Mesures hiérarchiques pondérées pour l'évaluation d'un système semi-automatique d'annotation de génomes utilisant des arbres de décision

L. Gentils*, J. Azé*, C. Toffano-Nioche*, V. Loux**, A. Poupon***, J-F. Gibrat**, C. Froidevaux*

* LRI UMR 8623 CNRS, Univ. Paris-Sud 11 F-91405 Orsay France
(Lucie.Gentils,Claire.Toffano-Nioche,Jerome.Aze,Christine.Froidevaux)@lri.fr
http://www.lri.fr
** MIG INRA, Domaine de Vilvert 78352 Jouy-en-Josas Cedex France
(Valentin.Loux,Jean-Francois.Gibrat)@jouy.inra.fr , http://mig.jouy.inra.fr
*** IBBMC UMR 8619 CNRS, Univ. Paris-Sud 11 F-91405 Orsay France
anne@rezo.net , www.ibbmc.u-psud.fr

Résumé. L'annotation d'une protéine consiste, entre autres, à lui attribuer une classe dans une hiérarchie fonctionnelle. Celle-ci permet d'organiser les connaissances biologiques et d'utiliser un vocabulaire contrôlé. Pour estimer la pertinence des annotations, des mesures telles que la précision, le rappel, la spécificité et le Fscore sont utilisées. Cependant ces mesures ne sont pas toujours bien adaptées à l'évaluation de données hiérarchiques, car elles ne permettent pas de distinguer les erreurs faites aux différents niveaux de la hiérarchie. Nous proposons ici une représentation formelle pour les différents types d'erreurs adaptés à notre problème.

1 Introduction

Aujourd'hui de nombreux génomes séquencés sont disponibles du fait du développement continu des technologies à haut débit et des procédures expérimentales[1]. Les experts biologistes jouent un rôle central dans l'analyse et l'annotation de cette quantité massive de données brutes. Pour annoter un nouveau génome, ils doivent intégrer plusieurs types d'informations en provenance de sources variées, ce qui prend entre 12 et 18 mois à une équipe de 2 à 4 personnes pour un petit génome bactérien contenant environ 2000 gènes. Pour faire face au déluge des nouvelles données génomiques, le processus d'annotation doit être le plus automatisé possible. Dans le contexte du projet RAFALE[2], nous proposons aux biologistes utilisant la plate-forme AGMIAL[3], un système semi-automatique d'annotation fonctionnelle de protéines. Nous proposons un système semi-automatique car le processus est collaboratif : pour chaque protéine, une annotation est suggérée par le système et les biologistes décident de l'annotation finale.

[1] http ://www.genomesonline.org
[2] http ://www.lri.fr/RAFALE/
[3] http ://genome.jouy.inra.fr/agmial

Les annotations sont issues d'arbres de décision obtenus par deux approches différentes : TILDE (Blockeel et Raedt, 1998) et Clus-HMC (Blockeel et al., 2006). Ces arbres sont appris sur les données d'un génome puis testés sur un autre génome (tous deux disponibles dans AGMIAL). Les classes fonctionelles que l'on propose (appelées prédictions) et les classes fonctionnelles données par l'expert aux protéines (appelées annotations) sont classées dans une hiérarchie fonctionnelle dérivée de celle de Subtilist (Moszer et al., 2002).

2 Mesures d'évaluation hiérarchiques

Les prédictions obtenues par les arbres de décision sont pondérées par un indice de confiance qui est égal au pourcentage d'exemples arrivant dans la feuille de l'arbre permettant d'effectuer la prédiction, arbre appris sur l'ensemble d'apprentissage.

Cet indice de confiance est utilisé pour contrôler la qualité des prédictions *via* un seuil d'élagage défini par l'expert.

Afin de mieux évaluer la pertinence des prédictions effectuées, nous avons exhaustivement recensé les types de couples annotation/prédiction possibles, en fonction des différents niveaux de la hiérarchie. Ce classement peut être comparé à celui de TABS (Iliopoulos et al., 2003) qui associe un score à chaque type de différences observées entre deux annotations d'un même génome. Comme nous ne travaillons pas dans le même contexte d'étude, certains types de différences de TABS ne s'appliquent pas à notre système (erreurs de typographies, annotations non répertoriées, erreurs de domaines, prédictions sans annotations)[4].

