Protein-protein docking based on shape complementarity and Voronoi fingerprint

Thomas Bourquard, Jérôme Azé, Anne Poupon, David Ritchie

Protein-protein docking based on shape complementarity and Voronoi fingerprint

Thomas BOURQUARD[1], Jerôme AZÉ[2], Anne POUPON[3] and David W. RITCHIE[1]

[1] INRIA Nancy-Grand Est, LORIA, 615 Rue du Jardin Botanique, 54600 Villers-lès-Nancy, France
{Thomas.Bourquard, Dave.Ritchie}@inria.fr

[2] INRIA AMIB group, Équipe Bioinformatique, CNRS UMR8623 Laboratoire de Recherche en Informatique, Université Paris-Sud, 91405 Orsay Cedex, France
Jerome.Aze@lri.fr

[3] Bios group, INRA, UMR85, Unité de physiologie de la Reproduction et des Comportements, F-37380 Nouzilly, France ; CNRS, UMR6175, F-37380 Nouzilly, France ; Université François Rabelais, 37041 Tours, France
Anne.Poupon@inra.fr

Abstract *Predicting the three-dimensional structures of protein-protein complexes is a major challenge for computational biology. Using a Voronoi tessellation model of protein structure, we showed previously that it was possible to use an evolutionary algorithm to train a scoring function to distinguish reliably between native and non-native docking conformations. Here, we show that this approach can be further improved by combining it with rigid body docking predictions generated by the Hex docking algorithm. This new approach is able to rank an acceptable or better conformation within the top 10 predictions for 7 out of the 9 targets available from rounds 8 to 18 of the CAPRI docking experiment.*

Keywords Protein-protein Docking, Evolutionary Algorithms, Hex, CAPRI.

Amarrage protéine-protéine par couplage de la complémentarité de forme et des empreintes Voronoï

Résumé *La prédiction de la structure tri-dimensionnelle des complexes protéine-protéine est un enjeu majeur pour la bioinformatique. Nous avions montré dans des travaux précédents que grâce à la modélisation par un diagramme de Voronoï de la structure des protéines, et à l'utilisation d'algorithmes évolutionnaires, il était possible d'optimiser des fonctions de score permettant de distinguer avec une bonne fiabilité les conformations natives des conformations non-natives. Nous montrons dans cet article que cette approche peut être sensiblement améliorée en combinant celle-ci avec des modèles en corps rigide générés par l'algorithme de docking Hex. Cette nouvelle approche, testée sur les cibles CAPRI des rounds 8 à 18, permet de classer dans les 10 meilleures, une conformation quasi-native pour 7 cibles sur les 9 disponibles.*

Mots-clés Amarrage protéine-protéine, Algorithmes Évolutionnaires, Hex, CAPRI.

1 Introduction

L'intégration des signaux extra-cellulaires en une réponse biologique adaptée repose en grande partie sur l'association de complexes protéine-protéine. La détection et la détermination de l'organisation structurale de ces assemblages moléculaires représente donc une étape essentielle pour la compréhension de ces mécanismes et de leur régulation. Si les techniques qui permettent la détermination expérimentale des structures protéiques ont connu des avancées fondamentales, notamment grâce aux projets de génomique structurale, cette détermination reste délicate voire impossible, surtout lorsque l'objet étudié est un complexe. De plus, il a été démontré expérimentalement que le nombre de complexes existant *in vivo* était bien supérieur au nombre de protéines, rendant inenvisageable le recours systématique à l'expérimentation. L'amarrage protéine-protéine, qui consiste à prédire la structure tridimensionnelle de ces assemblages macromoléculaires à partir des structures des partenaires isolés, serait donc un outil crucial dans l'étude du fonctionnement de la cellule [14].

Les différentes procédures existantes traitent généralement le problème en deux étapes : (i) une première phase au cours de laquelle un grand nombre de conformations sont générées (étape limitante en temps de calcul), (ii) puis une seconde phase au cours de laquelle ces différentes conformations sont évaluées afin d'en extraire un sous-ensemble de conformations proches de la conformation native, que nous appellerons conformations quasi-natives.

