Exploration informationnelle et construction de connaissances en génomique

Gabriel Gallezot

Prépublication de :
Gallezot Gabriel, « Exploration informationnelle et construction des connaissances en génomique, Les Cahiers du numérique », Hermes, vol. 3, n°3-2002, pp 121-136.

Exploration informationnelle et construction de connaissances en génomique.

Fondé sur un exemple de réalisation de dispositif informationnel, nous présentons un mode de navigation qui, par l'exploration des données factuelles et textuelles, sert la construction de connaissances en génomique.

L'impulsion du programme de séquençage complet du génome humain (le HGP), puis de celui d'autres génomes, et enfin la généralisation et l'évolution concomitante des séquenceurs, ont fait croître de manière exponentielle la production des séquences d'ADN. Cette recherche d'exhaustivité, liée au fait que toute l'information génétique nécessaire à un organisme est contenue dans son ADN[1], a propulsé la biologie moléculaire dans l'ère de la génomique. Pour faciliter l'accès et le traitement des séquences biologiques, il y a eu nécessité de les enregistrer dans les banques, désormais en ligne sur Internet. Ces banques de séquences sont filtrées par des équipes de bioinformaticiens pour réaliser des entrepôts de données dédiés à des recherches spécialisées sur un organisme, un type de molécule…[2]

Aussi, à travers des pages HTML ou des *imagemaps* générées par des scripts CGI à partir de bases de donnée, à travers une fédération de liens inter-banques rendue possible par des *identifieurs* uniques et à travers une

---

[1] Pour un rapide rappel en biologie moléculaire le lecteur pourra consulter : D.O.E. ; [consultée en 02/98] " To Know Ourselves " [en ligne] URL : http://www.ornl.gov/hgmis/publicat/tko/index.html, en particulier : http://www.ornl.gov/hgmis/publicat/tko/03_introducing.html

[2] Gallezot G., « La recherche in silico » In : Chartron G. (sous la dir.) *Les chercheurs et la documentation électronique : nouveaux services, nouveaux usages*, Edition du cercle de la Librairie, Coll. Bibliothèque,Juin 2002

fédération de bases de données devenue réalisable par des approches d'interopérabilité de composants logiciels, c'est tout un espace réticulaire qui offre aux génomistes de nouvelles explorations informationnelles[3].

Nous montrons, à travers l'exemple de la détection de transfert de gènes, une possibilité de navigation dans un corpus de données textuelles et factuelles libre d'accès sur Internet et comment cette exploration informationnelle est source de créativité. Cet exemple de *data mining* s'appuie sur l'inspection de deux cartes superposées. La carte physique d'un chromosome et une carte d'états statistiques (homogénéité) de la séquence d'ADN. Ainsi la mise en évidence, par visualisation, de régions spécifiques, permet aux *génomistes* de créer de nouvelles connaissances.

1- La construction de connaissances, le document au cœur du processus.

Nous distinguons deux pôles, la productivité et la créativité. Où la productivité a trait aux connaissances réalisées à partir de processus standardisés, de protocole établis, de techniques éprouvées et où la créativité relève de processus nouveau, de protocoles expérimentaux et d'appropriations originales de techniques. Concernant l'exploration informationnelle et plus précisément la navigation nous insistons sur la créativité liée à la présentation de nouveaux documents[4].

Les idées ne peuvent se former que sur des constructions cognitives antérieures présentes sous forme d'information dans les documents. Nous inscrivons donc, la construction de connaissance dans un processus de transformation de l'information où la connaissance est la formation des idées, l'information est la mise en forme des connaissances (in-formation) et l'information inscrite sur un support constitue un document. Ce processus s'inscrit à son tour dans le cycle de l'Information Scientifique et Technique

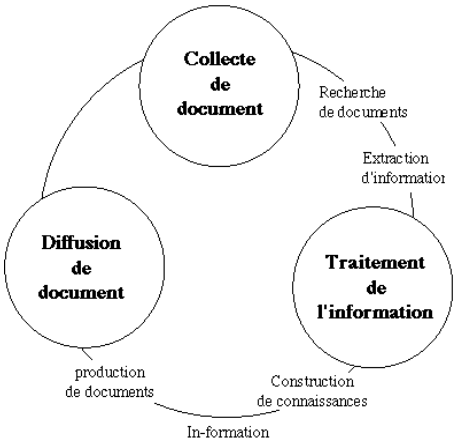
---

[3] Gallezot G., Sansom F., Brunaud V., Gas, S. et Bessière P., « Normes et standards dans le processus de traitement du document numérique en biologie moléculaire », *Solaris*, n°6, 2000. [En ligne]. http://www.info.unicaen.fr/bnum/jelec/Solaris/d06/6gallezo.html.

