# RAFALE

(Projet de Recherche 2005)

# Règles pour l'Annotation Fonctionnelle semi-Automatisée de la LEvure

**http://www.lri.fr/RAFALE/**

## CONTEXTE

Il y a 10 ans le séquençage des premiers génomes bactériens requérait les efforts conjugués de plusieurs dizaines de laboratoires. A l'heure actuelle le séquençage du génome d'un organisme est à la portée d'une équipe de biologistes. Dans ces études, le but poursuivi par les biologistes est de comprendre comment le génome d'un organisme permet d'expliquer ses propriétés biologiques. Ce processus d'extraction de connaissances biologiques à partir des données brutes est appelé **annotation**. L'annotation vise dans un premier temps à caractériser la fonction des gènes et des protéines, puis à décrire leurs interactions, qui sous-tendent les différents processus biologiques.

Aujourd'hui, pour obtenir une annotation d'excellente qualité du génome d'un organisme, il est nécessaire que des experts humains, spécialistes de la biologie de l'organisme en question, supervisent l'annotation. Pour une équipe de taille réduite c'est un travail considérable qui constitue clairement le goulot d'étranglement de l'analyse des données génomiques.

Ce travail a été facilité par le développement de plates-formes d'annotation. Le rôle de ces plates-formes est de permettre aux experts humains de se concentrer uniquement sur les tâches de haut niveau, en automatisant les autre tâches, en centralisant et facilitant l'accès aux données et résultats, en croisant diverses sources d'information, en fournissant une synthèse pertinente des résultats obtenus. Néanmoins le processus d'annotation repose entièrement sur les épaules des experts qui doivent valider, croiser, vérifier et intégrer les données et résultats générés par le système pour produire une annotation cohérente.

## OBJECTIFS

MIG a développé une plate-forme d'annotation, appelée AGMIAL, qui a été utilisée pour annoter les génomes de *L. sakei, L. bulgaricus* et *F. psychrophilum*. Le traitement automatique des données prend environ 5 jours de calcul puis il faut plusieurs mois de travail pour qu'une équipe de 2-3 annotateurs annotent un génome. L'objectif du projet RAFALE est donc de faciliter la tâche des experts humains en leur fournissant une **aide semi-automatique** à l'annotation. Pour cela, nous concevons et nous mettons au point d'un système à base de règles qui reflète la connaissance des experts et qui permette l'annotation des données génomiques. Dans un premier temps nous nous concentrons sur les bactéries qui sont plus étudiées et plus simples que les levures. Nous appliquerons ensuite notre stratégie aux levures.

## MOTS CLÉS

Annotation fonctionnelle, levures, bactéries, outil d'aide aux experts biologistes, méthode semi-automatique, apprentissage, programmation logique inductive, système à base de règles, système d'inférence, bases de données relationnelles, intégration de données.

## LABORATOIRES IMPLIQUÉS


• **IBBMC** Institut de Biochimie et de Biophysique Moléculaire et Cellulaire, CNRS, UMR 8619, Université Paris-Sud, Orsay
• **LRI** Laboratoire de Recherche en Informatique, CNRS UMR 8623, Université Paris-Sud, Orsay
• **MIG** Mathématiques Informatique et Génome, INRA Jouy-en-Josas


## PUBLICATIONS


• K. Coeytaux, A. Poupon : Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics* 2005 21: 1891 – 1900
• R. Bossy, V. Loux, K. Bryson, P. Nicolas, P. Bessières and J.-F. Gibrat : AGMIAL: a distributed genome annotation based on cooperative Web Services. *JOBIM* 2005 371-381


## CALENDRIER

• Depuis octobre 2004 : 8 réunions de travail (une par mois).
• 25/11/2004 à l'IGM, journée Annotation du Programme PluriFormation Bioinformatique et Génomique (Université Paris-Sud) : présentation du projet RAFALE (Ch. Froidevaux) et présentation de la plate-forme d'annotation AGMIAL (J-F. Gibrat).
• 17/03/2005 au LRI, séminaire Langage du Génome : exposé de C. Médigue sur MicroScope (projet ACI IMPBIO).

