# La recherche in silico

Gabriel Gallezot

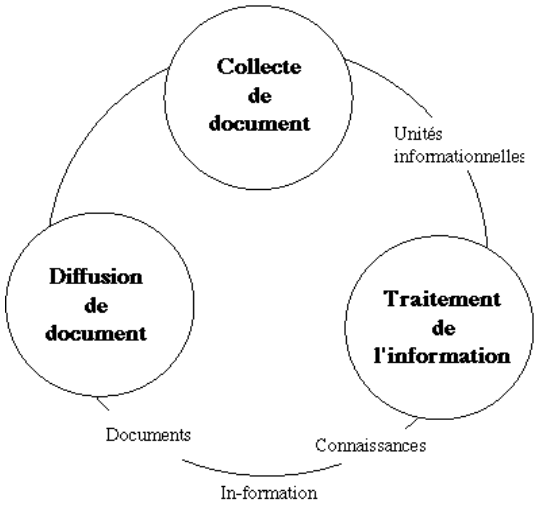
# La recherche *in silico*

Version de l'auteur > Gallezot G., "La recherche *in silico*", In : Chartron G. (sous la dir.) "*Les chercheurs et la documentation numérique : nouveaux serviceset usages*", Edition du cercle de la Librairie, Collections, **Juillet 2002**.

**Gabriel Gallezot**, gallezot@unice.fr
Chercheur au GRESI, Groupe de REcherche sur les Services d'Information, direction 2 : Information scientifique et numérique, ENSSIB, Lyon.

La recherche *in silico* [1]. Ce vocable indique le début et l'ampleur d'un phénomène en biologie moléculaire: les recherches ne sont plus seulement *in vivo* ou *in vitro*, mais ont un recours de plus en plus essentiel aux analyses informatiques. Il souligne ainsi l'importance des technologies de l'information et de la communication (TIC) dans le développement de cette discipline et en désigne surtout deux champs spécifiques : la génomique et la bioinformatique. Nous présentons ici un modèle explicatif du cycle de l'information scientifique et technique (IST) dans cette communauté par le biais d'une observation participante effectuée au sein d'un laboratoire de génétique des microbes à l'INRA[2]. Nous avons appréhendé l'appropriation de l'information numérique par les chercheurs à l'aide d'un modèle heuristique construit sur le cycle de vie du document (acquisition, recherche, archivage…) et de transformation de l'information (informations / documents / connaissances)[3] intitulé cycle de l'information scientifique et technique (*cf.* figure 1).

*figure 1 : le cycle de l'information scientifique et technique*




---

1 Néologisme qui signifie littéralement "dans le silicium". Une recherche sur les titres et sommaires des articles référencées dans Pubmed fait remonter sa première mention en 1991, par des microbiologistes danois dans une revues de l'institut pasteur (Hansen et al. In : *Research in microbiology*, vol. 142, n°2-3, pp161-167). Trois articles l'emploient en 1993 et son usage se généralise à partir de 1996.

2 Laboration de génétique microbienne, INRA de Jouy-en-Josas. Travail de thèse : Gallezot G., *Techniques de l'information, usages de l'IST et construction des connaissances des chercheurs en génomique*, doctorat, Université de Paris 1, sous la direction de S. Fayet-Scribe, 2000.

3 Pour plus de détails sur ces concepts nous renvoyons le lecteur à :

Blanquet M.F., S*cience de l'information et philosophie* Paris, ADBS éditions, 1997.

Le Coadic Y.-F., *La science de l'information*, Paris, PUF, Coll. Que sais-je, 1994.

Meadows A. J., *Communicating Research*, Academic Press, 1998.

L'activité informationnelle des chercheurs est circonscrite par ce modèle. Le “bouclage”[4] d’un cycle induit la construction de connaissances, d’informations et de documents. La collecte peut être réalisée à partir de banques de données, de sites web, d’expériences dans les laboratoires, de “butinage” (*browsing*) dans les rayonnages d’une bibliothèque... Le traitement correspond à l’activité cognitive des chercheurs ou à des manipulations par des outils informatiques. La diffusion est définie comme l’ensemble des opérations nécessaires à la propagation des connaissances. La transformation de l'information s'inscrit dans ce processus où la connaissance est la formation des idées, l'information est la mise en forme des connaissances (in-formation) et l'information inscrite sur un support constitue un document. Le concept d'unité informationnelle veut mettre en exergue la “granularité” de l'information et désigne une partie de document à collecter ou à traiter[5].

