Conception et développement d'une méthode de comparaison de surfaces appliquée aux protéines

Léa Sirugue

le cnam

ÉCOLE DOCTORALE Sciences des Métiers de l'Ingénieur
Génomique Bioinformatique et Chimie Moléculaire

THÈSE DE DOCTORAT

présentée par : **Léa SIRUGUE**
soutenue le : **14 décembre 2020**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline : **Informatique**

Spécialité : **Bioinformatique**

Conception et développement d'une méthode de comparaison de surfaces appliquée aux protéines

THÈSE dirigée par

Prof. Montes Matthieu	*Professeur des Universités, CNAM*

RAPPORTEURS

Dr. Guyon Frédéric	*Ingénieur de recherche, Université de Paris*
Prof. Lavoué Guillaume	*Professeur des Universités, École Nationale d'Ingénieurs de Saint-Etienne*

EXAMINATEURS

Dr. Andréani Jessica	*Ingénieure-chercheure, CEA*
Prof. Carbone Alessandra	*Professeur des Universités, Sorbonne Université*

Le plus intelligent de tous, à mon avis,
c'est celui qui au moins une fois par mois
se traite lui-même d'imbécile.

Fiodor Dostoïevski

Remerciements

Je tiens tout d'abord à grandement remercier le professeur Matthieu Montes d'avoir dirigé mes travaux de thèses avec compréhension, bienveillance et intelligence. Je le remercie pour la confiance qu'il m'a accordée, sa disponibilité pour m'aider dans tous les moments de ma vie et l'environnent qu'il a créé au sein de son équipe qui a permis de pleinement m'épanouir.
Je souhaite exprimer ma pleine reconnaissance au professeur Guillaume Lavoué et au docteur Frédéric Guyon d'avoir accepté de juger mon travail de thèse en tant que rapporteurs.
Je remercie énormément le professeur Alessandra Carbone et le docteur Jessica Andréani de m'avoir fait l'honneur de participer au jury de ma thèse.

Je remercie les équipes pédagogiques de l'UPMC. La licence PIMA de m'avoir permis de développer mon esprit de réflexion et critique sur des sujets complexes. Je souhaite également remercier l'équipe pédagogique du master BioInformatique et Modélisation pour m'avoir fait découvrir l'univers de la recherche et redécouvrir le domaine de la biologie sous un angle nouveau. Je remercie Nicolas Thome de m'avoir introduit avec simplicité à l'utilisation de méthode de Deep Learning.

Je souhaite remercier toutes les personnes du laboratoire GBCM et associées qui ont été là pour me soutenir, me faire réfléchir et échanger. Merci à tous pour votre bienveillance. Je tiens a remercier particulièrement : Benjamin pour avoir été là pour m'apprendre les ficelles de la recherche et partagé ma chambre d'hôtel à chaque déplacement ; Sigrid pour être toujours à l'écoute, ouverte et honnête ; Nathalie pour apporter de la bonne humeur à tous les moments de la journée et ta générosité ; Asma pour ta gentillesse et nos discussions dans le métro en rentrant du travail ; Taoufik pour tous nos échanges sur Whatsapp et pour toutes tes blagues qui égayent mes journées ; Florent pour tous tes conseils pertinents et d'être toujours là quand on a besoin d'aide ; Chloé pour ton énergie débordante

et de me laisser t'expliquer des concepts de sub-culture d'internet ; Max pour tous nos échanges sur le cinéma et sur les cultures SJW et alt-right ; Maxime et Simon pour toutes nos discussions autour de bonnes bières ; Machat pour ton intérêt sur tous les sujets que j'aborde ; Myriam pour tout ton soutien et nos soirées surprenantes ; Manon pour avoir partagé nos années de thèses et m'avoir soutenue durant celles-ci ; Josselin avec accent en anglais trop parfait pour que tu sois français ; Cedric vérificateur officiel des chiffres après chaque débat ; Marc pour ne jamais abandonner d'essayer de me surprendre ; Sophia et Christiane pour votre gentillesse et votre professionnalisme. Je souhaite également remercier toute l'équipe de chimie moléculaire et l'équipe de Cochin pour tous nos échanges et débats d'Ermenonville à Chissay et surtout au Léonard.

Je souhaite remercier tous mes amis qui n'ont toujours pas compris mon sujet de thèse. Merci à : Jules, Jérémie, Jonas, Clément et Marie-Sophie pour la grande influence positive que vous avez dans ma vie ; Mélanie, Céline, Théo, Etienne et Charles qui connaissent trop bien le lavoir de Fixin ; Louise avec qui j'ai partagé tellement de belles discussions autour d'un thé ; Ninon, Nally, Charline et Camille pour toutes nos soirées provinciales ; Romain, Gaëtan, Fazou, Laurent et David pour nos longues soirées parisiennes ; Léo qui a su calmer les élèves du collège Montgolfier pour que je travaille dans le calme ; Tom pour toutes nos discussions en voiture à sillonner la campagne ; Alice et Héloïse pour tout nos moments chills ; Carolane qui est ma nouvelle super amie post confinement ; Louis et Benjamin qui sont nés le bon jour ; Nina et Charlotte pour les bons restaurants que l'on a partagés ; Eunjin ma *eonni* préférée ; Simon et Julia pour votre énergie et votre gentillesse débordante ; Chloé, Fanny, Margaux, Elsa, Hugo, Emma, Alexandre, Sylvain et Hélène pour toute votre bienveillance ; Johanna, je te souhaite une thèse brillante ; Mattias, Eli, Loïc et Pierre-Antoine les gardiens du terroir ; Catarina pour toutes nos discussions passionnées et oubliés ; Karolina pour ta joie de vivre et ton grain de folie ; Linn pour toute l'attention que tu offres aux gens ; Sungjae pour ton indépendance et originalité. Enfin mon plus grand merci à Sinae qui est toujours pleine d'énergie pour me motiver et égayer tout les moments de ma vie, merci pour ton soutien inconditionnel et tous tes gestes attentionnés qui ne cessent de me surprendre.

Pour terminer, je remercie ma famille. Merci à mes parents d'avoir toujours cru en moi et de m'avoir aidée à partir étudier à Paris. Merci à mon père de m'avoir très tôt appris l'esprit de rationalité qui

fait la scientifique que je suis maintenant et merci à ma mère qui j'espère serait fière de moi et qui m'a transmis sa gentillesse et son altruisme. Merci à mes frères Guillaume et Didier qui ont été là pour croire en moi et me faire sourire. Merci à Nadine qui m'accueille avec bonne humeur et énergie quand je rentre à Dijon.

Résumé

Les interactions entre protéines jouent un rôle crucial dans les processus du vivant comme la communication cellulaire, l'immunité, la croissance, prolifération et la mort cellulaires. Ces interactions se font via leur surface et la perturbation des interactions entre protéines est à la base de nombreux processus pathologiques. Il est donc nécessaire de bien comprendre et caractériser la surface des protéines et leurs interactions mutuelles de manière à mieux comprendre les processus du vivant. Différentes méthodes de comparaison de la surface des protéines ont été développées ces dernières années mais aucune n'est assez puissante pour traiter l'ensemble des structures disponibles dans les différentes bases de données. Le projet de thèse est donc de développer des méthodes rapides de comparaison de surface et de les appliquer à la surface des macromolécules.

Mots-clés : Biologie structurale, Reconnaissance de formes, Docking, Traitement du signal, Géométrie spectrale

Abstract

Protein interactions play a crucial role in the living processes such as cell communication, immunity, cell growth, proliferation and death. These interactions occur through the surface of proteins and the disruption of their interactions is the start of many disease processes. It is therefore necessary to understand and characterize the surface of proteins and their interactions to better understand living processes. Different methods of protein surfaces comparison have been developed in the recent years but none are powerful enough to handle all the structures currently available in databases. The PhD project is to develop rapid methods of surface comparison and apply them to the surface of macromolecules.

Keywords : Structural biology, Shape recognition, Molecular docking, Signal processing, Spectral geometry

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

Contenu

1.1 La reconnaissance de formes

L'intérêt pour la reconnaissance, la classification et la comparaison de formes d'objets 3D est grandissant dû à ses nombreuses applications dans de nombreux domaines comme la médecine [10] [11], l'automatisation des voitures [12] [13] [14], la robotique [15] [16] ou la surveillance [17] [18].

Cet intérêt grandissant peut être expliqué par deux raisons majeures. La première étant qu'il y a de plus en plus de données disponibles permettant de développer et tester des méthodes de plus en plus précises. La base de données grand public 3D warehouse permet de rechercher et télécharger gratuitement des millions de modèles 3D conçus par ordinateur. Cette base de donnée est participative, tout le monde pouvant déposer leurs modèles 3D, ce qui est un signe d'une envie d'ouverture au grand public. PSB[19] est une base de données génériques de 1814 modèles développée pour la comparaison, la classification et la reconnaissance d'objets 3D. Les classes de ce jeu de données sont basées sur la définition donnée aux objets par l'être humain. TOSCA[20] est une base de données d'êtres vivants de haute-résolution comprenant 80 modèles. Les différents modèles de ce jeu de données ont différentes positions en faisant varier leurs articulations. Ces variations sont appelées mouvements non-rigides. Les classes de ce jeu de données sont définies par un modèle avec tous ses mouvements non rigides. ModelNet40[21] et ShapeNet[22] sont deux bases de données développées pour l'apprentissage profond (*deep learning*). Ces bases de données sont composées de plus de dix mille modèles. Les différents modèles sont annotés manuellement de différents mots pour permettre d'entraîner des algorithmes. Toutes ces bases de données à l'exception de 3D warehouse, possèdent une vérité-terrain permettant d'évaluer les méthodes testées sur ces jeux de données.

La seconde raison est l'évolution et l'optimisation technologique des outils de calcul permettant l'utilisation et l'amélioration de méthodes précédemment limitées par la technologie. L'exemple le plus emblématique est l'apprentissage profond qui, appliqué à la reconnaissance d'images ou d'objets 3D, est devenu dominant dans le domaine grâce aux évolutions techniques. [23] [24] [25] [26] [27]

L'accroissement des bases de données et les évolutions technologiques font que l'intérêt pour la reconnaissance de formes appliquée aux protéines [28] [29] croît aussi. On peut citer la base de données Protein Data Bank [30] référençant plus de 125 000 structures de protéines numérisées qui est la banque de référence de structures protéiques.

1.2 Les protéines

L'étude des protéines est important car les protéines sont au centre de toutes les interactions du vivant et ont divers rôles biologiques, on peut citer entre autres : un rôle enzymatique en catalysant des réactions chimiques comme par exemple la lactase qui dégrade la molécule de lactose en glucose et en galactose ; un rôle hormonal en transmettant un message dans le reste du corps comme la mélatonine qui permet de régler les rythmes circadiens ; un rôle moteur en permettant le mouvement du corps comme par exemple la myosine qui est centrale dans la contraction musculaire. L'activité protéique se fait principalement à travers l'interaction entre protéines. Comprendre ces interactions est donc essentiel pour la compréhension du vivant.

Les protéines sont composées d'acides aminés, un groupement d'atomes, reliés entre eux par des liaisons covalentes dites peptidiques pour former une chaîne appelée chaîne polypeptidique. La chaîne polypeptidique adopte un repliement dans l'espace nécessaire à l'activité de la protéine.

On peut représenter les protéines comme une séquence linéaire d'acides aminés, en écrivant à la suite les différents acides aminés composants la protéine. On appelle cette représentation la structure primaire. Cette représentation est compacte et permet de stocker l'information sur peu d'espace mais le désavantage est une perte d'information. L'information de la position relative dans l'espace aux autres acides aminés n'est pas disponible avec cette représentation.

La structure secondaire est l'organisation locale de la protéine en motifs récurrents dans l'espace 3D. Les deux motifs les plus courants sont les **hélices** α et **feuillets** β. Les hélices α sont l'enroulement en forme hélicoïdale d'une partie de la protéine maintenue par des liaisons hydrogène. Les feuillets β ont une forme de feuille pliée en accordéon et composés de brins β reliés entre eux par des liaisons hydrogène.

Les protéines sont souvent représentées par leur structure tertiaire. Cette représentation contient la position dans l'espace 3D des atomes et leurs liaisons covalentes. Utiliser une représentation dans l'espace 3D a le désavantage d'être plus lourd à stocker mais présente l'avantage d'inclure plus d'informations.

Une autre manière de caractériser une protéine dans l'espace 3D est avec sa surface qui est une

représentation découlant de la représentation structurale. Il y a trois types de surfaces lorsque l'on parle de surface protéique.

1) La première est la **surface de van der Waals** [31]. La surface de van der Waals est liée aux rayons de van der Waals qui est la mesure du rayon d'une sphère théorique permettant de modéliser un atome. C'est la surface formée par l'union des sphères définies par le rayon de Van der Waals et variant selon le type d'atome.

2) La seconde est la **surface accessible au solvant** (SAS) [32]. Cette surface est calculée en faisant parcourir une sphère, appelée sonde (*probe*), représentant le solvant sur la surface de van der Waals. La sonde est en général de 1.4 angströms qui est la taille approximative du rayon d'une molécule d'eau. Le centre de la probe en tout point du parcours définit la surface accessible au solvant.

3) La troisième est la **surface exclue au solvant** (SES) [33] ou **surface de Connolly** et est intimement lié à la SAS. On fait un parcours avec une sonde représentant le solvant et ce qui définit la surface est le point de contact entre la sonde et la surface de van der Waals. Si la sonde a deux points de contact avec la surface de van der Waals, la courbe sur la sonde reliant les deux points est prise en compte pour définir la SES ce qui permet d'éviter d'avoir une surface incomplète lorsque la cavité est trop étroite pour la sonde.

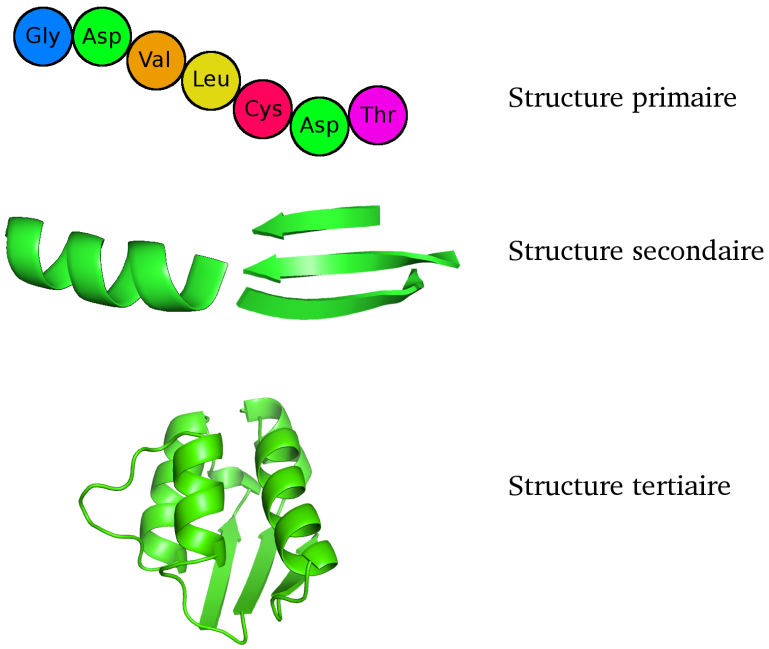


FIGURE 1.1 – De haut en bas, représentation de la structure primaire, secondaire et tertiaire d'une protéine

1.3 Acquisition des structures des protéines

Les protéines sont numérisées avec des outils de haute technologie utilisant la radiocristallographie, la spectroscopie par Résonance Magnétique Nucléaire (RMN) ou la cryo-microscopie électronique.

La première méthode à apparaître est la **cristallographie aux rayons X** ou **radiocristallographie**. Dans un premier temps il faut cristalliser la protéine en saturant le solvant dans lequel elle se trouve. Ensuite un rayon X est projeté sur le cristal de la protéine et les rayons diffractés passant par les atomes de la protéine sont capturés à l'aide d'un détecteur. Pour capturer la structure 3D de la protéine il faut à l'aide de la rotation du cristal, projeter le rayon X sous tous les angles du cristal. A partir du profil de diffraction obtenu, une carte de densité électronique est produite qui permettra de déterminer la structure protéique.

La seconde méthode est la **spectroscopie par Résonance Magnétique Nucléaire** (RMN). Le principe est de créer un champ magnétique intense autour de la protéine et d'appliquer un second champ de radiofréquence déterminée pour stimuler un transfert d'énergie entre le champ magnétique

et les atomes de la protéine. Certains noyaux se répartissent entre différents états énergétiques ce qui provoque un phénomène de résonance qui va pouvoir être mesuré. La fréquence de résonance d'énergie est propre à chaque type de noyaux et donc d'atomes. De plus la fréquence est influencée par l'environnement chimique du noyau (c'est-à-dire par les noyaux voisins) ce qui permet de déterminer sa position relative dans l'espace et donc la structure de la protéine.

La troisième méthode, plus récente, est la **cryo-microscopie électronique** (cryo-EM). Le principe est de congeler suffisamment rapidement la protéine pour éviter que le liquide autour de la protéine ne se cristallise et ne déforme la protéine. Une fois la protéine congelée il est possile de l'observer avec un microscope électronique en transmission. Cette technique de microscopie consiste à faire traverser un faisceau d'électrons à travers l'objet étudié. En compilant les images 2D prises sous différents angles par le microscope, la structure 3D de la protéine est obtenue.

La méthode de cristallographie aux rayons X a l'avantage de produire des structures sous hautes résolutions et n'est pas limitée par la taille de la protéine. Les inconvénients de cette méthode sont que toutes les protéines ne sont pas cristallisables et que la cristallisation altère la structure de la protéine. La spectroscopie par RMN permet d'obtenir des structures détaillées et de capturer différentes conformations de la protéine ou la liaison de petites molécules à la protéine. L'inconvénient principal est que cette méthode ne fonctionne que sur des petites protéines. La Cryo-EM a l'avantage d'observer en condition quasi naturelle la protéine et permet d'obtenir la structure des protéines de grandes tailles ou des complexes de protéines. L'inconvénient est qu'il est souvent difficile d'avoir une résolution proche de celles obtenues avec les deux autres méthodes sus-citées mais les avancées technologiques récentes permettent de réduire cet écart de résolution.

1.4 Objectif de la thèse

Dans le domaine de la biologie structurale la structure des protéines est plus souvent utilisée que la surface [34] [35] [36] [37] [38]. La principale raison est que les outils d'acquisition utilisés pour modéliser les protéines en 3D permettent d'obtenir les atomes et leurs positions dans l'espace. Pour obtenir la surface, une étape supplémentaire de calcul est nécessaire sur la structure. Au contraire la surface est plus souvent utilisée dans le domaine de la vision par ordinateur [39] [40] [41] [42] contrairement

aux représentations basées sur un squelette qui est proche de la représentation d'une protéine par sa structure. La raison est que les outils d'acquisition de formes dans ce domaine, étant basés sur la capture d'image, ne peuvent obtenir l'information à l'intérieur de l'objet [43] [44] [45] [46].

Nous avons décidé d'utiliser une représentation 3D car les autres représentations n'ont pas l'information spatiale qui est une donnée importante lors de l'étude des interactions protéine-protéine. Parmi les deux représentations 3D que sont la surface et la structure, la surface a été sélectionnée pour plusieurs raisons. La structure modélise l'entièreté des atomes or les atomes qui ne sont pas au contact du solvant n'ont que très peu d'influence sur les interactions protéine-protéine [47] [32] [48] [49]. La surface au contraire ne prend en compte que la forme de la protéine au contact du solvant et ignore donc la partie non-accessible par le solvant. Ceci permet de travailler sur une représentation plus abstraite de la protéine. Notre méthode utilise des techniques du domaine de la vision par ordinateur sur des bases de données de protéines. Sachant que ces méthodes sont développées pour des objets représentés par leur surface, cela constitue un argument supplémentaire pour travailler sur la surface des protéines.

Les travaux de cette thèse se concentrent sur la reconnaissance de formes d'objets 3D appliquée aux protéines. Un système de reconnaissance de formes se compose de deux parties, le calcul du descripteur qui permet d'encoder l'information des objets et la comparaison qui permet de classer les objets en comparant leurs formes encodées dans le descripteur précédemment calculé.

Un travail d'investigation a été fait pour trouver des méthodes de comparaison de formes génériques applicable aux protéines. Deux conditions sont posées pour développer la méthode de comparaison de protéines. La première condition est une méthode rapide pour pouvoir analyser en haut débit les protéines numérisées et dont la disponibilité en ligne augmente comme l'évolution de la taille de la Protein Data Bank[30] l'indique. En particulier optimiser le temps de comparaison est plus important que celui du calcul du descripteur car le descripteur a besoin d'être calculé qu'une seule fois. Le descripteur peut être ensuite stocké et réutilisé autant de fois que nécessaire pour la comparaison. La seconde est de maximiser le rappel (ou sensibilité), c'est-à-dire de minimiser le nombre d'objets classés comme faux négatifs. Le but étant de faire un premier tri pour éliminer les protéines dont on peut affirmer avec une confiance élevée qu'elles n'appartiennent pas à la même classe. Ensuite, une autre méthode avec une précision élevée peut être utilisée.

La comparaison des protéines peut être la première étape pour la découverte de nouvelles fonctions de protéine. Une fonction peut être inférée en comparant une protéine *query* ayant une fonction connue à d'autres protéines. Les protéines partiellement similaires dans la région où le *docking* a lieu peuvent posséder une fonction identique ou similaire à la protéine *query*. Ces protéines peuvent être ensuite étudiées en détail pour vérifier si elles possèdent effectivement cette fonction. Ce sont une partie des objectifs du projet ViDOCK dans lequel s'inscrit cette thèse. C'est pourquoi un descripteur local pouvant être utilisé pour de la reconnaissance globale ou partielle est développé.

Différents systèmes de reconnaissance de formes sont présentés. Tous possèdent un descripteur local basé sur une projection sur la sphère unitaire transformée dans l'espace 2D. Une projection sur la sphère unitaire implique de travailler avec des protéines de forme globulaire pour être efficaces, c'est pourquoi nous utilisons des jeux de données composées de protéines globulaires. La première représentation a pour valeur la convexité calculée sur le plan 2D et se nomme **Carte de Convexité**. La seconde représentation a pour valeur la courbure de l'objet 3D et est appelée **Carte de Courbure**. La troisième représentation ayant pour valeur le *Wave Kernel Signature* (WKS)[2] est appelée **Projected Wave Kernel Signature** (PWKS). La comparaison des objets est faite de manière exhaustive entre chaque sous-région recouvrante de chaque objet en utilisant l'absolu de la différence ou la fonction de score Earth Mover's Distance [50].

La thèse est organisée de la manière suivante. Dans la deuxième partie, un aperçu des méthodes de l'état de l'art sera donné. Dans la troisième partie, une méthode basée sur la convexité ainsi que sont analyse sera proposée. Dans la quatrième partie, différentes améliorations sont proposées suite à l'analyse des résultats de la troisième partie. Dans la cinquième partie, la nouvelle méthode découlant des améliorations précédentes sera présentée. Dans la sixième partie, une méthode alternative avec une autre valeur pour le descripteur sera proposée. Dans la septième partie, les objectifs futurs seront décrits.

Chapitre 2

Etat de l'art

Contenu

Trois grandes parties de l'état de l'art sont abordées, tout d'abord les méthodes liées à la reconnaissance de formes génériques, ensuite les méthodes d'alignement de structures de protéines qui peuvent être utilisées pour comparer des protéines et les méthodes du domaine de la vision par ordinateur utilisées sur un jeu de données composé de protéines.

2.1 Méthodes de reconnaissance de formes

Un système de reconnaissance de formes se décompose en deux parties. La première est la création de la représentation de l'objet comparé. Une représentation proche de la réalité visible peut être utilisée, des propriétés intéressantes de l'objet peuvent être extraites et représentées ou une transformation peut être appliquée à l'objet. La seconde partie est la comparaison d'un objet avec un autre. Pour cela, une fonction de distance permettant de donner une valeur comparative est appliquée.

Il existe trois catégories principales de descripteurs pour les méthodes de reconnaissance de formes : les descripteurs basés sur les **histogrammes**, sur le **spectre** d'une fonction et sur une **projection dans un espace 2D**. Il est courant qu'une méthode soit comprise dans plus d'une catégorie. Les méthodes seront ici classifiées selon leur caractéristique principale.

La comparaison des descripteurs peut se faire de différentes manières. La comparaison des descripteurs est souvent dépendante de la nature du descripteur. Un descripteur dont l'information est extraite du voisinage d'un point est différent d'un descripteur dont l'information est obtenue sur l'entièreté de l'objet étudié.

La comparaison peut donc être classifiée selon qu'elle se fait **globalement**, c'est-à-dire que l'on compare l'entièreté de l'objet à travers une seul instance d'un descripteur ou **localement** si l'on compare des régions de l'objet en utilisant plusieurs instances du descripteur. On peut donc classifier les descripteurs selon la comparaison qui va en découler en trois catégories : les **descripteurs globaux**, **les descripteurs locaux** et les **descripteurs globaux-locaux** qui peuvent aussi être appelés **hybrides**, ces derniers pouvant varier selon leur paramétrisation.

La première classification présentée permet de différencier les descripteurs selon le type d'outils mathématiques utilisé pour les concevoir tandis que la deuxième classification permet de différencier les descripteurs selon la nature du résultat souhaité lors de la comparaison.

Différentes méthodes de l'état de l'art sont présentées. Les méthodes basées sur le spectre : ShapeDNA [51][52], Global Point Signature [53], Heat Kernel Signature [54] et Wave Kernel Signature [2]. Spin Image [55], Point Feature Histograms [56] [4] et Signature of Histograms of OrienTations [5] sont des méthodes basées sur les histogrammes. Pour les méthodes basées sur la projection dans un espace 2D il y a Scale Invariant Feature Transform [57][6], PANORAMA [7] et Shape Similarity System driven by Digital Elevation Models [8]. Enfin une dernière méthode qui ne correspond à aucune des trois grandes catégories est le descripteur de Zernike 3D qui se base sur les moments du polynôme de Zernike [58][59].

2.1.1 Méthodes basées sur le spectre

La géométrie spectrale est la représentation de la géométrie ou topologie d'un objet par le spectre de **l'opérateur de Laplace-Beltrami**. L'opérateur de Laplace-Beltrami est la généralisation de l'opérateur laplacien aux variétés riemanniennes. L'opérateur de Laplace-Beltrami est la composition entre le gradient et la divergence d'une fonction f défini de la manière suivante : $\Delta f = div(grad(f))$. Une variété riemannienne est un espace topologique muni d'une distance riemannienne qui est une distance sur un espace courbé. Le spectre de l'opérateur de Laplace-Beltrami est une suite de valeurs propres découlant de l'opérateur de Laplace-Beltrami.

L'avantage d'utiliser un spectre pour descripteur est sa propriété d'invariance aux isométries [51].

2.1.1.1 ShapeDNA

Le premier descripteur spectral pour la reconnaissance de formes est la méthode ShapeDNA [51][52]. Le spectre d'un objet se base sur les valeurs propres de l'équation d'Helmoltz (ou le problème des valeurs propres laplaciennes), qui définit les ondes stationnaires (cf Figure 2.1) de l'équation de propagation des ondes. Plus précisément ce sont les vibrations naturelles qui sont des ondes stationnaires sur la surface sans forces extérieures.

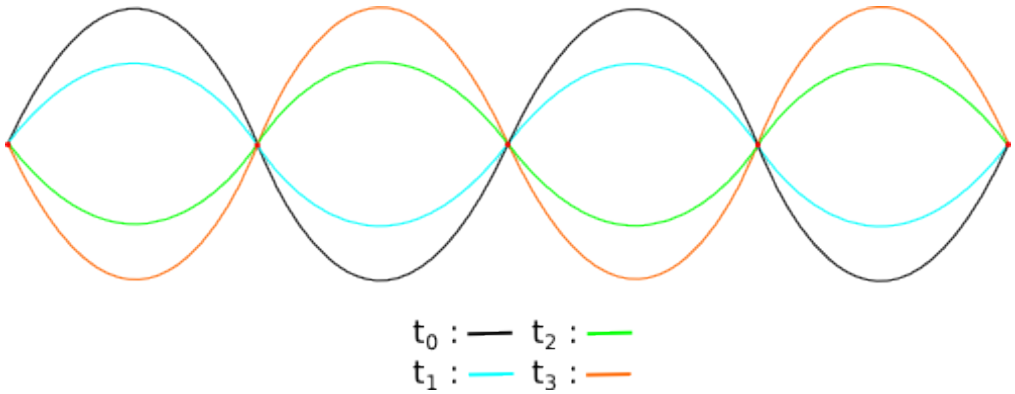


FIGURE 2.1 – Schéma de l'évolution d'une onde stationnaire pour différentes valeurs de temps.

L'équation d'Helmoltz est définie de la manière suivante :

$$\Delta f = -\lambda f \tag{2.1}$$

Avec $f \in C^2$ une fonction définie sur une variété riemmanienne et représentant les vibrations naturelles. λ est la valeur propre de l'équation représentant le nombre d'ondes qui est inversement proportionnel à la longueur d'onde.

Le spectre est donc l'ensemble des k premières valeurs propres normalisées λ_i tel que $0 \leq \lambda_1 \leq \lambda_2 \leq ...$, k étant un nombre choisi arbitrairement en fonction de la précision souhaitée. On obtient un descripteur global définissant avec le spectre l'entièreté de l'objet.

Le spectre d'un objet vérifie plusieurs propriétés [52].

— La première est l'invariance aux isométries dues au fait que les valeurs propres ne dépendent que de l'opérateur de Laplace-Beltrami. L'invariance aux isométries implique une invariance aux rotations et translations.

— La seconde propriété est l'invariance à la mise à l'échelle due à la normalisation des valeurs propres.

— La propriété de similarité (deux objets similaires ont un spectre similaire) est une propriété intrinsèque au spectre.

— La propriété de complétude établit que deux objets ayant un spectre identique sont le même objet à un mouvement non rigide près. Cette propriété n'est pas entièrement vraie car il existe

quelques rares exemples [60] d'objets non-congruents et pourtant isospectraux. Il est expliqué lors de la présentation de ShapeDNA [51] que malgré une propriété isométrique partielle les cas restent suffisamment anecdotiques pour que l'on puisse considérer la propriété d'isospectralité d'un spectre comme valide dans la pratique. Les cas ne vérifiant pas cette propriété sont des constructions artificielles complexes.

2.1.1.2 Global Point Signature

La seconde méthode spectrale est Global Point Signature (GPS) [53] qui en plus d'utiliser les valeurs propres dans un spectre, ajoute l'utilisation des fonctions propres associées à ces valeurs pour proposer un descripteur défini comme la signature d'un objet.

La signature GPS est définie de la manière suivante :

$$GPS(p) = (\frac{1}{\sqrt{\lambda_1}} f_1(p), \frac{1}{\sqrt{\lambda_2}} f_2(p), \frac{1}{\sqrt{\lambda_3}} f_3(p), ...)$$

Avec $f_i(p)$ représentant la $i$ème fonction propre au point p et λ_i la $i$ème valeur propre de l'opérateur de Laplace-Beltrami telles que définies dans subsubsection 2.1.1.1.

Ce descripteur a différentes propriétés :

— Deux points distincts d'une surface possèdent une signature différente.

— Le GPS étant issu de l'opérateur de Laplace-Beltrami est invariant aux isométries.

— L'utilisation de valeurs propres et vecteurs propres de Laplace-Beltrami induit la propriété de complétude.

— Le GPS est invariant aux translations et rotations du fait de l'invariance aux isométries qui est une propriété de l'opérateur de Laplace-Beltrami.

2.1.1.3 Heat Kernel Signature

Heat Kernel Signature (HKS)[54] est une méthode qui se base sur le noyau de chaleur, une équation modélisant la diffusion de la chaleur sur un objet en fonction du temps (cf Figure 2.3). L'équation de

la diffusion de la chaleur est la suivante :

$$\Delta_M u(x,t) = -\frac{\partial u(x,t)}{\partial t} \tag{2.2}$$

Avec Δ_M l'opérateur de Laplace-Beltrami défini sur la variété riemannienne compacte M et $u(x,t)$ l'équation de distribution de la chaleur au point x et au temps t.

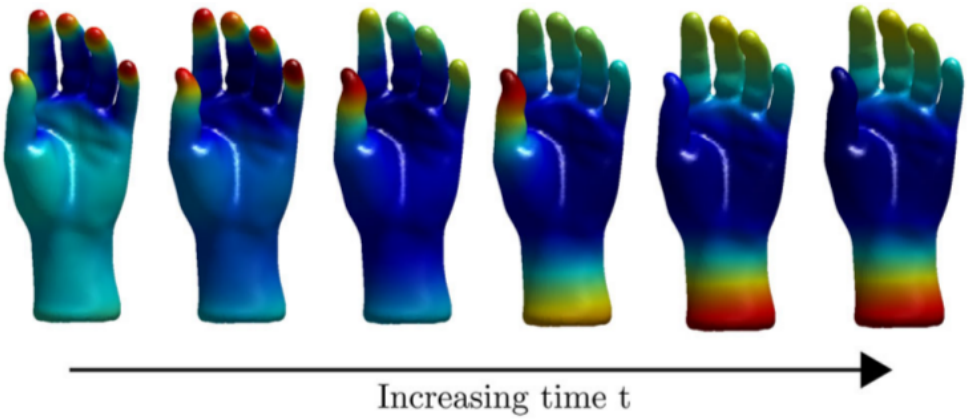


FIGURE 2.2 – Exemple de la diffusion de la chaleur en fonction du temps sur un objet 3D. Les couleurs variant de bleu (valeur faible) à rouge (valeur élevée). Illustration de [1].

Si on définit $f(x) = u(x,0)$ et $H_t(f)$ l'opérateur de chaleur tel que $lim_{t\to 0} H_t(f) = f$. La solution à l'équation de diffusion de la chaleur est :

$$H_t(f(x)) = \int_M k_t(x,y) f(y) dy \tag{2.3}$$

Les fonctions $k_t(x,y)$ qui vérifient cette équation sont appelées *kernel* de chaleur et représentent la quantité de chaleur transférée d'un point x à un point y au temps t. La décomposition en éléments propres de $k_t(x,y)$ est :

$$k_t(x,y) = \sum_{i=0}^{\inf} e^{-\lambda_i t} \phi_i(x) \phi_i(y) \tag{2.4}$$

Avec $lambda_i$ et ϕ_i respectivement la $i$ème valeur propre et fonction propre de l'opérateur de Laplace-Beltrami.

Pour simplifier le descripteur, sans pour autant perdre trop d'informations, il est proposé de ne prendre en compte que le kernel de chaleur lorsque l'on est sur le même point, c'est-à-dire de définir le Heat Kernel Signature de la manière suivante :

$$HKS(x,t) = k_t(x,x) \tag{2.5}$$

La méthode HKS possède trois propriétés :

— La première découle de l'utilisation de l'opérateur de Laplace-Beltrami qui est l'invariance isométrique

— La seconde est la propriété de multiéchelle. Plus la valeur t est grande plus la chaleur s'est diffusée sur la surface, on obtient donc un descripteur prenant en compte un plus grand voisinage qu'au temps précédent. En faisant varier le temps il est possible de sélectionner la taille des régions de la surface comparée.

— La troisième propriété est la propriété informative. Un objet est entièrement caractérisé par le HKS à une isométrie près.

— La dernière propriété mise en avant est la stabilité aux perturbations. La stabilité est proportionnelle au temps t et découle directement du mouvement brownien de la modélisation. La valeur de chaleur en un point x tend à s'uniformiser en fonction du temps écoulé.

Différentes variantes du HKS ont été proposées, on peut citer par exemple le Scale Invariant Heat Kernel Signature (SI-HKS) [61] qui ajoute l'invariance à la mise à l'échelle, Volumetric Heat Kernel Signatures [62] utilisant le volume plutôt que la surface d'un objet pour calculer le HKS ou Generalized Heat Kernel Signature (GHKS) [63] qui propose un HKS basé sur l'opérateur de Laplace-de Rham sur les formes différentielles de degré 1 (1-formes). L'opérateur de Laplace-de Rham est la généralisation de l'opérateur de Laplace-Beltrami aux r-formes. Une 1-formes est champ de formes linéaires sur une variété différentielle. Une forme linéaire étant un cas particulier d'application linéaire qui pour un vecteur fait correspondre une valeur scalaire.

2.1.1.4 Wave Kernel Signature

Wave Kernel Signature (WKS) [2] est basé sur la représentation de l'énergie de particules élémentaires sur la surface d'un objet. L'équation modélisant ceci est l'équation de Schrödinger :

$$\frac{\partial \psi}{\partial t}(x,t) = i\Delta\psi(x,t) \tag{2.6}$$

ψ étant l'équation d'onde.

L'équation décrivant l'énergie des particules élémentaires utilise l'opérateur de Laplace-Beltrami et il est ainsi possible d'obtenir un spectre à partir de cette équation.

Cette méthode est proche de la méthode HKS et répond à certains problèmes liés à l'HKS. Le premier problème est la suppression des hautes fréquences lorsque la valeur de temps t du HKS est élevée. De plus, le HKS est dominé par les basses fréquences qui représentent les propriétés macroscopiques de la surface, faisant qu'il est difficile de comparer des surfaces dans la situation où il est nécessaire de prendre en compte les petites formes de la surface pour un alignement de haute précision.

Une particule est sélectionnée en un point quelconque x sur la surface avec une approximation d'énergie E au temps $t = 0$. La distribution de la probabilité d'énergie est f_E^2 avec une espérance de E. La fonction d'ondes de la particule est la suivante :

$$\psi_E(x,t) = \sum_{k=0}^{\inf} e^{iE_k t}\phi_k(x) f_E(E_k) \tag{2.7}$$

Avec E_k et ϕ_k respectivement les valeurs propres et les fonctions propres de l'opérateur Δ. La probabilité de mesurer une particule au point x est $|\psi_E(x,t)|^2$.

Le WKS est défini comme la probabilité moyenne dans le temps de mesurer une particule en x :

$$WKS(E,x) = \lim_{T\to\inf} \frac{1}{T} \int_0^T |\psi_E(x,t)|^2 \tag{2.8}$$

Sachant que $e^{iE_k t}$ sont orthogonales avec la norme L^2, on obtient :

$$WKS(E,x) = \sum_{k=0}^{\inf} \phi_k(x)^2 f_E(E_k)^2 \tag{2.9}$$

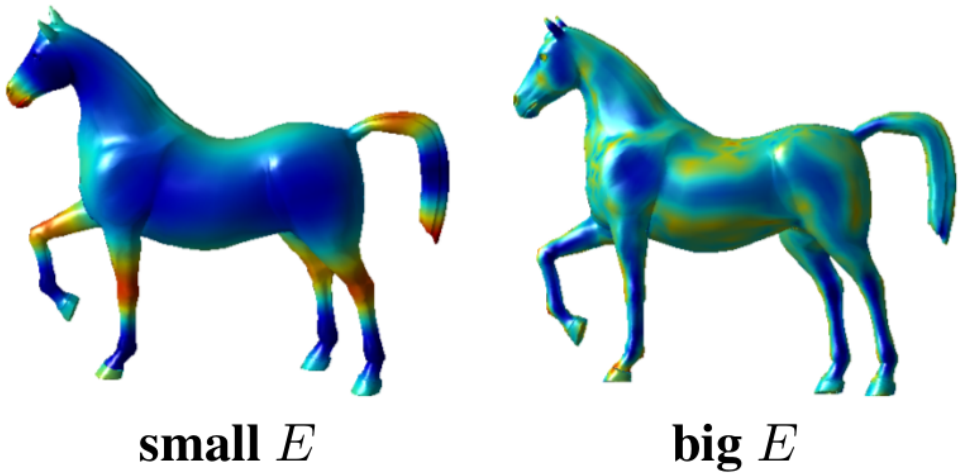


FIGURE 2.3 – Exemple de la probabilité de mesurer une particule sur un objet 3D pour deux valeurs initiales de E. Le rouge signifiant une probabilité élevée et le bleu un probabilité faible. Illustration de [2].

Cette méthode possède les propriétés suivantes :

— On retrouve l'invariance aux mouvements non-rigides car on utilise le spectre de l'opérateur de Laplace-Beltrami.

— Le WKS est informatif, c'est-à-dire que l'on peut retrouver un objet à une isométrie près à partir de sa signature.

— Le WKS possède une notion d'échelle claire, les parties avec une énergie haute correspondent à des zones influencées par la géométrie locale tandis que une faible énergie est influencée par la géométrie globale.

— La notion d'échelle permet d'avoir une stabilité aux perturbations de la surface.

Il est possible de créer une descripteur WKS invariant à l'échelle en suivant le même procédé que celui démontré pour le HKS [61].

2.1.2 Méthodes basées sur les histogrammes

Les histogrammes permettent d'agréger différentes informations géométriques d'un objet de manière concise, ainsi pouvant diminuer le temps de calcul de la comparaison.

2.1.2.1 Spin Image

La méthode de comparaison **Spin Image**[55] associe un histogramme 2D pour chaque point.

Dans un premier temps on définit un point orienté O qui est composé de la position p sur la surface et de la normale n à la surface. On définit la base 2D (p, n) en utilisant le plant tangent $\mathcal{P}$ passant par p et perpendiculaire à n et la droite $\mathcal{L}$ passant par p et parallèle à n. Les deux coordonnées de la base sont α et β respectivement la distance perpendiculaire à $\mathcal{L}$ et la distance perpendiculaire à $\mathcal{P}$. On obtient un système de coordonnées local tel qu'illustré en Figure 2.4.

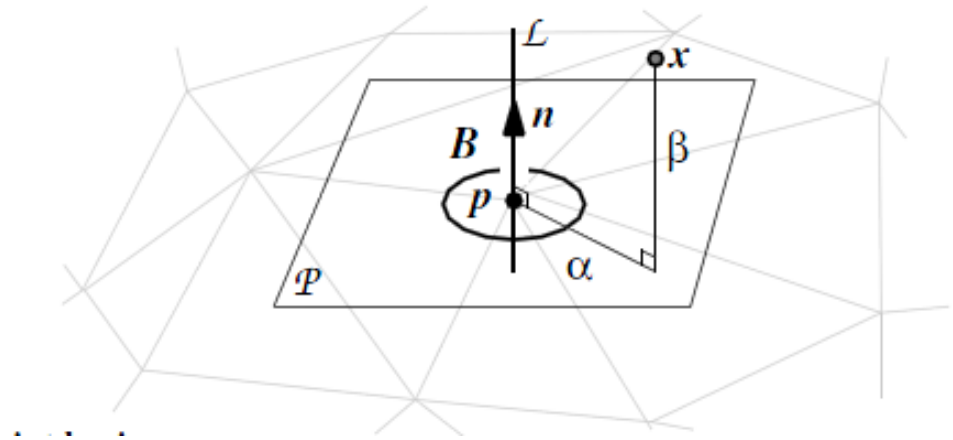


FIGURE 2.4 – Base formée par un point orienté. Illustration extraite de [3]

La fonction *spin-map* S_O projette un point 3D x dans la base (p, n) lié au point orienté O :

$$S_O(x) = (\alpha, \beta) = (\sqrt{\|x - p\|^2 - (n \cdot (x - p))^2}, n \cdot (x - p)) \tag{2.10}$$

Avec $\alpha \in \mathbb{R}_+$ et $\beta \in \mathbb{R}$.

On crée une *spin image* pour un point orienté O en appliquant la fonction spin-map sur tous les points de l'objet 3D. Ensuite un histogramme 2D est créé pour comptabiliser le nombre de points.

La taille de la spin image (i_{max}, j_{max}) est définie par les valeurs α et $|\beta|$ maximum appelées α_{max} et β_{max}. Ce sont les valeurs maximum parmi toutes les bases des points orientés. Ensuite est défini $i_{max} = \frac{2\beta_{max}}{b} + 1$ et $j_{max} = \frac{\alpha_{max}}{b} + 1$ avec b la taille du *bin* de l'histogramme. La position (i, j) dans l'histogramme d'un point (α, β) est : $i = \frac{\beta_{max} - \beta}{b}$ et $j = \frac{\alpha}{b}$

L'histogramme représente la position par rapport au point sélectionné dans la base du point orienté

O et peut être vu comme une rotation d'un plan autour de la normale au point.

— Les spin-images sont indépendantes aux transformations rigides car elles sont sont définies relativement à la position d'un point.

— La méthode des spin image est robuste lorsqu'une scène est encombrée ou lorsqu'il y a des occlusions, c'est à dire des parties incomplètes de l'objet 3D.

2.1.2.2 Point Feature Histograms

Point Feature Histograms (PFH) [56] et sa version améliorée appelée **Fast Point Feature Histograms** (FPFH) [4] définissent trois histogrammes pour chaque point de l'objet pour la méthode FPFH et quatre histogrammes pour PFH. Les histogrammes représentent la distribution dans le voisinage du point sélectionné de trois caractéristiques géométriques pour FPFH et quatre pour PFH.

Pour PFH, les k voisins d'un point p se trouvant à l'intérieur d'une sphère de rayon r sont sélectionnés. Pour chaque paire de points p_i et p_j, $i \neq j$, dans le k-voisinage de p, avec les normales estimées n_i et n_j, le repère de Darboux est défini comme :

$$u = n_i \tag{2.11}$$

$$v = (p_j - p_i) * u \tag{2.12}$$

$$w = u * v \tag{2.13}$$

Un repère de Darboux est un repère mobile suivant une courbe définie sur une surface.

