BULLETIN N° 101
ACADÉMIE EUROPEENNE INTERDISCIPLINAIRE DES SCIENCES

Séance du Mardi 10 janvier 2006

Quelques hypothèses sur les causes exogènes et environnementales des cancers et en particulier celui du col de l'utérus par :
Pr. J. Poirier, Dr. M.L. Labat, et le Dr. A.Lécu de la MNHN

Prochaine séance : le Mardi 14 février 2006 à 18h :

Conférence de notre Collègue le Pr. Gérard LEVY
« Les fouilles de données (data mining) »

Académie Européenne Interdisciplinaire des Sciences
Siège Social : Maison des Sciences de l'Homme 54, bd Raspail 75006 Paris
Site Web : http://aeis.ifrance.com/aeis/

ACADEMIE EUROPEENNE INTERDISCIPLINAIRE DES SCIENCES
MAISON DES SCIENCES DE L'HOMME

Janvier 2006

N°101

TABLE DES MATIERES

Prochaine séance : Mardi 14 février 2006,
MSH, salle 215 à 18h
Conférence de notre Collègue le Pr. Gérard LEVY

« Les fouilles de données (data mining) »

ACADEMIE EUROPEENNE INTERDISCIPLINAIRE DES SCIENCES
Maison des Sciences de l'Homme, Paris.

Séance du
Mardi 10 janvier 2006

Maison des Sciences de l'Homme, salle 215, à 18 h.

La séance est ouverte à 18 h. 00 sous la Présidence de Michel GONDRAN et en présence de nos collègues G. BELAUBRE, M. BERREBY, B.BLONDEL, N. CAGNARD, J.-P. FRANCOISE, I. HERPE-LITWIN, M.L. LABAT, E. NUNEZ, J. POIRIER et du conférencier Alexis LECU de la MNHN accompagné de J. LORILLEUX

Michel GONDRAN ouvre la séance en rappelant les faits marquants de l'année 2005 :

- Création de la section de METZ-NANCY à l'initiative du Pr. Pierre NABET
- Sortie du 1er TOME sur la Conscience réalisé à la suite du 1er congrès , « BIOLOGIE ET CONSCIENCE » qui avait été présidé par le Pr. G. EDELMAN en avril 2002
- Sortie du livre « La Science en Mouvement » réalisé par la section NICE-CÔTE D'AZUR
- La manifestation sur « COLLOQUE SUR LES PEURS » à Nice.
- La réussite exceptionnelle du Congrès « PHYSIQUE et CONSCIENCE » des 9 et 10 décembre 2005.

Il remercie par ailleurs G. BELAUBRE , I. HERPE et N. CAGNARD pour leurs contributions aux congrès « FRACTALES EN PROGRES » et « PHYSIQUE ET CONSCIENCE ».

Michel GONDRAN précise ensuite les projets de l'Académie pour 2006 :

- Consolider les avancées et les transformer
- Edition des Actes de « FRACTALES EN PROGRES » et du Tome 2 de la conscience (Congrès « PHYSIQUE ET CONSCIENCE »)
- Nouveaux projets de congrès à discuter : plasticité, biologie théorique…Il invite les nouveaux membres de l'Académie à donner leurs idées.
- Concernant la structure de l'Académie il invite à réfléchir sur les modalités de coopération entre les diverses sections de l'Académie, Nice, Nancy et bientôt à Neuchâtel-Genève

Après transmission de ses meilleurs vœux pour la nouvelle année et remerciements à notre collègue Gilles COHEN–TANNOUDJI pour son apport essentiel dans le dernier congrès, la parole est donnée à notre nouveau collègue J.P. FRANCOISE qui nous présente ses travaux sur le thème, « MODELISATIONS MATHEMATIQUES DES RYTHMES COMPLEXES DU VIVANT »[1]

[1] Voir p. 19 le résumé de cette présentation.

En conclusion, Jean Pierre FRANCOISE déclare avoir toujours été intéressé par les problèmes de mathématiques et leur intersection avec les autres disciplines scientifiques afin de pouvoir déceler des structures communes à ces diverses sciences.

Après quelques brèves questions , la parole est donnée au Pr. POIRIER qui nous entretient du lien entre les spermatozoïdes et le cancer du col de l'Utérus. Le Pr. POIRIER[2] a souhaité que des spécialistes viennent présenter des travaux qui relativisent la thèse généralement admise qui consiste à considérer le Papilloma virus comme agent principal de la cancérisation du col de l'utérus chez la femme. En se basant sur une comparaison entre certains mammifères dont l'activité sexuelle est relativement sporadique et qui ne manifestent pratiquement jamais de cancer du col utérin et l'espèce humaine chez laquelle ce cancer est très fréquemment observée. Les spermatozoïdes pourraient être les agents principaux de la cancérisation. Il observe également que les religieuses ainsi que les femmes soumises à certaines pratiques religieuses restrictives développent très peu de cancers utérins.

Avant de donner la parole à ses divers intervenants , le Pr. POIRIER rappelle que ses recherches ethnologiques et anthropologiques ont pu éclairer des enquêtes épidémiologiques sur certaines pathologies. Il donne comme exemple les travaux qu'il a réalisés sur les indigènes de l'Île de Pâques.

Après cet exposé, notre collègue Marie-Louise LABAT[3] nous présente sa thèse sur « imprégnation spermatique, inflammation, cellules souches et cancer du col de l'utérus ». M.L. LABAT inscrit ces recherches dans un cadre plus général de « LIEN ENTRE INFLAMMATION, STIMULATION DE CELLULES SOUCHES ADULTES PLURIPOTENTES PRESENTES DANS LE SANG ET CANCERISATION » . La cancérisation correspondrait à la perte d'émission par les cellules souches des substances entraînant la régulation de leur croissance par des lymphocytes spécialisés. L'inflammation, les spermatozoïdes, le papillomavirus pourraient ainsi induire de telles proliférations.

Le Dr. LECU [4] expose ensuite les statistiques des causes de décès observées chez les animaux du Parc zoologique de Paris. Il présente principalement les statistiques concernant les pathologies cancéreuses . Un seul cancer de l'utérus a été observé sur 10.000décès. On note globalement des taux très faibles des pathologies cancéreuses : 0,5% des décès sont imputables à des cancers.

Enfin notre collègue, le Pr. Emmanuel NUNEZ[5] insiste sur la combinatoire complexe des facteurs impliqués dans le développement des pathologies , en particulier cancéreuses. Il faut considérer dans le cas présenté du col utérin que les affections virales, les spermatozoïdes et les inflammations de toute nature sont des cofacteurs de développement du cancer utérin.

Après quelques questions , le débat prend fin et la séance est levée à 20 heures.

Compte rendu effectué par Noëlle CAGNARD
Secrétaire générale adjointe

[2] Voir dans le bulletin n°100 , résumé de la communication du Pr. POIRIER
[3] Voir texte correspondant dans le bulletin n°100
[4] Voir texte correspondant dans le bulletin n°100
[5] Voir texte correspondant dans le bulletin n° 100

Comptes-rendus de la Section Nice-Côte d'Azur

Le savoir est le seul bien qui s'accroisse à le partager. Comprendre est bien sans limite qui apporte une joie parfaite.
Baruch SPINOZA (1632-1677)

Compte-rendu de la séance du 15 décembre 2005

(89ème séance)

Présents :

Jean Aubouin, René Blanchet, Sonia Chakhoff, Pierre Coullet, Patrice Crossa-Raynaud, Guy Darcourt, René Dars, Jean-Pierre Delmont, Jean-Paul Goux, Maurice Papo, Jacques Wolgensinger.

Excusés :

Emile Girard, Gérard Iooss, Yves Ignazi, Jean Jaubert, Michel Lazdunski, Jean-François Mattéi.

1- Approbation du compte-rendu de la 88ème séance.

Le compte-rendu est approuvé sans modification à l'unanimité des présents.

2- Election de Monsieur le Recteur René Blanchet à l'Académie des Sciences.

Cette nouvelle nous a réjouis. Il a été élu aux deux sections : Sciences de l'Univers et Application des Sciences. Dans ces deux domaines, il est particulièrement qualifié et intéressé et cela va donc lui donner beaucoup de travail, mais il nous promet d'être toujours fidèle à nos activités de l'AEIS.

3- Accueil d'un nouveau membre.

Nous accueillons aujourd'hui un nouveau membre, Jacques Lebraty, professeur émérite des Universités (Economie), Président du Conseil scientifique et administrateur de l'A.P.M. (Association pour le Progrès du *Management*), coopté à l'unanimité.

4- Activités de l'Académie en 2006.

Pour préparer notre prochaine année, au cours de laquelle la trilogie Université – Centre Universitaire Méditerranéen – A.E.I.S. Nice Côte d'Azur devrait collaborer efficacement, nous envisageons **une série de conférences**, échelonnées de mars à juin, puis en octobre et novembre, avec des orateurs renommés.

Nous avons pensé à Claude J. Allègre, Philippe Taquet (le maître des Dinosaures), Henri de Lumley (le paléontologiste humain), André Vacheron (cardiologue), Annie Cazenave (les formes de la Terre), Reno Ruffini (Icranet et les trous noirs), Claude Lorius (glaciologue), André Brahic (de Mars à Titan …) : Jean-Pierre Delmont envisage de proposer un médecin qui parlerait des clonages humains ; Jean-Paul Goux souhaiterait un conférencier traitant des transplantations d'organes. D'autres noms ont été avancés : Emmanuel Le Roy-Ladurie (historien), Marc Fumaroli (historien et critique), Luc Ferry (en tant que philosophe).

Cette liste n'est pas exhaustive et nous devons la compléter en indiquant quelles personnalités nous pourrions inviter.

Nous pourrions mettre **la série** sous un thème (pas exemple *La science en mouvement*).

Pour pouvoir bénéficier de l'aide optimale du CUM, nous devons présenter un programme clair, précis et daté à la fin du mois de janvier.

Nous essaierons cependant de présenter également un colloque sur un thème que nous aurons défini ensemble.

Un des buts que nous nous sommes fixés, outre l'interdisciplinarité, était de créer à Nice une interface reconnue entre la Science et la Société. Nous l'avons réalisé avec nos publications, mais pas avec le public qui demeure trop limité.

Avec le C.U.M., cela pourrait devenir « Les mercredis (jour de congé des élèves) de l'AEIS » avec une périodicité mensuelle. Pour cela, il faut contacter Madame Rampal avec l'aide de Guy Darcourt et de Pierre Coullet.

5- Colloque « Les peurs de notre temps ».

Tous les exposés ont été réunis et seront prochainement préparés pour l'édition

6- Questions diverses.

Jean Aubouin : La foi dans le changement climatique à venir (cent ans ou même mille ans !) repose sur l'association entre l'augmentation conjointe du taux de gaz carbonique et de la température dans le passé. Or, les forages récents dans le dôme C de l'Antarctique, qui ont permis de recueillir la glace jusqu'à -650 000 ans, couvrant ainsi plusieurs glaciations du Quaternaire, ont montré que l'augmentation de la température précède toujours celle du gaz carbonique. Celui-ci n'est donc pas la cause des variations climatiques du passé qui sont d'ordre astronomique, comme Milankovitch l'a montré depuis longtemps.

L'augmentation actuelle de la température est-elle ainsi naturelle pour la double raison que nous sommes dans une période post-glaciaire (le dernier épisode glaciaire remonte à -18 000 ans) et dans une période postérieure au "Petit âge glaciaire" (période plus froide du 13ème siècle environ à la fin du 19ème siècle).Le taux particulièrement élevé du gaz carbonique dans l'atmosphère actuelle sera-t-il une cause d'un plus fort réchauffement ? C'est affaire de postulat et de modèles.

Mais ceux-ci sont bien variés ! Par exemple : "si la banquise polaire et la calotte grœnlandaise fondaient, les eaux arctiques, dès lors plus froides, pourraient faire dévier la branche septentrionale du Gulf Stream dans l'Atlantique nord. L'Europe connaîtrait alors un fort refroidissement … « Bref, le réchauffement provoquerait aussi le refroidissement … Tout est dans tout et réciproquement !