Nous définissons les types de différences, entre une annotation et une prédiction, répertoriés dans TABS et utilisés dans notre système à l'aide des notations suivantes :

Notons E l'ensemble des exemples étudiés (ici les protéines). Soit $x \in E$ une protéine et $f(x,i)$ (resp. $\widehat{f}(x,i)$) l'annotation (resp. la prédiction) de x, si elle existe, au niveau i de la hiérarchie utilisée. Soit $A^k_{i,j} = \{x \in E | \forall k' \in [1,k], f(x,k') = \widehat{f}(x,k') \, et \, \forall i' \in [1,i], f(x,i')$ est défini $; \forall j' \in [1,j], \widehat{f}(x,j')$ est défini $\}$ l'ensemble des protéines pour lesquelles annotation et prédiction sont en adéquation jusqu'au niveau k de la hiérarchie mais pas au niveau $k+1$, et il existe une annotation (resp. une prédiction) jusqu'au niveau i (resp. j).

Les types de différences de TABS pertinents pour notre approche sont les suivants : l'annotation et la prédiction existent et sont identiques jusqu'au niveau i (cas d'une protéine dans $A^i_{i,i}$) ; l'annotation est plus précise que la prédiction : elles sont identiques jusqu'au niveau j (cas d'une protéine dans $A^j_{i,j}$ avec $i > j$) et l'annotation est moins précise que la prédiction : elles sont identiques jusqu'au niveau i (cas d'une protéine dans $A^i_{i,j}$ avec $i < j$). De plus, nous ajoutons pour notre système le cas où i est le premier niveau où la prédiction diffère de l'annotation (cas d'une protéine dans $A^{i-1}_{i,i}$).

L'évaluation de la qualité des prédictions est effectuée avec différentes mesures hiérarchiques telles que : la précision, le rappel, la spécificité ou le Fscore.

2.1 Etude des prédictions plus précises que l'annotation

Le cas d'une protéine annotée 3.5.2 et prédite 3.5.1 (incluse dans l'ensemble $A^2_{3,3}$) entraînera, avec l'approche usuelle, un décompte de prédictions aux niveaux 1 et 2 justes (3.5 et 3)

[4] plus de détails sur `www.lri.fr/RAFALE/EGC08`

et d'une prédiction fausse au niveau 3 (3.5.1). De manière similaire, une protéine annotée 3.5 et prédite 3.5.1 (ensemble $A^2_{2,3}$) entraînera le décompte de prédictions justes aux niveaux 1 et 2 (3.5 et 3) et d'une erreur (3.5.1).

Or, si le premier cas correspond clairement à une erreur de prédiction, le second cas correspond à une sous-prédiction supplémentaire par rapport à l'existant qui pourrait permettre de raffiner une annotation, ce qui est potentiellement intéressant pour le biologiste.

La figure 1 présente différentes configurations d'intérêt pour notre étude.

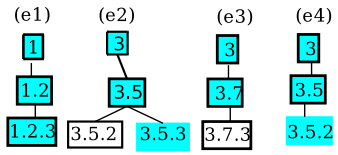


FIG. 1 – *Exemples de différences entre annotations (encadrées) et prédictions (grisées) dans le cadre de mesures hiérarchiques.*

- (e_1) correspond à l'adéquation parfaite entre l'annotation et la prédiction ($e_1 \in A^3_{3,3}$).
- (e_2) correspond au cas où l'annotation et la prédiction sont connues jusqu'au niveau 3 de la hiérarchie et ne sont en accord que jusqu'au niveau 2. Les prédictions de niveau 1 et 2 sont considérées comme justes, la prédiction de niveau 3 est considérée comme erronée par rapport à l'annotation ($e_2 \in A^2_{3,3}$).
- (e_3) correspond au cas où l'annotation est plus précise que la prédiction. Le système, bien que n'ayant pas commis d'erreur de prédiction, n'a pas su prédire suffisament finement la classe fonctionnelle ($e_3 \in A^2_{3,2}$).
- Enfin, le cas (e_4) illustre une prédiction plus précise que l'annotation courante. Le système a prédit la classe fonctionnelle 3.5.2 alors que l'annotation est 3.5. Ce cas est traditionnellement considéré comme érroné ($e_4 \in A^2_{2,3}$).

Les protéines possédant une prédiction plus précise (comme e_4)peuvent être classées de deux manières selon le sens que l'on décide de leur attribuer :

- elles interviennent dans le calcul du décompte des sous-prédictions erronées quand une sous-prédiction supplémentaire est comptée comme une erreur. Cette signification est couramment utilisée afin d'éviter de propager des erreurs en cas d'annotation d'un génome par rapport à un autre.
- elles interviennent à la fois dans le calcul du décompte des sous-prédictions justes et erronées pour moitié, en considérant qu'une sous-prédiction supplémentaire peut être aussi bien une erreur qu'une spécialisation justifiée (par exemple, dans le cas où de nouvelles informations sont disponibles depuis l'annotation).

