Une nouvelle méthode de combinaison d'outils d'identification non supervisés

Antoine Cornuéjols et Christine Martin

AgroParisTech, département MMIP et INRA UMR-518
16, rue Claude Bernard
F-75231 Paris Cedex 5 (France)
antoine.cornuejols,christine.martin@agroparistech.fr

Abstract

La connaissance de l'interactome, le réseau complet des interactions protéine-protéine d'un organisme donné, requiert la mise en œuvre de méthodes de détection d'interaction encore peu fiables. Une approche séduisante consiste à combiner les résultats de méthodes différentes afin de diminuer les taux de faux positifs et de faux négatifs.

Dans ce papier, nous présentons une méthode originale non paramétrique dans ce but. L'idée fondamentale est de considérer un ensemble de fonctions d'évaluation des données d'apprentissage et de comparer les tris qu'elles produisent. Grâce à une nouvelle mesure de corrélation entre paires de tris, nous pouvons mesurer la différence de corrélation observée sur des échantillons différents. Un processus itératif de construction de nouvelles fonctions d'évaluation à partir des fonctions de base permet progressivement de mettre les données correspondant à ces interactions en tête de classement et ainsi de les identifier.

1 Introduction

La connaissance de l'interactome, le réseau complet des interactions protéine-protéine d'un organisme donné, est la clé de la compréhension des processus biologiques que sont les voies de signalisation, les voies métaboliques ou les mécanismes de transcription [10]. Elle permet aussi la prédiction des fonctions de protéines non annotées à partir des fonctions des protéines connues avec lesquelles elles interagissent. La recherche de méthodes d'identification des interactions protéine-protéine d'un organisme donné a donc été et reste très active [5].

Une grande partie des approches proposées repose sur l'hypothèse que les protéines en interaction co-évoluent dans le temps (voir [7]). C'est le cas des méthodes expérimentales : *double-hybride* [4] et *TAP-TAG* [13], et des méthodes computationnelles : *Phylogenetic Profiles* [12], *Genomic Context* [2], ou encore de *mirrortree* [11]. Ces méthodes de détection sont non supervisées. Elles prennent en entrée une description de paire de protéines et calculent en sortie un score, d'autant plus élevé que l'interaction est vraisemblablement réelle. Le problème est d'identifier un seuil à partir duquel est décidé si la paire de protéines appartient, ou pas, à la classe des interactions vraies. Sachant que la proportion des interactions vraies parmi toutes les interactions potentielles est en général petite (de l'ordre de 0.5%), leur recherche s'apparente au problème de la recherche de « nouveautés » en apprentissage artificiel. L'objectif est de détecter des objets *nouveaux* (e.g. les vraies interactions) au sein d'un ensemble d'objets *normaux* ou *nominaux* (e.g. les interactions putatives mais fausses).

2 La détection de « nouveautés »

Dans le problème appelé « détection de nouveauté », l'objectif est d'étiqueter les exemples d'un échantillon $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ donné afin de distinguer les objets *nouveaux* de ceux qui sont

nominaux. Dans la suite du papier, on notera '–' l'étiquette *nominale* et '+' l'étiquette des objets *nouveaux*. Cette discrimination passe souvent par l'apprentissage d'une règle de décision ou d'un modèle statistique avant l'étape d'étiquetage de $\mathcal{S}$.

Une grande partie des travaux antérieurs se place dans un contexte dans lequel on suppose que seuls des exemples de la classe *nominale* sont disponibles pour l'apprentissage. Il existe alors deux familles d'approches : paramétriques et non paramétriques. L'idée essentielle est d'attribuer à la classe '+' les points situés dans les régions de l'espace d'entrée $\mathcal{X}$ pour lesquels la vraisemblance selon le modèle de distribution estimé est inférieure à un certain seuil (approche paramétrique), ou bien pour lesquels la règle de décision apprise ne donne pas l'étiquette '–' (approche non paramétrique). Les approches paramétriques impliquent donc le choix d'une famille de distributions de probabilités, souvent un mélange de Gaussiennes, et l'estimation des paramètres à partir des exemples d'apprentissage (voir [8] pour une revue générale). Les approches non paramétriques s'affranchissent de l'estimation d'un modèle génératif des données en cherchant directement une règle de décision séparant les régions de $\mathcal{X}$ où les exemples '–' sont abondants de celles où ils sont rares (voir par exemple [14, 3]).

