

# Stage M2

## Annotation fonctionnelle de protéines par apprentissage actif

Jérôme Azé et Christine Froidevaux  
*Équipe Bioinformatique*  
*LRI – Université Paris-Sud 11*

2009-2010

### Mots clés

Apprentissage supervisé, apprentissage actif, annotation fonctionnelle de protéines

## 1 Cadre général de l'annotation fonctionnelle de protéines

À l'heure actuelle, il y a 762 génomes de bactéries et d'archées qui ont été séquencés et publiés et il y a plus de 2000 projets de séquençage en cours. Les méthodes de séquençage deviennent de plus en plus puissantes induisant un coût de séquençage de plus en plus faible. Il existe ainsi une initiative du NIH visant à développer de nouvelles techniques permettant de séquencer le génome humain pour une somme de 1000\$. Dans ces conditions, on peut donc prévoir sans trop se tromper que le nombre de génomes séquencés va continuer à croître de façon exponentielle.

Les techniques de séquençage produisent des données brutes, des séquences nucléiques, qui en tant que telles ne sont que modérément intéressantes. Il est nécessaire d'analyser ces données pour en extraire des connaissances biologiques sur l'organisme étudié. Ce travail d'analyse, le passage de données à des connaissances, est ce que l'on appelle l'annotation (considérée au sens large) des données génomiques. L'annotation consiste à valider, croiser, intégrer des données provenant de sources diverses : les résultats des programmes bioinformatique d'analyse, les données stockées dans les grandes collections de données biologiques (Genbank/Embl, UniprotKB, Kegg, PDB, etc.), les résultats d'expériences à grande échelle (transcriptomique, protéomique, double-hybride, etc.) ainsi que les connaissances présentes dans la littérature du domaine. Même pour un génome bactérien qui ne contient qu'un nombre relativement limité de gènes (entre 2000 et 5000), c'est un travail considérable.

De façon à assister les biologistes dans cette tâche formidable un certain nombre de plates-formes d'annotation ont été développées. Les fonctionnalités particulières de ces plates-formes diffèrent selon les cas mais, en général, leur objectif est d'automatiser les traitements informatiques, de centraliser les données, de permettre de manipuler et de faire des recherches sur les données aisément, de fournir des interfaces homme-machine conviviales facilitant le travail des biologistes. En résumé, le rôle de ces plates-formes est de prendre en charge l'ensemble des tâches purement informatiques afin de permettre aux biologistes de se consacrer exclusivement aux tâches de haut niveau.

En effet, pour le moment, il est nécessaire que l'annotation automatique effectuée par la plate-forme soit supervisée par des experts humains si l'on désire obtenir une annotation d'excellente qualité.

L'unité MIG du centre INRA de Jouy-en-Josas a développé, depuis 2001, une telle plate-forme, nommée AGMIAL[2], pour aider à l'annotation de génomes bactériens d'intérêt agro-alimentaire[3, 4]. L'unité MIG est impliquée dans une douzaine de projets d'analyse de génomes de l'institut. MIG a donc une bonne expérience des problèmes pratiques qui se posent dans ce domaine. Entre autres, il est apparu très vite que l'annotation supervisée par des experts constituait le goulot d'étranglement principal de l'analyse d'un nouveau génome. Il faut 2 jours pour que la plate-forme fournisse l'annotation automatique et il faut entre 18 mois et 2 ans à une petite équipe d'annotateurs pour

produire l'annotation manuelle de qualité supérieure à celle obtenue automatiquement. Si cette situation perdure, on peut penser que dans un futur proche on disposera de grandes masses de données brutes très partiellement analysées et qui auront donc, de ce fait, un intérêt moindre.

Nous pensons donc qu'il est nécessaire de tenter d'automatiser le processus d'annotation lui-même afin d'améliorer la productivité des annotateurs. Comme indiqué ci-dessus ce processus est très complexe car il fait intervenir des données très nombreuses et très hétérogènes et nécessite des connaissances générales sur la biologie et des connaissances particulières sur l'organisme étudié. C'est pourquoi dans un premier temps nous désirons proposer un système semi-automatique qui permette d'augmenter la productivité des annotateurs tout en leur laissant la décision finale quant à l'annotation la plus appropriée. En outre, un tel système a aussi l'avantage d'assurer une meilleure cohérence entre les annotations de génomes effectuées par différents groupes d'annotateurs.