# I - De la biologie moléculaire à la génomique et la bioinformatique

## 1.1 - La biologie moléculaire

La biologie moléculaire occupe une place dominante dans les sciences biologiques. Cette “vision moléculaire” du vivant est née de la génétique et de la biochimie. On peut dater l’émergence de cette nouvelle discipline scientifique avec la mise en évidence de l’ADN comme vecteur de l’hérédité par O.T. Avery et C.M. McLeod (1944), puis la découverte de sa structure (double hélice) par J.D. Watson et F.H.C. Crick (1953).

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 dans son champ d’investigations. A 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’apport de la biologie moléculaire est souvent réduit à l’acquisition d’un ensemble de techniques : elle n’aurait fait qu’apporter aux biologistes des outils leur permettant d’étudier et de maîtriser le vivant. Cette conception est totalement erronée : c’est une illusion de penser que l’utilisation des techniques peut être dissociée de la nature des connaissances que ces techniques permettent d’acquérir. La biologie moléculaire est une révolution

---

4 Le terme “bouclage” désigne la réalisation des différentes étapes du cycle, le passage de la collecte au traitement et du traitement à la diffusion.

5 Guinchat et Skouri dans [Guinchat C.et Skouri Y. *Guide pratique des techniques documentaires*, Vanves, Edicef, 1996, vol.2, p.44.] introduisent la notion "d'unité documentaire" pour indiquer qu'une partie d'un document comme une figure ou un paragraphe peuvent être référencés pour peut que ces extraits portent la mention de son origine. Nous introduisons ici la notion d'unité informationnelle pour signifier qu'un document peut être perçu comme un agrégat d'unités de connaissances mises en forme (in-formation) que l'on peut décomposer à souhait pour peut que le document de référence puisse être retrouvé (traçabilité).

> scientifique, une nouvelle vision du monde du vivant, que la mise au point d'un ensemble de techniques a rendue efficace et opératoire[6].”

Comme Morange l'écrit ci-dessus, les techniques servent la connaissance qui sert la technique : c'est un processus récurrent (cyclique) qui inscrit dans le temps la construction de connaissances et la réalisation de techniques. Ainsi, il est impossible de dissocier la technique de la science en biologie moléculaire, quelque soient ces techniques: celles directement liées à la manipulation du vivant[7] et celles plus générales de l'informatique[8] (traitement automatique de l'information).

## 1.2 - De l'importance des techniques

A travers “les histoires parallèles”[9] de l'informatique et de la biologie moléculaire des périodes de “synchronie” mettent en évidence l'alliance des techniques de la génomique et de l'informatique.

Dans les années 1970, le développement des techniques de génie génétique comme le clonage (1974), ont fait “décoller” la biologie moléculaire née trente ans plus tôt, avec la découverte de l'ADN comme support d'information génétique. Cette période voit aussi la naissance du réseau Arpanet, premier composant du futur réseau Internet. Le premier micro-ordinateur et les systèmes de gestion de bases de données (SGBD) hiérarchique et réseau font leur apparition. Ainsi, on peut percevoir cette période comme “l'enfance” des techniques informatiques et génétiques.

Dans les années 1980, les techniques évoluent. Les micro-ordinateurs sont équipés de disques durs, de modems et les SGBD deviennent relationnels. La structure primaire (séquence) de l'ADN est analysée et des banques de données factuelles (la représentation de séquences nucléotidiques) internationales collectent ces résultats d'expériences. Ce dernier fait prend de l'ampleur avec la décision des éditeurs de revues scientifiques de ne plus publier le “texte” de la séquence analysée dans la publication. Ainsi, les biologistes peuvent disposer plus efficacement des travaux des autres. Cette période transitoire marque le début de l'interaction des techniques informatiques et de la biologie moléculaire. Elle est appelée “adolescence” pour signifier l'évolution opérée sur la précédente période et le début d'appropriation des techniques informatiques par les biologistes.

Dans les années 1990, l'ADN des gènes de chromosome(s)[10] est systématiquement analysé (les petits génomes en début de période, comme ceux des bactéries, puis en fin de période des génomes plus importants comme celui de la drosophile). C'est un changement important d'échelle, de point vue. L'analyse du génome est dorénavant une étape incontournable et essentielle dans l'étude du

---

6 Morange M., “Bréve histoire de la biologie moléculaire”, *Biofutur*, 1995, n°142, p.15.

7 Les enzymes de restriction, le clonage, l'électrophorèse, l'hybridation, la PCR (*Polymerase Chain Reaction*), le séquençage, le transcriptome (filtres à hautes densité, biopuces (DNAchips), et Micro Array), le protéome. (électrophorèse bi-dimensionnelle), le double hybride (construction d'une carte d'interaction entre les protéines, elle est basée sur un test biologique : l'activation ou non de la transcription).