[4] Concernant la productivité de connaissances le lecteur pourra consulter : Gallezot G., « La recherche in silico », op. cit., pp.229-253

(IST)[5] (Cf. Figure 1). L'activité des chercheurs est circonscrite par ce cycle. La collecte peut être réalisée à partir de banques de données, de sites web, d'expériences dans les laboratoires, de "butinage" (*browsing*) dans les rayonnages d'une bibliothèque... Le traitement correspond à l'activité cognitive des chercheurs ou à des manipulations par des outils informatiques. La diffusion est définie comme l'ensemble des opérations nécessaires à la propagation des connaissances.

Figure 1 :Le cycle de l'IST, contextualisation de la construction de connaissances



Un cycle d'IST produit des connaissances et des techniques. Le bouclage de plusieurs cycles avec, à chaque passage, l'introduction de nouvelles connaissances et de nouvelles techniques nourrit le contexte socio-technique. C'est un référent commun en perpétuelle évolution à partir duquel s'effectue l'ensemble des activités scientifiques. En fonction de critères sociologiques, économiques et culturels, les chercheurs accèdent de façon différente à ce réservoir cognitif et instrumental. Si les chercheurs disposent des mêmes innovations techniques, de la même littérature scientifique, ils n'ont pas forcément les mêmes possibilités de les utiliser (pays industrialisés/non-industrialisés, budgets de la recherche, de laboratoires, de fonctionnement, etc.) et s'ils les utilisent, ils se les

[5] Le cycle de l'IST un modèle heuristique construit sur le cycle de vie du document (acquisition, recherche, archivage…) et de transformation de l'information (informations / documents / connaissances). Pour plus de détails sur ces concepts nous renvoyons le lecteur à : M. F. Blanquet, *Science de l'information et philosophie,* ADBS Éditions, 1997 ; Y.-F. Le Coadic, *La Science de l'information*, PUF, Coll. « Que sais-je ? », 1994 ; A. J. Meadows, *Communicating Research*, Academic Press, 1998.

approprient différemment en fonction de leurs acquis (*background*) socio-culturels : culture générale, spécialité scientifique, culture technique, etc.
On peut dégager deux types d'appropriation du contexte socio-technique. Une appropriation que nous appelons "collective", relative à une simple [possibilité de] utilisation du réservoir cognitif et instrumental et une appropriation que nous appelons "individuelle" relative à une *lecture* spécifique, créative du réservoir cognitif et instrumental. Les chercheurs les plus en phase avec le contexte socio-technique favorisent ainsi leur créativité. La mise en œuvre d'artefacts informationnels qui permettent une vision heuristique révèle par exemple cette dernière appropriation (cf. point 4).

Aussi, la construction de connaissances dépend directement de la façon dont se déroule le cycle de l'IST : de la façon dont les documents sont collectés, de la manière dont est extraite l'information, des moyens de traitement de cette information, de la capacité à produire des connaissances et de les *traduire* en information, des possibilités d'édition et des canaux de communication utilisés.

L'inventivité prend naissance dans les achoppements du cycle de l'IST. Quand il est impossible de rendre compte de phénomènes, un saut qualitatif doit être réalisé. L'inventivité donne alors naissance à l'invention, à la création. La création de quelque chose de nouveau, qui ouvre de nouvelles pistes de recherche, des nouvelles possibilités de collecte, de traitement et de diffusion de l'IST, qui donne du *souffle* à la construction de connaissances.

Les idées se forment par distinction, par opposition ou par association de connaissances, c'est une appropriation individuelle des connaissances. Ces connaissances sont présentes sous forme d'informations dans les documents qu'ils s'agissent de données factuelles ou de publications. Leur manipulation (collecte, traitement et diffusion) est majoritairement informatisée. Nous retenons alors le vocable de technologies *procognitives* employé par Licklider[6] pour signifier l'importance des outils de traitement de l'information qui servent la connaissance et nous reprenons bien entendu Bush quand il présente les technologies qui aident à penser, un système qui couple un corpus documentaire et un accès rapide et sélectif à l'information.

---

[6] Cité par Schatz dans " Information Retrieval in digital Librairies : Bringing search to the net " In *Science*, vol 275, 1997, pp327-333.

