosphorothioate (A_P, G_P), l'adenosine protonée (A_ψ), le N6-methyladenosine (m^6A), le N6-methyladenosine incluant le phosphorothioate (m^6A_P), le O-methyladenosine (Am) et le O-methyladenosine incluant le phosphorothioate (Am_P). Tous ont été utilisés en conformations SYN et ANTI, pour un total de 1 000 poses (500 ANTI/500 SYN) par type de nucléotides. Un clustering a ensuite été appliqué à un seuil RMSD de 0.5Å , pour un nombre restant de 5 785 poses.

Pour générer des solutions de taille $k = 8$, nous avons considéré une valeur maximale de $\Delta_{max} := 3$, et graduellement augmenté Δ d'un pas unitaire, afin d'atteindre au maximum 100 solutions sans-clashes par coloration. Nous n'avons pas considéré initialement de contrainte de séquence. Le nombre unique de séquences obtenu a été de 75 sur 562 solutions sans-clashes. Nous illustrons ci-dessous le top 10 des solutions uniques, ainsi que leur énergie libre minimale :

A_ψ-A-A_P-G_P-A_P-m^6A-m^6A-A	-178.405	m^6A_P-m^6A-A-G_P-A_P-m^6A-m^6A-A	-178.404
A_ψ-A-G_P-G_P-A_P-m^6A-m^6A-A	-178.377	G-m^6A_P-m^6A-A-A_P-A_P-m^6A-m^6A	-178.055
A_ψ-A-A_P-G_P-A_P-Am-Am-m^6A	-178.052	A_{mP}-A_m-A-G_P-A_P-A_m-A_m-m^6A	-178.050
A_{mP}-A_m-A-A_P-A_P-A_m-A_m-A	-178.030	A_ψ-A-G_P-G_P-A_P–A_m-A_m-m^6A	-178.023
A_P-A_{mP}-A_m-A-G_P-A_P-A_m-A_m	-177.864	A_ψ-A-A_P-G_P-A_{mP}-A_m-A_m-A	-177.812

Il est intéressant de noter que ces séquences diffèrent de celles étudiées par Perzanowska et *al.*. En particulier, la paire de séquences ayant la meilleure affinité dans leur étude n'a pas été trouvée dans notre liste. Ceci n'est pas entièrement surprenant, car les auteurs se sont limités à un seul nucléotide

modifié par design dans leur étude.

Nous avons ensuite analysé leurs deux meilleures séquences, en exécutant une instance sous contrainte de séquence :

A-A-A-A-A-A-m^{6}A-A -161.252 m^{6}A-A-A-A-A-A-A-A -151.976

Nous constatons que leur MFE est significativement plus élevée (+16/+26 kcal.mol^{-1}), suggérant notre capacité à maitriser l'explosion combinatoire qui nous permet d'accéder à des alternatives prometteuses.

Estimateurs statistiques

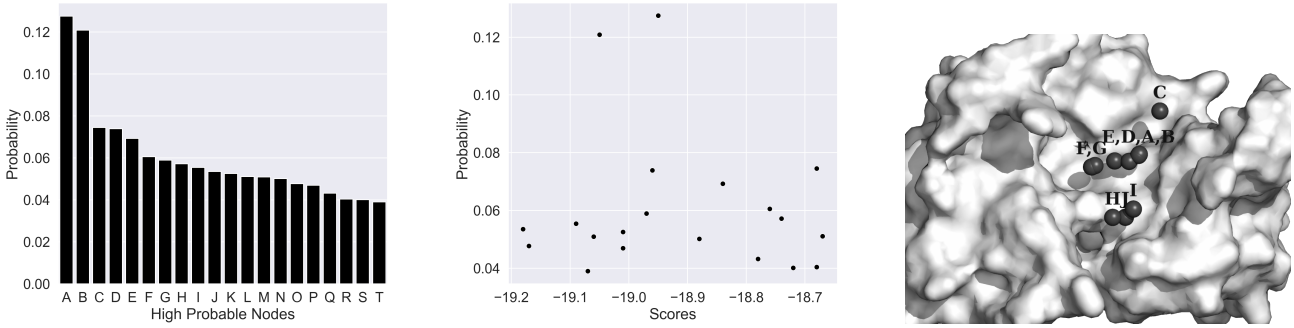
Identifier les "poses hautement probables" et les "profils de séquences"

Le design de molécules hautement spécifiques peut être conduit par des informations sur l'ensemble de structures ou de séquences, obtenues en utilisant les estimateurs introduits à la Section 2.2.4.

Nous avons étudié deux séries de features : i) identifier les poses hautement probables ; et ii) caractériser le profil de séquence induit par les complexes de faible énergie. Le premier correspond à un ensemble de fonctions binaires $\{f_\nu\}_{\nu \in V}$, chacune retournant 1 si le complexe utilise la pose ν, sinon 0, de sorte que l'espérance coïncide avec la probabilité de ν. Des features similaires permettent d'identifier la fréquence des nucléotides à chaque position dans le cadre du design.

Nous avons considéré `1CVJ` et un ensemble de poses restreint à une énergie libre en dessous de -18.74 kcal.mol^{-1}, pour un total de 6 262 poses. Un clustering de poses avec un seuil RMSD de 0.5Å a été réalisé. Nous avons recherché des chemins de taille $k = 5$, avec 1% de tolérance, et la pseudo-température β définie à 1.

Poses hautement probables. La Figure 2.3a montre les 20 poses les plus probables. Nous pouvons voir que deux poses (A et B) ont une probabilité supérieure à 10%. Les 18 autres poses ont une probabilité autour de 5%. Il est intéressant de noter que la localisation des poses A et B sont virtuellement indistinguable à la surface de la protéine (Figure 2.3c), et que les deux correspondent à une Adénine. Un complexe a ainsi 24% de probabilité de passer par cette région. Cette observation n'est pas triviale, et résulte d'un compromis délicat entre les énergies des poses et la combinatoire des complexes. En particulier, A et B sont chacune deux fois plus probables que les autres poses dans le top 20, malgré une contribution de l'énergie libre similaire (Figure 2.3b). Cela illustre donc le fait que les statistiques à l'équilibre fournissent des informations non-triviales.



(a) Probabilité des poses (Top 20)

(b) Énergie libre vs Probabilité (Top 20)

(c) Positionnement sur 1CVJ (Top 10)

FIGURE 2.3 – Étude statistique des poses hautement probables. (a) montre la probabilité du top 20 des poses, (b) montre le score du top 20 associé à leur probabilité, (c) montre la position du centre de masse des 10 premières poses hautement probables

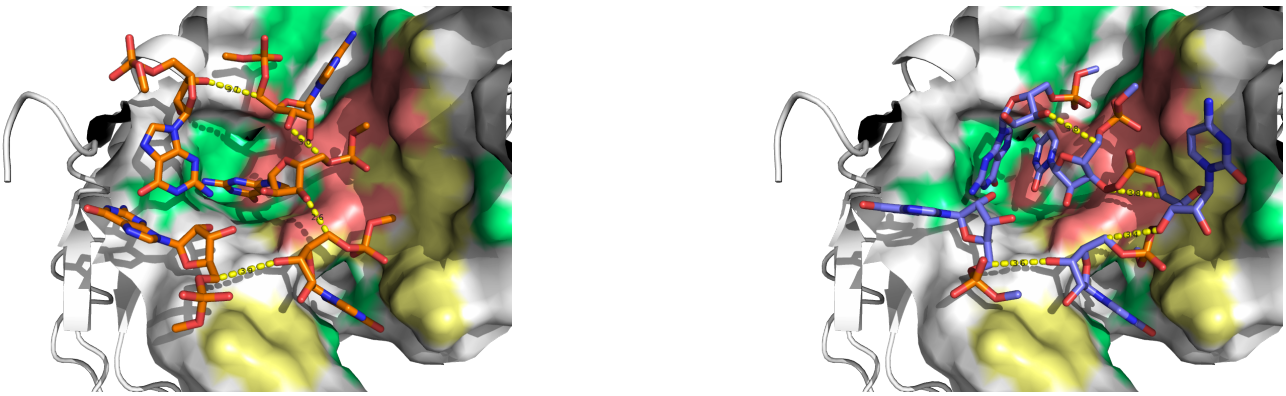
(a) Probabilité des nucléotides (Top 20)

FIGURE 2.4 – Étude statistique du profil de séquence. À la première position, les probabilités sont A :0.26, C :0.19, G :0.40, U :0.15. À la seconde position : A :0.35, C :0.15, G :0.39, U :0.11. À la troisième position : A :0.36, C :0.12, G :0.39, U :0.13. À la quatrième position : A :0.38, C :0.08, G :0.39, U :0.15. À la cinquième position : A :0.31, C :0.11, G :0.42, U :0.16.

Profil de séquences. À chaque position du 5-mer, nous avons estimé la probabilité d'observer le nucléotide le plus probable. Le profil de séquence résultant (Figure 2.4) montre la prédominance de la Guanine et de l'Adénine à chaque position. Ceci n'est pas une surprise, car le jeu de données comporte 42% de G et 31% de A. Plus généralement, nous observons une corrélation entre le nombre de poses pour un nucléotide et sa probabilité. Cependant, la probabilité de C (15% de poses) et de U (12% de poses) montre une tendance intéressante, avec C plus probable que U pour les deux premières positions du 5-mer (=4% et +3.7%), et étant ensuite de plus en plus dominée pour les trois dernières positions (-1.4%, -7.7% et -5.3%). Ceci confirme la capacité de l'estimateur à révéler des effets coopératifs non-triviaux.

Design basé sur l'estimateur statistique.

Une stratégie pour le design est d'exploiter des résultats de l'étude statistique (2.2.5) pour concevoir des solutions de taille $k = 5$. À partir du profil de séquence, nous avons recherché un 5-mer dont la séquence est composée soit d'une Guanine soit d'une Adénine à chaque position. Comme il est montré à la Figure 2.5, la séquence de meilleure énergie trouvée avec cette stratégie est GAGGG (Figure 2.5a). Il est intéressant de noter que cette MFE correspond à la même séquence et la même conformation obtenue avec une stratégie à l'aveugle, c'est-à-dire sans contrainte de séquence. Le score énergétique est égal à -117.37 kcal.mol^{-1}.



(a) GAGGG (-117.37 kcal.mol $^{-1}$) (b) GUCGG (-102.85 kcal.mol $^{-1}$)

FIGURE 2.5 – Représentation 3D de la poche de liaison pour le design de 5-mers inspiré par l'estimation statistique de l'espace de séquence.

Une autre solution a été proposée à partir des poses hautement probables, composée d'une Uracile à une position où la fréquence de ce fragment est égal à 7%. L'énergie libre associée est de -102.85 kcal.mol^{-1}. Ces résultats montrent que les propriétés statistiques peuvent être exploitées pour suggérer différentes solutions, et comparer leur conformation et leur énergies. Par exemple, la conformation

GAGGG montrée à la Figure 2.5a serait globalement meilleure en terme d'affinité que la conformation GUCGG montrée à la Figure 2.5b. Néanmoins, les régions d'interaction ne sont pas réellement identiques, et notre méthode fournit une information clé sur les interactions préférentielles entre une sous-région et un nucléotide particulier (ou famille de nucléotides).

2.2.6 . Conclusions

Nous avons introduit une nouvelle approche algorithmique basée sur la technique du color-coding pour une application de docking et de design d'*in-silico*-mers (ARNsb) basée sur l'approche par Fragments. La complexité de notre algorithme exacte est linéaire sur le nombre de poses connectées par paires, et exponentielles que sur la longueur `k` de l'ARNsb cible. Ainsi, notre algorithme peut être vu comme un algorithme de complexité paramétré.

Nous avons appliqué le `ColorDocking` sur 7 complexes ARNsb/protéines, en montrant l'utilité des différentes extensions que sont le "docking", le "design" et "l'estimateur statistique" ; et de son applicabilité. Cette nouvelle méthode peut donc faire l'objet d'une évaluation plus poussée sur de réelles études de cas, pour le design thérapeutique d'ARNsb contre une cible thérapeutique, et pour lequel nous proposons une première application dans le Chapitre 3.

3 - Application sur le modèle d'étude la Beta-Sécrétase 1 (BACE1)

3.1 . Rappels

Comme mentionnée dans l'introduction et les objectifs de thèse, la β-Sécrétase 1 (BACE1) a été sélectionnée comme modèle d'étude afin de démontrer la faisabilité de la conception de pseudo-aptamère par l'application de l'approche `ColorDocking`. Ce modèle d'étude inclut également l'enzyme homologue à BACE1, à savoir la β-Sécrétase 2 (BACE2). Cette dernière sert *in-fine* de validation négative pour s'assurer que les *in-silico*-mers prédits comme spécifiques vis-à-vis de BACE1 interagissent avec une affinité moindre vis-à-vis de BACE2, au travers par exemple d'une Dynamique Moléculaire ou du module "Docking" du `ColorDocking`.

Néanmoins, au vu du nombre de nucléotides différents présents dans la bibliothèque et au vu du nombre de modes de liaison pouvant être prédits, une analyse de spécificité peut être réalisée en amont au travers de deux études.

La première étude repose sur la compréhension de la relation structure-activité entre des ligands considérés comme affins vis-à-vis de BACE1 et de BACE2. Ces ligands correspondent à des dérivés de N-Phenylpicolinamide, un pharmacophore identifié pour l'approche QSAR. L'approche QSAR est une méthode appartenant aux approches LBDD (Figure 1.3, section 1.1.3), qui consiste à concevoir des molécules thérapeutiques à partir de la connaissance unique de ligands interagissant avec la cible biologique, et en absence de données 3D sur cette dernière. L'approche QSAR permet de créer des modèles mathématiques pour lesquels des descripteurs, définis au travers de ces modèles, permettent d'expliquer la spécificité associée à des propriétés physico-chimiques d'une famille de ligands. Bien évidemment, la modélisation d'un modèle QSAR pour une famille de ligands donnée nécessite d'avoir une quantité de données importantes, c'est-à-dire avoir suffisamment de ligands appartenant à cette famille interagissant avec la ou les protéines d'intérêt. Dans le cadre de BACE1 et BACE2, seule la famille de composés N-Phenylpicolinamide contient suffisamment d'inhibiteurs inhibant à la fois avec BACE1 et BACE2 pour permettre une étude QSAR. Néanmoins, comme nous le verrons succinctement dans la section 3.2.1, les modèles QSAR décrivant les propriétés spécifiques des ligands vis-à-vis de BACE1 et de BACE2 sont similaires[101], et n'apportent pas d'informations claires sur les contacts que peuvent établir ces ligands avec les deux enzymes.
Ainsi, nous proposons une analyse complémentaire afin de vérifier à la fois la cohérence des modèles QSAR en termes de sélectivité, mais aussi d'apporter d'autres éléments qui permettraient d'expliquer la spécificité des composés au travers d'une étude de docking. Ces résultats pourront être utiles pour le design d'*in-silico*-mers.

La deuxième étude repose principalement sur l'étude des régions favorables aux nucléotides et à

l'identification de nucléotides spécifiques de BACE1 et de BACE2. Ces données apporteront des informations sur les nucléotides à incorporer ou à prioriser pour le design d'*in-silico*-mers spécifiques de BACE1 ; et des groupes nucléotidiques à éviter qui seraient spécifiques à BACE2. Nous proposons aussi une première application du `ColorDocking` pour le design d'*in-silico*-mers en incorporant les nucléotides spécifiques à BACE1.

Ces résultats (sous-sections 3.2.3 et 3.3.8) serviront de bases pour démontrer en perspective l'applicabilité du `ColorDocking` dans la conception thérapeutique d'*in-silico*-mers par approche computationnelle.

Enfin, nous verrons dans les perspectives à envisager en termes d'analyses ou d'études futures pour le design d'*in-silico*-mers (Section 4.1.2).

3.2 . Application sur la β-sécrétase 1/2 : Étude d'interactions au travers du docking d'inhibiteurs spécifiques de BACE1 et BACE2

3.2.1 . Sélection de composés spécifiques vis-à-vis de BACE1 et BACE2

Une étude QSAR a été réalisée sur une série de 76 dérivés de N-Phenylpicolinamide interagissant avec les sites actifs de BACE1 et BACE2. Le choix de ce pharmacophore est dû principalement à la disponibilité des données. En effet, une étude QSAR repose sur un apprentissage de données existantes. La construction d'un modèle QSAR vis-à-vis de BACE1 et BACE2 nécessite alors de choisir des composés interagissant à la fois avec BACE1 et BACE2, et pour lesquels les activités sont connues. Or, seule la famille des composés N-Phenylpicolinamide réunit suffisamment de composés interagissant avec les deux cibles.

Le but de cette étude QSAR est de définir un modèle capable de discriminer les ligands sélectifs de BACE1 ou de BACE2 à partir de descripteurs définis (Figure 3.1)[101]. Ces descripteurs consistent à décrire des propriétés afin de prédire si un composé dérivé du pharmacophore est spécifique de BACE1 ou de BACE2 (voir Tableau 3.1). Néanmoins, les modèles BACE1 et BACE2 disposent de descripteurs globalement similaires, reposant sur des descripteurs de lipophilie, stérique ou électronique. Ainsi, aucun descripteur associé spécifiquement à une sous-structure du pharmacophore ne permet réellement d'expliquer la spécificité des composés vis-à-vis de BACE1 ou de BACE2, comme nous l'avons vu à la Section 3.1.

Ainsi, pour compléter ces modèles QSAR, une étude de docking est menée pour identifier les différentes interactions (*i.e* les acides aminés jouant un rôle clé) pouvant expliquer la spécificité de ces

Descripteurs Modèle BACE1 |	Descripteurs Modèle BACE2
Électronique	Stérique
Stérique	Stérique
Électronique	Lipophilicité-Stérique-Électronique
Lipophilicité	Lipophilicité-Stérique-Électronique
Lipophilicité-Stérique-Électronique	Lipophilicité-Stérique-Électronique

TABLE 3.1 – Les descripteurs des modèles QSAR pour BACE1 et BACE2

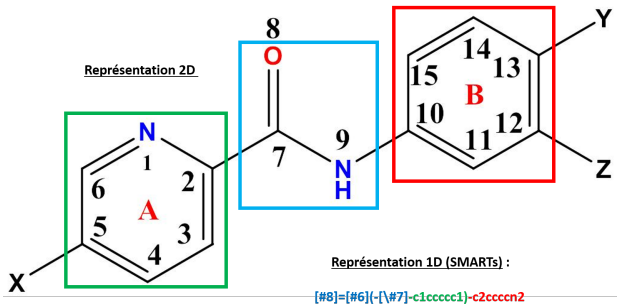


FIGURE 3.1 – Pharmacophore N-Phenylpicolinamide. Il est composé de deux cycles (hétéro)aromatiques, une pyridine et un phényl, liés par un linker amide. Le format 1D SMARTs correspondant au format 2D est aussi spécifié.

dérivés vis-à-vis de BACE1 ou de BACE2.

Pour ce faire, parmi la liste des dérivés utilisés pour l'étude QSAR, 5 composés spécifiques de BACE1 et 5 composés spécifiques de BACE2 ont été sélectionnés sur la base du rapport de leur pIC50 ($\frac{\text{pIC50(BACE1)}}{\text{pIC50(BACE2)}}$) (voir tableau 3.2). Cette métrique désigne le logarithme négatif de l'IC50, la concentration inhibitrice en molaire à partir de laquelle 50% de l'activité est inhibée. Ainsi, si le rapport de leur pIC50 est supérieur à 1, alors le composé est spécifique à BACE1 ; sinon il est spécifique à BACE2.

Le pharmacophore a été converti en format 1D SMARTs ("SMILES arbitrary target specification") (Figure 3.1), obtenu en utilisant le sketcher de PubMed (https://pubchem.ncbi.nlm.nih.gov//edit3/index.html). Le SMARTs correspond à un format 1D d'une sous-structure ("pattern") d'une molécule organique, ce qui permet de rechercher des composés incorporant la sous-structure recherchée (https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html). À partir de l'ensemble de ligands complexés à BACE1/BACE2 contenu dans la base de données PDB, nous pouvons sélectionner les complexes en interaction avec des ligands dérivés du pharmacophore en utilisant l'outil OpenBabel.

Composés	pIC50(BACE1)	pIC50(BACE2)	$\frac{\text{pIC50(BACE1)}}{\text{pIC50(BACE2)}}$
B11	9.22	7.40	1.25
B13	7.82	8.52	0.92
B19	8.21	9.30	0.88
B45	8.21	8.92	0.92
B50	7.33	6.52	1.12
B51	9.00	7.85	1.15
B61	8.04	9.00	0.89
B69	8.66	7.78	1.11
B70	8.01	7.18	1.12

TABLE 3.2 – Les 5 meilleurs composés spécifiques à BACE1 et les 5 meilleurs composés spécifiques à BACE2, avec leur pIC50 ainsi que leur rapport indiqués.

Ces complexes serviront de récepteurs (de références) pour l'amarrage des 10 composés indiqués dans le tableau 3.2. Il est important de noter ici que le SMARTs permet une sélection de ligands composés strictement du motif recherché.

