Normes et standards dans le processus de traitement du document numérique en biologie moléculaire

Gabriel Gallezot, Franck Samson, Véronique Brunaud, Shahinaz Gas, Philippe Bessières

Revue SOLARIS
Janvier 2000
ISSN : 1265-4876

Normes et standards dans le processus de traitement du document numérique en biologie moléculaire

Gallezot Gabriel
Université Paris 1 - Panthéon-Sorbonne -17 rue de la Sorbonne - 75005 - Paris.
Mathématiques, Informatique et Génome - INRA - 78350 - Jouy-en-Josas.
gallezot@biotec.jouy.inra.fr

Samson Franck
Génoplante - INRA - 2 rue Gaston Crémieux - 91000 - Évry.
samson@evry.inra.fr

Brunaud Véronique
Génoplante- INRA - 2 rue Gaston Crémieux - 91000 - Évry.
brunaud@evry.inra.fr

Gas Shahinaz
Génoscope - Centre National de Séquençage - 2 rue Gaston Crémieux - 91000 - Évry.
sgas@genoscope.cns.fr

Bessières Philippe
Mathématiques, Informatique et Génome - INRA - 78350 - Jouy-en-Josas.
philb@biotec.jouy.inra.fr

Résumé

L'article présente les ruptures et les convergences dans le processus de traitement de l'Information Scientifique et Technique (IST), consécutives à l'émergence de normes et de standards des documents numriques. Le champ d'application concerne l'activité des chercheurs en biologie moléculaire, plus particulièrement la génomique. Pour appréhender les changements dans le processus de traitement de l'IST, trois types d'objets sont à considérer : les objets biologiques (au sens d'entités physiques, de concepts et de leurs relations : une fonction physiologique, une régulation, un gène, une molécule, *etc* .), les langages et les outils informatiques, avec la bioinformatique, et pour finir, les documents numériques. Pour signifier les ruptures et les convergences dans le processus de traitement de l'IST, les objets indiqués sont analysés selon trois phases : (i) Une phase de standardisation en amont, relative à l'émergence de standards ou de normes créant une rupture du processus de traitement de l'IST préexistant. (ii) Une phase de standardisation liée au processus de traitement de l'IST, qui montre un traitement possible de documents numériques au regard d'une certaine convergence des standards de la précédente phase. (iii) Une phase de standardisation en aval, qui met en évidence de nouveaux usages et de nouveaux besoins de normalisation, orientés vers l'intégration des connaissances, les environnements coopératifs, avec la création de projets et de groupes *ad hoc*. . Nous concluons cet article en soulignant certaines des caractéristiques des changements opérés.

Abstract

This article presents ruptures and convergences in the processing of Scientific and Technical Information (STI), issued from emerging standards and norms for digital documents. The area of application relates to the activity of the researchers in molecular biology, and more precisely genomics. To apprehend the changes in the processing of STI, three kinds of objects are to be considered: biological objects (that include physical entities, concepts, and their relationships: a physiological fonction, a regulation, a gene, a molecule, *etc* .), the computer languages and tools, and finally, the digital documents. To mean the ruptures and convergences in the treatment of STI, the above-quoted objects are analyzed according to three phases : (i) An upstream phase of standardization, relating to the emergence of standards or norms, creating a rupture in the pre-existing processing of STI. (ii) A phase of standardization related to the processing of STI, which shows a possible processing of digital documents taking into consideration convergence from standards of the previous phase. (iii) A downstream phase of standardization, which shows new uses and standardization needs, directed towards knowledge integration, cooperative environments, with *ad hoc* projects and groups creation. We conclude this article by outlining some features of the changes that operated.

Introduction

Normes et standards

Si le Petit Robert renvoie " norme " à " standard " et vice-versa, Éric Sutter (1997), dans le dictionnaire encyclopédique de l'information et de la documentation, n'utilise pas le mot " standard " (en français) dans sa définition de " norme ". Seule la traduction anglaise de norme renvoie à *standard* . Est-ce à dire que seule l'utilisation de " norme " est acceptée en français, ou bien que ces deux termes peuvent être employés indifféremment au gré d'une rédaction et ainsi éviter les répétitions ? Pour être en phase avec les acceptions courantes de "norme" et " standard ", il semble qu'il faille réaliser une distinction. " Norme " est définie par la publication (édition et prescription) d'un ensemble de règles, d'opérations, de protocoles, de caractéristiques généralement édifiées sous l'égide d'organismes institutionnels, comme l'ISO au niveau international et l'AFNOR en France [1]. " Standard " renvoie à " standard de fait " et peut se définir comme l'ensemble de règles, d'opérations, de protocoles, de caractéristiques couramment employés et consacrés par l'usage. Une autre terminologie reprend cette distinction avec norme *de jure* et norme *de facto* (CEVEIL, 1995) :

> *Une norme* de jure *(ou formelle) désigne une norme qui a été établie par un organisme légalement (ou formellement) constitué et mandaté pour élaborer et développer des normes. [...] Une norme* de facto *désigne une norme qui a été élaborée par une organisation autre qu'un organisme formel de normalisation. Une norme* de facto *peut provenir de plusieurs sources d'activités, variées par le type d'influence technologique, et variées par la nature diverse de leurs intérêts.*

Nous traitons trois types de standards : **les standards liés aux objets biologiques**, ceux **des langages et des outils informatiques** et ceux relatifs à **un processus de traitement de documents numériques**. Ce dernier type relève d'une démarche, d'une propagation de l'usage, et repose sur la mise en oeuvre et l'assemblage des précédents standards.

Ces standards sont généralement spécifiés, proposés et promus par des groupes d'intérêt. Dans le cas de **l'informatique**, domaine où ce processus est le plus apparent, il s'agira de l'IETF pour Internet, de W3C pour WWW, de l'OMG pour les technologies à objets et les environnements distribués (CORBA), et une partie fait l'objet d'un processus de normalisation par l'ISO (ex : HTML). D'autres, comme SGML pour les documents structurés, ou le modèle OSI pour les réseaux, ont été promus plus directement par l'ISO, ce qui ne préjuge pas de leur adoption. Les protocoles réseau de l'OSI en particulier n'ont pas remplacé Internet comme il était prévu. Par contre, ils fournissent le cadre de référence attendu pour comparer les différentes architectures réseaux en cours, et réaliser les interfaces.

Dans le contexte de la **biologie moléculaire**, les classifications, les nomenclatures, les protocoles, les notices des banques relèvent plutôt de standards. Par exemple, les problèmes de nomenclature sont pris en charge par des sociétés savantes comme l'IUPAC (pour la chimie) et l'IUBMB (pour la biochimie et la biologie moléculaire), et fédérées au niveau international par l'ICSU. Il existe ainsi des représentations standardisées pour les molécules, leurs noms et leurs abréviations, comme les acides aminés et les nucléotides.

Document numérique

Pour mettre en évidence les différents standards qui interviennent dans le processus de traitement des documents numériques en biologie moléculaire, nous avons choisi de présenter un type de document numérique : les notices des banques de données en génomique. Les exemples traités concernent essentiellement deux points de vue : le monde des microbes et les séquences d'ADN. Le premier illustre la démarche d'intégration des connaissances du fonctionnement d'un organisme. Le deuxième est transversal à tous les êtres vivants, car l'ADN est le support du texte de l'information génétique de toutes les espèces (ou presque...), et cette propriété lui confère un rôle central en biologie moléculaire et en génomique. D'autres documents numériques existent dans ce domaine, comme les documents personnels des chercheurs, certaines revues en ligne, la littérature grise ou encore des images scientifiques. L'objectif de ce point est de montrer que ces notices représentent plus que de simples enregistrements d'une banque relatifs à des documents primaires, elles sont elles-mêmes des documents primaires numériques qui constituent non seulement l'aboutissement d'une phase d'expérimentation, mais aussi une source primaire d'information.

Dans son article " What is a document? ", Buckland (1997) passe en revue les différentes définitions du document et conclut en commentaire par :

> *Briet's notion of document as evidence can occur in at least two ways. One purpose of information systems is to store and maintain access to whatever evidence has been cited as evidence of some assertion. Another approach is for the person in a position to organize artifacts, samples, specimens, texts, or other objects to consider what each could tell one about the world that produced it, and then, having developed some theory of its significance to place the object in evidence, or presented. In this manner, information systems can be used not only in finding material that already is in evidence, but also in arranging materiel so that someone may be able to make use of it as (new) evidence for some purpose.*

Cette notion du document de Susanne Briet est issue du célèbre aphorisme de l'antilope. Briet (1951) voit un animal dans l'antilope vivant dans son milieu naturel, et la considère comme un document quand elle est dans un zoo. Susanne Briet met ainsi la preuve (*evidence*) au coeur de la notion de document. Aussi, nous définissons les notices des banques génomiques [2] comme des documents, car elles sont la(es) preuve(s) de l'existence de séquences nucléotidiques spécifiques d'un organisme vivant. Elles renferment donc des informations premières issues de résultats de la recherche. Dès lors, elles ne doivent pas être perçues uniquement comme un enregistrement, mais en tant que documents primaires. Ces derniers sont stockés dans des banques de données, donc codés numériquement et manipulés par des outils informatiques, il est donc légitime de les nommer documents numériques (ou électroniques). Néanmoins, nous proposons avec Linda Schamber une définition du document numérique. Dans son article " What is a document? Rethinking the Concept in Uneasy Times " (SCHAMBER, 1996), insiste sur la notion d'usage et de " propriétés " du document :

> *Consisting of dynamic, flexible, nonlinear content, represented as a set of linked information items, stored in one or more physical media or networked sites ; created and used by one or more individuals in the facilitation of some process or project.*

Un document numérique peut donc être défini comme une représentation numérique d'une preuve, qui doit pouvoir être réutilisée (*reused*) dans un autre processus de traitement. Une troisième notion essentielle à la définition d'un document concerne sa localisation, il devra renfermer l'information nécessaire à son repérage dans une collection.

▲Traitement du document numérique

Nous pensons le processus de traitement des documents en termes génériques, c'est-à-dire en termes de collecte, de traitement et de diffusion de l'Information Scientifique et Technique (IST). Il définit, sous forme de chaîne ou de cycle, les étapes essentielles de la construction des connaissances d'une discipline. La collecte peut être réalisée directement à partir d'expériences dans les laboratoires, de banques de données, ou de recherches dans les rayonnages d'une bibliothèque. Le traitement correspond à l'activité cognitive des chercheurs qui, dans le cas d'un document numérique, repose sur l'emploi d'outils informatiques. Ces derniers concernent à la fois la mise en forme et l'accès à l'information, mais aussi les moyens d'analyse qui permettent de produire de nouvelles connaissances, qui relèvent dans ce domaine de la bioinformatique. Enfin, la diffusion est définie comme l'ensemble des opérations nécessaires à la propagation de ces connaissances.

▲Synergie des standards

Alors, quelles sont les ruptures et les convergences dans le processus de traitement de l'Information Scientifique et Technique (IST), consécutives à l'émergence de normes et de standards des documents numériques pour la biologie moléculaire et la génomique ? Le nouveau substrat de dispositifs informationnels et techniques rend possible le déploiement de nouvelles techniques et d'une nouvelle technologie [3] qui changent l'organisation du traitement de l'IST en biologie, et conséquemment, le processus de construction de connaissances, où les objets biologiques peuvent être manipulés *in silico* [4].

Nous avons décidé de distinguer trois phases : **la standardisation en amont, le processus de traitement numérique**, et **la standardisation en aval**. Chacune de ces phases est décomposée selon les **standards ou les normes relatifs aux objets biologiques**, aux **langages et outils informatiques** et aux **usages dans les processus de traitement**. Le découpage des standards dans ces trois phases suit une logique diachronique. Il met en évidence leur évolution au service de la mécanisation de niveaux d'abstraction de plus en plus élevés de l'activité humaine. Il montre les synergies qui opèrent entre les standards dans la conception de systèmes de traitement de l'information, tandis que ces derniers sont à l'origine de nouveaux dispositifs qui font émerger de nouvelles pratiques.

La première phase, que nous appelons " **standardisation en amont** ", constitue un point de rupture dans le processus de traitement de l'IST. Elle est circonscrite par des usages et des formalismes qui font l'objet de standardisations :

- emploi de nomenclatures, établies pour nommer et décrire les propriétés des molécules et des concepts biologiques (par exemple pour les noms de gènes), et dont les règles sont émises par différentes communautés scientifiques, organisées par champs de recherches (espèces, mécanismes biologiques) ; elles évoluent dans le but de favoriser les processus d'intégration, et offrent de nouvelles perspectives pour le traitement de l'IST, comme l'analyse automatique de texte.
- numérisation de la lecture du texte de l'ADN à la sortie des séquenceurs, dernière étape de protocoles opératoires de plus en plus automatisés ;
- collecte dans des banques de données, sous forme de documents numériques, de représentations des objets et des objets biologiques [5], et des notices catalographiques enrichies [6] qui leur sont associées (PROVANSAL, 1997) ;
- accord entre les éditeurs de banques de données et de journaux scientifiques, pour imposer l'enregistrement des séquences biologiques dans les collections internationales, en préalable et de concert avec la publication des résultats de la recherche ;
- généralisation des accès à Internet et des ressources informatiques, qui accompagnent les expériences, la mise en forme, la consultation et l'analyse des résultats.

La deuxième phase, " **processus de traitement du document numérique** ", montre comment la disponibilité publique (gratuite) de documents numériques, impulsée par les standardisations suscitées, a fait émerger un nouveau processus de traitement de l'IST. Les documents numériques, structurés dans des banques de données, sont traduits et intégrés dans des modèles de bases de données, proposant ainsi des plates-formes versatiles et puissantes pour la chaîne de collecte, traitement et diffusion de l'IST. Elles composent le noyau autour duquel s'organisent des coopérations à grande échelle pour analyser les génomes. La description d'un dispositif technique élaboré au sein du laboratoire de Génétique Microbienne de l'INRA de Jouy-en-Josas (BIAUDET *et al* ., 1997), permet d'illustrer le rôle des standards dans la construction de ce type de système d'information.