Les variations angulaires et la distance euclidienne est calculée à partir de ce repère, ce sont les 4 caractéristiques géométriques de PFH :

$$\alpha = v \cdot n_i \tag{2.14}$$

$$\phi = \frac{(u \cdot (p_j - p_i))}{\|p_j - p_i\|} \tag{2.15}$$

$$\theta = \arctan(w \cdot n_j, u \cdot n_j) \tag{2.16}$$

$$d = \|p_j - p_i\| \tag{2.17}$$

α et ϕ sont le produit scalaire qui représente le cosinus de l'angle entre les deux vecteurs. θ est l'angle entre u et n_j projetés sur le plan formé par les vecteurs w et $n_j = u$. d est la distance euclidienne entre p_i et p_j.

FPFH se sépare de la distance euclidienne d car cela n'affecte pas la qualité des résultats [4]. Dans un premier temps, les caractéristiques géométriques entre le point *query* p et ses voisins p_k sont calculées et réunies en un histogramme appelé $SPFH$ (pour Simplified Point Feature Histogram). Ensuite les histogrammes $SPFH(p_k)$ entre les voisins de p et leurs propres voisins sont calculées (cf Figure 2.5). Les différents histogrammes $SPFH(k)$ sont sommés avec l'histogramme $SPFH(p)$ en étant chacun pondérés par l'inverse du coefficient ω_k qui représente la distance entre p et p_k.

$$FPFH(p) = SPFH(p) + \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\omega_k}SPFH(p_k) \tag{2.18}$$

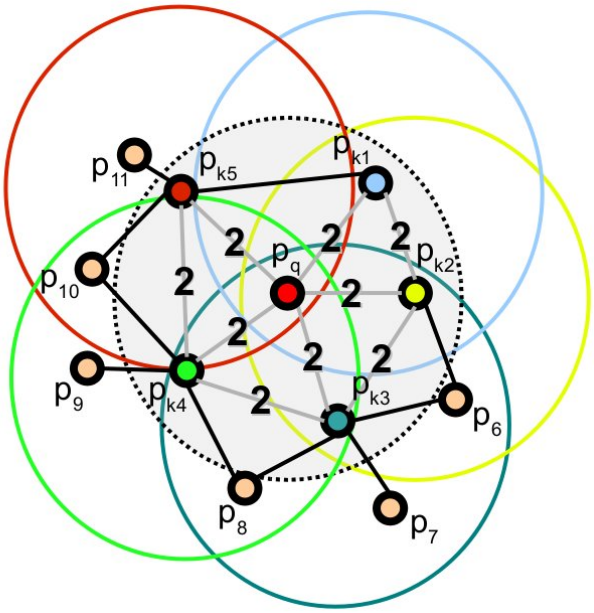


FIGURE 2.5 – Les voisins du point p_q sont dans le cercle en pointillés noirs du centre et les voisins des voisins de p_q sont dans les cercles de couleurs. L'histogramme SPFH est calculé pour le point p_q et les points p_{ki} qui sont les points dans le voisinage de p_q. Illustration extraite de [4]

(F)PFH possède les propriétés suivantes :

— Les deux descripteurs sont invariants aux transformations rigides car les caractéristiques géomé-

triques sont calculées dans un repère local

— Elles sont également invariantes à la mise à l'échelle du fait de l'utilisation d'un repère local pour chacun des points.

2.1.2.3 Signature of Histograms of Orientations

Signatures of Histograms of OrienTations (SHOT) [5] propose un référentiel local unique permettant ainsi de diminuer le temps de calcul, l'espace de stockage et les erreurs dues à de multiples référentiels.

Le calcul du référentiel unique se base sur une estimation des moindres carrés de la direction des normales [64] [65] en appliquant une décomposition des valeurs propres de la matrice de covariance M des points p_i du plus proche k-voisinage :

$$M = \frac{1}{k}\sum_{i=0}^{k}(p_i - \hat{p})(p_i - \hat{p})^T, \hat{p} = \frac{1}{k}\sum_{i=1}^{k} p_i \tag{2.19}$$

Le signe de la normale, l'un des trois axes du référentiel, est ensuite déterminé de manière heuristique et globale. Avec la méthode SHOT il est proposé pour améliorer la robustesse au bruit, de considérer tous les points à l'intérieur de la sphère de rayon R plutôt que les k plus proches voisins :

$$M = \frac{1}{\sum_{i:d_i \leq R}(R - d_i)} \sum_{i:d_i \leq R} (R - d_i)(p_i - p)(p_i - p)^T \tag{2.20}$$

Avec p un point quelconque de la surface.

Le signe des trois axes est déterminé localement[66] en utilisant le signe des vecteurs singuliers des matrices à gauche et à droite de la décomposition. Si les vecteurs à droite et à gauche sont de signe différents, on sélectionne le signe du vecteur dont la somme de l'absolu est la plus élevée. On divise la sphère autour du point *query* en 8 zones selon l'axe azimutal (une coupe du plan tel que toute droite dans le plan soit parallèle à z) , l'axe radial (une coupe en fonction de la taille du rayon) et l'axe d'élévation (une coupe du plan passant tel que tout point du plan est la même valeur pour la coordonnée z). 8 divisions azimutales, 2 divisions radiales et 2 divisions d'élévation sont suffisantes pour le bon fonctionnement de la méthode [5].

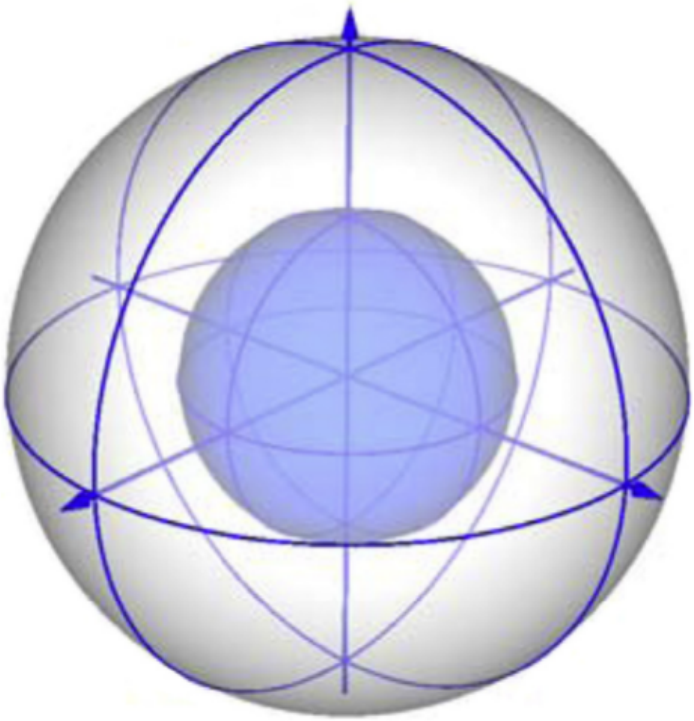


FIGURE 2.6 – Sphère avec 4 divisions azimutales, 2 divisions radiales et 2 divisions d'élévation. Illustraton extraite de [5]

Dans chacune de ces divisions un histogramme de taille n et divisé équitablement est calculé. L'histogramme possède en abscisse la valeur du cosinus entre l'axe z et la normale d'un point q appartenant au voisinage de p et en ordonnée le cardinal des valeurs de cosinus. Chacune des valeurs en abscisse est égale à $\cos_q = z_p \cdot n_q$.

— L'utilisation d'un référentiel local unique permet d'avoir une invariance aux transformations rigides (translations et rotations).

— Avec les évaluations expérimentales il a été prouvé que la méthode est robuste au bruit et aux occlusions.

Une variante de SHOT existe ou les couleurs sont prises en compte dans un autre histogramme [5] ainsi qu'une variante binaire [67] où les valeurs de SHOT sont remplacées par 0 ou 1 pour avoir une méthode plus rapide et prenant moins d'espace de stockage.

2.1.3 Méthodes basées sur une projection 2D

Les méthodes basées sur une projection 2D emploient une ou plusieurs représentations 2D de l'objet 3D étudié. L'avantage d'utiliser des représentations 2D est un stockage des données plus léger et un traitement des données plus rapide qu'avec une représentation 3D.

2.1.3.1 Scale Invariant Feature Transform

Scale Invariant Feature Transform (SIFT) [57][6] est une méthode appliquée sur une image 2D d'un environnement 3D. SIFT est une méthode de reconnaissance d'objets 3D en utilisant une seule projection dans l'espace 2D qui est une image (photographie, capture d'écran). La méthode associe entre eux des points clés détectés sur deux images. Ces points clés sont définis par rapport aux objets 3D de la scène représentée par une image 2D.

La première étape est de faire une convolution d'une image I avec un filtre Gaussien G.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2.21}$$

Ou x et y sont les coordonnées cartésiennes de l'image et σ le paramètre du filtre Gaussien.

La deuxième étape est de définir les différences de gaussiennes (DoG pour Difference of Gaussians) appelé D.

$$D(x, y, \sigma) = L(x, y, k_{i+1}\sigma) - L(x, y, k_i\sigma) \tag{2.22}$$

Avec k_{i+1} et k_i deux facteurs tel que $k_{i+1} = k * k_i$, $k_0 = 1$ et k une constance. Les différences gaussiennes sont donc la différence de deux images consécutives avec la convolution d'un filtre gaussien incrémenté comme schématisé dans la Figure 2.7.

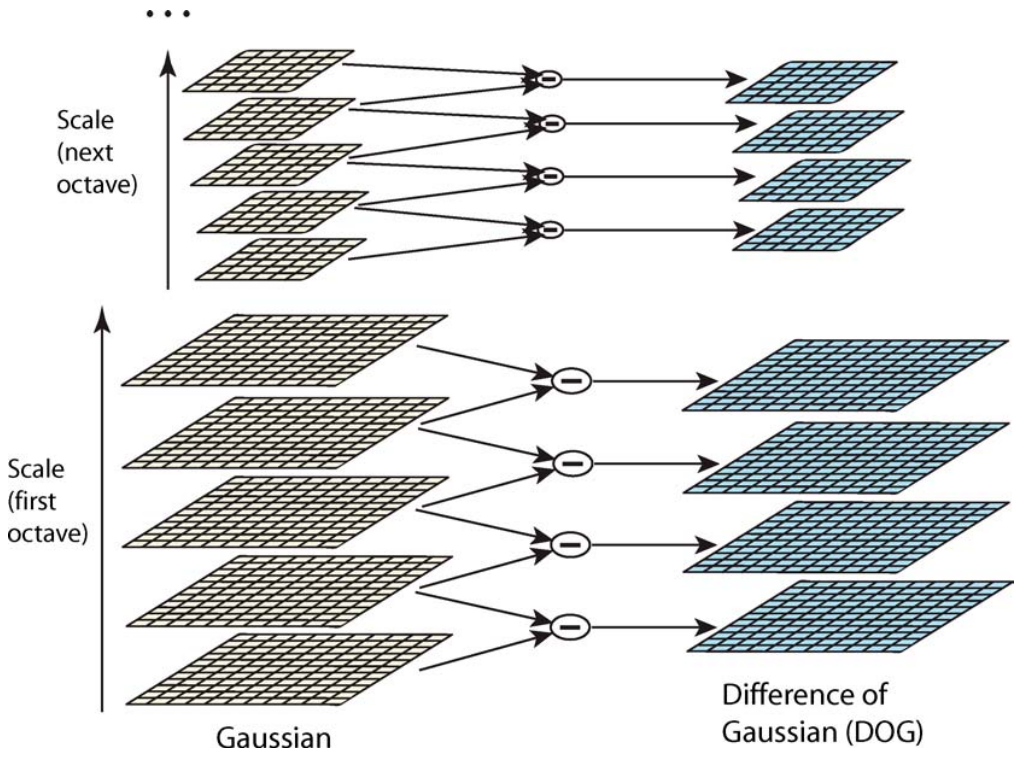


FIGURE 2.7 – Soustraction d'image avec l'application d'un filtre gaussien incrémenté (à gauche) pour obtenir la différence de gaussiennes (à droite) et pour différentes octaves. Illustration extraite de [6]

L'utilisation d'une pyramide, qui divise la résolution par deux de l'image, et dont chaque niveau est appelé octave, permet de traiter plus rapidement les points-clé. Pour chaque passage d'une octave à une autre, la valeur de base σ est multipliée par deux. L'idée est de chercher les points-clé sur l'image avec la plus basse résolution puis de remonter les octaves pour vérifier si ce sont toujours des points-clé. Ceci permet d'augmenter le temps de traitement des images.

Les points-clé sont définis comme les maxima ou les minima locaux. Un extremum local est la valeur de DoG par rapport à ses voisins dans un carré de 3×3 pixels centré sur cet extremum mais aussi par rapport aux voisins des DoG voisines. Un point (x, y) est un extremum local si :

$$\begin{cases} D(x,y,\sigma) \leq D(x+d_x, y+d_y, d_\sigma\sigma), d_x, d_y \in \{-1,0,1\}, d_\sigma \in \{k^{-1},1,k\} \\ \text{ou} \\ D(x,y,\sigma) \geq D(x+d_x, y+d_y, d_\sigma\sigma), d_x, d_y \in \{-1,0,1\}, d_\sigma \in \{k^{-1},1,k\} \end{cases} \tag{2.23}$$

Pour un point-clé détecté en (x_0, y_0, σ_0), l'amplitude $m(x, y)$ et l'orientation $\theta(x, y)$ de chaque point

(x, y) dans le voisinage du point-clé sont calculées sur l'image lissée L

$$\begin{cases} m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \\ \theta(x,y) = \tan^{-1}(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}) \end{cases} \tag{2.24}$$

On crée un histogramme des orientations des points du voisinage, divisé en 36 intervalles représentant chacun 10 degrés d'angle, pondéré par l'amplitude m et une fenêtre gaussienne centrée sur le point-clé avec pour paramètre $1.5\sigma_0$. La fenêtre gaussienne permet de donner moins d'importance aux points aux extrémités. Le pic le plus élevé permet de donner l'orientation du point-clé pour avoir l'invariance aux rotations et l'utilisation de σ_0 permet l'invariance à l'échelle.

Les coordonnées et les valeurs de gradient du point-clé sont recalculées selon l'orientation de celui-ci fourni par l'histogramme. Une fenêtre gaussienne centrée sur le point-clé avec pour paramètre $1.5\sigma_0$ est appliqué sur tous les points du voisinage de taille 16×16 pixels. Ensuite un histogramme des orientations de 8 intervalles est calculé pour des sous régions de 4×4 pixels dans la région de 16×16 pixels centré sur le point-clé (cf Figure 2.8). Le descripteur créé est un vecteur comprenant 4×4 histogrammes concaténés de 8 intervalles, il y a donc 128 éléments pour chaque points-clé.

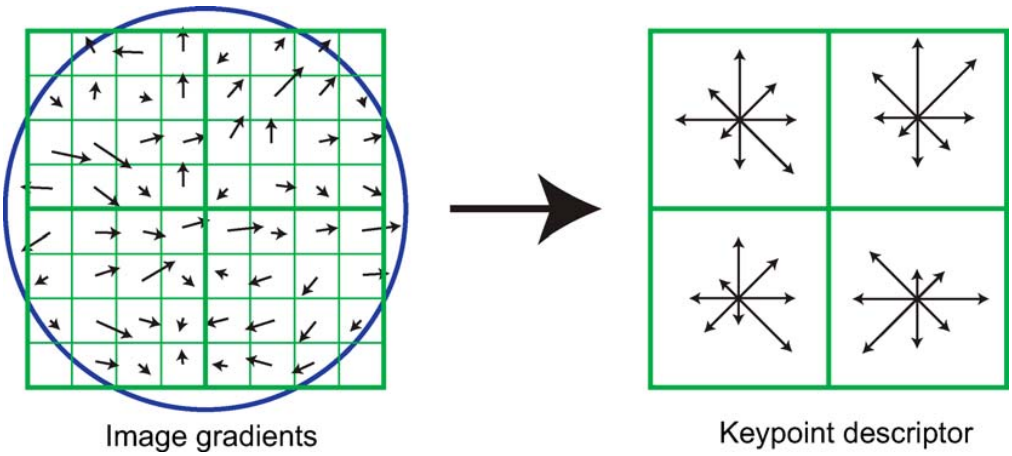


FIGURE 2.8 – Creation du descripteur SIFT. Le schéma représente une région de 8×8 pixels et un descripteur de $2 \times 2 \times 8$ éléments et le cercle bleu est la fenêtre gaussienne. Illustration extraite de [6]

Les différentes propriétés de SIFT sont :

— L'invariance aux translations et rotations est due à respectivement la création d'un descripteur localement et au calcul de l'orientation du gradient du point-clé

— L'utilisation de DoG permet l'invariance à la mise à l'échelle.

— Plus généralement il a été démontré expérimentalement que SIFT est robuste aux transformations affines mineures

2.1.3.2 PANORAMA

La méthode PANORAMA[7] projette l'objet 3D sur trois cylindres, ces cylindres étant orientés selon les trois axes de coordonnées. La projection cylindrique est ensuite transformée en une image 2D.

La première étape est la normalisation de la pose de l'objet 3D. La normalisation de la translation se fait en calculant le centroïde du maillage de l'objet 3D avec l'analyse en composantes principales continues (CPCA) [68]. Le centroïde est la moyenne entre tous les centroïdes des triangles composant le maillage avec un poids proportionnel à la surface du triangle. Pour normaliser la rotation la CPCA puis une PCA appliquée sur les normales des faces du maillage [69] (NPCA) est utilisé. Ainsi deux versions d'alignement de l'objet 3D de base sont obtenues.

L'étape suivante est d'obtenir les panoramiques. Une panoramique étant dans ce contexte ci, la représentation dans le plan 2D d'un objet 3D suite à une projection cylindrique. Trois cylindres sont centrés sur l'origine, avec l'axe du cylindre parallèle à respectivement l'un des trois axes des coordonnées que l'on appelle axe directeur. Ensuite une discrétisation des cylindres (cf Figure 2.9) est produite en faisant B coupes du cylindre perpendiculaire à l'axe directeur. Ces coupes sont divisées en $2B$ points sur sur leur circonférence. Un rayon est projeté à partir du centre du cercle c_v de la v-ème coupe dans la direction du u-ème point de la coupe. On obtient ainsi un ensemble de point (ϕ_u, y_v) où ϕ_u est l'angle du u-ème points sur le plan de la coupe v-ème coupe et y_v est la hauteur de la v-ème coupe. La valeur zéro est attribuée en chacun des points de la discrétisation.

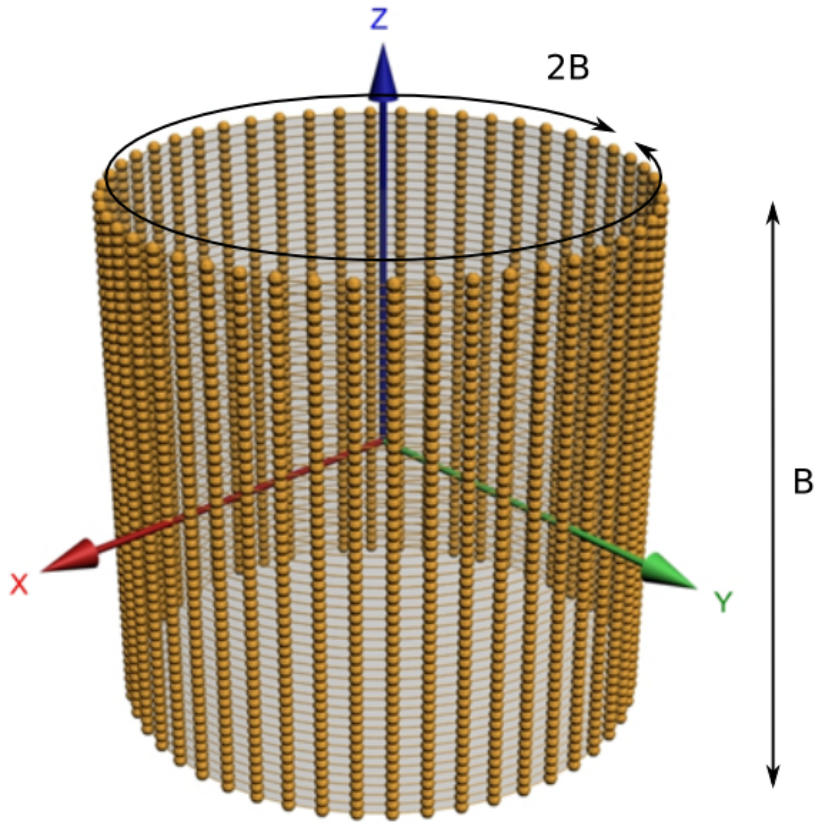


FIGURE 2.9 – Discrétisation du cylindre en B points sur la hauteur et $2B$ points sur la circonférence du cylindre. Illustration extraite de [7]

Deux valeurs sont calculées, s_1 et s_2. La première $s_1(\phi_u, y_v)$ est la distance maximum entre le centre c_v et l'intersection d'un rayon allant dans la direction du u-ème point de la coupe v. Si d est une fonction de distance, alors $s_1(\phi_u, y_v) = d(\phi_u, y_v)$. La seconde, $s_2(\phi_u, y_v)$, est basée sur l'angle entre le rayon partant de c_v dans la direction de ϕ_u et la normale du point le plus loin intersecté par le rayon appelé $A(\phi_u, y_v)$. Alors $s_2(\phi_u, y_v) = |\cos(A(\phi_u, y_v))|^n$ avec $n \geq 2$. Sachant que s_1 et s_2 sont calculés selon l'un des trois axes x, y, z on obtient 6 projections cylindriques pour $k \in \{1, 2\}$ et $t \in \{x, y, z\}$: $s_{k,t}$.

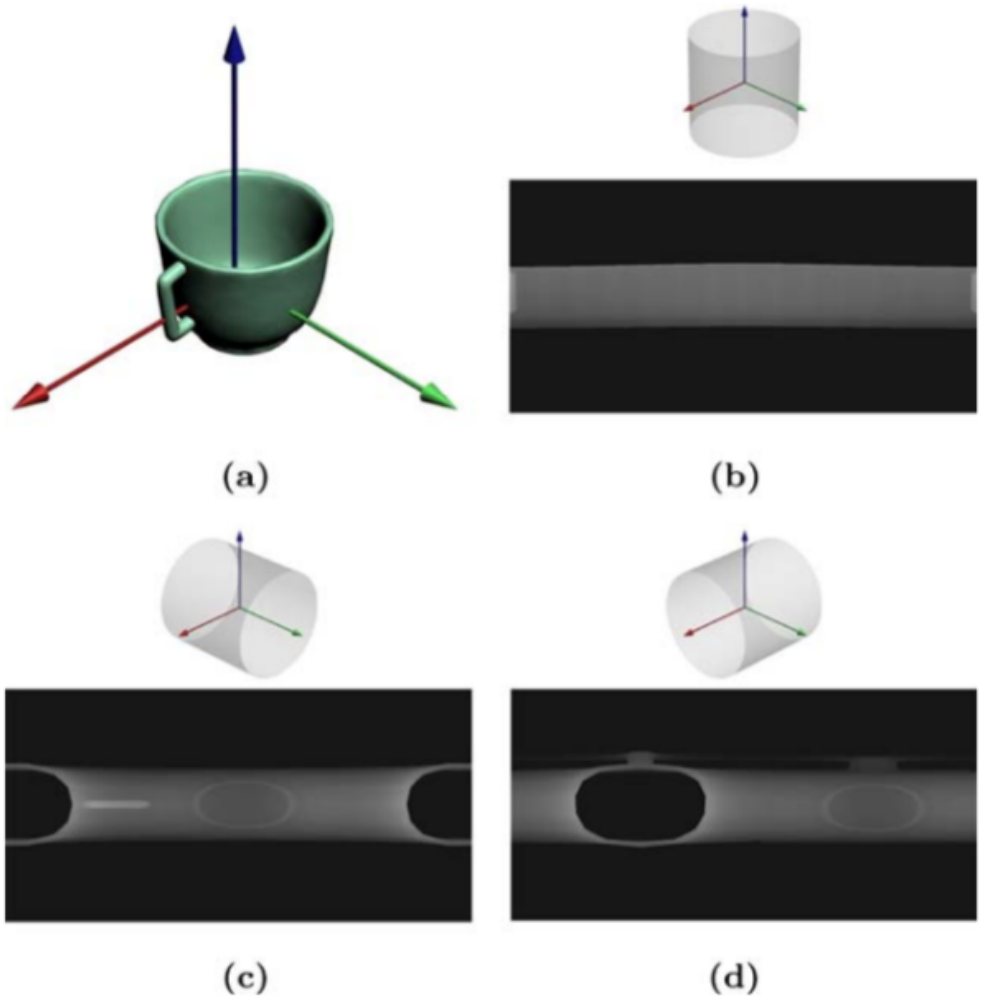


FIGURE 2.10 – Exemple de la projection $s_1(\phi_u, y_v)$ d'une tasse (a) sur les 3 cylindres orientés selon les axes x, y et z (b) - (d). Illustration extraite de [7]

Ensuite la transformée de Fourier discrète (DFT) 2D et la transformée en ondelettes discrète (DWT) 2D de chacune des 6 projections est calculée. La transformée de Fourier discrète 2D F est obtenue de la manière suivante :

$$F_{k,t}(m,n) = \sum_{u=0}^{2B-1} \sum_{v=0}^{B-1} s_{k,t}(\phi_u, y_v) e^{-2j\pi(\frac{mu}{2B} + \frac{nv}{B})} \tag{2.25}$$

avec $m \in [0, 2B-1]$ et $n \in [0, B-1]$. Le descripteur est épuré des valeurs centrales de l'image 2D formé par la transformée. Ces valeurs ne contiennent que peu d'informations. L'ensemble de transfor-

mées $\tilde{F}$ obtenu est :

$$s_F = (\tilde{F}_{1,x}, \tilde{F}_{2,x}, \tilde{F}_{1,y}, \tilde{F}_{2,y}, \tilde{F}_{1,z}, \tilde{F}_{2,z}) \tag{2.26}$$

La transformée en ondelettes discrète 2D W est la suivante :

$$W^{\phi}_{k,t}(j_0, m, n) = \frac{1}{\sqrt{2B * B}} \sum_{u=0}^{2B-1} \sum_{v=0}^{B-1} s_{k,t}(\phi_u, y_v)\phi_{j_0,m,n}(u, v)$$
$$W^{\psi}_{k,t}(j, m, n) = \frac{1}{\sqrt{2B * B}} \sum_{u=0}^{2B-1} \sum_{v=0}^{B-1} s_{k,t}(\phi_u, y_v)\psi_{j,m,n}(u, v)$$

avec $m \in [0, 2B-1]$ et $n \in [0, B-1]$ et $j \geq j_0$. A partir des images 2D formées par la DWT sont calculés la moyenne, l'écart-type et le coefficient d'asymétrie des images. Ces valeurs obtenues sont appelées $\tilde{V}_{k,t}$ et forment la base du descripteur s_W :

$$s_W = (\tilde{V}_{1,x}, \tilde{V}_{2,x}, \tilde{V}_{1,y}, \tilde{V}_{2,y}, \tilde{V}_{1,z}, \tilde{V}_{2,z}) \tag{2.27}$$

Deux alignements ont été produits au début, le descripteur pour un alignement l est $p_l = (s_{F,l}, s_{w,l}), l \in \{cpca, npca\}$ et le descripteur final P est $P = (p_{cpca}, p_{npca})$.

PANORAMA possède trois propriétés majeures :

— Invariance à la translation et à la rotation par la normalisation de pose

— La méthode est invariante à la mise à l'échelle en normalisant les coefficients des transformées

— Il a été démontré expérimentalement que la méthode est robuste au bruit

2.1.3.3 Shape Similarity System driven by Digital Elevation Models

La méthode Shape Similarity System driven by Digital Elevation Models (SSS-DEM) [8] s'inspire d'une représentation employée en cartographie, les modèles numeriques de terrain (Digital Elevation Model, DEM). Une DEM étant une représentation dans l'espace discret de l'élévation d'un terrain. Dans un premier temps l'objet 3D est projeté sur une sphère unitaire. La projection utilise l'opérateur de Laplace-Beltrami pour avoir une projection conforme, c'est-à-dire une projection qui conserve les angles. Le nuage de points de la sphère unité est ensuite projetée sur une grille 2D sphérique panoramique. La panoramique sphérique est une projection cylindrique dans le plan 2D d'une sphère.

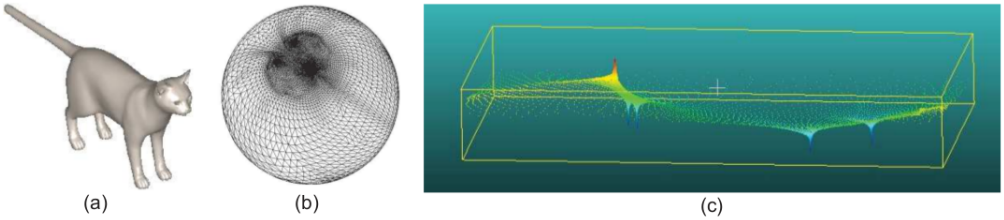


FIGURE 2.11 – Aperçu de la conception de DEM. (a) Maillage en entrée. (b) Projection du maillage sur la sphère unitaire. (c) DEM du maillage. Illustration extraite de [8]

Les coordonnées de la panoramique sont les coordonnées sphériques θ et ϕ de la sphère unitaire. Pour obtenir une panoramique rectangulaire de taille $[\theta_{min}, \theta_{max}] \times [\phi_{min}, \phi_{max}]$, un pas angulaire constant $d\theta$ et $d\phi$ est défini, de valeur $d\theta = d\phi = 0.01$. La valeur associée à chacun des points est la valeur de la coordonnée cartésienne de l'axe z du maillage en entrée. Pour une sphère unitaire P définie sur $\mathbb{R}^3$ et la panoramique m définie sur $\mathbb{R}^2$ alors la projection de la sphère unitaire à la panoramique sphérique est :

$$P(\theta_i, \phi_i, \rho_i) \rightarrow m(u_i, v_i) \tag{2.28}$$

avec $(\theta_i, \phi_i, \rho_i) = (\theta_i, \phi_i, 1)$ les coordonnées des points sur la sphère unitaire et (u_i, v_i) les coordonnées sur la panoramique sphérique. $i = 1...N_p$, avec N_p le nombre de points $p = (p_x, p_y, p_z)$ du maillage en entrée. Alors $m(u_i, v_i) = p_z$.

SSS-DEM possède les propriétés suivantes :

— Invariance à la translation et à la mise à l'échelle de part la projection sur une sphère unitaire centrée sur le centroïde de l'objet.

— Un temps rapide de comparaison et une taille de stockage faible

2.1.4 Descripteur 3D de Zernike

3D Zernike Descriptor (3DZD) [58] est la norme du vecteur des moments de Zernike. Les polynômes de Zernike 3D d'ordre n, de degré l et de répétition m sont les suivants :

$$Z_{nl}^m(r, \vartheta, \phi) = R_{nl}(r) Y_l^m(\vartheta, \phi) \tag{2.29}$$

(r, ϑ, ϕ) sont les coordonnées sphériques de la surface. On pose comme contrainte :

$$-l < m < l \tag{2.30}$$

$$0 \leq l \leq n \tag{2.31}$$

$$(n-l) = 2k \tag{2.32}$$

$R_n l(r)$ est le polynôme radial [70] et Y_l^m sont les harmoniques sphériques.

Un passage des coordonnées sphériques aux cordonnées cartésiennes X est appliqué en utilisant les polynômes harmoniques e_l^m :

$$Z_{nl}^m(X) = \sum_{v=0}^{k} q_{kl}^v |X|^{2v} e_l^m(X) \tag{2.33}$$

q_{kl}^v est défini de telle sorte que l'orthonormalité des fonctions dans la sphère unité soit vérifié.

Les moments de Zernike 3D Ω_{nl}^m d'une fonction $f(X)$ sont définis par :

$$\Omega_{nl}^m = \frac{3}{4\pi} \int_{|X|\leq 1} f(X) \bar{Z}_{nl}^m(X) dX \tag{2.34}$$

La norme des moments de Zernike 3D permet d'obtenir F_{nl} qui compose le vecteur du descripteur 3D de Zernike :

$$F_{nl} = ||\Omega_{nl}|| \tag{2.35}$$

Le descripteur 3D de Zernike est la combinaison pour chaque valeur de n et l de F_{nl}.

3DZD a différentes propriétés :

— L'utilisation de la norme permet l'invariance à la rotation

— L'invariance à la translation est obtenue en calculant le centre de gravité de l'objet

— L'invariance à la mise à l'échelle découle de la normalisation de l'objet dans un sphère unitaire

— L'utilisation d'un descripteur sous forme de vecteur permet une représentation compacte.

2.1.5 Caractérisations des descripteurs

Un descripteur peut être décrit comme global ou local. Si un descripteur encode l'information dans son entièreté ou une grande région de l'objet, le descripteur sera dit global. Au contraire, s'il encode

l'information du voisinage d'un point de l'objet, le descripteur sera défini comme local. Caractériser de cette manière les descripteurs permet de connaître les possibilités de comparaisons. Par exemple il est possible de faire de la reconnaissance partielle avec des descripteurs locaux. Un descripteur global est un descripteur ne prenant en général que peu d'espace de stockage et de temps de comparaison.

2.1.5.1 Descripteurs avec des caractéristiques globales

Un descripteur global peut être défini comme un descripteur contenant les informations de l'objet dans son entièreté. Les valeurs du descripteur sont affectées par n'importe quel changement ayant lieu sur la surface de l'objet.

Les valeurs propres de l'opérateur de Laplace-Beltrami sont considérées comme un descripteur global dû au fait que le spectre de l'opérateur de Laplace-Beltrami est défini à partir de l'ensemble des points formant un objet 3D.

ShapeDNA [52] et **Global Points Signature** [53] sont des méthodes employant le spectre de l'opérateur de Laplace-Beltrami comme descripteur et donc leur descripteur est un descripteur global.

PANORAMA [7] est basé sur la transformée discrète de Fourier et la transformée discrète d'ondelette. Ces deux transformées sont des descripteurs globaux tels que définis précédemment.

2.1.5.2 Descripteurs avec des caractéristiques locales

Les descripteurs locaux décrivent les sous-parties d'un objet. La valeur d'un descripteur local en un point est invariante ou très peu affectée par les perturbations sur la surface qui ne sont pas dans le voisinage de ce point.

Spin Image [55] produit un histogramme pour chaque point d'intérêt. Cet histogramme représente l'information spatiale de la relation entre le point d'intérêt et les points de son voisinage. Chaque point d'intérêt encode l'information spatiale locale autour de ce point d'intérêt ce qui fait de Spin Image une méthode utilisant un descripteur local.

Fast Point Feature Histogram (FPFH) [4] résume différentes informations géométriques autour d'un point en un histogramme. Les caractéristiques en un point étant calculées par rapport au voisinage de ce point, FPFH est une méthode avec un descripteur local.

2.1.5.3 Descripteurs avec des caractéristiques hybrides

Les descripteurs de caractéristiques globales et locales sont composés d'une partie décrivant la forme dans sa globalité qui peut être influencée par tous changements sur la surface et d'une autre partie affectée seulement par les changements du voisinage de la sous-partie que l'on souhaite décrire. Ce qui permet à un descripteur de passer de local à global est l'agrandissement de la région au point que cette région englobe la totalité de l'objet. C'est souvent à travers les paramètres que l'utilisateur définit si le descripteur est global ou local.

La méthode **Heat Kernel Signature** (HKS) peut être considérée comme un descripteur global ou local selon la valeur de la variable de temps de l'équation de chaleur [54]. Le HKS est une signature avec un vecteur pour chaque point de l'objet et dépendant du temps. Plus la valeur de la variable de temps est élevée, plus la signature de la méthode HKS décrit l'objet 3D dans sa globalité.

Wave Kernel Signature (WKS) a été présentée comme une méthode comblant les défauts de la méthode HKS. L'avantage par rapport à HKS est la distinction claire entre les caractéristiques globales et locales de la signature de la méthode WKS [2]. La signature varie selon l'énergie modélisée par l'équation utilisée par la méthode WKS. Une énergie élevée correspond à des caractéristiques locales tandis que une énergie faible est induite par la géométrie globale.

2.2 Méthodes de comparaison de protéines

Différentes méthodes de comparaison de structures de protéines ont été développées. Ces méthodes prennent en compte la séquence d'acides aminés et les structures secondaires. Ces méthodes peuvent aussi prendre en entrée les mutations évolutives de la protéine et des structures moléculaires tronquées d'un certain nombre d'atomes.

En alignant deux structures, il est possible de calculer la distance entre les deux structures. La distance obtenue permet de quantifier la ressemblance entre deux protéines pour ainsi les comparer.

2.2.1 Distance-matrix alignment

La méthode Distance-matrix ALIgnment (DALI) [71] [72] se base sur les matrices de distance 2D. DALI ne prend en compte que le carbone alpha (C_α) qui est l'un des carbones de l'acide aminé faisant partie du squelette de la protéine. La matrice de distance 2D représente la distance entre tous les

C_α de la protéine. L'avantage de cette représentation est de diminuer le temps de comparaison et de stockage des protéines.

DALI divise la protéine en morceaux de 6 acides aminés appelés hexapeptides. Ensuite une paire d'hexapeptides est sélectionnée dans la protéine A et une paire d'hexapeptides similaire est recherchée dans la protéine B. La chaîne d'acides aminés de la protéine *query* (ici A) est parcourue. La paire d'hexapeptides (a, b) de la protéine A est comparée à toutes les paires d'hexapeptides de la protéine B, puis la paire d'hexapeptides (b, c) de la protéine A est comparée à toutes les paires de la protéine B. Ce processus est répété jusqu'à avoir comparé toutes les paires d'hexapeptides consécutives de la chaîne d'acides aminés composant la protéine A.

Pour calculer la similarité entre deux sous-structures (paires d'hexapeptides) on utilise le score suivant :

$$S = \sum_{i=1}^{L} \sum_{j=1}^{L} \phi(i, j)$$

Avec i et j représentant chacun une paire d'acides aminés qui ont matché ensemble (par exemple $i = (i_A, i_B)$), L est la longueur de la sous-structure et ϕ est une mesure de similarité entre les acides aminés qui ont matché. Le but est de maximiser le score S.

La mesure de similarité proposée et appelée **score de similarité élastique** est la suivante :

$$\phi^E(i,j) = \begin{cases} (\theta^E - \frac{|d_{ij}^A - d_{ij}^B|}{d_{ij}^*}) w(d_{ij}^*), i \neq j \\ \theta^E, i = j \end{cases} \tag{2.36}$$

Avec d_{ij}^A et d_{ij}^B les valeurs entre le $i$ème résidu et le $j$ème résidu dans la matrice de distance de respectivement A et B. d_{ij}^* est la moyenne de d_{ij}^A et d_{ij}^B. θ^E est le seuil de similarité. $w(r) = \exp(-r^2/\alpha^2)$ est une fonction permettant de diminuer le poids des matchs entre paires éloignées et α est une constante calibrée sur la taille moyenne d'un domaine.

La deuxième étape est de rassembler les paires d'hexapeptides pour trouver l'alignement maximum. Pour faire ceci, la fonction de score S est maximisée. C'est un processus qui demande beaucoup de calcul. Pour gérer la complexité combinatoire de cette partie, un algorithme de Monte-Carlo est utilisé.

L'idée est de faire une marche aléatoire pour optimiser la fonction de score S. A chaque étape de la marche aléatoire, la probabilité de sélectionner la prochaine étape est défini par $p = \exp(\beta * (S' - S))$ avec S l'ancien score, S le nouveau score et β un paramètre.

Utiliser une marche aléatoire peut entraîner des maxima locaux mais permet un temps de maximisation de S significativement plus court.

2.2.2 Combinatorial Extension

La méthode Combinatorial Extension (CE)[73], utilise le même principe que DALI en calculant une matrice de similarité par rapport à la comparaison de sous-structures alignées appelées Aligned Fragments Pairs (AFP). Dans cette méthode les fragments sont une chaîne de 8 acides aminés. Les AFP sont une paire composée de 8 acides aminés venant d'une première protéine et une chaîne composée d'au moins 8 acides aminés venant d'une autre protéine. La paire d'acides aminés composant un AFP est alignée. Une chaîne peut être plus longue que 8 acides aminés dans l'éventualité où des *gap*, des trous, sont insérés dans la chaîne d'acides aminés pour avoir un meilleur alignement entre les deux fragments. Le nombre de *gap* est limité à 30 *gap* maximum par AFP. Le but est de trouver le chemin le plus long d'AFP.

La distance utilisée est la suivante :

$$\begin{aligned} D_{ij} = \frac{1}{m}(&|d^A_{p_i^A,p_i^A} - d^B_{p_i^B,p_i^B}|+ \\ &|d^A_{p_i^A+m-1,p_j^A+m-1} - d^B_{p_i^B+m-1,p_j^B+m-1}|+ \\ &\sum_{k=1}^{m-2} |d^A_{p_i^A+k,p_j^A+m-l-k} - d^B_{p_i^B+k,p_j^B+m-l-k}|) \end{aligned}$$

Avec p_i^A désignant la position du résidu à la position i de la protéine A, et de façon similaire pour la protéine B. D_{ij}^A représente la distance entre les résidus à la position i et j de la protéine A basé sur la coordonnée de l'atome C_α.

Trois contraintes sont données pour décider si l'on doit agrandir la séquence d'AFP consécutive, appelée chemin, pour obtenir l'alignement maximum. Les trois conditions sont les suivantes :

$$\begin{aligned} D_{nn} &< D_0 \\ \frac{1}{n-1}\sum_{i=0}^{n-1} D_{in} &< D_1 \\ \frac{1}{n^2}\sum_{i=0}^{n}\sum_{i=0}^{n} D_{ij} &< D_1 \end{aligned}$$

Avec D_{ij} la distance entre les AFP en position i et j dans le chemin et n la position de l'AFP à ajouter au chemin de taille $n-1$. D_0 et D_1 sont des seuils de similarité valant respectivement 3Å et 4Å.

Une optimisation est faite pour les protéines ayant un z-score faible (3.5). Parmi les 20 meilleurs chemins pour l'alignement de deux protéines, celui avec le meilleur RMSD (Root Mean Square Deviation) est choisi. Les *gap* sont relocalisés si une amélioration du RMSD est constatée. Ensuite en utilisant l'algorithme de programmation dynamique de **Needlman and Wunsch** [74], un meilleur alignement est recherché. L'utilisation de programmation dynamique permet de ne pas avoir un impact significatif sur le temps de calcul.

2.2.3 TM-align

Template Modeling Align (TM-align) [75] est une méthode d'alignement de protéine basée sur le Template Modeling Score (TM-score) décomposé en une matrice de similarité. Le TM-score a été proposé pour pallier au fait que le RMSD n'est pas indépendant de la longueur relative entre les deux protéines comparées.

Dans un premier temps, un alignement de structure initial est fourni. Trois méthodes d'alignement sont sélectionnées.

1) La première méthode d'alignement initial utilise l'algorithme d'alignement de Needleman et Wunsch [74] puis dans un second temps un algorithme d'alignement de structures secondaires fonctionnant de la manière suivante : un résidu d'une protéine va être assigné à l'une des trois structures secondaires (helice alpha, feuillet beta ou *coil*). Pour ce faire, la distance moyenne entre les résidus voisins d'une structure secondaire est comparée avec les distances entre les résidus de la protéine. Si la valeur de cette comparaison correspond à un profil de structure secondaire connue, il est classé comme tel. Ensuite, une matrice de score est créée avec pour valeur 0 ou 1 selon que les éléments alignés possèdent la même propriété de structure secondaire. Pour finir le chemin de cette matrice de score est maximisé.

2) Le second alignement a pour principe d'aligner la protéine sans utiliser de *gap* dans la plus petite des protéines. Le TM-score est utilisé pour décider du meilleur alignement.

3) Le troisième alignement est un hybride entre les deux alignements précédents, en combinant les

deux scores des alignements précédents.

Dans un deuxième temps une rotation est effectuée sur les structures [76] puis la matrice de score est calculée avec la formule suivante :

$$S(i,j) = \frac{1}{1 + d_{ij}^2 / d_0(L_{min})^2}$$

Avec d_{ij} est la distance entre le $i$ème atome C_α de la protéine 1 et le $j$ème atome C_α de la protéine 2. $d_0(L_{min}) = 1.24\sqrt[3]{L_{min} - 15} - 1.8$ avec L_{min} la longueur de la plus petite protéine.

Avec la nouvelle matrice un nouvel alignement est produit en utilisant un alignement par programmation dynamique. On répète l'étape de rotation et d'alignement jusqu'à ce qu'on atteigne une stabilité. Deux à trois itérations sont suffisantes pour trouver le meilleur alignement [75].

2.2.4 Deep-align

La méthode DeepAlign [77] utilise non seulement la proximité spatiale comme les méthodes précédentes mais aussi les liens d'évolution et les similarités entre les liaisons hydrogène.