3.2.2 . Sélection des différentes conformations de BACE1 et BACE2

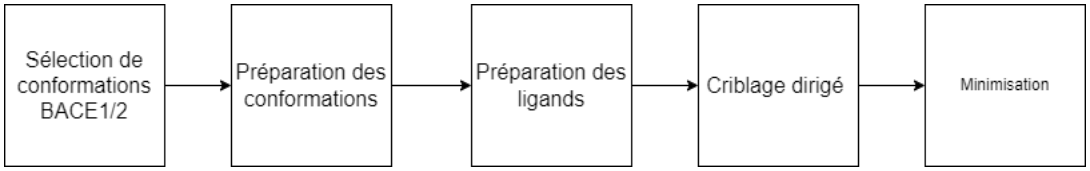


FIGURE 3.2 – Workflow appliqué pour l'étude de Relation Structure-Activité de ligands dérivés N-Phenylpicolinamide

Sélection des différentes conformations de BACE1 et BACE2

La totalité des conformations de BACE1 d'une part et de BACE2 d'autre part, issues de la base de données PDB, a été obtenue par une recherche avancée en renseignant les mots-clés "Beta-Sécrétase 1" ou "Beta-Sécrétase 2" dans le champ "Macromolécule".

Pour discriminer et sélectionner les conformations de BACE1 et de BACE2 en interaction avec un ligand dérivé du pharmacophore, différents filtres ont été appliqués sur tout ligand (au format SDF, 2D) n'appartenant pas au tableau 3.3. Cette liste définie toute molécule présente dans la structure cristalline (molécules de solvant par exemple), et n'étant pas à un inhibiteur.

Le premier filtre consiste à sélectionner uniquement les ligands dérivés du pharmacophore. Pour ce faire, l'outil OpenBabel est utilisé afin de comparer chaque ligand au format SDF (2D) avec le pharmacophore au format SMARTs (1D). Rappelons que seuls les ligands possédant strictement le motif recherché sont conservés dans la sélection (**Sélection A**).

Le deuxième filtre est la résolution avec un seuil de 2.0Å pour ne conserver, parmi cette sélection A, que les complexes qui ont été résolus à haute résolution (*i.e.* une résolution inférieure ou égale au seuil) (**Sélection B**).

Le troisième filtre est le coefficient de similarité, appelé également le coefficient de Tanimoto. Ce coefficient de similarité permet d'identifier le ligand de référence le plus similaire (*i.e.* coefficient de similarité le plus proche de 1) aux composés du tableau 3.2. En effet, l'idée consiste à trouver une référence parmi la sélection B pour les composés du tableau 3.2 sous l'hypothèse suivante : la conformation du récepteur de référence adopte une conformation similaire s'il interagit avec le composé du tableau 3.2. Ce faisant, l'amarrage des composés du tableau 3.2 n'en sera que plus précise et pertinente. Ce filtre a été appliqué avec OpenBabel. (**Sélection C**)

Enfin, le dernier filtre consiste à vérifier si des coordonnées ne sont pas manquantes au niveau du site actif des récepteurs issus de la sélection C. Le site actif est défini comme tout acide aminé à 6.0Å tout-atome du ligand. Dans le cas où des coordonnées seraient manquantes, une conformation alternative du récepteur le plus proche est sélectionnée parmi la liste de toutes les conformations avec une résolution inférieure ou égale à 2.0A en se basant sur la métrique du score de clashes entre la conformation alternative et le ligand de référence obtenue avec le serveur web MolProbity[102].

Par des questions pratiques, on appellera dans la suite du document "**référence**", le ligand RX sur lequel les composés dérivés N-Phenylpicolinamide sont alignés.

Définition de l'état de protonation des histidines, acides aspartiques et acides glutamiques du récepteur

L'état de protonations de la chaine latérale des histidines, des acides aspartiques et des acides glutamiques ont été déterminés. En effet, l'état protoné ou non-protoné peut jouer un rôle indirect lors de l'étape de minimisation (par des effets longues distances) ou jouer un rôle dans l'interaction clés avec les ligands s'ils constituent le site actif.

Les états de protonation ont été déterminés au travers d'un rapport du pKa prédit (obtenu avec l'outil Propka en version 3.4.0, `https://github.com/jensengroup/propka`,[103]) par rapport au pH de

Code des résidus	Nom complet
NA	SODIUM ION
CL	CHLORIDE ION
IOD	IODIDE ION
ACT	ACETATE ION
ZN	ZINC ION
MG	MAGNESIUM ION
NI	NICKEL ION
PCA	PYROGLUTAMIC ACID
TLA	L(+)-TARTARIC ACID
TAR	D(-)-TARTARIC ACID
PEG	DI(HYDROXYETHYL)ETHER
DMS	DIMETHYL SULFOXIDE
GOL	GLYCEROL
POL	N-PROPANOL
IPA	ISOPROPYL ALCOHOL
ACE	ACETYL GROUP
PO4	PHOSPHATE ION
SO4	SULFATE ION
EDO	1,2-ETHANEDIOL
ASP	ASPARTIC ACID
URE	UREA
LPD	L-PROLINAMIDE
1OL	5-amino-4-hydroxy-2,7-dimethyloctanoic acid
STA	STATINE
EV0	2-amino-6-propylpyrimidin-4(3H)-one
DPR	D-PROLINE
NH2	AMINO GROUP

TABLE 3.3 – Liste d'exclusion de molécules et ions présents dans les différentes conformations de BACE1 et BACE2

résolution de chaque conformation. Les règles suivantes ont été appliquées :

- **Pour les acides aspartiques et glutamiques**. Quel que soit le pH, si le pKa prédit est inférieur à 0.1 unité de pH, alors la chaine latérale est non-protonée ; sinon la chaine latérale est protonée. Si le pKa prédit est compris dans un intervalle de 0.1 unité de pH, alors la méthode par RMSD

locale suite à une minimisation est appliquée. Elle consiste à réaliser une minimisation pour chaque état de protonations possible, et de calculer une RMSD locale comprenant les acides aminés à 6.0Å, de l'acide aspartique ou glutamique cible. L'état de protonations pour lequel la RMSD locale est le plus faible est sélectionné.

- **Pour les histidines**. À pH acide, si le pKa prédit est supérieur à 1 unité de pH, alors l'histidine est protonée. Sinon, il peut être protoné ou non-protonée. À pH neutre, si le pKa est inférieur à 1 unité de pH, alors l'histidine n'est pas protonée. Sinon, il peut être protoné ou non-protonée. Dans le cas d'une ambiguité entre deux états de protonations possible, la méthode par RMSD locale suite à une minimisation est appliquée. Afin de prendre en compte un échange de proton entre des histidines voisines, une matrice de distance a été calculée entre chaque histidine. Une paire d'histidines à une distance inférieure de 10.0Å l'une de l'autre doit être traitée comme dépendante.

Remodélisation des boucles manquantes du récepteur

Une remodélisation des résidus manquants du récepteur est faite avec le serveur web CHARMM-GUI (`www.charmm-gui.org`), en utilisant le module **PDB Reader & Manipulator**. Le choix de cet outil repose sur le fait que les résidus manquants ne sont pas en contact direct avec le site actif, et ne correspondent pas à des acides aminés chargés. Seule l'option "Model missing residues" a été sélectionnée/cochée afin de reconstruire uniquement les résidus manquants en milieu de chaîne. Autrement dit, les résidus manquants aux extrémités N-ter et C-ter n'ont pas été remodélisés.

Définition des états de solvatation autour du récepteur

L'état de solvatation est défini ici comme le nombre de molécules d'eau autour d'une protéine, et dépendant de la "force de liaison" de ces dernières vis-à-vis du récepteur. Deux états de solvatation ont été définis : "basse solvatation" considérant uniquement les molécules d'eau fortement liées au récepteur, et "haute solvatation" considérant les molécules d'eau faiblement liées au récepteur en plus des molécules d'eau définies à basse solvatation.

Pour obtenir ces deux états de solvatation, l'approche EDIA ("Electron Density Score for Individual Atoms") a été appliquée afin d'estimer la densité électronique d'un atome individuel au travers d'un score compris entre 0 et 1.2[104]. Dans le cas des molécules d'eau, le score EDIA est associé à l'atome d'oxygène. Au plus cette valeur est élevée, et au plus la molécule d'eau est considérée comme très stable. Ainsi, on considère :

- un état à basse solvatation toute molécule d'eau ayant un score EDIA supérieur ou égal à 0.81
- un état à haute solvatation toute molécule d'eau ayant un score EDIA supérieur ou égal à 0.73

Le choix de ces valeurs est expliqué dans la section 3.3.8. Dans le cadre de cette étude, seule la haute solvatation est considérée.

Afin de générer ces états de solvatation, le module EDIA du serveur web Protein Plus (https://proteins.plus/) est utilisé. Pour chaque conformation, un fichier PDB est généré dans lequel les scores EDIA sont indiqués dans la colonne du b-facteur. À l'aide d'un script PyMol, la sélection des molécules d'eau suit les étapes suivantes :

1. Élimination des molécules d'eau dont la distance tout-atome est supérieure à 4.0Å du récepteur
2. Élimination des molécules d'eau avec un score EDIA inférieur à 0.73

Criblage dirigé et minimisation

Criblage. Le criblage dirigé est défini comme la reproduction du mode de liaison du ligand de référence pour un dérivé de N-Phenylpicolinamide. Nous avons vérifié, en amont du docking, la géométrie de la liaison amide (isomérie trans-) et du substituant R2 pour chaque dérivé par rapport à son ligand de référence, afin de s'assurer que la configuration correspond à l'orientation observée aux références (ligands expérimentaux) liés au site actif.

Un docking des composés sélectionnés est ensuite réalisé sur leur ligand de référence avec l'outil ShaEP (http://users.abo.fi/mivainio/shaep/download.php)[105]. Une visualisation est ensuite effectuée pour confirmer le bon docking des composés.

Minimisation des complexes. Les complexes (protéine et ligand) sont optimisés par minimisation avec le champ de force CHARMM36. Cette minimisation incorpore trois étapes successives :

1. Minimisation en maintenant rigide le squelette polypeptidique de la protéine et le ligand,
2. Minimisation en maintenant rigide le squelette polypeptidique de la protéine ; relâchement du ligand,
3. Relâchement du système complet (protéine et ligand).

Pour les étapes 1 et 2, le critère de convergence est de 10.0 pour SD et 1.0 pour ABNR ; tandis que le critère de convergence est de 0.1 pour ABNR concernant l'étape 3 de minimisation.

Analyse des interactions

L'analyse des interactions est réalisée avec l'outil OpenEye (https://www.eyesopen.com/) afin de générer la liste des interactions sous format TSV, JSON et SVG. Les interactions considérées sont :

- la liaison hydrogène (avec pour seuil de distance 3.2Å pour une liaison hydrogène idéale et pour une liaison hydrogène assistée par charge, 3.8Å pour une liaison hydrogène non-idéale),
- la liaison halogène (seuil de distance 3.2Å),
- les interactions π et T-stacking (avec pour seuil respectif 5.0Å et 5.35Å),
- le π-cation (avec pour seuil 5.5Å),
- les ponts salins (avec pour seuil 5.0Å),
- les contacts hydrophobes (avec pour seuil 1.2Å)
- les clashs (avec pour seuil 0.8Å).

Ces valeurs correspondent aux valeurs par défaut de OpenEye pour la version utilisée.

Les interactions analysées étant au niveau du site actif de BACE1, l'information concernant les sous-sites contactés est aussi ressortie en se basant sur les acides aminés constituant les sous-sites définis par Hu et al.[83].

3.2.3 . Résultats

Choix des références pour chaque composé

Beta-Sécrétase 1. Sur 474 ligands extraits des conformations de BACE1 issues de la base de données PDB, 36 sont composés strictement du pharmacophore, parmi lesquels 15 ont été résolus avec une résolution inférieure ou égale à 2.0Å. Le score de similarité entre les composés mentionnés dans le tableau 3.2 et ces 15 ligands cristallographiques ont été calculés. Le tableau 3.4 rapporte les références sélectionnées pour chaque composé, et pour lesquels le nombre de résidus à remodéliser est le plus petit possible.

Beta-Sécrétase 2. Sur 15 ligands extraits des conformations de BACE2 issues de la base de données PDB, 4 possèdent strictement le pharmacophore avec une résolution inférieure à 2.0Å. Le calcul de similarité a été réalisé entre ces 4 ligands cristallographiques et les composés issus du tableau 3.2. Deux références ont été sélectionnées, `7D5U` et `7F1G`. Néanmoins, nous avons remarqué que `7F1G` présente 12 acides aminés manquants au niveau du site actif.

Ce faisant, nous avons recherché un substitut (*i.e.* une conformation alternative) parmi les 13 conformations de BACE2 pour lesquelles aucun acide aminé n'est manquant au niveau du site actif avec une résolution inférieure à 2.0Å. Pour ce faire, nous avons procédé à (i) un alignement structural de toutes les conformations sur `7F1G`, (ii) créé pour chaque conformation un complexe avec le ligand de `7F1G` (code ligand : `0QQ`) et (iii) calculé un score de clashes, qui définit le niveau de clashes entre la protéine et le ligand (tableau 3.6). Excepté la conformation `7D5U`, les résultats ont montré que la conformation `7N4N` fut celle pour laquelle la conformation rentre la moins en clashes avec `0QQ`, avec un score de 1.54.

Une comparaison a été réalisée entre le ligand de 7F1G et le ligand cristallographique liant initialement la conformation 7N4N (code ligand : 0BK). Nous avons remarqué que le pattern principal de ces deux ligands est presque similaire, et que le ligand 0BK diffère très légèrement du pharmacophore avec la substitution de la pyridine par une pyrimidine (Figure 3.3). Ainsi, le choix de 7N4N comme substitut à 7F1G reste pertinent, car liant malgré tout un pharmacophore similaire à celui du N-Phenylpicolinamide.

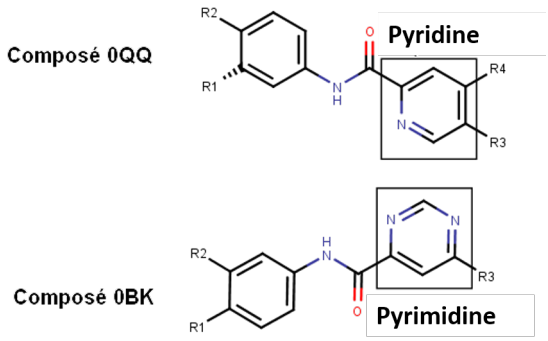


FIGURE 3.3 – Différence du pattern/sous-structure entre le ligand 0BK (PDB : 7N4N) et 0QQ (PDB : 7F1G)

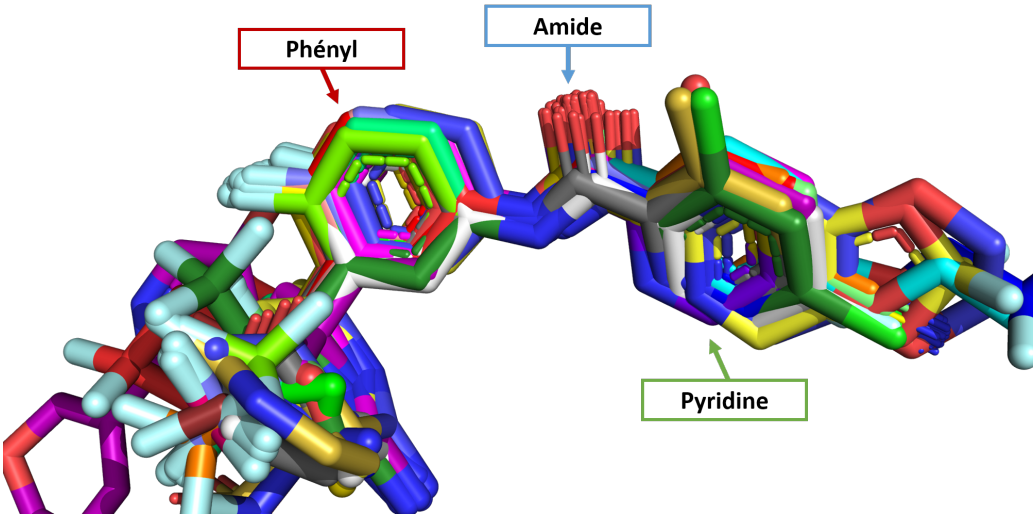
Résultats du Criblage par Alignement

L'"hypothèse émise est que le mode de liaison du pharmacophore est globalement conservé quelque le soit le dérivé. Cette hypothèse est confirmée par la visualisation (Figure 3.4), où nous avons observé des modes de liaisons fortement similaires. Ce mode de liaison implique, entre autres, une conservation d'interactions entre les résidus D32 et D228 (Numérotation BACE1) avec respectivement les amines tertiaire et primaire du substituant Z du groupement phényl du pharmacophore (Figure 3.4).
De ce fait, un docking "par alignement" (*i.e* alignement d'un composé sur la référence) nous est apparu pertinent comme méthodologie.

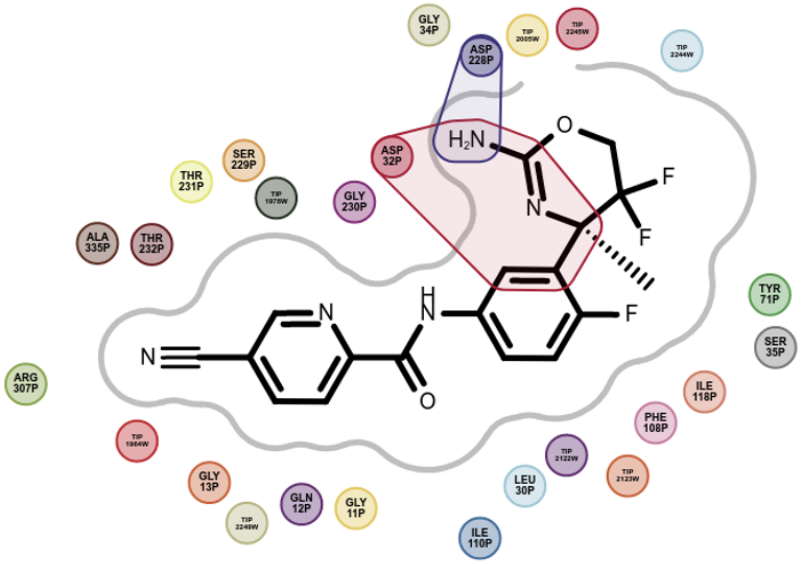
Les résultats de ce docking sont présentés dans le tableau 3.8, où le score d'alignement fourni par le logiciel ShaEP donne une indication sur la qualité de l'alignement entre le composé cible et la référence. Une minimisation d'énergie avec le champ de force CHARMM36 a été réalisée pour optimiser l'interaction de chaque composé dans le site actif.

Identification de contacts pouvant jouer un rôle dans la spécificité

Contacts avec les résidus spécifiques décrits par la bibliographie. Nous avons tout d'abord vérifié si des contacts étaient établis entre les composés criblés et les résidus considérés comme jouant un rôle dans la spécificité décrits dans le Tableau 1.30. Rappelons que ces résidus correspondent à



(a) Représentation de ligands RX dérivés du pharmacophore en complexe avec le site actif de BACE1. Les ligands correspondent à des inhibiteurs en complexe avec le site actif de BACE1, extraits de la base de données PDB.



(b) Représentation 2D de la caractéristique principale du mode de liaison. Le mode de liaison implique une interaction hydrogène entre l'amine primaire et D228 (surlignement en bleu), et entre l'amine tertiaire et D32 (surlignement en rouge) du substituant du groupement phényl (substituant "Z" d'après la figure 3.1)

FIGURE 3.4 – Mode de liaison des composés dérivés du Pharmacophore.

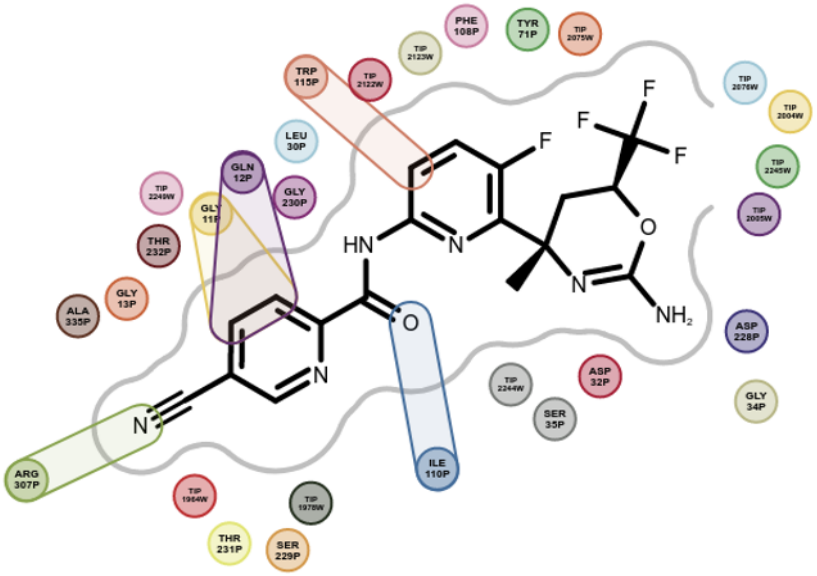
des résidus mutés sur BACE2.
Hormis I110 contacté par tous les composés, seul le composé B11 contacte les résidus P70 et R128 par l'intermédiaire de son groupement oxazine (substituant `Z` du pharmacophore). De manière intéressante, ce composé est celui qui a le meilleur ratio pIC50 vis-à-vis de BACE1 (le plus spécifique des composés vis-à-vis de BACE1), avec une valeur de 1.25. On peut ainsi s'attendre à ce que cette valeur de pIC50 soit expliquée notamment par les contacts avec ces trois résidus, par rapport aux autres composés. Ainsi, cette observation montre une cohérence entre la spécificité d'un composé et les résidus considérés par la bibliographie comme jouant un rôle dans la spécificité.