Une autre approche suppose l'existence d'un échantillon d'apprentissage $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{x}_{m+1}, \ldots, \mathbf{x}_{m+n}\}$ contenant m exemples *nominaux* (classe '–') et n exemples non supervisés pouvant contenir des exemples *nouveaux* (classe '+'). La plupart des techniques proposées [9] se réduisent à un problème de classification avec une étape préalable non supervisée, souvent de nature heuristique, d'identification des points '+' les plus probables, puis un apprentissage supervisé d'une règle de décision à partir de l'échantillon ainsi étiqueté. Dans une publication récente, Blanchard, Lee et Scott [1] montrent que les exemples non supervisés sont, en un certain sens, équivalents à la donnée d'exemples *nouveaux* (classe '+') si on utilise un critère inductif de Neyman-Pearson, ce qui montre l'intérêt de l'échantillon non étiqueté.

Cependant, ces approches supposent qu'il est possible, lors de l'apprentissage, d'isoler les exemples d'une classe, par exemple la classe des exemples « nominaux » (les '−'). Or ce n'est pas toujours le cas, comme le montre l'exemple de la recherche d'interactions protéine-protéine pour lequel on ne peut jamais exclure la présence de « vraies » interactions, ni leur absence. On peut néanmoins obtenir des échantillons de données dans lesquels la proportion d'exemples '+' diffère. Ainsi, la proportion de vraies interactions protéine-protéine diffère certainement dans un échantillon de paires de protéines prises aléatoirement et dans un échantillon de paires de protéines sélectionnées comme les plus prometteuses par une première méthode de sélection, même peu performante.

Nous présentons ici une méthode originale de combinaison de méthodes d'évaluation (scoring) pour sélectionner les exemples '+' à partir de données non étiquetées ou encore non supervisées, comme c'est le cas dans le problème d'identification des vraies interactions protéine-protéine. Cette méthode s'appuie de manière fondamentale sur une nouvelle mesure de corrélation entre fonctions d'évaluation (section 3). Elle permet, à l'instar du *boosting*, de construire itérativement un nouvel espace de description des données qui tend à faciliter la distinction entre exemples '+' et exemples '−' (section 4).

3 Mesure de corrélation entre fonctions d'évaluation

Dans la suite, on parlera de *fonction d'évaluation* pour désigner une fonction définie sur les exemples $\mathbf{x} \in \mathcal{X}$ et prenant sa valeur dans $\mathbb{R}$, soit $f : \mathcal{X} \to \mathbb{R}$.

Une mesure de corrélation entre fonctions d'évaluation mesure à quel point une information sur la valeur ou sur le classement d'un exemple par l'une des fonctions d'évaluation fournit une information sur la valeur ou sur le classement par l'autre fonction d'évaluation. Dans le cas des

classements, les mesures les plus utilisées sont celles de Spearman et de Kendall [6]. Elles ont l'inconvénient, pour notre problème, de mesurer la corrélation des classements de l'ensemble des exemples, alors que nous voudrions pouvoir nous concentrer sur les parties de classements où sont susceptibles d'apparaître les exemples '+'.

Dans la suite, nous appellerons ${top_n}^i$ du classement d'un ensemble d'exemples $\mathcal{S}$ par une fonction d'évaluation f_i, les n exemples les mieux classés par cette fonction. Nous noterons $\cap_n$ l'intersection des top_n de deux classements par deux fonctions d'évaluation : $\cap_n^{i,j} = top_n^i \cap top_n^j$. Ainsi si $top_5^i = \{a, b, c, d, e\}$ et $top_5^j = \{g, a, f, e, d\}$, alors $top_n^{i,j} = \{a, d, e\}$.

