Daniel DE MURAT - Madeleine DE SOUSA-VIOLANTE - Mei-Shiue KUO - Sonia TIEO
Equipe 2 - Amont- Université Paris Diderot (M2BI)

- Rapport Equipe 2 -

Introduction

Les protéines sont les composants de base de tout être vivant et remplissent différentes fonctions essentielles. Elles se caractérisent par leur séquence en acides aminés, formant une chaîne qui se replie pour former une structure tridimensionnelle. Connaître la structure d'une protéine est capital pour comprendre sa fonction.

Actuellement, les structures protéiques sont principalement résolues par cristallographie aux rayons X qui peut être complétée par la résonance magnétique nucléaire (RMN). La microscopie électronique est peut aussi être utilisé. Malgré leur efficacité, ces méthodes s'avèrent coûteuses et ne permettent pas de résoudre toutes les structures protéiques. En effet, sur les 100 millions de séquences protéiques connues, seulement 140 000 ont leurs structures résolues [1] .

Pour pallier ces inconvénients, des méthodes de prédiction de structures protéiques ont vu le jour. Ces techniques se basent sur les travaux de Christian Anfisen [2] qui part du principe que la structure d'une protéine peut être déterminée uniquement par sa séquence en acides aminés.

Il existe trois grandes catégories de méthodes de prédiction de structures protéiques: les méthodes *ab initio*, les méthodes de modélisation comparative et les méthodes *de novo*.

Tout d'abord, les méthodes *ab initio* sont des méthodes qui n'utilisent que la seule information de la séquence. Elles sont basées sur des principes physiques et utilisent des procédures qui miment le repliement des protéines ou des méthodes stochastiques de recherche de solutions possibles. Cependant, ce sont des procédures généralement coûteuses en temps de calcul et valables uniquement sur des petites protéines.

Les méthodes *de novo* utilisent des petits fragments pour construire un modèle de structure de protéine qui sont assemblés par exploration Monte Carlo [3]. Ces méthodes restent coûteuses et sont dans la majorité des cas pas assez précises, plus particulièrement pour les protéines d'une taille supérieure à un centaine d'acides aminés. Par ailleurs, elles prennent beaucoup de temps.

Enfin, les méthodes de modélisation comparatives se basent sur l'utilisation de structures déjà résolues expérimentalement comme point de départ pour la résolution d'une structure protéique inconnue. En effet, on sait que la structure d'une protéine est mieux conservée au cours de l'évolution que sa séquence [4]. Il est donc possible de se servir de la structure déjà résolue d'une protéine apparentée comme modèle, pour prédire la structure de la protéine d'intérêt. Ces méthodes sont performantes et ont permis la prédiction de la structure de 50% des protéines dont la structure a été résolue actuellement. Cependant, pour 50% des séquences protéiques, les méthodes classiques ne sont pas assez sensibles pour détecter les protéines homologues [5] , et la prédiction structurale se trouve donc compromise.

Afin de remédier à ce dernier problème, un concours inter-université a été mis en place : Meet-U. Cette édition se présente en deux étapes : (i) une annotation de domaine basée sur une

comparaison profil-profil (partie amont) et (ii) une identification du “fold” 3D le plus stable par “threading” (partie aval). Chaque équipe est en charge d'une des deux étapes.

Ici nous présentons les travaux de l’équipe 2, une équipe amont de l’université Paris Diderot qui travaille sur l’annotation de domaine basée sur une comparaison profil-profil. La stratégie ici, détaillée plus loin, consiste à construire un profil à partir d’une séquence en assemblant des informations séquentielles et de prédiction de structure secondaires. Ensuite, ce profil est comparé à des profils contenus dans la base de données HOMSTRAD [6], composée de 405 familles sous-dossiers correspondant aux metafolds dont la structure a été résolue, par alignement semi-global. Un score est attribué à chaque alignement. Les scores obtenus sont ensuite classés et stockés dans un fichier .foldrec (voir **Figure 1**).

Afin d’identifier le “fold” 3D le plus stable, il a été décidé de travailler avec le groupe 4 (Fold_U). Leur méthode a été récupérée et intégrée dans un pipeline qui permet de partir d’un fichier fasta, de créer un .foldrec et d’obtenir ensuite un .csv qui contient les résultats du threading.

Après évaluation de la performance des scores du jeu de donnée test de 21 séquences requêtes, nous avons fait tourné le pipeline entier sur les 11 séquences de nos collègues en Master Biologie-Santé de l’Université Paris Saclay.