Contacts identifiés spécifiques vis-à-vis de BACE1 et BACE2. Par définition, nous considérons ici un "contact spécifique" comme un contact observé uniquement entre BACE1 (ou BACE2) et les ligands. Par exemple, sur les 10 criblages, nous avons observé que les composés contactent le résidus I110 chez BACE1, mais que le pendant chez BACE2 (L110) n'est nullement contacté. Cette observation est en cohérence avec la bibliographie, décrivant l'I110 comme étant un résidu spécifique. Néanmoins, nos observations nous ont montré que le contact réalisé est un contact très faible, de nature "pseudo-hydrogène", entre la chaine latérale de l'I110 et l'oxygène du linker amide (figure 3.5). Cette observation est confirmée avec les ligands de référence `3YS` et `6Z0`. Une hypothèse serait que le contact avec le résidu 110 chez BACE1 puisse avoir un effet indirect sur la spécificité, mais non jouer un rôle majeur pour l'activité des composés de la famille N-Phenylpicolinamide.

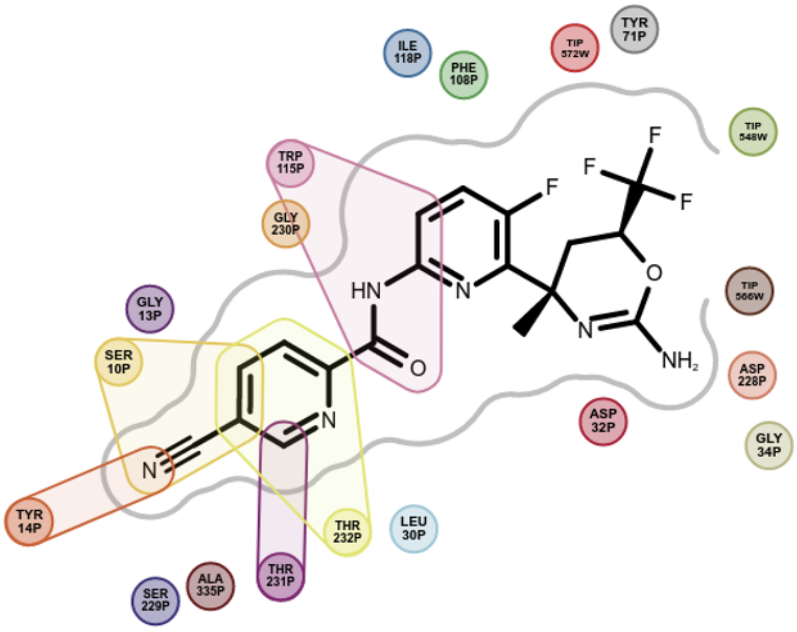
Parmi les autres contacts globalement spécifiques pour BACE1, nous avons relevé les résidus suivants : **G11** (SS9), **Q12**, **S35** (SS8) et **R307**. Notons ici que ces acides aminés ne sont pas mutés chez BACE2. Bien que les composés ne contactent pas, pour la plupart, ces résidus chez BACE2, nous avons malgré tout remarqué que certains contacts pouvaient être équivalents, avec néanmoins quelques différences. Par exemple, pour le composé B45, illustré par la figure 3.5 :

- le résidu R307 (BACE1) contacte le groupe nitrile par sa chaine latérale. Le résidu Y14 est le résidu équivalent sur BACE2, contactant aussi le groupe nitrile par sa chaîne latérale.
- le résidu I110 contacte l'oxygène de la liaison amide. L'équivalent avec BACE2 est W115, en contact avec la liaison amide et le groupement phényl. Notons que nous observons ce même résidu sur BACE1 en contact avec uniquement le groupement phényl de ce ligand.
- Les résidus G11 et Q12 contactent les carbones de la pyridine. Dans le cas de BACE2, nous pouvons notamment citer les résidus S10 et T232 ayant une portée plus élargie avec la pyridine.
- Le résidu S35 contacte le groupe méthyl de l'oxazine. Aucun contact équivalent n'est retrouvé sur BACE2

Tout comme avec BACE1, nous avons noté quelques contacts globalement spécifiques des composés avec BACE2 (**S10**, **Y14**, **W115** et **T231**).



(a) Principaux contacts déterminés comme spécifiques entre les composés criblés et BACE1



(b) Principaux contacts déterminés comme spécifiques entre les composés et BACE2, et contacts considérés comme similaires à ceux établis avec BACE1

FIGURE 3.5 – Contacts spécifiques des composés avec BACE1 et BACE2 illustrés avec le composé B45

Bien que certaines sous-structures du composé soient contactées à la fois par BACE1 et par BACE2, les résidus impliqués ne sont pas strictement équivalents : la nature et les propriétés peuvent jouer un rôle pour expliquer l'activité des composés.

Contacts identiques. Les composés peuvent également contacter des résidus identiques, sur BACE1 et sur BACE2. D'un point de vue quantitatif, nous avons observé que le nombre de contacts pour un résidu donné peut différer, pour un composé donné criblé dans le site actif de BACE1 et BACE2. Parmi les résidus qui pourraient faire l'objet d'une attention particulière pour une étude plus poussée, nous avons noté les résidus Y71 et F108, pour lesquels le nombre de contacts diffère souvent entre BACE1 et BACE2.
Néanmoins, bien que les contacts soient similaires, la nature de ces contacts peut éventuellement différer. De plus, un aspect dynamique devra être pris en compte puisqu'en effet, la durée de contact avec un tel résidu peut différer en fonction de l'enzyme.

3.2.4 . Conclusions

Parmi les résidus spécifiques décrits par la bibliographie, seul le résidu 110 contacte tous les composés lorsque criblés sur BACE1 et non sur BACE2. Néanmoins, ce contact est jugé très faible, de nature "pseudo-hydrogène", et devrait jouer un rôle indirect sur la spécificité des composés. Par ailleurs, le composé B11 est celui qui contacte le plus de résidus considérés comme jouant un rôle dans la spécificité (I110, P70, R128), et qui correspond au composé le plus spécifique vis-à-vis de BACE1. Ces résultats sont donc en cohérence avec les données de la bibliographie.

Par ailleurs, le criblage a révélé des contacts pouvant être spécifiques de BACE1 (G11, Q12, S35, R307) établis particulièrement avec le groupement pyridine et son substituant. Des contacts similaires ont été observés avec BACE2 par l'intermédiaire d'autres résidus (S10, T232, Y14). Néanmoins, la nature de l'interaction et le nombre d'interactions établis peut différer, et pourrait donc aussi jouer un rôle sur le pIC50.

De plus, nous avons aussi relevé des interactions avec des résidus identiques contactés à la fois sur BACE1 et sur BACE2, mais pour lesquels le nombre de contacts établi peut différer. Particulièrement, nous pouvons noter le résidu Y71 et F108 qui peuvent faire l'objet d'une attention particulière.

Composés	Références (CODES PDB::LIGAND)	Score de Similarité
B11	4X7I::3YS	0.50
B13	4X7I::3YS	0.64
B19	4X7I::3YS	0.67
B45	3ZMG::6Z0	0.70
B50	3ZMG::6Z0	0.83
B51	3ZMG::6Z0	0.68
B55	3ZMG::6Z0	0.73
B61	3ZMG::6Z0	0.70
B69	3ZMG::6Z0	0.50
B70	4X7I::3YS ; 7B1Q::SLK	0.58 ; 0.62

TABLE 3.4 – Les références issues de BACE1 associées à chaque composé dérivé du pharmacophore

Composés	Références (CODES PDB::LIGAND)	Score de Similarité
B11	7F1G::0QQ	0.52
B13	7D5U::GX6 ; 7F1G::0QQ	0.41 ; 0.49
B19	7D5U::GX6	0.48
B45	7D5U::GX6	0.67
B50	7D5U::GX6	0.66
B51	7D5U::GX6	0.72
B55	7D5U::GX6	0.73
B61	7D5U::GX6	0.60
B69	7D5U::GX6 ; 7F1G::0QQ	0.45 ; 0.45
B70	7D5U::GX6	0.51

TABLE 3.5 – Les références issues de BACE2 associées à chaque composé dérivé du pharmacophore. Notons ici que la référence 7F1G est remplacée par 7N4N

PDB BACE2	Score de Clashes avec 0QQ
7F1G	0.00
3ZKG	8.46
3ZKM	11.82
3ZKN	2.51
3ZKQ	2.14
4BEL	5.45
7D5B	1.36
7D5U	0.00
7N4N	1.74

TABLE 3.6 – Scores de clashes calculés avec MolProbity, entre le ligand RX de 7F1G et les conformations de BACE2 sans résidus manquants au site actif

Composés	Références (CODES PDB::LIGAND)	Score d'Alignement
B11	3YS	0.76
B13	3YS	0.79
B19	3YS	0.67
B45	6Z0	0.77
B50	6Z0	0.77
B51	6Z0	0.78
B55	6Z0	0.76
B61	6Z0	0.79
B69	6Z0	0.69
B70	3YS ; SLK	0.76 ; 0.81

TABLE 3.7 – Scores d'alignements obtenus avec ShaEP entre les composés dérivés du pharmacophores et les références issues de BACE1

Composés	Références (CODES PDB::LIGAND)	Score de Similarité
B11	0QQ	0.78
B13	GX6 ; 0QQ	0.80 ; 0.82
B19	GX6	0.80
B45	GX6	0.79
B50	GX6	0.82
B51	GX6	0.80
B55	GX6	0.79
B61	GX6	0.77
B69	GX6 ; 0QQ	0.75 ; 0.78
B70	GX6	0.77

TABLE 3.8 – Scores d'alignements obtenus avec ShaEP entre les composés dérivés du pharmacophores et les références issues de BACE2

3.3 . Application sur la β-sécrétase 1 : Criblage virtuel de la bibliothèque de nucléotides pour l'étude de spécificité et de design d'oligonucléotides

3.3.1 . Présentation du workflow général

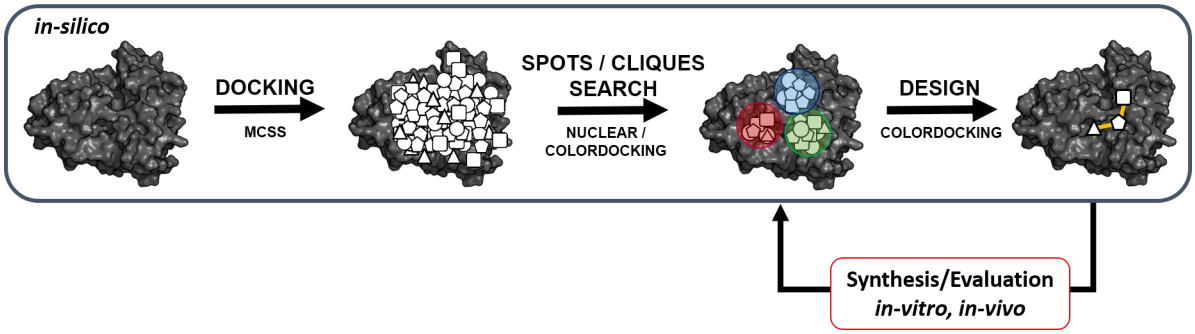


FIGURE 3.6 – Workflow général pour le design d'*in-silico*-mers

Le design d'*in-silico*-mers repose sur un workflow général intégratif, montré à la figure 3.6. Toute prédiction de solutions *in-silico*-mers par le `ColorDocking` nécessitera une évaluation expérimentale *in-vitro* et/ou *in-vivo*. Les données recueillies expérimentalement pourront être intégrées pour la prédiction de solutions avec des caractéristiques spécifiques. Bien que cela ne fasse pas intérêt dans le cadre de cette thèse, nous pouvons citer quelques méthodologies pour l'évaluation expérimentale des prédictions telles que la Calorimétrie de Titrage Isotherme (ITC) pour la mesure du K_D et donc de l'affinité, ou la Cristallographie par rayon X (RX) pour étudier le mode de liaison et analyser les contacts établis.

La méthodologie *in-silico* repose sur la connaissance de la structure 3D du récepteur et sur une bibliothèque de fragments constitués de mononucléotides modifiés, comme vu dans la sous-section 1.1.4 de l'introduction. Les étapes sont les suivantes :

1. Sélection d'une ou plusieurs conformations du récepteur à partir de la base de données PDB
2. Constitution de la bibliothèque de fragments de mononucléotides
3. Criblage virtuel avec MCSS de cette bibliothèque sur le récepteur
4. Recherche de spots et/ou de cliques avec NUCLEAR/`ColorDocking` respectivement
5. Design avec le logiciel `ColorDocking` d'*in-silico*-mers de taille k

Dans le cadre cette thèse, l'étude s'est focalisée principalement sur l'analyse de hotspots (voir sous-section 1.1.4). Nous proposons également une première application du `ColorDocking` sur BACE1.

3.3.2 . Sélection des différentes conformations de BACE1 et BACE2

Une protéine est une macromolécule dynamique et flexible. Elle peut ainsi adopter une multitude de conformations en fonction de l'environnement dans lequel elle se trouve (pH, température, etc.) et de son état de complexité (libre ou apo, liée ou halo). Afin de prendre en compte cette flexibilité, nous avons sélectionné quatre conformations de BACE1 et trois conformations de BACE2 issues de la base de données PDB, respectant les critères énoncées suivantes :

- Au minimum, si possible, trois états de complexités différents nommés `APO` (forme libre), `ACT` (forme liée avec une molécule occupant le site actif) et `EXO` (forme liée avec une molécule ou anticorps monoclonal interagissant avec l'exosite). On suppose indirectement que l'état de complexité induira un changement conformationnel au niveau du site actif et/ou de l'exosite,
- Des conformations pour lesquelles la résolution soit idéalement inférieure à 2.0Å,
- Des conformations pour lesquelles aucune coordonnée n'est manquante afin de ne pas induire de biais due à la modélisation ; ou au minimum ne posséder aucune coordonnées manquantes au niveau du site actif et de l'exosite,
- Avoir, si possible, une conformation à pH acide pour BACE1 due à son activité optimal à un tel pH,
- Avoir un niveau de solvatation suffisant autour de la protéine (idéalement supérieur à 100 molécules d'eau).

Afin de satisfaire le premier critère, une première analyse quantitative a été réalisée afin de déterminer le nombre de conformations d'états `APO`, `ACT` et `EXO` à la fois pour BACE1 et BACE2. Parallèlement à cette étude, un clustering R/Ψ par HBDSCAN sur le site actif[18] et un clustering "k-mean" sur l'exosite ont été réalisés uniquement sur BACE1 dû à un nombre élevé de conformations. En effet, à la date de cette étude, 424 conformations de BACE1 et 18 conformations de BACE2 ont été résolues.

Sélection des structures basée sur la conformation du site actif par clustering *R/Psi* par HDBSCAN pour BACE1

Le site actif de BACE1 est un site particulier caractérisé par la présence d'un flap (Section 1.5.4). De ce fait, un couple de paramètres, R et Ψ, a été défini. Le paramètre R correspond à une valeur de distance entre le carbone CG de l'acide aspartique 32 (CG-Asp32) et le groupement hydroxyle OH de la tyrosine 71 (OH-Tyr71) située sur le flap. Le paramètre Ψ correspond à un angle pseudo-dièdre défini par C-Trp76, N-Val69, CA-Thr72 et CA-Gln73. Pour obtenir différents clusters, l'algorithme de clustering basé sur la densité HBDSCAN a été utilisé.

Sélection des structures basée sur la conformation de l'exosite par clustering *k-mean* par MMTSB pour BACE1

Sur le nombre total de conformations de BACE1, nous avons dénombré 108 conformations pour lesquelles aucune coordonnées ne sont manquantes. Cette pré-sélection a été utilisée dans le cadre du clustering basée sur le "k-means".

Contrairement au site actif possédant une caractéristique particulière, l'exosite de BACE1 est une large surface définie notamment par trois boucles (Section 1.5.4). De ce fait, le clustering appliquée se base sur l'algorithme "k-means" avec le package Perl MMTSB (`http://blue11.bch.msu.edu/mmtsb/cluster.pl`).

Le clustering k-means est réalisé uniquement sur une sélection élargie de résidus (Ile248 à Phe322) de l'exosite. Le mode de clustering utilisé est basé sur la déviation de la racine de la moyenne des carrés (RMSD) en prenant en compte uniquement les carbones α de la sélection. Afin d'obtenir un nombre de clusters égal à 3, le rayon de clustering a été défini à une valeur supérieure ou égale à 0.9.

3.3.3 . Préparation des conformations retenues

La préparation des conformations retenues pour le criblage virtuel avec `MCSS` est réalisée au travers de différentes étapes :

1. Nettoyage des fichiers PDB : toutes molécules, hors les molécules d'eau et le récepteur, sont éliminées du fichier PDB
2. Remodélisation des coordonnées manquantes (uniquement appliquée aux conformations de BACE2)
3. Détermination des états de protonations des histidines, des acides glutamiques et des acides aspartiques (Section 3.2.2 pour le protocole)
4. Génération de deux états de solvatation pour chaque conformation (Section 3.2.2 pour la définition)
5. Minimisation des conformations avec CHARMM36 pour obtenir la conformation de plus basse énergie localement

Remodélisation des boucles non-résolues

Application de DaReUS-Loop. Toutes les conformations de BACE2 possèdent des résidus manquants, à proximité des sites d'intérêts ou sur les régions d'intérêts. De ce fait, la remodélisation de cette boucle est réalisée avec le serveur web **DaReUS-Loop** (`https://bioserv.rpbs.`

univ-paris-diderot.fr/services/DaReUS-Loop/)[106]. Cet outil possède la particularité de générer 10 modèles pour une conformation, permettant ainsi de sélectionner celle qui nous semble la plus pertinente.

Pour chaque conformation de BACE2, le mode "ReModelling" est utilisé. Ce mode prend pour données d'entrée la conformation sous format PDB et la séquence humaine complète de BACE2 sous format FASTA issue de UniProt (https://www.uniprot.org/uniprotkb/Q9Y5Z0/entry). En sortie, on obtient un fichier sous format Multi-PDB avec 10 modèles différents représentant respectivement une conformation différente de la boucle remodélisée par DaReUS-Loop.

Choix d'un modèle parmi dix. Pour choisir un modèle parmi les dix proposés par DaReUS-Loop pour une conformation donnée de BACE2, nous avons appliqué une méthodologie basée sur la minimisation du modèle suivie d'un calcul de RMSD local avant et après minimisation. La miminisation inclut trois étapes, dont :

1. La fixation de la conformation, sauf la boucle à remodéliser (Ala160 à Asn170 selon la numérotation BACE2)
2. Fixation du squelette polypeptidique uniquement
3. Relâchement total du système

Pour chaque étape, les algorithmes SD (Steepest Descen) et ABNR (Adopted Basis Newton-Raphson) ont été appliqués, et les critères de convergence appliqués pour les algorithmes SD et ABNR sont respectivement : (1) 10.0 et 1.0, (2) 10.0 et 1.0, et (3) 10.0 et 0.1.

Le RMSD local est défini comme le calcul de la déviation de la racine de la moyenne des carrés d'une région locale de BACE2. Cette région locale correspond soit aux molécules d'eau à 6.0Å de la boucle manquante (donc à 6.0Å de Ala160 à Asn170), soit de tout atome du récepteur à 6.0Å des extrémités de la boucle manquante. L'hypothèse admise est que le modèle pour lequel la déviation des molécules d'eau autour de la boucle est minime correspond au modèle le plus proche de la réalité.

Définition des états de solvatation

Comme mentionné dans la Section 3.2.2, nous avons défini deux états de solvatations différents nommés **basse solvatation** et **haute solvatation**. Dans la même logique, le serveur web EDIA a été utilisé pour générer les scores EDIA associés à chaque molécule d'eau. Nous rappelons les définitions suivantes :

- un état à basse solvatation est composée de toute molécule d'eau ayant un score EDIA supérieur ou égal à 0.81 (molécules d'eau fortement liée à la protéine)

- un état à haute solvatation est composée toute molécule d'eau ayant un score EDIA supérieur ou égal à 0.73

Néanmoins, une légère différence subsiste dans le protocole de sélection des molécules d'eau en comparaison à la Section 3.2.2. Cette différence s'explique par l'absence de ligand dans le site de liaison. En effet, contrairement au protocole de la section 3.2.2 où les molécules d'eau à proximité du ligand sont conservées, ces dernières sont éliminées ici. Le protocole, au travers d'un script PyMol, est donc le suivant :

1. Élimination de toute molécule d'eau dont la distance tout-atome est supérieure à 4.0Å du récepteur
2. Élimination de toute molécule d'eau dont le score EDIA est inférieur au seuil cible (0.73 ou 0.81)
3. Élimination de toute molécule d'eau dont la distance tout-atome est plus proche du ligand que la protéine (uniquement pour les conformations `ACT` et `EXO`)

Optimisation des conformations par minimisation

La dernière étape est l'optimisation des conformations par minimisation avec CHARMM36. Elle inclut deux étapes de minimisation qui sont successivement :

1. une étape de minimisation avec la fixation du squelette polypeptidique,
2. une étape de minimisation avec le système entièrement relâché.

Pour chaque étape, les algorithmes SD et ABNR ont été utilisés. Pour la première étape, la convergence fixée pour chacun des algorithmes est 10.0 et 1.0 respectivement. Pour la deuxième étape, la convergence fixée pour chacun des algorithmes est 10.0 et 0.1 respectivement.

3.3.4 . Criblage virtuel des nucléotides standards et nucléotides modifiés avec MCSS

Après la préparation des conformations, le criblage virtuel est exécuté avec la version 36 de `MCSS`. Trois données d'entrées sont fournies à `MCSS` : la bibliothèque de fragments de nucléotides ("groupes"), les coordonnées de la boîte déterminant la zone de criblage, et le fichier sous format CRD des conformations préparées et minimisées.

