Projet ABC : Analyse et base de connaissances

Claude Pasquier - Karim Jevardat de Fombelle - Fabrice Girardot - Richard Christen

UMR6543 Institut de signalisation, biologie du développement et cancer - Centre de Biochimie, Parc Valrose 06108 NICE

Introduction

La technologie des puces à ADN permet de mesurer des milliers de variables simultanément et de comparer leurs valeurs sur des centaines d'expériences. Cette nouvelle manière de procéder nécessite d'analyser une quantité d'information qui est bien supérieure à ce qu'un expert humain "normal" est capable d'appréhender. Des systèmes de classification sont couramment utilisés pour, par exemple, mettre en évidence des regroupements correspondant à des phénomènes ou des processus biologiques (corégulation ...). Mais, comme il existe quantité de méthodes permettant de calculer des partitionnements qui souvent ne reposent sur aucun fondement biologique, au final, on obtient des données, certes plus ordonnées, mais toujours aussi nombreuses. Le biologiste doit alors utiliser sa propre expertise combinée aux informations stockées dans les bases de données publiques ou disponibles dans la littérature pour interpréter les données.

Ce processus d'interprétation des données visant à l'élaboration de nouvelles connaissances ne peut pas, en l'était actuel des techniques informatiques, être complètement automatisé. Cependant, l'expert peut être grandement assisté dans sa tâche. Nous proposons d'utiliser une ontologie des connaissances biologiques pour automatiser le passage de la mesure ponctuelle au concept. Une ontologie fournit un modèle conceptuel qui peut être utilisé comme un réseau sémantique permettant de guider les tâches de stockage, d'accès aux données et d'analyse. Schématiquement, c'est une modélisation particulière des données dans laquelle tous les concepts utilisés sont eux même définis. Le but est de construire un modèle fermé sur lequel il soit possible de raisonner de manière autonome.

ABC : un système informatique intégré combinant deux fonctionnalités essentielles

Annotation des données issues d'une classification avec des informations biologiques provenant d'une base de connaissance

ABC fournit trois types d'aides à l'interprétation des résultats :

I. Exploration de Gene Ontology
- Quels transcrits sont associés à un ou plusieurs concepts biologiques ?
- Quels sont les termes biologiques caractérisant les transcrits ?

II. Exploration globale du partitionnement
- Les transcrits associés à une fonction donnée sont ils regroupés dans le partitionnement ?
- Des corrélations entre plusieurs fonctions sont elles visibles ?
- Existe-t-il des corrélations entre les niveaux d'expression de certains gènes et leur localisation sur le génome ?

III. Focalisation sur l'étude d'un groupe de gènes
- Peut-on nommer des groupements de gènes déjà connus ?
- Est-il possible d'identifier des co-expressions entre transcrits qui ne reflètent pas la connaissance biologique ?

