Identification de complexes protéine-protéine par combinaison de classifieurs. Application à *Escherichia Coli*

Thomas Bourquard[*,**], Damien M. de Vienne[***,****], Jérôme Azé[♯]

[*] BIOS group, INRA, UMR 85, Unité Physiologie de la Reproduction et des Comportements, Nouzilly, France
[**]CNRS, UMR 6175, Nouzilly, France ; Univ. François Rabelais, Tours, France
Thomas.Bourquard@tours.inra.fr
[***]Centre for Genomic Regulation, C/Dr. Aiguader 88, 08003 Barcelona, Spain
[****]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
Damien.de-Vienne@crg.es
[♯]LRI, CNRS UMR 8623, AMIB INRIA Team, Univ. Paris-Sud, Orsay, France
Jerome.Aze@lri.fr, http://www.lri.fr/∼aze

Résumé. Nous proposons une approche permettant de prédire des complexes impliquant trois protéines (appelés trimères) à partir de combinaison de classifieurs appris sur des complexes n'impliquant que deux protéines (dimères). La prédiction de ces trimères repose sur deux hypothèses biologiques : (i) deux protéines orthologues présentent des caractéristiques fonctionnelles similaires; (ii) deux protéines interagissant sous la forme d'un complexe sous-tendent une fonction biologique essentielle à l'espèce concernée. Ces deux hypothèses sont exploitées pour décrire chaque paire de protéines par l'ensemble des espèces pour lesquelles elles possèdent un orthologue. Un ensemble de mesures de qualité classiquement utilisées pour évaluer l'intérêt des règles d'association est utilisé pour évaluer la force du lien entre les deux protéines. L'organisme modèle *Escherichia Coli* a été utilisé pour évaluer notre approche.

1 Introduction

L'étude des génomes nous apprend beaucoup sur le vivant. Chaque organisme possède un génome dont la composition est le résultat de l'histoire évolutive propre à cet organisme. Cette information biologique codée par l'ADN est divisée en unités discrètes, **les gènes**. Ces gènes codent pour **les protéines**, véritables "rouages" des réseaux biologiques au niveau cellulaire. Le projet de séquençage du génome humain ($3,4$ Gb, 1 Gb représentant un milliard de paires de bases ou acides nucléiques A, C, G et T), initié en 1990, a duré 13 ans pour un coût total de $2,7$ milliards de dollars. Les techniques récentes (2nd Generation Sequencing) ou futures (3rd Generations sequencing) permettent des vitesses de séquençage bien plus élevées, de l'ordre de 50 Gb/ jour, soit un génome comme celui de l'Homme entièrement séquencé en moins de 3 heures (HiSeq Systems, Illumina Inc.).

On comprend aisément le déluge de données brutes qu'il faudra savoir stocker et analyser. Nous nous intéressons à l'exploitation de ces données pour en extraire des connaissances,

notamment sur les protéines codées par ces gènes et leurs rôles dans la réalisation d'une action biologique.

Ces informations devraient permettre à terme, de concevoir de nouveaux médicaments spécifiques à un individu donné, individu pour lequel nous aurons identifié précisément les interactions entre les protéines de son génome. La plupart des interactions fonctionnelles entre protéines sont réalisées sous la forme de complexe protéique (assemblage macro-moléculaire de plusieurs protéines).

Nous proposons d'apprendre un modèle permettant de prédire un sous-ensemble de ces assemblages : les complexes protéine-protéine impliquant trois protéines ou **trimères**.

Cet article est organisé de la manière suivante : la section 2 présente le contexte de l'étude que nous avons réalisé ainsi que le protocole mis en œuvre pour apprendre notre modèle prédictif. Les données utilisées pour évaluer et valider notre approche seront également présentées dans la section 2. Les résultats obtenus sont analysés selon deux visions : l'une purement informatique (sections 2.1 et 2.2) et l'autre plus orientée biologie (section 3). Enfin, nous présentons plusieurs perspectives à ce travail en fin d'article.

2 Contexte biologique de l'étude

Le postulat essentiel de ce travail est que deux protéines qui interagissent dans une espèce donnée vont être soumises, de manière corrélée (ou conjointe), aux contraintes évolutives imposées à l'une ou à l'autre. En d'autres termes, les deux protéines vont co-évoluer, permettant ainsi à l'interaction d'être maintenue et à la fonction biologique d'être conservée dans le temps. L'exemple extrême de coévolution entre protéines en interaction est le gain ou la perte conjointe de deux protéines : si l'association entre deux protéines est nécessaire pour qu'une fonction biologique soit réalisée, la perte d'un des deux partenaires (resp. le gain) entraînera la perte (resp. le gain) de l'autre. La comparaison des profils de présence/absence de protéines (nommés profils évolutifs ou profils phylogénétiques, (Pellegrini et al., 1999)) chez différentes espèces permet donc de détecter la co-évolution entre protéines et donc de prédire si deux protéines sont susceptibles d'interagir. Ce type d'approche est au cœur de plusieurs démarches similaires (Pellegrini et al. (1999), Marcotte et al. (1999), de Vienne et Azé (2012)).