Pour la bibliothèque de fragments utilisée, nous renvoyons le lecteur ou la lectrice à la sous-section 1.1.4. Nous mentionnons juste ici que les 444 groupes constituant la bibliothèque de fragments ont été criblés sur chacune des conformations sélectionnées.

Détermination des coordonnées de la boite en fonction de la nature du criblage virtuel

Criblage virtuel non-focalisé. Un premier criblage consiste à un criblage non-focalisé (ou "à l'aveugle"). Cela signifie que le criblage virtuel est réalisé sur toutes les régions de la protéine. Ce type de criblage permet de donner une vision plus globale sur les régions interagissant avec les différents nucléotides. Les résultats sont donc principalement utilisés pour une étude de hotspots. Elle est donc moins précise d'un point de vue de l'échantillonnage au vu de la zone explorée.

Afin de réaliser ce criblage, une boite englobant complètement le récepteur a été définie. À partir des coordoonnées extrèmes du récepteur, une valeur tampon de 7.5Å est ajoutée à ces coordonnées. Cette zone tampon permet de s'assurer que les plus gros fragments puissent interagir avec les régions extrêmes du récepteur.

Criblage virtuel focalisé. Le criblage focalisé est un criblage où la boite englobe spécifiquement la région de liaison ciblée. L'échantillonnage ne sera donc que meilleur par rapport à un criblage non-focalisé. En l'occurrence, nous avons défini deux boîtes, l'une focalisée au site actif et l'autre focalisée à l'exosite. Le travail de cette thèse s'est restreint uniquement sur le site actif.

La boite pour le criblage focalisé au site actif englobe tous les acides aminés qui sont à 6.0Å des deux aspartates 32 et 228 (selon la numérotation BACE1). Pour déterminer ces coordonnées, le module Xtractor de NUCLEAR est utilisé.

3.3.5 . Étude des hotspots avec NUCLEAR

On définit ici un **hotspot** comme une sous-région de la protéine **interagissant au minimum avec 10 groupes nucléotidiques distincts**.

L'analyse des hotspots est réalisé avec NUCLEAR (Section 1.1.4). Les paramètres utilisés sont les suivants :

- une analyse sur le **top 10** des poses pour chaque groupe nucléotidique
- une distance de contacts de 3.5Å (`inter_dist = 3.5`)
- la prise en compte de tous les atomes de la protéine et des poses (`prot_sel & frag_sel = all`)
- un pré-clustering avec un seuil de RMSD de 2.0Å (`rmsd_cut = 2.0`)
- le paramètre de densité fixé à 0.10 (`density_cut = 0.10`) pour la sélection des clusters les plus représentatifs
- le paramètre de "fusion" fixé à 0.40 (`merge_cut = 0.40`), qui permet d'assembler des clusters (et donc des hotspots) similaires.

Pour chaque conformation de BACE1 et de BACE2, cette analyse est réalisée avec, d'une part, les 222 groupes (conformations ANTI+SYN) de type PO et, d'autre part, avec les 222 groupes de type PS_2. Cette analyse est réalisée à partir de l'exploration issue du criblage virtuel non-focalisé.

En sortie, NUCLEAR génère, entre autres, un diagramme qui indique le nombre de groupes distincts contactés par chaque hotspot identifié, comme illustré à la figure 3.7.

Nous avons considéré qu'un hotspot appartient au site actif (ou partiellement au site actif) si 50% des résidus le constituant appartient à la boîte définie pour le criblage focalisée.

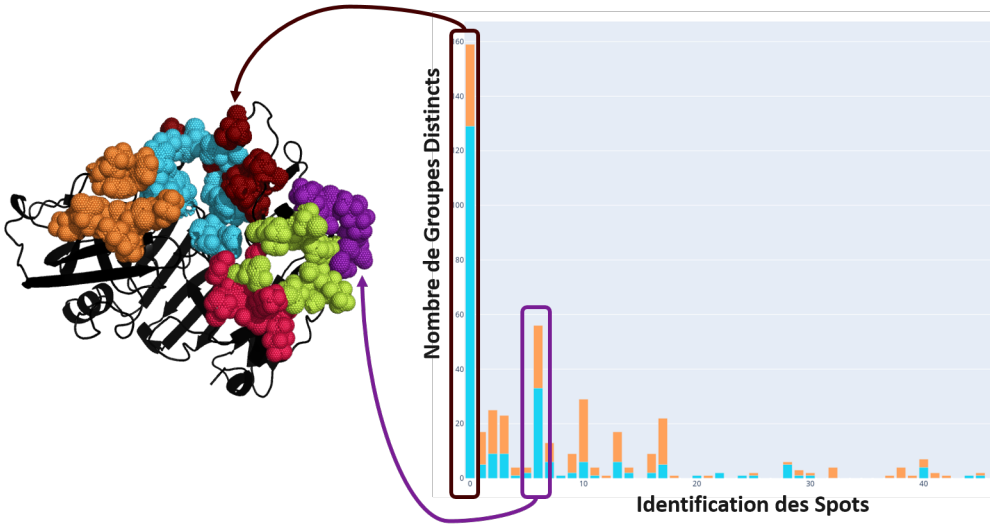


FIGURE 3.7 – Exemple de diagramme généré par NUCLEAR (à gauche), avec une représentation 3D associée (à droite) générée avec PyMoL. Sur le diagramme, en cyan représente le nombre de groupes dans le top 5, et en orange le nombre de groupes dans le top 10.

3.3.6 . Étude des groupes spécifiques dans le site actif entre BACE1 et BACE2

Afin d'étudier les groupes spécifiques pour chaque sous-site, nous avons utilisé le module Xtractor de NUCLEAR. Ce module permet, entre autres, d'extraire les groupes (et les poses associées) qui contactent une région spécifiée de la protéine.

Cette analyse est réalisée avec les paramètres suivants :

- une analyse sur le **top 50** des poses,
- une distance de contact de 3.0Å,
- une sélection de tous les atomes des poses en excluant le patch terminal

Cette étude est menée sur l'exploration issue du criblage focalisé au site actif.

3.3.7 . Design d'un *in-silico*-mer vis-à-vis de BACE1 et BACE2

Une première application du design d'*in-silico*-mer est proposée avec le `ColorDocking` sur le site actif de `1SGZ` et `3ZKM`, à partir des analyses ayant permis l'identification de groupes spécifiques vis-à-vis de BACE1. Le choix de ces deux conformations se justifie par l'absence de ligand dans le site actif des structures RX (Tableau 3.10 et 3.11).

Pour réaliser ce design, nous avons utilisé NUCLEAR pour générer les matrices de connectivité, de clashes, et de scores ; à partir de la distribution MCSS focalisée sur le site actif. Nous avons considéré 100 poses par groupe (50 ANTI/50 SYN), et pour lesquels 20 groupes ont été choisis et indiqués dans le Tableau 3.12. Le choix de ces groupes est discuté dans la Section 3.3.8. Le critère de distance considéré pour créer la matrice de connectivité est de 4.5Å entre les atomes C5′ et O3′, avec un pre-clustering basé sur le RMSD de 0.5Å.

Au niveau du `ColorDocking`, nous avons recherché des séquences de longueurs $k = 4, 5, 6$ nucléotides. Nous avons considéré une valeur maximale de $\Delta_{max} := 10 \mathrm{kcal.mol}^{-1}$, et graduellement augmenté Δ d'un pas unitaire, afin d'atteindre au maximum 100 solutions sans-clashes par coloration. Nous avons défini la tolérance à $\varepsilon = 0.01$.

3.3.8 . Résultats

Définition de la région de l'exosite

Bien que des études aient défini l'exosite par les boucles C, D et F de BACE1, nous avons considéré l'exosite comme une région plus étendue de 51 acides aminés (248-280 et 305-322 selon la numérotation BACE1), que nous appellerons respectivement les sous-régions ExoA et ExoB (Figure 3.8). Cette sélection permet de recouvrir une plus large zone utile pour le criblage virtuel, et rentre en continuité avec le site actif défini par ses 10 sous-sites.

Nous avons réalisé une comparaison de cette séquence de l'exosite avec celle de BACE2. Les principaux points relevés sont les suivants :

- La séquence de l'exosite de BACE2 est plus courte d'un acide aminé par rapport à celle de BACE1, avec une délétion de V309
- Sur la région ExoA, 17 résidus sur 33 sont substitués sur BACE2 (Tableau 3.9)
- Sur la région ExoB, 13 résidus sur 18 sont substitués sur BACE2 (Tableau 3.9)

Ces observations mettent donc en avant le fait que l'exosite est moins conservé entre BACE1 et BACE2, et peut donc justifier un intérêt de cibler cette région en plus de celle du site actif afin d'avoir

un composé spécifique vis-à-vis de BACE1.

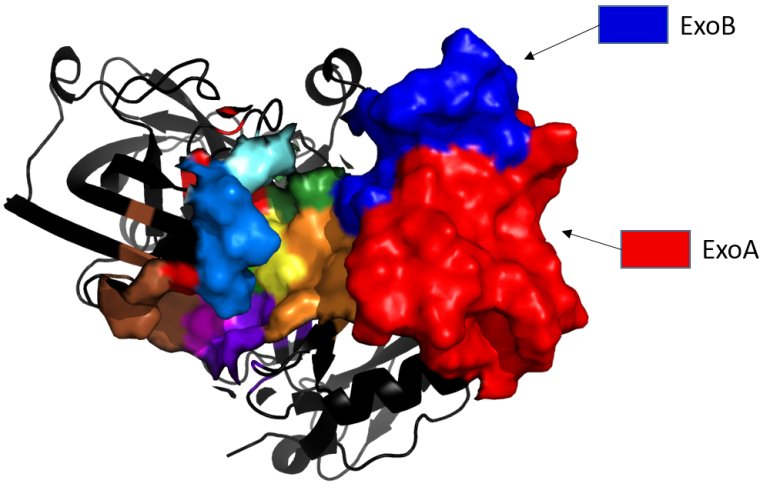


FIGURE 3.8 – Représentation en surface des deux sous-régions de l'exosite (ExoA et ExoB, respectivement en rouge et en bleu), avec une représentation en surface des 10 sous-sites du site actif

Sous-régions de l'exosite	**Résidus spécifiques sur BACE1**
ExoA	I248, K249, A250, S253, T254, E255, K256, P258, L263, E265, V268, Q271, A272, G273, T274, N278, I279
ExoB	L306, R307, V309, E310, D311, V312, A313, T314, S315, Q316, D317, D318, K321

TABLE 3.9 – Résidus de l'exosite (ExoA et ExoB) mutés chez BACE2

Sélection des conformations de BACE1 et BACE2

Analyse quantitative des états de complexité Comme mentionné dans la sous-section 3.3.2, nous avons dénombré 424 conformations de BACE1 et 18 conformations de BACE2 dans la base de données PDB. Parmi les 424 conformations de BACE1, on relève 4 conformations `APO` et 3 `EXO`. Parmi les 18 conformations de BACE2, on relève 2 conformations `APO`, 10 conformations `ACT` et 10 conformations `EXO`. Notons ici que certaines des conformations considérées comme `ACT` ou `EXO` peuvent être les deux à la fois.

Choix des conformations pour BACE1. Parmi les 4 conformations APO et les 3 conformations EXO de BACE1, seules deux conformations, respectivement 1SGZ et 5MCQ, ne possèdent aucune coordonnées manquantes et satisfont les cinq critères mentionnés dans la section 3.3.2.

Pour s'assurer que ces deux structures PDB aient bien une conformation différente de l'exosite, un clustering par RMSD local centré sur l'exosite est réalisé sur les 108 conformations de BACE1 pour lesquelles aucune coordonnées ne sont manquantes. Trois clusters ont été obtenus : 1) le cluster 1 regroupant 105 conformations, 2) le cluster 2 regroupant 2 conformations et 3) le cluster 3 regroupant 1 conformation. Les conformations 1SGZ et 5MCQ se retrouvent, respectivement, dans les clusters 1 et 3. Ce faisant, le choix dans le cluster 2 se relève très limité pour une conformation de type ACT, contenant les conformations 4GID et 6UVP.

Gonzalez-Aleman[18], dans sa thèse, a sélectionné trois conformations différentes de BACE1 basées sur la conformation du site actif. En effet, ce dernier a sélectionné initialement les conformations 1SGZ, 4GID et 6UVV.

En se basant donc aussi sur ses résultats, nous avons sélectionné quatre conformations de BACE1 qui différent au niveau du site actif et de l'exosite et qui respectent les critères énoncés dans la sous-section 3.3.2 : 1SGZ, 4GID, 5MCQ et 6UVV (voir tableau 3.10).

Code PDB	**État conformationnel**	**Résolution**	**Nombre H2O**	**pH**	**Nb Résidus manquants**
1SGZ	APO	2.00Å	253	6.5	0
4GID	ACT	2.00Å	350	7.4	0
6UVV	ACT	1.63Å	625	6.5	0
5MCQ	ACT+EXO	1.82Å	586	4.5	0

TABLE 3.10 – Caractéristiques des conformations de BACE1 retenues. Le nombre de molécules H_2O correspond ici au nombre de molécules d'eau initiale à 6.0Å tout-atome du récepteur

Choix des conformations pour BACE2. Parmi les 18 conformations de BACE2, il ne subsiste aucune conformation pour laquelle il n'y a pas de coordonnées manquantes. De plus, hormis 3 conformations ayant uniquement qu'une seule boucle manquante à proximité du site actif, la majorité possède des coordonnées manquantes sur l'exosite ou le site actif. Par conséquent, les trois conformations 2EWY, 3ZKM et 3ZKN ont été retenues. Leurs caractéristiques sont résumées dans le tableau 3.11.

Code PDB	État conformationnel	Résolution	Nombre H2O	pH	Nb Résidus manquants
2EWY	ACT	3.10Å	42	6.8	9
3ZKM	EXO	1.85Å	198	7.5	9
3ZKN	ACT+EXO	1.63Å	178	7.5	9

TABLE 3.11 – Caractéristiques des conformations de BACE2 retenues. Le nombre de molécules H_2O correspond ici au nombre de molécules d'eau initiale à 6.0Å tout-atome du récepteur

Choix des seuils EDIA pour les états de solvatation

Justification des seuils EDIA. Comme nous l'avons indiqué dans la sous-section 3.2.2, deux états de solvatation ont été définis. On parle de "basse solvatation" et de "haute solvatation", pour lesquels les seuils EDIA sont respectivement de 0.81 et 0.73. La fixation de ce seuil a été réalisée au travers d'une analyse quantitative et visuelle initialement sur deux conformations de BACE1 (`5MCQ`, `4GID`). Les seuils EDIA proviennent respectivement de la valeur médianne des scores EDIA de la totalité des molécules d'eau présentes dans la structure cristalline, respectivement $\mathrm{EDIA}_{\mathrm{mediane}}(\texttt{5MCQ}) = 0.73$ et $\mathrm{EDIA}_{\mathrm{mediane}}(\texttt{4GID}) = 0.81$.

Une première analyse sur `5MCQ` montre une légère tendance entre la distance des molécules d'eau par rapport au récepteur et leur score EDIA : au plus le score EDIA augmente, au plus la distance diminue. Néanmoins, nous observons des fluctuations importantes, comme par exemple de fortes valeurs EDIA pour des molécules d'eau situées à plus de 4Å de la surface du récepteur. Cette observation met probablement en évidence l'existence de réseaux de molécules d'eau, qui pourrait ainsi stabiliser des molécules d'eau distantes de la protéine. Ainsi, les valeurs EDIA des molécules d'eau peuvent être expliquées par le micro-environnemnt auxquelles elles appartiennent (Figure 3.9).

En effet, avec une valeur EDIA de 0.73, nous remarquons visuellement la présence d'un réseau de molécules d'eau au niveau de l'exosite, qui pourrait hypothétiquement empêcher des interactions directes entre les poses et la protéine, dû à une difficulté de déplacer ces molécules d'eau lors du docking (Figure 3.10).
À une valeur de 0.81, ce réseau est moins important, ce qui pourrait faciliter le déplacement des molécules d'eau lors du criblage virtuel. Ce constat a été aussi fait sur le site actif. Une analyse visuelle sur les autres conformations de BACE1 et BACE2 nous ait paru satisfaisante lorsque les deux seuils sont appliqués.
Par ailleurs, une augmentation du seuil au-delà de 0.81 n'apporte pas de grand changement au niveau

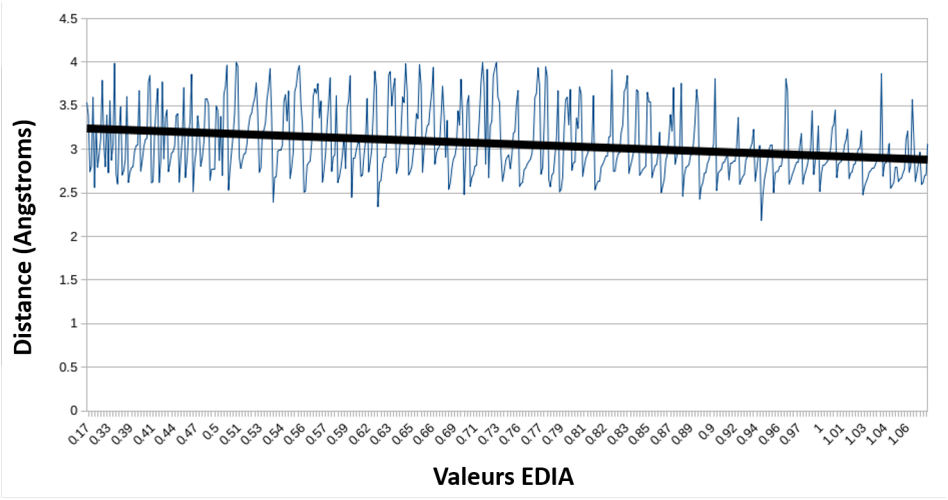


FIGURE 3.9 – Corrélation entre la distance et le facteur EDIA, analyse réalisée sur la structure `5MCQ`. La droite de régression linéaire (en noire) montre la tendance.

de l'exosite. C'est pourquoi les deux seuils 0.73 et 0.81 ont été considérés pour la suite.

Nous avons néanmoins rencontré deux exceptions avec `1SGZ` et `2EWY`.

Cas de `1SGZ`. Comme il a été décrit dans le protocole EDIA 3.2.2, l'outil prend pour en entrée une conformation PDB avec sa carte de densité électronique. Dans le cas de `1SGZ`, la carte de densité électronique n'est pas fournie dans la base de données PDB. De ce fait, deux solutions s'offraient à nous : soit reconstruire la carte de densité au travers d'un protocole informatique, soit appliquer une corrélation entre le facteur EDIA et le b-facteur. La première solution se révélait compromise, puisque la version OpenSource installable en local est une version plus ancienne que celle du serveur web. Par conséquent, nous avons appliqué la deuxième solution. En effet, le b-facteur peut être considéré comme l'inverse du facteur EDIA. Par définition, au plus le b-facteur est élevé, au plus les molécules d'eau sont flexibles, et correspondent en principe à des molécules d'eau faiblement liées au récepteur.

En sachant que la résolution structurale de `1SGZ` et de `4GID` sont identiques (Tableau 3.10), nous avons vérifié s'il subsistait une corrélation entre le b-facteur et le facteur EDIA sur 4GID (Figure 3.11). Et en effet, en traçant la droite de régression linéaire, nous observons une corrélation entre le b-facteur et le facteur EDIA : au plus le facteur EDIA est élevé, au plus le b-facteur est faible.

Fort de ce constat, les valeurs du b-facteur équivalents aux scores EDIA 0.73 et 0.81 sont respectivement de 44.17 et 37.81 sur `4GID`. Ainsi, pour `1SGZ`, toutes les molécules d'eau inférieures à 44.17 ou 37.81 sont retenues.

(a) EDIA = 0.73 (b) EDIA = 0.81

FIGURE 3.10 – Représentation 3D de BACE1 (5MCQ) en contacts avec des molécules d'eau selon le facteur EDIA : (a) 0.73 (en cyan) et (b) 0.81 (en vert). L'exosite est représenté en bleu.

Cas de 2EWY. Dans le cas de 2EWY, le nombre de molécules d'eau entourant le récepteur est très faible avec 42 molécules d'eau. Une application du protocole EDIA conduirait une élimination totale de la solvatation autour de 2EWY. Ce phénomène s'explique par la résolution structurale mauvaise de 2EWY, qui est de 3.10Å. Par conséquent, aucun filtre n'a été appliqué sur les molécules d'eau de 2EWY.

Le criblage virtuel a été réalisé sur ces deux états de solvatation pour BACE1 et BACE2, que ce soit pour un criblage non-focalisé ou un criblage focalisé.

Analyse des hotspots à partir du criblage non-focalisé

Comme nous en avons brièvement parlé dans la sous-section 3.3.4, le criblage non-focalisé consiste à un criblage de la bibliothèque de mononucléotides sur toutes les régions du récepteur. Ce criblage permet ainsi de faire une étude de hotspots.

Vérification des régions de liaisons préférentielles des nucléotides. BACE1 n'est pas biologiquement une enzyme de liaison à des acides nucléiques. Bien que la bibliographie ait montré que des aptamères pouvaient inhiber l'activité de BACE1, aucune donnée structurale n'existe à ce jour pour connaître les régions de liaison de ces aptamères. De ce fait, une question subsiste : "Est-ce que les nucléotides, et de surcroît les nucléotides modifiés, interagissent principalement avec les régions modulatrices de l'activité catalytique de BACE1 ; ou peuvent-ils aussi interagir préférentiellement avec d'autres régions ?". La réponse à cette question permettrait à la fois de s'assurer que la conception d'*in-silico*-mers, interagissant avec le site actif et/ou l'exosite, est possible ; mais également de proposer de potentielles optimisations avec des régions de liaison insoupçonnées.

