Construction ontologique à partir de séquences d'expression de champignons

Houda Fyad*, Karim Bouamrane*, Baghdad Atmani*, Claire Toffano-Nioche**, ***
*Département d'Informatique, Faculté des Sciences, Université d'Oran,
BP 1524, El M'naouer 31000 Oran, Algérie
{houdafyad82, kbouamrane, atmani.baghdad}@gmail.com
**Université de Paris-Sud XI, IGM UMR8621, Orsay, F-1405, France
***CNRS, Orsay, F-91405, France
claire.toffano-nioche@u-psud.fr

1 Introduction

Un des problèmes majeurs rencontré par le biologiste, est l'extraction et l'exploitation des données qui l'intéressent à travers les multiples ressources disponibles sur le Web. Ce problème existe en raison de la multiplicité des ressources, l'hétérogénéité et la variabilité des formats, les mises à jour inégales et la redondance des nomenclatures, etc… Une approche de fouille de données apporte une solution à notre objectif d'exploiter les données d'expression des gènes, les EST (Expressed Sequence Tags), en fonction des conditions expérimentales de l'organisme étudié. Ces EST sont exploités pour leur partie séquence mais les informations textuelles associées renseignant le protocole expérimental sont ignorées. Or, la souche de l'organisme, les conditions de culture, ou encore le stade de développement lors du séquençage, modifient l'expression et cela devrait influencer les analyses ultérieures. Ainsi, nous avons construit une ressource ontologique à partir d'un corpus composé des EST de deux champignons multicellulaires, *Neurospora crassa* et *Podospora anserina*, choisis car ils sont suffisamment proches évolutivement pour partager leur cycle de vie.

2 Principe

Nous avons choisi d'effectuer une extraction statistique des termes-clés. Le corpus est constitué des 277147 (resp. 51286) fiches d'EST de *N.crassa* (resp. *P.anserina)* issues de 22 (resp. 7) expériences issues de Genbank (NCBI). Nous avons utilisé l'outil KEA, Automatique Keyphrase Extractor (Jones et *al*, 2002), et exploité les métriques calculées pour chaque terme-clé, « TF*IDF » et « First occurrence », afin de filtrer les termes extraits qui proviennent principalement des informations associées aux EST. Ces termes représentent alors les concepts, propriétés ou valeurs de la ressource ontologique que nous avons établie ensuite manuellement.

3 Résultats

KEA a permis l'extraction automatique de 3,94 +/- 1,03 termes candidats par fiche d'EST. Pour ne sélectionner que les termes pertinents, un filtrage a été réalisé à partir des valeurs « TF*IDF » comprises entre 0.00000264 et 0.17922504, et entre 0.00000264 et 0.17750744