Afin de déterminer le profil évolutif d'une protéine d'intérêt, l'ensemble des protéines qui lui sont *similaires*, dans un ensemble d'espèces de référence, est calculé. Ces protéines similaires à une protéine donnée mais dans une autre espèce sont nommées des **orthologues**. Pour une protéine donnée, le profil évolutif que nous manipulons représente l'ensemble des espèces dans lesquelles notre protéine possède au moins un orthologue. Il peut être représenté sous la forme d'un vecteur de booléens comme montré sur la Figure 1.

Étant donné un ensemble de complexes protéiques avérés, nous réalisons l'extraction, pour chaque complexe, des interactions binaires entre protéines. Dans le cadre d'un apprentissage supervisé, ces interactions binaires entre protéines représentent l'ensemble des *exemples positifs*. À partir de l'ensemble des protéines présentes dans ces complexes, nous générons l'ensemble des paires de protéines n'appartenant pas à des complexes. Ces paires de protéines représentent alors les *exemples négatifs*.

Nous pouvons alors apprendre un modèle permettant de prédire l'interaction de deux protéines à partir de la connaissance de leur profil évolutif.

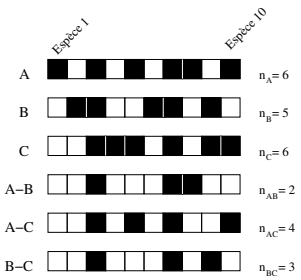

FIG. 1 – *Exemples de profils évolutifs pour l'ensemble des protéines* $\{A, B, C\}$ *et 10 espèces de référence. Dans chaque profil évolutif, une case noire indique que la protéine d'intérêt possède au moins un orthologue dans l'espèce considérée.*

Nous pouvons considérer les profils évolutifs associés aux différentes protéines comme des 1-motifs fréquents. L'étude des paires de protéines via leurs profils évolutifs, correspond alors à l'étude des 2-motifs fréquents que nous obtenons en calculant simplement l'intersection des profils associés à chaque protéine (voir Figure 1).

Rappelons que notre objectif est de pouvoir prédire si deux protéines vont interagir. Nous avons donc besoin de définir un ensemble de critères permettant d'associer, à partir des profils évolutifs des protéines, un score à chaque paire de protéines étudiée. Pour cela, nous proposons d'utiliser des métriques classiquement utilisées dans l'évaluation des règles d'association pour définir l'ensemble des critères numériques associés à une paire de protéines. Ces métriques, souvent appelées *mesures de qualité*, sont pour certaines d'entre elles non symétriques, i.e. elles n'évaluent pas de la même manière les règles $A \rightarrow B$ et $B \rightarrow A$. Dans notre cadre de travail, les interactions entre protéines ne sont pas orientées et nous ne devons donc pas prendre en considération cette information. Ainsi, pour une paire de protéine (A, B) donnée, nous calculons les mesures de qualité associées aux règles $A \rightarrow B$ et $B \rightarrow A$ dans le cas des mesures dites non symétriques.

Les critères retenus pour évaluer l'interaction entre deux protéines A et B sont :

- n_A, n_B, n_{AB} qui représentent respectivement le nombre d'espèces ayant au moins un orthologue pour la protéine A, B et pour les protéines A et B
- les mesures non symétriques : la confiance (et son symétrique : le rappel), la confiance centrée, la moindre contradiction, l'indice de Jaccard, Loevinger, TEC, LAP, GAN, Zhang, Pearl
- les mesures symétriques : Lift, Dice, Pearson, GiniIndex, IQC

Le Tableau 1 présente les expressions analytiques des mesures de qualité symétriques et non symétriques. Dans ce tableau, les notations suivantes sont utilisées : $P(X) = \frac{n_X}{N}$, $n_{\overline{X}} = N - n_X$ et $n_{X\overline{Y}} = n_X - n_{XY}$ avec X (resp Y) qui est égal à A, B ou AB. Où N représente le nombre d'espèces considérées.

Ainsi, une paire de protéines (A, B) est décrite par 28 valeurs réelles qui caractérisent le lien entre A et B. Afin de déterminer si ce lien est fonctionnel, i.e. se traduit par une interaction entre A et B, un modèle prédictif est appris.