Nous nous sommes focalisés sur l'analyse des protéines correspondant au cas (e_4). Les résultats des expérimentations que nous avons réalisées sur les génomes *L.bacillus* et *L.sakei* annotés avec la plateforme AGMIAL sont présentés dans la section suivante.

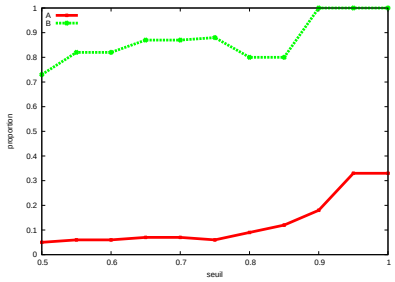
Dans la suite de cet article, nous choisissons la deuxième signification et considérons qu'une sous-prédiction supplémentaire est à la fois une erreur et une spécification justifiée.

3 Résultats expérimentaux

Les résultats sont analysés en fonction du seuil d'élagage appliqué aux arbres de décision (évoqués à la section 2).

Les courbes de la figure 2 présentent la proportion de prédictions plus précises entre les niveaux 1 et 2 (courbe A) et les niveaux 2 et 3 (courbe B). Dans l'approche usuelle, ces protéines sont indifférenciables des protéines ayant une prédiction réellement erronée.

D'après ces courbes, nous pouvons remarquer que nous identifions des prédictions plus précises que l'annotation, essentiellement au troisième niveau de la hiérarchie (jusqu'à 100% pour *L.sakei* à partir d'un seuil d'élagage 0.90, courbe B, correspondant aux cas des différences entre les niveaux 2 et 3).



apprentissage *L.bulgaricus*, tests *L.sakei*

FIG. 2 – *Proportion de prédictions plus précises par rapport à l'annotation mise en valeur par rapport au seuil d'élagage pour les différences entre le premier et le deuxième niveau (courbe A, en bas) et entre le deuxième et le troisième niveau (courbe B, en haut)*

Dans le cas d'un apprentissage réalisé sur *L.bulgaricus*, testé sur *L.sakei*, et avec un seuil d'élagage de 0.75, nous distinguons 28 protéines au niveau 2 (4 en $A^2_{1,2}$ et 24 en $A^2_{2,3}$) pour lesquelles les prédictions plus précises ne sont peut-être pas à comptabiliser comme des erreurs (voir courbe B).

De manière similaire, nous observons (courbe A) pour le même seuil d'élagage, uniquement 4 sous-prédictions plus fines que les annotations au niveau 2. Cela est potentiellement dû à la qualité des annotations. En effet, la plupart des protéines des génomes étudiés sont annotées aux niveaux 2 ou 3 de la hiérarchie et rares sont celles qui sont uniquement annotées au niveau 1.

Voici par exemple, les règles concluant à la prédiction de la classe 1.2.5 de la protéine 157 de *L.sakei* qui est annotée en 1.2 :

```
- si blastmatchGo(A,GO :0006810,C,D),D>0.6 alors c1 (acc 97)

- si blastmatchGo(A,GO :0006810,C,D) et
   non blastmatchGo(A,GO :0016469,E,F),F>0.625 alors c12 (acc95)

- si non blastmatchGo(A,GO :0009401,C,D) et
   non blastmatchGo(A,GO :0003824,E,F) et
   non blastmatchSw(A,Lipoprotein,G,H) et
```

`blastmatchGo(A,GO :0006865,I,J) alors 1.2.5 (acc 91)`

Ces règles sont des chemins issus de plusieurs arbres appris par TILDE à partir d'informations utilisées sur les protéines de *L.bulgaricus*. La hiérarchie que nous utilisons est organisée en trois niveaux, nous prédisons donc une classe fonctionnelle en trois étapes correspondant aux niveaux successifs.

