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

les niveaux intermédiaires d'organisation : les processus cellulaires, les tissus, les organes, les processus physiologiques… La marche vers la biologie intégrative se caractérise par la production d'une masse considérable de données et résultats hétérogènes, dont beaucoup alimentent des bases de données factuelles structurées et normalisées (séquences nucléiques et protéiques, structures 3D des protéines…). Mais la majeure partie se trouve sous forme textuelle, faiblement structurée et normalisée, dans les bases de résumés bibliographiques, comme Medline ou Pascal, ou les articles eux-mêmes, lesquels sont de plus en plus accessibles en ligne, librement ou sur abonnement.

On peut identifier trois niveaux au cours de l'annotation du génome d'un organisme (Stein 2001) : le premier est l'annotation de la séquence nucléique avec pour but principal de détecter les gènes et les signaux associés aux gènes (sites de fixation des ribosomes, promoteurs, terminateurs, etc). Le deuxième niveau s'intéresse en priorité aux protéines, produits des précédents gènes. Le but de ce niveau est de fournir une première description de la fonction des protéines (il s'agit ici de la fonction moléculaire, dans la plupart des cas). Dans ces deux niveaux on s'attache à décrire les objets fondamentaux de la biologie. Cette description fournit une vue essentiellement statique du génome. De façon à étudier l'influence du génome sur les propriétés biologiques générales de l'organisme il est impératif de disposer d'une vision dynamique : comment les gènes et les protéines interagissent-ils les uns avec les autres pour former des « modules fonctionnels » accomplissant une tâche particulière dans la cellule (voies métaboliques, cascades de signalisation, processus de régulation, organisation structurale de la cellule, etc.) ? On peut supposer que ces modules fonctionnels établissent aussi des interactions entre eux pour former des « modules supra-fonctionnels ». Ce processus pourrait se poursuivre sur plusieurs niveaux, générant une hiérarchie complexe qu'il reste à découvrir. Quoi qu'il en soit, c'est l'ensemble de ces réseaux interconnectés, dans leurs dimensions à la fois temporelle et spatiale, qui constitue l'objet d'étude du troisième niveau : *l'annotation des processus*.

À l'heure actuelle beaucoup des techniques bioinformatiques s'intéressent à des questions relatives aux deux premiers niveaux. Tout en continuant à améliorer les techniques existantes, ou en développer de nouvelles, on peut penser que les centres d'intérêt de la communauté bioinformatique vont se déplacer de plus en plus vers la résolution de problèmes posés par le troisième niveau.

Les processus biologiques sont d'une complexité et d'une imbrication extrême : pour ne citer qu'un exemple, il est établi que beaucoup de gènes et de produits de ces gènes sont polyfonctionnels (on pourrait presque dire : polysémiques, comme la majorité des mots de la langue…). Leur description relève pour l'instant de phrases en langage naturel, et échappe largement aux formalismes existants (matrices, graphes, hypergraphes, réseaux de Pétri…), en attendant les progrès, qui tardent, dans la formalisation mathématique des systèmes complexes. Il est donc clair que pour un bon nombre d'années encore, la source majeure de collecte massive de résultats biologiques partiels sera l'exploitation de la littérature scientifique, en particulier des résumés d'auteur répertoriés par Medline ou Pascal.

2 Problématique et état de l'art

Le courant dominant en extraction de connaissances textuelles tente de convertir directement des phrases isolées, écrites dans la langue naturelle, en informations factuelles structurées, par exemple en repérant deux noms de gènes séparés par un verbe d'action tel

« inhibits » ou « activates », approche que (Blaschke 1999) (Jenssen et al. 2001) ont initiée à grande échelle – pour une synthèse, cf. (Yandel, Majoros 2002). Un premier niveau de difficulté réside dans l'analyse correcte des constructions syntaxiques souvent complexes présentes dans ces textes rédigés sous forte contrainte de concision ; il n'est pas rare non plus que des chercheurs étrangers maîtrisent mal l'anglais…

Un autre niveau de difficultés est lié au caractère contextuel de la langue : le sens des mots d'un texte, fût-il scientifique, est spécifié par de multiples contextes ; dans notre cas : revue qui édite l'article, état des connaissances du moment, courant scientifique dans lequel il s'inscrit… A l'échelle du résumé, le sens de chaque phrase d'un texte est spécifié par le contexte créé par les phrases précédentes. Par exemple, pour exprimer une simple interaction génique directe, telle que *la protéine A produite par le gène a, active le gène b qui produit la protéine B*, de nombreuses formulations sont possibles. Sur le plan biologique, il est fait par exemple référence aux conditions expérimentales (mutation, stress, médicament), aux observations (variation du niveau de protéine, phénotype), à la temporalité (phase de la sporulation) sans que l'interaction soit clairement et explicitement mentionnée. Pour nous cette approche se situe clairement dans une perspective à long terme : comment pourrait-elle échapper à la prise en compte, pour interpréter une phrase donnée, de l'empilement des contextes que spécifient successivement la revue, l'auteur (socialement inséré dans un réseau de co-auteurs et co-citations), le titre, et les phrases précédent la phrase courante ?