Présentation de la stratégie

Une vue d’ensemble de la stratégie est représentée en **Figure 1** et chaque étape est détaillé.

Recherche de séquences homologues: Psi-Blast

La première étape de notre stratégie consistait à trouver des séquences homologues à la query. Pour cela, la version PSI-BLAST [7] 2.2.31 du package legacy blast a été utilisée contre la base de données UniRef90 [8] (plus coûteuse en mémoire que UniRef50 mais nécessaire pour la suite), avec un nombre maximal d’itérations de 3, un coût d’extension de gaps de 1 et en utilisant la matrice BLOSUM62 [9].

Concernant les bases de données UniRef (UniProt Reference Clusters), elles fournissent des ensembles de séquences en cluster à partir de connaissances UniProt et de certains enregistrements UniParc. UniRef90 et UniRef50 sont construits en regroupant des séquences UniRef100 aux niveaux d’identité de séquence de 90% ou 50%.
Une fois cette étape effectuée, les séquences obtenues ont été extraites. Parallèlement, l’outil PSIPRED[10] a été utilisé sur la query, afin de prédire sa structure secondaire qui sera utilisée ultérieurement.

Alignement multiple: MAFFT

L’étape suivante est un alignement multiple global des séquences trouvées grâce à PSI-BLAST avec la séquence cible. Pour ce faire, le logiciel MAFFT [11] a été utilisé

principalement pour sa vitesse dans le processus d'alignement multiple. Il a la capacité d'aligner de très grandes séquences tout en conservant une très bonne qualité de l'alignement. Les séquences sont représentées par des vecteurs de résidus selon leur volume et leur polarité puis un arbre guide est généré en utilisant des transformées de Fourrier.

