Exploration et analyse immersives de données moléculaires guidées par la tâche et la modélisation sémantique des contenus

Mikael Trellet

Numéro NNT : 2015SACLS262

THÈSE DE DOCTORAT DE
L'UNIVERSITÉ PARIS SACLAY

Spécialité :
Informatique

Présentée par :

Mikael Trellet

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ PARIS SACLAY

Sujet de la thèse :

Exploration et analyse immersives de données moléculaires guidées par la tâche et la modélisation sémantique des contenus

Soutenue le vendredi 18 décembre 2015

devant le jury composé de :

Indira Thouvenin	Enseignant-chercheur HDR / UMR 7253 Heudiasyc	Rapporteur
Serge Pérez	Directeur de recherche émérite CNRS / UMR 5063 DPM	Rapporteur
Alexandre Bonvin	Professeur Utrecht university / Bijvoet Center	Examinateur
Nicolas Sabouret	Professeur CNRS / LIMSI	Examinateur / Président
Patrick Bourdot	Directeur de Recherche CNRS / LIMSI	Directeur de thèse
Marc Baaden	Directeur de Recherche CNRS / LBT	Directeur de thèse
Nicolas Férey	Maître de Conférence Université Paris-Sud / LIMSI	Co-encadrant

Groupe VENISE
LIMSI-CNRS
B.P. 133
91403 Orsay Cedex, France

LBT
IBPC-CNRS
13 Rue Pierre et Marie Curie
75005 Paris, France

Remerciements

Conclure plusieurs mois de travail d'un manuscrit, comme paraphe de plusieurs années de travail, devrait constituer un moment un peu plus léger et un exercice facile après toutes ces pages écrites. Ce moment de remerciements va, semble-t-il, se révéler beaucoup plus difficile que prévu ! En témoignent les 34 changements de phrase d'accroche qui sont intervenus avant d'en arriver à ce point de votre lecture...

Puisque j'ai appris avec ces derniers mois d'écriture l'importance de l'organisation, et que ce concept me tient tant à cœur, j'ai décidé de ne pas l'appliquer pour cette partie, concentrons-nous sur le fond et non la forme. Et comme j'ai maintenant le droit de commencer par la fin, je m'excuse déjà pour les personnes que je vais omettre dans cette petite dédicace, vous qui me côtoyez quotidiennement, de façon hebdomadaire ou semestrielle, vous comptez évidemment pour beaucoup dans la réalisation de cette grande étape de ma vie de chercheur !

C'est naturellement que mes collègues, devenu rapidement des camarades et maintenant des amis me viennent à l'esprit. Je ne vais pas tout de suite les remercier, mais d'abord m'excuser platement. Durant de vagues et courts moments de lucidité, je pense m'être rendu compte à quel point travailler dans mon bureau peut être éprouvant, que ce soit pour les oreilles ou pour la concentration. Je conçois difficilement plus de 3h de travail sans une petite pause détente au détour d'un article sur le nouveau casque de Réalité Virtuelle transparent ou du rachat de la dernière technologie de tracking olfactive par Apple . Ces moments d'échange sont précieux pour alléger un peu le temps de travail au bureau, mais doivent certainement constituer une excuse des plus justifiées pour me détester. Alors Pierre, Gabriel, Remy, Weiya, Yuji et Xavier, désolé ! Ok, du coup je peux rayer ces noms de ma liste, un beau tir groupé pour commencer..! Sauf que techniquement je ne les ai pas encore remerciés... Ça compte de s'excuser ? Bon, ça ne va pas être très dur de les remercier en fait. Merci les gars (oui vous noterez le sexisme évident que la RV subit...) pour votre patience, mais aussi votre aide, vos rires (forcés ou non ;) à mes quelques calembours de tous les jours ! Je mettrais en avant Weiya, jamais un dérapage pour nous dire de nous taire ou nous calmer, toujours à l'écoute et même humoriste ponctuel, tel un sniper, qui arrive à dérider (jeu de mots très involontaire) le plus triste de ses collègues. À l'opposé on trouve Xavier, moins ponctuel, plus régulier dans ses interventions, mais il faut pouvoir raconter les 142 spectacles des Nuls, les 6 saisons de Kaamelott et les 19 albums de Chantal Goya. Je ne sais pas quels seraient nos scores de compatibilité si on appelait le 3618 Compatibility, mais on peut dire sans sourciller qu'on a passé de très bons moments tous les deux, studieux ou non ! Du coup il est tout naturel de passer à Alexandre ! Nous n'étions pas colocataires, mais il était plus prudent pour notre productivité respective que ce soit le cas..! Les quelques voyages en Xsara coupé/cabriolet m'ont confirmé, si jamais j'avais pu avoir des doutes, que nos discussions n'avaient de fin que quand le panneau de la gare de Juvisy illuminait le pare-brise avant de la voiture. Je crois que seuls les trajets en vélo du camping de Bidart jusqu'au lieu de l'école d'été de RV en 2014 à Biarritz ont pu couper nos échanges géopolitico-ludiques.

Une pensée amicale pour Pierre en passant, l'"Ancien", qui a fini sa thèse une bonne année et demie avant moi et qui fut avant Xavier mon compagnon de travail ! Le seul avec qui je pouvais parler un peu de notre passion commune, l'OM, et de sa passion commun-iste !

Je suis resté dans le Sud francilien jusqu'à maintenant, je n'oublie évidemment pas mes collègues et amis parisiens ! Alors je ne peux jurer qu'ils se souviennent tous de moi tant ces derniers mois mes visites furent rares... Mais le proverbe ne ment pas, elles avaient beau être brèves, elles étaient souvent intenses et je découvris mes premiers afterwork autour de légers cocktails sans alcools dans certains bars usités du voisinage de l'université de Jussieu à ces occasions-là. Lord Sir Dr Alex Tek, Yoann et puis Benoist et Samuel, les experts de GLIC, et accessoirement les doyens des non-permanents du LBT si je ne me trompe pas et surtout deux papas jonglant entre famille et science comme GLIC jongle avec ses molécules d'eau, de manière plutôt efficace !

Récent papa également, je ne peux aller plus loin sans remercier Marc, dont les projets et la motivation ont su me ramener en France et avec qui les échanges professionnels ou non étaient toujours enrichissants ! Son calme et sa pédagogie resteront des exemples dont j'essayerai de m'inspirer dans ma vie professionnelle future, car je sais que ce ne sont pas mes plus grandes qualités...

Et quand on parle d'enfants, difficile de ne pas en venir à Nico, excellente transition puisqu'il a lui aussi transité entre Paris et Orsay..! Quel homme ! S'occuper de 3 enfants à la maison est loin d'être une tâche facile, mais alors quand en plus vous en avez un 4e au boulot... Tout mon respect pour celui qui a tout simplement permis que ces 3 ans et quelques mois finissent sur un travail dont je pense être fier, lui qui, contre vents et marées, distille toujours son optimisme débordant quand vous avez un petit coup de mou, mais qui sait également vous ramener les pieds sur terre quand vous pensez avoir résolu le théorème de Fermat alors que vous avez juste compris pourquoi votre matrice de projection est fausse depuis le début. L'homme aux 10 idées par secondes, 8 quand il a eu une nuit un peu courte, une source d'inspiration quand on pense qu'on a beaucoup trop travaillé dans la semaine et qu'on se rend compte qu'on a tout juste atteint la moitié de ce que lui a fait... Merci Nico, tu es et resteras l'un des principaux maîtres d'oeuvre de mes modestes travaux, ton nom serait également sur la première page que je n'en serais qu'encore plus fier.

Nous sommes revenus à Orsay, et ce n'est pas plus mal, je n'avais pas été exhaustif, loin de là ! Le 2d Nico, par ordre d'apparition et d'éloignement du lieu de naissance (:p), est un peu le grand frère, seul postdoc de la bande, il est au four et au moulin pour que tout se passe bien dans EVE, et dieu sait que l'informatique peut être capricieuse quand elle s'y met... Alors merci de ton dévouement, le Scale-1 te le rendra... peut-être un jour !

Mon cher Jean-Marc, j'ai attendu un petit peu avant de m'occuper de ton cas, qui est très sérieux, le docteur, que je ne suis pas encore, le voit très bien. Je pense que tu es une sorte d'OVNI, une sorte de François Truffaut qui se serait perdu dans Matrix, et tel Jim Carrey dans Bruce-tout-puissant, tu avances sans te retourner dans la vie, sauf pour regarder dans le rétro de ta Kangoo si Xavier tient toujours fermement les sangles de ton armoire récupérée par on ne sait quel stratagème machiavélique dans les sous-sols d'une certaine université Cristollienne... Peu de septuagénaires m'ont paru aussi jeunes et frais que toi, et je le dis sans arrières pensés, nos échanges autour du sacro-saint café quotidien font, sans hésiter, partis de mes meilleurs moments au LIMSI. Il me tarde bien souvent de tailler la bavette (à l'exception de celle du CESFO cependant) avec le puits de curiosité et de science que tu es. Au moins autant pour glisser une ou deux conneries monumentales que pour t'écouter nous commenter de ta vue acérée les dernières actualités. Merci pour tous ces moments et ta bonne humeur légendaire, ça fait plaisir d'avoir quelqu'un avec qui échangé à toute heure de la journée !

Patrick, directeur de thèse et capitaine du voilier VENISE, on ne peut pas dire que tu y ailles avec le dos de la cuillère pour que ton/notre paquebot ne coule pas ! Aussi sanguin que Jean-Marc est calme, il ne te faut pourtant pas quelques secondes pour rebondir sur une remarque humoristique ou un jeu de mots qui passe dans le couloir séparant nos bureaux. Nos échanges scientifiques ont toujours l'effet escompté, mettre des mots et des explications sur les tortueux méandres de la Réalité Virtuelle et de ses nombreuses facettes. Merci beaucoup d'avoir accepté de prendre le risque de recruter un jeune bio-informaticien pour évoluer dans ton précieux coin de paradis virtuel répondant au doux nom d'EVE...

Je n'oublie évidemment pas Hwang et ses poignées de main énergiques, Rémy, le passionné traversant régulièrement ces deux mondes que sont le scoutisme et la RV, qu'on pourrait vouloir opposer, mais qui, je pense, ont, ayant connu les deux, certainement un peu plus de points communs qu'on ne pourrait le penser !

Vous avez tous permis que je puisse m'épanouir au travail, mais, même avec toute la meilleure volonté du monde, il est difficile de profiter pleinement de son boulot quand sa vie privée n'est pas au top... Heureusement pour moi, ce n'est pas le cas et je le dois à des amis et une famille en or, comblant au centuple ce que quelqu'un peut espérer de sa petite vie d'étudiant en thèse. Pas de noms ce serait trop long, mais merci à vous tous, je sais que vous vous reconnaîtrez, déjà parce que vous aurez lu ce quelques lignes, et ensuite parce que j'aurais tenu à partager avec vous la joie d'avoir fini ce doctorat !

I could not continue without a brief but deep thought about my ex-coworkers, now my friends, the NL crew as I like to call them. I did not make my thesis among you, first because most of you finished it way before me, but also because the French fresh air was irresistible at the time of my decision to come bakc here. However, you all played a major role in my decision to accept the challenge of a thesis, inspired by your wonderful experiences..! I will never regret it, so thank you Adrien, Alex, Charleen, Ezgi, Gydo, Joao, Marc, Panos for showing me the good way and sharing, still today, so much good times all together ! Your are examples, in and out the science, I cannot value more the bonds we have ;)

Bon et comme je suis moi et que depuis que j'ai 2 ans je ne peux m'empêcher de garder le meilleur pour la fin, je dédie ces dernières lignes nocturnes à ma chérie, soutien indéfectible de ces 3 ans, source inépuisable de gentillesse et d'attention et surtout brin de femme qui a su partager et supporter avec patience, et ce tous les jours, les hauts et bas de ces derniers 39 mois ! Merci du fond du coeur !!!

Table des matières

Introduction

Contexte et problématique

Il est possible de diviser la biologie structurale et donc l'étude théorique de structures moléculaires en trois activités principales organisées selon le processus séquentiel suivant : (1) l'**expérimentation**, (2) la **modélisation moléculaire** et (3) la **simulation moléculaire** des structures 3d. Chacune de ces étapes met en jeu des outils de visualisation et d'analyse permettant d'étudier les données générées, de les interpréter pour les étapes suivantes et enfin de produire de nouvelles connaissances et d'établir de nouvelles hypothèses scientifiques. Plusieurs itérations de ce processus sont nécessaires pour caractériser les mécanismes en jeu dans un complexe moléculaire dans son environnement, chaque nouveau processus prenant comme données d'entrée une partie des résultats générés par l'analyse des données du processus précédent.

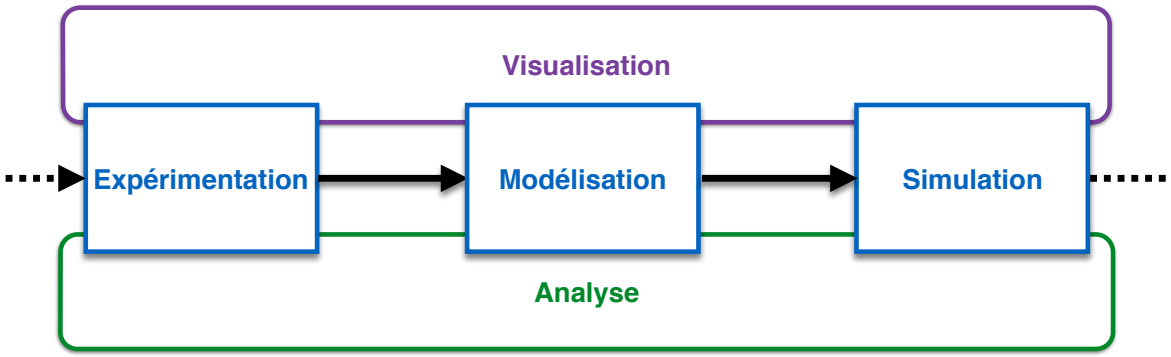


FIGURE 1 – *Description schématique des différentes étapes dans l'étude de structures moléculaires en biologie structurale. Les trois grandes étapes sont représentées en bleu, les outils de visualisation et d'analyses présents tout au long du processus sont représentés respectivement en violet et vert.*

Au sein de ce processus, la performance croissante des outils de calcul informatique entraîne aujourd'hui la génération de volumes de données considérables, en particulier dans la phase de simulation. De ce fait, les modèles de complexes moléculaires étudiés sont beaucoup plus gros et sont caractérisés par des résultats de plus en plus précis et détaillés. Aujourd'hui, l'efficacité croissante des programmes de simulation numérique moléculaire conduit à la production de données de trajectoires moléculaires, suite de modèles 3d décrivant l'évolution temporelle de structures moléculaires, pouvant atteindre plusieurs millions de particules à une précision atomique [144]. En parallèle, la performance des outils de manipulation et de visualisation moléculaire n'a pas progressé de manière suffisamment significative pour absor-

ber l'évolution des ressources de calcul et de stockage dédiées à la simulation moléculaire. L'étape de visualisation n'est pas la seule victime de l'amélioration des moyens de calcul. En effet, cette efficacité croissante des outils de simulation n'a pas été compensée par une augmentation suffisante des moyens d'analyse et la quantité de données générée est bien souvent très supérieure à la quantité de données pouvant être traitées par les experts scientifiques. De même, les capacités de calcul ont depuis longtemps dépassé les capacités de stockage, pourtant elles-mêmes en croissance forte [185]. **Visualiser, modéliser, et analyser des complexes moléculaires de très grande taille sont donc devenus un enjeu crucial en biologie moléculaire**. Enfin, la taille des données impacte directement l'efficacité et la rapidité des échanges de données entre les centres de calcul et les laboratoires de biologie. La minimisation de la quantité de données à échanger, à stocker et à analyser constitue une problématique importante pour les laboratoires.

L'approche dite *In Situ* répond à cette dernière problématique, en déportant les étapes de rendu graphique et d'analyse dans les centres de calcul [98, 107]. Cette approche suppose cependant de prévoir les analyses et les représentations visuelles qui seront nécessaires à la compréhension du phénomène. Les résultats de simulation, les représentations visuelles et les résultats d'analyse de la trajectoire sont ensuite analysés par l'expert en restant sur les centres de calcul. Cette méthodologie est cependant très difficile à mettre en pratique, car le choix des analyses requises dépend du résultat de la simulation que l'expert visualise en choisissant des modalités de visualisation, par ailleurs spécifiques à chaque phénomène étudié. Planifier les analyses et les rendus graphiques pourrait donc techniquement résoudre cette problématique de stockage et de transfert des données massives, mais ne correspond pas aux usages et aux contraintes dans un domaine dans lequel l'expert intervient nécessairement à chaque étape du processus. Pour minimiser les données échangées, **il s'agit donc de rapprocher les étapes de simulation, de visualisation, d'analyse tout en permettant à l'expert d'effectuer des visualisations et des analyses à la demande sur le lieu de simulation.**

La biologie structurale a toujours su intégrer au sein de chacune de ses étapes les résultats du domaine des sciences et technologies de l'information. Du traitement des signaux utilisé dans les méthodes expérimentales jusqu'à la visualisation de structures 3d, les avancées technologiques en informatique ont toujours été rapidement intégrées au coeur de les outils de la biologie moléculaire. Des concepts issus de la Réalité Virtuelle, comme la stéréoscopie, ont rapidement été appliqués pour mieux appréhender des contenus moléculaires intrinsèquement tridimensionnels [173, 165, 124, 77]. La Réalité Virtuelle est aussi associée à un espace de travail très important avec un point de vue dynamique et adaptatif facilitant l'exploration de complexes moléculaires de grande taille. Enfin, des dispositifs d'interaction 3d permettant d'interagir et de ressentir les forces en jeux dans un système moléculaire, hier confidentiels, sont désormais utilisés et intégrés dans les usages et les outils de la biologie moléculaire. On peut donc s'attendre à ce que le domaine de la biologie moléculaire continue d'intégrer dans ses usages les nouvelles technologies et dispositifs de Réalité Virtuelle, en particulier les dispositifs financièrement très abordables donnant accès à une immersion de grande qualité sur le lieu de travail des experts.

Cependant, l'immersion comporte des limites qu'il convient d'adresser avant de pouvoir constituer un outil récurrent en biologie structurale. La navigation au sein de dispositifs immersifs dans des données abstraites est un premier obstacle. Le mal du simulateur, souvent provoqué par l'activité de navigation dans les scènes virtuelles immersives réalistes et amplifié lors de l'exploration de données abstraites comme les données scientifiques, dégrade l'expérience de l'utilisateur, en particulier dans les dispositifs immersifs récents et grand public comme les HMD. **Des paradigmes de navigation adaptés aux contenus moléculaires**

abstraits et à la tâche de l'expert sont un prérequis à l'intégration de l'immersion en biologie moléculaire.

Par ailleurs, les interactions dans un environnement immersif ne peuvent être médiatisées par les dispositifs classiques comme la souris et le clavier. Elles sont mises de côté au profit d'interactions directes avec les objets virtuels présentés via des gestes et/ou par des commandes vocales. Par ailleurs, les environnements immersifs imposent un contexte d'interaction unique et homogène, avec un espace de travail non fenêtré. Pour répondre à ces contraintes propres aux environnements immersifs et les rendre opérationnels pour le domaine de la biologie moléculaire, **il s'agit de rapprocher les étapes de simulation, de visualisation, d'analyse dans un contexte interactif unique, en favorisant les interactions directes**.

Approche générale

Motivé par le manque de paradigmes de navigation spécifiques pour le domaine de la biologie structural, notre premier axe de travail s'est concentré sur le **développement de méthodes de navigation immersive basées sur le contenu et la tâche en biologie structurale**. Ces méthodes et paradigmes s'inspirent de particularités géométriques souvent observées dans la plupart des complexes moléculaires : un agencement des sous-unités de façon symétrique [66]. Notre approche n'est néanmoins aucunement contrainte à des complexes moléculaires symétriques puisque toute structure particulière retrouvée au sein d'une molécule nous permet de mettre en place nos paradigmes de navigation. Dans nos paradigmes, tout au long de son exploration, l'utilisateur 1) garde un point de vue contrôlé sur son complexe, 2) possède des repères spatiaux fixes afin de se situer dans la scène, 3) peut utiliser des chemins préférentiels basés sur la nature des objets observés, et qui respectent les deux conditions précédentes pour se déplacer pour atteindre des régions d'intérêt. Ces apports répondent respectivement aux problèmes de mal du simulateur pouvant être ressentis à cause: 1) d'une variation trop rapide de l'orientation de l'utilisateur par rapport à sa cible visuelle, 2) d'une dégradation de la conscience spatiale de l'utilisateur dans sa scène, 3) d'un trop grand nombre d'étapes ou de temps pour atteindre une région d'intérêt tout en gardant une bonne conscience spatiale de la scène virtuelle. En contraignant la navigation autour de chemins répondant aux objectifs d'exploration, via une restriction des directions du mouvement et des orientations de la caméra, nous fournissons des guides de navigation adaptés aux interactions pouvant prendre place au sein des environnements immersifs. De la même manière, nous anticipons certaines tâches d'exploration en fournissant des paradigmes automatisant certains déplacements considérés comme habituels en exploration moléculaire. Comme l'immersion sera probablement intégrée de manière progressive dans les usages, nous avons souhaité parallèlement au développement de paradigmes de navigation immersive dédiés, profiter de la démocratisation des périphériques mobiles en une démarche parallèle afin de **concevoir des solutions de visualisation en immersion réduite et avec une perception de la profondeur des contenus moléculaires**.

Par ailleurs, pour adresser la problématique liée à la quantité de données, nous avons travaillé sur une nouvelle approche afin **de regrouper des espaces de visualisation et d'analyse dans un contexte interactif homogène grâce à une modélisation sémantique**, afin de raccourcir la boucle d'étude des données de simulation moléculaire et pour répondre à la fois aux nouveaux enjeux de la biologie moléculaire et aux contraintes des environnements immersifs. Nous avons été inspirés par les techniques de *Visual analytics* visant à fournir de l'interactivité entre plusieurs représentations d'un phénomène ou de ses analyses

pour faciliter pour l'expert les corrélations [94] et la mise en relation de résultats de visualisation et d'analyse. Pour ce faire, notre approche s'est basée sur la mise place d'une couche d'abstraction décrivant les données manipulées usuellement par les experts du domaine. Cette couche d'abstraction a pu être mise en place par la construction d'une description sémantique des concepts mis en jeu lors de la visualisation et l'analyse de structures moléculaires. Cette description fut l'objet de la définition d'une ontologie, afin de formaliser l'ensemble des concepts manipulés lors d'une session de travail en intégrant les connaissances que possède l'expert scientifique en biologie moléculaire [17]. Cette représentation sémantique nous permet d'effectuer des liens interactifs entre les différentes représentations analytiques ou scientifiques d'objets d'intérêt présentés selon différentes modalités.

Plan du manuscrit

Notre travail étant à la frontière entre la biologie moléculaire et l'informatique, nous avons porté une attention particulière pour présenter aux lecteurs des chapitres de présentation, nous espérons clairs et concis, des différents domaines concernés, qui ne font pas nécessairement partie du domaine d'expertise du lecteur. Nous aborderons l'ensemble de nos contributions et les contextes dans lesquels ils s'inscrivent au travers de 5 chapitres distincts, le premier chapitre pouvant être lu rapidement par les lecteurs familiarisés avec la biologie moléculaire, le second chapitre pouvant être lu rapidement par les lecteurs familiarisés avec la Réalité Virtuelle.

Le **premier chapitre** fera un retour sur notre domaine d'application, la biologie structurale, à travers sa définition, ses usages et ses enjeux. Nous identifierons tout d'abord les acteurs et processus du domaine que constitue la biologie moléculaire, nous concentrant principalement sur le coeur de la transformation de l'information génétique en acteurs du fonctionnement d'une cellule vivante. Cela nous amènera à cerner la place de la biologie structurale au sein de la biologie moléculaire. Les différentes méthodes et outils caractérisant la biologie structurale seront ensuite énumérés pour finalement énoncer ses limites et anticiper ses futurs usages.

Le **second chapitre** abordera la notion d'immersion de l'utilisateur sous-jacente à la Réalité Virtuelle. Nous définirons tout d'abord ce domaine à travers trois de ses principaux axes : l'immersion, l'interaction et la navigation. Ses apports à la biologie structurale seront ensuite discutés pour finalement identifier les limites actuelles de son application en biologie structurale.

Nous développerons dans le **troisième chapitre** nos contributions pour amener la dimension immersive au coeur de trois activités importantes de la biologie structurale : l'exploration et la navigation dédiées à l'étude de complexes moléculaires, la diffusion au sein de la communauté de structures 3d, le contrôle rapide de l'état d'une simulation numérique moléculaire.

Le **quatrième chapitre** sera consacré à la présentation de pistes scientifiques pouvant apporter des réponses concrètes aux limites actuelles de l'utilisation de la Réalité Virtuelle en biologie structurale. Parmi ces pistes, une approche de type *Visual Analytics* mobilisant conjointement plusieurs méthodes et techniques empruntées à la représentation des connaissances et à l'interaction homme machine. Nous mettrons en avant l'apport de l'interactivité qu'elle introduit pour les tâches d'analyses et de visualisation moléculaire. Nous introduirons plus particulièrement comment l'interactivité entre ces deux tâches hétérogènes peut être offerte grâce à la création d'un cadre sémantique homogène.

Finalement, le **cinquième chapitre** sera consacré aux modalités de conception et d'implémentation d'un prototype d'application rapprochant les espaces d'analyses et de visualisa-

tion au sein d'un même contexte interactif, adapté aux environnements immersifs en privilégiant les interactions directes. Nous montrerons comment les outils issus du web sémantique en modélisant conjointement les contenus moléculaires et les interactions possibles sur ces contenus, peuvent supporter le rapprochement des phases d'analyse et de visualisation.

Chapitre 1

Contextes, usages et enjeux en biologie structurale

Dans ce chapitre sont exposés les fondements de la biologie structurale à travers sa place dans le monde de la biologie, son champ d'études et les outils et techniques qu'elle met en jeu au travers des spectres expérimentaux et théoriques. Cette définition de notre domaine d'application permettra de mettre en avant ses besoins spécifiques et ses enjeux. A travers ceux-ci nous verrons comment la Réalité Virtuelle et l'immersion peuvent apporter une approche complémentaire et pertinente pour les experts en biologie structurale.

Sommaire

Introduction

La biologie et les grands sous-domaines qui la composent ont pour objectif de caractériser les organismes vivants afin de comprendre leur fonctionnement, du niveau d'échelle macroscopique jusqu'à l'échelle moléculaire et atomique. Ainsi, l'*anatomie* a pour but de décrire les différents systèmes et appareils participant aux grandes fonctions des organismes. L'*histologie* concerne l'étude des organes et des tissus. La *biologie cellulaire* s'attelle à faire progresser la connaissance et la compréhension des mécanismes régissant le fonctionnement de la cellule. La *biologie moléculaire* s'intéresse de manière plus spécifique à la nature et à la structure des biomolécules, aux interactions avec leur environnement, ainsi qu'à leur synthèse dans le milieu cellulaire. Le but essentiel de la biologie moléculaire est d'expliquer des phénomènes biologiques de grande échelle à partir d'observations théoriques ou expérimentales effectuées à l'échelle moléculaire, ou à partir de l'étude des réactions chimiques et métaboliques induites par les interactions entre ces biomolécules et leur environnement. Enfin, la *biologie structurale* s'intéresse plus exclusivement aux structures tridimensionnelles des biomolécules et aux étapes permettant à une biomolécule d'acquérir cette structure. En effet, c'est la structure tridimensionnelle qui conditionne à la fois les interactions entre les biomolécules avec leur environnement et leur comportement dynamique, desquels découlent leur rôle et leur fonction dans la cellule. La biologie structurale mobilise des approches expérimentales, théoriques et computationnelles, basées sur des méthodes expérimentales et des lois physiques choisies en fonction de l'échelle ou de la nature des phénomènes étudiés.

Nous décrirons en premier lieu dans ce chapitre les différents acteurs et processus de la biologie moléculaire pour ensuite citer les méthodes expérimentales et théoriques contemporaines du domaine de la biologie structurale et enfin, dans la dernière partie, nous présenterons les enjeux et nouveaux usages en biologie structurale avant d'introduire les perspectives dans lesquelles s'inscrivent nos contributions.

1.1 Acteurs et processus de la biologie moléculaire

Les biomolécules sont impliquées dans le fonctionnement des organismes vivants et plus particulièrement de leur sous-unité la plus importante : la cellule. On retrouve parmi ces biomolécules, les **molécules d'eau**, qui constituent souvent la part majoritaire dans la composition des organismes, les **lipides**, qui sont les composants de base des membranes cellulaires permettant de créer des cloisons et compartiments, les **acides nucléiques** qui sont les constituants de l'ARN et de l'ADN, support de l'information génétique, les **acides aminés** qui forment les protéines, principaux acteurs du fonctionnement cellulaire, les **sucres** qui jouent un rôle fondamental dans de nombreux processus, puis diverses autres molécules, par exemple des cofacteurs comme l'hème.

Nos travaux de recherche portent sur les biomolécules de plus grande taille, appelées également macromolécules, comme l'ADN et les protéines. L'information génétique stockée dans l'ADN est transmise et conservée de génération en génération grâce à la reproduction. Les protéines sont à la fois les ouvriers, les briques et les messagers impliqués dans le fonctionnement cellulaire. Nous porterons une attention particulière aux édifices macromoléculaires composés de plusieurs protéines appelés complexes protéiques, même si les résultats de nos travaux de recherches s'appliquent à une large gamme de biomolécules, des plus simples aux plus complexes.

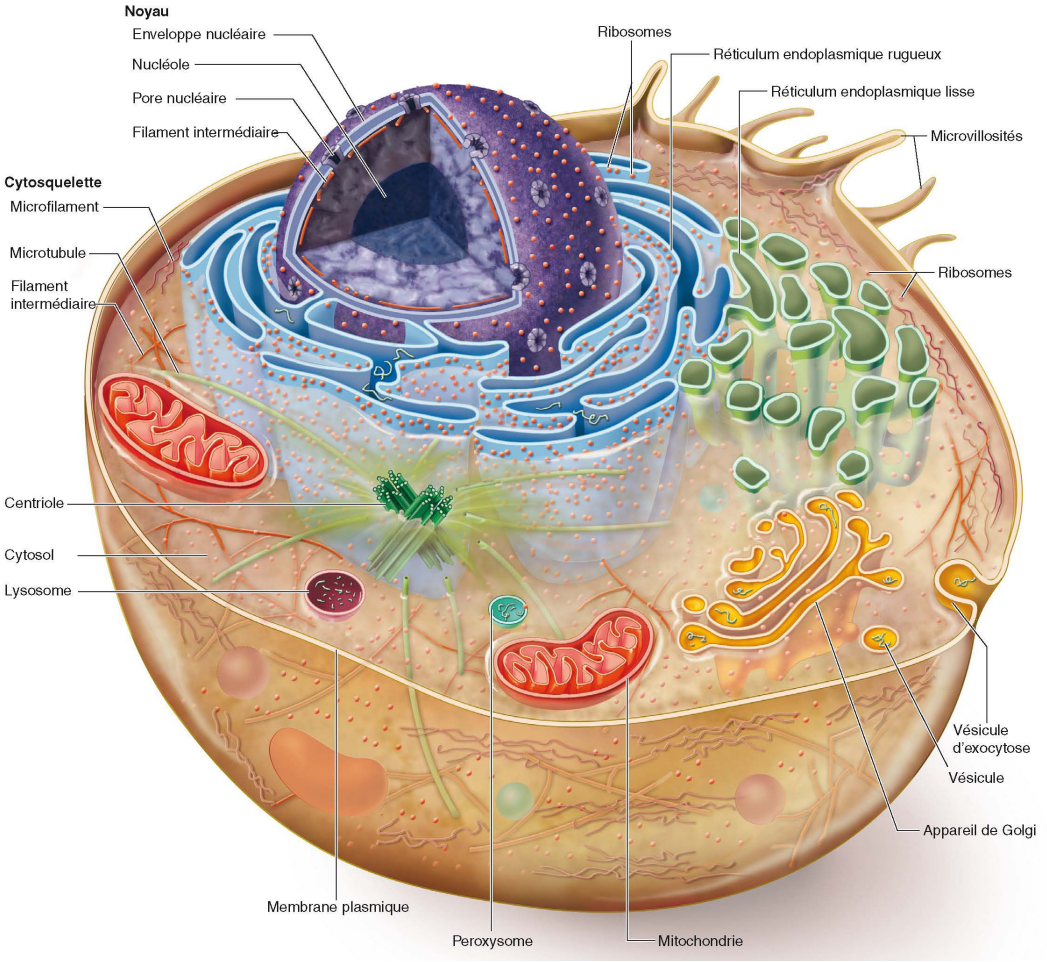
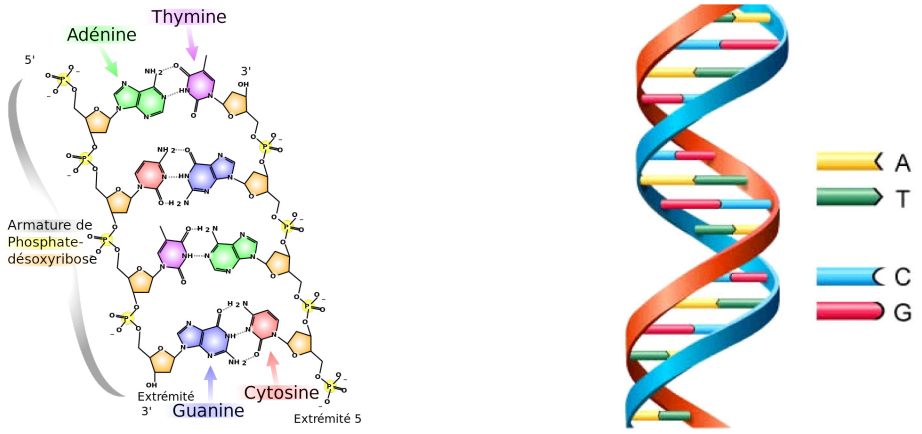


FIGURE 1.1 – *Cellule eucaryote en coupe horizontale [83].*

1.1.1 Les biomolécules au coeur de la machinerie cellulaire

L'information génétique (génotype) est un plan dont l'exécution conditionne l'apparence d'un être vivant, son fonctionnement, et son comportement dans son environnement (phénotype). Cette information génétique est stockée de manière pérenne et reproductible sur un support de nature moléculaire : l'**ADN** pour Acide Desoxyribo-Nucléique ou l'**ARN** pour Acide Ribo-Nucléique. L'ADN est porté par les chromosomes, situés dans le noyau de la cellule (voir Figure 1.1). L'exécution de ce plan s'effectue à l'échelle moléculaire et commence par la lecture de l'information génétique dans le noyau et se termine par la production de toutes les protéines dans le cytoplasme (voir Figure 1.1). Ce processus de transformation de l'information génétique en des composants fonctionnels est commun à tous les êtres vivants.

Le fonctionnement d'une cellule vivante implique aussi d'autres acteurs. Parmi ces molécules, les **polysaccharides** et les **lipides** ne sont pas générés par le code génétique, mais jouent un rôle prépondérant dans la structuration, notamment de la membrane cellulaire (voir Figure 1.1), d'autres stockent l'énergie nécessaire à la cellule et enfin certaines fonctionnent comme messagers inter- et intracellulaires.



(a) (b)

Figure 1.2 – *(a) Structure chimique de l'ADN mettant en jeu la liaison de nucléotides Adénine (en vert) avec Thymine (en rose) et Cytosine (en rouge) avec Guanine (en violet). Le squelette de l'ADN est composé par la liaison de Phosphates (en jaune) avec un groupement désoxyribose (en orange). (b) Représentation simplifiée de l'ADN en 3d, chaque type de nucléotides étant représenté par une couleur différente.*
Source : (a) Madprime - https://commons.wikimedia.org/wiki/File:DNA_chemical_structure-1-.fr.svg

1.1.1.1 L'ADN

L'Acide Désoxyribo-Nucléique (ADN) est une biomolécule pouvant être considérée comme le plan de construction de tous les êtres vivants. Le support moléculaire contenant l'information génétique est une longue séquence de nucléotides, de quatre types : l'Adénine, la Thymine, la Guanine et la Cytosine. Ces nucléotides partagent une structure moléculaire commune constituée d'un sucre, le désoxyribose, et d'un groupe phosphate (voir Figure 1.2a). À cette partie commune se lie une base azotée spécifique à chacun des 4 types de nucléotides. Les nucléotides s'organisent en séquence de deux brins en complémentarité en établissant des liaisons hydrogènes spécifiques. L'Adénine et la Guanine se liant respectivement à la Thymine et la Cytosine.

Les deux brins de l'ADN adoptent une structure hélicoïdale. Leur complémentarité permet tout d'abord d'assurer une certaine résistance de la structure à la dégradation et, en cas d'endommagement d'un des deux brins, la redondance de la complémentarité permet la réparation du brin intact. L'ADN est lui-même structuré de façon plus complexe, d'abord compacté par les histones (voir Figure 1.1), protéines structurantes possédant une forte affinité avec l'ADN qui s'enroule autour, puis finalement organisé en superstructures, les chromosomes. Chez les eucaryotes, organismes pluricellulaires possédant un noyau dans la cellule, l'ADN est stocké dans le noyau [83].

1.1.1.2 L'ARN

L'Acide RiboNucléique (ARN) est une biomolécule structurellement proche de l'ADN comportant néanmoins quelques différences. La première se retrouve au niveau de la séquence d'acides nucléiques qui, contrairement à l'ADN, très majoritairement composé d'un double brin sous forme d'hélice, s'organise en simple brin. Une seconde différence concerne les sucres constituant chacun des nucléotides puisque le désoxyribose de l'ADN est remplacé par un

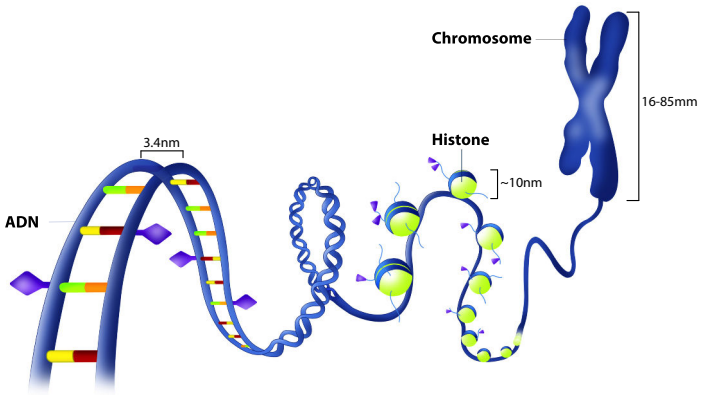


FIGURE 1.3 – Les différentes échelles de structuration d'un chromosome, forme condensée de l'ADN. Ce chromosome est lui-même composé d'ADN enroulé autour de protéines structurantes, les histones.
Source : NIGMS - http://images.nigms.nih.gov/index.cfm?event=viewDetail&imageID=2563

ribose pour l'ARN. La différence entre les deux groupements est illustrée dans la Figure 1.4. De plus, la Thymine présente dans l'ADN n'existe pas dans l'ARN et est remplacée par l'Uracile, complémentaire, comme la Thymine, à l'Adénine.

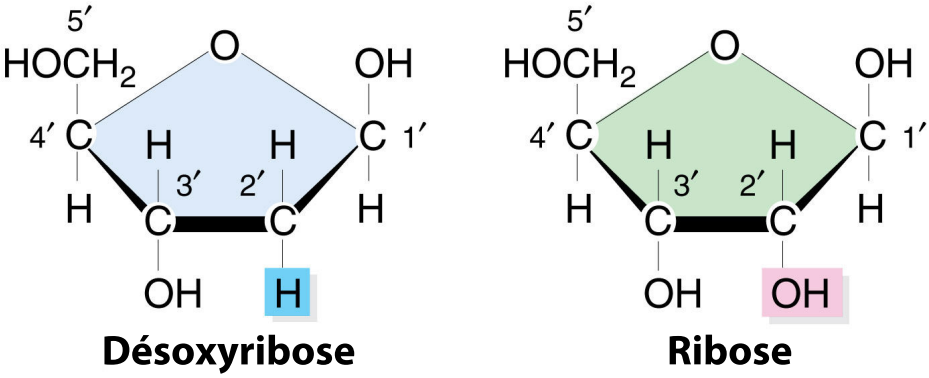


FIGURE 1.4 – *Composition chimique d'un désoxyribose, présent dans l'ADN, mis en parallèle avec un ribose, présent dans l'ARN. La seule différence provient du groupement hydrogène présent sur le carbone 2 du désoxyribose qui est remplacé par un groupement hydroxyle dans le ribose. Source: [140]*

1.1.1.3 Les protéines

Les protéines sont les biomolécules considérées comme les acteurs moléculaires fonctionnels de la cellule. Les protéines sont à la fois les briques, les ouvriers et les messagers participant au fonctionnement cellulaire. Les règles qui régissent la production de protéines à partir de la lecture de l'information génétique sont décrites dans le code génétique universel, commun à tous les organismes vivants. Certaines assurent un rôle **structurel**, en étant notamment impliquées dans la construction et la structure du squelette de la cellule, comme

l'actine et le collagène qui assurent le maintien physique et structurel de la cellule ainsi que la résistance de la matrice extracellulaire. Certaines sont impliquées dans la **mobilité** des cellules et des organismes, comme les myosines qui permettent la contraction musculaire, transformant l'énergie chimique en énergie mécanique. Certaines jouent un rôle dans le **conditionnement de l'ADN**, l'ADN étant enroulé autour de protéines appelées les histones (voir Figure 1.3), d'autres sont impliquées dans la **régulation de l'expression génétique**, comme les facteurs de transcription accompagnant l'ARN polymérase lors de la transcription. Certaines font office de **transporteuses** du matériel cellulaire d'un point à un autre, comme la kinésine évoluant sur des structures de microtubules (cf. Figure 1.5).

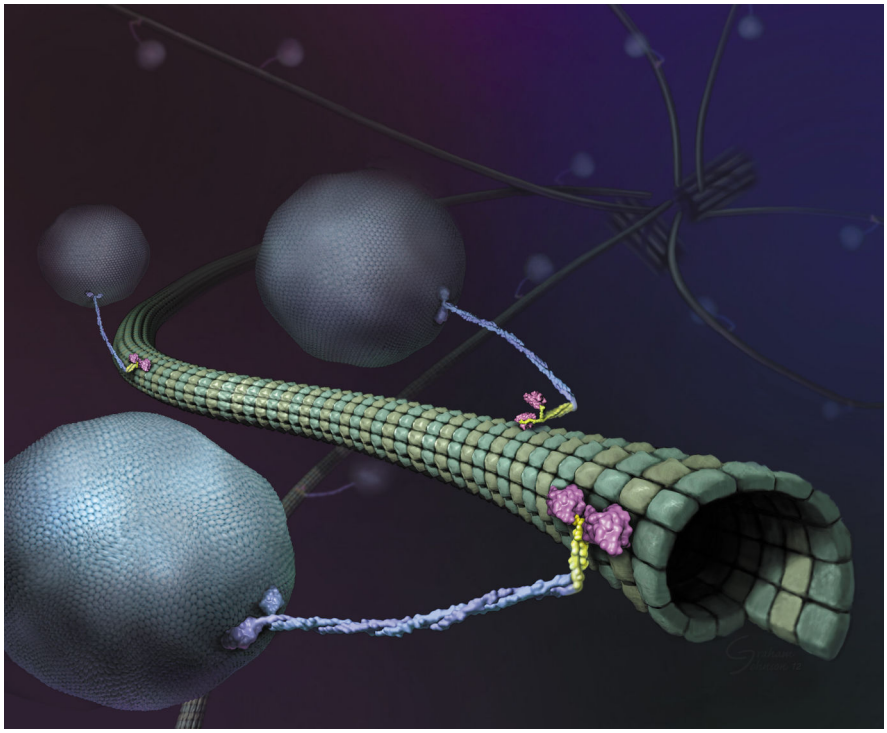
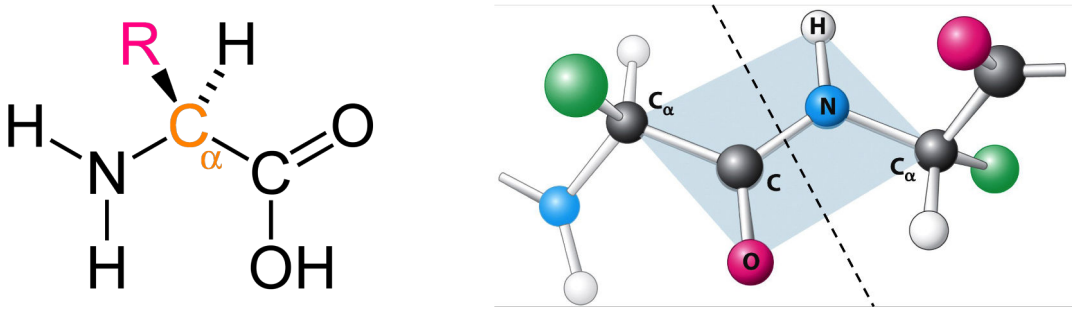


Figure 1.5 – *Vue d'artiste du transport de vésicules contenant du matériel cellulaire (boules blanches et bleues) le long de microtubules (tube de couleur verte) par une protéine appelée kinésine. La kinésine est composée de plusieurs domaines : en rose les deux domaines constituant les «jambes» de la protéine, en bleu un filament dont le domaine terminal se lie aux vésicules proches. À la frontière entre les domaines jaunes et roses, la transformation de l'énergie chimique en énergie mécanique permet à la protéine de «marcher» sur les microtubules.*
Source: Graham Johnson (graham@grahamj.com) / modifié de "Journal of Cell Biology, November 27, 2000"

Une protéine est constituée d'une succession d'acides aminés liés entre eux dont il existe 22 sortes différentes. Les acides aminés sont composés d'atomes de carbone, d'hydrogène, d'oxygène et d'azote, certains intégrant aussi un atome de soufre ou de sélénium. Ces acides aminés possèdent une partie commune, le squelette, et une partie spécifique appelée la *chaîne latérale*, qui caractérise le type d'acide aminé. C'est au niveau de la partie commune que les acides aminés sont liés par une *liaison peptidique*, la séquence des parties communes constituant la *chaîne principale* (ou squelette) de la protéine (cf. Figure 1.6a et 1.6b). La chaîne latérale spécifique à chaque type d'acide aminé donne lieu à des propriétés physico-

chimiques différentes. Chaque acide aminé peut être représenté par la formule générique H_2N-HCR-COOH, dans laquelle R désigne la chaîne latérale (cf. Figure 1.6a).



(a) (b)

FIGURE 1.6 – *(a) Structure chimique d'un acide aminé. En jaune est représenté le Carbone alpha où est liée la chaîne latérale dont la composition atomique est différente suivant le type d'acide aminé. (b) Illustration d'une liaison peptidique entre deux acides aminés (à gauche et à droite de la ligne pointillée). Cette liaison est une liaison covalente entre l'azote (N) du groupe amide de l'acide aminé de droite et le carbone (C) du groupe carboxyle de l'acide aminé de gauche. En vert sont représentées les chaînes latérales simplifiées.*
Source: [14]

Il est possible de classer les acides aminés selon plusieurs critères, depuis leur taille jusque leur propriété hydrophile (affinité avec l'eau) ou leur polarité. Il existe cependant un classement commun qui les regroupe en six groupes fonctionnels : Les acides aminés **aliphatiques** (Glycine, Alanine, Valine, Leucine et Isoleucine), les acides aminés avec groupement **hydroxyle, sulfurique ou sélénique** (Sérine, Thréonine, Méthionine, Cystéine et Sélénocystéine), les acides aminés **cycliques** (Proline), les acides aminés **aromatiques** (Phénylalanine, Tyrosine et Tryptophane), les acides aminés **basiques** (Histidine, Lysine et Arginine) et enfin les acides aminés **acides** et leurs **amides** (Aspartate, Glutamate, Asparagine et Glutamine) (voir Figure 1.7).

La séquence des acides aminés, pouvant être représentée par une suite de lettres choisies parmi un alphabet de 22 lettres correspondantes chacune à un type d'acide aminé, est appelée la **structure primaire** d'une protéine.

La protéine va adopter, contrainte par les interactions physiques et chimiques entre les différents atomes de la chaîne principale et latérale, des structurations locales particulières. Ces motifs structuraux formés sont au nombre de 3 : *hélices*, *feuillets* et *coudes* et leur enchaînement est appelé la **structure secondaire** de la protéine (voir Figure 1.8a).

Enfin, les protéines possèdent également des motifs plus importants, souvent le résultat de l'agencement dans l'espace des motifs de structures secondaires cités précédemment. C'est la **structure tertiaire** ou structure tridimensionnelle de la protéine (cf. Figure 1.8b). Cette structuration est due aux interactions proche et longue distance formées par les chaînes latérales des acides aminés. Parmi ces interactions, on retrouve les attractions/répulsions électrostatiques des acides aminés chargés électriquement, l'effet hydrophobe est le phénomène d'enfouissement et de regroupement des régions dont le ratio d'acides aminés hydrophobes est important. Ces régions vont se retrouver à l'intérieur de la protéine alors qu'à l'inverse, les régions dites hydrophiles vont majoritairement se situer en surface de la protéine.

Cette structuration, bien que primordiale, n'apparait pas de façon si précise chez toutes les protéines. En effet, certaines d'entre-elles sont désordonnées et ne vont se structurer qu'au

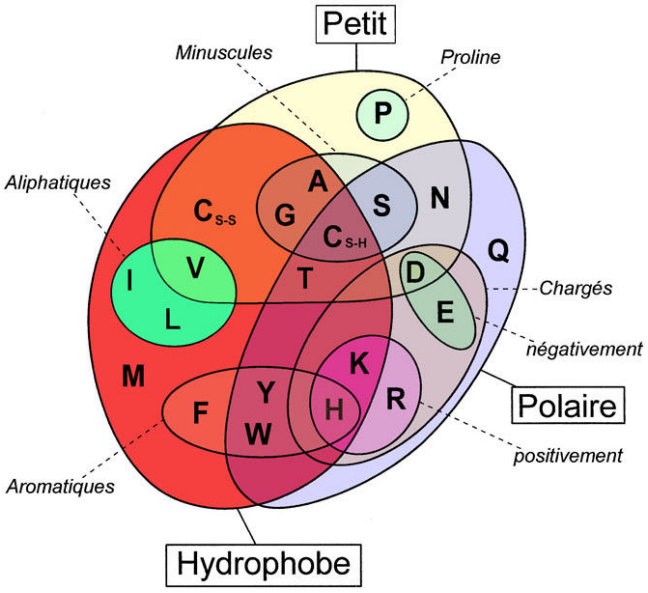


Figure 1.7 – *Diagramme de Venn illustrant les différentes classifications servant à définir les acides aminés. Ces propriétés peuvent être structurelles (minuscules, petit) ou physico-chimiques (polaire, hydrophobe, etc.).*
Source : Pixeltoo - https://commons.wikimedia.org/wiki/File:Acides_amin%C3%A9s_propri%C3%A9t%C3%A9s_diagramme_Venn.svg

contact avec d'autres molécules ou parfois ne jamais adopter de structures précises.

1.1.1.4 Les polysaccharides

Les polysaccharides, appelés aussi glucides complexes, sont des séquences répétées de sucres (ou oses). On peut distinguer deux groupes de polysaccharides qui se distinguent par leur fonction biologique. Les polysaccharides **de réserve** constituent une source d'énergie pour les êtres vivants. Ils se retrouvent principalement sous forme de *glucose*. Les polysaccharides **structuraux** participent eux à la formation des structures organiques comme la *cellulose* dans les tissus de soutien chez les végétaux ou la *chitine* chez les animaux (cf. Figure 1.9).

À la différence des biomolécules citées précédemment, les polysaccharides, bien qu'indispensables aux cellules, ne sont pas synthétisées par ces dernières, mais doivent être assimilés à travers la nourriture par exemple. De nombreuses informations à propos des polysaccharides, en particulier à propos de leur représentation et leurs structures, sont rassemblées au sein du site Glycopedia [1].

1.1.1.5 Les lipides

Les lipides constituent la matière grasse de l'organisme et englobent de nombreuses molécules différentes. Leur point commun réside dans la présence d'une partie hydrophobe même si celle-ci peut être liée à une partie hydrophile.

De la même manière que les polysaccharides, ils constituent une **source d'énergie** importante pour la cellule. Ils ont l'avantage de pouvoir être stockés, au contraire des glucides.

1. http://www.glycopedia.eu/

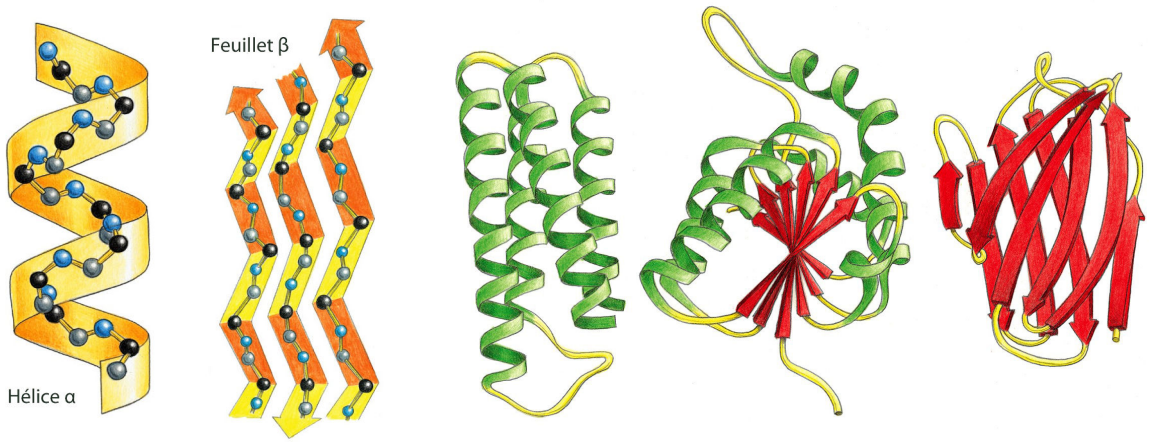


FIGURE 1.8 – *(a) Représentation en rubans des motifs de structure secondaire les plus communs, à gauche une hélice α et à droite un feuillet β, résultats de la configuration spatiale des atomes de la chaîne principale des protéines. (b) Exemples de motifs communs de structures tertiaires. À gauche, représentation d'un paquet de 4 hélices α agencées autour du motif hélice-coude-hélice. Au centre, enchevêtrement d'hélices α et feuillets β. À droite, double couche de feuilles β. Les hélices sont en vert, les feuillets en rouge et les coudes ou boucles en jaune. Source : [2]*

Ils sont également apportés par l'alimentation. Ils jouent également un rôle primordial dans la constitution des **membranes cellulaires** dont ils sont le principal composant (cf. Figure 1.9). Certains lipides ont un rôle de messager intercellulaire et intracellulaire. Parmi les principaux rôles qu'ils possèdent, ils sont aussi utilisés comme substrat dans des réactions métaboliques complexes. À la manière des polysaccharides, plusieurs sites rassemblent de façon centralisée des ressources se rapportant aux lipides, l'un des plus connus étant Lipidmaps [2].

1.1.2 De l'information génétique aux unités fonctionnelles

L'information génétique portée par l'ADN doit être transformée en unités fonctionnelles, les protéines. Cette transformation se déroule en plusieurs étapes, chacune assurant la préservation de l'information et sa bonne interprétation. Ces étapes de transformation sont le résultat d'un jeu de régulations positives et négatives des équilibres de concentrations moléculaires de ces acteurs. Ce sont des étapes se déroulant toutes simultanément et de façon parallèle.

1.1.2.1 La transcription, de l'ADN à l'ARN

Lors de la transcription, un ensemble d'enzymes composé de l'ARN polymérase et de protéines initiatrices et régulatrices, appelées facteurs de transcription, viennent reconnaître une séquence particulière de l'ADN appelée site promoteur. La liaison sur ce site engendre un clivage (une séparation) des deux brins complémentaires constituants l'ADN. L'ARN polymérase peut alors commencer à parcourir l'un des deux brins et initie la génération d'un pré-ARN messager (pré-ARNm). Ce pré-ARNm va être constitué d'une séquence de nucléotides complémentaires à la séquence du brin d'ADN que l'ARN polymérase parcourt. La

2. http://www.lipidmaps.org/

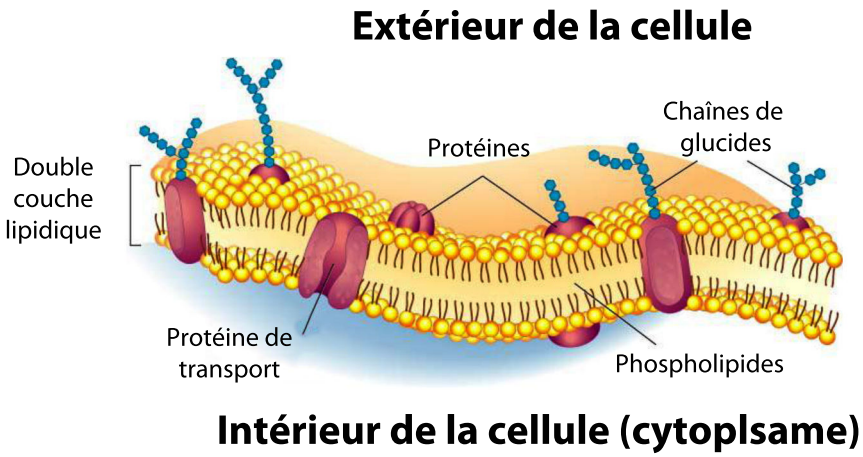


Figure 1.9 – *Schéma d'une membrane cellulaire composée d'une double couche lipidique sur laquelle sont liés différents types de protéines et des polysaccharides (ou glucides).*
Source : http://www.acpfg.com.au/blog/?p=143

terminaison de l'élongation de l'ARNm intervient lorsqu'une séquence spécifique de nucléotides est atteinte (AAUAAA). L'ARN polymérase se détache et libère le pré-ARNm ainsi formé qui va subir une étape de maturation. Cette étape de maturation est une étape importante de régulation de l'expression des gènes puisqu'elle peut influer sur la stabilité de l'ARN, sa capacité à être traduit ou bien avoir un impact sur la séquence traduite. L'une des modifications post-transcriptionnelles courantes implique des enzymes venant couper les séquences non codantes et non nécessaires à la formation de protéines dans l'ARN. Cette étape d'épissage va couper l'ARN a différents endroits et potentiellement générer plusieurs chaînes d'ARN. Autre modification importante, l'ajout d'une coiffe à une extrémité de l'ARN et une polyadénylation (ajout d'environ 200 résidus adénosine) à l'autre extrémité. Ces deux dernières modifications jouent un rôle important pour la reconnaissance de l'ARN par le ribosome pendant la traduction, mais également pour sa stabilité. Ces modifications constituent d'ailleurs la dernière étape de la transcription et précèdent la sortie de l'ARNm mature du noyau pour rejoindre le cytoplasme de la cellule, lieu de la traduction.

1.1.2.2 La traduction, de l'ARN à la protéine

Lorsque l'ARNm a rejoint le cytoplasme, l'étape de traduction peut commencer. La première phase de la traduction est la reconnaissance par la sous-partie du ribosome (30S chez les bactéries, 40S chez les eucaryotes) de la région en amont de l'ARNm. Cette reconnaissance entraîne le parcours de l'ARNm par le ribosome jusqu'au codon d'initiation qui est le premier triplet de nucléotide qui sera traduit. Chaque triplet de nucléotide code pour un acide aminé particulier. Puisqu'il existe 4 nucléotides différents, 64 combinaisons de triplets sont possibles et en moyenne, un acide aminé est codé par 3 combinaisons de triplets de nucléotides différents, c'est le caractère redondant du code génétique, comme illustré dans la Figure 1.10. Le premier triplet de la chaîne d'ARNm et la petite sous-unité du ribosome vont alors recruter le premier ARN de transfert (ARNt, chargé de se lier aux acides aminés) ayant formé un complexe avec une méthionine et la grande sous-partie du ribosome (50S chez les bactéries, 60S chez les eucaryotes). À la suite de cette étape de reconnaissance, la construction du reste de la protéine est initiée et se caractérise par la lecture séquentielle des

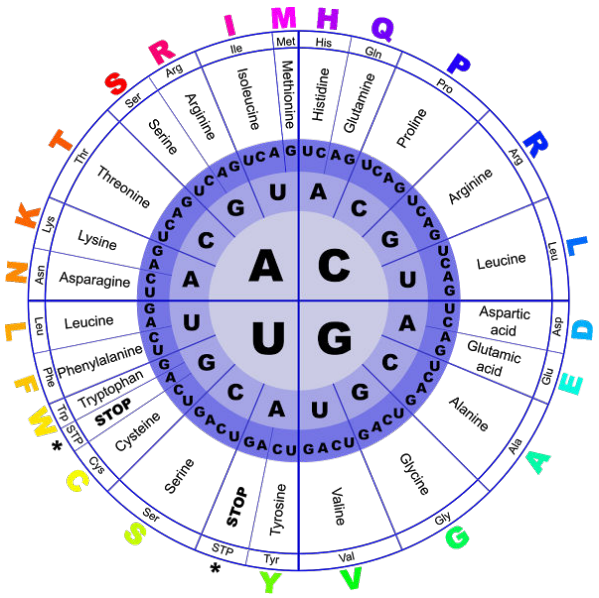


FIGURE 1.10 – *Tableau de correspondance des triplets d'acides nucléiques et les acides aminés pour lesquels ils codent. «Initiation» désigne les codons pouvant être retrouvés comme initiateur de la traduction, «Stop» désigne les codons responsables de l'arrêt du processus de traduction par le ribosome.*

triplets de nucléotides de l'ARNm par le ribosome. Le ribosome fait l'interface entre l'ARNm et les complexes ARNt complémentaires/acide aminé. Les acides aminés sont liés entre eux de façon séquentielle jusqu'à ce qu'un triplet codant pour un codon-STOP (ou d'arrêt) soit atteint par le ribosome. À ce moment-là, le ribosome libère la protéine et l'ARNm et l'étape de traduction se termine. À ce moment-là, la protéine n'est pas encore fonctionnelle et doit encore subir des modifications post-traductionnelles afin de pouvoir remplir sa fonction. Un schéma simplifié de la transcription et de la traduction est illustré dans la Figure 1.11

1.1.2.3 Maturation et acquisition de la fonction protéique

Afin de permettre leur stabilité au sein de la cellule, de les guider vers leur lieu d'action ou simplement d'assurer leur efficacité fonctionnelle, les protéines subissent plusieurs modifications post-traductionnelles dont la nature et le nombre dépendent de la nature des protéines. Ces modifications peuvent intervenir aux deux extrémités des protéines ou bien au niveau des acides aminés individuellement. Parmi les modifications post-traductionnelles, on retrouve l'ajout de groupes fonctionnels à certains acides aminés modifiant leurs propriétés, des étapes de clivage peuvent également intervenir afin d'éliminer des parties de la protéine qui étaient importantes lors des étapes précédentes, mais inutiles pour sa fonction finale. De la même façon, certaines modifications, effectuées par des molécules dites *chaperonnes* ont pour rôle de donner à la protéine sa structure tertiaire finale et donc fonctionnelle. C'est le processus spécifique du ***folding*** (ou **repliement**) moléculaire de la protéine. Enfin, certaines modifications permettent le bon signalement cellulaire de la protéine, à savoir le codage lui permettant de rejoindre son lieu d'action.

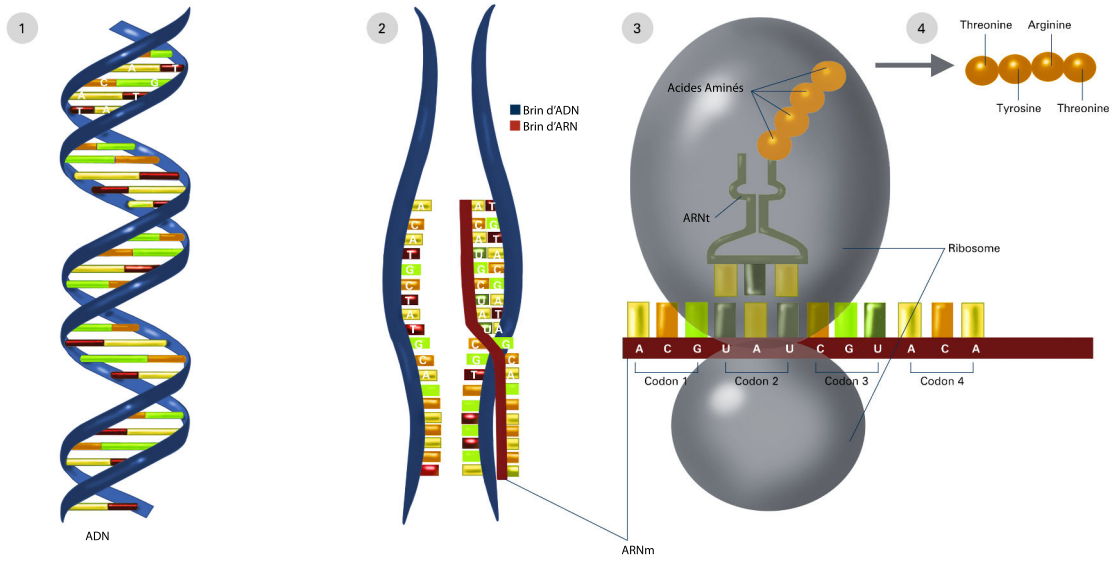


Figure 1.11 – *Schéma simplifié des étapes de transcription et de traduction.*
Source : NIGMS - http://images.nigms.nih.gov/index.cfm?event=viewDetail&imageID=2549

1.2 Méthodes et outils de la biologie structurale

La majorité des protéines ne peut pas être fonctionnelle sans structure 3d, car leur fonction découle de cette structure, en particulier de ses propriétés biomécaniques et biophysiques conditionnant des interactions avec des molécules partenaires. C'est la raison pour laquelle la biologie structurale accorde une importance primordiale à la structure, élément central dans la caractérisation de la fonction des protéines.

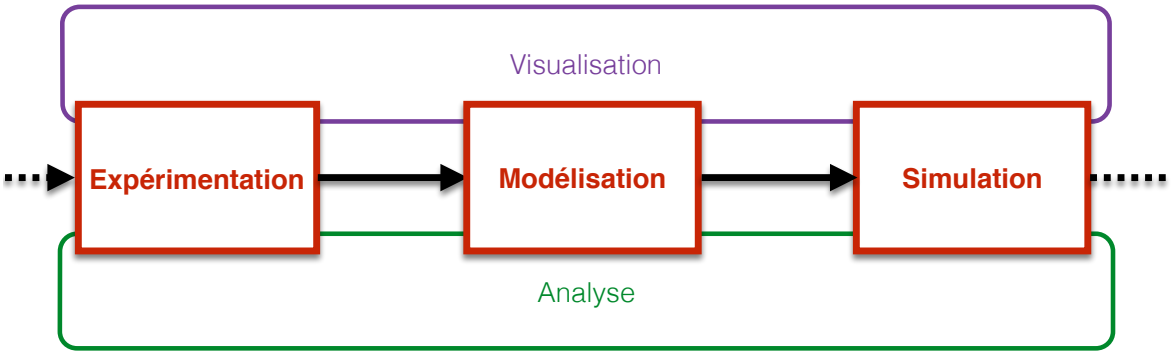


Figure 1.12 – *Les différentes étapes de l'étude d'un système moléculaire en biologie structurale.*

Il est possible de diviser le processus d'étude de la biologie structurale en plusieurs étapes distinctes, néanmoins étroitement liées. La Figure 1.12 donne un aperçu de ce découpage et identifie 3 grandes tâches : la collecte d'informations structurelles via des techniques expérimentales et théoriques, la modélisation de structures à partir des informations récoltées et

la simulation moléculaire qui permet d'ajouter une dimension dynamique indispensable pour l'étude de phénomènes et enfin l'analyse et l'interprétation de résultats permettent de postuler des hypothèses quant aux phénomènes étudiés et d'affiner les étapes de modélisation. Ces étapes ne mettent pas en jeu les mêmes outils et méthodes. Nous nous concentrerons dans cette section sur la collecte de données aboutissant à l'obtention de structures 3d.