D'autres approches prennent acte de l'état de l'art courant en matière de traitement de la langue naturelle et construisent à partir de la seule extraction, dans chaque résumé, de termes pertinents et d'entités nommées (gènes et protéines, espèces étudiées, auteurs…) ; cette extraction est un objectif raisonnable, rempli par les outils actuels avec une bonne précision pour qui se contente d'une couverture non exhaustive, en général suffisante pour appliquer des traitements statistiques sur un matériau textuel qu'on trouve en abondance. Le choix du résumé comme unité sémantique élémentaire paraît plus naturel que celui de la phrase, l'auteur délivrant généralement dans ce résumé un message unique et cohérent, de niveau de complexité élevé, scientifiquement original, donc distinct par définition de celui des autres articles. Dans cette voie on peut citer (Iliopoulos et al. 2001) (Wilbur 2002) qui extraient, par analyse multidimensionnelle non supervisée, des concepts (ou « thèmes ») à partir d'une sélection de résumés Medline soumis à une indexation générale, non typée, par l'ensemble des substantifs, verbes et expressions à caractère biologique qu'ils contiennent.

Notre hypothèse dans le présent travail est que les relations créées par les cooccurrences des seuls noms de gènes et protéines – donc par indexation *typée* de l'ensemble des résumés sélectionnés - donnent lieu à une structuration en thèmes beaucoup plus stable, incontestable, « objective », évaluable, que celle issue de l'ensemble des termes : pour un résumé donné l'identification des noms d'entités « gènes et protéines » peut donner lieu à un consensus d'experts plus fort que celle des autres termes biologiquement pertinents, particulièrement les multitermes, dont l'appréciation du degré de figement peut être très subjective. Une structuration en thèmes des seuls noms de gènes et protéines correspond nécessairement à des réalités – de laboratoire ou des espèces étudiées – bien identifiables. Ce qui n'empêche pas de « projeter » les autres termes sur la structure obtenue, à titre de variables illustratives, passives, pour faciliter ou confirmer l'interprétation experte des thèmes.

Un pas dans cette direction se trouve dans (Wilkinson, Huberman 2002) qui utilisent une méthode fondée sur les graphes pour analyser les relations entre les gènes et en isoler des « communautés », méthode qu'ils complètent de façon quelque peu *ad hoc* pour doter chaque gène d'un score d'appartenance nuancé à sa communauté propre, mais aussi aux autres –

caractéristique tout à fait louable pour nous qui l'avons implantée de longue date dans notre algorithme des K-Means Axiales. Mais leur méthode est muette sur le côté « dual » des thèmes : un résumé peut se trouver à l'articulation de plusieurs « communautés », ou typique d'une seule ; et elle ne permet pas d'introduire des variables illustratives.

Outre notre choix de principe d'une indexation typée, notre deuxième principe est de considérer ce type de techniques non comme une extraction automatique de réseaux d'interactions géniques, mais comme une aide informatisée au biologiste[1] attelé à cette tâche, par la fourniture 1) de « fils conducteurs », d'angles d'attaque initiaux pour aborder une masse d'informations sans structure à l'origine, 2) de la possibilité à tout moment, pour tout titre de résumé qu'il désire, de revenir au texte, voire à l'article lui-même, pour approfondissement, 3) de faire de même pour tout nom de gène et protéine, par un lien immédiat vers les bases de données factuelles et bibliographiques.

3 Choix du corpus

Notre ambition est d'isoler des contextes de recherche fins au sein d'un domaine bien délimité de la génomique fonctionnelle, qui concerne les mécanismes de la transcription chez Bacillus. Les travaux dans ce domaine décrivent, entre autres, une grande quantité d'interactions entre gènes et/ou protéines hors de portée de la mémoire d'un seul biologiste. Notre corpus est constitué des 2300 résumés Medline rassemblés par la requête : Bacillus subtilis *and* transcription *and* (transcription *or* promoter *or* sigma) pour les années antérieures à 2002. Ce corpus nous sert de banc d'essai pour nos travaux d'extraction de connaissances. Du fait de choix faits dans le passé, il ne comporte pas les titres d'articles.

4 Indexation typée des résumés

Insatisfaits de l'indexation de base fournie par Medline (termes MESH ou texte intégral) et fidèle à notre principe de « présence d'un expert dans la boucle des traitements informatiques » nous avons décidé de mettre à profit les compétences et l'expérience accumulées par l'équipe INIST/URI en matière de chaîne d'extraction terminologique et d'indexation automatique humainement validée (Royauté et al. 2001) (Jacquemin et al. 2002). La possibilité ici offerte de typer les termes est fondamentale : elle permet d'isoler les noms de gènes et protéines, pour leur appliquer ensuite nos traitements spécifiques.