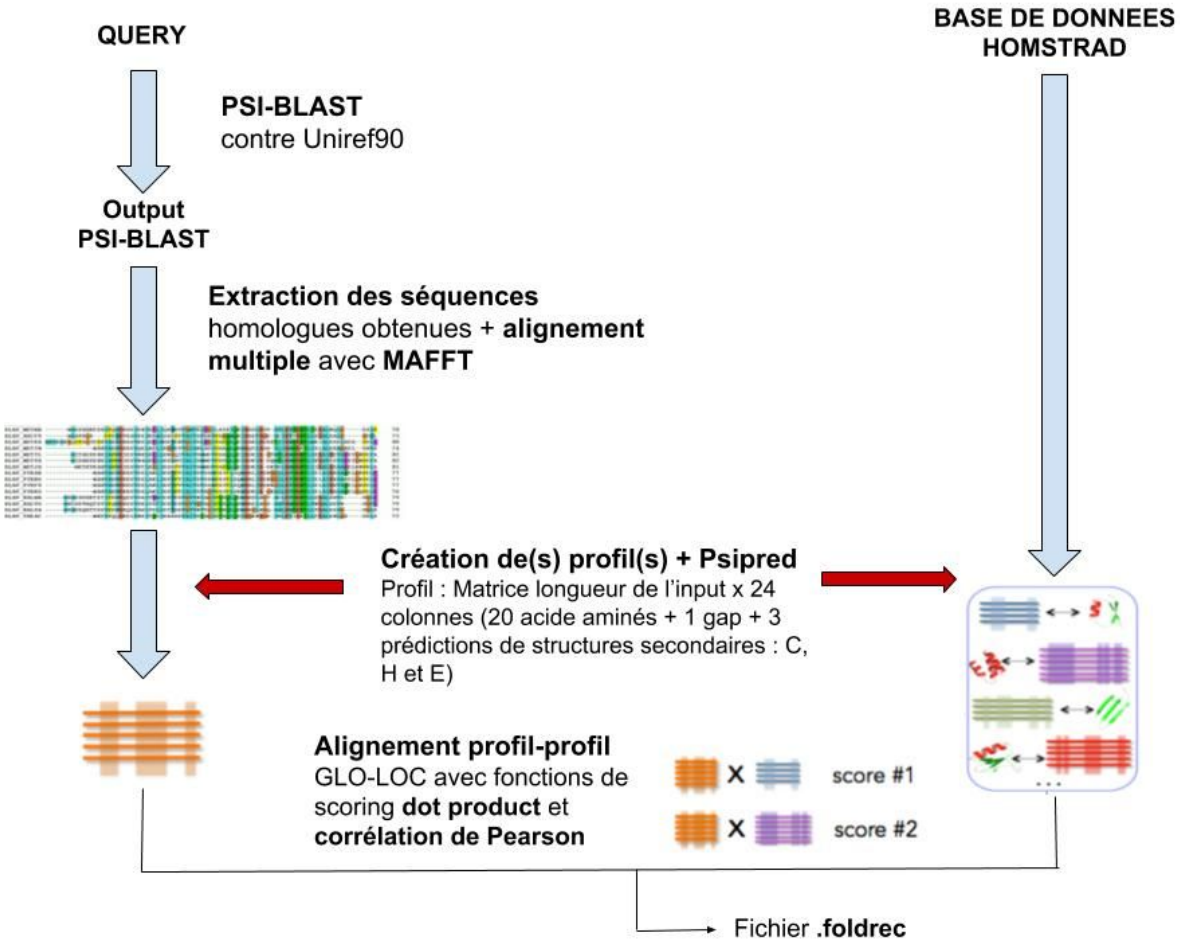

Figure 1 : Schéma explicatif de la stratégie utilisée

Construction du profil de la query et des templates

L'alignement multiple issu de MAFFT et la structure secondaire produite par PSIPRED sont ensuite utilisés pour construire un profil. Ce profil a été construit pour la query ainsi que pour toute la base de données HOMSTRAD grâce aux alignements .map fournis, qui sont des alignements multiples des représentants de chaque superfamille.

Ce profil comprend une matrice PSSM (position-specific scoring matrix) auquel est concaténé la structure secondaire produite par PSIPRED. Ce profil prend la forme d'une matrice dont le nombre de lignes est égal au nombre de positions retenues de l'alignement multiple et avec 24 colonnes. En effet, seulement les positions correspondant aux résidus de la séquence cible ou de la séquence contenue dans la base de donnée, et pour lesquelles le nombre de brèches ne dépassent pas la moitié des séquences totales de l'alignement, sont retenues pour la matrice PSSM. Les 24

colonnes correspondent à : 20 colonnes pour les acides aminés (ordre: ARNDCQEGHILKMFPSTWYV) , 1 colonne pour la fréquence des brèches (représenté par '-') et 3 colonnes pour les structures secondaires (ordres: C = coli, H = helix and E = strand). Un système de poids a été ajouté afin de pénaliser les séquences qui sont surreprésentées et favoriser les séquences rares. Nous avons également inséré un pseudo-count pour chaque acide aminé (qui égale à 1/20) pour ne pas avoir de fréquence à 0 dans la matrice.

Alignement profil-profil

Une fois le profil créé, il est comparé avec tous les autres profils contenus dans la base de donnée HOMSTRAD. Pour cela, les profils sont alignés selon un alignement semi-global, c'est-à-dire que les brèches ne sont pas pénalisés en début et en fin de séquence. De plus, les extensions de brèches sont moins pénalisées que les ouvertures de brèches. Cet alignement est compatible avec ce que l'on cherche à faire, c'est-à-dire aligner la séquence cible sur la séquence de la structure support mais en considérant qu'il puisse y avoir des délétions. Un score est ainsi attribué à chaque alignement engendré, classé puis stocké dans un fichier .foldrec. Pour le choix du score nous avons opté pour le Dot Product [12] et le coefficient de corrélation de Pearson [12].

Partie aval : groupe 4

Enfin, pour chaque alignement, la séquence cible est "enfilée" sur la structure résolue des séquences contenus dans la base de donnée, générant un score de threading utilisant la matrice DOPE. Modeller est ensuite utilisé pour générer un modèle 3D par homologie (grâce à l'alignement entre la query et le template), retournant un score de potentiel statistique DOPE. Le modèle obtenu permet ensuite de générer d'autres scores tels que ceux de co-évolution, d'accessibilité au solvant et de structures secondaires avec DSSP. Après normalisation, ces scores sont classés et stockés dans un fichier score.csv et le top N des structures pdb est généré.