## 2 Contexte du stage

Dans le cadre du projet RAFALE de l'ACI IMPBio, le LRI et l'IBBMC de l'université Paris-Sud 11, Centre scientifique d'Orsay, ainsi que MIG ont développé un système semi-automatique d'annotation se basant sur des arbres de décision. À l'origine de ce choix était la volonté de produire des règles d'annotation qui soient facilement interprétables par les experts biologistes. Il est important de noter que le prototype développé s'intéresse à un problème particulier de l'annotation qui consiste à attribuer aux protéines du génome une (ou plusieurs) classe particulière d'une hiérarchie fonctionnelle (un nœud ou une feuille). Ce type de hiérarchie fonctionnelle permet d'utiliser un vocabulaire contrôlé et de synthétiser l'information génomique en "classant" les protéines selon des catégories générales décrivant les grandes fonctions biologiques de la cellule. Le prototype produit a fourni une série de résultats intéressants[1].

Nous souhaitons, dans le cadre du projet AFON<sup>1</sup>, étendre ces résultats en :

- modifiant la nature des modèles appris :  $k$  plus-proche-voisins, bayésien naïf, SVM, ... (moins expressifs mais souvent plus rapides à évaluer)
- combinant différents modèles
- incluant l'expert dans le processus d'apprentissage (apprentissage actif)
- intégrant plus finement la hiérarchie fonctionnelle
- concevant un système d'annotation actif permettant de rendre compte des scénarios d'annotation propres à chaque expert

Un système d'annotation fonctionnelle par apprentissage actif a déjà été élaboré et testé sur les données du projet RAFALE. Les résultats obtenus ont permis de montrer la faisabilité et l'intérêt d'un tel système. Celui-ci n'intègre actuellement pas toutes les fonctionnalités listées ci-dessus.

## 3 Travail à effectuer

L'objectif du stage est d'étendre ce système pour y intégrer des mesures hiérarchiques plus souples qui, compte tenu de l'implication de l'expert dans le processus d'annotation actif, permettraient de mieux s'adapter aux données nouvellement disponibles.

Un processus d'annotation par apprentissage actif suppose que le système soit capable de :

1. sélectionner les exemples les plus susceptibles d'améliorer les performances du système d'annotation en cours d'apprentissage
2. solliciter l'expert pour obtenir la véritable annotation de tout ou partie de ces exemples
3. mettre à jour le système d'annotation
4. itérer sur le point 1

Le choix des exemples à proposer à l'expert est non trivial car il fait intervenir deux caractéristiques *a priori* décorréliées :

1. l'"utilité" des exemples choisis pour le système d'annotation en cours de construction
2. l'adéquation des exemples avec les scénarios d'annotation de l'expert

---

<sup>1</sup>Annotation FONctionnelle, projet UniverSud, Pôle Thématique "Biomédical - Santé" "Biologie Systémique et Synthétique"

### 3.1 “Utilité” des exemples choisis pour le système d’annotation en cours de construction

L’utilité des exemples peut être mesurée par le nombre de prédictions en conflit pour chaque exemple. *A priori*, plus un exemple est consensuel et moins il apportera d’informations pour améliorer le système d’annotation en cours d’apprentissage. Inversement, plus un exemple est conflictuel et plus il est susceptible de se trouver à la frontière entre plusieurs classes. La connaissance de la classe de ce type d’exemples peut donc s’avérer très utile pour le système en cours d’apprentissage.

### 3.2 Adéquation des exemples avec les scénarios d’annotation de l’expert

Du point de vue de l’annotateur, le fait de se voir proposer une séquence d’exemples, en apparence, totalement décorrélés du point de vue des fonctions biologiques (classes fonctionnelles) peut être très déstabilisant.