4.1 Les ressources terminologiques :

Notre ressource terminologique de base est le vocabulaire multidisciplinaire de l'INIST (le MX), mis au point et alimenté par les documentalistes qui indexent la littérature scientifique pour la base de données PASCAL. Cette ressource est partiellement organisée sous forme de thésaurus et intègre des vocabulaires structurés. En particulier cette ressource intègre la nomenclature internationale des enzymes selon le « Nomenclature Committee of the International Union of Biochemistry and Molecular Biology » (IUMBM) et la classification des bactéries suivant les recommandations de la « Society for General Microbiology » dont les mises à jour paraissent dans la revue « International Journal of Systematic and Evolutionary Microbiology ». A cette ressource, nous avons intégré les

[1] De par la nature de l'activité scientifique, il est capable de contextualiser, relativiser, situer les informations fournies par l'auteur dans toutes leurs nuances, et juger leur degré de fiabilité.

candidats termes proposés par les documentalistes de l'INIST spécialistes du domaine de la microbiologie et de l'agronomie dont un certain nombre sont régulièrement intégrés au MX lors de mises à jour annuelles. Au total l'ensemble du vocabulaire est constitué de 108 344 termes répartis en 90 764 préférentiels ou concepts et 17 580 synonymes (dont 3239 candidats préférentiels et 312 candidats synonymes).

4.2 Préparation des ressources terminologiques

Après mise au format SGML, la ressource a été divisée en deux parties : 1) concepts susceptibles de subir les variations linguistiques du langage naturel: 56 389 termes répartis en 55 163 préférentiels et 1226 synonymes, 2) concepts concernant essentiellement les composés chimiques, biochimiques, et la systématique : 51 955 termes répartis en 35 601 préférentiels et 16 354 synonymes. Les synonymes sont inclus dans la ressource terminologique. Cette ressource INIST ne contient pas d'acronymes de noms de gènes et protéines, extraits des textes par expressions régulières (voir § 4.3, section 3).

4.3 Outils d'indexation :

Trois procédures complémentaires d'indexation, mis au point à l'URI ont été utilisées. Les deux premières réalisent une extraction des termes à l'aide des ressources terminologiques, nous parlons alors d'indexation contrôlée. La troisième réalise une extraction des termes sur la base d'expressions régulières. Celles-ci permettent de capter des termes qui ne sont pas dans les ressources terminologiques et de compléter l'indexation des gènes, des protéines et de la systématique dont la syntaxe est relativement bien normalisée dans le domaine de la Microbiologie. De plus les pluri-termes issus de néologismes spécifiques du domaine tel que les opérons ou les fusions de gènes avec des marqueurs sont extraits sur la base de la présence de séparateurs dans leur construction morphologique.

1 - La plate forme d'ingénierie linguistique ILC (Information, Langage et Connaissance) de l'INIST/URI a été utilisée pour l'indexation des concepts subissant les variations du langage naturel (Royauté et al. 2001) (Royauté et al. 2004).

2 - Un programme écrit en PERL, appelé IRC3 (Indexation par Recherche et Comparaison de Chaînes de Caractères) mis au point à l'URI pour l'indexation des concepts ne subissant pas les variations du langage naturel, souvent avec des caractères de ponctuation pouvant être interprétés comme des éléments de la phrase par les outils linguistiques d'ILC. Ici, il s'agit essentiellement des composés chimiques et biochimiques (Royauté et al. 2004).

3 - Un programme de recherche de motifs spécifiques par expressions régulières, appelé RGM (Recherche de Gènes en Microbiologie) a été mis au point à l'URI spécifiquement pour les acronymes de gènes, protéines et autres entités génétiques, dont la structure lexicale est normalisée en microbiologie et respectée par les auteurs dans la plupart des publications scientifiques. Ce programme capte également les séquences textuelles séparées par des tirets comme les fusion de gènes ou les opérons par exemple, ainsi que le terme situé immédiatement derrière l'acronyme, permettant à l'expert chargé de la validation des termes de les désambiguer en fonction du contexte sémantique. Voici quelques exemples de séquences textuelles captées par ce programme pour les gènes *spoIIR* et *sigX*, et pour les protéines correspondantes (*spoIIR expression*; *spoIIR gene*; *SpoIIR protein*; *spoIIR-lacZ fusion*; *sigX gene*; *SigX protein*; *sigX-containing DNA* ; *sigX-ypuN control*). Les séquences textuelles précédents des termes laissant présager une entité génétique susceptible de ne pas respecter la règle utilisée sont également captées. Ainsi les termes spécifiques d'entités génétiques telles que les opérons, gènes, protéines, mutations, mRNAs, transcripts …

4.4 Procédure d'indexation

Après extraction, les termes sont répartis en trois catégories : 1) les gènes et les protéines, 2) la systématique, 3) le reste du vocabulaire. Ceci répond à une procédure conçue et proposée à partir de l'expérience réalisée dans le projet Inter EPST 2000-2002 associant l'URI-INIST et l'Institut Gustave Roussy sur la génomique du cancer de la thyroide (Royauté et al. 2004). L'ensemble du vocabulaire d'indexation est conservé, mais à ce stade du travail, les classifications n'ont été réalisées que sur les gènes et les protéines pour caractériser les principaux groupes fonctionnels. Cette étape de structuration sémantique du vocabulaire est réalisée automatiquement dans un premier temps à partir des ressources terminologiques de l'INIST pour les protéines bien identifiées telles que les enzymes. Ensuite, une ventilation manuelle est réalisée par un expert pour les protéines non étiquetées dans la ressource telles que « DNA binding protein » ou « Decay accelerating factor ». Les expressions extraites par les programmes RGM sont directement ventilées dans la catégorie sémantique des gènes et protéines. Ensuite l'expert du domaine valide les termes captés et gère les renvois de synonymie non répertoriés dans les ressources terminologiques. En retour, cette procédure permet d'enrichir les ressources terminologiques. Enfin un programme de dédoublonnage permet de nettoyer les index, notice par notice, selon deux principes : 1) les doublons stricts sont éliminés, 2) les termes (uni- et pluri-) contenus dans une chaîne de caractères plus longue sont éliminés. Ceci permet de conserver les concepts les plus spécifiques.