L'implémentation de l'algorithme de complémentarité de formes Hex sur cartes graphiques (GPU) a permis de réduire considérablement le temps nécessaire pour l'échantillonnage statistique des quelques 10^9 modes d'associations possibles pour deux protéines de taille moyenne [11]. Cet algorithme est capable de générer et évaluer en quelques secondes plusieurs millions de conformations candidates afin d'en extraire un ensemble réduit de conformations d'intérêt [15]. Cependant la fonction d'évaluation intégrée dans Hex ne permet pas d'identifier de manière fiable une solution quasi-native dans cet ensemble.

Dans des travaux précédents, nous avons pu montrer que la représentation des structures protéiques par un modèle "gros-grain" basé sur la tessellation de Voronoï décrivait particulièrement bien les propriétés physico-chimiques aux interfaces protéine-protéine [3]. Ce modèle, couplé à un algorithme évolutionnaire, permet d'optimiser des fonctions de score pour l'amarrage protéine-protéine [4,3,6]. Néanmoins, ces fonctions de score ne sont pas suffisamment sensibles pour envisager l'exploration de l'interactome à grande échelle.

Dans ce travail, nous montrons que la génération de conformations candidates et l'évaluation de la complémentarité de forme par Hex, couplées à l'évaluation des caractéristiques physico-chimiques par les empreintes Voronoï, permettent une prédiction particulièrement efficace de la conformation de complexes protéine-protéine. Afin d'évaluer cette approche, nous nous sommes placés dans le cadre de l'expérience CAPRI[1] [9]. L'objectif de CAPRI est l'évaluation des méthodes d'amarrage protéine-protéine. Des complexes dont la structure tridimensionnelle a été résolue, mais pas encore rendue publique, sont proposés aux prédicteurs. Le processus se déroule en deux étapes : les prédicteurs proposent 10 candidats. Puis ils déposent une centaine de candidats, qui sont alors proposés aux "scoreurs", ce qui permet de tester les fonctions d'évaluation indépendamment de la génération des conformations candidates [10]. Nous présentons dans cet article les résultats obtenus par notre méthode pour les rounds 8 à 18 de cette expérience de "scoring".

2 Méthodes

2.1 Base d'apprentissage des complexes protéine-protéine

Les complexes utilisés pour les procédures d'apprentissage correspondent à ceux utilisés précédemment [6], auxquels nous avons ajouté les complexes des benchmarks 3.0 et 4.0 proposés par *Hwang et al.*[7,8] qui n'étaient pas déjà présents. Ce jeu d'apprentissage comprend 231 complexes liés-non liés ou non liés-non liés (complexes pour lesquels la structure d'au moins un des partenaires isolé est connue). Tous les complexes retenus ont été comparés deux-à-deux suivant la classification SCOP [12] afin d'éviter toute redondance.

Le jeu d'apprentissage est composé de structures natives, correspondant aux structures expérimentales, et de structures non-natives associées. Les conformations non-natives ont été générées avec le logiciel Hex. Pour un complexe donné, Hex recherche la conformation dans laquelle la complémentarité géométrique est la meilleure. Cela permet de définir un axe de référence reliant les centres de gravité des deux partenaires. Les solutions explorées sont alors celles pour lesquelles l'axe reliant les centres de gravité se trouve dans deux cônes dont les sommets sont les centres de gravité, et dont l'axe central est cet axe de référence. Les angles définissant ces cônes peuvent être choisis par l'utilisateur entre 0 et 180 $^\circ$. Dans cette étude, ces deux angles ont été fixés à 45 $^\circ$ car nous avons pu constater que des valeurs supérieures n'augmentaient pas la probabilité de générer des conformations quasi-natives, mais augmentaient très fortement le nombre de conformations non-natives. Afin d'éliminer les modèles trop proches les uns des autres, le seuil de clustering de Hex a été fixé à $9.0\mathring{A}$ Root Mean Square Deviation ($RMSD$). Les examples négatifs du jeu d'apprentissage ont été choisis dans cet ensemble de conformations, et correspondent aux conformations non-natives (ayant un RMSD avec la conformation native supérieur à $10\mathring{A}$) de plus basse énergie trouvés par Hex, et ayant une surface d'interface supérieure à $400\mathring{A}^2$.

[1] **C**ritical **A**ssessment of **PR**edictions of **I**nteractions

10 structures non natives pour chaque structure native ont été incluses dans le jeu d'apprentissage (19 dans la comparaison cœur-couronne).

