Thème 4 : prot-prot docking

Art 1 : Recent Progress and Future Directions in Protein-Protein Docking 2008

Abstract :

Le docking prot-prot : expliquer comment deux sous-unités protéiques se lient pour former un complexe.

Les algorithmes actuels utilisent des fonctions de score qui prennent en compte la désolvation, l'hydrophobicité et les interactions électrostatiques mais pour le moment il est toujours difficile de discerner la bonne solution des faux-positifs et des erreurs.

Cet article va expliquer les progrès qu'on a pu faire pour améliorer l'identification et les futures directions de recherches.

Le but est d'avoir des explications plus précises des interactions qu'il peut y avoir entre plusieurs protéines et de pouvoir prédire les interactions.

Introduction :

Il existe plusieurs types d'interactions entre protéines : **les short-times** (enzymes qui catalysent des protéines) ou **les long-times** (systèmes multimériques, récepteurs-inhibiteurs).

Protein docking : arriver à calculer la structure 3D d'un complexe protéique en partant des sous-unités. A priori, sans information quant à la forme et les interactions du complexe… Nous ne nous focalisons plus sur **quelle** protéine interagit avec quelle autre mais bien sur **COMMENT** elles interagissent.

On a de plus en plus de modèle de protéines mais la structure des complexes reste encore difficile à mettre en évidence pour des raisons pratiques.

Exemple : Cristalliser un complexe peut le faire changer de forme…

Pourquoi est-ce si important d'avoir une meilleure vision de ces interactions ?
Notamment pour pouvoir synthétiser des médicaments qui cibleront les protéines responsables de certaines maladies…

Le premier algorithme de docking a 30 ans.

Les algorithmes actuels utilisent déjà une large gamme de stratégie de score, il affine leurs fonctions en ajoutant les modèles de désolvation, d'hydrophobicité, électrostatique,… MAIS c'est toujours difficile d'identifier les faux-positifs et les erreurs.

Comment peut-on améliorer ces fonctions de score ?

1. Ab initio rigid body docking

On considère les sous-unités comme des corps simplifiés et rigides qu'on insère dans une grille en 3D. Cette technique se base sur la complémentarité de forme.

On fait la distinction entre le cœur de la protéine, la surface et la zone d'interactions.

Lorsqu'on procède au docking, on regarde le recouvrement qu'il y a entre les 2 sous-unités et on calcule le degré de recouvrement pour plusieurs orientations (en tenant comptes des distinctions).

!!! Cela prends du beaucoup temps donc simplification

Cela dit, si lors de la liaison les protéines changent de forme, l'approche en corps rigide peut être complètement erronée. Et c'est en effet la difficulté principale.

2. Soft docking Techniques

Dans ce cas-ci, on tient compte de la flexibilité.

Comme dit précédemment, avant, on faisait l'hypothèse que la protéine était un corps rigide par soucis de temps de calcul. A présent, on commence à tenir compte de cette flexibilité au moins pour la chaine principal (back bone de la protéine).

Certains algorithmes arrivent déjà à tenir compte de cette flexibilité.

3. Prédire les interactions protéiques de surface

Les approches ab initio prennent beaucoup de temps, car elles testent énormément de possibilités.

Prédire la localisation de l'interaction prot-prot reste compliqué mais des études ont été faites.

Ils ont effectué des mutations sur certains résidus d'alanine, et ont observé que certains de ces résidus contribuaient à la majorité de l'énergie de liaison du complexe.

- ⇨ Les « hot spot/O-ring » : les résidus du hot spot sont entourés par un anneau d'autre résidus qui contribuent de manière moins importante à la liaison mais empêchent le solvant de se lier au hot spot.

La désolvation de site de liaison est une condition nécessaire à la liaison !

Les régions de hot spot contiennent souvent de l'alanine. Les régions autour sont denses et semblent conservées.

Même si on peut mettre en évidence certaines zones qui pourraient potentiellement être des zones d'interaction, il est difficile d'émettre des règles strictes. Ils utilisent des techniques de machine learning qui se base sur les zones enfuies, l'énergie d'interaction électrostatique et de désolvation, les scores d'hydrophobicité et les scores de conservation de résidu pour tenter d'en sortir des théories.

Des études sont quand même arrivées à prouver que la fente de désolvation pouvait prédire fortement la zone d'interaction.

4. Base de données interaction structural prot-prot

Les bases de données deviennent importantes dans la prédiction de structure de complexes. Ces bases de données permettent de tester des algorithmes.

Malgré tout, il n'y a pas énormément d'interaction répertoriée.

5. Retirer des informations des interactions

Le but est d'arriver à retirer des informations des interactions.

Exemple : - Des chercheurs ont établit le potentiel de force moyenne qui est basé sur l'hydrophobicité et la capacité de se lier à un atome d'Hydrogène.

- Ils ont pu mettre en évidence que les forces électrostatiques n'étaient pas si importantes.

6. (PCA of knowledge …)

(Ce chapitre parle de FOURIER, nous n'avons pas abordé ça en cours et c'est assez compliqué.)