Résultats:

Analyse quantitative : partie amont

Dans un premier temps, le programme de notre partie a été exécuté sur les séquences requêtes fournies pour le benchmark (21) et nous avons voulu comparer deux critères différents:

- La méthode de scoring pour l'alignement profil-profil: Dot Product contre le coefficient de corrélation de Pearson (avec la base de donnée UniRef50)
- Avec la base de donnée UniRef50 contre la base de donnée UniRef90 (avec un scoring Dot Product)

Afin de quantifier l'efficacité des 3 méthodes, un score d'enrichissement est calculé et représenté en **Figure 2**. En pratique, les rangs sont parcourus (de i=1 jusqu'à 405, ie: le nombre de séquence support de HOMSTRAD). Pour chaque rang i, on compte le nombre de fois qu'une famille correspondant à celle de la séquence requête est retrouvée à ce rang ou à un rang antérieur. D'après les résultats, à partir de la base de données UniRef50, 50% des séquences requêtes retrouvent leur famille respective dans les 100 premiers rangs. En revanche avec un coefficient de

corrélation de Pearson, il faut attendre le rang 200 pour que 50% des séquences retrouvent leur famille. Nous avons donc choisi de continuer avec le Dot Product pour la fonction de scoring.

Concernant le choix de la base de données, nous avons comparé UniRef90 et UniRef50. UniRef90 est certe plus plus précise mais elle requiert plus d'espace de stockage. D'après la **Figure 2,** il n'y a pas de différence pour le score d'enrichissement entre les deux bases de données. Cependant, pour la partie aval, il était nécessaire d'obtenir un nombre seuil conséquent de séquence aligné sur la séquence cible. Comme UniRef50 ne permettait pas d'obtenir assez de séquence alignés, UniRef90 est donc choisi pour permettre à notre méthode de fonctionner sur la partie aval.

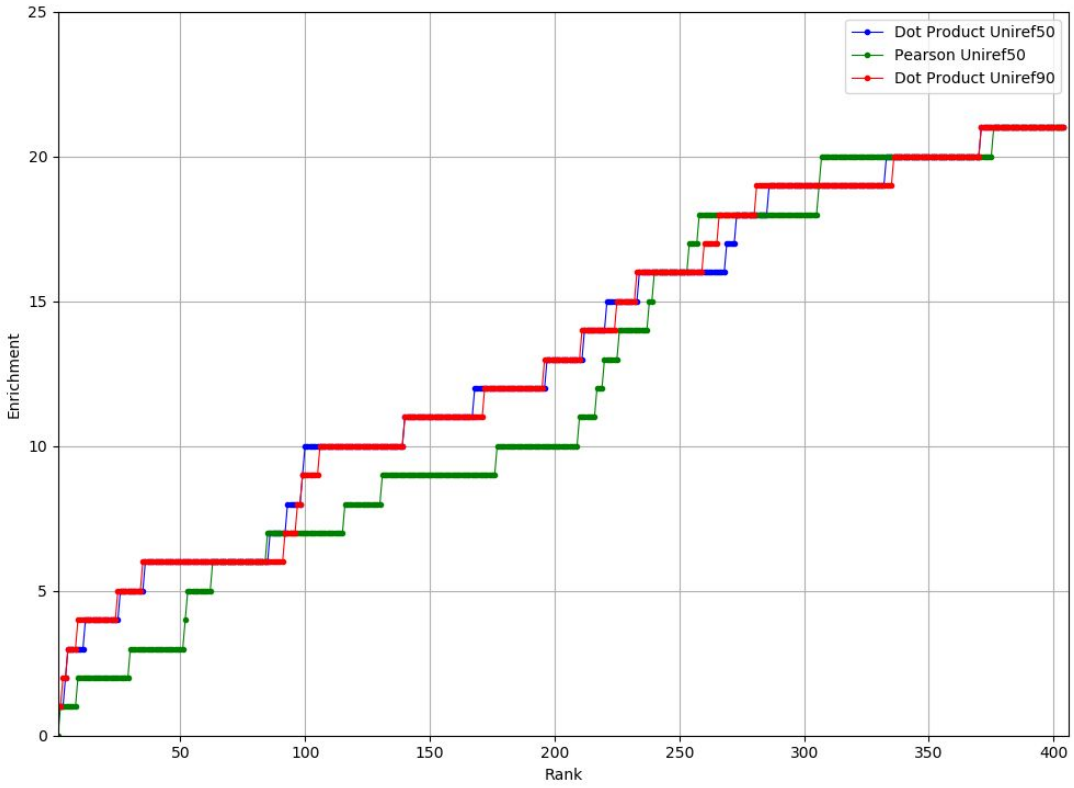

Figure 2 : Courbe d'enrichissement sur les 21 séquences requêtes en fonction du rangs. Chaque courbe représente une de la partie amont (Dot Product avec UniRef50, Coefficient de corrélation de Pearson avec UniRef50 et Dotproduct avec UniRef90).