La troisième phase, appelée " **standardisation en aval** ", constitue un point de convergence dans le processus de traitement de l'IST. Elle met en évidence l'émergence de standards et de nouveaux besoins de normalisations :

- valorisation des savoir-faire du processus de traitement du document numérique, par une intégration progressive de l'ensemble des connaissances biologiques, relative au caractère encyclopédique induit par la nécessité d'une approche globale pour l'analyse des génomes (la génomique) ;
- spécification de standards à vocation universelle pour les objets, les concepts biologiques et leurs relations, favorisée par l'apparition de nouvelles normes de médiations informatiques ayant une portée générique, comme XML, CORBA, et encore plus récemment, l'intérêt porté aux ontologies ;
- les *Digital Libraries* (DL) et les Collaboratoires illustrent la synergie des standards dans les nouveaux usages des documents numériques, liés à leur traitement et à leur partage.

Si les deux premières phases de standardisation sont assimilables aux paradigmes " lexical " et " syntaxique ", la troisième tient plutôt du paradigme " sémantique ", en référence à un article de Schatz (1997). Il distingue 3 phases de progression pour la recherche d'informations en rapprochant les types d'objets (texte, document, concept) " retrouvables " et les générations de techniques *ad hoc* . Notre expérience au sein d'une communauté de chercheurs en génétique moléculaire des microbes montre que les pratiques professionnelles évoluent avec, et font évoluer les standards et les normes. Les normalisations relatives aux objets biologiques et aux procédés techniques associés concourent à des changements de paradigmes qui bouleversent ou font émerger de nouvelles pratiques dans les processus de collecte, traitement et diffusion de l'IST.

▲Présentation du contexte

Biologie moléculaire

Il existe plusieurs définitions de la biologie moléculaire, de la plus large à la plus restrictive. On pourrait parler de biologie moléculaire au sens littéral, dès lors qu'une activité de recherche en biologie aborde le niveau moléculaire de son champ d'investigations. À l'opposé, cette discipline s'est structurée et définie autour de l'étude de l'expression de l'information génétique, et de ses régulations, ce qui aurait tendance à la faire apparaître comme un domaine de la génétique. Entre les deux, elle est décrite comme l'étude des macromolécules biologiques : les acides nucléiques, dont l'ADN, support des gènes et de l'information génétique, et les protéines, produits de ces gènes et " ingénieurs " de la cellule biologique.

L'avancée des techniques en biologie moléculaire, notamment la lecture du texte de l'ADN (séquençage) qui compose les chromosomes et porte l'information génétique, peut être présentée comme un des points de rupture dans le traitement des documents en biologie moléculaire. L'évolution du procédé de séquençage a conduit à son automatisation, avec l'apparition des appareils de séquençage, appelés séquenceurs. Cet outil de " production de masse ", associé à et contrôlé par l'ordinateur, permet la numérisation directe des séquences sous la forme de documents, condition préalable à la poursuite de l'automatisation des tâches dans leur traitement. La numérisation est le point nodal, les techniques se créent et s'organisent autour de cette étape devenue aujourd'hui prépondérante. Les standards ou les normes en définissent les représentations, et accompagnent ce processus pour le rendre plus performant dans la collecte, le traitement et la diffusion de l'information.

À la sortie des séquenceurs, l'information est numérisée et stockée sur un support magnétique (disque dur). La séquence fait ensuite l'objet d'analyses bioinformatiques qui cherchent à identifier les signaux associés aux gènes et à leur expression, et comparent une nouvelle séquence avec celles déjà publiées, afin de tenter d'en élucider la fonction, sur la base d'une relation de similarité de leurs textes. Il existe deux niveaux de lecture, le texte de l'ADN, composé de quatre lettres, les bases ou nucléotides (A, T, G et C), et celui des protéines, produites par une " traduction " des gènes que portent l'ADN, et résultat de leur expression, en un alphabet à vingt lettres (les acides aminés). Ces deux alphabets composent le premier niveau standardisé de représentation de l'information génétique. Les résultats des analyses informatiques sont ensuite associés à leurs séquences, sous forme d'annotations textuelles structurées, qui en décrivent les propriétés, par exemple, les coordonnées d'un gène, sa fonction prédite, et les signaux (mots, motifs) qui conditionnent et régulent son expression.

La séquence et ses annotations sont ensuite déposées dans des banques de données internationales, par l'intermédiaire du courrier électronique ou de WWW. Les administrateurs formatent les données et renvoient un document avec un numéro d'enregistrement (" *accession number* " pour GenBank par exemple) au chercheur pour vérification. Le chercheur, après ce contrôle, décide de publier le document dans la banque, soit immédiatement, soit avec un temps de latence, pour l'associer à la publication d'un article. L'article relatif à une séquence ne sera accepté par les éditeurs que s'il existe le numéro d'enregistrement, preuve de sa prise en compte par les administrateurs des banques. Ce dernier point a pour origine la limite en volume de la revue sur papier, en effet l'impression d'un document relatif à une séquence peut se compter aujourd'hui en dizaines de pages. L'impulsion du programme de séquençage complet du génome humain, 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 par son ADN, 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.

Génomique

Une fois le séquençage de l'ADN rendu possible (1977), l'idée de lire toute l'information génétique chez l'homme émerge de plusieurs propositions successives, issues de colloques organisés entre 1984 et 1987. Ils contribuent à l'élaboration du Human Genome Project (HGP). Le colloque de l'Université de Californie à Santa Cruz, en 1985, se conclue sur la proposition de commencer à séquencer le génome humain. En 1986, une réunion à Santa Fe détermine les avantages de construire la carte génétique du génome humain. Le rapport *Human Genome Initiative* de l'*Health and Environnement Research Advisory Commitee* (HERAC, du DOE), en 1987, et celui de l'*Office*

of Technologies Assessment en 1988, intitulé *Mapping Our Gene* s, résument les différentes propositions qui serviront à la mise en place du HGP en 1988, et son démarrage en 1990 (WATSON, 1990).

Le terme " génomique " émerge de ces débats [7], et le HGP aura un effet mobilisateur, tant sur le plan des techniques, que sur l'impulsion de projets de séquençage d'autres génomes. Du fait de la taille réduite de leurs chromosomes, les microbes sont les premiers organismes vivants et autonomes qui ont fait l'objet d'une lecture complète de leur information génétique. Le premier résultat général et inattendu, était l'importance de la part des gènes de fonction inconnue détectés par cette approche exhaustive, et qui n'avaient pas été mis en évidence auparavant par les méthodes classiques de la génétique (DUJON, 1996). Cette observation a été à l'origine d'une deuxième génération de programmes de génomique, visant à élucider de la manière la plus systématique possible la fonction de ces " nouveaux gènes ". Elle a également contribué à entériner au sein de la communauté le concept de " génomique fonctionnelle " (HIETER et BOGUSKI, 1997). Il regroupe les approches biochimiques et physiologiques, adaptées pour des analyses à l'échelle d'un génome entier, et qui vont compléter les informations apportées par les séquences d'ADN. Ainsi sont mis en place de nouveaux moyens de production de grandes quantités d'informations, sur le même mode que le séquençage, et qui vont devoir être croisées pour prédire les fonctions des gènes.

Bioinformatique

Le phénomène sans doute le plus marquant, relatif au processus de traitement des documents numériques en biologie moléculaire est l'émergence de la bioinformatique :

> *[...] a variety of specialists - including geneticists, molecular biologists, informatics specialists, computer scientists, mathematicians, and statisticians - have worked together and expended the knowledge base of genetic information. (WELLER, 1996)*

Les efforts ont d'abord porté sur l'analyse des séquences et des structures, par des approches algorithmiques et mathématiques. C'est une fois de plus avec la possibilité de lire le texte de l'ADN que cette activité a pris de l'ampleur, et que ses outils se sont généralisés chez les biologistes. En même temps, l'organisation et la gestion de l'information, à savoir les données factuelles sur les objets biologiques, devenaient une nécessité. Aujourd'hui, le passage de l'imprimé à l'électronique commence à dessiner d'autres traitements des documents numériques. Encore à la marge de cette nouvelle discipline, et relevant des sciences de l'information, les données non factuelles, par exemple le texte des publications, commencent à être exploitées pour enrichir les connaissances en génomique.

La bioinformatique, traitement automatique de l'information biologique, s'est définitivement imposée avec les programmes d'analyses de génomes (BENTON, 1996). Cette discipline reprend tous les thèmes de l'informatique : l'acquisition, l'organisation de l'information, l'analyse, la visualisation, la modélisation, pour les appliquer à la génomique. Plusieurs revues lui sont spécifiquement dédiées, *Bioinformatics* (*CABIOS* ou *Computer Applications in the Biosciences* jusqu'en 1997) et *Journal of Computational Biology* , tandis que *Nucleic Acids Research* , non content d'en publier des articles, consacre des numéros spéciaux aux banques et aux bases de données en biologie moléculaire. Enfin, des revues généralistes comme *Nature* , *Science* , ou les séries des *Trends* et des *Current Opinion* , lui consacrent régulièrement des colonnes et des rubriques. Ce qui illustre l'effet " bioinformatique ", c'est le passage de l'imprimé à l'électronique. D'une simple lecture " au hasard " des revues disponibles dans leur centre de documentation, les biologistes sont passés à une recherche sélective et exhaustive de données factuelles, ainsi qu'à leur manipulation *in silico* .

1 - Normalisation et standardisation en amont

1.1 - Banques de données et Internet

Les séquences biologiques, dès qu'elles ont pu être établies, ont très tôt fait l'objet d'une compilation dans des banques de données, et ce dans le but explicite de fournir une plate-forme pour la comparaison de leur texte. Ce sont d'abord les séquences des protéines (les produits de l'expression des gènes) qui ont été collectées dans l'*Atlas of Protein Sequences* par Margaret Dayhoff (1965), et qui comprenait à son début 50 entrées. Il fut imprimé jusqu'en 1978, après quoi le volume des informations à intégrer imposait le recours à un support électronique. Ce premier travail est d'une importance capitale, en effet, il pose les règles de base pour les futures compilations de séquences :

- toutes les séquences sont numérisées et distribuées publiquement,
- les séquences sont vérifiées et corrigées (près de 15% de données révisées),
- les ordinateurs sont utilisés pour supporter cette réalisation,
- les séquences sont accompagnées de notes des auteurs, rapports, *preprints* , corrections de précédents travaux, *etc.* (WELLER, 1996).

L'organisation sous forme de banques de données numériques se généralise avec l'apparition du séquençage de l'ADN, et la création en 1980 de la banque de données européenne EMBL Data Library, puis en 1982 de son homologue américain GenBank, mis en place par le LANL, à la demande du NIGMS. Aujourd'hui, ces deux banques de séquences d'ADN sont maintenues respectivement par l'EBI et le NCBI, elles s'échangent systématiquement leurs contenus, et constituent une source d'informations primaires pour les biologistes. Il faut rajouter à ce chapitre deux autres réalisations importantes pour les séquences de protéines, PIR du NBRF, banque américaine qui voit le jour en 1986 et se présente comme la continuité de l'Atlas de Margaret Dayhoff, et celle plus récente de SwissProt à l'Université de Genève (1993), qui propose des séquences protéiques annotées dans un format EMBL. Il est significatif, à propos des problèmes de qualité de l'information, que SwissProt se soit constituée sur la base d'un travail de correction des séquences et des annotations de PIR, d'où son premier nom, PIR+ (BAIROCH, communication personnelle). Enfin, deux autres collections associées à ces banques, PDB pour les structures tridimensionnelles des protéines, et MedLine pour les références bibliographiques, composent le gisement universel d'informations pour la biologie moléculaire et la génomique.

La création de ces banques de données numériques en ligne est permise par les progrès importants réalisés, durant les années 80, dans la capacité de stockage, de traitement et de communication des ordinateurs. C'est durant cette période que le réseau BIONET est mis en place par le NIH (1984) et il fédère en 1987 quelques 489 laboratoires (KRISTOFFERSON, 1987). Toujours en activité sur Internet, le coeur de BIONET est composé d'un système organisé hiérarchiquement pour les discussions et les échanges d'informations, sous la double forme de listes de messageries et de conférences USENET (les *newsgroups*). Leur fonction couvre aussi bien l'échange d'idées, de recettes sur l'emploi de logiciels spécialisés, que la mise à jour automatique des banques de données installées sur des sites " miroir ". L'apparition de WWW, et sa généralisation en 1993, notamment grâce au nouveau client graphique Mosaic, va étendre les possibilités de connexion à de nombreux laboratoires disséminés dans le monde. Puis la généralisation des navigateurs et la standardisation de HTML permet aux banques de données de proposer des interfaces conviviales de saisie des informations : BankIt et Sequin pour GenBank, et Webin pour EMBL Data Library.

L'innovation technique, représentée par l'informatique, en particulier le réseau Internet et les différents services qui lui sont liés, a changé le processus de traitement des documents issus de la biologie moléculaire. L'usage d'applications comme le courrier électronique, la copie de fichiers avec FTP et la navigation hypermédia sur WWW, qui reposent sur le protocole de communication d'Internet, sont des standards qui forment les composants de base pour l'édification des systèmes

d'information employés aujourd'hui par les biologistes. Commencée avec les systèmes de messagerie, l'appropriation de ces moyens de communication par la communauté des biologistes a été précoce. Leur usage s'est rapidement généralisé pour l'accès aux collections d'informations (HARPER, 1994), et le phénomène a été particulièrement frappant avec l'apparition de WWW (HARPER, 1995).