Les protéines sont des unités biologiques de taille nanoscopique, leur taille oscillant entre quelques dizaines et quelques centaines d'Angströms (Å = 10^{-10} m) et donc invisibles pour des systèmes de photographies standards. En effet, les techniques de microscopie optique ne peuvent permettre aujourd'hui une observation à l'échelle atomique, quel que soit l'objet d'étude. Notons que cette taille en Angströms est indicative et inappropriée pour désigner la grosseur des protéines. On préfère utiliser le kilodalton (kDa) qui est une mesure de masse, 1 Da correspondant à la masse d'un atome d'hydrogène. Un acide aminé représente environ 110 Da et une protéine entre 15 et plusieurs millions de kilodaltons pour les complexes multimériques les plus importants.

L'ensemble des approches expérimentales et théoriques permettant d'étudier la structure des protéines peut être divisé en 3 grandes approches : les **techniques expérimentales**, regroupant les techniques utilisant des extraits naturels des protéines étudiées et cherchant à obtenir leur structuration par des outils de mesure *in vitro*. Les approches de **modélisation/simulation** mettent en place des modèles informatiques de structures 3d à partir d'informations physico-chimiques et/ou statistiques utilisés comme paramètres de logiciels de calcul. Enfin, les programmes de **visualisation moléculaire** vont permettre d'observer et d'analyser les structures obtenues à partir des deux approches précédentes grâce à des codes graphiques précis dans le but d'extraire, d'illustrer ou de communiquer de nouvelles connaissances scientifiques.

1.2.1 Expérimentations

La nature des protéines et l'environnement dans lequel elles sont naturellement structurées impliquent l'utilisation de techniques expérimentales devant à la fois préserver leur nature et ne pas les dégrader ou les détruire, mais également assurer une précision minimum pour extraire des informations structurales et géométriques qui seront utilisées pour la caractérisation de leur fonction. Les techniques expérimentales utilisées en biologie structurale pour obtenir la structure tridimensionnelle de biomolécules sont coûteuses en raison de leur complexité. Ce sont des méthodes physiques indirectes qui sont principalement utilisées afin d'obtenir des informations sur la structure 3d d'une biomolécule à l'échelle atomique. La plupart de ces techniques mettent en jeu des instruments de mesure de haute technologie et demandent souvent une préparation complexe et longue de l'échantillon pour qu'il soit suffisamment concentré en protéines. Ils nécessitent également des conditions de stockage adaptées et dépendantes de la technique utilisée.

1.2.1.1 Cristallographie à rayons X ou radiocristallographie

Parmi ces techniques expérimentales, la plus ancienne et la plus utilisée, principalement pour sa précision, est la cristallographie aux rayons X ou radiocristallographie. Cette technique consiste à envoyer un faisceau de rayons X sur un cristal composé exclusivement de la biomolécule étudiée. La mesure des angles et de l'intensité des rayons diffractés permet d'obtenir une image tridimensionnelle de la densité électronique dans le cristal. À partir de cette densité, il est possible d'obtenir la position moyenne des atomes présents dans le cristal

ainsi que les liens existants entre eux en superposant la séquence d'atomes, connue, dans la carte de densité électronique ainsi obtenue. Le processus est schématisé dans la Figure 1.13.

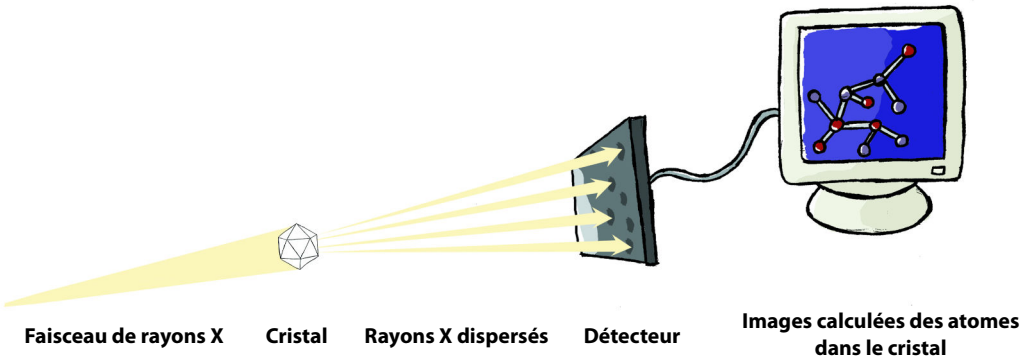


Figure 1.13 – *Schéma simplifié de la cristallographie à rayons X où les rayons X traversent un échantillon cristallisé avant d'être diffractés et capturés par un détecteur avant d'être interprétés informatiquement.*
Source : NIGMS - http://images.nigms.nih.gov/index.cfm?event=viewDetail&imageID=2512

L'avantage de la radiocristallographie est sa très grande précision et l'absence de limite de taille pour le cristal, permettant l'observation de structures moléculaires de très grande taille comme les ribosomes (environ un million d'atomes). Suivant les biomolécules observées, il est possible d'obtenir des structures tridimensionnelles à des résolutions minimales en dessous de 1Å pour une précision sur la position des atomes à moins de 0.5Å. 1Å constituant une précision atomique. Les cartes de densité et les modèles atomiques créés à partir de ces cartes sont illustrés dans la Figure 1.14.

Certaines biomolécules sont néanmoins très difficiles à cristalliser et on évalue à environ un quart la proportion de macromolécules permettant de créer un cristal de taille et de qualité suffisante de diffracter suffisamment les rayons X. Cette cristallisation est une étape difficilement automatisable et qui se révèle majoritairement empirique, demandant un travail de préparation spécifique en plus de l'application de la technique. Les hydrogènes présents dans les macromolécules sont très difficiles à percevoir du fait de leur très faible densité électronique. L'obtention d'une structure 3d par radiocristallographie est fastidieuse et nécessite en moyenne plusieurs mois de travail. Certains laboratoires sont cependant maintenant capables d'automatiser certaines étapes de préparation et d'execution de la cristallographie grâce afin d'analyser des échantillons de tissus ou de liquides possédant plusieurs centaines ou milliers de protéines différentes. Cette technique de cristallographie haut-débit passe par la mise en place de robot afin d'effectuer un nombre important de cristallisation dans un laps de temps réduit afin de générer un nombre important d'échantillons [103].

Il est possible de s'intéresser aux noyaux des atomes au lieu du nuage électronique et d'utiliser des techniques de cristallographie neutroniques afin de mettre en évidence la position d'élements légers comme l'hydrogène et de différencier des éléments chimiques possédant des numéros atomiques (nombre de neutrons dans le noyau) proches. La cristallographie nécessite d'observer les protéines dans un environnement non naturel, le cristal et se déroule à basse température, ne reflètant qu'un état unique de la protéine à cette température. Il est donc difficile d'extraire une dynamique structurelle des données cristallographiques. Enfin, même s'il est possible d'observer des complexes moléculaires de grande taille, la résolution obtenue est souvent moindre que celle des protéines composées d'entre 100 et 1000 acides aminés.

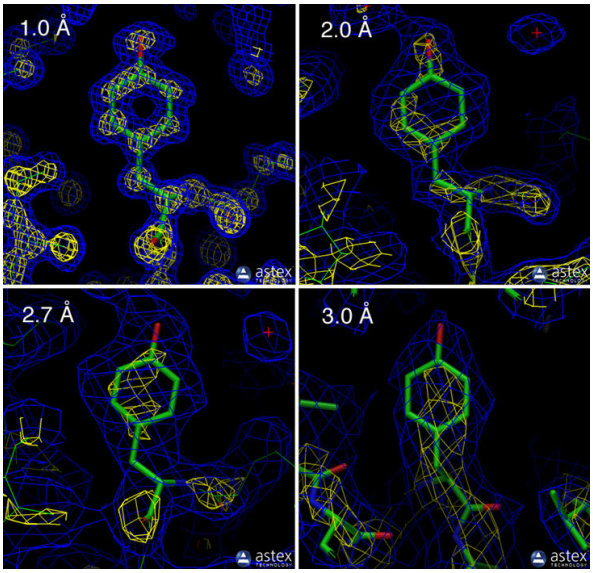


FIGURE 1.14 – *Rendu graphique d'un acide aminé représenté par des bâtonnets verts et rouges au sein de sa carte de densité obtenue par cristallographie et représentée par un mesh de triangles bleus. Plusieurs précisions de cartes de densité électronique sont présentées. Source : RCSB - http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/Looking-at-Structures/resolution.html*

1.2.1.2 Spectroscopie à Résonance Magnétique Nucléaire - RMN

Également très utilisée, la Spectroscopie à Résonance Magnétique Nucléaire ou RMN consiste à envoyer une séquence d'impulsions électromagnétiques sur une molécule en présence d'un champ magnétique [182]. La fréquence et la séquence des impulsions électromagnétiques sont propres à chaque type d'atome. Cette technique se base sur les mouvements de rotation naturels des atomes, créant un mini champ magnétique (ou spin). Seuls quelques atomes sont détectables en spectroscopie RMN, car ils possèdent un spin nucléaire particulier de 1/2. En biologie structurale, c'est principalement le proton H^1 qui est ciblé. Les noyaux atomiques possédant des spins sont excités par les impulsions électromagnétiques et absorbent l'énergie ainsi reçue. Lors de l'étape de relaxation suivant l'impulsion, les noyaux atomiques relâchent de l'énergie sous forme de résonance à différentes longueurs d'onde, calculée et reportée par les instruments de mesure (cf. Figure 1.15). Cette résonance varie en fonction de la nature de l'atome excité et de son environnement. Il est donc possible, grâce à ces résonances, pour un atome donné, d'avoir des informations sur la nature et le nombre d'atomes voisins, la liaison chimique dans laquelle il est impliqué, sa distance à d'autres atomes, sa mobilité, etc. À la différence de la cristallographie où la protéine est cristallisée, la RMN peut s'appliquer sur une molécule solubilisée (mise en solution liquide). Il est de fait plus aisé de préserver la structure et la fonctionnalité d'une molécule en solution que sous forme de cristal.

Il est également possible d'avoir des informations sur la dynamique de la molécule puisqu'au contraire de quand elle se trouve dans un cristal, une molécule en solution n'est pas statique. L'état dynamique de la molécule est aussi un inconvénient pour obtenir une structure 3d fixe puisque la précision de la mesure sera perturbée par les changements structurels de la molécule. L'un des autres inconvénients de la spectroscopie RMN est la limitation en taille des molécules, observée du fait de la complexité du traitement des signaux de résonance magnétique, ces derniers se superposant tous sur un spectre borné dont les limites ne varient

pas avec la complexité de la protéine étudiée. Ainsi, plus le nombre d'atomes est important, plus le nombre de signaux augmentera et s'accumulera sur un spectre de largeur constante. La taille optimale pour la RMN varie entre 10 et 50 kilodaltons (kDa) correspondant à des molécules de 1.500 à 10.000 atomes. De plus, pour les plus grosses molécules, il est nécessaire d'effectuer un marquage isotopique permettant d'obtenir un spectre non plus 2d, mais 3d où les atomes ciblés, en plus de H^1, sont le C^{13} et le N^{15} après enrichissement spécifique de la biomolécule observée.

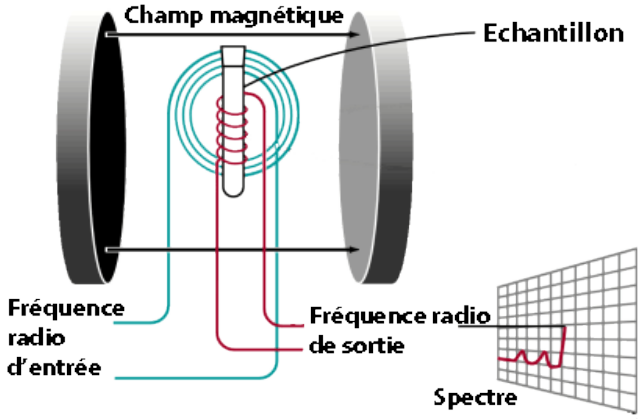
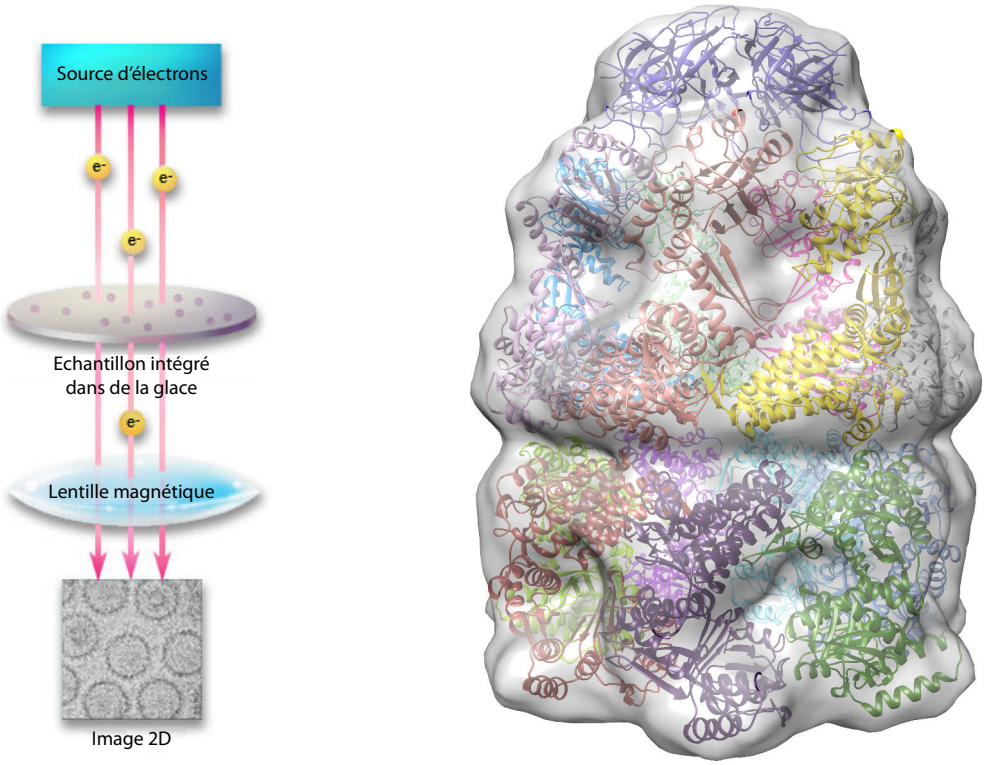


Figure 1.15 – *Schéma de la technique de spectroscopie RMN. Un échantillon est placé dans un champ magnétique et excité par des impulsions radio successives. La relaxation des atomes de l'échantillon induit un signal radio interprété par puissance de signal sur un spectre.*
Source : http://www.mhhe.com/physsci/chemistry/carey5e/Ch13/ch13-nmr-1b.html

1.2.1.3 Cryo-microscopie électronique - Cryo-EM

Une autre technique plus récente en biologie structurale, la Cryo-microscopie électronique (Cryo-EM), consiste à utiliser le principe de la microscopie électronique, et donc d'utiliser un faisceau de particules d'électrons au lieu de rayonnements électromagnétiques (lumière) comme en microscopie optique, sur un échantillon préalablement cryogénisé lors de l'étape de fixation. Le faisceau de particules d'électrons passe au travers de lentilles électrostatiques et électromagnétiques puis au travers de l'échantillon cryogénisé où il est modifié pour former une image électronique finalement amplifiée par d'autres lentilles et projetée sur un scintillateur (cf. Figure 1.16a). La protéine est ainsi visualisée sous plusieurs angles et les images générées sont ensuite traitées sur ordinateur afin de rassembler les différents points de vue et de créer une carte de densité électronique 3d rapportant le volume global de la protéine (voir Figure 1.16b).

La particularité de la cryofixation de l'échantillon, par opposition à la fixation chimique ou la déshydratation, est que l'échantillon est amené très rapidement à la température de l'azote liquide (−195.79 °C) ou de l'hélium liquide (−269 °C) afin que la glace formée ne soit pas cristalline et que le spécimen étudié conserve son état naturel. Pour des biomolécules, cela assure une conservation d'une structure 3d stable de la molécule. La cryo-EM permet d'étudier des structures moléculaires de grandes tailles, à partir de 300kDa jusqu'à des tissus de plusieurs centaines de nanomètres, en passant par les ribosomes, les virus ou les composants cellulaires.



(a) (b)

FIGURE 1.16 – *(a) Schéma illustrant la technique de cryo-microscopie électronique. (b) Carte de densité (en rendu surfacique transparent blanc) d'un complexe protéique (GroEL-GroES) composé de multiples sous-unités de la famille des protéines chaperonnes.*
Source : (a) `http://www.eicn.ucla.edu/cryoem`*, (b) [78]*

Bien qu'en constant développement, la cryo-EM possède une résolution relativement faible comparée aux deux techniques précédentes puisque les cartes de densité obtenues par cette technique ne permettent pas une résolution de moins de 4Å pour les cartes les plus précises[184]. L'exposition de l'échantillon à un faisceau de particules d'électrons, combinée à sa cryogénisation, dommage significativement l'échantillon qui ne peut être réutilisé ensuite.

1.2.1.4 Diffusion des rayons X - SAXS

La technique SAXS est basée sur les interactions élastiques entre les photons et les nuages électroniques [68]. Cette technique s'inspire directement de la diffraction des rayons X quand ils traversent un cristal (diffraction de Bragg) entraînant la diffusion de ces rayons à différents angles de l'ordre de la dizaine de degrés. L'angle de diffraction est inversement proportionnel à la distance interatomique des atomes présents dans le cristal. Or, les informations nécessaires pour constituer la structure 3d de macromolécules concernent des distances trop grandes pour que les angles de diffraction soient facilement détectables. La gamme d'angles pouvant être mesurés permet de rapporter des distances d'au minimum quelques nanomètres (environ 10Å de résolution). Il est nécessaire d'utiliser un rayonnement X monochromatique très proche de l'échantillon constitué de la biomolécule d'intérêt. De la même manière que pour les précédentes techniques, une carte de densité, ici densité électronique, est ainsi générée après interprétation des angles de diffusion calculés par l'instrument de mesure placé à la

suite du faisceau émis à travers l'échantillon. (voir Figure 1.17).

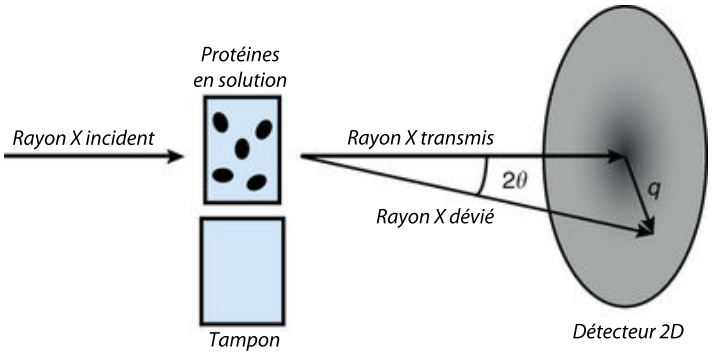


Figure 1.17 – *Schéma simplifié illustrant la technique de diffusion des rayons X. Source : [158]*

L'un des avantages de la technique SAXS par rapport à la cristallographie est l'absence de cristal pour effectuer l'expérience, l'échantillon étant mis en solution. Mais cet avantage est terni par la résolution de la technique qui ne permet pas d'obtenir une description atomique d'une biomolécule. Elle est souvent utilisée pour obtenir une structure 3d approximative de complexes moléculaires ou de composants cellulaires de grandes tailles grâce à une étape d'assemblage où la représentation atomique des composants individuels de ces complexes est intégrée dans une carte de densité. Une de ces forces repose aussi sur la rapidité de l'expérience qui, en combinant l'ensemble des étapes de préparation, expérimentation et analyses des résultats, peut générer une carte de densité en quelques jours.

L'ensemble de ces techniques possède une étape d'interprétation de signal basée sur des méthodes informatiques, autrement impossible à interpréter par un être humain. La complexité de l'interprétation d'un signal provient à la fois du bruit propre au signal, majoritairement dû à la complexité des échantillons observés aujourd'hui, et du nombre de signaux, corrélé directement à la taille des protéines étudiées. L'outil informatique joue donc un rôle très important dans les processus dits expérimentaux, mais ne se cantonne pas à de simples interprétations de résultats et permet également de générer des structures 3d grâce à des méthodes qui seront exposées ci-après (voir section 1.2.2).

1.2.1.5 Banques de structures protéiques

Les bases de données biologiques permettent de regrouper l'ensemble des informations scientifiques obtenues au cours de l'histoire et les mettre à disposition de la communauté scientifique. Les bases de données structurales regroupent plus particulièrement des structures obtenues par techniques expérimentales dont il est possible d'obtenir la représentation 3d sous forme de fichier informatique. Elles voient de plus en plus leur intégration au sein de portails biologiques gérant la conformité des données aux standards du domaine, leur accès à l'ensemble des scientifiques et éventuellement leur association avec des données d'autres domaines scientifiques proches.

L'une des bases de données les plus utilisées en biologie structurale est la **Protein Data Bank** (PDB) [15] qui regroupe l'ensemble des structures 3d de protéines publiées et vérifiées dans plusieurs formats standards, formats utilisés en entrée de la majorité des outils de bio-informatique structurale. Les structures 3d présentes dans la PDB proviennent majoritairement de cristallographie rayon X ou de spectroscopie RMN (89% pour la cristallographie, 10% pour la RMN). Aujourd'hui, environ 105 000 structures de protéines, 5 200 structures

de complexes mixtes protéine/acide nucléique et 2 800 structures d'acides nucléiques ont été déposées dans la PDB.

On peut également noter d'autres banques de données de structures protéiques :

— **SCOP** (Structural Classification of Proteins)[3] est une banque de données regroupant les protéines de la PDB présentant une relation de similarité structurale et d'évolution [121]. Les structures sont ainsi regroupées en 3 niveaux hiérarchiques : famille, superfamille et repliement.

— **CATH** (Class Architexture Topology and Homology)[4] regroupe les protéines dont la structure a été déterminée par RMN ou par cristallographie avec une résolution de détermination supérieure à 3Å [156]. Elle est composée de 4 niveaux hiérarchiques basés sur la structure des protéines, du plus générique au plus précis : classe, architecture, topologie et superfamilles homologues.

— **FSSP** (Fold Classification based on Structure-Structure alignement of Proteins)[5] regroupe les structures représentatives de la PDB [79]. Elle filtre les protéines dont les structures sont considérées comme redondantes dans la PDB, avec plus de 25% d'identité au niveau des séquences et de la structure après une comparaison des séquences et des motifs structuraux les composant. Elle se base sur le programme DALI d'alignement structural, consistant à comparer l'enchaînement des motifs structuraux entre les protéines, pour obtenir ces structures non redondantes [80].

— **MMDB** (Molecular Modeling Database)[6] est un sous-ensemble de la PDB excluant les modèles théoriques [108]. Elle héberge des données structurelles conventionnelles, mais non figées et pouvant être enrichies en ajoutant d'autres informations structurelles complémentaires obtenues par des technologies comme la microscopie électronique.

Parmi les informations stockées dans des bases de données et utilisées en biologie structurale en dehors de données purement structurales on retrouve : les profils biologiques des protéines avec leur structure primaire/secondaire/tertiaire, leur environnement cellulaire, leur rôle, etc. (SWISSPROT+TrEMBL [18], PDB, etc.); les réseaux d'interactions moléculaires (interactome) mettant en avant les partenaires moléculaires déjà identifiés (STRING [160], CCSB Interactome Database[7], etc.); les évolutions génomiques des séquences d'ADN codantes identifiant les régions évoluant rapidement au sein des protéines (séquence primaire changeante) et les régions plus stables et donc potentiellement importantes pour la fonction métabolique de la protéine (USCS [90], Ensembl [81], etc.).

1.2.2 Modélisation et simulation

Comme il n'est pas possible expérimentalement de filmer une protéine dans son environnement, en complément des techniques expérimentales, il est nécessaire d'utiliser des approches computationnelles, notamment pour accéder aux propriétés dynamiques et mécaniques des protéines. Il s'agit aussi de compléter les informations structurelles recherchées ou de générer des modèles pour des protéines dont la nature empêche leur étude expérimentale (mauvaise solubilité, instabilité en solution, dégradation par les rayons X, etc.). Ces approches dites *in silico* sont moins coûteuses que les approches expérimentales, mais ont parfois un facteur de confiance légèrement moins important que les approches énoncées précédemment, car elles

3. http://scop.berkeley.edu/
4. http://www.cathdb.info/
5. http://protein.hbu.cn/fssp/ekhidna.biocenter.helsinki.fi/dali/start.html
6. http://www.ncbi.nlm.nih.gov/structure/
7. http://interactome.dfci.harvard.edu/

dépendent uniquement d'un set de paramètres pour générer des modèles structuraux. Les approches computationnelles ont cependant beaucoup évolué ces deux dernières décennies, portées par l'essor de l'informatique. Elles intègrent des paramètres physico-chimiques de plus en plus précis permettant de modéliser des structures 3d avec de plus en plus de précisions. Leurs limitations actuelles se situent davantage au niveau du temps de simulation qui doit respecter un rapport étroit entre le temps nécessaire pour l'observation d'un phénomène biologique et le temps alloué aux experts pour leurs expérimentations [152]. De la même manière, la nature différente des protéines et de leur environnement nécessite un réglage précis des paramètres utilisés au sein de ces techniques. La reconnaissance de l'apport des méthodes théoriques pour la modélisation moléculaire est d'ailleurs très présente aujourd'hui et fut mise en exergue en 2013 par l'obtention du prix Nobel de chimie par ses pionniers (Martin Kaplus, Michael Levitt et Arieh Warshel) [8].

La différence entre la précision des techniques expérimentales et computationnelles est donc aujourd'hui très fine et ces deux approches sont utilisées de façon complémentaire au sein de nombreuses études.

La modélisation moléculaire consiste à étudier *in silico* les caractéristiques et le comportement des molécules par des techniques théoriques et computationnelles variées. Son champ d'applications regroupe les méthodes permettant d'étudier la structure, la dynamique, les propriétés de surfaces et la thermodynamique des systèmes biologiques. Parmi les activités biologiques pouvant être décrites par les outils de modélisation moléculaire, on retrouve le repliement des protéines aboutissant à leur structure 3d, les catalyses enzymatiques, la stabilité des protéines, les changements conformationnels associés aux fonctions biomoléculaires et la reconnaissance moléculaire des protéines, ADN et autres complexes membranaires. La modélisation moléculaire regroupe entre autres les méthodes capables de générer des modèles de structures 3d de biomolécules à partir d'un ensemble de données décrivant un ensemble de connaissances. Ces données peuvent être simplement la structure primaire d'une protéine ou bien provenir d'autres résultats expérimentaux ou computationnels. Ces deux approches se basent sur des simulations de mécanique moléculaire prenant en compte des paramètres physico-chimiques ou mécaniques, prédisant l'évolution de chaque particule dans son environnement (cellulaire, extracellulaire, membranaire, etc.).

La modélisation doit faire face à l'hétérogénéité importante qui existe entre les protéines. Suivant leur taille, leur complexité et leur environnement différentes méthodes répondent à différents besoins. De la même manière, le niveau de représentation des systèmes moléculaires varie au sein des méthodes. Ces différents niveaux permettent d'obtenir un équilibre être les ressources de calcul nécessaires et la résolution demandée. Un retour sur les différentes méthodes et représentations peut être trouvé dans l'article de Marc Baaden et Richard Lavery [9].

1.2.2.1 Modèles théoriques des biomolécules

La représentation du niveau énergétique d'une protéine peut être faite de différentes manières suivant la précision recherchée et les ressources de calcul à disposition. L'un des enjeux de la modélisation moléculaire est la recherche d'un équilibre entre une précision suffisante des modèles 3d générés et la puissance de calcul nécessaire à leur génération qui doit rester dans une échelle temporelle et économique raisonnable pour l'expert scientifique. L'ensemble des phénomènes physico-chimiques utilisé pour simuler une structure moléculaire dans son environnement constitue un **champ de force** et permet de calculer l'énergie potentielle du

8. http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/

système (voir section 1.2.2.1). Comme dans la nature, la simulation cherche à minimiser cette énergie potentielle.

Modèle quantique

Nous ne ferons qu'une brève description sur les modèles *quantiques* biologiques du fait de leur complexité et l'impossibilité de leur application sur des systèmes moléculaires de plus de quelques centaines d'atomes. Il est cependant important de noter que les modèles quantiques sont les modèles les plus précis pour décrire un système moléculaire. Ils prennent en compte les effets quantiques du système étudiés en se plaçant à une échelle subatomique puisqu'ils intègrent les contributions énergétiques des électrons et du noyau, constituants des atomes. Ils se basent sur l'équation de Schrödinger [147] qui décrit le mouvement d'une particule grâce à sa masse, son énergie, la constante de Planck et son énergie potentielle. Cette équation est trop complexe pour admettre une solution analytique et est donc résolue de façon approchée et/ou numérique. L'approximation de Born-Oppenheimer [20] est une première tentative pour réduire la complexité de ces résolutions en utilisant le fait que les masses des électrons sont très petites par rapport aux masses des nucléons. Elle permet la décomposition de l'équation de Schrödinger en deux étapes, mais ne suffit pas à réduire la complexité de façon assez importante pour que son application soit envisageable sur des systèmes composés de plus de quelques atomes.

Nous nous concentrerons donc sur les méthodes de modélisation utilisant des modèles dits *classiques* décrivant les particules non plus à l'échelle quantique, mais à l'échelle atomique. Les calculs quantiques ne sont cependant pas laissés de côté et sont utilisés pour dériver des valeurs énergétiques moyennes décrivant les interactions entre deux ou trois particules. Ces valeurs fixes serviront de base pour les paramètres des modèles *classiques*.

À la frontière des modèles quantiques et classiques se situent les approches mixtes de QM/MM [178]. Ces approches rassemblent au sein des simulations, la précision du modèle quantique et la rapidité du modèle classique. Elles se basent sur une définition multi échelle de la protéine qui contiendra seulement quelques régions décrites par des équations quantiques, le reste étant décrit par un modèle classique. Les régions choisies comme décrites quantiquement correspondent à des régions d'intérêt et impliquées dans le phénomène biologique étudié.

Une autre incursion des modèles quantiques dans la biologie se retrouve dans les approches Car Parrinello [31], utilisées par exemple avec succès à l'ADN. Cette approche est une approximation de la méthode Born-Oppenheimer citée précédemment, elle maintient les électrons dans un état stable qui permet d'éviter l'utilisation de minimisation à chaque pas de temps d'une simulation.

Modèle classique

Les modèles classiques sont eux empiriques, découlant du modèle quantique, ils décrivent une protéine à travers les propriétés physico-chimiques des atomes la composant. L'ensemble des équations régissant l'énergie d'un système moléculaire et de ses paramètres associés est regroupé dans un **champ de force**. Cette notion de champ de force est primordiale dans le modèle classique et est utilisée par la majorité des approches de simulation. L'ensemble des contributions énergétiques décrites dans les champs de force va permettre d'évaluer les modèles moléculaires que génèrent les calculs lors de la phase de modélisation ou de simulation.

D'un point de vue calculatoire, un champ de force est un ensemble de règles et de paramètres imposant aux particules de la protéine les lois physiques s'appliquant à l'échelle molé-

culaire. Ces règles peuvent inclure de nombreux paramètres comme les conditions physiques de l'environnement (température, pression, etc.) et les propriétés physiques des particules (polarité, charge, rayon, etc.). Les paramètres physiques comme la température, la pression et les paramètres de la mécanique newtonienne sont utilisés au sein de fonctions de calcul, au cours du temps et permettent donc au système d'évoluer vers différents niveaux énergétiques au cours de la simulation. La dynamique moléculaire, simulant les atomes de la biomolécule selon des lois newtoniennes, et les méthodes de Monte-Carlo qui permettent d'échantillonner les agencements spatiaux 3d de la protéine via des changements successifs aléatoires de la position des atomes, se reposent sur ces lois physico-chimiques décrites dans chaque champ de force (voir section 1.2.2.2).

Un champ de force est le résultat de contributions énergétiques différentes qui, combinées, vont permettre le calcul d'une énergie potentielle pour l'ensemble du système moléculaire étudié. La somme de ces contributions peut être représentée de la manière suivante :

$$E = E_{liée} + E_{nonliée}$$

où les énergies des contributions des liaisons covalentes (interactions liées), résultats de liaisons fortes entre les atomes et difficiles à défaire, et des liaisons non covalentes (interactions non liées), liaisons à longue distance énergétiquement plus faibles que les précédentes, sont données par les calculs suivants :

$$E_{liée} = E_{liaison} + E_{angle} + E_{dièdres}(+E_{impropres})$$

$$E_{nonliée} = E_{électrostatique} + E_{vanderwaals}$$

Les paramètres influençant ces interactions regroupent les énergies de liaison entre atomes et l'énergie des angles plus ou moins complexes formés par des atomes voisins. Ils dépendent exclusivement de la configuration électronique et de la charge électrique des atomes qui varient suivant leur nature.

Les énergies de liaison et d'angle entre particules sont souvent modélisées par des potentiels harmoniques centrés autour de la valeur d'équilibre de la liaison/angle considéré et dérivés des calculs expérimentaux. Pour davantage de précision, mais à un coût computationnel plus important, on utilise parfois le potentiel de Morse qui décrit plus précisément les états de vibration d'une structure, car il prend en compte les effets d'absence de liaison ainsi que l'existence d'états non liés. La différence de contribution de ces deux potentiels est représentée dans la Figure 1.18. Les angles dièdres possèdent eux plusieurs minimas et leur contribution énergétique ne peut donc pas être représentée par ces potentiels. Il est commun d'ajouter un terme décrivant les angles de torsion impropres afin de contraindre la géométrie de certains plans. On cherchera par exemple à forcer la planéité des cycles aromatiques de chaînes latérales.

Les termes d'énergie non liée sont plus difficiles à calculer, car chaque atome interagit de façon non liée avec l'ensemble des atomes du système alors qu'il agit de façon liée avec un ensemble limité d'atomes avec qui il possède une liaison covalente (4 atomes maximum). Parmi les interactions non liées, les forces de Van der Waals sont rapidement négligeables lorsque la distance entre deux atomes est trop importante. Lorsque la contribution du terme de Van der Waals est modélisée par un potentiel de Lennard-Jones 6-12, un des potentiels les plus utilisés, elle décroît proportionnellement à r^{-6} pour les forces attractives et r^{-12} pour les forces répulsives où r représente la distance entre les deux atomes considérés (cf. Figure 1.19a). Cette modélisation est cependant inexacte pour les forces répulsives lorsque la distance est réduite puisque celles-ci augmentent de façon exponentielle. Afin d'accélérer les

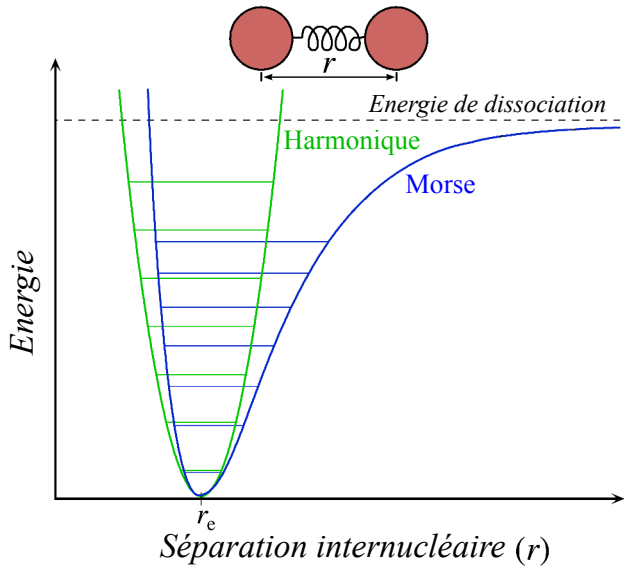


FIGURE 1.18 – *Contributions énergétiques des liaisons représentées par un potentiel harmonique (en vert) ou un potentiel de Morse (en bleu) en fonction de la distance entre deux atomes.*
Source : Mark Somoza - https://commons.wikimedia.org/wiki/File:Morse-potential.png

calculs, on introduit usuellement un seuil pour la distance au-dessus duquel la contribution des interactions de Van der Waals est égale à 0.

Les forces électrostatiques ne sont pas si aisées à calculer du fait de leur contribution non négligeable à des distances considérées comme grandes au sein d'une protéine et impliquant de nombreux atomes. Elles sont habituellement représentées grâce au potentiel de Coulomb qui ne décroît que proportionnellement à r^{-1} comme illustré sur la Figure 1.19b. Plusieurs méthodes existent pour réduire la complexité de calcul induite par cette modélisation. Une solution repose, à la manière de ce qui est fait pour les contributions de Van der Waals, sur la mise en place d'un seuil au-delà duquel l'interaction est considérée comme négligeable. Cependant, cette technique provoque des artefacts importants du fait du changement brutal de la contribution électrostatique avant et après le seuil choisi. Une solution pour limiter ces artefacts est la mise en place d'une fonction de commutation qui introduit un facteur d'échelle compris entre 0 et 1 à l'extérieur et l'intérieur du seuil de distance. Il existe également d'autres méthodes plus coûteuses que les seuils, mais plus précises comme la méthode de *Particle mesh Ewald* (PME) qui s'appuie sur des approximations des valeurs électrostatiques au-delà du seuil en considérant l'espace comme un espace de Fourier et nécessite seulement le découpage de l'espace en une grille régulière.

Les paramètres utilisés au sein des champs de force sont dérivés de données expérimentales, mais peuvent également être le résultat de calculs computationnels de mécanique quantique. Parmi les champs de force, certains décrivent la protéine de façon atomique en considérant chaque atome comme une particule et en appliquant des paramètres propres à chaque atome dans les équations. Ces champs de force, appelés **tout-atome**, malgré leur précision, souffrent d'un temps de calcul de leurs équations très important qui est un frein significatif pour l'évaluation énergétique de systèmes moléculaires de plus de quelques centaines/milliers d'atomes. Afin de permettre l'étude de protéines dépassant cette taille, il est possible de sim-

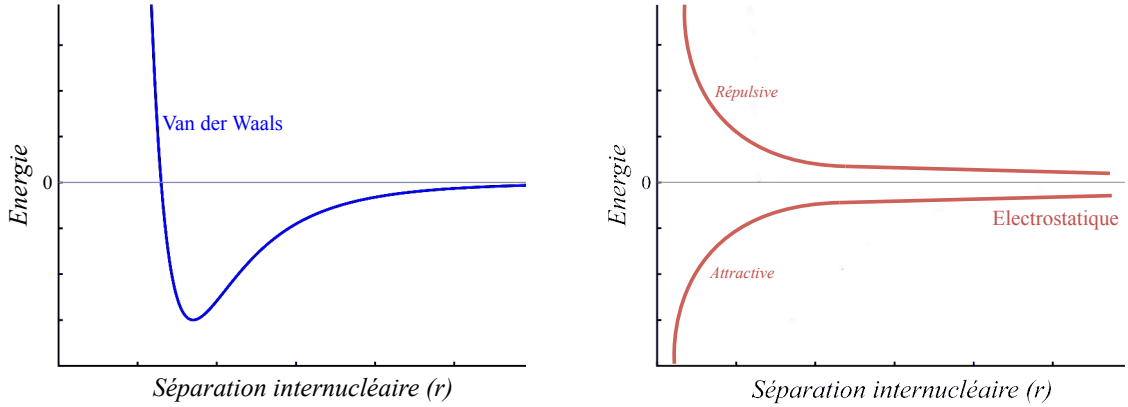


Figure 1.19 – *(a) Contributions énergétiques des forces de Van der Waals représentées par un potentiel de Lennard-Jones. (b) Contributions énergétiques des forces électrostatiques représentées par un potentiel de Coulomb en fonction de la distance entre deux atomes.*

plifier le champ de force. Dans ces champs de force simplifiés, appelés **gros grain**, le modèle de la protéine est différent puisque certains groupes d'atomes, groupes dont les propriétés physiques sont bien connues, sont considérés comme des particules uniques, réduisant ainsi le nombre de particules du système. La différence entre deux modèles de champs de force tout-atome et gros grain est illustrée dans la Figure 1.20. Une particule étant régie par un ensemble d'équations, réduire le nombre de particules revient à réduire le nombre de calculs à effectuer. Les chaînes latérales des acides aminés ainsi que les groupements méthyles sont deux exemples de groupements d'atomes dont la contribution énergétique est suffisamment approximée pour pouvoir être simplifiée. Dans le même but de simplification des calculs, certains atomes dont la contribution est considérée comme négligeable sont enlevés des champs de force, c'est souvent le cas des atomes d'hydrogène par exemple.

Le choix d'un champ de force dépend à la fois de l'environnement dans lequel est modélisé le complexe moléculaire d'intérêt ainsi que du degré de précision utilisé pour décrire les particules du système (tout-atome, atomes unifiés, gros grains, etc.). Parmi les champs de force les plus utilisés, nous pouvons citer CHARMM [27], AMBER [128], GROMOS [125] ou OPLS [86].

Les différentes valeurs d'énergie potentielle d'une protéine sont obtenues par la résolution de l'ensemble des équations des champs de force. Ces équations dépendent directement de la configuration spatiale de la protéine et constituent le paysage énergétique de cette protéine. Suivant l'objectif de la simulation, on cherchera soit à identifier le minimum global de ce paysage, soit à trouver des minima locaux traduisant des états structurels différents, mais stables d'une protéine (voir Figure 1.21).

1.2.2.2 Simulation moléculaire

La simulation moléculaire cherche à modéliser l'état ou l'évolution d'un système de particules biologiques. Elle peut tenter de décrire un changement conformationnel ou le déroulement d'un phénomène impliquant plusieurs molécules. Une protéine est rarement simulée dans le vide et il est commun d'ajouter au moins un modèle de solvant afin de se rapprocher au maximum des conditions réelles.

Plusieurs approches répondent aux différents besoins et applications de la simulation

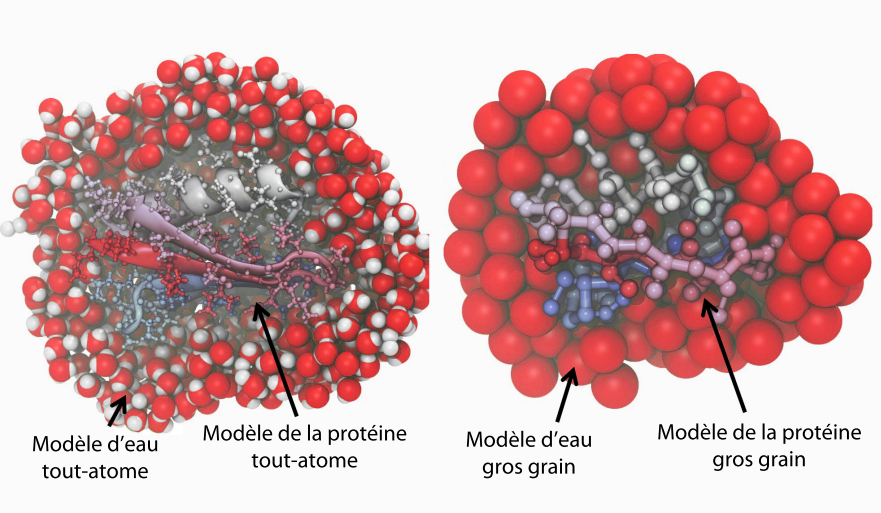


FIGURE 1.20 – *Modèles de représentation d'un système moléculaire. À gauche, chaque atome est considéré indépendamment par une particule unique. À droite, chaque particule représente un ensemble d'atomes constituant un groupe particulier du système (molécules d'eau, chaînes latérales, chaînes principales, etc.)*
Source : http://www.ks.uiuc.edu/Research/cgfolding/

moléculaire. Les méthodes de mécaniques moléculaires simples ne prennent pas en compte l'énergie cinétique de l'environnement et sont plus rapides et adaptées à la recherche d'un état d'équilibre. Au contraire, la dynamique moléculaire ou les méthodes basées sur Monte-Carlo voient la simulation des conditions de température, pression, etc. influençant le comportement des particules pour calculer l'énergie du système, mais là où la notion de temps est présente en dynamique moléculaire, elle est absente de l'approche basée sur Monte-Carlo. La dynamique sera donc très adaptée pour l'étude de phénomènes moléculaires au cours du temps alors que les méthodes de Monte-Carlo seront davantage utilisées pour échantillonner les configurations possibles d'une protéine et parcourir son paysage énergétique.

Mécanique moléculaire

La mécanique moléculaire fait abstraction des paramètres thermodynamiques régissant les évolutions conformationnelles d'une structure protéique pour décrire de façon plus simple un système de particules en condition statique par la seule mécanique newtonienne. Elle repose sur l'utilisation de champs de force afin de calculer l'énergie potentielle d'une protéine. Ces différentes propriétés sont appliquées pour les modèles tout-atome :

— Chaque atome est représenté par une particule unique
— Chaque particule possède un rayon, une polarisation et une charge nette constante
— Les interactions liées sont considérées comme des ressorts

De façon plus précise, le rayon des particules est bien souvent le rayon de *van der Waals* qui est une sphère géométrique théorique censée représenter l'espace occupé par un atome. La charge nette est dérivée de calculs expérimentaux ou quantiques afin d'être la plus proche possible de la réalité. Enfin, la distance à l'équilibre des ressorts représentant les liaisons est égale à la longueur expérimentale de la liaison.

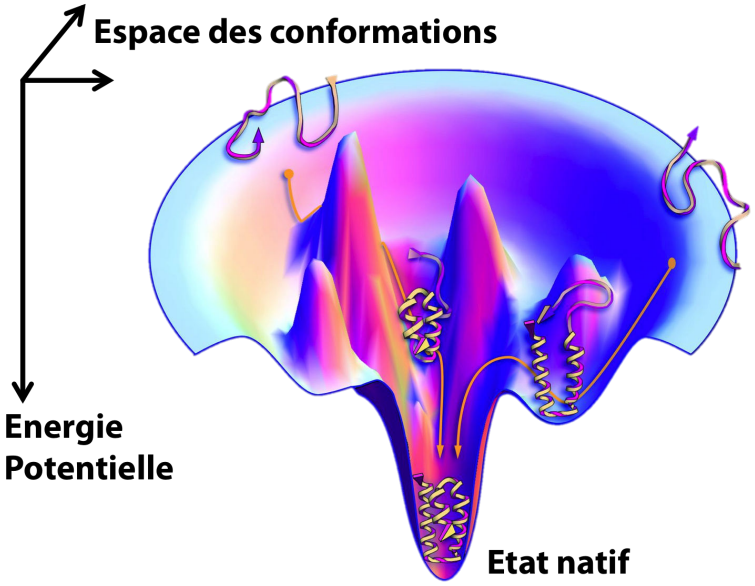


Figure 1.21 – *Représentation d'un paysage énergétique d'une protéine suivant sa conformation spatiale. Plus l'énergie potentielle est haute, plus la protéine est considérée comme désordonnée et donc instable. Un seul minimum global existe correspondant à l'énergie de la conformation la plus stable de la protéine : état natif de la protéine.*
Source : [53]

L'application la plus directe et courante de la mécanique moléculaire est la minimisation énergétique où l'on va chercher à obtenir des modèles avec une énergie de plus en plus faible afin de trouver la structure 3d la plus stable pour une protéine donnée. Les algorithmes de minimisation moléculaires sont divers, mais ils se basent très souvent sur les champs de force de mécanique moléculaire.

Dynamique moléculaire

Le principe de la dynamique moléculaire (MD) repose sur la résolution d'équations thermodynamiques respectant les lois énoncées en mécanique moléculaire afin de guider les particules décrivant la protéine vers des configurations spatiales différentes. Son but est donc de simuler les mouvements de particules dans le temps pour des conditions thermodynamiques (température, pression, etc.) précises tout en prenant en compte la mécanique newtonienne régissant les contributions énergétiques de ces particules.

Au sein des publications scientifiques de ces dernières années, les échelles de temps des processus biologiques simulés vont de quelques nanosecondes (10^{-9}s) à quelques microsecondes (10^{-6}s). Ces échelles de temps biologiques correspondent de plusieurs jours à plusieurs années de calculs computationnels. Cependant, la parallélisation au sein de CPU et maintenant de GPU permet de diviser ces temps par un facteur allant de 10 à quelques milliers afin d'obtenir des temps de calcul s'étalant sur quelques semaines au maximum.

Une dynamique moléculaire introduit, à la différence des deux méthodes précédentes, la notion de mouvement atomique qui se caractérise par la formule donnée par la 2de loi de Newton :

$$F_i = m_i.a_i$$

où F_i représente la somme des forces exercées sur l'atome i de masse m_i et d'accélération a_i. On obtient la force localisée sur un atome en dérivant l'énergie du système entier par rapport à la position de l'atome. À chaque pas de temps on obtiendra ainsi une nouvelle position de l'ensemble des atomes et il est possible d'évaluer l'énergie potentielle du système grâce au champ de force utilisé.

Le facteur de temps dans les simulations de MD évolue dans un espace discret et tout au long de la simulation, les coordonnées 3d de la structure sont sauvegardées à des pas de temps réguliers afin de pouvoir rejouer la trajectoire de la simulation à tout moment au moyen de programmes de visualisation 3d. Chaque pas de la simulation correspond à un ensemble de coordonnées 3d qui est associé à une valeur d'énergie potentielle.

Les programmes de dynamique moléculaire les plus connus regroupent plusieurs programmes basés exclusivement sur les champs de force qu'ils utilisent : AMBER, GROMACS ou CHARMM tiennent leur nom de leur champ de force (AMBER, GROMOS et CHARMM respectivement) et sont les plus utilisés dans la communauté de modélisation moléculaire. Nous pouvons aussi citer NAMD [9] [130] spécialisé dans la dynamique moléculaire parallélisée souvent utilisé pour simuler de grands systèmes moléculaires de plusieurs millions d'atomes.

Minimisation moléculaire

Il s'agit de trouver le minimum de la fonction qui permet de retrouver l'énergie potentielle de la molécule en fonction de sa géométrie. Il s'agit plus précisément de trouver des puits d'énergies le long du paysage énergétique de la protéine reflétant les conformations structurelles les plus stables de la protéine (voir Figure 1.21). Cette minimisation peut emprunter plusieurs algorithmes pour parvenir à ses fins, mais le principe central reste le même. Il s'agit de bouger les atomes d'un système afin de minimiser l'énergie à chaque déplacement d'atomes. Chaque pas de la minimisation va évaluer l'impact énergétique du déplacement et ainsi décider si celui-ci est stabilisant pour la structure ou perturbant. Cette évaluation est faite via la résolution des fonctions du champ de force et du calcul de l'énergie potentielle.

Algorithme de Monte-Carlo

La méthode de simulation basée sur Monte-Carlo est l'une des plus usitées [118] consiste en un échantillonnage aléatoire des différentes configurations spatiales d'un système moléculaire. Les changements entre les configurations sont des changements induits de façon statistique et se caractérisent par la perturbation de la longueur d'une liaison, d'un angle, d'un angle dièdre ou le léger déplacement d'un atome ou d'un groupe d'atomes. Il est possible de décrire le principe général d'une simulation utilisant la méthode de Monte-Carlo de la manière suivante :

1. Calcul de l'énergie du système dans son état n grâce aux équations du champ de force utilisé => U(n)
2. Changement aléatoire pour obtenir une nouvelle conformation spatiale n+1
3. Calcul de l'énergie du système dans son état n+1 => U(n+1)
4. Test d'acceptabilité: Si U(n+1) est inférieur à U(n) alors le changement est accepté et la conformation n+1 est gardée. Si U(n+1) est supérieur à U(n) alors on évalue le

9. http://www.ks.uiuc.edu/Research/namd/

changement selon l'équation : $acc(n- > n+1) = min(1, exp(-(U(n+1) - U(n))/kT))$ où k est la constante de Boltzmann et T la température du système. Dans cette équation, le facteur de Boltzmann ($exp(-(U(n+1) - U(n))/kT)$) est comparé avec un nombre aléatoire compris entre 0 et 1. Si le nombre aléatoire est plus grand que le facteur de Boltzmann alors la conformation est rejetée, sinon, elle est acceptée.

5. Le processus reprend à l'étape 1 avec l'état n ou n+1 suivant le résultat de l'étape précédente.

1.2.2.3 Folding moléculaire ou prédiction de structure tertiaire

La prédiction de structure tertiaire d'une protéine consiste à retrouver la structure 3d naturelle d'une protéine à partir de sa séquence primaire. Cette prédiction de structure est également appelée *folding* moléculaire (ou repliement en français), car elle peut permet d'obtenir des informations sur le processus de repliement des protéines

Lorsqu'aucune information ne peut être utilisée pour cette prédiction, on parle de prédiction *ab initio*, prédiction passant souvent par la prédiction préalable de la structure secondaire de la protéine. Cette première étape de prédiction peut passer par : (1) des méthodes statistiques attribuant une probabilité de structuration en feuillet/hélice/coude par acide aminé, (2) des approches physico-chimiques en cherchant à calculer les interactions interatomiques et donc appliquant les forces d'attraction/répulsion pour obtenir une structure ou (3) un alignement et comparaison évolutive en protéines de la même famille dont la structure secondaire est connue. Lorsque la structure secondaire est connue, il est possible d'effectuer une simulation moléculaire afin d'utiliser des paramètres physico-chimiques pour diriger le repliement des motifs de structures secondaires identifiés. Les changements conformationnels menant de la structure de départ au modèle 3d final sont souvent importants puisqu'aucune information préalable n'est utilisée.

Il est évidemment possible de s'appuyer sur des informations expérimentales existantes afin de prédire la structure 3d d'une protéine. Parmi les informations utilisées, les bibliothèques de fragments ou alphabets structuraux permettent d'obtenir la structuration la plus retrouvée pour différentes séquences d'une dizaine d'acides aminés. Il est également possible de s'intéresser aux protéines dont la structure 3d est déjà connue. Grâce à un alignement global de la séquence de la protéine étudiée, il est possible d'identifier des protéines possédant une similarité de séquence significative (> à 30%) et ainsi structurer la protéine à prédire à partir des motifs structuraux connus; il s'agit de la modélisation par homologie. Une dernière méthode de prédiction est la méthode de *threading* qui voit l'enfilage de courtes séquences d'acides aminés au sein de structures 3d répertoriées comme courantes au sein de banques de données de structures 3d de protéines. Chaque enfilage donne lieu à un calcul d'énergie des acides aminés contraints spatialement et les structures 3d de plus basses énergies sont considérées comme les candidats les plus probables pour structurer la protéine.

1.2.2.4 Docking moléculaire ou amarrage protéine-protéine

Le docking moléculaire (ou *amarrage* moléculaire en français) consiste à prédire l'orientation et la configuration spatiale d'une molécule par rapport à une seconde lorsqu'elles se lient pour former un complexe moléculaire stable. Ce domaine de la modélisation moléculaire est de grande importance, car la quasi-exclusivité des réactions métaboliques se déroulant au sein d'un organisme se traduit par la liaison de deux ou plusieurs molécules entre elles. C'est notamment le cas des protéines dont les fonctions sont dirigées par leur capacité à se lier à des partenaires moléculaires afin d'induire différents effets. Ces partenaires peuvent

être de différentes natures : acides nucléiques (en particulier pendant les étapes de transcription et traduction), autres protéines, petites molécules (avec souvent un rôle activateur ou inhibiteur), etc.

Parmi les applications les plus importantes du docking moléculaire, la conception de médicaments dont les constituants vont se lier à une cible moléculaire pour un effet agoniste ou antagoniste est un domaine où les méthodes de docking jouent un rôle primordial.

Une image souvent utilisée pour représenter le concept de docking moléculaire est le problème du verrou et de la clé. L'idée de ce problème est de trouver le trou, site de liaison de la clé, puis l'orientation correcte de la clé pour ouvrir le verrou. La protéine serait le verrou dont le trou se trouverait à sa surface et le ligand constituerait donc la clé.

Lorsqu'on s'intéresse au docking *ab initio*, où aucune information n'est utilisée sur le potentiel site de liaison, deux approches peuvent être identifiées afin de générer un modèle 3d de complexe fonctionnel entre la protéine et son ligand.

La **complémentarité de forme** considère la protéine et le ligand du point de vue des propriétés géométriques et physiques de leur surface. Parmi les descripteurs utilisés pour représenter la protéine et son ligand, la surface accessible au solvant (SAS) pour la protéine et des descriptifs de complémentarité de surface pour le ligand peuvent être utilisés [155]. Un autre descripteur décrit les paramètres hydrophobes de la protéine. Des patchs hydrophobes et leurs opposés hydrophiles peuvent ainsi aider à identifier les sites de liaison et mettre de côté les régions peu propices à être en contact avec une autre protéine [84]. Une fois que les paramètres géométriques sont analysés et utilisés pour rapprocher les deux partenaires, une minimisation moléculaire peut être effectuée afin de retrouver un état énergétique stable.

La **simulation** utilise les propriétés physico-chimiques et les contraintes géométriques des liaisons atomes pour diriger l'interaction entre la protéine et son ligand. Des poses sont générées tout au long du temps et un calcul de score, basé entre autres sur des calculs énergétiques d'interaction, permettra de connaître la pertinence de la structure du complexe et sa stabilité. Les méthodes basées sur des simulations permettent, à la différence des méthodes par *complémentarité de forme* qui mettent en jeu des partenaires rigides, d'introduire différents niveaux de flexibilité pour le ligand. Cette flexibilité permettra de rendre compte d'éventuels changements conformationnels entre la structure sous forme non liée d'un ligand et sa structure sous forme liée.

Il est également possible d'incorporer des informations expérimentales au sein des méthodes de docking afin d'orienter le processus [54]. Ces informations dirigeront le processus de docking vers des solutions validées comme cohérentes par les méthodes expérimentales réduisant ainsi le champ de conformations à explorer.

1.2.2.5 Évaluations des résultats théoriques

Chacune des méthodes de simulation moléculaire permet de générer des modèles censés représenter, finalement, un état stable et fonctionnel d'une protéine relié à un phénomène observé ou observable de manière expérimental. Cependant, lorsqu'aucune vérification expérimentale de ces modèles n'est possible, il est important de pouvoir évaluer chacun des modèles afin d'identifier les plus pertinents. Le choix de la méthode d'évaluation est important, car elle décidera des modèles qui seront considérés comme les plus proches de la structure native des protéines. Cette évaluation peut se faire soit par l'évaluation énergétique du système grâce à la résolution des équations définissant le champ de force utilisé, soit par l'utilisation de bases de connaissances permettant d'évaluer la pertinence de la structure 3d par rapport à l'ensemble des structures déjà résolues.

Les algorithmes de dynamique moléculaire et de modélisation sont nombreux, et il est nécessaire de posséder des outils d'évaluation fiables afin de connaître leur précision et leur efficacité. Il est aussi intéressant de savoir quels programmes possèdent les approches les plus pertinentes pour traiter certains cas biologiques. Afin d'évaluer leurs programmes, les équipes scientifiques peuvent participer à l'expérience CASP (Critical Assessment of protein Structure Prediction) qui propose de prédire la structure 3d d'une protéine à partir de sa seule séquence primaire. En plus de la structure tertiaire, certaines catégories supplémentaires peuvent être mises en place (prédiction de fonction, prédiction de structure secondaire, prédiction de contact résidu-résidu, etc.). CASP fournit également les critères d'évaluations utilisés pour juger de la précision des modèles générés par les algorithmes de docking. Ils se basent sur des calculs de distance appelés GDT_TS (Global Distance Test Total Score) afin de calculer la distance moyenne entre les modèles prédits et la structure expérimentale considérée comme naturelle.

De la même manière que pour les simulations moléculaires, les modèles de complexes moléculaires générés par les techniques de docking vont être évalués afin de décider de leur pertinence et stabilité biologique. En plus des énergies propres à chaque partenaire évalué individuellement, des paramètres comme les contacts hydrophobes ou le nombre de liaisons hydrogènes entre les deux partenaires, le nombre des molécules ou la surface accessible au solvant à l'interface vont influer le calcul de **scores** d'un complexe. Ces scores vont permettre de juger de la pertinence des complexes moléculaires entre eux. De nombreuses méthodes de *scoring* existent, basées sur des paramètres physico-chimiques (provenant de champs de force), empiriques ou statistiques (extraits de connaissances).

À la manière de CASP, CAPRI est une expérience de prédiction aveugle dont le but est de prédire la structure d'un complexe protéine-protéine à partir des coordonnées atomiques des deux partenaires impliqués sous leur forme non liée. Cette expérience permet aux équipes de comparer leurs méthodes de docking, mais elle est aussi à l'initiative de standards d'évaluation permettant d'évaluer les modèles de complexes prédits par rapport à un complexe obtenu expérimentalement et servant de référence.

1.2.2.6 Analyse post-simulation moléculaire

L'étape d'analyse est indispensable, au même titre que l'évaluation des modèles, à la production de connaissances en biologie structurale. La méthodologie d'analyse choisie n'est jamais systématique et dépend largement du complexe protéique étudié, du type de phénomène observé ou des données générées de manière expérimentale et théorique. De manière générale, le processus d'analyse des résultats de simulation et de dynamique moléculaire implique de prendre en compte la dimension spatiale, temporelle et énergétique du phénomène modélisé.

L'analyse des résultats selon ces trois dimensions est complexe, car chacune peut faire l'objet d'interprétations contradictoires, difficilement vérifiables sans expérimentation. Ainsi, une énergie potentielle élevée à un instant t d'un système moléculaire peut être le marqueur d'une étape de transition entre deux conformations plus stables, donc énergétiquement plus basse, mais peut également constituer une structuration non naturelle, voire même le résultat exclusif des biais computationnels.

Divers graphiques, nuages de points ou autres techniques visuelles de mise en relation permettent d'extraire de premières informations précieuses pour la compréhension du déroulement d'une simulation. Ces techniques se couplent particulièrement bien aux analyses de l'évolution conformationnelle d'une structure tout au long de la simulation. Nous avons vu dans les sections 1.2.2.5 que ces analyses sont parfois partie intégrante de l'algorithme

régissant le processus d'évolution de la simulation lorsque l'évaluation s'appuie sur des facteurs statistiques pour évaluer la stabilité énergétique d'une structure, en particulier dans les simulations de docking moléculaire.

Les approches analytiques ne passent pas exclusivement par l'analyse stricte des données de la simulation et nous avons vu qu'il était possible d'évaluer certaines étapes de modélisation grâce à des informations statistiques (à différencier de physico-chimiques). Il est intéressant par exemple de connaître la stabilité d'une protéine au cours du temps à la suite de mutations particulières de sa séquence en acides aminés [114]. Il est donc important d'associer un caractère statistique aux données brutes de simulation. Ces données permettent entre autres de vérifier que les parties conservées au cours de l'évolution sont préservées, de savoir à quel point certaines régions de la protéine varient de leurs homologues et si d'éventuels sites d'action et de liaison sont conservés favorisant la compréhension globale du phénomène étudié pour l'utilisateur.

Dans le cas d'analyses de docking, en plus des préoccupations analytiques précédentes, la présence de plusieurs molécules censées former une liaison implique des approches supplémentaires. La vérification des contraintes structurales et énergétiques de la protéine respectées après liaison d'un ligand est une étape indispensable, de la même façon que l'observation d'une modification conformationnelle significative de la protéine entre son état lié et son état non lié. La stabilité d'un complexe pourra également être reflétée par le nombre de liaisons covalentes ayant été formées, la nature de ces liaisons et leurs contributions énergétiques locale et globale.

Les analyses post-simulation permettant d'extraire des informations de la trajectoire sont souvent intégrées dans la suite d'outils de simulation (AMBER [128], CHARMM [27], GROMACS [132] pour les plus connus d'entre eux). Cependant, plusieurs bibliothèques ont également été conçues afin de permettre d'utiliser certains algorithmes et outils d'analyses de simulation moléculaire en dehors des suites logicielles dédiées (BioPython [38], MDAnalysis [119], MDTraj [116] par exemple).

1.2.3 Représentation et visualisation moléculaire

Il est important de stocker les structures 3d de protéines résolues de façon théorique ou expérimentale, mais il est tout autant important de pouvoir les visualiser. La visualisation moléculaire est indissociable des progrès effectués par la biologie au cours de son histoire. Tout d'abord utilisée comme moyen de communication et parfois de vulgarisation, elle est aujourd'hui un des vecteurs majeurs de génération d'hypothèses et de connaissances.

1.2.3.1 Evolution des représentations moléculaires

Les premières tentatives de représentation de structures moléculaires ont correspondu aux principales découvertes expérimentales ayant permis d'obtenir des indications claires sur la structure des protéines étudiées. On peut distinguer deux courants de représentation de ces structures, courants complémentaires et donc la séparation est principalement due à l'évolution des technologies et des moyens de communication.

Représentations physiques

Les modèles physiques constituent, les premières tentatives de représentation de structures de biomolécules. On parle ici de représentation à des échelles allant de quelques centimètres à plusieurs mètres de hauteur.

Une des toutes premières représentations exposées est particulièrement connue puisque c'est celle de l'ADN, créée par Watson et Crick et récompensée à travers ses auteurs par un prix Nobel en 1953 [179] (voir Figure 1.22).