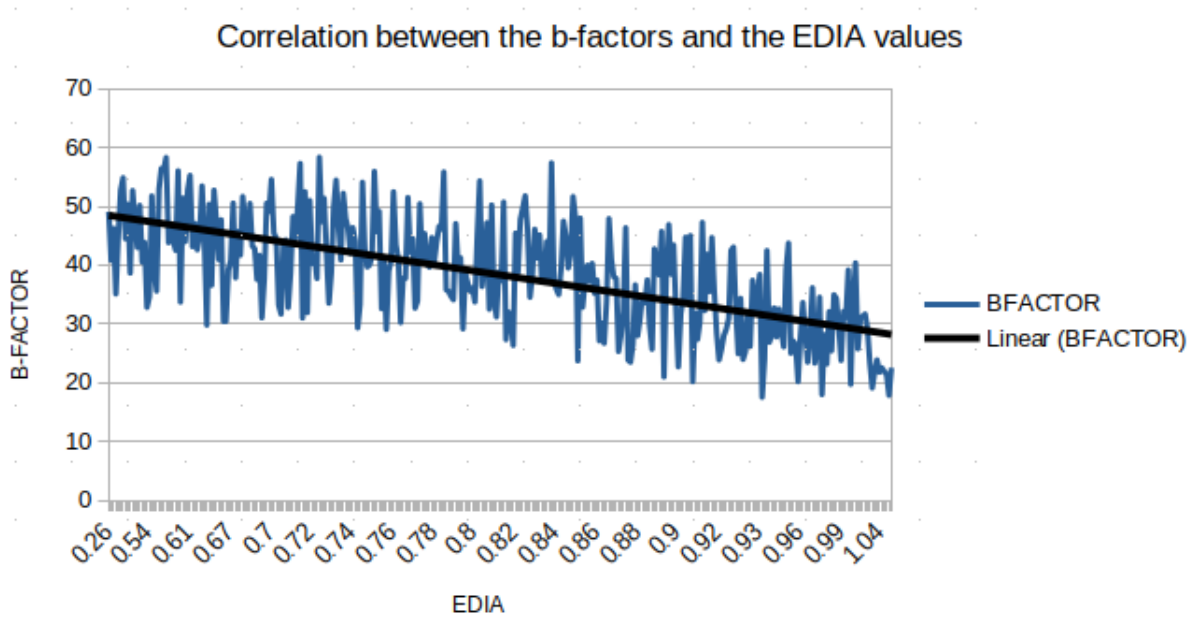


FIGURE 3.11 – Corrélation entre le b-facteur et le facteur EDIA, analyse réalisée sur la structure `4GID`

Nous rappelons qu'un hotspot, appliqué dans le cadre de cette thèse, est défini comme une région de la protéine interagissant au minimum avec 10 groupes nucléotidiques différents. Pour plus de clarté, nous définissons la nomenclature suivante : **PO** pour les hotspots identifiés à partir des groupes ayant un groupe phosphate et **PS**$_2$ pour les hotspots identifiés à partir des groupes ayant un groupe phosphorothiate.

D'un point de vue des données quantitatives, nous avons compté une dizaine de hotspots. En général, les hotspots sont localisés au niveau de la région du site actif et de l'exosite (Figure 3.12). Cette observation est positive car elle signifie que ces régions modulant l'activité catalytique sont favorables à des interactions avec des acides nucléiques.

Étude l'impact de facteurs sur les hotspots. Trois facteurs peuvent avoir un impact sur la présence et la localisation d'un hotspot :

1. la conformation,
2. la solvatation,
3. la modification (PO et PS$_2$)

D'après une étude visuelle des hotspots entre différentes conformations de BACE1, nous observons que des hotspots peuvent différer (absence/présence). C'est par exemple le cas entre la conformation

1SGZ et 5MCQ (Figure 3.12). De ce fait, la conformation et donc la **flexibilité joue un rôle dans les régions de liaisons aux nucléotides**. Cette observation conforte donc l'idée de sélectionner différentes conformations pour BACE1 et BACE2, et de prédire des solutions conformations dépendantes.

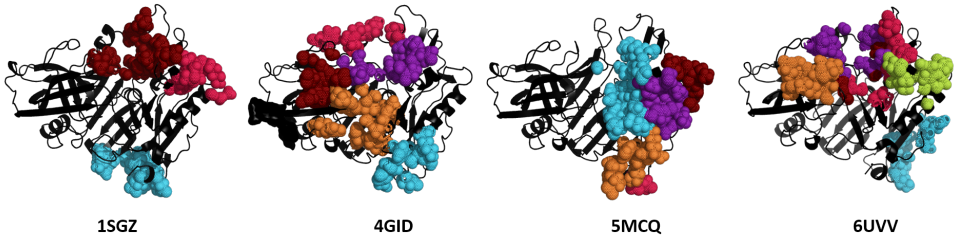


FIGURE 3.12 – Illustration de hotspots en fonction des 4 conformations de BACE1. Les hotspots ont été obtenus à haute solvatation, à partir des groupes PO.

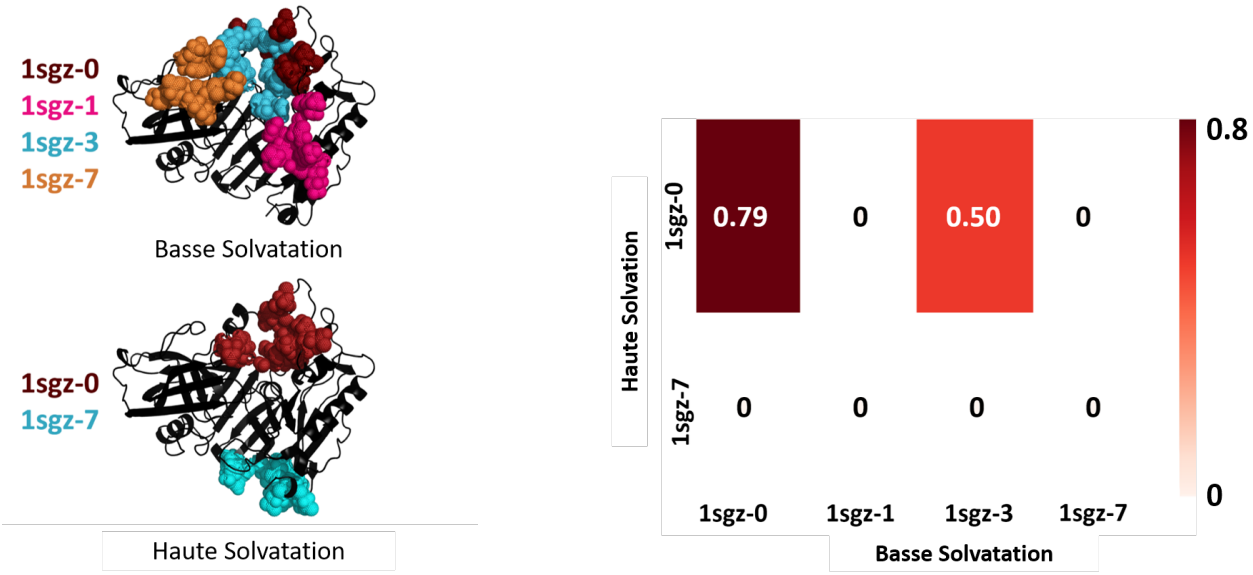
Le deuxième facteur est la solvatation. Bien qu'une analyse visuelle nous permette d'avoir une idée, nous avons réalisé une analyse quantitative. En effet, pour une conformation donnée entre deux états de solvatation, nous avons comparé chaque hotspot en mesurant le taux de recouvrement (défini comme le taux de résidus similaire entre deux hotspots). Un taux de recouvrement de 0 signifie que deux hotspots ne sont pas localisés dans la même région, tandis qu'un taux de recouvrement de 1 indique que deux hotspots sont identiques. Un exemple sur 1SGZ est donné à la figure 3.13. Sur le site actif, 2 hotspots ont été identifiés à haute solvatation, contre 4 hotspots à basse solvatation :

- Deux hotspots sont présents à haute solvatation, mais absent à basse solvatation
- Un hotspot est partiellement conservé entre les deux états (50%)
- Un hotspot est fortement conservé avec un chevauchement de 79%

Nous observons un phénomène similaire quelle que soit la conformation de BACE1, sur l'exosite ou le site actif. Ainsi, au vu de ces résultats, nous avons émis l'hypothèse que la **solvatation peut impacter l'interaction** des nucléotides avec la protéine de deux manières :

1. Soit les molécules d'eau bloquent une interaction directe entre des poses et une sous-région de la protéine ;
2. Soit les molécules d'eau jouent le rôle d'intermédiaire pour une interaction indirecte entre les poses et une sous-région de protéine.

Une analyse similaire est réalisée pour étudier l'impact des groupes PO et PS_2. Le soufre étant un atome plus volumineux que l'oxygène, nous pouvons nous attendre à un phénomène équivalent à ce



(a) Représentation 3D des hotspots en fonction de la solvatation.

(b) Chevauchement des hotspots en fonction de la solvatation

FIGURE 3.13 – Représentation des hotspots sur le site actif de 1SGZ en fonction de la solvatation : haute solvatation et basse solvatation. (A) Représentation 3D des spots entre les deux états de solvatation, (B) Heatmap représentant le chevauchement des hotspots au site actif : la couleur rouge foncé indique une conservation du hotspot ; la couleur blanche indique une non-conservation du hotspot.

qui est observé avec la solvatation. C'est le cas par exemple de 4GID (Figure 3.14), à haute solvatation. 5 hotspots ont été identifiés sur le site actif à partir du groupe PO, contre 6 hotspots à partir du groupe PS_2. Comme montré par la heatmap, nous faisons les observations suivantes :

- Les résidus du hotspot "4gid-4" contactent préférentiellement les nucléotides avec un groupe PO,
- les résidus des hotspots "4gid-21" et "4gid-22" interagissent préférentiellement avec les nucléotides incorporant la modification PS_2,
- Certains hotspots sont partiellement conservés, avec des taux de conservations supérieurs à 60%.

Ainsi, la modification sur l'acide phosphorique peut aussi jouer un rôle pour moduler l'affinité des molécules, en plus de sa propriété de résistance vis-à-vis des nucléases. Ces phénomènes observés sur le site actif sont également observables sur l'exosite de BACE1, comme en témoignent les visualisations 3D des Figures 3.12 et 3.13 par exemple.

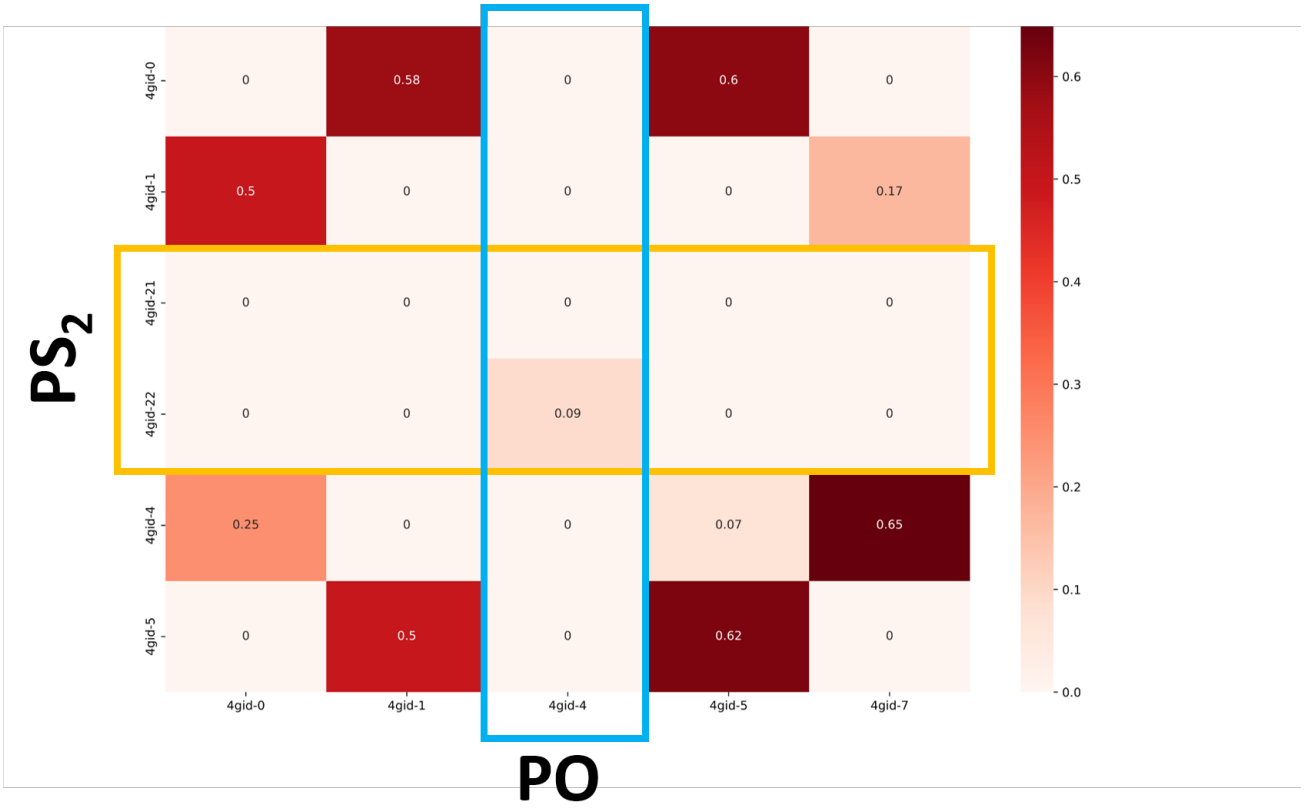


FIGURE 3.14 – Heatmap représentant le chevauchement des hotspots au site actif sur `4GID` : la couleur rouge foncé indique une conservation du hotspot ; la couleur blanche indique une non-conservation du hotspot. Le rectangle jaune met en évidence des résidus (deux hotspots : 4gid-21 et 4gid-22) contactés uniquement par les groupes de type PS_2. Le rectangle bleu met en évidence des résidus (un hotspot : 4gid-4) contactés uniquement par les groupes de type PO.

En conclusion de cette partie, la prise en compte de la flexibilité et de l'état de solvatation est nécessaire dans le processus de design d'*in-silico*-mers. Autrement dit, nous préconisons de prédire différents *in-silico*-mers pour chaque conformation et chaque état de solvatation. Cette prise en compte permettra d'obtenir des solutions plus diverses, avec potentiellement des caractéristiques physico-chimiques et mode de liaison différents.

Détermination des groupes spécifiques vis-à-vis de BACE1 et BACE2

Le design d'*in-silico*-mers spécifiques peut reposer sur la stratégie suivante : sélectionner les groupes considérés spécifiques vis-à-vis de BACE1, et écarter les groupes considérés comme spécifiques vis-à-vis de BACE2. On définit ici un groupe "spécifique" comme étant un groupe interagissant uniquement avec l'une des deux structures, soit globalement, soit restreint à un sous-site/hotspot donné. Autrement dit, est exclu de la définition tout groupe interagissant à la fois avec au moins une conformation de BACE1 et une conformation de BACE2.

Ainsi, nous avons étudié les différents groupes spécifiques de BACE1 et BACE2 associés à chaque sous-site, à partir de l'exploration à haute solvatation. Les résultats sont présentés dans les tableaux 3.14.

Pour presque la totalité des sous-sites du site catalytique, nous remarquons globalement un nombre de groupes spécifiques plus élevés vis-à-vis de BACE2 que BACE1. Par exemple, les sous-sites 6 et 9 en sont les parfaits exemples où nous retrouvons respectivement 40 et 66 groupes à backbone PO spécifiques vis-à-vis de BACE2, contre 11 et 1 groupes spécifiques vis-à-vis de BACE1. Cette observation est similaire avec les groupes à backbone PS_2 (Figure 3.15).

En revanche, nous pouvons également observer la tendance inverse, par exemple pour le sous-site 7 où nous retrouvons 25 groupes à backbone PS_2 spécifiques vis-à-vis de BACE1 contre 2 groupes spécifiques vis-à-vis de BACE2. Cette observation est équivalente pour les groupes à backbone PO interagissant avec le sous-site 7. (Figure 3.15)

Par ailleurs, nous notons également deux autres observations :

- Certains sous-sites n'accueillent pas de groupes spécifiques vis-à-vis de BACE1 et de BACE2, comme c'est le cas de SS8 et SS3 respectivement,
- Des groupes peuvent interagir avec au moins une conformation de BACE1 et au moins une conformation de BACE2,
- Quelques rares groupes ne contactent aucune des conformations de BACE2 tout sous-site confondu.

(a) Sous-site 7 (b) Sous-site 9

FIGURE 3.15 – Diagramme de Venn représentant la proportion de groupes spécifiques vis-à-vis de BACE1 (en rouge) et vis-à-vis de BACE2 (en vert), pour les sous-sites (a) SS7 et (b) SS9. Le rond marron indique la proportion de groupes contactant à la fois BACE1 et BACE2 sur ces deux sous-sites.

Ainsi, en conclusion, nous observons des groupes qui peuvent être spécifiques d'une conformation de BACE1 et de BACE2. Afin de s'affranchir de la dépendance vis-à-vis de la conformation, nous avons établi une liste de groupes spécifiques vis-à-vis de quatre conformations de BACE1 par rapport au trois conformations de BACE2, et inversement. Cette liste donne une indication sur les groupes à éviter potentiellement ou à choisir préférentiellement pour concevoir un *in-silico*-mer spécifique de BACE1. Néanmoins, l'origine de ces observations n'a pas fait l'objet d'une étude approfondie, et différentes hypothèses peuvent être envisagées :

- Rappelons que ces études sont réalisées sur des conformations fixes. L'accessibilité à des sous-sites de BACE1/BACE2 est favorable ou moins favorable en fonction du volume des groupes et de la conformation. Ainsi, une étude de la surface accessible à chaque sous-site associée à la détermination du volume des groupes permettrait d'obtenir une réponse pour expliquer pourquoi certains groupes ont tendance à contacter certains sous-sites d'une conformation donnée de BACE1 et/ou de BACE2,
- Bien que nous puissions observer des groupes contactant à la fois un sous-site donné de BACE1 et BACE2, il est nécessaire d'analyser les contacts établis. En effet, la définition de spécificité peut être élargie au-delà de la simple présence/absence du groupe dans un sous-site donné.

Analyse des scores des groupes spécifiques vis-à-vis de BACE1 en fonction de la conformation

Bien que nous ayons déterminé une liste de groupes spécifiques de BACE1, nous nous attendons à ce que la plage de scores d'énergie du top 50 des poses, associée à chaque groupe, diffère en fonction de la conformation. Ainsi, deux hypothèses sont envisagées :

- Pour un sous-site donné, le meilleur groupe est conservé entre au moins deux conformations de BACE1.
- Pour un sous-site donné, le meilleur groupe diffère entre au moins deux conformations de BACE1.

Ces deux hypothèses ont été vérifiées, en comparant par exemple les sous-sites SS2 et SS7 entre les conformations `1SGZ` et `4GID`. En effet, nous observons que le meilleur groupe contactant le sous-site SS1 de `1SGZ` et `4GID` est le groupe `R2C` (Figure 3.16) ; tandis que pour le sous-site SS7, nous avons relevé les groupes `MOU` et `DAG`, respectivement pour les conformations `1SGZ` et `4GID` (Figure 3.17).

Cette observation indique que la conformation du site actif impacte donc l'énergie de liaison des groupes avec les résidus du site actif. La nature des contacts ou les résidus contactés peuvent être deux hypothèses pour expliquer cette différence. Ainsi, une analyse plus fine est nécessaire afin de vérifier ces hypothèses.

Exemple de Design sur 1SGZ

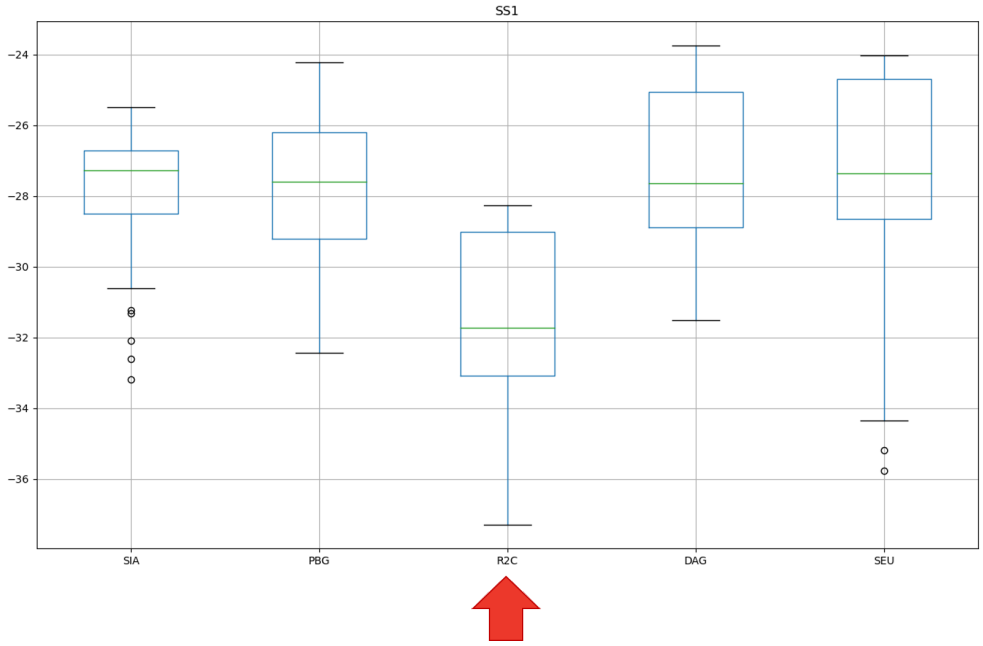
Design sur `1SGZ`. Nous proposons ici un exemple de conception d'un *in-silico*-mer, spécifique de BACE1, reposant sur le rationnel suivant, pour chaque sous-site :

- Identifier les groupes spécifiques vis-à-vis de BACE1 (Tableau 3.3.9),
- Parmi ces groupes, sélectionner le meilleur groupe de type PO et de type PS_2 vis-à-vis de BACE1 : le meilleur groupe est considéré comme celui ayant les poses de meilleurs scores (ce qui correspond aux "outlier" dans la Figure 3.16 par exemple).