Nous présenterons dans la suite les détails liés à la mise en œuvre de l'apprentissage su-

Mesure de qualité	Description / formule
Mesures symétriques	
Lift	$\frac{nn_{AB}}{n_A n_B}$
Dice	$\frac{2\times n_{AB}}{n_A+n_B}$
Pearson	$\frac{nn_{AB}-n_A n_B}{\sqrt{n_A n_B n_{\overline{A}} n_{\overline{B}}}}$
GI	$log\left(\frac{n_{AB}n}{n_A n_B}\right)$
IQC	$2 \times \frac{P(AB)-P(A)P(B)}{P(A)P(\overline{B})+P(\overline{A})P(B)}$
Mesures non symétriques	
Confiance	$\frac{n_{AB}}{n_A}$
Rappel	$\frac{n_{AB}}{n_B}$
Confiance centrée	$\frac{nn_{AB}-n_A n_B}{nn_A}$
Moindre Contradiction	$\frac{n_{AB}-n_{A\overline{B}}}{n_B}$
Indice de Jaccard	$\frac{n_{AB}}{n_{A\overline{B}}+n_B}$
Loevinger	$1-\frac{nn_{A\overline{B}}}{n_A n_{\overline{B}}}$
TEC	$\frac{n_{AB}-n_{A\overline{B}}}{n_{AB}}$
LAP	$\frac{n_{AB}+1}{n_A+2}$
GAN	$\frac{2*n_{AB}}{n_A}-1$
Zhang	$\frac{P(AB)-P(A)\times P(B)}{max\left(P(AB)\times P(\overline{B}),P(A\overline{B})\times P(B)\right)}$
Pearl	$P(AB) \times \left\lvert \frac{P(AB)}{P(A)-P(B)} \right\rvert$

TAB. 1 – *Mesures de qualité utilisées pour décrire les profils évolutifs. Nous renvoyons le lecteur à (Lallich et al., 2007) et (Lenca et al., 2003) pour une étude détaillée de chacune de ces mesures. Les métriques n_A, n_B et n_{AB} également utilisées pour décrire un profil évolutif ne sont pas rappelées dans ce tableau.*

pervisé de ce modèle. Nous avons retenu une approche de type "ensemble learning" où un ensemble de classifieurs binaires sont combinés pour construire *un méta-classifieur*.

Notons que le choix des descripteurs (Tableau 1) et leur utilité pour décrire les interactions protéine-protéine ont été décrits et évalués dans un article précédent (de Vienne et Azé (2012)). Nous ne discutons donc pas ces choix dans le présent article.

2.1 Combinaison de classifieurs

Des travaux antérieurs (Juan et al. (2008), de Vienne et Azé (2012)) ont montré que les hypothèses biologiques sur lesquelles reposent notre travail permettent d'apprendre des modèles qui s'avèrent efficaces en prédiction d'interaction protéine-protéine. Ces modèles permettent d'ordonner efficacement les paires de protéines par probabilité décroissante d'être en interaction.

Par contre, ces approches ne permettent pas de reconstruire nativement les complexes protéiques associés à ces paires de protéines. Cette reconstruction n'est immédiate que dans le cas de complexes n'impliquant que deux protéines. Pour un trimère ABC, le nombre d'interactions potentiels est égal à 3 : AB, BC, AC et les approches actuelles ne garantissent pas que la

totalité de ces interactions soient prédites, ni qu'elles aient des scores comparables permettant ainsi de les identifier facilement. Le travail présenté ici tente de répondre à cette question.

Pour traiter ce problème, nous avons utilisé comme données d'apprentissage l'ensemble des complexes binaires avérés de l'organisme modèle *Escherichia Coli* (noté *E. coli* dans la suite de l'article).

Cet organisme contient 4078 protéines réparties en de multiples monomères, 66 dimères, 45 trimères et quelques complexes impliquant plus de trois protéines.

Parmi l'ensemble des approches existantes pour calculer les orthologues, nous avons retenu celle proposée par Moreno-Hagelsieb et Latimer (2008). Ils ont proposé une combinaison de plusieurs méthodes afin de détecter avec une grande fiabilité les orthologues des protéines de *E. coli* chez $N = 1050$ espèces et ont rendu ces données publiques.

Nous proposons d'étendre les travaux de Juan et al. (2008) et de de Vienne et Azé (2012) en apportant les deux modifications suivantes : (i) l'ensemble des classifieurs utilisé est modifié et (ii) nous combinons un large ensemble de modèles appris à partir d'échantillons des données obtenus par tirage aléatoire sans remise.

(i) Les six classifieurs suivants ont été utilisés pour apprendre des méta-classifieurs : règles de décision (JRip, PART), arbres de décision (J48 et RandomForest), Bayésien Naïf et Régression Logistique. Ces classifieurs ont été appris avec la boîte à outils Weka (Hall et al., 2009).