7. Data-driven docking

Parfois, la structure 3D d'un complexe n'est pas disponible mais il parfois intéressant d'avoir la position d'un site fonctionnel.
On peut utiliser par exemple, Evolution trace (ET). On fait l'hypothèse que le résidu/site fonctionnel est conservé durant l'évolution. On peut dès lors trouver des zones intéressantes.

Ils ont aussi pu rajouter des informations biologiques dans leurs algorithmes de docking. Exemple : des résidus bloquant qui sont des résidus que se désolvent pas.

8. Rescoring Docking decoys (compliqué)

Des études ont montré que de faible résolution de fonction de score peuvent parfois indiquer la localisation d'un site de liaison.

9. Modéliser la flexibilité de la chaine latérale

Lorsqu'il y a formation d'un complexe, il y a un réarrangement structural. La chaîne transversale peut changer d'angle. (Les aa flexibles sont Lys et Arg).

Des chercheurs ont avance que d'un point de vue dynamique, le temps de collision est trop court pour qu'un réarrangement total ait lieu. Les résidus spécifiques agissent comme des motifs prêts à l'emploi, les résidus sont déjà dans une conformation favorisée.

D'autre proposent que le docking se fait en 2 étapes, d'abord la protéines se lie via un « anchor residue », peut être un hotspot, qui représente un proportion importante de

l'énergie libre de liaison et ensuite la périphérie agit en comme loquet et les résidus s'ajustent.

10. Modéliser la flexibilité de la chaine principale

On a pu observer que lorsque un complexe se forme une partie des sous-unités peut modifier légèrement sa chaîne principale.

Pour bien faire, il faut tenir compte des 2 types de flexibilités.

11. Modélisation de l'interface avec l'eau

La solvation et la désolvation sont thermodynamiquement importants. Beaucoup d'algorithmes ne tiennent pas comptent du solvant pourtant, parfois le solvant intervient dans la formation de l'interface. Cette approche n'a cependant pas encore été incorporée dans les algorithmes actuels.

12. Conclusion et résumé

La formation de complexe se fait en 2 étapes : collision initiale avec reconnaissance après désolvation et enfuissement dans la fente du hot spot et ensuite la latching phase (loquet).

Lorsqu'on prend en compte la dynamique moléculaire et la flexibilité on se rapproche de plus en plus de la réelle manière dont le complexe est formé.

En utilisant les nouvelles découvertes en biologie, les informations physico-chimique des interactions, on pourra améliorer les prédictions structurales, mieux comprendre les interactions au sein des cellules.

Art 2 : Sampling & scoring : a marriage made in heaven – 2013

Abstract :

Ils mettent en évidence que pour le protéine-docking, il est parfois préférable plus intéressant d'utiliser directement une fonction de score plus précise sans procéder à un échantillonnage préliminaire. C'est une technique plus performante mais qui prend aussi plus de temps.

Introduction :

En Général

Les algorithmes de docking consistent souvent en un échantillonnage de l'espace conformationnel – recherche de solution partielle grossière – et ensuite ils utilisent une fonction de score plus précise pour affiner la recherche.

Dans cet article, ils se concentrent sur le docking mais pour eux leur technique pourrait très bien être utilisé pour la prédiction de structure de protéine.

Un des challenges dans le docking est d'arriver à prédire la structure d'un complexe à partir de sous-unités en n'ayant aucune information sur la structure du complexe.

Lorsqu'on a deux sous-unités, l'échantillonnage des conformations doit explorer un très large espace conformationnel en utilisant une fonction d'E relativement simple pour que ce soit possible pour l'ordinateur (au niveau du temps).

L'algorithme va évaluer comment les 2 partenaires vont interagir sans qu'il y ait trop de chevauchement et qu'il y ait si possible des propriétés avantageuse comme par exemple la complémentarité chimique.

Attention, l'échantillonnage requiert aussi une fonction d'énergie mais ce qui diffère du scoring c'est que le but de l'échantillonage est : de générer un ensemble de structure qui comprend un nombre élevé de conformation « near-native ». L'échantillonnage n'est cependant pas très précis, il y a pas mal d'erreur et de faux positifs (structure loin de la conformation near-native mais dont les caractéristiques physico-chimique sont bonnes).
Une fois qu'on a généré notre échantillon, il y a un ensemble d'erreur (decoy set) qui peuvent être stockée pour tester la fonction de score et affiner la recherche.
Le but ultime étant d'identifier la structure avec le plus petit RMSD (root mean square deviation).

ICI

Ils vont intégrer les 2 étapes en une et tenter de montrer que les résultats peuvent être meilleur.

Les erreurs dans le docking prot-prot

Les decoys sont souvent utilisé dans le développement d'algorithme de prédiction notamment pour les tester et faciliter le développement de fonction de score.

On peut aussi simuler des decoys, en ajoutant des perturbations au docking.

Les ensembles de decoys doivent avoir un nombre de conformation near-native ou si possible le complexe natif.

Les résultats de CAPRI ont montré que le découpage des 2 étapes est moins performant.