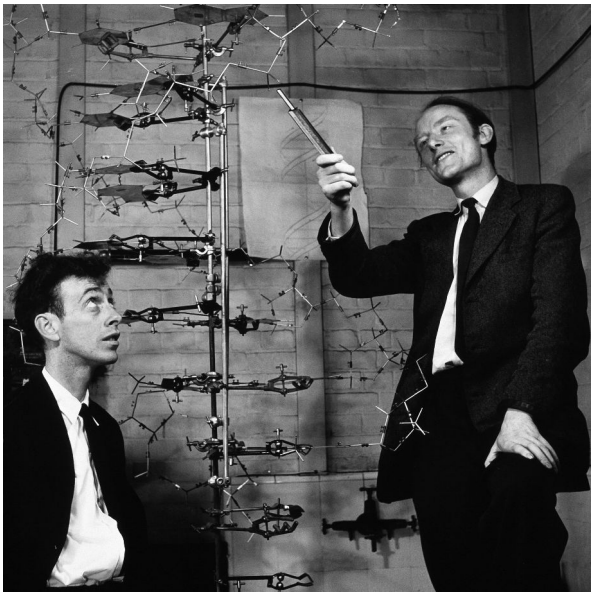
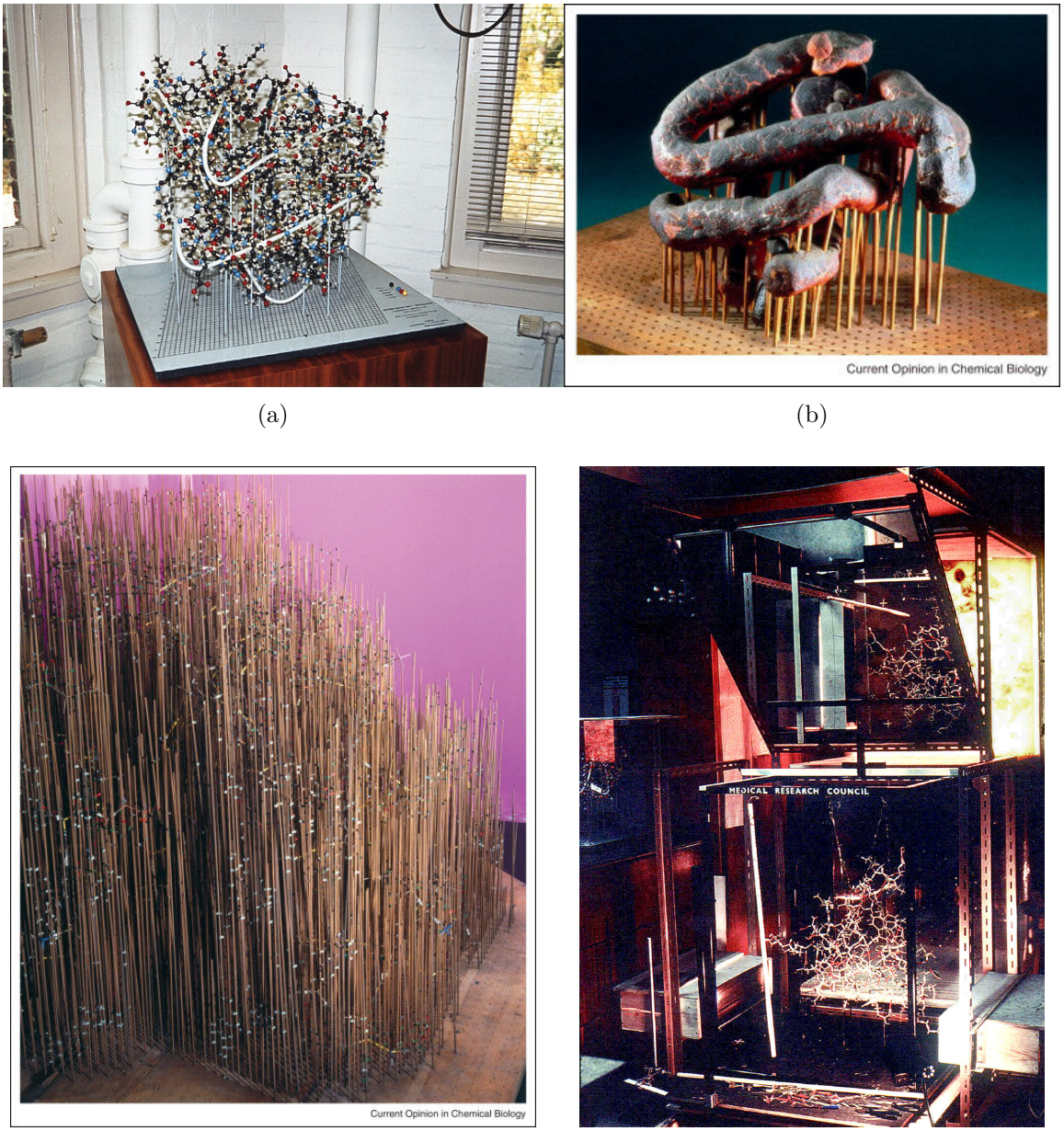


Figure 1.22 – *Modèle physique de l'ADN par Watson et Crick publiée en 1953 dans Nature. Le modèle physique créé pour l'occasion est constitué de pièces en ferraille récupérées au sein de leur laboratoire.*
Source : http://www.thehistoryblog.com/archives/25193

La première représentation physique notoire d'une protéine est le fait de J.C. Kendrew qui, pour illustrer la résolution de la structure de la myoglobine par cristallographie [89] à rayon X, a créé plusieurs modèles de cette structure. Il créa en collaboration avec Max Perutz le premier modèle, en 1957, composé exclusivement de plasticine (voir Figure 1.23b). La qualité du modèle était en corrélation directe avec la résolution expérimentale de la structure cristallographique qui avait une basse résolution d'environ 6Å, rapportant ainsi grossièrement la structure tertiaire, mais pas davantage de la protéine. Ce modèle leur permit tout de même d'obtenir en 1962 le prix Nobel de chimie. Il fut le prédécesseur d'un autre modèle construit par les deux mêmes scientifiques et appelé *forest of rods* (littéralement forêt de tiges), qui, avec une échelle de 5cm pour 1Å, était composé de 2500 tiges verticales remplissant un cube de 2m de côté (voir Figure 1.23c). Des attaches de couleur étaient attachées aux tiges afin de représenter la densité d'électrons et guider la construction du modèle. Il fit suite à une amélioration de la résolution des cartes de densité de cristallographie atteignant 2Å de précision. Les chaînes latérales n'étaient cependant pas encore bien discernables et il fallut attendre 1965 pour que les résultats expérimentaux permettent de construire un modèle plus précis.

Pour ce faire, Kendrew approcha A.A Baker et ils mirent au point un modèle composé de balles et de bâtons afin de représenter les atomes et leurs liaisons (voir Figure 1.23a). L'échelle de ces modèles était plus petite (entre 2.5 et 1cm / Å) et rendait ces modèles plus facilement transportables, mais cependant de taille encore conséquente. La construction de ces modèles se faisait grâce à la projection des cartes de densités électroniques dessinées sur des plaques de verre dans un dispositif innovant appelé *Richard's box* et illustrés dans la Figure 1.23d.

Dans le début des années 1970, le cristallographe Byron Rubin mit au point une machine

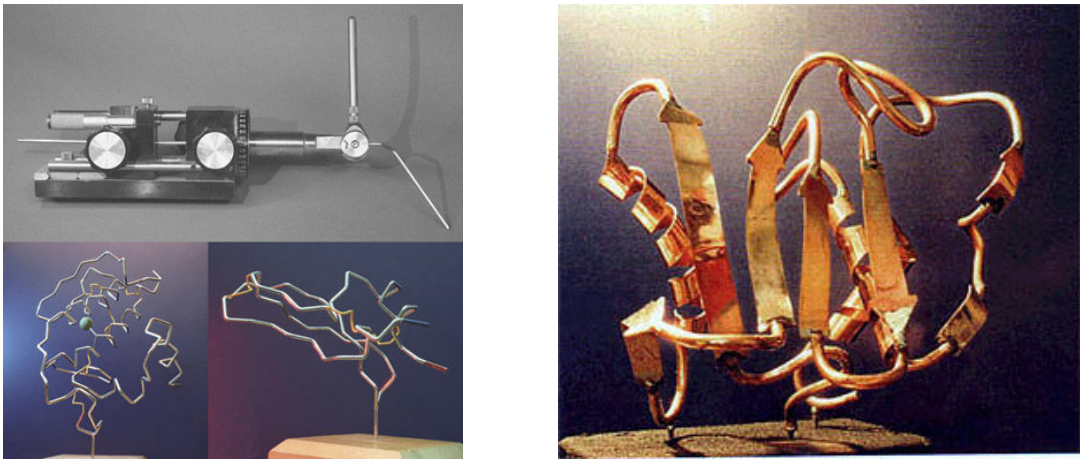


(a) (b)

(c) (d)

FIGURE 1.23 – *(a) Modèle balle et bâtons de Kendrew et Baker représentant la myoglobine. Un fil blanc représente la chaîne principale au milieu des atomes (boules de couleur) et de leurs liaisons (bâtons). (b) Premier modèle physique de la myoglobine, conçu par Sir John Kendrew et Max Perutz. (c) «Forest of rods», modèle basé sur des tiges et des attaches de couleurs représentant également la myoglobine. (d) Dispositif de visualisation appelé «Richards box» (ou «Fred's Folly»).*
Source : (a) http://www.umass.edu/molvis/francoeur/barker/barker.html*, (b)(c) Science Museum/Science & Society Picture Library, (d) [104]*

permettant de plier un câble selon la forme de la chaîne principale d'une protéine (voir Figure 1.24a). Les modèles générés par cette machine, appelée *Byron's Bender*, sont de plus petites tailles que les autres modèles physiques existants à cette époque-là et permettaient donc leur



(a) (b)

Figure 1.24 – *(a) En haut, «Byron's Bender» l'outil mit au point par Byron Rubin. En dessous sont exposés les modèles de la phospholipase pancréatique A2 (à gauche) et de l'inhibiteur de la trypsine pancréatique bovine (à droite). (b) Sculpture moléculaire de Byron Rubin représentant une collagénase neutrophile humaine.*
Source : (a) http://jbiocommunication.org/issues/31-2/feature2.html*, (b)*http://www.umass.edu/microbio/rasmol/history.htm

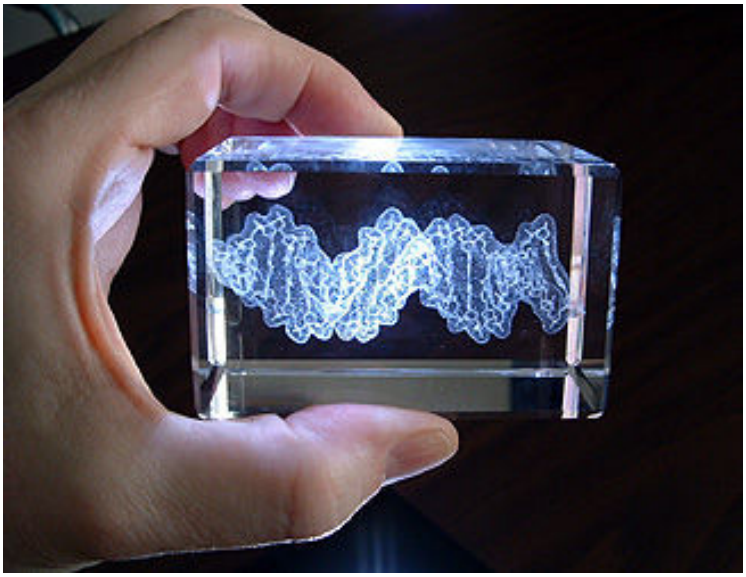
utilisation lors de réunions scientifiques. C'est d'ailleurs à l'occasion de l'une d'entre-elles, en 1976, alors que moins de deux douzaines de structures avaient été résolues, que deux modèles physiques de Bender représentant le fragment Fab de l'immunoglobuline et la dismutase superoxyde furent ramené par deux groupes distincts de chercheurs. Après comparaison des modèles, ils émirent l'hypothèse que ces deux protéines adoptaient probablement les mêmes méthodes de structuration malgré seulement 9% d'identité de leurs séquences primaires [139]. Cet événement, pouvant être qualifié de hasard de l'histoire, fut la première reconnaissance de l'existence de la superfamille des domaines immunoglobulines non liées par leur séquence.

Au-delà de l'importance de l'aspect tactile et visuel des modèles de *Bender*, une autre facette importante de ces modèles est leur possibilité de vibrer au toucher, simulant ainsi le mouvement thermique. Cet aspect est d'importance quand comparé aux modèles virtuels qui font parfois oublier à leurs utilisateurs la flexibilité que possède une protéine dans son environnement cellulaire.

Byron Rubin s'associa également, de façon plus étonnante, avec un garage de voitures local afin d'utiliser leur machine de pliage de pot d'échappement, fonctionnant sur le même principe que sa machine, afin d'atteindre une plus grande échelle comme pour la collagénase de la Figure 1.24b. Il créa de nombreuses sculptures de molécules considérées comme importantes à partir de ce temps-là et jusqu'aux années 90.

Concurrencés par les modèles sur ordinateurs, les modèles physiques surent néanmoins évoluer avec les technologies, après le métal, le bois ou le plastique furent le support de plusieurs modèles. À la fin des années 90, grâce aux efforts de Michael Bauler et Tim Herman d'utiliser les nouvelles technologies d'ingénierie, les lasers et imprimantes 3D permirent la création de modèles à moindre coût et moindre temps ainsi qu'à une échelle toujours plus fine, plusieurs d'entre eux sont représentés dans les Figures 1.25a et 1.25b. L'ensemble de ces modèles prit une part importante dans l'éducation et voit maintenant plusieurs projets se servir de ces modèles comme interfaces tangibles pour des expériences de simulations molé-

(a)



(b)

FIGURE 1.25 – *(a) Modèle en bois de Julian Voss-Andreae illustrant la chaîne principale d'une phéromone. (b) Modèle en verre de l'ADN creusé grâce à des rayons lasers fins. Source : (a)* http://julianvossandreae.com/works/protein-sculptures-indoor-works/, *(b)* http://www.umass.edu/microbio/rasmol/history.htm

culaires interactives permettant ainsi à l'utilisateur de conduire à travers une représentation physique de la molécule qu'il étudie [64].

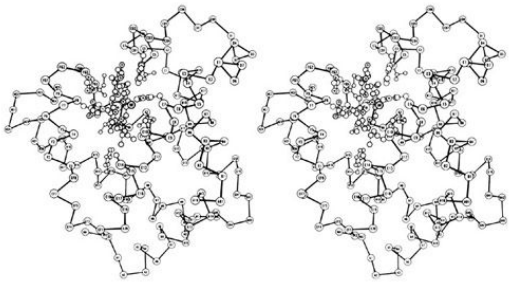
Représentations *in silico*

Les premiers modèles de structures moléculaires réalisés sur ordinateurs ne sont pas vraiment postérieurs aux premiers modèles physiques que nous avons cités. En effet, c'est en 1964 que Cyrus Levinthal et ses collègues du MIT ont développé un système qui permettait d'afficher, sur un oscilloscope, une représentation filaire d'une structure macromoléculaire qu'il était possible de tourner (voir Figure 1.26a). En 1965, Carroll K. Jonhson publia un programme créé en Fortran et appelé **ORTEP** (*Oak Ridge Thermal Ellipsoid Plot Program*) permettant de générer des illustrations de structures. Les cristallographes l'adoptèrent rapidement et s'en servirent pour publier leurs structures lors de conférences ou au sein d'articles de journaux. Sa capacité à générer des images stéréoscopiques des structures fut l'une des grandes forces d'ORTEP qui a vu sa seconde version être sortie en 1976 et sa troisième en 1996, cette dernière étant toujours disponible à l'utilisation [10]. ORTEP fut notamment utilisé pour illustrer la structure de la myoglobine et ces illustrations firent écho auprès de John C. Kendrew qui félicita personnellement Carroll K. Johnson pour son travail (voir Figure 1.26b).

À ce moment-là, les ordinateurs n'étaient pas encore capables d'interpréter seuls les coordonnées obtenues par cristallographie et il était nécessaire de construire un modèle de Kendrew, d'en mesurer les liaisons, avant de les rentrer dans l'ordinateur pour générer une représentation virtuelle.

10. http://web.ornl.gov/sci/ortep/

(a)

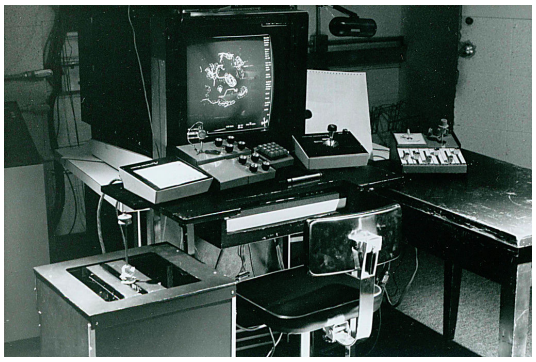


(b)

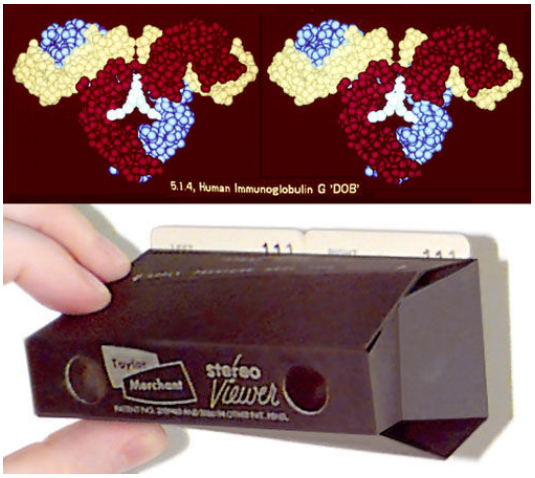
Figure 1.26 – *(a) Le premier système d'ordinateur pour visualiser des molécules développé par Cyrus Levinthal au MIT. (b) Vue stéréoscopique d'un dessin du squelette de la myoglobine générée par ORTEP.*
Source : (a) http://www.umass.edu/microbio/rasmol/history.htm*, (b) Source :* http://jbiocommunication.org/issues/31-2/feature2.html

Cela changea au milieu des années 70 quand, pour la première fois, une structure obtenue par cristallographie fut résolue et visualisée uniquement par des ordinateurs sans qu'un modèle physique existe. Pour ce faire, David et Jane Richardson utilisèrent un ordinateur permettant d'ajuster directement la densité électronique afin de générer un modèle virtuel de la structure d'une superoxyde dismutase [168]. Ce système informatique, appelé **GRIP** fut le prédécesseur de plusieurs programmes permettant ce passage d'une carte de densité électronique à une structure 3d moléculaire sur ordinateur (voir Figure 1.27a). Parmi ces programmes, en 1974 le laboratoire de biographie à l'université A&M du Texas développèrent un programme appelé **FIT** permettant de résoudre la structure de la nucléase *Staphylococcus* [40]. À la fin des années 70, le programme **FRODO** fut publié par T. Alwyn Jones et devint l'outil standard pour construire des modèles virtuels à partir de cartes de densité électronique [85]. Il est même estimé que 95% des structures cristallographiques résolues depuis le début des années 80 ont été construites en utilisant FRODO ou son successeur «O». Il permet une interaction de l'utilisateur avec le modèle en proposant de le tourner selon les trois axes de rotation X, Y et Z.

À la fin des années 70, de plus en plus de cristallographes franchirent le pas de la construction de modèles de nouvelles protéines résolues non plus par modèles physiques, mais grâce à des modèles par ordinateur. L'un des principaux avantages de ces derniers est la possibilité de garder en mémoire les coordonnées atomiques alors que lors de l'utilisation de modèles de Kendrew, elles devaient être mesurées à la main, atome par atome. En 1977, un atlas des structures macromoléculaires sous forme filaire fut publié. À la fin des années 70, Thomas K. Porter mit au point des algorithmes de représentations volumiques avec ombres. Cela révolutionna la représentation des molécules, mais n'était accessible qu'à un nombre limité de spécialistes possédant les plus puissants ordinateurs de l'époque. Le National Institutes of Health (NIH), qui employait Thomas K. Porter à l'époque, trouva cependant trop onéreux la publication d'un atlas de représentations volumiques de structures moléculaires en couleur. Leur idée de publication fut cependant de nouveau d'actualité quand ils prirent connaissance



(a)



(b)

FIGURE 1.27 – *(a) Système d'ordinateur GRIP créé par David et Jane Richardson. (b) En haut, exemple d'une des 116 vignettes présentes dans le set TAMS et permettant la vue stéréoscopique d'une immunoglobuline humaine grâce à un visualiseur de poche pliable, en bas.*
Source : (a) Jslipscomb - https://commons.wikimedia.org/wiki/File:Molecular-Graphics-GRIP-75-Console.jpg, (b) http://jbiocommunication.org/issues/31-2/feature2.html

de l'existence d'un lecteur de carte pour vignettes stéréoscopiques à moindre coût. Ce visualiseur en carton pouvait se plier et possédait donc un avantage certain pour la distribution de 116 vignettes, éditées par Richard Feldman grâce aux algorithmes de rendu volumique. Ce set de vignettes fut appelée TAMS pour *Teaching Aids for Macromolecular Structure*. Parmi les vignettes, plusieurs représentations étaient utilisées et plusieurs informations structurelles pouvaient être représentées, incluant un paragraphe d'explication associé. Un exemple de vignette est présenté dans la Figure 1.27b.

Jusqu'au début des années 90, l'utilisation de programmes de visualisation était exclusivement associée à des ordinateurs particuliers financièrement peu abordables et accessibles à une partie seulement des chercheurs. Cela changea en 1992 quand David et Jane Richardson décrivirent le format *kinemage* et leurs deux programmes associés **MAGE** et **PREKIN** [138]. Grâce à leur implémentation sur Macintosh, il fut le premier programme à pouvoir être utilisé par une large communauté de scientifiques, d'enseignants et d'étudiants. Le format *kinemage* permettait aux utilisateurs de générer une image animée de la structure ou sous-partie de structure qu'ils voulaient mettre en avant. Se rapprochant d'un GIF animé et bien qu'ayant disposé d'une grande popularité, en particulier en fournissant plus d'un millier d'illustrations pour les articles de *Protein Science*, il manquait à ce format la possibilité de diriger l'exploration d'une structure de façon neutre et interactive.

Ce manque fut comblé par RasMol en 1993. Ce programme fut la continuité d'un premier programme développé par Roger Sayle en 1989 et permettant le *shadowing* (ombrage en français) rapide d'objets virtuels permettant leur rotation en temps réel. Avec le mentorat d'un cristallographe, Andrew Coulson appliqua son algorithme sur des structures moléculaires et publia en 1992 une suite logicielle complète de visualisation moléculaire appelée **RasMol** possédant, entre autres, un nombre important de paramètres de rendus graphiques (rendu volumique, réflectivité des surfaces, etc.) et une interface graphique soignée [146]. Les sources

en C de ce programme furent rapidement mises à disposition de la communauté par Roger Sayle à l'obtention de son doctorat et put ainsi être adapté pour de multiples plateformes. Dés 1993 de nombreuses illustrations scientifiques s'appuyaient sur RasMol et les enseignants en ont rapidement fait également un outil de choix. Preuve de sa popularité, on estimait en 2005 le nombre d'utilisateurs à plus d'un million à travers le monde [113].

L'avènement du web fit apparaître quelques années plus tard un programme basé sur le portage du code C de RasMol vers C++. Développé par Bryan van Vliet et Tim Maffett, **Chime** (pour CHemical mIME) fut publié sous forme de plug-in pour le navigateur web Netscape. La possibilité d'associer cet espace de visualisation à des contenus web et de télécharger des modèles depuis des bases de données en ligne furent deux des principales raisons de son succès, en particulier dans le milieu académique et éducatif. Plusieurs centaines de tutoriels furent publiés sur le web en utilisant Chime et servirent de support à des étudiants, mais également à des cristallographes experts. Ses applications furent donc à la fois la communication de structures à travers le web, mais également la diffusion des connaissances et le support d'apprentissage.

1.2.3.2 La visualisation moléculaire contemporaine

Les programmes de visualisation moléculaire 3d permettent d'observer, dans un espace graphique 3d virtuel, des représentations abstraites de molécules à partir de fichiers de leurs coordonnées 3d. Ces coordonnées 3d sont communément inscrites dans des fichiers PDB (format provenant de la base de données *Protein Data Bank* fournissant un standard pour la représentation des données de structures macromolécules dérivées de RMN ou cristallographie [11]).

Les logiciels de visualisation permettent principalement de mettre en avant certaines des particularités structurelles des molécules au moyen de jeux de couleurs ou de formes plus ou moins adoptés dans la communauté. Il est compliqué de parler de standards sans un consensus officiel clair, n'existant pas actuellement, et plusieurs initiatives cherchent encore à faire évoluer les méthodes de représentation graphiques. Cependant, un certain accord s'est trouvé quant au jeu des couleurs et le schéma de couleur CPK (Corey-Pauling-Koltun) est certainement le schéma le plus utilisé [44].

Ces programmes s'appuient sur différents formats de fichier pour ces coordonnées même si la plupart sont capables de lire au moins le format PDB. Au-delà de la visualisation statique de structures 3d, il est souvent possible d'utiliser ces programmes afin de lire la trajectoire d'une simulation numérique. Ainsi, il est possible d'observer l'évolution structurelle d'un complexe moléculaire au cours du temps ou d'observer des phénomènes biologiques (passage de molécules au travers d'une membrane, repliement d'une région protéique suite à son interaction avec un partenaire biologique, etc.). Parmi les programmes de visualisation 3d largement utilisés aujourd'hui nous pouvons citer :

— **PyMol** [50], basé sur le langage Python et offrant une API pour ce langage, il dispose de nombreux rendus visuels et est, selon l'auteur original, utilisé dans plus d'un quart des images de structures moléculaires présentes dans les articles scientifiques. Il s'appuie sur la génération d'images de haute qualité grâce à des techniques de lancer de rayons. Son utilisation comme API est particulièrement adaptée au portage au sein de plateformes multi composantes et il est la base de plusieurs travaux de l'équipe pour la visualisation moléculaire dans différents environnements immersifs.

11. http://www.rcsb.org/pdb/static.do?p=file_formats/pdb/index.html

— **VMD** [82], spécialisé dans la visualisation et l'analyse de résultats de simulations de dynamiques moléculaires, cet outil est très usité du fait de ses nombreuses extensions et la possibilité de le relier à NAMD, un programme de dynamique moléculaire, permettant ainsi d'effectuer des simulations moléculaires interactives. Il est de plus bien adapté au rendu temps réel de larges modèles moléculaires grâce à des méthodes de rendus graphiques directement effectués et optimisés sur GPU.

— **Jmol** [76], basé sur Java et disponible sous forme d'application indépendante ou intégrée dans des pages web, ce visualiseur est populaire du fait de sa facile intégration au sein de contenus web. Il offre de plus plusieurs rendus graphiques très performants.

— **YASARA** [97], suite logicielle permettant à la fois la visualisation, mais aussi l'exécution de simulations moléculaires interactives, YASARA est depuis sa création dédiée pour le rendu stéréoscopique 3d et l'implication de l'utilisateur dans le suivi de sa simulation offrant même des solutions pour diriger la simulation en y ajoutant des forces.

— **SweetUnityMol** [129], est une extension de UnityMol [106], lui-même utilisé comme base d'une partie de nos travaux, et s'intéresse plus particulièrement aux représentations 3d des glucides et des polysaccharides. Il utilise différentes méthodes de représentation acceptées au sein de la communauté scientifique des glycosciences et propose de nouvelles techniques de rendus graphiques.

Cette liste est non-exhaustive, mais recense quatre des logiciels de visualisation moléculaire les plus utilisés dans le monde scientifique et plus particulièrement en biologie structurale.

Ces programmes sont donc tous capables de représenter des modèles 3d de protéines grâce aux standards de représentation mis au point à travers les années. Parmi ceux-ci, il est possible d'extraire 7 modes de représentations souvent utilisés par la communauté de biologie structurale, à la fois dans leur travail quotidien, mais également au travers de leurs communications (conférences, articles, etc.).

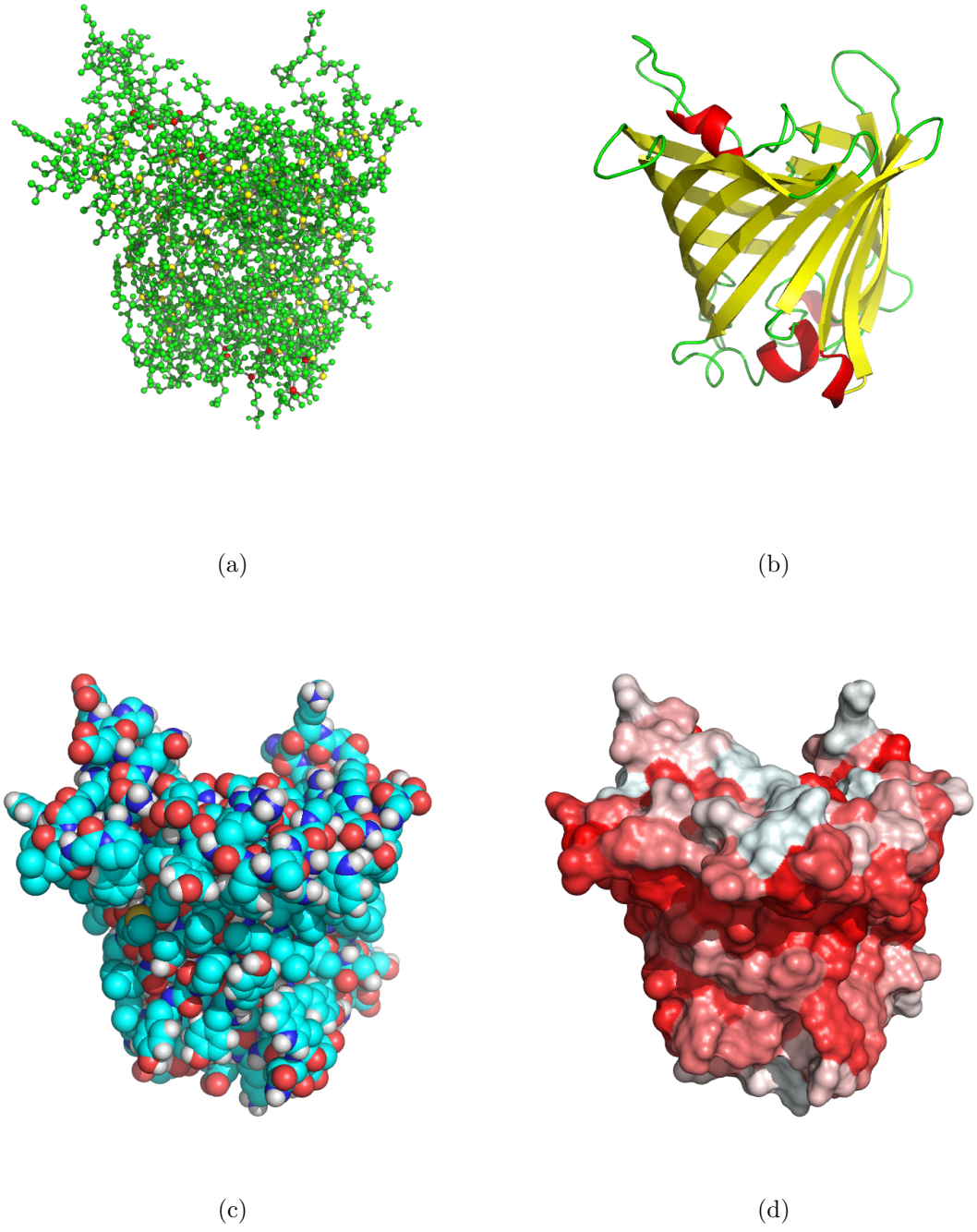
— Représentation par **ligne** : seules les liaisons covalentes entre atomes sont représentées par une ligne de longueur équivalente à la distance entre les deux centres des atomes situés de part et d'autre de la liaison. Cette représentation est souvent associée à une coloration précise, chacune de ses moitiés étant colorée de la couleur associée au type d'atome qu'elle relie.

— Représentation par **trace** : seule la chaîne principale de la protéine est représentée dans ce mode, par des bâtons droits dont les extrémités correspondent au centre d'un atome qu'elles relient.

— Représentation par **balles et bâtons** : alors que les atomes sont représentés par des sphères, les liaisons covalentes sont des bâtons. C'est l'une des représentations les plus courantes, car simple et permettant de rapidement cerner l'agencement d'atomes d'une même molécule. Elle est cependant plus adaptée aux molécules de tailles limitées, car elle engrange rapidement une surcharge visuelle pour les molécules de plus grande taille.

— Représentation par **rubans** : de la même manière que la représentation par trace, cette représentation suit les carbones alpha composant la chaîne principale de la protéine et affiche un ruban large le long de la chaîne principale.

— Représentation par ***cartoon***: ici, les différentes structurations composant la structure secondaire d'une protéine : hélices, feuillets et coudes, sont mises en avant respectivement par un ruban enroulé le long du squelette de la protéine, une flèche large pointant dans la direction du feuillet identifié et enfin un tube pour les coudes et boucles de la protéine rassemblant les parties non structurées.



(a) (b)

(c) (d)

Figure 1.28 – *Différentes représentations obtenues avec PyMol de la protéine OMPLA (pour Outer Membrane PhosphoLipase A) enzyme présente à la surface de certaines bactéries. Elle est successivement représentée : (a) par balles et bâtons, (b) cartoon, (c) sphères, et (d) représentation surfacique.*

— Représentation par **sphères** : chaque atome est représenté par une sphère dont le rayon est égal au rayon de Van der Waals de l'atome.

— Représentation par **surface** : dans ce mode, la SAS est représentée par une couche uniforme autour de la protéine.

Chaque type de représentation répond à des besoins de visualisation différents. Alors que l'agencement de la structure secondaire et tertiaire pourra être rapidement visualisé avec une représentation par *cartoon*, il sera nécessaire d'ajouter une représentation par lignes afin d'observer la proximité des atomes entre eux. De la même façon, la détection visuelle des trous et potentiels sites de liaisons de la protéine sera facilitée par la représentation surfacique, mais le cœur de la protéine ne pourra être visualisé correctement qu'avec des représentations par ligne ou rubans. Ces différentes représentations ne sont pas cloisonnées et il n'est pas rare de les combiner afin de visualiser les caractéristiques propres à chaque représentation en même temps. Les calculs permettant chacune de ces représentations ne demandent pas les mêmes ressources computationnelles et même si la majorité d'entre elles peuvent aujourd'hui être appliquées de façon presque instantanée sur des ordinateurs de bureau, certaines comme la représentation par surface demandent des calculs coûteux sur des protéines de grande taille.

La représentation de surface fait justement partie des techniques de visualisation ayant pu profiter des avancées rapides provenant de certains domaines non scientifiques comme le jeu vidéo ou le cinéma. Certains rendus appelés *photo-réalistes* et inspirés de techniques d'animation récentes sont maintenant retrouvés au sein de logiciels de visualisation de données scientifiques, améliorant sensiblement la perception des utilisateurs vis-à-vis de leurs données. La visualisation moléculaire n'a pas échappé à ces nouvelles techniques de rendus et a su implémenter certaines approches basées sur la **transparence**, les rendus **lumineux**, les effets de **profondeur**, etc. au sein des logiciels les plus utilisés dans la communauté. Ces nouvelles façons de représenter des structures protéiques permettent de jouer sur les capacités cognitives et perceptuelles des utilisateurs. Parallèlement, ces développements demandant des ressources computationnelles bien plus importantes que les anciennes implémentations graphiques, il a fallu également franchir le pas du calcul graphique sur GPU [34]. La parallélisation des programmes et leur utilisation des ressources graphiques présentes au sein des ordinateurs ont permis de conserver des performances d'affichages suffisantes pour afficher des molécules de taille raisonnable. Dans cette même dynamique, certains projets s'appuient sur l'utilisation de moteurs de jeux vidéos pour développer des environnements de visualisation profitant des derniers progrès du domaine, dont la spécialisation et l'investissement associé dépassent de loin les moyens attribués à la seule visualisation moléculaire dans le monde de la science [3, 105]. Ces moteurs de jeu ouvrent de nouvelles perspectives pour la visualisation, mais également l'exploration des structures de protéines.

1.2.3.3 Perspectives de la visualisation moléculaire

La présence de GPUs performants sur une grande partie du parc informatique scientifique permet d'envisager sérieusement leur utilisation intensive pour le développement d'outils de visualisation performants. La prochaine grande étape est la mise à l'échelle de ces outils pour répondre aux enjeux de taille, au sens propre comme au figuré, inhérents aux structures étudiées. Nous avons vu que ces dernières peuvent maintenant être composées de plusieurs millions d'atomes et donc occuper un espace d'affichage conséquent. Alors que la visualisation de quelques centaines d'atomes sur un écran d'ordinateur de bureau permet d'aborder la structure d'un point de vue détaillé tout en gardant une idée précise de sa structure générale et plus spécifiquement de la position de la caméra par rapport à la protéine affichée,

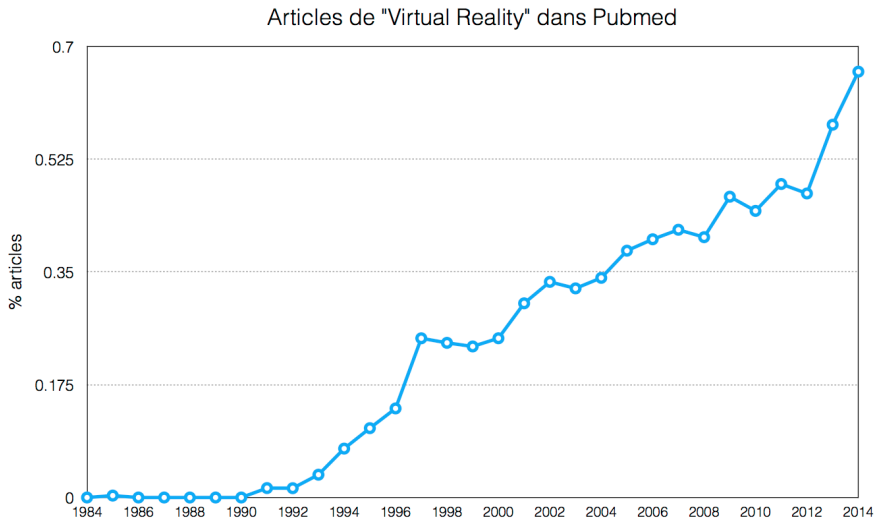


Figure 1.29 – *Évolution du pourcentage d'articles de PubMed où le terme «Virtual Reality» est retrouvé soit dans le titre soit dans le résumé au cours des 20 dernières années.*

les structures de plus grandes tailles nécessitent une résolution beaucoup plus importante et des moyens d'explorer ces structures de façon contrôlée et sans pertes de repères pour l'utilisateur. Or, parmi les différentes techniques de visualisation évoquées précédemment, si nous pouvons facilement constater les efforts effectués dans le rendu graphique des protéines, aucune contribution significative n'a été publiée pour les paradigmes de navigation et d'exploration depuis plusieurs années. Ces aspects sont, à l'opposé, centraux dans les sujets d'étude propres à la Réalité Virtuelle (RV), qui accorde une attention particulière à la place de l'utilisateur au sein des mondes virtuels avec lesquels il interagit [126, 92, 74].

Ce dernier point, couplé au rapprochement entre les méthodes provenant de disciplines spécialisées dans les rendus visuels de haute performance et l'application scientifique, peut expliquer la progressive arrivée de la visualisation moléculaire dans le monde de la RV. Il est d'ailleurs intéressant de constater dans la Figure 1.29 le nombre d'articles présents dans la base de données PubMed [12] et comportant le terme «Virtual Reality» dans le titre ou le résumé. Même si ces résultats bruts sont à prendre avec recul, l'évolution croissante de ce terme au sein d'articles biologiques et de médecine semble clairement indiquer un intérêt croissant de ces deux disciplines scientifiques pour la RV.

1.3 Conclusion

1.3.1 Perspectives et nouveaux usages de la biologie structurale

Par rapport au séquençage permettant d'obtenir la séquence de l'ADN d'un organisme de ses gènes et donc de ses séquences protéiques, la biologie structurale a pris du retard pour caractériser la structure tridimensionnelle des protéines dont la séquence a déjà pu être identifiée de façon systématique et fiable. Pour preuve l'écart entre le nombre de protéines dont la séquence est déposée dans UniProtKB [13] et le nombre de protéines possédant effectivement

12. http://www.ncbi.nlm.nih.gov/pubmed
13. http://www.uniprot.org/uniprot/

une structure déposée dans PDB [14]: Aujourd'hui, seul 1% des protéines séquencées possèdent une structure répertoriée dans la PDB.

Pour combler ce retard, les techniques expérimentales et computationnelles ont franchi des paliers, générant de grandes quantités de modèles d'une complexité en constante progression. En effet, la taille des structures moléculaires observées et étudiées augmente régulièrement et il est maintenant possible de simuler des systèmes de plusieurs millions d'atomes pendant des durées biologiques suffisamment longues permettant d'observer des phénomènes biologiques auparavant inaccessibles. La taille et la vitesse du flux de données généré par les outils de modélisation moléculaire dépassent désormais les capacités d'analyses des experts et des outils qu'ils utilisent. Même si les progrès portés sur la précision des techniques de résolution de structures de protéines ne cessent d'augmenter, une attention toute particulière est maintenant portée sur les étapes d'analyse et de visualisation de structures.

La visualisation de capsides virales, de ribosomes ou de complexes membranaires entiers demande des performances importantes de la part des logiciels de visualisation. Ces derniers ont donc dû s'adapter et évoluer pour répondre à ces nouveaux défis en terme de rendu visuel. Le calcul sur processeur graphique a permis de franchir une grande étape en terme de performance et plusieurs algorithmes de rendus graphiques largement utilisés au sein des logiciels de visualisation en ont profité [34]. En parallèle, de nouvelles techniques de rendus graphiques, inspirées pour certaines du monde du jeu vidéo, ont vu le jour afin de rapprocher les représentations de molécules de représentations d'objets réels, facilitant ainsi la perception des objets observés [105]. Les progrès effectués en terme de performance ont également permis de rendre la visualisation moléculaire interactive, accordant ainsi aux utilisateurs davantage de possibilités quant à leurs options de rendus et dépassant ainsi la simple image statique et précalculée de structure 3d.

Cependant, ces récentes avancées ne sont pas suffisantes pour appréhender la complexité des complexes moléculaires étudiés aujourd'hui. La limitation n'est pas dans l'efficacité du rendu, mais dans la façon d'explorer les structures 3d d'intérêt d'interagir avec elles.

1.3.2 Contributions

Il s'avère que la Réalité Virtuelle dispose d'une large gamme de techniques pour naviguer et interagir dans des scènes 3d complexes, comportant plusieurs millions de particules, dont certaines sont déjà appliquées pour faciliter notamment l'étape de modélisation et de simulation. Notre première contribution s'inscrit donc dans le développement d'outils de RV pour améliorer l'étape d'exploration et de visualisation des structures 3d de protéines. Nous verrons dans le chapitre suivant les actuelles limites du processus d'exploration de structures 3d de protéines dans des dispositifs immersifs en RV et quelles sont les pistes pour lever ces limites. L'étape suivante sera de répondre à la complexité de l'étape d'analyse dans le processus d'étude de structures 3d de protéines. Cette étape d'analyse, nous l'avons évoqué dans ce chapitre, ne peut être dissocier de l'étape d'exploration et il est donc nécessaire de les coupler afin de permettre à l'expert scientifique de profiter des informations fusionnées de chacune ces deux étapes dans un même processus de réflexion. Pour ce faire, il est possible de profiter de l'apport des interactions directes et naturelles que met en place la Réalité Virtuelle. Notre seconde contribution s'inspire de cela et est exposée dans un troisième chapitre. Nous verrons que ces interactions directes ont besoin de se baser sur une représentation précise du domaine d'application afin d'être mises en place.

14. http://www.rcsb.org/

Chapitre 2

La Réalité Virtuelle et la Biologie Moléculaire : usages, enjeux et perspectives

Nous mettons en avant dans ce chapitre les différentes caractéristiques de la Réalité Virtuelle à travers trois piliers qui la caractérisent : l'immersion, l'interaction et la navigation. Nous rapporterons dans une seconde partie les échos que ce domaine a trouvé en biologie structurale pour finalement identifier dans une dernière section les éléments précis à améliorer pour permettre l'utilisation de la Réalité Virtuelle en biologie structurale.

Sommaire

Les contributions de la Réalité Virtuelle (RV) pour résoudre des problématiques industrielles et scientifiques sont depuis quelques années de plus en plus nombreuses. Cet essor peut être expliqué par deux facteurs principaux : (1) la démocratisation des dispositifs de visualisation et d'interaction issus de la réalité virtuelle et augmentée et (2) l'apport de la 3d pour observer et manipuler des objets massifs complexes intrinsèquement tridimensionnels. Le besoin conjoint de mieux percevoir et de manipuler les objets tridimensionnels que constituent les structures moléculaires a abouti à une appropriation assez rapide de techniques de RV comme la stéréoscopie par la communauté de biologie moléculaire. Dans ce chapitre, après quelques définitions nous décrirons plus en détail les apports de la RV pour la biologie structurale, en particulier en terme de modélisation et de simulation moléculaire.

2.1 La Réalité Virtuelle

Plusieurs définitions de la Réalité Virtuelle ont été proposées depuis son émergence dans les années 90. Sherman et Craig définissent la RV comme le fait d'être immergé dans un monde virtuel interactif [153]. Brooks formule cette définition de façon presque similaire en disant que la RV est une expérience où l'utilisateur est efficacement immergé dans un monde virtuel réactif [28]. De façon légèrement différente, Burdea décrit la RV comme une simulation dans laquelle les graphismes générés par informatique sont utilisés pour créer un monde au rendu réaliste qui n'est pas statique, en répondant aux sollicitations de l'utilisateur [30]. On retrouve dans ces définitions les trois piliers qui définissent la RV selon Heim : Immersion, Interaction, Information [75]. Bien qu'il soit difficile d'extraire une définition simple et unique de la RV, l'idée principale est bien de mettre l'utilisateur au centre d'un environnement dynamique et réactif, créé artificiellement et qui viendra se supplanter au monde réel le temps de l'expérience. Cette définition est très proche de la définition de la RV que nous considérons au sein de l'équipe VENISE du LIMSI-CNRS [23]:

La Réalité Virtuelle vise à mettre au point des systèmes informatiques qui donnent à l'humain la capacité de percevoir et d'interagir de façon multi-sensori-motrice avec des données numériques ou mondes virtuels. Quand en plus, ces données numériques intègrent une virtualisation d'une partie de l'univers réel et permettent ainsi de gérer des interactions entre des objets réels et des objets virtuels, on parle alors de Réalité Augmentée, voire de Réalité Mixte.

2.1.1 Immersion

L'immersion se caractérise par la mise en place de techniques donnant à percevoir à l'utilisateur des retours sensoriels artificiels suffisamment réalistes et écologiques d'un point de vue perceptif pour lui donner l'illusion d'être immergé dans un monde virtuel. L'interaction constitue une dimension primordiale de ce réalisme, puisque le rendu visuel doit être dynamique en fonction de la direction du regard et de la position de l'utilisateur, condition minimum pour assurer la sensation d'immersion en mimant la perception écologique du monde réel.

Les informations présentées dans la scène virtuelle durant l'expérience immersive peuvent être adressées à plusieurs canaux sensoriels, majoritairement les canaux visuel et auditif, mais peuvent aussi mobiliser d'autre sens. De la richesse de ces retours et de la qualité d'intégration dépendront la qualité et le degré d'immersion de l'utilisateur. Bowman et McMahan notent qu'il n'est cependant pas nécessaire de gérer l'ensemble des sollicitations sensorielles d'un utilisateur pour assurer une sensation d'immersion [24]. Dans cette même étude, les auteurs

proposent un découpage de l'immersion en plusieurs composants, mettant en avant le fait qu'une immersion n'est ni complètement absente, ni parfaite, mais qu'elle peut être considérée au sein un continuum. L'immersion, concept essentiellement perceptif, induit chez l'utilisateur le sentiment d'être présent et incarné dans le monde virtuel, appelé sentiment de présence. Par opposition à l'immersion qui provient de la manipulation des sens de l'utilisateur avec une dimension essentiellement perceptive, la présence est un concept essentiellement cognitif puisqu'il s'agit d'un état de conscience dans lequel le sujet a le sentiment d'évoluer dans le monde virtuel et d'y être l'acteur. Bowman et McMahan mettent également en avant des applications de RV où le concept de présence de l'utilisateur n'est pas central. Parmi ces applications, la visualisation de données scientifiques, qui nous intéresse dans notre étude, met en effet davantage l'accent sur le contenu que sur la qualité d'immersion de l'utilisateur dans son monde virtuel.

Il est important de faire la distinction entre l'immersion perceptive et l'immersion cognitive. Alors que l'immersion perceptive est le résultat de la sollicitation sensorielle d'un individu pour qu'il se sente immergé, l'immersion cognitive se base principalement sur l'aspect psychologique de l'individu. L'immersion perceptive est donc plus directe et s'appuie sur un rapport étroit avec la réalité puisque des stimuli réels pourront être utilisés pour piéger les sens d'un individu en immersion. L'immersion cognitive s'intéressera par contre à l'imagination et à la représentation mentale de l'utilisateur qui provoque l'état d'immersion. Ces deux facettes de l'immersion sont présentes à des degrés différents suivant les applications, mais ne sont rarement ni complètement présentes ni complètement absentes. L'immersion cognitive tend à prendre une place de plus en plus importante au sein des nouvelles technologies. Les vidéos 360 degrés récemment rendues très facilement accessibles via la plateforme YouTube sont un exemple de moyen d'immersion à la fois cognitif et immersif. En effet, l'utilisateur manipule son smartphone doté d'un accéléromètre et/ou d'un gyroscope, qui fait office de fenêtre de visualisation sur le monde virtuel, dont il peut changer la position et l'orientation pour observer un monde virtuel tout autour de lui. Il y a alors construction d'une carte mentale de ce monde virtuel, à partir d'une perception parcellaire de ce monde virtuel à travers cette fenêtre dynamiquement contrôlée par le sujet.

2.1.1.1 Visuelle

Les interfaces visuelles sont de loin les plus utilisées pour l'immersion et bien que d'autres interfaces soient également utilisées dans des cas ponctuels, elles le sont rarement sans couplage à des interfaces visuelles. Ces dernières stimulent le sens de la vue de l'utilisateur et doivent parvenir à fournir un confort visuel le plus proche possible des conditions écologiques.

Stéréoscopie ou rendu 3d

On ne peut évoquer les interfaces visuelles immersives sans s'arrêter sur le concept de stéréoscopie, participant à la qualité de l'immersion. La stéréoscopie se caractérise par la génération de deux images distinctes, une pour chaque œil, avec un décalage de point de vue correspondant à l'image que verrait l'œil si le second était fermé. Lorsque chaque image est présentée à l'œil correspondant, le cerveau traite ces images afin de faire percevoir de la profondeur et du relief dans la scène observée. Ces images sont restituées par un écran ou un projecteur, mais doivent ensuite être séparées afin d'atteindre indépendamment les deux yeux. Cette séparation peut se faire soit de façon *active*, soit de façon *passive*. Lorsque la séparation est *active*, on occulte la vision des deux yeux de manière alternée et à haute fréquence. Une succession d'images alternant le point de vue de l'œil droit et le point de

vue de l'œil gauche est ainsi diffusée. Il est donc nécessaire d'afficher la bonne image pour le bon œil au moment où celui-ci n'est pas occulté, impliquant donc une forte synchronisation entre le générateur d'images et la lunette à occultation utilisée. Dans le cas d'une séparation *passive* des images, on utilise un filtre polarisant à la fois sur les écrans et les lunettes. Les deux images sont ici affichées en même temps, mais selon deux polarités différentes ce qui permet leur distinction par le filtre polarisé correspondant, un différent à chaque œil. Une polarité basée sur les couleurs donnera lieu par exemple à la diffusion d'une image rouge pour l'œil gauche, une image bleue pour l'œil droit, qui seront toutes deux superposées, le filtre bleu ou rouge des deux verres de lunettes permettant de faire la séparation.

Vision adaptative grâce au *tracking*

La **stéréoscopie adaptative** désigne la technique permettant de faire varier un rendu graphique 3d en fonction de la position et de l'orientation de l'utilisateur. Elle se base principalement sur la capture de cette position de façon instantanée et continue. Chaque nouvelle position de l'utilisateur déclenche ainsi un changement dans le contenu 3d affiché permettant de retranscrire le changement visuel attendu, conséquence du mouvement de l'utilisateur.

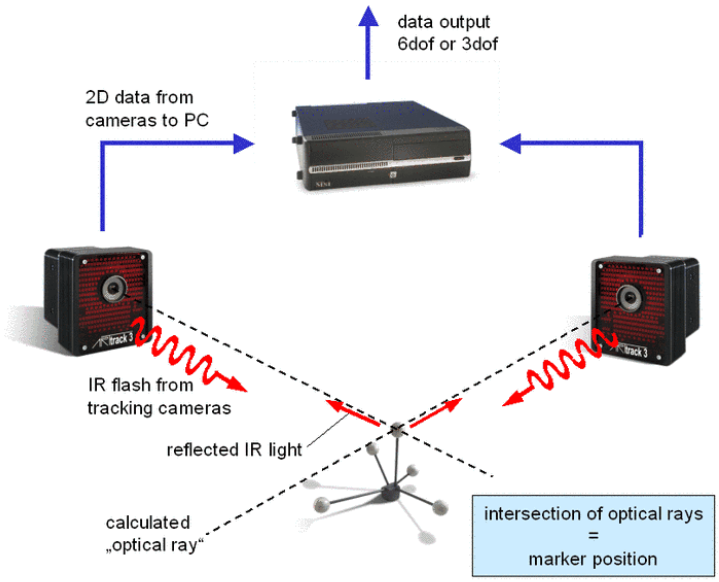


Figure 2.1 – *Schéma simplifié d'un système de tracking optique ART basé sur des signaux infrarouges lancés par des caméras et se reflétant sur des marqueurs spécifiques.*

Parmi les dispositifs permettant de récupérer les positions de la tête ou du corps de l'utilisateur, le **tracking optique** est l'un des plus usités. Il fonctionne au moyen de caméras infrarouges et de marqueurs réfléchissants (voir Figure 2.1). Des motifs de marqueurs vont servir de cibles pour les caméras infrarouges qui vont opérer une triangulation afin d'en extraire leur position. Chaque motif différent peut être associé à un objet particulier, une partie du corps ou un dispositif d'interaction. Le suivi de la tête est particulièrement utile dans les environnements immersifs puisqu'il permet de connaître la direction exacte du regard de l'utilisateur au sein du monde virtuel. Grâce à cela, les ordinateurs responsables du rendu graphique peuvent adapter les images affichées sur les écrans pour qu'elle corresponde au

point de vue de l'utilisateur dans le monde virtuel.

Le tracking optique demande la mise en place d'un système encombrant et spécifique, pour des conditions de luminosité et d'espace de travail limitées, et n'est parfois donc pas envisageable. Il est alors possible d'utiliser la vision adaptative grâce à des instruments de mesure embarqués dans un dispositif mobile porté par l'utilisateur. Ces derniers ne permettent pas un suivi absolu de la position de l'utilisateur dans l'espace, mais un suivi relatif de l'orientation de son regard. Les **gyroscopes**, présents dans l'intégralité des smartphones par exemple, permettent cette interprétation et, à la manière du tracking optique, envoient à l'application les informations d'orientation afin que le rendu graphique soit adapté.

Ces deux approches demandent une certaine efficacité dans la rapidité d'interprétation des positions/orientations de l'utilisateur et du calcul graphique 3d afin de réduire au maximum le temps de latence entre un mouvement et son effet sur le rendu de la scène.

Nous avons vu les principales techniques permettant de générer de la stéréoscopie 3d, statique ou adaptative, intéressons-nous maintenant aux dispositifs permettant de diffuser ces images.

Dispositifs

Il est difficile de décrire la RV sans décrire les dispositifs qui lui servent de support. Conçus pour solliciter l'ensemble des canaux sensori-moteurs de l'homme, ces dispositifs doivent permettre à l'utilisateur de percevoir le monde virtuel dans lequel il évolue de la manière la plus nette et claire possible afin qu'il y interagisse efficacement. Afin de répondre à chacun des prérequis que l'immersion génère, il est possible de distinguer 3 types d'interfaces comportementales venant exploiter la motricité ou les perceptions de l'homme issues de son comportement dans le monde réel. Nous retrouvons : les *interfaces sensorielles* qui vont permettre à l'homme de percevoir le monde virtuel, les *interfaces motrices* qui vont permettre à l'homme de se déplacer et d'agir dans le monde virtuel et enfin les *interfaces sensori-motrices* qui vont permettre une communication bidirectionnelle entre l'homme et le monde virtuel. Nous ne donnerons pas une liste exhaustive des dispositifs de RV, que l'on peut retrouver dans le Traité de la Réalité Virtuelle - Volume 2. [60].

Figure 2.2 – *Exemple de WILDER, mur d'écrans de 6 m sur 2 composé de 75 écrans, présent au sein du bâtiment Digiteo sur le campus de l'université Paris-Sud.*

Parmi les dispositifs fixes utilisés en RV, on retrouve les murs d'écrans (cf. Figure 2.2).

Évolution naturelle des écrans d'ordinateur classiques, ce sont de larges surfaces composées de plusieurs écrans aux bords fins collés les uns aux autres. Les murs d'écrans peuvent être de différentes natures, même si tous les écrans d'un même mur sont identiques, ils peuvent être rétro projetés ou non, 2d ou 3d, tactiles ou non et de résolutions différentes. Ils permettent l'affichage de larges images à de hautes résolutions et sont souvent utilisés pour les études scientifiques mettant en jeu de nombreuses données à des échelles très différentes. Ils permettent également de diffuser des contenus à une audience élargie grâce à leur surface d'affichage étendue.

L'immersion peut atteindre un niveau supérieur lorsque sont utilisés des environnements immersifs de type CAVE [46] comme illustré dans la Figure 2.3. Le système CAVE consiste en une série de projecteurs permettant la projection d'images sur 3 à 6 faces d'un cube de la taille d'une pièce. Les projecteurs peuvent être remplacés par des écrans plats dans certains cas, mais cette configuration est plus rare, car bien plus coûteuse à espace d'affichage similaire. Bien qu'il n'existe pas de tailles limites inférieures ou supérieures lorsqu'on évoque la taille des CAVEs, il est commun de pouvoir s'y tenir debout et donc d'y évoluer. Les projecteurs sont usuellement situés derrière les écrans afin de ne subir aucune entrave dans le rayon de projection. Le contenu affiché est 3d et de haute résolution. Les utilisateurs portent des lunettes 3d pour percevoir la stéréoscopie et leurs mouvements sont généralement capturés par un système de tracking optique (voir section 2.1.1.1). Il est possible dans un CAVE d'afficher des objets virtuels flottants autour desquels l'utilisateur pourra se déplacer.

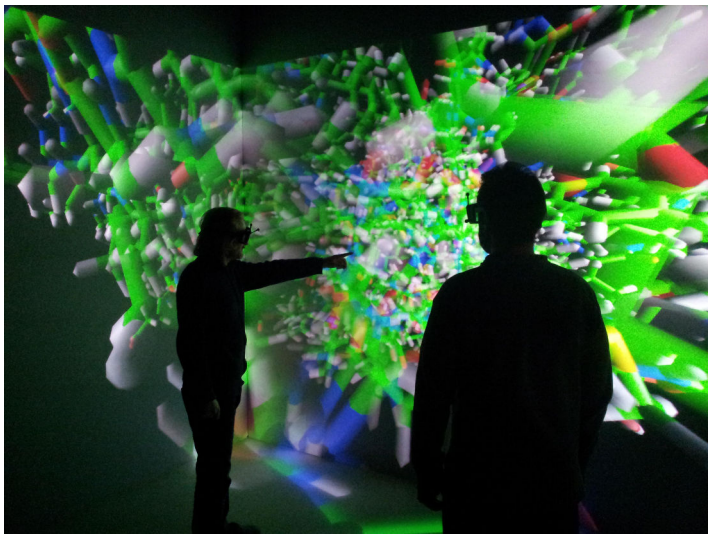


Figure 2.3 – *Exemple de EVE, système CAVE composé de 4 écrans (gauche, face, droite et sol), présent au LIMSI-CNRS sur le campus de l'université Paris-Sud.*

Des systèmes audio 3d peuvent être associés aux CAVEs. Ces systèmes audio profitent du large espace de la pièce pour disposer de multiples enceintes à l'image des systèmes WFS détaillés dans la section 2.1.1.2. La taille des CAVEs donne la possibilité de se déplacer dans un monde virtuel de la taille d'une pièce simplement en marchant au sein du dispositif et sans paradigme de navigation spécifique. L'intégration de dispositifs d'interaction est également facilitée par une absence de contrainte de taille. Certaines CAVEs utilisent plus d'un projecteur par écran leur permettant d'afficher des images pour deux points de vue distincts, en 3d et simultanément, c'est le cas d'EVE [1] (Evolutive Virtuel Environments) du groupe VENISE

1. http://www.limsi.fr/venise/EVEsystem

(a) (b)

FIGURE 2.4 – *(a) Casque de réalité virtuelle, l'Oculus Rift, appartenant à Facebook et constituant l'un des premiers casques de RV ayant été disponible sur le marché. (b) Capture d'écran du logiciel PyMol exécuté en mode stéréoscopique «cross-eyes»: l'image de la protéine de droite correspond à l'œil gauche et inversement.*

du LIMSI-CNRS. La multi stéréoscopie permet de mettre en place des scénarios collaboratifs entre les utilisateurs partageant la même scène virtuelle, mais selon leur propre point de vue.

Les visiocasques (appelés également casques immersifs ou casques de RV ou encore casques HMD for *Head-Mounted Display* en anglais) sont des dispositifs se portant à la manière de casques standards et composés de deux écrans situés en face des yeux de l'utilisateur (cf. Figure 2.4b). Grâce à un système de lentilles, l'utilisateur est capable de distinguer les images affichées sur chacun des écrans de façon nette. L'affichage peut être soit monoculaire et donc perçu en 2d, soit binoculaire et donc perçu en 3d. Dans ce dernier mode, le contenu affiché sur les écrans est découpé de telle sorte que chaque œil perçoit une image différence correspondant au point de vue qui devrait être le sien dans la réalité comme illustrée dans la Figure 2.4b. Le contenu affiché évolue suivant l'orientation de la tête de l'utilisateur. Cela est rendu possible par la présence d'un gyroscope calculant l'orientation relative de la tête par rapport à un point d'origine calibré au début de l'expérience. Bien que les premiers visiocasques binoculaires commercialisés datent de 1995, le domaine des casques immersifs a connu depuis quelques années un intérêt significatif de la part du monde de la recherche et du jeu vidéo. Cet intérêt peut s'expliquer par la résolution atteinte par les écrans de petite taille (type smartphone) et la précision des capteurs d'orientation utilisés pour adapter l'image à l'orientation de l'utilisateur. Certains acteurs de la RV ont d'ailleurs profité de la démocratisation du smartphone et de leurs capacités d'affichage de plus en plus performantes pour proposer des supports type casque permettant de positionner un smartphone devant les yeux et de rendre une image stéréoscopique grâce à deux lentilles intégrées au support. L'exemple le plus connu est le *cardboard* de Google permettant de mettre un smartphone d'environ 5 pouces de diagonale au sein d'un support en carton pliable comportant un jeu de lentilles optiques.

Du point de vue du coût, ces dispositifs sont très peu onéreux comparés aux systèmes fixes. Cependant, l'un de leurs inconvénients est le fait qu'ils dissocient l'utilisateur du monde extérieur, rendant difficile le travail collaboratif autour d'un objet virtuel partagé, possible dans les CAVE ou les murs d'écrans.

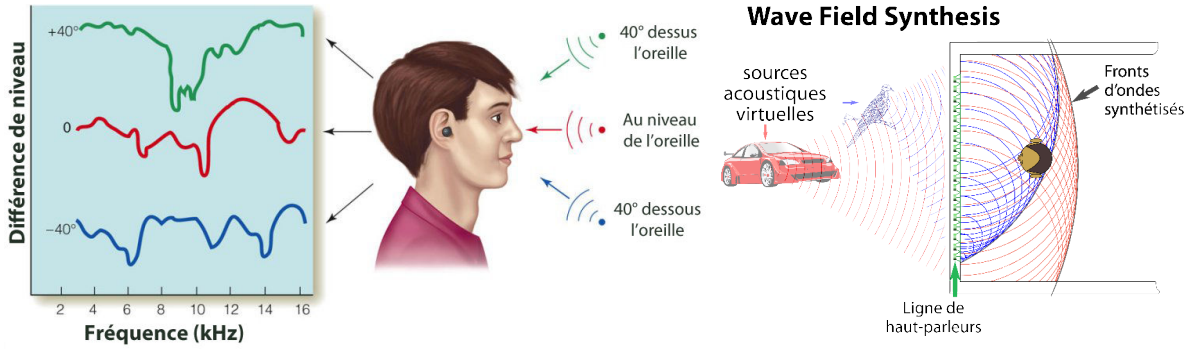
2.1.1.2 Auditive

Bien que l'aspect visuel de l'immersion soit considéré comme central pour immerger un individu dans une scène virtuelle, l' immersion visuelle doit être combinée à l'immersion auditive afin d'atteindre un degré d'immersion qui se rapproche des conditions écologiques, en particulier dans les scènes réalistes, dans lesquelles un environnement sonore est souvent attendu.

Parmi les méthodes d'immersion auditives, appelée aussi spatialisation sonore, on peut identifier deux méthodes principales :

Le **son binaural** est le fruit de l'application de filtres sur une source sonore qui une fois restituée permet au cerveau de traiter le son plus seulement de façon qualitative, mais également spatiale. Cette technique s'inspire des traitements habituels que fait le cerveau pour identifier la localisation d'une source sonore. Suivant plusieurs paramètres morphologiques d'un individu, le cerveau peut identifier de manière fiable la direction d'un son perçu. La position d'un son émis contrôlé grâce à l'application d'un filtre, appelé *Head-Related Transfer Function* (voir Figure 2.5a). Cette technique est être associée à des techniques de tracking de positon afin de rendre de façon fidèle la complexité sonore tridimensionnelle d'une scène virtuelle. La difficulté d'application de cette technique provient du fait que le filtre HRTF varie suivant les morphologies crâniennes des sujets et est donc propre à chaque individu.

La technique de ***Wave Field Synthesis*** désigne un procédé permettant de capter ou de synthétiser une scène sonore, par analogie aux scènes visuelles, en préservant les informations spatiales de distance et de direction des sources sonores (cf. Figure 2.5b). L'idée est d'émettre des vagues d'ondes synthétisées par un ensemble de haut-parleurs. À la différence des systèmes audio classiques types stéréo ou 5.1, la WFS reconstruit un champ sonore où il n'est pas nécessaire de rester au centre du dispositif pour garder une perception cohérente de la localisation des sources. Cette technique nécessite un nombre important de haut-parleurs spécifiques à cette technique qui est sensible à l'acoustique de la salle. C'est donc un dispositif assez coûteux et qui est retrouvé généralement associé à des dispositifs visuels de grande taille.



(a) (b)

Figure 2.5 – *(a) Illustration de la différence de fréquence et de niveau des ondes sonores interprétées par le cerveau suivant la localisation de leurs sources. (b) Schéma simplifié de la technique de Wave Field Synthesis.*
Source : (a) 2007 Thomson Higher Education, (b) Helmut Oellers - https://commons.wikimedia.org/wiki/File:Principle_wfs_2.svg

2.1.2 Interaction

Nous avons vu que les moyens dédiés à l'immersion sont nombreux et permettent de provoquer susciter chez le sujet au coeur de la scène virtuelle, faisant de lui un observateur actif autour des objets explorés dans le cas de la stéréoscopie adaptative.

Pour que le sujet puisse devenir l'acteur de ce monde virtuel, il faut lui fournir la possibilité d'interagir avec l'environnement virtuel ce qui assure par ailleurs une plus grande implication de l'utilisateur [163]. Les techniques et méthodes d'interaction n'ont pas connu la même évolution rapide que les systèmes d'immersion et il est encore difficile de mettre en avant des techniques d'interactions opérationnelles dans tous les Environnements Virtuels (EV) et pour tous les contenus et tâches virtuels.