Nous proposons ici une nouvelle mesure de corrélation entre deux fonctions d'évaluation f_i et f_j qui est caractérisée par la fonction $\mathrm{Corr}_{i,j}^n(\mathcal{S}) = | \cap_n^{i,j} |$ pour $1 \leq n \leq m$ si $\mathrm{Card}(\mathcal{S}) = m$.

Cette mesure est inspirée de la *loi hypergéométrique* qui donne la loi probabiliste de la taille k de l'intersection de deux tirages indépendants et sans remise de n éléments parmi m :

$$\mathbf{p}(| \cap_n^{i,j} | = k) \;=\; \frac{\binom{n}{k} \cdot \binom{m-n}{n-k}}{\binom{m}{n}}$$

Si les résultats observés pour deux tirages s'éloignent de ceux prédits par la loi hypergéométrique, on peut suspecter que les deux tirages ne sont pas indépendants. Dans un cas extrême, l'un des tirages est une copie de l'autre, et la taille de l'intersection est donnée par : $| \cap_n^{i,j} | = n, (\forall n \leq m)$. Dans le cas extrême inverse, le deuxième tirage tire autant que possible des boules non tirées par le premier. Ce cas est analogue à celui de deux méthodes de tri qui rangent les éléments en sens inverse. La loi de la taille d'intersection est alors donnée par une taille nulle jusqu'à $n = m/2$ puis par une croissance selon $\frac{2(n-(m/2))}{n}$. Tout un spectre de comportements est ainsi possible selon le degré de corrélation des tirages.

La fonction de corrélation décrite s'applique à deux tris d'un *même* ensemble d'éléments.

Afin de calculer la corrélation entre deux fonctions d'évaluation indépendamment de l'échantillon étudié, on considère une collection de ℓ échantillons $\{\mathcal{E}_r\}_{1 \leq r \leq \ell}$ de m éléments chacun et on mesure la *corrélation moyenne* entre les tris induits par les deux fonctions d'évaluation :

$$\overline{\mathrm{Corr}}_{i,j}^{\,n} \;=\; \frac{1}{\ell} \sum_{r=1}^{\ell} \mathrm{Corr}_{i,j}^n(\mathcal{E}_r)$$

Deux fonctions d'évaluation f_i et f_j sont totalement décorrélées *a priori* si la corrélation moyenne entre les tris induits par ces fonctions est nulle en moyenne : $\overline{\mathrm{Corr}}_{i,j}^{\,n} = 0$. En revanche, la corrélation entre une fonction d'évaluation et sa copie $\overline{\mathrm{Corr}}_{i,j}^{\,n} = 1$.

Dans la suite, nous nous intéresserons aux paires de fonctions d'évaluation dont la différence de corrélation sur l'échantillon $\mathcal{S}$ et *a priori* est maximale :

$$\underset{f_i, f_j \in \mathcal{F}^2}{\mathrm{Argmax}} \left\{ \underset{n \leq |\mathcal{S}|}{\mathrm{Argmax}} \left[\mathrm{Corr}_{i,j}^n(\mathcal{S}) - \overline{\mathrm{Corr}}_{i,j}^{\,n} \right] \right\} \tag{1}$$

4 Une méthode d'ensemble utilisant la mesure de surcorrélation

À l'instar du boosting utilisé pour la classification supervisée, nous proposons une méthode d'ensemble construisant itérativement un nouvel espace de représentation $\Phi(\mathcal{X})$ qui est tel que les points '+' de l'échantillon considéré $\mathcal{S}$ tendent à se distinguer des points '−' (voir figure 1).