8 Ordinateurs, langages informatiques et systèmes d'exploitation, Internet (techniques, langage, protocoles), les Systèmes de Gestion de Bases de Données (SGBD), ...

9 Flichy P. *L'innovation technique : récents développements en sciences sociales. Vers une théorie de l'innovation*. Paris, Ed. la découverte, 1995.

10 La totalité de l'ADN des chromosomes propres à chaque espèce constitue le génome. Le génome porte la quasi totalité de l'information requise pour définir les propriétés d'un organisme.

vivant. Le séquençage réalisé avec des séquenceurs informatisés[11] et des projets internationaux d'analyse de génomes impulsés par le *Human Genome Project* (HGP) contribuent à l'accroissement rapide du nombre de séquences dans les banques internationales. Parallèlement, l'informatique fait, elle aussi, un saut quantitatif : vitesse de traitement des processeurs, capacité des mémoires vives et de stockage,… La loi de Moore qui dit que la vitesse des processeurs double tous les 18 mois est toujours valable. Elle peut, dans une certaine mesure, s'appliquer à l'ensemble des techniques informatiques. L'ordinateur est présent un peu partout et l'informatique, le traitement automatique de l'information, n'a jamais été aussi prépondérant. De machine qui servait principalement à la gestion comptable dans les années 1970, l'ordinateur s'est imposé dans tous les champs de l'activité humaine. Cette dissémination des ordinateurs et des micro-ordinateurs, doublée de l'apparition des techniques liées au Web (langage HTML, serveur HTTP et navigateurs) participent à la mondialisation du réseau Internet. Cette quatrième période, appelée "une certaine maturité", montre une appropriation affirmée des techniques par les biologistes et l'alliance croissante de l'informatique et de la biologie moléculaire.

## 1.3 - La génomique

L'impulsion du programme de séquençage complet du génome humain (le HGP), puis de celui d'autres génomes, et enfin la généralisation et l'évolution concomitante des séquenceurs, ont fait croître de manière exponentielle la production des séquences d'ADN. Cette recherche d'exhaustivité, liée au fait que toute l'information génétique nécessaire à un organisme est contenue dans son ADN, 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.

Le terme "génomique"[12] émerge de débats relatifs aux différents colloques organisés entre 1984 et 1987. 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, élément qui n'avait pas été mis en évidence auparavant par les méthodes classiques de la génétique[13]. Cette observation est 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"[14], qui regroupe les approches biochimiques et physiologiques adaptées à des analyses d'un génome entier et complétant 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, qui devront être croisées pour prédire les fonctions des gènes.

---

11 A la sortie des séquenceurs, l'information est numérisée et stockée sur un support magnétique (disquette ou disque dur).

12 Editorial du premier numéro de la revue *Genomics*, en 1987, "Genomics: Structural and Functional Studies of Genomes", *Genomics*, n°45,1997, p. 244-249.

13 Dujon B., "The Yeast Genome Project: What Did we Learn?", *Trends in Genetics*, vol. 12, n° 7, 1996, p. 263-270.

14 Hieter P., Boguski M., "Functional genomics : it's all how you read it", *Science*, Vol.278, 1997, p. 601-602.

## 1.4 - La bioinformatique

Le phénomène sans doute le plus marquant, relatif à l'informatique et à la biologie moléculaire, est l'émergence de la bioinformatique :

> […] [divers] spécialistes - généticiens, spécialistes de la biologie moléculaire, ingénieur informatique, chercheurs en informatique, mathématiciens et statisticiens – ont travaillé ensemble et drainé [utilisé] les bases de connaissances en information génétique..[15]

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. Encore à la marge de cette nouvelle discipline, et relevant des sciences de l'information, les données non factuelles, par exemple le texte des articles scientifiques, commencent à être exploitées (avec l'informatique) pour enrichir les connaissances en génomique.

La bioinformatique, traitement automatique de l'information biologique, s'est définitivement imposée avec les programmes d'analyse de génomes[16]. 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 de 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 la série des *Trends*, lui consacrent régulièrement des colonnes et des rubriques.