2.1.2.1 Périphériques de tracking pour l'interaction

L'utilisation de périphériques de tracking pour interagir avec un EV immersif est la solution la plus communément utilisée au sein des EV immersifs. À la différence des moyens d'interaction en conditions classiques de bureau, les techniques de tracking sont adaptées à des conditions d'utilisation en position debout et dans des conditions d'obscurité. À l'aide de ces techniques, la souris des stations de travail a fait place à des périphériques tels que la souris 3d, appelée aussi *Flystick* (cf Figure 2.6). Le *Flystick* se caractérise par un périphérique à une main, composé de boutons et dont les contrôles directionnels sont déportés sur un mini-joystick ou une *tracking ball* manipulables au pouce. La manette reste un dispositif très utilisé du fait de sa courte courbe d'apprentissage due à sa démocratisation et également du fait de son nombre de commandes possibles important via ses nombreux boutons et combinaisons de bouton, mais est plutôt utilisée comme dispositif de contrôle de navigation, car ce dispositif n'est pas situé dans l'espace.

On retrouve également des dispositifs spécialisés reprenant la forme des outils qui sont censés être utilisés dans le monde virtuel permettant plus de réalisme dans l'interaction. Ces outils sont aussi suivis par des dispositifs de tracking optique permettant de les intégrer à travers leurs avatars dans les scènes virtuelles. Ces périphériques sont particulièrement utilisés dans des situations d'apprentissage à l'aide d'environnements virtuels.

Figure 2.6 – *Le Flystick2 d'ART, ou souris 3d, permettant d'interagir avec un environnement 3d en position fixe ou mobile.*

2.1.2.2 Interfaces sensori-motrices

Ces interfaces regroupent les dispositifs permettant à la fois d'interagir avec l'environnement, mais également de recevoir un stimulus sensoriel en réponse. La plus commune des interfaces sensori-motrices est l'interface haptique qui va permettre à l'utilisateur de ressentir l'effet de son interaction.

Les bras à retour d'efforts permettent par exemple de manipuler des objets en 3d tout en ressentant leur poids ou d'éventuelles collisions avec d'autres objets. Plus généralement, les systèmes à retour d'effort impliquent une plus grande précision des gestes et des manipulations de l'utilisateur.

Certains gants haptiques disposent de moteurs répartis sur l'ensemble de la main et des doigts permettant de faire ressentir à un utilisateur un objet particulier comme s'il l'attrapait réellement.

Le domaine médical et plus généralement des opérations télé robotisées profitent de ces dispositifs afin de garantir un ressenti proprio kinesthésique et tactile complémentaire à l'information visuelle. Ils sont courants en conception ou simulation de pilotages, car tentant de reproduire fidèlement les conditions réelles. Dans les sciences, ils sont utilisés pour guider l'utilisateur à l'aide de retour de force lors des manipulations de systèmes moléculaires, afin de lui faire percevoir les forces en jeu dans la simulation du phénomène observé.

2.1.2.3 Interactions gestuelles

Afin d'augmenter la sensation d'immersion de l'utilisateur et améliorer son expérience, il existe des interactions dites gestuelles qui vont chercher à s'inspirer des techniques. Les interactions gestuelles, regroupant les interactions mettant en jeu des mouvements de l'utilisateur ou des commandes vocales pour interagir avec son environnement virtuel, font opposition aux interactions indirectes comme peuvent l'être le clavier, la souris ou les dispositifs d'interaction immersifs comme la souris 3d ou les manettes évoquées précédemment. Ces interactions induisent une présence moins importante des menus, peu compatible avec l'immersion. Il faut cependant noter que ces interactions ont souvent une courbe d'apprentissage plus longue que les interactions indirectes, avec une robustesse plus faible. Elles sont aussi moins contraignantes pour l'utilisateur puisqu'elles ne se basent pas sur une charge matérielle supplémentaire dans les mains de l'utilisateur autre que des supports de marqueurs pour la capture de gestes [21] et un micro, déporté ou non, pour la reconnaissance vocale.

La combinaison des différentes méthodes d'interaction que nous avons vue est un domaine de recherche à part entière [111, 109]. On parle de **multimodalité** lorsqu'on propose aux utilisateurs des interactions au travers de leurs différents canaux sensori-moteurs. Cette combinaison va permettre de gérer des actions complexes et complémentaires réduisant ainsi le besoin de simplification des tâches entre les versions non immersive et immersive d'un programme.

2.1.3 Navigation

La navigation dans un monde virtuel permet d'étendre les limites de l'immersion au-delà des contraintes physiques imposées par les dispositifs immersifs. C'est ce qui la distingue de la vision adaptative qui se restreint à ces limites physiques.

2.1.3.1 Définition

La navigation est un concept qui nécessite une clarification terminologique, car comme le constatent Darken et Peterson, spécialistes en cognition spatiale, il existe une certaine confusion dans la littérature sur les termes employés [47]. Ainsi, certains auteurs emploient comme synonyme les termes de « navigation », « déplacement » ou encore « exploration ». Ces trois termes sont utilisés tout trois dans notre étude d'où l'importance de définir clairement ces termes. La navigation dans des environnements réels ou virtuels se définit à la fois par des composantes **motrices** et à la fois par des composantes **cognitives** [25].

- La composante motrice définit le mouvement/déplacement réel d'un utilisateur dans l'espace. Il existe plusieurs modes ou techniques de déplacement. Bowman (2002) les classe en deux catégories, selon la métaphore employée pour se déplacer dans l'espace virtuel :
  - Les métaphores réelles qui font appel à des comportements réalistes et/ou naturels comme le déplacement en mode « marche », « vol », « à vélo » ou encore « en conduisant ».
  - Les métaphores « virtuelles » où les chercheurs utilisent le potentiel du virtuel pour imaginer et créer des métaphores pour permettre le déplacement dans l'espace.
- La composante cognitive ou « wayfinding » est un processus cognitif de définition d'un chemin à travers un environnement. Le « wayfinding » a pour principal objectif de permettre au sujet de se construire une « carte cognitive » de l'espace exploré.

Dans un environnement virtuel, les utilisateurs peuvent disposer de plusieurs objectifs lors de l'activité de navigation. Trois tâches de navigation ont été identifiées par Darken et Sibert [48] de façon générique alors que Van dam et al. [174] ainsi que Bowman [25] en proposent une quatrième prenant tout son sens dans le cadre de la navigation pour la visualisation scientifique :

1. L'**exploration** c'est-à-dire une navigation sans cible explicite à atteindre. Le but étant uniquement de connaître et comprendre le nouvel environnement exploré. L'exploration peut aussi être psychologiquement active si le sujet doit suivre des indications. Dans le cas contraire, l'exploration est dite psychologiquement passive.
2. La **recherche d'une cible inconnue** où le sujet cherche une cible/destination particulière, mais ne connaît pas la position de celle-ci.
3. La **recherche d'une cible dont la position est connue** dans lequel La tâche/objectif est de retrouver une cible.
4. La **manoeuvre** consiste en des mouvements courts permettant à l'utilisateur de se positionner et de s'orienter.

La dissociation entre les distances pouvant être parcourues dans un monde virtuel et les contraintes dimensionnelles des dispositifs immersifs a rapidement obligé les experts en RV de mettre au point des méthodes de navigation répondant à cette problématique. Ces méthodes ne répondent pas seulement au besoin de changement d'échelle entre l'espace réel d'interaction et l'espace virtuel de navigation, elles doivent également prendre en compte les effets induits par l'activité de navigation, notamment la une perte de repères spatiaux.

Cette perte de repères spatiaux n'est pas le seul fait de paradigmes de navigation, elle sera également accentuée par le caractère abstrait et donc non écologique des données observées. Alors que la navigation au sein d'une ville ou d'une pièce peut permettre de garder pour le sujet une conscience de sa position et de son orientation suffisante, la navigation dans une

scène représentant des informations abstraites et/ou non orientées diminuera cette capacité à savoir se situer par rapport au contenu et au monde virtuel. Les degrés de liberté de l'utilisateur pour naviguer dans sa scène virtuelle sont également un facteur pouvant compromettre sa bonne conscience spatiale. Alors que les scènes réalistes vont souvent induire une navigation avec 2 degrés de liberté parallèlement au sol virtuel, les scènes possédant des données abstraites nécessitent une navigation en 3 dimensions dans l'ensemble du volume composant la scène virtuelle.

La réduction ou l'absence de repères spatiaux n'a pas seulement une conséquence sur le fait qu'un utilisateur puisse se sentir perdu au milieu d'une scène virtuelle. Elles peuvent également déclencher ou favoriser l'apparition d'un sentiment de malaise, communément appelé *cybersickness*, ou mal du simulateur.

2.1.3.2 Mal du simulateur ou *cybersickness*

Le *cybersickness* peut s'apparenter au mal des transports, transposé aux mondes virtuels, et se caractérisant par plusieurs effets indésirables et désagréables pour l'utilisateur. En plus de simples sensations d'inconfort, on retrouve comme symptômes de la fatigue excessive, des vertiges, des maux de tête ou des nausées, dégradant de manière importante la qualité de l'expérience de l'utilisateur [96, 101]. Le mal du simulateur est susceptible de diminuer l'efficacité de l'utilisateur lors de l'exécution de ses tâches et induit également un phénomène de rejet vis-à-vis du dispositif immersif. Ce phénomène fut donc étudié de près afin d'en identifier les causes et d'en trouver des solutions. Les causes principales mises en évidence lors des expériences de RV menées dans le but d'induire ce phénomène passent presque exclusivement par la dissociation des canaux perceptifs du corps humain. Le découplage et l'incohérence des informations fournies par canal visuel avec celles du système vestibulaire est notamment identifié comme étant une cause probable du *cybersickness* [137].

Ainsi, une scène virtuelle impliquant un déplacement non contrôlé de l'utilisateur et dont les paramètres de vitesse, d'accélération et de rotations ne sont pas contrôlés, entraîne dans beaucoup de cas le mal du simulateur. Ces situations de déplacements non contrôlés par l'utilisateur sont également identifiées comme responsables du mal des transports. L'absence d'implication d'un usager durant la navigation, peut également être un facteur déclencheur du *cybersickness*. Ce phénomène a été observé lors du visionnage de film ou de contenu vidéo impliquant des déplacements non contrôlés par le sujet même sur un rendu non stéréoscopique. L'immersion sans contrôle amplifie le phénomène de *cybersickness*, effet observé lors du visionnage du film «The Walk» [2], sorti au cinéma en format 3D en septembre 2015, un nombre très significatif de personnes ayant rapporté souffrir de nausées et de vomissements [3]. Il n'est cependant pas envisageable de réduire l'immersion afin de réduire le *cybersickness*, il est donc important de jouer sur les facteurs impliqués dans le *cybersickness* est proposant des paradigmes de navigation qui permettent de minimiser le mal du simulateur.

Nous décrivons ci-dessous les différents paradigmes de navigation, incluant ceux basés sur les technologies de *tracking* et cherchant à répondre aux besoins d'exploration des mondes virtuels en prenant en compte les facteurs inhérents au phénomène de mal du simulateur.

2.1.3.3 Navigation au sein de scènes virtuelles réalistes

Il existe plusieurs degrés de contrôle de la navigation dans des environnements virtuels 3d, allant d'un contrôle absolu de l'utilisateur sur ses déplacements au sein de la scène virtuelle

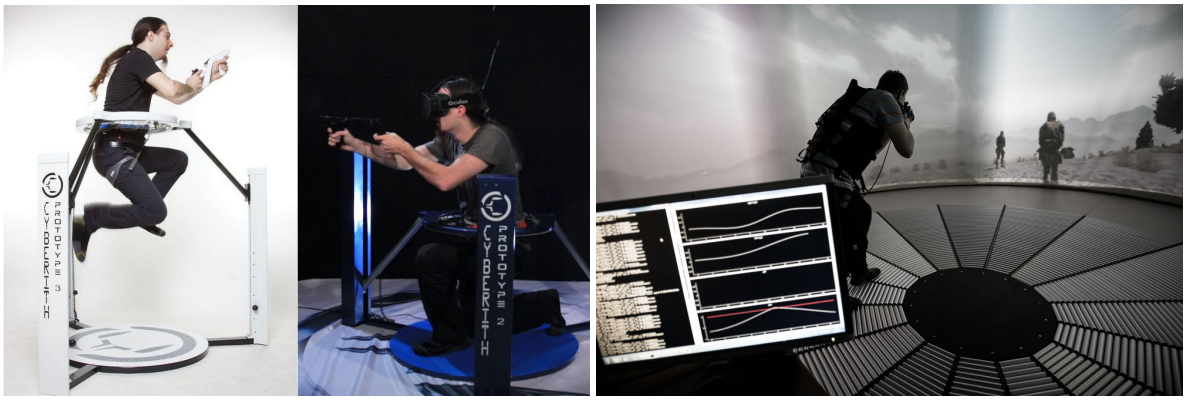
2. https://en.wikipedia.org/wiki/The_Walk_%282015_film%29
3. http://www.theguardian.com/film/2015/sep/30/robert-zemeckis-3d-the-walk-audiences-vertigo

explorée à un calcul logiciel automatisé des chemins de navigation imposés à l'utilisateur en fonction du contenu virtuel.

Navigation libre

On regroupe dans ces méthodes les approches permettant à l'utilisateur de diriger ses déplacements au sein d'un EV. Ce contrôle peut passer par des techniques s'inspirant du contexte écologique, comme dans le cadre de technique mimant la marche, ou de techniques non écologiques caractérisées par des métaphores utilisant des dispositifs d'interaction tels que des manettes, des joysticks ou des souris 3d.

Pour mimer les conditions écologiques, on peut citer les **tapis roulants multidirectionnels** qui sont des dispositifs mécaniques permettant de simuler plusieurs actions physiques du monde réel, dont la marche, la course, le fait de s'accroupir, et parfois même de s'asseoir (voir Figure 2.7a). Leur fonctionnement est similaire aux tapis roulants standards, mais la grande différence provient de leur capacité à bouger dans toutes les directions, offrant ainsi à l'utilisateur une possibilité de déplacement dans toutes les directions en restant dans un espace de travail physique contraint. Cette technique permet également de libérer les mains de l'utilisateur de tout périphérique dédié à la navigation. Enfin, l'implication physique de l'utilisateur permet de réduire le phénomène de *cybersickness* comme évoqué dans la section 2.1.3.2. Ces périphériques se couplent aussi bien à des dispositifs immersifs type CAVE qu'à des dispositifs mobiles comme les HMD.



(a) (b)

Figure 2.7 – *(a) Tapis roulant multidirectionnel statique (Cyberith Virtualizer) permettant de reconnaître, en plus de la marche et la course, des mouvements verticaux comme le saut ou l'accroupissement. (b) Premier tapis roulant multidirectionnel large et statique (Omnifinity Omnideck), utilisé en couplage avec une technologie de tracking optique afin de garantir un suivi précis des mouvements de l'utilisateur.*

À l'opposé des paradigmes de navigation les plus écologiques, les techniques de navigation très bon marché et les plus communément usitées, basées sur des dispositifs d'interaction tels que des manettes et des joysticks, inspirées des jeux vidéos, ne nécessitent pas une véritable implication physique du sujet. Cependant l'implication physique est un facteur permettant de diminuer le *cybersickness*. Pourtant, dans le cadre de casques virtuels, les dispositifs d'interaction, utilisables sans que l'utilisateur ait besoin de percevoir visuellement ses mains, sont pour l'instant privilégiés.

De manière intermédiaire, d'autres techniques basées sur le *tracking* nécessitent une implication du système vestibulaire de l'utilisateur. Cette implication étant un facteur permettant la réduction du *cybersickness*, ces techniques constituent un bon compromis entre les techniques de navigation très réalistes, mais très coûteuses, et les techniques qui ne requièrent pas d'implication physique. Des paradigmes ont découlé du tracking de l'utilisateur au sein de son environnement immersif, ils sont décrits dans la section 2.1.3.3

Navigation automatique ou semi-assistée

La navigation automatique dans une scène virtuelle peut s'apparenter à une navigation dans un véhicule sur lequel l'utilisateur n'aurait aucun contrôle [72]. Cette navigation complètement automatique, où les déplacements de l'utilisateur seront dirigés par le programme, peut se faire selon deux méthodes principales :

— Méthodes de ***path finding*** : ces méthodes demandent la définition de points de passage dans la scène virtuelle. Ces points de passage, considérés comme les meilleurs points de vue, peuvent être définis manuellement ou automatiquement. En cas de définition automatique, on se servira de la nature de la scène à explorer afin de les définir. Plusieurs méthodes existent pour trouver les points singuliers de la scène. Certaines méthodes se basent sur des analyses de l'entropie de la scène et cherchent les points de vue permettant de visualiser un maximum d'objets 3D en même temps[175]. Il est également possible de se baser sur les informations lumineuses afin d'extraire les meilleurs points de vue . Dans ce cas-là, ce seront les points de vue maximisant l'illumination de la scène qui seront retenus et constitueront les points de passage de la caméra pendant l'exploration automatique de la scène [69].

— Méthodes de contrôle **basées sur les images**: il est ici question de déplacer la caméra en optimisant une fonction de coût dont les paramètres sont définis en fonction des propriétés des images. Ces approches sont particulièrement appropriées afin de suivre les modifications et des singularités de la scène, comme un objet en mouvement par exemple [45].

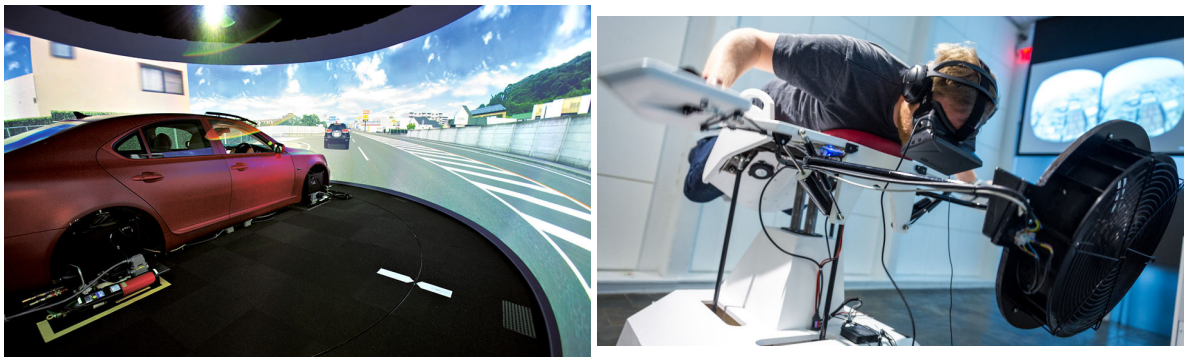
Dans une situation de navigation automatisée, la seule liberté de l'utilisateur est la direction de son regard, le *tracking* de tête ou les informations du système gyroscopique associé au dispositif utilisé permettant de déterminer cette direction.

Il est également possible de mettre en place une navigation semi-assistée ou semi-contrôlée où cette fois l'utilisateur pourra contrôler une partie des paramètres de navigation. Parmi ces paramètres, soit la direction, la vitesse ou l'accélération seront le fait d'interactions de l'utilisateur, les autres paramètres étant dirigés par le programme. Il est ainsi possible de créer des chemins de navigation précalculés à partir d'une position de l'utilisateur, ce dernier devant choisir le chemin qu'il considère optimal pour sa tâche. À chaque nouvelle position, de nouveaux chemins optimaux seront calculés et soumis au choix de l'utilisateur. Cela passe également par la mise en place de contraintes, soit d'orientation, soit de direction, ou encore de vitesse qui viendront influencer le déplacement de l'utilisateur vers un point donné ou autour d'un objet d'intérêt. Ces contraintes, lorsqu'elles sont mises en place, reflètent souvent la volonté de la part du créateur de la scène virtuelle de garder l'attention de l'utilisateur sur le coeur de sa tâche et de simplifier sa navigation pour qu'il se concentre sur l'exécution de cette tâche [142].

Paradigmes

Au-delà des périphériques d'interaction utilisés pour la navigation, les paradigmes permettant de traduire une commande issue de ces dispositifs par un déplacement dans le monde virtuel sont nombreux. Les professionnels du jeu vidéo conçoivent depuis longtemps des paradigmes de navigation variés et appliqués la plupart du temps à des contenus réalistes, 2d ou 3d, utilisant des dispositifs spécifiquement conçus pour un usage domestique, comme les manettes ou les joysticks.

Dans des domaines plus industriels, les paradigmes propres aux simulations de transports comme le vol ou la conduite de véhicule sont utilisés dans les scènes virtuelles constituées de grandes étendues à explorer (cf. Figure 2.8a). Proches des conditions réelles de pilotage ou de conduite, ils permettent d'assurer une certaine identité de l'expérience réelle/virtuelle et possèdent une courbe d'apprentissage courte, puisque basée sur l'expérience des utilisateurs.



(a) (b)

FIGURE 2.8 – *(a) Simulateur de conduite dans un système de rétroprojection sur écran courbé. (b) Exemple du simulateur de vol Birdly où les déplacements virtuels sont la conséquence des mouvements réels de l'utilisateur adoptant une posture de chute libre.*
Source : (a) `http://www.lexus-int.com/design/virtual-reality.html`

En réalité virtuelle, au-delà des approches qui tentent de reproduire de manière stricte un déplacement dans le monde réel, des paradigmes de navigation s'appuient sur le *tracking* de tête ou du corps. Parmi ces approches, certaines considèrent chaque position de l'utilisateur par rapport à une zone spatiale de référence. Un déplacement dans la direction de la droite reliant la zone de référence à la position de l'utilisateur, à la manière d'un joystick où l'utilisateur serait le sommet du manche et la zone de référence constituerait la base de ce manche (voir Figure 2.9) [22, 37].

Le *tracking* des mouvements de l'utilisateur est aussi utilisé afin de traduire le mouvement naturel de la marche en un équivalent dans le monde virtuel.

D'autres techniques plus avancées, comme la **marche redirigée,** exploitent l'imprécision chez l'homme de la perception de la rotation de sa tête, pour induire une trajectoire courbe de son déplacement physique pour le contraindre dans espace de travail physique restreint, alors que le sujet a l'impression de marcher en ligne droite dans le monde virtuel [29].

Il existe donc une large variété de paradigmes permettant l'exploration et la navigation au sein de scènes écologiques.

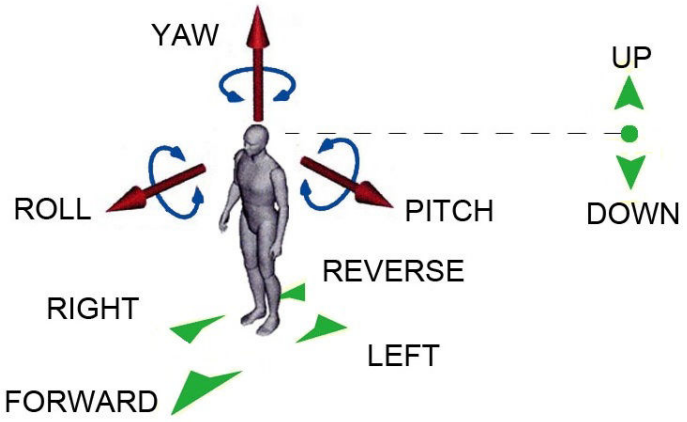


Figure 2.9 – *Exemple de l'utilisateur des déplacements relatifs d'un utilisateur par rapport à une zone de référence pour contrôler son déplacement dans le monde virtuel. L'orientation de son regard décidera également des rotations dans le monde virtuel. Il est donc possible d'avoir un contrôle de 6 degrés de liberté pour la navigation (3 en translation et 3 en rotation).*

2.2 Apports et usages de la Réalité Virtuelle en biologie structurale

La RV possède plusieurs facettes répondant naturellement aux problèmes posés par l'analyse scientifique. Rappelons la problématique actuelle de la visualisation de données scientifiques. Les données générées excèdent de loin les capacités d'interprétation humaines. De plus, la complexité et la quantité des données sont telles que leur rendu 2D ou 3d sur des écrans d'ordinateur ne sont plus suffisants pour rapporter l'ensemble des informations que les données contiennent.

2.2.1 L'immersion dédiée à la visualisation moléculaire

Dans le cadre de la visualisation scientifique, on peut considérer que la capacité d'affichage stéréoscopique doublée à une surface d'affichage à 360 degrés est la caractéristique qui la rend la plus attractive pour l'exploration de données scientifiques. Plusieurs études ont par exemple démontré que la perception de la profondeur lors de la représentation de structures moléculaires apportait une aide non négligeable pour leur compréhension structurelle [174, 165, 124]. Les complexes moléculaires et les protéines sont par nature structurés en 3d et c'est cette structuration qui est au cœur des études en biologie structurale. La stéréoscopie s'est donc imposée comme une technique adaptée pour l'observation de structures protéiques. Mais l'étude de la structure seule ne peut suffire lors de l'étude d'un complexe moléculaire, nous avons souligné auparavant la présence de nombreuses données accompagnant la génération de modèles 3d lors d'une simulation moléculaire.

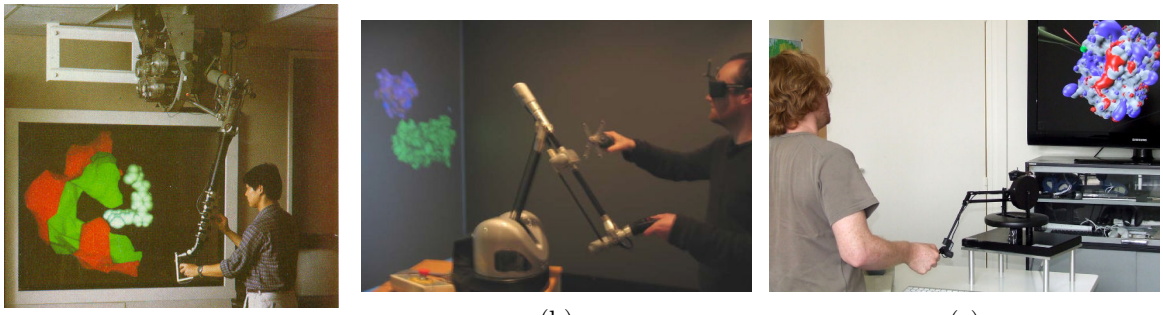
2.2.2 Les interactions multimodales

Nous avons vu que ces retours sensoriels participent à l'immersion ressentie par l'utilisateur dans un monde virtuel, mais ils peuvent également servir de repères ou de vecteurs d'informations utilisés pour compléter les informations visuelles. Même s'il est possible de mettre en place certaines sollicitations autres que visuelles lors d'un travail sur un poste de

travail standard, il est très rare de trouver des retours sonores ou haptiques lors d'une session de travail. Au-delà des limites matérielles qui peuvent exister, un retour haptique impliquant par exemple la nécessité de posséder un dispositif muni d'un système de retour d'effort, les limites sont souvent logicielles. Peu de programmes implémentent des retours sensoriels autres que visuels dans la conception de leurs outils d'analyses de données. Il est au contraire très commun de prendre en compte ces retours sensoriels lors du développement de solutions logicielles dédiées à la RV.

La RV est par essence définie par l'implication de l'utilisateur. Elle a donc dû développer très tôt des moyens pour retranscrire un maximum de sensations aux utilisateurs lors de leurs expériences virtuelles. Les systèmes haptiques ont su rapidement venir de la RV jusqu'aux laboratoires de biologie structurale dans le but permanent d'améliorer la perception 3d des structures moléculaires d'intérêt. La possibilité de toucher une molécule, en combinaison de sa visualisation 3d, offre une complémentarité importante pour la compréhension de certaines particularités structurelles [164].

Le domaine du docking moléculaire fut assez friand de la technologie de bras haptique afin de permettre le ressenti des forces d'interaction prenant place entre deux partenaires [122, 145]. L'ajout de la flexibilité dans les techniques de docking et l'utilisation de techniques de simulations interactives (voir section 2.2.4) où des forces peuvent être ajoutées par l'utilisateur au sein d'une simulation en cours fut aussi un terrain propice à l'utilisation de systèmes de retours haptiques [166]. Ces derniers permettent en effet un contrôle fin de la force appliquée et imposent une limite physique à celle-ci que l'utilisateur ne peut dépasser. Le matériel a pu évoluer avec le temps afin de répondre aux besoins évoluant des équipes de recherche et un aperçu de cette évolution est illustré dans la Figure 2.10a.



(a) (b) (c)

Figure 2.10 – *(a) Système haptique appelé the Docker conçu et construit par Ming Ouh-young et permettant de simuler les forces et les torsions dues aux interactions électrostatiques entre les deux molécules. (b) Système haptique Phantom utilisé dans la salle immersive d'un laboratoire de réalité virtuelle dans le cadre d'une expérience mettant en jeu le contrôle d'un docking par interaction haptique. (c) Système haptique permettant la manipulation d'une protéine au sein d'un laboratoire de biologie structurale dans le cadre d'une expérience de simulation moléculaire interactive (voir section 2.2.4).*

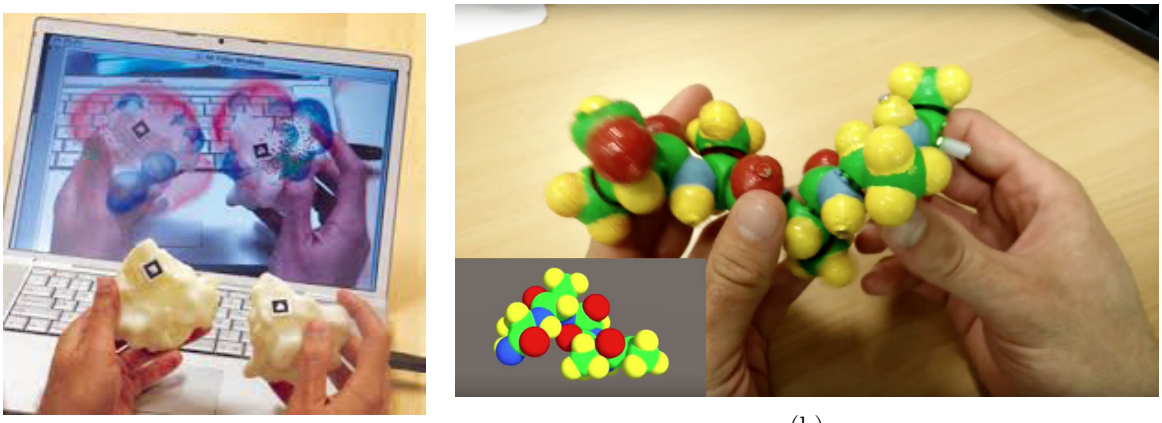
En visualisation de données abstraites et/ou scientifiques, l'utilisation de retours sensoriels a donc pu être détournée de son but premier, l'immersion, pour communiquer des informations supplémentaires à l'utilisateur pendant ses phases d'interactions. La possibilité par exemple de déclencher un événement sonore lors de la sélection de données critiques ou extrêmes dans un set de données est l'un des exemples de l'utilisation d'un retour auditif pour transmettre une information [61]. De la même façon, le domaine de la conception assistée par ordinateur

(CAO), très présent en RV, utilise des dispositifs de retour de force afin de juger de la résistance de matériaux ou de limites de torsion/translation des objets [167]. La chirurgie est également demandeuse de solutions précises de retours haptiques au sein de ses récentes applications de RV dédiées à l'entraînement des chirurgiens à des opérations spécifiques ou développées pour le contrôle de robots pour des opérations sur des patients réels [100]. Étendre ces moyens de fournir des informations par d'autres canaux que les canaux visuels pour la visualisation de données scientifiques en RV est donc une solution réaliste et concrète, simplifiée par les méthodes de RV déjà existantes.

2.2.3 Interfaces moléculaires tangibles et réalité augmentée

Nous avons vu dans le chapitre précédent la place majeure qu'avaient les modèles physiques comme moyens de représentation en premier lieu, les représentations par ordinateurs ne permettant pas au départ une complexité équivalente aux modèles physiques, puis comme moyens de communication, moyens les plus simples à transporter et présenter lors de congrès. Désormais, plusieurs études considèrent leur utilisation comme vecteurs d'interaction. Ces études s'appuient sur l'évolution conjointe des techniques de Réalité Augmentée (RA) et d'impression 3d.

Avant tout approfondissement des cas d'usages de ces modèles physiques, nous revenons rapidement sur les concepts de Réalité Augmentée et d'impression 3d. La RA se définit comme l'ajout en temps réel de contenus virtuels 2d ou 3d au sein du monde réel grâce à un dispositif informatique (smartphone, casques de RV, lunettes de RA). De nombreuses applications de la RA existent, du rendu de meubles virtuels dans des pièces d'habitation aux labels flottants d'information sur certains lieux touristiques. L'impression 3d est quant à elle une technologie permettant de fabriquer des objets à partir de leur modèle 3d informatique.



(a) (b)

Figure 2.11 – *(a) Tracking par marqueurs de modèles monobloc de protéines et rendu vidéo en réalité augmentée ajoutant les résidus chargés en bleu et rouge. Travail effectué au sein de l' «Olson Laboratory». (b) Reconstruction 3d d'un modèle complexe de peptide formé de groupes physiques distincts et possédant des aimants afin de rapporter les forces rotationnelles de liaison grâce à un retour haptique. Un rendu temps réel est effectué en parallèle grâce à un tracking par analyses d'image. Travail effectué au sein de VENISE, groupe du LIMSI-CNRS. Source : (a)* http://mgl.scripps.edu/

Combinées ensemble, ces techniques permettent par exemple, après la conception et l'impression d'un modèle physique de protéine, son traitement visuel en temps réel pour le suivre

(par *tracking* de marqueurs spécifiques ou *tracking* de forme/couleurs) et permettre un rendu visuel du modèle et l'affichage de propriétés ou d'informations de façon dynamique et toujours en temps réel [64]. Les modèles physiques de protéine peuvent être des modèles statiques monobloc, mais peuvent aussi être des modèles physiques complexes et résultats d'assemblages de sous-unités physiques faisant l'analogie des acides aminés. Au-delà de la simple analogie, ce découpage permet de mettre en place des solutions techniques visant à imiter les énergies rotationnelles inter acides aminés à l'échelle des sous-unités physiques [33]. Ces modèles physiques avancés peuvent être envisagés comme moyens d'interactions pour la reconstruction assistée de peptides ou l'estimation d'énergie potentielle [112]. Leurs changements conformationnels de modèles informatiques 3d simulés seraient induits par les déformations du modèle physique manipulé (voir Figure 2.11b).

2.2.4 Simulation moléculaire interactive

Certaines simulations moléculaires, couplées à de rendus graphiques très performants et des dispositifs d'interactions 3d, permettent à l'utilisateur de contraindre certaines parties d'une structure moléculaire pendant sa simulation en injectant dans les calculs physiques des forces s'appliquant sur les particules manipulées [19]. Le jeu sérieux Foldit [4] s'appuie sur les stratégies et l'esprit d'analyse des utilisateurs, experts ou non du domaine, pour guider les changements conformationnels de protéines par des commandes simples. Un score est calculé en temps réel permettant de rendre compte de la stabilité énergétique de la structure obtenue. Alors que le chemin énergétique suivi par un algorithme standard va favoriser les changements conformationnels peu coûteux, l'esprit humain est capable d'anticiper un changement qui va contraindre la protéine de façon importante, mais permettant d'obtenir une structure plus stable par la suite [93].

Dans le même objectif de coupler des simulations à des outils de visualisation performants, SAMSON [5] est un logiciel de visualisation pour la nanoscience permettant de coupler un rendu graphique de haute performance, stéréoscopique, et plusieurs plug-ins de simulation ou docking. Il permet entre autres l'utilisation de modèles quantiques pour des simulations de petites molécules [71] et peut être couplé à des périphériques d'interactions de RV tels que des bras haptiques ou des dispositifs de tracking de doigts comme le «*Leap Motion*» par exemple [6].

Il est également possible de générer des représentations volumiques de certaines informations expérimentales comme des cartes de densité de données cristallographiques ou SAXS afin de permettre à l'utilisateur d'agir sur les domaines flexibles de sa protéine afin de la faire entrer dans ces volumes. Certains travaux utilisent même directement l'aide de l'utilisateur afin de replier une protéine représentée grâce à des paramètres de mécanique moléculaire et déplacée/orientée dans une carte de densité électronique SAXS [120].

La possibilité d'influer ainsi sur une simulation moléculaire demande la mise en place d'une structure logicielle de haute performance, car des modules de simulation, de visualisation, d'interactions homme-machine doivent fonctionner de façon synchronisée. La représentation simplifiée de la protéine permet une certaine rapidité de la simulation et donc la prise en compte rapide de l'effet des changements opérés par l'utilisateur. Dans cette optique, le projet BioSpring permet la représentation de protéines par des réseaux de ressorts pour les interactions liées et des paramètres de champ de force configurables pour les interactions non liées [58]. Cette représentation semi-simplifiée permet de contraindre les parties rigides de

4. https://fold.it/portal/
5. http://www.samson-connect.net/
6. https://www.leapmotion.com/

la protéine tout en assurant une certaine flexibilité au niveau des régions moins structurées (boucles et coudes).

Dans le cas de dynamiques moléculaires sur des systèmes de taille plus importante, calculées dans des clusters de calcul déportés, les défis sont autres. Ils passent par la mise en place de communications privilégiées entre le centre de calcul et le lieu de visualisation et d'interaction. Des suites logicielles performantes permettent aujourd'hui de traiter une simulation de façon interactive alors même que cette dernière est déportée [56], à l'aide d'approche de synchronisation et de parallélisation très avancées compatibles avec le calcul haute performance.

2.2.5 Outils et applications

La biologie structurale n'a su que tardivement se placer par rapport à l'immersion apportée par la RV. Si elle a utilisé certains de ses outils (systèmes haptiques par exemple), ce n'est que très récemment que certains programmes dédiés à l'exploration moléculaire ont commencé à être utilisés au sein de dispositifs immersifs de RV [124].

YASARA [97] ou VMD [165] sont deux exemples de programmes de visualisation moléculaires disponibles pour un rendu stéréoscopique et adapté aux environnements immersifs. Ils permettent l'interfaçage de leur solution de visualisation avec de nombreuses librairies utilisées en RV dans les systèmes immersifs de type CAVE ou mur d'écran. Parmi ces librairies, nous pouvons citer VRPN [170], FreeVR [127] ou VR Juggler par exemple.

D'autres initiatives ponctuelles ont aussi vu le portage d'autres logiciels dans des environnements immersifs, mais ces développements spécifiques ont rarement dépassé l'extension du rendu graphique depuis le 2d jusqu'en 3d [7] ou la mise en place de librairies comme base de développement d'applications RV pour la visualisation moléculaire [143]. Afin d'améliorer l'apprentissage de certains concepts biologiques, il est parfois utile d'en améliorer la perception, en particulier quand ces concepts sont par certains aspects abstraits. Un projet éducatif a cherché à évaluer l'impact pour les étudiants de visualiser des objets moléculaires en RV dans le cadre de cours de biologie structurale [169]. Selon leurs conclusions, les dispositifs immersifs ont permis un meilleur apprentissage des notions abordées ainsi qu'une compréhension globale plus complète de la biologie structurale.

L'usage des contextes immersifs dans le cadre de la biologie structurale est donc limité et de nombreuses pistes d'améliorations sont possibles.

2.2.6 Limites et perspectives

Nous avons dessiné dans la section 2.1 les contours de la RV et mis en avant les apports de son utilisation pour la biologie structurale dans la section 2.2.

La notion de monde virtuel implique la possibilité pour l'utilisateur d'évoluer au sein d'un environnement étendu de la même manière qu'il évoluerait dans la vie réelle. Or, si cette navigation virtuelle peut être inspirée de paradigmes utilisés dans un cadre de navigation réelle au sein d'environnements virtuels écologiques où le contenu virtuel proposé est une copie artificielle d'éléments réels, la question est tout autre lorsque la scène observée n'est plus écologique et implique des éléments abstraits.

Conçus autour de contenus réalistes, les paradigmes cités précédemment ne sont donc pas nécessairement transposables à des scènes abstraites. Leur efficacité est limitée du fait de la nature très différente des données à observer. Alors que la navigation aura tendance à

7. http://www.rug.nl/science-and-society/centre-for-information-technology/research/hpcv/vr_visualisation/mol_visualisation?lang=en

permettre l'exploration d'une surface virtuelle horizontale étendue dans des scènes réalistes, les scènes abstraites scientifiques, et plus particulièrement les scènes moléculaires, concentrent les informations dans une zone centrale autour de laquelle l'utilisateur va évoluer. L'échelle de visualisation des données est capable d'augmenter la distance (toujours à l'échelle) des données observées, mais la nature même de l'exploration est souvent différente. Sheidermann décrit la visualisation de données comme un processus où l'exploration est l'étape préliminaire avant les étapes de zoom et de filtre qui précèdent elles-mêmes l'étape finale de l'obtention des détails à la demande [154]. Les différentes échelles de précision mises en avant dans cette description sont rarement retrouvées dans les paradigmes de navigation au sein de scènes virtuelles réalistes.

Bien qu'il existe de nombreux portages de logiciels experts de visualisation moléculaire dans des environnements virtuels, la navigation au sein de données scientifiques dans ces derniers se contente encore largement de la manipulation d'objet retrouvée dans les logiciels de visualisation moléculaire courants (voir Figure 2.12) [59]. Les seules tâches de navigation pouvant être identifiées au sein de logiciels experts de visualisation moléculaire se rapportent à des transitions progressives permettant de rejoindre une position spécifique depuis la position actuelle de l'utilisateur. Enfin, ces logiciels ne présentent aucune adaptation de la manipulation suivant le type de molécule observé et les paradigmes mis en jeu sont identiques, que l'objet observé soit une protéine de quelques acides aminés ou un virus de plusieurs millions d'atomes.

Ils permettent une navigation totalement libre autour de la molécule et n'imposent aucune contrainte, au détriment de la conscience spatiale de l'utilisateur. La visualisation moléculaire met en scène des représentations artificielles d'atomes, non observables à l'œil humain en temps normal, et dont les couleurs et formes respectent des standards décidés par le domaine, non dictés par des observations réelles. Les repères spatiaux sont donc peu nombreux et souvent trop sommaires pour assurer une orientation acceptable de l'utilisateur. Ils se limitent donc au seul objet intérêt, son environnement (skybox, paysage, etc.) qui se révèle être majoritairement vide avec un arrière-plan de couleur unie.

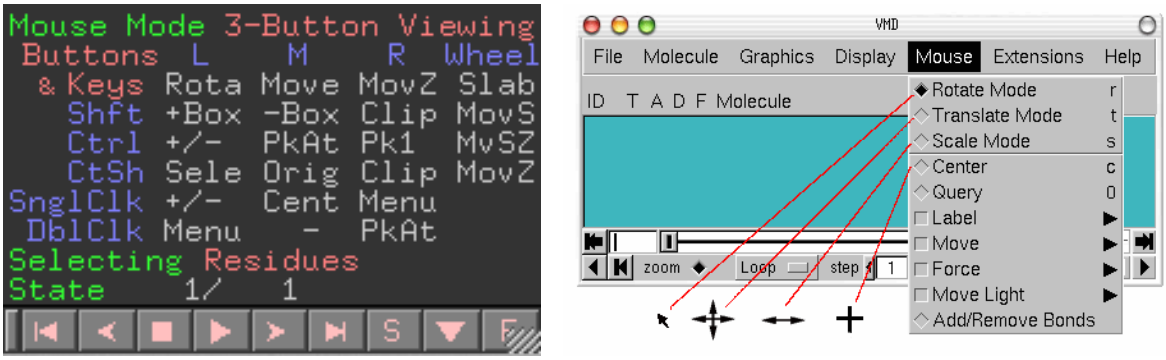


FIGURE 2.12 – *Capture d'écran des interfaces de manipulation offertes par PyMol (à gauche) et VMD à droite. Ils sont constitués de nombreuses combinaisons souris/clavier pour permettre d'utiliser l'ensemble des possibilités de manipulation disponibles.*

2.2.7 Évaluation des usages et tâches expertes

Nous sommes partis du constat que l'évaluation systématique de tâches métier est très complexe à mettre en place à cause de plusieurs facteurs: (1) la spécialisation de l'outil évalué est très importante et nous faisons face à des outils, en particulier de visualisation moléculaire,

dont la nature (PyMol, VMD, Yasara) varie suivant les usages usuels et individuels des experts scientifiques interrogés. (2) L'implémentation et l'adaptation de développements sur un ensemble représentatif d'outils spécialisés est compliquée et très chronophage. (3) Notre démarche est biaisée en raison de son approche basée sur l'exécution de tâches expertes par rapport à des solutions proposées actuellement dans les outils de visualisation ou d'analyses. (4) Afin d'appliquer au mieux les méthodes d'évaluation classiques, il est nécessaire d'avoir un nombre statistiquement élevé de participants. Or la population d'experts du domaine est relativement limitée.

Nous proposons dans cette partie de reprendre une méthode d'évaluation plus théorique qu'empirique pour cerner l'efficacité de notre approche. La méthode HTA (pour «Hierarchical Task Analysis») consiste en une subdivision d'une tâche principale en plusieurs sous-tâches précises [4]. Chaque sous-tâche peut à son tour être subdivisée jusqu'à ce que les sous-tâches atteignent un degré de précision suffisant pour qu'il soit possible de donner une approximation du temps d'exécution qu'elle nécessite pour être accomplie.

C'est une méthode très usitée par les ergonomistes qui la dédient à plusieurs fins, dont l'optimisation de tâches, la conception d'interfaces, etc [61, 52, 13]. Elle implique l'apport d'un unique expert pour évaluer les temps d'execution des sous-tâches identifiées et permet une lecture rapide et simple des différences de performances, mais également de méthodologies entre deux conditions de travail différentes pour la même tâche experte.

Nos développements seront donc analysés au moyen d'une décomposition hiérarchique de tâches expertes que nous analyserons tant en terme de performances que de méthodologies afin de mettre en avant les différences avec des outils classiques (voir sections 3.1.6 et 5.3.8).

2.3 Conclusion

L'apport de l'immersion et de la perception de profondeur pour la représentation de structures moléculaires est indéniable. Cependant, le caractère abstrait des biomolécules en biologie structurale influe négativement sur la qualité des repères spatiaux que l'utilisateur possède pour naviguer au seins de scènes virtuelles. Comparé aux scènes mettant en jeu des mondes réalistes, les scènes de données scientifiques ne comportent pas de repères spatiaux standards utilisés naturellement par les gens pour évoluer au sein d'un espace 3d virtuel. Les paradigmes de navigation habituels qui permettent d'évoluer dans un monde réaliste et donc inspirés des techniques de navigation naturelles, ne sont plus adaptés lorsque l'objet d'intérêt est unique, central et sans orientation propre. Les repères spatiaux utilisés pour orienter l'utilisateur dans le monde virtuel qu'ils utilisent ne peuvent être repris à l'identique dans des scènes moléculaires.

De plus, les techniques actuelles d'exploration de structures moléculaires dans les logiciels experts 2d ne permettent pas de gérer efficacement les structures de grande taille et de grande complexité (intégration de l'environnement, multiples composants, etc.). Absentes en 2d, il n'est donc pas possible de s'inspirer des techniques de navigation des logiciels experts actuels.

Afin de combler ce manque, préjudiciable pour l'expérience utilisatrice lors de ses futures tâches en biologie structurale, nous avons développé une série de paradigmes de navigation, basées sur le contenu observé et les tâches usuelles en visualisation/exploration moléculaire. Ces paradigmes sont présentés dans le prochain chapitre et doivent permettre l'exploration d'une scène moléculaire, dans des dispositifs immersifs, de façon contrôlée et la moins perturbante possible pour l'utilisateur.

Chapitre 3

Exploration interactive de données moléculaire en immersion

Dans ce chapitre sera présentée, dans une première partie, la mise en place de nouveaux paradigmes de navigation permettant à l'utilisateur d'explorer, dans un contexte immersif, une scène virtuelle moléculaire, contraint par le contenu qu'il explore et la tâche métier qu'il effectue. Nous présenterons dans une seconde partie un travail visant à améliorer la façon de communiquer et de visualiser des structures moléculaires sur des périphériques mobiles en s'inspirant de la perception de profondeur offerts par les environnements immersifs utilisés dans la première partie.

Sommaire

Introduction

La visualisation de modèles 3d de protéines ou de complexes moléculaires est une étape indispensable en biologie structurale (voir Figure 3.1). Le rôle historique de la visualisation moléculaire pour communiquer et comprendre le fonctionnement des complexes moléculaires a été mis en avant dans le premier chapitre (voir section 1.2.3) et a été récemment l'un des sujets d'une revue complète [87].

Nous avons effectué un inventaire rapide des techniques de visualisation moléculaire, dans la section 1.2.3, permettant de mettre en avant des informations structurelles ou physico-chimiques au moyen de représentations graphiques aujourd'hui standardisées (formes, couleur, etc.) et de techniques de rendu photo-réaliste (ombres, éclairage, texture, etc.).

Le chapitre précédent nous a montré que la RV et les technologies s'y rapportant constituent un complément à la visualisation afin d'améliorer notamment la perception de la profondeur, en offrant un espace de travail plus large, et des modalités d'interactions plus adaptées aux objets manipulés, intrinsèquement tridimensionnels.

Nous avons cependant mis en exergue le manque de paradigme de navigation adapté à l'exploration de données scientifiques par nature abstraite par opposition aux nombreux paradigmes retrouvés pour parcourir des scènes virtuelles réalistes.

La première contribution de notre travail eut pour principal objectif de combler les lacunes en terme de navigation dans les données moléculaires, lacunes qui constituent aujourd'hui un frein important à l'ergonomie des dispositifs immersifs, même bon marché. La conscience spatiale dégradée et la perte de repère dans les scènes virtuelles, induites par des techniques de navigation inadaptées aux contenus et à la tâche, constituent la principale cause du mal du simulateur, et réduisent la qualité de l'expérience et la performance de l'utilisateur. Le mal du simulateur a été étudié de manière récurrente en RV en faisant l'objet de nombreuses études [101, 49]. Ces études ont cependant ciblé principalement l'exploration de mondes virtuels réalistes, les études concernant l'exploration de données abstraites et scientifiques étant beaucoup plus rares. Nous étudierons donc les méthodes de navigation dans des scènes réalistes pour identifier les progrès à réaliser dans les scènes de données abstraites, puis nous présenterons nos contributions pour répondre au besoin de paradigmes de navigation adaptés aux données moléculaires [171]. Nous évaluerons ensuite ces paradigmes à l'aide d'une méthodologie théorique d'évaluation de tâches. Nos paradigmes seront appliqués à la réalisation de tâches métier complexes, conçues avec l'aide des experts du domaine et de spécialistes en ergonomie, dont Julien Nelson, maître de conférences à l'université Paris Descartes et ayant déjà contribué à plusieurs études ergonomiques pour la biologie structurale en conditions immersives [61, 176].

La seconde contribution eut pour objectif de fournir aux experts des techniques d'exploration et de navigation leur permettant de suivre une simulation moléculaire, grâce à des outils resituant l'expert au cœur du processus de simulation, et non plus seulement en aval de ce processus. En effet, les simulations moléculaires de grande échelle s'exécutant souvent sur plusieurs jours ou semaines sur les centres de calcul font fréquemment l'objet d'une perte de temps de calcul importante liée à un manque d'outils de supervision pendant une simulation. Or cette simulation peut éventuellement avoir évoluée vers des états aberrants ou incohérents d'un point de vue biologique ou physico-chimique et donc présenter des résultats peu exploitables pour les experts. Pour répondre à cette problématique, la prise en compte des contraintes organisationnelles des centres de calcul et des limites des ressources des laboratoires est cruciale. D'une part, les accès aux centres de calcul sont souvent très contraignants du fait d'une exigence de sécurité très élevée. D'autre part, les laboratoires disposent de res-

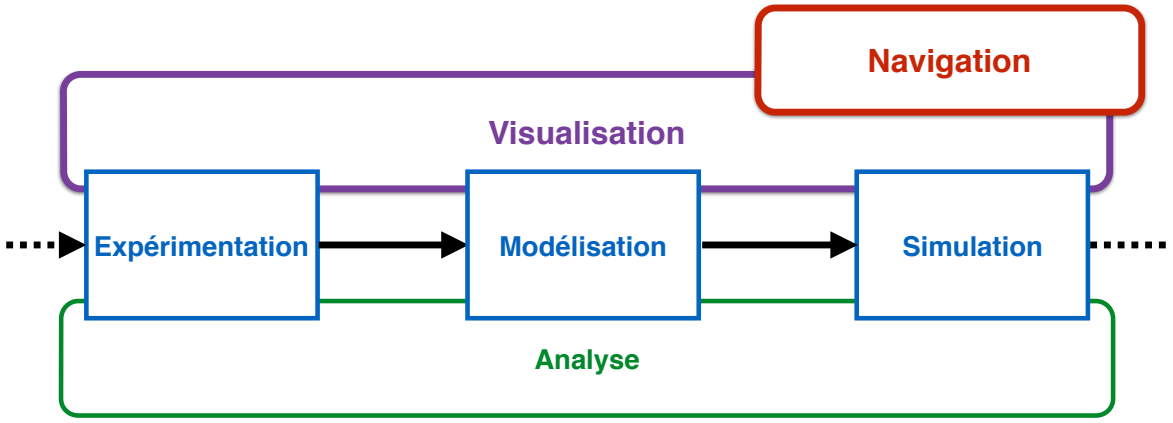


FIGURE 3.1 – *Schéma du processus itératif d'étude d'une structure moléculaire en biologie structurale. Nous nous intéressons ici à l'outil de visualisation et plus spécifiquement à l'une des méthodes qu'elle introduit, la navigation.*

sources de communication réseau et de stockage limitées au regard de la quantité de données produites par les centres de calcul. Dans l'optique d'apporter des solutions à cette problématique tout en respectant ces lourdes contraintes, nous avons travaillé sur une approche offrant aux experts la possibilité d'accéder au quotidien à des résumés visuels de leur résultat de simulation en cours avec un niveau de rendu et de contrôle de leur scène virtuelle moléculaire suffisant pour être utilisés comme supervision des résultats de simulation. Cette approche fut aussi l'occasion d'exploiter les nouvelles ressources de visualisation et d'interaction mobiles propices au partage de résultats moléculaires dans un domaine où la communication scientifique est très centrée autour du rendu visuel de complexes biomoléculaires.

3.1 Paradigmes de navigation pour l'exploration de complexes moléculaire

Nous avons vu dans le chapitre précédent (voir section 2.1.3) que la navigation dans des EV immersifs se caractérisait par la volonté d'étendre les capacités d'exploration de mondes virtuels en dépit des limites physiques imposées par les EV immersifs.

Nous avons mis en évidence le besoin de mettre en place des paradigmes de navigation qui répondent aux problématiques posées par la visualisation moléculaire qui met en jeu des objets scientifiques abstraits, et de plus en plus complexes.

À l'image de ce qui a été fait de manière implicite en terme de navigation dans les scènes réalistes, il est important de prendre en compte le contenu virtuel à explorer pour mettre en place un paradigme de navigation efficace. De la même façon, la navigation étant le support d'un nombre important de tâches en visualisation moléculaire, il nous a paru important de prendre également en compte l'activité des experts.

Ce travail [171] a été réalisé en collaboration avec des biologistes travaillant sur des exemples concrets de modélisation de phénomènes biologiques et utilisant l'outil de la visualisation moléculaire quotidiennement. Nous avons par ailleurs travaillé avec des ergonomes dont les approches d'analyse et de recueil des informations au travers d'entretiens encadrés furent très utiles à l'identification des besoins des experts. Il fut important dans notre ap-

proche d'associer les experts du domaine dans toute la phase de conception de paradigmes afin d'éviter un écart trop important entre leurs attentes et le résultat final. Nous avons ensuite, après la conception et le développement de chaque paradigme de navigation, mis en place un protocole d'évaluation sous la forme d'une décomposition hiérarchique en utilisant la méthodologie de *Hierarchical Task Analysis* (HTA) afin d'évaluer la performance de nos paradigmes de navigation par rapport à une navigation libre [4].

3.1.1 Symétrie moléculaire et axes remarquables comme ancrage visuel

Les complexes moléculaires sont souvent le résultat d'agencements de monomères (chaînes d'acide aminé répétées) créant des multimères de taille importante et souvent constitués de plusieurs domaines. Les progrès des simulations moléculaires permettent désormais de simuler ces complexes moléculaires au sein de leur environnement, à l'image des protéines transmembranaires et des membranes dans lesquelles elles sont enfouies.

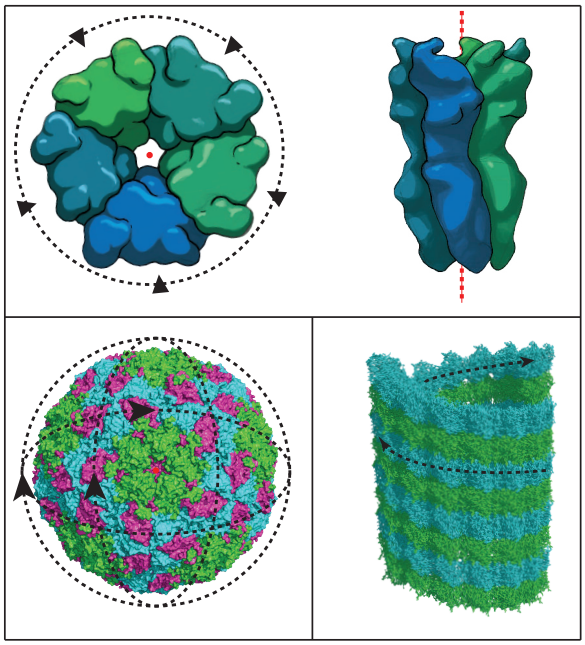


Figure 3.2 – *Différents types de symétries retrouvés dans des complexes moléculaires. En haut, GLIC, une protéine transmembranaire composée de 5 monomères avec un axe de symétrie correspondant au centre du pore formé par l'agencement des 5 monomères. Vue du dessus (à gauche) et vue de côté (à droite). En bas à gauche, une capside de virus composée de 4 types de monomères présentant un centre de symétrie superposé au centre de gravité du virus. En bas à droite, enchevêtrement de tubulines agencées autour d'une symétrie hélicoïdale (axiale + translation verticale) formant un microtubule.*

Ces assemblages moléculaires font souvent l'objet d'axes ou de centres de symétrie jouant un rôle primordial dans la fonction même des complexes moléculaires [66]. Plusieurs exemples d'arrangements symétriques trouvés dans des complexes biologiques sont illustrés dans la Figure 3.2. Ces axes de symétrie, obtenus de façon automatisée ou renseignés manuellement, peuvent donc permettre de définir un sens arbitraire (haut et bas) de la scène virtuelle moléculaire ainsi qu'être la base de la génération automatique de chemins de navigation.

3.1.2 Des indices visuels stables pour améliorer la conscience spatiale de l'utilisateur

Lors de la phase d'exploration non détaillée d'un complexe moléculaire, la navigation consiste à se déplacer autour de l'objet d'intérêt. Les axes de symétrie constituant une orientation préférentielle pérenne tout au long de l'expérience virtuelle, ils sont à la base de la conception nos paradigmes de navigation afin d'offrir des repères spatiaux durant toute cette phase de la navigation. Lors de la phase d'exploration plus détaillée, l'utilisateur étant immergé au milieu de son complexe moléculaire, l'orientation préférentielle de la structure est plus difficilement perceptible, et il s'agit de fournir des indices visuels plus éloignés. Pour ajouter d'autres indices, nous avons donc choisi d'utiliser une approche par *skybox*. Il s'agit de présenter à l'utilisateur un paysage distant dont la perception lors de l'activité de navigation détaillée permet à l'utilisateur de se construire des repères visuels éloignés améliorant sa conscience spatiale. Le choix de la skybox doit respecter un certain nombre de prérequis afin de remplir complètement son rôle [177], de permettre à l'utilisateur de mieux mieux percevoir son orientation dans la scène virtuelle lors de la navigation.

3.1.3 Exploration guidée

Nous savons que la forme générale d'une protéine peut fournir des indices importants sur son rôle fonctionnel. Souvent associée aux premières étapes de la visualisation, comme énoncé par Shneidermann [154], l'exploration de complexes moléculaires passe par un parcours extérieur du complexe observé. Nous avons mis en place des chemins de navigations circulaires préférentiels autour de la molécule d'intérêt construits en fonction des symétries de la structure observée.

Dans notre technique de navigation, durant l'étape d'exploration externe, l'utilisateur manipule 3 degrés de liberté afin de modifier la position de son point de vue. Le premier lui permet de tourner autour de la molécule en imposant à la caméra une trajectoire circulaire à une distance constante et autour de l'axe de symétrie. Le second permet de se rapprocher de la molécule en modifiant la distance de cette trajectoire circulaire. Le troisième permet de manipuler l'altitude de cette trajectoire circulaire. Tout au long du déplacement de son point de vue, l'orientation du point de vue de l'utilisateur est stable et co-linéaire par l'axe de symétrie du complexe (voir Figure 3.3. Cette propriété constitue un repère stable important participant à une meilleure conscience spatiale du sujet. Lorsque l'utilisateur atteint en terme d'altitude des extrémités hautes et basses du complexe observé, d'autres contraintes sont progressivement appliquées de manière à contraindre le focus de la caméra sur le barycentre de la protéine.

Concernant une exploration plus détaillée pour observer des phénomènes se situant à l'intérieur des complexes moléculaires, nous proposons d'autres chemins préférentiels, toujours construits autour des axes ou des centres de symétrie, mais adaptés à une modalité d'exploration non plus externe, mais interne du complexe.

Par ailleurs, ce paradigme a été adapté à d'autres types de symétrie notamment pour les symétries centrales. Dans ce cas le degré de liberté permettant de modifier la distance du point de vue par rapport à l'objet d'intérêt est conservée, avec deux degrés de liberté pour la rotation autour de la molécule à la place d'un seul pour les symétries axiales, la notion d'altitude le long d'un axe n'ayant plus de sens dans le cas de la symétrie centrale. Nous avons mis en place ces calculs de chemins préférentiels en nous inspirant de techniques utilisées dans d'autres cadres [92, 73].

Les deux modes d'exploration externes et internes sont illustrés dans les Figures 3.4a et

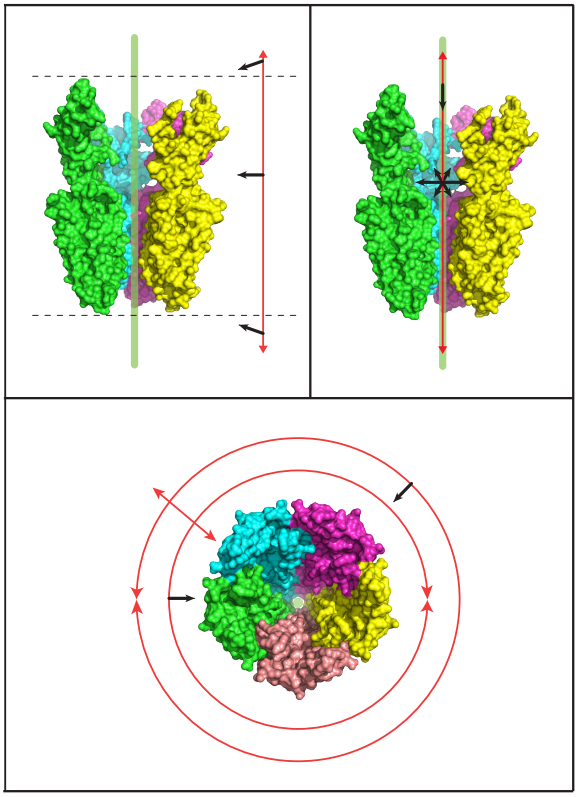
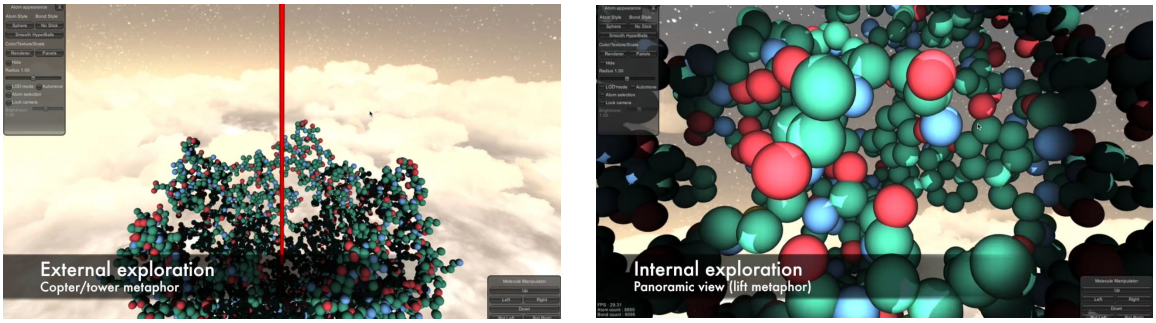


Figure 3.3 – *Illustration des contraintes de navigation imposées par nos paradigmes. Les mouvements de translation possibles sont représentés en rouge, les points de vue autorisés sont représentés par des flèches noires et l'axe de symétrie est représenté en vert. Vue transversale du complexe en haut et vue du dessus en bas.*

3.4b à partir de notre implémentation des paradigmes au sein du logiciel expert UnityMol [106].

3.1.4 Optimisation du parcours des régions répétées

La symétrie, et plus généralement l'agencement des monomères d'un complexe moléculaire, induit souvent l'apparition d'événements structuraux comme la liaison d'un ligand, l'ouverture d'une poche de liaison ou la simple variation structurelle d'une région de façon répétée pour chaque monomère identique. Ainsi, la liaison d'un ligand sur une première chaîne A signifiera certainement que ce ligand peut aussi se lier à une chaîne identique B, C, D, etc. Comparer ces régions d'intérêts est très essentiel pour détecter d'éventuels comportements asynchrones ou des effets différents, et il est difficile de naviguer d'une région à une autre pour effectuer cette comparaison avec des paradigmes classiques de navigation libre sans guide de navigation. Grâce à la symétrie, nous avons mis en place un mode de navigation adapté à cette problématique permettant d'automatiser complètement le passage d'une région d'intérêt à la région répétée sur un autre monomère. Ce calcul est rendu possible par l'utilisation de l'axe de symétrie comme axe de rotation pour la caméra dont l'angle de rotation sera simplement dépendant de l'angle de rotation ayant été calculé entre les monomères. Un assemblage standard verra cet angle être de 360 degrés divisé par le nombre de monomères identiques du complexe. Il est cependant également possible d'utiliser des scripts de calcul



(a) (b)

FIGURE 3.4 – *Implémentation de nos paradigmes de navigation guidée au sein de UnityMol pour l'exploration de GLIC (représenté en sphères) autour d'un axe de symétrie (en rouge, invisible dans la version publique). (a) Capture d'écran d'un point de vue de l'utilisateur pendant une exploration externe de la protéine GLIC. (b) Capture d'écran d'un point de vue de l'utilisateur pendant une exploration externe de la protéine GLIC.*

plus fins afin d'obtenir les valeurs d'angle précises entre les monomères impliqués. Lorsque cette information est connue (elle est par défaut calculée lors de la phase d'initialisation de la scène virtuelle) et que l'utilisateur a choisi son point de vue sur la région d'intérêt, alors une simple interaction avec un bouton permet de parcourir les différents monomères, préservant le point de vue relatif au monomère observé initialement. Il est possible de parcourir les monomères dans les deux sens. Chaque déplacement est effectué automatiquement et une série d'interpolations au cours de la trajectoire de la caméra permet une transition douce entre les différentes positions sans perte de repère pour le sujet.