2.2 Empreintes Voronoï et paramètres d'apprentissage

Le modèle "gros-grain" défini dans [6], basé sur la tesselation de Voronoï, a été utilisé pour représenter les structures des complexes. Pour chaque conformation candidate, la triangulation de Delaunay (duale de la tesselation de Voronoï) est construite par utilisation de la CGAL [5]. L'interface est définie comme l'ensemble des acides aminés d'un partenaire en contact avec l'autre partenaire. Cette interface est, soit restreinte aux acides aminés qui ne sont pas en contact avec le solvant : interface cœur, soit non restreinte : interface cœur plus couronne.

Pour chaque conformation, un vecteur de 96 paramètres est calculé et utilisé dans les procédures d'apprentissage ou de test. Ce vecteur comprend le nombre total de résidus à l'interface, l'aire de l'interface, les fréquences et volumes moyens des cellules de Voronoï de chaque type de résidu, les distances et fréquences de paires de résidus regroupés en six catégories physico-chimiques (hydrophobe (IFMLV), aromatique (FYW), polaire (NQ), chargé positivement (HKR), chargé négativement (DE) et petits (AGSTCP)), les fréquences et les volumes moyens de chaque catégorie de résidus (voir [6]).

2.3 Algorithme évolutionnaire et procédure d'apprentissage

À l'aide des attributs d'apprentissage décrits plus haut, des algorithmes évolutionnaires ont été utilisés afin de trouver un ensemble de fonctions permettant de discriminer les conformations quasi-natives et non-natives. La fonction d'adaptation utilisée est l'aire sous la courbe de ROC (Receiver Operating Characteristic). Les fonctions de score apprises dans cette étude sont de la forme :

$$S_j(conf) = \sum_{i=1}^{96} w_i \, |x_i(conf) - c_i|$$

où pour chaque attribut d'apprentissage X_i, x_i, w_i et c_i représentent respectivement les valeurs, poids et valeurs de centrage associés, w_i et c_i étant optimisés au cours de l'apprentissage. L'algorithme évolutionnaire est de type $\lambda + \mu$, avec $\lambda = 20$ parents $\mu = 120$ enfants. Le maximum de générations a été fixé à 500. Les performances ont été évaluées en validation croisée. Un apprentissage correspond à l'optimisation de 30 fonctions de score, et le rang final d'une conformation correspond à la somme des rangs obtenus après application de chacune des 30 fonctions apprises.

Les fonctions de score sont évaluées par la précision et le rappel :

$$\text{Précision} = \frac{VP}{VP+FP} \qquad \text{Rappel} = \frac{VP}{VP+FN}$$

Où VP : vrais positifs, FP : faux positifs et FN : faux négatifs.

2.4 Gestion des valeurs manquantes et Normalisation

Dans des travaux précédents, nous avions constaté que les valeurs manquantes ont un impact négatif très important sur les performances des fonctions apprises. Afin de limiter cet impact, nous avons testé plusieurs méthodes de gestion des valeurs manquantes. Nous avons retenu les méthodes les plus fréquemment utilisées pour gérer des valeurs manquantes [1] : remplacement par une valeur constante (0), par une valeur dépendant des données manipulées (valeur minimale, maximale, médiane ou moyenne de l'attribut considéré) ou par des valeurs obtenues sur un sous-ensemble des données manipulées (calcul des exemples les plus proches et remplacement par la valeur moyenne : $k$nn ou $k$means).

Le remplacement des valeurs manquantes par l'approche $k$nn est réalisée de la manière suivante : pour chaque exemple ayant au moins une valeur manquante, ses k plus proches voisins sont recherchés en utilisant une distance euclidienne calculée uniquement entre les valeurs renseignées de l'exemple considéré et le reste

des données disponibles. Puis, pour chaque attribut non renseigné, la valeur manquante est remplacée par la valeur moyenne de cet attribut dans ses k plus proches voisins. Si l'ensemble des plus proches voisins est vide (trop de valeurs manquantes par exemple), ou que pour un attribut les plus proches voisins sont tous non renseignés, alors les valeurs manquantes sont remplacées par les valeurs moyennes calculées sur l'intégralité des données.