Sélectionner une fonction de score nécessite d'avoir des informations sur la stratégie d'échantillonnage.

Il faut savoir comment l'ensemble de decoy a été construit.

Les structures liés et déliés sont souvent différentes si elles forment un complexe ou non. En sachant que le terme de Vander Waals est très sensible aux changements de coordonnées atomiques. Même une conformation très proche de l'état « near-native » peut avoir beaucoup d'énergie. En sachant que l'échantillonnage évalue l'énergie de 10^9 conformations et n'en garde que 1000. Il est possible de perdre des structures étant proches de la conformation recherchée.

Le decoy set dépend de la méthode utilisée pour échantillonner et affecte la fonction de score et la stratégie d'affinement.

Il vaut mieux échantillonner avec un score plus précis que d'échantillonner et puis scorer.

En utilisant une fonction de score plus précise pour l'échantillonnage, on augmente le temps de calcule.

Ils ont testé les 2 manières. Il s'attendait à avoir le même résultat mais en intégrant directement la fonction d'E plus précise dans l'étape d'échantillonnage, ils ont vu que leur manière était meilleure. (Mais il n'y a pas de résultat dans leur papier :s)

Exemple d'intégration de scoring par sampling.

L'idée générale est d'utiliser une fonction de score pour polariser une échantillon et ensuite ranker les conformations en fonction de la fonction de score.

1) Scoring en se basant sur la taille du cluster (cluspro)

Dans cette technique, nous nous focalisons l'étape d'échantillonage. Cluspro modélise les 1000 solutions et fait des clusters en fonction de leurs ressemblances structurales. Il renvoie non pas une solution mais un cluster de solution, le cluster le plus peuplé. Ne choisit pas nécessairement le cluster où il y a une conformation avec une fonction de score la plus petite.

On peut montrer que les clusters larges contiennent plus de structures natives.

L'échantillon peut-être considée comme un ensemble avec une fonction de partition Z (grandeur fondamentale qui englobe les propriétés statistiques d'un système à l'équilibre thermodynamique)

$$Z=\Sigma_j \exp(-E_j/RT),$$

where E_j is the energy of the jth pose(structure dans un cluster), and we sum over all poses.

Pour le cluster k, $Z_k = \Sigma_j \exp(E_j/RT)$

En partant de ça, la probabilité de kth cluster est donnée par, $P_k = Z_k/Z$

Au sein d'un cluster, on ne peut pas faire la distinction entre les structures qui sont trop similaires. Ej=E

$$P_k= (\exp(-E/RT) * N_k)/ Z$$

Pk est proportionnel à Nk, nk est le nombre de structure dans kth cluster.

Ils ont montré que les plus gros clusters contenait la majorité des structures near-native.

<u>Conclusion</u>

-Séparer la partie échantillonnage et scorage peut faciliter la méthode mais les intégrer améliore les résultats.

-toutes les fonctions de score basées sur un decoy-set sont affectées par le decoy set… et donc peut entraîner des erreurs…

-on peut obtenir créer des decoys set en ajoutant des perturbations au complexe near-native.

-Il vaut mieux avoir une fonction de score qui tienne compte de plusieurs contributions énergétique et l'intégrer directement à l'étape de sampling.

<u>Résumé du protocole des TP's.</u>

La structure expérimentale de blot 5 et derp 5 est disponible.

Blot 5 : possède une structure monomérique composée de 3 hélices alpha.
Derp 5 : possède une structure dimérique ayant une grande cavité hydrophobe. La cavité pourrait être lié des ligands hydrophobes et être à l'origine de la réaction

allergène. Chaque monomère de derp 5 possède 3 hélices alpha avec l'hélice N-terminal légèrement coudé contrairement à Blot 5.

Ce qu'on veut faire : étudier la possibilité des allergènes der f 5 et Blo t 5 à former un dimère similaire der p 5 avec la grande cavité hydrophobe.

- On a réalisé un alignement de séquence et on a regardé quels résidus jouaient un rôle important dans la dimérisation.

Derp5 : on a constaté qu'il y avait des interactions hydrophobiques, il y a aussi de valine zipper (similaire au leucine zipper)
Blot 5 : il y a 4 répétitions heptad (résidu hydrophobes).

Il y a environ 40 % d'identité de séquence entre les prtotéines.

L'alignement de séquence a été réalisé avec mobyle pasteur.

- On a généré un modèle structural pour derf 5 en utilisant blot 5 ou la chaîne A de derf 5 à l'aide de Modeller.

On a regardé les scores pour la modélisation de derf 5 si on prenait soit un modèle avec ou sans hélice coudés.
Les scores étant similaires, on a pris arbitrairement le modèle à hélice coudée.

On a ensuite réalisé le docking. On a été voir sur CAPRI pour voir lequel des serveurs étaient le meilleur => Cluspro

Le cluster le plus peuplé ne conduit pas nécessairement aux structures les plus correctes, ici nous avons pris la structure dont l'énergie la plus faible.

A la fin, nous aurions pu essayer de muter la valine pour voir si le dimère était toujours aussi stable (mutations qui aurait été réalisées avec pymol).