A revolution must be wrought in the ways in which we make, store, and consult the record of accomplishment.... It is not just a problem for the libraries, although that is important. Rather, the problem is how creative men think, and what can be done to help them think. It is a problem of how the great mass of material shall be handled so that the individual can draw from it what he needs instantly, correctly, and with utter freedom. Compact storage of desired material and swift selective access to it are the two basic elements of the problem.[7]

Si, les dispositifs de repérage de l'information servent l'inventivité et aident l'homme à devenir créatif, les possibilités de traitement et de diffusion aussi. L'utilisation de dispositifs informationnels spécifiques permet non seulement un repérage et un accès aisé à l'information, mais permet aussi des visualisations qui servent l'expertise des connaissances. Les images ou les cartes proposées par des artéfacts informationnels, comme celui présenté au point 4, sont des construits scientifiques issus du traitement de résultats antérieurs contenus dans des documents.

2 - Corpus de documents numériques pour la construction des connaissances : informations factuelles, informations textuelles et qualité des données.

Le type d'IST à traiter peut être résumé en deux catégories : les informations issues de la paillasse, de l'expérimentation et celles issues de la littérature. Bien évidemment l'une n'existe pas sans l'autre. Néanmoins, la première catégorie relève de la production de données brutes et la deuxième de production de connaissances sur ces dernières. Aussi, deux types de documents sont à considérer : les données factuelles et les données textuelles.

Les données factuelles

[7] Bush V., Science Is Not Enough, 1967, cité en introduction par Buck A.M., Flagan R.C., Coles B. [consulté le 09/09/2002] " scolar's forum : a new model for scholarly communication " [en ligne] URL : http://library.caltech.edu/publications/scholarsforum/

Elles sont issues de la paillasse ou des banques de séquences internationales. Ces banques, comme GenBank, EMBL ou DDBJ[8] pour les séquences d'ADN, donnent accès par FTP à leurs gisements de documents primaires, qui sont des fichiers informatiques de texte codé en ASCII[9]. Ces fichiers sont dits *à plat*, c'est-à-dire des fichiers bruts fournis sans outil d'organisation. Néanmoins, ils possèdent une nomenclature de description et constituent ainsi les enregistrements, les notices des banques. Chaque enregistrement[10] est organisé en champs, pour lesquels des descripteurs spécifient une information relative aux propriétés d'un objet biologique. Si chaque banque possède ses descripteurs ou ses étiquettes pour coder l'information suivant un format qui lui est propre, le contenu informationnel intrinsèque de l'objet biologique reste inchangé.

Les données textuelles

Elles sont représentées par la littérature scientifique au sens large du terme. Dans cet ensemble on trouve les articles des revues, les ouvrages scientifiques, les actes de colloques... Mais aussi les notices catalographiques des banques de données bibliographiques comme Medline (Pubmed) . Au même titre que les notices Genbank, la notice Medline[11] est un fichier ASCII avec des descripteurs spécifiques. On peut distinguer les éléments qui relèvent d'une notice catalographique classique (auteurs, titre, résumé, revue, date, descripteurs du thésaurus...) et ceux qui ont trait aux liens avec les objets biologiques (les gènes, l'enregistrement dans une banque de séquence). Récemment, ces notices ont fait l'objet d'un formatage XML[12] .

---

[8] GenBank : Genetic sequences data Bank, EMBL :European Molecular Biology Laboratory, DDBJ : DNA Data Bank of Japan

[9] American Standard Code for Information Interchange

[10] Exemple d'une notice EMBL [consulté le 09/09/2002] : http://www.ebi.ac.uk/cgi-bin/emblfetch?L09228

Exemple d'une GenBank via l'interface Entrez [consulté le 09/09/2002] : http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=nucleotide&list_uids=410114&dopt=GenBank

[11] Exemple de notice Medline/Pubmed [consulté le 09/09/2002] : http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Display&DB=PubMed

[12] eXtensible Markup language

Ainsi, avec la DTD[13] et les filtres *ad hoc*, il est devenu plus aisé de les manipuler. Nous insistons particulièrement sur la notice catalographique car des développements en bioinformatique l'utilisent dans le cadre de projets d'extraction automatique d'informations. Par exemple, comme la majeure partie des connaissances biologiques sur les interactions géniques n'est pas décrite dans les banques mais dans les articles scientifiques, l'exploitation des résumés des notices et plus généralement des textes scientifiques constitue un enjeu central dans la construction des modèles d'interaction entre gènes ou encore dans la contextualisation de données issues de l'expérimentation[14].