1.2 - Processus de double publication

Pour publier un article sur une séquence dans une revue scientifique, il est impératif de " déposer " les données relatives à cette séquence dans des banques de données internationales, et ce, afin de recevoir un numéro d'enregistrement [8] qui permettra la publication dans une revue.

En 1984, dans l'"Instructions to Authors" du *Journal of Biological Chemistry* , il fut demandé, aux auteurs, pour la première fois, de publier leurs séquences directement dans une banque de données, comme GenBank ou EMBL. La même revue, un an plus tôt, informait les potentiels auteurs que leurs séquences ne seraient pas publiées dans la revue et qu'il n'était pas nécessaire de soumettre la séquence avec la proposition d'article :

> *In view of increasing numbers of submitted papers describing nucleic acid sequence...data obtained by well-established techniques... will generally not be published and should not be submitted with the manuscript*

On remarquera une précision d'ordre méthodologique, sur l'obtention de données réalisée par des " techniques bien établies ". Ce qui est mis en avant ici, c'est la qualité des données en rapport avec les techniques de séquençage.

En 1987, la revue *Nucleic Acids Research* tient les mêmes propos, mais demande aux auteurs de publier leur séquence dans EMBL, pour équilibrer la quantité de séquences avec GenBank. En particulier, la revue demande aux auteurs de posséder un *accession number* , pour pouvoir publier leur article : pas de dépôt de séquence, pas d'article. Le dépôt peut être " confidentiel ", c'est-à-dire que la séquence ne sera pas publique tant que l'article n'est pas paru.

En 1988 les banques de données, GenBank, EMBL mais aussi DDBJ, PIR et MIPS s'échangent leurs données. *Nucleic Acids Research* propose alors aux auteurs un *accession number* , en échange de l'envoi de leur séquence sous forme électronique, et que l'éditeur de la revue soumet lui-même aux banques. L'innovation qui consiste à déposer les séquences dans les banques en même temps que la soumission d'un article ne se fait pas sans controverse.

> *Mc Gourty (1989), in a letter to Nature addressed the fact that years of refinement often are necessary before a structure is ready to be entered into a database. She was concerned about rumor in the crystallography community, which claimed that researchers were selling coordinates to private companies instead of making them publicly avaible and she argued that journals should play a rôle in the fast deposition of coordinates (WELLER, 1996)*

Ainsi, l'évolution de la publication des données électroniques requiert une coopération entre les producteurs des banques de données et les éditeurs. Les auteurs des séquences n'ont finalement qu'à s'exécuter. Par exemple, GenBank demande aux chercheurs dont les travaux sont financés par des fonds publics (*federal fund* aux USA) de soumettre leur séquence en même temps que leur article. Cette obligation de double publication peut heurter individuellement, mais ce processus présente l'avantage de rendre publiques les données issues de fonds publics, donc collectives. Les entreprises privées qui produisent des séquences génétiques ne sont bien sûr pas soumises à cette obligation, mais peuvent bénéficier des séquences publiques. Néanmoins certaines entreprises privées participent à rendre disponible de plus en plus d'information génétique :

For example, Merck & Co. announced in October 1994 its intention of developping the human mRNA sequences as a public ressource with access to the data unrestricted (WELLER, 1996)

Ce processus de double publication s'impose comme un standard pour les séquences biologiques. Il oblige les chercheurs à respecter un processus de dépôt des données factuelles, avant de pouvoir publier dans les revues scientifiques. Bien qu'imposé au départ comme une contrainte, ce schéma de fonctionnement semble maintenant admis par tous, et il devrait se répéter pour d'autres problèmes de standardisation, notamment les nomenclatures des objets biologiques. Aujourd'hui, il crée les conditions d'un dialogue entre les communautés de spécialistes, les administrateurs de banques de données et les éditeurs des revues.

1.3 - Nomenclatures biologiques

Il existe de nombreuses nomenclatures pour nommer et représenter les objets biologiques, certaines sont bien établies et d'autres en pleine évolution. Elles sont définies pour répondre aux besoins de représentations, d'échanges, et de traitements informatiques. Certaines sont édifiées par des sociétés savantes qui font autorité, d'autres sont émises par des communautés scientifiques d'un domaine spécifique. L'existence d'une nomenclature n'induit pas automatiquement son respect, en général " nécessité fait loi ". Nous présentons deux types de nomenclatures indispensables au champ de la biologie moléculaire et de la génomique, et qui illustrent bien ces différences. Le premier est relatif à la représentation du texte des macromolécules biologiques, l'ADN et les protéines, et le deuxième concerne les noms de gènes.

Les séquences d'ADN et de protéines

Les deux nomenclatures proposées pour les nucléotides (IUPAC-IUB et MOSS , 1970 ; CORNISH-BOWDEN, 1985), qui forment l'alphabet des séquences d'ADN, et les acides aminés (JCBN, 1983), celui des protéines, sont spécifiées par une commission commune aux deux sociétés savantes de chimie (IUPAC) et de biochimie et biologie moléculaire (IUBMB) : l'IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Du fait de leur universalité, ces deux nomenclatures sont couramment utilisées et reconnues par la communauté des biologistes, et elles s'avèrent indispensables pour les analyses informatiques des séquences.

Figure 1 : Représentation des nucléotides

Symbol	Meaning
a	a; adenine
c	c; cytosine
g	g; guanine
t	t; thymine in DNA; uracil in RNA
m	a or c
r	a or g
w	a or t
s	c or g
y	c or t
k	g or t
v	a or c or g; not t
h	a or c or t; not g
d	a or g or t; not c
b	c or g or t; not a
n	a or c or g or t

Figure 2 : Représentation des acide aminés

Abbreviation		Amino acid name
Ala	A	Alanine
Arg	R	Arginine
Asn	N	Asparagine
Asp	D	Aspartic acid (Aspartate)
Cys	C	Cysteine
Gln	Q	Glutamine
Glu	E	Glutamic acid (Glutamate)
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Leu	L	Leucine
Lys	K	Lysine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
Val	V	Valine
Asx	B	Aspartic acid or Asparagine
Glx	Z	Glutamine or Glutamic acid
Xaa	X	Any amino acid.
TERM		termination codon

Noms de gènes

Les règles définies pour les noms de gènes, contrairement à celles des alphabets des séquences biologiques, ne sont pas universelles. Elles sont élaborées pour une, ou un groupe d'espèces, et quand elles existent, ce sont les spécialistes du domaine qui en sont les auteurs (WHITE *et al.* , sd). Les problèmes se posent à deux niveaux différents. Le premier concerne la spécification des formes lexicales adoptées pour les noms de gènes. Par exemple pour les microbes, le nom d'un gène est commencé par trois lettres minuscules, suivi d'une majuscule (*ex : uvrA* , *uvrB* , *hisJ*). Un gène de fonction inconnue commencera par la lettre " *y* " (*ytvP*). Dans la pratique, il est fréquent de rencontrer des noms avec seulement trois minuscules, ou terminés par une combinaison de majuscules et de chiffres, voire dans le pire des cas, des noms qui n'obéissent pas à cette définition. La situation est très variable suivant les espèces, et dépend du degré de coopération des communautés concernées.

Le deuxième niveau de règles vise à favoriser les comparaisons entre espèces, sur la base des noms de gènes, et dans cette situation, une espèce aura un rôle de référence pour les espèces cousines. Ainsi, les noms de gènes des bactéries s'alignent sur ceux employés pour l'espèce " modèle " la mieux connue, *Escherichia coli* . C'est de cette manière que les noms de gènes d'autres espèces, comme ceux de la deuxième bactérie modèle, *Bacillus subtilis* , ont été changés pour satisfaire cette condition. Néanmoins, cette situation n'est pas sans poser des problèmes, notamment pour conserver les liens vers la littérature très abondante qui employait les anciens noms. Un mouvement similaire s'amorce pour les mammifères qui vont s'aligner sur les noms de gènes humains. C'est notamment le cas pour les animaux d'élevage comme les bovins ou les porcins, mais il demeure des exceptions, dont celle notable de la souris. En effet, les scientifiques de cette communauté n'ont pas adhéré à cette règle et gardent leur propre système de nomenclature. Une raison profonde tient à ce que la génétique de cet animal est la plus avancée parmi les mammifères. Il est étudié depuis longtemps et facile à manipuler, en conséquence les chercheurs ne voulaient pas prendre le risque de perturber des usages solidement établis.

La question centrale soulevée par ces changements de nomenclature est relative au repérage de l'information dans les données factuelles et la littérature, et plus précisément, à la cohérence des différents noms de gènes attribués à la même séquence codante au cours du temps. Il se pose alors le problème de la " traçabilité " des documents qui relèvent de ces gènes. La situation est similaire pour l'attribution de nouveaux noms aux gènes inconnus dont la fonction est élucidée. Et le phénomène va prendre de l'ampleur, puisque des programmes d'analyse à grande échelle sont lancés sur les organismes séquencés, pour la détermination systématique de la fonction des gènes inconnus (les projets d'analyse fonctionnelle). Une solution pour les collections d'informations numérisées consiste à attribuer un numéro d'identification unique pour chacun des gènes, évitant ainsi d'employer des liens instables pour les interconnecter. C'est de cette manière que la communauté de la souris opère pour offrir des passerelles vers les autres génomes de mammifères.

2 - Processus de traitement du document numérique

Le développement de collections d'informations a précédé et accompagne depuis leur début les programmes d'étude 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 évoluent vers un accès global aux données, la sélectivité des recherches et la facilité des interrogations (BAKER et BRASS, 1998). Le fait que l'information à gérer soit spécialisée ne demande pas de techniques particulières, mais une appropriation particulière des techniques, des standards et des normes en cours.

2.1 - Intégration et bases de données

Un enjeu des programmes génomes vise à relier l'ensemble des 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 nous pouvons poser aux génomes, et la mise en oeuvre concerne aussi bien des analyses automatiques, que la génération d'outils de visualisation interactifs. 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 dans un même système pour l'interrogation.

Aujourd'hui, plusieurs types de SGBD sont mis en oeuvre dans la communauté de biologie moléculaire :

- des produits " maison ", qui sans proposer toutes les fonctions d'un système commercial, sont adaptés, efficaces et gratuits (ACNUC) et ont innové dans la convivialité, comme les accès par des interfaces graphiques (ACeDB) ;

- des systèmes relationnels, qui sont parmi les plus utilisés des SGBD actuellement en ligne ; ils ont l'avantage de la maturité, sont capables de supporter de grandes quantités d'informations, et possèdent SQL, un langage normalisé de définition et d'interrogation des données ;
- des systèmes à objets, leur représentation des données est plus expressive que celle des modèles relationnels, mais ils sont plus exigeants en ressources matérielles et souffrent de la jeunesse de leurs standards, en cours d'élaboration au sein de l'ODMG ;
- des systèmes plus spécialisés, comme les modèles à *frames* (EcoCyc), dérivés de l'intelligence artificielle et de la programmation fonctionnelle ; ils sont orientés vers les processus de raisonnement et de découverte de connaissances (*knowledge discovery*), mais limités en volume d'intégration des données.

La construction de bases de données apporte des réponses à des problèmes particuliers, mais étant donné l'universalité de l'outil, des standards existent, tant en termes de méthodes de conception, qu'au niveau des langages, des protocoles et des systèmes choisis pour les implémentations. Cette démarche produit des systèmes d'information adaptables, extensibles et pérennes, propriétés indispensables pour supporter des projets d'analyse de génomes en permanente évolution.

Certaines des banques de données " primaires " [9] présentées plus haut ont fait l'objet d'une implémentation sous des SGBD relationnels, notamment à l'EBI. Le NCBI privilégie Entrez (COCKERILL, 1994), son système " maison ", pour l'accès public, et basé sur ASN.1, un format standard pour l'échange de données. Cette représentation est dans la logique du standard SGML pour les documents structurés, car elle fournit une description de la structure de l'information, sous forme de grammaire formelle, à l'équivalent des DTD de SGML. Entrez ne réalise pas une véritable intégration dans un modèle de données unique, mais les liens entre les collections primaires sont assurés *via* des relations d'homologies entre les séquences, et le voisinage des termes pour l'information textuelle, établi par des méthodes statistiques (WILBUR, 1992). Enfin, il faut mentionner ACNUC (GOUY *et al.* , 1984 ; GOUY *et al.* , 1985), qui a été le premier système d'interrogation de séquences biologiques en France, et sur lequel étaient interfacés les logiciels d'analyse de séquences, accédés par les biologistes dès les années 80 sur le serveur BISANCE (DESSEN*et al.* , 1990).

Le deuxième champ d'application des bases de données concerne les collections d'informations " secondaires " [10]. En général, ces bases de données importent une partie de leurs informations depuis les banques primaires, où elles sont structurées et intégrées autour d'un génome, d'un groupe de génomes, ou de concepts et d'objets biologiques transversaux aux espèces (par exemple une famille de protéines). Leur conception s'accompagne souvent de l'intégration de données spécifiques au domaine, et d'un travail de contrôle et d'enrichissement des annotations. Nous trouvons actuellement dans cette catégorie l'emploi de bases de données relationnelles, aussi bien sur les micro-ordinateurs, pour en exploiter les capacités graphiques et la convivialité (Colibri ; MEDIGUE *et al.* , 1993), que sur des systèmes serveurs multi-utilisateurs, qui cette fois-ci parient sur la puissance et les standards de l'accès en réseau (Micado ; BIAUDET *et al.* , 1997). ACeDB (DURBIN et THIERRY-MIEG, 1991) a été spécifiquement conçu pour gérer les informations afférentes à un génome, à l'origine le ver nématode *Caenorhabditis elegans* , et a suscité de nombreux dérivés adaptés aux génomes des microbes et des plantes. Ces bases de données secondaires ont émergé à la demande de communautés spécifiques, et font en général l'objet d'un contrôle accru de la qualité des données, par rapport aux collections primaires. Il s'agit de profiter des collaborations entre spécialistes pour nettoyer les informations, corriger des erreurs d'annotations, éliminer la redondance, et enrichir le contenu (EMGLib ; PERRIÈRE *et al.* , 1999).