Le score permettant de décider si deux résidus de deux protéines doivent être alignés est le suivant :

$$S(i,j) = (max(0, BLOSUM(i,j)) + CLESUM(i,j)) * v(i,j) * d(i,j)$$

BLOcks SUbstitution Matrix (BLOSUM) est une matrice donnant une valeur permettant d'évaluer la substitution d'un acide aminé par un autre lors d'une mutation dans un processus évolutif. Conformation LEtter SUbstitution Matrix (CLESUM) est une matrice donnant une valeur permettant d'évaluer la substitution d'une conformation autour d'un atome C_α par une autre lors d'une mutation dans un processus évolutif. $d(i,j)$ mesure la proximité spatiale entre deux résidus en utilisant le TMscore. $v(i,j)$ mesure la similarité de liaison hydrogène.

La méthode est définie en trois grandes étapes. La première étape est de détecter des groupes de paires d'acides aminés similaires (SFP pour Similar Fragment Pairs). Deux types de SFP sont définis. Les SFP courts composés de 6 à 8 acides aminés et les SFP longs composés de 9 à 18 résidus. Une similarité évolutive est recherchée pour définir les SFP. Le score de similarité est le suivant :

$$Similarity(i,j) = max(0, BLOSUM(i,j)) + CLESUM(i,j)$$

Avec un seuil de similarité d'au moins 0 pour les SFP courts et de 10 pour les SFP longs.

La deuxième étape est de minimiser le RMSD à l'intérieur d'un SFP en produisant des rotations sur les protéines. Puis un premier alignement est soumis en utilisant la programmation dynamique permettant de maximiser la fonction de score.

La troisième étape est, à partir de l'alignement initial fourni, d'utiliser la programmation dynamique pour maximiser la fonction de score S.

2.3 Méthodes de reconnaissance de formes appliquées aux protéines

Les méthodes d'alignement et de comparaison de protéines prennent en entrée principalement une représentation de la structure chimique de la protéine tandis que les méthodes de vision par ordinateur utilisent une représentation géométrique de l'objet. La comparaison de protéines reste un problème complexe, c'est pour cette raison qu'un intérêt grandissant pour cette problématique est apparu de la part du domaine de la vision par ordinateur. Récemment des méthodes de vision par ordinateur ont été testées sur des jeux de données de protéines comme lors du challenge SHREC sur la reconnaissance de protéines [78] [79] [80] [81] [82]. Des méthodes ont aussi été proposées spécifiquement pour répondre à ce problème comme 3D Zernike Descriptor (3DZD) [59] un descripteur basé sur les polynômes de Zernike ou MoleculAr Surface Interaction Fingerprint (MaSIF) [83], une méthode de *deep learning* se basant sur les caractéristiques géométriques et chimiques de la protéine.

2.3.1 Descripteur 3D de Zernike appliqué aux protéines

3D Zernike Descriptor (3DZD) appliqué aux protéines [59][84] est le même que celui défini précédemment [58] qui est la norme du vecteur des moments de Zernike. Les polynômes de Zernike 3D d'ordre n, de degré l et de répétition m tel que défini précédemment sont :

$$Z_{nl}^{m}(r,\vartheta,\phi) = R_{nl}(r)Y_{l}^{m}(\vartheta,\phi) \tag{2.37}$$

Les moments de Zernike 3D Ω_{nl}^m d'une fonction $f(X)$ sont définis par :

$$\Omega_{nl}^m = \frac{3}{4\pi} \int_{|X|\leq 1} f(X) \bar{Z}_{nl}^m(X) dX \tag{2.38}$$

La norme des moments de Zernike 3D permet d'obtenir F_{nl} qui est le descripteur 3D de Zernike :

$$F_{nl} = \sqrt{\sum_{m=-l}^{m=l} (\Omega_{nl}^m)^2} \tag{2.39}$$

Ce descripteur a été adapté pour comparer des protéines entres elles, en particulier sur des protéines avec une structure de faible résolution. En associant les propriétés physico-chimiques des protéines au descripteur 3D de Zernike, des régions de la protéine peuvent être comparées. Ceci permet de comparer les poches d'une protéine à des poches de liaison de ligand connues pour ainsi découvrir de nouveaux domaines de liaison de ligand. En étendant ce principe à une comparaison globale, il est possible de prédire un docking protéine-protéine en utilisant la complémentarité de surface.

2.3.2 MaSIF

Molecular Surface Interaction Fingerprint (MaSIF) [83] est une méthode basée sur le *deep learning* de comparaison et de docking de protéines.

La première étape de la méthode est de calculer la Surface Exclue au Solvant (SES) sous forme de maillage et d'associer à chaque vertex (point) des caractéristiques géométriques et chimiques. Les caractéristiques géométriques sont le *shape index* et la courbure dépendante de la distance. *Shape index* est une valeur dépendant des courbures principales κ_1 et κ_2, avec $\kappa_1 \geq \kappa_2$:

$$\frac{2}{\pi} \tan^{-1} \frac{\kappa_1 + \kappa_2}{\kappa_1 - \kappa_2} \tag{2.40}$$

La courbure dépendante de la distance k_{ij} pour deux vertex v_i et v_j est obtenue de la manière suivante :

$$k_{ij} = \Theta(|r_j + n_j - r_i - n_i| - d_{ij}) \frac{|n_j - n_i|}{d_{ij}} \tag{2.41}$$

n_i, n_j, r_i et r_j sont respectivement les normales et les coordonnées de v_i et v_j. Θ est une fonction escalier, $d_{ij} = |rj - ri|$ est la distance entre v_i et v_j.

Les caractéristiques chimiques sont l'indice d'hydropathie, le continuum électrostatique et les protons/électrons libres.

L'indice d'hydropathie[85] est une échelle dont le but est de quantifier l'hydrophilie d'un acide aminé.

Le continuum électrostatique est calculé à partir de l'équation de Poisson-Boltzmann définie comme :

$$\Delta V(\overrightarrow{r}) - \frac{2ec_S^0}{\epsilon_d}\sinh(\frac{eV(\overrightarrow{r})}{k_BT}) = 0 \tag{2.42}$$

$V(\overrightarrow{r})$ est le potentiel électrique, e la charge d'un électron, c_S est la concentration en ions de la solution, ϵ_d est la permittivité diélectrique du solvant, k_B est la constante de Boltzmann et T la température.

Les protons et électrons libres donneurs représentent les liaisons hydrogènes possibles. Les atomes possiblement donneurs ou receveurs sont détectés. Une valeur définissant si le vertex est donneur ou receveur est attribuée.

En parallèle du calcul des caractéristiques, des patchs chevauchants sont extraits de la surface de la protéine. Les coordonnées sont des coordonnées géodésiques et polaires définies par rapport au centre du patch avec un rayon géodésique r.

K rotations sont appliquées sur le patch pour permettre l'invariance à la rotation. On applique des couches de convolution géodésique [86] sur les K rotations et une couche finale donnant le maximum parmi les K rotations. Cette procédure est appliquée pour chacun des patchs et l'ensemble de ces résultats donne l'empreinte (*fingerprint*) de la protéine selon la méthode MaSIF.

L'empreinte de MaSIF peut être ensuite exploitée ainsi ou transformée par un réseau de neurones additionnel pour la prédiction de sites de liaison de ligand, de sites d'interactions protéine-protéine et d'interactions protéine-protéine.

2.4 Mesures d'évaluation

2.4.1 Nearest Neighbor, First Tier et Second Tier

Ces trois métriques sont définies dans le cas où tous les objets du jeu de données sont comparés entre eux. L'identité, c'est à dire le match de l'objet *query* avec lui même, est ignoré lors des évaluations avec ces métriques car les méthodes de l'état de l'art retrouvent systématiquement l'identité.

Le Nearest Neighbor (NN) est le pourcentage de matchs le plus proche appartenant à la même classe. Comme expliqué dans le paragraphe précédent, le NN exclu l'identité et est donc calculé parmi le second meilleur match.

Le First Tier (FT) est le pourcentage d'objets appartenant à la même classe que l'objet *query* parmi ceux classés dans les K_1 premiers. Si la taille de la classe de l'objet *query* est C alors $K_1 = C - 1$. La valeur K_1 n'est pas égal à C car l'identité n'est pas prise en compte. Le FT représente le rappel (ou sensibilité) avec K_1 la plus petite valeur permettant au FT de potentiellement atteindre 100%. Le Second Tier (ST) est similaire au FT mais est plus permissif. Cette mesure représente le pourcentage d'objets appartenant à la même classe que l'objet *query* parmi les K_2 premiers objets du classement avec $K_2 = 2 * (C - 1)$.

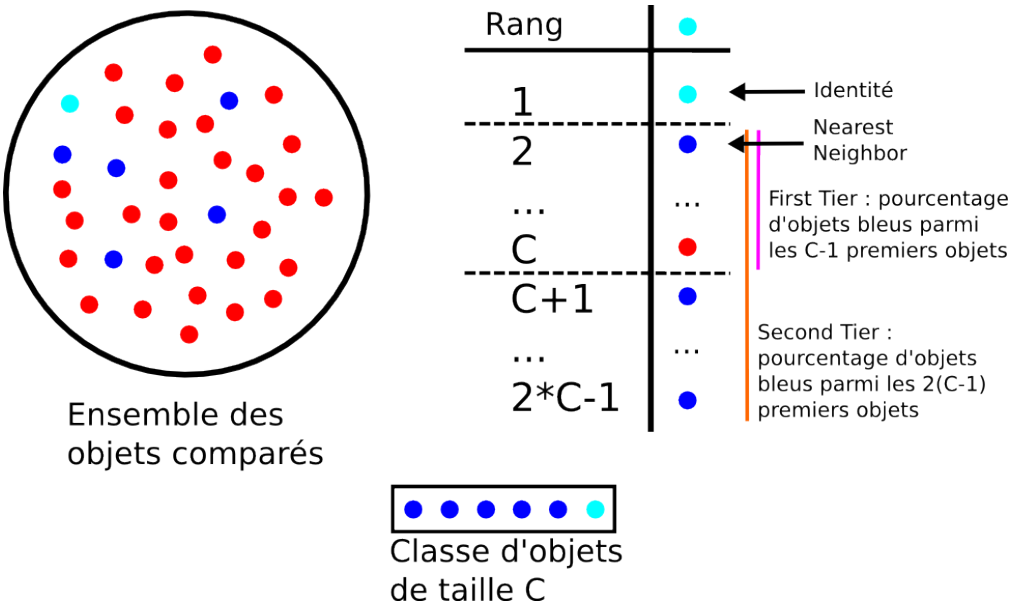


FIGURE 2.12 – Schéma de classification d'un objet et du Nearest Neighbor, First Tier et Second Tier

2.4.2 La courbe de précision-rappel

La courbe de précision-rappel trace la valeur de précision en fonction du rappel aussi appelé sensibilité. Le rappel varie en fonction de la limite entre les objets classés comme positifs et ceux classés comme négatifs (cf Figure 2.13). Cette limite est la valeur K qui est comprise entre 1 et N, N étant le nombre d'objets comparés. La valeur K est le K-ème objet parmi les N objets classés par ordre de similarité par la méthode et pour un objet *query*. C'est en faisant varié la valeur K et par interpolation qu'il est possible d'obtenir des valeurs de rappel régulières. La précision est ensuite calculée pour chacune des valeurs de K déterminées.

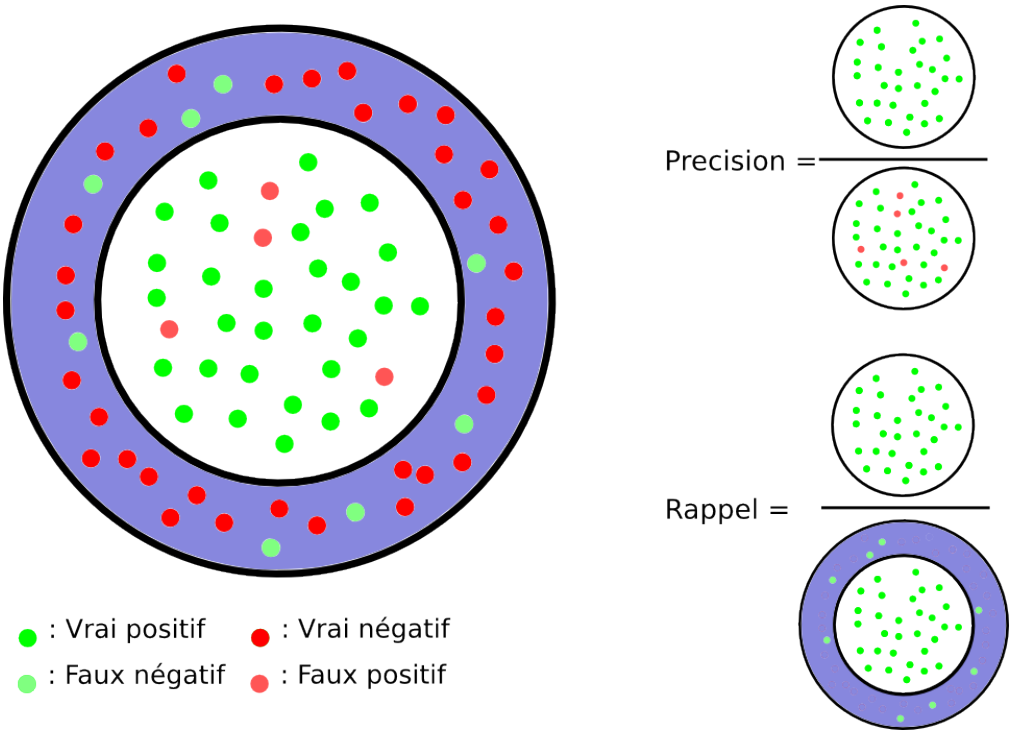


FIGURE 2.13 – Schéma du calcul de la précision et du rappel.

2.5 Comparaison de méthodes de l'état de l'art

2.5.1 Résumé

Dans cet article, différentes méthodes de l'état de l'art de la vision par ordinateur et de la biologie structurale sont comparées. Les méthodes de reconnaissance de formes utilisé sont **3D-SURFER**, **PANORAMA**, **ShapeDNA**, **Viewpoint Feature Histogram** (VFH). Ces méthodes sont des

descripteurs globaux, ce qui permet une comparaison rapide mais plus difficilement adaptable à une comparaison partielle. Les méthodes d'alignement de structures de protéines sont **Combinatorial Extension** (CE), **DeepAlign** et **TM-align**. Deux jeux de données sont utilisés, le premier jeu de données $\mathcal{A}$ est composé du jeu de données utilisé dans l'article *SHREC19 Protein Shape Retrieval Contest* [81] composé de 5298 conformations de protéines. Le second jeu de données $\mathcal{B}$ est composé de 16 structures de protéines [87]. Le but étant d'évaluer les méthodes de vision par ordinateur sur un jeu de données composé de protéines avec pour références les méthodes de comparaison de structures de protéines.

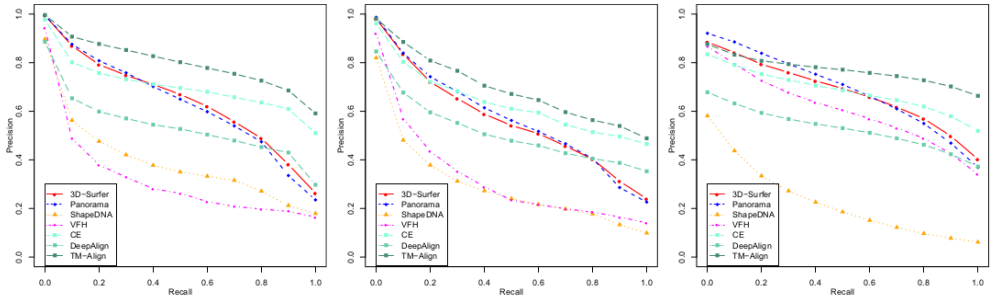
Les métriques d'évaluation utilisées sont le Nearest Neighbor, le First Tier, le Second Tier, le Mean Average Precision et la courbe de précision-rappel ainsi que l'identité de séquence. De plus le temps de calcul est pris en compte pour déterminer les méthodes permettant un traitement haut débit.

Les courbes de précision-rappel (cf Figure 2.15) de 3D-surfer et PANORAMA suivent la même tendance. C'est également le cas pour ShapeDNA et VFH. Les valeurs d'évaluations (cf Figure 2.14) sont élevées pour 3D-Surfer et PANORAMA et un peu plus faible pour Shape-DNA et VFH. Les méthodes de comparaison de structure CE et TM-Align ont un NN, FT, ST et MAP similaires à 3D-Surfer et PANORAMA. DeepAlign possède des mesures d'évaluation plus proche de VFH et ShapeDNA. Les courbes de précision-rappel diminuent plus lentement en fonction du rappel pour les méthodes d'alignements de structures.

Les courbes de précision-rappel sont dominées par celles de TM-Align. Le NN est le plus élevé pour des méthodes de comparaison de surfaces, 3D-surfer pour les classes au niveau hiérarchique de la protéine et de l'espèce et PANORAMA au niveau du domaine. Les autres mesures d'évaluation sont le plus élevées pour TM-Align tous niveaux hiérarchiques de classes confondus.

Method	Hierarchy	NN	FT	ST	MAP
3D-Surfer	Protein	**0.993016**	0.591094	0.725946	0.65347
	Species	**0.979237**	0.565988	0.656099	0.593331
	Domain	0.791808	0.688918	0.839672	0.720762
Panorama	Protein	0.988109	0.576079	0.7161	0.629772
	Species	0.976972	0.540468	0.64445	0.565733
	Domain	**0.805776**	0.647282	0.792324	0.683539
Shape-DNA	Protein	0.81578	0.34838	0.534274	0.365649
	Species	0.709702	0.272602	0.420454	0.270242
	Domain	0.41544	0.235546	0.321569	0.21337
VFH	Protein	0.899962	0.27136	0.44334	0.287025
	Species	0.880143	0.289058	0.413764	0.301598
	Domain	0.787844	0.572687	0.689267	0.59925
CE	Protein	0.95319	0.675288	0.828753	0.695678
	Species	0.939977	0.598196	0.692926	0.625466
	Domain	0.740091	0.647606	0.780736	0.675805
DeepAlign	Protein	0.677803	0.513238	0.679353	0.515659
	Species	0.662703	0.485843	0.623079	0.489423
	Domain	0.447339	0.480719	0.665951	0.50041
TM-Align	Protein	0.990751	**0.74868**	**0.896931**	**0.792712**
	Species	0.97282	**0.648212**	**0.746878**	**0.684902**
	Domain	0.797093	**0.732519**	**0.85881**	**0.758466**

FIGURE 2.14 – Mesures d'évaluations les méthodes de reconnaissance de formes (3D-surfer, Panorama, ShapeDNA et VFH) et les méthodes d'alignements de structures (CE, DeepAlign et TM-Align) pour différents niveaux hiérarchique de classe sur le jeu de données $\mathcal{A}$.



(a) *Protein* hierarchical level (b) *Species* hierarchical level (c) *Domain* hierarchical level

FIGURE 2.15 – Courbe de précision-rappel pour les méthodes de reconnaissance de formes (3D-surfer, Panorama, ShapeDNA et VFH) et les méthodes d'alignements de structures (CE, DeepAlign et TM-Align) pour différents niveaux hiérarchique de classe sur le jeu de données $\mathcal{A}$.

Les performances sont globalement supérieures pour 3D-Surfer pour l'ensemble des valeurs d'évaluation des méthodes de vision par ordinateur. 3D-surfer est la seule méthode parmi les 4 méthodes de reconnaissance de formes à avoir été développée pour la comparaison de surfaces de protéines. PANORAMA possède des valeurs d'évaluation proches de 3D-Surfer bien que cette méthode n'est pas été développée pour la comparaison de surfaces de protéines. Les valeurs d'évaluation de ces deux méthodes sont aussi similaires aux valeurs d'évaluation des méthodes de comparaison de structures. TM-align domine les différentes mesures d'évaluation à l'exception du NN.

Bien que les méthodes de comparaison de structure de protéines aient des performances en moyenne légèrement supérieures aux méthodes de reconnaissance de formes, ces dernières ont permis de trouver 6 protéines similaires de par leur surface malgré une identité de séquence faible.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Structural Bioinformatics

Comparative Evaluation of Shape Retrieval Methods on Macromolecular Surfaces: An Application of Computer Vision Methods in Structural Bioinformatics.

Mickael Machat [1], Florent Langenfeld [1], Daniela Craciun [1], Léa Sirugue [1], Taoufik Labib [1], Nathalie Lagarde [1], Maxime Maria [2,1] and Matthieu Montes [1,*]

[1]Laboratoire GBCM, EA 7528, Conservatoire National des Arts et Métiers, Hesam université, 2 rue Conté, 75003 Paris, France
[2]Laboratoire XLIM, UMR CNRS 7252, Université de Limoges, 123 avenue Albert Thomas, 87000 Limoges, France

*To whom correspondence should be addressed

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The investigation of the structure of biological systems at the molecular level gives insight about their functions and dynamics. Shape and surface of biomolecules are fundamental to molecular recognition events. Characterizing their geometry can lead to more adequate predictions of their interactions. In the present work, we assess the performance of reference shape retrieval methods from the computer vision community on protein shapes.
Results: Shape retrieval methods are efficient in identifying orthologous proteins and tracking large conformational changes. This work illustrates the interest for the protein shape as a higher-level description of the protein structure that 1) abstracts the underlying protein sequence, structure or fold, 2) allows the use of shape retrieval methods to screen large database of protein structures to identify surficial homologs and possible interacting partners 3) opens an extension of the protein structure-function paradigm towards a protein structure-surface(s)-function paradigm. **Availability:** All data are available online at http://datasetmachat.drugdesign.fr
Contact: Prof. Matthieu Montes. Mail: matthieu.montes@cnam.fr. Phone: (33) 1 40 27 28 09.
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Proteins are macromolecules involved in most biological processes that interact through their molecular surfaces [1]. The protein molecular surface representation is considered as a higher-level description of the protein structure [2] that abstracts the underlying protein sequence, structure and fold into a continuous shape with geometric and chemical features that fingerprint its interactions [3, 4]. Functionally related proteins often share similar surface properties despite a potentially low sequence and/or backbone conformation similarity [5, 6], allowing the identification of distant homologs or interacting partners using their surface [4, 6]. Since protein shape is a major descriptor of protein molecular surface, comparing protein structures using their shapes allows to apply the shape retrieval methods that have been extensively used in the computer vision field, notably for military, civil security or medical imaging applications [7].

Different shape retrieval methods have been developed to compare the shape of proteins and ligands [4, 8, 9, 10, 11, 12]. The most recent method is a local shape comparison method based on geometric deep learning to generate a molecular surface interaction fingerprint [4]. The evaluation of the performance of shape retrieval methods in the literature is performed classically during the SHREC community benchmark [13] where joint efforts between the structural bioinformatics and the computer vision communities have been performed to develop benchmarking datasets on protein shapes [14, 15, 16, 17, 18].

In this work, we evaluate the performance of four different shape retrieval methods (3D-Surfer [19], Panorama [20], ShapeDNA [21] and VFH [22]) on the complete cross-comparison of the SHREC 2019 protein

shapes benchmarking dataset (5298 shapes) [17]. These methods have shown top performances on non-protein shapes benchmarks [23, 24, 25, 26]. As a reference, we include different protein structure comparison methods (CE [27], DeepAlign [28], and TM-Align [29]).

2 Methods

2.1 Datasets

Set $\mathcal{A}$ has been designed for the evaluation of the performance of shape retrieval methods on protein shapes for the community benchmark SHREC 2019 [17]. The dataset comprises 5298 experimental conformations of protein domains extracted from 211 PDB entries resolved by NMR. The dataset classification relies on the Structural Classification of Proteins-extended (SCOPe) database [30, 31]. The lowest hierarchical level –called *Domain* hierarchical level– links the SCOPe database to the Protein Data Bank (PDB) [32]. The following inclusion procedure was applied on all SCOPe entries. A PDB structure was included if 1) its conformers display the same number of atoms, 2) it belongs to the α, $\alpha + \beta$ or α/β structural classes of the SCOPe database, 3) at least four orthologous protein structures exist and satisfy the previous inclusion rules. A total of 211 PDB entries satisfying all these criteria were selected and assigned to 17 classes (following the SCOPe *Protein* hierarchical level).Apart from the *Protein* hierarchical level, the dataset contains sub-classes along two hierarchical sublevels. The *Species* hierarchical level contains 54 classes corresponding to the different species. The *Domain* hierarchical level is composed of 241 classes corresponding to the initial SCOPe classification. For each structure of the dataset, the solvent excluded surface (SES) [1] was computed using EDTSurf [33] with default parameters.
Set $\mathcal{B}$ consists in 16 protein structures that were studied in [5]. The following protein couples in set $\mathcal{B}$ display high surface shape similarity and low sequence identity: 1jzn_A - 1g1q_A, 1bar_A - 1rro, 1ryp_B - 1gwz, 1a31 - 1cy0, 1tbp - 1t7p, 1b3t - 1adv, 2nwl - 2bbh and 2b2i - 2cfp where the first four characters correspond to the PDB ID and the last character (if present) corresponds to the chain ID.

2.2 Shape retrieval methods

In **3D-Surfer**, the protein global surface information is represented with 3D Zernike Descriptors (3DZD), mathematical moment-based invariants of 3D functions [5]. The molecular surface of the protein is triangulated using MSROLL [34] and mapped onto a 3D cubic grid from which 3DZD descriptors are calculated for each protein. The similarity between two given protein surfaces is quantified by the Euclidean distance between their two respective descriptors. 3D-Surfer is available on-line [19].

In Panorama [20], the panoramic views are acquired through cylindrical projections, in order to capture two characteristics of the 3D model's surface; 1) the position of the model's surface in 3D space and 2) the orientation of the model's surface. The feature extraction relies on the use of 2D Discrete Fourier Transform and 2D Discrete Wavelet Transform. Once the descriptor is extracted for each object of the dataset, the Manhattan and the Canberra distances are used to compare the overall similarity between two descriptors.

ShapeDNA is a framework [35, 21] that provides an effective solution for invariance to non-rigid deformations [24]. In Shape DNA, the Laplace-Beltrami operator is applied on the molecular surface. The resulting spectrum (eigenvalues) represent an isometry invariant numerical fingerprint of the molecular surface. The similarity between two given protein surfaces is quantified by comparing their spectra using the Euclidean distance.

The **Viewpoint Feature Histogram (VFH)** [22] is a geometric descriptor invariant to scale [36]. In VFH, a two-components descriptor is calculated from the normal at each point of the point cloud of the molecular surface shape and the normal of the centroid of the point cloud of the molecular surface. VFH is available in the PCL library [37].

2.3 Protein structure comparison methods

CE (Combinatorial Extension) [27] represents proteins as a set of octameric fragments. Each pair of octameric fragments that can be aligned within a given threshold is considered an aligned fragment pair (AFP). CE uses a combinatorial extension algorithm to identify and combine the most similar AFPs between the compared structures. A Z-score is computed for the final alignment using a reference set of alignments [38].

DeepAlign [28] performs automatic pairwise protein structure alignment using evolutionary relationships and hydrogen-bonding similarity, in addition to spatial proximity of equivalent residues. The scoring function is composed of amino acid mutation score, local substructure substitution potential, hydrogen-bonding similarity and geometric similarity.

TM-Align [29] identifies the best structural alignment between protein pairs independently from their sequences. It first generates optimized residue-to-residue alignment based on structural similarity using heuristic dynamic programming iterations. Then, the scoring function TM-score [39] is used to scale the structural similarity. TM-score outputs a score s in $(0, 1]$, where 1 indicates a perfect match between two structures. Output Scores below 0.2 usually correspond to unrelated proteins, while those higher than 0.5 assume generally the same fold in SCOP/CATH [40, 41].

2.4 Shape retrieval performance evaluation

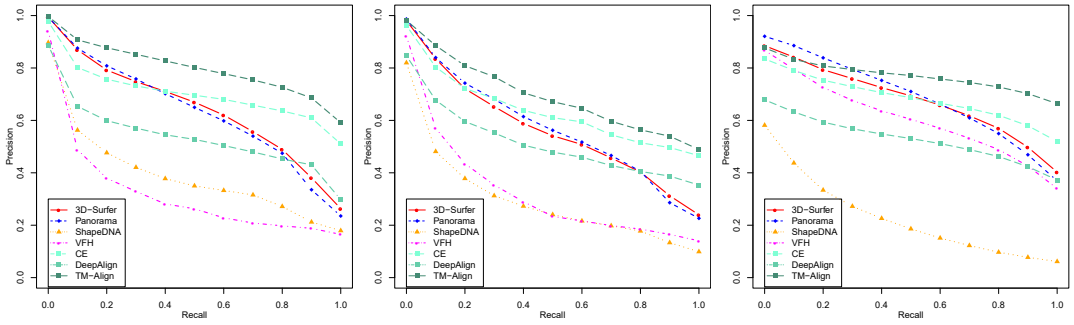
The performances in retrieval of each method were evaluated using Precision-Recall, Nearest Neighbor (NN), First-tier (FT), Second-tier (ST), and Mean Average Precision (MAP).

2.5 Runtime

To evaluate the computation time of the shape retrieval methods, we considered the sum of the runtimes required to compute 1) the method's descriptor for the biggest mesh, 2) the method's descriptor for the smallest mesh, and 3) the distance between these two descriptors. The biggest and smallest meshes contained respectively 271,854 and 53,696 vertices. All calculations were performed on an Intel Core i7-6700HQ CPU@2.60 GHz with 32 GB of RAM except for the 3D-Surfer web-service for which we used the mean runtimes reported in the original publication [5].

2.6 Identifying distant surficial homologs

In order to identify hits with low sequence identity but similar surface shapes (*i.e* distant surficial homologs), we compared the dissimilarity matrices to a sequence aligning matrix. Clustal Omega [42] was used to align all the sequences of set $\mathcal{A}$. A homology matrix $H_{N,N} = 5298$ was constructed, such as $H_{[i,j]},(i,j) \in [\![1, 5298]\!]^2$ enrolls the sequence identity ratio between protein i and protein j. Then, the dissimilarity matrix output M_N^k, $k = 1, 2, 3, 4$, of each method was normalized, and we computed $H + M_N^k$. For each method, $\min_j H + M_N^k$ represents the protein target j combining the least sequence identity and the highest shape similarity for protein query i. Afterwards, $\min_j H + M_N^k$ for each method k was compared. If at least two methods k_1 and k_2 bring the same $\min_j H + M_N^k$, the protein pair i and j are considered distant surficial homologs.



(a) *Protein* hierarchical level (b) *Species* hierarchical level (c) *Domain* hierarchical level

Fig. 1: Precision-Recall plots for the shape retrieval methods (3D-Surfer, Panorama, ShapeDNA, VFH) and structure comparison methods (CE, DeepAlign, TM-Align) over the different hierarchical levels of set $\mathcal{A}$.

3 Results

First, we present and compare the performances in retrieval of the shape retrieval methods (3D-Surfer, Panorama, ShapeDNA, VFH) on the hierarchical protein shapes set $\mathcal{A}$. Then, we illustrate their performance in 1) tracking conformational changes on selected subsets and 2) identifying distant surficial homologs (low sequence identity with high shape similarity). The performance of widely used structure comparison methods (CE, DeepAlign, TM-Align) is presented as a reference.

3.1 Protein shape retrieval

Table 1 summarizes the quantitative statistics values for each method on the three hierarchical levels, *Protein*, *Species*, and *Domain*. Figure 1 presents the precision-recall curves for each method and each hierarchical level.

3.1.1 *Protein* hierarchical level

The precision-recall curves in Figure 1a show similar performances between 3D-Surfer and Panorama, and between ShapeDNA and VFH. The high performances of 3D-Surfer and Panorama are corroborated by Table 1, where the NN statistics display values greater than 0.98 and MAP statistics greater than 0.5 for both methods. For FT and ST statistics, 3D-Surfer and Panorama surpass ShapeDNA and VFH as well. 3D-Surfer and Panorama outperform the structure comparison methods CE and DeepAlign for recall values below 0.5. In particular, 3D-Surfer displays the best NN followed by TM-Align, Panorama, CE, VFH, ShapeDNA, and DeepAlign, respectively.

3.1.2 *Species* hierarchical level

As the hierarchical level goes down –from the *Protein* hierarchical level to the *Species* hierarchical level, the orthologous proteins are separated into disjoint classes. The overall performances decrease for the shape retrieval methods, except for VFH in FT and MAP (Table 1, Figure 1b). For the NN performance metric, as previously observed in the *Protein* hierarchical level, the best performances are associated with 3D-Surfer, followed by Panorama, TM-Align, CE, VFH, ShapeDNA, and DeepAlign, respectively.

3.1.3 *Domain* hierarchical level

In this hierarchical level, the classes are the least populated (from 2 to 160 protein objects). We observe a flattening of the precision-recall curves (Figure 1c) for the shape retrieval methods, except for ShapeDNA whose performances are in decay *w.r.t* the *Species* and the *Protein* hierarchical level. Regarding the other methods, only the NN statistic drops down compared to the higher hierarchical levels, with Panorama displaying the best value, followed by TM-Align, 3D-Surfer, VFH, CE, DeepAlign, and ShapeDNA, respectively. Except for ShapeDNA, all methods displayed increased performances in retrieval on this level in the FT, ST and MAP compared to the *Species* hierarchical level.

Table 1. Retrieval statistics computed for each method and each hierarchical level of set $\mathcal{A}$. Bold numbers represent the best value in each category.

Method	Hierarchy	NN	FT	ST	MAP
3D-Surfer	Protein	**0.993016**	0.591094	0.725946	0.65347
	Species	**0.979237**	0.565988	0.656099	0.593331
	Domain	0.791808	0.688918	0.839672	0.720762
Panorama	Protein	0.988109	0.576079	0.7161	0.629772
	Species	0.976972	0.540468	0.64445	0.565733
	Domain	**0.805776**	0.647282	0.792324	0.683539
Shape-DNA	Protein	0.81578	0.34838	0.534274	0.365649
	Species	0.709702	0.272602	0.420454	0.270242
	Domain	0.41544	0.235546	0.321569	0.21337
VFH	Protein	0.899962	0.27136	0.44334	0.287025
	Species	0.880143	0.289058	0.413764	0.301598
	Domain	0.787844	0.572687	0.689267	0.59925
CE	Protein	0.95319	0.675288	0.828753	0.695678
	Species	0.939977	0.598196	0.692926	0.625466
	Domain	0.740091	0.647606	0.780736	0.675805
DeepAlign	Protein	0.677803	0.513238	0.679353	0.515659
	Species	0.662703	0.485843	0.623079	0.489423
	Domain	0.447339	0.480719	0.665951	0.50041
TM-Align	Protein	0.990751	**0.74868**	**0.896931**	**0.792712**
	Species	0.97282	**0.648212**	**0.746878**	**0.684902**
	Domain	0.797093	**0.732519**	**0.85881**	**0.758466**

3.1.4 Runtime

The runtimes are presented in Table 2, with a distinction between the runtimes for the descriptors calculation and the runtimes for the descriptors comparison. In total runtime, we observe that VFH is the fastest, followed by Panorama, 3D-Surfer and ShapeDNA, respectively. The structure comparison methods mean computation times for CE, DeepAlign, and TM-Align are $2s$, $0.2s$, and $0.1s$ respectively.

Table 2. Runtimes of the evaluated shape retrieval methods in seconds to compute the descriptor for the largest mesh, the descriptor for the smallest mesh, and the distance between the two descriptors.

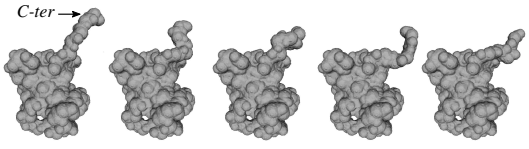
Method	3D-Surfer	Panorama	Shape-DNA	VFH
Largest mesh	3.48	5.05	19.05	2.31
Smallest mesh	3.48	1.04	2.98	0.44
Distance	0.06	0.27	0.02	0.03
Total	7.02	6.36	22.05	2.78

3.2 Tracking conformational changes

In order to evaluate if conformational changes affecting protein shapes could be tracked by shape retrieval methods, we identified two proteins in set $\mathcal{A}$ displaying low and high amplitude conformational changes, respectively.

Human ubiquitin displays a very flexible C-terminal (Figure 2) that modifies a subpart of its global shape. We focused on the retrieval results for this class (class 2n2k_A), to see if the shape retrieval methods are able to rank the 70 conformers of the human ubiquitin available in set $\mathcal{A}$ as the first 70 matches (Table 3, Supplementary Figure S1). All shape retrieval methods retrieved on average more than 67 conformers within the top 70 for each of the 70 queries of class 2n2k_A (69.12, 67.04, 69.72 and 70 for 3D-Surfer, Panorama, ShapeDNA and VFH respectively).

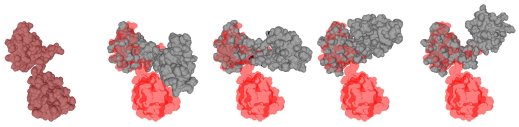
Using structure comparison methods, none of the human ubiquitin conformers were ranked within the first 70 matches with CE and DeepAlign for each query of class 2n2k_A except the identity. TM-Align retrieved on average 24.91 conformers. Structure comparison methods mostly retrieved conformations of class 2n2k_B instead, another protein class containing the human ubiquitin structure without the five C-terminal residues.



(a) 4 (b) 24 (c) 30 (d) 54 (e) 70

Fig. 2: Illustration of the mobility of the C-terminal (*C-ter*) of human ubiquitin (PDB ID 2n2k, chain A). The numbers correspond to the conformer number in the PDB entry 2n2k.

***Xenopus laevis* calmodulin** is composed of two domains linked by a three-residue coil (Figure 3) that allows an ample motion of one domain *w.r.t* the other (Supplementary Table S1), resulting in very different conformations (class 1dmo_A, 30 conformers). In order to investigate whether the selected shape retrieval methods are able to track these high amplitude non-rigid transformations, we enumerated the number of 1dmo_A conformers retrieved for each query within the first 30 retrieved shapes (Table 3, Supplementary Figure S1). Shape retrieval methods retrieved on average at least 7.13 conformers within the top 30 for each of the 30 queries (9.60, 10.70, 7.13 and 14.46 for 3D-Surfer, Panorama, ShapeDNA and VFH, respectively). Structure comparison methods retrieved on average less than 5.86 conformers for each query (2.3, 5.86 and 2.03 for CE, DeepAlign and TM-Align, respectively).



(a) 17 (b) 12 (c) 18 (d) 22 (e) 24

Fig. 3: Illustration of the conformational changes of *the Xenopus laevis* calmodulin (PDB ID 1dmo, chain A). The numbers correspond to the conformer number in the PDB enry 1dmo. (b), (c), (d) and (e) are superimposed on (a).

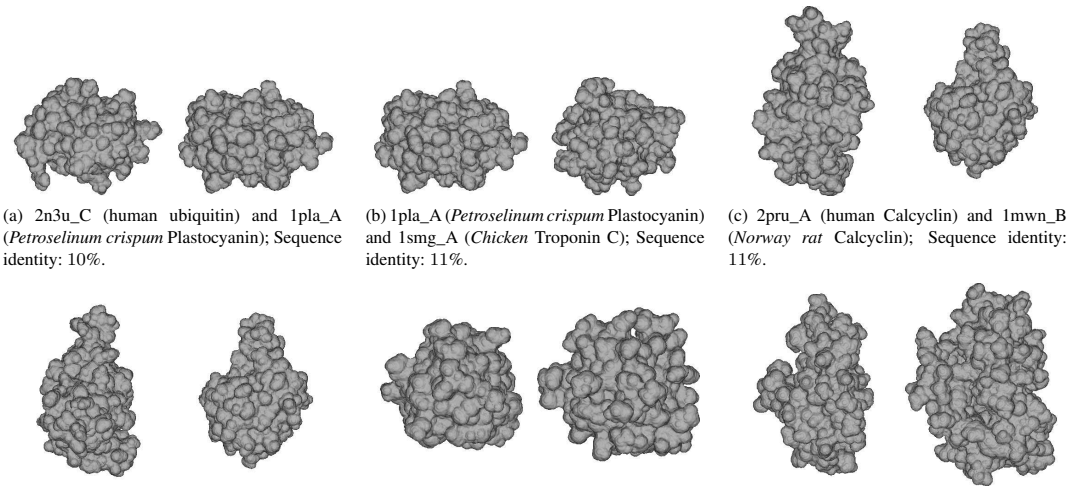
Table 3. Number of retrieved conformations within the top 70 for the class 2n2k_A and the top 30 for the class 1dmo_A with the different shape retrieval (top) and structure comparison (bottom) methods. SD is standard deviation.

Methods	2n2k_A		1dmo_A	
	Mean	SD	Mean	SD
3D-Surfer	69.12	6.69	9.60	2.45
Panorama	67.04	4.49	10.70	3.41
ShapeDNA	69.72	2.87	7.13	3.28
VFH	70	0	14.46	5.28
CE	1	0	1.91	2.3
DeepAlign	1	0	5.86	2.55
TM-Align	24.91	14.46	2.03	2.05

3.3 Identifying distant surficial homologs

Shape retrieval methods allow to compare protein structures regardless of their sequences, secondary structures or fold. We identified in set $\mathcal{A}$, 6 pairs of protein shapes (Figure 4, a-f) sharing up to 19% sequence identity but displaying similar surface shapes (*i.e* distant surficial homologs). From the biological function point of view, 3 out of the 6 pairs share different biological functions. The pairs b, f and e respectively associate a Calcium binding protein and an electron binding protein; a Calcium binding protein and a metal binding protein; a metal binding protein and an electron transport protein.

We compared the distances obtained for these protein pairs with the distances obtained within distant surficial homologs retrieved from the literature [5] (set $\mathcal{B}$, Table 4). Using the maximum distances observed for the selected protein pairs from set $\mathcal{A}$ as a similarity threshold for the protein pairs in set $\mathcal{B}$, 1, 0, 3, and 7 out of the 8 pairs from set $\mathcal{B}$ were identified by 3D-Surfer, Panorama, ShapeDNA, and VFH, respectively. On the same task, considering 2Å as a typical similarity threshold for protein structures [43], CE retrieved zero pair and DeepAlign one pair.



(a) 2n3u_C (human ubiquitin) and 1pla_A (*Petroselinum crispum* Plastocyanin); Sequence identity: 10%.

(b) 1pla_A (*Petroselinum crispum* Plastocyanin) and 1smg_A (*Chicken* Troponin C); Sequence identity: 11%.

(c) 2pru_A (human Calcyclin) and 1mwn_B (*Norway rat* Calcyclin); Sequence identity: 11%.

(d) 1dx1_A (*Cow* Prion) and 1mwn_B (*Norway rat* Calcyclin); Sequence identity: 18%.

(e) 1k3g_A (*Bacillus pasteurii* Cytochrome c6) and 1skt_A (*Chicken* Troponin C); Sequence identity: 9%.

(f) 1sym_A (*Norway rat* Calcyclin) and 2n27_A (human calmodulin); Sequence identity: 10%.

Fig. 4: Distant Surficial homologs identified using the shape similarity search methods (Sequence identity between 1smg_A (b) and 2n3u_C (a) is 8%).

Seven out of the eight pairs were considered to be similar using TM-Align (TM-score > 0.5).

4 Discussion

4.1 Protein shape retrieval

The evaluation of the performance of the shape retrieval methods on the protein shapes hierarchical dataset (set $\mathcal{A}$) highlighted the high performance of 3D-Surfer, the only shape retrieval method evaluated here originally designed to compare protein shapes. It was tightly followed by Panorama a shape retrieval method that was never applied to protein shapes to date. Despite displaying high performances in different non-molecular shapes benchmarks [23, 24, 25, 26], ShapeDNA and VFH were outperformed by 3D-Surfer and Panorama. This suggests that their shape descriptors could be less adapted to protein shape and highlights the complexity of molecular shapes compared to the 3D objects classically used in computer vision (furniture, buildings, animals, human faces, …). Proteins and molecular objects are considered feature-less compared to classical 3D-objects that usually display easily extractable and matchable features (wheels, ears, nose, legs, …). The extraction of standard 3D descriptors for a homogeneous surface such as a protein could result in ambiguous correspondences unless the descriptor is able to scale-up with a higher level of detail, notably in the *Protein* and the *Species* hierarchical level.

At the *Species* hierarchical level, orthologous proteins are separated in different classes. Most shape retrieval methods classified orthologous proteins within the same class, resulting in a loss of performance in retrieval compared to the *Protein* hierarchical level.

4.2 Tracking conformational changes

Recognizing the conformational changes of a given protein is very useful, notably in cryo-Electron Microscopy (cryo-EM) and cryo-Electron Tomography (cryo-ET) where detected macromolecular shapes can be identified using shape retrieval methods [6].

We explored the ability of the different shape retrieval methods to track protein conformational changes using the human ubiquitin and the *Xenopus Levis* calmodulin systems.

All shape retrieval methods succesfully retrieved most of the conformers of human ubiquitin (more than 67 conformers out of the 70). Except for TM-Align that retrieved less than 25 conformers out of the 70 within the first ranks on average, structure comparison methods were not able to retrieve a single conformer.

Concerning the *Xenopus Levis* calmodulin class, Shape retrieval methods outperformed structure comparison methods.The best performing VFH was designed to track the mobility of objects on camera snapshots over time [37]. This highlights one of the advantages of comparing proteins through their surface shapes since surface-based protein comparison abstracts this layer of complexity, *i.e* the secondary structures encoded in the backbone atoms 3D coordinates. The poor performances of TM-Align in this task could be explained by its residue to residue optimization that could have failed with the large motion of the second domain of the *Xenopus Levis* calmodulin.