3.1.5 Trouver un point de vue optimal

Même si l'exploration peut apporter un grand nombre d'informations, certains phénomènes moléculaires ne concernent qu'un sous-ensemble restreint d'atomes ou de résidus. Au sein de larges structures de plusieurs millions d'atomes, ces zones d'intérêt deviennent difficiles à visualiser à cause de la densité des particules voisines. De plus, certaines régions sont profondément enfouies dans le complexe moléculaire, transformant la simple tâche de visualisation en un défi pour l'utilisateur. Afin de résoudre cette problématique, nous avons développé un algorithme capable de calculer le meilleur point de vue pour la caméra, connaissant les coordonnées de la cible et la distance à laquelle la caméra devra être située. Cet algorithme prend en compte les atomes entourant la cible et calcule les cônes de vue les plus larges pour identifier le meilleur point de vue. L'utilisateur aura la possibilité de choisir entre plusieurs cônes de vue afin d'obtenir un angle de vue différent sur la cible choisie.

3.1.5.1 Algorithme de recherche de meilleur point de vue au sein d'un environnement dense

L'algorithme utilisé pour obtenir les cônes de vue sans occultations sur une cible au sein d'un environnement dense en atomes peut être décomposé de la sorte:

1. L'utilisateur fournit les coordonnées 3d de la cible soit en entrant les valeurs (x,y,z) au sein d'une fenêtre de saisie, soit via une sélection directe. Si la cible comporte plus d'un atome alors les coordonnées du centre de masse de la sélection seront calculées.

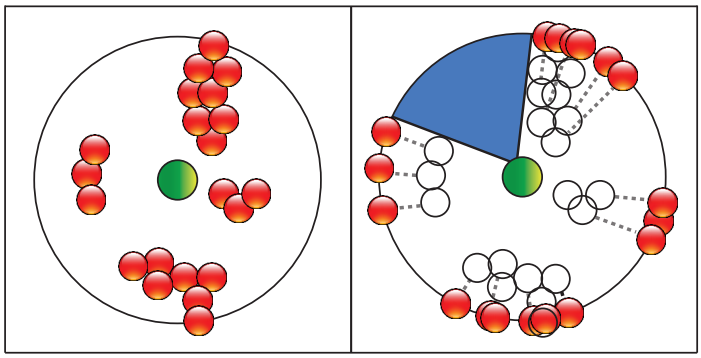


Figure 3.5 – *Schéma de l'étape 4 de l'algorithme de point de vue optimal, représenté en 2d. À gauche, environnement atomique initial. À droite, situations des atomes après projection sur un cercle de diamètre connu.*

La hauteur désirée du cône de vue d est également demandée, elle correspondra à la distance de la caméra par rapport à la cible.

2. Tous les atomes situés au sein d'une sphère de rayon d et centrée sur les coordonnées de la cible sont considérés comme voisins et comme potentielles occlusions, ou obstacles, pour le cône de vue. Leurs coordonnées 3d sont stockées.

3. L'ensemble des coordonnées 3d des atomes voisins stockés précédemment est décalé dans l'espace d'un vecteur correspondant aux coordonnées 3d de la cible. Cela entraîne la création d'une sphère centrée sur les coordonnées de l'atome ciblé.

4. L'ensemble des coordonnées cartésiennes des atomes contenus dans la sphère est transformé en coordonnées sphériques. Les coordonnées sphériques sont composées de 3 paramètres : une distance radiale à l'origine r, un angle polaire θ (thêta) et un azimut ϕ (phi). En mettant de côté la distance radiale, nous obtenons pour chaque atome une information de direction par rapport à l'origine donnée par les angles θ et ϕ. Il est possible de fixer la distance radiale de tous les atomes voisins à la valeur de d et ainsi mettre en place une projection de ces atomes sur la sphère de rayon d. Lorsque la projection est effectuée, la recherche du plus large cône de vue passera par la recherche du plus grand cercle vide à la surface de la sphère. Nous illustrons cette étape en 2d dans la Figure 3.5 où la cible est représentée en vert, les voisins en rouge et le plus large cône de vue obtenu en bleu.

5. Puisque la valeur de la distance radiale est identique pour l'ensemble des atomes voisins, il est possible de créer un nuage de points correspondant aux valeurs de ϕ en fonction des valeurs de θ afin d'obtenir une représentation aplatie de la surface de la sphère en 2d contenant l'ensemble des points projetés (un point = un atome). La matrice 2d ainsi obtenue est étendue de 50% le long des ordonnées et des abscisses afin de prendre en compte la périodicité de la sphère. Il n'est pas nécessaire de l'étendre au-delà de 50% puisque la taille maximum d'un cercle vide sera la taille du plus long côté de la matrice. Si aucune extension n'est effectuée, nous pourrions manquer les solutions qui apparaîtraient à l'extrémité de la matrice où deux régions de la sphère aplatie se rejoindraient. Nous pouvons d'ailleurs voir dans la Figure 3.6 que l'absence d'extensions nous aurait fait manquer deux solutions. Pour synthétiser, ces nuages de points 2d peuvent être considérés comme une carte 2d de la distribution à la surface de la sphère des atomes ayant été projetés.

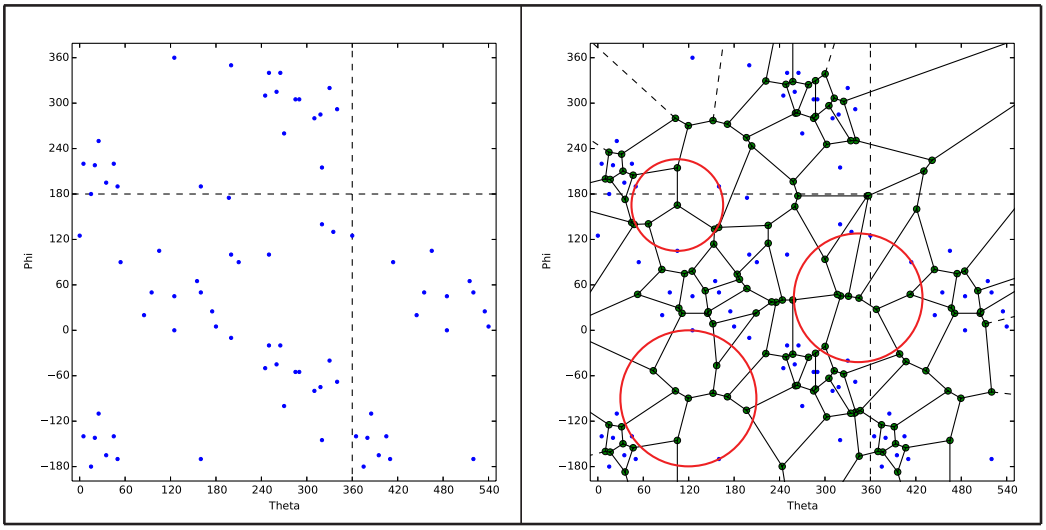


FIGURE 3.6 – *A gauche, graphique représentant les angles ϕ en fonction des angles θ des atomes voisins de la cible, chaque atome est représenté par un point bleu. A droite, même graphique contenant en plus les sommets (en vert) et les segments (lignes noires) du diagramme de Voronoi ayant été calculé sur les coordonnées sphériques des atomes voisins (en bleu). Les cercles rouges représentent les plus larges cercles vides pouvant être trouvés au sein des points bleus.*

6. À partir de la liste des atomes formés par les paires θ/ϕ, un diagramme de Voronoi est calculé. Nous obtenons une liste de sommets de Voronoi associés chacun à un triplet des trois atomes les plus proches.
7. La distance entre chaque sommet de Voronoi et les trois points les plus proches est calculée et les plus grandes valeurs de cette distance indiqueront les plus grands cercles vides présents sur la surface de la sphère. Ces cercles auront pour centre un sommet de Voronoi et passeront obligatoirement par les trois points les plus proches puisque l'algorithme de Voronoi impose que chaque sommet soit équidistant d'au moins 3 points, ceux-ci étant également les plus proches. Le cercle avec le rayon le plus large sera considéré comme le cercle vide le plus grand (cf. Figure 3.6). Il est important de noter qu'en dépit de l'extension de la matrice, nous ne sélectionnons pas de centres de cercles situés à l'extérieur de la boîte entourant les premières coordonnées sphériques tracées (en pointillé dans la Figure 3.6) et correspondant aux extensions. Cependant les points appartenant au cercle peuvent eux se situer dans les parties étendues du graphe.
8. Les valeurs θ/ϕ du centre des plus larges cercles vides sont transformés en coordonnées cartésiennes. Le rayon des cercles donne un angle d'ouverture qui sera utilisé par la suite pour calculer une trajectoire optimale (voir ci-dessous).
9. Les coordonnées cartésiennes peuvent ensuite être choisies itérativement par l'utilisateur pour se positionner par rapport à la cible, la caméra sera contrainte de façon à toujours faire face à la cible.

Cet algorithme a une complexité faible en $\theta(n.\log n)$, avec n désignant le nombre d'atomes autour de la cible, puisque la recherche des voisins de la cible est calculée en n-1 itérations (constant) et que le diagramme de Voronoi et la recherche du cercle le plus large est effectué en $\theta(n.\log n)$ dans le pire des cas où tous les atomes du système sont considérés comme voisins

et situés dans la sphère de rayon d. L'algorithme est capable de fournir une liste des vues les plus dégagées sur une cible de façon très performante et permet à l'utilisateur de changer de point de vue en temps interactif.

3.1.5.2 Atteindre les points de vue optimaux

La recherche d'un point de vue optimal peut être effectuée sur n'importe quelle région du complexe moléculaire étudié, nous avons cependant mis en avant son apport dans le cas de régions d'intérêt à observer enfouies dans le complexe et très difficilement atteignables. Le chemin entre la position d'un utilisateur au moment où il décide de rentrer dans un mode de recherche de point de vue optimal et la position optimale trouvée passe souvent par le passage au travers du complexe, dans des régions potentiellement denses en atomes et perturbantes pour le point de vue de l'utilisateur. Il nous a donc paru nécessaire de mettre en application les recommandations énoncées dans la section 2.1.3.2 qui rappelaient l'importance des transitions entre points de vue, afin tout d'abord de minimiser l'impact négatif que pourrait avoir ce déplacement sur l'utilisateur et ensuite pour permettre à l'utilisateur de préserver une bonne conscience spatiale de l'environnement pendant le déplacement et ainsi pouvoir se situer rapidement lors de son positionnement final devant la cible.

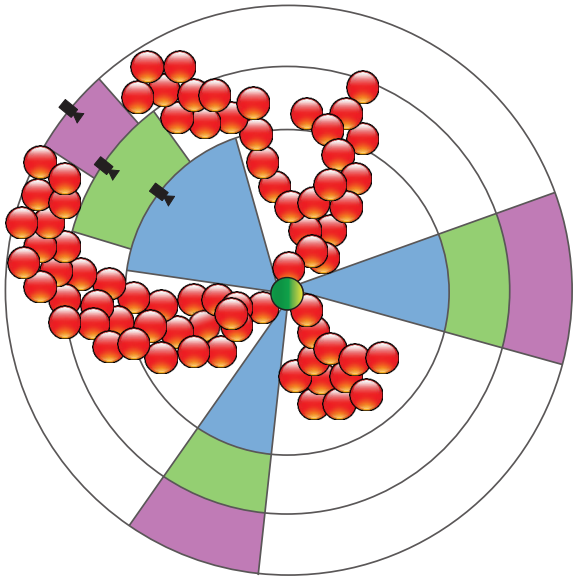
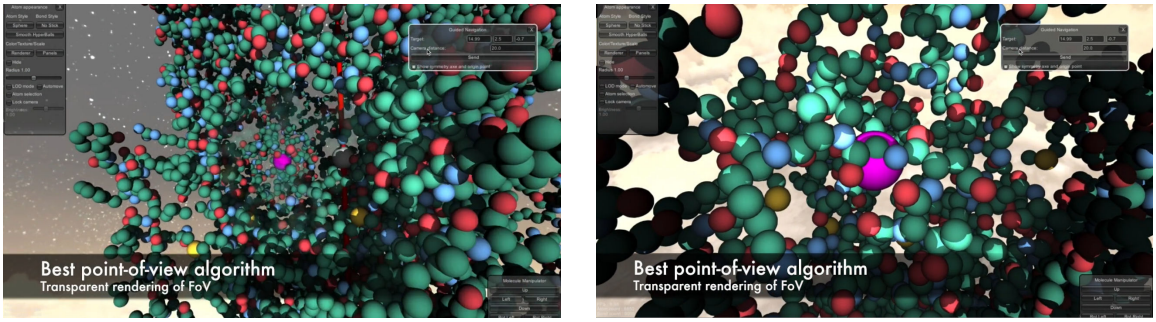


Figure 3.7 – *Schéma de l'algorithme de recherche du meilleur chemin de caméra pour atteindre un point de vue optimal. Le chemin qu'empruntera la caméra pour atteindre le point de vue optimal est représenté grâce aux symboles de couleur noire.*

Une transition progressive et automatique est ainsi mise en place. Pour calculer les différents points du chemin reliant la position initiale de l'utilisateur à la position optimale de visualisation de sa cible, l'algorithme développé précédemment est utilisé de façon répétée à différentes valeurs de d et appliqué à une aire restreinte de la surface de la sphère correspondant aux aires délimitées par les cercles vides identifiés lors de la première itération de l'algorithme. Comme illustré dans la Figure 3.7 nous obtenons un chemin de navigation constitué de positions obtenues après une succession de calculs des aires non obstruées (en couleur bleue, verte et magenta) à différentes distances de la cible. La cible est représentée en vert et ses voisins en rouge.

Cela permet également de laisser à l'utilisateur la possibilité de choisir entre la position offrant le plus large cône de vue sur la cible ou la position dont le chemin d'accès est le plus dégagé parmi les positions optimales recensées.

Le passage d'un point de vue optimal à un autre est le fruit de techniques d'interpolation classiques respectant s'il y a lieu, l'orientation préférentielle relative au type de symétrie de la structure.



(a) (b)

FIGURE 3.8 – *Implémentation de nos paradigmes de navigation guidée au sein de UnityMol pour l'exploration de GLIC (représenté en sphères) autour d'un axe de symétrie (en rouge, invisible dans la version publique). (a) Capture d'écran d'un point de vue de l'utilisateur pendant une exploration externe de la protéine GLIC. (b) Capture d'écran d'un point de vue de l'utilisateur pendant une exploration externe de la protéine GLIC.*

3.1.5.3 Grande densité atomique

Dans le cas où l'algorithme de point de vue optimal n'est pas suffisant pour permettre la visualisation d'une cible sans occlusions devant la caméra, nous avons mis en place une étape supplémentaire dans le processus de visualisation. Cette étape est une étape de modification structurelle du complexe moléculaire. Elle consiste à découper le complexe moléculaire en sous-régions distinctes et de permettre leur dissociation pendant l'exploration. Nous avons de nouveau utilisé la géométrie du complexe moléculaire comme base de décision pour ce découpage. Puisque la symétrie définit les monomères d'un complexe, nous pouvons facilement découper la structure pour que ces monomères puissent bouger de façon contrôlée et de manière à ne pas perdre la forme générale de la structure (voir Figure 3.9). L'utilisateur a de plus la possibilité, à chaque instant, d'étendre, rétrécir ou simplement déplacer les monomères du complexe en fonction de l'axe ou du centre de symétrie qu'il choisit (dans le cas où plusieurs axes/centres ont été renseignés). Un mode automatisé a également été ajouté permettant d'étendre ou rétrécir la structure suivant la position de l'utilisateur par rapport au complexe. Une distance maximum d'extension peut être imposée afin de préserver les informations d'interaction entre les monomères ou les sous-unités découpées. Cette adaptation structurelle permet la mise en lumière des interfaces enfouies entre les sous-unités et permet de calculer de nouveaux points de vue optimaux. Pour éviter toute surcharge dans le processus de modification structurelle, la configuration spatiale initiale du complexe peut être retrouvée à chaque instant.

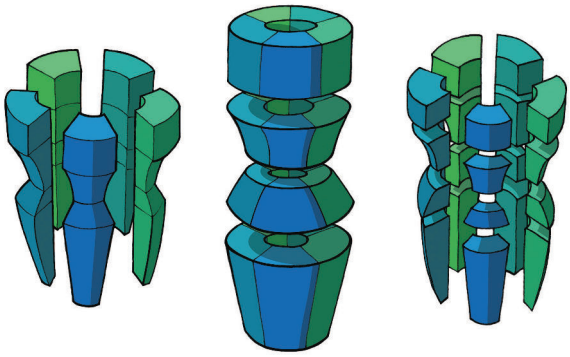


Figure 3.9 – *Illustration d'une protéine présentant un pore central constituant son axe de symétrie. À gauche, les monomères du complexe moléculaires sont écartés de l'axe de symétrie. Au centre, différentes régions d'intérêt ont été identifiées le long des monomères et le complexe est écarté de façon verticale. À droite sont combinées la séparation des monomères et des régions d'intérêt. Illustration de Davide Spalvieri.*

3.1.6 Évaluation par analyse hiérarchique de la tâche via la méthode HTA

Nous avons choisi d'évaluer nos paradigmes de navigation au moyen de la décomposition hiérarchique de deux tâches identifiées comme habituelles pour les experts scientifiques. Le choix de ces tâches a pu se faire grâce à la contribution d'ergonomes et au moyen de méthodes d'entretiens supervisés.

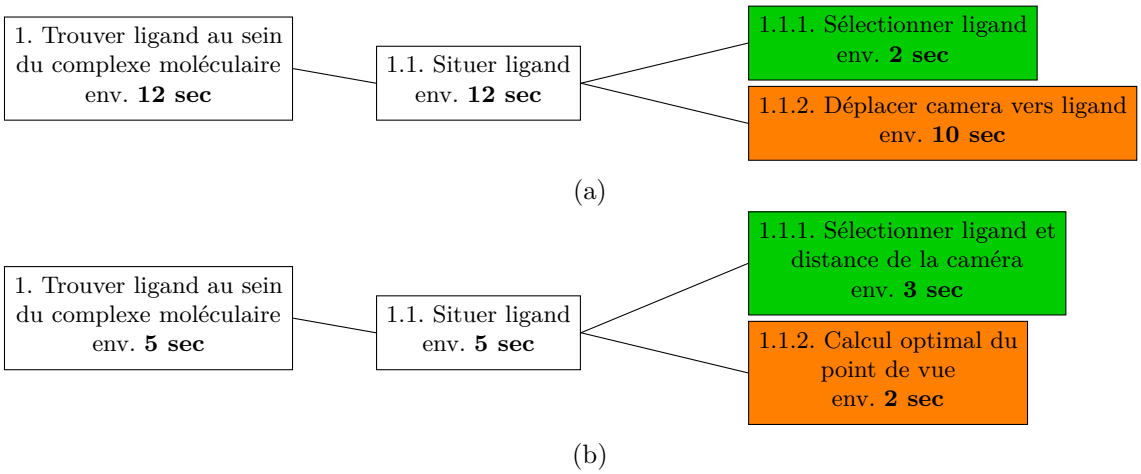


Figure 3.10 – Subdivision par HTA d'une tâche experte réalisée en visualisation moléculaire (a) classique et (b) guidée par la tâche et le contenu.

L'arbre de tâche HTA présenté ci-dessus décrit la manière de répondre à une tâche courante de visualisation moléculaire, la recherche d'un ligand au sein d'un complexe moléculaire. Nous utiliserons l'exemple de la liaison du bromoforme avec GLIC pour illustrer ce cas. Ce modèle, extrait grâce aux experts du domaine, donne une subdivision progressive de la tâche énoncée. Les subdivisions les plus externes pouvant finalement être évaluées d'un point de vue temporel. Nous comparons ici des conditions de navigation libres avec des conditions de navigation guidée utilisant les paradigmes expliqués précédemment. Nous ferons l'hypothèse d'une situation initiale identique dans le même logiciel de visualisation moléculaire pour les

deux conditions évaluées.

Cette première tâche peut être considérée comme simple et ne comporte après sa décomposition en sous-divisions que deux étapes. Tout d'abord la sélection du ligand pour le mettre en avant au sein de la scène moléculaire puis le déplacement de la caméra pour une vision rapprochée sur sa position et son environnement proche. Alors que cette tâche prend environ **12 secondes** lors d'une session de navigation sans contraintes, elle est réduite à **4 secondes** avec les paradigmes de navigation mis à disposition. L'algorithme de point de vue optimal est particulièrement utile ici puisqu'il permet, grâce à la sélection du ligand et une distance de caméra émise par l'utilisateur, de calculer automatiquement le meilleur point de vue sur le ligand ainsi que le chemin optimal pour y accéder. La sélection du ligand et de la distance de la caméra, dans le cas de l'utilisation de l'algorithme, sont des sous-tâches de temps équivalent, mises en avant en vert dans la Figure 3.10. La différence se fait sur la seconde sous-tâche (en orange) qui vient automatiser la recherche du meilleur point de vue et le positionnement de la caméra.

Nous avons identifié une seconde tâche métier type, dans la continuité de la première, et consistant à identifier, au sein d'un complexe protéique multimérique composé de 5 monomères, les différents acides aminés impliqués dans la liaison avec le ligand identifié précédemment. Cette identification des résidus en contact avec le ligand devra être effectuée pour chaque monomère du complexe.

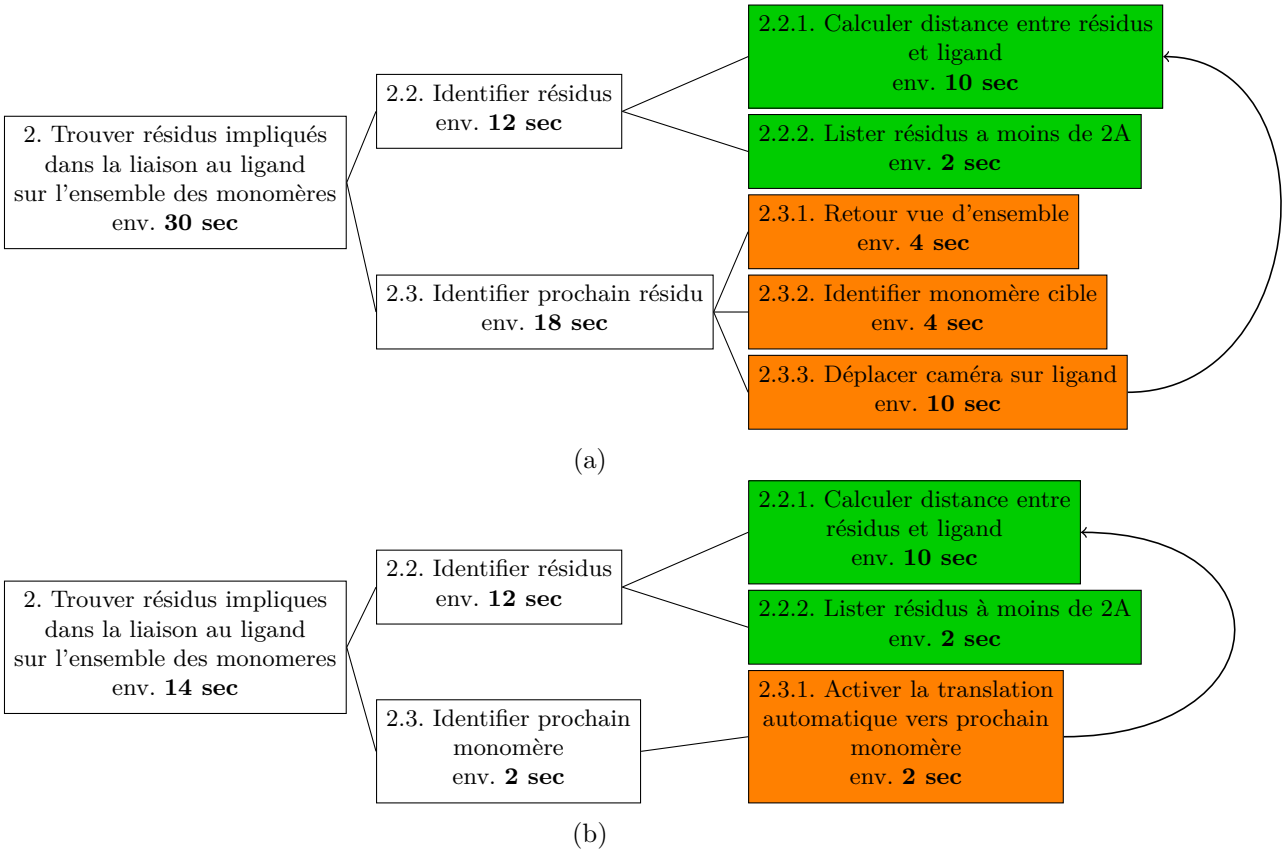


FIGURE 3.11 – Subdivision par HTA d'une tâche experte réalisée en visualisation moléculaire (a) classique et (b) guidée par la tâche et le contenu.

Cette tâche est divisible en deux étapes principales: L'identification des résidus en interaction et l'accès au monomère suivant. Suivant les conditions d'expérience initiales, la

subdivision de chacune de ces deux étapes est différente et décrite dans les deux arbres hiérarchiques ci-dessus. On obtient un temps d'exécution pour cette tâche plus élevé dans les conditions normales (environ **2 min 12 s**) que dans les conditions intégrant nos paradigmes de navigation (environ **1 min 8 s**). Alors que l'identification des résidus voisins est une étape constante (sous-tâche en vert dans la Figure 3.11), en terme de temps d'exécution, entre les deux conditions, l'accès à la cible (ici le ligand) et la recherche du monomère suivant sont les deux étapes cruciales où les paradigmes de navigation permettent un gain de temps substantiel, elles sont identifiées en orange dans l'arbre de la Figure 3.11. De façon moins tangible car difficilement mesurable, nos paradigmes permettent également d'assurer une vision sans occultations de la zone cible et donc pourraient être considérés comme facilitant l'étape 2.2.

3.1.7 Conclusion

La visualisation de données est un processus dans lequel l'expert est central et où l'observation et l'exploration des données amènent de nouvelles connaissances ou hypothèses au scientifique.

L'acquisition d'informations passe donc par un processus de navigation qui dirige l'utilisateur vers les événements qu'il désire observer ou vers des régions avec lesquelles il désire interagir. La navigation est un processus important en immersion car il conditionne le confort de l'utilisateur pour effectuer ces tâches. Le mal du simulateur, lié en partie à la perte de repères spatiaux est un phénomène qu'il est important de prendre en compte, mais qui fut absent des approches de navigation développées dans les logiciels de visualisation moléculaire.

Forts du constat de l'absence de développements de navigation spécifiques à la visualisation de données scientifiques et plus particulièrement moléculaires, nous avons proposé plusieurs paradigmes ou modes de navigation pouvant répondre à ce besoin. Nos efforts se sont portés sur l'utilisation de l'objet d'intérêt, ici un complexe moléculaire, comme base intelligente et dynamique de la création de chemins de navigation contraints, semi-automatisés ou automatisés suivant les modes de navigation considérés. Notre approche peut s'inscrire dans le processus de visualisation de données évoqué par Shneidermann et introduisant une échelle de granularité partant de l'exploration globale des données à l'observation précise de certains phénomènes ponctuels et locaux.

Nos développements se sont basés sur le contenu et la tâche, mais se sont voulus indépendants de toute considération de contexte et nos paradigmes peuvent être utilisés sur n'importe quelle plateforme et sont complètement indépendants des périphériques d'interaction utilisés. Cette généricité permet une utilisation à la fois au sein d'EV immersifs, mais également de stations de travail fixes.

La généricité de notre approche peut également être soulignée par le nombre d'applications potentielles auxquelles elle peut répondre. En effet, les domaines mettant en jeu une visualisation de données abstraites présentant une densité importante d'informations pourraient bénéficier de notre étude. La mécanique des fluides, les sciences des matériaux, la climatologie, la médecine ou l'astronomie sont des exemples de disciplines où les données ne comportent pas toujours une orientation spatiale précise, mais où la visualisation joue un rôle central.

Des méthodes ergonomiques comme la HTA nous ont finalement permis de mettre en évidence les apports en terme de performance et de simplification des processus de nos paradigmes de navigation. L'apport de ces paradigmes est indéniable pour les tâches explorées, il est certainement extensible à d'autres tâches expertes impliquant une exploration avancée des caractéristiques structurelles d'une protéine.

3.2 La visualisation adaptative au service de la visualisation moléculaire

Nous avons souligné dans le chapitre 1 l'importance que prenait la visualisation et la représentation de structures 3d de molécules pour la compréhension de leur fonctionnement. En plus de permettre d'appréhender leur rôle biologique, la visualisation moléculaire est également un excellent vecteur de communication dans le monde scientifique. Fort de ce constat, nous nous sommes intéressés aux améliorations possibles que nous pourrions apporter (1) aux méthodes de contrôle visuelles de simulations moléculaires distantes, (2) aux moyens de communication scientifique actuels qu'offre la visualisation moléculaire.

3.2.1 Rapprocher l'expert de sa simulation moléculaire

Une simulation moléculaire, consistant en une modélisation d'un complexe moléculaire de façon informatique, peut parfois s'étaler sur un laps de temps conséquent (plusieurs jours voire semaines pour les plus longues). Les ressources informatiques mobilisées sont importantes et une mauvaise paramétrisation de la simulation peut ainsi entraîner une perte de temps et de ressources. Des moyens de contrôle peuvent permettre de modifier d'éventuels mauvais paramètres ou de terminer une simulation avant son temps d'exécution total initialement prévu. Parmi ces moyens de contrôle, la visualisation de la structure du complexe moléculaire est une aide importante pour juger de l'état d'une simulation à un instant t. Cette visualisation passe par la génération d'une photographie du complexe moléculaire au moyen de logiciels de rendu experts comme ceux cités dans la section 1.2.3.

3.2.2 L'évolution des méthodes de communication du monde scientifique

Les moyens de communication scientifique se sont justement, en parallèle des moyens de communication standards, développés de façon importante sur les appareils mobiles type smartphone et tablettes. Les journaux physiques et papiers sont de plus en plus rares et la grande majorité des articles scientifiques sont numérisés et accessibles en ligne. Quelques évolutions ont enrichi les contenus pour les rendre interactifs [7]. Les structures moléculaires peuvent retrouver une 3e dimension grâce à plusieurs méthodes [99, 136] qui dépendant cependant de prérequis logiciels minimums pour être complètement fonctionnels ou requièrent des programmes spécifiques (comme la version 3d améliorée d'Adobe Acrobat Reader [1] ou les plug-ins spécifiques comme le navigateur ICM [2]).

Les éditeurs scientifiques ont d'ailleurs rapidement commencé à publier sur les périphériques mobiles: Cell Press [3], ACS [4], Nature Publishing [5], Science [6] ou PloS [7] fournissent des applications mobiles dédiées pour leurs publications, dont certaines se révèlent avoir une utilité scientifique significative [131].

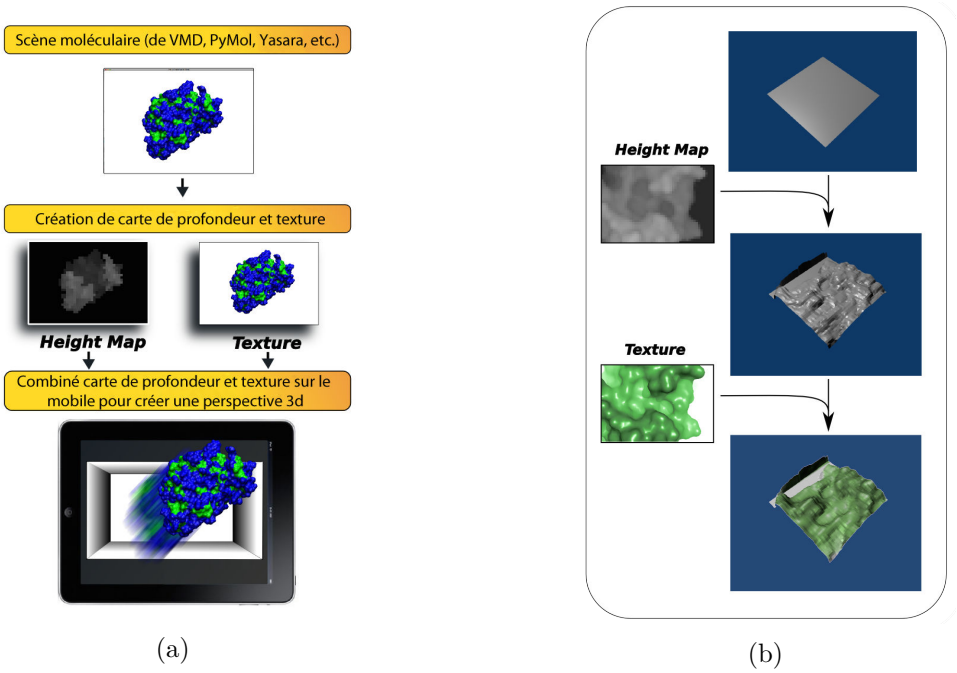
1. http://get.adobe.com/reader/
2. http://www.molsoft.com/icm_browser.html
3. http://www.cell.com/journalreader
4. http://pubs.acs.org/page/tools/acsmobile/index.html
5. http://www.nature.com/mobileapps
6. http://content.aaas.org/mobile
7. blogs.plos.org/everyone/2010/09/16/plos-reader-2-0/

3.2.3 Vers une immersion sur dispositifs mobiles

Pour répondre à ces deux objectifs conjoints, d'une part de fournir un outil de *reporting* de résultats d'une simulation en cours, mais aussi un outil dédié à la communication scientifique, nous avons conçu une approche permettant de réduire et de résumer les données de simulation et de modélisation volumineuses. Notre approche se distingue par le fait que nous utilisons les fonctionnalités de tracking des dispositifs mobiles, d'une part pour produire des rendus dont l'utilisateur peut percevoir la profondeur grâce à la visualisation adaptative, et d'autre part pour immerger l'utilisateur à l'intérieur d'une scène moléculaire virtuelle.

3.2.3.1 Donner à percevoir la profondeur sur dispositif mobile grâce à la visualisation adaptative

Notre approche se base sur l'utilisation d'images 2d, beaucoup plus légères que des modèles 3d et permettant un certain degré de personnalisation de la part des utilisateurs. Indépendante des formats de fichier utilisés dans les logiciels de visualisation, notre application ne nécessite qu'un périphérique mobile pour fonctionner et donner à percevoir la profondeur de contenus 3d moléculaires.



(a) (b)

Figure 3.12 – *(a) Schéma du fonctionnement de notre application mobile. (b) Processus d'obtention d'un objet 3d à partir de 2 images.*

Dans notre approche, deux images sont construites pour donner à percevoir la profondeur, une image de **texture** et une **carte de profondeur** (cf. Figure 3.12a). L'image de *texture* est une capture simple de la scène 3d moléculaire alors que la *carte de profondeur* embarque les informations de distance à l'écran pour l'image de texture. L'information de distance est codée grâce à un gradient de couleur grise. Ce gradient de gris correspond à la distance entre chaque point de la scène moléculaire 3d et l'écran. De ce fait, plus un objet est distant de la caméra (ou plus simplement de l'écran de la station de travail) et plus la couleur est sombre. Une fois que les deux images ont été générées, elles doivent être transférées sur le périphérique

mobile où elles sont lues et traitées par notre application. Un navigateur de fichier intégré permet à l'utilisateur de trouver ses images et de les charger.

Le traitement effectué par l'application permet tout d'abord de créer un objet 3d à partir de la carte de profondeur, objet dont la surface est ensuite colorée par la texture associée (3.12b). Le niveau de gris de chaque pixel de la carte de profondeur est ainsi interprété en tant que hauteur et correspond à un sommet de la surface 3d générée.

Sans disposer de rendu stéréoscopique sur dispositif mobile, la profondeur de la scène 3d générée est donnée à percevoir à l'utilisateur à l'aide de techniques de visualisation adaptative. L'orientation du périphérique, mesurée par un accéléromètre ou un gyroscope embarqués dans la majorité des périphériques mobiles d'aujourd'hui permet de présenter à l'utilisateur d'observer de manière active sa molécule selon plusieurs points de vue. Il est possible par ailleurs d'utiliser jusque 4 images (2 fois deux images contenant une carte de profondeur et une texture associée) afin de créer des scènes plus complexes avec plusieurs plans. Cette possibilité permet notamment l'ajout d'informations complémentaires comme des annotations ou de situer la molécule d'intérêt dans son environnement présenté en arrière plan. Plusieurs exemples sont fournis avec l'application et permettent à l'utilisateur d'appréhender les rendus 3d possibles et d'avoir une première prise en main des possibilités qu'offrent notre approche, ces exemples sont présentés dans l'Annexe A. Des captures d'écran de l'application sont aussi disponibles dans l'Annexe B.

3.2.3.2 Vers une véritable immersion sur dispositif mobile

Un deuxième mode de visualisation permet d'importer directement des fichiers d'objets 3d dans l'application. Avec ce type de données 3d, un seul fichier est nécessaire pour créer une perspective 3d complète de l'objet. Après importation, il est possible de manipuler l'objet, de le tourner, le déplacer en avant ou arrière. Cette visualisation se rapproche de ce qui peut se retrouver dans les visualiseurs moléculaires pour périphériques mobiles. Pour aller plus loin dans l'immersion, l'utilisateur aura également la possibilité de changer l'échelle de la molécule et de plonger à l'intérieur et d'activer une visualisation plus interactive. Il utilisera en effet son périphérique comme une fenêtre sur le monde virtuel et le manipulera en orientant le périphérique à la main et en l'utilisant comme une vue adaptative sur un monde virtuel présent tout autour de lui. Basés sur le gyroscope du périphérique, l'objet 3d est visualisable à 360 degrés, mais à partir d'une position fixe.

Pour une étape supplémentaire vers une immersion plus importante, un rendu stéréoscopique peut être effectué au sein d'un support de casque, à la manière de ce que proposent de nombreux acteurs de la RV aujourd'hui n'ayant pas franchi le pas de la conception de casques de RV, mais fournissant des supports pour l'utilisation de smartphones (voir section 2.1.1.1). Dans cette configuration, un simple mouvement de tête permet de regarder une zone différente de la molécule (voir Figure 3.13). Cette approche est rendue possible par un mode stéréoscopique *side-by-side* intégré au sein de l'application.

3.2.3.3 Bilan

La préparation des images utilisées au sein de l'application est simple et rapide, des tutoriels sont en plus préparés pour aider les utilisateurs à effectuer les actions nécessaires. Notre approche est indépendante de la plateforme de visualisation et la majorité des visualiseurs moléculaires permettant de générer un rendu graphique de structures permettent de générer les images demandées. Il est également possible d'utiliser des outils génériques tels que

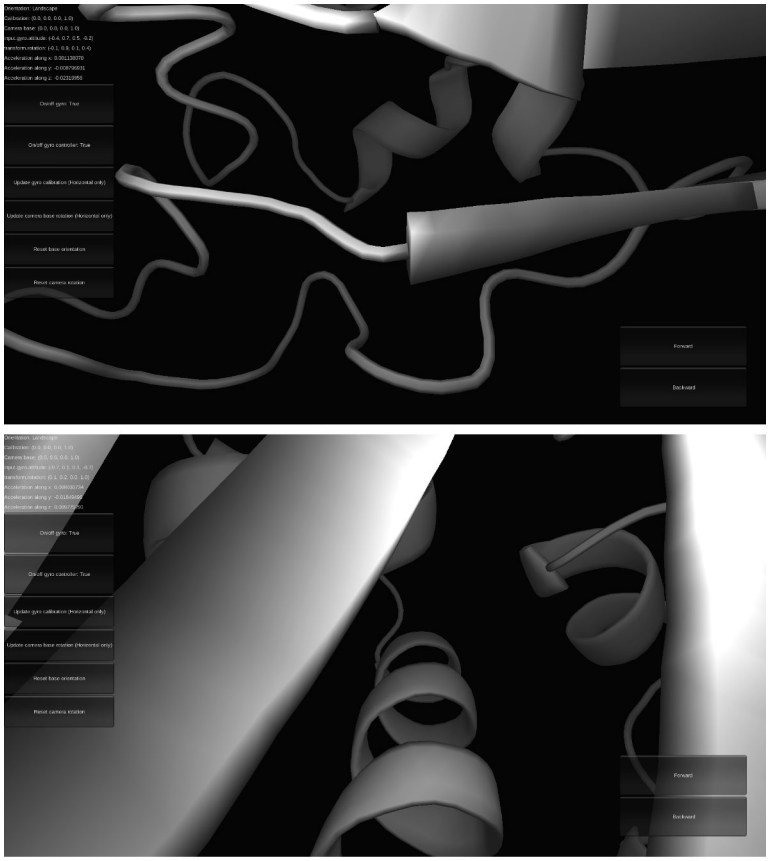


Figure 3.13 – *Captures d'écran du mode d'exploration de modèles 3d au sein de l'application. La différence de point de vue entre les deux captures d'écran est le résultat d'une rotation du périphérique mobile tenu par l'utilisateur.*

Povray [8] ou Paraview [9]. La création de vues complexes est rendue possible grâce à la notion d'arrière-plan et de premier plan permettant ainsi de mettre en valeur certains aspects d'une scène virtuelle où la seule structure ne serait pas assez illustrative d'un phénomène particulier.

La performance de l'application permet son utilisation sur la plupart des smartphones et tablettes Android et iOS de moins de 5 ans et la quasi-intégralité des derniers smartphones/-tablettes. Notre application est gratuite et le code source est disponible à la demande.

Les limitations de notre application résident tout d'abord dans la résolution des écrans des périphériques. Les cartes de profondeur sont pour le moment limitées à des résolutions assez basses et peuvent faire apparaître des artefacts quand les détails de la scène sont trop fins. La limitation de la résolution des cartes de profondeur considérées dans l'application est due à la limitation technique du moteur Unity3D qui ne peut gérer des objets 3d de plus de 65 536 sommets. Ceci implique donc l'utilisation de cartes de profondeur de 255*255 pixels de résolution au maximum. Il n'existe par contre aucune limitation dans la résolution de la texture tant que celle-ci est aux mêmes dimensions que la carte de profondeur. L'utilisation d'images 2d ne permet évidemment pas un contrôle total de l'objet 3d généré et de sa manipulation dans l'espace. Notre mode d'importation d'objets 3d permet de pallier en partie cette limitation, mais la performance du rendu est moindre, tout autant que la facilité de

8. www.povray.com
9. www.paraview.org

création de nouvelles structures.

3.3 Conclusion

Notre travail s'est axé autour de deux axes forts, mis en avant dans les sections 3.1 et 3.2 de ce chapitre. Nous venons tout juste de voir comment de simples techniques empruntées à la RV pouvaient fournir d'utiles outils de représentation améliorée permettant d'appréhender les structures moléculaires complexes 3d de façon plus intuitive que les représentations 2d actuelles. Nous avons également mis en avant nos efforts pour pallier certains manques méthodologiques préjudiciables au sein des scènes moléculaires pour une expérience immersive optimale. La navigation était pour cela au cœur de notre approche et nous avons mis en place une navigation contrainte par le contenu et la tâche.

Maintenant que certains outils ont été mis en place, nous pouvons nous intéresser à l'intégration du processus d'étude de biologie structurale au sein de dispositifs immersifs. Cette intégration ne peut pas se faire seulement par la mise en place d'outils de navigation et doit passer par une réflexion autour de l'homogénéité des espaces qu'elle met en jeu.

Chapitre 4

Visual Analytics et approches sémantiques pour la biologie moléculaire

Nous exposerons dans ce chapitre les pré-requis identifiés comme indispensables pour l'intégration d'outils de visualisation et d'analyse de biologie structurale au sein d'un même espace de travail immersif. Nous définirons pour cela l'un des domaines mettant en jeu cette intégration, le domaine du *Visual Analytics*, et identifierons des axes de développement pouvant profiter à notre travail. Parmi ces axes de développement, nous démontrerons la nécessité d'utiliser un formalisme de représentation des connaissances en biologie structurale. Plusieurs formalismes de représentation seront finalement discutés et le plus pertinent pour notre domaine d'application sera décrit.

Sommaire

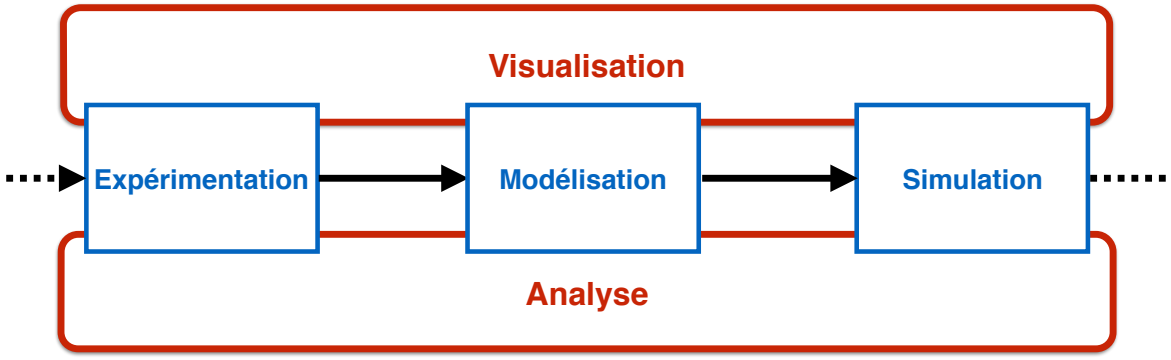


Figure 4.1 – *Schéma du processus itératif d'étude d'une structure moléculaire en biologie structurale. Notre focus est ici les liens unissant les outils de visualisation et d'analyses.*

Introduction

Les tâches conjointes d'analyse et de visualisation des données générées par simulation moléculaire ont pour objectifs, d'une part de confirmer la cohérence et la validité d'un résultat théorique de simulation au regard des résultats expérimentaux scientifiques établis, et d'autre part d'aider l'expert à formuler de nouvelles hypothèses à confirmer ensuite de manière expérimentale. Nous avons vu qu'au sein du processus d'étude de structures moléculaires, ces deux espaces de travail mobilisant des outils variés sont omniprésents dans chacune des étapes du processus, depuis l'expérimentation, jusqu'à la simulation en passant par la modélisation et la visualisation (voir Figure 4.1).

Nous rappelons que les objectifs de ce travail sont de minimiser les données échangées entre les étapes de visualisation et d'analyse, en rapprochant les étapes de simulation, de visualisation, d'analyse tout en permettant à l'expert d'effectuer des visualisations et des analyses à la demande sur le lieu de simulation. Il s'agit aussi de répondre aux contraintes propres aux environnements immersifs pour les rendre opérationnels pour le domaine de la biologie moléculaire, en favorisant les interactions directes dans un contexte interactif commun. Pour répondre à ces objectifs, nous avons emprunté des concepts issus des approches de type *Visual Analytics*, en mobilisant des concepts de représentation sémantique des connaissances. Nous présentons donc dans ce chapitre un aperçu des concepts et des techniques relatifs à ces deux domaines, de leurs applications aux domaines de la biologie, prérequis nécessaire pour présenter notre contribution dans le dernier chapitre adressant les deux objectifs précités.

4.1 Visual Analytics : définition, outils et applications

Lorsque l'on s'intéresse à la place des approches par *Visual Analytics* dans les disciplines de biologie et de médecine (voir Figure 4.2), on s'aperçoit que ce domaine émergent connaît une progression identique à celle de la RV. Nous expliquons cette croissance par l'observation partagée de la communauté du besoin de trouver de nouvelles approches pour répondre à la problématique centrale relative à la quantité de données à visualiser et à analyser.

L'association par des liens interactifs, de plusieurs représentations visuelles de données souvent hétérogènes, dans un même espace de travail, nécessite l'introduction de nouvelles techniques de visualisation définies dans le cadre d'un domaine de recherche émergent appelé *Visual Analytics*.

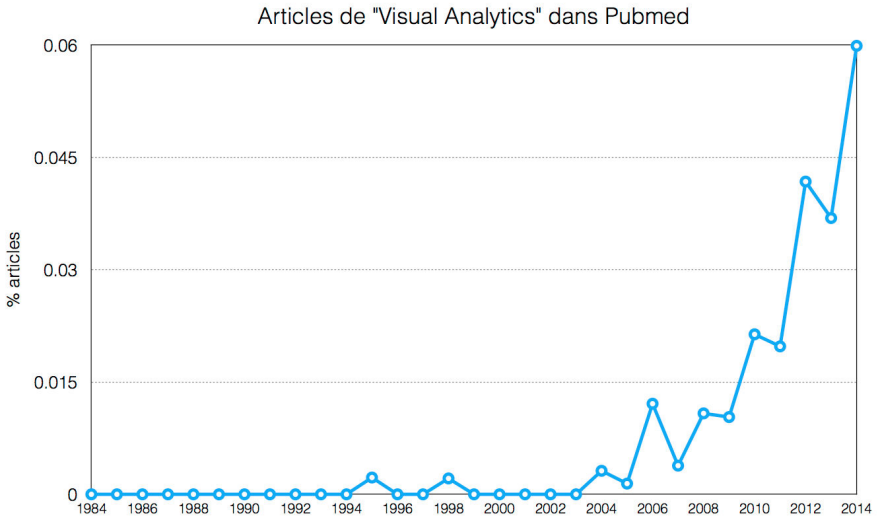


FIGURE 4.2 – *Évolution du pourcentage d'articles de PubMed où le terme «Visual Analytics» est retrouvé soit dans le titre soit dans le résumé au cours des 20 dernières années.*

4.1.1 Définition

Cette discipline récente a pour but de faciliter l'analyse visuelle de données complexes et/ou scientifiques et se définit par le «raisonnement analytique au travers d'interfaces visuelles interactives» [42]. Elle se place à la frontière de nombreux domaines, comme celui de la visualisation, de l'interaction homme-machine et de perception, afin de visuellement mettre en avant des relations entre les informations, difficiles à percevoir lors de l'utilisation cloisonnée de techniques classiques issues de ces différents domaines. Elle s'inspire donc par plusieurs niveaux aux études d'interactions homme-machine (IHM) dont la réalité virtuelle s'inspire également [5]. De la même manière qu'en IHM, elle se repose sur des outils et des techniques basées sur des considérations ergonomiques et perceptuelles. Son principal but est de mettre l'être humain au centre d'une boucle de décision qui sera facilitée par la mise en relation de données de différentes natures et de différentes sources. Ce contexte de travail généré par le *Visual Analytics* doit être cohérent pour le chercheur. C'est à ce niveau que les domaines de la perception et des études cognitives interviennent afin d'assurer une pleine compréhension et utilisation de l'espace de travail. La VA se place comme un catalyseur de domaines comme le raisonnement analytique, l'interaction, la transformation et la représentation des données pour leur visualisation et analyses. Elle a, par beaucoup de facettes, des points communs importants avec la **visualisation d'information** et la **visualisation scientifique**. Leurs limites respectives et leurs frontières sont assez floues, mais elles peuvent être distinguées de la manière suivante :

— **La visualisation scientifique** s'intéresse aux données possédant une structure géométrique 3d intrinsèque (images médicales, évolution de fluides, structures atomiques par exemple).

— **La visualisation d'informations** concerne la représentation de données abstraites la plupart du temps à travers des représentations graphiques 2d.

— **Le *Visual Analytics*** concerne le couplage par des méthodes interactives de plusieurs représentations issues soit de la visualisation scientifique soit de la visualisation d'informations.

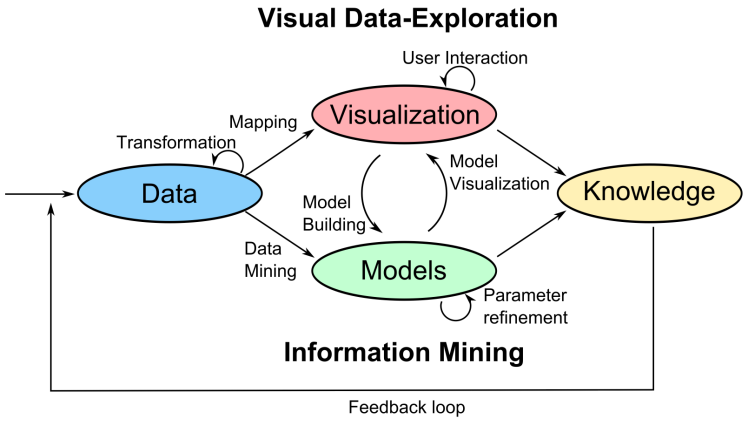


Figure 4.3 – *Schéma illustratif du processus de Visual Analytics tel que proposé par Keim et al.. Les différentes étapes sont illustrées par des ovales, les transitions par des flèches.*

L'approche par *Visual Analytics* vise à combiner différentes modalités de visualisation afin de faciliter l'extraction de connaissances par l'expert. Ce processus, expliqué par Keim et al. [88], forme ainsi une boucle d'analyse fermée (voir Figure 4.3) susceptible de raccourcir considérablement le processus standard et plus séquentiel des étapes de visualisation moléculaire et d'analyse de biologie structurale, présenté dans la Figure 4.1. Le schéma proposé par Keim peut être transposé dans sa quasi-intégralité au sein de notre conception du processus de biologie structurale, il est très utile pour comprendre la relation étroite et bilatérale entre *visualisation* et *analyse* (respectivement «**Visualisation**» et «**Models**» dans le schéma de Keim) que met en place le *Visual Analytics*.

Plusieurs aspects de ce schéma trouvent donc des échos au sein de notre travail. Nous retrouvons d'abord la volonté de mettre l'utilisateur au centre de la boucle décisionnelle au travers, entre autres, des processus de «**User Interaction**» et de «**Parameter Refinement**» qui placent l'utilisateur comme modulateur des données qu'il désire mettre en avant au sein des étapes de *visualisation* et d'*analyse*. C'est une caractéristique importante du *Visual Analytics* que nous avons également mis en avant précédemment comme pré-requis dans l'établissement d'un processus d'étude de données moléculaires simplifié et optimisé pour la prise de décision de l'expert et l'extraction de nouvelles connaissances.

Certaines données et méthodes interactives de visualisation (détaillées dans la section 4.1.2) permettent l'*extraction de nouvelles connaissances* («**Knowledge**» dans le schéma) sans formatage particulier des données d'entrée. Cependant l'étape de *visualisation* seule n'est souvent pas suffisante pour extraire de telles connaissances et elle doit être couplée à une étape d'*analyse* (caractérisé par les «**Models**» d'analyse que Keim décrit). Les communications entre ces deux étapes apparaissent dans le schéma sous les noms «**Model Building**» et «**Model Visualisation**» et décrivent une communication étroite entre les deux étapes qu'elles relient. L'extraction de connaissances passe ici par des allers-retours entre les outils de visualisation et ceux d'analyses. Chaque analyse produira des informations qui seront alors soit directement interprétées par l'utilisateur, soit envoyées aux outils de visualisation pour aider à leur compréhension («**Models Visualisation**» dans le schéma).

Le processus de «**Transformation**» des données hétérogènes en données homogènes présenté lors de l'étape «**Data**» du schéma de Keim est une étape indispensable en biologie structurale. En effet, les données en entrée peuvent être de nature géométrique (coordonnées 3d des atomes de la structure) ou brutes (valeurs d'énergie, de polarité, etc.), elles doivent

être transformées de façon à être utilisées simultanément. Nous verrons par la suite que cette transformation des données peut être effectuée grâce à une approche par représentation des connaissances (voir section 4.2).

Les étapes d'*analyse* et de *visualisation* impliquent donc toutes deux l'intervention de l'expert et implémentent donc une couche d'interaction permettant de guider les méthodes utilisées vers le cercle d'information que l'utilisateur considère comme important. Lorsque de nouvelles connaissances sont extraites, elles sont retournées comme données d'entrée dans la boucle de *Visual Analytics* («**Feedback Loop**»)et les étapes précédentes sont répétées.

La création de nouvelles connaissances passe par une exploitation plus efficace des capacités cognitives de l'humain, supportées par des techniques de *Visual Analytics* à travers 6 moyens principaux : [32, 42] :

1. en exploitant plus efficacement les ressources cognitives du sujet
2. en réduisant le temps de recherche, grâce à la représentation résumée de grandes quantités de données
3. en exploitant les capacités humaines de reconnaissance de motifs, en organisant spatialement l'information de manière appropriée
4. en soutenant l'inférence perceptive pour faciliter par la mise en relation des données,
5. par la surveillance perceptive d'un grand nombre d'événements potentiels,
6. en fournissant des représentations graphiques manipulables et dynamiques, à la différence des graphes statiques, permettant une exploration interactive permettant de couvrir de manière interactive tout l'espace dans lequel sont définies les données analysées.

Ces recommandations ergonomiques guident le développement de nouvelles générations d'applications dédiées au *Visual Analytics*.

4.1.2 Outils et techniques

L'implémentation de ce couplage interactif étroit entre différentes modalités de visualisation en *Visual Analytics* passe par plusieurs classes d'approches :

1. «Overview+Detail»: Cette technique met en relation plusieurs vues, synchrones ou asynchrones et dont l'une présente une vue globale du sujet,le tout dans des espaces visuels distincts. L'absence de synchronisation peut se retrouver dans la vue globale ou dans la vue détaillée. Une interaction dans l'une ou les autres des vues n'entraîne pas obligatoirement un changement dans les autres vues. Cependant, dans la plupart des cas, les vues sont synchrones et une cohérence est assurée entre les données affichées afin de pouvoir lier leur contenu. Une application connue de cette technique est retrouvée dans les programmes de visualisation de cartes comme illustrée dans la Figure 4.4. Ces visualiseurs de cartes géographiques présentent deux échelles spatiales différentes dans deux fenêtres distinctes. Seule l'interaction dans la fenêtre de contexte globale (la plus grande / principale) a un effet dans la seconde fenêtre.
2. «Zooming»: Cette technique se base cette fois sur une séparation temporelle des vues. Un zoom avant amènera une vue plus détaillée d'une scène alors qu'un zoom arrière apportera une vue plus globale. La transition entre les différentes échelles de vue peut être continue ou discrète et présenter une animation ou non. L'un des problèmes majeurs de cette technique est la difficulté pour l'utilisateur d'inverser une action de zoom. En effet, la notion de défaire ou d'annuler une action en informatique fait souvent

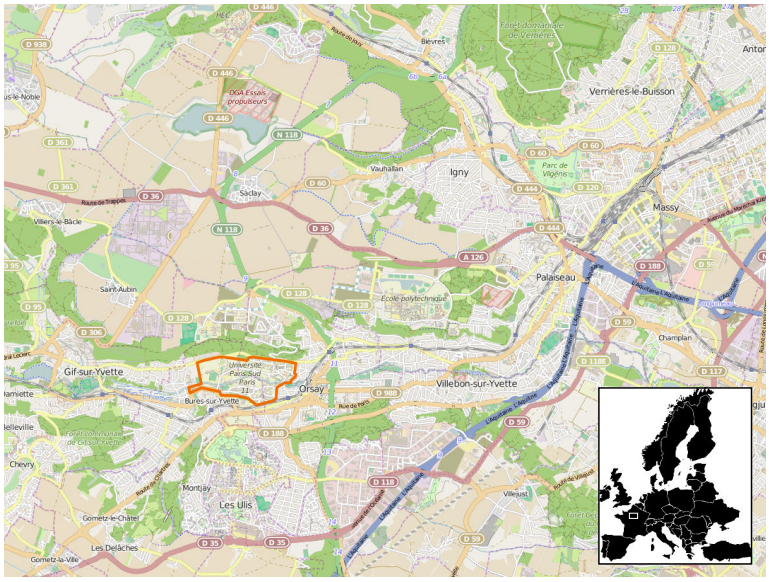


Figure 4.4 – *Exemple d'application utilisant la technique d'«Overview+Detail» où un fond de carte principal possédant une certaine concentration d'informations et ciblant une région particulière (en couleur) est mis dans un contexte plus large et moins détaillé sur un deuxième fond de carte annexe (en noir et blanc).*

référence à des actions ayant eu pour effet de modifier l'état des données et non la vue de l'utilisateur.

3. «Focus+Context»: Cette technique permet à l'utilisateur de se concentrer sur une partie intéressante des données visualisées sans perdre le contexte global dans lequel s'inscrivent ces données. Les informations présentées dans le contexte global ne sont pas nécessairement identiques à celles présentées en détail, mais les deux échelles d'information peuvent être associées à travers un affichage dynamique simple. Par exemple, si différents sets de données ont leurs entrées liées par au minimum une propriété équivalente, la sélection d'un point dans un set particulier sélectionne également tous les points correspondants à ce point dans les autres sets. La Figure 4.5 nous montre par exemple la sélection d'un ensemble de modèles dans un set de données de simulation moléculaire. Au moment de la sélection, toutes les représentations des modèles Y dans les espaces de visualisation ou d'analyse sont également sélectionnées et visuellement mises en avant. Cet exemple spécifique d'utilisation du focus+context est appelé *brush-and-link* et se retrouve dans de nombreux logiciels de visualisation des données sous forme de graphiques simples.

4. «Cue-based technique»: Cette technique vise à mettre en avant un sous-ensemble de données intéressantes à l'utilisateur en intervenant sur la façon dont ces données sont affichées. À la différence des autres schémas décrits précédemment, elle n'intervient pas sur la taille des données, mais peut être utilisée en conjonction de chacune des techniques précédentes. Elle regroupe des méthodes de brouillage visuel des données, d'ajout de labels ou de mise en place d'éléments décoratifs divers pour attirer l'attention de l'utilisateur.

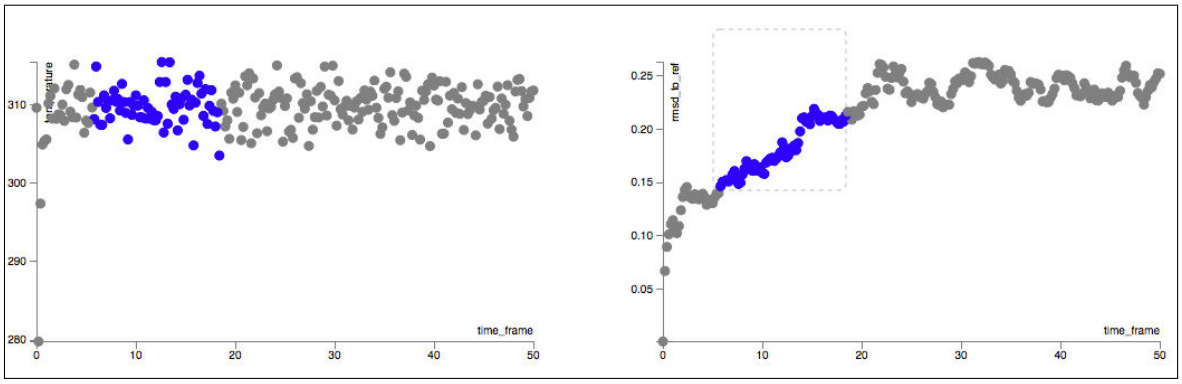


FIGURE 4.5 – *Illustration de la sélection simultanée et synchronisée d'un ensemble d'individus dans les deux graphiques exposés. La sélection est effectuée dans le graphique de droite, mais se répercute dans le graphique de gauche.*

4.1.3 Applications en biologie structurale

Les approches de type *Visual Analytics* ont été ponctuellement appliquées à la biologie structurale, mais possèdent encore un certain retard dans leur association systématique au processus d'analyse. Le *Visual Analytics* propose pourtant différentes modalités de couplage qui pourrait être adapté à la visualisation scientifique et à la visualisation d'information issues des analyses *a posteriori* des résultats de simulation moléculaires [67]. En reprenant le schéma du *Visual Analytics* proposé par Keim *et al.* [88], exposé dans la Figure 4.3, il est facile de considérer la sortie d'une simulation moléculaire (des coordonnées 3D de modèles + propriétés physico-chimiques) comme données d'entrée directement exploitable en utilisant des rendus visuels adaptés à la nature des données à observer (rendu 3D navigable pour les coordonnées 3D des complexes moléculaires, plots et graphes pour les données d'analyses, etc.). Une telle approche suppose cependant d'utiliser un formalisme commun à toutes les données manipulées, afin d'être en mesure d'appliquer des techniques de Focus+Context ou de Overview+Detail, supposant la mise à jour synchronisée des espaces de visualisation et d'analyse lors des événements d'interaction [148, 91].

On retrouve quelques caractéristiques de ce qu'apporte le *Visual Analytics* dans des applications comme KING [36] ou DIVE [141] cherchant à accroître la quantité d'informations disponibles pour l'utilisateur tout en maintenant un lien fort entre les différentes représentations des données affichées. Une bonne présentation des informations dans leur contexte et la mise en place de connexions interactives apportent des environnements puissants d'aide à la décision telle que Cytoscape [151, 55]. Cette plateforme logicielle permet la visualisation de réseaux complexes, notamment biologiques, associée à de nombreux types de données tels des structures 3d protéiques, graphiques d'analyses, etc. Le prérequis de ces environnements particuliers suppose une description formalisée homogène des données manipulées afin d'être en mesure de construire et de fournir à l'utilisateur des liens interactifs entre des concepts analogues présentés selon des modalités différentes.

Il n'est pas possible de générer de tels liens entre des données provenant de sources tant hétérogènes sans devoir posséder un certain niveau d'organisation et de structuration des concepts qui seront utilisés. Cette structuration passe par une étape préliminaire de représentation des connaissances venant poser une définition des différentes notions manipulées. Nous verrons dans la prochaine section comment se caractérise une telle représentation des connaissances et quels sont les formalismes à notre disposition.

4.2 Représentation des connaissances

Dans les sciences informatiques, la représentation des connaissances est souvent associée à la notion d'ontologie. Une ontologie se définit comme un ensemble structuré de concepts et de relations permettant de décrire tout ou une partie d'un domaine.

La mise en place d'ontologies afin de standardiser les connaissances dans des domaines de recherche scientifique a connu un développement spontané et important à partir de la fin des années 90 [149, 11]. Une ontologie est *l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances.*[1]. Une ontologie se définit comme un vocabulaire cherchant à classer des concepts et définir des relations entre ces concepts pour un domaine spécifique. Ce classement hiérarchique se doit d'être interprétable à la fois par les machines et les humains. Tout d'abord afin d'être intégré dans des traitements informatiques et afin de permettre aux experts d'élaborer ces ontologies.

4.2.1 Choix du formalisme

Afin de correctement manipuler les connaissances et les données sous-jacentes, il convient de choisir un formalisme adapté à notre objectif. Le formalisme utilisé pour représenter des connaissances dépend à la fois du domaine d'application (ici la biologie structurale), des opérations à mettre en oeuvre sur ces connaissances (lier les faits entre eux, extraire de nouvelles connaissances, obtenir des descripteurs pertinents pour l'aide à la décision). Il est possible de distinguer deux classes d'approches, les approches non logiques et les approches logiques.

Les approches non logiques regroupent des logiques descriptives comme les *réseaux sémantiques* et les *Graphes Conceptuels* (GC). Les approches logiques regroupent les *logiques classiques* comme les logiques propositionnelles, de 1er ordres ou de 2ème ordres ainsi que les *logiques de description*.