Pour l'approche $k$means, les données sont préalablement réparties en k clusters les plus homogènes possibles. La distance intra-cluster est calculée de la même manière que pour l'approche $k$nn. Les clusters sont initialisés avec les exemples contenant le moins de valeurs manquantes (moins de 10%). Puis, dans chaque cluster, les valeurs manquantes des exemples sont remplacées par les valeurs moyennes calculées sur les exemples du clusters. De manière similaire à l'approche $k$nn, si un attribut n'est jamais renseigné dans le cluster, alors la valeur moyenne globale est utilisée pour remplacer les valeurs manquantes de cet attribut.

Enfin, les intervalles de valeurs admissibles pour chaque paramètre sont par définition très hétérogènes. Bien que ces différences d'échelles soient en partie capturées par l'algorithme évolutionnaire via les valeurs de centrage c_i, l'ensemble des attributs dont les valeurs admissibles sont élevées peuvent atténuer voire complètement masquer les attributs ayant des valeurs plus faibles.

Deux procédures de normalisation des données ont été mises en œuvre afin de réduire ce biais :
- la procédure **minMax**, qui normalise les attributs en fonction du minimum et du maximum observés pour le paramètre :

$$x_i(conf) = \frac{x_i(conf) - min(X_i)}{max(X_i) - min(X_i)}$$

- la procédure **meanStd**, qui normalise les attributs en fonction de la moyenne et l'écart-type :

$$x_i(conf) = \frac{x_i(conf) - \bar{X_i}}{\sigma_i}$$

À l'issue de ces deux étapes de pré-traitement, une dernière étape de sélection d'attributs aurait pu être mise en place et ainsi réduire ce problème de valeurs manquantes. Notre choix de représentation des complexes, et notamment le grand nombre de paramètres utilisés, implique nécessairement qu'une partie de ces paramètres soient non renseignés pour un example donné. Cependant, mis à part les paramètres concernant les acides aminés les plus représentés dans les protéines, les paramètres non renseignés varient d'un exemple à l'autre, reflètant la diversité des modes d'interaction, elle-même liée à la diversité des protéines. Considérons un complexe dont l'interface comporte un tryptophane. L'attribut "volume moyen du tryptophane" est essentiel pour la prédiction de cette interface. Or, le tryptophane est un acide aminé très peu représenté dans les protéines, et les attributs correspondant seraient très certainement éliminés par une sélection de paramètres, rendant difficile la prédiction correcte de la structure de ce complexe.

Ainsi, une phase de sélection d'attributs risquerait de nous faire perdre la capacité de représenter efficacement des complexes faisant intervenir, dans leur interface, des résidus peu fréquents dans l'ensemble des complexes étudiés.

2.5 Classification des conformations

Pour classer les conformations nous avons utilisé les critères définis dans l'expérience CAPRI :
- Haute qualité : $[fnat \geq 0.5$ et $(I_{RMSD} \leq 1$ ou $L_{RMSD} \leq 1)]$
- Moyenne qualité : $[(fnat \geq 0.3$ et $fnnat < 0.5)$ et $(I_{RMSD} \leq 2.0$ ou $L_{RMSD} \leq 5.0)]$ ou $[fnat > 0.5$ et $(I_{RMSD} > 1.0$ ou $I_{RMSD} > 1.0)]$
- Acceptable : $[(fnat \geq 0.1$ et $fnnat < 0.1)$ et $(I_{RMSD} \leq 4.0$ ou $L_{RMSD} \leq 10.0)]$ ou $[fnat > 0.3$ et $(L_{RMSD} > 5.0$ ou $I_{RMSD} > 2.0)]$

Où $fnat$ est la fraction de contacts natifs présents dans la prédiction, $fnnat$ est la fraction de contacts de la prédiction qui sont natifs, I_{RMSD} est le RMSD entre l'interface prédite et l'interface native, L_{RMSD} est le RMSD entre le ligand prédit et le ligand natif, les récepteurs étant superposés.

3 Résultats

3.1 Interface cœur vs Interface couronne

La première question que nous adressons ici est de savoir si les résidus de la couronne, à savoir les résidus de l'interface qui sont en contact avec le solvant, doivent ou non être pris en compte dans l'apprentissage et l'évaluation. Ne pas inclure ces résidus revient à éliminer environ $2/3$ de l'aire de l'interface, et surtout augmente considérablement le nombre de valeurs manquantes. En effet, le pourcentage de valeurs non renseignées pour les structures natives passe de $12,63\%$ en ne considérant que le cœur, à seulement $3,63\%$ en ajoutant la couronne, et de $16,9\%$ à $5,1\%$ pour les conformations non-natives.