EThis suggests that the structure comparison methods assign small distance values even if only a subpart of a protein is similar to the second protein to be compared with.

4.3 Identifying distant surficial homologs

The molecular surface is an abstraction of the primary, secondary, and tertiary structure representations. Functionally related proteins often share similar surface properties despite a low sequence and/or backbone conformation similarity [5, 6]. Identifying distant surficial homologs *i.e* proteins with similar surface shapes and low sequence identity, is of a major interest. It underlines the usefulness of shape retrieval methods and beyond, tackling protein structure comparison through surface shape instead of protein main chain orientation, especially when structural methods fail to identify such similarity. Shape retrieval methods could

Table 4. Distance values for the protein couples of set $\mathcal{B}$ [5] according to the shape similarity search and the structure comparison methods. SeqID is the sequence identity between the proteins within the couple.. **Reproduced from Sael et al. 2008 [5] for information.

Method	1jzn_A / 1g1q_A	1bar_A / 1rro	1ryp_B / 1gwz	1a31 / 1cy0	1tbp / 1t7p	1b3t / 1adv	2nwl / 2bbh	2b2i / 2cfp
3D-Surfer	7.60	8.62	11.88	4.72	7.46	8.07	7.57	5.80
Panorama	0.0263	0.0255	0.0251	0.0278	0.0237	0.0264	0.0247	0.0230
Shape-DNA	0.59	0.60	2.19	0.51	1.74	2.01	1.93	1.42
VFH	29	15	547	87	18	90	71	17
CE	2.48	5.48	6.54	6.47	3.35	5.88	7.77	5.04
DeepAlign	1.97	5.07	5.65	7.01	3.51	4.30	2.83	6.05
TM-Align	0.35	0.70	0.74	0.76	0.86	0.79	0.81	0.69
3DZD**	52.6	12.6	12.7	5.58	7.25	7.65	6.04	7.28
CE*,**	2	6.7	5	6.3	4.9	6.7	8.1	4.9
SeqID**	23.5%	3.6%	9.7%	5.8%	2%	9%	5.7%	7.8%

Table 5. Output distances for each method on the identified distant surficial homologs couples illustrated in Figure 4.

Protein couple	a	b	c	d	e	f
3D-Surfer	4.26	5.26	3.87	3.37	4.02	3.57
Panorama	0.0219	0.0227	0.0217	0.0213	0.0221	0.0222
Shape-DNA	1.09	0.22	0.57	0.73	0.37	0.79
VFH	180	57	44	33	79	31
CE	5.75	5.54	4.76	4.94	5.36	3.35
DeepAlign	4.46	3.21	3.55	3.43	3.24	3.51
TM-Align	0.31	0.29	0.24	0.27	0.23	0.79
SeqID	10.11%	11.23%	11.26%	18.42%	9.85%	10.14%

be used to identify proteins with similar molecular surfaces despite a low sequence identity which could be beneficial to the protein structure prediction community, notably in threading where folds could be enriched with surface shapes. This could also be useful for identifying possible interacting partners [4] since molecular shape plays a crucial role in binding [3, 44]. Shape retrieval methods could then be used for creating a structural classification of proteins based on their surfaces [6], rather than evolutionary distances or fold categories as in SCOP [40] or CATH [41] opening the possibility to extend the protein structure-function paradigm towards a protein structure-surface(s)-function paradigm.

Here, we extended set $\mathcal{B}$ by using a consensus of the shape retrieval methods evaluated in this study to screen set $\mathcal{A}$. 6 protein pairs were identified in set $\mathcal{A}$ displaying similar surface shapes with sequence identity below 19%. These protein pairs from sets $\mathcal{A}$ and $\mathcal{B}$ could constitute a useful resource for the evaluation of the performance of future shape retrieval methods to identify distant surficial homologs.

5 Conclusion

In this work, we evaluated the performance of four shape retrieval methods from the computer vision field (3D-Surfer, Panorama, Shape-DNA, and VFH) on a protein shapes dataset. On this dataset, 3D-Surfer and Panorama outperformed Shape-DNA and VFH. On selected proteins displaying large conformational changes, all methods displayed a high performance in recognizing their different conformations.

Different structure comparison methods were used as a reference in this study (CE, DeepAlign and TM-Align). TM-Align slightly outperformed shape retrieval methods in the retrieving task, but failed in tracking large conformational changes. We also identified 6 pairs of distant surficial homologs that could be used for future studies on protein surficial similarity search.

This work confirms the interest of protein shape as a higher-level description of the protein structure that 1) abstracts the underlying protein sequence, structure or fold, 2) allows the use of shape retrieval methods to screen large databases of protein structures to identify surficial homologs and possible interacting partners, 3) opens an extension of the protein structure-function paradigm towards a protein structure-surface(s)-function paradigm.

Acknowledgements

We thank Prof. D. Kihara for providing support with the 3D-Surfer server. M. Machat thanks Mr J. Saint-Jean for his help with the Meshlab software.

Funding

This work is funded by the European Research Council Executive Agency under the research grant number 640283.

References

[1]Michael L Connolly. Analytical molecular surface calculation. *Journal of applied crystallography*, 16(5):548–558, 1983.

[2]Frederic M Richards. Areas, volumes, packing, and protein structure. *Annual review of biophysics and bioengineering*, 6(1):151–176, 1977.

[3]A Shulman-Peleg, R Nussinov, and H J Wolfson. Content-based 3d object retrieval. *J. Mol. Biol.*, 339:607–633, 2004.

[4]Pablo Gainza, Freyr Sverrisson, Frederico Monti, Emanuele Rodola, Michael M Bronstein, and Bruno E Correia. Deciphering interaction fingerprints from protein molecular surfaces. *Nature Methods*, 17:184–192, 2020.

[5]Lee Sael, Bin Li, David La, Yi Fang, Karthik Ramani, Raif Rustamov, and Daisuke Kihara. Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins: Structure, Function, and Bioinformatics*, 72(4):1259–1273, 2008.

[6]Xusi Han, Atilla Sit, Charles Christoffer, Siyang Chen, and Daisuke Kihara. A global map of the protein shape universe. *PLoS computational biology*, 15(4):e1006969, 2019.

[7]Benjamin Bustos, Daniel A Keim, Dietmar Saupe, and Tobias Schreck. Content-based 3d object retrieval. *IEEE Computer graphics and Applications*, 27(4), 2007.

[8]Apostol Gramada and Philip E Bourne. Multipolar representation of protein structure. *BMC bioinformatics*, 7(1):1–13, 2006.

[9]Lee Sael, David La, Bin Li, Raif Rustamov, and Daisuke Kihara. Rapid comparison of properties on protein surface. *Proteins: Structure, function, and bioinformatics*, 73(1):1–10, 2008.

[10]Lora Mak, Scott Grandison, and Richard J Morris. An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *Journal of Molecular Graphics and Modelling*, 26(7):1035–1045, 2008.

[11]Daniela Craciun, Jeremy Sirugue, and Matthieu Montes. Global-to-local protein shape similarity system driven by digital elevation models. In *Proceedings of the 2nd International Conference on Bio-engineering for Smart Technologies*, pages 1–4. IEEE, 2017.

[12]Zhanheng Gao, Reihaneh Rostami, Xiaoli Pang, Zhicheng Fu, and Zeyun Yu. Mesh generation and flexible shape comparisons for biomolecules. *Computational and Mathematical Biophysics*, 1(open-issue), 2016.

[13]Remco Veltkamp, Remco Ruijsenaars, Michela Spagnuolo, Roelof van Zwol, and Frank ter Haar. Shrec2006 3d shape retrieval contest. In *Proceedings of the Workshop on 3D Object Retrieval*, pages 1–9. UU WINFI Informatica en Informatiekunde, 2006.

[14]Lazaros Mavridis, Vishwesh Venkatraman, et al. Shrec-10 track: Protein models. In *3DOR: Eurographics Workshop on 3D Object Retrieval*, pages 117–124, 2010.

[15]Na Song, Daniela Craciun, et al. Protein shape retrieval: Shrec'17 track. In *Proceedings of the Workshop on 3D Object Retrieval*, pages 67–74. Eurographics Association, 2017.

[16]Florent Langenfeld, Apostolos Axenopoulos, et al. Shrec 2018–protein shape retrieval. In *Eurographics Workshop on 3D Object Retrieval*, pages 53–61, 2018.

[17]Florent Langenfeld, Apostolos Axenopoulos, et al. Shrec19–protein shape retrieval contest. In *Eurographics Workshop on 3D Object Retrieval*, pages 25–31, 2019.

[18]Florent Langenfeld, Yuxu Peng, et al. Shrec2020 track: Multi-domain protein shape retrieval challenge. *Computers & Graphics*, 91:189 – 198, 2020.

[19]David La, Juan Esquivel-Rodríguez, Vishwesh Venkatraman, Bin Li, Lee Sael, Stephen Ueng, Steven Ahrendt, and Daisuke Kihara. 3d-surfer: software for high-throughput protein surface comparison and analysis. *Bioinformatics*, 25(21):2843–2844, 2009.

[20]Panagiotis Papadakis, Ioannis Pratikakis, Theoharis Theoharis, and Stavros Perantonis. Panorama: A 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval. *International Journal of Computer Vision*, 89(2-3):177–192, 2010.

[21]Martin Reuter, Franz-Erich Wolter, and Niklas Peinecke. Laplace–beltrami spectra as 'shape-dna' of surfaces and solids. *Computer-Aided Design*, 38(4):342–366, 2006.

[22]Radu Bogdan Rusu, Gary Bradski, Romain Thibaux, and John Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2155–2162. IEEE, 2010.

[23]Zhouhui Lian, Afzal Godil, et al. A comparison of methods for non-rigid 3d shape retrieval. *Pattern Recognition*, 46(1):449–461, 2013.

[24]Chunyuan Li and A Ben Hamza. Spatially aggregating spectral descriptors for nonrigid 3d shape retrieval: a comparative survey. *Multimedia Systems*, 20(3):253–281, 2014.

[25]Bo Li, Yijuan Lu, et al. A comparison of 3d shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Computer Vision and Image Understanding*, 131:1–27, 2015.

[26]Z. Lian, J. Zhang, et al. Shrec'15 track: Non-rigid 3d shape retrieval. *3DOR*, 2015.

[27]Ilya N Shindyalov and Philip E Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein engineering*, 11(9):739–747, 1998.

[28]Sheng Wang, Jianzhu Ma, Jian Peng, and Jinbo Xu. Protein structure alignment beyond spatial proximity. *Scientific reports*, 3:1448, 2013.

[29]Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

[30]Naomi K Fox, Steven E Brenner, and John-Marc Chandonia. Scope: Structural classification of proteins—extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42(D1):D304–D309, 2013.

[31]John-Marc Chandonia, Naomi K Fox, and Steven E Brenner. Scope: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic acids research*, 47(D1):D475–D481, 2018.

[32]HM Berman, J Westbrook, et al. The protein data bank nucleic acids research, 28: 235-242. *URL: www. rcsb. org Citation*, 2000.

[33]Dong Xu and Yang Zhang. Generating triangulated macromolecular surfaces by euclidean distance transform. *PloS one*, 4(12):e8140, 2009.

[34]Michael L Connolly. The molecular surface package. *Journal of molecular graphics*, 11(2):139–141, 1993.

[35]Martin Reuter. Hierarchical shape segmentation and registration via topological features of laplace-beltrami eigenfunctions. *International Journal of Computer Vision*, 89(2):287–308, Sep 2010.

[36]Aitor Aldoma, Zoltan-Csaba Marton, et al. Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation. *IEEE Robotics & Automation Magazine*, 19(3):80–91, 2012.

[37]Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[38]Marc Marti-Renom, Emidio Capriotti, I. Shindyalov, and Philip Bourne. Structure comparison and alignment. *Structure Comparison and Alignment. edn 2*, 01 2009.

[39]Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[40]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.

[41]Christine A Orengo, Alex D Michie, Sue Jones, David T Jones, Mark B Swindells, and Janet M Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109, 1997.

[42]Fabian Sievers, Andreas Wilm, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1), 2011.

[43]Ram Samudrala and Michael Levitt. A comprehensive analysis of 40 blind protein structure predictions. *BMC Structural Biology*, 2(1):3, Aug 2002.

[44]Guillaume Levieux, Guillaume Tiger, et al. Udock, the interactive docking entertainment system. *Faraday Discuss.*, 169:425–441, 2014.

Chapitre 3

Méthode basée sur les cartes de convexité

Contenu

3.1 Motivations

L'un des objectifs du **projet ViDOCK** est la comparaison rapide des surfaces macromoléculaires. L'idée développée dans le cadre du projet ViDOCK pour pouvoir comparer rapidement les protéines est une projection sur une sphère unitaire de la surface macromoléculaire pour obtenir un descripteur en deux dimensions [8]. Un descripteur 2D permet ainsi de diminuer la taille de stockage et de comparer plus rapidement deux objets entre eux. La comparaison entre les deux protéines se fait ensuite en comparant deux à deux les sous-régions de la représentation 2D en utilisant une fonction de distance avec un temps de calcul négligeable par rapport à la comparaison générale de deux objets.

Avec ces prérequis, différentes approches ont été expérimentées pour trouver une méthode du domaine de vision par ordinateur pouvant être adaptés aux surfaces macromoléculaires.

Les entrées de l'algorithme sont des fichiers pdb, représentant la structure de la protéine. Une première étape est de calculer la surface macromoléculaire avec l'aide d'EDTSurf [9]. Les surfaces obtenues sont ensuite projetées sur la sphère unité en utilisant l'opérateur de Laplace-Beltrami avec une projection conforme [88]. La sphère unité est ensuite projetée sur un plan 2D. Les étapes du procédé sont résumées en Figure 3.1.

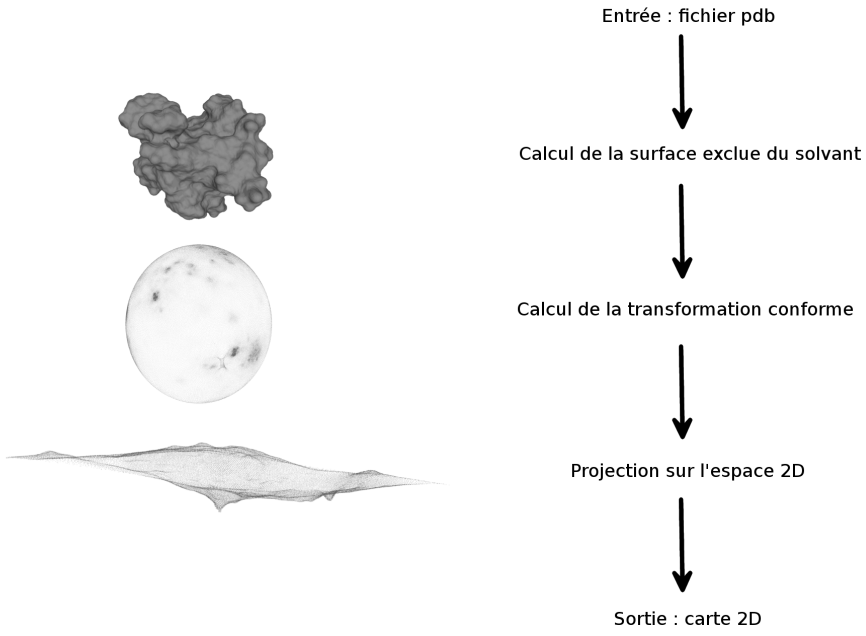


FIGURE 3.1 – *Workflow* pour la projection d'une protéine sur un espace 2D.

3.2 Conception de la méthode

3.2.1 Passage de la 3D à la 2D

Dans un premier temps on utilise la méthode EDTSurf [9] permettant de passer d'une représentation de la protéine par sa structure à une représentation de la protéine par sa surface. EDTSurf peut générer les trois surfaces triangulées couramment utilisé en bioinformatique structurale, c'est-à-dire la triangulation de la surface de Van der Waals (VdW), la surface accessible au solvant (SAS) et la surface exclue au solvant (SES).

En utilisant une carte de distances 3D (ou transformée de distance) avec la distance euclidienne sur la structure de la protéine, les trois types de surface peuvent être obtenus. L'idée est d'utiliser une *bounding box* basée sur la division de l'espace en voxel pour déterminer la surface en calculant la distance minimum entre les atomes de la structure et la *bounding box*. L'utilisation d'une transformée de distance sur les coordonnées des atomes permet d'obtenir un nuages de points représentant la surface de la protéine. Une fois la surface obtenue un dérivé de l'algorithme **Marching Cube** (MC)

[40] est appliqué sur la surface pour obtenir la triangulation de la surface. Ce *marching cube* est appelé Vertex-Connected Marching Cube (VCMC) qui utilise les voxels de la grille divisant l'espace contenant l'objet 3D. MC prend en compte les points de l'intersection de la surface avec la grille tandis que VCMC utilise directement les points de la grille. Cette variante permet de diminuer par deux le nombre de vertices [9] et donc de diminuer l'espace de stockage des surfaces. De plus le temps de calcul de VCMC est 1.4 fois plus rapide que MC sans pour autant impacter significativement le résultat final [9].

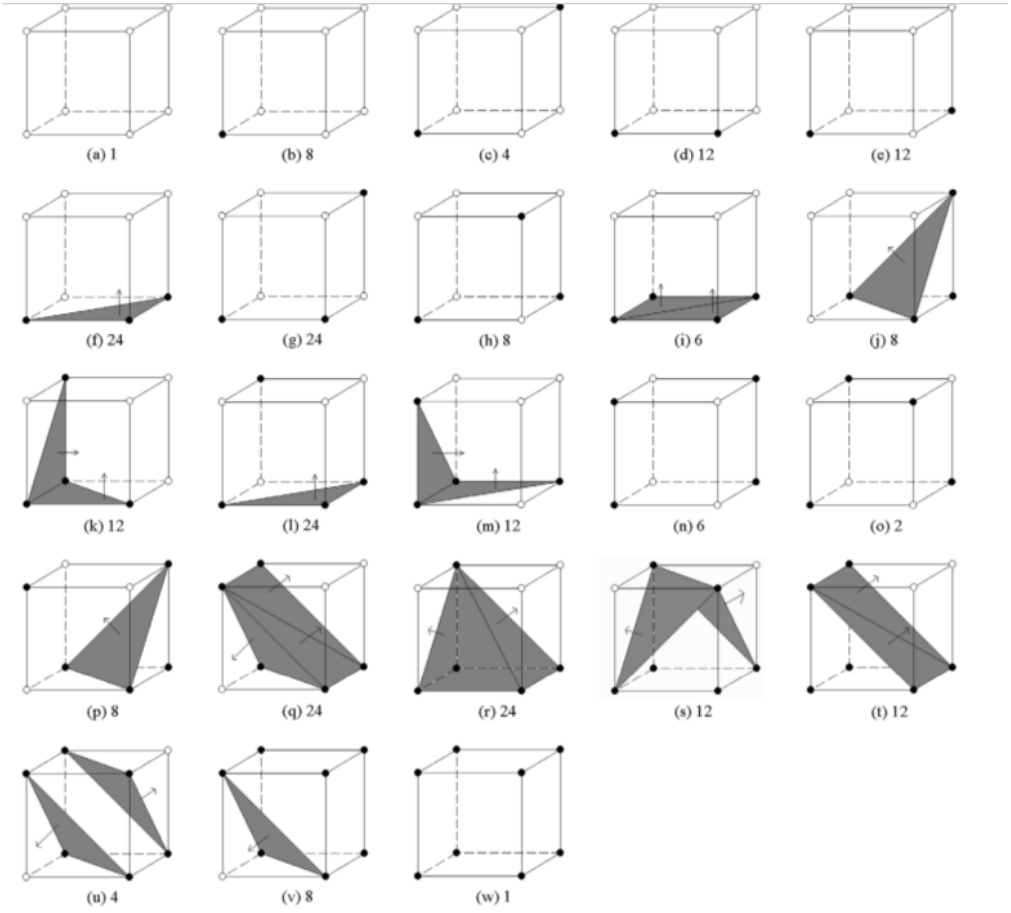


FIGURE 3.2 – Cas possible de triangles à partir de voxels lors de la triangulation. Les points noirs sont les voxels à l'intérieur de la surface et les voxels blancs sont à l'extérieur. Illustration extraite de [9].

La surface est projetée sur une sphère unitaire [88]. La transformation conforme est obtenue avec la résolution d'une équation aux dérivées partielles elliptiques de second ordre pour obtenir une fonction z appelée *équivalence conforme*. Une équivalence conforme est une transformation bijective qui conserve

localement les angles permettant de définir la position des points de l'objet 3D sur la sphère unitaire. L'équation aux dérivées partielles (EDP) à résoudre est la suivante :

$$\Delta z = (\frac{\partial}{\partial u} - i\frac{\partial}{\partial v})\partial_p$$

Avec p un point de la surface choisi arbitrairement, u et v les coordonnées conformes définies dans le voisinage de p, ∂_p est la fonction de Dirac au point p.

Pour résoudre l'EDP la méthode des éléments finis est utilisée qui est une méthode de résolution approchée et numérique d'EDP dans un espace discret. La transformation conforme projette l'objet 3D sur le plan complexe. Ensuite une projection stéréographique inverse passe du plan complexe à la sphère unitaire. Une projection stéréographique est une projection d'une sphère sur un plan.

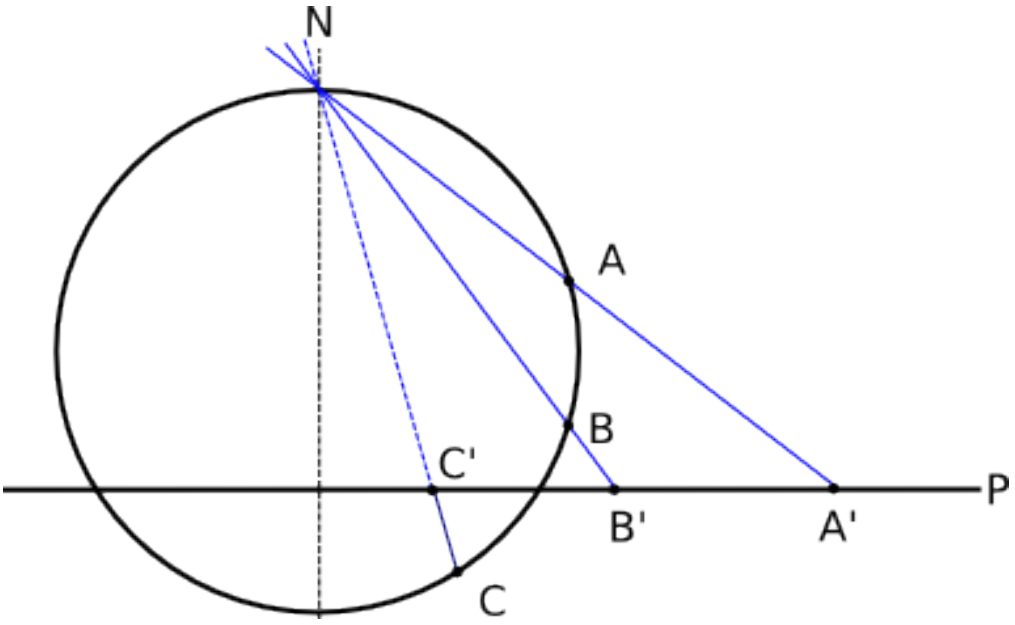


FIGURE 3.3 – Schéma d'une coupe d'une projection stéréographique du pôle nord N sur le plan P

Une sphère unitaire S avec les points du maillage 3D de départ est obtenue. Les coordonnées polaires θ et ϕ correspondant respectivement à la longitude et la latitude sur la sphère unitaire. La valeur du rayon r est constamment égal à 1.

La transformation sur le plan 2D de la méthode SSS-DEM[8] génère une grille avec une variation δ constante selon l'axe des coordonnées polaires θ et ϕ. Les coordonnées de la grille selon l'axe x et y sont respectivement les coordonnées θ et ϕ de la sphère unitaire. Une ou plusieurs valeurs sont ajoutées en

chacun des points de la sphère unitaire projetés sur la grille pour former le descripteur appelé **carte**. Ces valeurs sont l'information qui est jugée pertinente pour être comparée.

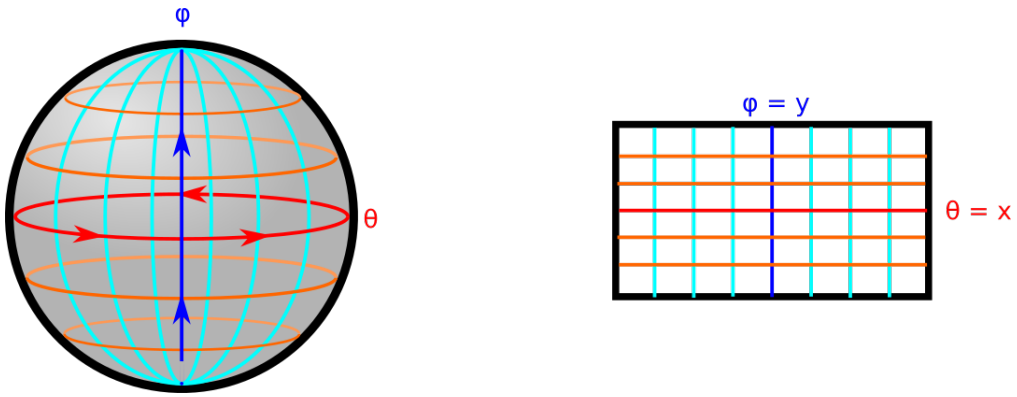


FIGURE 3.4 – Schéma de la projection de la sphère unitaire dans le plan 2D pour former une carte.

L'avantage de la projection conforme est de conserver localement la forme de l'objet qui peut être ensuite comparée localement avec par exemple des patches sans que la distorsion soit un problème [89].

Un désavantage d'utiliser une projection sur la sphère unitaire est que cette projection est adaptée aux formes globulaires des protéines. Plus les formes des protéines sont allongées, comme par exemple les protéines fibreuses, plus la projection sera déséquilibrée du fait de la différence de forme entre la protéine et la sphère.

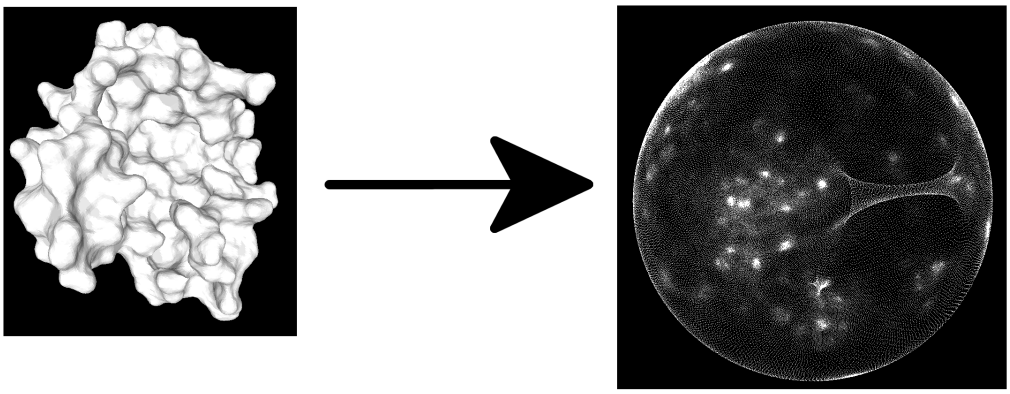


FIGURE 3.5 – Exemple de la projection sur la sphère unitaire d'une protéine

3.2.2 Carte convexité

Le premier descripteur développé est appelé Carte de Convexité (CCvx) et permet de représenter les zones de convexité et de concavité sur les Digital Elevation Model (DEM) [8] d'une protéine.

Dans un premier temps la valeur d'élévation p_z est attribuée aux points de la carte 2D et est appelée DEM. Cette valeur est la coordonnée selon l'axe z de l'objet 3D.

Les DEM sont une représentation utilisée en topographie [90] [91] [92]. Les travaux de topographie ont pour objectif de classifier une zone géographique dans le but de pouvoir étudier ensuite les cartes de cette zone géographique. Les descripteurs utilisés en topographie produits à partir de DEM sont souvent liés au gradient de la pente en chaque point de la carte. La pente et l'élévation permettent de définir si un point de la surface désigne un pic, une colline, un plat, une vallée ou une dépression.

La convexité sur les DEM est calculée pour obtenir une **carte de convexité**. Les cartes de convexité s'inspirent des descripteurs de topographie pour caractériser les zones "creuses" et les zones "bombées", c'est-à-dire de caractériser les zones **concaves** et **convexes** d'une DEM en utilisant les valeurs de gradient en chacun des points.

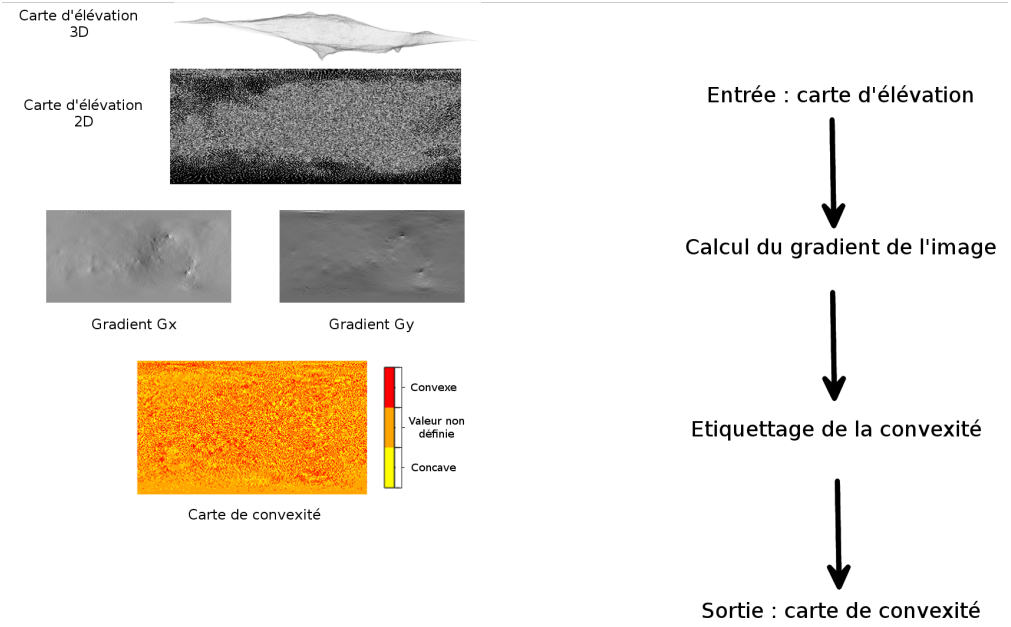


FIGURE 3.6 – Workflow de la création de cartes de convexité

A partir d'une DEM I on calcule le gradient selon l'axe x et y pour obtenir respectivement les

cartes de gradient G_x et G_y.

On fait une convolution de l'image avec le filtre pour chacune des deux cartes de gradients :

$$G_x = I * [-1 \ +1]$$

$$G_y = I * \begin{bmatrix} -1 \\ +1 \end{bmatrix}$$

Pour en déduire la convexité on définit ce qu'on appelle un point critique. Un point critique est un point où le gradient est égal à zéro. Dans notre cas, un point critique est l'ensemble $\{p_i, p_{i+1}\}$ composé de deux points consécutifs p_i et p_{i+1} tel que la multiplication de la valeur de gradient g_{p_i} au point p_i avec la valeur de gradient $g_{p_{i+1}}$ au point p_{i+1} soit négative : $g_{p_i} g_{p_{i+1}} < 0$.

Pour chacun des axes et suivant leur sens on étiquette deux cartes de convexité intermédiaire, C_x et C_y selon respectivement l'axe x et y. Un point est étiqueté *convexe* autour du point critique si, dans le sens de l'axe x ou y, le point critique est composé d'une valeur de gradient positive puis négative. A l'inverse, un point est étiqueté *concave* si, dans le sens de l'axe, le point critique est composé d'une valeur de gradient négative puis positive (cf Figure 3.7).

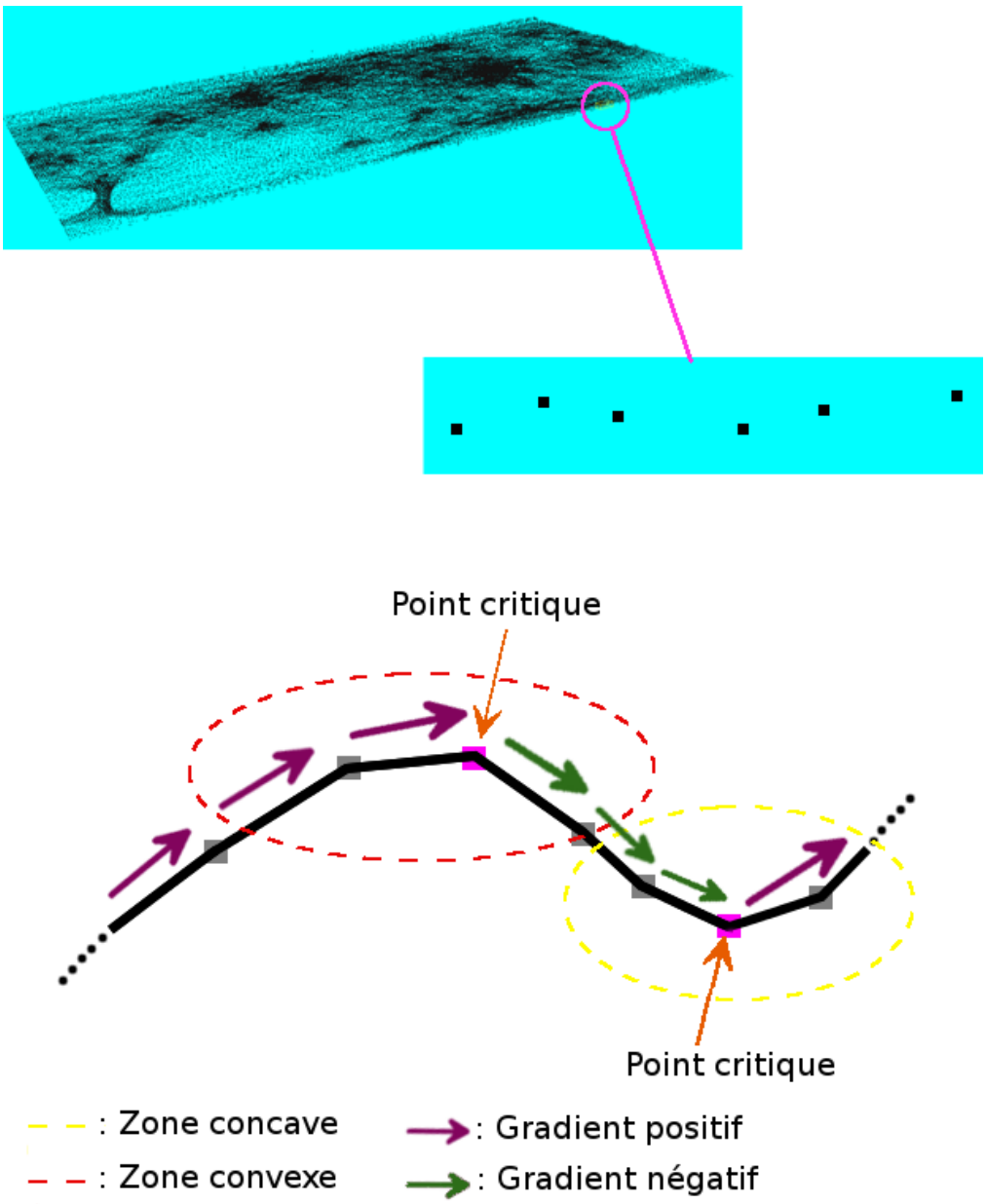


FIGURE 3.7 – En haut, nuage de points d'une DEM. En bas, schéma du calcul de la convexité à partir d'une DEM

Les deux cartes de convexité intermédiaires sont ensuite réunies en une seule en les superposant. Si deux points ont la même étiquette, alors celle-ci est attribuée à la nouvelle carte de convexité. Si les étiquettes sont différentes, le point est appelé un point col (ou point-selle). L'étiquette des points col est déterminée à partir de l'angle du gradient. La valeur de l'angle indique si le vecteur gradient

est plus proche de l'axe x ou y. La valeur de la carte de convexité représentant l'axe dont le vecteur est le plus proche est sélectionnée. L'angle θ du gradient est calculé de la manière suivante :

$$\theta = \arctan(\frac{g_x}{g_y}) \ \ \forall \ gx \in G_x \ \ \text{et} \ \ \forall \ g_y \in G_y$$

Lors de la projection sur la carte 2D, du fait de la déformation induite par le passage d'une sphère à un plan, certaines parties se retrouvent sans valeur attribuées. Dans ce cas là on attribue une valeur nulle à cet emplacement qui ne sera pas prise en compte lors de l'analyse des cartes de convexité.

Une carte qui représente tous les points d'une DEM selon une valeur topographique dépendante du gradient de la DEM est ainsi obtenue.

L'avantage de cette représentation est que l'on a besoin de seulement deux valeurs (concave et convexe) permettant ainsi de minimiser la taille de stockage des cartes de convexités en les représentants sous forme binaire. Cette représentation permet l'emploi d'opérateurs binaires lors de la comparaison permettant ainsi de réduire le temps de comparaison des cartes de convexité.

Autour de chaque pixel de la panoramique de convexité, on va définir une zone de $N * N$ pixels. Cette zone est appelée patch et définit le voisinage d'une point. A partir de ces patchs est extrait un histogramme qui possède deux *bins*, une pour les valeurs concaves et une pour les valeurs convexes. Cet histogramme donne le cardinal des deux valeurs (cf Figure 3.8).

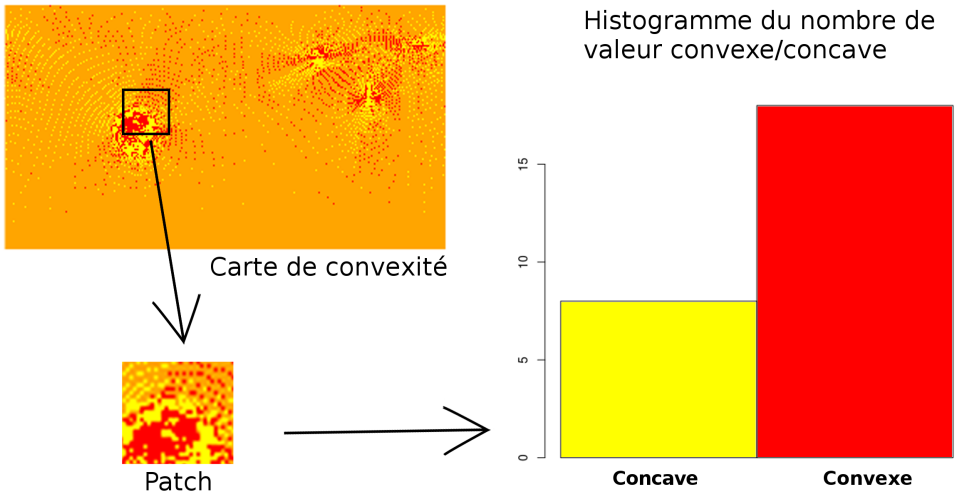


FIGURE 3.8 – Schéma du calcul d'un histogramme à partir d'une carte de convexité

3.2.3 Comparaison des cartes de convexité

Les patchs de la carte de convexité *query* I sont comparés un à un avec tous les patchs de la carte de convexité cible J (cf Figure 3.9). La comparaison se fait en calculant le score défini comme la somme de l'absolu de la différence de chacune des classes de l'histogramme.

Deux objets sont comparés en recherchant les meilleures correspondances entre deux patchs de leurs cartes de convexité I et J. Les patchs des cartes de convexité I et J sont respectivement appelés P_I et P_J. Pour comparer les deux *bins* respectivement convexe et concave $H_{cvx}(a)$ et $H_{ccv}(a)$ au point a, la **distance de Minkowski** est utilisée sur chacun des *bins* des histogrammes de I et J. Une somme est appliquée sur les différentes valeur de distance calculée. Est ainsi obtenu un **score de dissimilarité** entre deux patchs. Le score, noté S_p, entre deux points $k_I = (i, j)$ et $k_J = (k, l)$, respectivement un point de I et J, est défini de la manière suivante :

$$S_p(P_I(k_I), P_J(k_J)) = |H_{cvx}(k_I) - H_{cvx}(k_J)| + |H_{ccv}(k_I) - H_{ccv}(k_J)|$$

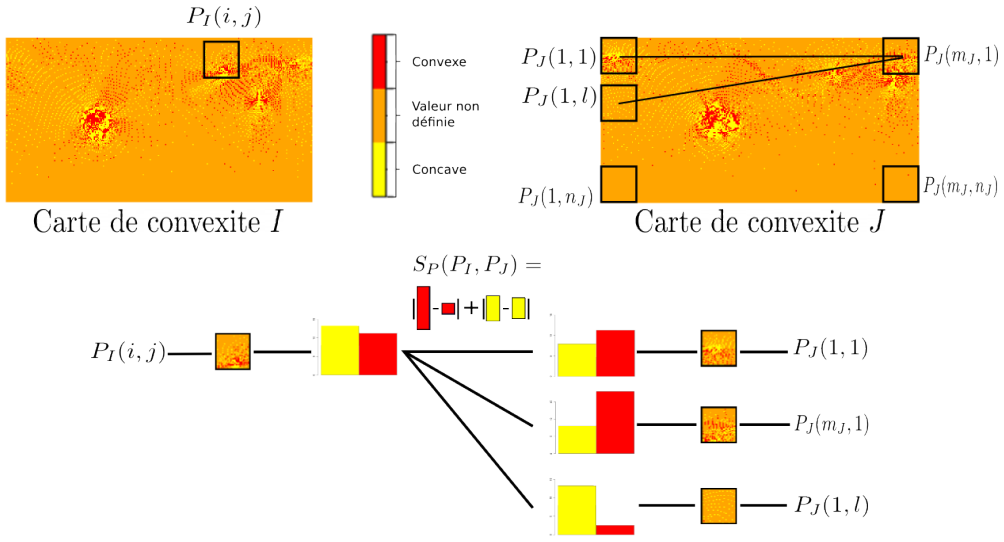


FIGURE 3.9 – Schéma de comparaison d'un patch P_I d'une image I contre tous les patchs P_J d'une image J

Pour chaque point de l'objet *query* I, le meilleur score est sélectionné, c'est-à-dire le score minimum

entre un point de I et tout les points de J :

$$\min_{k_J \in J} S_p(P_I(k_I), P_J(k_J)) \tag{3.1}$$

Ensuite, tous les scores sélectionnés sont sommés. Si I est composé de $N_I = (m_I, n_I)$ points et J de $N_J = (m_J, n_J)$ points, alors le score $S(I, J)$ de dissimilarité entre I et J est :

$$S(I, J) = \sum_{k_I=(1,1)}^{N_I} \min_{k_J \in J} S_p(P_I(k_I), P_J(k_J))$$

3.3 Jeux de données

Le jeu de données appelé Tosca (cf Figure 3.10), est un jeu de données de formes anatomiques, comprenant des humains et des animaux. Il est composé de 9 classes, les objets étant en haute résolution et sous différentes conformations. Les classes sont un même objet avec différents mouvements non-rigides.

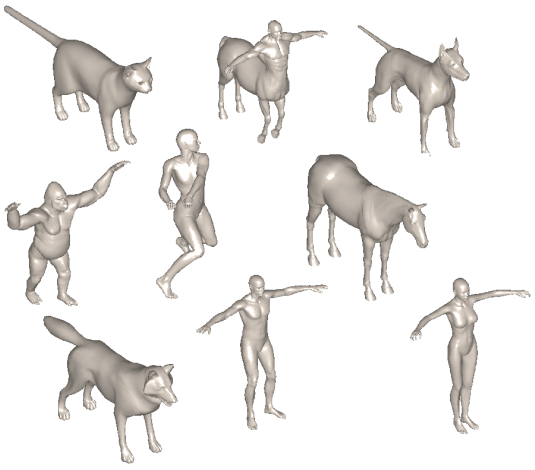


FIGURE 3.10 – Exemple des objets du jeu de données Tosca

3.4 Analyse de la méthode sur le jeu de données TOSCA

La méthode locale a été testée sur le jeu de donnée TOSCA pour avoir une validation de celle-ci. La validation permet d'introduire la méthode locale utilisée par l'approche Globale-Locale présentée

ensuite.

3.4.1 Résultats

La matrice de score de dissimilarité permet d'avoir une évaluation qualitative de la méthode locale. L'intersection entre la ligne I et la colonne J représente le score de dissimilarité entre l'objet I et J. Un score de dissimilarité faible (rouge sur la Figure 3.11) indique des objets similaires et au contraire si la dissimilarité est forte (blanc sur la Figure 3.11) alors les objets sont peu similaires.

La reconnaissance intra-classe est visible sur la matrice de score (cf Figure 3.11). En particulier, les classes du cheval, du loup et du centaure ont un très bon score intra-classe comparé à leur score inter-classe. Le score inter-classe entre David et Michael et entre le centaure et le loup est proche de leur score intra-classe.