A noter qu'on conserve dans la structure SGML des documents indexés la trace de l'origine de la séquence textuelle captée (ressource terminologique, outil d'extraction).

Après mise en ligne des résumés indexés sous l'interface VISA de l'INIST (cf. < http://uri.inist.fr >) avec surlignage des termes retenus (cf. exemple encadré ci-après), nous avons constaté que l'indexation était précise, mais pas toujours exhaustive : des améliorations sont possibles au niveau des expressions régulières, ou par l'apport de ressources terminologiques supplémentaires (base Swissprot…), et par l'utilisation d'outils linguistiques réalisant des extractions à partir de patrons morpho-syntaxiques. C'est désormais l'objectif recherché dans la poursuite de ce travail.

Endospores of the Gram-positive bacterium Bacillus subtilis are encased in a tough protein shell known as the coat. The coat is composed of a dozen or more different structural proteins. We report the identification of and studies on the regulation of promoters governing the expression of coat protein (cot) genes designated B to E encoding polypeptides of 59, 12, 11 and 24 kDa, respectively. We show that transcription of genes B, C and D is governed by single promoters and that transcription of gene E is governed by tandem promoters designated P1 and P2. In extension of recent work on the transcription of cot gene A and the mother-cell regulatory genes gerE, sigK and spoIIID, we show that genes involved in coat formation are turned on in a regulatory cascade of at least four co-ordinately controlled gene sets. The cascade consists of: cotE as transcribed from its P1 promoter and spoIIID, which are turned on during hours three to four of sporulation; cotE as transcribed from its P2 romoter and sigK, which are turned on during hour five by the appearance of the product (a small DNA-binding protein) of spoIIID; cotA, cotD and gerE, which are turned on during hours five to six by the appearance of the product (sigma factor sigma K) of sigK; and cotB and cotC, which are turned on during hour seven by the appearance of the product (an inferred DNA-binding protein) of gerE. The cascade is hierarchical in that the first three gene sets each contain the regulatory gene that turns on the expression of the next gene set in the pathway. We also show that the level of expression of a member (cotC) of the terminal class of gene expression is strongly influenced by medium and that this effect directly or indirectly depends on the product of sporulation gene spoIV A.

TAB 1 – *Noms de gènes et protéines extraits dans le résumé PMID: 1 691 789 de Medline.*

5 Extraction de thèmes homogènes à partir des cooccurrences des noms de gènes et protéines.

La méthode de classification automatique K-Means Axiales (Lelu 1994) dont nous présentons ici une application au domaine biologique a été conçue au confluent de trois ordres de préoccupations, pour viser des résultats sûrs, stables et nuancés :

. Opérer dans un espace de données pourvu de la même propriété d'équivalence distributionnelle que celui de l'Analyse Factorielle des Correspondances – AFC - (Benzécri et al. 1981), garante d'une bonne stabilité de l'analyse au regard des fusions / éclatements de descripteurs dont les significations sont voisines.

. Ne pas classer en tout ou rien, mais doter chaque individu classé d'un indicateur de centralité dans sa classe - tout autant que dans les classes voisines. Comme les descripteurs sont également dotés de tels indices de centralité, notre méthode s'apparente à des travaux antérieurs (analyses factorielles obliques : psychologie et géologie) ou postérieurs (méthodes de mélanges probabilistes à base d'algorithme EM – pour une revue, cf. (Buntine 2002)).

. Permettre la projection de points-variables ou points-documents en éléments supplémentaires, à titre illustratif, une fois l'analyse effectuée.

5.1 Distance et cosinus distributionnels

Toute méthode d'analyse des données est caractérisée à la base par trois choix : 1) une transformation opérée sur les vecteurs-données bruts, 2) une métrique, ou pondération des dimensions dans lesquels ces vecteurs sont définis, 3) une pondération de ces vecteurs. Sur le nuage de points ainsi défini, de nombreuses techniques de synthèse d'information et réduction des dimensions peuvent être appliquées : classification ascendante ou descendante hiérarchique, classification à centres mobiles, décomposition aux valeurs singulières, incluant toutes les méthodes d'analyse factorielle qui en sont les variantes. Notre méthode de K-Means Axiales n'échappe pas à la règle ; voyons tout d'abord la transformation des données et la métrique choisies, qui définissent notre distance :

Une lignée ancienne de travaux (Matusita 1955) (Escofier 1978) (Domengès et Volle 1979) (Fichet et Gbegan 1985) s'est intéressée à ce que ces derniers auteurs appellent *distance distributionnelle*, qu'on explicitera ici dans notre cadre applicatif d'analyse d'un ensemble de textes décrits par les fréquences de leurs mots :