Insistons pour terminer sur le fait suivant : ce qui, dans d'autres disciplines scientifiques, est distinct et payant (données factuelles et références bibliographiques) est en génomique gratuit et groupé. Cette situation particulière et privilégiée rend la pratique de collecte singulièrement aisée et principalement réalisée sur Internet. Ce phénomène est relatif au principe de double publication. [15]

De la qualité des données

La gratuité des ressources informationnelles au travers des banques de séquences, des banques de documents secondaires et l'apparition du texte intégral (ex. Pubmed Central) confère à la génomique un statut particulier en terme de partage d'information par hyperliens. Une grande majorité des documents contenus dans les banques de données internationales sont liés par un numéro d'identification (*identifieur*). Le biologiste peut ainsi passer d'une référence bibliographique aux descriptions des séquences nucléotidiques[16] sur lesquelles s appuient la recherche présentée dans

---

[13] Definition Type Document : <!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle, 9th May 2000//EN"
"http://www.nlm.nih.gov/databases/dtd/pubmed_000509.dtd">

[14]Bessières P., Nazarenko, Nedellec C., *Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques. texte accepté pour* CIDE'2001. Nedellec C., Ould Abdel Vetah M., Bessières P., " Sentence filtering for information extraction in genomics, a classification problem ", *PKDD'01 - Proceedings of the Conference on Practical Knowledge Discovery in Databases*, Springer Verlag, Freiburg, September 2001.

[15] Gallezot G., Sansom F., Brunaud V., Gas S., Bessière P., op. cit.

[16] Quelques lignes de A.T.G.C., la représentation des nucléotides : Adénine, Thynine, Cytosine et Guanine

l'article, puis naviguer d'*identifieur* en *identifieur* pour atteindre un objet biologique ou une référence bibliographique.

Néanmoins la qualité des données des banques de séquences n'est pas sans reproche, les erreurs se propagent de collections primaires (banques internationale comme Genbank, EMBL) en collections secondaires (base de données spécifiques, sur un animal, une bactérie...) et rejaillissent par liens sur l'ensemble des autres entrepôts de documents. Elles doivent être contrôlées et nettoyées par des traitements de base, il faut détecter les erreurs et éliminer la redondance, pour donner la meilleure précision possible aux calculs. Les sources d'information et leurs mises à jour doivent être identifiées et tracées: comment ont-elles été produites, par quel type d'approche et par qui ? Ces erreurs relèvent d'annotations "mal réalisées", de problèmes de standardisation de noms d'entités biologiques. Bork et Bairoch[17] décrivent des erreurs de synonymie, d'homonymie, de saisie, de contamination dans les séquences biologiques et les annotations qui les accompagnent.

3 – Intégration, hyperliens, hypermedia et interopérabilité

Ce point indique brièvement comment l'IST est gérer en génomique. Nous présentons ainsi les conceptions techniques qui sous-tendent le système d'information qui a servi de plateforme pour la recherche présentée au point 4.

Intégration et SGBD

Le développement de collections d'informations a précédé et accompagne depuis leurs débuts les programmes d'études des génomes, pour constituer aujourd'hui une mémoire indispensable, partagée par les communautés de chercheurs en biologie. Leurs moyens d'accès ont connu des changements importants cette dernière décennie, notamment avec l'usage généralisé de systèmes de gestion de bases de données (SGBD), du réseau et des interfaces graphiques. Grâce à ces outils, les systèmes d'information

---

[17]Bork P., Bairoch A., "Go Hunting in Sequence Database But Watch Out for the Traps", *Trends in Genetics*, Vol. 12, n°10, 1996, p. 425-427

évoluent vers une meilleure sélectivité des recherches et une plus grande facilité d'interrogations.
L'un des enjeux des « programmes génomes » vise à relier l'ensemble les différents types de données expérimentales qu'ils produisent, d'abord entre elles, et ensuite avec les autres sources d'information disponibles. Ce processus contribue à l'enrichissement des connaissances sur les génomes, et les bases de données servent cet objectif. Elles structurent l'information dans un modèle de données interrogeable, et permettent d'en croiser la variété, pour élucider la fonction des gènes inconnus.
L'intégration de connaissances hétérogènes dans une représentation unifiée des données offre une plate-forme générique, qui permet de formuler des requêtes globales sur l'ensemble des informations disponibles dans le système. Elles s'appliquent de cette manière à tous les types de questions que l'on peut poser aux génomes, et la mise en œuvre concerne aussi bien des analyses automatiques, que la génération d'outils de visualisation interactifs. L'un des buts assignés à ce processus d'intégration consiste à rendre possible la détection de nouvelles corrélations, parmi une masse de données qui n'étaient jusqu'alors pas reliées entre elles dans un même système d'interrogation.