Un autre aspect des modes de pensée "politiquement correcte" est l'Ecologie. Celle-ci s'est développée, heureusement d'ailleurs, pour compenser l'abandon, dans la biologie moderne, des "vieilles" disciplines descriptives, botanique, zoologie, systématique etc., jugées démodées au temps de la biologie moléculaire. L'Ecologie a ainsi permis de redécouvrir le prix de la biodiversité.

Mais, très vite, les excès sont venus qui ont montré aux écologistes "politiques" leur puissance extraordinaire : pour un coléoptère (bien sûr endémique), on bloque la construction d'une autoroute.

On est alors passé au "scientifico-médiatique". C'est ainsi que l'on a eu les pluies acides, la caulerpe, le trou d'ozone, etc. ; et maintenant le réchauffement climatique. Et quoi demain ?

A tous il faut rappeler que la pensée scientifique consiste à admettre que "les faits ont raison par rapport à l'idée qu'on en a". Ce qui ne veut pas dire qu'il ne faut pas avoir d'idées ; mais qu'il convient de les soumettre à l'épreuve de la réalité.

Pour illustrer ce propos, **Pierre Coullet** rappelle qu'un objet lâché tombe sur le sol. C'est un fait évident pour lequel on n'a toujours pas d'explication ; il est seulement décrit.

Et pourtant, Newton fut un bien plus grand génie qu'Einstein qui a travaillé dans une continuité avec ses prédécesseurs, notamment H. Poincaré.

Bonne année à tous

★★

Prochaine réunion
le jeudi 19 janvier 2006 à 17 heures
au siège
Palais Marie-Christine
20 rue de France
06000 NICE

INVITATION

P. Centre Cavaillès - Ecole normale supérieure

Perspectives nouvelles en biologie théorique

Journée d'étude - 27 Février 2006
Salle Dussane, 45 rue d'Ulm, 75005, Paris

Cinq ans après le séquençage du génome humain, jamais le besoin d'une biologie théorique n'a été aussi sensible. Il se manifeste par l'émergence rapide de la biologie des systèmes. Cette journée sera l'occasion de confronter et de discuter quelques travaux allant dans ce sens.

9H Charles Auffray : SYSTEMOSCOPE : un programme trans-disciplinaire de biologie systémique

9H45 Carlos Sonnenschein : Fondements de la théorie du champ d'organisation tissulaire: La prolifération, état par défaut universel de toutes les cellules

10H30 Ana Soto : Fondements de la théorie du champ d'organisation tissulaire: Le champ d'organisation tissulaire

11H15 Michel Morange : Quelle 'niche' pour la biologie théorique aujourd'hui?

14H Denis Noble : Introduction à l'étude de la médecine théorique

14H45 Olivier Gandrillon : Vers une biologie des systèmes : émergence de structures multi-protéiques dans un système multi-agents

15H30 Jean-Jacques Kupiec : Du programme génétique au darwinisme cellulaire

16H15 Jean-Pascal Capp : La nature aléatoire de l'expression génique à l'origine du cancer ? Eléments pour une nouvelle perspective sur la pathologie tumorale.

Résumés des interventions

SYSTEMOSCOPE : un programme trans-disciplinaire de biologie systémique

Charles Auffray, Equipe Genexpress, Génomique Fonctionnelle et Biologie Systémique pour la Santé, CNRS et Université Pierre et Marie Curie- Paris VI, LGN - UMR 7091, Villejuif

Pour appréhender la complexité des systèmes biologiques, il apparaît nécessaire d'intégrer les approches analytiques de la découverte avec les théories des systèmes et de la dynamique non linéaire. Dans ce but, les membres du Consortium SYSTEMOSCOPE ont entrepris de repenser leur stratégie de recherche et de collaborer au sein d'un réseau national et international pour développer un programme de recherche et de formation trans-disciplinaire en biologie systémique, fondé sur leur expérience étendue en mathématique, informatique, physique, biologie et médecine.

Ce programme vise à concevoir et implémenter un projet pilote de démonstration en biologie systémique afin de mesurer les dysfonctions du métabolisme énergétique et les modulations de profil d'expression dans les muscles squelettiques de patients atteints de maladies pulmonaires ou d'insuffisance cardiaque, et dans des cas de transplantation pulmonaire ou cardiaque avant et après réhabilitation de la myopathie systémique; d'identifier les réseaux fonctionnels de régulation sous-jacents en utilisant les outils développés pour la biologie systémique; de mesurer l'influence des systèmes immunitaire et nerveux par l'étude de polymorphismes génétiques; de formuler des hypothèses de travail et les tester; d'itérer le processus et évaluer l'amélioration des modèles et des méthodes.

Nous utiliserons des technologies de mesure, d'annotation et de biovalidation standardisées pour permettre aux technologies de génomique fonctionnelle de fournir les données précises et fiables requises pour capter les multiples fluctuations modérées biologiquement importantes échappant aux méta analyses.

Un objectif important de ce programme est de contribuer au cadre mathématique et conceptuel pour la biologie systémique en utilisant des outils mathématiques et informatiques pour inférer, modéliser et simuler des réseaux fonctionnels à partir de jeux étendus de données d'expression combinés avec des données génétiques et phénotypiques; en développant des mesures robustes pour les fonctions et la complexité biologique; en simulant la différentiation cellulaire par le modèle sélectif d'expression stochastique régulée par des contraintes dynamiques et des modifications épigénétiques; en construisant un cadre théorique pour modéliser les attracteurs à nombre variable de dimensions dans les systèmes biologiques; en combinant les éléments de modélisation de la relativité d'échelle pour décrire les effets induits par des lois d'échelle non linéaires, afin de relier les propriétés des systèmes biologiques aux principes physiques premiers sous-jacents.

Ce programme favorisera la formation trans-disciplinaire, la dissémination des résultats par des publications librement accessibles, la protection des inventions, et l'émergence de nouveaux programmes de recherche en biologie systémique à l'interface des disciplines concernées.

Auffray, C., Imbeaud, S., Roux-Rouquié, M. and Hood, L. (2003) Self-organized living systems: conjunction of a stable organization with chaotic fluctuations in biological space time. Phil. Trans. R. Soc. Math. Phys. Eng. Sci. 361, 1125-1139.

Fondements de la théorie du champ d'organisation tissulaire: La prolifération, état par défaut universel de toutes les cellules

Carlos Sonnenschein - Tufts University - Boston

La théorie du champ d'organisation tissulaire est fondée sur l'hypothèse que l'état par défaut de toute cellules est la prolifération, et sur le fait que la carcinogenèse a son origine au niveau tissulaire. Dans une perspective évolutive, et d'après nos propres données expérimentales et notre interprétation des données de nos collègues, nous proposons donc :

(i)La prolifération est une propriété constitutive des cellules, qu'il s'agisse d'organismes unicellulaires ou pluricellulaires. Leur état par défaut est la prolifération.

(ii) La prolifération est contrôlée par des inhibiteurs.

(iii) Le cycle cellulaire est une séquence d'algorithmes (ou de réactions, ou d'événements) pour fabriquer deux cellules à partir d'une; il a lieu de manière automatique et répétitive, tant que la disponibilité en nutriments est assurée et qu'il n'y a pas d'inhibiteurs de la prolifération présents et effectivement reconnus. Une fois le cycle engagé, il ne peut s'arrêter naturellement avant de s'être terminé.

(iv) Les inhibiteurs conduisent les cellules à la quiescence (phase G0). Le contrôle de la prolifération fonctionne comme un interrupteur ouvert/fermé.

Quand les cellules sont dissociées, elles sont aussi libérées des contraintes de l'organisme qui les amènent à exprimer un phénotype particulier en fonction de leur position. Ainsi, elles vont utiliser leur capacités propres à proliférer, et vont pouvoir exprimer de nouveaux phénotypes, notamment leurs capacités propres à migrer.

Sonnenschein C, Soto AM 1999 The Society of Cells: Cancer and Control of Cell Proliferation.Springer Verlag, New York

Fondements de la théorie du champ d'organisation tissulaire: Le champ d'organisation tissulaire

Ana Soto - Tufts University - Boston

En affirmant que la carcinogenèse a lieu à l'échelle du tissu, les oncogènes et les gènes suppresseurs ne sont plus considérés comme des causes car ils opèrent au niveau cellulaire/subcellulaire. Cette théorie se prête à être étudiée en utilisant le paradigme du champ morphogénétique. A l'âge adulte, les interactions entre le stroma et le parenchyme continuent de réguler la maintenance et la réparation tissulaires. Elle propose en plus que le stroma et le parenchyme soient les cibles des changements induits par les carcinogènes. Dans cette perspective, l'étude de la théorie du champ d'organisation tissulaire de la carcinogenèse nécessite de mettre en œuvre des expériences de recombinaison et de transplantation de tissus, comparables à celles qui ont fournit des informations aux embryologistes sur les influences inductives et permissives pendant l'organogenèse.

Soto AM, Sonnenschein C. The somatic mutation theory of cancer: growing problems with the paradigm? BioEssays 26:1097–1107, 2004

Quelle 'niche' pour la biologie théorique aujourd'hui?

Michel Morange - Centre Cavaillès - Ecole normale supérieure

En dépit de l'ancienneté de ses racines, la biologie théorique a toujours eu beaucoup de mal à trouver une place parmi les sciences du vivant. Le titre de cette réunion, et de beaucoup d'autres, suggère que ses chances de développement seraient plus grandes aujourd'hui qu'elles n'ont jamais été.

Ma communication visera d'abord à définir la niche écologique dans laquelle la biologie théorique pourrait croître aujourd'hui, c'est-à-dire ce qui en favoriserait le développement : les difficultés de l'approche moléculaire et la déception conséquente ou, à l'inverse, la connaissance enfin sûre des bases moléculaires sur lesquelles une modélisation pourrait être tentée de manière efficace ; les difficultés rencontrées dans l'interprétation moléculaire des phénomènes pathologiques, tel le cancer, et dans les approches thérapeutiques qui en sont dérivées.

La deuxième partie visera à rechercher ce qui pourrait constituer le coeur de cette nouvelle biologie théorique : les outils mathématiques, informatiques, utilisés dans les nouvelles approches de la post-génomique ? une attention plus grande à l'histoire évolutive des phénomènes observés ? ou le basculement d'une vision réductionniste à une vision plus globale ?

Michel Morange (2005) Les secrets du vivant : contre la pensée unique en biologie (Paris : La Découverte)

Introduction à l'étude de la médecine théorique

Denis Noble - Oxford Cardiac Electrophysiology Group - Université d'Oxford

Sans doute le succès de la génétique moléculaire tient-il à ce qu'elle permet d'aller des gènes aux protéines à l'aide d'un simple code. Mais pour comprendre le Vivant, il n' y a pas d'autre voie que de se projeter à des niveaux d'explication plus élevés. C'est en explorant la logique du Vivant à ces niveaux supérieurs que nous pourrons finalement développer une théorie biologique. Voilà le grand défi de la biologie actuelle : rendre compte du phénotype en termes d'interactions des protéines des systèmes. Sur ce point, la génétique moléculaire ne nous est guère utile. C'est même le contraire tant il est sûr que c'est en accomplissant cela que nous pourrons correctement « annoter » le génome, c'est-à-dire déterminer les fonctions de chacune de ses séquences. Le génome sera lu à partir du phénotype, et non l'inverse. J'emploierai nos études expérimentales et théoriques sur le coeur pour illustrer ces principes.

Denis Noble, The Music of Life, Biology beyond the genome, Oxford University Press, sous-presse, Juin 2006.

Vers une biologie des systèmes : émergence de structures multi-protéiques dans un système multi-agents.

