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

pour « First occurrence ». Certains termes ont ensuite été exclus car i) non significatifs pour le domaine (ex : « micromol », « Mm », « stage », « meiose ») ou ii) correspondant à un niveau de détail trop profond et dont les éléments seront représentés par la hiérarchie de l'ontologie (ex : « Vogel without nitrate », « Vogel without glucose »). Cette méthode a permis l'extraction de 108 termes auxquels nous avons ajoutés d'autres termes obtenus en suivant la même approche mais à partir de matériaux spécialisés (Raju, 2009 ; Gaad, 2005). L'ensemble des termes a été reparti et organisé en 4 ontologies parallèles à l'image de la stratégie eVOC (Kelso *et al*, 2003) : 27 termes pour les étapes du cycle cellulaire, 17 pour les types cellulaires, 13 pour les caractéristiques des souches de champignons, et 51 termes pour les conditions de culture. Ainsi, la description d'un EST recourt à ces ontologies parallèles et complémentaires et la liaison entre elles se réalise à l'usage, lors de la caractérisation d'un EST par la liste des termes issue de chacune des ontologies.

4 Conclusion

Nous avons présenté une contribution au domaine ontologique par la construction d'un modèle de connaissance à partir d'une extraction statistique des termes issus d'un corpus composé de données biologiques. Après organisation de la ressource terminologique, la représentation des connaissances qui en résulte est modulaire car répartie en plusieurs ontologies complémentaires, ce qui permet flexibilité et facilité des mises à jour. La spécificité du domaine de connaissance établi réside dans le choix des champignons d'étude mais aussi, dans la gestion du contexte des expériences des données d'expression dont la sémantique associée est capturée par la méthodologie que nous avons définie.

Références

Gaad, M.V (2005). Genomic conflicts in Podospora anserina. PhD, Univ. Wageningen.

Jones, S. and Paynter, G.W. (2002) *Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications*. *Journal of the American Society for Information Science and Technology*.

Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien, S., Smedley, D., Otgaar, D., Greyling, G., Victor Jongeneel, C., McCarthy, M.I., Hide, T., Hide, W (2003). eVOC: A Controlled Vocabulary for Unifying Gene Expression Data. *Journal of Genome Research*. 13:1223–1227.

Raju, N.B (2009). Neurospora as a model fungus for studies in cytogenetics and sexual biology at Stanford. *Journal of Biosci*. 139–142.

Summary. This paper describes the construction of four ontology's based on a statistical extraction of terms from a corpus of biological data. It aims at characterize the experimental context (developmental state, culture conditions...) used to get expressed sequence of two fungi.