La bioinformatique peut être assimilée au passage de l'imprimé à l'électronique. D'une simple lecture de représentations de séquences nucléotidiques (quelques lignes de A.T.G.C., la réprésentation des nucléotides : Adénine, Thynine, Cytosine et Guanine) en sélectionnant ou en feuilletant des revues disponibles dans leur centre de documentation, les biologistes peuvent désormais, pour des données factuelles plus conséquentes, faire une recherche sélective et exhaustive, puis les collecter *via* Internet, les manipuler *in silico* et en proposer une représentation graphique sur écran.

La bioinformatique est la résultante de la nécessité de traiter l'information génomique et d'une forte appropriation des technologies informatiques par des chercheurs. De développements ponctuels, isolés et réalisés par des chercheurs en génomique ayant de bonnes compétences en informatique, la bioinformatique " s'institutionnalise ". Des formations universitaires apparaissent avec le label "bioinformatique ", des équipes regroupant des informaticiens, des mathématiciens et des biologistes se forment et des chercheurs en science de l'information apparaissent dans ces équipes. Du traitement d'un petit ensemble de données factuelles, la bioinformatique a désormais pour tâche

---

15 Weller A. C.," The Human Genome Project " In : Crawford S. Y., Hurd J. M., Weller A. C., *From Print to Electronic, The Transformation of Scientific Communication*, Medford, Information today Inc (Assis monograph series), Washington, USA, 1996, p.70

16 Benton D., " Bioinformatics -Principles and Potential of a New Multidiciplinary Tool ", *Trends in Biotechnology*, vol. 14, n° 8, 1996, p. 261-272.

de rassembler l'ensemble de l'information d'un domaine, d'un champ spécifique. Cela engendre deux conséquences :

- les développements informatiques sont d'une plus grande ampleur, ils convoquent et combinent différentes techniques,
- les informations traitées ne sont plus uniquement des données expérimentales, mais aussi des données "textuelles " (non factuelles) issues de la littérature scientifique.

On considère donc la génomique comme un aspect de la biologie moléculaire et la bioinformatique comme une activité scientifique adjacente à la génomique. Chacun de ces champs disciplinaires a sa spécificité, des revues dédiées, et un objectif commun : l'analyse des génomes.

# II - L'information en génomique

## 2.1 - L'information scientifique et technique en génomique

Trois types de documents sont à considérer : les données factuelles, les données textuelles et des informations "communautaires". Cette distinction est usuelle, mais révèle, en trois points, l'activité du chercheur. Schématiquement, les données factuelles représentent la partie expérimentation de son activité, les données textuelles sont le contexte cognitif et les informations communautaires constituent le liant de la vie scientifique.

### 2.1.1 - Les données factuelles

Elles sont issues de la paillasse ou des banques de séquences internationales. Ces banques, comme GenBank, EMBL ou DDBJ[17] pour les séquences d'ADN, donnent accès par FTP à leurs gisements de documents primaires, qui sont des fichiers informatiques de texte codé en ASCII *(American Standard Code for Information Interchange*). 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 (*cf.* figure 2) est organisé en champs, pour lesquels des descripteurs spécifient une information relative aux propriétés d'un objet biologique. Si chaque banque possède ses descripteurs, ou ses étiquettes, pour coder l'information suivant un format qui lui est propre, le contenu informationnel intrinsèque de l'objet biologique reste inchangé. Dans une notice, on peut distinguer quatre grandes 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,

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

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*. Il 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 notices évoluent avec les connaissances et le contexte de l'activité scientifique. Ainsi, un descripteur similaire à Medline concernant PubMed Central[18],'Open Archive Initiative ou toute autre archive numérique de textes scientifiques pourrait apparaître. De plus, si les banques de séquences d'ADN représentent une même information, mais dans un format propriétaire, le champ FEATURES fait l'objet d'une standardisation commune à toutes les banques (DDBJ/EMBL/GenBank[19]). 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.

*Figure 2 :* Notices Genbank (Extrait, la notice complète représente une vingtaine de pages, le pointillé indique les coupures)

```
GenBank (116.0, 03/10/2000)
Accession: L09228

GenBank (NCBI, Bethesda, Md. USA)

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"
```

18 Varmus H. [Consulté en mai 1999]. *E-BIOMED : A Proposal for Electronic Publication in the Biomedical Sciences.* [En ligne] http://www.nih.gov/welcome/director/ebiomed/ebi.htm

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

```
                     /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]
```