(ii) L'échantillonnage des données est réalisé d'une part de manière à contrôler le nombre de positifs et de négatifs dans les données et d'autre part, pour palier au faible nombre d'exemples positifs disponibles.

De nombreux travaux ont montré l'intérêt des approches de type "Ensemble Learning" par rapport à l'utilisation de classifieurs "classiques" (voir Quinlan (1996), Opitz et Maclin (1999), Bauer et Kohavi (1999) pour quelques approches de références). Nous avons utilisé comme données d'apprentissage les 66 dimères d'*E. coli*, soit 132 protéines différentes. À partir de ces 132 protéines, nous construisons l'ensemble des 8580 paires de protéines représentant des interactions négatives. Cet ensemble est réduit à 8279 paires de protéines négatives après application de la contrainte $n_{AB} \geq 2$ à chaque profil évolutif. Cette contrainte minimale permet d'assurer une certaine cohérence avec notre hypothèse de travail reposant sur la coévolution (dans au moins deux espèces différentes) des protéines en interaction.

Nous procédons de la même manière pour créer l'ensemble de test constitué uniquement des 45 trimères. Ces 45 trimères sont constitués à partir de 135 protéines différentes. Notre ensemble de test est donc constitué de 135 exemples positifs (paires de protéines impliquées dans un trimère) et de 8910 négatifs (le filtre $n_{AB} \geq 2$ ne rejette aucune paire).

Les six classifieurs listés précédemment sont utilisés pour apprendre deux méta-classifieurs différents :

M_1 ce premier méta-classifieur est construit de la manière suivante : (i) l'ensemble des données d'apprentissage (les dimères) est utilisé pour apprendre chacun des six classifieurs, (ii) chacun des six classifieurs obtenu est appliqué sur le jeu de test (les trimères), et (iii) pour chaque exemple, le nombre de votes positifs est utilisé pour évaluer l'appartenance à la classe positive. Ainsi, une valeur variant de 0 à 6 sera associé à chaque exemple.

M_2^n ce second méta-classifieur ne diffère que sur le point (i) où un ensemble de n échantillons des données d'apprentissage est construit par tirage sans remise et les six classifieurs

sont appris sur cet échantillon. Chaque échantillon contient 50 exemples positifs et 1000 exemples négatifs.

Afin d'éviter tout biais dû à l'échantillonnage aléatoire dans la construction des échantillons utilisés par l'approche M_2^n, nous avons itéré le processus 100 fois et les votes ont été moyennés. Ainsi, une valeur variant de 0 à $6 \times n$ est associée à chaque exemple du jeu de test.

Ces deux méta-classifieurs sont ensuite utilisés pour deux tâches : (A) élaguer l'ensemble des paires potentiellement en interaction et (B) détecter les trimères.

L'élagage des paires potentiellement en interaction est simplement réalisée en utilisant la contrainte suivante : si $score(exemple) = 0$ alors élagage($exemple$), où $score(exemple)$ représente le nombre de votes positifs associés à l'$exemple$. Les performances des différents classifieurs pour la tâche (A) sont présentées dans le Tableau 2.

Le méta-classifieur M_1 se comporte comme le Bayésien Naïf, car le Bayésien Naïf est le classifieur qui prédit le plus de positifs et tous les positifs prédits par les différents classifieurs sont également prédits par le Bayésien Naïf.

Concernant le méta-classifieur M_2^n, nous ne présentons que les résultats pour $n = 50$ car nous avons constaté qu'à partir de $n = 50$ échantillons utilisés pour composer le méta-classifieur, les performances observées ne variaient plus (par manque de place, nous ne pouvons discuter plus précisément cet aspect de nos résultats). Les performances de M_2^{50} sont moins bonnes en termes de quantité d'exemples élagués mais meilleures en terme d'erreurs commises lors de l'élagage (exemples positifs élagués)

	VP	FP	FN	VN
Bayésien Naïf, M_1	99	542	36	8368
Régression Logistique	33	15	102	8895
PART	31	6	104	8904
Ripper	30	17	105	8893
C4.5	42	50	93	8860
Random Forest	38	29	97	8881
M_2^{50}	**116**	**1377**	**19**	**7533**

TAB. 2 – *Comparaison des performances des classifieurs avec et sans échantillonnage aléatoire des données pour l'apprentissage. Les évaluations sont réalisées sur les données relatives aux trimères. Les notations utilisées sont les suivantes : VP = Vrais Positifs, FP = Faux Positifs, FN = Faux Négatifs et VN = Vrais Négatifs.*

Pour la tâche (B), nous proposons l'algorithme $pTri$, présenté dans la section suivante, qui permet de reconstruire les trimères, à partir des paires non élaguées.