Le centaure a un score intermédiaire avec David, le chien et le cheval. Victoria possède les scores de dissimilarité les plus élevés avec David et Michael.

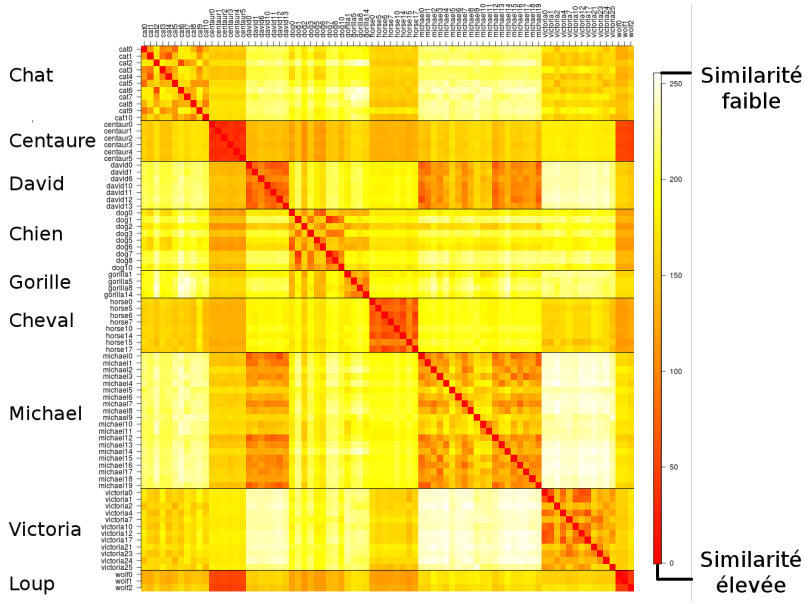


FIGURE 3.11 – Matrice de score de la méthode locale testée sur le jeu de données Tosca

Les performances de Nearest Neighbour (NN), First Tier (FT) et Second Tier (ST), listé dans le tableau ci-dessous (cf Table A.2) sont étudiées.

Le Nearest Neighbour est égal à 1 pour toutes les classes à l'exception des deux modèles masculins,

David et Michael, ou le NN est de respectivement 0.714 et 0.9.

Le First Tier est supérieur à 0.75 pour le centaure, gorille, le cheval et le loup. Le FT de David, Michael et Victoria est supérieur à 0.5. Le chat et le chien ont un FT inférieur à 0.5.

Le Second Tier donne un meilleur pourcentage de vrai positif que le ST pour chacune des classes. Le centaure et le loup ont un ST de 1. David, le gorille, le cheval, Michael et Victoria ont un ST supérieur à 0.8. Le chat et le chien ont respectivement un ST de 0.564 et 0.694.

Classes	Nearest Neighbour	First Tier	Second Tier
cat	1.000	0.464	0.564
centaur	1.000	0.933	1.000
david	0.714	0.548	0.881
dog	1.000	0.431	0.694
gorilla	1.000	0.750	0.833
horse	1.000	0.821	0.893
michael	0.900	0.587	0.963
victoria	1.000	0.652	0.841
wolf	1.000	1.000	1.000

TABLE 3.1 – Performances sur le jeu de données TOSCA de 80 objets

La comparaison du NN, FT et ST global de CCvx aux méthodes FPFH, USC et WKS est résumée dans la Table 3.2. Les différentes méthodes de l'état de l'art possèdent des mesures faibles variant entre 0 et 0.2 pour le NN, entre 0.09 et 0.3 pour le FT et 0.2 et 0.584 pour le ST.

Classes	Nearest Neighbour	First Tier	Second Tier
CCvx	0.938	0.711	0.893
FPFH	0.038	0.176	0.378
USC	0.025	0.097	0.200
WKS	0.188	0.295	0.584

TABLE 3.2 – Performances sur le jeu de données TOSCA de 80 objets avec les méthodes CCvx, FPFH, USC et WKS

La courbe de précision rappel (cf Figure 3.12) diminue de façon linéaire entre une valeur de précision de 1 pour 0.05 de rappel et 0.51 pour une valeur de 1 de rappel.

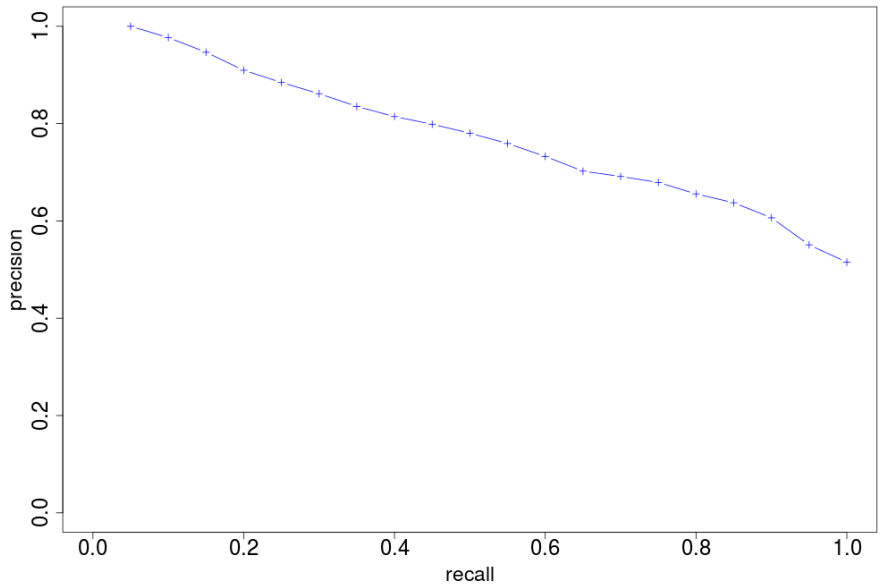


FIGURE 3.12 – Courbe de précision rappel sur le jeu de données Tosca

Les courbes de précision-rappel pour les 3 méthodes de l'état de l'art sont faibles avec une diminution rapide vers la valeur 0.4 pour le WKS et 0.3 pour FPFH et USC.

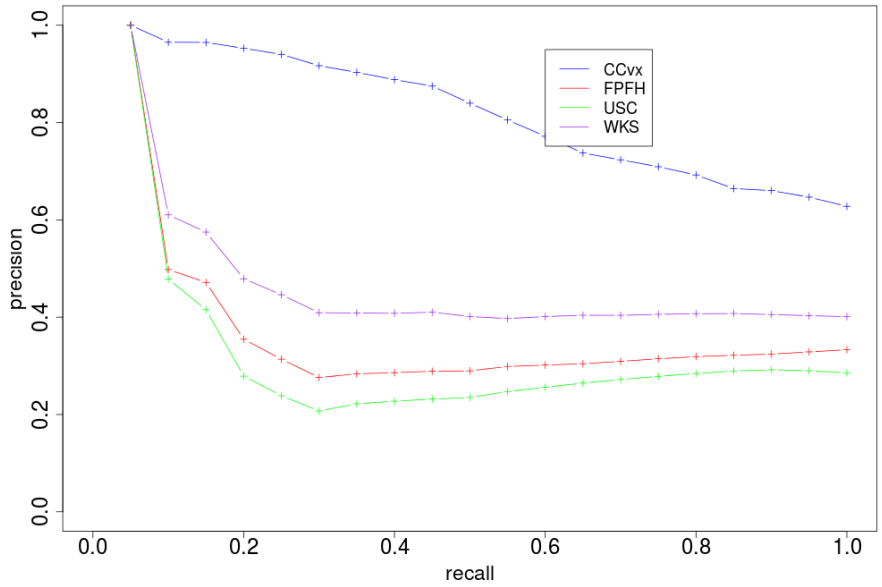


FIGURE 3.13 – Courbe de précision rappel sur le jeu de données Tosca pour les méthodes CCvx, FPFH, USC et WKS

Le temps de calcul du descripteur et le temps de comparaison de CCvx, FPFH, USC et WKS est détaillé dans la Table 3.3. CCvx possède le temps de calcul d'un descripteur le plus rapide avec 4.7 secondes mais la comparaison moyenne la plus lente avec 1 minute et 39 secondes. Au contraire WKS est la plus lent pour calculer son descripteur avec 48 secondes et le plus rapide pour comparer avec 0.4 secondes. Les fichiers les plus légers sont ceux du descripteur de CCvx avec 11 kilo octet et les plus lourd sont les fichiers du descripteur de WKS avec 3.3 mega octet.

Descripteur	Temps de calcul du descripteur	Temps de comparaison	Taille des fichiers
CCvx	4.7 sec	1 min 39 sec	11 kB
FPFH	7.1 sec	2.7 sec	1.1 mB
USC	19 sec	2.3 sec	923 kB
WKS	48 sec	0.4 sec	3.3 mB

TABLE 3.3 – Temps de calcul moyens et taille des fichiers des descripteurs de CCvx FPFH, USC et WKS

3.4.2 Discussion

Michael et David sont tous les deux des objets représentant un être humain masculin et sont de corpulence proche, il est donc facile d'avoir une confusion entre ces deux classes.

De nombreux objets qui ont des scores intermédiaires peuvent s'expliquer par une forme partiellement commune entre eux. Les objets possédant quatre pattes ont un score intermédiaire. Le cas du centaure est intéressant car il partage un score intermédiaire avec David qui ont en commun un buste humanoïde et avec le cheval qui possèdent tout les deux un corps de cheval.

Le loup et le centaure ont un score inter-classe élevé, ce sont deux classes ne possédant pas beaucoup de points comme on peut le voir sur les panoramiques de convexité (cf Figure 3.14) ce qui peut expliquer d'éventuels faux positifs accentués par le fait qu'ils ont une base commune, les quatre pattes.

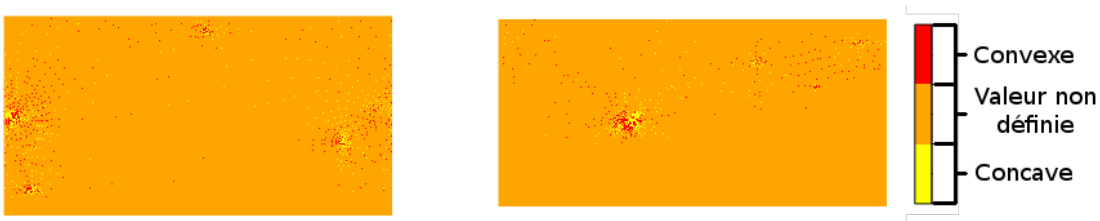


FIGURE 3.14 – Panoramique de convexité du centaure - Centaur0 (gauche) et du loup - Wolf0 (droite)

Victoria a des scores de dissimilarité sensiblement supérieurs à la moyenne en particulier avec David et Michael. Il est surprenant que Victoria possède des scores de dissimilarité élevés avec David et Michael alors que tout les trois sont des humains. Si l'on observe les cartes de convexité, l'une des régions denses en points est coupée en deux pour David et Michael contrairement à victoria. (cf Figure 3.15). Cette région ne se retrouve pas au centre d'un patch dans le cas de David et Michael et donc elle n'est pas comparée dans sa totalité contrairement à celle de Victoria. Une solution serait de faire varier la bordure, ce qui est une solution proposée dans le chapitre suivant avec le *multiview*.

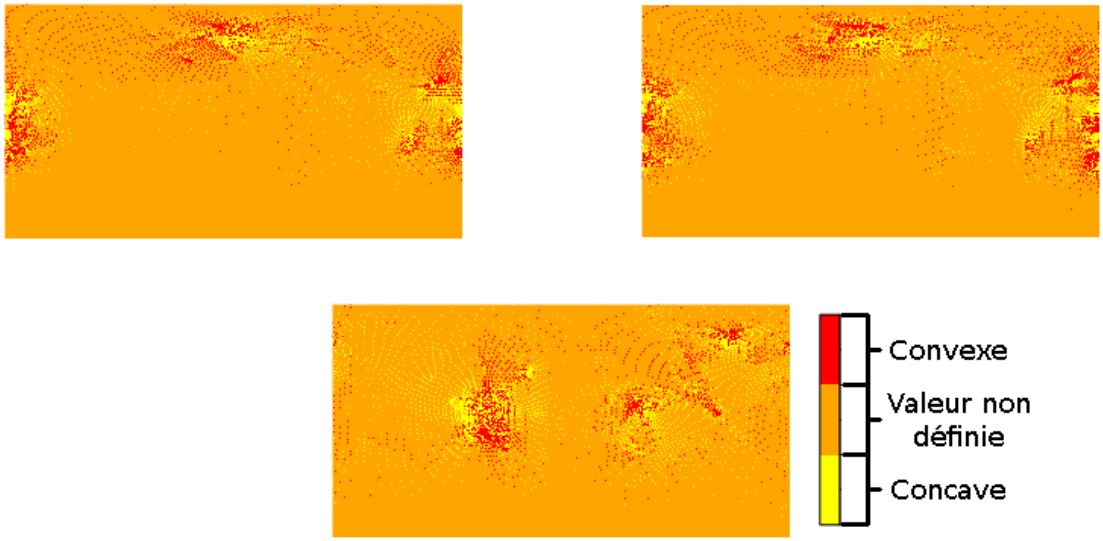


FIGURE 3.15 – Panoramique de convexité du David - David0 (haut, gauche), de Michael (haut, droite) et de Victoria - Victoria7 (bas)

Les valeurs obtenues avec les méthodes de l'état de l'art sont surprenantes. Ceci peut s'expliquer par le jeu de données TOSCA. La vérité terrain de ce jeu de données est basée sur le nombre de points des objets. Des objets d'une même classe possède le même nombre de points. Or le nombre de points des objets de TOSCA peut varier de cinquante mille points à cinq mille. Les trois méthodes de l'état de l'art calculent en chaque point un descripteur géométrique et comparent les points définies entre eux. Notre méthode produit une représentation qui est considérée comme une image, c'est à dire que même les points sur la représentation 2D qui n'ont pas une valeur définie sont parcourue lors de la comparaison et un histogramme est produit avec ses voisins. De plus le descripteur calcul le cardinal des valeurs convexes et concaves. En d'autres termes le descripteur CCvx est corrélé au nombre de points. Ces deux raisons font que notre méthode est robuste à la comparaison d'objets avec

des résolutions différentes. Il est important de noter que ce problème n'apparaît pas pour les méthodes avec une comparaison globale car le descripteur n'est pas définie sur chacun des points.

Malgré un descripteur simple, ne possédant que deux valeurs entières, CCvx est la méthode la plus lente pour comparer. Ceci est en partie dû au calcul de l'histogramme lors de la comparaison. Cette approche avait été choisie pour pouvoir stocker un grand nombre de descripteurs pré-calculés sur une base de donnée en ligne. Les fichiers de stockage des cartes de convexité sont effectivement léger avec 11kB comparés aux autres fichiers des autres méthodes proches de 1mB. Le temps de calcul étant trop long comparé aux autres méthodes, dans la suite cette optimisation de stockage est abandonnée pour pouvoir diminuer le temps de comparaison.

3.5 Méthode globale à locale sur un jeu de données de protéines

3.5.1 Résumé

La méthode globale Shape Similarity System driven by Digital Elevation Models[8] a montré de bons résultats sur le jeu de données TOSCA. L'idée est donc de faire une première classification du jeu de données avec la méthode globale puis affiner les résultats avec la méthode locale décrite ci-dessus sur le top 4 des objets donné par la méthode globale. Ceci permettant aussi de palier au temps de comparaison des cartes de convexité.

Le jeu de données, appelé Shrec 2017, est un jeu de données de protéines développé pour le *workshop* 3DOR [79] contenant 10 protéines *query* (cf Figure 3.16). Les protéines *query* sont comparées à 5484 autres protéines comprenant 2 des 10 protéines *query*, la protéine *query* q11 et q14.

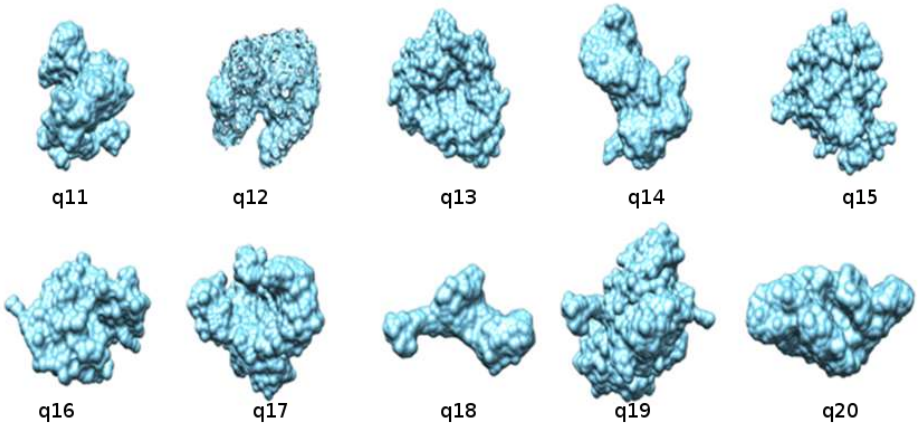


FIGURE 3.16 – Les 10 protéines *queries* du jeu de données Shrec 2017

Le graphique du coefficient de corrélation de premier rang pour chacune des protéines (cf Figure 3.17) montre que l'approche globale à locale permet de retrouver l'identité lorsqu'elle se trouve dans le jeu de données de SHREC17 ce qui était une caractéristique manquante de la méthode globale. En revanche la valeur du coefficient corrélation reste similaire à la méthode globale pour les autres protéines *query*.

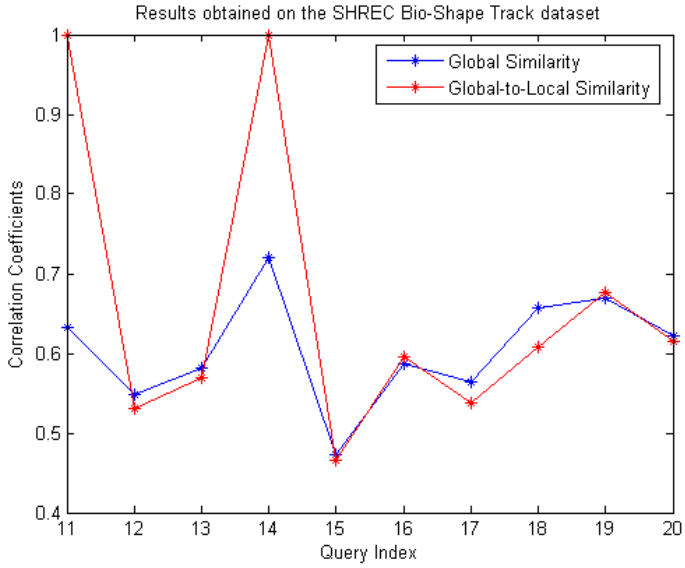


FIGURE 3.17 – Coefficient

Le besoin de combiner l'approche locale à l'approche globale est dû à une comparaison lente de

la méthode locale. C'est un des points qui est amélioré avec les méthodes présentées dans la suite de ces travaux. Le descripteur de convexité ne possède que deux valeurs pour discriminer les objets. Ces deux valeurs permettent une taille de stockage minime mais ne suffisent pas pour discriminer les protéines. Ceci s'explique par le fait que la surface d'une protéine possède moins de caractéristiques discriminantes qu'un objet classiquement utilisé dans le domaine de la vision par ordinateur tel une chaise, une main ou un visage. C'est ce qui motive l'utilisation de descripteurs à valeurs continues dans les autres méthodes présentées dans la suite de ces travaux.

Global-to-Local Protein Shape Similarity System driven by Digital Elevation Models

Daniela Craciun*, Jeremy Sirugue* and Matthieu Montes*
*Conservatoire National des Arts et Métiers, Laboratoire GBA, EA 4627
2 rue Conté, 75003 Paris, France, Email: firstname.lastname@cnam.fr

***Abstract*—We are currently developing a bio-shape similarity system for supplying high-throughput protein shape similarity applications within massive datasets. The proposed system is powered by a global-to-local shape similarity system which exploits shape elevation and local convexity attributes. In the first step, a global similarity is computed between the shape descriptors associated to each protein input. The procedure outputs best N similarities chosen by the user, within a *query-to-cluster* approach. The second stage is a patch-based local similarity computation method which is designed to find the best similar target from the cluster for supplying *query-to-target* protein retrieval applications. The local patch-based similarity comparison benefits of a multi-CPU implementation, offering thus fast query search capabilities within massive datasets. Experimental results on the SHREC 2017 BioShape dataset [4] composed of 5484 models, illustrate the effectiveness of the proposed system.**

I. Introduction and Motivation

Structural biologists face a rapid growing of the number of protein structures in the Protein Data Bank [1] (130000 structures in 2017) which induces a considerable need for fast protein similarity search methods able to screen such large databases in a reasonable amount of time. Since protein 3D structures are more conserved during evolution than their respective amino acid sequences [7], sequence-based protein similarity search methods fail to retrieve structural homologs that share low sequence homology. Protein structural similarity search methods such as DALI [11], CE [12], FAST [13] or FATCAT [14] that use different methods to align protein structures and compute their similarity, notably using the Root Mean Square Deviation of their alpha carbons. Most of the existing protein comparison methods share the major limitation that they are too computationally expensive to search a large protein structure database in a reasonable amount of time [15].

In this paper, we introduce a shape-based method which allows to perform high-throughput protein classification without relying on human operator intervention. In addition, the proposed method is designed to perform fast query search within massive datasets. The present research work introduces a global-to-local framework designed in a complementary fashion, along with a multi-CPU implementation, in order to cope with rapidity constraints. Our paper is organized as follows: Section II describes the proposed global-to-local **P**rotein **S**hape **S**imilarity **S**earch **S**ystem (**PS4**), followed by Section III which presents experimental results and the performance evaluation on the SHREC 2017 (SHape REtrieval Contest) BioShape dataset [4]. Finally, Section IV concludes the present research work and gives main perspectives.

II. Proposed Protein Shape Similarity Search System (PS4)

The proposed **P**rotein **S**hape **S**imilarity **S**earch **S**ystem (**PS4**) is composed of two main stages: the first stage is performed for each shape and consists in the **M**acromolecular **S**hape (**MS**) representation as a **D**igital **E**levation **M**odel (**DEM**), encoded over a 2D grid. The second stage corresponds to the shape comparison phase which is supplied via a global-to-local framework relying on the MS-DEMs descriptors. Both stages are summarized through the following two sections.

A. Representing Macromolecular Shapes as Digital Elevation Models

Macromolecular triangular surface computation. The shape representation algorithm applies the EDTSurf [2] technique to generate the macromolecular surface (MS) from the input data. The algorithm exploits the Vertex Connected Marching Cubes and the Euclidean Distance Transform to generate the triangular mesh which is kept for further processing.

Digital Elevation Model descriptor computation. The present work exploits the DEM concept traditionally employed in cartography for representing Earth's surface from terrain elevation data [8]. The algorithm starts by applying the mesh flattening procedure introduced in [3], which maps the mesh onto the unit sphere using the Laplace-Beltrami operator [10]. The spherical mapping provides a valid solution for any genus-0 triangle meshes, being adapted in our current research work. In the second step, the unit sphere is projected onto a 2D spherical panoramic grid and the elevation values of the input mesh are assigned to each 2D location of the panoramic grid. This results in a global descriptor which encodes shape's elevation, while providing topology and fast comparison over a 2D grid space. The DEM descriptor is stable under rotations and translations variations of the input mesh and varies in presence of scale transformations. A detailed description of the algorithm can be found in [5]. The final output is the digital elevation model associated to the macromolecular surface, noted MS-DEM. Figures 1 (a)-(d) illustrate the results obtained for a target belonging to the protein pool of the SHREC 2017 BioShape track [4].

978-1-5386-0706-0/17/$31.00

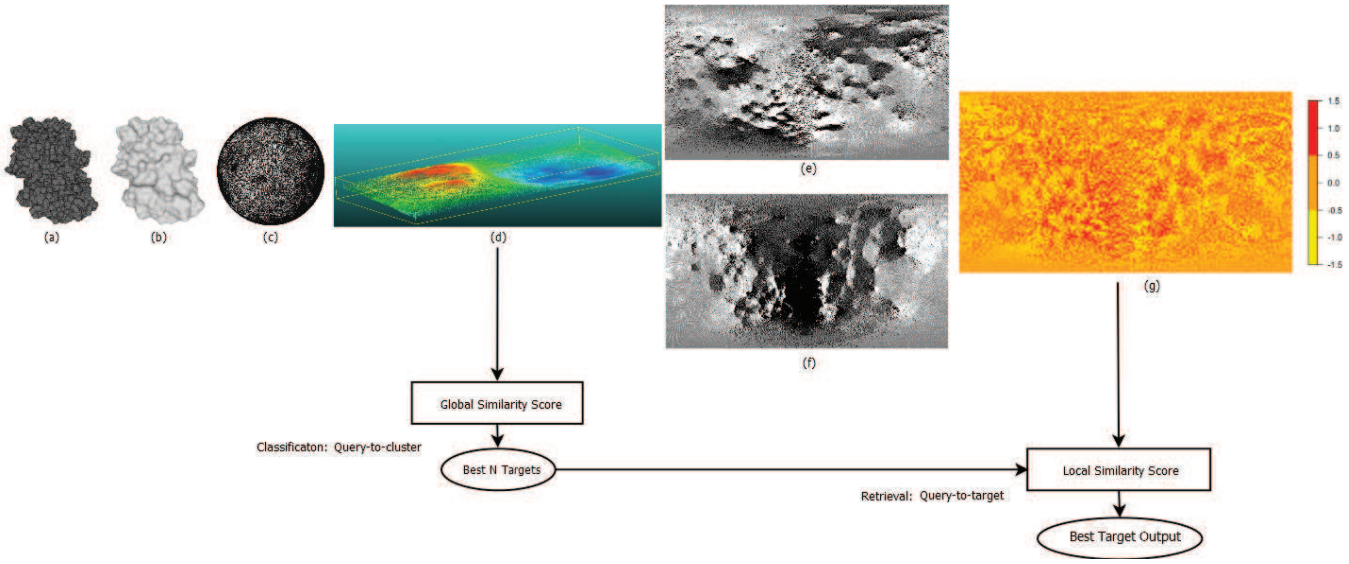


Fig. 1. Overview of the global descriptor computation stage (input model: $m10001$) and the integration within the global-to-local similarity computation framework of the **PS4** prototype: (a) input data $N_p = 187866$ points, $N_t = 357840$ triangles; (b) macromolecular mesh generated by EDTSurf [2]: $N_p = 86079$ points, $N_t = 172154$ triangles; (c) spherical mapping output [3]: $N_p = 86079$ points, $N_t = 172154$ triangles; (d) MS-DEM output: $N_p = 86089$ points, bounding box dimensions: $[472, 257, 36.112]$, (e) Gradient G_x computed for the MS-DEM illustrated in Figure (d); (f) Gradient G_y computed for the MS-DEM illustrated in Figure (d); (g) Convexity map result corresponding to the MS-DEM illustrated in Figure (d).

B. Global-to-Local Protein Shape Similarity Computation

The proposed protein similarity search system relies on a global-to-local shape similarity search framework designed in a complementary fashion. The global similarity stage allows to perform fast comparisons, being therefore employed as a first stage to search for best similar candidates w.r.t. the query. In exchange, the local stage provides a finer comparison, being suitable for selecting the best rank similarity among targets clustered at the global comparison stage.

Global Comparison of MS-DEMs. The MS-DEM shape descriptor is used along with different global distances for supplying the protein shape similarity computation stage. The present research work evaluates the Mean Absolute Differences (d_{MAD}) and the Root Mean Square Deviation (d_{RMSD}) distances. They are measured over the points belonging to the 2D grids. For input meshes with different number of points, distances are computed over the minimum number of points computed between the query and the target meshes. In absence of scale variation, the similarity score is valid for meshes with a similar number of points, belonging to the same class. In this configuration, the global comparison stage was compared w.r.t. state-of-the-art shape retrieval algorithms and a detailed peformance evaluation can be found in [5]. In the present research work, the shape comparison stage outputs the dissimilarity matrix which is exploited for extracting the best N similarities chosen by the user w.r.t. the query. Figure 2 illustrates an example of the *query-to-cluster* procedure output which provides the best $N = 4$ similar targets w.r.t. the query q_{11}.

Patch-based Local Comparison of MS-DEMs via Convexity Maps. The local shape comparison stage takes as input the best $N = 4$ similarities output by the global comparison stage and finds the best rank similarity w.r.t. the query. The local similarity computation is performed through the use of convexity maps which are computed from each MS-DEM descriptor. The MS-DEMs are exploited for computing the gradient along X and Y directions, noted G_x and G_y, respectively. The convexity coefficients are computed by identifying gradient extremum values (minimas and maximas) and by assigning convexity labels to each point belonging to the MS-DEM descriptor. Figures 1 (e), (f) and (g) illustrate an example of the G_x, G_y and the associated convexity map, noted C_{map}, respectively.

The pairwise patch-based comparison is performed by extracting patches from the convexity maps corresponding to the query and the target meshes. For each patch extracted in the query, a local dissimilarity measure is computed w.r.t. each patch extracted from the target mesh. The d_{MAD} distance is computed between each query patch and all patches belonging to the target. Similar patches selected w.r.t a threshold value are further considered for computing the overall dissimilarity score between the query and the target meshes. In our experiments, it was observed that a patch ray of 7 pixels provides accurate results in a reasonable amount of time. Moreover, the results let us concluded that higher patch ray values lead to a computationally expensive framework without improving considerably the accuracy. Selected similar patches have less than 20% dissimilarity compared to the maximum d_{MAD} distance computed over all the compared patches.

In order to avoid the computational burden of the local comparison stage, the convexity maps are computed from MS-DEMs with a reduced resolution (by a factor of 2). In addition,

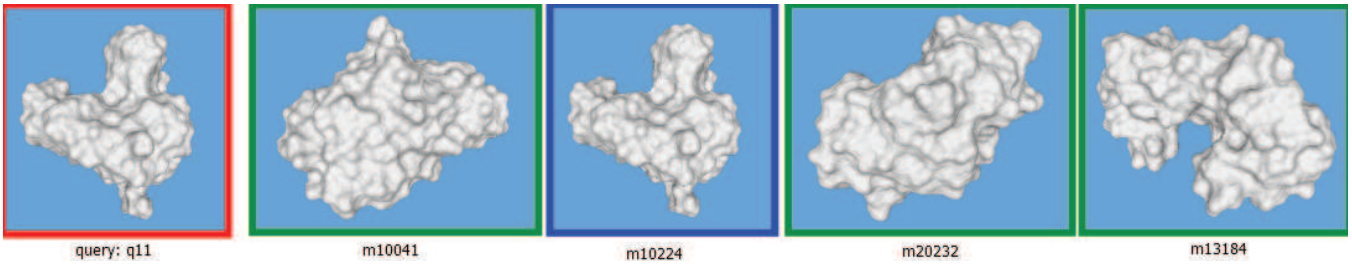


Fig. 2. Output generated by the *query-to-cluster* procedure for query q_{11}: Global similarity output obtained using the d_{MAD} distance. The procedure outputs $N = 4$ best similarities (illustrated from left to right), the identity query (model: $m10224$) is found as the 2^{nd} top similarity (contour emphasized in blue color).

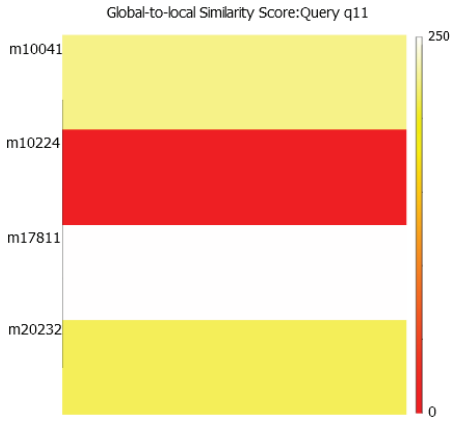


Fig. 3. Output generated by the *query-to-target* procedure for query q_{11} via the patch-based local similarity comparison of convexity maps. Input: $N = 4$ top similarities generated by the global comparison stage (*query-to-cluster*) via the d_{MAD} distance. Output: best similar target obtained for q_{11}: $m10224$.

the local patch-based comparison benefits of a multi-CPU implementation. Figure 3 illustrates an example of the best similarity output found at the local comparison stage for the first query, q_{11}. As shown in Figure 2, the global similarity phase outputs the best $N = 4$ similar targets, with the identity query found as the 2^{nd} top-rank similarity (blue contour). Figure 3 illustrates that the identity target of query q_{11}, noted $m10224$, was found as the best similar target by the local similarity comparison phase.

III. Results and Performance Evaluation

This section presents the performance evaluation of the proposed framework, **PS4**, on the dataset made available for the SHape REtrieval Contest (SHREC 2017) BioShape track [4]. We analyse the proposed system in terms of accuracy, runtime and memory usage.

Dataset. The dataset consists in 10 queries (q_{11},...,q_{20}) (selected from the molecule of the Month collection [9]) and 5484 targets. In order to allow accurate validation, for two queries (q_{11} and q_{14}), identic protein were included in the target set.

Evaluation measures. In the SHREC 2017 BioShape track, the protein similarity evaluation relies on the 3DZM method [6] which measures the accuracy by comparing the correlation coefficients computed between the query and each target. More details about the dataset and the evaluation protocol can be found in the SHREC 2017 BioShape track [4].

A. Accuracy Evaluation

As presented in the SHREC 2017 BioShape track [4], the accuracy of the proposed framework is evaluated w.r.t. the correlation coefficients obtained by the 3DZM method [6]. In order to analyse the behaviour of the local stage, we provide an evaluation of the **PS4** framework employed in both modes: global similarity computation and global-to-local computation mode. Figure 4 illustrates the results generated by the global stage (employing the d_{RMSD} distance) and the global-to-local framework. It can be observed that for queries q_{11} and q_{14}, the global-to-local approach retrieved successfully identity shapes as the best rank similarity.

The patch-based local comparison stage improves the average correlation coefficient, attaining 0.6598, compared to 0.6051 provided by the global comparison stage alone. In addition, while for some queries (i.e. $q_{12}, q_{13}, q_{17}, q_{18}$) the global minimum was lost by the local comparison stage, for the remaining queries ($q_{15}, q_{16}, q_{19}, q_{20}$), the global minimum was correctly maintained. While providing a rapid protein retrieval framework, there is still room for improving the local patch-based similarity computation stage by searching more stable and intrinsic features w.r.t. the dataset (shape, size, resolution).

B. Runtime and memory usage

The proposed algorithm is implemented in C/C++ and runs on a 64b Linux machine equipped with 32Gb of RAM memory and an Intel Xeon running at 2.3 GHz. Less computationally expensive stages, i.e. the descriptors' computation (MS-DEM, C_{map}) and the global comparison, are designed for simple-CPU implementation. The most expensive stage, i.e. the pair-wise patch-based local comparison of convexity maps benefits of a multi-CPU implementation.

Simple-CPU global comparison of MS-DEMs. The first row of Table I resumes the average runtime for extracting the MS-DEM descriptor for one model belonging to the protein pool of the SHREC 2017 BioShape track [4]. The computation time for comparing one query against the entire protein pool

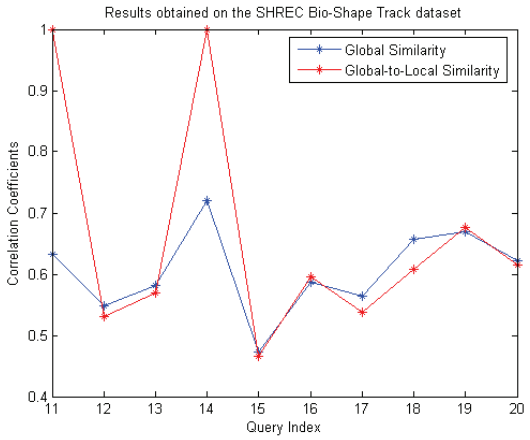


Fig. 4. Experimental results generated by the proposed system **PS4** on the SHREC 2017 BioShape dataset [4], results comparison: Global similarity vs. Global-to-Local similarity computations; Global similarity output generated by the d_{RMSD} distance computed between MS-DEM descriptors.

TABLE I
AVERAGE RUNTIME (SECONDS) OBTAINED FOR SIMPLE-CPU IMPLEMENTATION OF EACH MODULE COMPOSING THE SHAPE DESCRIPTOR EXTRACTION PROCEDURE ILLUSTRATED IN FIGURE 1.

Module	(b)	(c)	(d)	(e), (f)	(g)
CPU (s)	3.34	2.65	0.12	0.03	0.015
RAM (Mb)	8.08	6.6	**1.47**	2.3	**1.14**

takes is in average of 2.3502 seconds and 3.3518 seconds for d_{MAD} and d_{RMSD} distances, respectively.

Multi-CPU Patch-based Local Comparison of Convexity Maps. In order to avoid the computational burden, the pairwise local comparison procedure is implemented on $N_{CPU} = 24$ cores, taking in average 1 min 15 sec for providing the best similar target. When compared to the simple-CPU implementation, the parallelization allows to reduce the runtime by an average factor of 46. This result emphasizes the fast query search capability detained by the proposed protein shape similarity search system, **PS4**.

Memory usage. The average memory usage for storing the MS-DEM descriptor is 1.058 kb. The second row of Table I illustrates the memory usage for the target $m10001$ depicted in Figure 1. The overall memory usage required by the MS-DEM descriptor and the associated convexity map C_{map} is 2.61 Mb. When compared to the input mesh storage, (Figure 1 b)), the proposed descriptors reduce the memory usage by a factor of 3, being therefore suitable for processing massive datasets.

IV. CONCLUSIONS AND FUTURE WORK

This paper presented the **P**rotein **S**hape **S**imilarity **S**earch **S**ystem, (**PS4**), a global-to-local geometric-based protein similarity framework designed for supplying fast query search within massive datasets. The main features which ensure the rapidity of the proposed system operate at two levels: (i) software architecture: global and local shape similarity computation for fast *query-to-target* computation, and (ii) software implementation: multi-CPU optimization of local comparison. This gives rise to a protein similarity system designed in a complementary fashion: the global comparison stage allows rapid selection of best candidates, while the local stage provides best target selection among them. Experimental results on the SHREC 2017 BioShape dataset [4], containing 5484 models, demonstrate that our approach detains fast query search capabilities for supplying high-throughput *query-to-target* protein retrieval applications. Nevertheless, while solving the rapidity issue, there is still room for improving the accuracy of the local similarity stage. Research perspectives are concerned with the accuracy improvement of the local stage, in presence of various shape types, while still maintaining the fast query search capability detained by the proposed system, **PS4**.

ACKNOWLEDGMENT

The present research work is supported by the ERC ViDOCK Grant no. ♯640283 from the European Research Council Executive Agency.

REFERENCES

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N Shindyalov and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1), pp. 235-242, 2000.
[2] D. Xu and Y. Zhang. Generating Triangulated Macromolecular Surfaces by Euclidean Distance Transform. *PLOS ONE*, 4(12), pp. 1–11, 2009.
[3] S. Angenent, S. Haker, A. Tannenbaum and R. Kikinis. On the Laplace-Beltrami operator and brain surface flattening. In *IEEE Transactions on Medical Imaging*, 18(8), pp. 700–711, 1999.
[4] N. Song, D. Craciun, C. W. Christoffer, X. Han, D. Kihara, G. Levieux, M. Montes, H. Qin, P. Sahu, G. Terashi and H. Liu. SHREC'17 (SHape REtrieval Contest) BioShape Track: Protein Shape Retrieval. In *Eurographics Workshop on 3D Object Retrieval*, 2017.
[5] D. Craciun, G. Levieux and M. Montes. Shape Similarity System driven by Digital Elevation Models for Non-rigid Shape Retrieval. In *Eurographics Workshop on 3D Object Retrieval*, 2017.
[6] M. Novotni and R. Klein: 3D Zernike descriptors for content based shape retrieval. In *Proceedings of the 8th ACM Symposium on Solid Modeling and Applications*, pp. 216-225, 2003.
[7] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. In *The EMBO Journal*, 5(4), pp. 823–826, 1986.
[8] C. L. Miller and R. A. Laflamme. The digital terrain model - theory and application. In *Photogrammetric Engineering*, pp. 433–442, 1958.
[9] D. S. Goodsell, S. Dutta, C. Zardecki, M. Voigt, H. M. Berman and S. K. Burley. The RCSB PDB Molecule of the Month: Inspiring a Molecular View of Biology. In *PLOS Biology*, 13(5), pp. 1–12, 2015.
[10] G. Craig, G. Xianfeng and A. Sheffer. Fundamentals of Spherical Parameterization for 3D Meshes. In *ACM SIGGRAPH*, pp.358–363, 2003.
[11] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. In *Journal of Molecular Biology*, 233(1), pp. 123–138, 1993.
[12] I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. In *Protein Eng*, 11(9), 739–747, 1998.
[13] J. Zhu and Z. Weng. FAST: a novel protein structure alignment algorithm. In *Proteins*, 58(3), pp. 618–627, 2005.
[14] Y. Ye and A. Godzik. Flexible structure alignment by chaining aligned fragment pairs allowing twists. In *Bioinformatics*, Suppl. 2, pp. 246–255, 2003.
[15] S. Mezulis, M. J. Sternberg and L. A. Kelley. PhyreStorm: A Web Server for Fast Structural Searches against the PDB. In *Journal of Molecular Biology*, 428(4), pp. 702–708, 2016.

Chapitre 4

Améliorations de la méthode basée sur les cartes de convexité

Contenu

L'utilisation de cartes de convexité sur les protéines n'a pas apporté de résultats satisfaisant comparé à la méthode globale. De plus le temps de calcul est un problème si l'on souhaite comparer des jeux de données de la taille de celui de SHREC 2017. C'est suite à la recherche de solutions à ces problèmes que les cartes de convexité sont devenues les cartes de courbures.

Ce chapitre présente différents problèmes identifiés et la solution proposée. Il est divisé en trois catégories : les changements dans la représentation, les modifications du descripteur et la stratégie de comparaison du descripteur. L'évaluation de ces changements se fera dans le chapitre suivant avec les cartes de courbures.

4.1 Représentation

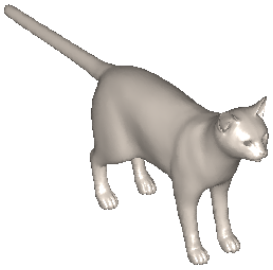
Une comparaison dense, c'est-à-dire une prise en compte de tous les pixels, entre deux CCvx dure environ une minute ce qui est un inconvénient de la méthode dans une perspective de *big data*. C'est pourquoi on diminue la résolution des de la grille lors de la projection 2D.

La taille de la grille est divisée par 4 (divisé par 2 pour la longueur et par 2 pour la largeur) pour éviter de perdre trop d'informations avec la diminution de la taille.

4.2 Descripteur

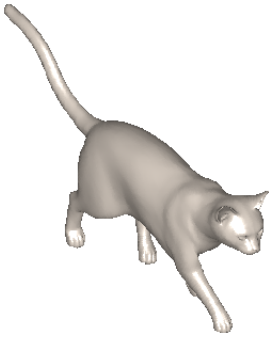
4.2.1 Moyenne des valeurs des DEMs

Un problème apparaît en regardant les DEM des trois chats *cat0*, *cat1* et *cat10* du jeu de données TOSCA (cf Figure 4.1). En une coordonnée d'une DEM, il est possible d'avoir plusieurs valeurs d'élévation (cf Figure 4.1e). Dans la version précédente de la méthode, la première valeur accessible en une coordonnée était la valeur sélectionnée. Ceci est dû au pas de projection des coordonnées sphériques de la sphère unitaire sur la grille. Dans cette nouvelle méthode, la **moyenne des valeurs en une position particulière** est calculée puis sélectionnée. Les CCvx possèdent maintenant des parties plus homogènes (cf Figure 4.2d et Figure 4.2d) comparé aux CCvx de l'ancienne version (cf Figure 4.2a et Figure 4.2c), ceci est particulièrement visible dans les zones denses. Ces changements se combinent bien avec la diminution de la résolution qui va concentrer de nombreuses valeurs dans les zones denses.