---

```
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
```

### 2.1.2 - Les données textuelles

Elles sont représentées par la littérature scientifique au sens large du terme. Dans cet ensemble on trouve les articles des revues, les ouvrages scientifiques, les actes de colloques... Mais aussi les notices catalographiques des banques de données bibliographiques comme Medline (*cf.* figure 3). Au même titre que les notices Genbank, la notice Medline présentée ci-dessous est un fichier ASCII avec des descripteurs spécifiques. On peut distinguer les éléments qui relèvent d'une notice catalographique classique (auteurs, titre, résumé, revue, date, descripteurs du thésaurus...) et ceux qui ont trait aux liens avec les objets biologiques (les gènes, l'enregistrement dans une banque de séquence). Récemment, ces notices ont fait l'objet d'un formatage XML[20] . Ainsi, avec la DTD[21] et les filtres *ad hoc*, il est devenu plus aisé de les manipuler. Nous insistons particulièrement sur la notice catalographique car des développements en bioinformatique l'utilisent dans le cadre de projets d'extraction automatique d'informations. Par exemple, comme la majeure partie des connaissances biologiques sur les interactions géniques n'est pas décrite dans les banques mais dans les articles scientifiques, l'exploitation des résumés des notices et plus généralement des textes scientifiques constitue un enjeu central dans la construction des modèles d'interaction entre gènes ou encore dans la contextualisation de données issues de l'expérimentation[22].

*Figure 3* : Notice Medline

---

20 eXtensible Markup language : http://www.w3.org/XML/1999/XML-in-10-points

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

22 Bessières P., Nazarenko, Nedellec C., *Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques. texte accepté pour* CIDE'2001 -

Nedellec C., Ould Abdel Vetah M., Bessières P., " Sentence filtering for information extraction in genomics, a classification problem ", *PKDD'01 - Proceedings of the Conference on Practical Knowledge Discovery in Databases*, Springer Verlag, Freiburg, September 2001.

```
UI  - 95020538
AU  - Sorokin A
AU  - Zumstein E
AU  - Azevedo V
AU  - Ehrlich SD
AU  - Serror P
TI  - The organization of the Bacillus subtilis 168 chromosome region between
      the spoVA and serA genetic loci, based on sequence data.
LA  - Eng
MH  - Amino Acid Sequence
MH  - Bacillus subtilis/*genetics
MH  - Base Sequence
MH  - *Chromosomes, Bacterial
MH  - Cloning, Molecular
MH  - Comparative Study
MH  - Genes, Bacterial/*genetics
MH  - Molecular Sequence Data
MH  - Operon/genetics
MH  - Sequence Analysis, DNA
MH  - Sequence Homology, Amino Acid
MH  - Support, Non-U.S. Gov't
PT  - JOURNAL ARTICLE
DA  - 19941109
DP  - 1993 Oct
IS  - 0950-382X
TA  - Mol Microbiol
PG  - 385-95
SB  - M
CY  - ENGLAND
IP  - 2
VI  - 10
JC  - MOM
AA  - Author
EM  - 199501
AB  - Three different lambda phage clones with overlapping inserts of Bacillus
      subtilis DNA, which cover the region from spoIIAA to serA, have been
      [...]
      basis of the sequences of their transcription terminators, promoters and
      regulatory elements.
AD  - Laboratoire de Genetique Microbienne, Institut National de la Recherche
      Agronomique, Jouy en Josas, France.
PMID- 0007934829
GS  - spoIIA
GS  - aroC
GS  - spoV
GS  - spoA
GS  - spoF
GS  - lysA
GS  - ppiB
GS  - sipS
GS  - ribG
GS  - ribH
GS  - serA
SI  - GENBANK/L09228
SI  - GENBANK/M15349
SI  - GENBANK/D90189
SI  - GENBANK/X17013
SI  - GENBANK/Z11847
SI  - GENBANK/X51510
SI  - GENBANK/M84227
EDAT- 1993/10/01 00:00
MHDA- 1993/10/01 00:00
SO  - Mol Microbiol 1993 Oct;10(2):385-95
```

### 2.1.3 - Les informations “communautaires”

Elles sont représentées par l'ensemble des documents issus des listes de discussion (*mailing list)*, forums (*news groups)*, sites et portails web d'une communauté scientifique (au sens large ou restreint). Leurs contenus sont de nature très différente : des informations circonstancielles (*call for papers*, annonces de colloques, …), des informations de type savoir-faire (“dépannage” technique, utilisation de tel ou tel produit, *etc*), groupe de travail (sur une norme, un projet, *etc*.)...

D'autres éléments peuvent contribuer à affiner cette typologie, notamment la distinction document primaire, secondaire et tertiaire ou encore information gratuite et payante.