La distance distributionnelle entre 2 textes est la distance euclidienne, classique (dimensions équipondérées), entre les 2 points t_1 et t_2, de coordonnées les vecteurs $\mathbf{z}_{t1}$ et $\mathbf{z}_{t2}$, situés sur l'hypersphère unité dans l'espace des I mots et représentant chacun une unité textuelle, définis par la transformation suivante sur les données :

$\mathbf{z}_{t1} : \{\sqrt{}(x_{it1}/ x_{.t1}) \}$; $\mathbf{z}_{t2} : \{\sqrt{}(x_{it2}/ x_{.t2}) \}$

où l'accolade désigne l'ensemble des composantes numériques d'un vecteur, x_{it} désigne la fréquence du mot N° i dans le document N° t, et $x_{.t}$ le nombre total de mots du document t

La distance distributionnelle Dd(t1, t2) entre les textes t1 et t2 est donc :

$Dd(t_1, t_2) = \| \mathbf{z}_{t1} - \mathbf{z}_{t2} \|$

où $\| \mathbf{z} \|$ désigne la norme du vecteur $\mathbf{z}$.

Cette distance est la longueur de la corde correspondant à l'angle ($\mathbf{z}_{t1}$, $\mathbf{z}_{t2}$) - égale au plus à 2 quand ces 2 vecteurs sont opposés, égale à $\sqrt{2}$ quand ils sont orthogonaux. Cette distance semble triviale et arbitraire en apparence (pourquoi cette normalisation insolite plutôt que la normalisation classique $\{x_{it}/ \|\mathbf{x}_t\| \}$?) , mais elle jouit de propriétés remarquables :

. Elle est liée à certaines mesures de gain d'information (Renyi 1966).

. Elle est rapide à calculer dans le cas des données textuelles (vecteurs $\mathbf{x}_t$ très « creux »).

. *Last but not least* Escofier et Volle ont montré qu'elle satisfaisait à la propriété d'équivalence distributionnelle caractéristique de la distance du khi-deux utilisée en AFC : si on fusionne deux descripteurs de mêmes profils relatifs, les distances entre les unités textuelles sont inchangées. En d'autres termes, cette propriété assure la stabilité du système des distances entre textes au regard de l'éclatement ou du regroupement de mots de sens voisins, dont les textes qu'ils indexent se répartissent de façon un tant soit peu équivalente.

5.2 Décomposition aux valeurs singulières dans l'espace distributionnel

Si l'on transforme le tableau (textes × mots) des données brutes comme suit :

. Coordonnées : - des vecteurs-colonnes $\mathbf{x}_t : \{ x_{it} \} \rightarrow \mathbf{y}_t : \{\sqrt{} x_{it} \}$

- des vecteurs-lignes $\mathbf{x}_i : \{ x_{it} \} \rightarrow \mathbf{y}_i : \{ \sqrt{} x_{it} \}$

. Poids de ces vecteurs : unité

. Métrique : euclidienne standard (équipondération des coordonnées)

les cosinus entre vecteurs-colonnes $\mathbf{y}_t$ (resp. entre vecteurs-ligne $\mathbf{y}_i$) sont liés à la notion de distance distributionnelle Dd par la relation :

$Dd(t_1,t_2)^2 = 2(1-\cos(t_1,t_2))$ [resp. $Dd(i_1,i_2)^2 = 2(1-\cos(i_1,i_2))$]

Si on calcule les K directions propres $\mathbf{U} = \{ \mathbf{u}^{(k)} \}$ (resp $\mathbf{W} = \{ \mathbf{w}^{(k)} \}$) du nuage des points-colonnes $\mathbf{y}_t$ (resp, des points-lignes $\mathbf{y}_i$) défini plus haut, au moyen de la décomposition aux valeurs singulières[2] du tableau des racines carrées, on démontre que les cosinus ci-après se déduisent des directions propres :

$\text{Cos}(\mathbf{y}_t, \mathbf{u}^{(k)}) = w_t^{(k)} \sqrt{(\lambda^{(k)} / x_{.t})}$ (1) $\text{Cos}(\mathbf{y}_i, \mathbf{w}^{(k)}) = u_i^{(k)} \sqrt{(\lambda^{(k)} / x_{i.})}$ (2)

Ces cosinus peuvent être considérés comme les facteurs d'un cas particulier et simple d'analyse factorielle sphérique, pour reprendre la terminologie de M. Volle, dite centrée sur le « tableau nul ». Nous les nommerons respectivement $F_t^{(k)}$ et $G_i^{(k)}$. Ils sont liés entre eux et avec les $w_t^{(k)}$ et $u_i^{(k)}$ par des formules de transition, en particulier :

$F_t^{(k)} = \Sigma_i u_i^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$ $G_i^{(k)} = \Sigma_t w_t^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$ (3)

Ce qui permet de les calculer à partir de l'extraction des éléments propres du nuage de points de cardinalité minimale.