2.2 Extraction des trimères à partir de l'évaluation des paires de protéines par le méta-classifieur

Sachant que nous nous focalisons sur la recherche de trimères, nous proposons l'algorithme $pTri$ (voir Algorithme 1) pour identifier ces trimères. $pTri$ exploite le nombre de votes positifs associés aux paires de protéines pour identifier les paires les plus prometteuses. Nous considérons la liste des paires ordonnées par nombre de votes positifs décroissant.

$pTri$ parcourt cette liste par valeur décroissante du nombre de votes positifs. Les paires de protéines sont extraites de la liste une par une pour reconstituer des trimères. L'étape initiale consiste à extraire la première paire de la liste. Puis la liste est parcourue en appliquant la règle suivante : lorsqu'une nouvelle paire de protéines AB est prise en considération :

- soit il existe une protéine X telle que la paire AX (resp. BX) a déjà été rencontrée précédemment (avec un nombre de votes positifs plus élevé que AB) alors :
  - s'il n'existe pas de protéine Y tel que BY (resp. AY) ait été précédemment rencontré, alors le trimère AXB est prédit et toutes les paires impliquant l'une des trois protéines A, B ou X sont supprimées de la liste ordonnée des paires considérées.
  - soit il existe Y tel que BY (resp. AY) ait été rencontrée alors la paire AB est simplement supprimée. L'hypothèse sous-jacente est que les paires AX (resp. BX) et BY (resp. AY) appartiennent chacune à un autre trimère.
- sinon, la paire AB est ajoutée à l'ensemble des paires prédites. Cette paire représente alors la première paire d'un trimère dont aucune des protéines n'a déjà été identifiée.

Nous avons appliqué $pTri$ sur les données de test où les interactions entre protéines ont été évaluées d'une part en utilisant le méta-classifieur composé des 6 classifieurs appris sur l'ensemble des données relatives aux dimères et d'autre part, sur le méta-classifieur obtenu par combinaison des $6 \times n$ classifieurs appris sur des échantillons aléatoires des données relatives aux dimères.

Le Tableau 3 présente les résultats obtenus pour les approches $pTri(M_1)$ et $pTri(M_2^{50})$.

Les performances de $pTri$ peuvent être résumées selon les deux axes suivants.

2.2.1 Réduction du nombre de paires de protéines à analyser

La quantité de paires à analyser a été réduite de manière drastique pour les deux méthodes. Avant application de $pTri$, l'approche M_1 permet d'élaguer $92,9\%$ des paires de protéines candidates avec un taux d'erreur de $0,43\%$ et une perte de $26,7\%$ des paires représentant de vraies interactions. L'approche M_2^{50} permet elle d'élaguer seulement $83,5\%$ des paires avec un taux d'erreurs plus faible que l'approche M_1 ($0,25\%$). Par contre, les approches M_1 et M_2^{50} ne permettent pas seules de reconstruire les trimères.

$pTri$ se comporte comme un filtre appliqué sur les paires identifiées comme potentiellement en interaction par les approches M_1 et M_2^{50}. Les résultats obtenus montrent que $pTri$ permet de sélectionner efficacement les paires représentant de vraies interactions. Ainsi, après application de $pTri(M_1)$, $94,74\%$ des données initiales sont élaguées, avec $0,79\%$ d'erreur, et $47,4\%$ des interactions sont correctement identifiées. Alors qu'après l'application de $pTri(M_2^{50})$, $94,9\%$ des données initiales sont élaguées, avec $0,4\%$ d'erreur, et $73,33\%$ des interactions sont correctement identifiées.

L'approche M_2^{50} combinée avec l'algorithme $pTri$ permet donc d'augmenter le taux d'élagage de manière importante, en conservant un taux d'erreur relativement faible.

2.2.2 Augmentation du nombre de trimères correctement identifiés

Outre le nombre de paires de protéines correctement identifiées, l'approche $pTri(M_2^{50})$ permet également de prédire correctement un plus grand nombre de trimères que l'approche $pTri(M_1)$. En effet, en utilisant les 64 paires de protéines correctement identifiées par l'ap-