2.2 - Standards des notices de banques

Les banques de données primaires, comme GenBank, EMBL ou DDBJ, pour les séquences d'ADN, donnent accès par FTP à leurs gisements d'informations, qui sont des fichiers informatiques de texte codé en ASCII. Ces fichiers sont dits " à plat " (*flat file*), 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 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 étiquettes, pour coder l'information suivant un format qui lui est propre, le contenu informationnel intrinsèque de l'objet biologique reste inchangé.

Figure 3 : *Notice GenBank (extrait)*

```
LOCUS       BACDIA      28206 bp    DNA             BCT       26-MAY-1995
DEFINITION  Bacillus subtilis spoVA to serA region.
ACCESSION   L09228
NID         g410114
VERSION     L09228.1  GI:410114
KEYWORDS    3-dehydroquinate dehydratase; aroC gene; diaminopimelate
            decarboxylase; lysA gene; penicillin-binding protein;
            peptidyl-prolyl isomerase; phosphoglycerate dehydrogenase; ppiB
            gene; response regulator; response regulator kinase; ribA gene;
            ribB gene; ribD gene; ribG gene; ribH gene; ribT gene; riboflavin
            biosynthesis operon; serA gene; signal peptidase; sipS gene; spoA
            gene; spoVAF gene.
SOURCE      Bacillus subtilis (strain 168, sub_species Marburg) DNA.
  ORGANISM  Bacillus subtilis
            Bacteria; Firmicutes; Bacillus/Clostridium group;
            Bacillus/Staphylococcus group; Bacillus.
----------------------------------------------------------------------
REFERENCE   3  (bases 1 to 28206)
  AUTHORS   Sorokin,A., Zumstein,E., Azevedo,V., Ehrlich,S.D. and Serror,P.
  TITLE     The organization of the Bacillus subtilis 168 chromosome region
            between the spoVA and serA genetic loci, based on sequence data
  JOURNAL   Mol. Microbiol. 10 (2), 385-395 (1993)
  MEDLINE   95020538
FEATURES             Location/Qualifiers
     source          1..28206
                     /organism="Bacillus subtilis"
                     /strain="168"
                     /sub_species="Marburg"
                     /db_xref="taxon:1423"
     gene            1..1239
                     /gene="spoVAF"
     CDS             <1..1239
                     /gene="spoVAF"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="AAA67472.1"
                     /db_xref="PID:g410115"
                     /db_xref="GI:410115"
                       /translation="VDIVENRLLNAQVEKVKTLDETTDQVLSGLVAVIVEGAGFAFII
                       DVRSYPGRNPEEPDTEKVVRGARDGFVENIVVNTALLRRRIRDERLRVKMTKVGERSK
                       TDLSICYIEDIADPDLVEIVEKEIASIDVDGLTMADKTVEEFIVNQSYNPFPLVRYTE
                       RPDVAANHVLEGHVIIIVDTSPSVIITPTTLFHHVQHAEEYRQAPSVGTFLRWVRFFG
                       ILASTLFLPIWFLFVLQPDLLPDNMKFIGLNKDTHIPIILQIFLADLGIEFLRMAAIH
                       TPTALSTAMGLIAAVLIGQIAIEVGLFSPEVILYVSLAAIGTFTTPSYELSLATNEPS
                       CPHDTRCFISYKRARHRLYSANYAMASIKSLQTPYLWPLIPFNGKALWQVLVRTAKPG
                     AKVRPSIVHPKNRLRQPTNS"
     conflict        1158..1164
                     /gene="spoVAF"
                     /citation=[1]
     -35_signal      1245..1250
                     /gene="lysA"
                     /note="putative"
     gene            1245..2664
                     /gene="lysA"
     -10_signal      1266..1272
                     /gene="lysA"
                    /note="putative"
----------------------------------------------------------------------------
BASE COUNT     8529 a    5565 c    6530 g    7582 t
ORIGIN
        1 gtcgacatcg tcgaaaacag gctgcttaac gcccaggtcg aaaaagtaaa aaccttggat
       61 gaaaccaccg accaagtgct gtccgggctc gtcgctgtca ttgttgaagg tgcaggcttc
      121 gcatttataa ttgatgtcag aagctatccg ggcagaaacc cggaagaacc tgatacggaa
----------------------------------------------------------------------------
```

* Les pointillés indiquent des coupures dans la notice

Exemples de notice de banques :

- L09228 *via* GenBank : http://www2.ncbi.nlm.nih.gov/cgi-bin/birx_by_acc?genbank+L09228
- L09228 *via* EMBL : http://www.ebi.ac.uk/cgi-bin/emblfetch?L09228
- L09228 *via* Entrez : http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=410114&form=6&dbn&Dopt=g

Ainsi, nous distinguons deux grands types de standards dans un enregistrement. Un standard que l'on peut dénommer "

classificatoire ", puisqu'il permet d'attribuer les propriétés à la séquence, à l'aide des descripteurs, et un autre standard que l'on peut dénommer " structural ", puisqu'il organise le document pour une " lecture " informatique. Le premier est relatif à l'IST biologique, le contenu; le second a trait au format du document, le contenant. Ce dernier précise la position des descripteurs dans les colonnes du texte en ASCII.

Le standard classificatoire d'un enregistrement de séquence d'ADN peut se diviser en quatre parties :

Identité biologique. Du descripteur `LOCUS` au descripteur `ORGANISM`. Ce sont des informations générales qui renseignent " l'état civil " de la séquence : son nom, le type de molécule, son affiliation biologique, la date de son entrée (`LOCUS`), son numéro d'accès (`ACCESSION`) comme identificateur unique de l'enregistrement dans la banque, une brève définition et des mots-clés pour la caractériser, et enfin son origine (`SOURCE`) et son affiliation biologique à une espèce (`ORGANISM`).

Références. Du descripteur `REFERENCE` au descripteur `MEDLINE`. Ce sont les références bibliographiques des publications relatives à la production de la séquence. Mais plus précisément, cette partie est une notice catalographique enrichie : en plus de renseigner sur les auteurs, le titre, la revue, elle localise le document dans la banque MedLine.

Propriétés de la séquence. Le champ `FEATURES`, où figurent les annotations qui décrivent la séquence, c'est-à-dire qui spécifient précisément la fonction de chacune des sous-séquences de l'enregistrement. Il est formé de plusieurs sous-champs donnant la position (`Location`) et les attributs spécifiques (`Qualifiers`) de chacune des sous-séquences, correspondant à une fonction identifiée.

Texte de la séquence d'ADN. Débute par le descripteur `ORIGIN`. C'est la représentation de la séquence nucléotidique à l'aide des symboles `ATGC`, sur laquelle toutes les expériences ont été réalisées.

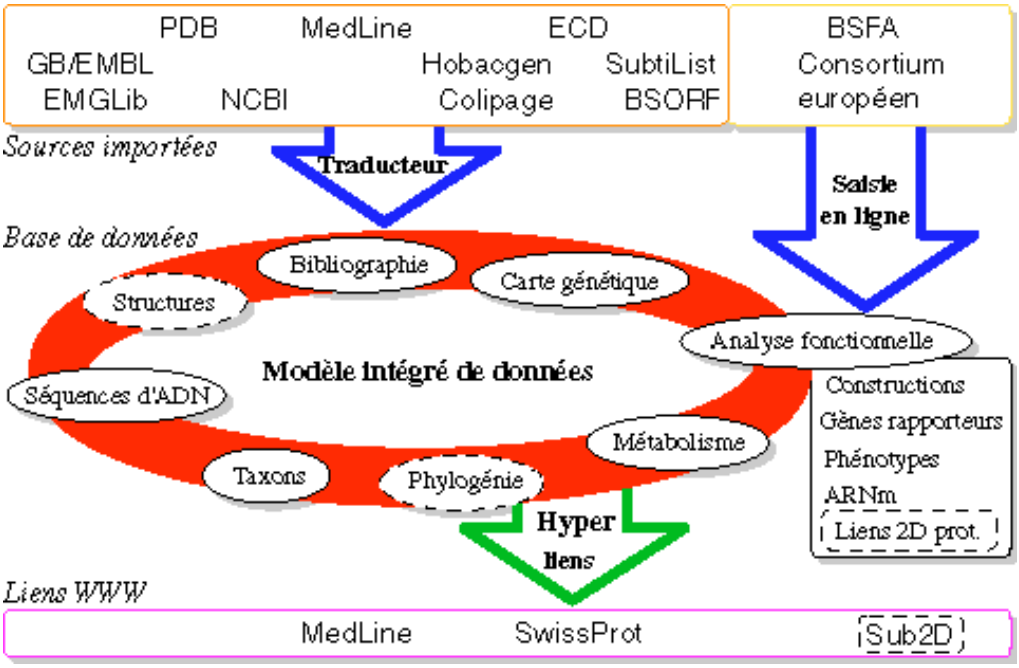
Cette standardisation permet donc de coder l'information et garantit une utilisation pérenne du document. Les spécifications de ces standards sont bien entendu accessibles à travers Internet sur les serveurs des organismes qui maintiennent les banques primaires [11]. Elles évoluent avec les connaissances et le contexte de l'activité scientifique. Ainsi, un descripteur similaire à MedLine concernant PubMed Central (ex- E-biomed, BUTTLER, 1999 ; VARMUS, 1999) [12] pourrait, par exemple, apparaître. De plus, si les banque de séquences d'ADN représentent une même information, mais dans un format propriétaire, le champ `FEATURES` fait l'objet d'une standardisation commune à elles toutes (DDBJ/EMBL/GenBank staffs,1999). De nouveaux *features* et de nouveaux *qualifiers* sont introduits de cette manière, dont la liste et les définitions sont décrites par un document collectif et consultable en ligne. Il compose en quelque sorte une DTD pour cette partie des enregistrements de séquences, une grammaire hors-contexte y est spécifiée en langage BNF pour la syntaxe du champ, et il en garantit l'extensibilité.

▲2.3 - Une base de données de génomes microbiens

Nous allons illustrer le rôle des bases de données pour la biologie moléculaire, en présentant Micado (MICrobial Advanced Database Organization), réalisée au laboratoire de Génétique Microbienne, à l'INRA de Jouy-en-Josas (BIAUDET *et al.* , 1997). Micado a été impulsée par les programmes d'analyse du génome de la bactérie modèle *Bacillus subtilis*, auxquels participait le laboratoire. Le premier programme visait le séquençage complet de l'ADN de la bactérie (KUNST *et al.* , 1997), tandis que le deuxième consistait à élucider la fonction des gènes inconnus, découverts

par le séquençage systématique du chromosome. Micado gère les informations de ce deuxième programme d'analyse fonctionnelle (EHRLICH et OGASAWARA, 1999), qui sont rentrées par un consortium Européen de 17 laboratoires, auquel est associé un consortium d'industriels (BACIP). 1200 gènes inconnus ont ainsi été étudiés, tandis qu'un consortium Japonais s'intéressait à 800 autres, sur les 4100 gènes que porte le chromosome de la bactérie.

Cette activité de génomique fonctionnelle, qui implique un travail coopératif à grande échelle, appelle une intégration des informations disponibles sur le génome de la bactérie. La base de données a été créée en 1993, elle fonctionne avec le SGBD relationnel ORACLE et un serveur UNIX. La structuration des données microbiennes offre un moyen d'interrogation précis et élaboré avec le langage SQL (MELTON et SIMON, 1993). Une partie de Micado réalise une intégration " verticale " de l'information génomique de *B. subtilis* , c'est-à-dire du niveau moléculaire de la séquence jusqu'aux caractères observables, les phénotypes. Elle associe aux données de l'analyse fonctionnelle, la séquence d'ADN complète et annotée du chromosome (MOSZER, 1998), mise à jour et distribuée depuis la base de données SubtiList (MOSZER *et al.* , 1995), la carte génétique produite au laboratoire (BIAUDET *et al.* , 1996), et une classification des gènes en fonction de leur implication dans le métabolisme de la bactérie. La finalité de ces projets consiste à décrypter tous les niveaux intermédiaires, biochimiques et physiologiques, qui relient le gène à son phénotype.



cf. idem format png

Figure 4 : *Modèle de données de Micado. Les différentes parties du modèle sont mises en correspondance avec leurs sources d'informations. Une partie est directement importée par traduction automatique (parsing) depuis des collections internationales. Les données d'analyse fonctionnelle de la bactérie Bacillus subtilis sont saisies directement depuis les laboratoires, à travers une interface WWW. Les tirets dénotent des réalisations en cours ou planifiées.*

Un deuxième axe d'intégration, " horizontal " celui-ci, est poursuivi pour les séquences d'ADN, qui sont importées depuis GenBank et EMBL, au moyen de traducteurs automatiques. Ils sont basés sur des grammaires formelles "hors-contexte" qui décrivent la structure syntaxique de leurs enregistrements L'intégration concerne à ce jour toutes les bactéries et les archées (auparavant appelées archéobactéries), et inclut toutes les séquences complètes actuellement publiées pour ces génomes. Le modèle des séquences d'ADN possède des propriétés de généricité, traduction intégrale des enregistrements EMBL/GenBank, atomicité sémantique, extensibilité à de nouvelles définitions de *features* et de *qualifiers* , qui lui permettent d'être aujourd'hui utilisé dans des bases de données génomiques pour les animaux (Bovmap et dérivés ; GAS *et al.* , 1996) et les plantes (Génoplante). Une partie du modèle, le niveau moléculaire et les références bibliographiques, sera prochainement enrichi par les structures des protéines en provenance de PDB. Cette partie peut servir à terme de plate-forme commune aux projets d'analyse des génomes d'autres espèces (animal, plante, microbe). Elle s'intègre dans les bases de données génomiques en s'associant à un deuxième niveau génétique et physiologique, qui est par contre

spécifique de l'organisme étudié.