Olivier Gandrillon - Université Claude Bernard Lyon1 - Centre de Génétique Moléculaire et Cellulaire UMR 5534. Groupe de travail transdisciplinaire BSMC

Au delà des mots et des incantations, le développement attendu d'une « Biologie des Systèmes » nécessitera le développement d'approches appropriées à son objet (les systèmes vivants dans leur globalité). Pour cela deux visions s'opposent : l'une, orientée données, pousse les feux sous la production de données de plus en plus exhaustives et d'une qualité croissante. L'autre, orientée modèle, pose de façon symétrique la question de trouver les « bonnes » façon de modéliser et simuler dans un cadre systémique. Le travail présenté s'inscrit dans cette deuxième approche. Il consiste à simuler informatiquement un monde tridimensionnel dans lequel les règles physiques de bases les plus réalistes possibles sont implémentées. Des agents, représentant des domaines protéiques, sont ensemencés dans ce monde, accompagnés de quelques règles simples d'interaction. Les résultats préliminaires montrent que sous ces règles très simples d'interaction des structures dynamiques peuvent émerger qui miment des structures biologiques connues. Depuis ces premiers résultats, des améliorations importantes ont permis d'augmenter le réalisme de la simulation, tout en autorisant la formation de structures parfois inattendues. Les systèmes multi-agents semblent donc bien être idéalement adaptés pour des simulations informatives sur l'analyse des règles génériques qui sous-tendent les systèmes vivants.

Soula H, Robardet C, Perrin F, Gripon S, Beslon G, Gandrillon O, Modeling the emergence of multi-protein dynamic structures by principles of self-organization through the use of 3DSpi, a multi-agent-based software. BMC Bioinformatics. 2005 Sep 19;6:228

Du programme génétique au darwinisme cellulaire

Jean-Jacques Kupiec - Centre Cavaillès - Ecole normale supérieure

Dans *Qu'est-ce que la vie ?* Schrödinger a tracé une séparation très nette entre la physique et la biologie. Alors qu'en physique un *principe d'ordre à partir du désordre* opère, la biologie dépendrait *d'un principe d'ordre à partir de l'ordre*. Alors qu'en physique l'ordre au niveau macroscopique est produit par un comportement probabiliste des molécules au niveau microscopique, en biologie, les protéines échapperaient au hasard brownien grâce à l'information génétique. Cette vision, à la base de la théorie du programme génétique, a profondément influencé la biologie moléculaire. Cependant, les données récentes démontrant l'importance des phénomènes aléatoires dans les interactions moléculaires et dans l'expression des gènes contredisent le *principe d'ordre par l'ordre* de Schrödinger. Une théorie alternative peut être proposée dans laquelle l'ordre biologique provient d'une extension de la Sélection Naturelle à l'intérieur de l'organisme. L'expression probabiliste des gènes génère une diversité d'états cellulaires et un mécanisme de sélection cellulaire dirige l'embryon vers le stade adulte. Cette théorie a été l'objet de simulations numériques qui démontrent sa pertinence.

Laforge, B., Guez, D., Martinez, M., Kupiec, J.J., 2005. Modeling embryogenesis and cancer : an approach based on an equilibrium between the autostabilization of stochastic gene expression and the interdependance of cells for proliferation. Progress Biophys. Mol. Biol. 89: 93-120.

La nature aléatoire de l'expression génique à l'origine du cancer? Eléments pour une nouvelle perspective sur la pathologie tumorale.

JP Capp – Equipe instabilité génétique et cancer – IPBS – CNRS UMR 5089 - Toulouse

Depuis plusieurs décennies, l'accumulation de données décrivant les altérations du génome des cellules tumorales a conduit à une vision purement génétique du cancer. Toutefois, un nombre croissant de résultats indique que le micro-environnement cellulaire joue un rôle primordial dans l'initiation et la progression de la maladie. La prise en compte de la nature aléatoire de l'expression génique et du rôle du micro-environnement permet d'établir un modèle de cancérogénèse basé sur la perturbation des interactions cellulaires, qui normalement assurent la stabilisation de l'expression génique, la conservation de l'état de différentiation et la quiescence des cellules. Ce modèle permet de prédire de nombreux phénotypes tumoraux, allant de la présence de cellules souches ou dé-différenciées à la génération d'instabilités génétique et épigénétique**. »**

Jean-Pascal Capp , Stochastic gene expression, disruption of tissue averaging effects and cancer as a disease of development , Bioessays 2005 Vol 27 (12) p 1277-1285

Documents

Modélisations mathématiques des rythmes complexes du vivant.

Résumé de la présentation des travaux de notre Collègue Jean-Pierre FRANCOISE

Les rythmes comptent parmi les phénomènes les plus caractéristiques des organismes vivants.

Ils se produisent à tous les niveaux de l'organisation biologique, depuis les organismes unicellulaires jusqu'aux plus évolués, avec des périodes allant du centième de seconde à l'année. Par exemple, un des modes d'expression de l'activité électrique des neurones alterne des phases quiescentes avec des oscillations très rapides regroupées sur une phase dite active de leur période. On ignore pour l'instant comment ce type d'activité électrique permet d'encoder une information transmise à d'autres neurones. Mises en évidence pour la première fois par A. Arvanitaki au collège de France en 1939 dans les neurones, ce type d'oscillation , appelé oscillations en salves, s'est par la suite révélé ubiquitaire en physiologie et plus généralement dans les sciences du vivant. Avec cette complexité des phénomènes temporels mis en jeu se posent au biologiste, physiologiste, médecin de nouvelles questions. Quels faits signifiants faut-il extraire de cette expression périodique ?. La forme des salves, l'amplitude, la fréquence des salves, la période relative des phases actives et quiescentes, sont-ils des données importantes ? La tentative contemporaine de constituer une biologie systémique avec laquelle seront définis des standards reproductibles passe par une analyse approfondie de tels exemples. Pour l'instant, on dispose de modèles biophysiques qui permettent d'expliquer la présence de telles oscillations dans la suite de l'approche de Hodgkin et Huxley[6] à l'électrophysiologie. Que fait le mathématicien ?

Le mathématicien fournit d'abord un langage (celui de la théorie des bifurcations créé par H. Poincaré en 1880) qui permet d'interpréter les données de la simulation numérique, de classer les portraits de phase et de prévoir (parfois au-delà des valeurs possibles pour l'expérimentation) leur changement en fonction de paramètres. Il intervient aussi pour construire des systèmes réduits, plus accessibles à l'analyse, dont les solutions gardent les caractéristiques essentielles du modèle biophysique. C'est ainsi, par exemple que le modèle de FitzHugh-Nagumo décrit par deux équations différentielles

$$x' = y - f(x)$$
$$y' = e(x - c)$$

où f est un polynôme cubique et e est un nombre petit qui représente l'échelle de temps entre la variable x (rapide) et y (lente), permet de comprendre la transition d'un mode pulsatile à un mode excitable d'une cellule. Lorsqu'on considère un forçage de l'équation (c est cette fois une fonction du temps c=c(t)), on peut reproduire des oscillations qui sont du type oscillations en salves[7]. Le mathématicien intervient alors avec l'arsenal de techniques développées dans des contextes complètement différents pour rattacher ce type de phénomènes à d'autres tels, dans cet exemple, la résonance paramétrique.(*)

On peut maintenant considérer le cas où l'entrée c(t) est elle-même solution d'un système de FitzHugh-Nagumo et décrire ainsi une situation de deux oscillateurs (ou de l'évolution moyenne de deux populations d'oscillateurs). Le modèle permet ainsi de prédire l'apparition de patterns temporels constitués de l'alternance de phase pulsatile et de phase décharge (surge). Il évoque donc des situations

[6] J.-P. Françoise, Oscillations en biologie, Collection Mathématiques & Applications, 46, 2005.

[7] C. Doss-Bachelet, J.-P. Françoise, C. Piquet Bursting oscillations in two coupled FitzHugh-Nagumo equations. ComPlexUs 2, (2003), 101-111.

fréquemment observées dans les sécrétions hormonales. Dans un article récent [8](3), en collaboration avec Frédérique Clément (INRIA), nous avons suggéré d'expliquer l'alternance de phase pulsatile et de décharge des neurones à GNRH par l'implication de deux populations de neurones (ou d'une population de neurone et d'une autre constituée d'interneurones). Les neurones à GNRH interviennent dans l'axe hypothalamo-hypophysaire et semblent jouer un rôle important dans la physiologie de la reproduction chez les mammifères. Ces travaux théoriques bénéficient de l'apport expérimental de collègues de l'université de Tours et de l'INRA de Nouzilly. Ils s'inscrivent dans un programme plus vaste de compréhension de l'ontogénèse de ces rythmes complexes et de leur lien possible avec des dynamiques calciques.

[8] F. Clément, J.-P. Françoise Mathematical modeling of the GNRH-pulse and surge generator. Soumis à publication, disponible sur ArXiv

(*) L'encensoir de Saint Jacques de Compostelle, appelé Botafumeiro, fonctionne depuis le 13ème siècle sur le principe de la résonance paramétrique

A la recherche d'un plan B

Deux équipes proposent des méthodes pour produire des cellules souches sans détruire l'embryon. Elles cherchent à résoudre les problèmes éthiques sans pourtant y parvenir complètement.

Extrait du journal Science janvier 2006

Devant les nombreuses objections éthiques liées au clonage thérapeutique, les chercheurs ont développé deux méthodes permettant de créer - du moins chez les souris - des cellules souches pluripotentes, ou embryonnaires, sans détruire en même temps l'embryon.

La. première de ces méthodes, dite par transfert de noyau modifié *(altered nudear transfer,* ou ANT), consiste à injecter dans un ovocyte un noyau modifié pour créer des cellules incapables de former un embryon normal, mais susceptibles de produire des cellules souches embryonnaires. Dans la seconde méthode, les chercheurs obtiennent une lignée de cellules souches embryonnaires à partir d'une cellule prélevée sur un embryon aux premiers stades de développement, tout en laissant les autres cellules se développer dans l'organisme d'une souris vivante.

Jusqu'à présent, le débat sur les méthodes alternatives restait théorique. Rudolf Jaenisch et Alexander Meissner, du Massachusetts Institute of Technology (MIT), ont voulu vérifier s'il était possible de transposer là "théorie dé l'ANT dans la pratique. Ils ont désactivé la fonction d'un gène nommé cdx2 dans une cellule provenant de l'épiderme d'un donneur et fusionné cette cellule avec un ovocyte. L'opération a généré des cellules qui ont pu former une sorte d'embryon précoce, le blastocyte, et donner ainsi naissance à des cellules souches embryonnaires, mais qui étaient incapables de s'implanter dans un utérus, et n'avaient donc aucune chance de se développer en un organisme complet. William Hurlbut, médecin et éthicien à l'université Stanford, membre du Council of Bioethics créé par le président Bush pour étudier ces questions, souligne que, sans cdx2, les ensembles de cellules ne disposent pas de la capacité organisationnelle élémentaire qui leur vaudrait l'appellation d' *"organisme vivant"*. L'absence de cdx2, explique-t-il, *"n'est pas une déficience, mais une insuffisance. Je pense qu'il est raisonnable d'affirmer que les cellules ainsi créées ne constituent pas un être humain."*

Pour d'autres cependant, la méthode soulève plus de questions qu'elle n'en résout, tant sur le plan politique que scientifique. Désactiver cdx2 crée un embryon certes diminué, mais un embryon quand même, estime Tadeusz Pacholczyk, du National Catholic Bioethics Center de Philadelphie.

L'autre solution est l'œuvre de Robert Lanza et ses collègues d'Advanced Cell Technology (ACT), une entreprise américaine de biotechnologie. Ils ont montré qu'il était possible de prélever une seule cellule sur un embryon précoce de souris et de produire une lignée de cellules souches à partir de celle-ci. La technique utilisée par ces scientifiques est similaire à celle employée dans les diagnostics génétiques préimplantatoires pratiqués dans les cliniques de procréation médicalement assistée du monde entier. Les scientifiques prélèvent une ou deux cellules sur des embryons à un stade précoce afin de détecter la présence de certains gènes. Un embryon n'est implanté que si l'on a la certitude qu'il n'est porteur d'aucune maladie génétique.