Entrées : $E = \{(p_A, p_B, n_{AB}^{pos})\}$ où n_{AB}^{pos} représente le nombre de votes 'positifs' associés à la paire de protéines $p_A - p_B$
Sorties : T : l'ensemble des couples $p_A - p_B$
début
  $T \leftarrow \emptyset$; $L \leftarrow \emptyset$;
  pour tous les $(p_A, p_B, n_{AB}^{pos}) \in E$ **faire**
    si *(*$n_{AB}^{pos} > 0$*)* **alors**
      /* Ajout du triplet dans la liste L maintenue triée selon le critère n_{AB}^{pos} décroissant */
      $L = L \cup \{(p_A, p_B, n_{AB}^{pos})\}$;
    fin Si
  fin
  tant que *(*$L \neq \emptyset$*)* **faire**
    $(p_A, p_B) = first(L)$;
    /* L'opérateur $first(L)$ renvoie le premier élément de L en le supprimant de la liste */
    si *((*$\exists p_X \; tq \; (p_A, p_X) \in T$*) et (*$\exists p_Y \; tq \; (p_B, p_Y) \in T$*))* **alors**
      p_A et p_B n'appartiennent pas au même trimère;
    sinon
      $p_C \leftarrow NULL$;
      si *(*$\exists p_X \; tq \; (p_A, p_X) \in T$*)* **alors**
        $p_C = p_X$, la protéine associée à p_A dans T
      sinon
        si *(*$\exists p_X \; tq \; (p_B, p_X) \in T$*)* **alors**
          $p_C = p_X$, la protéine associée à p_B dans T
        fin Si
      fin Si
      si *(*$p_C \neq NULL$*)* **alors**
        Supprimer toutes les paires (p_X, p_Y) de L telles que p_X ou p_Y est égale à p_A, p_B ou p_C;
      fin Si
      Ajouter la paire $(p_A - p_B)$ dans T;
    fin Si
  fin Tq
fin

Algorithme 1 : Algorithme $pTri$: $p$rédiction des trimères à partir des paires de protéines évaluées par le méta-classifieur.

	VP	FP	FN	VN
$pTri(M_1)$	64	37	71	8873
$pTri(M_2^{50})$	99	21	36	8889

TAB. 3 – *Comparaison des performances de $pTri$ sur les deux approches permettant d'évaluer la force de l'interaction entre deux protéines.*

proche $pTri(M_1)$ nous pouvons reconstruire sans erreur 16 des 45 trimères, alors qu'avec l'approche $pTri(M_2^{50})$, 30 des 45 trimères sont reconstruits [1].

3 Application aux données biologiques

Nous nous sommes focalisés ici sur la prédiction de complexes hétérotrimèriques impliquant des protéines de structures et fonctions différentes. Ces trimères sont impliqués dans plusieurs rôles clés de la cellule : régulation des concentrations en différents substrats assurant le maintien de la cellule dans un état quasi-stationnaire, dégradation et synthèse de l'ADN/ARN.

Le méta-classifieur M_2^{50} nous permet d'identifier correctement les trois protéines de 30 des 45 trimères. Ces 30 prédictions se décomposent en 27 trimères parfaitement prédits (l'intégralité des interactions) et 3 trimères partiellement prédits (2 interactions sur 3), ce qui s'avère parfaitement suffisant pour reconstruire le trimère original.

Pour 12 trimères, nous n'arrivons à identifier qu'une seule des trois interactions, enfin pour 3 trimères, aucune des interactions n'est identifiée.

Parmi les 19 paires de protéines incorrectement élaguées par le méta-classifieur M_2^{50} (voir Tableau 2), nous retrouvons 8 des 9 paires impliquées dans les 3 trimères non identifiés. Ces 8 paires ayant été élaguées avant l'application de l'algorithme $pTri$, il devient impossible d'identifier correctement les trimères concernés.

3.1 Prises en compte des spécificités structurales : cas des transporteurs membranaires type ABC

L'objectif initial de cette étude concerne la prédiction des interactions pour un trimère ABC, pour lequel le nombre d'interactions potentiels est égal à trois : AB, BC, AC. Néanmoins quels renseignements pouvons nous retenir des prédictions partielles de ces assemblages protéiques, i.e lorsque seule une voire deux interactions du trimère original sont correctement prédites ? le corollaire de cette question étant : quels enseignements pour la suite, nous fournissent ces trimères "atypiques" ?

Parmi les complexes pour lesquels la prédiction est incomplète (de 1 à 2 interactions sur les 3 potentielles), figurent les transporteurs ABC. Ces transporteurs ABC, situés au niveau des membranes de la cellule, permettent l'assimilation ou l'élimination au niveau cellulaire d'une large variété de métabolites et constituent une cible thérapeutique de choix comme les traitements antibactériens (Cangelosi et al., 1990).

Ce transporteur, qui est ATP-dépendant (i.e. nécessitant une source d'énergie ATP pour fonctionner), est représenté schématiquement Figure 2. Il est constitué d'une structure extra-membranaire, récepteur des métabolites provenant de l'environnement, d'une sous-structure membranaire ou perméase permettant le transfert des métabolites au niveau de la membrane hydrophobe sans altérer leurs structures et enfin d'une partie intra-membranaire correspondant au site de fixation des molécules d'ATP (Dawson et Locher, 2006).