2.4 - Accès, visualisation et WWW

Le premier accès convivial programmé pour les utilisateurs de Micado était composé de menus de texte. Mais il allait rapidement évoluer vers des interfaces graphiques en réseau, notamment avec l'arrivée de WWW et du langage Perl, puis du langage Java et son principe de machine virtuelle, et enfin CORBA pour l'interconnexion des applications en réseau (SAMSON *et al.* , 1998). Leur mise en oeuvre permet de programmer des accès conviviaux et interactifs, et le succès de WWW a offert un environnement standard et portable pour le développement des interfaces, notamment pour l'accès aux bases de données. L'information dans Micado 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 de 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, et 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 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, elles sont le support indispensable à l'élaboration de stratégies systématiques d'exploration des données (*data mining*), qui entrelacent des chaînes d'analyse et de classification automatique, avec des étapes interactives de traitement et de visualisation (WESTPHAL et BLAXTON, 1998). L'exemple d'un tel usage de Micado 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 superposé l'affichage des états statistiques calculés sur les séquences (BIZE *et al.* , 1999).

L'emphase de l'information relative à la génomique nécessite un repérage accru et efficient des connaissances, qui explique l'intérêt pour les techniques de visualisation (SCHNEIDERMAN, 1997). L'exploration de données 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 un système d'information offrent des " machines de vision " (BALZT, 1998) capables de générer de nouvelles connaissances, de nouveaux projets de recherche ou d'autres éléments de réflexion.

3 - Normalisation et standardisation en aval

L'évolution des bases de données génomiques est guidée par la recherche de la fonction des gènes inconnus. C'est

maintenant un problème universel, ou du moins commun à toutes les espèces partiellement ou totalement séquencées. Nous avons vu qu'ils composent une fraction significative, voire majoritaire, de l'ensemble des gènes mis en évidence sur l'ADN des chromosomes, par les programmes de séquençage complet. La deuxième génération de programmes d'analyse de génomes vise à répondre à cette question par des expériences menées à grande échelle, auxquelles doivent être combinées les outils d'analyse de séquences de la bioinformatique, et les collections d'information ayant trait au métabolisme et à la physiologie des organismes.

Cette question générale appelle à de nouvelles extensions des bases de données génomiques, et pousse ainsi à leurs limites les techniques de l'intégration des informations dans un modèle de données, sur un site physique et par une équipe unique (DAVIDSON *et al.* , 1995). Le problème est d'autant plus aigu que de nouvelles collections, susceptibles de contribuer à la recherche des réponses, continuent toujours à apparaître. Concrètement, vouloir réaliser une intégration la plus exhaustive possible pour Micado revient à développer et maintenir un traducteur (*parser*) pour chaque collection importée dans la base de données. Il faut rajouter à cette contrainte la création et la mise à jour des extensions correspondantes dans le modèle de données, des nouveaux liens exploitables sur WWW, ainsi que des interfaces conviviales d'interrogation et de visualisation. Ce type de tâche est accessible dans le cadre d'environnements coopératifs en réseau, et c'est pourquoi avec sa généralisation, la communauté bioinformatique s'intéresse depuis quelques années au concept de l'interopération des bases de données (KARP, 1996).

3.1 - Fédérations navigables sur WWW

Les hyperliens à travers WWW constituent un premier niveau d'alternative à l'intégration physique dans un modèle de base de données. Ils sont susceptibles d'exploiter des relations définies dans les enregistrements des objets biologiques par les curateurs [13] des banques, par exemple pour un gène, entre l'enregistrement de sa séquence d'ADN et celui de la séquence de sa protéine. La notice GenBank (Fig. 3) propose des liens depuis la partie des références bibliographiques vers les notices catalographiques enrichies de Pubmed [14]. La version EMBL du même enregistrement de séquence effectue des liens systématiques vers la banque de séquences de protéines SwissProt, qui cette fois-ci reposent sur la définition d'un nouveau *qualifier* , `/db_xref`, associé au *feature* `CDS`, signifiant *Coding Sequence* (*cf.* 2.2). La collection de séquences de protéines SwissProt est par ailleurs très riche en liens qui pointent vers les séquences d'ADN correspondantes, les structures tridimensionnelles, lorsqu'elles ont été déterminées et enregistrées dans PDB, l'activité enzymatique, *etc* ., toutes choses qui expliquent qu'elle ait été une des premières interfacées à WWW (APPEL *et al.* , 1994).

Les liens entre les collections d'informations biologiques ont très tôt figuré dans les enregistrements des banques de données. Leur existence précoce a nourri les premières réflexions sur le futur des systèmes d'information pour la biologie moléculaire, facilitant la prise de conscience du problème de l'intégration, et de la nécessité d'une évolution technologique (FUCHS *et al.* , 1992). La présence d'informations de liens entre les enregistrements de différentes collections a aussi permis la création de SRS, Sequence Retrieval System (ETZOLD et ARGOS, 1993a ; ETZOLD et ARGOS,1993b), avant même que se généralise leur emploi avec WWW. La question des liens, notamment leur absence dans certains cas, a même suscité un des premiers projets d'analyse automatique de texte en bioinformatique, qui vise à extraire l'information contenue dans les notices des banques (ACHARD et DESSEN,1998). À l'origine pour la création automatique de liens entre GenBank et GDB (Genome DataBase), la base de données pour la cartographie des gènes humains, ce travail est aujourd'hui consultable dans la base de données Virgil (ACHARD *et al.* , 1999).

L'emploi des hyperliens sur WWW propose un premier niveau d'interaction entre les bases ou les banques de données biologiques, et elles constituent des fédérations navigables de manière transparente pour les utilisateurs. Les hyperliens sont faciles à mettre en place, si tant est que les références croisées préexistent, et ne nécessitent pas la coopération des responsables des collections. Cependant, ils présentent des limites, au sens où ils ne remplacent pas la structuration de l'information dans une base de données (LETOVSKY, 1995 ; KARP, 1996). Dans le cas de Micado, le modèle de données contient les références bibliographiques relatives aux séquences d'ADN figurant dans la base. Ces références ont des liens WWW vers les notices PubMed correspondantes, qui offrent en supplément du contenu actuel de Micado un accès aux résumés (*abstracts*) et aux termes du thésaurus de MedLine (MeSH, Medical Subject Headings), qui ont servi

à indexer l'article. Il est possible d'exploiter le nom des auteurs, le titre des articles ou le nom des revues, comme critères de sélection des séquences d'ADN dans une interrogation de Micado. Par contre, les parties de la notice catalographique enrichie, comme le résumé ou les termes du thésaurus ne sont pas utilisables pour chercher une séquence, puisque seules figurent dans le modèle de la base de données les liens vers ces notices, et non la totalité de leurs contenus.

3.2 - Bases de données partagées

De nouvelles approches d'interopération des logiciels sont apparues conjointement à la généralisation du réseau au cours de la décennie 90. L'une d'elles, CORBA, offre une interface de communication de haut niveau entre les " composants " d'un logiciel, indépendante de leurs langages de programmation et de leur implémentation physique. Ce standard est proposé par l'OMG, et repose sur les technologies orientées objet, dont la promotion est la première raison d'être de l'organisation. L'OMG rassemble l'ensemble des acteurs de l'informatique d'aujourd'hui, y compris Microsoft. Ce dernier avance son propre standard, OLE/COM, mais il est plus adapté à l'interopération de composants logiciels sur une même plate-forme, son système d'exploitation Windows, et par ailleurs il est maintenant interopérable avec CORBA. La communication repose sur un bus logiciel, l'ORB (Négociateur de Requêtes d'Objets), qui permet aux composants d'une application de s'échanger des services et des informations de manière transparente, tandis qu'ils peuvent s'exécuter sur des systèmes d'exploitation différents, reliés par des réseaux hétérogènes, et ne pas avoir été programmés dans le même langage. Cette indépendance est réalisée par l'emploi d'un langage de spécification d'interface, IDL, à partir duquel les définitions sont écrites, puis traduites (compilées) dans les langages cible de programmation des composants.

Il est important de percevoir que le but de CORBA, avant de convertir au paradigme de la programmation à objets, est de se mettre au service de l'interopération et de l'intégration des logiciels. En effet, l'utilisation d'objets pour spécifier les interfaces, isole, par encapsulation, le contexte de la création et du fonctionnement du composant, vis-à-vis de sa mise en oeuvre dans l'application à laquelle il coopère. En d'autres termes, dans cette logique modulaire des composants, il offre à leurs concepteurs/assembleurs la souplesse d'intégrer des solutions adaptées qui exploitent la diversité informatique. CORBA ne restreint pas à l'utilisation d'un langage, d'un logiciel et d'une plate-forme, comme c'est souvent le cas pour un développement classique d'application. Seul le futur écornera cette vision idéale, le présent témoigne de ses maladies de jeunesse, et de ses limites, dans l'exigence d'un savoir-faire récent et d'une bonne logistique en informatique, spécialement pour un déploiement en réseau. En tout état de cause, CORBA pose désormais les bases d'une interopérabilité universelle des logiciels, et trouve son usage dans des développements intégratifs à grande échelle, comme l'est depuis les années 80 la norme de documents structurés SGML. De plus, comme en son temps ce standard concentrait les logiques de la programmation structurée autour de la mise en forme des documents et de leur réutilisation, CORBA exprime les plus récents acquis de cette pensée modulaire, dans une vision objet, et propulse le document numérique dans les environnements partagés.

Aussi, CORBA a suscité l'attention de la communauté bioinformatique, comme une réponse générale à l'interconnexion des bases de données biologiques, et de celles-ci à des outils d'interface ou d'analyse. L'EBI, à la recherche dès sa création de solutions d'intégration, a joué un rôle essentiel dans cette sensibilisation (WILLIAMS, 1997a)., mais c'est la pression des industriels qui en a entériné l'adoption (WILLIAMS, 1997b). À l'heure actuelle, l'OMG organise la standardisation pour l'interopérabilité des logiciels par domaines de métiers, qualifiée d'intégration verticale, en suscitant la création de Domain Task Forces (DTF). La récente Life Sciences Research DTF, est consacrée à la biologie moléculaire et la génomique (HAYTER, sd). Le groupe travaille au sein de l'OMG à l'élaboration de représentations communes pour l'interconnexion des logiciels de bioinformatique. La motivation est forte pour disposer de représentations des concepts et des objets biologiques qui fassent l'objet d'un consensus de la part de la communauté, au moins pour ceux d'entre les plus génériques, comme les séquences biologiques et les cartes génétiques.

Dans le cas de Micado, une spécification d'interface en IDL est écrite pour les séquences d'ADN et leurs annotations, qui sont contenues dans le modèle relationnel de la base de données. Le modèle est lui-même alimenté par la traduction des enregistrements des collections EMBL/GenBank. On perçoit ainsi aisément la versatilité des documents numériques. Les

standards évoqués organisent plusieurs représentations d'un même contenu documentaire : un texte balisé, un modèle de base de données, une représentation en objets. La dernière encapsule les données et les méthodes pour y accéder, et fournit un service de séquences microbiennes annotées, qui peut être invoqué par des programmes clients, à la condition d'inclure dans leur code la même définition d'interface IDL que le serveur. Un client de visualisation graphique, un navigateur de génomes, est interfacé de cette manière à la base de données, *via* le composant serveur d'objets. Cette première application a permis d'évaluer la pertinence de CORBA pour l'interface des outils, et elle préfigure une première étape vers l'intégration de Micado dans un ensemble interrogeable unifié de bases de données biologiques fédérées.

3.3 - Ontologies

L'apparition de solutions standards pour l'interopération des logiciels a trouvé un écho dans une communauté bioinformatique sensibilisée au problème de l'intégration des collections d'informations en biologie. Une première réunion internationale était organisée en 1994, MIMBD'94 (Meeting for the Interconnection of Molecular Biology Databases), suivie de deux autres les années suivantes, en association avec un des trois grands congrès de bioinformatique, ISMB (Intelligent Systems in Molecular Biology). Cette première réunion permettait pour la première fois de confronter des expériences d'intégration physique des bases de données, et leurs limites, mais aussi des solutions " propriétaires " d'intégration virtuelle, comme OPM. Assez logiquement, la dernière réunion de 1996 était dédiée au monde de l'objet, avec CORBA, le nouveau langage Java, et les bénéfices de leur emploi sur la portabilité et la réutilisation des composants logiciels. Cette première série de réunions donnait ensuite naissance aux conférences OIB (Objects In Bioinformatics), tenues annuellement depuis 1997, et dont certains des participants animent aujourd'hui LSR (Life Sciences Research DTF).

Ces échanges ont permis de poser les problèmes de l'intégration virtuelle, à la lumière des outils désormais disponibles, et fait apparaître les barrières qui constituent encore un obstacle à l'intégration des bases de données dans des ensembles à grande échelle. Ces barrières sont de nature sémantique, les auteurs des bases de données peuvent utiliser des définitions, des attributs différents, ou des nomenclatures non normalisées, pour la représentation des entités biologiques. En prenant les gènes comme exemple, deux niveaux de définitions sont à observer, le premier relatif au concept même de gène, le second ayant trait à son implémentation informatique. Des points de vue expérimentaux différents vont d'abord amener à des définitions divergentes. Un gène dans le contexte d'une séquence d'ADN sera considéré comme un élément codant une protéine ou un ARN. Sur la carte génétique, il sera perçu comme un caractère observable, dont la transmission héréditaire pourra être suivie, et il englobe dans ce cas des séquences d'ADN non codantes, mais conférant certaines propriétés à l'organisme (typiquement, un site d'attachement d'un virus sur le chromosome). Deux définitions peuvent également s'opposer pour un même contexte. Ainsi, pour en revenir à la séquence d'ADN, soit le gène est défini strictement par la région codante, soit par l'ensemble région codante avec les signaux qui permettent son expression (le codage) et la régulation de cette expression. Enfin, à supposer que deux concepteurs de bases de données partent de la même définition, il est à peu près certain que leurs implémentations, c'est-à-dire la représentation informatique, présenteront des divergences.