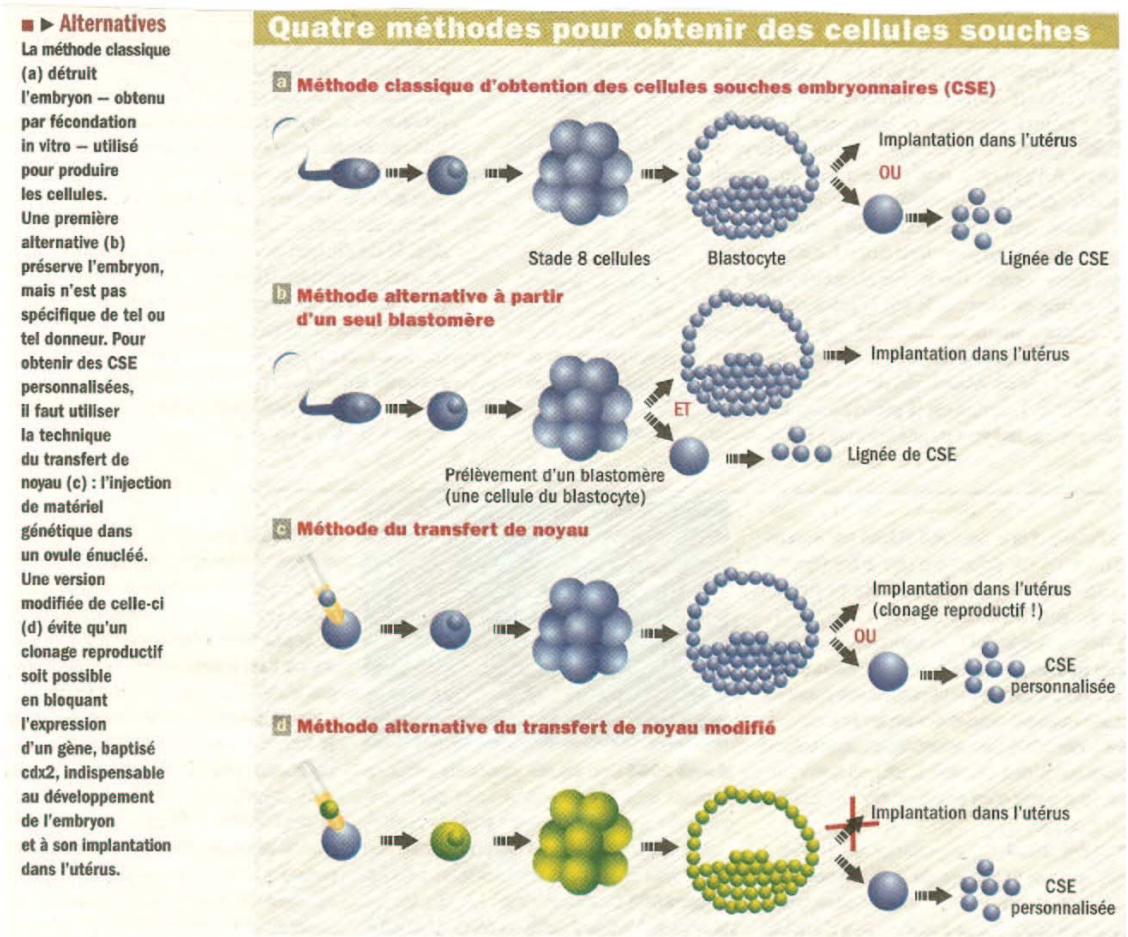
CHAQUE MÉTHODE SOULÈVE SON LOT DE NOUVELLES QUESTIONS

L'équipe d'ACT a montré que, en .ce qui concerne les embryons de souris, une seule cellule prélevée sur un embryon de huit cellules peut produire des lignées de cellules souches embryonnaires. La technique n'est pas aussi efficace que lorsqu'on obtient des cellules souches embryonnaires sur des embryons plus développés, mais Robert Lanza rappelle qu'il poursuit ses recherches sur ce point. Sur 125 tentatives, son équipe n'a pu obtenir que cinq lignées de cellules souches embryonnaires, alors que le taux habituel de réussite se situe autour

des 30 %. Il a montré que les embryons formés de sept cellules avaient autant de chances de survivre après leur implantation dans l'utérus de la mère porteuse que n'en avaient les embryons témoins non manipulés [et donc de huit cellules]. Le chercheur pense que les centres de procréation assistée pourraient utiliser des techniques similaires afin d'obtenir des lignées de cellules souches embryonnaires humaines dans le cadre des réglementations et des limites actuellement en vigueur.

Pourtant, cette méthode soulève aussi son lot de questions, souligne le spécialiste de la fécondation et des cellules souches embryonnaires John Gearhart, de l'université Johns Hopkins, de Baltimore. On ne sait toujours pas si, aussi bien chez les souris que chez les êtres humains, la cellule prélevée sur l'embryon précoce pourrait être capable de se développer en un organisme complet T- un jumeau génétique de l'embryon original -. Si c'était le cas, affirment certains, la technique serait elle aussi susceptible de détruire une vie potentielle. En outre, le recours à la biopsie recèle *"un risque mineur mais connu"*, pour-suit John Gearhart, non seulement pour l'embryon mais aussi pour la mère potentieïle, qui pourrait subir de nouvelles implantations après fécondation in vitro dans le cas où l'embryon ne réussirait pas à se développer. Partisans et adversaires de ces deux méthodes conviennent que la solution idéale serait de trouver une façon de reprogrammer une cellule de l'épiderme pour la transformer directement en cellule souche embryonnaire, sans impliquer d'embryon.*"Bien entendu, ce serait le Saint-Graal"*, assure Robert Lanza. George Daley, spécialiste des cellules souches au Children's Hospital de Boston, prédit que, grâce à la compréhension croissante des gènes qui contrôlent les cellules souches embryonnaires, une telle méthode finira par être Utilisable, *"Nous aurions alors une solution technique raisonnable"*, conclut-il. Une méthode sur laquelle tout le monde s'accorderait.

Gretchen VOGEL



10 % de tricheurs chez les scientifiques

Les affaires de fraude se multiplient chez les chercheurs. Le *New Scientist* s'en inquiète et exige un sursaut de la part de l'ensemble des acteurs.

NEW SCIENTIST (extraits) janvier 2006.

I1 y a deux mois encore, tout le monde pensait que 2005 resterait comme l'année du grand bond en avant pour le clonage thérapeutique. Aujourd'hui, il est probable que cette année restera dans les mémoires comme celle du plus grand scandale scientifique des Temps modernes, scandale qui plonge dans le désarroi l'ensemble de la recherche sur le clonage. L'invalidation des travaux du D^{r} Hwang entraîne des dégâts considérables pour la recherche sur le clonage thérapeutique.

Mais le pire, ce sont les dégâts causés à la science elle-même. Celle-ci repose sur une relation de confiance. Les gouvernements financent les chercheurs pour que ceux-ci utilisent les fonds de manière honnête et rapportent fidèlement leurs résultats. Les vérificateurs chargés de valider les articles soumis partent du principe que ce qu'ils ont *à* juger est le reflet fidèle de ce qui s'est réellement passé ; il n'entre pas dans leurs attributions de faire le tri parmi des données douteuses. En l'absence de confiance, c'est l'ensemble du projet scientifique qui risque de s'effondrer. Certains diront qu'il s'agit là d'une réaction excessive et que l'on ne peut formuler de si sombres prédictions à partir d'un cas isolé.

Or il ne s'agit pas d'un cas isolé. Il y a trois ans à peine, Hendrik Schön a été exclu de la communauté scientifique pour avoir présenté de fausses preuves dans 17 articles de physique. Cette année encore, le Massachusetts Instituts of Technology a exclu l'extravagant biologiste Luk Van Parijs après qu'il eut admis avoir contrefait des preuves. Des doutes subsistent au sujet d'au moins 3 des 40 articles qu'il a publiés au cours des huit dernières années, il semble que le milieu scientifique soit victime de façon endémique de comportements qui, s'ils ne sont pas aussi graves que ces cas emblématiques, n'en restent pas moins immoraux. Au début de l'année 2005, Brian Martinson et quelques collègues de Minneapolis ont publié dans *Nature* une étude portant sur plus de *3* 000 chercheurs financés par le National Institute of Health, l'institut de recherche médicale américain. Ils ont découvert qu'au moins 10 % d'entre eux reconnaissaient avoir dissimulé certains détails dans la méthodologie ou les résultats, s'être sciemment crédités de découvertes faites par d'autres ou encore avoir négligé certaines observations ou données sous prétexte qu'ils étaient *"persuadés"* qu'elles étaient erronées.

Dans ces conditions, que faire ? A tout le moins, il faudrait accorder une plus grande priorité à l'éthique de la recherche, tant dans la formation des jeunes scientifiques que dans les labos. Brian Martinson a également constaté que, parmi les scientifiques ayant répondu à son enquête, 1 sur 8 reconnaissait avoir négligé de relever les erreurs ou les interprétations douteuses de données dans les travaux d'autres chercheurs. Il est grand temps que les scientifiques adoptent une attitude plus offensive face à de tels faits, à l'instar de ce groupe de jeunes chercheurs sud-coréens qui ont dénoncé l'affaire Hwang.

La vérification, des publications par des membres de la communauté scientifique doit également être renforcée. Repérer les erreurs dans le papier invalidé de Hwang eût sans doute nécessité un examen particulièrement minutieux, mais les directeurs de revues scientifiques devraient exiger des preuves supplémentaires, surtout dans le cas d'articles annonçant des avancées révolutionnaires. Il est par exemple stupéfiant que *Nature* ait accepté l'article de Hwang sur Snuppy, le premier chien cloné, sans avoir consulté les données ADN originales.

Expression des gènes et cancer : une question de probabilité ?

Sait-on pourquoi un gène s'exprime ou non, au sein d'une cellule ? Depuis les années 60, l'explication la plus courante fait appel à la notion de programme génétique : il existe dans toute cellule des gènes dits répresseurs ou activateurs qui commandent ou répriment la production de protéines indispensables à la cellule. Depuis une quarantaine d'années, plusieurs démonstrations expérimentales sont venues remettre en question cette approche déterministe (une molécule-signal au niveau du gène détermine une réponse de la cellule) au profit d'une nouvelle théorie dite probabiliste : un gène a simplement une probabilité de s'exprimer ou pas à tout moment. Les travaux de chercheurs de l'Inserm, du CNRS et de l'Université Pierre et Marie Curie[9] apportent une contribution supplémentaire à l'édification de cette nouvelle théorie du développement embryonnaire. Le changement de paradigme que sous-tendent ces nouvelles données permet notamment de considérer les phénomènes de la cancérogenèse, sous un angle totalement nouveau. Ces travaux, publiés ce mois-ci dans la revue Progress in Biophysics & Molecular Biology, sont accessibles en ligne[10].

L'objectif du travail consiste à mieux cerner les règles qui gouvernent le comportement des cellules, notamment au cours des différentes étapes de l'embryogenèse, un processus qui aboutit à un organisme adulte, via une multitude d'interactions entre tissus.

Grâce à l'outil informatique, ces chercheurs ont pu tester plusieurs hypothèses formulées sur le comportement d'une ou de plusieurs cellules au cours de leurs différenciations. Leur hypothèse de départ : le mécanisme qui régit le comportement des cellules est fait, non pas de signaux programmés génétiquement mais d'événements aléatoires qui déclenchent l'activation des gènes contrôlant la différenciation d'une cellule. Les interactions entre cellules qui interviennent ensuite ne sont plus la cause de la différenciation comme admis jusqu'à présent via l'émission et la réception de molécules entre cellules, elles servent à stabiliser les cellules dans le phénotype qu'elles ont d'abord acquis aléatoirement.

Le modèle d'étude est constitué de la simulation de 2 types cellulaires A ou B choisis aléatoirement pour chaque cellule. La simulation informatique permet de complexifier très fortement les paramètres du système probabiliste tout en vérifiant que le modèle ainsi créé aboutit bien à une organisation cellulaire reproductible, comme c'est le cas au cours de l'embryogenèse d'un être vivant. Les paramètres suivis au sein de chacune des cellules et au moment de chaque simulation sont le nombre de molécules synthétisées, le nombre de molécules dégradées, la diffusion de ces molécules, la consommation de molécules par la cellule. À chaque étape de simulation, les événements cellulaires (différenciation, division ou mort) sont enregistrés.

L'aspect probabiliste du modèle est inclus par le biais des probabilités associées à chacun des phénotypes, lesquels apparaissent d'une couleur différente sur l'écran de l'ordinateur ; tandis que l'aspect stabilisateur provient de l'environnement qui agit en retour sur les cellules en modifiant ces probabilités associées à chaque phénotype.