Interfaçage, hyperliens, hypermédia

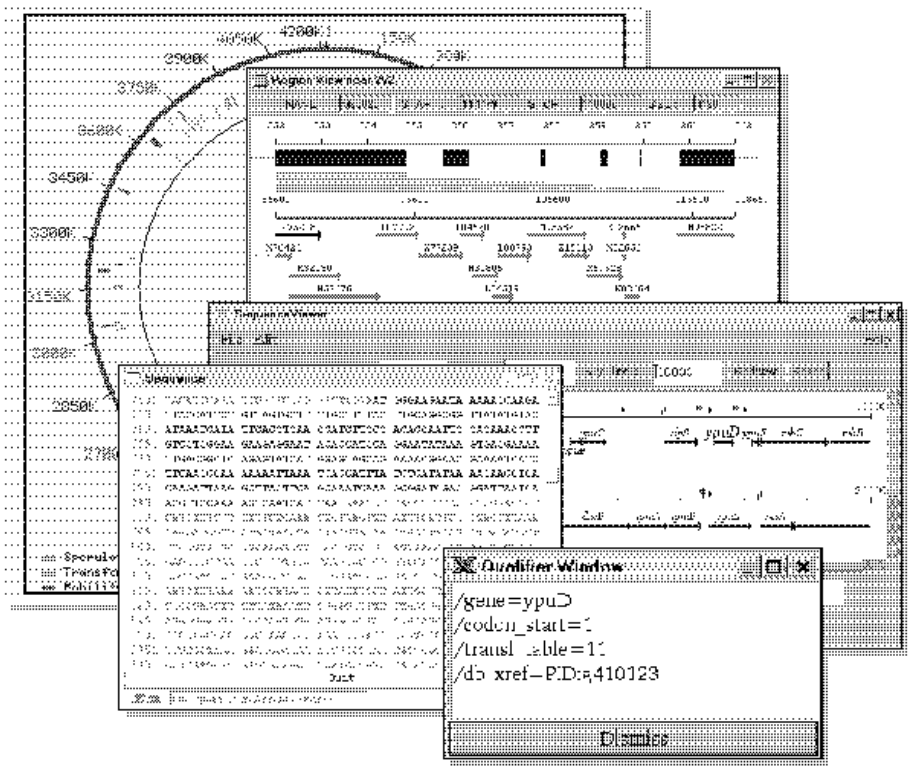
Si les réseaux internationaux existent depuis les années 70, leur utilisation ne s'est généralisée et banalisée qu'avec la naissance du WWW et des clients graphiques. Cet essor à naturellement conduit les biologistes a penser la diffusion du contenu de leur base de données sur Internet. l'interfaçage WWW/SGBD permet dans un premier temps l'interrogation de la base de données, puis dans un deuxième temps la génération, à partir des informations retournées de pages HTML, d'*imagemaps*, etc. Ces hyperliens permettent de se connecter, par un simple *clic* de souris à des banques de données internationales (GenBank, MedLine, SwissProt, etc.), mais aussi à d'autres éléments de la base de données. L'interfaçage permet ainsi d'obtenir une plus grande intéractivité, comme des hypermédia avec synchronisation des affichages, par exemple entre la carte physique et la séquence nucléotidique et la présentation des *qualifiers* correspondants (Cf. Figure 2).

Interopérabilité

Chacune des bases de données ne pouvant pas continuer à intégrer de plus en plus d'objets biologiques, il y a donc nécessité de fédérer ces bases. L'emploi des hyperliens sur WWW propose un premier niveau d'interopération entre les bases de données biologiques, et elles constituent des fédérations navigables de manière transparente pour les utilisateurs. Cependant, les liens présentent des limites, et ils ne remplacent pas la structuration de l'information dans une base de données. Aussi, de nouvelles approches d'interopérabilité des logiciels sont apparues conjointement au développement du réseau au cours de la décennie 90. Il s'agit d'effectuer une interface de communication de haut niveau entre les *composants* d'un logiciel, indépendante de leurs langages de programmation et de leur implémentation physique (exemple CORBA standard de l'OMG[18]) L'obstacle majeur de ce type d'interopérabilité ne relève pas de la mise en œuvre technique mais d'un problème de nomenclature, de classification, de relation, de définition des objets biologiques, en un mot d'un problème d'ontologie.

Figure 2 :Exemple d'interface utilisateur (le système d'information Micado)

[18]Object Management Group (http://www.omg.org/)



4 – Exploration informationnelle et création de connaissances

L'exemple de recherche proposé ci-dessous est issu de l'observation participante réalisée au sein d'un laboratoire de génétique des microbes[19]. La recherche présentée utilise un système d'information appelé Micado qui s'appuit sur l'ensemble des dispositifs informationnels de la génomique définis aux points 2 et 3[20].