4.2.1.1 Réseaux sémantiques

Les réseaux sémantiques sont destinés à la modélisation hiérarchique de connaissances sous forme de graphes marqués. Un réseau sémantique contient des concepts, représenté par des noeuds et des relations, représentés par des liens (ou arcs) étiquetés entre chaque noeud. Ils ont été dans les années 60 par Quillian et Collins [39] comme base pour la représentation taxonomique. Ils se caractérisent par des relations binaires entre les concepts de type *est-un*, *a* ou *type-de*.

4.2.1.2 Graphes Conceptuels

Les graphes conceptuels, introduits par Sowa en 1984 [161], tiennent leur origine des réseaux sémantiques [102] et sont un moyen de formaliser les connaissances [35]. Ce modèle est mathématiquement fondé sur la logique et la théorie des graphes. Cependant, pour raisonner à l'aide de GC, deux approches peuvent être distinguées : (1) considérer les GC comme une interface graphique pour la logique et donc raisonner à l'aide de la logique et (2) considérer les GC comme un modèle de représentation à part entière disposant de ses propres mécanismes de raisonnement fondés sur la théorie des graphes. Dans ce modèle, les concepts et les relations sont des noeuds reliés par des arcs orientés comme illustrés dans la Figure 4.6.

1. https://fr.wikipedia.org/wiki/Ontologie_%28informatique%29

FIGURE 4.6 – *Exemple de représentation sous forme de graphe conceptuel de deux concepts et une propriété. Ce sous-graphe est orienté (ici un acide aminé appartient à une chaîne, l'inverse n'est pas vrai) et étiqueté (les concepts sont Amino – acid et Chain et la propriété belongs*$_{t}o$*).*

Les connaissances sont donc représentées par des graphes étiquetés dont les mécanismes de raisonnement se basent sur des opérations de graphes et en particulier sur l'*homomorphisme* de graphes. Un des inconvénients des GC réside dans leur faible utilisation au sein de la communauté scientifique.

4.2.1.3 Logique classique

Les logiques de 1er ordre, 2e ordre et propositionnelles sont 3 propositions de logiques dites «classiques» présentation une première formalisation du langage et du raisonnement mathématique. Cette logique est caractérisée par des postulats qui la fondent:

— Le *tiers exclu* énonce que pour toute proposition mathématique considérée, elle-même ou sa négation est vraie.

— Le *raisonnement par l'absurde* qui veut prouver qu'une proposition est vraie non pas en la démontrant, mais on démontrant que la proposition contraire est absurde.

— La *contraposition* qui consiste à dire qu'une implication de type «A implique B» permet de dire que «si non-B alors non-A». B est donc une condition nécessaire de A.

— L'*implication matérielle* ou le «seulement si» ou «si ..., alors ...» qui en logique classique se caractérise par la volonté de donner une valeur de vérité à toute proposition. Par exemple «s'il pleut, alors mon gazon est mouillé.» Cette proposition ne permet pas de dire, «si mon gazon est mouillé, alors il pleut».

— Les *lois de Morgan* qui sont des identités entre propositions logiques énonçant que «non(A et B) est (non A) ou (non B)» et «non(A ou B) est (non A) et (non B)».

Parmi les logiques classiques, la logique du 1er ordre (ou calcul des prédicats) introduit les symboles mathématiques permettant de formuler des modèles de relations au sein d'ensembles mathématiques. On retrouve dans ces symboles : les variables, les prédicats (ou relations), les connecteurs logiques (et, ou, etc.) et deux quantificateurs universel $\forall$ («Quel que soit», «Pour tout») et existentiel $\exists$ («il existe au moins un ... tel que»). Les formules logiques déduites des énoncés de calculs de prédicats ont pour but de s'appliquer à n'importe quel modèle où l'on retrouve des variables, des fonctions et des prédicats représentant respectivement les éléments de l'ensemble, les fonctions de l'ensemble et les parties (ou sous-ensembles) de l'ensemble.

4.2.1.4 Logique de description

Ces logiques se rapportent à la fois à la description des concepts décrivant un domaine et à la fois à la sémantique basée sur la logique. Cette logique est un apport par rapport aux réseaux sémantiques qui ne possèdent pas de sémantique basée sur la logique et sont donc

exclusivement descriptifs. C'est une famille de langage permettant d'une part la description des connaissances d'un domaine de façon structurée et formelle et d'autre part elles possèdent une sémantique formelle définie en logique du 1er ordre. Elles sont utilisées pour de nombreuses applications, dont le web sémantique et la bio-informatique, au travers d'ontologies associées au domaine. Similairement aux logiques classiques, les logiques de description utilisent les notions de *concept*, *rôle* et *individu* [10]. Les *concepts* désignent les sous-ensembles d'élément dans un univers étudié, les *rôles* correspondent aux liens entre les éléments et enfin les *individus* sont les éléments de l'univers.

4.2.2 Logiques et ontologies en biologie

La génomique fut le domaine biologique qui a intégré rapidement des ontologies afin d'uniformiser les données hétérogènes produites en grande quantité [150]. En génomique et protéomique, de nombreuses études référence *Gene Ontology* [6], une «bio-ontologie» née de la nécessité d'intégrer des données génomiques massives issues du séquençage, et de partager les connaissances relatives à ces données de manière standardisée sur le web. *Gene Ontology* n'est pas totalement considéré comme une ontologie selon la définition informatique stricte du terme, car les relations entre les concepts se réduisent à des relations de type *est-un* ou *est-une-partie-de*. Il est davantage mis en avant comme un vocabulaire standardisé des concepts mis en jeu dans les recherches afin de mettre en place un espace commun de termes précis et définis, possédant une hiérarchie établie. Il permet donc de mettre en relation des bases de données hétérogènes respectant ses codes ontologiques et donc d'effectuer des opérations de requêtes croisées ou de comparaison. Progressivement de nombreuses autres «bio-ontologies» sont apparues dans la lignée de *Gene Ontology* et plusieurs d'entre elles ont permis de mettre en avant, au moyen du simple mécanisme d'inférence et de mise en relations des données hétérogènes, de nouvelles avancées en biologie [183, 162, 159].

Enfin, plusieurs programmes de *Visual Analytics* se sont appuyés sur la mise en place d'une sémantique décrivant les concepts du métier [141]. Cette représentation participe à la formalisation des concepts analysés et instaure un premier lien entre les différents modules d'un programme cherchant à partager visualisation et analyses au sein d'un même espace de travail. DIVE constitue un premier exemple de programme possédant un procédé logiciel incluant la mise en place automatique d'une ontologie. Cette ontologie permet de classer les données et de définir le modèle de données utilisé au sein des librairies de l'application. Le module de traitement de fichier mis en place dans DIVE permet la transformation d'une structure de données hiérarchique pour l'ontologie via une architecture .NET [2] en entrée.

Chaque valeur de donnée est partagée entre l'architecture .NET et les représentations correspondantes dans DIVE permettant une modification ou évolution dynamique de celle-ci pendant l'exécution du programme. Cette solution permet également de mettre en place des méthodes de requêtes sur les valeurs ou les relations entre les objets présents dans l'ontologie.

DIVE permet donc la génération d'ontologies de façon souple et dépendante des données manipulées, assurant ainsi une grande généricité de l'approche. Une représentation approximative ou erronée des données au sein de l'architecture .NET ne pourra cependant pas être contrôlée. Les raisonnements possibles sur l'ontologie créée à partir des données sont limités à la notion d'héritage multiple (d'objets, types et propriétés), étendant les modèles orientés objet retrouvés dans certains langages de programmation (Java, C#, VB.NET, etc.). Il permet l'introduction des relations ontologiques simples de type *est-un*, *contient*, *fait-parti-de* et *lié-par* qui vont permettre d'enrichir les relations entre les données. Les relations seront

2. https://microsoft.com/net

cependant contraintes à ces 4 types de relation, ne comporte pas de notion de cardinalité ou d'opérateurs logiques pour la définition de classes.

Notre application met en jeu des finalités et des besoins légèrement différents que nous exposons ci-dessous.

4.2.3 Formalisme pour une représentation sémantique des données moléculaires en *Visual Analytics*

Le formalisme employé au sein de notre application doit répondre à trois critères principaux:

1. Représentation des données de façon hiérarchique au moyen de concepts et de propriétés.
2. Possibilité de raisonner sur les données et d'extraire de nouvelles règles à partir des règles existantes
3. Requêtes sur les données performantes respectant la contrainte du temps interactif

Notre première implémentation d'une représentation sémantique des connaissances en biologie moléculaire s'est d'abord appliquée sur les graphes conceptuels, formalisme familier de l'équipe, au sein su logiciel de Cogitant [63]. L'utilisation des GC à travers l'API Cogitant s'est rapidement révélée incompatible avec les contraintes du contexte interactif. Cette limitation avait déjà mis en lumière par le travail Yannick Dennemont [51] avec l'API Prolog CG, limitations confirmées par notre propre expérience avec la librairie Cogitant en C++. Le besoin de haute performance imposé par le contexte interactif nous a amené sur la voie des logiques de description et le web sémantique pour la représentation de connaissances et l'extraction performante d'informations au sein d'une base de faits massive pour supporter les fonctionnalités de *Visual Analytics* en biologie moléculaire.

4.3 Web sémantique et formalismes à base de graphes

Le web sémantique a été initié par le *World Wide Web Consortium* et vise à développer des méthodes communes pour l'échange de données sémantiques sur le web. Le but du web sémantique est de structurer et lier les connaissances disponibles sur internet. Selon la définition même du W3C, *le web sémantique fournit un modèle qui permet aux données d'être partagées et réutilisées entre plusieurs applications, entreprises et groupes d'utilisateurs.* Historiquement, le web sémantique doit ses premiers pas à Tim Berners-Lee, le directeur actuel du W3C, qui pose les premières bases d'un environnement où les données seraient classées afin de pouvoir être rapidement mises en relation et partagées [17]. Plusieurs chercheurs ont travaillé à son utilisation et les conséquences d'un passage de l'ensemble des acteurs du web à un tel concept [57]. Nous pouvons citer comme initiatives notables DBpedia, un effort pour publier les données extraites de Wikipedia sous format RDF et interrogeables grâce à SPARQL que nous décrivons ci-après [8] ou le projet Data.bnf.fr [3] qui intègre des données provenant de formats divers (Itermarc [4], XML-EAD et Dublin Core [180] pour la bibliothèque numérique), les regroupant et les formalisant par des traitements automatiques et les publiant dans divers standards du web sémantique basé sur le RDF *Resource Description Framework* (RDF-XML, RDF-N3 et RDF-NT, voir section 4.3.1).

Le web sémantique se base sur un formalisme grandement inspiré des logiques de description. À l'image de plusieurs modèles du web et de l'internet, le web sémantique est caractérisé

3. http://data.bnf.fr
4. http://www.bnf.fr/fr/professionnels/f_intermarc/s.format_intermarc_biblio.html

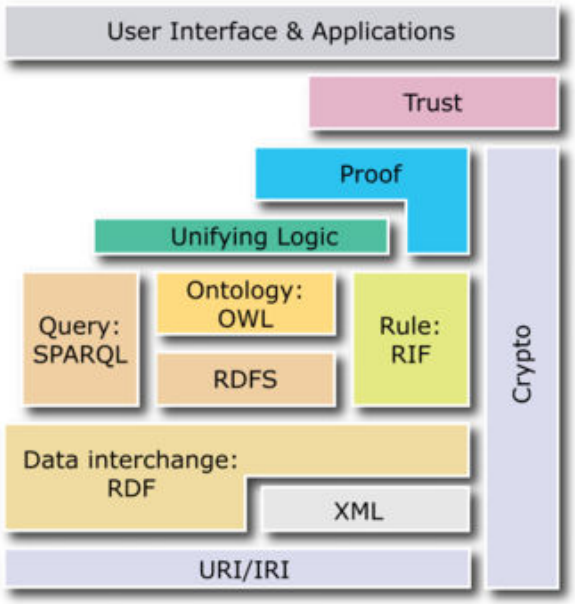


Figure 4.7 – *Architecture du web sémantique et représentation de ses différentes couches. Cette illustration reprend notamment les standards adoptés en web sémantique et utilisés pour chaque couche d'implémentation: RDF, RDFS, OWL, SPARQL, etc..*
Source : http://www.w3.org/2001/sw/

par plusieurs couches de définition selon le schéma présenté dans la Figure 4.7. Le modèle de graphe associé au web sémantique est le RDF [95]. Il est structuré par le RDFS (RDF Schema) [26] qui décrit les vocabulaires (ontologies) sur lesquelles des ressources RDF se basent à la manière d'une DTD (Document Type Definition) pour le XML (eXtensible Markup Language) qui permet de mettre en place une hiérarchie au niveau des balises utilisées dans un document XML. L'ensemble des caractéristiques principales du RDFS est repris dans un langage ontologique plus expressif appelé OWL (Web Ontology Language)[117]. OWL fonctionne de la même manière que RDFS en étendant certaines logiques sémantiques au RDF. La structuration des ressources RDF avec RDFS et OWL permet l'interrogation de ces ressources par le langage de requête SPARQL (SPARQL Protocol and RDF Query Language) [134]. RDF, RDFS, OWL et SPARQL sont tous les 4 des recommandations du W3C pour le web sémantique et ils disposent d'une intégration de plus en plus élargie au sein des technologies et contenus web destinés au partage de connaissances et à leur traitement.

4.3.1 Modèle RDF, formats et langages

Le langage RDF est le langage de base du web sémantique qui l'utilise afin de mettre en place un réseau de données partagées et libres. C'est un modèle de graphe destiné à décrire de façon formelle des ressources et leurs métadonnées associées. Le modèle RDF se base sur une représentation des connaissances à partir de triplets comme illustrés dans la Figure 4.8 par Douglas R. Hofstader [65]. Le triplet est la plus petite division de connaissances en RDF et toute description de données est un ensemble de triplets comprenant :

```
(sujet, prédicat, objet)
```

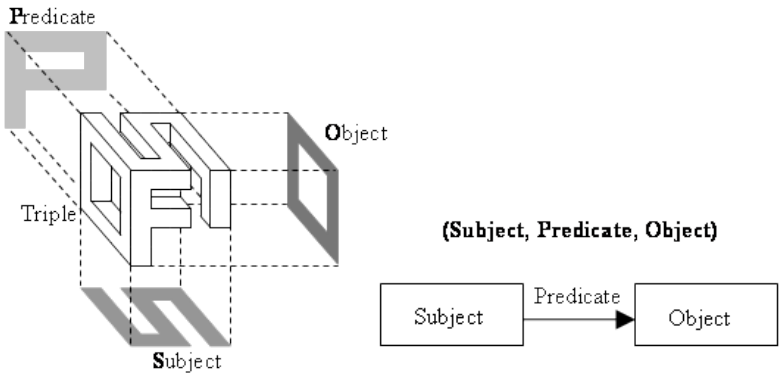


FIGURE 4.8 – *Illustration de la décomposition de base du langage RDF : sujet, prédicat, objet.*

Le *sujet* représente la ressource que nous cherchons à décrire. Le *prédicat* représente une propriété pouvant être associée à la ressource. L'*objet* peut représenter soit une ressource, soit une donnée, cela correspond à la valeur de la propriété. Chaque ressource est identifiée par un URI (Uniform Ressource Identifier) alors qu'une donnée est anonyme puisque pouvant être dupliquée (valeur numérique, chaîne de caractère, etc.). Un exemple de triplet pourrait être:

```
http://mon-ontologie.fr/#Pierre http://mon-ontologie.fr/#âge xsd:int^^26
```

Les ressources de ce triplet (*sujet* et *prédicat*) sont décrites par leur URI qui est constitué d'une partie fixe (*http://mon-ontologie.fr/#*) identifiant le moyen de localiser la ressource et une partie variable (*Pierre* et *âge*) correspondant au nom de la ressource identifiée. Les URI les plus connues et utilisées sont les URL permettant d'accéder à des ressources internet, mais les numéros ISBN référençant les livres sont également des URI. La notion d'URI est importante dans le modèle RDF, car ce sont ces URI qui permettent de désigner de façon non ambiguë une ressource que l'on souhaite décrire. Lors de la mise en place d'une nouvelle base de données, il est courant de créer son propre domaine de définition ou préfixe afin d'identifier toute ressource, classe ou propriété nouvelle décrite. Dans la suite de ce chapitre, nous considérerons *http://mon-ontologie.fr/#* comme notre ensemble de définitions et nous le préfixerons grâce au mot-clé *my*. En langage RDF, cela se caractériserait par la ligne suivante:

```
@prefix my: <http://mon-ontologie.fr/#>
```

La présence de cette ligne en en-tête d'un document RDF permet de réécrire l'exemple précédent:

```
my:Pierre my:âge xsd:int^^26
```

Dans des termes propres aux bases de données relationnelles SQL, RDF peut être considéré comme une table composée de 3 colonnes, sujet, prédicat, objet. À la différence de SQL cependant, la colonne objet est hétérogène et le type de donnée par cellule est sous-entendue (ou spécifié dans l'ontologie, voir 4.3.2) par la valeur du prédicat.

De la même manière que les graphes conceptuels, les documents RDF décrivant un ensemble de données peuvent être représentés par un ensemble de graphes orientés étiquetés. Chaque triplet est représenté par un arc orienté dont les extrémités sont le sujet et l'objet et le label de l'arc correspond au prédicat. L'exemple précédent est illustré sous forme de graphe orienté ci-après:

```
GRAPHE Pierre --âge--> 26
```

Fabien Gandon met très bien en avant le fait que le modèle RDF possède d'ailleurs plusieurs similarités avec le modèle des graphes conceptuels [62]. Ils partagent en effet la distinction entre sémantique (support pour les GC et schémas RDFS/OWL pour RDF) et la connaissance factuelle ou assertionnelle. Dans les deux modèles, cette connaissance est positive, conjonctive et existentielle, pouvant être représentée par des graphes orientés étiquetés. Les schémas RDF et GC possèdent une hiérarchie similaire des classes/propriétés pour RDF et concept/relation pour GC. Les relations et propriétés sont d'ailleurs indépendantes et définies en dehors des classes/concepts et se définissent donc comme de 1ers ordres dans la hiérarchie de ces deux modèles. Enfin, les deux modèles implémentent un mécanisme de déduction équivalent basé sur la subsomption en RDFS et la projection en GC. Les CG et le modèle RDF se distinguent cependant sur plusieurs points. Alors que RDF permet la multi-instanciation, GC ne possède pas d'équivalent. Une déclaration de propriété en RDF peut indiquer plusieurs domaines de définition et d'ensembles d'arrivée ce qui n'est pas le cas en GC. Les GC permettent par des relations d'arité supérieures à deux alors que les graphes RDF sont binaires. Les GC proposent également des extensions supérieures à l'expressivité de RDF/S. Des travaux ont donc cherché à rapprocher ces deux formalismes afin de profiter de leurs avantages respectifs. C'est le cas du projet CORESE [43] qui est un moteur de recherche s'appuyant sur le formalisme de RDF(S)/XML pour exprimer et partager ses données, mais utilise les mécanismes de requête et d'inférence fournis par les GC.

Les données RDF peuvent être représentées sous différents formats, du plus lisible pour l'être humain au plus optimisé pour le traitement informatique, souvent au détriment de sa lisibilité. De la même façon, ils peuvent répondre à différentes applications logicielles et ont été déclinés dans d'autres formats courants.

La syntaxe usuelle du RDF est le XML appelé **RDF/XML**. Le format par balises du XML et les attributs associés ainsi que sa possibilité de définir des hiérarchies sont des caractéristiques du RDF/XML particulièrement appréciés pour représenter des ressources et leurs propriétés. Cette syntaxe est définie par le W3C et fut le premier standard de format pour le modèle RDF [12].

Le **RDFa** (RDF dans des Attributs) est un ensemble d'éléments et d'attributs permettant d'ajouter des données RDF à des documents HTML ou XML [1]. Il se base sur une partie de la syntaxe HTML qu'il reprend dans l'attribut *classe*, qui va préciser le type de l'objet, l'attribut *id*, qui va permettre de définir l'URI de l'objet et les attributs *rel*, *rev* et *href* qui vont spécifier des relations avec d'autres ressources. Le RDFa possède également ses propres attributs dont les plus importants sont *about*, permettant d'ajouter un URI décrivant la ressource décrite par les métadonnées, *property* qui va permettre d'ajouter des propriétés à la ressource décrite.

Le **GRDDL** (pour Gleaning Resource Descriptions from Dialects of Languages) est un langage, recommandé par le W3C [41], permettant d'extraire des données RDF depuis un document XML ou XHTML grâce à des algorithmes de transformation et peut fonctionner comme implémentation du XSLT (Extensible Stylesheet Language Transformations, langage de transformation du XML en formats plus lisibles comme le HTML, le PDF ou le PNG par exemple).

La **Notation3** ou **N3** est une norme de sérialisation non-XML pour le modèle RDF. Elle a été mise en place afin de fournir une syntaxe plus compacte et plus facile à lire. Elle a été développée, entre autres, par Tim Berners-Lee au sein de la communauté du web sémantique [16]. La norme N3 est davantage qu'une norme de sérialisation pour RDF puisqu'elle supporte également les règles basées sur RDF. Elle permet l'utilisation de symboles spécifiques pour éviter la répétition de motifs communs entre les triplets. Par exemple:

```
my:Alanine rdf:type my:Acide-aminé ;
  my:a-une-charge my:Positive ;
  my:est-composé-de my:Carbon_alpha, my:Carbon_beta
```

Ici le «;» est utilisé afin de conservé le sujet (<Alanine>) alors que la «,» est utilisée pour conservé le sujet et le prédicat (<Alanine> et my:est-composé-de).

Le **Turtle** est un sous-ensemble de la syntaxe N3 n'existant pour le moment que sous forme de recommandation du W3C [133]. Sa différence majeure avec le N3 est syntaxique. Il se rapproche beaucoup du format utilisé dans le langage de requête SPARQL (voir 4.3.4).

N-Triplet est un autre sous-ensemble de N3 beaucoup plus explicite puisque ne comportant pas de symboles particuliers simplifiant les descriptions de ressources. Son utilisation se limite souvent aux jeux de données RDF restreints, car son traitement a un coût computationnel assez élevé et la taille des fichiers générés est conséquente.

JSON-LD (ou JavaScriptObjectNotation for Linked Data) est une méthode de transmission des données liées RDF basée sur JSON. Pour rappel, JSON est un format standard ouvert largement utilisé dans les protocoles de communication asynchrones entre serveurs et clients web. Il se base sur la déclaration de structures clé/valeur, particulièrement adaptées à la description de concepts. Dans le cadre de JSON-LD, les clés sont les propriétés du concept, associées à leur valeur. Les valeurs peuvent être des références à d'autres concepts, des nombres ou encore du texte simple [5]. Deux propriétés, *context* et *type*, sont obligatoires et définissent respectivement l'ontologie utilisée et le meta-concept du concept courant (la ressource décrite). Un exemple de document JSON-LD est présenté ci-dessous:

```
<script type="application/ld+json">
{
  "@context": "http://schema.org/",
  "@type": "Movie",
  "name": "Brannigan",
  "genre": "Thriller",
  "trailer": "https://www.youtube.com/watch?v=gAhzli9jbKU",
  "actors":
    {
    "@type": "Person",
    "name": "John Wayne",
    "birthDate": "1907-05-26"
    }
}
</script>
```

JSON-LD tend à devenir la nouvelle norme, majoritairement grâce à son interopérabilité avec les bases de données non relationnelles. Le domaine du *Big Data* connaît un essor très important et beaucoup de grands acteurs du web (Google, Amazon, Twitter, etc.) migrent vers des approches non relationnelles, plus flexibles. Les API qu'ils mettent à disposition sont accessibles à travers le protocole HTTP et utilisent le format JSON pour communiquer. Trois raisons principales pour l'utilisation de ce format : sa clarté, sa structuration et son faible poids. JSON-LD est un protocole que facilite donc grandement l'interopérabilité entre plusieurs composants logiciels distants qui doivent échanger des données sémantiques sur le réseau.

5. http://www.woptimo.com/2015/02/json-ld-la-revolution-semantique-des-microdonnees/

Nous avons vu que le modèle de représentation RDF s'appuie sur un formalisme précis et standardisé qui permet l'uniformisation des données représentées et la mise en place de lien entre ces données. Ce formalisme n'est qu'une méthode de représentation et ne constitue pas une logique de raisonnement qui permettrait d'extraire des connaissances de ces données. C'est dans ce but qu'ont été créés RDFS et OWL, des langages formels servant à décrire des ontologies. À la différence des données RDF qui sont de l'ordre du *factuel*, une ontologie permet de décrire les concepts et les relations entre ces concepts et constitue la partie *structurelle* du web sémantique. OWL et RDFS reprennent les critères standards des ontologies existantes et qui sont le support de nombreuses ontologies recensées dans le portail officiel de bio-ontologies largement utilisées dans la communauté scientifique [159].

4.3.2 RDF Schema

RDFS fut la première extension permettant d'ajouter une couche sémantique au modèle de ressource RDF. Il définit les notions de classes, sous-classes, propriétés et sous-propriétés desquelles dépendront les ressources RDF identifiées. C'est donc un ensemble de concepts haut-niveau définissant les différents individus d'un domaine et permettant leur hiérarchisation. Par exemple, RDF permet de décrire qu'une <Lysine> *a une charge* <positive> grâce aux individus <Lysine> et <positive>, respectivement sujet et objet, et au prédicat *a une charge*. RDFS permet d'ajouter les concepts décrivant les individus comme <Acide-aminé>, <Charge>, <Hydrophobicité>, <Molécule>, etc. et de préciser des relations entre ces groupes de façon à pouvoir émettre des déductions simples. Par extension de l'exemple précédent, si l'on ajoute que tout <Acide-Aminé> est une <Molécule> qui revient à dire en RDFS, <Acide-aminé> *sous-classe de* <Molécule>, et <Lysine> *instance de* <Acide-aminé>. Cela nous permet d'ajouter un niveau d'information induite des deux affirmations précédentes qui serait: <Lysine> *instance de* <Molécule>. Cette déduction se fait grâce au système d'implication introduit par RDFS et permettant de déduire des informations complémentaires à partir d'un minimum de données.

RDFS introduit également la notion de domaine de définition (*rdfs:domain*) et d'ensemble d'arrivée (*rdfs:range*) pour les propriétés. Le domaine de définition correspond à la classe des sujets liés à une propriété alors que l'ensemble d'arrivée correspond à la classe ou au type de données des valeurs de la propriété. Il est par exemple possible de préciser que la propriété <identifiant> doit avoir comme domaine de définition seulement des individus de classe <Objets> et comme ensemble de définition des valeurs de type <xsd:integer>. Le système d'implication de RDFS fonctionne également avec les notions de domaine d'application et d'ensemble de définitions. De ce fait, ces 3 affirmations:

```
<my:Alanine> <my:est-un> <my:Acide-aminé> .
<myProtéine> <my:contient> <my:Lysine> .
<my:contient> <rdfs:range> <my:Acide-aminé>
```

permettent d'induire l'affirmation suivante:

```
<my:Lysine> <my:est-un> <my:Acide-aminé>
```

4.3.3 OWL

OWL est donc un standard informatique visant également à jouer le rôle de grammaire pour le langage RDF en complément du RDFS à qui il reprend ses bases. Le langage OWL se base sur des éléments des logiques de description et constitue un standard informatique

permettant de vérifier que les données sont cohérentes, de déduire de nouvelles connaissances de ces données ou d'en extraire certaines informations. Plus expressif que RDFS, OWL permet d'adjoindre à la définition des relations entre objets par des assertions fournies par RDFS, des propriétés reliant les classes à travers des relations de symétrie, d'équivalences, de cardinalité, etc. entre les classes. Il est donc possible de mettre en place des associations de classes et de propriétés plus complexes et basées sur une fondation logique. Si l'on étend l'exemple précédent avec les notions apportées par OWL au sein des triplets suivants:

```
<est-composé-de> \textit{est} <owl:TransitiveProperty>,
<Protéine> <est-composé-de> <Acide-aminé>,
<Acide-aminé> <est-composé-de> <Atome>
```

nous pouvons donc en déduire, grâce au caractère transitif de la propriété <est-composé-de>, que:

```
<Protéine> <est-composé-de> <Atome>
```

De la même manière, il est possible de définir des propriétés comme symétriques, asymétriques, inverses, réflexives, etc.[6] OWL est composé de 3 sous-langages classés du moins expressif au plus expressif: OWL-Lite, OWL-DL et OWL-Full. Parmi ces sous-langages, seul OWL-Lite est implémenté sous forme d'algorithmes décidables dans la majorité des moteurs d'inférence utilisés lors de l'interrogation de bases de données RDF possédant une ontologie OWL associée. La simplicité d'OWL-Lite lui permet une complexité réduite et donc un temps de calcul également réduit lors de son utilisation. Il regroupe les principales relations et descriptions de classe amenée par OWL et permet ainsi une mise en place de logiques descriptives suffisamment évoluées pour permettre l'extraction de nouvelles connaissances à partir de données simples.

4.3.4 SPARQL

Nous venons d'évoquer la possibilité d'utiliser un moteur d'inférence afin d'extraire des données d'une base de données RDF. Cette extraction peut se faire grâce à l'utilisation d'un protocole et langage de requête appelé SPARQL. Ce langage permet à la fois de récupérer des données stockées sous format RDF dans une base de données, mais également de les éditer, d'en ajouter ou bien d'en supprimer. L'accès aux bases de données se fait grâce à une interface d'accès (en anglais *endpoint*) géré par un service capable de recevoir des requêtes SPARQL et de renvoyer des résultats sous différents formats. À la différence du SQL, SPARQL se base essentiellement sur le format en triplets des bases de données RDF et la majorité de ses requêtes repose sur la mise en place d'un schéma de correspondance entre triplets sujet/prédicat/objet. Il n'y a pas de contraintes de typage pour la colonne objet qui est habituellement implicite ou définie par l'ontologie. Dans le même esprit, l'ontologie est directement intégrée dans les résultats de requêtes et le schéma de données n'a donc pas besoin d'être appliqué de façon séparée. SPARQL fournit également plusieurs opérations sur les résultats comme SORT, JOIN, DISTINCT, qui permettent un traitement direct des résultats afin de les classer ou filtrer suivant les besoins...[7] Certains mots-clés de SQL ont été conservés tels que SELECT, FROM, WHERE, etc. Une requête SPARQL se retrouve sous la forme suivante:

```
SELECT ?x ?id
```

6. http://www.w3.org/2007/OWL/wiki/Quick_Reference_Guide
7. http://www.cambridgesemantics.com/semantic-university/sparql-vs-sql-intro

```
FROM <http://my\_database.com>
WHERE {
  ?y my:est-composé-de ?x .
  ?y rdf:type my:Chain .
  ?y my:chain\_id "B" .
  ?x my:res\_id ?id
}
```

Dans son fonctionnement, SPARQL va effectuer deux opérations successives sur la base de données. Une première étape est la recherche des triplets correspondant au modèle de triplet énoncé dans la requête. C'est l'étape de restriction qui va retourner l'ensemble des lignes de la base de données construites sur ce motif et comportant les valeurs spécifiées quand elles le sont (ici *my:a-une-charge*, *my:positive* et *my:res_id*). Cette étape se rapporte donc à ce qui se trouve dans le «WHERE». Lorsque les triplets sont identifiés, SPARQL va sélectionner uniquement les variables demandées au niveau du mot-clé «SELECT», ici **?x** et **?id**. C'est l'étape de projection qui va venir sélectionner les colonnes que l'utilisateur a demandé parmi les lignes sélectionnées dans l'étape précédente. Le résultat de la requête SPARQL précédente sur le jeu de données fourni en Annexe A pourrait donc être:

?x	?id
RES_11	1
RES_12	2
RES_13	3
RES_14	4
RES_15	5

Les résultats sont présentés par défaut sous forme de tableaux par souci de clarté, mais la plupart des services SPARQL permettent leur conversion automatique en un nombre important de formats comme le XML, le JSON, JavaScript, Turtle, etc. De nombreuses librairies utilisent des points d'accès SPARQL afin d'y présenter des requêtes et d'en récupérer leurs résultats. Cette implémentation présente dans de nombreux langages de programmation assure une certaine généricité du travail d'interfaçage nécessaire dans une suite logicielle multi composante.

4.3.5 Implémentations et outils

Nous avons évoqué la présence de librairies ou plateformes implémentant les standards du web sémantique et permettant d'utiliser ce modèle de représentation de données pour des activités hétérogènes. Nous pouvons citer plusieurs d'entre eux, associés fortement au web sémantique et cherchant à se développer parallèlement au domaine afin de lui offrir un support aussi complet que possible. Parmi ceux-ci:

— **Jena**[8] [115] est l'un des moteurs actuels les plus complets et propose une persistance en mémoire ou en base de données. C'est une structure logicielle Java gratuite et ouverte implémentant RDF, RDFS et OWL ainsi que les requêtes SPARQL et propose un moteur en chaînage avant (RETE), arrière (programmation logique) et hybride. Ce moteur est utilisé pour implanter la sémantique de RDFS et OWL et son modèle repose sur une structure prédéfinie de bases de données.

8. https://jena.apache.org/index.html

— **Protégé**[9] [123] est un outil basé sur Java et utilisé pour construire et éditer des ontologies. Il supporte RDF et OWL de façon native et intègre différents outils comme un visualiseur de réseau ontologique sous forme de graphe orienté ou une interface de requête SPARQL. Il est capable d'utiliser certains moteurs de logique de description afin d'inférer une ontologie ou de la valider.

— **Corese**[10] [43] est un moteur de recherche basé sur l'ontologie et rapprochant les approches de graphes conceptuels et de RDFS afin de répondre à des requêtes faites sur des annotations RDF. Il étend ainsi les règles d'inférence RDFS en s'appuyant sur certaines opérations propres aux graphes conceptuels.

— **Triple** [157] est un langage de requêtes basé sur la logique de Horn permettant d'interroger des données du web sémantique et principalement des données RDF. Il est capable de raisonner grâce à des moteurs d'inférence et intègre les primitives RDF comme les espaces de nommage, les ressources et la notion d'individu (à opposer à concept ontologique). Son principal atout réside dans sa capacité à interroger différents modèles de données grâce à différentes logiques de description.

— **Fact** et son successeur **Fact++** [172] est un moteur à base de logiques de descriptions permettant de générer des requêtes conjonctives qui vont identifier, classifier ou valider des données d'une base de connaissances à partir de son ontologie. Ce moteur intègre les ontologies OWL et donc capable d'appliquer les règles d'inférences qui lui sont associées. Il est par exemple implémentable dans l'outil Protégé cité précédemment.

— **jOWL**[11] est une librairie jQuery permettant de naviguer et de visualiser des documents OWL-RDFS au sein d'un environnement JavaScript et donc pouvant être intégré au sein de pages web.

4.4 Conclusion

Dans le but de réduire les données à interpréter par l'utilisateur pendant un processus d'étude, le domaine du *Visual Analytics* apporte un processus de travail où la visualisation de données scientifiques et géométriques peut se mêler à la représentation de données abstraites. Basé sur des techniques d'interactivité forte entre l'espace de visualisation et d'analyse, il répond au besoin de mettre l'utilisateur au cœur de son processus d'étude des données. Nous avons vu en transposant le schéma du *Visual Analytics* proposé par Keim (dans la section 4.1.1) que le processus de travail de biologie structurale trouve tout son sens quand intégré dans ce cadre.

Notre second point d'intérêt, pour l'implémentation du processus de *Visual Analytics* dans un contexte immersif et moléculaire, est la façon de mettre en place l'axe principal autour duquel tourne le *Visual Analytics*, l'interactivité. Cette interactivité est également l'une des pierres angulaires de l'immersion, qui met en jeu des environnements où les techniques d'interaction sont complexes, de part le caractère simplifié et limité des périphériques d'interaction disponibles. L'interaction entre les espaces de visualisation et d'analyses en biologie structurale met en jeu des données hétérogènes puisqu'elles sont à la fois de nature géométrique et de nature analytique. Des informations 3d (distance au ligand, poches de liaisons, etc.) doivent donc être recoupées avec des informations 2d (énergie potentielle, charge, etc.). Pour les regrouper au sein d'un même contexte interactif, nous avons montré que la mise en place

9. http://protege.stanford.edu/
10. http://wimmics.inria.fr/corese
11. http://jowl.ontologyonline.org/

un formalisme commun permettant de représenter l'ensemble des connaissances du métier est une réponse adaptée et pertinente du fait de sa généricité, ses capacités d'extension et son utilisation déjà présente dans plusieurs domaines biologiques.

Parmi les formalismes à notre disposition pour cette représentation, la logique de description présente l'approche la plus en adapté pour implémenter les fonctionnalités d'interaction de *Visual Analytics* en répondant aux contraintes du contexte interactif. Elle introduit la notion d'ontologie, utilisée dans de nombreux domaines d'application, en particulier biologiques, pour mettre en place une description des données utilisées et permettre leurs réutilisation, comparaison et enrichissement autour d'une terminologie de concept formalisée dans l'ontologie. Les logiques de description permettent également d'étendre les connaissances d'une base de connaissances via des relations logiques définies dans l'ontologie sans que ces relations factuelles soient explicitement présentes dans la base de connaissances.

La construction d'ontologies est la première étape du processus de formalisation sémantique des connaissances. Le web sémantique est une initiative présente dans de plus en plus d'institutions du web, nous avons choisi d'utiliser ses outils et modèles comme base de développement pour la description d'une plateforme cherchant à rassembler au sein d'une même représentation haut-niveau des concepts hétérogènes comme ceux mis en jeu au sein de la visualisation et de l'analyse de simulations moléculaires. Notre plateforme s'appuie sur une représentation complète du contenu, du contexte interactif et de la tâche afin de permettre la liaison des données et leur manipulation conjointe dans un contexte interactif commun favorisant les interactions directes et adaptées à un usage dans un contexte immersif.

Chapitre 5

Conception, architecture et implémentation d'une application de Visual Analytics Moléculaire adaptée au contexte immersif

Nous présentons dans ce chapitre comment la création d'une ontologie décrivant les différents concepts mis en jeu dans les outils de visualisation et d'analyse peut permettre leur intégration au sein d'un même espace de travail. Après une introduction conceptuelle de l'articulation que nous avons imaginé entre représentation des connaissances et biologie structurale dans une première partie, nous introduirons une première implémentation de notre approche à travers la mise en place d'un moteur d'interprétation de mots-clés, pouvant être couplé à un logiciel de reconnaissance vocale, comme solution d'interaction directe avec l'environnement de travail. Dans une dernière partie nous décrirons notre principale implémentation basée sur la sémantique et les outils de *Visual Analytics* à travers une plateforme d'étude de structures moléculaires nous permettant de lier et mettre en relation des données hétérogènes provenant d'outils de visualisation et d'analyses utilisés en biologie structurale.

Sommaire

Introduction

La définition du *Visual Analytics* introduite dans le chapitre précédent a mis en évidence la possibilité d'intégrer des outils de visualisation 3d et des représentations analytiques 2d au sein d'un même espace de travail (voir Figure 5.1).

L'extension des approches de *Visual Analytics* par une représentation des concepts mis en jeu permet le regroupement et l'analyse conjointe d'informations hétérogènes. Parmi ces informations, les modèles 3d et les métadonnées associées issues d'analyses décrivant ces modèles sont celles que nous chercherons à manipuler au sein d'environnements immersifs. Cette articulation entre la représentation des connaissances et les espaces de travail usuels en biologie structurale est décrite dans la prochaine section.

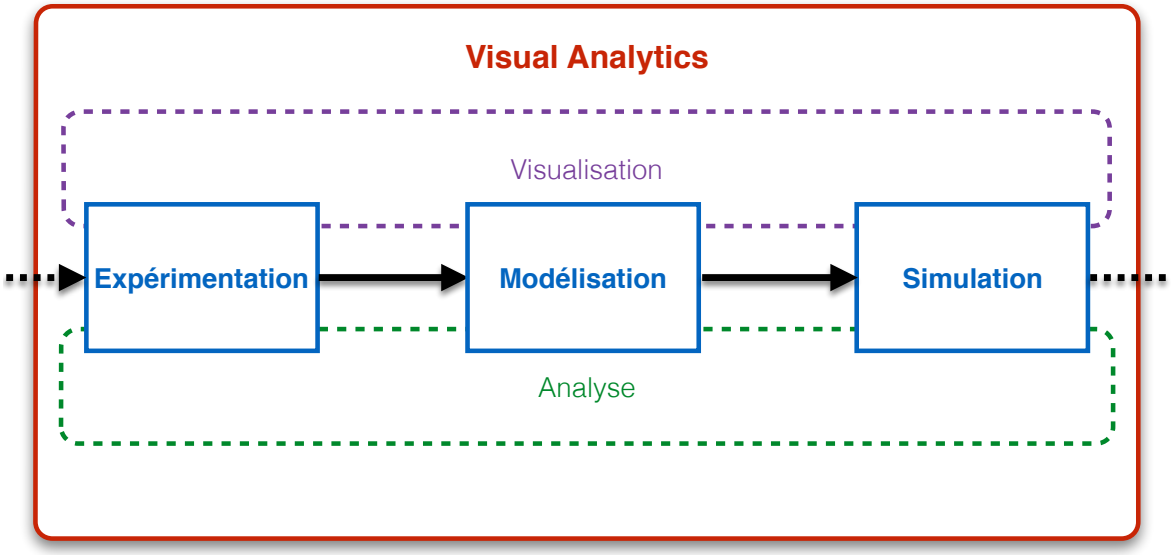


Figure 5.1 – *Schéma du processus itératif d'étude d'une structure moléculaire en biologie structurale. Nous illustrons dans ce chapitre l'utilisation du Visual Analytics pour favoriser la visualisation des données d'analyse.*

5.1 Conceptualisation de la représentation de connaissances pour le *Visual Analytics*

Les outils du web sémantique décrits dans le chapitre précédent constituent un environnement logiciel propice pour offrir et intégrer des données liées au sein d'une plateforme composée de modules communicant grâce au protocole d'échange des connaissances du domaine, interrogeable grâce au langage SPARQL.

5.1.1 Des données hétérogènes aux données liées

Nous savons qu'une des limites pour la mise en place d'une communication bilatérale entre la visualisation et l'analyse est la nécessité de traiter des données possédant des liens ou pouvant être relier par des dénominateurs communs. Lors de la mise en place d'une ontologie, la hiérarchie créée va permettre de mettre en place un ensemble de concepts, hiérarchisés et possédant un nombre important de possibilités de relations.

La définition d'un concept au sein de l'ontologie se fait à travers la mise en place de ses propriétés mais également à travers son positionnement hiérarchique et les relations avec l'ensemble des autres concepts de l'ontologie. L'ensemble de ces relations et propriétés va constituer la définition ontologique de nos données biologiques mais également la définition des espaces dans lequel ces données vont être représentées. Un individu de la base de fait, instance d'un concept donné, possédera donc à la fois des propriétés propres à chaque espace de représentation dans lesquels il est visualisé et/ou manipulé.

En plus de lier étroitement les espaces partageant les représentations des mêmes individus, la représentation sémantique sous forme d'ontologie et la base de fait formalisée grâce à cette ontologie apportent les avantages suivants :

— L'identification lors de chaque interaction avec les représentations des objets d'intérêt pour la proposition d'actions adaptées à la nature et aux propriétés de ces objets,

— La représentation hiérarchisée des concepts permettant l'inférence et donc la production de nouvelles connaissances, facilitant la mise en relation de données concernant des objets de même nature,

— Le partage des données entre les différents composants de la plateforme,

— Une ontologie évolutive avec des mécanismes d'interprétation haut niveau des requêtes métier.

Cette approche construite autour de la mise en place d'ontologies, bien qu'inspirée par notre domaine d'application, ne se limite pourtant pas à celle-ci. Tout domaine mettant en jeu des espaces de représentation différents ou des jeux des données hétérogènes peut profiter de la mise en place d'ontologie pour fournir un carcan sémantique rassemblant les concepts du domaine et les liant au sein d'une représentation globale des connaissances. L'approche par représentation est en plus indépendante de l'application métier rajoutant encore à sa généricité.

Au sein de notre application, lors du processus d'analyse de la structure ou de résultats d'analyse numérique d'une simulation dans un espace de travail conjoint, chaque élément avec lequel interagit l'utilisateur correspond à un individu unique dans la base de fait dont la nature ou l'une de ses propriétés est mise en avant, soit à travers une représentation visuelle 3d, soit à travers une ou plusieurs représentations analytiques 2d. Chaque interaction avec cet individu devra déclencher un événement visuel dans tous les espaces représentant cet individu ou une de ses propriétés.

5.1.2 Ontologie pour la modélisation des concepts de biologie structurale

Comme le montre en partie la Figure 5.2, nous avons essayé de définir systématiquement l'ensemble des concepts que l'utilisateur aurait à manipuler lors de ses activités de visualisation et d'analyses. La figure introduit deux concepts, un type d'acide-aminé, «Arginine», et un concept pouvant désigné à la fois une propriété géométrique et une représentation visuelle, «Secondary_structure». Au sein de l'ontologie, les concepts sont donc déjà définis au travers de propriétés appartenant au rang des connaissances expertes. Ces définitions sont invariables et considérées comme acquises pour la plateforme.

Nous l'avons vu auparavant, plusieurs bio-ontologies ont été mises en place ces dernières années. Afin d'avoir une description des concepts biologiques mis en jeu, nous avons étendu une ontologie déjà existante et disponible en ligne décrivant de façon complète les acides aminés et leurs propriétés biophysiques et géométriques [1]. Cette ontologie nous a permis de

1. http://bioportal.bioontology.org/ontologies/AMINO-ACID

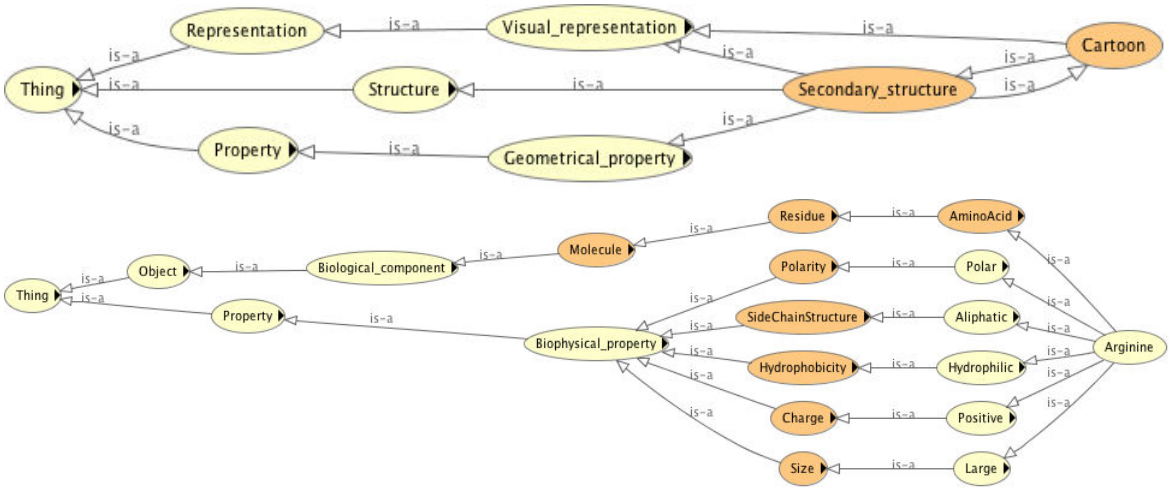


Figure 5.2 – *Extrait de l'ontologie OWL créée pour la mise en place de notre représentation sémantique de la plateforme logicielle. Les concepts «Arginine» et «Secondary_structure» sont mis en avant. Les concepts sont représentés par des cercles jaunes, pour les classes primitives, et en orange pour les classes possédant au moins une classe équivalente. Les relations entre les classes sont symbolisées par des flèches labellisées*

décrire les connaissances biologiques support des principaux concepts de biologie structurale. En rassemblant des informations telles que la taille, l'hydrophobicité ou bien la charge de chaque acide-aminé il est ainsi possible d'extraire rapidement des groupes d'acides aminés possédant les mêmes propriétés et ainsi utiliser ces propriétés pour des raisonnements complexes lors de l'interrogation de la base de données. Il est cependant important de noter qu'une extension de notre ontologie est possible en fonction des besoins spécifiques des experts par rapport aux phénomènes étudiés. Cette extension pourrait par exemple compléter et étoffer certaines propriétés biologiques omises mais intéressantes pour des études ponctuelles ou spécialisées.

L'ontologie a été conçue autour de 5 ensembles de définitions distincts, s'adressant à 5 éléments différents composant notre plateforme :

— **Connaissance biomoléculaire** - Regroupe les concepts métier et les objets de la biologie structurale

— **Représentation 3d des structures** - Regroupe les concepts liés à la représentation et à la visualisation des complexes moléculaires 3d

— **Représentation 2d** - Regroupe les moyens de représentation des analyses numériques 2d en biologie moléculaire

— **Interactions 3d** - Regroupe les concepts génériques liés à l'interaction dans les environnements 3d

— **Interactions 2d** - Regroupe les concepts génériques liés à l'interaction dans les environnements 2d

La distinction des ensembles ne signifie pas qu'il n'existe pas de relation entre deux ensembles. La concept de «Secondary Structure» fait par exemple partie de l'ensemble de définition **Connaissance biomoléculaire** mais est également rattaché au concept «Cartoon» de l'ensemble de définition **Représentation 3d**. C'est l'ensemble de ces connections qui permettront par la suite de raisonner sur l'ontologie afin de supporter le niveau d'interactivité

avancée requis en *Visual Analytics*.

Les concepts et les propriétés présentes au sein des ensembles de définition de **Représentation 2d** et de **Représentation 3d** désignent les éléments graphiques permettant de représenter les concepts de l'ensemble des **Connaissances biomoléculaires**. Les notions de forme, de couleur mais également de types de graphes 2d sont des notions qui seront définies dans ces ensembles. Il est à noter que les concepts analytiques sont définis par les éléments graphiques ou abstraits qui rentrent en jeu dans la création et la visualisation d'un résultat d'analyse, notions définies dans les deux ensembles de définition de **Représentations**. Nous ne définissons cependant pas, volontairement, les différents calculs et analyses des données de simulation car ceux-ci peuvent être de natures très différentes et possèdent un champ de définition bien trop large et complexe pour le sujet de cette thèse. Cela ne signifie pas que certains processus d'analyses ne seront pas utilisé au sein de la plateforme logicielle, il n'est simplement pas pertinent de les intégrer à l'ontologie, ne participant à l'uniformisation des rendus 3D et d'analyses.

En plus des concepts biologiques et de représentations cités précédemment, nous avons donc également cherché à définir tous les concepts d'interaction qui seront en jeu lors de la session d'utilisation de notre plateforme. Les interactions rassemblent toutes les actions possibles de l'utilisateur sur les données qu'il manipule, de façon directe ou indirecte. Elles reprennent aussi les commandes proposées par la plupart des applications de visualisation moléculaire et les outils d'analyse utilisés.

Une fois l'ontologie mise en place, il est possible d'alimenter la base de faits en ajoutant les informations biologiques regroupées par l'expert scientifique. Ces informations vont devoir respecter le vocabulaire et la classification définie par les règles présentes dans l'ontologie OWL.

5.1.3 Base de faits moléculaires

La description d'un environnement d'intérêt passe par l'analyse de toutes les informations biologiques identifiables par une chaîne de caractère ou une valeur et qui correspondent à un concept ou une propriété identifiée dans l'ontologie OWL. Chaque information alimentera une base de données RDF regroupant de manière exhaustive toutes les connaissances sous forme de triplets de type Sujet/Propriété/Objet.

Dans le cas qui nous intéresse pour ce travail de thèse, les simulations numériques moléculaires peuvent être découpées en une succession de modèles statiques 3D successifs. Chaque modèle correspond au concept *Model* décrite dans notre ontologie. C'est le groupe structurel le plus large que nous ayons défini dans les composants structurels biologiques. Plusieurs triplets présents dans notre base de données finale sont illustrés dans le Listing 5.1. Chaque atome présent dans un modèle possède une position 3d ainsi qu'un résidu auquel il appartient et indirectement, de la même manière, une chaîne et son modèle de référence grâce aux règles d'inférence. Ces informations n'ont pas besoin d'être spécifiées explicitement puisque les règles d'inférence définies dans l'ontologie instaurent, entre autres, le fait qu'un atome est sous-composant d'un résidu qui est un sous-composant d'une chaîne elle-même sous-composante d'un modèle. Cette information n'est donc pas présente dans la base de données sous forme d'un triplet explicite de type *ATOM_1234 my belongs_to CHAIN_12* mais ce triplet pourra être tout de même obtenu lors d'une requête cherchant à obtenir l'identifiant de la chaîne dont l'atome est l'un des composants.

Toutes les propriétés géométriques (position, angles, distance, etc.), physico-chimiques

(accessibilité, charge partielle, liaisons, etc.) ou analytiques (énergie, RMSD [2], température, etc.) sont donc intégrés dans la base de données de faits et surtout associés aux individus créés à partir de résultats de structures 3d (atomes/résidus/chaîne/modèles) à chaque pas de temps de la simulation. Tous les individus sont des instances des concepts définis dans l'ontologie. Ces individus associés toutes leurs propriétés forment la population de la base de données de faits.

Au-delà du stockage et de la mise à disposition des données, la base de donnée RDF et plus particulièrement l'ontologie OWL utilisé comme support permettent de mettre en place des moteurs d'interprétation de commandes ou requêtes qui vont avoir pour objectif de récupérer une liste d'individus respectant un certain nombre de propriétés requises.

Listing 5.1 – Exemple de triplets RDF présents dans notre base de données

```
<owl:ObjectProperty rdf:about="&my;hasSize">
    <rdf:type rdf:resource="&owl;FunctionalProperty"/>
    <rdfs:domain rdf:resource="&my;AminoAcid"/>
    <rdfs:range rdf:resource="&my;Size"/>
    <rdfs:subPropertyOf rdf:resource="&my;has_property"/>
</owl:ObjectProperty>

<owl:DatatypeProperty rdf:about="&my;chain_id">
    <rdfs:range rdf:resource="&xsd;int"/>
</owl:DatatypeProperty>

<rdf:Description rdf:about="my:ATOM_36795">
        <my:atom_id rdf:datatype="http://www.w3.org/...#int">125</my:atom_id>
        <my:time_frame rdf:datatype="http://www.w3.org/...#int">190</my:time_frame>
        <rdf:type rdf:resource="my:Atom"/>
        <my:pos_z rdf:datatype="http://www.w3.org/...#double">22.33</my:pos_z>
        <my:atom_type>CB</my:atom_type>
        <my:pos_x rdf:datatype="http://www.w3.org/...#double">21.86</my:pos_x>
        <my:uniq_id rdf:datatype="http://www.w3.org/...#int">36795</my:uniq_id>
        <my:pos_y rdf:datatype="http://www.w3.org/...#double">31.6</my:pos_y>
        <my:belongs_to rdf:resource="my:RES_3622"/>
</rdf:Description>
```

5.1.4 Requêtes et interrogation pour l'interaction directe

Lorsque les données ont été intégrées au sein de la base de faits RDF, sous forme d'instances ou d'individus des concepts définis au sein de l'ontologie OWL, il est nécessaire de mettre en place un système d'interrogation et de récupération des données pour leur visualisation et traitement en réaction aux événements d'interaction dans l'espace de travail de l'expert. Nous avons vu dans le chapitre précédent que RDF et OWL pouvaient être interrogés grâce au langage de requête SPARQL. A la manière du langage SQL pour les bases de

2. Le Root Mean Square Deviation est un calcul des distances moyennes entre deux groupes d'atomes identiques provenant de deux structures 3d différentes. Après une superposition de certains atomes pour minimiser leurs distances, on calcule la moyenne des distances du reste des atomes deux à deux. C'est un critère très utilisé pour comparer deux modèles de molécules.

données relationnelles, SPARQL est optimisé pour l'interrogation de bases RDF sous forme de triplets.

Notre implémentation de SPARQL a pris plusieurs formes afin de répondre à plusieurs besoins au sein de notre plateforme. La richesse des requêtes pouvant être faites par SPARQL Au delà du fait que SPARQL est utilisé pour implémenter les fonctionnalités de *Visual Analytics* ciblé, un des premiers besoins identifié dans les environnement immersifs, illustrant la puissance de la modélisation par ontologie OWL et base de données RDF, fut la mise en place d'un moteur d'interprétation de mots-clés en commandes applicatives.

5.2 Une approche opérationnelle d'interprétation de commande vocale

Une des techniques d'interaction plébiscitée au sein des environnements immersifs est la commande vocale. Elle traduit une phrase ou un groupe de mots édicté par l'utilisateur en une commande interprétable par l'application déclenchant ainsi une action appropriée. Les commandes vocales ont également pour intérêt de pouvoir être combinées à des commandes gestuelles. Un effet serait par exemple d'apporter un filtre supplémentaire sur le champ de sélection des objets virtuels concernés par une commande vocale, prémices de commandes multimodale.

Les actions identifiées au sein de notre programme impliquent pour une majorité d'entre elles une activité appliquée à un groupe structurel désigné par l'utilisateur dans l'ensemble moléculaire observé. Or ces groupes structurels peuvent être caractérisés par des identifiants ayant un sens biologique (les acides aminés sont, de façon conventionnelle, numérotés séquentiellement au sein d'une chaîne de la partie N-terminale de la protéine vers la partie C-terminale), des identifiants uniques au sein de la base de données RDF ou bien leurs propriétés. Ainsi, pour interpréter les commandes émises en langage naturel par l'utilisateur, utilisant un vocabulaire spécifique du domaine avec un haut niveau d'abstraction, nous avons besoin d'une représentation qui peut porter la complexité des connaissances du domaine et relier les objets désignés par l'utilisateur aux objets virtuels impliqués dans l'interaction.

A travers la mise en place de raisonnements sur l'ontologie créée, grâce à l'expressivité du langage OWL, il nous a été possible d'élaborer un moteur d'interprétation traduisant une succession de mots-clés, extraite d'une commande vocale dans notre exemple, afin de générer une commande spécifique et applicative utilisée de façon synchrone dans l'espace de visualisation et l'espace d'analyses.

L'ensemble du processus permettant de répondre à une verbalisation d'un utilisateur pour déclencher une action dans ses espaces de travail peut être décomposé en 3 parties :

1. Reconnaissance de mots-clés vocalisés(5.2.1)
2. Classification des mots-clés dans une structure décomposée de commande (5.2.2)
3. Construction de la commande finale opérationnelle (5.2.2.6)

Notre effort de conceptualisation et l'utilisation de l'ontologie ont porté essentiellement sur les parties 2. et 3. Les parties 1. et 3. sont davantage orientées d'un point de vue implémentation et décrite dans le but de mettre en avant l'application finale.

5.2.1 Reconnaissance de mots-clés métiers via Sphinx

La reconnaissance vocale en elle-même se fait via Sphinx. Ce logiciel de reconnaissance vocale base son processus de reconnaissance sur l'implémentation de dictionnaires créés au-

paravant et listant l'ensemble des termes pouvant être utilisés lors d'une session de travail, désignant les concepts modélisés dans l'ontologie. Ce dictionnaire se doit d'être le plus complet possible afin de prendre en compte l'ensemble du vocabulaire spécialisé que pourrait utiliser l'expert. L'ontologie nous fournit justement une liste exhaustive des concepts manipulés au sein de la plateforme. Une liste complète des concepts et propriétés peut donc en être extraite afin de créer le dictionnaire métier sur lequel se basera Sphinx dans notre application.

Sphinx possède un module dédié nous permettant de l'intégrer aisément dans une architecture logicielle hiérarchisée. Sphinx analyse donc chaque commande vocale de l'utilisateur en la faisant correspondre au dictionnaire chargé afin d'en extraire textuellement les mots ou groupes de mots reconnus. Ces mots sont ensuite classés en groupes syntaxiques dans une première partie du moteur d'interprétation que nous avons développé.

5.2.2 Classification des mots-clés

Une fois la réception des mots-clés effectuée, notre moteur va catégoriser chaque mot reçu. Cette classification se base sur le découpage de notre base de données qui différencie cinq catégories de mots pouvant être retrouvés dans une commande vocale, modélisés sémantiquement :

- **Action**
- **Composant**
- **Identifiant**
- **Propriété**
- **Représentation**

Cette classification est effectuée par des requêtes SPARQL distinctes pour chaque catégorie et interrogeant l'ontologie OWL.

Alors que les catégories *Action*, *Composant*, *Propriété* et *Représentation* possèdent leurs propres concepts et peuvent être identifiées par un mot unique (*Hide*, *Chain*, *Charged*, *Sphere*, etc.), la catégorie *Identifiant* est elle liée à une instance d'un concept de la catégorie *Composant*. Étant donné la possible redondance des identifiants dans une base de donnée de simulation moléculaire puisque chaque modèle 3d décrit le même système moléculaire autant de fois qu'il y a d'unités de temps découpant la trajectoire, l'association entre un identifiant et un composant est obligatoire, quel que soit son niveau structurel. Un identifiant ne pourra donc être classé comme tel que si un composant existe, et seulement si ce composant possède effectivement l'identifiant demandé. Ce pré-requis constitue la première règle de notre classification de mots-clés.

Les commandes SPARQL permettant de dire si un mot-clé appartient ou non à une catégorie, utilisent l'opérateur *ASK* qui prend en argument un ou plusieurs triplets et retourne un booléen *true* si l'ensemble des triplets se vérifie (existe) dans la base de données. Les cinq requêtes SPARQL formulées pour la classification sont donc construites sur la même forme et illustrées dans le Listing 5.2.

Listing 5.2 – *Requêtes SPARQL effectuées pour tester la nature des mots-clés entrés par l'utilisateur*

```
ASK {my:Hide rdfs:subClassOf my:Action}
ASK {my:Alanine rdfs:subClassOf my:Biological_component}
ASK {my:Cartoon rdfs:subClassOf my:Representation}
ASK {my:Red rdfs:subClassOf my:Colors}
ASK {my:Aliphatic rdfs:subClassOf my:Property}
```

Les règles de raisonnement et d'inférence sont utilisées lors de l'ensemble des requêtes SPARQL. Par exemple, la requête

```
ASK {my:Alanine rdfs:subClassOf my:Biological_component}
```

renverra *true* malgré l'absence de lien direct, les concepts *AminoAcid*, *Residue* et *Molecule* se situant entre ces deux concepts (voir Figure 5.3).

FIGURE 5.3 – *Extrait de l'ontologie OWL décrivant la hiérarchie du concept Alanine.*

5.2.2.1 Action

C'est le concept le plus simple à identifier parmi les mots-clés car il ne nécessite aucune association avec d'autres mots-clés dans notre représentation ontologique. Une liste précise d'actions a été identifiée et une correspondance simple nous permet de savoir si le mot employé est une action ou non.

5.2.2.2 Composant

Un composant est un ensemble d'atomes ou un unique atome identifié dans notre ontologie, soit par son niveau hiérarchique (modèle, chaîne, résidu, atome), soit par sa désignation directe (carbone, alanine, eau). Lorsqu'un composant est identifié, on cherchera toujours l'éventuelle présence d'un ou de plusieurs identifiants associés au sein de la liste de mots-clés extraite de la commande vocale.

5.2.2.3 Identifiant

Un identifiant ne peut être trouvé seul, il est toujours associé à un composant qu'il désigne. Les identifiants ont aussi la particularité de ne pas être recherché à un niveau ontologique mais au niveau de la base de données RDF puisqu'il n'existe pas de listes d'identifiants fixes dans notre ontologie, ces derniers variant entre les systèmes moléculaires étudiés. Lorsqu'un couple composant/identifiant existe au sein de la base de données de fait, ces deux éléments sont regroupés et constituent un groupe syntaxique indépendant.

5.2.2.4 Propriété

Une propriété met en avant la particularité chimique, physique, biologique ou géométrique d'un composant. Les propriétés, à la différence des identifiants, constituent une liste finie dans notre ontologie et n'ont donc pas besoin d'être associé à un composant afin d'être identifiées. Cependant, certaines propriétés sont directement associées à un niveau hiérarchique précis au sein d'une protéine. On parlera par exemple d'un résidu hydrophobe mais jamais d'une chaîne hydrophile. La cohérence d'une propriété avec les concepts *Composant* identifiés dans la commande est donc vérifiée.

Les propriétés vont agir comme un filtre direct sur le sous-ensemble structurel sur lequel l'utilisateur veut agir. Ce filtre agira soit en unique sélecteur, rassemblant tous les individus

appartenant à un groupe hiérarchique possédant la propriété identifiée, soit en filtre supplémentaire d'un sous-ensemble structurel ayant été identifié en parallèle grâce à d'autres concepts *Composant*.

5.2.2.5 Représentation

Une représentation est une propriété visuelle associée directement et presque exclusivement à l'espace de visualisation. Les mots-clés désignant des concepts *Représentation* sont associés aux actions centrées sur les changements de visualisation. Ils décrivent les états visuels sur lesquels l'utilisateur voudrait intervenir. Ils rassemblent par exemple les méthodes de représentation utilisés dans le visualiseur moléculaire, les couleurs utilisées, etc.

Une fois que chaque mot-clé est validé, c'est à dire identifié comme un concept (ou un individu pour les identifiants) et éventuellement associé à un autre mot-clé, il constitue un groupe syntaxique. Chaque groupe syntaxique possède une information qui correspond à différentes parties de la commande métier. Chaque type de commande est construite autour d'un agencement précis de groupes syntaxiques. On remarque que l'étape de classification possède une forte tolérance à l'ordre des mots-clés, ceux-ci pouvant être introduits dans n'importe quel ordre.

5.2.2.6 De la commande vocale par mots-clés à la commande applicative

Au sein de notre programme, une commande vocale est composée d'une succession de groupements syntaxiques distincts mais liés les uns aux autres pour former une requête d'action. C'est l'unique type de commande présent pour le moment au sein de notre plateforme mais ses différentes déclinaisons permettent d'appliquer un large choix d'actions différentes sur les espaces de visualisation et d'analyses. Il est possible de décrire le type de commande que nous avons défini de la façon suivante :

$$action\ [paramètre]^+,\ (\ groupe\_structurel\ [identifiant]^+\ )^+$$

Les groupes syntaxiques sans crochets [] sont obligatoires à l'inverse de ceux entre crochets qui sont optionnels. Le $^+$ désigne la possibilité d'avoir 0, 1 ou plusieurs occurrences du groupe syntaxique. Enfin les parenthèses désignent un bloc de groupes syntaxiques.

Cette représentation des groupes syntaxiques est présente dans notre ontologie. En effet, le concept *Action* comporte dans sa définition ontologique un certain nombre de concepts associés, sous forme de pré-requis, nécessaires à son exécution. Ces pré-requis sont indispensables car ils se comportent comme des paramètres de la commande exécutée. Une action de type *Color* nécessite par exemple la présence d'une propriété de type *Colors* et d'une suite de mots-clés désignant un sous-ensemble structurel.

De la même façon que pour l'action, la désignation d'un sous-ensemble structurel est obligatoire et peut être obtenu de différentes manières, dépendant directement du critère choisi par l'utilisateur pour filtrer les données sur lesquelles il souhaite appliquer son action. Les différentes manières d'obtenir un sous-ensemble structurel sont :

1. Composant seul, l'ensemble des individus appartenant au concept indiqué sera pris en compte.
2. Combinaison d'un composant moléculaire et d'un ensemble d'identifiants (unique, listé ou sous forme de plages continues). Cohérence entre identifiant et composant requis.

