Extraction de processus fonctionnels en génétique des microbes à partir de résumés MEDLINE

Alain Lelu*, Philippe Bessières *
Alain Zasadzinski **, Dominique Besagni **

* INRA / MIG, Domaine de Vilvert, 78352 Jouy en Josas Cedex
alain.lelu@jouy.inra.fr, philb@diamant.jouy.inra.fr
http://www-mig.jouy.inra.fr/mig/index.html

** INIST / URI 2 Allée du Parc de Brabois, 54514 Vandoeuvre lès Nancy Cedex
Zasadzin@inist.fr, Besagni@inist.fr
http:// www.inist.fr/uri/accueil.htm

Résumé. Après l'ère du décodage des génomes, les biologistes sont de plus en plus confrontés à l'intégration de myriades de connaissances parcellaires, stockées majoritairement sous forme textuelle. Nous montrons, à travers un exemple concret, que la conjonction de deux chaînes de traitement faisant appel de façon modérée à l'expertise humaine offre au biologiste une aide utile pour parcourir cette littérature, à partir d'une structuration sans *a priori* de son corpus ; il s'agit ici de résumés Medline indexés par les gènes et protéines qu'ils citent, et que l'algorithme structure (sans superviseur) en principales voies métaboliques et de régulation présentes dans le corpus choisi. 1) Une chaîne d'indexation par les noms de gènes et protéines inclut un expert pour valider, 2) Un environnement interactif de clustering thématique attribue des valeurs graduées de centralité dans chaque thème aux résumés comme aux noms, comme à toute autre variable illustrative (autres termes bio., MeSH, …).

1 Introduction : la biologie devient intégrative.

On assiste depuis une dizaine d'années à un changement majeur en biologie, où les techniques d'analyse de masse (séquençage du génome, puce à ADN, …) ont permis une inversion complète de la perspective :

. On part de plus en plus du génome pour aller vers le phénotype (observable). La séquence du génome peut même, dans certains cas, être la première donnée acquise sur une espèce, alors que des centaines de génomes sont séquencés ou en passe de l'être.

. L'autre caractéristique majeure de cette révolution est son caractère encyclopédique. A la limite, on embrasse simultanément tous les objets d'une même collection (tous les gènes, toutes les protéines, l'ensemble des réseaux métaboliques, …) et toutes les échelles d'organisation. *A minima*, l'étude d'une fonction ou d'une régulation particulière d'une espèce donnée exige de la situer dans tout ce que l'on en sait chez les autres espèces, et dans le contexte des autres fonctions de la cellule. Un effet « boule de neige » est en route… tant que le chercheur arrive à maîtriser l'information antérieurement produite.

Le but ultime de la biologie nouvelle, dite intégrative, est d'expliquer comment le génome spécifie les propriétés des organismes (le phénotype) en explorant et décrivant tous