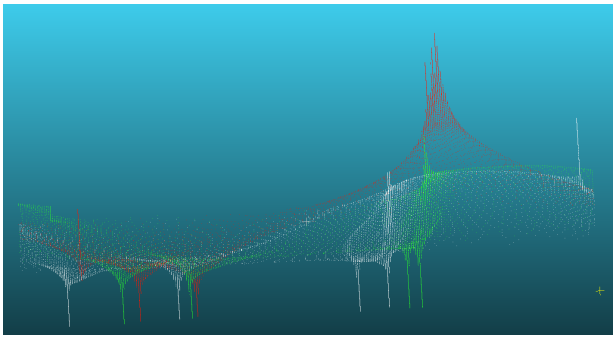


(a) Cat0

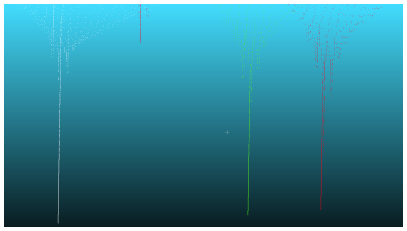
(b) Cat1



(c) Cat10

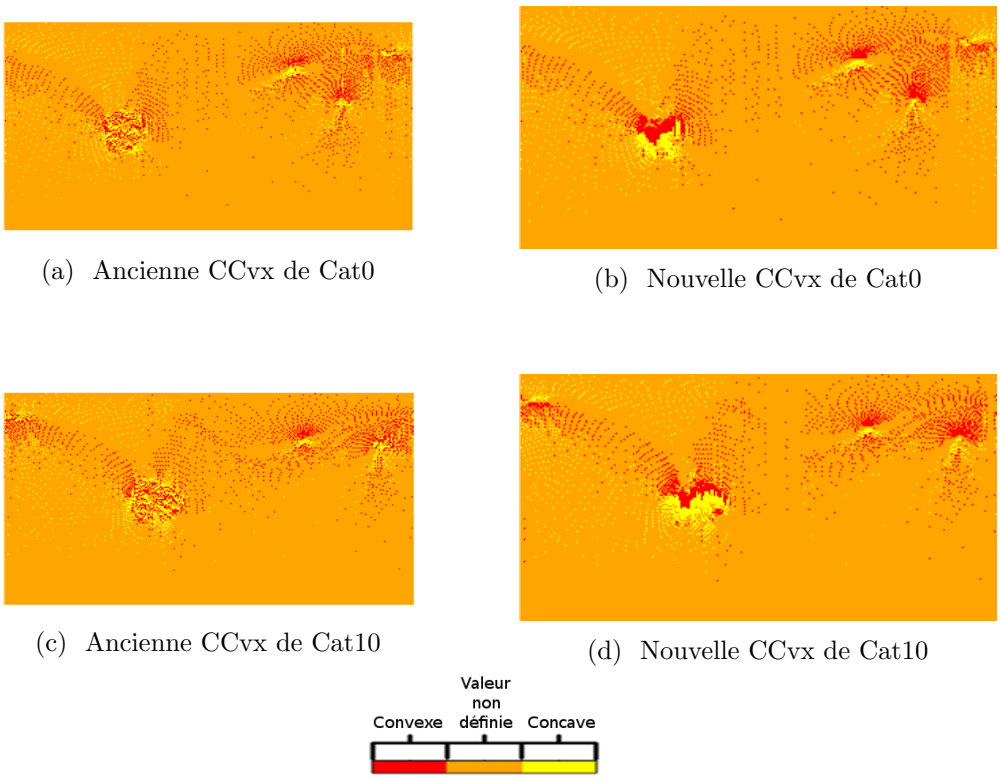


(d) Les DEM pour cat0 (vert), cat1 (rouge), cat10 (blanc)



(e) Zoom sur les DEM pour cat0 (vert), cat1 (rouge), cat10 (blanc)

FIGURE 4.1 – Trois objets *cat* du jeu de données TOSCA et leur DEM associées



(a) Ancienne CCvx de Cat0

(b) Nouvelle CCvx de Cat0

(c) Ancienne CCvx de Cat10

(d) Nouvelle CCvx de Cat10

FIGURE 4.2 – Ancienne et nouvelle CCvx de cat0 et cat10 du jeu de données TOSCA

4.2.2 Histogramme des valeurs continues

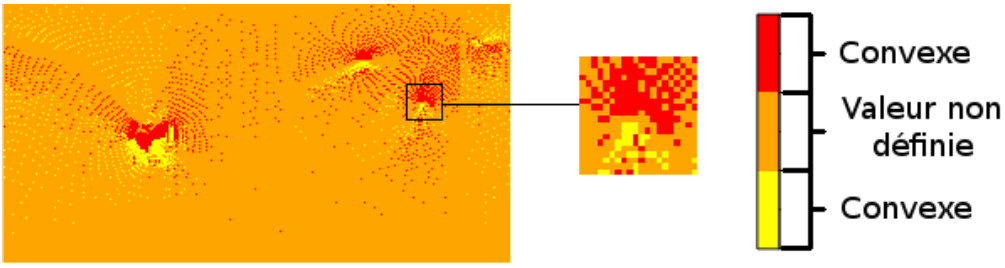
Un autre problème majeur du descripteur proposé est le manque de discrimination entre les patchs. Ce manque de discrimination est expliqué par l'utilisation de nombres entiers qui sont ensuite résumés sous forme d'histogramme. Cette représentation élimine un trop grand nombre d'informations par rapport aux DEM.

L'utilisation de valeurs binaires et entières étaient justifiées par la recherche d'une manière de stockage concise et de plus la distance de Hamming avait été envisagée comme une distance possible.

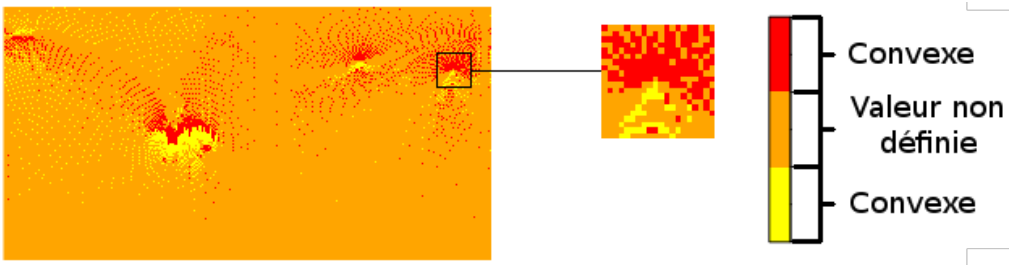
Certains patchs ont plus de correspondances que d'autres et ceci est expliqué en partie par l'utilisation d'entiers et d'histogrammes qui diminuent la discrimination des patchs. La perte d'information est trop importante pour utiliser à la fois des entiers et des histogrammes.

L'utilisation de valeurs entières et d'histogrammes a les avantages de diminuer la taille de stockage et le temps de calcul. Un bénéfice de l'utilisation d'histogramme est d'atténuer les erreurs de petites variations non-rigides de la forme. L'utilisation d'une comparaison point à point à l'intérieur d'un patch

peut créer un score plus bas qu'il ne devrait l'être si un décalage a eu lieu. Un décalage est produit par une transformation non rigide locale ou par une erreur de précision lors de l'échantillonnage comme on peut le voir en Figure 4.3. L'histogramme d'un patch ne possède pas la propriété de spatialité des points ce qui apporte l'avantage de ne pas être affecté par un décalage de point à l'intérieur d'un patch. C'est pour cette raison que les histogrammes sont préférés aux valeurs discrètes entières qui ont seulement l'avantage de diminuer la taille de stockage.



(a) Zoom sur la CCvx de Cat0



(b) Zoom sur la CCvx de Cat10

FIGURE 4.3 – Zoom sur la CCvx cat0 et cat10 du jeu de données TOSCA. Zoom sur la jambe de l'objet *cat*

4.2.3 Valeur de courbure

Les histogrammes sont utilisés pour diminuer l'erreur produite par une méthode point à point mais comme noté précédemment, la valeur binaire de la convexité crée un manque d'information. Une valeur continue a donc été introduite pour servir de coefficient sur les histogrammes tandis que les histogrammes restent sur une valeur discrète binaire. La valeur de l'élévation de la DEM a été choisie comme première valeur de coefficient mais cette amélioration ne résout pas le problème de précision. Les raisons sont que le coefficient est un ajout sur l'information binarisé sous forme de convexité, donc le problème de perte d'informations sur la binarisation est toujours présent. De plus ce coefficient

repose trop sur les valeurs des DEM et donc l'utilisation de CCvx serait un ajout superflu. C'est pourquoi la valeur de coefficient a été testée puis abandonnée.

La courbure signée est une valeur continue proche du descripteur de convexité proposé. Une courbure positive représente une partie convexe et une courbure négative est la partie concave d'un objet. De plus la courbure apporte l'information de l'intensité de la convexité sur la surface.

La valeur de courbure est donc introduite et permet de passer d'une valeur binaire à une valeur continue. Cette valeur permet aussi de garder l'idée de convexité développée.

4.2.4 Suppression de la DEM

Les valeurs de courbures sont calculées à partir de la carte de DEM. Les valeurs d'élévation des DEM et les valeurs de courbure sont deux valeurs de nature proche, ce sont des valeurs continue réelles extraites de la topologie de l'objet. Calculer la courbure à partir des valeurs d'élévation des DEM ajoute une étape supplémentaire où des erreurs d'imprécision peuvent s'ajouter diminuant la valeur de l'information finale.

C'est pourquoi la courbure est directement calculée sur l'objet 3D étudié puis projetée dans l'espace 2D.

4.3 Comparaison

4.3.1 Meilleur score réciproque

Un score réciproque entre paires de patchs a été implémenté. Pour être inclus dans le score final, un patch P_1, possédant un meilleur *match* avec un second patch P_2, doit aussi être le meilleur *match* du patch P_2 (cf Figure 4.4).

Soit $S_p(P_1(a), P_2(b))$ le score entre deux patchs P_1 et P_2 centré en respectivement en a et b.

Soit k_T et l_T deux points de la représentation 2D T et k_V un point de la représentation 2D V vérifiant

$$k_V = \arg\min_{k \in V} S_p(k_T, k)$$

et

$$l_T = \arg\min_{l \in T} S_p(l, k_V)$$

Si $k_T = l_T$ alors c'est un meilleur match réciproque.

Le meilleur score réciproque permet de diminuer le nombre d'erreurs en empêchant un patch de se retrouver associer à plusieurs patchs. Dans certains cas, un patch déséquilibre trop le score en ayant un grand nombre de *matchs*. Il permet aussi d'apporter l'unicité de *match* pour chaque patch.

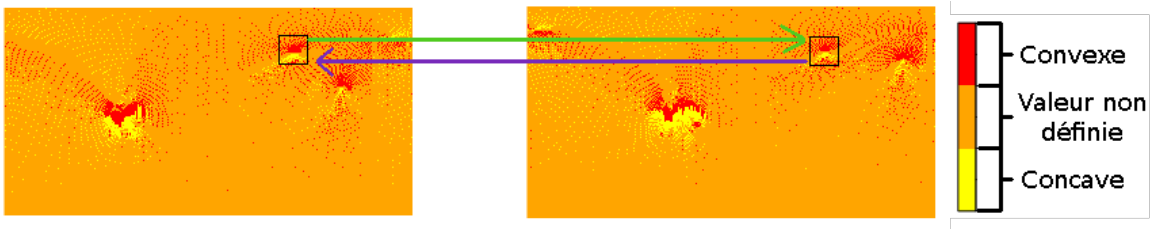


FIGURE 4.4 – Un *match* réciproque entre deux patchs des CCvx de cat0 et cat10 du jeu de donnée TOSCA

4.3.2 Multiview

Une variante de la méthode décrite est la comparaison de plusieurs vues de la sphère. Pour limiter la déformation aux pôles lors du passage de la sphère au plan, plusieurs vues de la carte son calculées. Pour obtenir ces différentes vues, une rotation est appliquée selon l'axe x, y et z de l'espace cartésien de la sphère (cf Figure 4.5). Cette rotation de $\frac{\pi}{4}$ permet pour chaque vues de centrer une partie différente de la surface autour de la ligne coupant horizontalement en deux les cartes de courbures. Cette ligne peut être comparée à l'équateur sur une carte.

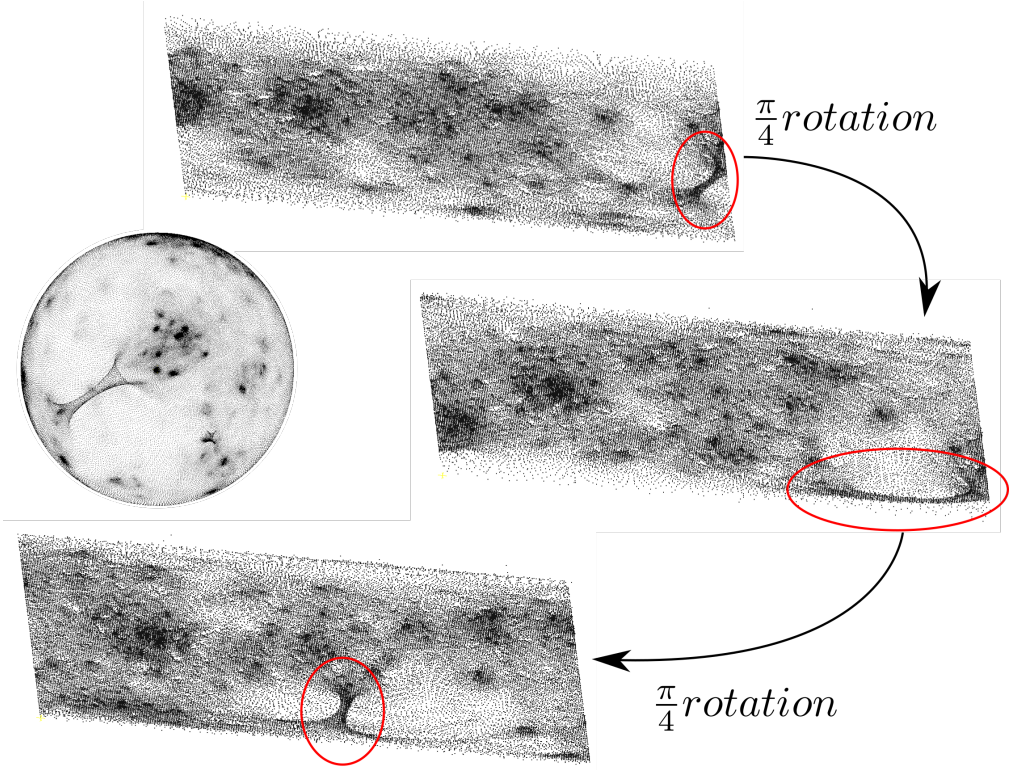


FIGURE 4.5 – Schéma de la rotation des pôles de la sphère selon les trois axes (ici selon l'axe y) utilisés pour le **multiview**

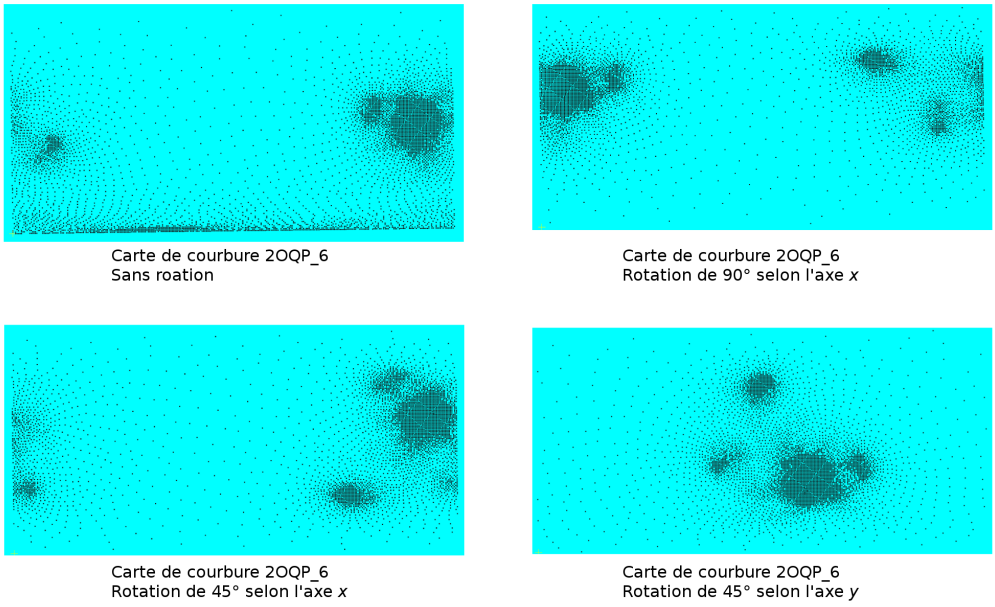


FIGURE 4.6 – Différentes cartes de courbure avec une rotation selon l'un des trois axes

Chapitre 5

Méthode basée sur les cartes de courbure

Contenu

5.1 Carte de courbure

Le descripteur développé et appelé **carte de courbure** (CM pour Curvature Map) qui est dans la continuité avec la convexité utilisée dans la version précédente de la méthode développée. En effet, la convexité peut être obtenue en se basant sur le signe de la courbure [93].

La convexité s'inspire de la topographie [94] [95] [96]. Les travaux de topographie ont pour objectif de classifier une zone géographique selon sa topologie dans le but de pouvoir étudier ensuite les cartes de cette zone géographique selon sa topographie. L'idée de la courbure est d'étendre le concept de convexité en donnant des informations topographiques [41] sur la surface tout en proposant une valeur continue.

Une CM est l'équivalent d'une DEM, la valeur d'élévation ayant été remplacée par la valeur de courbure de la surface 3D en entrée. Les CM sont obtenues en deux grandes étapes tel que décrit dans le schéma Figure 5.1. Dans un premier temps la courbure gaussienne K est calculée sur la surface. Ensuite la valeur de courbure est assignée à son point correspondant sur la carte 2D. La courbure gaussienne est calculée sur un maillage M. Pour un vertex (point) p de M la courbure gaussienne K est :

$$K(p) = \frac{A(p)}{3}(2\pi - \sum_{(i,j)\in V_p, i\neq j} \gamma_p(i,j)) \tag{5.1}$$

$A(p)$ est l'aire autour de p qui est la somme des triangle qui ont p pour point. V_p sont les points voisins (connectés par une arrête) de p et $\gamma_p(i,j)$ l'angle formé par les droites (i,p) et (j,p)

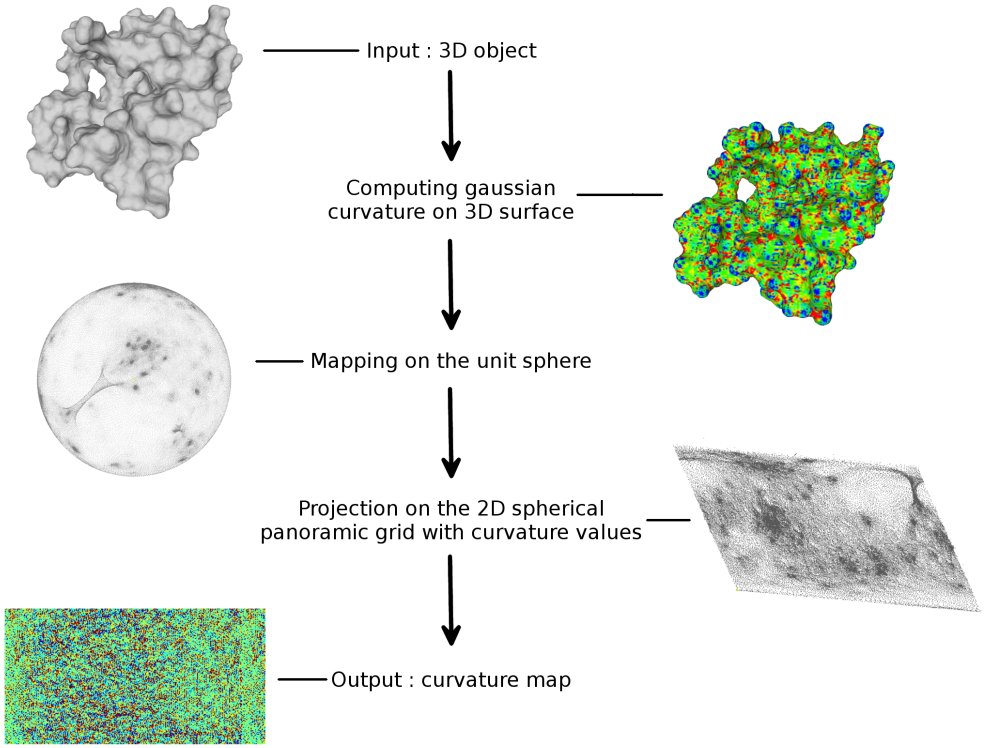


FIGURE 5.1 – Workflow de la création de cartes de courbure

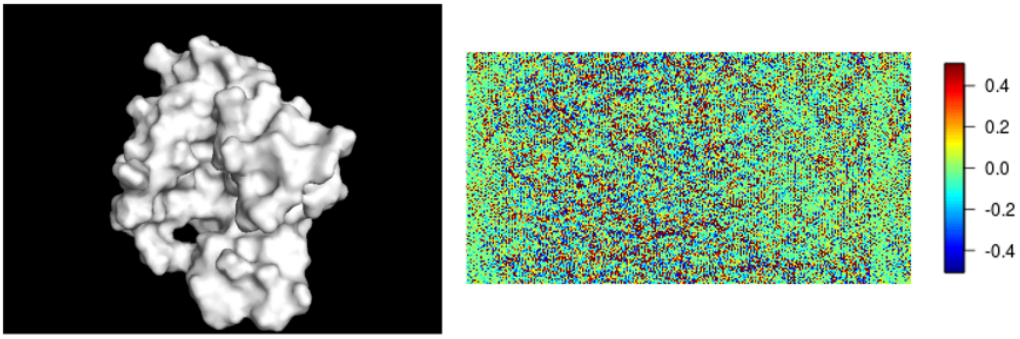


FIGURE 5.2 – Carte de courbure de la surface de la protéine CRABPII

Pour chacun des points de la carte de courbure, une fenêtre, appelée **patch**, centrée sur ce point est définie. Un histogramme est ensuite calculé pour ce point. On crée N intervalles réguliers, les *bins*, chacun représentant $\frac{1}{N}$ de toutes les valeurs de courbure de toutes les CM comparées. Les extrema de tous les histogrammes sont le minimum et le maximum des valeurs de courbures parmi toutes les cartes de courbures comparées. Chaque ***bin*** comptabilise les valeurs de courbure appartenant à l'intervalle

défini par le *bin* dans la zone du patch.

Une représentation de la création d'un histogramme est présentée dans la Figure 5.3.

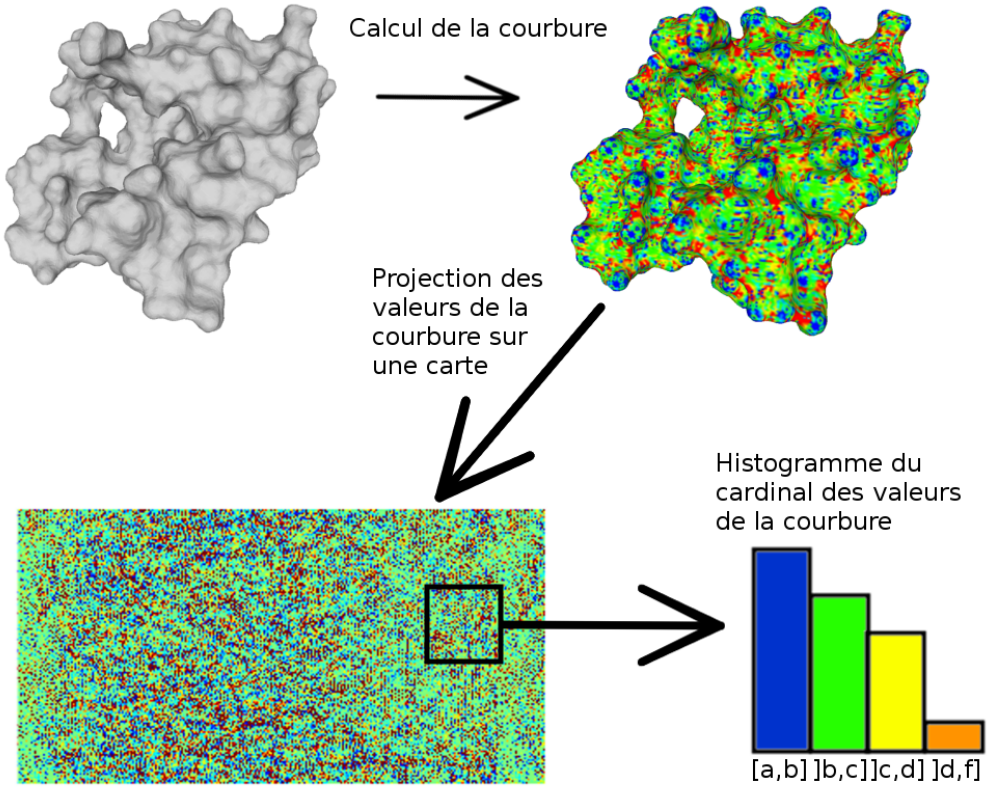


FIGURE 5.3 – Schéma du calcul d'un histogramme à partir d'une carte de courbure

5.2 Comparaison des cartes de courbure

La méthode proposée encode une caractéristique locale et donc implique un descripteur comparé localement, c'est-à-dire que chaque descripteur est comparé à tous les autres descripteurs. La comparaison des deux objets est similaire à la comparaison proposée pour les cartes de convexité.

Deux objets T et V sont comparés en recherchant les meilleures correspondances entre deux histogrammes de leur cartes de courbure respectives C_T et C_V. Pour comparer deux histogrammes H_{k_T} et H_{k_V} la **distance de Minkowski** est utilisée sur chacun des *bins* des histogrammes de l'objet *query* et cible. Le score, noté S_p, entre deux points k_T et k_V, respectivement un point de C_T and C_V, est

défini de la manière suivante :

$$S_p(k_T, k_V) = \sum_{i=1}^{N} |H_{k_T}(i) - H_{k_V}(i)|$$

Avec i le ième intervalle de l'histogramme.

Pour un point de la CM *query*, le point de la CM cible donnant le meilleur score est conservé. Tous les scores sélectionnés ainsi sont sommés pour donner le score final entre deux CM. Si C_T est composé de N_T points et C_V de N_V points, alors le score $S(T, V)$ de dissimilarité entre T et V est :

$$S(T, V) = \sum_{k_T=1}^{N_T} \min_{k_V \in C_V} S_p(k_T, k_V)$$

Le score peut être pondéré par un coefficient qui est l'inverse du nombre de points de la carte. Le score est alors le suivant :

$$S(T, V) = \frac{1}{N_T} \sum_{k_T=1}^{N_T} \min_{k_V \in C_V} S_p(k_T, k_V)$$

Le coefficient permet de normaliser le score. La normalisation permet de comparer des cartes de courbure avec une différence significative de nombre de points.

5.3 Traitement des données

5.3.1 Jeux de données

Pour travailler sur la méthode développée nous utilisons un jeu de données créé au sein du laboratoire (cf Figure 5.4). Ce jeu de données, appelé **DSV** pour DataSet Vidock, est composé de deux protéines avec une surface très proche (**CRABPII** et **LFABP**), d'une troisième protéine avec une structure similaire aux deux premières (**Synaptotagmin**) et d'une quatrième protéine n'ayant pas de similarité avec les trois premières protéines (**hIL**).

Les différentes structures des protéines obtenues par Résonance Magnétique Nucléaire (RNM) fournissent plusieurs transformations non-rigides de la protéine appelée conformations. Le jeu de données est donc composé de 4 protéines et 82 surfaces en prenant en compte les différentes conformations.

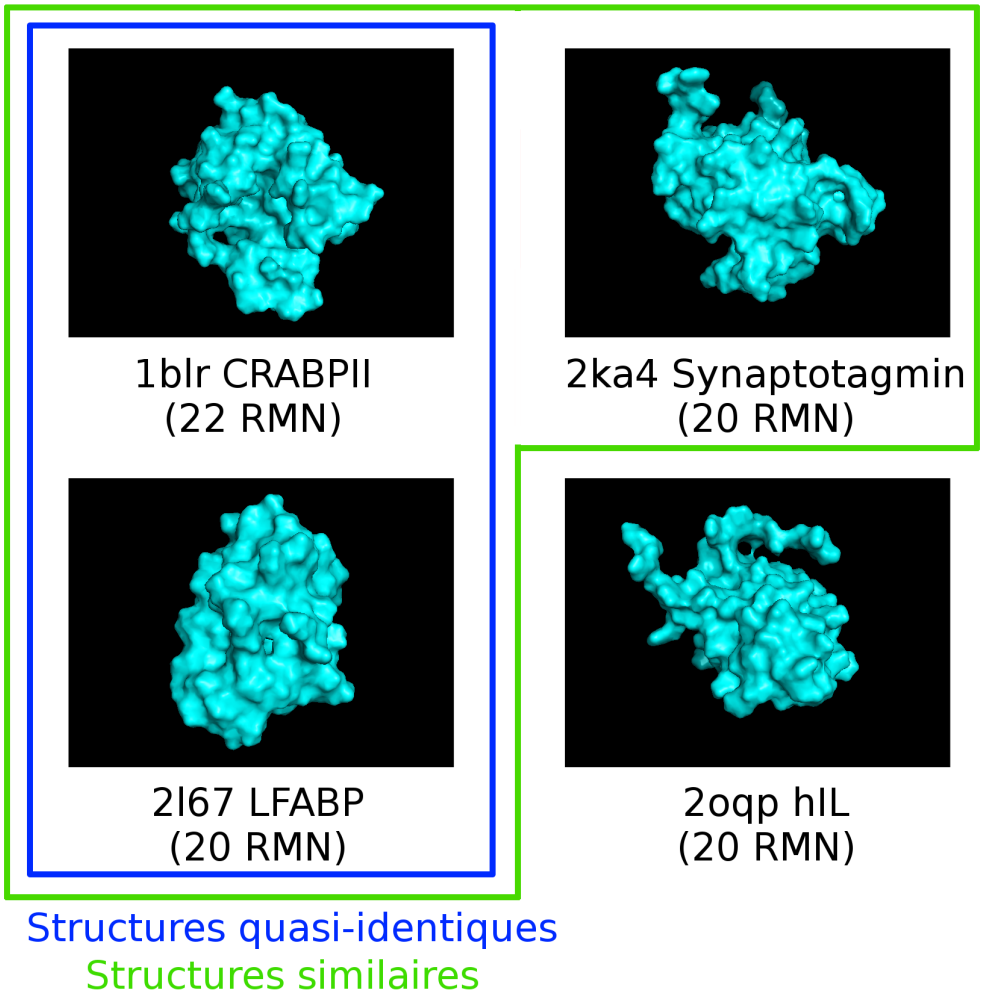


FIGURE 5.4 – Les 4 protéines du jeu de données **DSV** composé de 82 surfaces

Par nécessité un sous jeu de données est crée, appelé **DSV16**, composé des 4 protéines et de 16 conformations (4 conformations par protéine).

5.3.2 Représentation des résultats

La matrice de score de dissimilarité permet d'avoir une évaluation qualitative de la méthode de comparaison locale. L'intersection entre la ligne i et la colonne j représente le score de dissimilarité entre le i-ème objet et le j-ème objet. Un score de dissimilarité faible (rouge sur la matrice de score) indique des objets similaires et au contraire si la dissimilarité est élevée (blanc sur la matrice de score)

alors les objets sont peu similaires.

5.3.3 Mesures d'évaluations

Les mesures utilisées pour avoir une vision des performances sont le **Nearest Neighbour** (NN), **le First Tier** (FT) et le **Second Tier** (ST)[19] avec l'outil **Princeton Shape Benchmark**. Les courbes de **précision rappel** représentent l'évolution de la precision en fonction du rappel.

5.4 Résultats avec la méthode monoview

La méthode avec une seule vue de la sphère unitaire est étudiée ici. Deux approches sont comparées dans cette section, la comparaison avec ou sans le coefficient de normalisation.

5.4.1 Résultats sur DSV

Les résultats avec la méthode *monoview* et sans coefficient (cf Figure 5.5) semblent tendre vers une homogénéité des scores à l'exception de deux conformations, les conformations 5 et 6 de la protéine hIL-21 (2oqp). L'identité est toujours le premier résultat obtenu.

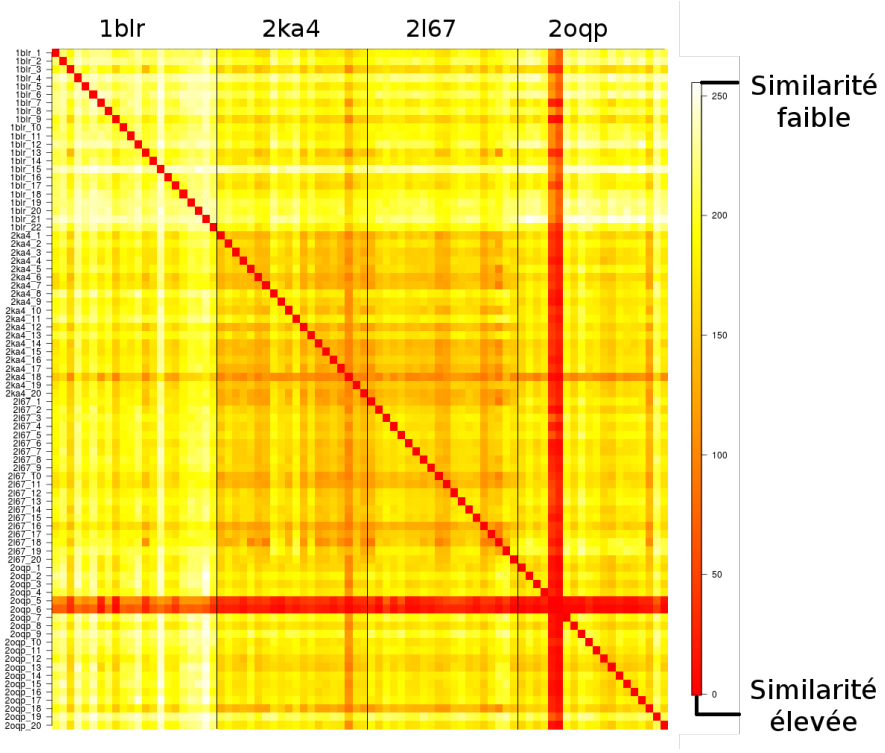


FIGURE 5.5 – Matrice de score de la méthode monoview sur le jeu de données DSV

Les deux conformations de 2oqp n'ont pas une surface qui diffère des autres conformations de 2oqp (cf Figure 5.6).

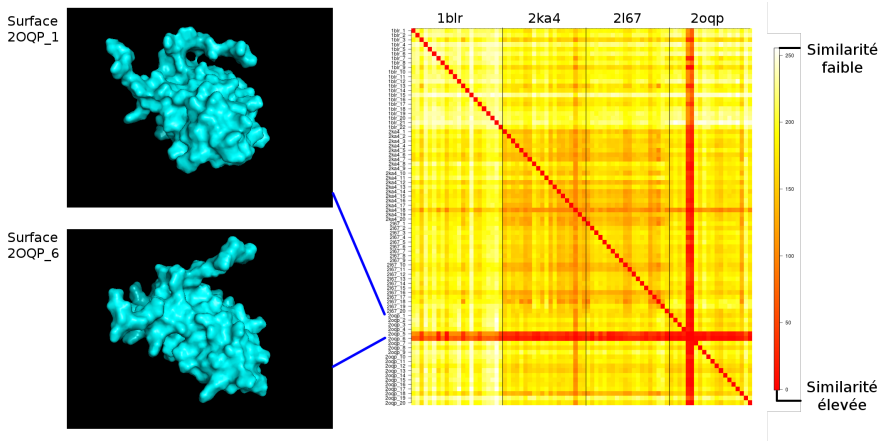


FIGURE 5.6 – Surface de 2oqp (conformation 1) et 2oqp (conformations 5) à gauche et matrice de score à droite de la méthode monoview sur le jeu de données DSV

Les cartes de courbure de 2oqp (conformation 5) et 2oqp (conformation 6) ont une plus faible

densité de points que les autres cartes de courbure (cf Figure 5.7).

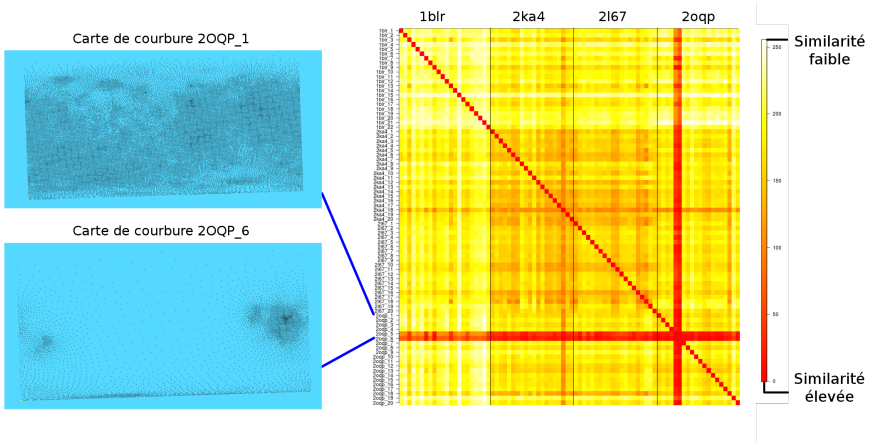


FIGURE 5.7 – Cartes de courbure de 2oqp (conformation 1) et 2oqp (conformation 5) à gauche et matrice de score de la méthode monoview sur le jeu de données DSV à droite

Le tableau Table 5.1 regroupe le NN, FT et ST.

Les valeurs de NN sont de 0 à l'exception de hIL-21 (2oqp) qui est de 1.

La protéine CRABPII (1blr) a un FT et ST de respectivement 0.013 et 0.15. Synaptotagmin1 (2ka4) possède le FT le plus élevé avec 0.505. Le ST de Synaptotagmin1 et L-FABP est supérieur à 0.7 et hIL-21 à un ST de 0.642.

Classes	Nearest Neighbour	First Tier	Second Tier
1blr	0.000	0.013	0.117
2ka4	0.000	0.505	0.787
2l67	0.000	0.276	0.747
2oqp	1.000	0.429	0.642

TABLE 5.1 – Performances sur le jeu de données DSV

La précision de la courbe précision-rappel Figure 5.8 varie entre **100%** et **30%** de précision. La courbe décroît abruptement dès la première valeur pour se stabiliser entre **50%** et **30%**.

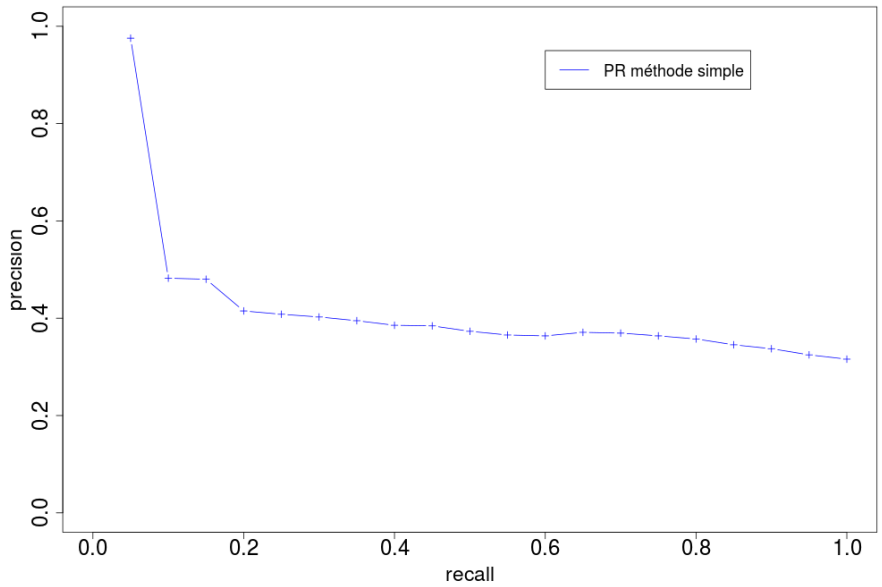


FIGURE 5.8 – Courbe de précision rappel de la méthode monoview sur le jeu de données DSV

5.4.2 Résultats avec coefficients sur DSV

A l'opposé des résultats précédents, les deux conformations 5 et 6 de 2oqp ont des scores de dissimilarité élevés contrairement aux autres conformations (cf Figure 5.9).

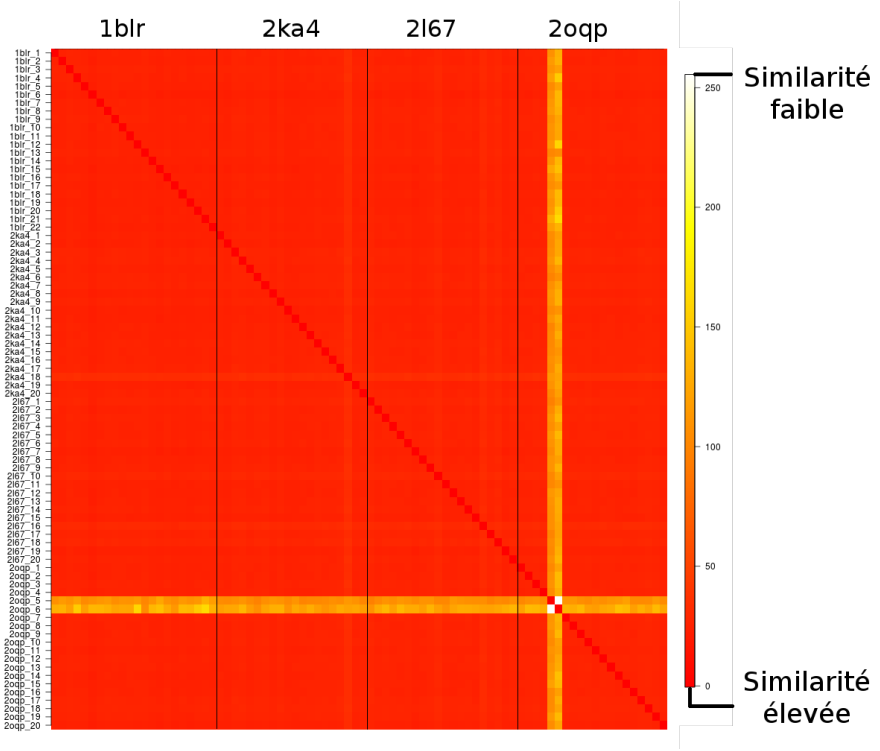


FIGURE 5.9 – Matrice de score de la méthode monoview avec coefficients sur le jeu de données DSV

Les valeurs de NN pour Table 5.2 sont d'environ 0 sauf pour CRABPII (1blr) qui est proche de 1.

Le FT varie entre 0.3 et 0 et le ST entre 0.5 et 0.2 pour Synaptotagmin1, hIL-21, L-FABP. CRABPII a un FT de 0.422 et son ST est 0.777.

Classes	Nearest Neighbour	First Tier	Second Tier
1blr	0.909	0.422	0.777
2ka4	0.050	0.261	0.453
2l67	0.000	0.176	0.400
2oqp	0.000	0.079	0.289

TABLE 5.2 – Performances sur le jeu de données DSV avec l'utilisation de coefficients sur la méthode monoview

La précision de la courbe précision-rappel Figure 5.10 varie entre **100%** et **25%** de précision. Comme pour la méthode sans coefficient, la courbe décroît abruptement dès la première valeur et se stabilise entre **40%** et **20%**.

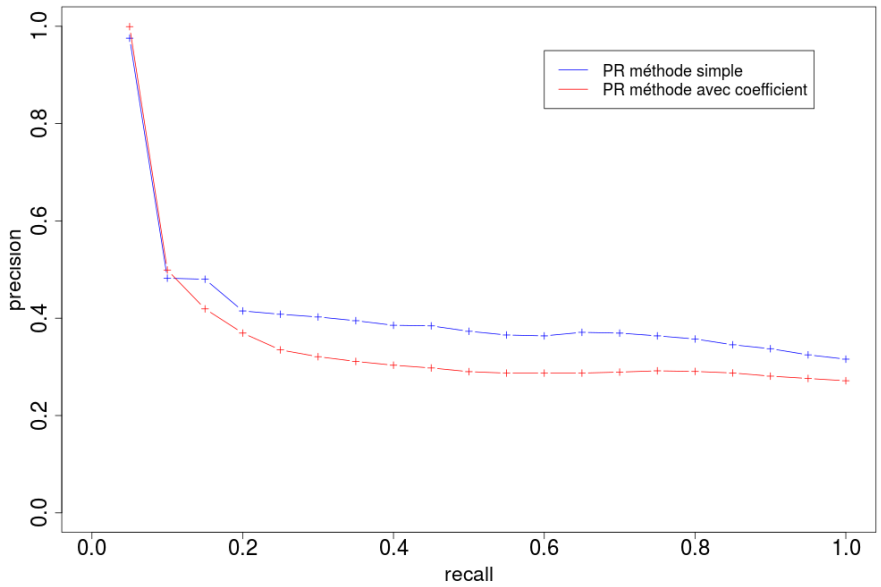


FIGURE 5.10 – Courbe de précision-rappel de la méthode monoview avec un coefficients sur le jeu de données DSV

5.5 Résultats avec la méthode multiview

Le multiview comprend 18 vues différentes et demande un temps de calcul significativement plus long, c'est pourquoi on utilise le sous jeu de données DSV16. Pour essayer de diminuer le temps de calcul, une comparaison éparse est proposée en deuxième partie de cette section. La première partie compare tous les points des cartes de courbure tandis qu'avec la comparaison éparse seulement un point sur quatre est comparé.

5.5.1 Résultats avec multiview sur DSV16

Sur la carte de score (cf Figure 5.11) l'identité est toujours le meilleur match. On remarque que les valeurs des deux conformations 5 et 6 de 2oqp sont légèrement inférieures aux autres valeurs et que les conformations 1 et 2 de 2oqp sont les valeurs de dissimilarité les plus élevées.