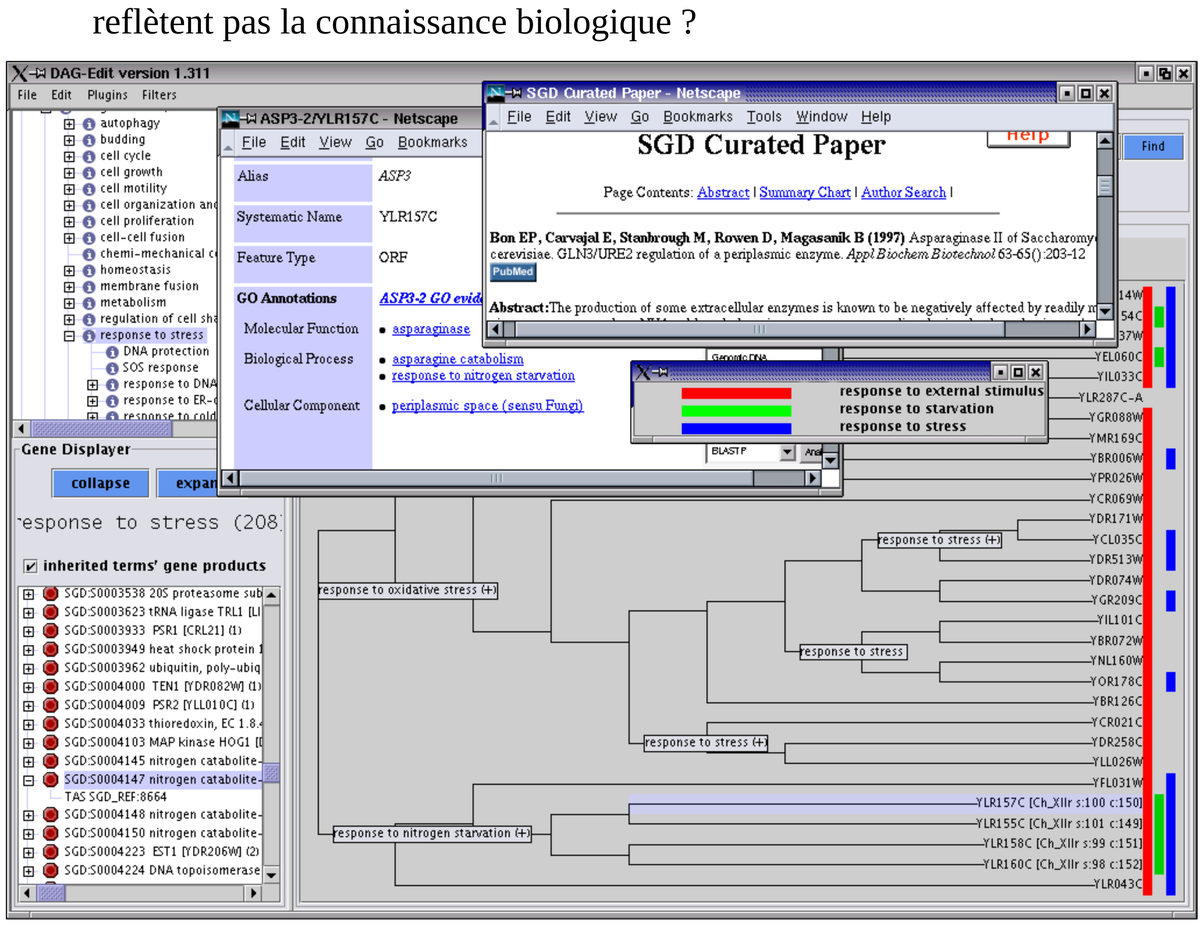

Figure 1 : Analyse avec ABC des données extraites de « Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14863-8. » Le sous arbre affiché regroupe 30 transcrits majoritairement impliqués dans le processus biologique de réponse au stress (cf. sélection bleue affichée à droite de la classification, nommage de la racine par le terme de gene ontology (GO) « response to oxidative stress » et visualisation de l'arborescence GO en haut à gauche). Les 208 transcrits de *Saccharomyces cerevisiae* associés au terme « response to stress » ou à l'un de ses sous-termes sont affichés en bas à gauche. La fiche du gène ASP3 sélectionné (identifiant S0004147 ou YMR157C) ainsi que le résumé de l'article justifiant son rattachement au terme GO sont facilement accessibles depuis l'interface. Les 4 copies du gène ASP3 (YLR157C, YLR155C, YLR158C et YLR160C), co-exprimés dans les expériences effectuées mais aussi co-localisés sur le chromosome XII sont mis en évidence dans l'arbre de classification.

Construction et enrichissement d'une base de connaissances en fonction des résultats d'expériences