Ainsi, il est très vite apparu le besoin de partager les mêmes définitions formelles pour les concepts et les entités biologiques qui sont représentées et manipulées. C'est pourquoi se posait rapidement la question d'un recours à la spécification d'ontologies pour le domaine des connaissances en biologie moléculaire.

> *A definition of the term [ontology] as currently used is given : an explicit specification of a " world " that is to be represented in a computer system (VICKERY, 1997).*

Une ontologie consiste en une spécification formelle des termes et des associations entre ces termes pour un domaine de connaissances précis. Autrement dit, c'est un catalogue sémantique, dont les descriptions sont à la fois concises, non ambiguës, et qui se doit d'être exploitable par un logiciel comme par un opérateur humain. Une ontologie se différencie d'un thésaurus par le fait qu'elle propose une description littéraire (*human readable*) et une description formelle

(*machine readable*) d'un concept, qu'elle ne propose pas forcément une classification alphabétique des concepts, qu'elle ne sert pas uniquement à l'analyse d'un corpus mais aussi à la médiation de corpus, enfin sa perception se réalise plutôt sous forme de graphe que d'arbre.

L'intérêt de réaliser l'ontologie d'un domaine de connaissances dépasse le seul cadre de l'interopération des bases de données. En effet, elle revient à définir et partager des vocabulaires spécifiques, qui pourront relever de toute activité d'ingénierie des connaissances mise en oeuvre dans le domaine considéré. On peut penser par exemple à l'extraction d'informations à partir d'analyse automatique de texte. C'est ainsi qu'un groupe s'est créé depuis 1998, et organise une conférence annuelle sur les ontologies pour la biologie moléculaire, Bio-Ontologies, un autre surgeon de MIMBD, et que les premiers travaux appliqués à la biologie sont apparus (SCHULZE-KREMER, 1998 ; BAKER *et al.* , 1999).

3.4 - Bibliothèque électronique et collaboratoire

Cette partie illustre le rôle et la synergie des standards à travers les concepts de Digital Library (DL) (SCHATZ, 1997) et Collaboratory (POOL, 1993). Ces concepts mettent en avant des processus de traitement des documents numériques. Nous éclairons ces concepts dans le contexte de la génomique et des standards présentés plus haut.

La Digital Library Federation (DLF) définit les DL ainsi :

> *Digital libraries are organizations that provide the ressources, including the specialized staff, to select, structure, offer intellectual access to interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.(WATERS, 1998)*

Turner (1995) définit les collaboratoires ainsi :

> *Les nouvelles structures sociales de la production de la science*
>
> *Le vocable qui convient à la description de ces nouvelles structures est celui de collaboratoires de recherche. [...] Dans le document servant à décrire les objectifs de l'administration Clinton, la notion de collaboratoire désigne les "centres de production scientifique et technique "sans murs" dans lesquels les chercheurs réalisent leurs recherches sans être limités par leur localisation géographique. Grâce aux réseaux, ils peuvent tout à la fois interagir interactivement avec leurs collègues dans d'autres universités; accéder à des instruments à distance; partager leurs données et leurs ressources computationnelles et, enfin, accéder aux informations réunies dans les bibliothèques électroniques.*

L'essence de ces concepts est relatif à l'échange et la gestion d'informations médiatisées par l'informatique. Les DL oeuvrent pour la diffusion et l'utilisation de collections spécialisées ou non, et les collaboratoires sont la version électronique des " collèges invisibles " et utilisent entre autres les DL. Ces concepts dans le domaine de la génomique mettent en oeuvre les standards précédemment évoqués (interopérabilité, la constitution de gisements d'information, la structure des documents, les interfaces WWW).

Les banques de données internationales, comme GenBank, EMBL, DDBJ, constituent des collections électroniques de travaux en biologie moléculaire. Complétées par les revues scientifiques en ligne, les notices catalographiques enrichies de PubMed, et d'autres ressources informationnelles disponibles sur Internet, cet ensemble de documents électroniques peut être perçu comme une DL et un collaboratoire. En effet, même si les collections sont disséminées, elles forment une fédération navigable (*Cf.* 3.1) et circonscrivent la communauté des génomistes. Certes, c'est une acception large des deux concepts, mais elle constitue un premier niveau qui ne peut être négligé au regard d'autres disciplines scientifiques. Les nomenclatures biologiques, le processus de double publication, les notices des banques de données (factuelles et

bibliographiques) sont des standards qui structurent fortement cette discipline, et les DL et les collaboratoires en sont la résultante.

Une acceptation plus restrictive des DL fait référence à la réalisation de systèmes d'information ou à leur fédération par l'interopérabilité logicielle (*Cf.* 3.2), capables d'organiser des collections et de proposer des services à des usagers bien identifiés (HAKALA et HORMIA, 1997). À cet égard, Micado est un exemple de DL. Ce système d'information collecte les documents numériques des banques de séquences internationales, structure et gère cette collection à travers des outils et langages informatiques standardisés et robustes, propose une interface d'interrogation et de consultation efficiente et conviviale *via* WWW, et les usagers ciblés sont des chercheurs de la communauté des microbiologistes (*Cf.* 2)

Une acception plus restreinte de collaboratoire relève du " travail coopératif " (*cooperative work*) et des DL. Elle emprunte au " *CSCW* ", au " *groupware* ", au " *workflow* ", le principe de travail collectif en réseau, avec échange de données, tableau blanc et noir, collecticiels (outils logiciels collectifs) (CHARTRON, 1994), et aux bibliothèques numériques, l'accès à des ressources informationnelles organisées, manipulables et interrogeables. Il s'agit d'une plate-forme de travail délocalisée pour des chercheurs. À l'instar du *Worm Community System* (WCS) qui représente le collaboratoire de l'organisme modèle *Caenorhabditis elegans* (SCHATZ, 1993), Micado abrite un programme européen de génomique fonctionnelle des gènes de *Bacillus subtilis* (*Cf* . 2.3). Les données factuelles issues de l'expérimentation de 17 laboratoires sont saisies à travers une interface WWW et présentées sous forme de " cahiers " de laboratoires. Ces données sont gérées par le système d'information qui génère dynamiquement des documents. Ainsi chaque laboratoire peut consulter les travaux des autres et apprécier l'avancement du projet. Le courrier électronique sert aux échanges interpersonnels.

Les DL et les collaboratoires ne doivent pas uniquement être définis par des questions d'ordre technique, des considérations économiques, sociales et légales sont à soulever. On peut notamment s'interroger sur la constitution d'un gisement d'information à partir d'informations gratuites ou payantes, la diffusion gratuite ou payante des informations intégrées dans le système d'information, le droit d'auteur, la validité des informations, *etc* .

Il faut indiquer que Micado dispose d'une partie publique et d'une partie privée. Schématiquement on peut considérer que la partie publique constitue la DL et que la partie privée est relative au collaboratoire. La partie publique diffuse les documents issus de l'activité scientifique sur les génomes d'archées et de bactéries disponibles dans les banques de séquences internationales. L'accès à la partie privée, dans laquelle les données du consortium européen sont entreposées, est réservé au collaboratoire. Ce contexte répond en partie aux questions d'ordre économique et juridique : l'accès à la DL est gratuit et profite à toute la communauté scientifique, les données du collaboratoires ne seront disponibles gratuitement qu'après un temps de latence, pendant lequel les industriels associés au consortium bénéficient de la primeur des résultats.

Micado, DL et collaboratoire spécialisés en génomique des microbes, exerce une fonction cohésive sur l'information issue des génomes d'archées et de bactéries. Ses capacités techniques et l'utilisation de standards donnent à ses usagers des informations à " valeur ajoutée " en transformant (organiser, manipuler, et diffuser) un gisement d'informations gratuit, disponible sur Internet ou issu d'une plate-forme de travail coopératif.

Conclusion

A travers les phases de standardisation présentées, nous avons finalement mis en évidence une diachronie de la génomique, des outils informatiques et des traitements des documents. Cette diachronie montre dans ces trois domaines un changement de paradigme. La génomique passe de l'analyse du génotype à l'étude du phénotype (de la description des nucléotides de la séquence d'ADN qui compose le génome au caractère observable sur l'organisme vivant) ; les outils

informatiques, notamment représentés par les systèmes d'information, passent de systèmes de gestion de fichiers " à plat " (banques de données) aux bases de connaissances ; le traitement des documents, au-delà du fait d'être passés de l'imprimé à l'électronique, ne sont plus manipulés comme un ensemble indissociable, mais comme une composition d'unités informationnelles, d'unités sémantiques extractibles ou référençables.

On peut donc considérer deux paradigmes. Le premier est relatif à la syntaxe, il représente : le génotype qui correspond à 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 second est relatif à la sémantique, il 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 (de l'analyse statistique à l'analyse sémantique de texte).

Primauté de la sémantique

Ce glissement de paradigme est commun à tous les domaines de l'activité humaine qui font appel aux technologies de l'information. Il correspond à un changement d'intérêt et à de nouvelles possibilités de traitement relatifs au document. Le document n'existe pas sans support et informations, et l'information est la mise en forme de connaissances. Si la recherche et la collecte de documents sont essentielles, et l'extraction d'informations primordiales, la production de sens devient cruciale. " De la syntaxe à la sémantique " exprime ce changement de point de vue (RAPPAPORT, 1997), un recentrage sur la connaissance, sur l'essence du document. L'engouement récent pour les concepts de gestion des connaissances (*knowledge management*) et le terme de " société de la connaissance " révèle ce propos.

La boucle de rétroaction dans les processus de standardisation

L'évolution des standards et la création de groupes de standardisation pour l'information en génomique, et les techniques qui la traitent, font émerger de nouvelles pratiques dans le processus de collecte, traitement et diffusion des documents numériques. Cette assertion doit être aussi inversée. En effet, il faut aussi considérer que ce sont des nouvelles pratiques qui font émerger les standards et les groupes de standardisation. Ces deux propositions sont intimement liées. Si la première insiste sur une utilisation créative, dans un processus de traitements, de documents et de techniques standardisés, la seconde met en évidence la nécessité de standardiser pour rendre plus efficient un processus de traitement des documents numériques. Avec ces nouveaux dispositifs, le document devient réparti, collectif et versatile.

La qualité à l'épreuve

Le partage actuel de ressources informationnelles, au travers des banques primaires et secondaires, confère à la génomique un statut singulier. La création de " PubMed Central " (ex E-biomed) en janvier 2000 renforcera cette situation privilégiée. Néanmoins la qualité des données des banques de séquences n'est pas sans reproche, et les erreurs se propagent de collections primaires en collections secondaires. 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 (1996) décrivent des erreurs de synonymie, d'homonymie, de saisie, de contamination dans les séquences biologiques et les annotations qui les accompagnent.

Ainsi, des réflexions et des travaux de standardisation sont à mettre en place pour améliorer la qualité des données. Ces travaux sont tributaires d'une meilleure organisation de la communauté scientifique, et ils sont même nécessaires. En effet, alors que sont envisagés des outils sophistiqués de raisonnement automatique, il n'y a pas lieu d'en espérer des résultats satisfaisants, s'ils doivent procéder à partir d'ensembles de données dont une fraction significative est erronée ou approximative. La réponse à ces problèmes de qualité devrait être un bon " marqueur " de l'état d'avancement des collaborations dans la communauté des génomistes.

Les nouveaux standards technologiques

De nombreux facteurs contribuent à l'adoption ou au rejet de nouveaux standards, et parmi ceux-ci, la complexité de leur mise en oeuvre. La création de bases de données, ou le développement d'un environnement partagé avec CORBA, requièrent un investissement humain et font appel à des compétences qui demandent un effort important et soutenu de la part des communautés intéressées, avant que celles-ci soient payées en retour. Le succès de WWW et de HTML tient en partie à leur simplicité. Au-delà, par leur lisibilité, ils ont popularisé le concept de balisage de texte et de document structuré. Un nouveau standard, XML (MICHARD, 1999), prend la suite de HTML, et se place dans la logique de SGML, en introduisant une DTD qui décrit la syntaxe du langage, et lui confère par-là même des propriétés d'extension dont le premier était dépourvu. Il ouvre la voie à des extensions spécialisées, par exemple pour la représentation des formules mathématiques ou des molécules chimiques, qui ne sont pas possibles avec HTML, autrement que sous forme d'images numérisées. De plus, un langage de requêtes en cours d'élaboration, XQL, va procurer aux documents XML des caractéristiques de bases de données. On peut espérer que XML, à la manière de HTML pour le balisage, sensibilise un plus large public aux grammaires informatiques.

XML apparaît déjà pour certains comme une alternative séduisante à CORBA. Mais la plus grande richesse et la puissance de ce dernier, notamment sur le plan de l'interactivité, de la sécurité et d'autres services annexes progressivement intégrés dans le standard, fait qu'ils doivent être perçus comme des produits complémentaires, ce qui explique leur association récente. Néanmoins, se pose la question si dans le futur CORBA se restreindra à des applications à haut niveau d'exigence, un peu comme SGML a été le plus souvent pratiqué dans certains milieux professionnels de l'édition et de la documentation, notamment pour la documentation technique.