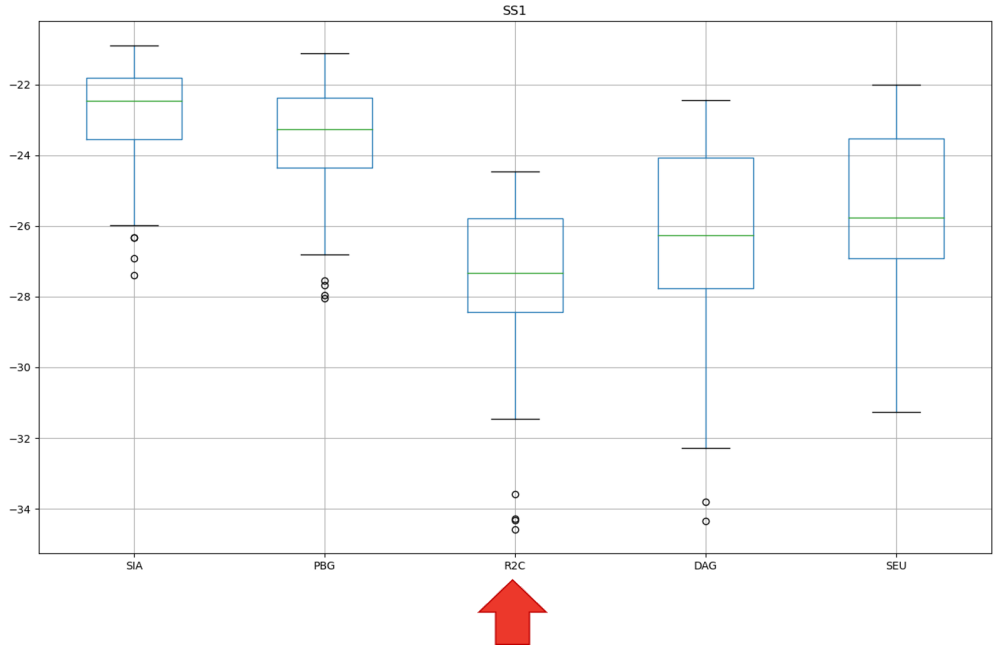
Cela revient ainsi à sélectionner 2 groupes par sous-site si possible, et donc à un maximum de 20 groupes distincts.

Une analyse des scores (sur le TOP50) sur `1SGZ` des différents groupes pour chaque sous-site a permis de considérer les groupes suivants pour le design (voir Tableau 3.12) :

- 6 dérivés de l'Uracile et 2 dérivés de la Cytosine pour les groupes PO ;
- 3 dérivés de l'Adénine, 2 dérivés de la Cytosine, 3 dérivés de l'Uracile et 2 dérivés de la Guanine

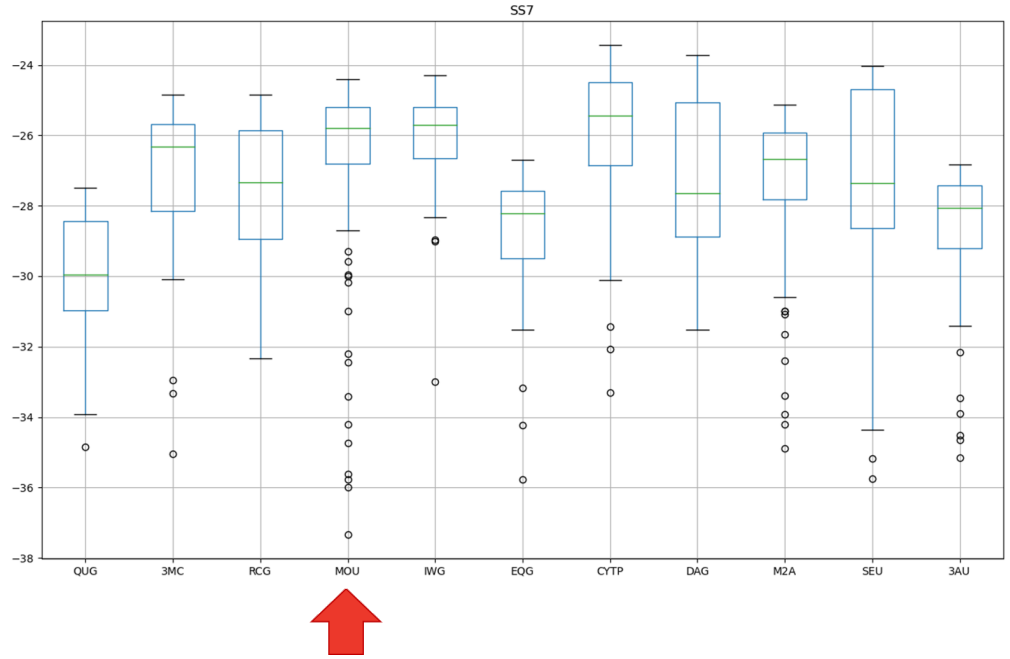


(a) Plage de scores pour chaque groupe spécifique de BACE1 pour le sous-site SS1 sur 1SZG.

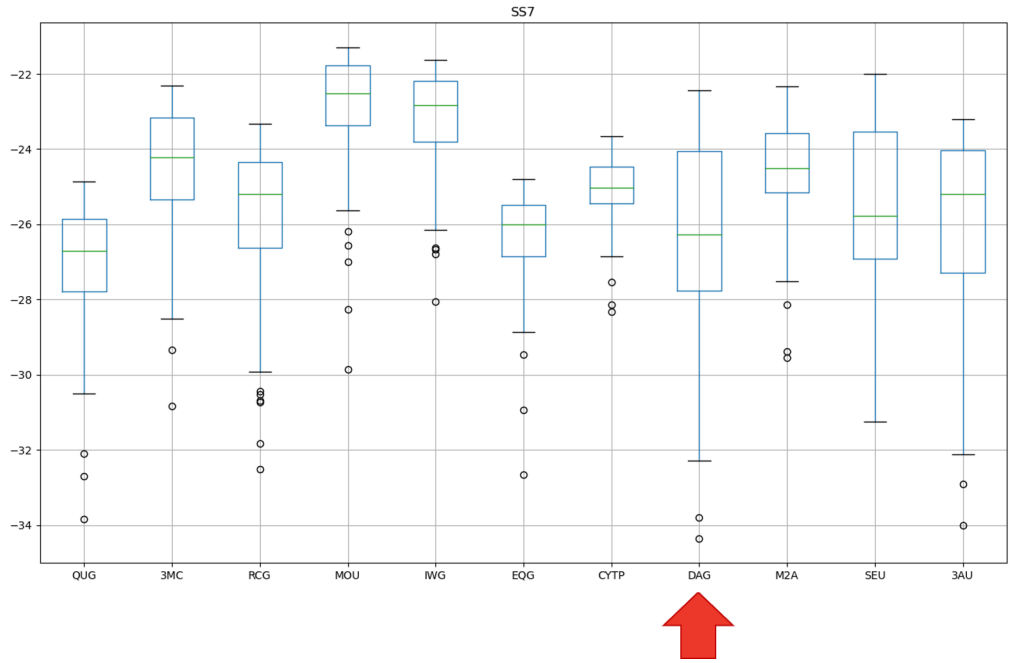


(b) Plage de scores pour chaque groupe spécifique de BACE1 pour le sous-site SS1 sur 4GID

FIGURE 3.16 – Plage de scores pour chaque groupe spécifique de BACE1 pour le sous-site SS1 sur 1SZG et 4GID. L'axe des abscisses correspond aux différents groupes spécifiques, tandis que l'axe des ordonnées correspond aux scores énergétiques pour ces groupes en kcal.mol^{-1}. La flèche rouge indique le meilleur groupe par la présence de poses de meilleurs scores représentées par les outliers.



(a) Plage de scores pour chaque groupe spécifique de BACE1 pour le sous-site SS7 sur 1SZG.



(b) Plage de scores pour chaque groupe spécifique de BACE1 pour le sous-site SS7 sur 4GID

FIGURE 3.17 – Plage de scores pour chaque groupe spécifique de BACE1 pour le sous-site SS1 sur 1SZG et 4GID. L'axe des abscisses correspond aux différents groupes spécifiques, tandis que l'axe des ordonnées correspond aux scores énergétiques pour ces groupes en kcal.mol^{-1}. La flèche rouge indique le meilleur groupe par la présence de poses de meilleurs scores représentées par les outliers.

Backbone	Groupes sélectionnés
PO	• SS1 : R2C (C) • SS2 : GCU (U) • SS3 : - • SS4 : R2C (C) • SS5 : OAU (U) • SS6 : 5HU/U8U (U) • SS7 : MOU (U) • SS8 : - • SS9 : CMU (U) • SS10 : SEU (U)
PS_2	• SS1 : 12A (A) • SS2 : 6GA (A) • SS3 : CYTP (C) • SS4 : GCU (U) • SS5 : GAU (U) • SS6 : DAG (G) • SS7 : 6GA/R2C (A/C) • SS8 : - • SS9 : EQG (G) • SS10 : SAU (U)

TABLE 3.12 – Sélection des meilleurs groupes spécifiques PO et PS_2 pour un exemple de design

pour les groupes PS_2

Nous avons réalisé 3 design sur BACE1 (conformation 1SGZ) pour lesquels la taille de la chaine est respectivement : k = 4, k = 5 et k = 6. Respectivement, nous avons pu générer 401 4-mers, 325 5-mers et 255 6-mers ; parmi lesquels nous dénombrons respectivement 10 séquences uniques dans chaque cas, et pour lesquels les scores énergétiques sont présentés dans le Tableau 3.13. Dans les trois cas, nous remarquons que la différence énergétique des MFE associées aux séquences est non-significative, n'excédant pas 3 kcal.mol^{-1}. Autrement dit, cela signifie que chacune de ces solutions pourraient être envisageables. Rappelons néanmoins que le score d'énergie associée à chacune de ces solutions correspond à un score avant minimisation. Ainsi, une étape de minimisation devrait être envisagée avec un re-calcul du score énergétique.
Par ailleurs, nous observons que certaines séquences passent deux fois par le même groupe nucléotidique, pour des séquences de taille k = 5 et k = 6 (R2C-CYT-SAU-SAU-OAU ou SAU-OAU-U8U-OAU-5HU-SAU) en particulier. Ceci semble signifier qu'il subsiste la possibilité de créer des solutions avec deux groupes tout au plus pouvant interagir avec un même sous-site.

Design et Docking sur 3ZKM. Afin de faire une comparaison avec BACE2, nous avons sélectionné la conformation 3ZKM, qui partage la particularité de ne pas avoir de ligands interagissant initialement avec son site actif. Nous avons utilisé la distribution focalisée sur le site actif générée sur 3ZKM.
Nous avons commencé par une étude de docking (contrainte de séquence), en criblant les 4-mer, 5-mer et 6-mer générés vis-à-vis de BACE1. Pour cette étude, en maintenant strictement les mêmes paramètres du `ColorDocking` que ceux utilisés pour BACE1, aucune solution n'a pu être générée. Dès lors, nous avons utilisé une distribution de 1 000 poses par groupe (500 ANTI/500 SYN), en maintenant les mêmes paramètres du `ColorDocking`. Nous avons généré des solutions, pour lesquels le score de la MFE pour chaque séquence est supérieur de plus de 30 kcal.mol^{-1} environ. Ces résultats indiquent donc que les solutions générées sur BACE1 interagiront plus favorablement avec cette enzyme qu'avec BACE2.

Nous avons complété notre étude avec un design sur BACE2. Les paramètres utilisés sont strictement identiques que ceux utilisés pour le design sur BACE1. Les résultats ont été les suivants :

- Pour k = 4, la meilleure solution générée est SEU-SAU-DAG-OAU avec un score de -113.0 kcal.mol^{-1}, soit un score énergétique plus élevé de 19 à 23 kcal.mol-1 par rapport aux solutions de BACE1.
- Pour k = 5, la meilleure solution générée est OAU-R2C-1MG-R2C-R2C avec un score de -138.8 kcal.mol^{-1}, soit un score énergétique plus élevé de 14 à 18 kcal.mol-1 par rapport aux solutions

Taille k *in-silico*-mer	**MFE Séquence**	**Energie (kcal.mol−1)** (BACE1)	**Energie (kcal.mol−1)** (BACE2)
4	OAU-SAU-6GA-U8U	-136.7	-97.2
4	OAU-SAU-OAU-U8U	-134.8	-108.4
4	OAU-SAU-GAU-U8U	-134.1	-98.6
4	OAU-SAU-GCU-U8U	-134.1	-100.2
4	OAU-SAU-OAU-SEU	-132.6	-104.9
4	MDU-OAU-SEU-GCU	-132.5	-103.4
4	OAU-SAU-12A-SEU	-132.5	-99.6
4	SAU-OAU-SEU-GCU	-132.4	-96.8
4	MDU-OAU-U8U-6GA	-132.2	-98.2
4	MDU-OAU-U8U-GCU	-132.2	-103.5
5	MDU-OAU-SEU-SAU-GCU	-156.7	-124.8
5	SAU-OAU-SEU-SAU-GCU	-156.6	-109.1
5	MDU-OAU-U8U-OAU-MOU	-155.9	-121.8
5	SAu-OAU-U8U-OAU-MOU	-155.8	-115.7
5	R2C-CYT-SAU-SAU-OAU	-155.1	-112.3
5	MDU-12A-SAU-GCU-SEU	-153.9	-96.9
5	OAU-U8U-OAU-MOU-SAU	-153.9	-121.5
5	MDU-OAU-U8U-OAU-5HU	-153.8	-123.6
5	MDU-OAU-SEU-OAU-MOU	-153.7	-122.7
5	SAU-OAU-U8U-OAU-5HU	-153.7	-123.4
6	MDU-OAU-U8U-OAU-MOU-SAU	-180.6	-141.7
6	SAU-OAU-U8U-OAU-MOU-SAU	-180.5	-126.0
6	GCU-U8U-CYT-SAU-SAU-OAU	-179.1	-
6	OAU-SEU-OAU-MOU-DAG-1MG	-178.7	-129.1
6	MDU-OAU-U8U-OAU-5HU-SAU	-178.5	-136.3
6	MDU-OAU-SEU-OAU-MOU-SAU	-178.4	-141.6
6	SAU-OAU-U8U-OAU-5HU-SAU	-178.4	-
6	SAU-OAU-SEU-OAU-MOU-SAU	-178.3	-137.1
6	OAU-SEU-OAU-MOU-SAU-1MG	-178.1	-134.1
6	GAU-U8U-CYT-SAU-SAU-OAH	-177.2	-

TABLE 3.13 – MFE de *in-silico*-mers prédites à partir de la distribution focalisée sur le site actif de 1SGZ pour générer des 4-mer, 5-mer et 6-mer, avec un docking de ces solutions sur 3ZKM. Le "-" indique qu'aucune solution n'a été trouvée pour la séquence donnée.

de BACE1.

- Pour $k = 6$, la meilleure solution générée est R2C-R2C-12A-OAU-SEU-OAU avec un score de -173.0 kcal.mol^{-1}, soit un score énergétique plus élevé de 4 à 8 kcal.mol-1.

Ainsi, les solutions incorporant des nucléotides considérés comme spécifiques vis-à-vis de BACE1 auront tendance à avoir un meilleur score énergétique vis-à-vis de BACE1 que BACE2.

3.3.9 . Conclusions

En conclusion, nous avons montré que les régions modulant l'activité de BACE1 (site actif et exosite) sont des régions favorables à des interactions avec des acides nucléiques. En effet, nous avons mis en évidence la présence de hotspots principalement au niveau de ces régions. Ces hotspots permettront, en plus des sous-sites du site actif, de concevoir une molécule qui peut interagir à la fois avec le site actif et l'exosite. En effet, le "rationnel drug design" consisterait à avoir différents hotspots se chevauchant les uns avec les autres, de la région du site actif à l'exosite.

De plus, une étude plus approfondie sur le site actif a permis de mettre en évidence des groupes spécifiques vis-à-vis des différents sous-sites de BACE1 et de BACE2. Ainsi, cette liste de groupes peut servir de bases afin de sélectionner ou exclure des groupes pour la conception d'*in-silico*-mers. Nous proposons néanmoins qu'une analyse d'interactions entre les différentes poses et le site actif puisse être réalisée afin de comprendre les raisons de cette spécificité. Parmi les hypothèses, nous pouvons citer (i) l'impact du volume de la cavité du site actif, et (ii) la présence de résidus différents entre BACE1 et BACE2 pour certains sous-sites. Cette deuxième hypothèse est d'autant plus forte que l'étude de Relation-Structure Activité confirme une corrélation possible entre l'activité d'un composé et des interactions avec de tels résidus du site actif, comme ce fut le cas pour le composé B11.

Par ailleurs, nous avons montré l'impact des groupes PO et PS_2, à la fois sur la présence de hotspots que sur sa capacité à avoir un rôle dans la spécificité. En effet, nous pouvons observer que certains groupes contactent un sous-site donné, uniquement avec la présence d'un groupement PO ou PS_2. Par exemple, pour le SS1, nous observons que les groupes `PBG`, `DAG` et `SEU` sont spécifiques de BACE1 uniquement avec le groupement PO, tandis que leur pendant avec le groupement PS_2 interagit avec le SS1 de BACE1 et de BACE2.

Sur la base de ces résultats, nous avons réalisé une application de design avec le `ColorDocking`, en générant des solutions spécifiques vis-à-vis du site actif de BACE1. Néanmoins, des analyses devront être réalisées plus loin, telles que : une analyse d'interactions entre ces composés et le site actif, une étude statistique qui puisse apporter des informations supplémentaires pour le design (par exemple,

les groupes et poses hautement favorables).

Enfin, uniquement avec un nombre total de 1500 poses environ et de 15 000 poses environ, nous avons pu réaliser nos études de docking et de design avec des temps de calculs d'une durée de quelques secondes à quelques minutes, pour des tailles $k = 4, 5, 6$. Ainsi, ces résultats laisse suggérer la possibilité de réaliser un design avec un nombre de groupes plus important. Un protocole peut donc être le suivant :

1. Générer un ligand de taille $k < 7$ focalisé sur le site actif
2. Générer un ligand de taille $k < 7$ focalisé sur l'exosite
3. Lier les deux parties en une seule molécule.

Enfin, nous suggérons de réaliser des tests avec des longueurs de k compris entre 7 et 10, afin de tester la possibilité d'un design d'un *in-silico*-mer interagissant à la fois avec le site actif et l'exosite.