Par définition, les articles et ouvrages scientifiques constituent les documents primaires d'une discipline, en génomique il faut aussi intégrer dans cet ensemble les documents des banques de données factuelles puisqu'ils sont la représentation, la description première d'un objet biologique.[23]. Les données bibliographiques issues des banques bibliographiques ou celles contenues dans les documents primaires des banques sont des documents secondaires qui font référence aux articles scientifiques[24]. Enfin, les documents générés à partir des systèmes d'information qui intègrent les documents primaires et secondaires des banques peuvent être considérés comme des documents tertiaires, ils réalisent une synthèse de ressources.

En génomique, les données factuelles sont généralement numérisées sous forme de notices dans des banques accessibles gratuitement par internet. Les documents secondaires sont eux aussi principalement sous forme électronique et gratuite: les références bibliographiques sont disponibles à partir de banques bibliographiques comme Medline (PubMed) accessibles gratuitement par le web. Seuls, les articles publiés dans les revues relèvent encore d'une économie marchande avec toutefois des pressions actuelles d'ouverture perçues notamment par le développement d'archives ouvertes donnant accès aux numéros anciens de ces revues.

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

## 2.2 - Le principe de double publication

La publication d'un article relatif à une séquence nucléotidique dans une revue scientifique est soumise à une obligation préalable : la publication de la séquence.
Plus précisément, cette publication est réalisée par le “dépôt” des données de représentation de la séquence dans des banques *ad hoc* [25]. En échange de ce dépôt le chercheur reçoit un numéro

---

23 Avec Suzanne Briet [*qu'est-ce que la documentation*, Paris, EDIT, 1951] la notion de “preuve” est au cœur de la définition du document. Ainsi les notices des banques sont des documents parce qu'elles sont la preuve (première) de l'existence de séquences nucléotidiques ou protéiques d'un organisme vivant.

24 DDBJ : DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/), EMBL :European Molecular Biology Laboratory (http://www.embl-heidelberg.de/), GenBank : Genetic sequences data Bank (http://www2.ncbi.nlm.nih.gov/Genbank/index.html), MIPS : Munich Information Center for Protein Sequences (http://www.mips.biochem.mpg.de/)

25 Les notices des banques de données factuelles. A la sortie des séquenceurs, l’information est numérisée et stockée sur un support magnétique (disquette ou disque dur). La séquence fait ensuite l’objet d’analyses informatiques 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 porte l’ADN, et résultat de leur expression, en un alphabet à vingt lettres (les acides aminés). 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 du 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

d'enregistrement[26] qui permettra la publication de son article dans une revue. Le terme proposé "double publication" exprime donc clairement le processus de publication en biologie moléculaire.

En 1984, dans les instructions aux auteurs du "*Journal of Biological Chemistry*", il fut demandé pour la première fois aux auteurs de publier leurs séquences directement dans une banque de données comme GenBank (américaine) ou EMBL Data Library (européenne). La même revue, un an plus tôt, informait les auteurs potentiels 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.

> Au regard du nombre croissant d'articles soumis qui décrivent des séquences nucléotditiques ... Les données obtenues par des techniques bien établies ... ne seront généralement pas publiée et ne devront plus être soumises avec le manuscrit.

En 1987, la revue "*Nucleic Acids Research*" tient les mêmes propos mais demande aux auteurs de publier leur séquence dans EMBL Data Library pour équilibrer la quantité de séquences avec GenBank. Surtout, 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", la séquence ne sera pas rendue publique tant que l'article n'aura pas été publié. En 1988, les banques de données, Genbank, EMBL Data Library mais aussi DDBJ, PIR, MIPS et JIPIDS s'échangent leurs données[27]

## 2.3 - Le débat de l'accès public/privé

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 Goutry (1989), dans une lettre à Nature évoque le fait que des années d'affinage sont souvent nécessaires avant qu'une structure soit intégrée dans une base de données. Elle s'inquiétait de la rumeur dans la communauté de la crystallographie, qui prétendait que les chercheurs vendaient les coordonnées à des compagnies privées plutôt que de les rendre publique et elle soutenait que les revues devaient jouer un rôle dans le dépôt rapide des coordonnées.[28]

Au-delà du questionnement sur la nature des supports ou encore le temps d'intégration des données dans les banques, ce qui est finalement mise en avant c'est le débat sur le financement de la recherche et la valeur économique des résultats.

La mise en place du principe de double publication qui concourt largement à l'évolution de la publication des données électroniques relève d'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 issus de fonds publics (*federal fund* aux USA) de soumettre leur séquence en même temps que leur article. Si cette obligation de double

---

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. Ainsi, 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.