---

[19] Gallezot G., Techniques de l'information, usages de l'IST et construction des connaissances des chercheurs en génomique, doctorat, Université de Paris 1 Panthéon-Sorbonne, sous la direction de S. Fayet-Scribe, septembre 2000.

[20] Samson F., Biaudet V., Gas S., Dervyn E., Gallezot G., Duchet S., Batto J-M., Ehrlich D. et Bessières P., « Micado, an Integrative Database Dedicated to the Functional Analysis of Bacillus subtilis and microbial genomics » In : Schumann W. .- *Functional Analysis of Bacterial Genes*, John Wiley & Sons, Ltd., 2000, pp.45-52

L'information dans Micado[21] est cherchée par les annotations (ex : gènes et autres *features* de l'ADN, références associées), des comparaisons de séquences (les programmes BLAST et FASTA), le parcours d'arbres de classification, et enfin la navigation sur les cartes génomiques, qui représentent le chromosome à différentes échelles. Les annotations figurent sur les cartes sous la forme de symboles graphiques cliquables. De cette manière la navigation autorise l'extraction d'une information précise à partir de grandes quantités de données. Ce moyen de visualisation fournit un aperçu global et fin des connaissances disponibles sur les génomes étudiés. La consultation des cartes en ligne a, par exemple, déjà servi à faire contrôler la qualité de la carte génétique de Bacillus subtilis par la communauté des chercheurs, avant sa publication. Les interfaces graphiques sont ainsi un composant essentiel du développement du système d'information. Combinées aux outils d'analyse des séquences, elles sont le support indispensable à l'élaboration de stratégies systématiques d'exploration des données, qui entrelacent des chaînes d'analyse et de classification automatique, avec des étapes interactives de traitement et de visualisation.
Un bon exemple concerne la recherche de transferts de gènes chez *Bacillus subtilis*, c'est-à-dire l'identification d'ADN étranger inséré dans le chromosome de la bactérie. Elle s'appuie sur l'inspection d'une carte physique graphique du chromosome. Celle-ci est construite avec l'ADN annoté extrait de la base de données, à laquelle est superposée l'affichage des états statistiques calculés sur les séquences[22] (Cf. Figure 3).

Cette recherche est un exemple de *data mining*, c'est à dire d'exploration de données à l'aide outils d'analyse statistique ou sémantique et d'interface de visualisation de résultats permettant d'expertiser un gisement de documents. Les segments flèchés avec les annotations de type *yxxX* (par exemple : *yvaW* ou *yvbW*) oureprésentent les gènes de la carte physique produit par Micado et la courbe noire superposée est obtenue à partir d'un traitement statistique (chaînes de Markov cachées) sur la totalité du génome[23]. Cette étude statistique est réalisée par le logiciel RHOM

---

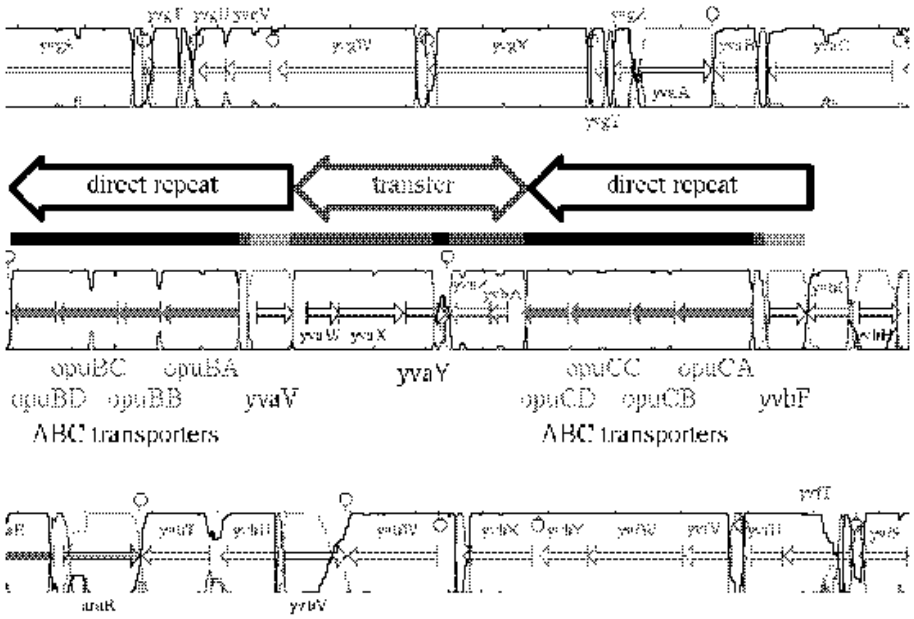
[21] http://locus.jouy.inra.fr/cgi-bin/genmic/madbase/progs/madbase.operl

[22] Bize, L., Muri, F., Samson, F., Rodolphe, F., Ehrlich, S.D., Prum, B. & Bessières, P. " Searching Gene Transfers on *Bacillus subtilis* Using Hidden Markov Chains ". In : *RECOMB'99 - 3rd Ann. Intl. Conf. on Computational Molecular Biology*, Lyon, France, 1999, pp.43-49.

