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