3. Propriété seule, l'ensemble des individus possédant la propriété indiquée sera pris en compte.

4. Combinaison d'un composant et d'une propriété. Cohérence nécessaire entre propriété et composant.

Par soucis d'homogénéité, chaque sous-groupe structurel, même s'il est désigné par un concept sans identifiant, est ramené à la liste d'individus du composant. SPARQL permet de récupérer directement les individus désignés par le sous-groupe structurel. Cette étape permet de désambiguïser les résultats entre les commandes. La commande générée sera par contre plus complexe puisqu'elle comportera à chaque fois la liste détaillée des individus concernés par la commande. Le niveau hiérarchique des individus de la liste sera déterminé par défaut par le niveau structurel hiérarchique du moment où est interprétée la commande vocale. Ainsi, la commande générée pour la sélection d'un modèle pendant une étape d'analyses d'acides aminés comportera une liste de l'ensemble des résidus du modèle sur lesquels appliqués la sélection. Cela permet une simplification de notre moteur mais entraîne une légère complexification de la commande générée, sans impact sur l'utilisateur et sans conséquences sur l'exactitude de la commande.

Par défaut, le sous-ensemble structurel désigné dans la commande vocale dépendra directement du jeu de données affiché dans l'espace de visualisation afin d'éviter une surcharge de précision à donner par l'utilisateur. Par exemple, si seulement deux modèles sont affichés dans l'espace de visualisation, la commande *Alanine 147* désignera uniquement les alanines 147 des deux modèles affichés, pas celles des modèles présents mais non affichés.

Si tous les pré-requis associés à une action sont respectés, alors la commande est ordonnée et transformée en fonction de l'environnement de visualisation utilisé. La dernière étape est une simple transformation des concepts et des individus en une formulation compréhensible par l'application de visualisation moléculaire utilisée.

Performances

Les performances de notre moteur d'interprétation ont été testé sur plusieurs commandes vocales, simples et complexes, et les temps d'exécution ont été relevés. Par soucis de clarté du tableau de résultat, nous avons fait les tests sur une base de données RDF contenant les informations d'une simulation moléculaire d'un peptide de 19 acides aminés dont la séquence primaire est KETAAAKFERQHMDSSTSA. L'ontologie utilisée est celle créée pour notre plateforme. Nous nous plaçons dans un contexte où le niveau hiérarchique structural de l'environnement est l'acide aminé, principalement pour profiter des nombreuses propriétés associées à ce niveau hiérarchique dans l'ontologie et ainsi pouvoir énoncer des commandes complexes. La syntaxe des commandes est adaptée pour être interprétable par le logiciel PyMol. Enfin, ces tests ont été effectué indépendamment du logiciel Sphinx afin de pouvoir les comparer entre-eux sans effets de bord des performances de l'interpréteur vocal.

Table 5.1 – Liste des commandes testées pour évaluer les performances du moteur d'inférence.

Mots-clés	Commande attendue	Commande générée	Temps d'exécution
Hide, Lines, Model, 128	hide lines, residue 1+2+3+4+5+6+7+8+ 9+10+11+12+13+14+ 15+16+17+18+19 and model 128	hide lines, residue 1+2+3+4+5+6+7+8+ 9+10+11+12+13+14+ 15+16+17+18+19 and model 128	env. **54** millisecondes
Color, Alanine, Blue	color blue, residue 4+5+6+19	color blue, residue 4+5+6+19	env. **72** millisecondes
Show, Secondary_structure, Residue, [2,5], Cartoon	show cartoon, residue 2+3+4+5	show secondary_structure, residue 2+3+4+5	env. **56** millisecondes
Show, Positive, Residue, Hydrophobic, Ribbon, Chain, A	show ribbon, residue 1+7+12 and chain A	show ribbon, residue 1+4+5+6+7+8+10+ 12+13+19 and chain A	env, **550** millisecondes

Les différentes commandes testées ont été choisies pour leur complexité croissante ainsi que pour représenter l'ensemble des catégories que l'utilisateur peut avoir à énoncer lors de sa commande. La première commande implique de cacher la représentation graphique *lines* pour le modèle 128. La seule difficulté est de trouver les résidus dépendants de l'individu de plus haut niveau hiérarchique indiqué dans les mots-clés. La commande attendue est identique à la commande générée, pour un temps d'exécution convenant parfaitement aux contraintes de temps interactif.

La seconde commande demande de colorer les alanines en bleu. L'action et sa propriété ont été correctement associés et si nous reprenons la séquence primaire, nous trouvons bien les 4 alanines aux positions énoncées par la commande générée dans un temps moyen d'exécution légèrement supérieur à l'exemple précédent, mais toujours en dessous de nos contraintes.

La troisième commande demande de représenter en *cartoon* la structure secondaire des résidus 2 à 5. Une erreur s'est introduite dans la commande générée qui a bien identifié les concepts de *Secondary_structure* et *cartoon* comme équivalents, mais qui n'a pas gardé le bon mot-clé de représentation pour le logiciel de visualisation PyMol. L'ajout d'un filtre plus fin sur la nature des représentations autorisées par le logiciel serait nécessaire. Le temps d'exécution est toujours satisfaisant pour une action interactive. La commande a su parfaitement interpréter le concept d'intervalle d'identifiants que nous lui avons fourni.

La quatrième commande demande au logiciel de montrer une représentation en ruban des acides aminés hydrophobes et positifs de la chaîne A. On observe une différence entre les deux commandes générées qui ne comportent pas la même liste de résidus. En effet, alors que la commande attendue rassemblait les acides aminés à la fois hydrophobes et positifs (lysine K et histidine H), la commande a généré une liste d'identifiants des résidus positifs (lysine K, arginine R et histidine H) ou hydrophobes (alanine A, phénylalanine F et méthionine M). Cette erreur met en avant la difficulté d'interprétation par mot-clé dans le cas d'utilisation des connecteurs logiques «et/ou». Il est nécessaire de prendre en compte ces deux possibilités et d'ajouter leur interprétation au sein du moteur d'inférence. Enfin, le temps d'exécution de la requête, plus complexe, a augmenté d'un facteur 10 par rapport aux commandes précédentes. Les traitements successifs des propriétés et composants utilisés pour filtrer les résidus concernés participent à cette augmentation.

Limites et perspectives

Notre moteur d'inférence est capable de convertir un large panel de listes de mots-clés,

désordonnées ou non, en une commande logicielle fonctionnelle et interprétable par un logiciel de visualisation moléculaire cible. Il présente cependant quelques limites apportant d'intéressantes opportunités d'extension pour un travail futur. Nous avons vu que l'implémentation de la notion de connecteurs logiques était indispensable afin de pouvoir gérer les situations de filtres multiples sur les individus. Ces connecteurs logiques peuvent difficilement s'intégrer à notre ontologie métier, n'appartenant vraiment à aucun des 5 ensembles de définition autour desquels s'est axée sa construction. Le traitement de ces connecteurs doit donc se faire d'abord au niveau de la reconnaissance vocale afin de les identifier au sein d'une commande utilisateur pour être intégrer ensuite au sein du moteur lors de la structuration de la commande.

Il est important de noter que l'efficacité du moteur d'inférence dépend aussi de la qualité des mots-clés recueillis par l'étape de reconnaissance vocale dans notre implémentation exemple, mais plus généralement par l'étape de génération de ces mots-clés. Une absence d'un ou plusieurs mots-clés ou la reconnaissance d'un mot-clé erroné constituent des erreurs que pouvant être considérées comme courantes. Afin de permettre un processus d'énonciation de commande plus pédagogique et intelligent que le simple retour d'erreur invitant à répéter la commande, il est possible de se servir de la connaissance métier accumulée dans l'ontologie pour proposer à l'utilisateur un sous-ensemble contrôlé de mots-clés pertinents pour compléter la commande. Cette caractéristique participe à l'effort de fournir un mode d'interaction informé entre l'expert et son espace de visualisation, facilitant ainsi l'expérience utilisatrice. Dans le même ordre d'idée, la possibilité de fournir à l'expert un nombre fini d'identifiants pour exécuter sa sélection pourrait anticiper certaines erreurs utilisatrices. Il serait donc possible de désambiguïser un mot-clé identifier comme non conforme à ce qui serait attendu ou compléter une commande partielle dont un ou plusieurs mots-clés seraient manquants.

5.3 Implémentation logicielle de la sémantique pour la biologie structurale immersive

L'ontologie métier définie dans la section 5.1.2 et la base de données générée à partir de données de simulation moléculaire dans la section 5.1.3 nous ont permis de mettre en place un espace de travail combinant visualisation et analyses qui repose exclusivement sur une interrogation de la base de données grâce à SPARQL.

La conception de notre programme repose sur une architecture logicielle complexe. Dans le schéma illustré dans la Figure 5.4, nous l'avons volontairement placé au centre d'une boucle bi-latérale de communication reliant l'espace de visualisation à l'espace d'analyse.

Notre base de données est hébergée dans un serveur local accessible depuis le réseau afin de garantir un accès privilégié et optimisé à nos données.

Notre cas d'étude est une simulation moléculaire test reprenant l'évolution structurelle et énergétique d'une protéine transmembranaire, GLIC, pendant 2 ns ($2 * 10^{-9}$s). Des modèles statiques de cette évolution ont été générés tous les 8 ps ($2 * 10^{-12}$s) créant ainsi 251 fichiers PDB. Le système contient également un ligand identifié de GLIC, le bromoforme, dont on cherche à connaître le mode de liaison et surtout son impact sur la forme de GLIC au cours du temps. Le solvant utilisé pour la simulation est un modèle simplifié d'eau.

5.3.1 Création de la base de donnée RDF

Comme nous l'avons évoqué précédemment, nous avons mis en place une ontologie OWL regroupant l'ensemble des concepts que les experts seraient amenés à manipuler ou visualiser

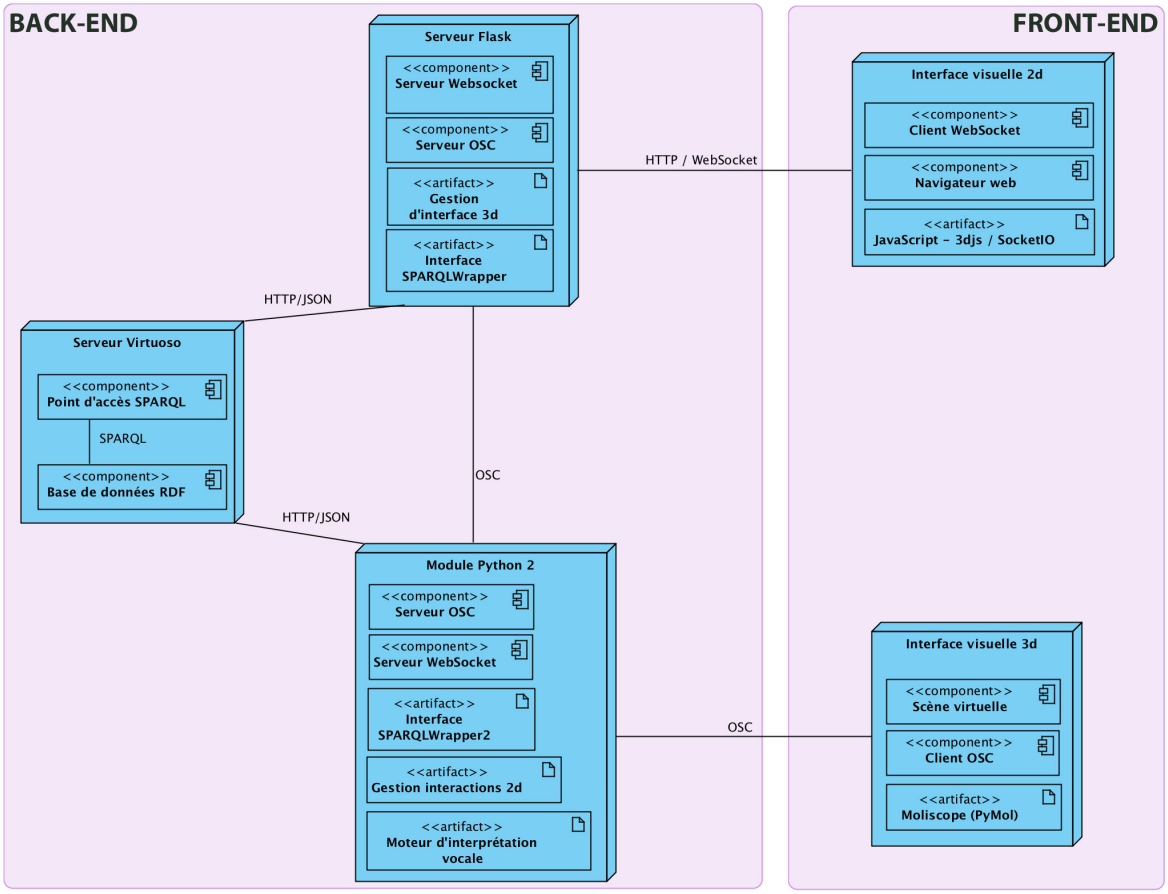


Figure 5.4 – *Diagramme de déploiement des modules et leur connectivité au sein de notre plateforme immersive de Visual Analytics.*

durant leur travail au sein de notre programme. Cette ontologie a été créée grâce à Protégé[3], une structure logicielle et éditeur d'ontologies regroupant plusieurs outils pour la création, l'édition et l'interrogation d'ontologies OWL.

Concernant la partie expérimentale, nous sommes partis de données réelles de simulations moléculaires générées par des experts du domaine. La trajectoire étudiée correspond à une dynamique moléculaire d'une protéine transmembranaire, GLIC, et d'un ligand connu pour être un de ses inhibiteurs, le bromoforme. Cette simulation de 2 nanosecondes a été effectué en solvant explicite grâce à un modèle d'eau standard et met en évidence la présence du bromoforme dans plusieurs sites de liaisons de la protéine. GLIC est une protéine responsable de la transmission de signaux de la cellule puisqu'elle régule le passage de cations entre la partie extra- et intra-cellulaire des cellules. Plusieurs anesthésiques ont été identifié comme possédant une affinité suffisante avec GLIC pour qu'une liaison s'effectue, le bromoforme étant un des ligands concernés. La structure générale de GLIC ainsi que du bromoforme peut être observée dans la Figure 5.5. Le choix de GLIC ne fut pas anodin, trois grandes raisons ont dictées notre décision :

1. C'est une protéine bien connue des experts avec lesquels nous travaillons, il est ainsi aisé d'identifier les analyses pertinentes que nous pourrons associer par la suite.

3. http://protege.stanford.edu/

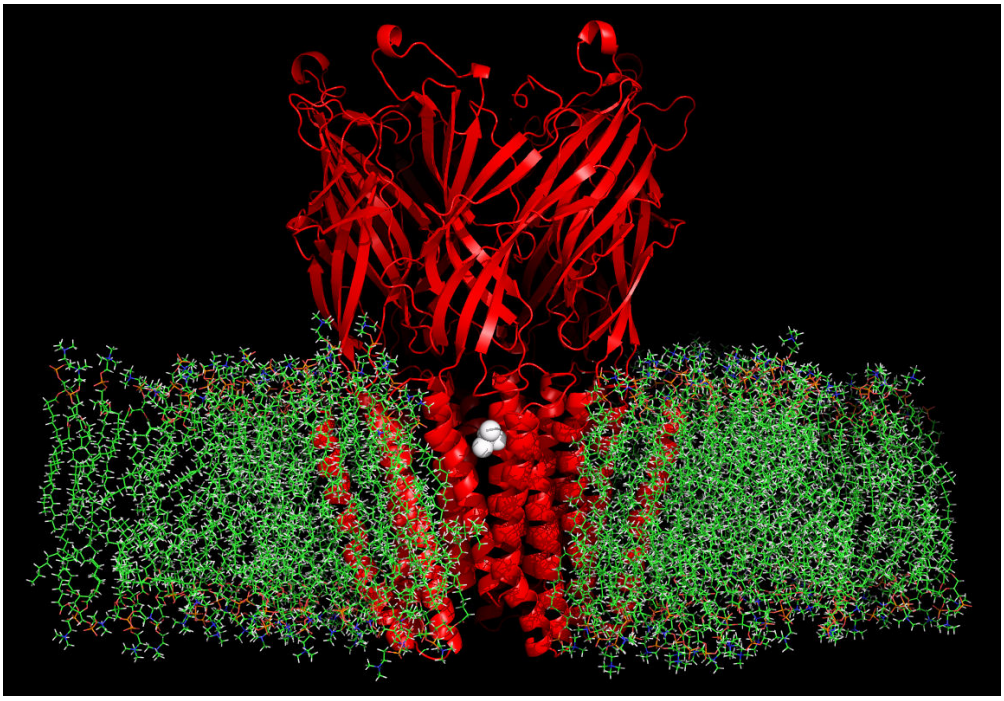


FIGURE 5.5 – *Rendu graphique de la liaison d'un bromoforme (sphères blanches) sur la protéine GLIC («cartoon» rouge) au sein d'une membrane cellulaire (lignes vertes).*

2. La taille de GLIC est relativement importante, en particulier lorsque la membrane est représentée, nous considérons donc ce complexe comme un cas d'étude intéressant pour tester les limites de nos méthodes en terme de performance.
3. L'environnement de GLIC est très hétérogène, permettant de manipuler des concepts comme les membranes et les ligands et de vérifier la robustesse de notre programme pour des concepts particuliers.

La transformation des données expérimentales en triplets RDF s'est faite grâce à Jena, une librairie Java permettant la manipulation de données RDF/OWL via le langage Java. Cette librairie permet de charger une ontologie sous différents formats, dont plusieurs supportés par Protégé lors de l'export, et de s'appuyer dessus afin de créer les individus correspondants aux concepts ontologiques définis. Chaque individu découle d'un ensemble de données obtenu à partir des fichiers caractérisant la simulation moléculaire étudiée. Dans notre cas, nous nous appuyons essentiellement sur des fichiers PDB décrivant la structure 3d des modèles à chaque pas de temps de la simulation. Ces fichiers possèdent les identifiants des différents groupes structurels qui seront manipulés (atomes/acides-aminés/chaînes), les coordonnées 3d des atomes ainsi que les liens des groupes structurels entre-eux. Des fichiers regroupant des données expérimentales telles que les énergies ou les températures ou bien des données analytiques comme les distances inter-atomes sont également utilisés en entrée de notre programme Java afin que leurs valeurs soient associés aux individus qu'ils décrivent. Les triplets créés dans Jena sont sauvegardées dans des fichiers texte sous différents formats : RDF/XML, N3 ou Turtle. Le processus est schématisé dans la Figure 5.6.

L'ontologie est faite de telle sorte qu'il soit possible d'ajouter *a posteriori* des valeurs physico-chimiques ou d'analyses que l'utilisateur voudrait associer à la trajectoire étudiée. Il est par exemple possible d'effectuer une analyse d'accessibilité des atomes ou des acides-aminés à la suite de la simulation et de charger les résultats pour les individus déjà existants

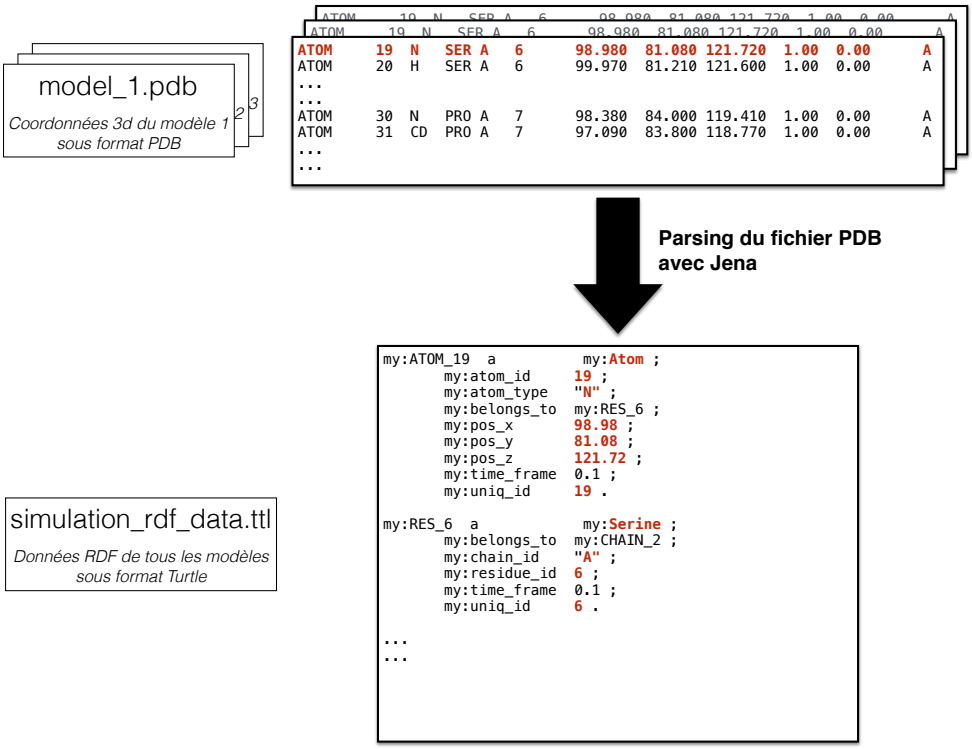


Figure 5.6 – *Schéma du processus de parsing effectué au sein d'un module Java utilisant la librairie de gestion de données OWL/RDF(S) Jena. Les fichiers PDB ont été obtenus à partir d'une simulation moléculaire et chaque PDB représente la conformation spatiale du système simulé à un instant t.*

dans la base de données. Cette possibilité nous permettra également de mettre en place des modules «optionnels» d'analyses à la demande que l'utilisateur pourra exploiter pendant sa session de travail et qui viendront ajouter les résultats obtenus directement dans la base de données pour une visualisation de ceux-ci de façon complètement intégrée. En plus du gain de temps et d'uniformisation, cela permet de sauvegarder toutes nouvelles données générées lors de la session de travail et donc de les réutiliser par la suite sans besoin de relancer les modules d'analyses ayant servi à leur création.

Les fichiers générés, celui de l'ontologie et celui des triplets RDF, sont ensuite chargés au sein d'un serveur *Virtuoso*[4] qui est un serveur de gestion de données dont l'architecture permet le stockage de données RDF de façon optimisée et met en place un point d'accès à ces données via un espace de requêtes SPARQL. Ce point d'accès se fait via une adresse URL que la plupart des librairies prenant en charge le SPARQL utilisent comme interface sur les données RDF. La version du serveur Virtuoso que nous utilisons permet la mise en place des règles d'inférence OWL-Lite pour chaque requête SPARQL effectuée, aspect très important pour notre plateforme.

4. http://virtuoso.openlinksw.com/

5.3.2 Gestion des données RDF et requêtes SPARQL

Dans le but de mettre en place une interface entre les différents modules et la base de données RDF, nous avons conçu un module spécifique contenant un ensemble de fonctions permettant de lancer des requêtes SPARQL sur la base de données. Cette interface sous forme de module Python utilise la librairie SPARQLWrapper[5] qui permet d'accéder à tout point d'accès SPARQL via une URL, de lancer une requête et de récupérer les résultats sous format JSON, ensuite facilement convertissables en dictionnaires Python. Les résultats sont ensuite transmis aux modules ayant appelé une fonction de cette interface. Dans notre cas, le serveur Virtuoso hébergeant la base de données fournit le point d'accès SPARQL nécessaire et permet ainsi l'interfaçage du module avec la base de donnée RDF que nous avons créé (cf. Figure 5.4).

5.3.3 Visualisation des données moléculaire 3d

Nous avons choisi PyMol [50] comme programme de visualisation moléculaire 3d. PyMol est un logiciel de visualisation très utilisé dans la communauté scientifique et qui possède une API complète qui facilite son intégration au sein d'une plateforme logicielle basée sur des modules hétérogènes. Nous avons utilisé l'API de Pymol pour automatiser plusieurs étapes du chargement des trajectoires de simulation. Au-delà du simple chargement visuel des structures 3d de molécules et de leur rendu, nous avons mis en place au sein de PyMol des processus spécifiques surveillant les actions de l'utilisateur dans l'espace de visualisation. Plusieurs actions, dont notamment celle de sélection, peuvent déclencher un événement qui se répercutera de façon synchrone avec l'espace d'analyse. Le portage de PyMol dans des environnements immersifs est un projet en cours sous le nom «Moliscope» et nous profitons des derniers développements de ce projet pour notre programme (voir Figure 5.7).

Nous avons ajouté notre moteur d'interprétation de commandes vocales vers des commandes logicielles. Notre moteur est relié à PyMol afin de donner la possibilité à l'utilisateur d'effectuer plusieurs actions pré-définies de façon simple et adaptée aux environnements immersifs. PyMol est la partie frontale de notre application, elle est reliée indirectement à un module de gestion générique qui s'occupe de recevoir et d'émettre les informations et les événements entre l'espace de visualisation et l'espace d'analyse. Ce module est aussi construit de façon à pouvoir être relié à n'importe quel logiciel de visualisation. Il est donc indépendant de PyMol et transmet les événements liés à la visualisation à un moteur spécifique qui va s'occuper de transmettre des commandes logicielles en commandes PyMol dont la syntaxe lui est propre. PyMol pourrait donc être remplacé par un autre logiciel de visualisation moyennant une simple modification du moteur de commandes afin de l'adapter aux commandes spécifiques du logiciel utilisé.

5.3.4 Visualisation des résultats d'analyses 2d

Afin de visualiser les différents graphiques résultant des analyses effectuées *a priori* ou *a posteriori* de la session de travail, nous avons mis en place une interface web de visualisation/interaction avec ces graphiques. Nous souhaitions garantir une utilisation générique et indépendante du type de dispositif d'interaction. Pour ce faire, nous nous appuyons sur une librairie JavaScript de création de graphiques au sein de pages HTML appelée d3js[6]. Cette librairie possède plusieurs outils de chargement de données via différentes sources, sous forme

5. https://rdflib.github.io/sparqlwrapper/
6. http://d3js.org/

Figure 5.7 – *Photo prise dans EVE, système CAVE du LIMSI/CNRS pendant une session de travail collaborative sur Moliscope.*

de fichiers ou à travers des flux de données. D3js permet de mettre en place des graphiques de façon simple mais extrêmement paramétrables dans leur apparence ainsi que leur nature. S'appuyant sur JavaScript, elle est directement intégrée dans le code HTML d'une page web et accessible à travers toutes les plateformes permettant de visualiser du contenu HTML.

Nous utilisons un serveur web basé sur une librairie python, Flask, afin de mettre en place notre environnement web. Plusieurs extensions existent dont Flask-SocketIO qui permet aux applications Flask d'accéder à une communication bi-directionnelle et de faible latence entre le serveur et le client web. Côté client, la librairie JavaScript SocketIO est suffisante pour pouvoir envoyer et recevoir des messages vers ou depuis le serveur web Flask. Une connexion permanente peut ainsi être mise en place entre le client et le serveur (voir Figure 5.4).

La plupart des événements d'interaction déclenchés par l'utilisateur au sein de la page HTML et des graphiques d3js peuvent ainsi être répercutés directement au serveur web. Suivant la nature des événements, le serveur renvoie ensuite les informations aux modules concernés. L'espace de visualisation des résultats d'analyses est divisé en niveaux hiérarchiques structuraux qui correspondent à ceux référencés dans l'ontologie : *Model* / *Chain* / *Residue* / *Atom*. Les graphiques dessinés dans chaque niveau correspondent à des individus du niveau hiérarchique représenté. Afin de garder une cohérence entre les graphiques représentés, chaque sélection effectuée dans un graphique entraîne une mise à jour des graphiques présents dans les niveaux hiérarchiques inférieurs. Cette mise à jour se fait grâce à une requête émanant du script JS relayée ensuite au module de gestion RDF qui lancera une requête SPARQL pour les propriétés affichées dans les graphes mis à jour mais seulement pour les individus sélectionnés par l'utilisateur (voir Figure 5.8).

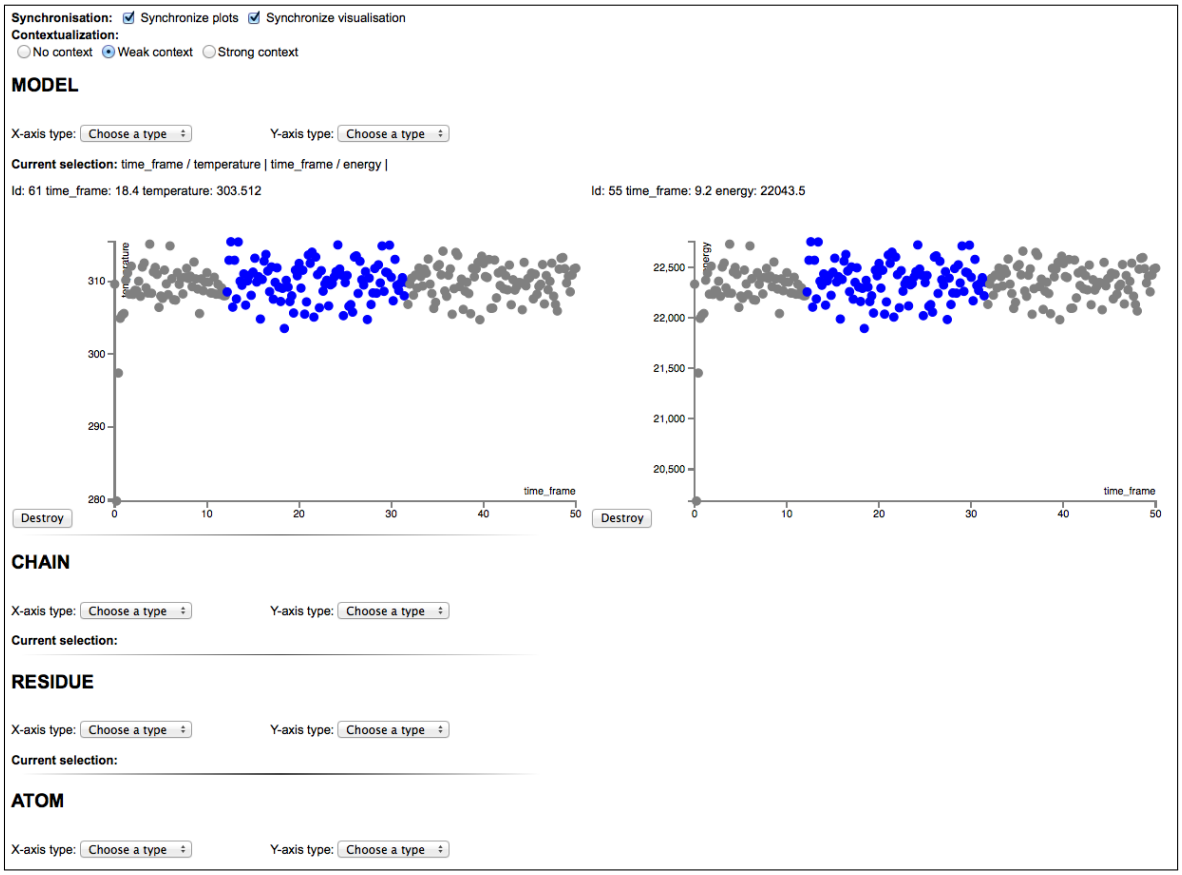


FIGURE 5.8 – *Interface web 2d permettant la mise en place des graphes d'analyse sur les donnée de la base RDF. Les individus sélectionnés (en bleu) dans un graphe apparaissent également dans les graphes du même niveau hiérarchique.*

5.3.5 Analyses semi-automatiques

Bien que la majorité des données soient présentes au sein de la base de données créée par l'utilisateur, une session de travail habituelle requiert souvent des données résultantes de calculs post-simulation et donc absentes de la base de données de départ. Ces calculs sont habituellement gérés au sein de scripts, liés ou non aux outils de simulation, et exécutés en dehors de la boucle de visualisation suite à l'observation de phénomènes particuliers lors de l'exploration des structures de la simulation ou suite à d'autres analyses déjà effectuées auparavant. Afin de ne pas surcharger la base de données et laisser l'utilisateur maître des analyses qu'il veut effectuer, nous avons profité de notre connaissance métier pour mettre en place la possibilité de lancer certaines analyses semi-automatisées pendant la session de travail.

La force du langage de requête SPARQL est qu'il permet, en plus d'interroger une base de données, de modifier ces données, les supprimer ou en ajouter. Nous utilisons la possibilité d'ajouter des données pour permettre à l'expert d'alimenter la base de données avec des résultats d'analyses lancées pendant sa session de travail. Une liste d'analyses a été construite et chacune des analyses proposées possède sa définition ontologique. Cette définition nous permet de connaître le type de données utilisé en entrée de l'analyse et le type de données en sortie. Ainsi, suivant l'analyse désirée, la plateforme proposera un choix filtré d'individus à sélectionner, correspondant au type de donnée attendu. De la même manière, les valeurs

générées en sortie de l'analyse sont automatiquement introduites dans la base de données en respectant leur définition ontologique.

L'outil *Distance* nécessite par exemple deux individus de même niveau hiérarchique structural ainsi qu'une sélection d'individus de niveau hiérarchique supérieur au sein desquels seront calculés ces distances.

Il est possible de classer ces analyses en deux catégories : Les analyses simples regroupent les analyses générant une valeur pouvant être ajoutée directement aux propriétés des individus concernés. Parmi celles-ci nous pouvons citer l'accessibilité au solvant, l'hydrophobicité, l'énergie, etc. Les analyses complexes sont le résultat d'une propriété décrivant un rapport entre deux individus et nécessitant donc une connaissance de ces individus pour être pertinentes. Nous pouvons citer parmi les analyses complexes reliant deux individus : la distance entre deux atomes, le RMSD entre deux ensembles d'individus, l'angle entre deux chaînes, etc.

Alors que les analyses simples ajoutent simplement à un individu une propriété et la valeur associée, les analyses complexes doivent elles créer une instance particulière d'un des concepts *Analyse* de l'ontologie qui regroupera les informations nécessaires à sa compréhension.

Le concept *Distance* (de type *Analyse*) de l'ontologie permettra par exemple de stocker toute distance calculée entre deux individus pour une sélection de structures parentes définies. La valeur de la distance, l'URI des deux individus impliqués ainsi que l'ensemble des structures au sein desquelles le calcul s'est effectué seront des propriétés d'une instance *Distance* et seront accessibles seulement à travers cette instance.

Cela ne complexifie donc pas seulement le stockage mais également l'accès aux informations puisque les valeurs d'analyses complexes ne sont, au contraire des analyses simples, plus directement associées aux individus concernés mais à travers un concept *Analyse* intermédiaire. Il fut donc nécessaire d'adapter les requêtes SPARQL utilisées pour la génération des graphes afin de prendre en compte cette complexité. La différence entre une requête SPARQL accédant aux valeurs d'une analyse simple et la requête SPARQL permettant d'accéder à celles d'une analyse complexe est illustrée dans le Listing 5.3.

Listing 5.3 – Deux requêtes SPARQL : 1. Accès à la température d'un modèle 2. Accès à la distance entre deux résidus

```
SELECT DISTINCT ?temp WHERE {my:MODEL_161 my:temperature ?temp}

SELECT DISTINCT ?distance WHERE {?indiv rdf:type my:Distance . ?indiv
  my:objectA my:RES_3622 . ?indiv my:objectB my:RES_3626 . ?indiv
  my:distance ?distance}
```

La liste des analyses implémentées jusqu'à présent peut être retrouvée dans le tableau 5.2.

Table 5.2 – Liste des analyses semi-automatiques présentes. A=Atom, R=Residue, C=Chain, M=Model.

	Distance	**RMSD**	**Angle**
Description	Distance entre deux points 3d (coordonnées pour les atomes / centres de masse pour les groupes d'atomes). Groupes possibles : **A,R,C,M**	RMSD entre deux groupes d'atomes identiques (mêmes identifiants biologiques) après superposition des atomes du squelette de la protéine. Groupes possibles : **R,C,M**	Angle entre deux segments formés par deux paires de deux atomes ou deux groupes d'atomes. Groupes possibles : **A,R,C**

5.3.6 Synchronisation

La nature hétérogène des modules de notre application nous empêche d'adopter une communication directe entre les différentes instances Python créées par les différents modules. La synchronisation entre l'espace visuel et l'espace d'analyses, représenté respectivement par PyMol et le serveur web Flask, se fait au travers de communications entre leurs modules de gestion respectifs. Cette communication s'appuie sur des messages utilisant le protocole OSC (*Open Sound Control*) envoyées par les modules à chaque événement déclenché dans un des espaces de travail. Chacun des modules est ainsi associé à un serveur OSC opérant en arrière-plan dans un sous-processus et surveillant tout message arrivant sur le port qui lui est dédié. Nous utilisons la librairie Python pyliblo[7] qui est une interface Python pour liblo[8], une implémentation du protocole OSC. OSC est un format de transmission de données conçu pour être utilisé en temps réel. Pyliblo, en plus de permettre la création d'un serveur OSC, fournit également les fonctions nécessaires pour envoyer un message. Ces messages sont envoyés localement sur un port ouvert référencé par la machine qu'un serveur OSC «surveille» afin de détecter tout message entrant.

Les communications constantes entre les modules à travers les ports de nos serveurs OSC permettent donc d'assurer une synchronisation des actions utilisateurs dans chacun des espaces de travail. Les actions identifiées comme synchronisables sont effectuées en coordination entre les espaces.

A chaque création de graphique dans l'espace de visualisation, l'information de création est envoyée afin de créer un objet *Selection* qui regroupera les informations de sélection effectuées au sein du graphique correspondant. Cet objet est référencé au sein de l'ontologie et contient : le niveau hiérarchique structural des individus, les individus sélectionnés dans le graphique, la représentation géométrique associée et le rendu visuel appliqué (couleur et transparence). Ce stockage permet de sauvegarder l'état de la visualisation à tout moment et de la corréler aux analyses présentées au sein de l'espace d'analyse. De plus, un découpage des graphiques en objet *Selection* permet à l'utilisateur de choisir quels groupes de sélection il désire afficher ou cacher au sein de l'espace de visualisation. Il est donc possible de superposer les rendus visuels de chaque graphique ou d'une partie seulement suivant la volonté de l'utilisateur. Ces objets *Selection* sont mis à jour à chaque interaction de l'utilisateur dans le graphique correspondant afin de refléter les individus sélectionnés. La possibilité de sauvegarder l'état de visualisation et des analyses présentes ainsi que des individus sélectionnées permet naturellement de charger un état précis d'une session de travail. Ce détail est particulièrement utile à la fin d'une

7. http://das.nasophon.de/pyliblo/
8. http://liblo.sourceforge.net/

longue session de travail ayant abouti à un rendu visuel satisfaisant.

5.3.7 Scénario métier comme exemple d'usage

Afin d'illustrer le fonctionnement de notre plateforme, nous avons choisi un court scénario métier mettant en jeu les différents développements décrits précédemment. Ce scénario consistera en l'analyse d'une trajectoire de simulation moléculaire d'un peptide de 19 acides aminés et de l'évolution des changements structurels en début de simulation.

Une première interrogation de la base de données permettra d'obtenir l'ensemble des valeurs numériques qu'elle contient pouvant être analysées au sein de graphes. Nous choisirons d'observer les modèles de la base de données à travers leur pas de temps associé et du RMSD calculé par rapport au modèle de départ de la simulation. Nous sélectionnerons ensuite l'ensemble des modèles possédant un RMSD proche de la structure de départ, à savoir l'ensemble des modèles à moins de 0.1Å de cette structure de référence. Cette sélection sera affichée automatiquement dans l'espace de visualisation. Dans cet espace, nous aurons la possibilité de nous intéresser à certains acides aminés considérés comme inhabituellement positionnés au sein des modèles et de les sélectionner. Cette sélection est envoyée à la base de données qui permettra d'identifier les individus correspondant aux acides aminés sélectionnés et permettra la création de graphes d'analyses les mettant en avant. Les valeurs numériques choisies par l'utilisateur et correspondant aux acides aminés sélectionnés sont alors affichées au sein des graphes.

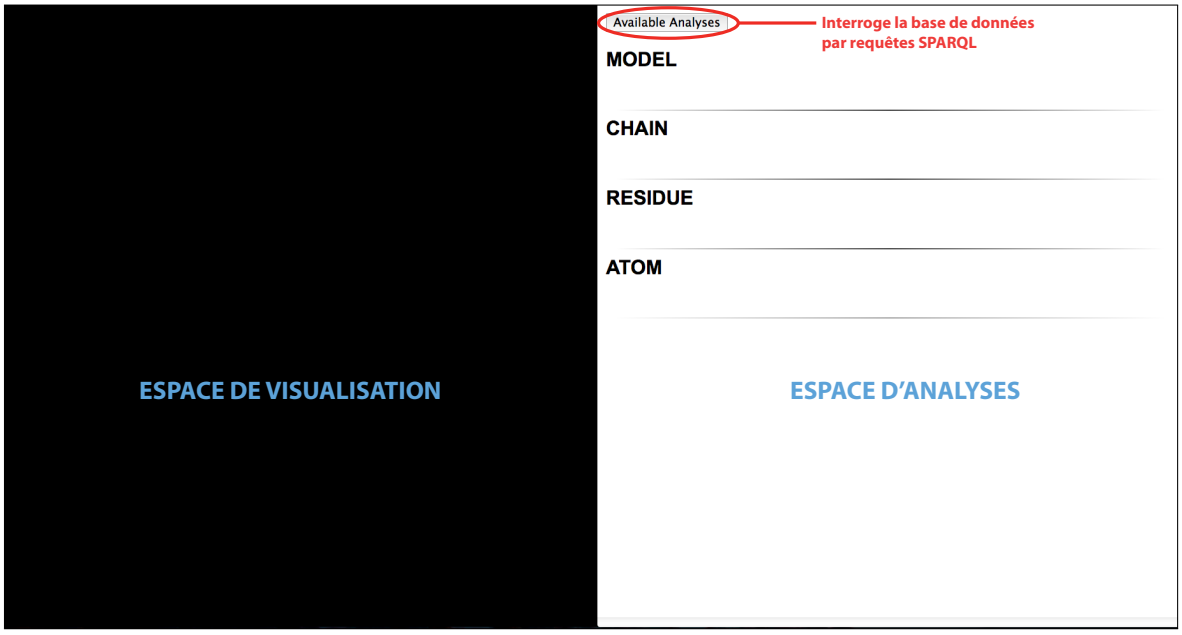


Figure 5.9 – *Première étape de connexion des espaces à la base de données. Ecran d'accueil comportant un simple bouton qui permet de lancer une requête vers la base de données pour rassembler les propriétés pouvant être analysées sur les graphes. A gauche se trouve l'espace de visualisation moléculaire, pour le moment vide, à droite se trouve l'espace d'analyse dans une page HTML.*

Récupération des valeurs numériques dans la base de données RDF

Dans le scénario choisi, le premier événement pris en compte au sein de l'application est le choix de l'utilisateur quant aux analyses qu'il désire afficher. Ce choix lancera une première paramétrisation de l'espace d'analyse qui pourra ainsi se connecter à l'espace de visualisation. Ces deux espaces sont illustrés dans la Figure 5.9. Il est cependant possible de démarrer une session de travail par une exploration de la trajectoire étudiée mais cette exploration ne déclenchera pas d'événements dans l'espace visuel, ce dernier n'étant pas paramétré.

Le choix des analyses se déroule sur la page HTML et va interroger la base de données RDF afin d'identifier les propriétés pour lesquelles une valeur numérique existe pour le groupe structurel choisi par l'utilisateur (cf. Figure 5.10), ici les modèles. Ces valeurs numériques peuvent provenir de données structurales générées par la simulation moléculaire ou bien d'analyses post-simulations intégrées au sein de la base de données. Toutes les propriétés possédant une valeur numérique sont affichées dans deux colonnes, la première permet de choisir la propriété à afficher sur l'axe des abscisses, la seconde désigne la propriété qui sera affiché sur l'axe des ordonnées. L'utilisateur a la possibilité de générer plusieurs combinaisons de propriétés à afficher sous forme de graphes (voir Figure 5.8). Nous choisissons ici de présenter un seul graphe affichant les modèles selon le pas de temps de la simulation qu'ils illustrent et leur valeur de RMSD par rapport au modèle utilisé comme point de départ de la simulation.

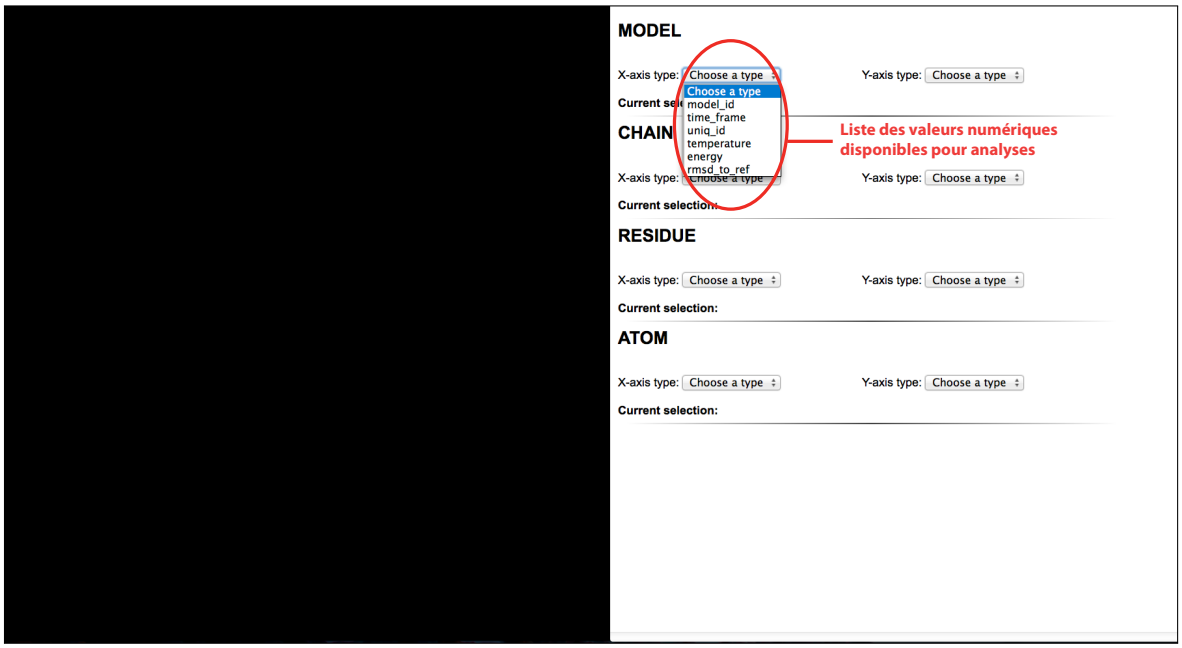


FIGURE 5.10 – *Seconde étape de sélection des valeurs à afficher sous forme de graphes, une liste de propriétés visualisables au sein des graphes est disponible pour chaque niveau hiérarchique structural.*

Quand le choix est terminé, des requêtes SPARQL sont envoyées à la base de données pour récupérer les valeurs des propriétés choisies. Ces données sont ensuite mises sous forme de graphes et structurées.

Lorsque les graphes sont générés, l'utilisateur a la possibilité d'interagir avec eux de différentes manières :

— Il peut sélectionner un point unique afin d'obtenir un ensemble d'informations provenant

de la base de données et liés à l'individu désigné.

— Il est possible de sélectionner un ensemble de points dans un des graphes et afficher un cadre d'information rassemblant plusieurs statistiques sur l'ensemble des propriétés des individus ainsi sélectionnés (moyenne, écart-type, pourcentages, etc.).

— Il est également possible de synchroniser la sélection d'un groupe de points afin de mettre en évidence les individus concernés dans l'ensemble des graphes de la page.

Sélection dans l'espace d'analyse

Toute sélection au sein d'un ou plusieurs graphes de la page déclenche un événement transmis depuis le client web jusqu'au module de gestion 3d entraînant une adaptation de la visualisation pour mettre en avant le ou les individu(s) ayant été sélectionnés. Il est aussi possible d'obtenir, sous forme de fenêtre indépendante, un résumé des informations disponibles sur le ou les individus. Dans le cas d'un individu unique, une liste de ses propriétés et de leurs valeurs sera affichée alors que pour plusieurs individus, une moyenne effectuée sur les valeurs de leurs propriétés remplacera les valeurs brutes. Nous voyons dans la Figure 5.11 qu'une sélection réduit donc le focus de l'utilisateur à un sous-ensemble d'individus, à la fois dans l'espace d'analyses (étape 1 de la figure) mais également de façon synchrone dans l'espace de visualisation (étape 2 de la même figure).

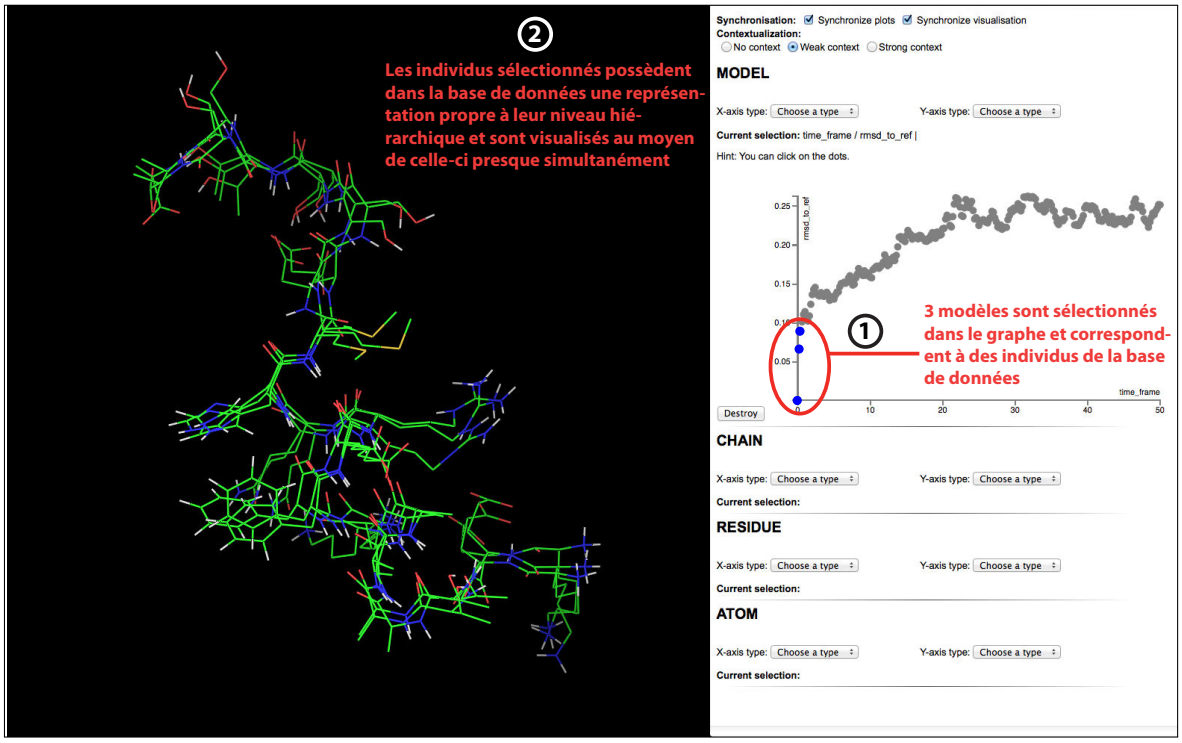


Figure 5.11 – *Sélection des modèles dont le RMSD par rapport au modèle de référence est inférieur à 0.1Å dans l'espace d'analyse (à droite). Synchronisation de l'espace de visualisation (à gauche) après requête SPARQL de la base de données pour identifier les individus sélectionnés et retrouver leur représentation associé.*

Il est possible d'adapter ce focus suivant les besoins de l'utilisateur en modifiant le niveau de contexte dans lequel il désire que sa sélection apparaisse. Trois niveaux de contextualisation sont possibles :

— Aucun contexte - La sélection d'individu(s) entraîne la visualisation unique de ces individus dans l'espace de visualisation et d'analyse et donc cache tout individu n'étant pas sélectionné

— Contexte faible - La sélection d'individu(s) met en avant ces individus dans les espaces de travail et réduit la perception des autres individus du jeu de données (couleur grise, transparence, rendu visuel simplifié, etc.)

— Contexte fort - La sélection d'individu(s) n'est perçu qu'à travers une mise en avant simple de ces individus dans les espaces de travail. Tout autre individu apparaîtra également avec des paramètres visuels proches des individus sélectionnés.

Ces différents niveaux permettent soit de mettre en évidence les différences entre la sélection et le reste du jeu de données, soit de mettre en place un environnement de travail épuré sur une sélection d'intérêt pour l'utilisateur. Ces niveaux s'appliquent à la fois pour la partie visuelle et analytique via des systèmes de rendus visuels propres à chaque espace.

Sélection dans l'espace de visualisation

Lorsque une sous-sélection d'individus est effectuée, l'utilisateur a toute liberté pour l'explorer dans l'espace de visualisation. Si, lors de son exploration, l'utilisateur désire sélectionner un sous-groupe plus précis au sein des individus affichés, il a à sa disposition deux manières de le faire : soit par une sélection directe via un dispositif d'interaction, soit par commande vocale en verbalisant le sous-groupe d'intérêt via la commande *Select*.

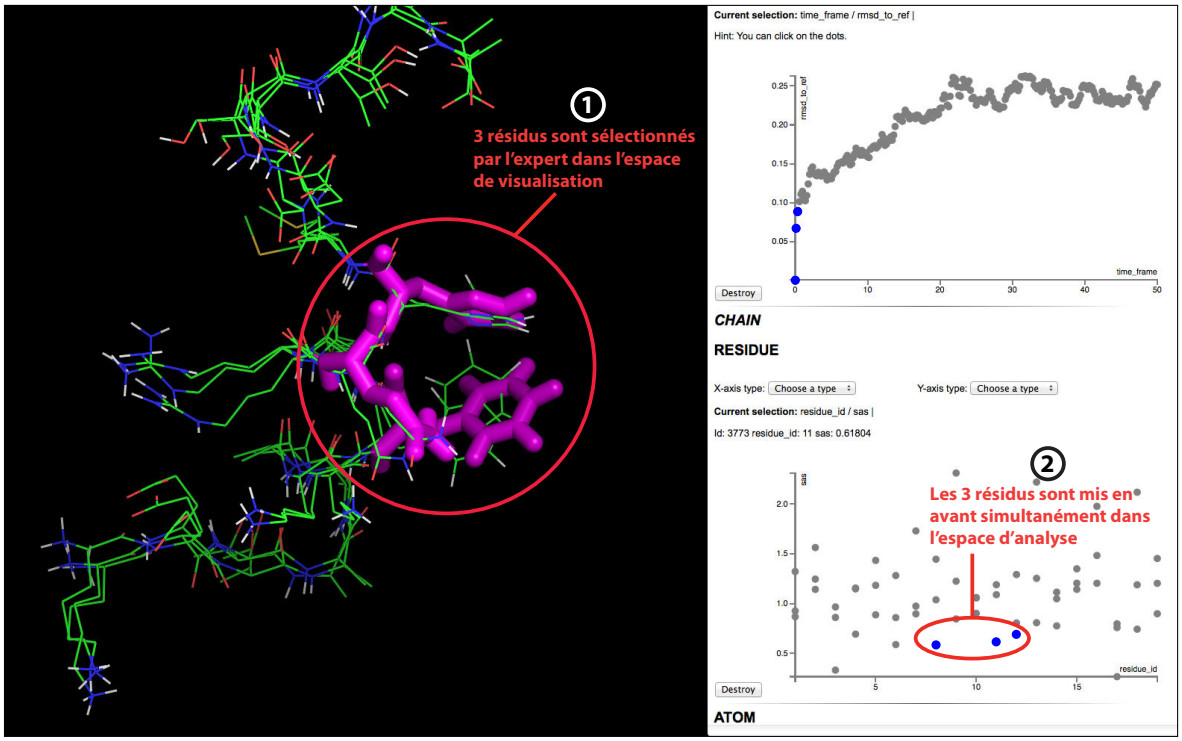


FIGURE 5.12 – *Sélection des acides aminés considérés comme remarquable dans l'espace de visualisation (à gauche). Synchronisation de l'espace d'analyse (à droite) après requête SPARQL de la base de données pour identifier les individus sélectionnés et les mettre en avant au sein des graphes.*

Au-delà du changement visuel induit par la sélection dans l'espace de visualisation, ce changement va également produire un rafraîchissement de l'affichage de l'espace d'analyses, de la même manière que la précédente sélection effectuée dans l'espace d'analyses induisait une synchronisation dans l'espace de visualisation. Des graphes correspondant aux individus sélectionnés seront affichés et mettront en avant les valeurs numériques associées à ces individus. Nous illustrons dans la Figure 5.12 la sélection d'acides aminés (sous-groupes) parmi les trois modèles affichés dans l'espace de visualisation en étape 1.

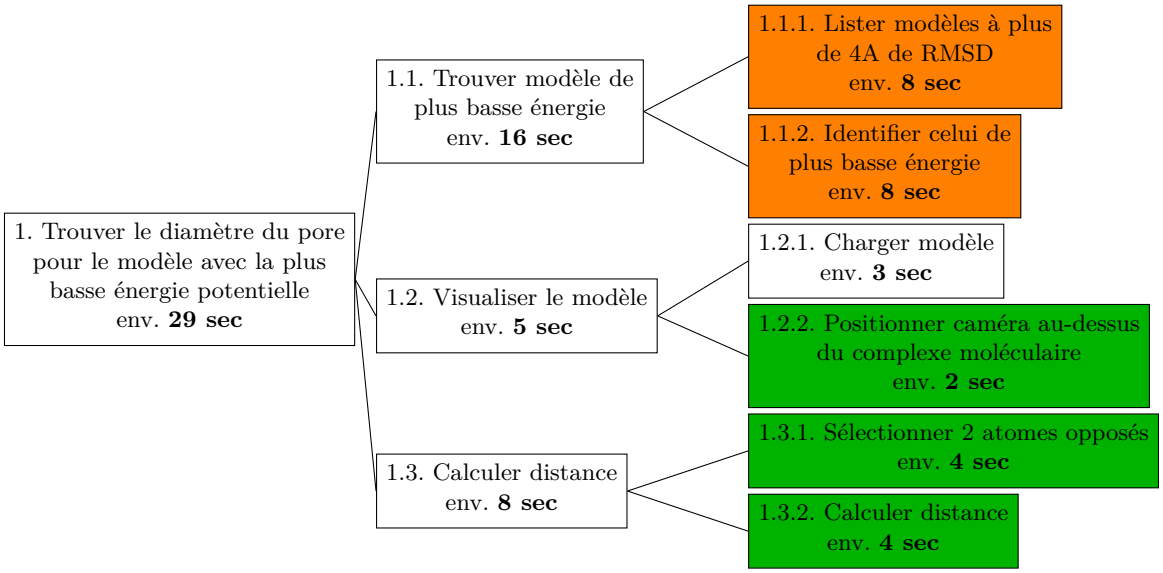
Pour que la synchronisation ait lieu, il est nécessaire qu'au moins un graphe corresponde au niveau hiérarchique structural des individus de la sélection. Si ce n'est pas le cas, la sélection n'aura pas d'effet. Sinon, les individus seront mis en avant dans les graphes concernés. Il est cependant aussi possible, malgré la présence de graphes du bon niveau hiérarchique structural, d'activer un mode asynchrone. Dans ce mode asynchrone, aucune action effectué dans chacun des espaces n'aura d'influence sur le second espace. La synchronisation pourra se réactiver à tout moment, aux conditions exposées précédemment, lorsque les niveaux structurels des deux espaces seront identiques. Ici nous avons choisi préalablement à la sélection dans l'espace de visualisation d'afficher les résidus selon leur identifiant dans chacun des modèles (de 1 à 19) et leur accessibilité au solvant (SAS). Lorsque les acides aminés sont sélectionnés lors de la 1ère étape illustrée dans la Figure 5.12 alors ils sont mis en évidence dans l'espace d'analyse (étape 2 de la même figure).

5.3.8 Évaluation par analyse hiérarchique de la tâche via la méthode HTA

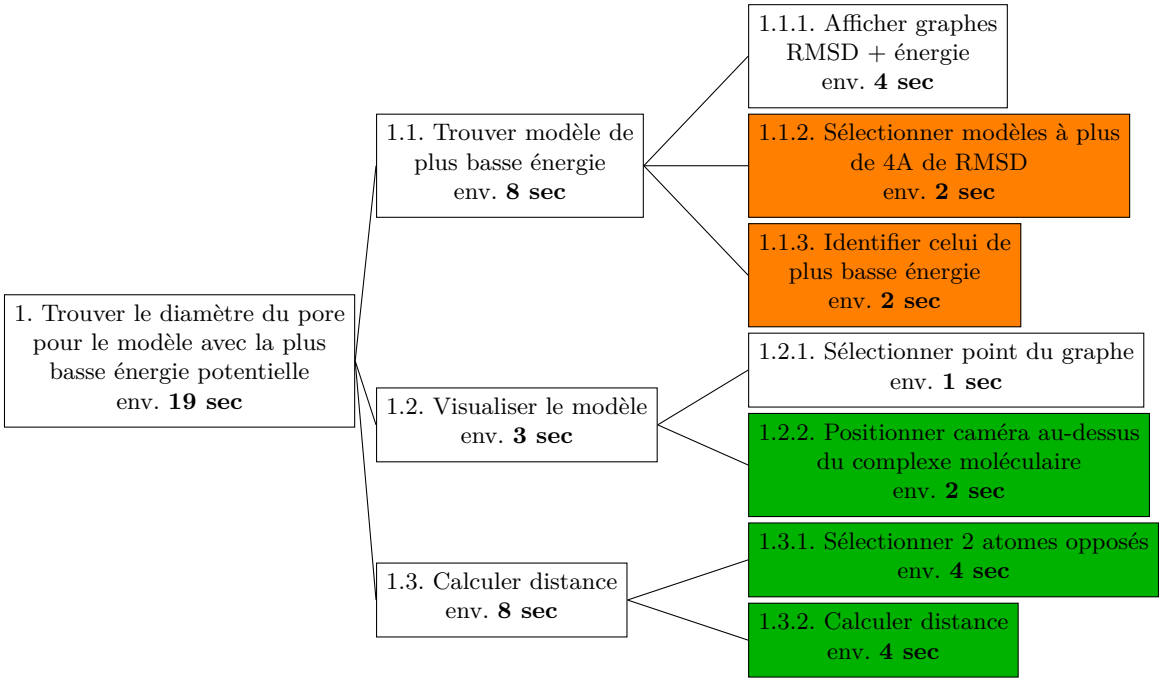
Nous avons repris la méthode HTA introduite dans la section 2.2.7 afin d'évaluer les apports de nos développements pour une tâche experte identifiée et commune en biologie structurale. De la même manière que pour l'évaluation de nos paradigmes de navigation, l'évaluation au moyen de méthodes empiriques d'une plateforme dont les développements se sont basés autour de tâche experte, nous semblait dénué de sens.

Grâce à l'approche hiérarchique HTA, nous comparerons donc deux conditions de travail différentes, la première sera une utilisation classique d'un logiciel de visualisation moléculaire et d'un terminal permettant l'exploration de fichiers d'analyses. Nous ferons dans ces conditions l'hypothèse que l'ensemble des valeurs d'analyse nécessaires ont été obtenues. La seconde condition se basera sur l'utilisation de notre programme, combinant l'utilisation d'un logiciel de visualisation moléculaire, éventuellement immersif et en interaction directe, et d'une interface web regroupant le traitement automatique des analyse en graphiques.

Le scénario de tâche que nous avons choisi consiste à trouver le diamètre du pore principal d'une protéine au sein d'une trajectoire de simulation moléculaire. On s'intéressera au modèle de la trajectoire de plus basse énergie. Cette tâche peut être divisée en trois étapes distinctes. Il nécessite une première étape de traitement des données d'analyse où l'on cherchera le modèle de plus basse énergie parmi les modèles distants de plus de 10Å de RMSD du modèle de référence, traduisant des changements conformationnels importants. Lorsque le modèle concerné est identifié, il convient de le visualiser le complexe moléculaire de façon à voir le pore et pouvoir sélectionner ses extrémités. La troisième étape consiste finalement à calculer la distance entre deux atomes de part et d'autre du pore. On retrouve un temps d'exécution significativement inférieur lors de l'utilisation de notre programme (**19 s**) comparé à une utilisation standard des outils d'analyse et de visualisation (**29 s**). La première étape d'analyse est l'étape où la différence est la plus importante et est mise en avant par les sous-tâches en orange dans le graphe HTA de la figure 5.13. Cette différence s'explique par l'utilisation de graphes interactifs pour visualiser les valeurs de RMSD et d'énergie pour l'ensemble des modèles. Cet outil nous permet de mettre en avant par un simple rectangle de sélection



(a)



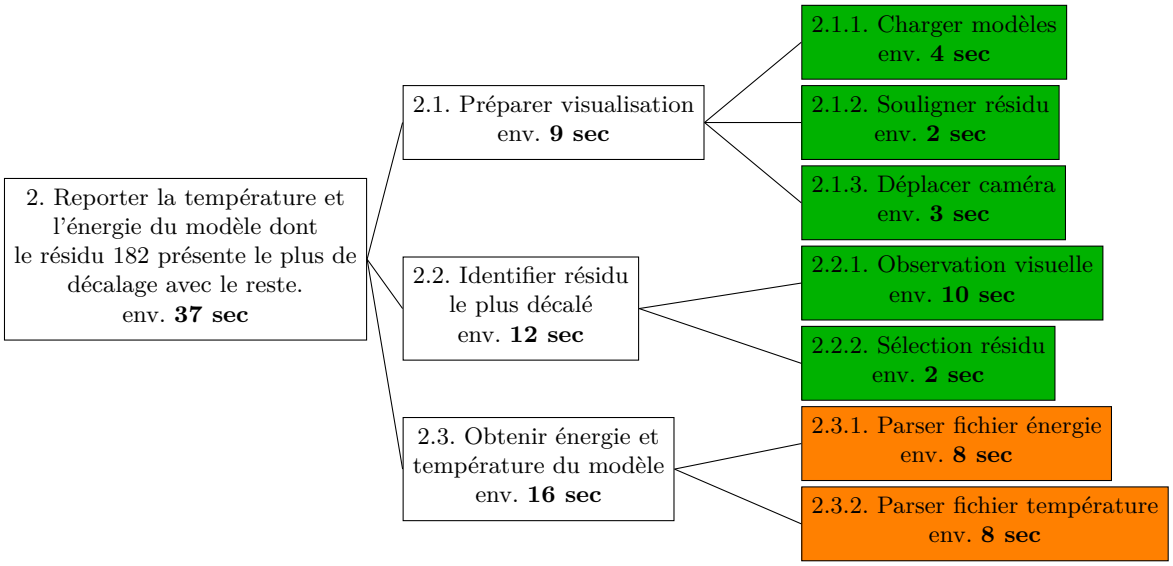
(b)

FIGURE 5.13 – *Subdivision par HTA d'une 1ère tâche experte réalisée (a) en conditions normales et (b) au sein de notre plateforme.*

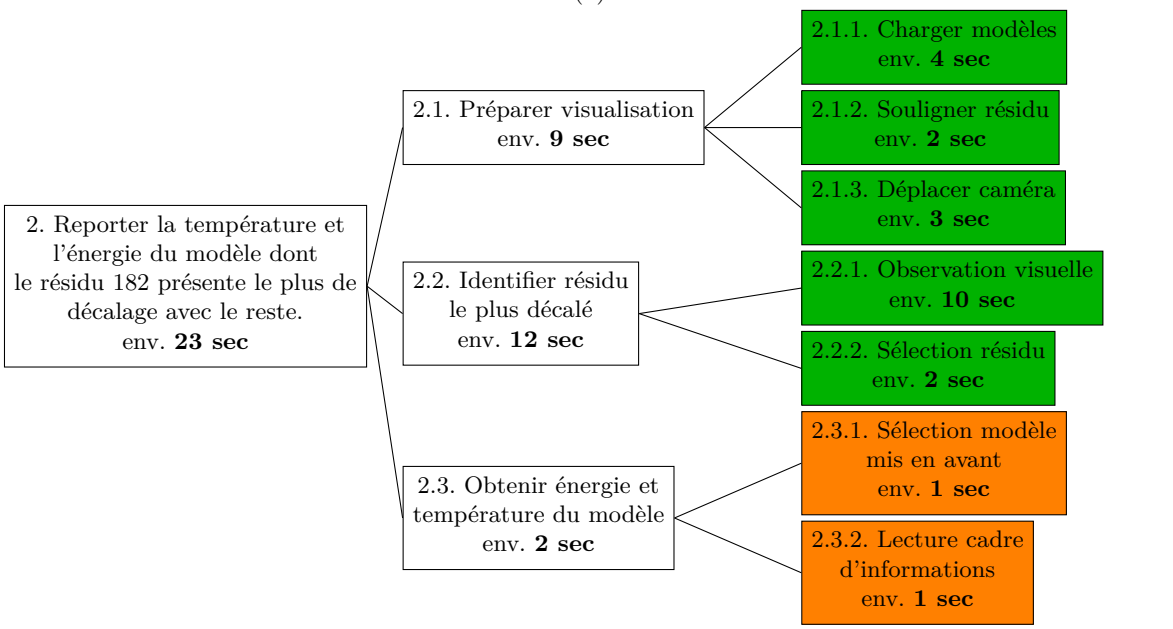
l'ensemble des modèles à plus de 10Å du modèle de référence. L'identification du modèle de plus basse énergie au sein des modèles mis en avant consiste ensuite en une simple analyse visuelle du graphe d'énergie. A l'opposé, l'utilisation d'outils en ligne de commandes est beaucoup plus fastidieuse car nécessite un traitement visuel plus complexe. Il est en effet beaucoup plus aisé de trouver une valeur minimale au sein d'un nuage de point qu'au sein d'un fichier texte simple. L'étape de chargement du modèle au sein du logiciel de visualisation est également plus simple puisque notre application permet d'interpréter une sélection du

modèle concerné dans un graphique directement dans le logiciel de visualisation grâce à un filtre affichant uniquement le modèle concerné. Les étapes similaires, en vert dans la figure 5.13, mettent elles en jeu des temps d'exécution et donc indépendants des conditions de travail dans lesquelles sont exécutées les sous-tâches.