5.3 Algorithme des K-Means Axiales

Le principe de notre algorithme KMA (Lelu 1994) est de réaliser la partition en classes des unités textuelles dans l'espace distributionnel : après initialisation au hasard de K axes de classes, on attribue chaque vecteur-résumé transformé à l'axe le plus proche angulairement, puis on extrait pour chaque classe les facteurs mots et documents définis ci-dessus. Chaque premier facteur est alors un indicateur de centralité (ou « typicité ») du texte ou du mot dans sa classe (cf. (Page et al. 1998) pour l'application d'une notion de centralité voisine, basée sur les liens et non sur les mots, en tant que mesure de notoriété globale d'une page Web dans le moteur de recherche Google). Les résumés et les mots sont projetés sur l'ensemble des axes, permettant une interprétation nuancée en termes de classes recouvrantes – nous parlerons de thèmes plutôt que de classes. Enfin les thèmes sont disposés sur une carte

[2] Toute matrice X réelle de T lignes, I colonnes et de rang K se décompose classiquement en :
X = U D W' où D est la matrice diagonale de ses K valeurs singulières ; U et W de dimensions respectives (T, K) et (I, K) sont les matrices rassemblant les vecteurs singuliers.

factorielle globale, où les cercles représentent les thèmes, d'importance mesurée par leurs inerties $\lambda^{(1)}$, et les arêtes représentent l'intensité de leurs liens, mesurée par leurs cosinus (Fig. 1).

5.4 Projection de variables illustratives

L'indexation des résumés par d'autres types de termes que les noms de gènes et protéines fournit des vecteurs de fréquences qu'il est possible de projeter en tant que variables explicatives sur les divers axes au moyen de la deuxième formule de transition (3). On trouvera ci-après (Tab. 2) un exemple de ces projections sur un thème issu du présent travail.

.48	5	oxygen_limitation
.41	8	electron_acceptors
.32	5	ResD_and_ResE
.30	312	Bacillus_Subtilis
.30	11	transduction_system
.17	50	structural_gene
.17	4	transcriptional_regulation
.16	4	nitrate_reductase
.16	7	regulatory_pathway
.16	7	regulatory_mechanism
.16	4	response_to_changes
.14	9	signal_transduction_system
.14	5	sensor_kinase
.14	24	response_regulator
.14	7	anaerobic_respiration

TAB 2 – *Projection sur le thème* Respiration sur nitrate *des multitermes biologiques extraits. (au milieu : fréquences totales dans le corpus)*

6 Traitements et interprétation : processus biologiques extraits.

Après introduction du corpus, indexé par les seuls gènes et protéines, dans notre environnement NeuroNav (Lelu et Aubin 1999) nous avons éliminé les noms de gènes et protéines « outils » (*LacZ*, *reporter gene*…) dont la présence bruitait l'analyse. Pour la même raison nous avons également éliminé itérativement ces noms s'ils n'apparaissaient que dans un seul des résumés sélectionnés, et les résumés indexés par moins de quatre noms, ce qui a réduit notre corpus à 581 résumés (indexés par ailleurs par des multitermes biologiques).

L'analyse obtenue par les K-Means Axiales avec 20 thèmes nous a paru la plus pertinente ; disposés sur la carte globale (cf. Fig. 1), les thèmes se répartissent ainsi :

- En bas de la carte, le thème général *Facteurs de transcription*, concerne les protéines responsables de l'expression des gènes en protéines. Ce thème est relié à un ensemble de thèmes dévolus à l'adaptation de la bactérie (réponses à des conditions adverses) :

. *Réponses générales aux stress* : un ensemble de protéines sont produites dans des situations de stress, quel que soit leur nature (variations brutales d'un paramètre physico-chimique de l'environnement : chaleur, pH, oxydants, pression osmotique).

. *Initiation de la sporulation* : une des réponses possibles de la bactérie à ces situations de stress, ou à des carences alimentaires, consiste à former des spores. Ceci lui permet de rentrer dans une phase de « dormance », et d'attendre le retour de conditions plus favorables à sa croissance, en mesure de quoi la spore « germinera » pour former à nouveau une bactérie fonctionnelle.