Pour en revenir aux scores prédits par notre méthode (avec comme base UniRef90 et comme fonction de scoring le Dot Product), le **Tableau 1** montre des précisions de prédiction de scores très variées. Nous avons eu de très bonnes prédictions pour 2 séquences requêtes, ETF_alpha et TBCA (leur familles se trouvent respectivement dans le 1er et 3ème classement). Par contre, les prédictions sont médiocres pour certaines séquences requêtes, par exemple DnaJ. Notre score indiqué sur son foldrec prédit que sa famille se trouve à la 371ème place. Pour mieux comprendre nos résultats, nous avons examiné l'identité entre la séquence de la requête et la séquence du représentant de sa famille HOMSTRAD. Nous avons remarqué que notre score peut permettre de bonnes prédictions lorsque l'alignement de deux séquences est long et s'ils partagent au moins 10% d'identité. De ce fait, nous avons vu que l'identité relative à la couverture de la longueur d'alignement d'une requête

joue un rôle déterminant pour la prédiction lorsque la longueur d'alignement est supérieur à 400. Par exemple, Lipoprotein_4 a seulement 11% d'identité avec sa représentant mais son identité relative est de 19.89% en raison de sa faible couverture (55.4% =277 acides aminés sur 500). Notre score a réussi à identifier sa famille dans le 9ème classement. Pour les 2 autres requêtes ayant des alignements longs, LRR et FAD-oxidase_NC, nous avons eu de prédiction moyennes car leurs identités relatives restent faibles, 8.6 % et 7 %. Par contre, cette identité relative à la couverture de la longueur d'alignement ne peut pas appliquer aux séquences de petite ou moyenne longueur. En général, plus la séquence est petite, plus la structure a du mal à être prédite même si elle partage une identité de 15.8% avec le représentant de sa famille, comme UBQ par exemple (voir **Tableau 1** pour plus de détails).

Tableau 1: Résumé des résultats des Foldrec de notre partie.

Query	Fold ou Famille du query	Foldrec No. 1	Classement de Foldrec (de famille)	Longueur d'alignement (aa)*	Identité (%)*	Identité relative à sa couverture (%)**
UBQ	protg Fold	GP120	**260**	**77**	**15.58 %**	**16.43 %**
DEP	myb_DNA-binding Fold	NiFeSe_Hases	**336**	**105**	**8.57 %**	**9.57 %**
igvar-h	Desulfoferrodox Fold	GP120	**35**	**156**	**16.03 %**	**20.01 %**
His_biosynth	OMPdecase Superfamille	SAM_decarbox	97	NA	NA	NA
Lum_binding	EFTU_C Fold	AMP-binding	92	133	6.67 %	9.54 %
Rib_hydrolayse	TIR Fold	Glu_syn_central	233	263	14.45 %	15.20 %
DnaJ	ATP-synt_DE_N Fold	transcript_fac2	**371**	**118**	**10.17 %**	**16.00 %**
Cohesin	Fimbrial Fold	CBD_6	**25**	**202**	**14.36 %**	**20.28 %**
PAC	GAF Fold	tyrosinase	**281**	**164**	**14.02 %**	**22.11 %**
Agglutinin	intb Fold	Cu_nir	196	NA	NA	NA
hemery	SRP54 Fold	annexin	99	NA	NA	NA
LRR	Recep_L_domain Fold	Glu_syn_central	212	**457**	**8.57 %**	**8.59 %**
svmp	Astacin Superfamille	Isochorismatase	140	NA	NA	NA
Lipoprotein_4	mofe Superfamille	ATP-sulfurylase	**9**	**500**	**11 %**	**19.86 %**
SSB	TIMP Fold	GP120	106	NA	NA	NA
TBCA	BAG Fold	Ribosomal_S7	**3**	**116**	**15.52 %**	**17.65 %**
FAD-oxidase_NC	MurB_C Superfamille	lipoxygenase	225	**527**	**6.83 %**	**6.99 %**
ETF_alpha	**Arginosuc_synth Superfamille**	**Arginosuc_synth**	**1**	**430**	**14.65 %**	**20.52 %**
histone	Arch_histone Superfamille	actin	172	NA	NA	NA
GA	B Superfamille	Rop	174	NA	NA	NA
PCNA	DNA_PPF Famille	Arginosuc_synth	**266**	**252**	**12.70 %**	**12.85 %**

*Longueur d'alignement et l'identité de séquences sont obtenues par l'outil "Sequence Manipulation Suite" (http://www.bioinformatics.org/sms2/ident_sim.html). Les meilleures prédictions sont marquées en bleues et les pires prédictions sont marquées en rouges.