Cependant, de nombreuses études ont montré que le cœur et la couronne présentent des caractéristiques physico-chimiques nettement distinctes, ce qui n'est pas favorable dans notre cas. La prise en compte de ces résidus conduit à définir des interfaces contenant plus de résidus polaires et chargés, des volumes moyens associés aux cellules de Voronoï plus importants ou encore des distances entre paires de résidus en interaction également plus grandes. De même, toutes les déviations standards sont plus élevées.

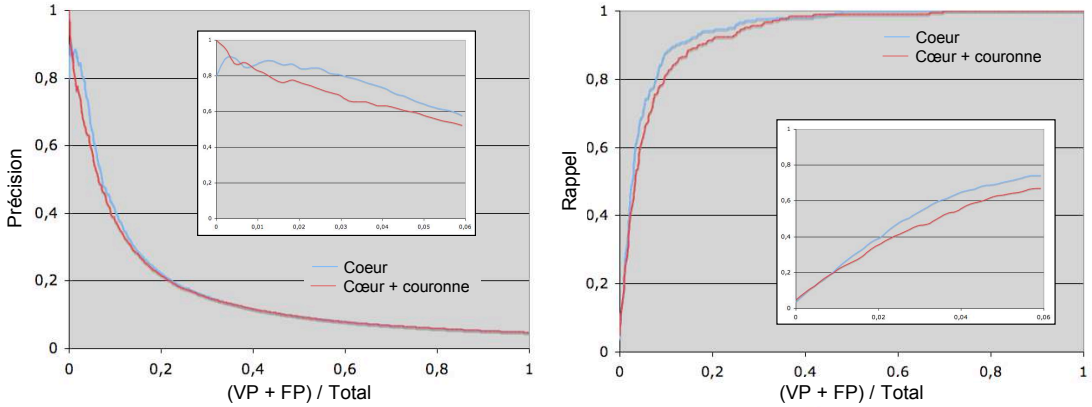

Figure 1. Précision et rappel en fonction de la fraction évaluée positive (VP + FP)/Total, pour un apprentissage en 10 validation croisée, en interface cœur (bleu) ou cœur+ couronne (rouge). La région correspondant aux fractions allant de 0 à 0,06 a été aggrandie (encadrés).

Nous avons réalisé, sur le jeu d'apprentissage en 10-validation croisée, une série d'apprentissages avec les résidus du cœur, et une série d'apprentissages avec les résidus du cœur et de la couronne. Les mesures de précision et rappel montrent que la prise en considération des résidus de cœur uniquement donne de meilleurs résultats (voir Fig. 1). Idéalement, étant donné que le jeu d'apprentissage contient 19 négatifs pour 1 positif, lorsque $(VP + FP)/Total = 0,05$, c'est-à-dire lorsque la fraction de conformations évaluées positives est égale à la fraction de conformations réellement positives, on devrait avoir une précision de 1 (toutes les conformations évaluées positives sont positives), et un rappel de 1 (toutes les conformations positives sont évaluées positives). À cette abscisse, nous obtenons une précision de $0,66$ en interface cœur, contre $0,6$ en interface couronne et des rappels de respectivement $0,69$ contre $0,62$. Ainsi, le "bruit" résultant de la prise en compte des résidus de la couronne à un impact négatif qui est plus important que l'impact positif résultant de la diminution du taux de valeurs manquantes. Par la suite, seuls les résidus du cœur de l'interface seront utilisés.

3.2 Gestion des valeurs manquantes et variants normalisés

Le fait que les résidus de la couronne ne puissent pas être utilisés rend la gestion des valeurs manquantes d'autant plus importante. Par ailleurs, les différents paramètres ayant des valeurs dans des ordres de grandeurs très différents, il est nécessaire de déterminer si les valeurs doivent être normalisées, et si oui par quelle méthode. Afin de répondre à ces deux questions, nous avons réalisé des apprentissages en 3-validation croisée en faisant varier la normalisation des données et le remplacement des valeurs manquantes.