Le rôle croissant des acteurs scientifiques pour la standardisation informatique

Au-delà de la définition des standards, la manière dont ils sont édifiés pour l'informatique, implique de plus en plus des communautés concernées par les applications, et contribue à les fédérer. On peut citer les *Domain Task Force* pour CORBA, qui ciblent les applications liées aux métiers, et font largement appel à leurs représentants. Les procédures d'adoption de nouveaux standards s'inspirent de celles pratiquées pour les RFC d'Internet, assimilables à un système de publication qui prévoit une conception par étapes, afin de garantir une spécification finale adéquate aux besoins du domaine. A coté des organismes officiels comme l'ISO, et des grands groupes d'intérêt comme l'OMG ou le W3C, nous assistons aujourd'hui à l'apparition d'associations qui reproduisent les mêmes démarches de standardisation autour de projets plus ciblés. On peut citer l'exemple de Bioperl (CHERVITZ *et al.* ,1997) qui se propose de coordonner les développements en langage Perl d'applications destinées à la biologie moléculaire et la bioinformatique.

Notes

1 Un glossaire en fin d'article développe les sigles et les acronymes. [Glossaire]

2 Analyse systématique et exhaustive des génomes, mettant en oeuvre des processus automatisés.

3 " Le terme technique décrit une manière de faire. Le mot technologie, à cause de son suffixe *logie* , désigne une science ou un ensemble de connaissances " Christian DeBresson, " comprendre le changement technique ", p. 26.

4 Néologisme qui signifie littéralement " dans le silicium ", concrètement " dans l'ordinateur ", par analogie à *in vivo* et *in vitro* . Une recherche sur les titres et sommaires des articles référencés par MedLine fait remonter sa première mention en 1991, par des microbiologistes danois dans une revue de l'Institut Pasteur (HANSEN *et al.* In : *Research in Microbiology* , vol. 142, nº2-3, pp. 161-167). Trois articles l'emploient en 1993, et son usage se généralise à partir de 1996.

5 " Objet biologique " désigne au sens large des entités physiques, des concepts biologiques et leurs relations, c'est-à-dire aussi bien une molécule (ADN, ARN, protéine), qu'un gène, une fonction physiologique ou une régulation.

6 Titre, auteur(s), résumé, revue, mots clés au sein d'une notice Genbank, avec la possibilité de pointer (hyperliens) la notice catalographique enrichie avec *abstract* sur PubMed.

7 Voir l'éditorial du premier numéro de la revue *Genomics* , en 1987. " Genomics: Structural and Functional Studies of Genomes. ". In : *Genomics* , vol 45, 1997, pp. 244-249.

8 Genbank *accession number* par exemple.

9 Appelées ainsi par les chercheurs, parce qu'elles contiennent les documents primaires que sont les enregistrements de séquences nucléotidiques (les notices).

10 Appelées ainsi par les chercheurs, parce qu'elles se construisent à partir des banques de données primaires, mais elles contiennent les mêmes types de documents primaires, seul leurs nombres diffèrent et sont fonction d'un projet ou d'une activité scientifique. L'objectif à atteindre est bien souvent la production des documents tertiaires qui synthétisent, sous forme de carte graphique par exemple, les documents primaires intégrés.

11 The DDBJ/EMBL/GenBank Feature Table Definition: http://www.ncbi.nlm.nih.gov/collab/FT/

12 PubMed Central, ex E-biomed, est un serveur de *preprints* qui devrait voir le jour en janvier 2000. Cette initiative est à comparer aux *preprints* de Ginsparg en physique ou de Harnad en science cognitive.

13 Gestionnaire de contenu.

14 Numéro d'enregistrement MedLine du champ MEDLINE, cliquable pour la version présente dans Entrez.

Glossaire/Glossary (et URL des sigles et acronymes)

- A, T, C, G : Adénine, Thymine, Cytosine, Guanine
- AceDB : A Caenorhabditis elegans DataBase (http://probe.nalusda.gov:8000/other/aboutacedb.html)
- ACNUC : ACide NUCléique (http://pbil.univ-lyon1.fr/databases/acnuc.html)
- ADN : Acide DésoxyriboNucléique
- AFNOR : Association Française de NORmalisation. (http://www.afnor.fr/)
- ARN : Acide Ribonucléique
- ASCII : American Standard Code for Information Interchange (http://www.kerryr.net/pioneers/lite/ascii1.htm)
- ASN.1 : Abstract Syntax Notation

- BACIP : *Bacillus* Industrial Platform
- Bioperl : http://bio.perl.org/
- BLAST : Basic Local Alignment Search Tool
- BNF : Backus-Naur Form (http://www.cs.dartmouth.edu/~samr/Courses/CS68.97W/Outlines/Nodes/intro-bnf.htm)
- CGI : Common Gateway Interface
- CORBA : Common Object Request Broker Architecture (http://www.omg.org, http://corba.ebi.ac.uk/)
- CSCW : Computer Supported Cooperative Work
- DBI : DataBase Interface for Perl
- DDBJ : DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/)
- DLI : Digital Library Initiative (http://ki.grainger.uiuc.edu/default.htm)
- DOE : Department Of Energy
- DTD : Document Type Definition
- EBI : European Bioinformatics Institute (http://www.ebi.ac.uk/)
- ECDC : E. Coli Data Collection
- EcoCyc : Encyclopedia. of E. Coli Genes and Metabolism (http://ecocyc.PangeaSystems.com/ecocyc)
- EMBL :European Molecular Biology Laboratory (http://www.embl-heidelberg.de/)
- Entrez : http://www.ncbi.nlm.nih.gov/Entrez/
- FASTA : Fast Alignment Search Tool (All)
- FTP : File Transfer Protocol
- GDB : Genome DataBase
- GenBank : Genetic sequences data Bank (http://www2.ncbi.nlm.nih.gov/Genbank/index.html)
- HGP : Human Genome Project
- HTML : HyperText Markup Language
- ICSU : International Council of Scientific Unions (http://www.icsu.org/)
- IDL : Interface Definition Language
- IETF : Internet Engineering Task Force (http://www.ietf.org/)
- IIOP : Internet Inter-ORB Protocol
- INRA : Institut National de la Recherche Agronomique (http://www.inra.fr/)
- ISO : International Standard Organization (http://www.iso.ch/)
- IUBMB : International Union Of Biochemistry And Molecular Biology (http://www.chem.qmw.ac.uk/iubmb/)
- IUPAC : International Union Of Pure And Applied Chemistry (http://www.chem.qmw.ac.uk/iupac/)
- JCBN : IUPAC-IUB Joint Commission on Biochemical Nomenclature (http://www.chem.qmw.ac.uk/iupac/jcbn/)
- LANL : Los Alamos National Laboratory (http://www.lanl.gov/)
- LSR : Life Science Research (Domain Task Force)
- MEDLINE : MEDlars onLINE, PubMed (http://www3.ncbi.nlm.nih.gov/PubMed/)
- MICADO : MICrobial Advanced Database Organization (http://locus.jouy.inra.fr/micado)
- MIPS : Munich Information Center for Protein Sequences (http://www.mips.biochem.mpg.de/)
- NBRF : National Biomedical Research Foundation
- NCBI : National Center for Biotechnology Information (http://www3.ncbi.nlm.nih.gov/)
- NC-IUBMB : Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (http://www.chem.qmw.ac.uk/iubmb/nomenclature/)
- NIGMS : National Institute of General Medical Sciences (http://www.nih.gov/nigms/)
- NIH : National Institutes of Health (http://www.nih.gov/)
- NLM : National Library of Medicine (http://www.nlm.nih.gov/)
- ODMG : Object Data Management Group (http://odmg.org/)
- OLE/COM : Object Linking and Embedding/Component Object Model
- OMG : Object Management Group (http://www.omg.org/)
- OQL : Object Query Language
- ORB : Object Request Broker
- OSI : Open Systems Interconnection
- PDB : Protein Data Bank (http://www.rcsb.org/pdb/)
- PDF : Portable Document Format
- PIR : Protein Identification Resource (http://www-nbrf.georgetown.edu/)
- RFC : Request For Comment (http://www.urec.cnrs.fr/ipv6/RFC-1880)
- SGBD : Système de Gestion de Base de Données
- SGML : Standard Generalized Markup Language (http://www.sgml.org)
- SMTP : Simple Mail Transfer Protocol
- SQL : Structured Query Language
- SRS : Sequence Retrieval System

- SWISS-PROT : protein sequence database (http://www.expasy.ch/sprot/sprot-top.html)
- TCP/IP : Transfer Control Protocol/Internet Protocol
- TrEMBL : Translations of EMBL (http://www.expasy.ch/sprot/sprot-top.html)
- W3C : World Wide Web Consortium (http://www.w3c.org)
- WCS : Worm Community System. (http://www.canis.uiuc.edu/)
- WWW : World Wide Web (http://www.w3c.org)
- XML : eXtensible Markup Language (http://www.xml.org et http://www.w3c.org/XML/)

Pour continuer la lecture

- To Know Ourselves : http://www.ornl.gov/hgmis/publicat/tko/index.html
- Nucleic Acids Research, Instructions to Authors. http://www3.oup.co.uk/nar/instauth/
- Guidelines for Human Gene Nomenclature (1997) : http://www.gene.ucl.ac.uk/nomenclature/guidelines.html
- The DDBJ/EMBL/GenBank Feature Table Definition: http://www.ncbi.nlm.nih.gov/collab/FT/
- Éditorial (1997). " Genomics: Structural and Functional Studies of Genomes. " In : *Genomics* , vol. 45, 1997, pp. 244-249.
- GRAY, P. ; KEMP, G. ; RAWLINGS, C. ; BROWN, N. ; SANDER, C. ; THORNTON, J. ; ORENGO, C. ; WODAK, S. & RICHELLE, J. (1996). " Macromolecular Structure Information and Database. The EU BRIDGE Database Project Consortium. ". In : *Trends in Biochemical Sciences* , vol. 21, 1996, pp. 251-256. [PubMed]
- HÉNAUT, A. (1997). " Information en Biologie. ". In : CACALY, S. (sous la dir.) (1997). *Dictionnaire Encyclopédique de l'Information et de la Documentation* . Paris, Nathan Université, 1997, pp. 303-306.
- LOMME, L. (1998). " L'Information Électronique en Biologie Moléculaire ". In : *Documentaliste - Science de l'Information* , vol. 35, nº3, 1998, pp. 179-185.
- SCHATZ, B.R. & CHEN, H. (1996). " Building Large-Scale Digital Libraries. " In : *IEEE Computer* [Theme Issue on the US Digital Library Initiative], May 1996. [Internet] consulté en septembre 1999 : IEEE Computer
- WOOD, R. (1998). "Genetic Nomenclature Guide". In : *Trends in Genetics* (supplement), 1998.
- DESSEN, Ph. (1995). "Les secrets de la séquences". In : *Biofutur* , nº146, (spécial génomes), pp.39-43.
- DANCHIN, A. *et al.* (1996). "*Bacillus subtilis* dévoile ses gènes". In : *Biofutur* , nº174, janvier 1998, pp.14-17.
- DANCHIN, A. (1993). "La séquences des petits génomes - Vers la description complète d'un organisme vivant". In : *La Recherche* , vol.24, nº251, fév 1993, pp.222-232.
- LOGOZE, C. et FIELDING, D. (1998). "Defining Collections in Distributed Digital Librairies". In : *Dlib magazine* , Nov 1998. [Internet] http://www.dlib.org/november98/lagoze/11lagoze.html
- CHEN, H. (1999). "Semantic Research for Digital Libraries". In : *Dlib magazine* , Oct. 1999. [Internet] http://www.dlib.org/october99/chen/10chen.html

Remerciements / Acknowledgments

Nous remercions S. Dusko EHRLICH, Directeur du Laboratoire de Génétique Microbienne de l'INRA, au sein duquel ont été effectués ces travaux, Étienne DERVYN, pour ses conseils vigilants en tant qu'expérimentateur et usager de la base de données Micado depuis sa création, Laurent BIZE, pour le déploiement d'applications d'exploration des données sur le système d'information, et Sandrine DUCHET, actuellement en charge du quotidien de ce dernier.

Bibliographie/references

ACHARD, F. & DESSEN, P. (1998). " GenXref. VI: Automatic Generation of Links between Two Heterogeneous Databases. " In : *Nucleic Acids Research* , vol. 27, nº1, 1999, pp. 113-114. [PubMed]

ACHARD, F. ; VAYSSEIX, G. ; DESSEN, P. & BARILLOT, E. (1999). " Virgil Database for Rich Links (1999 update). " In : *Bioinformatics* , vol. 14, nº1, 1998, pp. 20-24. [PubMed]

APPEL, R.D. ; BAIROCH A. & HOCHSTRASSER, D.F. (1994). " A New Generation of Information Retrieval Tools for Biologists: the Example of the ExPASy WWW Server. " In : *Trends in Biochemical Sciences* , vol. 19, nº6, 1994, pp. 258-260. [PubMed]

BAKER, P.G. & BRASS, A. (1998). " Recent Developments in Biological Sequence Databases. " In : *Current Opinion in Biotechnology* , vol. 9, nº1, 1998, pp. 54-58. [PubMed]

BAKER, P. G. ; GOBLE, C.A. ; BECHHOFER, S. ; PATON, N.W. ; STEVENS, R. & BRASS, A. (1999). " An Ontology for Bioinformatic Applications. " In : *Bioinformatics* , vol. 15, nº6, 1999, pp. 510-520. [PubMed]

BALZT, C. (1998). " Une Culture pour la Société de l'Information, Position Théorique, Définition, Enjeux. " In : *Documentaliste - Science de l'Information* , vol. 35, nº2, 1998, pp. 75-85.