Nous voyons ici que d'un point de vue structural, la partie extra-cellulaire (site de fixation des métabolites) et le site de fixation de l'ATP ne sont jamais connectés. Dès lors pour reprendre la notation utilisée précédemment pour un trimère ABC donné, (où A désigne le site

1. Nous considérons qu'un trimère est correctement reconstruit si les trois protéines impliquées sont correctement identifiées et qu'au moins deux relations entre protéines (sur les trois potentielles) sont identifiées.

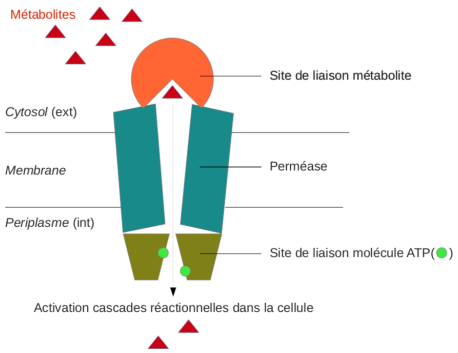

FIG. 2 – *Représentation schématique du transporteur ABC, et des sous-unités qui le constitue. En orange le site extracellulaire de fixation des métabolites, en bleu la perméase, en vert le site de fixation à l'ATP.*

extra-cellulaire, B la perméase et C le site de fixation de l'ATP), seules les interactions directes AB et BC s'avèrent pertinentes. L'information associée à la localisation de la protéine (intra ou extra cellulaire) peut être exploitée en post-traitement des prédictions effectuées par $pTri$.

Prenons l'exemple présenté sur la Figure 3 pour illustrer ce post-traitement. La Figure 3-(a) représente un sous-graphe du graphe des prédictions effectués par $pTri$. Nous pouvons y identifier une prédiction aberrante : $proX - modC$. Ces deux protéines ne peuvent pas interagir physiquement car $proX$ est une protéine intra-cellulaire et $modC$ est une protéine extra-cellulaire. Une information évidente pour un expert du domaine qui peut donc supprimer manuellement le lien prédit entre $proX$ et $modC$.

Si nous réappliquons $pTri$ uniquement sur les paires de protéines présentes dans ce sous-graphe (à l'exception de $proX - modC$), alors nous obtenons les quatre trimères présentés sur la Figure 3-(b).

Nous pouvons voir qu'une simple connaissance expert injectée dans les prédictions permet automatiquement à notre algorithme d'effectuer les bonnes prédictions d'interactions.

Ce type de configuration se reproduit pour un autre sous-graphe pour lequel 2 trimères sont prédits après une simple intervention de l'expert.

Après application de ce nouveau post-traitement, 9 des 12 trimères partiellement prédits (1 interactions sur 3) sont alors extraits avec deux interactions prédites sur les trois, permettant ainsi d'identifier 39 des 45 trimères.

4 Conclusion

Dans ce contexte spécifique de la biologie, cette méthode par combinaison de classifieurs s'avère performante. Elle permet de réduire de manière drastique le taux de faux négatifs, rejetant ainsi $94,74\%$ des paires de protéines. Les résultats montrent que l'échantillonnage aléatoire des données avec renforcement du taux de positifs a un impact positif sur la qualité des prédictions observées.

Concernant la reconstruction des trimères de protéines, l'extraction des 2-motifs fréquents en

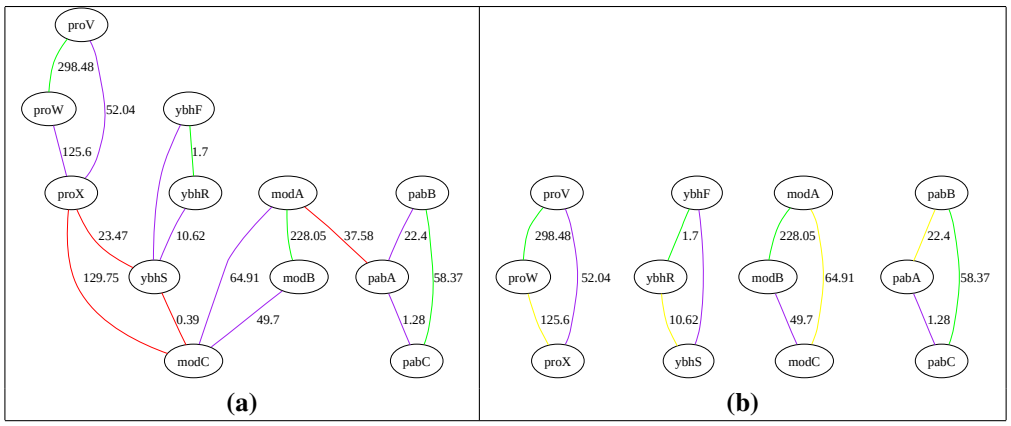

FIG. 3 – **(a)** *Trimères prédits par* $pTri$. **(b)** *Trimères obtenus après intervention de l'expert pour supprimer l'arc* $proX - modC$. *Les arcs verts représentent les prédictions justes (VP), les arcs rouges les prédictions incorrectes (FP), les arcs violets les interactions élaguées par* $pTri$ *(FN) et les arcs jaunes, les interactions automatiquement obtenues par ré-application locale de* $pTri$ *uniquement aux protéines de ce sous-graphe. Les valeurs indiquées sur les arcs correspondent aux scores associées aux paires de protéines. Un arc non valué correspond à une interaction élaguée par* M_2^{50}.