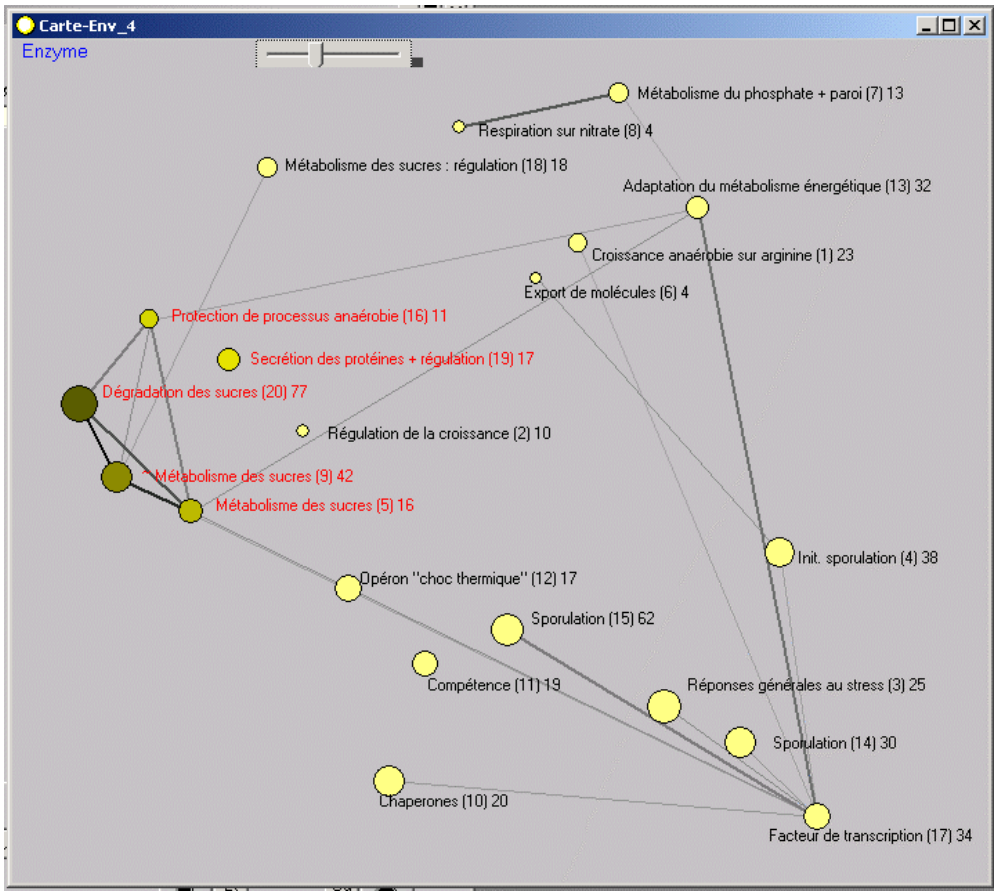

FIG 1 – *Carte globale des 20 thèmes, « éclairée » par le terme* Enzyme*. Les arêtes traduisent la proximité des thèmes dans l'espace de tous les noms de gènes et protéines.*

. Deux thèmes *Sporulation* : ces deux thèmes sont relatifs à la formation de la spore, dont l'un évoque explicitement les protéines qui forment l'enveloppe de la spore, et les gènes qui en régulent l'expression.

. *Compétence* (« sexe », échange de matériel génétique) : une autre des réponses adaptatives possible de la bactérie aux stress consiste à engendrer de la variabilité génétique par échange et recombinaison de segments d'ADN chromosomique entre individus d'une population. La stratégie consiste à essayer de générer des variants génétiques plus adaptés aux nouvelles conditions que les formes de la population existante.

. *Chaperones* : ces protéines ont généralement pour fonction d'aider d'autres protéines à conserver intacte leur structure tridimensionnelle, elles empêchent notamment les protéines de se déplier lors d'un stress thermique (résistance à la chaleur). Elles servent aussi de protéines d'« accompagnement » d'autres protéines nouvellement synthétisées, qui sont ainsi conservées dans un état déplié pour la traversée des membranes biologiques. Beaucoup de protéines sont en effet sécrétées dans le milieu extérieur (enzymes de dégradation, bactéricides).

Dans la même région de la carte on trouve un thème lié à cette problématique, *Opéron « Choc thermique »* : un opéron est un groupe de gènes contigus, dont l'expression est soumise à une régulation commune. Les gènes de cet opéron particulier codent pour des protéines protégeant la bactérie des chocs thermiques, notamment des chaperones.

- L'ouest de la carte, fortement éclairé par le terme « enzyme », est dominé par les processus impliqués dans le *métabolisme des sucres*. On observe un gradient d'Ouest en Est sur trois thèmes, dont le dernier relie ce métabolisme des sucres à la biosynthèse des acides aminés, en passant par le cycle de Krebs (cycle de l'acide citrique, source principale de l'énergie métabolique).

Les trois classes précédentes sont reliées à une classe concernant la *protection des processus anaérobies* (gènes de réponse à des stress oxydatifs, respiration sur nitrate).

Deux petites classes apparaissent dans la même région, concernant respectivement la *sécrétion des protéines* et les régulations associées, ainsi que les *régulations de la croissance* de la cellule.

- Le nord-est de la carte comporte des thèmes liés au métabolisme et à des modes de fonctionnement anaérobie :

- *Respiration sur nitrate* : en absence d'oxygène, la bactérie est capable de le remplacer par du nitrate, afin d'assurer la production d'énergie par l'oxydation des molécule carbonées.
- *Croissance anaérobie sur arginine* : la bactérie est également capable d'utiliser l'acide aminé arginine comme source d'énergie en absence d'oxygène.
- *Métabolisme du phosphate* : la molécule de phosphate est un composant important des acides nucléiques (ADN) et des molécules qui stockent sous forme chimique (ATP) l'énergie récupérée de l'oxydation des aliments.

On y trouve aussi : *Adaptation du métabolisme énergétique* : il s'agit des régulations qui permettent à la bactérie de s'adapter aux sources d'énergie disponibles ; ainsi que *Transport de molécules* : contrairement à la classe sur la sécrétion des protéines, cette dernière parle du transport de petites molécules à travers les membranes biologiques.