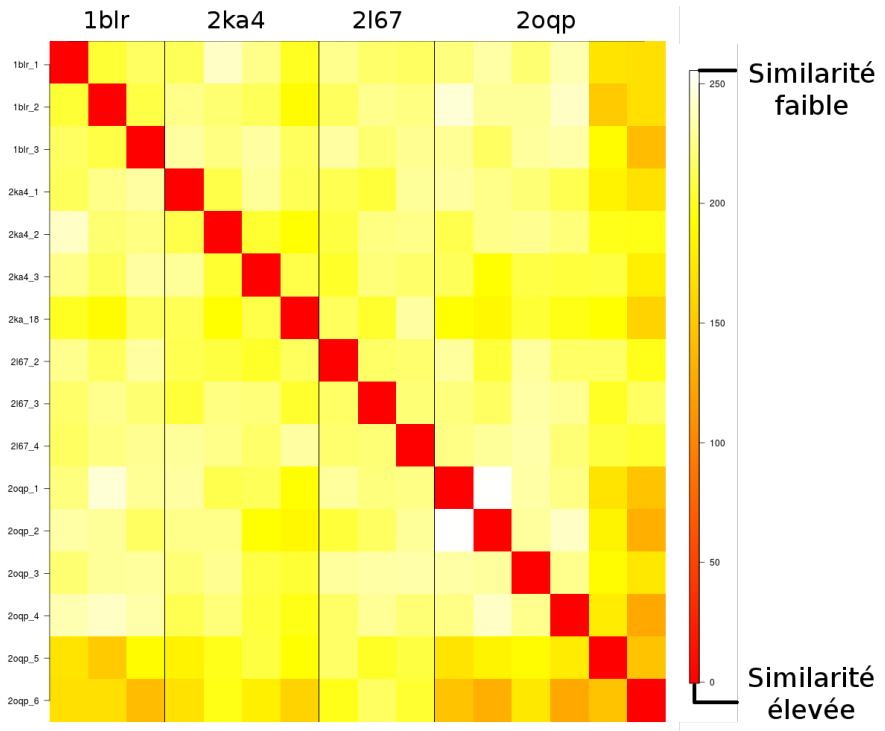


FIGURE 5.11 – Matrice de score de la méthode multiview avec coefficients sur le jeu de données DSV16

Les valeurs de NN pour Table 5.3 sont de 1 pour hIL-21 (2oqp), de 0.25 pour Synaptotagmin1 (2ka4) et de 0 pour CRABPII (1blr) et L-FABP (2l67).

Le FT est proche de 0 pour Synaptotagmin1, L-FABP, CRABPII. hIL-21 a un FT de 0.5

Le ST de L-FABP est 0, CRABPII et Synaptotagmin1 ont un ST de 0.5 et hIL-21 a un ST de 0.733.

Classes	Nearest Neighbour	First Tier	Second Tier
1blr	0.000	0.000	0.500
2ka4	0.250	0.083	0.500
2l67	0.000	0.000	0.000
2oqp	1.000	0.500	0.733

TABLE 5.3 – Performances sur le sous jeu de données DSV16 multiview

La précision de la courbe précision-rappel Figure 5.10 varie entre **100%** et **40%** de précision. La courbe décroit linéairement pour arriver à environ **40%** de précision pour un rappel supérieur à **60%**.

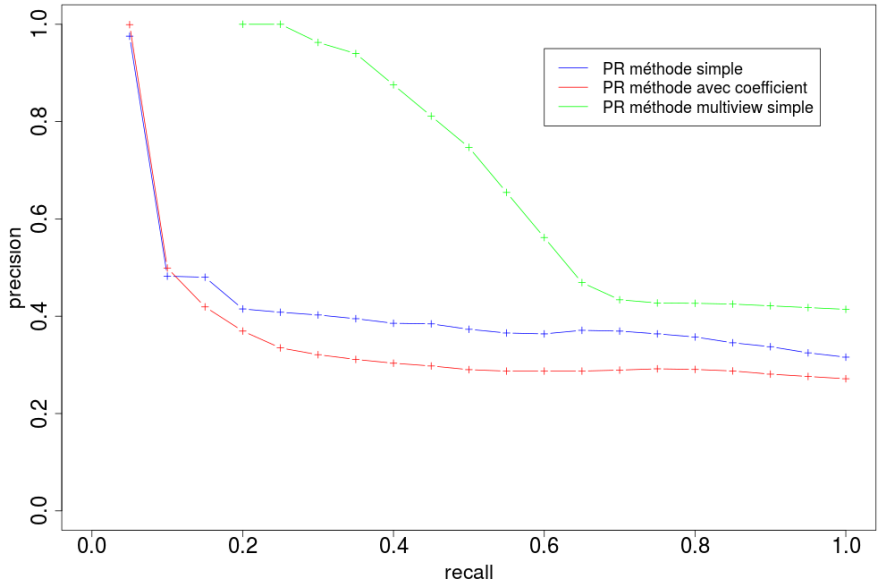


FIGURE 5.12 – Courbe de précision-rappel de la méthode multiview sur le jeu de données DSV16

5.5.2 Résultats avec multiview sur DSV16 avec une comparaison éparse

Les valeurs de NN pour Table 5.4 sont de 0.833 pour hIL-21 (2oqp), de 0.5 pour Synaptotagmin1 (2ka4) et de 0 pour CRABPII (1blr) et L-FABP (2l67).

Le FT est 0 pour L-FABP, CRABPII. Synaptotagmin1 a un FT de 0.167 et hIL-21 de 0.367.

Le ST de L-FABP est 0, CRABPII et Synaptotagmin1 ont un ST de 0.333 et hIL-21 a un ST de 0.667.

Classes	Nearest Neighbour	First Tier	Second Tier
1blr	0.000	0.000	0.333
2ka4	0.500	0.167	0.333
2l67	0.000	0.000	0.000
2oqp	0.833	0.367	0.667

TABLE 5.4 – Performances sur le sous jeu de données DSV16 avec multiview et une comparaison éparse

La précision de la courbe précision-rappel Figure 5.13 varie entre **100%** et **30%** de précision. La courbe décroît linéairement pour arriver à environ **30%** de précision pour un rappel supérieur à **70%**.

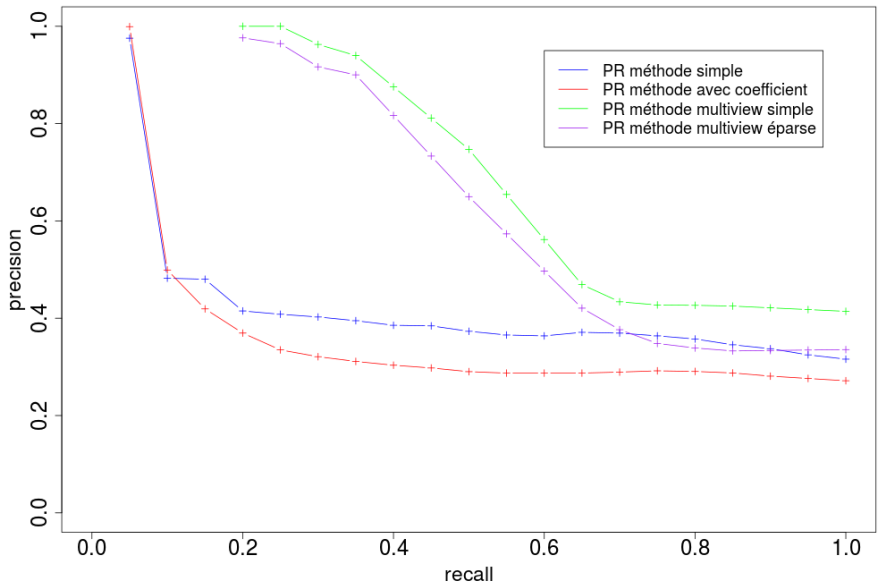


FIGURE 5.13 – Courbe de précision-rappel de la méthode multiview avec une comparaison éparse sur le jeu de données DSV16

5.6 Discussion

5.6.1 Comparaison avec la méthode monoview

Les scores de dissimilarité avec la méthode monoview sont assez homogènes ce qui peut indiquer un problème de discrimination du descripteur. Ceci peut s'expliquer par la surface des protéines qui est moins lisse que les formes habituellement décrites par la courbure. Ces formes moins lisses sont souvent des petites protubérances ou dépressions. Elles peuvent donc se retrouver dans beaucoup de cas et influencer le score. Les conformations 5 et 6 de 2oqp on un score de dissimilarité faible avec toutes les autres conformations. Ceci est la conséquence de cartes de courbure avec une faible densité de points. Ces faibles scores vont influencer négativement les mesures d'évaluations.

La courbe de précision-rappel possède des valeurs de précision basses et un NN et FT eux aussi bas. Seul le ST possède des valeurs plus élevées. Il faut aussi noter que ces résultats sont meilleurs lorsque l'on n'ajoute pas de coefficients.

L'ajout du coefficient au score final a inversé le *ranking*, faisant que les conformation 5 et 6 de 2oqp ne possèdent plus le meilleur score avec toutes les autres protéines. Ceci permet d'éviter que les

conformations 5 et 6 de 2oqp aient une influence sur l'entièreté des mesures d'évaluation. Le meilleur score avec toutes les protéines est maintenant la conformation 15 de 1blr qui auparavant était dans le bas du *ranking*. Le coefficient a seulement inversé la situation en classant en premier la carte de courbure de la conformation 15 de 1blr ayant le plus de points. De plus, le coefficient peut amener de faux positifs en rapprochant le score d'un objet ayant naturellement peu de points avec un objet ayant beaucoup de points mais ne se ressemblant pas. Il peut être préférable de conserver cette information sur le nombre de points.

On remarque que les deux conformations de hIL-21 (2oqp) ont, du fait de leurs cartes de courbure très éparses, des résultats qui diffèrent fortement des autres résultats. Ajouter des coefficients ne permettant pas de rééquilibrer ces différences. C'est pour ces raisons qu'il ne semble donc pas utile d'utiliser de coefficient basé sur le nombre de points pour la suite.

Le problème des deux conformations 5 et 6 de 2oqp et les résultats faibles du NN, FT et de précision-rappel sont dus à un manque de discrimination. Les deux conformations de 2oqp n'ont pas assez de points sur leur carte de courbure dont l'une des causes est la projection aux pôles.

Une des solutions est d'utiliser le multiview avec une rotation du pôle de la sphère unitaire. Cette rotation permet de limiter la déformation de la projection aux pôles.

5.6.2 Comparaison avec la méthode multiview

Le premier problème de la méthode multiview est le temps de comparaison de deux protéines qui est environ multiplié par 100. Pour comparer les 82 protéines entre elles, il faudrait compter environ 187 jours. Pour palier à ceci, un plus petit jeu de données a été produit et comprenant les deux conformations 5 et 6 de 2oqp.

On remarque que la différence entre les deux conformations de 2oqp et le reste des autres surfaces est atténuée mais pas complètement effacée. Il est difficile de distinguer des groupes à partir des matrices de score.

Les mesures NN, FT et ST n'offrent que des valeurs faibles voir nulles pour CRABPII (1blr) et L-FABP (2l67). Synaptotagmin1 (2ka4) a des résultats faibles ou moyens et hIL-21 (2oqp) a des résultats significatifs. Si l'on regarde les protéines 1blr et 2l67 ce sont les deux protéines quasiment identiques qui sont les protéines ayant des résultats très faibles. 2ka4 est similaire aux deux premières et a un NN,

FT et ST faible tandis que 2oqp qui n'est similaire à aucune des trois protéines a un pourcentage de classification élevée. Ceci indique que la méthode développée peut distinguer les surfaces suffisamment différentes, ce qui est confirmé par un ST globalement élevé. Mais lorsque les surfaces sont similaires il est plus difficile de les discerner.

On remarque que la courbe de précision rappel diminue de façon plus lente qu'avec la méthode monoview ce qui indique une meilleure précision grâce au multiview.

Chapitre 6

Méthode basée sur les cartes WKS

6.1 Résumé

Bien que la courbure et la convexité soient des notions empruntées à la topographie, domaine utilisant la projection utilisée, il est difficile d'extraire avec ces descripteurs suffisamment d'information pour une discrimination sur les protéines. Il a donc été décidé de rajouter la descripteur WKS qui a déjà montré de bons résultats [97] [81] [82] .

Cette méthode utilise la projection de la sphère unité sur le plan 2D. Les valeurs inscrites dans les cartes sont les valeur du descripteur WKS. Le multiview est appliqué car il est permet d'atténuer l'erreur due à la déformation aux pôles. Une interpolation des points est ajoutée pour compléter la carte avec les points n'étant pas projetés sur la grille extraite de la sphère unitaire. Ce descripteur est appelé Projected Wave Kernel Signature Map (PWKSM).

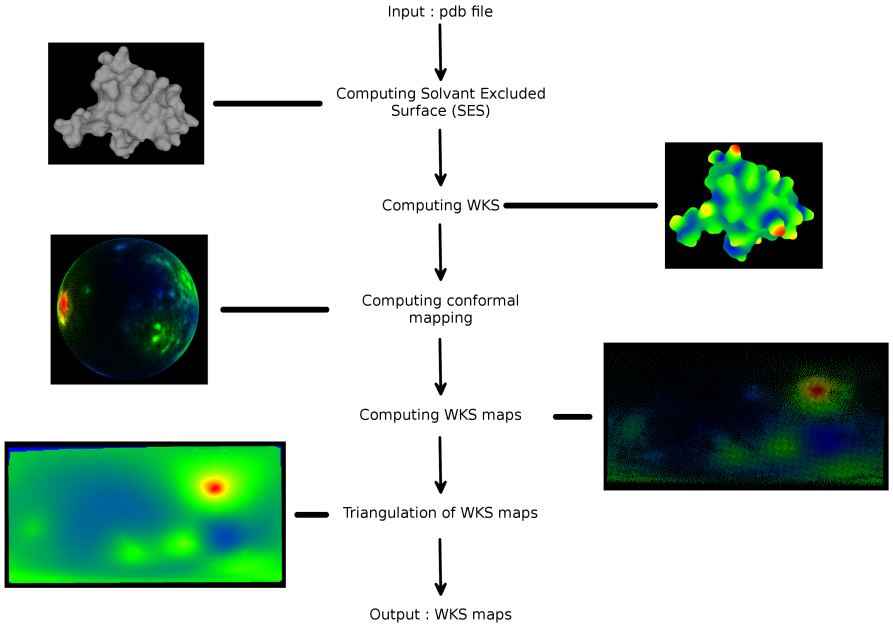


FIGURE 6.1 – Aperçu de la création du descripteur PWKSM.

Le jeu de données (cf Figure 6.2) est composé de 3 classes déterminées par le classement de la banque de données SCOPe : α, β et α/β. Chacune des classes contient 4 protéines ayant la même fonction mais venant d'espèces différentes. 10 conformations sont extraites de chacune des protéines. Une classe est donc composée de 40 objets.

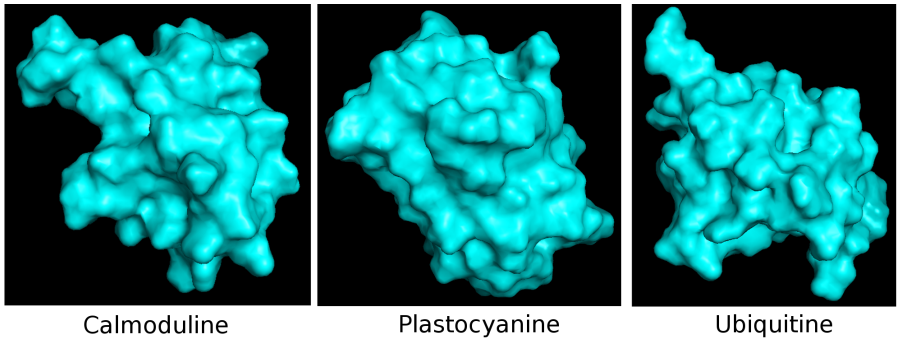


FIGURE 6.2 – Le jeu de données composé de 120 objets et divisé en 3 classes.

Le NN le plus élevé est celui de la méthode WKS avec 75.8% et le plus faible pour FPFH avec 20%. PWKSM possède un NN de 34.2%. Le FT et ST sont similaires pour les quatre méthodes, respectivement autour de 35% et 65%. Les courbes de précision-rappel pour les quatre méthodes

suivent la même tendance. La précision est différente pour un rappel faible variant de 87.5% pour le WKS à 41.8% pour FPFH. La précision est d'environ 40% pour les quatre méthodes lorsque le rappel est élevé.

	NN	FT	ST
PWKSM	0.342	0.346	0.672
FPFH	0.200	0.306	0.643
USC	0.517	0.361	0.680
WKS	0.758	0.425	0.687

TABLE 6.1 – Tableau du Nearest Neighbour (NN), First Tier (FT) and Second Tier (ST) pour PWKSM, FPFH, USC et WKS pour le jeu de données des 120 protéines.

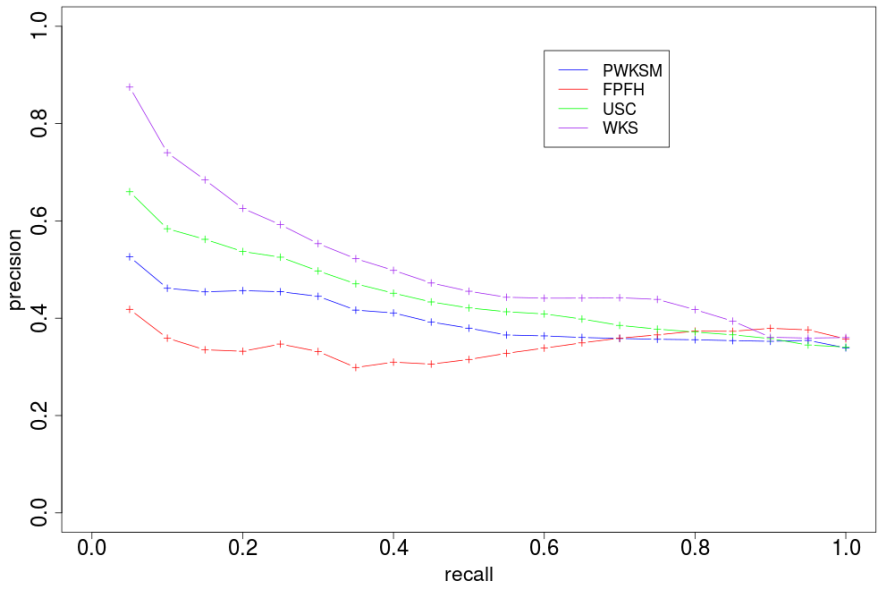


FIGURE 6.3 – Courbe de précision-rappel pour PWKSM, FPFH, USC et WKS pour le jeu de données des 120 protéines.

Le temps de calcul de descripteur est le plus rapide pour FPFH et le plus lent pour PWKSM avec respectivement 34 secondes et 23 minutes et 35 secondes. Le temps de comparaison le plus élevé est celui de FPFH qui est proche de celui de USC qui est de 14 minutes et 43 secondes. La comparaison la plus rapide est celle de PWKSM avec 7 secondes.

	Temps moyen de calcul du descripteur	Temps moyen de comparaison
PWKSM	23min 35s	7s
FPFH	34s	14min 58s
USC	1min 10s	14min 43s
WKS	9min 36s	1min 43s

TABLE 6.2 – Temps moyen de calcul du descripteur et de comparaison pour PWKSM, FPFH, USC et WKS sur le jeu de données des 120 objets.

Le point fort de notre méthode PWKSM est son temps de comparaison des objets qui est au minimum 15 fois plus rapide que les autres méthodes de l'état de l'art. Il est préférable d'avoir un temps de comparaison rapide contrairement à temps de calcul des descripteurs car les descripteurs n'ont besoin d'être calculés qu'une seule fois. Ils peuvent être ensuite stockés pour être réutilisés.

Cette méthode peut donc s'inscrire dans une optique de haut débit en diminuant la taille d'un jeu de données pour ensuite être traitée par des méthodes plus précises.

A System For Comparison Of Shapes Based On A Wave Kernel Signature Map applied to proteins

1st Lea Sirugue
Laboratoire GBCM, EA4627
Conservatoire National des Arts et Metiers
Paris, France
jeremy.sirugue@cnam.fr

2nd Matthieu Montes
Laboratoire GBCM, EA4627
Conservatoire National des Arts et Metiers
Paris, France
matthieu.montes@cnam.fr

***Abstract*—Many methods have been proposed to compare macroscopic objects in the computer vision fields. Few of these methods have been tested on proteins. Proteins have their own unique shape which can be a challenge to compare and differentiate. This article proposed a method based on the descriptor Wave Kernel Signature (WKS) aimed to compare rapidly proteins in the context of big data and called Projected WKS Map (PWKSM). The descriptor PWKSM is a descriptor projected on a unit sphere mapped on a 2D plane. Experimental results on a proteins dataset show performances similar to techniques from the State-of-the-Art with a fast comparison time.**

***Index Terms*—protein, comparison, computer vision, Laplace-Beltrami**

I. Introduction

The 3D object retrieval field has grown in interest the recent years. It has application in the medical field [1] [2], automated car [3] [4] [5], robotics [6] [7], surveillance [8] [9] and biology [10] [11] [12]. This growth can be explained by an increase of available shapes. The proteome informatics is one fields fitting this growing interest. The Protein Data Bank [13] is the main data bank for proteins which are digitalized with high-technology instruments and the information available is ranging from the chemical proprieties to the spatial connections of the atoms of the proteins.

Comparing proteins is a central question in structural bioinformatics for classifying proteins [14] [15] [16] [17] [18] or finding protein-protein interactions [19] [20] [21]. The two most common approach to compare proteins are through their sequence [22] [23] [24] or their structure [25] [26] [27] [28]. Another approach is to compare surface [29] [30] [31], especially in protein-protein interactions. Most interactions in proteomics are made on the surface of the protein making the surface a more compact and still informative representation of the protein compare to the structure.
Proteins surfaces have the particularity of being irregular due to their organic and microscopic nature making it complex to compare. It is a challenge in computer graphic field where the main shapes studied are man made or have a smooth surface.

In this research project a shape recognition system based on local features is proposed. This work is aimed at proposing a wave kernel signature projected on a 2D plane descriptor entitled *Projected Wave Kernel Signature Map* (PWKSM) and derived from the descriptor *Wave Kernel Signature* method [32].
Additionally, a simple method for comparing the PWKS maps by using a dense point-to-point comparison is described.

This paper is organized as follows. In the section 2 an overview of the previous works is given. In the section 3 the descriptor proposed and the comparison of it is explained. In section 4 we show results compared to other methods of the state-of-the-art. In section 5 the result are discussed. Finally, in section 6 we conclude and give an insight on the future works.

II. Previous works

A lot of methods have been proposed to compare proteins and most of them are based on biological or physico-chemical proprieties. Studying the proteins in a geometrical and a topological point of view is mostly new.
Descriptors with a property of invariance to isometry are relevant descriptors because proteins can be more flexible than other articulated objects and non-rigid movement is an important feature [33].

A. Histograms-based methods

One well-known histograms-based descriptor is Spin Image [34]. Spin Image encode a 2D-histogram for each keypoints. This histogram represents the position of the points in the neighborhood of the keypoint by spinning a plane around the normal of the keypoint.
The Unique Shape Context [35] which compute histogram of neighbors based on the normal of the frame of the point and each point inside the neighbor. The values of this histogram are the cosine of the angle of the normal of the frame and the normal of a point inside the neighbor.
Another histograms-based method is Point Feature Histograms [36] and its improved version Fast Point Feature Histograms [37]. For each point, three geometric characteristics are computed between the keypoint and the neighborhoods points. A histogram for each characteristic is encoding the

geometric information.

B. View-based methods

View-based methods are using one or several 2D representation of 3D shapes. The 2D representation allows a smaller file and a faster parsing than 3D representation.
PANORAMA method [38] is projecting 3D shape on three cylinders, cylinders oriented in parallel with the coordinate axes. The cylindrical projection are then mapped on a 2D image.
In the proposed method Shape Similarity System driven by Digital Elevation Models (SSS-DEM) [39], meshes are mapped onto the unit sphere using the Laplace-Beltrami operator then the unit sphere is projected onto a 2D spherical panoramic grid with the elevation values of the input mesh and called Digital Elevation Model (DEM).

C. Spectra-based methods

The spectra-based descriptors is one category of descriptor which can fit smoothly to the proteins comparison issues because of its invariance to isometry property [40] [41] [32]. Through the spectral geometry domain, the geometry and topology of a shape are represented with its spectrum which is the eigenvalues of the Laplace-Beltrami operator. The methods presented here are all based on the eigenvalues and eigenfunctions of the Laplace-Beltrami operator.
One of the points of spectral descriptors is its isometry which allows to recognize the same object with different deformation. Despite few examples of two non-congruents objects being isospectral [42], it has been explained that examples of isospectral and non-congruent shape are rare enough and artificial for not being an issue in shape recognition [40].
The first work using spectra for describing a shape is Reuter et al. [40] with the ShapeDNA method [43]. ShapeDNA is using the eigenvalues of the Laplace-Beltrami operator. Another work is Global Point Signature [44] which add the eigenfunctions to the descriptor and is called signature.
Heat Kernel Signature (HKS) [41] is derived from the Heat Kernel which represent the diffusion of heat on an object depending on the time. The HKS is based on the Laplace-Beltrami operator and its spectrum. HKS has the property of isometric invariance and is stable against perturbation [41]. The HKS method has been improved by many people, for citing a few : the Scale Invariant Heat Kernel Signature (SI-HKS) [45] which add the property of invariance to scaling or the Generalized Heat Kernel Signature (GHKS) [46] which generalize the HKS to r-forms.
Wave Kernel Signature (WKS) [32] is also based on the spectrum of the Laplace-Beltrami operator. It is describing the energy of quantum particles on the surface of the shape based on the wave equation which is a solution of the Schrödinger's equation. A signature is created with the solution of the wave equation. The idea is close to HKS and have also the propriety of isometry and scale invariance. The improvement is made through the time which is not take in consideration in this method and replace by energy of the particle. The size of the energy is related to the geometry size; large energy depict a local geometry and small energy a global geometry.

In this work we proposed a method called Projected WKS Maps (PWKSM) and based on conformal projection on the unit sphere. Then the unit sphere is mapped on the 2D plane with the WKS value computed on the 3D object. The final descriptor, PWKSM, is compared for each point of each PWKSM.

III. Our method

A Projected WKS Map (PWKSM) is a representation of the WKS computed on a 3D surface projected on a 2D plane. We proposed a method composed of two parts, the computation of the PWKSM descriptor and the comparison of this descriptor.

A. Generation of Projected WKS descriptor

In a first step the Wave Kernel Signature (WKS) is computed on the input 3D mesh M following the Aubry et al. method [32]. For each point on the surface of a 3D input mesh, a vector of size N, representing the WKS descriptor is computed. This descriptor, based on the eigenvalues of the Laplace-Beltrami operator has the proprieties of invariance to isometry and has been proved to be robust to perturbations.
On a second step the 3D input mesh is flattened on a unit sphere S as described by Angenent et al [47]. The unit sphere projection assigns a point called *north pole* which is the reference point defining the frame. This transformation is using the Laplace-Beltrami operator, conformal and bijective. Also, while distances and area are not preserved, it is only modified by a scaling factor.
Then the unit sphere is transformed onto the 2D plane based on the two spherical coordinates of the angles (θ, ϕ) according to the Shape Similarity System driven by Digital Elevation Models method [39] [48]. A map is created of size $(\theta_{max} - \theta_{min})/\delta, (\phi_{max} - \phi_{min})/\delta$. θ_{max} and θ_{min} are the maximum and minimum values of θ and same with ϕ. δ is the step for dividing the sphere into area representing a point in the discrete plan. To each point on the map the values associated to each point are the WKS descriptor. These maps are called WKS maps.
A final step before comparison is the interpolation of the point. The map is encoded as an image and each pixel is not fill with a value. It can create an unbalance if the neighborhood of a point is considered for comparison. So for each pixel with no value, the three points creating the triangle with the smallest area containing the pixel are used for interpolating the value of this pixel.
The main issue with this representation is the deformation in the neighbor of the poles while passing from the unit sphere to the 2D plane. To handle this, the pole axis is rotated of an angle α in the planes perpendicular to the three Cartesian

axis of the unit sphere. Then a WKS map is created as above-mentionned. The representation is then a set of WKS maps.

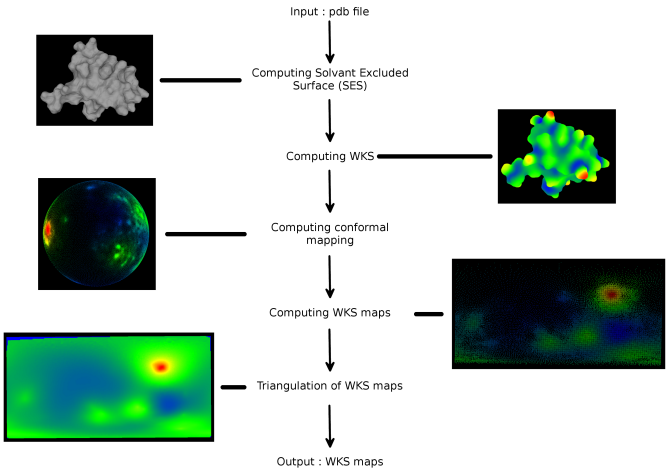


Fig. 1. Diagram for the generation of PWKS descriptor

B. Comparison

Around each points an area of size $S * S$ is defined. All the WKS descriptors inside the area are summed and normalize. Then a dense comparison is made on WKS maps, each point of one 2D WKS map is compared to each point of another 2D WKS map.

Two shapes T and V are compared by a search of the best matches of histograms of their WKS maps C_T and C_V. For comparing histograms of WKS maps H_{k_T} and H_{k_V} at points k_T and k_V of respectively C_T and C_V, the Earth Mover's Distance (EMD) [49] is used. EMD is a distance adapted to signatures because it take in account the nearness of the bins. EMD is used between the WKS of two points. The WKS is represented as a 1D array, which mean that the EMD equation can be simplified as the sum of the absolute difference between the cumulative values.

Given two WKS $X = \{x_1, ...x_n\}$ and $Y = \{x_1, ...x_n\}$, $d_{ij} = |i - j|$ is define and the values of f_{ij} is minimize in the following equation :

$$\sum_{i=1}^{n}, \sum_{j=1}^{n} f_{ij} d_{ij}$$

With the constraints $f_{ij} > 0$, $\sum_{j=1}^{n} = x_i$ and $\sum_{i=1}^{n} = y_i$. This equation can be rewriten as the sum of the absolute difference between the cumulative values of both WKS. Given $D_X(i) = \sum_{j=1}^{i} x_j$ and $D_Y(j) = \sum_{i=1}^{j} x_i$ then the EMD distance L between two WKS X and Y is defined as :

$$L(X, Y) = \sum_{l=1}^{n} |D_X(l) - D_Y(l)|$$

For having the score for two shapes, we sum the best distance of each point L. If C_T is composed of N_T points and C_V of N_V points, then the score $S(T, V)$ of dissimilarity between T and V is :

$$S(T, V) = \min(\sum_{k_T=1}^{N_T} \min_{k_V} L(X_{k_T}, Y_{k_V}), \sum_{k_V=1}^{N_V} \min_{k_T} L(X_{k_T}, Y_{k_V})$$

IV. Results

A. Material

All the computation had been made on a computer based on a 64-bit OS with an Intel Xeon CPU of 2.30GHz and 32GB of RAM.

B. Dataset

The data used have been made by ourself. It contains 3 classes, and each class contain 10 conformations of 4 proteins. The proteins are the **calmodulin**, the **ubiquitin** and the **plastocyanin**.

C. Evaluation measures

We have used the Princeton Shape Benchmark tools [50] for computing the Nearest Neighbor (NN), the First Tier (FT), the Second Tier (ST) and the Precision-Recall.

D. Experimental results

The Nearest Neighbor, the First Tier and the Second Tier are shown in I. The NN is the highest for the WKS with 75.8%, then the USC NN is 51.7%. On the lower part there is the PWKSM and the FPFH for respectively 34.2% and 20.0%.
The first tier is between 42.5% for WKS and 30.6% for the FPFH.
The ST is around 65%, with 68.7% for WKS, 68% for USC, 67.2% for PWKSM and 64.3% for FPFH.

	NN	FT	ST
PWKSM	0.342	0.346	0.672
FPFH	0.200	0.306	0.643
USC	0.517	0.361	0.680
WKS	0.758	0.425	0.687

TABLE I
TABLE OF NEAREST NEIGHBOUR (NN), FIRST TIER (FT) AND SECOND TIER (ST) FOR PWKSM, FPFH, USC AND WKS FOR THE 120 PROTEINS DATASET

The precision recall of the tested methods shown in 2. ALl methods is following a global similar trend which is a quick decreasing for the first value of the recall and then a slow decrease until reaching the value of 40% precision. The starting precision value for WKS is 87.5%, for USC it is 66%, for PWKSM it is 52.6% and FPFH 41.8%.

An overview of the average computation time of different methods tested has been made in II. The descriptor computation time includes the time of processing from the input mesh to the final descriptor. The fastest time for computing the descriptor it is the FPFH with 34 seconds and the slowest is 23 minutes and 35 seconds for PWKSM. The fastest comparison time is with method PWKSM with 7 seconds and the slowest is FPFH with 14 minutes and 58 seconds close to the USC method with 14 minutes and 43 seconds.

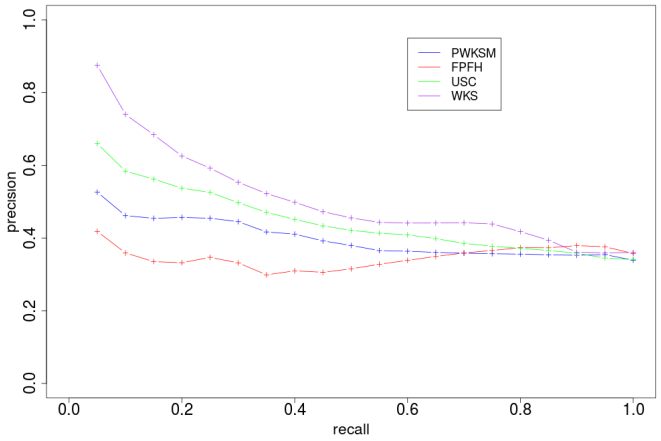


Fig. 2. Precision-recall graph for PWKSM, FPFH, USC and WKS on the 120 proteins dataset

	Descriptor ACT	Comparison ACT
PWKSM	23min 35s	7s
FPFH	34s	14min 58s
USC	1min 10s	14min 43s
WKS	9min 36s	1min 43s

TABLE II
AVERAGE COMPUTATION TIME (ACT) FOR ONE DESCRIPTOR AND COMPARISON FOR PWKSM, FPFH, USC AND WKS IN SECONDS ON THE 120 PROTEINS DATASET

V. DISCUSSION

With a high NN, 75.8%, WKS has a good precision for its top ranking objects which is corroborated by the precision recall plot and to a lesser extent USC too with an NN of 51.7%. On the other hand PWKSM and FPFH have a lower NN of repectively 34.2% and 20% and also a lower precision-recall than the two previous methods. All methods show an FT and ST similar to each other which indicate a precision similar for all method to the objects in the middle of the ranking. All this values show the difficulty to discriminate proteins compare to traditional computer vision's objects.

For a set of n objects, in an all versus all comparison, it required $n \times n$ comparison. For this reason, a faster comparison time is more important than a computing descriptor time. Because WKS and USC have a long comparison time of respectively 1 minute 43 seconds and 14 minutes and 43 seconds. They also have a good precision as shown before. Hence, these two methods can be used for a fine-grained comparison.

On the other hand, PWKSM has a fast comparison time of 7 seconds which is 15 times faster than WKS comparison and 126 times faster than USC comparison. PWKSM also produce an FT and ST similar to the two fine-grained method. This method can be used for a coarse-grained comparison which can produce a narrow set of proteins to compare with a fine grained method.

VI. CONCLUSION AND FUTURE WORK

We have developed a method composed of a 2D representation based on a conformal projection and a spectra based descriptor for local features of the shapes.

This prototype has to be improved on its sensitivity to produce a fast and coarse-grained method for narrowing the solution in a big dataset. Then the solution PWKSM method can be injected in a fine-grained method.

This method is part of the ViDOCK project, which aim at comparing and docking proteins. The next step is to use this method for finding similarities between proteins and infer proteins functions and docking by transitivity.

ACKNOWLEDGMENT

We thank the European Research Council (ERC) for the funding.

REFERENCES

[1] L. Mi, W. Zhang, J. Zhang, Y. Fan, D. Goradia, K. Chen, E. M. Reiman, X. Gu, and Y. Wang, "An optimal transportation based univariate neuroimaging index," in *Proceedings. IEEE International Conference on Computer Vision*, vol. 2017. NIH Public Access, 2017, p. 182.

[2] Q. Chen, R. Bise, L. Gu, Y. Zheng, I. Sato, J.-N. Hwang, N. Imanishi, and S. Aiso, "Virtual blood vessels in complex background using stereo x-ray images," *arXiv preprint arXiv:1709.07551*, 2017.

[3] M. Ahrnbom, M. B. Jensen, K. Aström, M. Nilsson, H. Ardö, and T. Moeslund, "Improving a real-time object detector with compact temporal information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 190–197.

[4] J.-T. Chien, C.-J. Chou, D.-J. Chen, and H.-T. Chen, "Detecting nonexistent pedestrians," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 182–189.

[5] S. Tsutsui, T. Kerola, and S. Saito, "Distantly supervised road segmentation," *arXiv preprint arXiv:1708.06118*, 2017.

[6] S. Paul, L. Vig, A. Manno-Kovacs, L. Kovacs, J. Suchan, M. Bhatt, M. Melis, A. Demontis, B. Biggio, G. Brown *et al.*, "Deterministic policy gradient based robotic path planning with continuous action spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 725–733.

[7] X. Wei, K. Jia, J. Lan, Y. Li, Y. Zeng, and C. Wang, "Automatic method of fruit object extraction under complex agricultural background for vision system of fruit picking robot," *Optik-International Journal for Light and Electron Optics*, vol. 125, no. 19, pp. 5684–5689, 2014.

[8] V. Gajjar, A. Gurnani, and Y. Khandhediya, "Human detection and tracking for video surveillance: A cognitive science approach," *arXiv preprint arXiv:1709.00726*, 2017.

[9] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*. IEEE, 2011, pp. 3153–3160.

[10] Y. Lu, L. Yang, H. Yuan, Y. Wang, H. Luo, and Y. Y. Tang, "A novel method for protein structure retrieval using tableau representation and sparse coding," in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4042–4046.

[11] A. Axenopoulos, D. Rafailidis, G. Papadopoulos, E. N. Houstis, and P. Daras, "Similarity search of flexible 3d molecules combining local and global shape descriptors," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 5, pp. 954–970, 2016.

[12] P. Gainza, F. Sverrisson, F. Monti, E. Rodola, D. Boscaini, M. Bronstein, and B. Correia, "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," *Nature Methods*, vol. 17, no. 2, pp. 184–192, 2020.

[13] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, Jan. 2000. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102472/

[14] J. Söding, "Protein homology detection by hmm–hmm comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005.

[15] K. Pawłowski and A. Godzik, "Surface map comparison: studying function diversity of homologous proteins," *Journal of molecular biology*, vol. 309, no. 3, pp. 793–806, 2001.

[16] L. Holm, S. Kääriäinen, C. Wilton, and D. Plewczynski, "Using dali for structural comparison of proteins," *Current protocols in bioinformatics*, vol. 14, no. 1, pp. 5–5, 2006.

[17] P. Daras, D. Zarpalas, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Three-dimensional shape-structure comparison method for protein classification," *IEEE/ACM transactions on Computational Biology and Bioinformatics*, vol. 3, no. 3, pp. 193–207, 2006.

[18] E. P. Costa, A. C. Lorena, A. C. Carvalho, A. A. Freitas, and N. Holden, "Comparing several approaches for hierarchical classification of proteins with decision trees," in *Brazilian symposium on bioinformatics*. Springer, 2007, pp. 126–137.

[19] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, "Detecting protein function and protein-protein interactions from genome sequences," *Science*, vol. 285, no. 5428, pp. 751–753, 1999.

[20] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[21] S. Jones and J. M. Thornton, "Analysis of protein-protein interaction sites using surface patches," *Journal of molecular biology*, vol. 272, no. 1, pp. 121–132, 1997.

[22] W. R. Pearson, "[5] rapid and sensitive sequence comparison with fastp and fasta," 1990.

[23] T. A. Tatusova and T. L. Madden, "Blast 2 sequences, a new tool for comparing protein and nucleotide sequences," *FEMS microbiology letters*, vol. 174, no. 2, pp. 247–250, 1999.

[24] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[25] A. Sali and T. L. Blundell, "Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming," *Journal of molecular biology*, vol. 212, no. 2, pp. 403–428, 1990.

[26] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of molecular biology*, vol. 233, no. 1, pp. 123–138, 1993.

[27] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (ce) of the optimal path." *Protein engineering*, vol. 11, no. 9, pp. 739–747, 1998.

[28] Y. Zhang and J. Skolnick, "Tm-align: a protein structure alignment algorithm based on the tm-score," *Nucleic acids research*, vol. 33, no. 7, pp. 2302–2309, 2005.

[29] L. Sael, D. La, B. Li, R. Rustamov, and D. Kihara, "Rapid comparison of properties on protein surface," *Proteins: Structure, function, and bioinformatics*, vol. 73, no. 1, pp. 1–10, 2008.

[30] B. Ma, T. Elkayam, H. Wolfson, and R. Nussinov, "Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5772–5777, 2003.

[31] A. P. Kornev, N. M. Haste, S. S. Taylor, and L. F. Ten Eyck, "Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism," *Proceedings of the National Academy of Sciences*, vol. 103, no. 47, pp. 17 783–17 788, 2006.

[32] M. Aubry, U. Schlickewei, and D. Cremers, "The wave kernel signature: A quantum mechanical approach to shape analysis," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov. 2011, pp. 1626–1633.

[33] J. Cortés, T. Siméon, V. Ruiz de Angulo, D. Guieysse, M. Remaud-Siméon, and V. Tran, "A path planning approach for computing large-amplitude motions of flexible molecules," *Bioinformatics*, vol. 21, no. suppl_1, pp. i116–i125, 2005.

[34] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 433–449, May 1999.

[35] F. Tombari, S. Salti, and L. Di Stefano, "Unique shape context for 3d data description," in *Proceedings of the ACM workshop on 3D object retrieval*. ACM, 2010, pp. 57–62.

[36] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*. IEEE, 2008, pp. 3384–3391.

[37] R. B. Rusu, N. Blodow, and M. Beetz, "Fast Point Feature Histograms (FPFH) for 3d registration," in *2009 IEEE International Conference on Robotics and Automation*, May 2009, pp. 3212–3217.

[38] P. Papadakis, I. Pratikakis, T. Theoharis, and S. Perantonis, "Panorama: A 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 177–192, 2010.

[39] D. Craciun, G. Levieux, and M. Montes, "Shape Similarity System driven by Digital Elevation Models for Non-rigid Shape Retrieval," in *Eurographics Workshop on 3D Object Retrieval*, I. Pratikakis, F. Dupont, and M. Ovsjanikov, Eds. The Eurographics Association, 2017.

[40] M. Reuter, F.-E. Wolter, and N. Peinecke, "Laplace-spectra as fingerprints for shape matching," in *Proceedings of the 2005 ACM symposium on Solid and physical modeling*. ACM, 2005, pp. 101–106.

[41] J. Sun, M. Ovsjanikov, and L. Guibas, "A Concise and Provably Informative Multi-Scale Signature Based on Heat Diffusion," in *Computer graphics forum*, vol. 28. Wiley Online Library, 2009, pp. 1383–1392.

[42] C. Gordon, D. Webb, and S. Wolpert, "Isospectral plane domains and surfaces via riemannian orbifolds," *Inventiones mathematicae*, vol. 110, no. 1, pp. 1–22, 1992.

[43] M. Reuter, F.-E. Wolter, and N. Peinecke, "LaplaceBeltrami spectra as Shape-DNA of surfaces and solids," *Computer-Aided Design*, vol. 38, no. 4, pp. 342–366, Apr. 2006.

[44] R. M. Rustamov, "Laplace-Beltrami Eigenfunctions for Deformation Invariant Shape Representation," in *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, ser. SGP '07. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2007, pp. 225–233. [Online]. Available: http://dl.acm.org/citation.cfm?id=1281991.1282022

[45] M. M. Bronstein and I. Kokkinos, "Scale-invariant heat kernel signatures for non-rigid shape recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1704–1711.

[46] V. Zobel, J. Reininghaus, and I. Hotz, "Generalized heat kernel signatures," 2011.

[47] S. Angenent, S. Haker, A. Tannenbaum, and R. Kikinis, "On the laplace-beltrami operator and brain surface flattening," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 700–711, 1999.

[48] D. Craciun, J. Sirugue, and M. Montes, "Global-to-local protein shape similarity system driven by digital elevation models," in *IEEE BioSmart*, 2017.

[49] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998, pp. 59–66.

[50] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The princeton shape benchmark," in *Shape modeling applications, 2004. Proceedings*. IEEE, 2004, pp. 167–178. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/1314504/

6.2 Comparaison avec des méthodes d'alignements de structures

Les quatre méthodes de comparaison de surfaces sont comparées à des méthodes d'alignement de structures.