Les résultats obtenus (Table 1) montrent que la normalisation améliore de manière très sensible les performances des fonctions de score. En effet, quelle que soit la méthode de gestion des valeurs manquante utilisée,

Normalisation	zéro	min	max	med	moy	$k$means	$k$nn $k=3$	$k$nn $k=5$	$k$nn $k=10$
aucune	0,62	0,63	0,78	0,72	0,68	0,72	0,68	0,69	0,70
minMax	0,80	0,80	**0,81**	**0,81**	0,80	0,80	0,80	0,80	0,80
meanStd	0,78	**0,81**	0,74	0,79	0,79	0,78	0,78	0,79	0,80

Table 1. Valeurs des critères de ROC obtenus pour différentes méthodes de gestion des valeurs manquantes et avec ou sans normalisation des données.

le critère de ROC est plus élevé avec normalisation que sans. Dans la majorité des cas, la méthode minMax semble plus performante, excepté lorsque les valeurs manquantes sont remplacées par le minimum observé. En ce qui concerne les méthodes de gestion des valeurs manquantes, trois des méthodes donnent des résultats équivalents : remplacement par le minimum, le maximum et la médiane. Dans la suite de cette étude nous avons appliqué le remplacement des valeurs manquantes par le maximum observé et la normalisation via l'approche minMax.

On peut noter ici que la précision obtenue sans normalisation, et en remplaçant les valeurs manquantes par la moyenne, correspondant à la configuration que nous utilisions précédemment, est supérieure à celle que nous avions obtenue sur notre précédent jeu d'apprentissage "(0.62)". Ceci est uniquement dû à l'utilisation des structures non-natives générées par Hex. L'amélioration de la précision est lié au fait que ces conformations non-natives sont plus "vraisemblables" que celles précédemment utilisées : bonne complémentarité géométrique et bonne énergie d'interaction en particulier.

3.3 Résultats sur les cibles CAPRI

Afin de vérifier la validité de notre approche, nous avons repris l'expérience de "scoring" CAPRI des rounds 8 à 18, et comparé les résultats avec ceux obtenus par les autres participants. Certaines cibles ont été éliminées de l'étude :

- Les cibles 23, 24, 26, 27 et 28 car les classements selon les critères CAPRI ne sont pas disponibles.
- La cible 30 car il s'agit d'un homodimère, le fait que ce dimère soit biologique est par ailleurs encore en discussion, et les auteurs n'ont pu le démontrer expérimentalement.
- La cible 31 car la structure native n'est pas disponible, il n'est donc pas possible d'évaluer les résultats.
- Les cibles 36 et 38 car aucune conformation au moins acceptable n'est présente dans les ensembles de conformations.
- Les cibles 33 et 34 car il s'agit de complexes protéine-ARN.

Les résultats obtenus pour les cibles restantes sont présentés dans la Table 2. Notre méthode est capable de classer une solution de qualité moyenne ou haute pour 7 des 9 cibles (voir Fig. 2), ce qui en fait la méthode la plus performante. Le cas de la cible 40 est particulièrement intéressant. En effet, ce complexe est un trimère constitué de l'inhibiteur de protéinase à sérine A, et de deux trypsines cationiques [2]. Il y a ainsi deux interfaces, notées T40A et T40B, et un seul ensemble de conformations candidates. Notre méthode classe une conformation de haute qualité pour l'interface A en première position et une conformation de haute qualité pour l'interface B en seconde position. Il y a par ailleurs dans le top 10 une autre conformation quasi-native pour chacune des deux interfaces.

Le cas de la cible 35[13], pour laquelle aucun groupe n'est parvenu à isoler une solution au moins acceptable dans le top10, est un peu particulier. En effet, il ne s'agit pas réellement de deux protéines, mais de deux domaines de la même protéine qui ont été artificiellement séparés, puis co-cristallisés. Or, nous avons déjà montré dans des travaux précédents que les valeurs moyennes de nos paramètres sont significativement différentes à l'intérieur des protéines et à l'interface entre deux protéines.

Pour la cible 39, aucun participant n'a été capable d'extraire une bonne solution parmi les conformations proposées par l'ensemble des prédicteurs. Ceci s'explique en grande partie par le fait que la structure de l'un des deux partenaire n'était pas connue, et a été modélisée avec un succès relativement mitigé. De ce fait, il y a seulement 4 conformations au moins acceptables dans l'ensemble proposé (3 de qualité moyenne et 1 acceptable) pour 1 296 conformations incorrectes.