[23] Dans la pratique la séquence du génome est segmentée pour un traitement efficient avec le logiciel.

(Research of HOMogeneous regions of DNA sequences) produit par les biométriciens de l'INRA. Il permet de détecter les régions homogènes (et par incidence hétérogènes) sur une séquences d'ADN. Les variations de la courbe (les pics) indiquent des "zones d'hétérogénéité" : les régions intergéniques, sauf dans le cas indiqué sur la figure par la double flèche < transfert >. Cette région n'est pas homogène par rapport au reste de la séquence ce qui permet aux biologistes de dire que les gènes de cette région ont potentiellement été transférés à partir d'un autre organisme.

Figure 3 : Visualisation de transferts potentiels de gènes



La recherche d'homologies de séquences sur ces derniers vient compléter cette supposition en indiquant que ces gènes sont des gènes de résistances à des facteurs toxiques (antibiotiques, métaux, etc.), enfin les régions de chaque coté de cette zone désignent des régions cibles dupliquées (*direct repeat*). Ainsi, l'expertise humaine, à partir des connaissances connues sur le transfert de gènes et la mise en évidence par visualisation de régions spécifiques de la séquence d'ADN, crée de nouvelles connaissances.

L'abondance de l'information génomique nécessite un repérage accru et efficient des connaissances, qui explique l'intérêt pour les techniques de

visualisation[24]. L'exploration informationnelle s'inscrit dans cette démarche. Plusieurs techniques peuvent bénéficier de cette approche, de l'analyse statistique sur du texte à la structuration et l'organisation de données dans des bases de connaissances (*knowledge bases*). Ce qu'il faut noter, c'est la généralisation de ce processus. L'extraction d'information sur un seul type de données en vue d'obtenir un résultat précis, ne suffit pas à rendre compte de situations complexes. La globalisation d'informations sur un sujet et la visualisation sous forme graphique des résultats d'un traitement réalisé par des techniques d'information offrent des *machines de vision*[25] capables de générer de nouvelles connaissances, de nouveaux projets de recherche ou d'autres éléments de réflexion… de créer. Ce sont des artefacts informationnels qui offrent une vision heuristique des résultats de la recherche en biologie moléculaire.
Les formes nouvelles de documents, que l'on peut qualifier de tertiaires, deviennent des adjuvants prépondérants et essentiels pour lire l'ensemble des documents disponibles. Ces construits sont la synthèse de résultats expérimentaux (l'ensemble des données factuelles) et de conceptions théoriques de la biologie (à travers la modélisation de la base de données et la création de liens). En revanche, les interprétations, les créations sont liées à des perceptions, des appropriations personnelles des représentations. Cette appréhension repose sur un *tiny bang*[26], un petit big bang individuel, une sorte d'explosion qui crée un nouvel espace dans lequel de nouvelles possibilités de connaissances sont offertes. Il dépend de la culture technique et informationnelle de chaque individu et sa capacité à l'intégrer dans son activité quotidienne, pour produire du sens sur les objets qu'il manipule.

Conclusion

Les modes d'exploration informationnelle

Pour tenter d'expliquer l'influence de l'exploration informationnelle sur la créativité en matière de construction de connaissances, il semble que deux

---

[24] Shneiderman, B. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley Publishing Company, Reading, MA, 1997.

[25]Virilio, P., *La machine de vision*, Ed. Galilée, Paris,1988.

[26] Balzt, C.. " Une culture pour la société de l'information ", *Documentaliste – Science de l'information*, Vol. 35, n°2, 1998, p.80.

dimensions soient à retenir : l'importance du contexte et le transfert de compétences dans une situation nouvelle. Le contexte est composé de dispositifs informationnels, c'est-à-dire de techniques manipulant de l'information sous forme essentiellement électronique. Le transfert de compétences dans une situation nouvelle est lié à la culture technique et informationnelle que les génomistes ont su acquérir.