Un second scénario a été étudié, comparant les mêmes conditions que précédemment.



(a)



(b)

Figure 5.14 – *Subdivision par HTA d'une 2nde tâche experte réalisée (a) en conditions normales et (b) au sein de notre plateforme.*

Nous nous sommes intéressés dans cette tâche au processus inverse que celui étudié précédemment. Alors que la démarche précédente était de chercher des informations d'analyses pour guider la navigation, nous sommes ici partis d'une observation faite pendant l'exploration afin ensuite de caractériser analytiquement le phénomène. Plus précisément, la tâche

consiste en la détection d'une anomalie structurelle sur le résidu 182 et la recherche des valeurs de température et d'énergie du modèle dont il provient. Un nouveau découpage en trois principales étapes peut être effectué. Il comprend d'abord la préparation de l'étape de visualisation. Même si cette étape peut être partiellement automatisée au sein de notre plateforme, nous faisons l'hypothèse que les étapes de préparation sont identiques. Les temps d'exécution sont donc identiques (**9 s**). La seconde étape est l'identification visuelle du résidu concerné. Cette étape ne présente pas non plus de différences entre les deux approches dont les temps sont une nouvelle fois égaux (**12 s**). Elle est étroitement liée au logiciel de visualisation et à l'exploration du complexe moléculaire 3d. Ces deux premières étapes, identiques en terme de sous-tâches sont identifiées en vert dans la figure 5.14. La dernière étape marque la plus grande différence de performance entre les deux conditions, les sous-tâches la caractérisant sont colorées en orange dans la figure 5.14. En effet, elle concerne la prise d'informations au sein de l'espace d'analyses après identification du modèle concerné dans l'espace de visualisation. Une simple sélection du résidu mettra en avant le modèle concerné dans l'espace d'analyse au sein de notre programme et que les informations liées à cet individu seront demandées instantanément à la base de données (**2 s**). Dans des conditions standards, le processus de recherche d'informations passe par le traitement manuel des fichiers concernés et n'est aucunement facilité par une quelconque interaction provenant du logiciel de visualisation (**16 s**).

On observe à travers ces deux scénarios un avantage significatif de l'utilisation de notre programme lorsque la sélection d'un sous ensemble d'espace de travail (visualisation ou analyse) permet le filtrage des données de l'espace complémentaire. Quand ces deux espaces sont séparés, comme c'est le cas dans des conditions de travail normales, les aller-retours entre les informations sont fastidieuses et prennent un temps plus important. Notre évaluation a porté sur des tâches courtes, ponctuelles mais fréquentes et représentatives, elle peut cependant être étendue à des tâches plus complexes ou des séquences de tâches similaires où on s'attendrait à une différence de temps d'exécution encore plus importante.

5.4 Résumé et conclusion

Afin d'optimiser le travail autour de l'étude de structures moléculaires au sein d'un environnement immersif, nous avons développé une plateforme logicielle basée sur une communication interactive entre deux espaces indispensables en biologie structurale : un espace de visualisation et un espace d'analyses. Cette communication entre deux espaces aux caractéristiques différentes est rendue possible grâce à la représentation sémantique de l'ensemble des concepts avec lesquels l'utilisateur interagit lors d'une session de travail. Ces concepts peuvent être aussi bien scientifiques, et concerner les données manipulées et observées, que logiciels, décrivant les composants et les actions de la plateforme choisie.

L'ontologie créée et le formalisme RDF/RDFS permettent une homogénéisation des données échangées entre l'ensemble des modules de calculs (rendu visuel, analyses, simulation) puisque chaque module peut accéder aux données qu'il utilise au sein de la base de données construite autour de cette ontologie. On retrouve cette notion d'homogénéité en dehors de l'utilisation même de l'application puisque la représentation ontologique des concepts permet également d'assurer une cohérence et un modèle commun de données à respecter par les utilisateurs lors de la mise en place de nouvelles bases de données de complexes moléculaires. Ainsi, il est possible de mettre en commun certaines informations et d'enrichir les connaissances grâce à des requêtes croisées SPARQL par exemple. La communication inter-modules est la clé de voûte de notre application et l'approche passant par représentation sémantique

des données favorise grandement sa mise en place. Cette approche n'a pas pour seul effet de garantir la cohérence des données échangées, elle permet aussi l'anticipation de certaines actions de l'utilisateur et donc de mettre en place des processus d'interaction directe requis dans les environnements immersifs dans lesquels nous travaillons.

Notre plateforme, malgré son architecture multi-composant impliquant un nombre important de communications, respecte les contraintes temporelles inhérentes à toute application interactive.

Les bases de données sémantiques sont un excellent moyen de stocker, de mettre à disposition et de recouper à tout instant les informations obtenues par l'utilisateur lors de ses sessions immersives de travail. Chaque nouvelle analyse met en jeu des résultats qui seront automatiquement stockés dans la base de données et réutilisables par la suite. Ceci assure une continuité dans le travail et permet de reprendre toute session à un point bien particulier. Les possibilités de partager ainsi une session de travail entre collaborateurs ou de reprendre une session de travail à l'état où elle se trouvait lors de sa dernière utilisation sont des caractéristiques que la majorité des scientifiques apprécient lors de l'utilisation de leurs outils. L'étude de l'évolution de structures moléculaires est un processus long et, comme évoqué précédemment, structuré en boucles successives, il est donc important d'être capable d'avoir un suivi stricte du travail effectué. La communication est également un enjeu majeur dans les sciences où le partage des informations entre collaborateurs et scientifiques du même milieu est primordial puisqu'elle constitue une partie des processus d'évaluation. Il est possible d'aller plus loin dans ce partage d'informations au sein de notre application en développant l'aspect collaboratif. L'une des forces de notre application repose également sur son développement autour d'une base de données construite autour d'une ontologie indépendante. Les modules au coeur de l'application n'ont aucune connaissance des logiciels ou librairies utilisés pour l'affichage 2d et 3d. Cette dissociation permet le couplage d'un nombre important de logiciels de visualisation 3d et 2d. Ce couplage passerait par un simple portage des modules de conversion des données en commandes spécifiques pour les logiciels de visualisation choisis.

Conclusion générale et perspectives

En réponse à la complexité grandissante des modèles 3d de structures moléculaires et à l'augmentation de la quantité de données générées, qui induisent de nouveaux problèmes non résolus de stockage et de transfert des données, nous proposons une nouvelle approche qui regroupe les processus de simulation, de visualisation scientifique et d'analyse des résultats. Dans cette approche, **une formalisation sémantique du domaine, du contenu scientifique et de l'interaction**, supporte **l'intégration de l'activité de visualisation et l'activité d'analyse des résultats dans un contexte interactif commun**, en répondant aux contraintes de performance relatives au temps interactif. Par ailleurs, persuadés que la Réalité Virtuelle continuera à modifier peu à peu les usages en biologie moléculaire, nous avons aussi adressé des problématiques liées à l'exploration des structures moléculaires complexes. Nous avons en particulier développé **des paradigmes de navigation basés sur le contenu moléculaire et la tâche des experts scientifiques**, génériques et donc compatibles avec tous les environnements de travail, qu'ils soient immersifs ou plus classiques. Pour évaluer l'apport de ces approches par rapport aux outils communément utilisés en biologie structurale, nous avons proposé une **méthodologie théorique d'évaluation des performances basée sur l'analyse ergonomique des tâches** en collaboration avec des ergonomes et les experts du domaine. Nous avons pour cette évaluation, modélisé ces tâches métier complexes par une décomposition hiérarchique de ses activités, les activités atomiques étant celles dont il est possible d'estimer le temps de complétion, avec les experts, ou à travers des évaluations expérimentales des paradigmes d'interaction déjà disponibles dans la littérature.

Contributions de la thèse

Nous avons mis en évidence, dans la première partie de ce manuscrit, **les usages, les enjeux et les perspectives en biologie structurale**, le domaine d'application de nos travaux de recherche. Le domaine de la biologie moléculaire, qui a pour principal objectif d'étudier les structures des complexes biomoléculaires afin d'en déduire leurs fonctions, a su intégrer les résultats de recherche du domaine informatique, pour prendre en compte l'évolution des méthodes expérimentales produisant des résultats de plus en plus hétérogènes et de plus en plus complexes. L'informatique a tout d'abord fait partie intégrante du traitement des données générées par les méthodes expérimentales aboutissant à la production de connaissances qui constituent le fondement d'approches plus théoriques de modélisation et simulation moléculaire. Le domaine de modélisation moléculaire n'a reçu que très récemment ses plus belles lettres de noblesse en 2013, ses pionniers, M. Karplus, M. Levitt et A. Warshel, ayant été honorés par le prix Nobel de chimie. Parmi les domaines dédiés à la compréhension de la structuration des biomolécules du vivant, la représentation moléculaire, d'abord physique puis informatique, fut un pilier de la production et de la transmission des connaissances,

et ce depuis l'émergence de la biologie structurale. Le domaine de la biologie moléculaire est aujourd'hui confronté à la complexité et à l'hétérogénéité croissante des données expérimentales et à une explosion du volume des résultats de simulation. La taille des complexes moléculaires et la variété des résultats expérimentaux et théoriques obligent à repenser les méthodes d'exploration usuelles de visualisation et de manipulation des représentations de biomolécules.

Par ailleurs, ces 10 dernières années ont été le témoin de l'essor de la Réalité Virtuelle qui n'est plus seulement cantonnée aux laboratoires de recherche de son domaine, en se démocratisant avec l'apparition de nombreux dispositifs immersifs et d'interaction mobiles et bon marché. Nous avons mis l'accent sur le fait que la majorité des travaux concernant la navigation dans les environnements virtuels concernent des scènes réalistes dont les repères spatiaux et les indices visuels écologiques sont nombreux, produisant donc des paradigmes de navigation qui ne sont que peu applicables aux données scientifiques abstraites. L'activité de visualisation et d'exploration de données scientifiques dans un contexte immersif passe par l'adaptation des **paradigmes de navigation au contenu moléculaire manipulé et aux tâches des experts, indépendants du contexte interactif**. Nous avons donc présenté dans une seconde partie un survol du domaine de **la Réalité Virtuelle, de ses concepts, et des supports technologiques** permettant notamment l'immersion du sujet au coeur des scènes virtuelles, l'exploration des contenus 3d, et l'interaction avec ces contenus dans un contexte immersif.

Forts du constat de l'absence de paradigmes pour l'exploration de structures moléculaires et plus généralement de données scientifiques abstraites dans des dispositifs immersifs, nous avons donc proposé dans une troisième partie plusieurs paradigmes de navigation basés sur le contenu moléculaire et les tâches que les experts scientifiques sont amenées à effectuer. Ces paradigmes répondent aux différents niveaux de granularité des complexes moléculaires, en outillant la tâche d'exploration à la fois globale, locale et détaillée des phénomènes étudiés. Notre réflexion a été particulièrement attentive à la problématique du mal du simulateur, induite par l'exploration de représentations visuelles abstraites et non réalistes que constituent les résultats de biologie moléculaire, qui ne comportent pas d'indices visuels écologiques. Nous avons basé la conception de ces paradigmes sur la prise en compte des particularités géométriques observées dans les complexes moléculaires, notamment de leur symétrie, pour améliorer la conscience spatiale de l'utilisateur, la désorientation étant un facteur du *cybersickness*, et plus généralement la performance lors des tâches supposant une activité de navigation. Grâce à l'utilisation d'axes et centres de symétrie des complexes moléculaires, nos paradigmes de navigation guident l'utilisateur autour de chemins de navigation préférentiels adaptés au contenu exploré. Nous avons considéré des chemins de navigation externes et internes aux complexes et mis en place des solutions de génération automatique de ces chemins préférentiels dans le cadre de tâches expertes spécifiques comme l'accès à des régions d'intérêt enfouies dans le complexe ou la comparaison de phénomènes biologiques répétés sur les différentes sous-unités constituant le complexe moléculaire. Nos paradigmes, centrés sur le contenu moléculaire et la spécificité symétrique du complexe observé ainsi que sur les tâches expertes en biologie structurale, sont par ailleurs indépendants du contexte d'interaction. Ce travail répond à la fois à une lacune de paradigmes de navigation adaptés au contenu moléculaire dans les outils communément utilisés, mais plus généralement à une lacune en terme de paradigme de navigation dédié à l'exploration de représentations abstraites de phénomènes scientifiques.

La quatrième partie du manuscrit rappelle la nécessité de resituer l'utilisateur au coeur de la boucle de de visualisation et d'analyse grâce à des concepts de ***Visual Analytics***

appliqué à la biologie moléculaire, en lui permettant de manipuler conjointement la représentation 2D et 3D de résultats et d'analyse relatifs aux structures moléculaires, pour réduire la quantité de données échangées entre les deux espaces de travail, aujourd'hui dissociés. Nous sommes revenus sur la nécessité de faire évoluer l'organisation du travail donnant lieu à cette dissociation des étapes de visualisation et d'analyse, pourtant tous deux étroitement liées, en **proposant une organisation qui rapproche ces deux activités dans un contexte interactif commun et homogène**.

La concrétisation conceptuelle et logicielle de ce rapprochement des espaces de travail de visualisation et d'analyse a été inspirée par des résultats de travaux de *Visual Analytics*. Ce domaine se base sur la mise en place d'une interactivité forte entre plusieurs représentations de données, éventuellement hétérogènes. Notre réflexion autour des techniques de *Visual Analytics* nous ont amené à considérer la sémantisation des contenus et de l'interaction comme un prérequis pour parvenir à relier interactivement les concepts identiques habituellement manipulés de manière séparée dans les espaces de visualisation et d'analyses, afin de créer un contexte interaction commun. Au-delà d'une plus grande disponibilité et simplicité d'accès à ces données représentées dans un formalisme sémantique homogène, il est aussi possible de raisonner sur celles-ci pour de mettre en place des liens interactifs dynamiques entre des objets moléculaires présentés selon différentes modalités, de la représentation des analyses à la visualisation 3d.

La cinquième et dernière partie de notre manuscrit présente **la conception, l'architecture et l'implémentation d'un prototype d'application de *Visual Analytics* moléculaire adapté aux environnements virtuels** mixant visualisation et analyse au sein d'un même espace de travail. Après avoir introduit les différents aspects de notre plateforme, nous l'avons évaluée au moyen de scénarii définis avec les experts en suivant la même approche méthodologie que nous avons proposée pour évaluer les paradigmes d'évaluation.

Perspectives

Vers des paradigmes de navigation plus avancée pour la biologie moléculaire

Les tâches expertes considérées comme prioritaires par les experts et qui ont été utilisées comme base pour la conception de nos paradigmes de navigation ne constituent pas une liste exhaustive des tâches pouvant être effectuées pendant une session d'exploration moléculaire. Parmi les tâches supplémentaires qui pourraient profiter de chemins de navigation créés au moyen de bases géométriques, le parcours de surfaces ou de structures moléculaires de grande taille (comme les membranes cellulaires) pourrait constituer une aide précieuse pour la recherche de singularités sur la membrane ou à l'interface des protéines membranaires. Le calcul de chemins de navigation pourrait aussi s'effectuer à partir des surfaces accessibles au solvant de la protéine ou à partir d'isosurfaces ou de lignes de champs électrostatiques. L'axe de développement autour de la réduction du *cybersickness* est passé par l'ajout d'indices visuels réalistes dans la scène moléculaire plus abstraite, par l'intermédiaire de *skybox* orientée, ou associé à une contrainte d'orientation tout au long de l'exploration, mais pourrait aller plus loin. La conception de scènes d'immersion réalistes utilise par défaut des repères spatiaux universels qui constituent des moyens de s'orienter naturellement. Il s'agit dans les scènes scientifiques non réalistes d'introduire des repères visuels pour faciliter l'orientation de l'utilisateur. Il serait aussi possible d'envisager une sonification spatialisée d'objet 3d dans la scène moléculaire qui servirait d'indices d'orientation si les indices visuels sont insuffisants, et pourrait servir de vecteurs d'informations pour des tâches de simulation moléculaires par

exemple [135]. Enfin si nous avons eu des retours informels positifs de la part des utilisateurs concernant nos paradigmes navigation, et que nous avons pu évaluer de manière théorique l'apport de ces paradigmes en terme de performance et d'adaptation à des tâches métier, nous n'avons cependant pas été en mesure d'évaluer de manière rigoureuse si l'utilisation de ces paradigmes permettait de diminuer le *cybersickness*. Au-delà de la difficulté d'évaluer le *cybersickness* de manière plus objective que l'utilisation de questionnaires standards, qui ne permettent qu'une évaluation globale de l'expérience utilisateur, une évaluation sérieuse nécessitera un nombre de sujets important, imposant de plus que les sujets soient des experts en biologie structurale.

Supervision multimodale

La mise en place d'une ontologie pour définir l'ensemble des concepts mis en jeu au sein de notre approche, nous a permis de construire un moteur d'interprétation de mot-clé métier qui, couplé à une reconnaissance vocale, permet de convertir ces mots-clés, en une commande exécutable par le logiciel de visualisation moléculaire utilisé. L'utilisation de la sémantique par le formalisme RDF/RDFS/OWL a permis de produire une solution de commandes vocales opérationnelles, du fait de sa simplicité pour l'expert (suite de mot-clés métier dans la terminologie du domaine, insensible à l'ordre). Les répercussions de ce travail concernent la supervision de la multimodalité en général. En effet, notre approche intègre la connaissance de la biologie moléculaire des tâches métiers relatives à la manipulation des complexes, en formalisant ces connaissances par une représentation homogène ainsi que tous les événements d'interactions génériques indépendants de l'application et du domaine ciblé. Les observateurs de contexte interactif dans des architectures de supervision de la multimodalité en entrée [110], produiraient des événements interactifs formalisés dans le même formalisme que celui utilisé pour modéliser la connaissance. Ensuite, le haut niveau de performance du langage SPARQL permet d'envisager d'effectuer des requêtes à la fois sur les contenus manipulés et sur la nature des interactions, permettant de construire et de déclencher des commandes multimodales appropriées au contenu manipulé (atome, résidu, structure...) et à la nature de l'interaction (focus, pointage, navigation, sélection, commande vocale...), dans la lignée de travaux de M. E. Latoschik [181] ou de [70] mais appliqué à la biologie moléculaire.

Le formalisme RDFS/RDF/OWL et le langage SPARQL permettent d'énoncer des règles d'inférences essentielles à la construction de ces commandes multimodales, pour répondre en particulier aux problématiques de prise de décision pour la multimodalité dans un contexte collaboratif [110]. Dans un tel contexte, deux utilisateurs peuvent chacun émettre une commande multimodale de manière conjointe, qu'il peut être difficile à interpréter sans règle, si les commandes sont incohérentes, ou si elles provoquent une incohérence de manipulation de contenu partagé. Il s'agira donc d'intégrer des règles, dans un futur superviseur de la multimodalité en entrée, basé sur ce formalisme, prenant en compte le fait que pour certains auteurs [110], un utilisateur dans un environnement collaboratif durant l'interaction multimodale peut être considéré comme une modalité.

Automatisation de la génération de graphiques adaptés à la nature des données d'analyse à manipuler et à représenter

Les modalités de représentations des résultats d'analyses sont actuellement choisies *a priori* par l'utilisateur. Les nuages de points et les courbes sont la modalité de représentation choisie par défaut pour visualiser les valeurs numériques caractérisant chaque objet moléculaire d'intérêt (les atomes, les résidus, les domaines de la protéine, la protéine dans

son intégralité). Les représentations analytiques peuvent cependant prendre d'autres formes, variantes suivant la nature des données à représenter et l'information à mettre en valeur. La fréquence d'apparition, au sein de l'ensemble des modèles d'une simulation, de l'association de valeurs de deux propriétés distinctes (les angles des domaines protéiques et les profondeurs d'insertion d'une protéine membranaire) est habituellement représentée grâce à un histogramme à deux dimensions, ou la fréquence d'apparition de chaque couple de valeurs dans la trajectoire est représentée par une couleur. Si nous avons défini dans l'ontologie tous les concepts liés à la visualisation moléculaire, nous n'avons pas modélisé tous les concepts liés à la visualisation d'information. Cette modélisation pourrait permettre, à partir du choix des données de l'utilisateur par sélection interactives, de proposer des représentations préférentielles de ses données adaptées à leur nature et aux usages des experts du domaine.

Annotations des objets 3d d'intérêt

L'accès rapide aux informations contenues dans la base de données nous permet d'ajouter une couche informative aux représentations 3d de l'espace de visualisation. En effet, il est facile d'accéder et de présenter sous forme d'annotations les connaissances concernant un objet 3d sélectionné par l'utilisateur stocké dans la base de faits. Ces informations, pouvant par exemple se présenter sous forme de fenêtres dynamiques, permettraient à l'utilisateur d'enrichir son expérience d'exploration en possédant des clés de compréhension complémentaires aux seules informations structurelles.

Bilan global

Nos approches ouvrent la porte à une nouvelle génération d'applications scientifiques, notre démarche ayant été plus spécifiquement consacrée au domaine de biologie structurale. Nos contributions intègrent les dernières avancées issues des domaines de la *Réalité Virtuelle*, du *Visual Analytics*, et de la *Représentation Sémantique des connaissances*. La Réalité Virtuelle n'est plus seulement cantonnée aux laboratoires de recherche de son domaine, se démocratise avec l'apparition de nombreux dispositifs immersifs et d'interaction mobiles et bon marché. Nous avons dans ce travail de thèse proposé des approches intégrant toutes les étapes de biologie structurale au sein d'un contexte interactif commun, et favorisant les interactions directes, deux prérequis d'une organisation du travail dans un contexte immersif. Ce travail pose les premières briques d'une plateforme dédiée à la biologie moléculaire où les espaces de visualisation de structures 3d et les espaces de représentation des résultats analytiques sont reliés par une interactivité bidirectionnelle supportée par une modélisation sémantique des contenus et de l'interaction, censés pour l'expert faciliter la mise en relation de ses résultats hétérogènes. Essentiellement motivé par une plus forte intégration de l'immersion dans les usages en biologie structurale, nos contributions n'en sont pas moins applicables aux stations de travail plus classiques proposant des paradigmes de navigation adaptés aux contenus et la tâche, indépendants du contexte et des dispositifs d'interaction, et de l'application métier, en raccourcissant la boucle d'étude d'un complexe moléculaire du fait du rapprochement des étapes de visualisation et d'analyse.

ANNEXES

Annexe A

Exemples de rendus 3d générés à partir de scènes 2d

Plusieurs exemples sont mis à disposition des utilisateurs pour juger des capacités de rendu de notre application mobile (voir section 3.2 pour les détails de l'application) :

A. Visualisation d'une poche de liaison à la surface de GLIC (Ligand-Gated Ion Channel), une protéine transmembranaire responsable du passage des ions et des molécules d'eau à travers la membrane cellulaire. Scène extraite de VMD.

B. Carte de basse résolution obtenue à partir de cryo-EM pour un filament d'actine. La position de la structure atomique dans l'enveloppe de basse résolution est en premier-plan alors que la carte électromagnétique est présentée en arrière-plan. Visualisation provenant de Chimera

C. Couple verticale du modèle de la capside du virus Influenza A exposant la membrane externe et l'intérieur du virus. GraphiteLifeExplorer a été utilisé pour visualiser la coupure.

D. Interaction entre une protéine SNARE et une membrane biologique lors d'une dynamique moléculaire (arrière-plan). Un graphe 3d rapportant le taux de courbure de la membrane est présenté en premier-plan. La surface 3d a été générée par Paraview (programme de visualisation scientifique)

E. Scènes extraites d'un programme d'animation, MolecularMaya, dédié à la création de films scientifiques pour une audience large. Les scènes extraites mettent en avant deux virus au milieu de villosités à gauche et de multiples virus à droite.

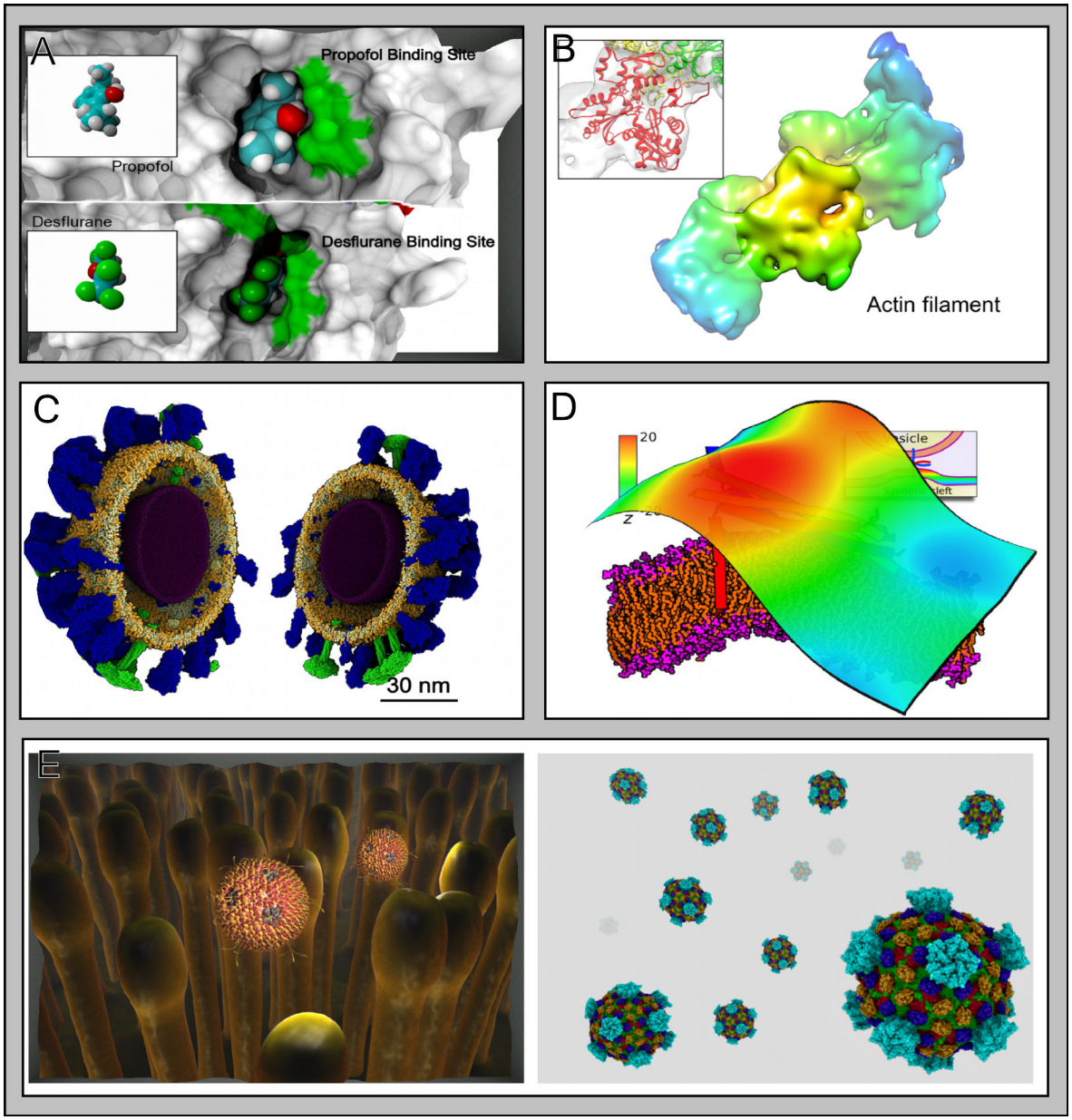

A
Propofol Binding Site
Propofol
Desflurane
Desflurane Binding Site
B
Actin filament
C
30 nm
D
20
E

FIGURE A.1

Annexe B

Captures d'écran de l'application mobile

Captures d'écran provenant de l'application mobile développée et mettant en avant le menu principal de l'application permettant de charger une image ou un objet 3d (en haut) et une scène 3d type représentant la surface d'une protéine avec les différentes options de paramétrisation (en bas). L'application est décrite dans la section 3.2.

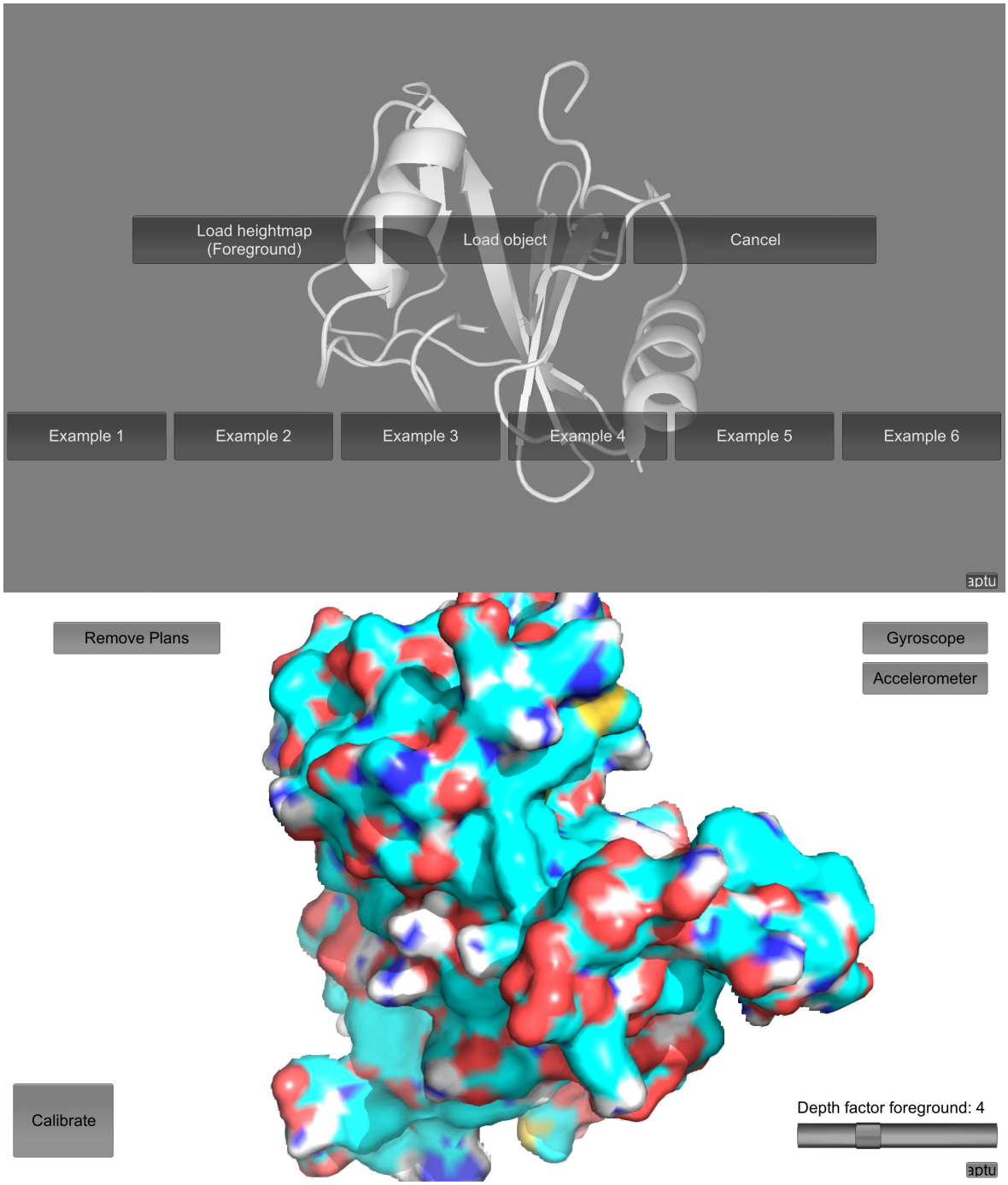

Load heightmap (Foreground)
Load object
Cancel
Example 1
Example 2
Example 3
Example 4
Example 5
Example 6
aptu
Remove Plans
Gyroscope
Accelerometer
Calibrate
Depth factor foreground: 4
aptu

FIGURE B.1

Annexe C

Publications associées

Articles de conférence

— Matthieu Dreher, Jessica Prevoteau-Jonquet, **Mikael Trellet**, Marc Piuzzi, Marc Baaden, Bruno Raffin, Nicolas Ferey, Sophie Robert et Sébastien Limet. ExaViz: a flexible framework to analyse, steer and interact with molecular dynamics simulations. *Faraday Discussion*, 169:119-142, 2014.

— **Mikael Trellet**, Nicolas Férey, Marc Baaden et Patrick Bourdot. Content-guided Navigation in Multimeric Molecular Complexes. *BIOIMAGING*, 76–81, 2014.

— **Mikael Trellet**, Nicolas Férey, Marc Baaden et Patrick Bourdot. Content and task based navigation for structural biology in 3d environments. *VARMS@ IEEEVR, 2015 IEEE 1st International Workshop on*, 31-36, 2015.

Table des figures

Notations et expressions

Acronyme ou notation	Signification / Traduction
ADN	Acide Desoxyribo-Nucléique
ARN	Acide Ribo-Nucléique
$CAPRI$	Critical Assessment of PRedictions of Interfaces
$CAVE$	Cave Automatic Virtual Environment
CPU	Computing Processor Unit
$Cryo-EM$	Cryo-Electron Microscopy
EV	Environnement Virtuel
GC	Graphe Conceptuel
$GLIC$	Ligand-Gated Ion Channel
GPU	Graphical Processor Unit
HMD	Head-Mounted Display
HTA	Hierarchical Task Analysis
MD	Molecular Dynamic
OWL	Web Ontology Language
PDB	Protein Data Bank
RA	Réalité Augmentée
$RDF(S)$	Ressource Description Framework (Schema)
RMN	Résonance Magnétique Nucléaire
$RMSD$	Root Mean Square Deviation
RV	Réalité Virtuelle
SAS	Surface Accessible au Solvant
$SAXS$	Small-Angle X-ray Scattering
$SPARQL$	SPARQL Protocol and RDF Query Language
URI	Uniform Ressource Identifier
WFS	Wave Field Synthesis

Bibliographie

[1] Ben Adida, Mark Birbeck, Shane McCarron, and Steven Pemberton. Rdfa in xhtml: Syntax and processing. *Recommendation, W3C*, 2008.

[2] Bruce Alberts, Dennis Bray, Karen Hopkin, et al. *Essential cell biology*. Garland Science, 2013.

[3] Raluca M Andrei, Marco Callieri, Maria F Zini, et al. Intuitive representation of surface properties of biomolecules using bioblender. *BMC bioinformatics*, 13(Suppl 4):S16, 2012.

[4] John Annett. Hierarchical task analysis. *Handbook of cognitive task design*, pages 17–35, 2003.

[5] Richard Arias-Hernández, John Dill, Brian Fisher, and Tera Marie Green. Visual Analytics and Human-computer Interaction. *interactions*, 18(1):51–55, January 2011.

[6] Michael Ashburner, Catherine A. Ball, Judith A. Blake, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[7] Teresa K Attwood, Douglas B Kell, Philip McDermott, et al. Utopia documents: linking scholarly literature with research data. *Bioinformatics*, 26(18):i568–i574, 2010.

[8] Sören Auer, Christian Bizer, Georgi Kobilarov, et al. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[9] M Baaden and R Lavery. There's plenty of room in the middle: multi-scale modelling of biological systems. *Recent Adv. In Structural Bioinformatics, Research Signpost India, AG de Brevern, ed*, 2007.

[10] Franz Baader. *The description logic handbook: theory, implementation, and applications*. Cambridge university press, 2003.

[11] P. G. Baker, C. A. Goble, S. Bechhofer, et al. An ontology for bioinformatics applications. *Bioinformatics*, 15(6):510–520, June 1999.

[12] Dave Beckett and Brian McBride. Rdf/xml syntax specification (revised). *W3C recommendation*, 10, 2004.

[13] David Benyon, Phil Turner, and Susan Turner. *Designing interactive systems: People, activities, contexts, technologies*. Pearson Education, 2005.

[14] Jeremy Mark Berg, John L Tymoczko, and Lubert Stryer. *Biochemistry*. W. H. Freeman and Company, New York, N.Y, 2012.

[15] H. M. Berman, T. N. Bhat, P. E. Bourne, et al. The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol*, 7 Suppl:957–9, 2000.

[16] Tim Berners-Lee, Dan Connolly, Lalana Kagal, Yosi Scharf, and Jim Hendler. N3logic: A logical framework for the world wide web. *Theory and Practice of Logic Programming*, 8(03):249–269, 2008.

[17] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[18] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370, 2003.

[19] Aude Bolopion, Barthélemy Cagneau, Stephane Redon, and Stéphane Régnier. Comparing position and force control for interactive molecular simulators with haptic feedback. *Journal of Molecular Graphics and Modelling*, 29(2):280–289, September 2010.

[20] Max Born and Robert Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389(20):457–484, 1927.

[21] B Bossard. Gestural interaction for virtual scene description. In *Proceedings of Gesture Workshop*, 2005.

[22] Patrick Bourdot and Damien Touraine. Polyvalent display framework to control virtual navigations by 6DoF tracking. In *Virtual Reality, 2002. Proceedings. IEEE*, pages 277–278, 2002.

[23] Bourdot, Patrick. *Reconstruction et interaction 3D : contribution à la réalité virtuelle et augmentée*. HDR, Université Paris-Sud (XI), LIMSI-CNRS, 2002.

[24] D.A Bowman and R.P. McMahan. Virtual Reality: How Much Immersion Is Enough? *Computer*, 40(7):36–43, July 2007.

[25] Doug A Bowman, Ernst Kruijff, Joseph J LaViola Jr, and Ivan Poupyrev. *3D user interfaces: theory and practice*. Addison-Wesley, 2004.

[26] Dan Brickley and Ramanathan V Guha. {RDF vocabulary description language 1.0: RDF schema}. 2004.

[27] Bernard R Brooks, Charles L Brooks, Alexander D MacKerell, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–1614, 2009.

[28] Frederick P Brooks Jr. What's real about virtual reality? *Computer Graphics and Applications, IEEE*, 19(6):16–27, 1999.

[29] G. Bruder, V. Interrante, L. Phillips, and F. Steinicke. Redirecting Walking and Driving for Natural Navigation in Immersive Virtual Environments. *IEEE Transactions on Visualization and Computer Graphics*, 18(4):538–545, April 2012.

[30] Grigore Burdea and Philippe Coiffet. Virtual reality technology. *Presence: Teleoperators and virtual environments*, 12(6):663–664, 2003.

[31] Richard Car and Mark Parrinello. Unified approach for molecular dynamics and density-functional theory. *Physical review letters*, 55(22):2471, 1985.

[32] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[33] Promita Chakraborty and Ronald N Zuckermann. Coarse-grained, foldable, physical model of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 110(33):13368–13373, 2013.

[34] Matthieu Chavent, Bruno Lévy, Michael Krone, et al. GPU-powered tools boost molecular visualization. *Briefings in Bioinformatics*, 12(6):689–701, November 2011.

[35] Michel Chein and Marie-Laure Mugnier. *Graph-based knowledge representation: computational foundations of conceptual graphs*. Springer Science & Business Media, 2008.

[36] Vincent B. Chen, Ian W. Davis, and David C. Richardson. KING (Kinemage, Next Generation): A versatile interactive molecular and scientific visualization program. *Protein Science*, 18(11):2403–2409, November 2009.

[37] Weiya Chen, Anthony Plancoulaine, Nicolas Férey, et al. 6dof navigation in virtual worlds: comparison of joystick-based and head-controlled paradigms. In *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology*, VRST '13, pages 111–114, New York, NY, USA, 2013. ACM.

[38] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, et al. Biopython: freely available Python tools for computational molecular biology and. *Bioinformatics*, 25(11):1422–1423, June 2009.

[39] Allan M Collins and M Ross Quillian. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2):240–247, 1969.

[40] DM Collins, FA Cotton, EE Hazen, EF Meyer, and CN Morimoto. Protein crystal structures: quicker, cheaper approaches. *Science*, 190(4219):1047–1053, 1975.

[41] Dan Connolly et al. Gleaning resource descriptions from dialects of languages (grddl). *W3C, W3C Recommendation*, 11, 2007.

[42] Kristin A. Cook and James J. Thomas. Illuminating the Path: The Research and Development Agenda for Visual Analytics. Technical Report PNNL-SA-45230, Pacific Northwest National Laboratory (PNNL), Richland, WA (US), May 2005.

[43] Olivier Corby, Rose Dieng-Kuntz, Fabien Gandon, and Catherine Faron-Zucker. Searching the semantic web: Approximate query processing based on ontologies. *Intelligent Systems, IEEE*, 21(1):20–27, 2006.

[44] Robert B Corey and Linus Pauling. Molecular models of amino acids, peptides, and proteins. *Review of Scientific Instruments*, 24(8):621–627, 1953.

[45] Nicolas Courty and Eric Marchand. Computer animation: A new application for image-based visual servoing. In *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, volume 1, pages 223–228. IEEE, 2001.

[46] Carolina Cruz-Neira, Daniel J. Sandin, Thomas A. DeFanti, Robert V. Kenyon, and John C. Hart. The CAVE: Audio Visual Experience Automatic Virtual Environment. *Commun. ACM*, 35(6):64–72, June 1992.

[47] Rudolph P Darken and Barry Peterson. Spatial orientation, wayfinding, and representation. *Handbook of virtual environments*, pages 493–518, 2002.

[48] Rudolph P Darken and John L Sibert. Navigating large virtual spaces. *International Journal of Human-Computer Interaction*, 8(1):49–71, 1996.

[49] Simon Davis, Keith Nesbitt, and Eugene Nalivaiko. A systematic review of cybersickness. In *Proceedings of the 2014 Conference on Interactive Entertainment*, pages 1–9. ACM, 2014.

[50] Warren L DeLano. The PyMOL molecular graphics system. 2002.

[51] Yannick Dennemont. *Une assistance à l'interaction 3D en réalité virtuelle par un raisonnement sémantique et une conscience du contexte*. PhD thesis, Université d'Evry-Val d'Essonne, 2013.

[52] Dan Diaper and Colston Sanger. Tasks for and tasks in human–computer interaction. *Interacting with Computers*, 18(1):117–138, 2006.

[53] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[54] C. Dominguez, R. Boelens, and A. M. Bonvin. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125:1731–7, 2003.

[55] Nadezhda T. Doncheva, Yassen Assenov, Francisco S. Domingues, and Mario Albrecht. Topological analysis and interactive visualization of biological networks and protein structures. *Nature Protocols*, 7(4):670–685, April 2012.

[56] Matthieu Dreher, Jessica Prevoteau-Jonquet, Mikael Trellet, et al. Exaviz: a flexible framework to analyse, steer and interact with molecular dynamics simulations. *Faraday discussions*, 169:119–142, 2014.

[57] Lee Feigenbaum, Ivan Herman, Tonya Hongsermeier, Eric Neumann, and Susie Stephens. The Semantic Web in Action. *Scientific American*, 297(6):90–97, 2007.

[58] Nicolas Ferey, Olivier Delalande, and Marc Baaden. Biospring: an interactive and multi-resolution software for exible docking and for mechanical exploration of large biomolecular assemblies. In *JOBIM 2012-Journées Ouvertes en Biologie, Informatique et Mathématiques*, pages 433–434. Inria Rennes-Bretagne Atlantique, 2012.

[59] Bernd Fröhlich, Stephen Barrass, Björn Zehner, John Plate, and Martin Göbel. Exploring geo-scientific data in virtual environments. In *Proceedings of the conference on Visualization'99: celebrating ten years*, pages 169–173. IEEE Computer Society Press, 1999.

[60] Philippe Fuchs and Guillaume Moreau. *Le traité de la réalité virtuelle*, volume 2. Presses des MINES, 2006.

[61] N. Férey, J. Nelson, C. Martin, et al. Multisensory VR interaction for protein-docking in the CoRSAIRe project. *Virtual Reality*, 13(4):273–293, December 2009.

[62] Fabien Gandon. *Graphes RDF et leur Manipulation pour la Gestion de Connaissances.* HDR, INRIA, 2008.

[63] David Genest and Eric Salvat. A platform allowing typed nested graphs: How cogito became cogitant. In *Conceptual Structures: Theory, Tools and Applications*, pages 154–161. Springer, 1998.

[64] Alexandre Gillet, Michel Sanner, Daniel Stoffler, and Arthur Olson. Tangible interfaces for structural molecular biology. *Structure*, 13(3):483–491, 2005.

[65] ESCHER GODEL. Bach: An eternal golden braid. *Douglas R. Hofstadter.(Originally published by Basic Books, 1979.) Basic Books*, 1999.

[66] David S. Goodsell and Arthur J. Olson. Structural symmetry and protein function. *Annual review of biophysics and biomolecular structure*, 29(1):105–153, 2000.

[67] Sebastian Grottel, Guido Reina, Jadran Vrabec, and Thomas Ertl. Visual verification and analysis of cluster detection for molecular dynamics. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1624–1631, 2007.

[68] ANDRE Guimer and Gérard Fournet. Small angle scattering of x-rays. *J. Wiley & Sons, New York*, 1955.

[69] Stefan Gumhold. Maximum entropy light source placement. In *Visualization, 2002. VIS 2002. IEEE*, pages 275–282. IEEE, 2002.

[70] Mario Gutierrez, Frederic Vexo, and Daniel Thalmann. Semantics-based representation of virtual environments. *International journal of computer applications in technology*, 23(2-4):229–238, 2005.

[71] Moritz P Haag, Alain C Vaucher, Maël Bosson, Stéphane Redon, and Markus Reiher. Interactive chemical reactivity exploration. *ChemPhysChem*, 15(15):3301–3319, 2014.

[72] Zaynab Habibi, Guillaume Caron, and El Mustapha Mouaddib. 3d model automatic exploration: Smooth and intelligent virtual camera control. In *Computer Vision-ACCV 2014 Workshops*, pages 612–626. Springer, 2014.

[73] Andrew J. Hanson and Eric A. Wernert. Constrained 3d navigation with 2d controllers. In *Visualization'97., Proceedings*, pages 175–182, 1997.

[74] Li-wei He, Michael F Cohen, and David H Salesin. The virtual cinematographer: a paradigm for automatic real-time camera control and directing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 217–224. ACM, 1996.

[75] Michael Heim. *Virtual realism*. Oxford University Press, 1998.

[76] Angel Herraez. Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34(4):255–261, 2006.

[77] Jonathan D Hirst, David R Glowacki, and Marc Baaden. Molecular simulations and visualization: introduction and overview. *Faraday discussions*, 169:9–22, 2014.

[78] Thai V Hoang, Xavier Cavin, and David W Ritchie. gemfitter: A highly parallel fft-based 3d density fitting tool with gpu texture memory acceleration. *Journal of structural biology*, 184(2):348–354, 2013.

[79] Liisa Holm and Chris Sander. Mapping the protein universe. *Science*, 273(5275):595–602, 1996.

[80] Liisa Holm and Chris Sander. Touring protein fold space with dali/fssp. *Nucleic acids research*, 26(1):316–319, 1998.

[81] T Hubbard, Daniel Barker, Ewan Birney, et al. The ensembl genome database project. *Nucleic acids research*, 30(1):38–41, 2002.

[82] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[83] Georges B Johnson, Peter H Raven, Kenneth A Mason, Jonathan B Losos, and Susan R Singer. *Biologie-Version luxe*. De Boeck Supérieur, 2011.

[84] Susan Jones and Janet M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 1996.

[85] T ALWYN Jones. A graphics model building and refinement system for macromolecules. *Journal of Applied Crystallography*, 11(4):268–272, 1978.

[86] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[87] J. Kehrer and H. Hauser. Visualization and Visual Analysis of Multifaceted Scientific Data: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):495–513, March 2013.

[88] Daniel A Keim, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann. *Mastering the information age-solving problems with visual analytics*. Florian Mansmann, 2010.

[89] John C Kendrew, G Bodo, Howard M Dintzis, et al. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.

[90] W James Kent, Charles W Sugnet, Terrence S Furey, et al. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[91] Andreas Kerren and Falk Schreiber. Toward the Role of Interaction in Visual Analytics. In *Proceedings of the Winter Simulation Conference*, WSC '12, pages 420:1–420:13, Berlin, Germany, 2012. Winter Simulation Conference.

[92] Azam Khan, Ben Komalo, Jos Stam, George Fitzmaurice, and Gordon Kurtenbach. HoverCam: interactive 3d navigation for proximal object inspection. In *Proceedings of the 2005 symposium on Interactive 3D graphics and games*, pages 73–80, 2005.

[93] Firas Khatib, Frank DiMaio, Seth Cooper, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.

[94] Joe Kielman, Jim Thomas, and Richard May. Foundations and frontiers in visual analytics. *Information Visualization*, 8(4):239, 2009.

[95] Graham Klyne and Jeremy J Carroll. Resource description framework (rdf): Concepts and abstract syntax. 2006.

[96] Eugenia M Kolasinski. Simulator sickness in virtual environments. Technical report, DTIC Document, 1995.

[97] Elmar Krieger and Gert Vriend. Yasara view—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics*, 30(20):2981–2982, 2014.

[98] T Kuhlen, R Pajarola, and K Zhou. Parallel in situ coupling of simulation with a fully featured visualization system. 2011.

[99] Pravin Kumar, Alexander Ziegler, Julian Ziegler, Barbara Uchanska-Ziegler, and Andreas Ziegler. Grasping molecular structures through publication-integrated 3d models. *Trends in biochemical sciences*, 33(9):408–412, 2008.

[100] Naoki Kusumoto, Taiji Sohmura, Shinichi Yamada, et al. Application of virtual reality force feedback haptic device for oral implant surgery. *Clinical Oral Implants Research*, 17(6):708–713, December 2006.

[101] Joseph J LaViola Jr. A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin*, 32(1):47–56, 2000.

[102] Fritz Lehmann. *Semantic networks in artificial intelligence*. Elsevier Science Inc., 1992.

[103] Dianfan Li, Coiln Boland, Kilian Walsh, and Martin Caffrey. Use of a robot for high-throughput crystallization of membrane proteins in lipidic mesophases. *J. Vis. Exp*, 67:e4000, 2012.

[104] Wendell A. Lim. Frederic M Richards 1925–2009. *Nature Structural & Molecular Biology*, 16(3):230–232, March 2009.

[105] Zhihan Lv, Alex Tek, Franck Da Silva, et al. Game On, Science - How Video Game Technology May Help Biologists Tackle Visualization Challenges. *PLoS ONE*, 8(3):e57990, March 2013.

[106] Zhihan Lv, Alex Tek, Franck Da Silva, et al. Game on, science-how video game technology may help biologists tackle visualization challenges. *PloS one*, 8(3):57990, 2013.

[107] Kwan-Liu Ma. In situ visualization at extreme scale: Challenges and opportunities. *Computer Graphics and Applications, IEEE*, 29(6):14–19, 2009.

[108] Thomas Madej, Christopher J Lanczycki, Dachuan Zhang, et al. Mmdb and vast+: tracking structural similarities between macromolecular complexes. *Nucleic acids research*, 42(D1):D297–D303, 2014.

[109] P. Martin, P. Bourdot, and D. Touraine. A reconfigurable architecture for multimodal and collaborative interactions in Virtual Environments. In *2011 IEEE Symposium on 3D User Interfaces (3DUI)*, pages 11–14, March 2011.

[110] P Martin, A Tseu, N Férey, D Touraine, and P Bourdot. A hardware and software architecture to deal with multimodal and collaborative interactions in multiuser virtual reality environments. In *IS&T/SPIE Electronic Imaging*, pages 901209–901209. International Society for Optics and Photonics, 2014.

[111] P. Martin, A. Tseu, N. Férey, D. Touraine, and P. Bourdot. A hardware and software architecture to deal with multimodal and collaborative interactions in multiuser virtual reality environments. volume 9012, pages 901209–901209–16, 2014.

[112] Xavier Martinez, Nicolas Ferey, Jean-Marc Vezien, and Patrick Bourdot. Virtual structure reconstruction and energy estimation of a peptide from a physical tangible interface. In *Virtual and Augmented Reality for Molecular Science (VARMS@ IEEEVR), 2015 IEEE 1st International Workshop on*, pages 41–42. IEEE, 2015.

[113] Eric Martz and Eric Francoeur. History of visualization of biological macromolecules. *University of Massachusetts Amherst. Sep*, 2004.

[114] Majid Masso and Iosif I Vaisman. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics*, 24(18):2002–2009, 2008.

[115] Brian McBride. Jena: A semantic web toolkit. *IEEE Internet computing*, (6):55–59, 2002.

[116] Robert T. McGibbon, Kyle A. Beauchamp, Christian R. Schwantes, et al. Mdtraj: a modern, open library for the analysis of molecular dynamics trajectories. *bioRxiv*, 2014.

[117] Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.

[118] Nicholas Metropolis and Stanislaw Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

[119] Naveen Michaud-Agrawal, Elizabeth J. Denning, Thomas B. Woolf, and Oliver Beckstein. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, 32(10):2319–2327, July 2011.

[120] A-E Molza, Nicolas Férey, Mirjam Czjzek, et al. Innovative interactive flexible docking method for multi-scale reconstruction elucidates dystrophin molecular assembly. *Faraday discussions*, 169:45–62, 2014.

[121] Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995.

[122] Hiroshi Nagata, Hiroshi Mizushima, and Hiroshi Tanaka. Concept and prototype of protein–ligand docking simulator with force feedback technology. *Bioinformatics*, 18(1):140–146, 2002.

[123] Natalya F Noy, Michael Sintek, Stefan Decker, et al. Creating semantic web contents with protege-2000. *IEEE intelligent systems*, (2):60–71, 2001.

[124] Seán I. O'Donoghue, David S. Goodsell, Achilleas S. Frangakis, et al. Visualization of macromolecular structures. *Nature Methods*, 7:S42–S55, March 2010.

[125] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred F Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6. *Journal of computational chemistry*, 25(13):1656–1676, 2004.

[126] Michael Ortega, Wolfgang Stuerzlinger, and Doug Scheurich. SHOCam: A 3d Orbiting Algorithm. 2015.

[127] Dave Pape, Josephine Anstey, and Bill Sherman. Commodity-based projection vr. In *ACM SIGGRAPH 2004 Course Notes*, page 19. ACM, 2004.

[128] David A Pearlman, David A Case, James W Caldwell, et al. Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1):1–41, 1995.

[129] Serge Pérez, Thibault Tubiana, Anne Imberty, and Marc Baaden. Three-dimensional representations of complex carbohydrates and polysaccharides—sweetunitymol: A video game-based computer graphic software. *Glycobiology*, 25(5):483–491, 2015.

[130] James C Phillips, Rosemary Braun, Wei Wang, et al. Scalable molecular dynamics with namd. *Journal of computational chemistry*, 26(16):1781–1802, 2005.

[131] Kendall Powell. A lab app for that. *Nature*, 484(7395):553–555, April 2012.

[132] Sander Pronk, Szilárd Páll, Roland Schulz, et al. Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, page btt055, 2013.

[133] Eric Prud'hommeaux, Gavin Carothers, Dave Beckett, and Tim Berners-Lee. Turtle–terse rdf triple language. *Candidate Recommendation, W3C*, 2013.

[134] Eric Prud'Hommeaux, Andy Seaborne, et al. Sparql query language for rdf. *W3C recommendation*, 15, 2008.

[135] Benjamin Rau, Florian Frieß, Michael Krone, Christoph Müller, and Thomas Ertl. Enhancing visualization of molecular simulations using sonification. In *International Workshop on Virtual and Augmented Reality for Molecular Science (VARMS@IEEEVR)*, volume 1, pages 25–30. IEEE, 2015.

[136] Eugene Raush, Max Totrov, Brian D Marsden, and Ruben Abagyan. A new method for publishing three-dimensional content. *PLoS One*, 4(10):e7394, 2009.

[137] James T Reason and Joseph John Brand. *Motion sickness.* Academic press, 1975.

[138] David C Richardson and Jane S Richardson. The kinemage: a tool for scientific communication. *Protein science*, 1(1):3–9, 1992.

[139] Jane S Richardson, David C Richardson, Kenneth A Thomas, Enid W Silverton, and David R Davies. Similarity of three-dimensional structure between the immunoglobulin domain and the copper, zinc superoxide dismutase subunit. *Journal of molecular biology*, 102(2):221–235, 1976.

[140] Peter J Russel. *iGenetics: A Molecular Approach. 3 uppl.* Pearson Education.(828 s). ISBN, 2010.

[141] S.J. Rysavy, D. Bromley, and V. Daggett. DIVE: A Graph-Based Visual-Analytics Framework for Big Data. *IEEE Computer Graphics and Applications*, 34:26–37, March 2014.

[142] Brian Salomon, Maxim Garber, Ming C Lin, and Dinesh Manocha. Interactive navigation in complex environments using path planning. In *Proceedings of the 2003 symposium on Interactive 3D graphics*, pages 41–50. ACM, 2003.

[143] Andrea Salvadori, Andrea Brogni, Giordano Mancini, and Vincenzo Barone. Moka: Designing a Simple Scene Graph Library for Cluster-Based Virtual Reality Systems. In Lucio Tommaso De Paolis and Antonio Mongelli, editors, *Augmented and Virtual Reality*, number 8853 in Lecture Notes in Computer Science, pages 333–350. Springer International Publishing, September 2014.

[144] Karissa Y Sanbonmatsu, Scott C Blanchard, and Paul C Whitford. Molecular dynamics simulations of the ribosome. In *Biophysical approaches to translational control of gene expression*, pages 51–68. Springer, 2013.

[145] Ganesh Sankaranarayanan, Suzanne Weghorst, Michel Sanner, Alexandre Gillet, and Arthur Olson. Role of haptics in teaching structural molecular biology. In *Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2003. HAPTICS 2003. Proceedings. 11th Symposium on*, pages 363–366. IEEE, 2003.

[146] Roger Sayle and Andrew Bissell. Rasmol: A program for fast, realistic rendering of molecular structures with shadows. In *Proceedings of the 10th Eurographics UK*, volume 92, pages 7–9, 1992.

[147] Erwin Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Physical Review*, 28(6):1049, 1926.

[148] Hans-Jörg Schulz, Adelinde M Uhrmacher, and Heidrun Schumann. Visual analytics for stochastic simulation in cell biology. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, page 48. ACM, 2011.

[149] Steffen Schulze-Kremer. Ontologies for molecular biology and bioinformatics. *In silico biology*, 2(3):179–193, 2002.

[150] Nadine Schuurman and Agnieszka Leszczynski. Ontologies for Bioinformatics. *Bioinformatics and Biology Insights*, 2:187–200, March 2008.

[151] Paul Shannon, Andrew Markiel, Owen Ozier, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, November 2003.

[152] David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[153] William R Sherman and Alan B Craig. *Understanding virtual reality: Interface, application, and design*. Elsevier, 2002.

[154] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *, IEEE Symposium on Visual Languages, 1996. Proceedings*, pages 336–343, September 1996.

[155] Brian K Shoichet, Irwin D Kuntz, and Dale L Bodian. Molecular docking using shape descriptors. *Journal of Computational Chemistry*, 13(3):380–397, 1992.

[156] Ian Sillitoe, Tony E Lewis, Alison Cuff, et al. Cath: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research*, 43(D1):D376–D381, 2015.

[157] Michael Sintek and Stefan Decker. Triple—a query, inference, and transformation language for the semantic web. In *The Semantic Web—ISWC 2002*, pages 364–378. Springer, 2002.

[158] Soren Skou, Richard E Gillilan, and Nozomi Ando. Synchrotron-based small-angle x-ray scattering of proteins in solution. *Nature protocols*, 9(7):1727–1739, 2014.

[159] Barry Smith, Michael Ashburner, Cornelius Rosse, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, November 2007.

[160] B. Snel, G. Lehmann, P. Bork, and M. A. Huynen. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Research*, 28(18):3442–3444, 2000.

[161] John F Sowa. Conceptual structures: information processing in mind and machine. 1983.

[162] Holger Stenzhorn, Stefan Schulz, Elena Bei\s swanger, et al. BioTop and ChemTop-Top-Domain Ontologies for Biology and Chemistry. In *International Semantic Web Conference (Posters & Demos)*. Citeseer, 2008.

[163] Jonathan Steuer, Frank Biocca, Mark R Levy, et al. Defining virtual reality: Dimensions determining telepresence. *Communication in the age of virtual reality*, pages 33–56, 1995.

[164] Matthew B Stocks, Steven Hayward, and Stephen D Laycock. Interacting with the biomolecular solvent accessible surface via a haptic feedback device. *BMC structural biology*, 9(1):69, 2009.

[165] J. Stone, A. Kohlmeyer, K. Vandivort, and K. Schulten. Immersive molecular visualization and interactive modeling with commodity hardware. *Advances in Visual Computing*, pages 382–393, 2010.

[166] John E Stone, Justin Gullingsrud, and Klaus Schulten. A system for interactive molecular dynamics simulation. In *Proceedings of the 2001 symposium on Interactive 3D graphics*, pages 191–194. ACM, 2001.

[167] Gang Sun, Ping Jun Xia, Yuan Li, Wang Min Yi, and Lei Huang. A haptic-based approach for cable design and routing in industrial complex products. In *Advanced Materials Research*, volume 139, pages 1356–1360. Trans Tech Publ, 2010.

[168] John A Tainer, Elizabeth D Getzoff, Karl M Beem, Jane S Richardson, and David C Richardson. Determination and analysis of the 2 å structure of copper, zinc superoxide dismutase. *Journal of molecular biology*, 160(2):181–217, 1982.

[169] Sandra Tan and Russell Waugh. Use of Virtual-Reality in Teaching and Learning Molecular Biology. In Yiyu Cai, editor, *3D Immersive and Interactive Learning*, pages 17–43. Springer Singapore, 2013.

[170] Russell M Taylor II, Thomas C Hudson, Adam Seeger, et al. Vrpn: a device-independent, network-transparent vr peripheral system. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 55–61. ACM, 2001.

[171] Mikael Trellet, Nicolas Férey, Marc Baaden, and Patrick Bourdot. Content-guided navigation in multimeric molecular complexes. In *BIOIMAGING*, pages 76–81, 2014.

[172] Dmitry Tsarkov and Ian Horrocks. Fact++ description logic reasoner: System description. In *Automated reasoning*, pages 292–297. Springer, 2006.

[173] Andries Van Dam, Andrew S Forsberg, David H Laidlaw, Joseph J LaViola Jr, and Rosemary M Simpson. Immersive vr for scientific visualization: A progress report. *Computer Graphics and Applications, IEEE*, 20(6):26–52, 2000.

[174] Andries Van Dam, Andrew S. Forsberg, David H. Laidlaw, Joseph J. LaViola Jr, and Rosemary M. Simpson. Immersive VR for scientific visualization: A progress report. *Computer Graphics and Applications, IEEE*, 20(6):26–52, 2000.

[175] Pere-Pau Vázquez, Miquel Feixas, Mateu Sbert, and Wolfgang Heidrich. Viewpoint selection using viewpoint entropy. In *VMV*, volume 1, pages 273–280, 2001.

[176] Jean-Marc Vézien, Bob Ménélas, Julien Nelson, et al. Multisensory vr exploration for computer fluid dynamics in the corsaire project. *Virtual Reality*, 13(4):257–271, 2009.

[177] Norman G. Vinson. Design Guidelines for Landmarks to Support Navigation in Virtual Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 278–285, New York, NY, USA, 1999. ACM.

[178] Arieh Warshel and Michael Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of molecular biology*, 103(2):227–249, 1976.

[179] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[180] Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. Dublin core metadata for resource discovery. Technical report, 1998.

[181] Dennis Wiebusch and Marc Erich Latoschik. Decoupling the entity-component-system pattern using semantic traits for reusable realtime interactive systems. In *IEEE VR Workshop on Software Engineering and Architectures for Realtime Interactive Systems*, IEEE VR, 2015.

[182] Kurt Wuthrich. *NMR of proteins and nucleic acids*. Wiley, 1986.

[183] Sumi Yoshikawa, Kenji Satou, and Akihiko Konagaya. Drug interaction ontology (DIO) for inferences of possible drug-drug interactions. *Studies in Health Technology and Informatics*, 107(Pt 1):454–458, 2004.

[184] Z. Hong Zhou. Atomic resolution cryo electron microscopy of macromolecular complexes. *Advances in Protein Chemistry and Structural Biology*, 82:1–35, 2011.

[185] Matthew D Zimmerman, Marek Grabowski, Marcin J Domagalski, et al. Data management in the modern structural biology and biomedical research environment. In *Structural Genomics and Drug Discovery*, pages 1–25. Springer, 2014.