6.2.1 Résultats

Les méthodes d'alignements de structures, DALI et TM-Align, possèdent un Nearest Neighbor de 100%, signifiant que pour toutes les protéines, la première protéine dans le classement fait partie de la même classe. Les méthodes de vision par ordinateur ont un NN inférieur à 100%, avec la valeur la plus haute de 75.8% pour la méthode WKS et la plus basse pour la méthode FPFH avec 20%. DALI et TM-Align possèdent un First Tier et Second Tier de 100%. Les méthodes de vision par ordinateur ont un FT proche variant entre 48.7% et 30.6%. Le ST est également similaire pour ces méthodes, avec un ST se situant entre 68.7% et 64.3%.

Classes	Nearest Neighbour	First Tier	Second Tier
FPFH	0.200	0.306	0.643
PWKSM	0.342	0.346	0.672
USC	0.517	0.361	0.680
WKS	0.758	0.425	0.687
DALI	1.000	1.000	1.000
TM-Align	1.000	1.000	1.000

TABLE 6.3 – Performances des méthodes DALI, FPFH, PWKSM, TM-Align, USC et WKS

DALI et TM-Align ont une précision constante de 100%. La précision décroît en suivant la même tendance pour les méthodes de vision par ordinateur. Pour une valeur de rappel de 0.05 la précision la plus élevée pour ces quatre méthodes est de 87.5% et la plus basse étant 41.8% pour FPFH. Ces méthodes tendent vers 40% lorsque le rappel est élevé.

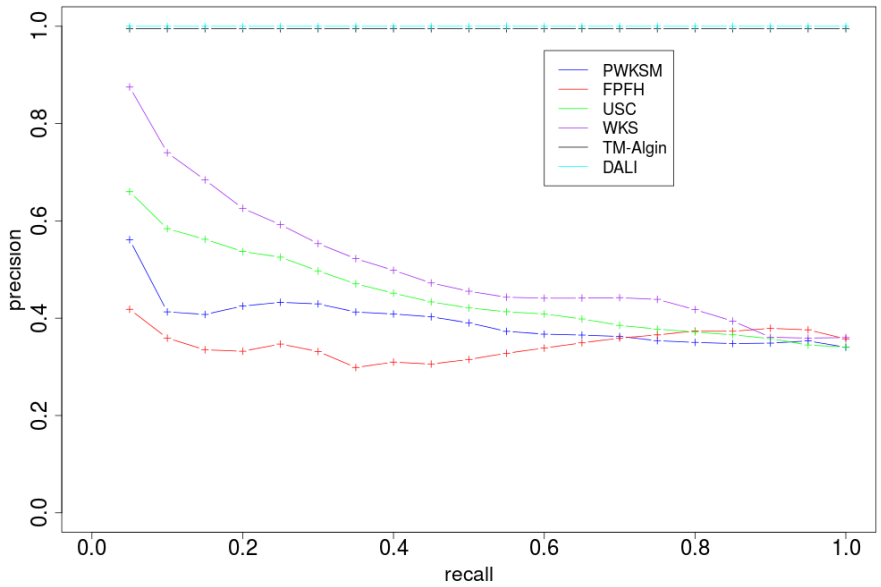


FIGURE 6.4 – Courbe de précision-rappel des méthodes DALI, FPFH, PWKSM, TM-Align, USC et WKS

6.2.2 Discussion

Les méthodes d'alignements de structures réussissent à classer toutes les protéines correctement contrairement aux méthodes de visions par ordinateur. Ces méthodes sont développées pour l'alignement de structure. De plus, le jeu de données possède des structures et des repliements distincts, les trois classes étant dans trois groupes de classe différentes d'après le classement de la base de données SCOPe, les groupes α, β et α/β. Ces raisons expliquent pourquoi la classification est correcte dans 100% des cas avec ces méthodes.

Au contraire les méthodes de comparaison de surfaces génériques ont des résultats plus faibles. Ces méthodes n'emploient pas des descripteurs développés spécifiquement pour les protéines ce qui peut expliquer ces résultats. L'objectif d'utiliser des méthodes de comparaisons de surfaces est de détecter des protéines avec une surface similaire qui est plus difficilement détectée lors d'alignement de structure. Cette approche peut être complémentaire aux méthodes d'alignement de structures.

Conclusion

L'objectif de ces travaux est la comparaison de surface de protéines en haut débit. La contrainte majeure est de réduire le temps de calcul pour permettre un traitement haut débit tout en conservant les informations les plus pertinentes. Optimiser le temps de comparaison est la priorité contrairement au temps de calcul du descripteur car le descripteur a besoin d'être calculé qu'une seule fois et peut être ensuite stocké pour être utilisé autant de fois que nécessaire pour la comparaison. Cette première étape permet de trouver des protéines avec des fonctions similaires pour ensuite vérifier si ces protéines similaires peuvent interagir avec les mêmes protéines. La comparaison de surfaces permettrait de compléter les résultats obtenus avec les méthodes d'alignements de structures en utilisant une représentation différente qui est la surface. Les méthodes d'alignement de structures utilisent la structure qui est fortement liée à la séquence. Les interactions protéine-protéine ont lieu à la surface des protéines, il est donc possible d'avoir des interactions protéine-protéine entre deux protéines ayant une faible identité de séquence mais une surface similaire.

Différentes méthodes ont été proposées pour comparer des protéines entre elles. La surface des protéines étant similaire à la surface de la terre ou à un cerveau, les outils de comparaison pour ces objets ont été utilisés [94] [95] [96] [88]. A la base des différentes méthodes présentées, une projection sur la sphère unitaire puis une transformation sur le plan 2D a été proposée. Cette représentation 2D avait pour objectif de diminuer la taille du stockage des objets à comparer ainsi que de diminuer la complexité du calcul fait lors de la comparaison.

La première méthode utilise une représentation avec des valeurs binaires. Le calcul de la convexité de manière binaire permet un stockage minimal et une comparaison rapide. La binarité des valeurs permet aussi d'inverser simplement les valeurs pour permettre de travailler sur la complémentarité de surface.

La seconde méthode apporte diverses modifications permettant d'augmenter la discrimination faite lors de la comparaison. La valeur binaire de convexité est remplacée par la courbure qui est une valeur continue et qui contient la notion de convexité. L'utilisation d'un matching aller-retour et du multiview permet d'améliorer la précision du résultat de la comparaison.

La dernière méthode améliore la consistance du descripteur en interpolant les valeurs qui ne sont pas projetées sur la sphère unitaire. La valeur de courbure est remplacée par les valeurs du WKS qui est un descripteur qui a prouvé son efficacité [80] [81] [82].

Ces différentes méthodes, bien qu'étant rapides, ont une précision faible mais possèdent en général une sensibilité élevée. Avoir une sensibilité élevée est ce qui est recherché dans ces travaux car nous souhaitons éviter d'éliminer les vrais positifs. Il est ensuite possible d'affiner les résultats avec une méthode ayant une précision élevée.

Pour le moment nous nous sommes limités à des protéines de formes globulaires car la projection sur une sphère unitaire implique d'avoir une forme globale à peu près similaire. Pour dépasser cette limitation une projection sur une forme basique plus proche de la protéine étudié peut être introduit. Par exemple dans le cas d'une protéine allongées tel un protéine fibreuse, une projection sur un cylindre peut être fait.

Les différentes méthodes proposées peuvent être améliorées en utilisant un *Bag of Features* (BoF) sur les cartes. L'utilisation d'un BoF permettrait d'avoir une représentation plus compacte des images et donc de diminuer le temps de comparaison qui reste l'un des facteurs limitant des différentes méthodes présentées. De plus, la création de BoF permettrait d'utiliser le vocabulaire créé dans un algorithme d'apprentissage.

Des tests ont été faits sur les cartes de convexité avec deux réseaux de neurones pré-entraînés sur des images, VGG16 et ResNet50. Lors de l'utilisation de ces deux modèles, aucun entraînement supplémentaire n'a été fait sur les cartes de convexités et la comparaison a été faite de manière globale sur les vecteurs finaux apportés par les réseaux de neurones. De bons résultats ont été obtenus en précision et en sensibilité. L'application d'un réseau de neurones sur les cartes de courbure et de WKS pourrait donc aussi apporter de bons résultats.

Après avoir amélioré les méthodes, il est possible d'ajouter les propriétés physico-chimiques des protéines aux descripteurs. Les propriétés physico-chimiques sont les descripteurs standards dans l'analyse des protéines. L'utilisation de ces propriétés permettra d'améliorer la reconnaissance de protéines similaires.

L'étape suivante est d'utiliser la comparaison de protéines pour pouvoir détecter de nouvelles fonctions de protéines et docker des protéines entre elles. Pour ce faire, une comparaison complète croisée, c'est-à-dire comparer toutes les protéines à toutes les protéines, peut être faite en haut débit pour pouvoir identifier de nouvelles fonctionnalités ou des zones de docking. Il est aussi possible de faire un criblage en haut débit, qui permet de chercher une fonctionnalité précise parmi un grand

nombre de protéines.

Les protéines ayant une surface similaire peuvent partager des fonctions similaires. Trouver une protéine similaire à des protéines avec une fonction connue permet d'inférer une nouvelle fonction à cette protéine. Trouver une protéine similaire à l'inverse de la surface d'une autre protéine peut permettre d'identifier de potentiels partenaires d'interaction. En particulier la comparaison locale peut être adaptée en une reconnaissance de formes partielle pour pouvoir ainsi détecter des parties de protéines similaires et donc des régions où peut avoir lieu l'interaction entre deux protéines.

Ce projet de thèse, à l'intersection de différents domaines, propose différentes approches pour apporter une solution à la comparaison de protéines en haut débit. L'idée principale est l'utilisation d'une représentation 2D pour faciliter la comparaison des protéines et fournir rapidement un classement de similarité des protéines. Ce classement n'a pas pour but d'être précis mais de réduire l'espace de recherche. Le deuxième point important est la conception de descripteurs locaux pour permettre d'adapter les travaux à une recherche partielle de formes et donc de prédire des sites de liaisons de molécules. Ces prédictions pourront être faites en comparant des protéines avec une fonction déjà connue ou en utilisant la complémentarité de surface.

Bibliographie

[1] I. Sipiran *et al.*, "Scalable 3d shape retrieval using local features and the signature quadratic form distance," *The Visual Computer*, vol. 33, n°. 12, p. 1571–1585, 2017.

[2] M. Aubry, U. Schlickewei et D. Cremers, "The wave kernel signature : A quantum mechanical approach to shape analysis," dans *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, nov. 2011, p. 1626–1633.

[3] A. E. Johnson et M. Hebert, "Surface matching for object recognition in complex three-dimensional scenes," *Image and Vision Computing*, vol. 16, n°. 9-10, p. 635–651, 1998.

[4] R. B. Rusu, N. Blodow et M. Beetz, "Fast Point Feature Histograms (FPFH) for 3d registration," dans *2009 IEEE International Conference on Robotics and Automation*, mai 2009, p. 3212–3217.

[5] F. Tombari, S. Salti et L. Di Stefano, "Unique signatures of histograms for local surface description," *Computer vision–ECCV 2010*, p. 356–369, 2010.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, n°. 2, p. 91–110, 2004.

[7] P. Papadakis *et al.*, "Panorama : A 3d shape descriptor based on panoramic views for unsupervised 3d object retrieval," *International Journal of Computer Vision*, vol. 89, n°. 2-3, p. 177–192, 2010.

[8] D. Craciun, G. Levieux et M. Montes, "Shape Similarity System driven by Digital Elevation Models for Non-rigid Shape Retrieval," dans *Eurographics Workshop on 3D Object Retrieval*, I. Pratikakis, F. Dupont et M. Ovsjanikov, édit. The Eurographics Association, 2017.

[9] D. Xu et Y. Zhang, "Generating triangulated macromolecular surfaces by euclidean distance transform," *PloS one*, vol. 4, n°. 12, p. e8140, 2009.

[10] L. Mi *et al.*, "An optimal transportation based univariate neuroimaging index," dans *Proceedings. IEEE International Conference on Computer Vision*, vol. 2017. NIH Public Access, 2017, p. 182.

[11] Q. Chen *et al.*, "Virtual blood vessels in complex background using stereo x-ray images," *arXiv preprint arXiv :1709.07551*, 2017.

[12] M. Ahrnbom *et al.*, "Improving a real-time object detector with compact temporal information," dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 190–197.

[13] J.-T. Chien *et al.*, "Detecting nonexistent pedestrians," dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 182–189.

[14] S. Tsutsui, T. Kerola et S. Saito, "Distantly supervised road segmentation," *arXiv preprint arXiv :1708.06118*, 2017.

[15] S. Paul *et al.*, "Deterministic policy gradient based robotic path planning with continuous action spaces," dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 725–733.

[16] X. Wei *et al.*, "Automatic method of fruit object extraction under complex agricultural background for vision system of fruit picking robot," *Optik-International Journal for Light and Electron Optics*, vol. 125, n°. 19, p. 5684–5689, 2014.

[17] V. Gajjar, A. Gurnani et Y. Khandhediya, "Human detection and tracking for video surveillance : A cognitive science approach," *arXiv preprint arXiv :1709.00726*, 2017.

[18] S. Oh *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," dans *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*. IEEE, 2011, p. 3153–3160.

[19] P. Shilane *et al.*, "The princeton shape benchmark," dans *Shape modeling applications, 2004. Proceedings.* IEEE, 2004, p. 167–178. [En ligne]. Disponible : http://ieeexplore.ieee.org/abstract/document/1314504/

[20] A. M. Bronstein, M. M. Bronstein et R. Kimmel, *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.

[21] Z. Wu *et al.*, "3d shapenets : A deep representation for volumetric shapes," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, p. 1912–1920.

[22] A. X. Chang *et al.*, "Shapenet : An information-rich 3d model repository," *arXiv preprint arXiv :1512.03012*, 2015.

[23] B. Shi *et al.*, "Deeppano : Deep panoramic representation for 3-d shape recognition," *IEEE Signal Processing Letters*, vol. 22, n°. 12, p. 2339–2343, 2015.

[24] H. Su *et al.*, "Multi-view convolutional neural networks for 3d shape recognition," dans *Proceedings of the IEEE international conference on computer vision*, 2015, p. 945–953.

[25] A. Sinha, J. Bai et K. Ramani, "Deep learning 3d shape surfaces using geometry images," dans *European Conference on Computer Vision*. Springer, 2016, p. 223–240. [En ligne]. Disponible : http://link.springer.com/chapter/10.1007/978-3-319-46466-4_14

[26] R. Klokov et V. Lempitsky, "Escape from cells : Deep kd-networks for the recognition of 3d point cloud models," dans *Proceedings of the IEEE International Conference on Computer Vision*, 2017, p. 863–872.

[27] C. R. Qi *et al.*, "Pointnet : Deep learning on point sets for 3d classification and segmentation," dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, p. 652–660.

[28] Y. Lu *et al.*, "A novel method for protein structure retrieval using tableau representation and sparse coding," dans *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*. IEEE, 2014, p. 4042–4046.

[29] A. Axenopoulos *et al.*, "Similarity search of flexible 3d molecules combining local and global shape descriptors," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, n°. 5, p. 954–970, 2016.

[30] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, n°. 1, p. 235–242, janv. 2000. [En ligne]. Disponible : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102472/

[31] A. v. Bondi, "van der waals volumes and radii," *The Journal of physical chemistry*, vol. 68, n°. 3, p. 441–451, 1964.

[32] B. Lee et F. M. Richards, "The interpretation of protein structures : estimation of static accessibility," *Journal of molecular biology*, vol. 55, n°. 3, p. 379–IN4, 1971.

[33] M. L. Connolly, "Shape complementarity at the hemoglobin $\alpha 1\beta 1$ subunit interface," *Biopolymers*, vol. 25, n°. 7, p. 1229–1247, 1986.

[34] C. Chothia et A. M. Lesk, "The relation between the divergence of sequence and structure in proteins." *The EMBO journal*, vol. 5, n°. 4, p. 823–826, 1986.

[35] L. Holm et J. Park, "Dalilite workbench for protein structure comparison," *Bioinformatics*, vol. 16, n°. 6, p. 566–567, 2000.

[36] J. M. Sauder, J. W. Arthur et R. L. Dunbrack Jr, "Large-scale comparison of protein sequence alignment algorithms with structure alignments," *Proteins : Structure, Function, and Bioinformatics*, vol. 40, n°. 1, p. 6–22, 2000.

[37] C. A. Rohl *et al.*, "Protein structure prediction using rosetta," dans *Methods in enzymology*. Elsevier, 2004, vol. 383, p. 66–93.

[38] S. Burley et G. A. Petsko, "Aromatic-aromatic interaction : a mechanism of protein structure stabilization," *Science*, vol. 229, n°. 4708, p. 23–28, 1985.

[39] J. W. Tangelder et R. C. Veltkamp, "A survey of content based 3d shape retrieval methods," *Multimedia tools and applications*, vol. 39, n°. 3, p. 441–471, 2008.

[40] W. E. Lorensen et H. E. Cline, "Marching cubes : A high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, n°. 4, p. 163–169, 1987.

[41] A. P. Mangan et R. T. Whitaker, "Partitioning 3d surface meshes using watershed segmentation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 5, n°. 4, p. 308–321, 1999.

[42] A. M. Bronstein, M. M. Bronstein et R. Kimmel, "Expression-invariant 3d face recognition," dans *international conference on Audio-and video-based biometric person authentication*. Springer, 2003, p. 62–70.

[43] N. A. Borghese et S. Ferrari, "A portable modular system for automatic acquisition of 3d objects," *IEEE Transactions on instrumentation and Measurement*, vol. 49, n°. 5, p. 1128–1136, 2000.

[44] S. Rusinkiewicz, O. Hall-Holt et M. Levoy, "Real-time 3d model acquisition," *ACM Transactions on Graphics (TOG)*, vol. 21, n°. 3, p. 438–446, 2002.

[45] D. G. Aliaga et Y. Xu, "A self-calibrating method for photogeometric acquisition of 3d objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, n°. 4, p. 747–754, 2010.

[46] F. Bernardini et H. Rushmeier, "The 3d model acquisition pipeline," dans *Computer graphics forum*, vol. 21, n°. 2. Wiley Online Library, 2002, p. 149–172.

[47] W. Kauzmann, "Some factors in the interpretation of protein denaturation," dans *Advances in protein chemistry*. Elsevier, 1959, vol. 14, p. 1–63.

[48] C. Chothia, "The nature of the accessible and buried surfaces in proteins," *Journal of molecular biology*, vol. 105, n°. 1, p. 1–12, 1976.

[49] H.-X. Zhou et Y. Shan, "Prediction of protein interaction sites from sequence profile and residue neighbor list," *Proteins : Structure, Function, and Bioinformatics*, vol. 44, n°. 3, p. 336–343, 2001.

[50] Y. Rubner, C. Tomasi et L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, n°. 2, p. 99–121, 2000.

[51] M. Reuter, F.-E. Wolter et N. Peinecke, "Laplace-spectra as fingerprints for shape matching," dans *Proceedings of the 2005 ACM symposium on Solid and physical modeling*. ACM, 2005, p. 101–106.

[52] ——, "Laplace–Beltrami spectra as 'Shape-DNA' of surfaces and solids," *Computer-Aided Design*, vol. 38, n°. 4, p. 342–366, avr. 2006.

[53] R. M. Rustamov, "Laplace-Beltrami Eigenfunctions for Deformation Invariant Shape Representation," dans *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, ser. SGP '07. Aire-la-Ville, Switzerland, Switzerland : Eurographics Association, 2007, p. 225–233. [En ligne]. Disponible : http://dl.acm.org/citation.cfm?id=1281991.1282022

[54] J. Sun, M. Ovsjanikov et L. Guibas, "A Concise and Provably Informative Multi-Scale Signature Based on Heat Diffusion," dans *Computer graphics forum*, vol. 28. Wiley Online Library, 2009, p. 1383–1392.

[55] A. E. Johnson et M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, n°. 5, p. 433–449, mai 1999.

[56] R. B. Rusu *et al.*, "Aligning point cloud views using persistent feature histograms," dans *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*. IEEE, 2008, p. 3384–3391.

[57] D. G. Lowe, "Object recognition from local scale-invariant features," dans *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, p. 1150–1157.

[58] M. Novotni et R. Klein, "3d zernike descriptors for content based shape retrieval," dans *Proceedings of the eighth ACM symposium on Solid modeling and applications*, 2003, p. 216–225.

[59] D. Kihara *et al.*, "Molecular surface representation using 3d zernike descriptors for protein shape comparison and docking," *Current Protein and Peptide Science*, vol. 12, n°. 6, p. 520–530, 2011.

[60] C. Gordon, D. Webb et S. Wolpert, "Isospectral plane domains and surfaces via riemannian orbifolds," *Inventiones mathematicae*, vol. 110, n°. 1, p. 1–22, 1992.

[61] M. M. Bronstein et I. Kokkinos, "Scale-invariant heat kernel signatures for non-rigid shape recognition," dans *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, p. 1704–1711.

[62] D. Raviv *et al.*, "Volumetric heat kernel signatures," dans *Proceedings of the ACM workshop on 3D object retrieval*, 2010, p. 39–44.

[63] V. Zobel, J. Reininghaus et I. Hotz, "Generalized heat kernel signatures," 2011.

[64] H. Hoppe *et al.*, "Surface reconstruction from unorganized points," dans *Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, 1992, p. 71–78.

[65] N. J. Mitra et A. Nguyen, "Estimating surface normals in noisy point cloud data," dans *Proceedings of the nineteenth annual symposium on Computational geometry*, 2003, p. 322–328.

[66] R. Bro, E. Acar et T. G. Kolda, "Resolving the sign ambiguity in the singular value decomposition," *Journal of Chemometrics : A Journal of the Chemometrics Society*, vol. 22, n°. 2, p. 135–140, 2008.

[67] S. M. Prakhya, B. Liu et W. Lin, "B-shot : A binary feature descriptor for fast and efficient keypoint matching on 3d point clouds," dans *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, p. 1929–1934.

[68] D. V. Vranic et D. Saupe, "3d model retrieval." Thèse de doctorat, University of Leipzig PhD thesis, 2004.

[69] P. Papadakis *et al.*, "Efficient 3d shape matching and retrieval using a concrete radialized spherical projection representation," *Pattern Recognition*, vol. 40, n°. 9, p. 2437–2452, 2007.

[70] N. Canterakis, "3d zernike moments and zernike affine invariants for 3d image analysis and recognition," dans *In 11th Scandinavian Conf. on Image Analysis*. Citeseer, 1999.

[71] L. Holm et C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of molecular biology*, vol. 233, n°. 1, p. 123–138, 1993.

[72] ——, "Mapping the protein universe," *Science*, vol. 273, n°. 5275, p. 595–602, 1996.

[73] I. N. Shindyalov et P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (ce) of the optimal path." *Protein engineering*, vol. 11, n°. 9, p. 739–747, 1998.

[74] S. B. Needleman et C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, n°. 3, p. 443–453, 1970.

[75] Y. Zhang et J. Skolnick, "Tm-align : a protein structure alignment algorithm based on the tm-score," *Nucleic acids research*, vol. 33, n°. 7, p. 2302–2309, 2005.

[76] ——, "Scoring function for automated assessment of protein structure template quality," *Proteins : Structure, Function, and Bioinformatics*, vol. 57, n°. 4, p. 702–710, 2004.

[77] S. Wang *et al.*, "Protein structure alignment beyond spatial proximity," *Scientific reports*, vol. 3, p. 1448, 2013.

[78] L. Mavridis *et al.*, "Shrec-10 track : Protein models," dans *3DOR : Eurographics Workshop on 3D Object Retrieval*, 2010, p. 117–124.

[79] N. Song *et al.*, "Protein shape retrieval," 2017.

[80] F. Langenfeld *et al.*, "Shrec 2018-protein shape retrieval," 2018.

[81] ——, "Shrec19 protein shape retrieval contest," 2019.

[82] ——, "Shrec2020 track : Multi-domain protein shape retrieval challenge," *Computers & Graphics*, 2020.

[83] P. Gainza *et al.*, "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," *Nature Methods*, vol. 17, n°. 2, p. 184–192, 2020.

[84] S. Daberdaku et C. Ferrari, "Exploring the potential of 3d zernike descriptors and svm for protein–protein interface prediction," *BMC bioinformatics*, vol. 19, n°. 1, p. 35, 2018.

[85] J. Kyte et R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of molecular biology*, vol. 157, n°. 1, p. 105–132, 1982.

[86] J. Masci *et al.*, "Geodesic convolutional neural networks on riemannian manifolds," dans *Proceedings of the IEEE international conference on computer vision workshops*, 2015, p. 37–45.

[87] L. Sael *et al.*, "Fast protein tertiary structure retrieval based on global surface shape similarity," *Proteins : Structure, Function, and Bioinformatics*, vol. 72, n°. 4, p. 1259–1273, 2008.

[88] S. Angenent *et al.*, "On the laplace-beltrami operator and brain surface flattening," *IEEE Transactions on Medical Imaging*, vol. 18, n°. 8, p. 700–711, 1999.

[89] S. Haker *et al.*, "Conformal surface parameterization for texture mapping," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, n°. 2, p. 181–189, 2000.

[90] L. E. Band, "Topographic partition of watersheds with digital elevation models," *Water resources research*, vol. 22, n°. 1, p. 15–24, 1986.

[91] D. H. Douglas, "Experiments to locate ridges and channels to create a new type of digital elevation model," *Cartographica : The International Journal for Geographic Information and Geovisualization*, vol. 23, n°. 4, p. 29–61, 1986.

[92] W. Zhang et D. R. Montgomery, "Digital elevation model grid size, landscape representation, and hydrologic simulations," *Water resources research*, vol. 30, n°. 4, p. 1019–1028, 1994.

[93] J. J. Koenderink et A. J. Van Doorn, "Surface shape and curvature scales," *Image and vision computing*, vol. 10, n°. 8, p. 557–564, 1992.

[94] S. Jacek, "Landform characterization with geographic information systems," *Photogrammetric Engineering & Remote Sensing*, vol. 63, n°. 2, p. 183–191, 1997.

[95] J. Iwahashi et R. J. Pike, "Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature," *Geomorphology*, vol. 86, n°. 3-4, p. 409–440, mai 2007. [En ligne]. Disponible : http://linkinghub.elsevier.com/retrieve/pii/S0169555X06004375

[96] J. F. White, "Convex-concave landslopes : a geometrical study," 1966. [En ligne]. Disponible : https://kb.osu.edu/dspace/handle/1811/5248

[97] F. A. Limberger et R. C. Wilson, "Feature encoding of spectral signatures for 3d non-rigid shape retrieval." dans *BMVC*, 2015, p. 56–1.

[98] K. Simonyan et A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv :1409.1556*, 2014.

Annexe A

Annexes

A.1 Annexe 1 : Résultats sur les cartes de convexité avec un apprentissage profond

Les cartes de convexité telles que présentées dans subsection 3.2.2 sont données en entrée du réseau de neurones VGG16 [98]. Aucun entraînement supplémentaire n'a été fait sur le réseau de neurone. Chaque carte de convexité produit un vecteur en sortie du réseau de neurones VGG16. C'est ce vecteur qui est utilisé pour comparer les cartes. La comparaison entre deux vecteurs est l'absolu de la différence.

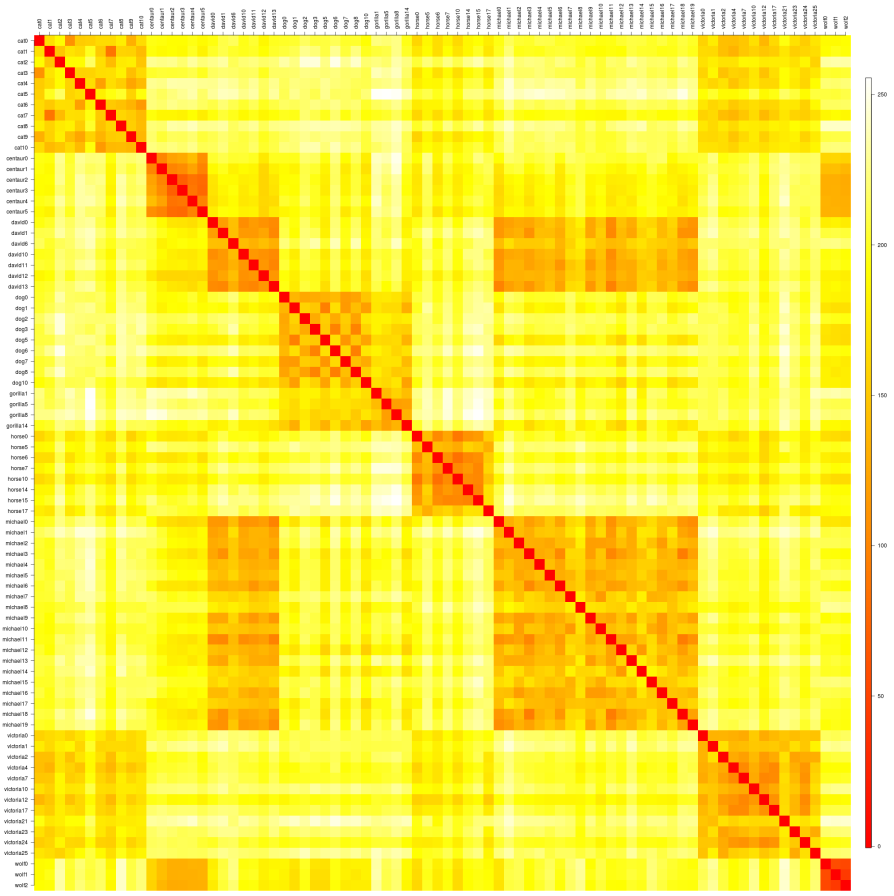


FIGURE A.1 – Matrice de score obtenue avec l'utilisation du réseau de neurones VGG16 sur les cartes de convexité du jeu de données TOSCA

Classes	Nearest Neighbour	First Tier	Second Tier
cat	1.000	0.555	0.845
centaur	1.000	0.967	1.000
david	0.857	0.524	0.857
dog	1.000	0.875	0.958
gorilla	1.000	0.875	0.958
horse	1.000	0.982	1.000
michael	0.900	0.679	1.000
victoria	1.000	0.818	0.992
wolf	1.000	1.000	1.000

TABLE A.1 – Performances obtenues avec l'utilisation du réseau de neurones VGG16 sur les cartes de convexité du jeu de données TOSCA

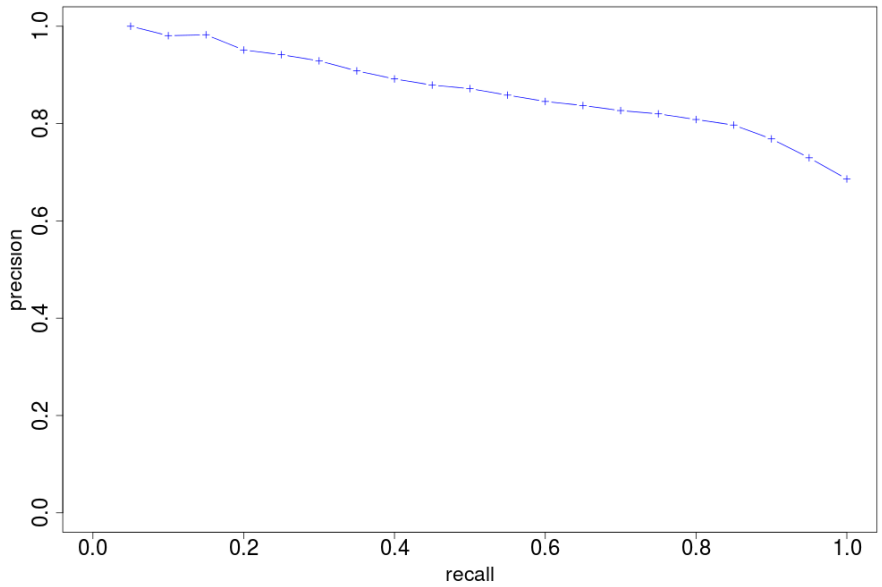


FIGURE A.2 – Courbe de précision rappel obtenue avec l'utilisation du réseau de neurones VGG16 sur les cartes de convexité sur le jeu de données TOSCA

Les différentes évaluations sont similaires à celle obtenues en faisant une comparaison de patchs. Le temps de comparaison est négligeable car c'est une comparaison globale sur des vecteurs de taille 512. Entraîner le réseau de neurones sur les cartes de convexité pourrait augmenter la précision.

A.2 Annexe 2 : Taille des fichiers des représentations

La taille des fichiers stockant les représentations ainsi que le temps de comparaison sont donnés dans le tableau ci-dessous. Le temps de comparaison est calculé pour la comparaison moyenne de deux protéines sur les protéines du jeu de données DSV (Figure 5.4 de la subsection 5.3.1). Ce sont les valeurs pour la version la plus récente et avec les paramètres donnant les performances les plus élevées.

Type de carte	Taille moyenne	Temps de comparaison
Carte de convexité	11 kB	99 sec
Carte de courbure	195 kB	2 sec
Carte WKS	424 kB	7 sec

TABLE A.2 – Taille moyenne des fichiers des différentes représentations proposées

Le temps de comparaison des cartes de convexité est plus élevé que pour les deux autres représentations. Ceci s'explique entre autres par le fait que la taille des patchs est plus grande (25×25) que les cartes de courbures (10×10). Une autre raison pouvant être que cette méthode n'a pas reçu les dernières optimisations faites après la création de la représentation par les cartes de courbure.

A.3 Annexe 3 : Exemple d'interpolation de carte de WKS

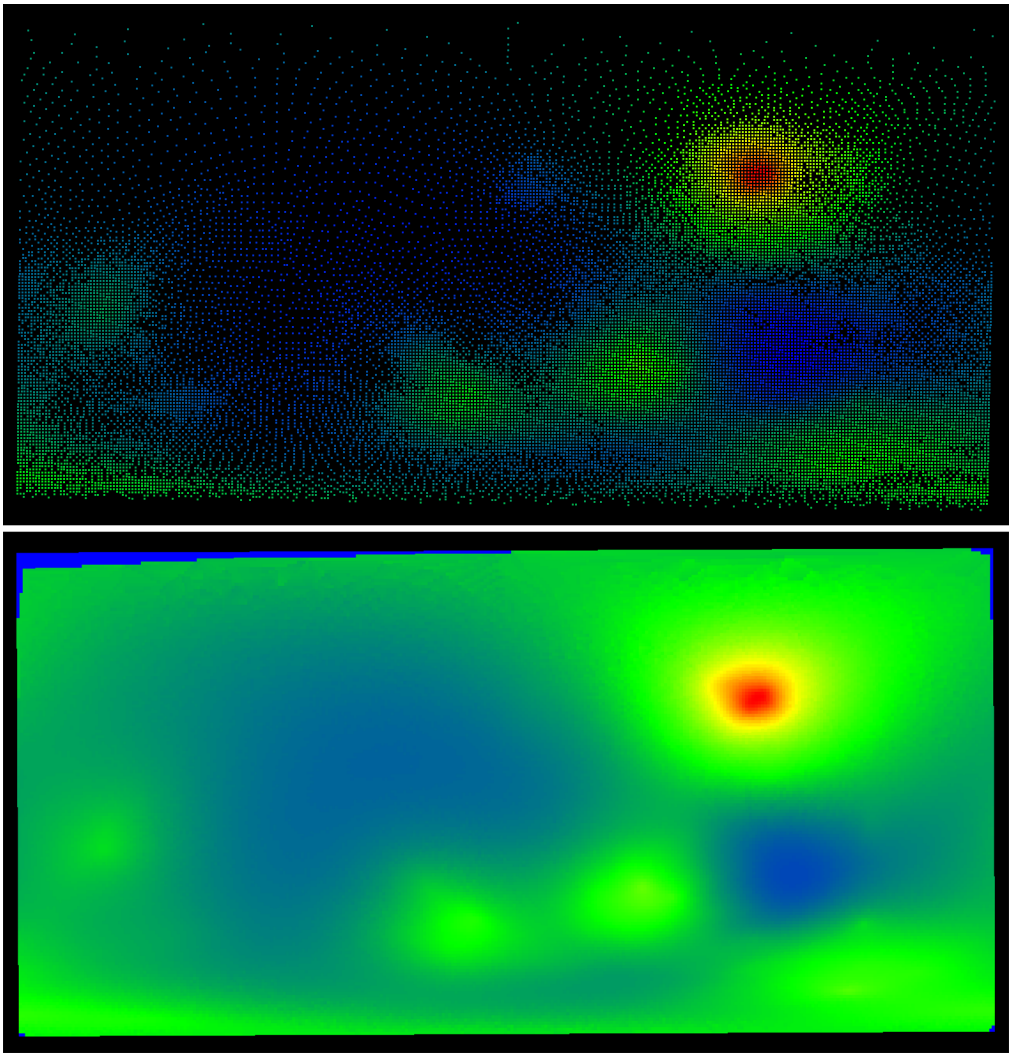


FIGURE A.3 – Exemple d'interpolation d'une carte de WKS

A.4 Annexe 4 : Calcul générique sur processeur graphique

Les comparaisons des différentes méthodes présentées dans ce manuscrit sont faites en utilisant la puissance de calcul de la carte graphique de l'ordinateur. Ceci permet de faire des calculs parallèles dans la mémoire locale de la carte graphique. La méthode utilisée est appelée **réduction parallèle** et est applicable sur des opérations binaires comme par exemple une addition ou un multiplication. Si $*$ est une opération binaire, T un tableau de taille n et V un tableau de taille $\frac{n}{2} \in \mathbb{N}$ alors :

$$\forall i \in [1, \frac{n}{2}], V[i] = T[i] * T[i + \frac{n}{2}] \tag{A.1}$$

Cette opération est répétée jusqu'à ce que la taille de V soit de 1. Si n est impaire, la valeur $T[n]$ est ajoutée selon l'opération binaire à une des autres valeurs de T durant l'initialisation.

L'implémentation de l'algorithme ce fait tel que décrit dans la Figure A.4. Un seul tableau est initialisé pour éviter une perte de temps en allocation mémoire. Lors de l'initialisation de ce tableau, l'opération binaire est réalisée une première fois. Ensuite, pour la k-ème étape, l'opération est répétée sur les $\frac{n}{2^k}$ premières cases du tableau jusqu'à obtenir le résultat final dans la première case.

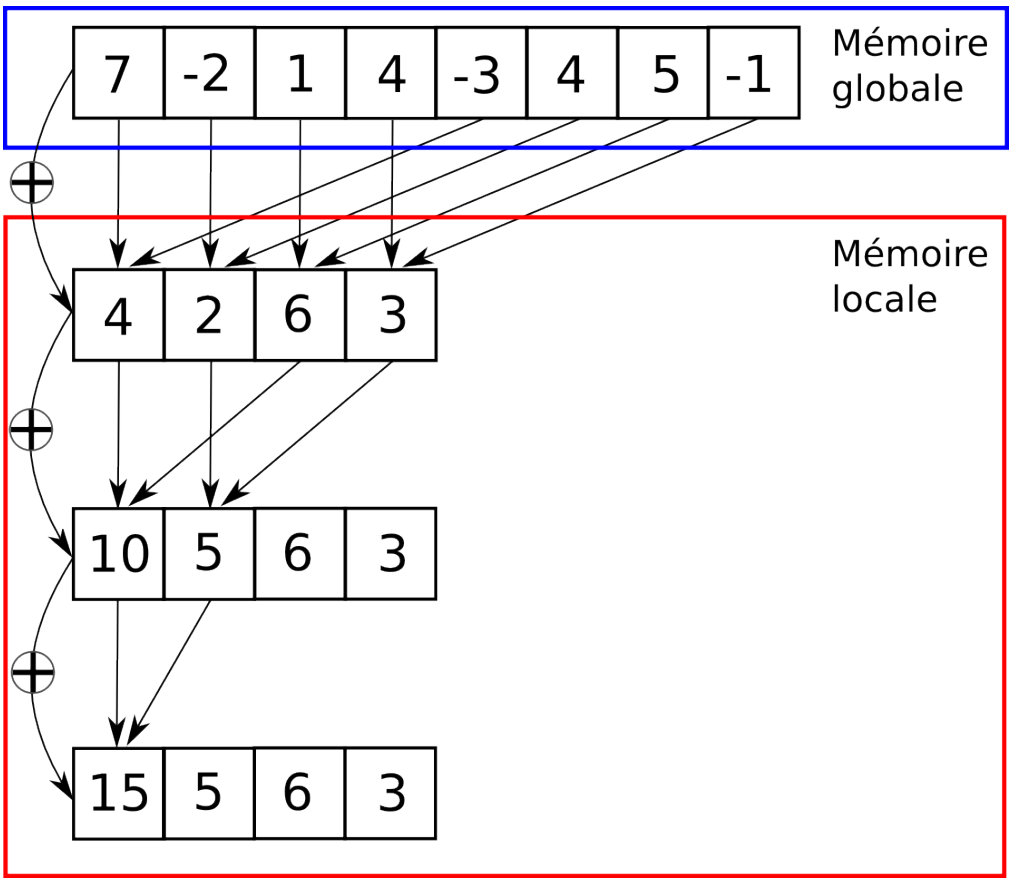


FIGURE A.4 – Schéma d'une réduction parallèle avec l'opération binaire de l'addition

Annexe B

Liste des acronymes

1D/2D/3D 1 Dimension / 2 Dimensions / 3 Dimensions.

3DZD *3D Zernike Descriptor*.

AFP *Aligned Fragments Pair*.

BLOSUM *BLOcks SUbstitution Matrix*.

BoF *Bag of Features*.

CCvx Carte de Convexité ou *Convexity Map*.

CE *Combinatorial Extension*.

CLESUM *Conformation LEtter SUbstitution Matrix*.

CM Carte de Courbure ou *Curvature Map*.

CPCA Analyse en composantes principales continues ou *Continuous Principal Component Analysis*.

Cryo-EM Cryo-microscopie électronique ou *cryo-electron microscopy*.

DALI *Distance-matrix alignment*.

DEM Modèle numérique de terrain ou *Digital Elevation Model*.

DFT Transformation / transformée de Fourier discrète ou *Discrete Fourier transform*.

DoG Différence de Gaussiennes ou *Difference of Gaussians*.

DWT Transformation / transformée d'ondelettes discrète ou *Discrete Wavelet Transform*.

EDP Equation aux Dérivées Partielles.

EDTSurf *Euclidean Distance Transform based Surface*.

EMD *Earth Mover's Distance*.

FPFH *Fast Point Feature Histograms*.

FT *First Tier*.

GHKS *Generalized Heat Kernel Signature*.

GPS *Global Point Signature*.

HKS *Heat Kernel Signature*.

MAP *Mean Average Precision*.

MASIF *MoleculAr Surface Interaction Fingerprint*.

MC *Marching Cube*.

NN *Nearest Neighbor*.

NPCA Analyse en composantes principales des normales ou *Normal Principal Component Analysis*.

PANORAMA *PANoramic Object Representation for Accurate Model Attributing*.

PDB *Protein Data Bank*.

PFH *Point Feature Histograms*.

PR Précision Rappel ou *Precision Recall*.

PSB *Princeton Shape Benchmark*.

PWKS *Projected Wave Kernel Signature*.

RMN Résonance Magnétique Nucléaire.

RMSD *Root Mean Square Deviation*.

SAS Surface Accessible au Solvant.

SES Surface Exclue au Solvant ou Surface de Conolly.

SFP *Similar Fragment Pairs*.

SHOT *Signature of Histograms of OrienTations*.

SIFT *Scale Invariant Feature Transform*.

SI-HKS *Scale Invariant Heat Kernel Signature*.

SPFH *Simplified Point Feature Histogram*.

SSS-DEM *Shape Similarity System driven by Digital Elevation Models*.

ST *Second Tier*.

TM-align *Template Modeling Align*.

TM-score *Template Modeling Score*.

VCMC *Vertex-Connected Marching Cube*.

VdW *Van der Waals*.

VGG *Visual Geometry Group*.

WKS *Wave Kernel Signature*.

le cnam

Léa SIRUGUE

Conception et développement d'une méthode de comparaison de surfaces appliquée aux protéines

le cnam

Résumé : Les interactions entre protéines jouent un rôle crucial dans les processus du vivant comme la communication cellulaire, l'immunité, la croissance, prolifération et la mort cellulaires. Ces interactions se font via leur surface et la perturbation des interactions entre protéines est à la base de nombreux processus pathologiques. Il est donc nécessaire de bien comprendre et caractériser la surface des protéines et leurs interactions mutuelles de manière à mieux comprendre les processus du vivant. Différentes méthodes de comparaison de la surface des protéines ont été développées ces dernières années mais aucune n'est assez puissante pour traiter l'ensemble des structures disponibles dans les différentes bases de données. Le projet de thèse est donc de développer des méthodes rapides de comparaison de surface et de les appliquer à la surface des macromolécules.

Mots clés : Biologie structurale, Reconnaissance de formes, Docking, Traitement du signal, Géométrie spectrale

Abstract : Protein interactions play a crucial role in the living processes such as cell communication, immunity, cell growth, proliferation and death. These interactions occur through the surface of proteins and the disruption of their interactions is the start of many disease processes. It is therefore necessary to understand and characterize the surface of proteins and their interactions to better understand living processes. Different methods of protein surfaces comparison have been developed in the recent years but none are powerful enough to handle all the structures currently available in databases. The PhD project is to develop rapid methods of surface comparison and apply them to the surface of macromolecules.

Keywords : Structural biology, Shape recognition, Molecular docking, Signal processing, Spectral geometry