** L'identité relative à la couverture de la longueur d'alignement = L'identité / (longueur de requête / longueur d'alignement)

Analyse quantitative : partie amont et aval

Afin d'évaluer les différents scores proposés par le groupe 4, les 21 séquences fournies comme "benchmark" ont été prédites. La **Figure 3** montre les résultats obtenus. Chaque séquence a été prédite à un certain rang selon un score. Le score alignement représente le score obtenu avec notre méthode. Les autres scores sont les différents scores que propose la méthode de l'équipe 4. Plus la courbe d'enrichissement augmente rapidement, plus les séquences sont prédites correctement à des bonne positions.

Sur les 21 séquences à tester, 5 sont prédites dans le top 50 de notre méthode. Ensuite, les séquences prédites sont retrouvées progressivement les unes après les autres (pas de palier) jusqu'au 405ème résultats.

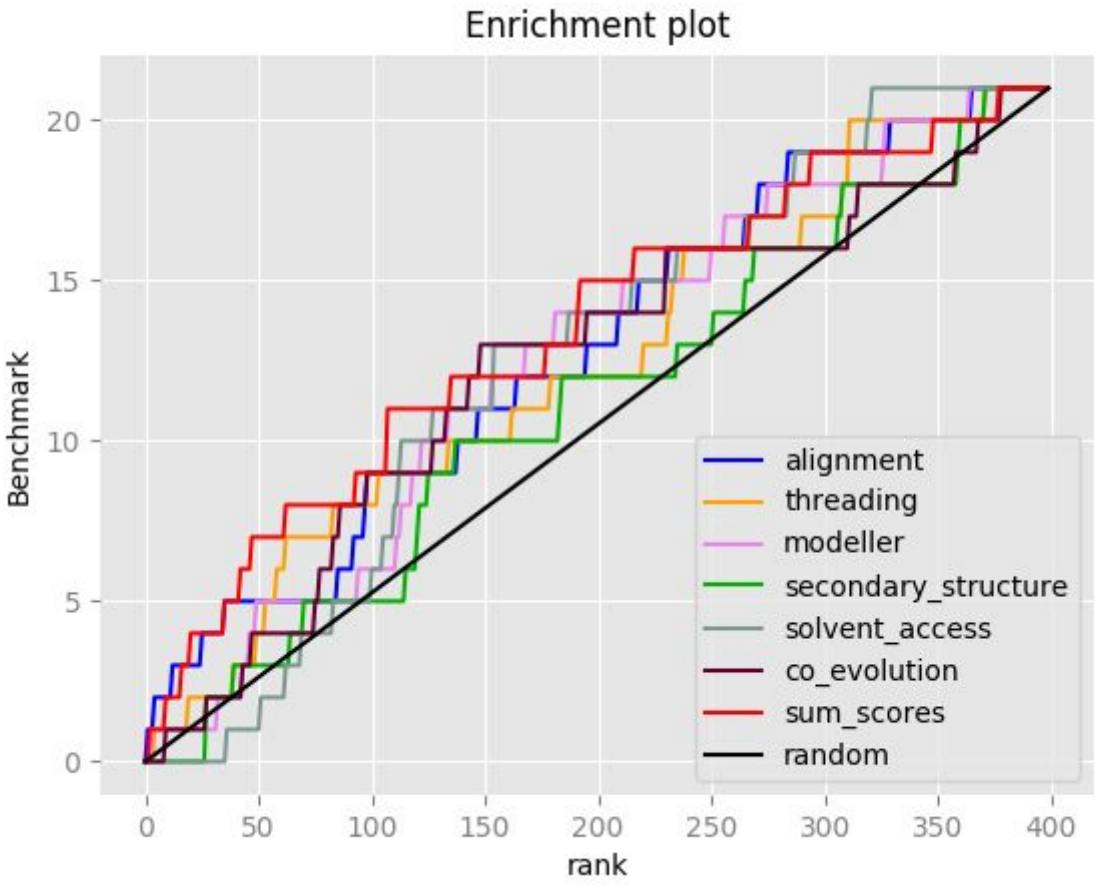

Figure 3 : Courbe d'enrichissement sur les 21 séquences requêtes en fonction du rangs. Chaque courbe représente une méthode de scoring différent de la partie aval (après fusion de la partie amont).