Parmi les résultats notables (l'article fait 30 pages !), les chercheurs s'aperçoivent que lorsque les conditions de simulation informatique suppriment la mort cellulaire (les cellules continuent à vivre même si la quantité de nutriments nécessaire n'existe plus), la structure cellulaire se forme

[9] Jean-Jacques Kupiec, chercheur Inserm au sein du Centre Cavaillés de l'Ecole normale supérieure, Bertrand Laforge, maître de conférence à l'Université Pierre et Marie Curie (Laboratoire de physique nucléaire et des hautes énergies, CNRS/IN2P3, Universités Paris 6 et 7), David Guez et Michael Martinez

[10] A l'adresse : http://dx.doi.Org/10.1016/j.pbiomolbio.2004.11.004

correctement mais avec un taux d'échec de l'ordre de 50% (contre 20% seulement dans le cas où la mort cellulaire existe). La mort cellulaire semble donc jouer un rôle positif dans la formation de cette structure en bi-couche, caractéristique des êtres multicellulaires. La mort cellulaire élimine les cellules non adaptées à leur environnement, au profit... des plus viables. En somme, on est en présence d'une sorte de darwinisme transposé au niveau de la cellule.

Plus généralement, les simulations permettent de comprendre comment un ordre cellulaire correspondant aux tissus d'un organisme peut apparaître et comment ce même ordre peut être déstabilisé et donner lieu *à* une croissance cellulaire incontrôlée. En effet, les données permettent de comprendre le cancer d'une nouvelle manière. Plutôt que de considérer que les cellules se mettent à proliférer anarchiquement (et deviennent cancéreuses) sous l'influence de mutations altérant les signaux d'un programme génétique de contrôle de la prolifération cellulaire, les chercheurs français proposent une autre explication. L'organisation cellulaire résulte d'un équilibre quantitatif entre différents paramètres. En effet, les cellules structurées en bi-couche par le modèle arrêtent de proliférer spontanément, sans qu'aucun signal inhibiteur n'ait pourtant été intégré au système de simulation informatique. Les auteurs montrent donc que cet arrêt de croissance est dû à un équilibre entre les effets autostabilisateurs du phénotype (les cellules différenciées stabilisent leur propre phénotype) et les effets d'interdépendance pour la prolifération (les cellules différenciées stimulent la prolifération de phénotypes autres que le leur) exercés via les interactions entre cellules. Dès qu'une modification quantitative intervient dans un de ces deux processus, la croissance cellulaire est désorganisée, le cancer se déclenche. Autrement dit, si l'organisation tissulaire résulte de la combinaison de multiples causes, l'origine du cancer peut être diffuse. En fart, la croissance d'une tumeur est l'effet visible localement d'un déséquilibre entre l'ensemble des facteurs stabilisateurs impliqués dans l'environnement de la cellule.

|

Cette nouvelle manière de comprendre le cancer peut ouvrir de nouvelles stratégies de lutte thérapeutique. Ainsi, au lieu de pallier le déficit d'une protéine qui affecte la prolifération en « réparant » la mutation qui en est à l'origine, le modèle proposé suggère qu'il faudrait agir en rétablissant le ratio [entre les paramètres d'autostabilisation et d'interdépendance pour la prolifération (des paramètres dej diffusion de la protéine, par exemple).

« *L'apport de la biologie moléculaire et de la génétique n'est pas remis en question par ces travaux, mais il est nécessaire maintenant d'enrichir nos connaissances en intégrant l'ensemble des composants de l'organisation biologique*, expliquent les auteurs. *L'organisation cellulaire résulte d'un équilibre entre les influences du génome et des interactions cellulaires. L'embryogenèse est l'évolution de l'embryon soumis à cet équilibre. Le cancer est la destruction de cet équilibre* », concluent-ils.

> Pour en savoir plus :

- Source

"Modeling embryogenesis and cancer: an approach based on an equilibrium between the autostabilization of stochastic gene expression and the interdependence of cells for prolifération"

Bertrand Laforge(a), David Guez(a), Michael Martinez(b), Jean-Jacques Kupiec(c)

(a) Laboratoire de Physique Nucléaire et des Hautes Energies (LPNHE), Université Paris VI-Pierre et Marie Curie, Bureau 227, Tour 33RdC, 4 Place Jussieu, 75252 Paris Cedex 5, France

(b) Laboratoire de Physique Théorique des Liquides (LPTL), Université Paris Vl-Pierre et Marie Curie, Tour 24 2eme étage, Boite 121;, 4 Place Jussieu, 75252, Paris Cedex 5, France

(c) Centre Cavaillès et inserm, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France ***Progress in Biophysics & Molecular Biology, accessible online:*** **http://dx.doi.org/1**

Oj.1016/j.pbiomolbio.2004.11.004
Publication papier Avril 2005, volume 89/1

• Contact chercheur Jean-Jacques Kupiec Centre Cavailles, Inserm Ecole normale Supérieure, Paris

Tél:(01)44 32 29 61

Mail : kupiec@canoe.ens.fr

Projection de requêtes pour une recherche d'information intelligente sur le Web

Lina F.Soualmia 1,2 et Stéfan J. Darmoni1,2
1Laboratoire PSI-CNRS FRE 2645-INSA de Rouen-76131 Mont Saint-Aignan
{lina.soualmia,stefan.darmoni }@chu-rouen.fr
2
CISMeF & L@STICS-CHU de Rouen- 76031 Rouen
http://www.cchu-rouen.fr/cismef/

Résumé : *La recherche d'information sur le Web demeure problématique malgré l'existence de nombreux moteurs de recherche et de sites catalogues. Le Web doit faire face aux problèmes d'exhaustivité et de précision en recherche d'information. Le projet CISMeF (Catalogue et Index de Sites Médicaux Francophones) a été développé afin de faciliter l'accès à l'information de santé disponible sur l'Internet. La problématique d'aujourd'hui se veut aussi être une recherche d'information intelligente dans l'infrastructure du Web Sémantique, une extension du web actuel qui permettrait de rendre interprétable le contenu des ressources par les hommes mais aussi par les machines, grâce à des ontologies et des méta-données. La recherche d'information dans CISMeF est fondée sur une terminologie semblable à une ontologie et un ensemble de méta-données qui nous permettent de placer le projet à cheval entre le Web actuel qui est informel, et le Web Sémantique de demain. Nous proposons dans cet article d'utiliser trois types de ressources (base de connaissances morphologiques, base de règles d'association et ensemble de règles d'inférence) afin de donner les moyens à la recherche d'information de devenir intelligente. Nous détaillons les traitements nécessaires pour la construction automatique des ces ressources qui se basent sur le traitement du langage naturel, le data mining et le raisonnement sur les descriptions de concepts.*

Mots-clés : Recherche d'information, Traitement du langage naturel, Extraction de connaissances, Ontologies, Web Sémantique.

1 Introduction

La quantité d'information disponible sur le Web est importante et elle ne cesse de croître. Les catalogues et les moteurs de recherche en ligne (Yahoo, Google,Lycos…etc) permettent d'effectuer des requêtes par mot clé et de les affiner à l'aide d'opérateurs booléens. Avec des requêtes plus fines, le moteur peut renvoyer plus de documents et ainsi augmenter le rappel. Cependant, la tâche la plus lourde revient à l'utilisateur qui doit fouiller dans cette masse d'information pour sélectionner les documents qui lui seront les plus utiles. Les résultats ne sont pas tous pertinents et l'information retrouvée n'est pas complète. La recherche plein texte n'est pas toujours efficace : les fautes de frappe, les variantes lexicales et les synonymes sont considérés comme étant des termes différents.

Aujourd'hui, la problématique qui se pose est celle d'une *recherche d'information intelligente* sur le Web. Les moteurs de recherche actuels ne peuvent pas traiter *intelligemment* les pages HTML, langage le plus répandu sur le Web. Le Web Sémantique (Berners-Lee et al. 2001) est un espace d'échange qui reste à construire. Un de ses intérêts est d'une part d'apporter suffisamment de renseignements sur les ressources, en ajoutant des annotations sous la forme de *méta-données* et d'autre part, de décrire leur contenu de manière à la fois formelle et signifiante à l'aide d'une *ontologie* pour être interprétables aussi bien par les humains que par les machines. Cet espace doit être formalisé, le Web actuel étant informel. En effet, il est composéprincipalement de pages HTML écrites à la main ou générées automatiquement pour un traitement humain. Les ontologies et les méta-données sont donc deux éléments principaux pour la construction de l'infrastructure du Web Sémantique (Laublet et al. 2002). Une ontologie est une modélisation partagée d'un domaine pour améliorer la communication et éliminer les ambiguïtés

entre personnes, entre personnes etapplications ou entre applications. Elle est composée d'une hiérarchie de concepts, de relations entre concepts et d'un ensemble de règles ou de contraintes. Les méta-données font référence à une information descriptive des ressources du Web. Leur première utilité est la recherche d'information.

Plusieurs projets plus ou moins récents se basent sur l'utilisation d'une ontologie

pour décrire formellement des ensembles de documents ou de ressources. SHOE (Luke & Heflin, 2000) est l'un des précurseurs du Web Sémantique. Des ontologies et un langage basé sur HTML sont utilisés pour annoter sémantiquement des pages Web. OntoSeek (Guarino et al. 1999) est un moteur de recherche de pages Web qui utilise l'ontologie terminologique WordNet. Les documents et les requêtes sont représentés à l'aide de graphes conceptuels. Dans CoMMA (Gandon et al. 2002) des annotations sous forme de fichiers RDF (Lassila & Swick, 1999) sont associées à des documents d'entreprise en fonction des concepts de l'ontologie O'CoMMA.

Le contexte d'application de notre problématique de recherche d'information est le projet CISMeF[11](Darmoni et al. 2001) qui se fonde sur un ensemble de méta-données et une terminologie du domaine médical. Afin d'améliorer la recherche d'information au sein du catalogue nous étudions différentes techniques pour la projection de requêtes sur la terminologie. Les requêtes considérées sont celles qui sont saisies via une interface par les utilisateurs. Pour cela nous avons enrichi la terminologie d'une base de connaissances morphologiques en utilisant le traitement du langage naturel, d'une base de règlesd'associations grâce aux techniques de data mining et enfin nous allons formaliser la terminologie, à l'aide d'un langage de représentation des connaissances, et modéliser une base de règles d'inférences pour permettre un raisonnement sur le contenu des documents. Cet article est organisé de la manière suivante : tout d'abord nous décrivons en section (2) les méta-données et la structure de la terminologie utilisée dans CISMeF. En section (3) nous expliquons le processus de recherche d'information dans le catalogue tel qu'en place aujourd'hui ainsi que les problèmes rencontrés. Avant de conclure, nous détaillons en section (4) les méthodes utilisées pour l'enrichissement de la terminologie et l'extraction de connaissances pour l'expansion de requêtes.

2 Vers un web sémantique médical

Le projet CISMeF a été développé en 1995 pour assister les professionnels de santé, les étudiants et le grand public dans leur quête d'information en santé sur le Web. CISMeF et Doc'CISMeF, le moteur de recherche associé, prennent en compte la diversité des utilisateurs et leur permettent de trouver des documents de qualité qui répondent à un besoin précis. De nombreuses ressources (n=11,600) sont sélectionnées en fonction de critères stricts par une équipe de documentalistes et sont répertoriées selon une méthodologie de mise à jour du catalogue. Une ressource peut être un site Web, une page Web, un document, un rapport : tout support qui contient des informations relatives à la santé. La description de ces ressources se fait à l'aide de *notices* en fonction d'un ensemble de méta-données et d'une terminologie

structurée du domaine médical. Cette structure nous permet de placer le projet à cheval entre le Web informel d'aujourd'hui et le Web Sémantique de demain.