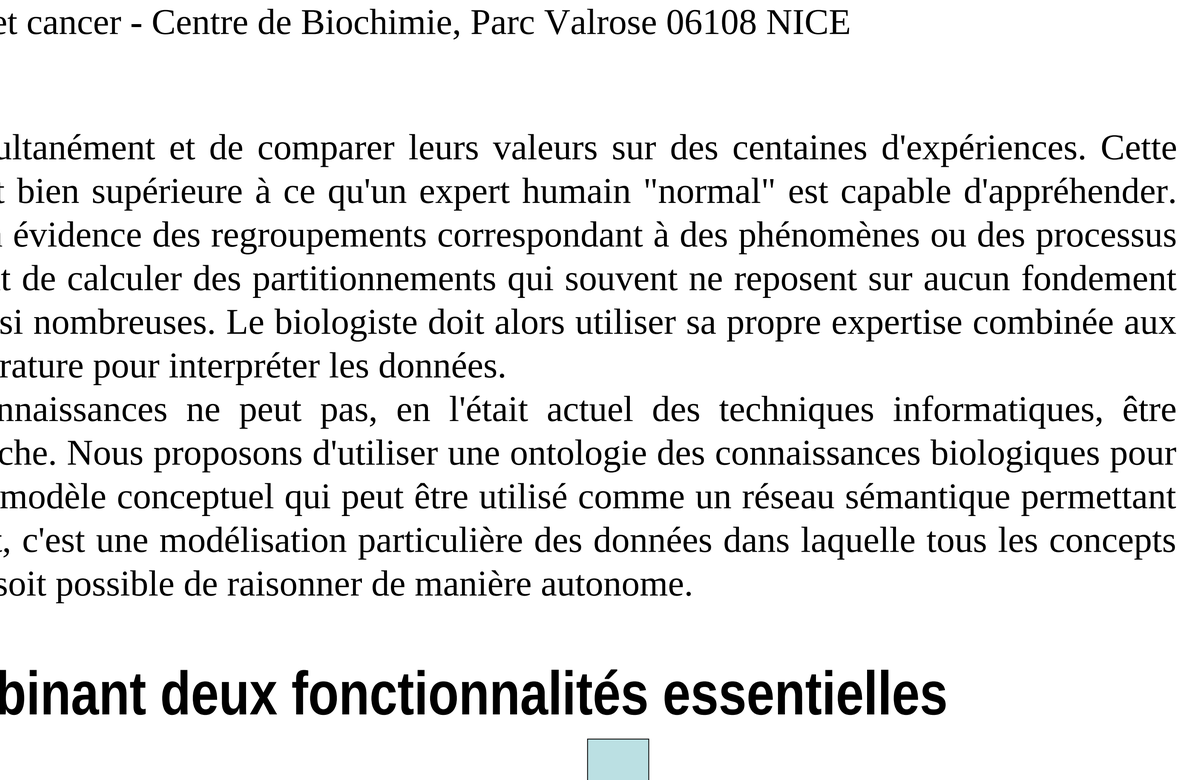

Figure 2 : Exemple de relations pouvant être modélisées dans une base de connaissances. La partie grisée représente les informations et relations actuellement représentées dans Gene Ontology.

Dans Gene Ontology, sont uniquement représentées des informations factuelles à l'aide de relations statiques du type « is-a » et « part-of » et cela est insuffisant. Il manque en particulier la représentation des réseaux de régulation, la localisation des transcrits dans les tissus ou la modélisation des interactions entre gènes, entre cellules ou entre organes dont la compréhension est essentielle. Modéliser des informations complexes comme celles-ci est une tâche difficile, non seulement du point de vue biologique, mais aussi informatique. Il faut dépasser la représentation arborescente des données qui est trop limitée par une modélisation de la connaissance sous forme de concepts et de relations entre ces concepts. Il faut pouvoir valoriser les relations pour représenter par exemple le fait que les interactions entre transcrits ont des niveaux d'intensité différents. Il faut également tenir compte d'une certaine variabilité ou incertitude dans la modélisation. Par rapport à Gene Ontology, nous généraliserons la représentation arborescente pour modéliser la connaissance sous forme de graphe. L'objectif final serait d'arriver à une modélisation très détaillée des connaissances biologiques comme dans la figure 2. Cette représentation de la connaissance comportera une partie publique, partagée par l'ensemble des scientifiques sur laquelle se greffera une connaissance locale élaborée par le chercheur manipulant l'outil. Les informations ajoutées localement pourront, soit être diffusées pour être intégrées aux bases de connaissances publiques soit partagées avec une communauté plus limitée de chercheurs ayant par exemple le même centre d'intérêt.

Développements futurs et perspectives

L'objectif poursuivi est de progresser vers une représentation des données cohérente et unifiée en s'attachant à donner du sens à toutes les données, aussi bien pour un acteur humain que pour un système informatique. Il sera nécessaire d'élever le niveau de représentation des données en utilisant des méta données permettant de modéliser des informations se rapportant aux données elles mêmes. Cette représentation de haut niveau, bien définie sémantiquement, aura comme premiers avantages de faciliter l'échange d'information entre les chercheurs de laboratoires différents, de contribuer aux bases de données publiques en leur proposant des informations correctement balisées et même d'établir une base d'information partageable sur le web qui pourra faire référence. Il sera également possible de faire des recherches sur un ensemble d'expériences passées de manière à induire de nouvelles connaissances qui seront, soit proposées aux chercheurs, soit mémorisées dans la base. L'élaboration automatique de nouvelles connaissances ("data mining") est un domaine en pleine expansion qui constitue un autre axe de recherche à long terme. On s'attachera par exemple à déduire des propriétés nouvelles sur les sondes à partir de résultat d'un ensemble d'expériences et ainsi, peut être, proposer, en fonction de l'expérimentation prévue, les sondes les plus appropriées (celles qui sont les plus spécifiques) à placer sur une puce. Alternativement, la même approche devrait permettre de rechercher des réseaux métaboliques encore mal connus, mettant par exemple en jeu des fonctions où interviennent uniquement des ARN non codants. Un autre axe de recherche, également à long terme, est la réutilisation de ces ontologies et des outils développés pour l'analyse de résultats d'expériences autres que celles effectuées avec les puces à ADN (protéomique) ou avec des puces à ADN dans d'autres domaines que la génomique (bactériologie clinique, environnement).