On remarque que c'est le sum-scores de l'équipe 4 qui permet d'obtenir un meilleur résultat à partir des fichiers d'alignements obtenus par notre méthode. En effet, c'est celui qui prédit

globalement le plus de séquences à tout rang donné, particulièrement à des rangs faibles (inférieurs à 100).

Ainsi, dans le pipeline finale qui regroupe les deux méthodes, c'est le sum-score qui a été choisi à la fin afin de prédire la structure qui pourra servir de structure support afin de modéliser la séquence cible.

Résultats d'annotation des 11 séquences:

Pour finir, le pipeline entier a été lancé sur les 11 séquences à annoter fournies par nos collègues de biologie santé, ce qui nous a permis de tester notre programme dans un cas concret. Nous avons remarqué que globalement les résultats étaient très différents d'une équipe à l'autre, exceptée pour deux séquences : trQ6HKV8 et G1G14-3311, où le groupe 5 a aussi obtenu la famille bacteriofer et COX-1 en rang 1.

De plus, parmi 9 familles prédites par notre programme, nous avons retrouvé deux fois COX1 et une fois COX3 qui sont les sous-unités du complexe Cytochrome c oxidase, le composant de la chaîne respiratoire mitochondriale qui catalyse la réduction de l'oxygène en eau. Le cytochrome b qui a été retrouvé deux fois est le composant du complexe ubiquinol-cytochrome c réductase et fait également partie de la chaîne respiratoire. Notre programme a réussi à prédire les familles qui partagent les même fonctions biologiques. Cependant, il est difficile de conclure la pertinence de nos résultats sans avoir plus de contexte biologique de leur expériences.

Discussion

Bien que notre scoring ait réussi à bien classer la famille de ETF_alpha au premier rang et la famille de TBCA à la 3ème place, notre stratégie (fusionnée avec l'équipe 4) n'a pas pu classer les familles correspondant aux 21 requêtes du benchmark dans les top10. Cela montre bien que la prédiction de la structure 3D d'une protéine reste difficile. Nous avons remarqué que la famille Syntaxin a été 7 fois dans les premières positions parmi 21 requêtes. Ceci peut être dû à un déséquilibre des échantillons présent dans les tests, ou un soucis au niveau de la méthode amont ou aval. Il est aussi intéressant de voir que sur les 11 séquences du master de bio-santé, on retrouve une fois la Syntaxin. Il faudrait investiguer avec plus de benchmark afin de comprendre pourquoi cette famille se retrouve souvent en tête de classement.

Améliorations possibles

<u>Prédiction de structure secondaire avec PSI-PRED:</u>

Pour la structure secondaire de la séquence cible, il aurait été plus judicieux de faire la prédiction de la structure secondaire par PSI-PRED sur l'alignement multiple produit par MAFFT que sur la séquence cible. Malheureusement, cela a été impossible car l'alignement multiple n'est pas compatible avec le format d'alignement multiple demandé par PSI-PRED. L'équipe PSI-PRED a été contacté afin d'avoir la possibilité de regarder leur scripts et de résoudre ce problème de compatibilité, mais sans succès (certains scripts sont binaires).

La structure secondaire de la séquence cible produite par PSI-PRED est donc basée sur une seule séquence protéique au lieu de l'alignement multiple de tous les séquences ayant l'homologie avec la query. Ceci réduit considérablement la précision de la prédiction de la structure secondaire de la séquence cible.

Prédiction de structure secondaire sur les séquences de HOSMTRAD avec DSSP:

Comme la structure des séquences supports de HOMSTRA est résolues, il a été pensé d'utiliser l'outil DSSP pour l'assignement de structure secondaire. Ceci a été confronté à deux difficultés.

Premièrement, DSSP assigne 8 types de structures secondaires (G, H, I, E, B, S, T et C) qui sont différents des 3 types de structures par PSI-PRED. Ceci rend la comparaison entre la séquence cible et la template compliquée. Nous pouvons résoudre ceci en regroupant certains types de structures dans les 3 groupes utilisés par PSI-PRED, par exemple regrouper G, H et I en groupe de l'hélice (H), regrouper B et E en groupe de strand (E).