2.1 Les Méta-données

Les méta-données sont par définition des données concernant les données et dans le contexte du Web elles font référence à une information descriptive de sesressources. LeWeb a été initialement constitué pour un traitement humain et pour cette raison qu'il est difficile de tout y automatiser. Le concept de méta-données existait avant l'avènement d'Internet mais son intérêt a grandi avec le nombre de publications électroniques et de bibliothèques virtuelles. La solution

[11] Catalogue et Index de Sites Médicaux Francophones ; http://www.chu-rouen.fr/cismef/

proposée par le World Wide Web Consortium (W3C) est d'utiliser les méta-données pour décrire les données disponibles sur le Web. Dans le contexte du Web Sémantique elles

constituent un module fondamental et permettent notamment de faciliter la recherche d'information. Elles garantissent l'interopérabilité en assurant le partage et l'échange d'information rendant son contenu lisible et compréhensible par les machines.

Nous utilisons dans CISMeF plusieurs ensembles de méta-données parmi lesquels celui du Dublin Core (Baker, 2000) composé de 15 éléments. Les ressources indexées dans CISMeF sont décrites à l'aide de 11 éléments du Dublin Core : *auteur*, *date*, *description, format, identifiant, langage, éditeur, type de ressource, droits, sujet* et *titre*. Le Dublin Core ne permet pas de rendre compte de la qualité ou de la localisation d'une ressource. Pour pallier ce problème, 8 éléments spécifiques àCISMeF ont été définis : *institution*, *ville*, *province ou département*, *pays*, *public ciblé*, *type d'accès*, *coût* et *parrainage* de la ressource. Le type d'utilisateur est pris en compte. Pour les ressources destinées aux professionnels de santé (les lignesdirectrices et consensus de bonne pratique clinique) deux champs supplémentaires sont définis : *indication du niveau de preuve* et la *méthode* utilisée pour le déterminer.

Pour les ressources pédagogiques ce sont onze éléments de la catégorie « Educational » du standard IEEE 1484 qui sont rajoutés. Le format de ces méta-données est passé du langage HTML en 1995, à XML en 2000 pour permettre l'interopérabilité avec d'autres plateformes (e-learning du projet UMVF[12]). Depuis décembre 2002, le format utilisé est RDF un langage basique du Web Sémantique, et ce dans le cadre du projet européen MedCIRCLE (Mayer et al. 2003) dont le but est de qualifier la qualité des ressources d'information en santé et de guider les utilisateurs vers une information de confiance. Le vocabulaire des méta-données HIDDEL[13]est contenu dans une ontologie (représentée à l'aide du langage RDFS) et les ressources décrites en RDF en fonction des concepts de cette ontologie.

2.2 La Terminologie CISMeF

Les ressources sont indexées en fonction de la terminologie CISMeF. Celle-ci a été construite à partir des concepts du thésaurus MeSH[14](développé depuis 1960) et de sa traduction en français fournie par l'INSERM[15]. Le MeSH dans sa version 2003 est composé d'environ 22,000 *mots clés* (comme *abdomen*, *hépatite*) et 84 *qualificatifs* (comme *diagnostic, complications, thérapeutique…*) regroupés sous la forme d'arborescences. Les mots clés correspondent à des concepts médicaux et sont organisés sous la forme de hiérarchie à 9 niveaux allant du terme le plus général en haut de la hiérarchie aux termes les plus spécifiques en bas de la hiérarchie. Par exemple le mot clé *aberration chromosomique* est plus général que le mot clé *trisomie*. Les qualificatifs, organisés également en hiérarchie, permettent de préciser le sens des mots clés en limitant leur étendue à certains aspects. Par exemple l'association du mot clé *lombalgie* et du qualificatif *diagnostic* (notée *lombalgie/diagnostic*) permet de restreindre la *lombalgie* au seul aspect *diagnostic*. Bien qu'il existe des ontologies médicales générales, comme GALEN (Rodrigues et al. 1998), ou spécifiques à un domaine comme MENELAS (Bouaud et al. 1995), c'est le MeSH qui a été choisi car il correspond aux attentes des documentalistes et il est connu des professionnels de santé.

Les mots clés ont été regroupés dans CISMeF en fonction de spécialités médicales (n=66) intitulés *métatermes* (Cancérologie). Ce sont des super-concepts qui permettent une vision plus globale concernant une spécialité en offrant un niveau supplémentaire d'abstraction. Les

[12] Université Médicale Virtuelle Francophone ; http://www.umvf.prd.fr

[13] Health Information Disclosure Description Evaluation Language ; http://www.medcircle.org

[14] Medical Subject Headings. Le MeSH est produit par la US-National Library of Medicine pour la base documentaire Medline.

[15] Institut National de la Santé et de la Recherche Médicale http://dicdoc.kb.inserm.fr:2010/basimesh/mesh.html

métatermes permettent de connaître l'ensemble des termes MeSH qui sont répartis dans plusieurs arborescences mais qui concernent une même spécialité. Une hiérarchie de *types de ressources* (*n*=127) a été modélisée et elle permet de décrire la nature de la ressource (*cours, information patient*). Les métatermes et les types de ressources permettent d'exprimer des requêtes complexes dans CISMeF comme des '*recommandations en cardiologie'* ou encore des *'cours en virologie'* ce qui n'est pas possible avec la structure actuelle du MeSH.

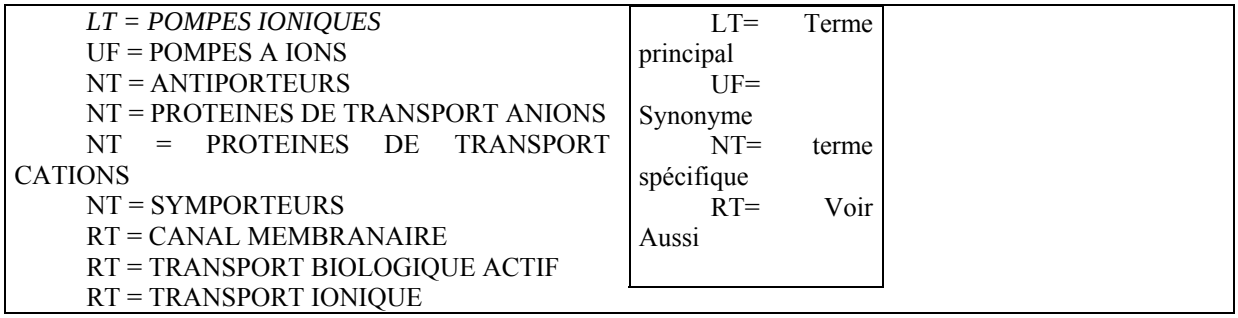


Fig. 1 – Exemple de termes et de relations dans les fichiers texte du MeSH

A partir du MeSH fournit sous la forme de fichiers texte (Fig.1) seules les relations du type '*est-un*' et *'partie-de'* sont utilisées pour définir des liens père-fils dans la hiérarchie des mots clés CISMeF (*n*=7,435 soit 34% du MeSH). Ces liens hiérarchiques sont exploités pour la recherche d'information et la navigation dans le catalogue. Par exemple le mot clé *Oreille* initialement défini comme étant *partie-de* du mot clé *Tête*, est défini dans la terminologie CISMeF par *Oreille* est fils de *Tête*. Les fichiers MeSH sont traités automatiquement pour renseigner la terminologie CISMeF afin qu'elle soit exploitable au niveau du site.

3 Recherche d'information

3.1 Processus de recherche d'information

La structure de la terminologie est exploitée pour l'indexation des ressources, la visualisation et la navigation dans les hiérarchies des termes du domaine, la recherche de ressources par le moteur Doc'CISMeF. Différents modes de recherche d'information sont possibles. La recherche *simple* permet à l'utilisateur de saisir une requête en texte libre en français ou en anglais. La recherche *avancée* engage des recherches plus pointues à l'aide d'un formulaire contenant des listes déroulantes et permet de combiner plusieurs champs (mots clés, titre, année…etc.) avec desopérateurs booléens (ET,OU,SAUF). La recherche *logique* s'effectue à l'aide d'un langage de requêtes associé, des opérateurs booléens et des caractères spéciaux.

La recherche simple telle qu'en place aujourd'hui se base sur les relations de hiérarchie entre termes. Si le terme (un mot ou une expression) saisi par l'utilisateur est un terme existant dans la terminologie, le résultat de la requête est l'union de toutes les ressources indexées par ce terme et par ses fils directs ou indirects de toutes les hiérarchies dans lesquelles il peut se trouver. Par exemple une requête sur le terme *tumeur* va renvoyer comme réponse l'ensemble des ressources rattachées à *tumeur* mais également celles rattachées à *tumeur colon*, *tumeur rectum.*…etc. De même qu'une requête sur *tête* va renvoyer les ressources rattachées à *tête* mais également à *oreille, nez…*etc. Si le terme saisi par l'utilisateur n'est pas un terme réservé, une recherche sur tous les autres champs de méta-données est effectuée, voire en plein texte sur tous les documents indexés en risquant de retomber dans les mêmes problèmes de bruit qu'un moteur de recherche plein texte. Ce type de recherche simple nécessite donc une bonne connaissance des termes de CISMeF, ce qui n'est pas évident pour un utilisateur novice.

3.2 Problèmes de la recherche d'information

La (ou les) requête(s) saisie(s) par l'utilisateur correspond(ent) rarement à la formulation exacte effectivement utilisée pour l'indexation. Nous avons extrait les requêtes des utilisateurs, à partir des logs du serveur http du moteur Doc'CISMeF, et déterminé le type de requête employé ainsi que le nombre de réponses obtenu, entre le 15/08/2002 et le 06/02/2003 (Table 1). 1,522,776 requêtes ont été extraites. 892,591 requêtes (58.62%) ont été soumises via l'interface de recherche simple et 365,688 (40.97% des requêtes simples) ne renvoient aucune réponse. Une analyse plus fine des requêtes simples (Table 2) nous a permis de déduire que 12.01% des réponses sont nulles, non pas du fait qu'elles correspondent à des requêtes erronées (ce sont bien des termes réservés), mais du fait qu'aucune ressource ne leur est rattachée

.

Table 1. Analyse des requêtes des utilisateurs du 15/08/2002 au 6/02/2003.

	Requêtes		**Requêtes Nulles**	
Type de Requête	**Nombre**	**Pourcentage**	**Nombre**	**Pourcentage**
Simple	892 591	58.62 %	365 688	40.97 %
Autre	630 175	41.38 %	144 790	22.97 %
Total	1 522 776		510 478	

Table 2. Répartition des requêtes simples à 0 réponse

	Nombre	**Pourcentage**
Expression reconnue	43 922	12.01 %
Expression non reconnue	321 766	87.99%
Total	365 688	

Afin d'améliorer ce type de recherche d'information qui est le plus utilisé dans le catalogue, nous proposons d'appliquer et d'évaluer la contribution de trois méthodes. Nous détaillons les pré-traitements de données nécessaires.

4 Améliorer la recherche d'information

4.1 Traitement du langage naturel

Les connaissances morphologiques sont utiles pour la recherche d'information et leur apport a été démontré dans plusieurs travaux (Gaussier et al. 2000) (Savoy, 2002) pour retrouver des expressions différentes qui dénotent des notions identiques ou proches. A partir d'un mot, on peut obtenir trois formes de variations. La *flexion* produit les pluriels, féminins lorsqu'il s'agit d'un nom et les conjugaisons lorsque c'est un verbe. La *dérivation* produit la forme adjectivale d'un nom. Enfin la *composition* combine plusieurs noms. Les connaissances morphologiques (flexions,

dérivations, compositions) d'un mot donné constituent sa *famille morphologique*. Il existe pour le domaine médical un lexique de mots dérivés en langue anglaise qui est le Specialist Lexicon de l'UMLS[16](Lindberg et al. 1993) mais aujourd'hui aucune ressource de ce type n'est disponible pour le français médical. Nous souhaitons utiliser ce type de connaissances pour améliorer la recherche d'information en réalisant une expansion de requêtes. Le but est de construire cette ressource morphologique pour la terminologie CISMeF.