26 *GenBank Accession Number*, par exemple.

27 DDBJ : DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/), EMBL :European Molecular Biology Laboratory (http://www.embl-heidelberg.de/), GenBank : Genetic sequences data Bank (http://www2.ncbi.nlm.nih.gov/Genbank/index.html), MIPS : Munich Information Center for Protein Sequences (http://www.mips.biochem.mpg.de/)

28 Weller A.C., *op. cit.* p.46

publication peut heurter individuellement, elle présente l'avantage de rendre publiques les données issues de fonds publics. Les entreprises privées qui produisent des séquences génétiques ne sont, quant à elles, pas soumises à cette obligation mais peuvent bénéficier des séquences publiques.
On peut alors, légitimement, s'interroger sur les effets de cette politique de publication. En effet, mettre les résultats de la recherche dans le domaine public c'est aussi donner gratuitement aux multinationales privées des informations qu'elles n'hésiteront peut-être pas à transformer en innovations brevetées. Les organismes scientifiques comme l'INRA, pour contrebalancer cette situation et favoriser la valorisation des résultats de la recherche publique, sont favorables, sous certaines conditions, à leur protection par brevets :

> "[...] la détention de tel brevet par des organismes publics doit permettre à la France et à l'Europe de tenir leur place -aux plans scientifique et économique- dans le concert des Nations et favoriser, par le biais de l'octroi de licences, l'accès à nos filières aux innovations concernées."[29]

Mais cette situation n'est souvent pas aussi contrastée et l'analyse ne peut pas s'arrêter à ce niveau. D'une part, certaines entreprises privées rendent publiques leurs données expérimentales après un délai de quelques mois et d'autre part, les programmes de recherche en génomique se réalisent généralement dans le cadre d'un financement conjoint entre le public et le privé. Sans investissement privé, de nombreuses recherches ne pourraient aboutir; fréquemment, la répartition public/privé se schématise ainsi : le public propose ses "cerveaux" et des locaux, le privé contribue à la logistique (matériel, frais de fonctionnement, CDD, …). Donc, il faut comprendre les exigences de retour sur investissement imposées par le secteur privé, traduites notamment ici par cette réserve de publication afin d'exploiter au mieux en interne les résultats obtenus.

Il n'en reste pas moins que cette mixité des fonds de financement d'un programme d'analyse de génome pose de réelles interrogations aux chercheurs quand il appartient au secteur public et désire publier ses résultats de recherche pour se valoriser. Selon les règles établies, il ne peut communiquer qu'en "double-publiant" une séquence et l'article afférent. Mais l'association public/privé stipule souvent que les données sont la propriété du privé avec généralement une clause de "temps de latence" qui donne la primeur des séquences et des résultats pendant quelques mois (souvent 6 mois) à(ux) l'entreprise(s) qui finance(nt) le projet. Ainsi le chercheur ne pourra partager ses résultats avec sa communauté scientifique que quelques mois plus tard. Quelques mois qui peuvent rendre obsolètes (pour la communauté scientifique) les connaissances construites, ainsi que le contenu de l'article.

## 2.4 - De la qualité des données

La gratuité des ressources informationnelles au travers des banques de séquences, des banques de documents secondaires et l'apparition du texte intégral (ex. Pubmed Central) confère à la génomique un statut particulier en terme de partage d'information par hyperliens.
Néanmoins la qualité des données des banques de séquences n'est pas sans reproche, les erreurs se propagent de collections primaires (banques internationale comme Genbank, EMBL) en collections secondaires (base de données spécifiques, sur un animal, une bactérie...) et rejaillissent par liens sur l'ensemble des autres entrepôts de documents. Elles doivent être contrôlées et nettoyées par des traitements de base, il faut détecter les erreurs et éliminer la redondance, pour donner la meilleure précision possible aux calculs. Les sources d'information et leurs mises à jour doivent être

---

29 Hervieu B., Guillou M., "la brevetabilité du vivant en débat" INRA Mensuel, n°110, mai-juillet 2001, p.20.

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[30] 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 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 révélateur de l'état d'avancement des collaborations dans la communauté des génomistes.[31]

# III - L'appropriation de l'information numérique par les chercheurs

Les résultats issus de l'analyse effectuée au sein d'une communauté de microbiologistes révèlent deux types d'appropriation de l'information numérique, découlant de deux activités, de deux types d'appropriation des technologies de l'information et de la communication (TIC)

- une appropriation profonde, une assimilation de l'informatique par les chercheurs en génomique (les génomistes) qui établit une symbiose entre la biologie et l'informatique. Cette symbiose donne lieu à une nouvelle discipline : la bioinformatique (*cf.* supra) ;
- une appropriation “superficielle” cantonnée au seul usage des outils de communication électronique, de bureautique et des produits de la bioinformatique.