7 Conclusion

Nous avons montré que la conjonction d'une chaîne d'indexation automatique humainement contrôlée, pour extraire les noms des gènes et de leurs produits, et d'un environnement interactif de classification non supervisée, permet de structurer la connaissance contenue dans un corpus de résumés Medline de façon utile à la pratique des biologistes, de plus en plus « intégrative ». Ce processus est général : seules les ressources textuelles utilisées sont spécifiques à la microbiologie. Nos efforts portent désormais 1) sur l'exhaustivité de l'extraction des noms de gènes et protéines, en particulier par l'utilisation plus poussée de patrons morpho-syntaxiques, 2) sur le perfectionnement de nos méthodes de classification non supervisée, elles-mêmes issus d'une longue expérience de l'intérêt et des limites des méthodes classiques (AFC, cartes de Kohonen…), en particulier pour les rendre dynamiques, afin de rendre compte des évolutions, 3) sur l'intégration d'informations extérieures pour illustrer l'analyse, 4) sur l'ergonomie de l'interface à la disposition des biologistes.

Références

Benzécri J.P. et coll. (1981). Pratique de l'Analyse des Données : Linguistique et Lexicologie. Dunod, Paris.

Blaschke, C., Andrade, M. A., Ouzounis, C. & Valencia, A. (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. Proc. AAAI Conf. Intell.Syst. Mol. Biol. 7, 60–67

Buntine W. (2002) – Variational extensions to EM and multinomial PCA. In proc. ECML 2002.

Domengès D., Volle M. (1979). Analyse factorielle sphérique : une exploration. Annales de l'INSEE, 35-1979.

Escofier B. (1978). Analyses factorielles et distances répondant au principe d'équivalence distributionnelle. Revue de Stat. Appliquée, 26(4):29-37, Paris

Fichet B. et Gbegan A. (1985). Analyse factorielle des correspondances sur signes de présence-absence Diday et al. eds, 4e Journées Analyse des données et Informatique. INRIA, Rocquencourt.

Iliopoulos I., Enright A.J., Ouzounis C.A. (2001), Textquest: document clustering of medline abstracts for concept discovery in molecular biology ; Pacific Symposium on Biocomputing pp.384-395

Jacquemin C., Daille B., Royauté J. et Polanco X. (2002) - In vitro Evaluation of a Program for Machine-Aided, Information Processing & Management, Vol. 38, Issue 6, Pages 765-792, 2002

Jenssen TK, Laegreid A, Komorowski J, Hovig E. (2001) A literature network of human genes for high-throughput analysis of gene expression. Nature Genet; 28 :21–28.

Lebart L., Morineau A., Tabard N. (1977). Techniques de la description statistique. Dunod, Paris.

Lelu A. (1994). Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In E. Diday, Y. Lechevallier & al. editors. New Approaches in Classification and Data Analysis, pages 241-248, Springer-Verlag, Berlin

Lelu A., Aubin S. (2001) – Vers un environnement complet de synthèse statistique de contenus textuels : Neuronav V2 - séminaire ADEST du 13/11/2001, Paris
www.upmf-grenoble.fr/adest/seminaires/lelu02/ADEST2001_SA_AL.htm

Matusita K. (1955). Decision rules, based on the distance for problems of fit, two examples, and estimation. Annals of the Institute of Statistical Mathematics, pages 631-640, Tokyo.

Page L., Brin S., Motwani R., Winograd T. (1998) - The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project

Renyi A. (1966). Calcul des probabilités. Dunod, Paris.

Royauté J., François C., Besagni D., (2001) - Apport d'une méthodologie de recherche de termes en corpus dans un processus de KDD : application de veille en biologie moléculaire, VSST'2001, 15-19 octobre 2001, Barcelone, Org. FPC/UPC - SFBA-IRIT, Actes I : full paper, pp. 49-62

Royauté J., François C., Zasadzinski A., Besagni D.,Dessen P., Le Minor S.,Maunoury M.T, (2004). Relations entre gènes impliqués dans les cancers de la thyroïde. Actes EGC2004.

Stein, L. (2001). Genome annotation : from sequence to biology. Nat Rev Genet **2** 493–503.

Wilbur WJ. (2002) A thematic analysis of the AIDS literature. 1: Pac Symp Biocomput.pp.386-97.

Wilkinson D., Huberman D. A. (2002) Finding Communities of Related Genes, Quantitative Biology archive, http://arxiv.org/list/q-bio/0210

Yandell M.D., Majoros W.H. (2002), Genomics and natural language processing, Nature Reviews Genetics, vol.3, pp 601-610.

Summary

In the pathway to large-scale integrative biology, the challenge is to integrate millions of elementary facts mainly concentrated in text databases, such as Medline. Our experience on *Bacillus* AND *transcription* literature reported here shows that a 2-steps procedure can solve the problem for a given domain, provided that an expert with powerful computer interfaces is in the loop: 1) exhaustive mining for gene/protein names, 2) interactively structuring the corpus using our fuzzy and dual unsupervised Axial K-Means clustering method, giving any abstract and name a centrality measure in any thematic cluster. These clusters describe useful signaling and metabolic pathways extensively studied in the Medline abstracts corpus.