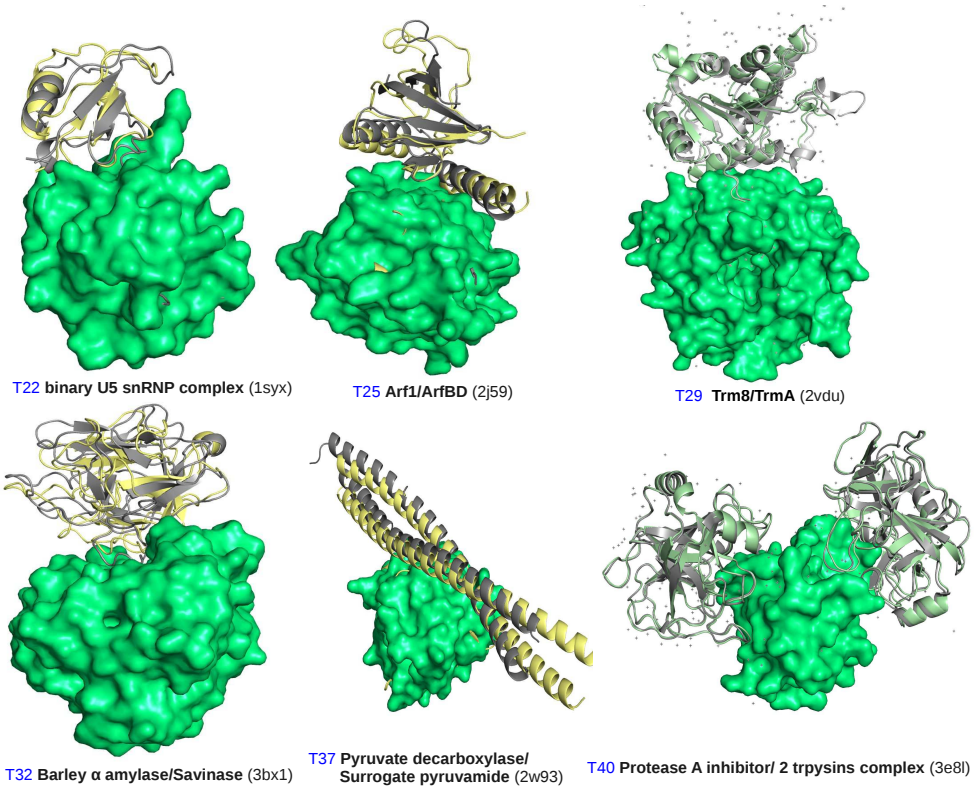

Figure 2. Superpositions des structures des complexes obtenues par cristallographie avec les structures au moins acceptables obtenues par évaluation lors de la seconde phase CAPRI présentes dans les 10 meilleures conformations ou "Top10". Les structures cristallines sont en gris, le recepteur en vert, le ligand représenté en ruban apparaît en jaune (Moyenne qualité) ou vert (Haute qualité).

Groups	T22	T25	T29	T32	T35	T37	T39	T40A	T40B
C Wang	0	**	0	0	*	**	0	***	**
A.M.J.J Bonvin	*	*	**	0	0	*	0	***	**
H. Wolfson	-	**	0	0	0	*	0	*	0
P. A. Bates	-	-	**	0	0	***	0	***	0
Z. Weng	-	-	**	0	0	***	0	***	0
J. F.-Recio	-	**	***	0	0	0	0	0	0
X. Zou	-	-	-	0	0	***	0	***	***
T. Haliloglu	-	-	-	-	-	**	0	***	**
C. J. Camacho	-	-	**	-	-	-	-	***	***
M. Takeda-Shitaka	-	-	0	0	0	-	-	***	**
I. Vakser	-	-	-	**	0	0	0	-	-
VDOCK	0	*	**	***	0	0	0	*	0
VDOCK-Hex Models	**	**	***	**	0	**	0	***	***
	(6)	(1)	(10)	(1)	(145)	(1)	(80)	(1)	(2)

Table 2. Meilleures conformations détectées dans les top 10 des différents scorers. **0** : Aucune solution au moins acceptable n'a été trouvée ; **-** : scorer n'ayant pas participé. Pour notre méthode (VDOCK-Hex Models) lorsqu'aucune conformation au moins acceptable n'est présente dans le top 10 le rang de la première conformation quasi-native est indiqué entre parenthèses.