1. Au premier niveau, comme plus de 60% des protéines qui ressemblent à la protéine considérée (sous certaines conditions), sont annotées avec le terme *GO :0006810 transport*, la protéine esa157 est classée dans la classe 1 (*Cell envelope and cellular processes*) avec un indice de confiance de 97%. L'indice de confiance, `(acc97)`, correspond au nombre de protéines dont l'annotation et la prédiction sont en adéquation lors de la phase d'apprentissage des arbres.
2. Au second niveau, comme il existe des protéines qui ressemblent à la protéine considérée, annotées avec le terme *GO :0006810 transport* et comme il n'y a pas plus de 62.5% des protéines qui ressemblent à la protéine esa157, et qui sont annotées avec le terme *GO :0016469 proton-transporting two-sector ATPase complex*, la protéine esa157 est classée en 1.2 *Transport/binding proteins and lipoproteins* (indice de confiance de 95%).
3. Au troisième niveau comme il n'existe pas de protéine qui ressemble à la protéine considérée et qui soit annotée avec le terme *GO :0009401 phosphoenolpyruvate-dependent sugar phosphotransferase system* ou *GO :0003824 catalytic activity* ou le mot clé Swiss-Prot *Lipoprotein*, et comme il existe des protéines qui ressemblent à la protéine considérée, qui sont annotées avec le terme *GO :0006865 amino acid transport* alors la protéine esa157 est classée en 1.2.5 *Transport/binding of amino-acids* (indice de confiance de 91%).

L'identification explicite des protéines ayant des prédictions plus précises que l'annotation permettra à l'expert de les traiter séparément, surtout dans le cadre de réannotation de génomes où les informations complémentaires apportées par l'annotation automatique sur ces protéines pourront être analysées plus rapidement. Ici, pour la protéine esa157, l'annotation initiale 1.2 peut être corrigée en $1.2.5$, comme prédit (après consultation de l'annotateur).

4 Conclusion et perspectives

Les mesures hiérarchiques usuelles prennent en compte la hiérarchie mais elles ne traduisent pas bien la statégie d'annotation des experts biologistes qui évitent de propager des erreurs.

De plus le cas de prédiction plus fine que l'annotation existante est traité comme une erreur par les mesures usuelles alors que dans un cadre d'annotation fonctionnelle, cette information doit impérativement être différenciée.

Nous avons répertorié les types de différences qui peuvent exister entre une annotation et une prédiction dans notre problème d'annotation fonctionnelle, et proposé une formalisation pour les représenter, valable quelque soit le niveau de profondeur de la hiérarchie.

Nous travaillons actuellement sur de nouvelles mesures hiérarchiques permettant de traiter ces cas à l'aide de cette formalisation. Nous pourrons ainsi traiter différemment les protéines ayant des prédictions plus précises et les protéines ayant une prédiction erronée.

En outre, nous pourrons évaluer différentes mesures selon le but recherché, en choisissant les poids à attribuer aux types de différences répertoriées entre une prédiction et une annotation.

Avec l'utilisation de ces mesures, nous pouvons comparer de façon plus appropriée les annotations et les prédictions des protéines d'un génome ou deux annotations d'un même génome, dans le cas d'une réannotation.

L'optimisation de la méthode de prédiction suivant ces mesures permettrait en outre d'être plus proche de la stratégie des annotateurs, et ainsi d'éviter ou de mieux prendre en compte la sur-annotation qui est souvent la source de propagation d'erreur, tout en assurant une bonne qualité des annotations au premier niveau.

Remerciements : Nous remercions l'ACI IMPBIO qui a soutenu le projet RAFALE.

Références

Blockeel, H. et L. D. Raedt (1998). Top-down induction of first-order logical decision trees. *Artificial Intelligence 101*(1-2), 285–297.

Blockeel, H., L. Schietgat, J. Struyf, S. Dzeroski, et A. Clare (2006). Decision trees for hierarchical multilabel classification : A case study in functional genomics. In *PKDD'06*, pp. 18–29.

Iliopoulos, I., S. Tsoka, M. A. Andrade, A. J. Enright, M. Carroll, P. Poullet, V. Promponas, T. Liakopoulos, G. Palaios, C. Pasquier, S. Hamodrakas, J. Tamames, A. T. Yagnik, A. Tramontano, D. Devos, C. Blaschke, A. Valencia, D. Brett, D. Martin, C. Leroy, I. Rigoutsos, C. Sander, et C. A. Ouzounis (2003). Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics 19*(6), 717–26.

Moszer, I., L. Jones, S. Moreira, C. Fabry, et A. Danchin (2002). Subtilist : the reference database for the *bacillus subtilis* genome. *Nucleic Acids Res 30*, 62–5.

Summary

One main goal of protein annotation is to associate a class in a functional hierarchy to the considered protein. This hierarchy allows to organize biological knowledge and to use a controlled vocabulary. To estimate the relevance of the annotations, measures such as precision, recall, specificity and Fscore are used. However these measures are not always well adapted to the evaluation of hierarchical data, as they do not allow to distinguish errors made on the various levels of the hierarchy. We propose, here, a formal representation for the various types of misannotation well-suited to our problem.