tenant compte du nombre de votes positifs associés à chaque paire de protéine, permet de retrouver le modèle mutlimérique dans $66,67\%$ des cas. L'étude des contacts structuraux directs a permis de montrer que ce taux pouvait atteindre $86,67\%$ de bonnes prédictions, en accord avec le principe même de coévolution où, en grande majorité, la modification évolutive d'une protéine "touche" principalement son interactant protéique direct. À très court terme, cette approche offre des perspectives remarquables dans les reconstructions des réseaux de signalisation cellulaire ou de docking, visant à prédire, au sein même de la cellule, quelles protéines interagissent entre elles et comment, d'un point de vue structural, celles-ci peuvent s'assembler.

Une généralisation de l'algorithme $pTri$ pour pouvoir l'étendre aux multimères de plus de 3 partenaires s'avère essentielle. En outre, afin de réduire un peu plus le taux de faux positifs, nous souhaiterions utiliser l'ensemble des classifieurs pour vérifier la qualité de l'annotation faite sur les données en isolant les interactions indirectes entre protéines prédites dans un même complexe.

Enfin à plus long terme, il serait intéressant d'introduire une procédure d'apprentissage actif en recoupant les votes associés au méta-classifieur et les données multi-sources mises à disposition par les experts en biologie.

Références

Bauer, E. et R. Kohavi (1999). An empirical comparison of voting classification algorithms : Bagging, boosting and variants. *Machine Learning 36*(Issue 1-2), 105–139.

Cangelosi, G. A., R. G. Ankenbauer, et E. W. Nester (1990). Sugars induce the Agrobacterium virulence genes through a periplasmic binding protein and a transmembrane signal protein. *PNAS 87*(17), 6708–6712.

Dawson, R. J. et K. P. Locher (2006). Structure of a bacterial multidrug ABC transporter. *Nature 443*(7108), 180–185.

de Vienne, D. M. et J. Azé (2012). Efficient prediction of co-complexed proteins based on coevolution. *PLoS One 7*(11), e48728.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The weka data mining software : An update. *SIGKDD Explorations 11*(1), 10–18.

Juan, D., F. Pazos, et A. Valencia (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *PNAS 105*(3), 934–939.

Lallich, S., B. Vaillant, et P. Lenca (2007). A probabilistic framework towards the parameterization of association rule interestingness measures. *MCAP, Methodology and Computing in Applied Probability 9*(3), 447–463.

Lenca, P., P. Meyer, P. Picouet, B. Vaillant, et S. Lallich (2003). Critères d'évaluation des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'information (RNTI) RNTI 1*, 123–134.

Marcotte, E., M. Pellegrini, H. Ng, D. Rice, T. Yeates, et D. Eisenberg (1999). Detecting protein function and protein-protein interactions from genome sequences. *SCIENCE 285*(5428), 751–753.

Moreno-Hagelsieb, G. et K. Latimer (2008). Choosing blast options for better detection of orthologs as reciprocal best hits. *Bioinformatics 24*(3), 319–324.

Opitz, D. et R. Maclin (1999). Popular ensemble methods : An empirical study. *Journal of Artificial Intelligence Research 11*, 169–198.

Pellegrini, M., E. Marcotte, M. Thompson, D. Eisenberg, et T. Yeates (1999). Assiging protein functions by comparative genome analysis : protein phylogenetic profiles. *PNAS 96*, 4285–4288.

Quinlan, J. (1996). Bagging, boosting, and c4.5. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 725–730.

Summary

We propose an approach to predict complexes with three proteins (trimers) by using classifiers learnt on protein-protein complexes (dimers). The prediction of trimers relies on two strong biological hypotheses: (i) two orthologous proteins share similar functional characteristics; (ii) two proteins interact as a complex to ensure an essential biological function for the studied species. These two hypotheses are used to describe each pair of proteins with the set of species for which they share an ortholog. A set of quality measures, initially developed for the evaluation of the interest of association rules, is used to evaluate the strength of the link between the two proteins. Our approach has been tested on *Eschericha Coli*.