4 Conclusion

La génération d'exemples négatifs de très basse énergie par Hex, la restriction de l'étude aux résidus appartenant au cœur de l'interface, la normalisation des valeurs des paramètres, et enfin la bonne gestion des valeurs manquantes, nous ont permis d'améliorer considérablement les performances de notre méthode. L'impact, au niveau de l'apprentissage, des exemples négatif générés par Hex est très important.

Cependant, si les choix faits à la suite de cette étude permettent une meilleure performance globale, dans certains cas particuliers ils ont un impact négatif. Par exemple, le remplacement des valeurs manquantes par 0 permet d'améliorer le classement de la première solution de haute qualité pour la cible 29. D'autre part, nous avons pu également montrer que dans certains cas l'utilisation des interfaces cœur plus couronne était plus performante. Il serait donc intéressant de mieux définir les cas dans lesquels ces méthodes alternatives sont plus performantes afin de permettre un choix de la méthode la plus adaptée en fonction du complexe à prédire.

Remerciements

Ce projet a été supporté par le programme ANR-08-CEXC-017-01.

Références

[1] E. Acuna and C. Rodriguez. The treatment of missing values and its effect in the classifier accuracy. *Classification, Clustering and Data Mining Applications*, pages 639–648, Feb 2004.

[2] R. Bao, J.C. Zhou, C. Jiang, S.X. Lin, C.W. Chi, and Y. Chen. The ternary structure of double-headed arrowhead protease inhibitor api-a complexed with two trypsins reveals a novel reactive site conformation. *J Biol Chem*, 284 :26676–84, 2009.

[3] J. Bernauer, J. Azé, J. Janin, and A. Poupon. A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics*, 23(5) :555–62, 2007.

[4] J. Bernauer, R. P. Bahadur, F. Rodier, J. Janin, and A. Poupon. DiMoVo : a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics*, 24(5) :652–8, 2008.

[5] J.-D. Boissonnat, O. Devillers, S. Pion, M. Teillaud, and M. Yvinec. Triangulations in CGAL. *Comput. Geom. Theory Appl.*, 22 :5–19, 2002.

[6] T. Bourquard, J. Bernauer, J. Azé, and A. Poupon. Comparing Voronoi and Laguerre tessellations in the protein-protein docking context ;. In *Sixth International Symposium on Voronoi Diagrams (ISVD)*, pages 225–232, 2009.

[7] H. Hwang, B. Pierce, J. Mintseris, and Z. Janin, J.and Weng. Protein-protein docking benchmark version 3.0. *Proteins*, 73(3) :705–9, 2008.

[8] H. Hwang, T. Vreven, J. Janin, and Z. Weng. Protein-protein docking benchmark version 4.0. *Proteins*, 78(15) :3111–4, 2010.

[9] J. Janin, K. Henrick, J. Moult, LT Eyck, MJ Sternberg, S Vajda, I Vakser, and SJ. Wodak. CAPRI : a Critical Assessment of PRedicted Interactions. *Proteins*, 52 :2–9, 2003.

[10] J. Janin and S.J. Wodak. The Third CAPRI Assessment meeting. *Structure*, 15 :755–759, 2007.

[11] J.C. Mitchell, R. Kerr, and LF. Ten Eyck. Rapid atomic density methods for molecular shape characterization. *J Mol Graph Model*, 19 :325–30, 2001.

[12] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. Scop : a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247 :536–540, 1995.

[13] S. Najmudin, BA Pinheiro, JAM Prates, MJ Romao, and CMGA Fontes. Putting an n-terminal end to the clostridium thermocellum xylanase xyn10b story : Crystallographic structure of the cbm22-1-gh10 modules complexed with xylohexaose. *Journal of Structural Biology*, 172 :353–362, 2010.

[14] D.W. Ritchie. Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci*, 9(1) :1–15, 2008.

[15] D.W. Ritchie and V. Venkatraman. Ultra-fast FFT protein docking on graphics processors. *Bioinformatics*, 26(19) :2398–405, 2010.