L'autre problème réside dans le fait que DSSP n'assigne pas la structure secondaire à toutes les positions. Certaines positions n'ont donc pas de structures correspondante car les structures résolues ne correspondent pas parfaitement à la séquence du métafold ou bien les structures ont des fragments non résolues.

Il a donc été décidé d'utiliser la même méthode de celle pour la séquence cible afin de produire la structure secondaire des structures résolues contenus dans la base de donnée. Encore une fois, ceci réduit considérablement la précision de la prédiction de la structure secondaire.

Il faut également noter que certains groupes avals utilisent DSSP dans leur fonction de score, et ils comparent leur DSSP avec le psipred prédit sur la séquence cible. La lacune ici est donc rattrapé par le groupe aval.

Accessibilité au solvant:

Par manque de temps, certaines améliorations qui étaient prévus à la base n'ont pas pu être mises en place. Afin d'obtenir de meilleurs résultats, nous voulions mettre en place un score d'accessibilité au solvant qui aurait été ajouté aux PSSM de la séquence cible et des séquences contenus dans la base de donnée.

Fonction de score pour l'alignement profil-profil:

Une meilleure fonction de score adaptée à un alignement pseudo-global pourrait améliorer nos résultats. En effet, seulement les scores Dot Product et corrélation de Pearsons ont été implémenté, et il a été montré que la corrélation de Pearson n'est pas adapté à notre méthode.

De plus, il a été observé que l'on retrouve des membres de la même super-famille de la séquence à prédire à de meilleures positions, et certains membres sont très proches au niveau de la famille et structurellement de la solution. Un scoring avec ces éléments de famille et super-famille pourrait être mis en place sur la partie amont, en parallèle avec les résultats de la partie aval pour obtenir de meilleurs résultats.

Amélioration de la matrice PSSM:

Nous avons vu que la longueur d'alignement pourrait jouer un rôle important dans la prédiction en utilisant notre méthode de scoring. Dans l'étape de construction de la matrice, nous avons supprimé certaines positions qui contiennent des brèches avant de calculer les poids de chaque séquences. Nous nous demandons si c'est peut-être plus judicieux de garder tous ces positions de brèches pour déterminer les poids de chaque séquences, puisque l'existence de ces brèches a un certain sens biologique dans l'évaluation. C'est aussi intéressant de voir si la matrice PSSM avec un nombre maximum de brèches améliore notre résultat de prédiction. Par manque de temps, nous n'avons pas pu vérifier ces 2 points intéressants.

Bibliographie:

1. Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne; "The Protein Data Bank" ; Nucleic Acids Research, Volume 28, Issue 1, 1,Pages 235–242; January 2000.
2. Anfinsen, C. B; "Studies on the principles that govern the folding of protein chains"; 1972
3. Karger, David R.; Stein, Clifford; "A New Approach to the Minimum Cut Problem". J. ACM. 43 (4); July 1996,601–640. doi:10.1145/234533.234534
4. Illergård, K., Ardell, D. H., & Elofsson, A.; "Structure is three to ten times more conserved than sequence—a study of structural response in protein cores." ; Proteins: Structure, Function, and Bioinformatics; 77(3), 499-508; 2009
5. Schwede T. Protein Modelling; "What Happened to the "Protein Structure Gap?"; Structure (London, England : 1993); 2013
6. Mizuguchi K, Deane CM, Blundell TL, Overington JP ; "HOMSTRAD: a database of protein structure alignments for homologous families."; Protein Science : A Publication of the Protein Society;1998
7. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J ; "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402; 1997
8. "UniProt: the universal protein knowledgebase - The Uniprot Consortium"; Nucleic Acids Research ; 02 / 2018
9. Henikoff, S.; Henikoff, J.G; "Amino Acid Substitution Matrices from Protein Blocks". PNAS. 89 (22): 10915–10919. doi:10.1073/pnas.89.22.10915; 1992
10. Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT ; "Scalable web services for the PSIPRED Protein Analysis Workbench" ; Nucleic Acids Research . 41 (W1): W340-W348; 2013
11. Katoh K, Misawa K, Kuma K-I, Miyata T . " MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform."Nucleic Acids Res; 30: 3059–3066; 2002
12. Wang, Guoli and Roland L Dunbrack. "Scoring profile-to-profile sequence alignments" Protein science : a publication of the Protein Society vol. 13,6 (2004): 1612-26.