## PROJET

Le déroulement du projet comporte 3 étapes. La première étape consiste à construire un **jeu de données** permettant de développer et tester le système de règles. La seconde étape nécessite de définir une **stratégie d'annotation** fondée sur l'expertise des biologistes et de fixer le langage de description des données. La troisième étape a trait à la mise en place d'un ensemble de règles expertes acquises en partie par **apprentissage automatique** ainsi que d'un module de raisonnement capable de mettre en oeuvre la stratégie d'annotation.

## Hiérarchie Fonctionnelle

***Données disponibles :***
• Nous disposons de 3 génomes annotés avec la plate-forme AGMIAL et utilisant une hiérarchie fonctionnelle dérivée de SubtiList.
• Le MIPS fournit 4 génomes (*B. subtilis, H. pylori,* et 2 listeria : *L. innocua* et *L. monocytogenes*) annotés avec la hiérarchie fonctionnelle FunCat.
• Nous désirons aussi inclure l'organisme modèle *E. coli* qui, lui, a été annoté avec la classification fonctionnelle de M. Riley.

***Travail à réaliser :***
• Établir des *relations* entre ces hiérarchies fonctionnelles afin de pouvoir disposer d'une annotation homogène de tous les génomes.
• *Intégrer* les 5 génomes qui n'ont pas été annotés avec la plate-forme AGMIAL: les décrire dans notre langage de description (voir ci-dessous).

## Stratégie d'annotation

La deuxième étape consiste à définir une **stratégie d'annotation**, c'est-à-dire, à identifier les différentes méthodes d'analyse bioinformatique qui fournissent des informations pertinentes sur la fonction. Ces méthodes sont de deux types:
• Celles qui établissent des *relations entre protéines* (par ex. les méthodes de recherche d'homologie, ou celles basées sur l'étude du contexte des gènes)
• Celles qui définissent des *propriétés des protéines* (le point isoélectrique, la présence d'un domaine particulier, la localisation cellulaire, la fonction, les mots-clés les caractérisant dans des collections de données biologiques, etc).

***Représentation des exemples*** :
Les résultats fournis par ces différentes méthodes permettent de décrire les protéines annotées des génomes choisis sous la forme d'un certain nombre de prédicats instanciés (atomes), qui constituent les *exemples*.

## Programmation Logique Inductive

La troisième étape consiste à exploiter les exemples ainsi définis pour développer un système de règles en utilisant des techniques d'apprentissage. Nous avons choisi de recourir à la **Programmation Logique Inductive** (PLI).

***Pourquoi ?***
• Pour l'*expressivité* : les règles sont exprimées dans un sous-ensemble de la logique des prédicats, il est donc possible d'avoir un raisonnement logique et de combiner des règles.
• D'un point de vue *applicatif* : la PLI offre un cadre formel pour l'apprentissage à partir de bases de données multi-relationnelles.

***Avantages de la PLI :***
• Les exemples donnés en entrée peuvent être *faiblement structurés.*
• Les règles apprises sont très *expressives* : elles comportent des variables qui peuvent être liées.
• Les règles obtenues sous ce format sont *compréhensibles* par les experts.

## RÉSULTATS

Forme des premières règles obtenues :

```
memsatcount(A,C) ?
+--yes: blast(A,D,E,F), interpro(A,G,H) ?
|       +--yes: [1.2.5 Transport/binding of amino-acids]
|       +--no:  [1.2.1 Transport/binding of proteins/peptides]
+--no:  [1.2.1 Transport/binding of proteins/peptides]
```

Traduction :
Si la protéine A possède C segments transmembranaires et si la protéine A est similaire à la protéine D avec une probabilité de E et un taux de ressemblance de F et A appartient à la famille de protéines G avec un score de H alors la fonction de la protéine A est 1.2.5 : transport ou liaison d'acides aminés, sinon elle est 1.2.1: transport ou liaison de protéines ou de peptides.