Une étude préliminaire (Zweigenbaum et al. 2001) a été réalisée sur un ensemble de requêtes lancées sur Doc'CISMeF. Les résultats ont montré que l'utilisation de connaissances morphologiques amélioraient sensiblement les résultats des requêtes en diminuant le nombre de réponses nulles. La base de connaissances morphologique a été construite automatiquement dans des travaux antérieurs et l'algorithme proposé consiste à corriger la requête de l'utilisateur (dans le cas de non réponse seulement) en éliminant les « mots vides » (*comment*, *alors*, *du* …etc.) et en remplaçant chaque terme de la requête par une disjonction de tous les termes de sa famillemorphologique. Par exemple le terme *Cœur* a comme flexion *Cœurs*, comme dérivation *Cardiaque*, et comme composition *Cardiovasculaire.* Si l'utilisateur saisit la requête *interaction entre médicaments et alimentation* l'algorithme permet de reconnaître le mot clé *interaction aliment médicament*. Cette première ressource n'a pas été construite en fonction des termes de CISMeF et elle contient 6,312 couples (mot | mot dérivé). Après comparaison avec le sous-ensemble des termes de CISMeF utilisés pour l'indexation des ressources, nous avons obtenu 646 termes dérivés qui couvrent 608 mots clés, 30 qualificatifs et 8 types de ressources.

Nous avons analysé au préalable la structure de la terminologie de CISMeF (Table 3) relative à la composition des termes réservés. Il nous a semblé plus logique pour cette étude de ne considérer en premier lieu que les termes utilisés pour l'indexation des ressources car même si la requête d'un utilisateur correspond à un terme réservé, le résultat sera nul s'il n'existe pas de ressource qui lui soit rattachée.

Table 3. Structure des termes utilisés pour l'indexation

Nombre de mots	**Mots Clés**	**Qualificatifs**	**Types de Ressources**	**Termes**
1	1 437	55	28	1 520
2	1 706	10	42	1 758
3	612	11	39	662
4	148	3	12	163
5	40	--	4	44
6	8	--	2	10
7	2	--	--	2
TOTAL	**3953**	**79**	**127**	**4 159**

[16] Unified Medical Language System

Dans un second temps nous avons complété nos données grâce à la ressource terminologique Lexique (New et al. 2001). Elle comporte tout le lexique du français contemporain déduit à partir d'un corpus de textes, écrits entre les années 1950 et 2000, en se basant sur des calculs de fréquences d'apparition des mots contenus dans des pages du Web. Grâce à cette ressource terminologique, nous avons obtenu 34,710 variantes morphologiques et couvert exactement 1,300 termes de CISMeF (1,222 mots clés, 53 qualificatifs et 25 types de ressources). Toutes les variantes, y compris les verbes et subjonctifs, sont présentes et la liste est relativement complète.

L'analyse des termes composés de 2 ou plusieurs mots nous a permis de déduire que 1,935 termes étaient 'semi-couverts' (1,899 mots clés; 8 qualificatifs; 28 types de ressources). On considère qu'un terme est semi -couvert si au moins un des mots qui le composent est couvert. Par exemple *accident circulation* est un mot clé composé d'un terme dérivé du mot clé *accidents* qui a comme famille : *{accident, accidents, accidenté, accidentés, accidentées, accidentel, accidentels, accidentelle, accidentelles, accidentellement, accidenter}*. Celle-ci existant déjà dans la base grâce à l'étape de reconnaissance des termes, on considère que le terme *accident circulation* est semi-couvert. Il nous reste donc à compléter cette base de connaissances morphologiques sachant que l'appariement à plusieurs mots est plus exigeant (Zweigenbaum et al. 2001).

Table 4. Couverture du vocabulaire

	Mots Clés	**Qualificatifs**	**Types de Ressources**	**Termes**
Nb termes couverts	1 405	54	25	**1 484**
Couverture 1 mot	97.77%	98.18 %	89.28 %	**97.63%**
Semi-couverture	83.58%	78.48 %	41.73 %	**77.59%**
Couverture exacte	35.54%	68.35 %	19.68 %	**35.68%**

Les résultats que nous avons obtenus dans (Grabar et al. 2003) montrent qu'une normalisation des requêtes et de la terminologie augmente sa couverture : en effet si le mot clé est *accidenté*, la requête *Accidentè* sera nulle. Nous avons donc désaccentué et mis en minuscule tous les termes dérivés obtenus. La gestion des accents et minuscules est également effectuée sur les requêtes au niveau du prototype de recherche développé pour la réalisation des différents tests. A présent, l'algorithme permet de déduire à partir de la requête simple *douleurs dorsales* la requête logique douleur.mc ET dorsalgie.mc, avec mc indiquant que le terme considéré a été reconnu par l'algorithme de recherche comme étant un mot clé.

Nous avons également extrait de Lexique tous les termes qui peuventcorrespondre à des mots vides. Les mots vides sélectionnés sont tous les adjectifs possessifs (*mon*), les conjonctions (*mais*), les déterminants (*du*), les interjections (*diantre*), les prépositions (*durant*), les pronoms personnels (*il*), les pronoms possessifs (*leur*) et les pronoms relationnels (*auquel*). Nous avons déterminé ainsi 873 mots vides supplémentaires aux 473 initiaux, nous donnant un total de 1,346 mots vides. Ce nombre est élevé vu que des termes comme *boum, bye, bravo* ou encore *sniff* sont

considérés comme vides. La requête *douleur du bas du dos* est ainsitransformée en douleur.mc ET dos.mc

.

En plus de connaissances morphologiques, des connaissances sémantiques sont nécessaires. Par exemple le terme médical correspondant à *fausse couche* est *avortement spontané*. Nous étudions actuellement les logs des utilisateurs et collaborons avec des associations de patients et la Ligue Nationale contre le Cancer pour compléter la liste des synonymes CISMeF.

4.2 Data mining

Nous souhaitons découvrir de «nouvelles» connaissances à partir de la base de données CISMeF (en particulier à partir des notices et des termes) qui seront exploitées dans le processus de recherche d'information. Nous appliquons une technique de Data Mining appelée *Règles d'Association* dans le but d'extraire des associations intéressantes, non triviales, précédemment inconnues à partir de la base. Les règles d'association ont été initialement utilisées en analyse des données puis en fouille de données dans les bases de données relationnelles de grande taille (Agrawal & Srikant, 1994). Ces règles d'association utilisées dans un contexte d'expansion de requêtes permettent d'améliorer les performances de recherche d'information(Haddad et al. 2000).

Nous nous intéressons à la découverte de règles d'association booléennes. Une règle d'association booléenne RA est de la forme :

RA : $a_1 \Lambda a_2 \Lambda \ldots \Lambda a_i \Rightarrow a_{i+1} \Lambda \ldots \Lambda a_n$ (1)

Elle s'interprète intuitivement de la manière suivante : si un objet possède les attributs $\{a_1,\ldots,a_i\}$ alors il a tendance à posséder également les attributs $\{a_{i+1},\ldots,a_n\}$. Le *support* d'une règle représente son utilité. Cette mesure correspond à la proportion d'objets qui contiennent à la fois l'antécédent et le conséquent de la règle. La *confiance* représente sa précision. Cette mesure correspond à la proportion d'objets contenant le conséquent de la règle parmi ceux contenant l'antécédent.

Le processus d'extraction de connaissances est composé de plusieurs phases : la préparation des données et du contexte (sélection des objets et des attributs),l'extraction des ensembles fréquents d'attributs (*itemsets* fréquents par rapport à un seuil de support minimum), la génération des règles d'association les plus informatives à l'aide d'un algorithme de Data Mining (par rapport à un seuil de confiance minimum) et enfin l'interprétation des résultats (ou déduction de nouvelles connaissances).

Notre contexte d'extraction est le triplet *C=(O,A,R)* avec *O* l'ensemble des objets,

A l'ensemble des items, *R* une relation binaire entre *O* et *A*. Les objets sont les notices utilisées pour décrire les ressources indexées. Elles ont un identifiant unique. La relation *R* correspond à la relation d'indexation entre un objet et un item. Nous considérons pour l'instant deux cas différents pour les items :

- *A*={Mots Clés} ; *A* est l'ensemble des Mots clés.
- *A*={(Mot Clé, Qualificatif)} ; *A* est l'ensemble des couples (Mot Clé, Qualificatif).

Un itemset est fréquent dans son contexte *C* si son support est supérieur à un seuil minimal défini au préalable. Le problème de l'extraction des itemsets fréquents est de complexité exponentielle dans la taille *n* de l'ensemble d'items, le nombre d'itemsets fréquents potentiels étant 2^n. Dans le premier nous avons n=7,435. Les itemsets forment un treillis (Davey & Priestley, 1994). Plusieurs algorithmes de découverte d'itemsets fréquents ont été proposés. Le plus connu est l'algorithme Apriori (Agrawal & Srikant, 1994). Nous utilisons l'algorithme A-Close (Pasquier, 2000) dans lequel l'extraction se fait par le calcul des itemsets *fermés* fréquents avec l'opérateur de fermeture de la connexion de Galois d'une relation binaire finie (Ganter & Wille 2000). L'espace des itemsets à étudier est ainsi réduit. L'algorithme calcule aussi les

générateurs des itemsets fermés fréquents. Les générateurs d'un itemset fermé I_f sont les itemsets de taille maximale dont la fermeture est égale à I_f

.

Les bases pour les règles d'association sont déterminées à partir des itemsets fermés fréquents et de leurs générateurs. L'union de ces bases est un ensemble de générateurs non redondants de toutes les règles d'association non redondantes, d'antécédents minimaux et de conséquences maximales qui ne représentent aucune perte d'information. Ce sont les règles les plus utiles et les plus pertinentes.

Pour tester l'algorithme nous avons fixé le support à 5 documents et la confiance à 100%. La première étape de l'algorithme (itemsets de taille 2) nous a permis de trouver les règles suivantes :

- *hépatite C ⊐ sida* ; support = 14 (*A*={Mots Clés}).
- *sida/prévention et contrôle ⊐ condom* ; support = 6 (*A*={(MotClé,Qualificatif)}).

La seconde étape de cette étude est de déterminer toutes les autres règles d'association qui seront exploitées dans le processus de recherche d'information.

4.3 Raisonnement sur le contenu des documents

La terminologie CISMeF a la même utilité et structure qu'une ontologie terminologique (Sowa, 2000) :

- Le vocabulaire est bien connu des documentalistes et des professionnels de la santé et il correspond à celui du domaine médical.
- Chaque concept (Fig.2) a un terme préférentiel (Descripteur) pour l'exprimer en language naturel, un ensemble de propriétés, la définition en langage naturel permet quelquefois le différencier des concepts le subsumant et de ceux qu'il subsume, un ensemble de synonymes et un ensemble de règles et de contraintes

(Fig.3)

- Les concepts sont organisés selon une relation de subsomption allant du concept le plus général en haut de la hiérarchie au plus spécifique en bas de la hiérarchie.

D'après l'exemple de définition de la Fig.2, le terme associé au concept ayant l'identifiant unique D006521 est *Hépatite Chronique*. Le code cat.MeSH indique à quel niveau ce concept est situé dans la hiérarchie : on peut déduire que *Hépatite Chronique* (C06.552.380.350) est subsumé par *Hépatite* (C06.552.380). La Fig.3 est un exemple de contraintes sous la forme de règles à appliquer sur les concepts. Par exemple l'association *Hépatite/induit chimiquement* est équivalente (≡) au concept *hépatite toxique*. On peut considérer cette équivalence comme une règle d'inférence qui remplacerait toute association du concept *hépatite* avec le qualificatif *induit chimiquement* par le concept *hépatite toxique*. Ces règles n'existent pas sous format électronique mais nous comptons les modéliser et les exploiter dans le processus de recherche d'information.