Considérons une hiérarchie permettant d’affecter à chaque “news” une thématique :

- 1 : politique
  - 1.1 : politique intérieure
  - 1.2 : politique internationale
- 2 : sport
  - 2.1 : rugby
  - 2.2 : football
  - 2.3 : tennis
  - 2.4 : cyclisme
  - 2.5 : golf
- 3 : culture
  - 3.1 : théâtre
  - 3.2 : cinéma
    - 3.2.1 : cinéma d’auteurs
    - 3.2.2 : science fiction
    - 3.2.3 : dessins animés
    - 3.2.4 : autres
  - 3.3 : littérature
    - 3.3.1 : bandes dessinées
    - 3.3.2 : romans
    - 3.3.3 : essais

Si, lors d’une itération de l’apprentissage actif, le système demande à un expert d’annoter une séquence de news telles que la première news relève de la politique intérieure (classe 1.1), la suivante des bandes dessinées (classe 3.3.1), puis une news de la classe cinéma d’auteurs, puis rugby, puis cinéma d’auteurs, puis golf, ... l’expert devra se livrer à une gymnastique intellectuelle relativement épuisante qui ne l’incitera pas à utiliser le système. De plus, dans le domaine de l’annotation fonctionnelle de protéines, plusieurs experts sont susceptibles d’intervenir dans le processus d’annotation. Chaque expert ayant des connaissances spécifiques à une partie de la hiérarchie fonctionnelle, mais pas à l’intégralité de celle-ci, il convient de pouvoir proposer, pour un expert donné, des protéines à annoter compatibles avec ses connaissances.

Le prototype n’intègre actuellement que des critères d’“utilité” et ne prends donc pas vraiment en considération les connaissances de l’expert telles qu’elles ont été présentées ici.

Il conviendra donc de concevoir une approche permettant de prendre en considération le point de vue de l’expert et de l’intégrer dans le prototype existant.

## 4 Connaissances requises

Tous les programmes existants actuellement sont écrits en C. Les analyseurs de données sont écrits, le plus souvent, en Perl. Une maîtrise de ces deux langages sera donc fortement appréciée dans le cadre de ce stage.

Aucun pré-requis en biologie n'est demandé pour pouvoir effectuer ce stage, par contre un intérêt et une curiosité pour le domaine de la bioinformatique est plus que souhaitable.

## Références

- [1] J. Azé, L. Gentils, C. Toffano-Nioche, J.-F. Loux, V. Gibrat, P. Bessières, A. Rouveirol, C. Poupon, and C. Froidevaux. Towards a semi-automatic functional annotation tool based on decision-tree techniques. *BMC Proceedings 2008*, 2((Suppl 4) :S3), 2008.
- [2] K. Bryson, V. Loux, R. Bossy, P. Nicolas, M. van de Chaillou, S. Guchte, S. Penaud, E. Maguin, M. Hoebeke, P. Bessières, and J.-F. Gibrat. AGMIAL : implementing an annotation strategy for prokaryote genomes as a distributed system. *Nucleic Acids Res*, 34(12), 2006.
- [3] S. Chaillou, M.-C. Champomier-Vergès, M. Cornet, A.-M. Crutz-Le Coq, A.-M. Dudez, V. Martin, S. Beaufils, E. Darbon-Rongère, R. Bossy, V. Loux, and M. Zagorec. The complete genome sequence of the meat-borne lactic acid bacterium *Lactobacillus sakei* 23k. *Nature Biotechnology*, 23 :1527–33, 2005.
- [4] E. Duchaud, M. Boussaha, V. Loux, J.-F. Bernardet, C. Michel, B. Kerouault, S. Mondot, P. Nicolas, R. Bossy, C. Caron, P. Bessieres, J.-F. Gibrat, S. Claverol, M.L. Dumetz, F. Henaff, and A. Benmansour. Complete genome sequence of the fish pathogen *Flavobacterium psychrophilum*. *Nat Biotechnol*, 25 :763–9, 2007.