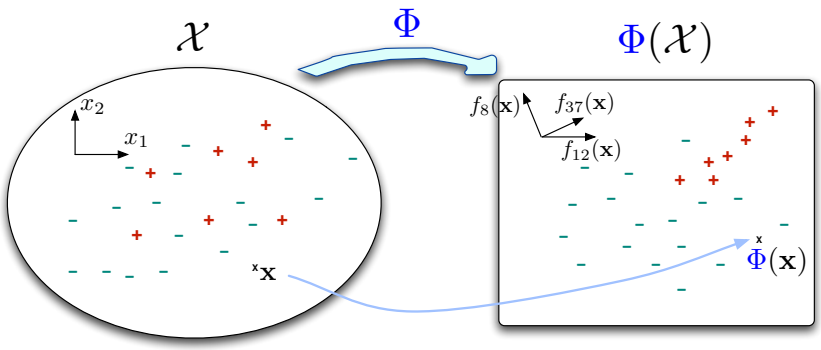

Figure 1: Les points de $\mathcal{S}$ décrits dans l'espace d'entrée $\mathcal{X}$ sont projetés dans un espace $\Phi(\mathcal{X})$ dont les axes, des fonctions d'évaluation, dépendent des caractéristiques de l'échantillon étudié $\mathcal{S}$. Les points '+' ont tendance à apparaître alignés selon une diagonale dans le nouvel espace.

La méthode sélectionne les fonctions d'évaluation qui sont les plus corrélées sur l'échantillon considéré $\mathcal{S}$ tout en étant les plus décorrélées *a priori* selon l'équation 1 ci-dessus. Cela nous permet d'obtenir un espace de fonctions d'évaluation « orthogonales » *a priori*, mais plutôt corrélées sur les points '+', ceux-ci étant donc proches de la diagonale principale dans ce nouvel espace.

5 Conclusion

Comme d'autres problèmes de détection de nouveautés, la recherche des interactions protéine-protéine d'un organisme donné s'appuie pour le moment sur des méthodes imparfaites d'évaluation de chaque paire protéine-protéine. Nous proposons une nouvelle méthode permettant la combinaison de telles méthodes « faibles » pour augmenter le pouvoir discriminant global. Cette méthode utilise une nouvelle mesure de corrélation entre fonctions d'évaluation permettant en particulier de se concentrer de manière adaptative sur les hauts des classements. Grâce à cette mesure, il est possible de sélectionner les fonctions d'évaluation les plus sensibles aux caractéristiques particulières de l'échantillon d'exemples (e.g. paires de protéines) étudié et de construire ainsi itérativement un nouvel espace de redescription des données. Dans cet espace, les points '+' (e.g. les vraies interactions protéine-protéine) sont mises en évidence le long de la diagonale principale.

References

[1] G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009, 2010.

[2] T. Dandekar, B. Snel, M. Huynen, P. Bork, et al. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences*, 23(9):324, 1998.

[3] R. El-Yaniv and M. Nisenson. Optimal single-class classification strategies. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, volume 19, page 377. MIT Press, 2007.

[4] S. Fields and O. Song. A novel genetic system to detect protein protein interactions. *Nature*, 340:245–246, 1989.

[5] V. Helms. *Principles of computational cell biology.* Wiley-VCH, 2008.

[6] M.G. Kendall. *Rank correlation methods.* Griffin, 1970.

[7] S.C. Lovell and D.L. Robertson. An integrated view of molecular coevolution in protein–protein interactions. *Molecular biology and evolution*, 27(11):2567–2575, 2010.

[8] M. Markou and S. Singh. Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[9] S. Marsland. Novelty detection in learning systems. *Neural computing surveys*, 3(2):157–195, 2003.

[10] F. Pazos, J.A.G. Ranea, D. Juan, M.J.E. Sternberg, et al. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *Journal of molecular biology*, 352(4):1002–1015, 2005.

[11] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein engineering*, 14(9):609–614, 2001.

[12] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

[13] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, B. Séraphin, et al. A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, 17(10):1030–1032, 1999.

[14] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.