Descripteur Francais: HEPATITE CHRONIQUE
Descripteur Americain: Hepatitis, Chronic
Code Cat MESH: C06.552.380.350
Synonymes Français: HEPATITE CHRONIQUE ACTIVE
Synonymes Américains: Chronic Hepatitis
Cryptogenic Chronic Hepatitis
Hepatitis, Chronic, Cryptogenic
Derives Americains: Hepatitis, Chronic Active
Active Hepatitides, Chronic
Active Hepatitis, Chronic
…
Hepatitis, Cryptogenic Chronic
MESH définition: A collective term for a clinical and pathological syndrome which has several causes and is characterized by varying degrees of hepatocellular necrosis and inflammation. Specific forms of chronic hepatitis include autoimmune hepatitis (HEPATITIS, AUTOIMMUNE), chronic hepatitis B; (HEPATITIS B, CHRONIC), chronic hepatitis C; (HEPATITIS C, CHRONIC), chronic hepatitis D; (HEPATITIS D,

CHRONIC), indeterminate chronic viral hepatitis, cryptogenic chronic hepatitis and drug-related chronic hepatitis (HEPATITIS, CHRONIC, DRUG-INDUCED).Numero NLM: D006521

Fig.2 – Exemple de définition de concept

Hepatitis : C06.552.380+
Viral Hepatitis = Hepatitis,Viral Human and Hepatitis,Viral Animal
/chemically induced = Hepatitis,Toxic
/veterinary = Hepatitis,Animal or Hepatitis,Viral Animal
hepatitis parenterally transmitted= Hepatitis C
hepatitis enterally transmitted = Hepatitis E
Non-A, Non-B hepatitis = probably Hepatitis C

Fig.3 – Exemple de contraintes sur les concepts.

Nous envisageons d'améliorer le moteur Doc'CISMeF pour lui permettre de réaliser une recherche intelligente de ressources dans le cadre du Web Sémantique. Pour cela nous proposons une approche semblable à celle de nombreux projets, qui est l'utilisation d'une ontologie formelle définissant les concepts et les relations entre concepts et un ensemble de ressources annotées en fonction des concepts et relations de l'ontologie. Il manque à notre terminologie une dimension formelle mais sa structure est telle qu'elle sera facilement traduite dans un langage de représentation des connaissances.

Nos données de départ sont un sous-ensemble de mots clés, qualificatifs et types de ressources ainsi qu'un ensemble de ressources. A cela nous ajoutons les règles et contraintes sur les mots clés, le réseau sémantique de l'UMLS qui est composé de concepts médicaux (n=134) et de relations (n=54) entre les concepts ainsi qu'un ensemble de *règles métier* (n=45). Celles-ci ont été recueillies auprès d'un médecin généraliste. Elles sont de la forme *Complications (Hépatite, Cirrhose)* indiquant que le concept *Hépatite* est relié au concept *Cirrhose* par la relation *Complications.* En analysant de plus près ces relations on remarque qu'elles correspondent aux qualificatifs du MeSH et que les concepts sont les mots clés du MeSH. Ces règles permettront entre autres d'enrichir l'ontologie car la seule information qui estdisponible dans la terminologie est que les concepts *Cirrhose* et *Hépatite* sont subsumés par le concept *Maladies du Foie.*

L'ontologie reposera sur un schéma RDFS qui définit les concepts (classes) qui sont les mots clés et types de ressources, les rôles (relations entre concepts) qui sont les qualificatifs et une relation de subsomption pour organiser les classes en hiérarchie. Les ressources seront annotées en fonction des concepts et des rôles de l'ontologie sous le format RDF. Les règles métier ne sont pas utilisées pour annoter les ressources et elles ne contribuent pas à la définition de concepts. Elles seront traduites sous la forme de règles d'inférence et utilisées pour un raisonnement sur le contenu des ressources dans le processus de recherche d'information. RDFS permet de définir des hiérarchies de classes et des propriétés mais il n'intègre pas de capacités de raisonnement, comme ceux qu'offrent les systèmes basés sur des langages formels comme les Logiques de Description. L'écriture des règles d'inférence n'étant pas possible en RDFS, nous utiliserons les fonctionnalités de l'outil TRIPLE (Sintek & Decker, 2001) qui a été développé pour une recherche d'information intelligente basée sur les connaissances. Il permet de réaliser des raisonnements complexes sur des ressources RDF instances de concepts en traduisant RDF en Horn-Logic mais aussi en DAML+OIL[17]. Un accès à des éléments externes comme le classifieur RACER (Haarslev & Möller, 2001) (basé sur les Logiques de Description) offre la possibilité de bénéficier de ses mécanismes de raisonnement. Pour la recherche de ressources c'est particulièrement utile. Par exemple si une ressource *a* est instance du concept *C:= ¤ Hepatite.Complications* et qu'un utilisateur recherche des ressources associées à *Cirrhose*, c'est à

[17] DAML+OIL joint committee. http://www.daml.org/2001/03/daml+oil-index.html.

dire instances du concept *Cirrhose*, le système déduira que la ressource *a* est également une réponse à la requête grâce à la règle d'inférence ¤ *Hepatite. Complications* ⊄ *Cirrhose*.

Fig.4

```
        // collection of resources
        @cismef:resources {
        cismef:doc1[
        meta:title->"Document 1 is related to Hepatitis and
        Aids";
        meta:author->"Toto";
        meta:keyword->HEPATITIS;
        meta:keyword->AIDS;
        meta:qualifier->COMPLICATIONS].
        cismef:doc2[
        meta:title->"Document 2 is related to Accidents";
        meta:author->"Doctor E in Risks";
        meta:keyword->ACCIDENT;
        meta:qualifier->RISKS].
        cismef:doc4[
        meta:title->"Document 4 is related to Cirrhosis";
        meta:author->"Titi";
        meta:keyword->LIVERCIRRHOSIS].
        }
        // domain ontology
        @cismefOntology {
        HEPATITIS[subClassOf -> LIVERDISEASES].
        LIVERCIRRHOSIS[subClassOf -> LIVERDISEASES].
        HEPATITIS[Complications->LIVERCIRRHOSIS].
        // ...
        }
        // requête toutes les ressources reliées à LIVERCIRRHOSIS//
        FORALL Resource, Author <-
search(Resource,LIVERCIRRHOSIS)@search
        (cismef:resources,
        cismefOntology) AND Resource[meta:author -> Author].
```

```
        compiling cismef.triple
        running cismef.triple
        ***
        Resource = cismef:doc1, Author
= 'Toto'
        Resource = cismef:doc4, Author
= 'Titi'
        done.
```

Exemple d'ontologie, ressources, requête et fichier résultat sous TRIPLE.

L'intégration de règles d'inférences à l'outil TRIPLE nous permettra de réaliser des requêtes de niveau supérieur dans CISMeF et les relations du réseau sémantique offriront une navigation sémantique plus riche que la navigation hiérarchique actuellement en place.

5 Perspectives et Conclusion

Nous avons abordé dans cet article la problématique de la recherche d'information sur le Web. Nous avons présenté certains aspects du projet CISMeF qui est notre contexte d'application pour la recherche intelligente d'information. Nous voulons pour cela lui donner les moyens d'être intelligente en construisant une base de connaissances morphologiques, une base de règles d'associations et en

formalisant la terminologie à l'aide d'un langage standard de représentation des connaissances. Les techniques de traitement automatique du langage naturel ont permis de construire une base de connaissances morphologiques. Le data mining permettra de découvrir des règles d'association entre concepts. Enfin le raisonnement sur les ontologies offrira un niveau supérieur tant au niveau de l'ontologie (vérification de la consistance et cohérence, exploitation du réseau sémantique de l'UMLS) qu'au niveau recherche d'information grâce à un ensemble de règles d'inférences. Nous pensons que la plus
value est dans la combinaison de ces différentes techniques.

L'évaluation de l'apport de chacune des méthodes se fera de deux manières. Tout d'abord par une expansion automatique (enrichissement) des requêtes pour élargir le champ de la recherche en utilisant chacune des ressources (base morphologique, règles d'association, et ontologie formelle) séparément puis conjointement. Les requêtes considérées sont celles du fichier log dont le nombre de réponses est nul.

Ensuite par une expansion interactive : nous demanderons à un échantillon d'utilisateurs abonnés au site d'évaluer, pour chaque requête qu'ils poseront,« l'utilité » des différentes suggestions de requêtes enrichies apportées par les différentes méthodes. Cette évaluation (expansion automatique ou interactive) à échelle réelle permettra d'établir une base de règles ou un protocole pour l'application des méthodes en fonction du type de requête posé.

Références

AGRAWAL R. & SRIKANT R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings VLDB Conference*, p.478-499.

BAKER T.(2000) A Grammar of Dublin Core. *Digital-Library Magazine.* 6(10).

BERNERS-LEE T., HENDLER J. & LASSILA O. (2001). The Semantic Web. *Scientific American* p.35-43.

BOUAUD J., BACHIMONT B., CHARLET J. & ZWEIGENBAUM P. (1995) Methodological Principles for Structuring an « Ontology ». *Proceedings of IJCAI conference*.

DARMONI SJ., THIRION B., LEROY JP., DOUYÈRE M. & al. (2001). A Search Tool based on 'Encapsulated' MeSH Thesaurus to Retrieve Quality Health Resources on the Internet. *Medical Informatics & the Internet in Medicine*, 26 (3):165-178.

DAVEY BA. & PRIESTLEY HA. (1994) Introduction to Lattices and Order. *Cambridge University*

GANDON F., DIENG-KUNTZ R., CORBY O. & GIBOIN A. (2002) Web Sémantique et Approche Multi-Agents pour la Gestion d'une Mémoire Organisationnelle Distribuée. *Journées Ingénierie des Connaissances,* p.15-26.

GANTER B. & WILLE R. (2000) Formal Concept Analysis : Mathematical Foundations. *Springer-Verlag*.

GAUSSIER E., GREFENSTETTE G., HULL D. & ROUX C. (2000) Recherche d'Information en Français et Traitement Automatique des Langues. *TAL,* 41(2) : p.473-493.

GRABAR N., ZWEIGENBAUM P., SOUALMIA LF., & DARMONI SJ. (2003) Matching Controlled Vocabulary Words. *Medical Informatics Europe* p. 445-450.

GUARINO N., MASOLO C. & VETERE G. (1999) Ontoseek : Content-Based Access to the Web. *IEEE Intelligent Systems* 14(3).

HAARSLEV V. & MÖLLER R. (2001) Description of the RACER System and its Applications. *Proceedings Internationl Workshop on Description Logics.*

HADDAD MH., CHEVALLET JP. & BRUANDET MF. (2000) Relations between Terms Discovered by Association Rules. *Practices of Knowledge Discovery in Databases.*

LASSILA O. & SWICK R. (1999) Resource Description Framework (RDF) Model and Syntax Specification. *W3C Candidate Recommendation 1999*.

LAUBLET P., REYNAUDC. & CHARLET J. (2002). Sur Quelques Aspects du Web Sémantique. *Actes des deuxièmes assises nationales du GdRI3*, p.59-78.

LINDGBERG DAB., HUMPHREYS BL. & McCRAY AT. (1993) The Unified Medical Language System. *Methods of Information in Medicine.*

LUKE S. & HEFLIN J. (2000) SHOE Project Specification.

MAYER MA., D ARMONI SJ., FIENE M., KÖHLER C., & al. (2003) MedCIRCLE -Modeling a Collaboration for Internet Rating, Certification, Labeling and Evaluation of Health Information on the Semantic World-Wide-Web. *Medical Informatics Europe* p.667-672.

NEW B., PALLIER C., FERRAND L. & MATOS R. (2001) Une Base de données Lexicales du Français Contemporain sur Internet: LEXIQUE, *L'Année Psychologique*, p. 447-462.

PASQUIER N. (2000) Data Mining, : Algorithmes d'Extraction et de Réduction des Règles d'Association dans les Bases de.Données *Thèse de doctorat*, Université Clermont-Ferrand II.

RODRIGUES JM., TROMBERT-PAVIOT B., BAUD R. & al. (1998) GALEN-In-Use : using Artificial Intelligence Terminology Tools to Improve the Linguistic Coherence of a National Coding System for Surgical Procedures. Cesnik et al. (eds). *MedInfo'1998*.

SAVOY J. (2002) Morphologie et Recherche d'Information. *Cahier de Recherche en Informatique,* CR-I-2002-01, Université de Neuchatel.

SINTEK M. & DECKER S. (2001) TRIPLE- An RDF Query, Inference and Transformation Language. *Proceedings of Deductive Databases and Knowledge Management Workshop.*

SOWA JF. (2000) Ontology, Metadata and Semiotics. *ICCS*

ZWEIGENBAUM P., GRABAR N. & DARMONI SJ. (2001) Apport de Connaissances Morphologiques pour la Projection de Requêtes sur une Terminologie Normalisée. *TALN.*