BENTON, D. (1996). " Bioinformatics - Principles and Potential of a New Multidisciplinary Tool. " In : *Trends in Biotechnology* , vol. 14, nº8, 1996, pp. 261-272. [PubMed]

BIAUDET, V. ; SAMSON, F. & BESSIÈRES, P. (1997). " Micado: a Network Oriented Database for Microbial Genomes. " In : *Computer Applications in the Biosciences* , vol. 13, nº4, 1997, pp. 431-438. [PubMed]

BIAUDET, V. ; SAMSON, F. ; ANAGNOSTOPOULOS, C. ; EHRLICH, S.D. & BESSIÈRES, P. (1996). " Computerized Genetic Map of Bacillus subtilis. " In : *Microbiology* , vol. 142, nº10, 1996, pp. 2669-2729. [PubMed]

BIZE, L. ; MURI, F. ; SAMSON, F. ; RODOLPHE, F. ; EHRLICH, S.D. ; PRUM, B. & BESSIÈRES, P. (1999). " 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.

BORK, P. & BAIROCH, A. (1996). " Go Hunting in Sequence Databases but Watch Out for the Traps. " In : *Trends in Genetics* , vol. 12, nº10, 1996, pp. 425-427. [PubMed]

BRIET, S. (1951). *Qu'est-ce que la Documentation* . Paris, Éditions Documentaires Industrielles et Techniques (ÉDIT), 47 p.

BUCKLAND, M.K. (1997). " What Is a Document? ". In : *J. Am. Soc. Information Science* , vol. 48, nº9, 1997, pp.804-809.

BUTTLER, D. (1999). " US Biologists Propose Launch of Electronic Preprint Archive. " In : *Nature* , vol. 397, 1999, p. 91.

CEVEIL (1995). " Préambule à la Normalisation. " In : BOURBEAU, L. & PINARD, F. (1995). *Normalisation et Internationalisation* . CEVEIL. [Internet] consulté en septembre 1999 : http://www.ceveil.qc.ca/Normes/preamb.html

CHARTRON, G. (1994). " Nouvelles Technologies et Organisations de Travail Coopératif : Quelques Repères. " In : *Solaris* , nº1, 1994. [Internet] consulté en mars 1998 : Solaris

CHERVITZ, S.A. ; FUELLEN, G. ; DAGDIGIAN, C. ; RESNICK, R. & BRENNER, S.E. (1997). " Bioperl: Object-Oriented Perl Modules for Bioinformatics. " In : *OIB'97 - Objects In Bioinformatics Conference, Cambridge* , UK. [Internet] consulté en octobre 1999 : http://industry.ebi.ac.uk/~alan/Meeting/SpeakerAbstracts/Chervitz.html

COCKERILL, M. (1994). " A Versatile Tool for Retrieving Molecular Sequences: Entrez. " In : *Trends in Biochemical Sciences* , vol 19, nº2, 1994, pp. 94-96.

CORNISH-BOWDEN, A. (1985). " Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences: Recommendations 1984. " In : *Nucleic Acids Research* , vol. 13, nº9, 1985, pp.3021-3030. [PubMed]

DAVIDSON, S.B. ; OVERTON, C. & BUNEMAN P. (1995). " Challenges in Integrating Biological Data Sources. " In : *J. Computational Biology* , vol. 2, nº4, 1995, pp. 557-572. [PubMed]

DAYHOFF, M.O. ; ECK, R.V. ; CHANG, M.A. & SOCHARD, M.R. (1965). *Atlas of Protein Sequence and Structure* ., vol. 1, Silver Spring, MD, National Biomedical Research Foundation.

DESSEN, P. ; FONDRAT, C. ; VALENCIEN, C. & MUGNIER C. (1990). " BISANCE: a French Service for Access to Biomolecular Sequence Databases. " In : *Computer Applications in the Biosciences* , vol. 6, nº4, 1990, pp. 355-356. [PubMed]

DUJON, B. (1996). " The Yeast Genome Project: What Did we Learn? " In : *Trends in Genetics* , vol. 12, nº7, 1996, pp. 263-270. [PubMed]

DURBIN, R. & THIERRY-MIEG, J. (1991). "A Caenorhabditis elegans Database". [Internet] consulté en octobre 1999 : ftp://lirmm.lirmm.fr, ftp://cele.mrc-lmb.cam.ac.uk & ftp://ncbi.nlm.nih.gov

EHRLICH, S.D. & OGASAWARA, N. (1999). " Functional Analysis of Bacillus subtilis. " In : *3rd Conf. on Microbial Genomes: Sequencing, Functional Characterization and Comparative Genomics* , Chantilly, VA, USA.

ETZOLD, T. & ARGOS, P. (1993a). " SRS - an Indexing and Retrieval Tool for Flat File Data Libraries. " In : *Computer Applications in the Biosciences* , vol. 9, nº1, 1993, pp. 49-57. [PubMed]

ETZOLD & ARGOS (1993b). " Transforming a Set of Biological Flat File Libraries to a Fast Access Network. " In : *Computer Applications in the Biosciences* , vol. 9, nº1, 1993, pp. 59-64. [PubMed]

FUCHS, R. ; RICE, P. ; CAMERON, GN. (1992) " Molecular Biological Databases - present and future ". In : *Trends Biotechno.* .vol. 10, N°1-2, 1992, pp.61-66 [PubMed]

GAS, S. ; EGGEN, A. ; SAMSON, F. ; CHRISTOPHE, C. ; MUNGALL, C. ; BESSIÈRES, P. & LEVÉZIEL, H. (1996). " The Bovmap Database. ". In : *XXVth. Intl. Conf. on Animal Genetics* , Tours, France.

GOUY, M. ; MILLERET, F. ; MUGNIER, C. ; JACOBZONE, M. & GAUTIER, C. (1984). " ACNUC: a Nucleic Acid Sequence Database and Analysis System. " In : *Nucleic Acids Research* , vol. 12, nº1, 1984, pp.121-127. [PubMed]

GOUY, M. ;????? (1985). " ACNUC - A Portable Retrieval System for Nucleic Acid Sequence Databases: Logical and Physical Designs and Usage.;" In : *Comput. Appl. Biosci.* , vol. 1, n??;3, 1985, pp.167-172. [PubMed]

HAKALA, J. & HORMIA, K. (1997). "Digital Library Initiative Projects". 1997. [Internet] consulté en septembre 1998 : http://linnea.helsinki.fi/meta/dli.html

HARPER, R. (1994). " Access to DNA and Protein Databases on the Internet. ". In : *Current Opinion in Biotechnology* , vol. 5, 1994, pp. 4-18. [PubMed]

HARPER, R. (1995). " World Wide Web Resources for Biologists. ". In : *Trends in Genetics* , vol. 11, nº6, 1995, pp. 223-228. [PubMed]

HAYTER, J. (SD). " Life Sciences Research Standards. " In : *BITS Journal* , [Internet] consulté en septembre 1999 : BITS Journal

HIETER, P. & BOGUSKI, M. (1997). " Functional Genomics : It's All How You Read It. ". In : *Science* , vol.

278, 1997, pp. 601-602. [PubMed]

IUPAC-IUB Commission on Biochemical Nomenclature (CBN) & MOSS G. P. (1970). "Abbreviations and Symbols for Nucleic Acids, Polynucleotides and their Constituents". [Internet] consulté en septembre 1999 : <http://www.chem.qmw.ac.uk/iupac/misc/naabb.html>

IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) (1984). " Nomenclature and Symbolism for Amino Acids and Peptides. Recommendations 1983. ". In : *European J. Biochemistry* , vol. 138, nº1, 1984, pp. 9-37.

KARP, P.D. (1996). " Database Links are a Foundation for Interoperability. " In : *Trends in Biotechnology* , vol. 14, nº8, 1996, pp. 273-279. [PubMed]

KRISTOFFERSON, D. (1987). " The BIONET Electronic Network. ". In : *Nature* , vol. 325, 1987, pp. 555-556.

KUNST, F. et al. (1997). " The Complete Genome Sequence of the Gram-positive Bacterium Bacillus subtilis. ". In : *Nature* , vol. 390, 1997, pp. 249-256. [PubMed]

LETOVSKY, S. (1995). " Beyond the Information Maze. ". In : *J. Computational Biology* , vol. 2, nº4, 1995, pp. 539-546. [PubMed]

MÉDIGUE, C. ; VIARI, A. ; HÉNAUT, A. & DANCHIN, A. (1993). " Colibri: a Functional Data Base for the Escherichia coli Genome. " In : *Microbiological Rev.* , vol. 57, 1993, pp. 623-54. [PubMed]

MELTON, J. & SIMON, A.R. (1993). *Understanding the New SQL : a Complete Guide* . San Francisco, Californie : Morgan-Kaufmann Publisher.

MICHARD, A. (1999). *XML, Langage et Applications* . Paris : Eyrolles, 361 p.

MOSZER, I. (1998). " The Complete Genome of Bacillus subtilis: from Sequence Annotation to Data Management and Analysis. ". In : *FEBS Letters* , vol. 430, 1998, pp. 28-36. [PubMed]

MOSZER, I. ; GLASER, P. & DANCHIN, A. (1995). " SubtiList: a Relational Database for the Bacillus subtilis Genome ". In : *Microbiology* , vol. 141, 1995, pp. 261-268. [PubMed]

PERRIÈRE, G. ; BESSIÈRES, P. & LABEDAN, B. (1999). " The Enhanced Microbial Genomes Library. ". In : *Nucleic Acids Research* , vol. 27, 1999, pp. 63-65. [PubMed]

POOL, R. (1993). " Beyond Databases and E-Mail. " In : *Science* , vol. 261, 1993, pp. 841-843.

PROVANSAL, A. (1997). " Notice. ". In : CACALY, S. (sous la dir.) (1997). *Dictionnaire Encyclopédique de l'Information et de la Documentation* . Paris, Nathan Université, pp. 429-431.

RAPPAPORT, B. (1997). " From Syntax to Semantics. ". In : *Nature Biotechnology* , vol. 15, 1997, p. 1228.

SAMSON, F. ; BIAUDET, V. & BESSIÈRES, P. (1998). " Viewing Microbial Genome Maps with Java and CORBA. " In : *OIB'98 - Objects In Bioinformatics Conference, Cambridge* , UK. [Internet] consulté en octobre 1999 : OIB'98

SCHAMBER, L. (1996). "What is a Document ? Rethinking the Concept in Uneasy Times ". In : *J. Am. Soc. Information Sciences* , vol. 47, nº9, 1996, pp. 669-671.

SCHATZ, B.R. (1993). " A Model Collaboratory - the Worm Community System. " In : *NRC report on National Collaboratories* , 1993. [Internet] consulté en septembre 1998 : http://www.canis.uiuc.edu/projects/wcs/national_collaboratories.html

SCHATZ, B.R. (1997). " Information Retrieval in Digital Librairies : Bringing Search to the Net ". In : *Science* , vol. 275, 1997, pp. 327-333. [PubMed]

SCHULZE-KREMER, S. (1998). " Ontologies for Molecular Biology ". In : *PSB 98 On-Line Proceedings* , 1998. [Internet] consulté en septembre 1999 : http://www.smi.stanford.edu/projects/helix/psb98/schulze-kremer.pdf

SCHNEIDERMAN, B. (1997). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* . Addison-Wesley Publishing Company, Reading, MA.

STAFFS, DDBJ/EMBL/GenBank. (1999). Feature Table Definition : http://www.ncbi.nlm.nih.gov/collab/FT/

SUTTER, É. (1997). " Norme. " In : CACALY, S. (sous la dir.) (1997). *Dictionnaire Encyclopédique de l'Information et de la Documentation* . Paris : Nathan Université, 1997, pp. 428-429.

TURNER, W.A. (1995). " Les Professionnels de l'Information Auront-ils une Place dans les Collaboratoires de la Recherche ? " In : *Solaris* , nº2, 1995. [Internet] consulté en mars 1999 : http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2turner.html

VARMUS, H. (1999). E-BIOMED. [Internet] consulté en mai 1999 : http://www.nih.gov/welcome/director/ebiomed/ebi.htm

VICKERY, B.C. (1997). " Ontologies. " In : *J. Information Science* , vol. 23, 1997, pp. 277-286.

WATERS, D.J. (1998). " What Are Digital Libraries? " In : *CLIR issues* , nº4, 1998. [Internet] consulté en octobre 1999 : http://www.clir.org/pubs/issues/issues04.html#dlf

WATSON, J.D. (1990). " The Human Genome Project: Past, Present, and Future. " In : *Science* , vol. 248, 1990, pp. 44-49. [PubMed]

WELLER, A.C. (1996). " The Human Genome Project. " In : CRAWFORD, S.Y. ; HURD, J.M. & WELLER, A.C. (1996). *From Print to Electronics, the Transformation of Scientific Communication* . Medford, Information Today Inc. (ASSIS Monograph Series), 1996, pp. 46-73.

WESTPHAL, C. et BLAXTON, T. (1998) " Data Mining Solutions - Methods and Tools for Solving Real-World Problems ". Wiley, USA, 1998, 617 p.

WHITE J. A. ; MALTAIS, L.J. & NEBERT D.W. (SD). " An Increasingly Urgent Need for Standardized Gene Nomenclature. " In : *Nature Genetics* . [Internet] consulté en septembre 1999 : http://genetics.nature.com/web_specials/nomen/nomen_article.html

WILBUR, W.J. (1992). " A retrieval system based on automatic relevance weighting of search terms ". In : *Shaw, D., Proceeding of the 55th American Society of Information Science Annual Meeting* ,Pittsburgh, PA : Learned Information 1992, pp. 216-220.

WILLIAMS, N. (1997a). " How to Get Databases Talking the Same Language? ". In : *Science* , vol. 275, 1997, pp. 301-302.

WILLIAMS, N. (1997b). " Drug Firms Back Move to Link Databases ". In : *Science* , vol. 277, 1997, p. 902.