Les modes d'exploration informationnelle qui influent sur la créativité relèvent plutôt des techniques de visualisation et de recherche de sens et sont souvent issues de réalisations personnelles (un ou plusieurs chercheurs). Ils peuvent être sériés selon deux paradigmes : *syntaxique* et *sémantique*. Ces derniers sont inscrits dans le temps et font référence aux possibilités de manipulations de documents proposées par Gardin[27] et Shatz[28]. Si leurs propos respectifs ont dix ans d'écart (87-97) et ne se situent pas exactement dans le même champ disciplinaire, leurs conclusions sont similaires. Trois phases sont distinguées :

- la réalisation de gisements d'informations à l'aide de banque de données, dans lesquels, ce sont les références au document qui sont cherchés. Il s'agit de stocker et de gérer des informations pour accélérer l'accès à l'information,
- l'approche statistique sur le texte créant des liens entre les textes ou entre les termes d'un texte (Schatz ajoute le concept d'hyperliens sur Internet). Il s'agit de classer, d'ordonner, de relier des documents, des informations, ou des données pour accélérer l'analyse de texte,
- l'extraction de connaissances dans des documents à l'aide de système expert (Gardin) ou de " *vocabulary switching* " (Schatz). Il ne s'agit plus de retrouver ou de classer des documents mais d'extraire du sens d'un gisement d'information.

Si les trois phases se préoccupent d'aider à la construction de connaissance, elles interviennent à des niveaux bien différents. Il y a un changement de paradigme : les outils informatiques passent de systèmes de gestion de fichiers *à plat* aux bases de connaissances ; le traitement des documents (au-delà du fait d'être passés de l'imprimé à l'électronique) ne sont plus

---

[27]Cité dans Le Coadic Y-F. "Creativity and Conduct of Research". In *Information technology and the research process* : proceedings of a conference held at Cranfield Institute of Technology, UK, 18-21 July 1989 / edited by Mary Feeney and Karen Merry. London ; New York : Bowker-Saur, 1990. pp.21-29.

[28] Schatz, B. R., op. cit.

manipulés comme un ensemble indissociable, mais comme une composition d'unités informationnelles, d'unités sémantiques extractibles ou *référençables*. Le paradigme *syntaxique* est relatif aux techniques avec lesquelles il est possible de traiter un texte à l'aide de termes ou d'expressions contenus dans ce texte, qu'il s'agisse d'une recherche de textes indexés dans une banque, d'analyse des cooccurrences de termes ou de balisage de texte. Le paradigme *sémantique* fait référence aux techniques avec lesquelles le sens des mots, des expressions d'un texte est analysé pour extraire des connaissances.

En génomique, le paradigme *syntaxique* représente l'analyse de la syntaxe des nucléotides sur un chromosome ; les banques de données qui organisent les relations entre les documents ; et la manipulation de documents qui s'effectue au mieux sur leur description et leur structure. Le paradigme *sémantique* représente l'appréhension globale d'un gène, de son génotype à son phénotype (caractères observables, c'est-à-dire le sens, la fonction des gènes) ; les bases de connaissances qui gèrent un ensemble de contenus de documents (des unités sémantiques) ; et la manipulation des documents qui tente d'accéder à la connaissance, au contenu du document (des représentations graphiques de l'analyse statistique à l'analyse sémantique de textes).

La serendipity dans l'exploration informationnelle

Il ne s'agit pas spécifiquement de naviguer ou déambuler[29], mais de fouiller les sédiments cognitifs accumulés depuis quelques années à la recherche d'information. Les visualisations proposées mettent en lumière des liens qui n'auraient pas pu être perçu autrement et peuvent faire sens auprès d'un expert.

Dans notre exemple, les hypermedia (les cartes superposées) qui aident à parcourir, ne relève pas exactement d'un choix, d'une sélection d'information, mais d'une composition aléatoire dirigée par des solidarités annotationnelles[30]. Le renouvellement des hypermédia proposés est lié à l'ajout de documents dans les banques séquences.

---

[29] Gasté D., « Navigation ou déambulation multimédia » In : *La communication Médiatisé par Ordinateur : un carrefour de problématiques*, Université de Sherbrooke, 15 et 16 mai, 2001.

[30] Bachimont B., « du texte à l'hypotexte les parcours de la mémoire documentaire », Technologie, Idéologies, Pratiques, n° spécial « Mémoires collectives », 1999.

Un agrégat d'informations intégré dans un artefact informationnel, comme Micado, n'étant pas connu *a priori,* la recherche d'information, *a posteriori,* peut retourner des documents auxquels le chercheur ne pensait pas. Plus encore, la mise en relation des unités informationnelles peut permettre de découvrir des corrélations insoupçonnées soit par lecture directe d'un résultat, soit par inspection d'un visuel. Le chercheur détecte des faits de façon quasi fortuite. Cette réécriture aléatoire et cette relecture fortuite relève de la *serendipity*.