Sous-site	Backbone	BACE1	BACE2
SS1	PO	SIA PBG R2C DAG SEU	3MC RCG 6GA IWG HCU MMC 6MA THY DWG M6A 13P 7MG SCU CMU
	PS_2	5AU SIA 12A CYTP R2C 6AA ISU 5FC	OMU IWG HCU MMA 1MP HWG 6MA DWG M6A 7MG MFC 3AU
SS2	PO	T6A 4MC GAU BUG 5MU HIA IAU SIA 5HU GCU 2SC 5MCP R2C 3MU PSU H2U	MMG MWG OMA 4OC BCU M2G IWG ADE MMA HWG OMT THY 2MG 52U 2MU TMC 3AU
	PS_2	T6A 4MC BUG 6GA GCU 12A PBG MMI 3MU PSU	5MU OMP 1MG MMA URA 1MP 6MA THY 1MA H2U TMC
SS3	PO	DAG	-
	PS_2	RCG CYTP R2C 5AU	-

SS4	PO	3MC MOU URA 2SC 5MCP PSU M2A MMI 6MA MCU 6IA HMC 5HU SEU 4MC 5MU IMG MDU R2C TMC	T6A 70U MWG OMA HWG OMT 27G SMA OMP MAC 5CU INO 6AA 52U MEU M7G 5AU MMG YYG OEU M2G GCU DCG OMI 2MG 13P 7MG MAU CMU N2G SAU BCU 4AC 5TU
	PS_2	4MC U8U 5HU MOU 1MG GCU CYU PBG 5MCP R2C 4SU	T6A MWG MMC CYT OMP 5CU OMC ADE IWG MAC MTA 2MU MCU 6IA OMU OEU MMA M3U OAU 2SU 3MP GAU 4OC 5MU MTG 4AC MDU SCU
SS5	PO	M2G CYT 6AA 52U SEU OAU	3MC PBG 5MCP DAG M2A 27G BUG K2C DWG MEU M7G 5AU HMC 5DU RCG GCU STU OMI M3U MAU N2G U8U MTG ADEP 12A SCU TMC
	PS_2	GAU CYTP R2C PSU 5MC SMA	HCU MMC DAG 27G BUG OMP OMC 6MA INO MIA MCU M7G IAU M2G STU OMI 7MG 1MA CMU N2G 3MP ADEP BCU EQG SCU TMC

SS6	PO	4MC 70U U8U 5HU IWG MTA CYT 1MP 4AC DWG MCU	T6A MSU MWG OMA SIA OMT M2A H2U 27G HIA OMP 5CU CYU 6AA 2MU MEU M7G 6IA QUG MMG HMC M2G GCU MMA DCG THY OMI 7MG MAU CMU N2G 3MP 4OC MTG SAU BCU 12A 5TU SCU TMC
	PS_2	N2G 5HU 1MG OCU DWG DAG MIA	T6A MSU 70U OMA 3MC SIA MMC PBG CYT 2SC PSU M2A H2U HNA SMA HIA MAC OMC ADE CYU MMI MTA 6MA 52U ISU 2MU M7G 6IA M1G OEU RCG M2G MMA GUA THY 3MU M6A 2MG 13P OMI CMU OAU 4MC 2SU 3MP BCU IMG 1MP 4AC 4SU 5MC SCU TMC 5FC 3AU
SS7	PO	QUG 3MC RCG MOU IWG EQG CYTP DAG M2A SEU 3AU	1MA MMG GUA
	PS_2	5AU RCG MOU 5HU HCU OCU 2SC OMT 5MCP M2A 7MG 1MA MFC SEU N2G U8U SAU ADEP 6GA ADE CYTP R2C INO 2MU M7G	MIA GUA
SS8	PO	-	ADEP 3MC 5MCP M6A M2A 27G M7G
	PS_2	-	-

SS9	PO	CMU	T6A MSU 70U OMA HCU OCU SIA 1MG 2SC 5MCP PSU DAG 27G BUG HIA OMP 6GA IWG ADE CYU K2C MMI MAC MTA DWG INO 6AA 2MU OMG MCU M7G 6IA 5DU MMG OMU QUG M1G YYG RCG 5HU M2G MMA STU 3MU M3U 13P 7MG 1MA MFC MAU N2G 2SU 4MC 4OC MTG 5MU IMG 12A 1MP 4AC 5TU 4SU 5MC SCU TMC 3AU
	PS_2	EQG	T6A MSU MWG OMA MOU SIA 1MG MMC CYT 2SC PSU M2A DAG H2U 27G SMA BUG HIA 6GA OMP IWG ADE CYU K2C OMC MAC MTA DWG INO 6AA 52U 2MU OMG MCU MEU MIA M7G 5DU QUG OMU MMG OEU RCG 5HU M2G MMA DCG STU 3MU M6A OMI 13P M3U MFC SEU MAU N2G 2SU U8U 4MC 4OC 5MU MTG SAU IMG 12A 1MP 4AC MDU 5TU 4SU TMC 5FC 3AU
SS10	PO	CYTP SEU SAU URA	IWG 5MCP 4SU DAG 27G
	PS_2	4OC SAU OMP 5HU M2G	MIA

TABLE 3.14 – Groupes spécifiques vis-à-vis de BACE1 et de BACE2, conformation indépendante, à haute solvatation

4 - Conclusions et Perspectives

4.1 . Conclusions et Perspectives

Au travers de cette thèse, nous avons montré la capacité d'approcher le design d'ARNsb par la méthode par fragments et la méthode computationnelle au travers du `ColorDocking`.

4.1.1 . Conclusions et Perspectives sur le `ColorDocking`

Conclusions. Le `ColorDocking` est un logiciel pour le design d'ARNsb, écrit par une collection de scripts `Python` à l'interface avec du code `C`. Nous avons implémenté deux algorithmes, que sont le MFEdock et l'Estimateur statistique.
Nous avons tout d'abord montré que l'un des points forts de notre algorithme exact réside dans sa complexité linéaire sur le nombre de poses connectées par paires, n'étant exponentielle que sur la longueur k de l'ARNsb. En tant que tel, il peut être considéré comme un algorithme de complexité paramétré, montrant que le problème MFEdock est FPT pour la longueur k de l'ARNsb. Par ailleurs, nous avons démontré la capacité de notre approche à s'adapter à différentes définitions de fragments, apportant ainsi une caractéristique plus universelle.

D'un point de vue pratique, nous avons montré l'intérêt de notre approche pour le design au travers de trois extensions pouvant être complémentaire :

- L'extension "Design" (sans contrainte de séquence) afin de concevoir des solutions liant favorablement la protéine ;
- L'extension "Docking" qui consiste à prédire une conformation à partir d'une séquence donnée. Ce module permettrait de comparer les énergies de liaison pour un ARNsb spécifique entre deux structures homologues (ou plus) ;
- l'extension "Estimateur Statistique" qui permet d'obtenir des informations non-triviales utiles pour guider le design d'ARNsb.

Perspectives. Comme nous l'avons mentionné, le `ColorDocking` est actuellement implémenté pour le design d'ARNsb interagissant avec une cible biologique, en supposant l'inexistence de motifs de structures secondaires. Ce non-repliement est notamment facilité par l'incorporation de nucléotides modifiés pouvant empêcher des appariements entre bases nucléiques. Pour cette implémentation, une perspective est d'appliquer le `ColorDocking` au travers de différentes réelles études de cas, afin de mieux comprendre l'impact des paramètres sur la qualité des résultats. Notamment, cette étude peut être réalisée sur l'enzyme BACE1, pour lequel une première application du `ColorDocking` a été proposée.

Parmi les autres points de perspectives, le `ColorDocking` pourrait inclure une implémentation pour tenir compte des structures secondaires pour le design d'ARN. Ces motifs peuvent être par exemple

des doubles hélices ou des tiges-boucles. Ainsi, deux aspects peuvent être envisagés :

1. Soit la structure 2D de l'ARN est connue, auquel cas, on peut utiliser une bibliothèque de motif 2D criblée sur la protéine. L'adaptation du `ColorDocking` serait faisable afin de reconstruire et prédire l'ARN natif (Docking) ;
2. Soit la structure 2D de l'ARN n'est pas connue, auquel cas cela demanderait un effort intellectuel pour mieux comprendre et intégrer le phénomène de coopération ARN/Protéine.

Enfin, le `ColorDocking` constitue avec NUCLEAR les seuls logiciels à des fins de design d'ARNsb thérapeutique à partir de l'approche par Fragments, en intégrant la connaissance de la structure 3D de la cible biologique. Une étude comparative entre les deux logiciels permettrait de mieux comprendre les similitudes et différences en termes de qualité de résultats.

4.1.2 . Conclusions et Perspectives sur BACE1

Conclusions. Nous avons mis en évidence des résultats montrant l'intérêt de concevoir un ARNsb vis-à-vis de cette enzyme. Les principaux résultats peuvent être résumés en ces différents points :

- le site catalytique et l'exosite de BACE1 sont des régions favorables à l'interaction avec des nucléotides standards et des nucléotides modifiés ;
- la prise en compte de différentes conformations (au niveau du site actif et de l'exosite) simulant la flexibilité de la protéine est nécessaire et peut impacter les régions favorables liant la bibliothèque de nucléotides ;
- la prise en compte de la solvatation est aussi importante : nous avons émis l'hypothèse que les molécules d'eau peuvent jouer le rôle d'intermédiaire entre les nucléotides et la protéine, ou empêcher des contacts directs entre des nucléotides et la protéine ;
- la prise en compte de différents types de la liaison phosphodiester (PO ou PS_2) est pertinente et peut jouer un rôle sur la spécificité des composés générés,
- la mise en évidence de groupes spécifiques associés à chaque sous-site de BACE1 et BACE2 ;
- la mise en évidence de contacts pouvant jouer un rôle dans la spécificité des composés.

Au travers de cette étude de cas, nous proposons ainsi un protocole de design basé sur l'approche par Fragments. Une première application du `ColorDocking` a été appliquée, avec la démonstration de notre capacité à proposer des *in-silico*-mers favorables au site actif de BACE1 par rapport à BACE2.

Perspectives. Notre analyse s'est restreinte sur quatre conformations de BACE1 et trois conformations de BACE2. Bien que le choix de différentes conformations tend à simuler la flexibilité, il n'est pas impossible que des *in-silico*-mers proposés ne puissent ne pas stabiliser une autre conformation

de BACE2. C'est pourquoi nous proposons à ce que des études de Dynamique Moléculaire soient réalisées comme validation *in-silico*, en amont ou en parallèle des études expérimentales.

Une étude statistique, avec le `ColorDocking`, pourrait également être réalisée afin de déterminer les poses hautement probables ou le profil de séquence, appliquée à la fois sur BACE1 et BACE2. Ceci apporterait des informations complémentaires utiles pour le design.

Par ailleurs, notre analyse sur la spécificité s'est restreinte au site actif. Nous proposons de réaliser une étude équivalente sur l'exosite afin de générer des *in-silico*-mers au niveau de cette région. Nous proposons qu'un criblage focalisée sur l'exosite soit réalisé, avec une boîte englobant la région de l'exosite telle que nous l'avons définie.

Bibliographie

[1] Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. *Discovering high-affinity ligands for proteins : SAR by NMR*. *Science*, 274(5292) :1531–1534, 1996. doi :10.1126/science.274.5292.1531.

[2] Bollag, G., Tsai, J., Zhang, J., Zhang, C., Ibrahim, P., Nolop, K., and Hirth, P. *Vemurafenib : the first drug approved for BRAF-mutant cancer*. *Nature Reviews Drug Discovery*, 11(11) :873–886, 2012. doi :10.1038/nrd3847.

[3] Tap, W. D., Wainberg, Z. A., Anthony, S. P., Ibrahim, P. N., Zhang, C., Healey, J. H., Chmielowski, B., Staddon, A. P., Cohn, A. L., Shapiro, G. I., Keedy, V. L., Singh, A. S., Puzanov, I., Kwak, E. L., Wagner, A. J., Hoff, D. D. V., Weiss, G. J., Ramanathan, R. K., Zhang, J., Habets, G., Zhang, Y., Burton, E. A., Visor, G., Sanftner, L., Severson, P., Nguyen, H., Kim, M. J., Marimuthu, A., Tsang, G., Shellooe, R., Gee, C., West, B. L., Hirth, P., Nolop, K., van de Rijn, M., Hsu, H. H., Peterfy, C., Lin, P. S., Tong-Starksen, S., and Bollag, G. *Structure-Guided Blockade of CSF1R Kinase in Tenosynovial Giant-Cell Tumor*. *New England Journal of Medicine*, 373(5) :428–437, 2015. doi :10.1056/nejmoa1411366.

[4] Perera, T. P. S., Jovcheva, E., Mevellec, L., Vialard, J., Lange, D. D., Verhulst, T., Paulussen, C., Ven, K. V. D., King, P., Freyne, E., Rees, D. C., Squires, M., Saxty, G., Page, M., Murray, C. W., Gilissen, R., Ward, G., Thompson, N. T., Newell, D. R., Cheng, N., Xie, L., Yang, J., Platero, S. J., Karkera, J. D., Moy, C., Angibaud, P., Laquerre, S., and Lorenzi, M. V. *Discovery and Pharmacological Characterization of JNJ-42756493 (Erdafitinib), a Functionally Selective Small-Molecule FGFR Family Inhibitor*. *Molecular Cancer Therapeutics*, 16(6) :1010–1020, 2017. doi : 10.1158/1535-7163.mct-16-0589.

[5] Schoepfer, J., Jahnke, W., Berellini, G., Buonamici, S., Cotesta, S., Cowan-Jacob, S. W., Dodd, S., Drueckes, P., Fabbro, D., Gabriel, T., Groell, J.-M., Grotzfeld, R. M., Hassan, A. Q., Henry, C., Iyer, V., Jones, D., Lombardo, F., Loo, A., Manley, P. W., Pellé, X., Rummel, G., Salem, B., Warmuth, M., Wylie, A. A., Zoller, T., Marzinzik, A. L., and Furet, P. *Discovery of Asciminib (ABL001), an Allosteric Inhibitor of the Tyrosine Kinase Activity of BCR-ABL1*. *Journal of Medicinal Chemistry*, 61(18) :8120–8135, 2018. doi :10.1021/acs.jmedchem.8b01040.

[6] Souers, A. J., Leverson, J. D., Boghaert, E. R., Ackler, S. L., Catron, N. D., Chen, J., Dayton, B. D., Ding, H., Enschede, S. H., Fairbrother, W. J., Huang, D. C. S., Hymowitz, S. G., Jin, S., Khaw,

S. L., Kovar, P. J., Lam, L. T., Lee, J., Maecker, H. L., Marsh, K. C., Mason, K. D., Mitten, M. J., Nimmer, P. M., Oleksijew, A., Park, C. H., Park, C.-M., Phillips, D. C., Roberts, A. W., Sampath, D., Seymour, J. F., Smith, M. L., Sullivan, G. M., Tahir, S. K., Tse, C., Wendt, M. D., Xiao, Y., Xue, J. C., Zhang, H., Humerickhouse, R. A., Rosenberg, S. H., and Elmore, S. W. *ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets*. *Nature Medicine*, 19(2) :202–208, 2013. doi :10.1038/nm.3048.

[7] Scott, D. E., Coyne, A. G., Hudson, S. A., and Abell, C. *Fragment-Based Approaches in Drug Discovery and Chemical Biology*. *Biochemistry*, 51(25) :4990–5003, 2012. doi :10.1021/bi3005126.

[8] Kumar, A., Voet, A., and Zhang, K. Y. J. *Send Orders of Reprints at reprints@benthamscience.org Fragment Based Drug Design : From Experimental to Computational Approaches*. Technical report, 2012.

[9] Murray, C. W. and Verdonk, M. L. *The consequences of translational and rotational entropy lost by small molecules on binding to proteins*. *Journal of Computer-Aided Molecular Design*, 16(10) :741–753, 2002. doi :10.1023/A:1022446720849.

[10] de Souza Neto, L. R., Moreira-Filho, J. T., Neves, B. J., Maidana, R. L. B. R., Guimarães, A. C. R., Furnham, N., Andrade, C. H., and Silva, F. P. *In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery*. *Frontiers in Chemistry*, 8, 2020. doi :10.3389/fchem.2020.00093.

[11] Agu, P. C., Afiukwa, C. A., Orji, O. U., Ezeh, E. M., Ofoke, I. H., Ogbu, C. O., Ugwuja, E. I., and Aja, P. M. *Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management*. *Scientific Reports*, 13(1) :13398, 2023. doi :10.1038/s41598-023-40160-2.

[12] Zoete, V., Grosdidier, A., and Michielin, O. *Docking, virtual high throughput screening and in silico fragment-based drug design*. *J. Cell. Mol. Med*, 13(2) :238–248, 2009. doi :10.1111/j.1582-4934.2009.00665.x.

[13] González-Alemán, R., Chevrollier, N., Simoes, M., Montero-Cabrera, L., and Leclerc, F. *MCSS-Based Predictions of Binding Mode and Selectivity of Nucleotide Ligands*. *Journal of Chemical Theory and Computation*, 17(4) :2599–2618, 2021. doi :10.1021/acs.jctc.0c01339.

[14] Yu, W. and Mackerell, A. D. *Computer-aided drug design methods*. In *Methods in Molecular Biology*, volume 1520, pages 85–106. Humana Press Inc., 2017. doi :10.1007/978-1-4939-6634-9{_}5.

[15] Bacilieri, M. and Moro, S. *Ligand-Based Drug Design Methodologies in Drug Discovery Process : An Overview*. Technical report, 2006.

[16] Schubert, C. R. and Stultz, C. M. *The multi-copy simultaneous search methodology : a fundamental tool for structure-based drug design*. *Journal of Computer-Aided Molecular Design*, 23(8) :475–489, 2009. doi :10.1007/s10822-009-9287-y.

[17] Moniot, A. *Modélisation 3D de complexes ARN-protéine par assemblage combinatoire de fragments structuraux*. Technical report, 2022.

[18] González-Alemán, R. *Computational Fragment-Based Design of Chemically Modified Oligonucleotides for Selective Protein Inhibition : BACE1 as a Case Study*. Ph.D. thesis, 2023.

[19] Moniot, A., Guermeur, Y., de Vries, S. J., and de Beauchene, I. *ProtNAff : protein-bound Nucleic Acid filters and fragment libraries*. *Bioinformatics*, 38(16) :3911–3917, 2022. doi :10.1093/bioinformatics/btac430.

[20] de˜Beauchene, I. C., de˜Vries, S. J., and Zacharias, M. *Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins*. *Nucleic Acids Research*, 44(10) :4565–4580, 2016. doi :10.1093/nar/gkw328.

[21] Cook, S. A. *The Complexity of Theorem-Proving Procedures* *. Technical report, 1973.

[22] Levin, L. A. *Universal sequential search problems*. *Problemy peredachi informatsii*, 9(3) :115–116, 1973.

[23] Scornavacca, C. and Weller, M. *Treewidth-based algorithms for the small parsimony problem on networks*. *Algorithms for Molecular Biology*, 17(1) :15, 2022. doi :10.1186/s13015-022-00216-w.

[24] Marchand, B., Ponty, Y., and Bulteau, L. *Tree diet : reducing the treewidth to unlock FPT algorithms in RNA bioinformatics*. *Algorithms for Molecular Biology*, 17(1) :8, 2022. doi : 10.1186/s13015-022-00213-z.

[25] Garey, M. R. and Johnson, D. S. *Computers and intractability : a guide to the theory of NP-completeness*. W.H. Freeman, 1979.

[26] Jörg Flum, M. G. *Parameterized Complexity Theory*. Springer Berlin, Heidelberg, Berlin, 1 edition, 2006.

[27] Alon, N., Yuster, R., and Zwick, U. *Color-coding*. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing - STOC 94*. ACM Press, 1994. doi :10.1145/195058.195179.

[28] Alon, N., Dao, P., Hajirasouliha, I., Hormozdiari, F., and Sahinalp, S. C. *Biomolecular network motif counting and discovery by color coding*. *Bioinformatics*, 24(13) :i241–i249, 2008. doi :10.1093/bioinformatics/btn163.

[29] Shlomi, T., Segal, D., Ruppin, E., and Sharan, R. *QPath : a method for querying pathways in a protein-protein interaction network*. *BMC Bioinformatics*, 7(1), 2006. doi :10.1186/1471-2105-7-199.

[30] Dost, B., Shlomi, T., Gupta, N., Ruppin, E., Bafna, V., and Sharan, R. *QNet : A Tool for Querying Protein Interaction Networks*. *Journal of Computational Biology*, 15(7) :913–925, 2008. doi :10.1089/cmb.2007.0172.

[31] Naor, M., Schulman, L. J., and Srinivasan, A. *Splitters and near-optimal derandomization*. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*. IEEE Comput. Soc. Press, 1995. doi :10.1109/sfcs.1995.492475.

[32] Sarfaraz, S., Muneer, I., and Liu, H. *Combining fragment docking with graph theory to improve ligand docking for homology model structures*. *Journal of Computer-Aided Molecular Design*, 34(12) :1237–1259, 2020. doi :10.1007/s10822-020-00345-7.

[33] Duesbury, E., Holliday, J., and Willett, P. *Comparison of Maximum Common Subgraph Isomorphism Algorithms for the Alignment of 2D Chemical Structures*. *ChemMedChem*, 13(6) :588–598, 2018. doi :https://doi.org/10.1002/cmdc.201700482.

[34] Smith, C. I. E. and Zain, R. *Therapeutic Oligonucleotides : State of the Art*. *Annual Review of Pharmacology and Toxicology*, 59(1) :605–630, 2019. doi :10.1146/annurev-pharmtox-010818-021050.

[35] Hammond, S. M., Aartsma-Rus, A., Alves, S., Borgos, S. E., Buijsen, R. A. M., Collin, R. W. J., Covello, G., Denti, M. A., Desviat, L. R., Echevarría, L., Foged, C., Gaina, G., Garanto, A., Goyenvalle, A. T., Guzowska, M., Holodnuka, I., Jones, D. R., Krause, S., Lehto, T., Montolio, M., Roon-Mom, W. V., and Arechavala-Gomeza, V. *Delivery of oligonucleotide-based therapeutics : challenges and opportunities*. *EMBO Molecular Medicine*, 13(4) :e13243, 2021. doi :10.15252/emmm.202013243.

[36] Xiang, J., Zhang, W., Cai, X. F., Cai, M., Yu, Z. H., Yang, F., Zhu, W., Li, X. T., Wu, T., Zhang, J. S., and Cai, D. F. *DNA Aptamers Targeting BACE1 Reduce Amyloid Levels and Rescue Neuronal Deficiency in Cultured Cells*. *Molecular Therapy - Nucleic Acids*, 16 :302–312, 2019. doi :10.1016/j.omtn.2019.02.025.

[37] Liang, H., Shi, Y., Kou, Z., Peng, Y., Chen, W., Li, X., Li, S., Wang, Y., Wang, F., and Zhang, X. *Inhibition of BACE1 Activity by a DNA Aptamer in an Alzheimer's Disease Cell Model*. *PLOS ONE*, 10(10) :e0140733–, 2015.

[38] Edwards, P. D., Albert, J. S., Sylvester, M., Aharony, D., Andisik, D., Callaghan, O., Campbell, J. B., Carr, R. A., Chessari, G., Congreve, M., Frederickson, M., Folmer, R. H. A., Geschwindner, S., Koether, G., Kolmodin, K., Krumrine, J., Mauger, R. C., Murray, C. W., Olsson, L.-L., Patel, S., Spear, N., and Tian, G. *Application of Fragment-Based Lead Generation to the Discovery of Novel, Cyclic Amidine β-Secretase Inhibitors with Nanomolar Potency, Cellular Activity, and High Ligand Efficiency*. *Journal of Medicinal Chemistry*, 50(24) :5912–5925, 2007. doi :10.1021/jm070829p.

[39] Darmostuk, M., Rimpelova, S., Gbelcova, H., and Ruml, T. *Current approaches in SELEX : An update to aptamer selection technology*. *Biotechnology Advances*, 33(6, Part 2) :1141–1161, 2015. doi :https://doi.org/10.1016/j.biotechadv.2015.02.008.

[40] Gasse, C., Zaarour, M., Noppen, S., Abramov, M., Marlière, P., Liekens, S., DeStrooper, B., and Herdewijn, P. *Modulation of BACE1 Activity by Chemically Modified Aptamers*. *ChemBioChem*, 19(7) :754–763, 2018. doi :https://doi.org/10.1002/cbic.201700461.

[41] Pfeiffer, F., Rosenthal, M., Siegl, J., Ewers, J., and Mayer, G. *Customised nucleic acid libraries for enhanced aptamer selection and performance. Current opinion in biotechnology*, 48 :111–118, 2017. doi :10.1016/j.copbio.2017.03.026.

[42] Verma, S. and Eckstein, F. *MODIFIED OLIGONUCLEOTIDES : Synthesis and Strategy for Users*. Technical report, 1998.

[43] Gupta, S., Hirota, M., Waugh, S. M., Murakami, I., Suzuki, T., Muraguchi, M., Shibamori, M., Ishikawa, Y., Jarvis, T. C., Carter, J. D., Zhang, C., Gawande, B., Vrkljan, M., Janjic, N., and Schneider, D. J. *Chemically modified DNA aptamers bind interleukin-6 with high affinity and inhibit signaling by blocking its interaction with interleukin-6 receptor. The Journal of biological chemistry*, 289(12) :8706–8719, 2014. doi :10.1074/jbc.M113.532580.

[44] Barage, S. H. and Sonawane, K. D. *Amyloid cascade hypothesis : Pathogenesis and therapeutic strategies in Alzheimer's disease*. *Neuropeptides*, 52 :1–18, 2015. doi :https://doi.org/10.1016/j.npep.2015.06.008.

[45] Liu, P.-P., Xie, Y., Meng, X.-Y., and Kang, J.-S. *History and progress of hypotheses and clinical trials for Alzheimer's disease*. *Signal Transduction and Targeted Therapy*, 4(1) :29, 2019. doi :10.1038/s41392-019-0063-8.

[46] Glenner, G. G. and Wong, C. W. *Alzheimer's disease : Initial report of the purification and characterization of a novel cerebrovascular amyloid protein*. *Biochemical and Biophysical Research Communications*, 120(3) :885–890, 1984. doi :https://doi.org/10.1016/S0006-291X(84)80190-4.

[47] *2023 Alzheimer's disease facts and figures*. *Alzheimer's & Dementia*, 19(4) :1598–1695, 2023. doi : https://doi.org/10.1002/alz.13016.

[48] Hardy, J. A. and Higgins, G. A. *Alzheimer's Disease : The Amyloid Cascade Hypothesis*. *Science*, 256(5054) :184–185, 1992. doi :10.1126/science.1566067.

[49] Chen, G.-f., Xu, T.-h., Yan, Y., Zhou, Y.-r., Jiang, Y., Melcher, K., and Xu, H. E. *Amyloid beta : structure, biology and structure-based therapeutic development*. *Acta Pharmacologica Sinica*, 38(9) :1205–1235, 2017. doi :10.1038/aps.2017.28.

[50] Bachurin, S., Bovina, E., and Ustyugov, A. *Drugs in Clinical Trials for Alzheimer's Disease : The Major Trends*. *Medicinal research reviews*, 37, 2017. doi :10.1002/med.21434.

[51] Selkoe, D. J. and Schenk, D. *Alzheimer's Disease : Molecular Understanding Predicts Amyloid-Based Therapeutics*. *Annual Review of Pharmacology and Toxicology*, 43(1) :545–584, 2003. doi : 10.1146/annurev.pharmtox.43.100901.140248.

[52] Weggen, S. and Beher, D. *Molecular consequences of amyloid precursor protein and presenilin mutations causing autosomal-dominant Alzheimer's disease*. *Alzheimer's Research & Therapy*, 4(2) :9, 2012. doi :10.1186/alzrt107.

[53] Mullan, M., Crawford, F., Axelman, K., Houlden, H., Lilius, L., Winblad, B., and Lannfelt, L. *A pathogenic mutation for probable Alzheimer's disease in the APP gene at the N–terminus of β–amyloid*. *Nature Genetics*, 1(5) :345–347, 1992. doi :10.1038/ng0892-345.

[54] Di Fede, G., Catania, M., Morbin, M., Rossi, G., Suardi, S., Mazzoleni, G., Merlin, M., Giovagnoli, A. R., Prioni, S., Erbetta, A., Falcone, C., Gobbi, M., Colombo, L., Bastone, A., Beeg, M., Manzoni, C., Francescucci, B., Spagnoli, A., Cantù, L., Del Favero, E., Levy, E., Salmona, M., and Tagliavini, F. *A Recessive Mutation in the APP Gene with Dominant-Negative Effect on Amyloidogenesis*. *Science*, 323(5920) :1473–1477, 2009. doi :10.1126/science.1168979.

[55] Johnston, J., O'Neill, C., Lannfelt, L., Winblad, B., and Cowburn, R. F. *The significance of the Swedish APP670/671 mutation for the development of Alzheimer's disease amyloidosis*. *Neurochemistry International*, 25(1) :73–80, 1994. doi :10.1016/0197-0186(94)90056-6.

[56] Yoshikai, S.-i., Sasaki, H., Doh-ura, K., Furuya, H., and Sakaki, Y. *Genomic organization of the human amyloid beta-protein precursor gene*. *Gene*, 87(2) :257–263, 1990. doi :https://doi.org/10.1016/0378-1119(90)90310-N.

[57] Wolfe, M. S., Xia, W., Ostaszewski, B. L., Diehl, T. S., Kimberly, W. T., and Selkoe, D. J. *Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and γ-secretase activity*. *Nature*, 398(6727) :513–517, 1999. doi :10.1038/19077.

[58] Shen, J. and Kelleher, R. J. *The presenilin hypothesis of Alzheimer's disease : Evidence for a loss-of-function pathogenic mechanism*. *Proceedings of the National Academy of Sciences*, 104(2) :403–409, 2007. doi :10.1073/pnas.0608332104.

[59] Bentahir, M., Nyabi, O., Verhamme, J., Tolia, A., Horré, K., Wiltfang, J., Esselmann, H., and De Strooper, B. *Presenilin clinical mutations can affect γ-secretase activity by different mechanisms*. *Journal of Neurochemistry*, 96(3) :732–742, 2006. doi :https://doi.org/10.1111/j.1471-4159.2005.03578.x.

[60] Ma, J., Yee, A., Brewer, H. B., Das, S., and Potter, H. *Amyloid-associated proteins α1-antichymotrypsin and apolipoprotein E promote assembly of Alzheimer β-protein into filaments*. *Nature*, 372(6501) :92–94, 1994. doi :10.1038/372092a0.

[61] Wisniewski, T., Castano, E. M., Golabek, A., Vogel, T., and Frangione, B. *National Institutes of Health grants AG 10953*. Technical Report 5, 1994.

[62] Castano, E. M., Prelli, F., Wisniewski, T., Golabek, A., Kumar, R. A., Soto, C., and Frangione, B. *Fibrillogenesis in Alzheimer's disease of amyloid β peptides and apolipoprotein E*. *Biochemical Journal*, 306(2) :599–604, 1995. doi :10.1042/bj3060599.

[63] Fondation Vaincre Alzheimer. *La recherche médicale spécialisée dans la maladie d'Alzheimer : État des lieux et perspectives d'espoir*. Technical report.

[64] Budd Haeberlein, S., Aisen, P. S., Barkhof, F., Chalkias, S., Chen, T., Cohen, S., Dent, G., Hansson, O., Harrison, K., von Hehn, C., Iwatsubo, T., Mallinckrodt, C., Mummery, C. J., Muralidharan, K. K., Nestorov, I., Nisenbaum, L., Rajagovindan, R., Skordos, L., Tian, Y., van Dyck, C. H., Vellas, B., Wu, S., Zhu, Y., and Sandrock, A. *Two Randomized Phase 3 Studies of Aducanumab in Early Alzheimer's Disease*. *The Journal of Prevention of Alzheimer's Disease*, 9(2) :197–210, 2022. doi :10.14283/jpad.2022.30.

[65] Kuller, L. H. and Lopez, O. L. *ENGAGE and EMERGE : Truth and consequences ? Alzheimer's & Dementia*, 17(4) :692–695, 2021. doi :https://doi.org/10.1002/alz.12286.

[66] van Dyck, C. H., Swanson, C. J., Aisen, P., Bateman, R. J., Chen, C., Gee, M., Kanekiyo, M., Li, D., Reyderman, L., Cohen, S., Froelich, L., Katayama, S., Sabbagh, M., Vellas, B., Watson, D., Dhadda, S., Irizarry, M., Kramer, L. D., and Iwatsubo, T. *Lecanemab in Early Alzheimer's Disease*. *New England Journal of Medicine*, 388(1) :9–21, 2022. doi :10.1056/NEJMoa2212948.

[67] Sims, J. R., Zimmer, J. A., Evans, C. D., Lu, M., Ardayfio, P., Sparks, J., Wessels, A. M., Shcherbinin, S., Wang, H., Monkul Nery, E. S., Collins, E. C., Solomon, P., Salloway, S., Apostolova, L. G., Hansson, O., Ritchie, C., Brooks, D. A., Mintun, M., Skovronsky, D. M., and Investigators, T.-A. . *Donanemab in Early Symptomatic Alzheimer Disease : The TRAILBLAZER-ALZ 2 Randomized Clinical Trial*. *JAMA*, 330(6) :512–527, 2023. doi :10.1001/jama.2023.13239.

[68] Wen, W., Li, P., Liu, P., Xu, S., Wang, F., and Huang, J. H. *Post-Translational Modifications of BACE1 in Alzheimer's Disease*. *Current Neuropharmacology*, 20(1) :211–222, 2021. doi :10.2174/1570159x19666210121163224.

[69] Zhang, X. and Song, W. *The role of APP and BACE1 trafficking in APP processing and amyloid-β generation*. *Alzheimer's Research & Therapy*, 5(5) :46, 2013. doi :10.1186/alzrt211.

[70] Ovsepian, S. V., Horacek, J., O'Leary, V. B., and Hoschl, C. *The Ups and Downs of BACE1 : Walking a Fine Line between Neurocognitive and Other Psychiatric Symptoms of Alzheimer's Disease*. *The Neuroscientist*, 27(3) :222–234, 2020. doi :10.1177/1073858420940943.

[71] Vassar, R., Kovacs, D. M., Yan, R., and Wong, P. C. *The β-secretase enzyme BACE in health and Alzheimer's disease : Regulation, cell biology, function, and therapeutic potential*. In *Journal of Neuroscience*, volume 29, pages 12787–12794. 2009. doi :10.1523/JNEUROSCI.3657-09.2009.

[72] Vassar, R., Kuhn, P.-H., Haass, C., Kennedy, M. E., Rajendran, L., Wong, P. C., and Lichtenthaler, S. F. *Function, therapeutic potential and cell biology of BACE proteases : current status and future prospects*. *Journal of Neurochemistry*, 130(1) :4–28, 2014. doi :https://doi.org/10.1111/jnc.12715.

[73] Lombardo, S., Chiacchiaretta, M., Tarr, A., Kim, W., Cao, T., Sigal, G., Rosahl, T. W., Xia, W., Haydon, P. G., Kennedy, M. E., and Tesco, G. *BACE1 partial deletion induces synaptic plasticity deficit in adult mice*. *Scientific Reports*, 9(1) :19877, 2019. doi :10.1038/s41598-019-56329-7.

[74] Satir, T. M., Agholme, L., Karlsson, A., Karlsson, M., Karila, P., Illes, S., Bergström, P., and Zetterberg, H. *Partial reduction of amyloid β production by β-secretase inhibitors does not decrease synaptic transmission*. *Alzheimer's Research & Therapy*, 12(1) :63, 2020. doi :10.1186/s13195-020-00635-0.

[75] Shippy, D. C. and Ulland, T. K. *Microglial Immunometabolism in Alzheimer's Disease*. *Frontiers in Cellular Neuroscience*, 14, 2020. doi :10.3389/fncel.2020.563446.

[76] Cai, Y., Liu, J., Wang, B., Sun, M., and Yang, H. *Microglia in the Neuroinflammatory Pathogenesis of Alzheimer's Disease and Related Therapeutic Targets*. *Frontiers in Immunology*, 13, 2022. doi : 10.3389/fimmu.2022.856376.

[77] Singh, N., Benoit, M. R., Zhou, J., Das, B., Davila-Velderrain, J., Kellis, M., Tsai, L.-H., Hu, X., and Yan, R. *BACE-1 inhibition facilitates the transition from homeostatic microglia to DAM-1*. *Science Advances*, 8(24) :eabo1286, 2023. doi :10.1126/sciadv.abo1286.

[78] Singh, N., Das, B., Zhou, J., Hu, X., and Yan, R. *Targeted BACE-1 inhibition in microglia enhances amyloid clearance and improved cognitive performance*. *Science Advances*, 8(29) :eabo3610, 2023. doi :10.1126/sciadv.abo3610.

[79] McDade, E., Voytyuk, I., Aisen, P., Bateman, R. J., Carrillo, M. C., De Strooper, B., Haass, C., Reiman, E. M., Sperling, R., Tariot, P. N., Yan, R., Masters, C. L., Vassar, R., and Lichtenthaler, S. F. *The case for low-level BACE1 inhibition for the prevention of Alzheimer disease*. *Nature Reviews Neurology*, 17(11) :703–714, 2021. doi :10.1038/s41582-021-00545-1.

[80] Leclerc, F. *La recherche sur la maladie d'Alzheimer, des raisons d'espérer . . .*. Technical Report 44000.

[81] Shimizu, H., Tosaki, A., Kaneko, K., Hisano, T., Sakurai, T., and Nukina, N. *Crystal Structure of an Active Form of BACE1, an Enzyme Responsible for Amyloid β Protein Production*. *Molecular and Cellular Biology*, 28(11) :3663–3671, 2008. doi :10.1128/mcb.02185-07.

[82] Xu, Y., Li, M.-j., Greenblatt, H., Chen, W., Paz, A., Dym, O., Peleg, Y., Chen, T., Shen, X., He, J., Jiang, H., Silman, I., and Sussman, J. L. *Flexibility of the flap in the active site of BACE1 as revealed by crystal structures and molecular dynamics simulations*. *Acta Crystallographica Section D*, 68(1) :13–25, 2012. doi :10.1107/S0907444911047251.

[83] Hu, H., Chen, Z., Xu, X., and Xu, Y. *Structure-Based Survey of the Binding Modes of BACE1 Inhibitors*. *ACS Chemical Neuroscience*, 10(2) :880–889, 2019. doi :10.1021/acschemneuro.8b00420.

[84] Atwal, J. K., Chen, Y., Chiu, C., Mortensen, D. L., Meilandt, W. J., Liu, Y., Heise, C. E., Hoyte, K., Luk, W., Lu, Y., Peng, K., Wu, P., Rouge, L., Zhang, Y., Lazarus, R. A., Scearce-Levie, K., Wang, W., Wu, Y., Tessier-Lavigne, M., and Watts, R. J. *A Therapeutic Antibody Targeting BACE1 Inhibits Amyloid-β Production in Vivo*. *Science Translational Medicine*, 3(84) :43–84, 2011. doi : 10.1126/scitranslmed.3002254.

[85] Wang, W., Liu, Y., and Lazarus, R. A. *Allosteric inhibition of BACE1 by an exosite-binding antibody*. *Current Opinion in Structural Biology*, 23(6) :797–805, 2013. doi :https://doi.org/10.1016/j.sbi.2013.08.001.

[86] Ruderisch, N., Schlatter, D., Kuglstatter, A., Guba, W., Huber, S., Cusulin, C., Benz, J., Rufer, A. C., Hoernschemeyer, J., Schweitzer, C., Bülau, T., Gärtner, A., Hoffmann, E., Niewoehner, J., Patsch, C., Baumann, K., Loetscher, H., Kitas, E., and Freskgård, P.-O. *Potent and Selective BACE-1 Peptide Inhibitors Lower Brain Aβ Levels Mediated by Brain Shuttle Transport*. *eBioMedicine*, 24 :76–92, 2017. doi :10.1016/j.ebiom.2017.09.004.

[87] Mirsafian, H., Mat Ripen, A., Merican, A. F., and Mohamad, S. B. *Amino Acid Sequence and Structural Comparison of BACE1 and BACE2 Using Evolutionary Trace Method*. *The Scientific World Journal*, 2014 :482463, 2014. doi :10.1155/2014/482463.

[88] Wyss, D. F., Wang, Y.-S., Eaton, H. L., Strickland, C., Voigt, J. H., Zhu, Z., and Stamford, A. W. *Combining NMR and X-ray crystallography in fragment-based drug discovery : Discovery of highly potent and selective BACE-1 inhibitors*. *Topics in Current Chemistry*, 317 :83–114, 2012. doi :10.1007/128-2011-183.

[89] Hernández-Rodríguez, M., Correa-Basurto, J., Gutiérrez, A., Vitorica, J., and Rosales-Hernández, M. C. *Asp32 and Asp228 determine the selective inhibition of BACE1 as shown by docking and molecular dynamics simulations*. *European Journal of Medicinal Chemistry*, 124 :1142–1154, 2016. doi :https://doi.org/10.1016/j.ejmech.2016.08.028.

[90] Hall, D., Li, S., Yamashita, K., Azuma, R., Carver, J. A., and Standley, D. M. *RNALIM : A novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure*. *Analytical Biochemistry*, 472 :52–61, 2015. doi :10.1016/j.ab.2014.11.004.

[91] Kappel, K. and Das, R. *Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking*. *Structure*, 27(1) :140–151.e5, 2019. doi :10.1016/j.str.2018.10.001.

[92] Itai, A., Papadimitriou, C. H., and Szwarcfiter, J. L. *Hamilton Paths in Grid Graphs*. *SIAM Journal on Computing*, 11(4) :676–686, 1982. doi :10.1137/0211056.

[93] Marx, D. *Parameterized Complexity of Independence and Domination on Geometric Graphs*. In M. A. Bodlaender Hans L.
}and Langston, editor, *Parameterized and Exact Computation*, pages 154–165. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[94] Alon, N., Yuster, R., and Zwick, U. *Color-Coding*. *J. ACM*, 42(4) :844–856, 1995. doi :10.1145/210332.210337.

[95] Schmidt, J. P. and Siegel, A. *The Spatial Complexity of Oblivious k-Probe Hash Functions*. *SIAM Journal on Computing*, 19(5) :775–786, 1990. doi :10.1137/0219054.

[96] Waterman, M. S. and Byers, T. H. *A dynamic programming algorithm to find all solutions in a neighborhood of the optimum*. *Mathematical Biosciences*, 77(1-2) :179–188, 1985. doi :10.1016/0025-5564(85)90096-3.

[97] Miranker, A. and Karplus, M. *Functionality maps of binding sites : A multiple copy simultaneous search method*. *Proteins : Structure, Function, and Genetics*, 11(1) :29–34, 1991. doi :10.1002/prot.340110104.

[98] González-Alemán, R. and Leclerc, F. *NUCLEotide AssembleR (NUCLEAR) package*, 2023.

[99] de Vries, S. J., Schindler, C. E. M., de Beauchêne, I. C., and Zacharias, M. *A Web Interface for Easy Flexible Protein-Protein Docking with ATTRACT*. *Biophysical Journal*, 108(3) :462–465, 2015. doi :https://doi.org/10.1016/j.bpj.2014.12.015.

[100] Perzanowska, O., Smietanski, M., Jemielity, J., and Kowalska, J. *Chemically Modified Poly(A) Analogs Targeting PABP : Structure Activity Relationship and Translation Inhibitory Properties*. *Chemistry – A European Journal*, 28(42) :e202201115, 2022. doi :https://doi.org/10.1002/chem.202201115.

[101] Ginarte, Y. , Aleman, R. G., Yacoub, T., Leclerc, F., Cabrera, L. A. M., González, R., Cabrera, M., and Montero, L. A. *Identification of selective inhibitors against BACE 1 over BACE 2 in Alzheimer's disease by Quantitative Structure-Activity Relationship (QSAR)*. Technical report, 2022.

[102] Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. *MolProbity : All-atom structure validation for macromolecular crystallography*. *Acta Crystallographica Section D : Biological Crystallography*, 66(1) :12–21, 2010. doi :10.1107/S0907444909042073.

[103] Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. *PROPKA3 : Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions*. *Journal of Chemical Theory and Computation*, 7(2) :525–537, 2011. doi :10.1021/ct100578z.

[104] Meyder, A., Nittinger, E., Lange, G., Klein, R., and Rarey, M. *Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures*. *Journal of Chemical Information and Modeling*, 57(10) :2437–2447, 2017. doi :10.1021/acs.jcim.7b00391.

[105] Vainio, M. J., Puranen, J. S., and Johnson, M. S. *ShaEP : Molecular Overlay Based on Shape and Electrostatic Potential*. *Journal of Chemical Information and Modeling*, 49(2) :492–502, 2009. doi :10.1021/ci800315d.

[106] Karami, Y., Rey, J., Postic, G., Murail, S., Tufféry, P., and De Vries, S. J. *DaReUS-Loop : a web server to model multiple loops in homology models*. *Nucleic Acids Research*, 47(W1) :W423–W428, 2019. doi :10.1093/nar/gkz403.