Ainsi, coexistent d'une part les bioinformaticiens qui font de la recherche et du développement sur les représentations des objets biologiques et mettent en place des artefacts informationnels ; d'autre part les biologistes qui utilisent des outils informatiques standards commercialisés ou distribués librement et les artefacts informationnels mis en place par les bioinformaticiens. On peut en effet distinguer un groupe qui manipule les outils informatiques de façon professionnelle et crée des produits, et un autre groupe constitué de simples utilisateurs.

Aussi, dans un premier temps nous scindons les dispositifs informationnels communs à l'ensembles des biologistes et ceux construits par les bioinformaticiens, pour mieux analyser ensuite l'ensemble de l'activité scientifique des génomistes.

## 3.1 - Le cycle de l'information scientifique et technique et les dispositifs informationnels

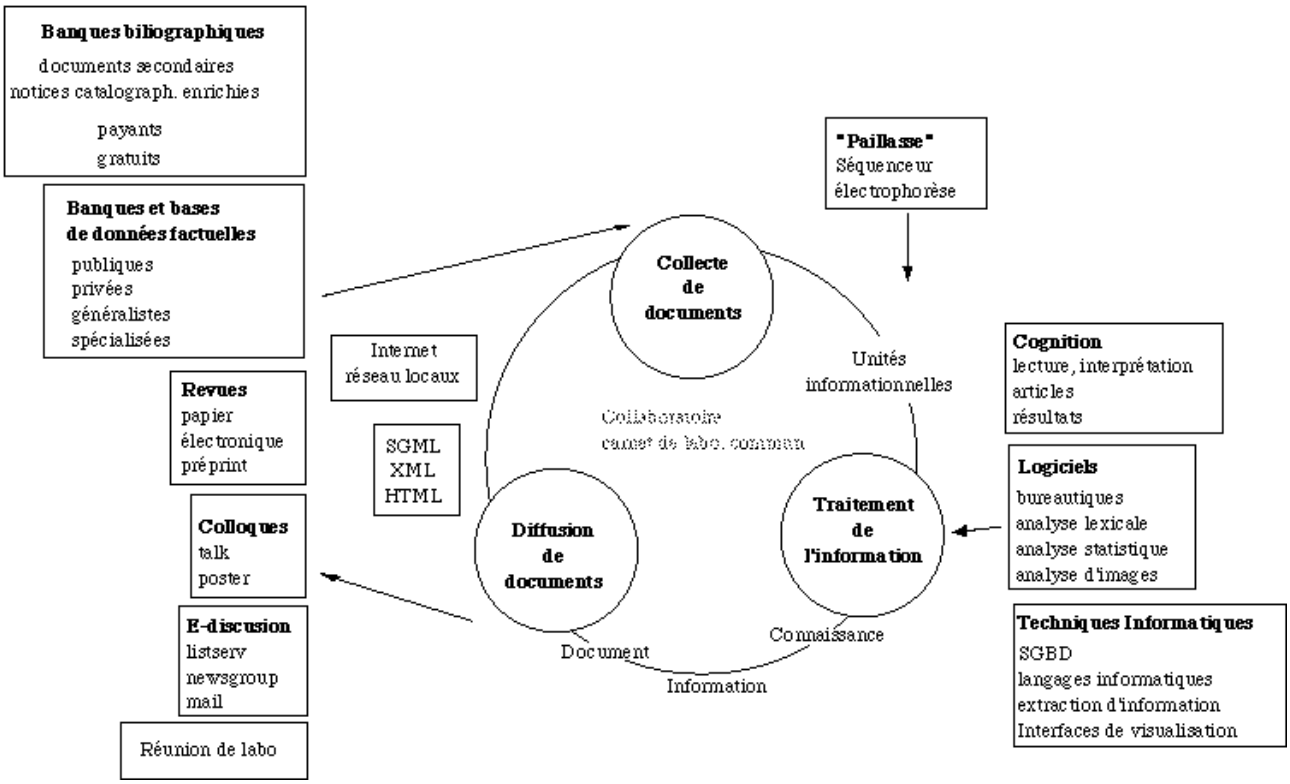
---

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

31 Gallezot G., Sansom F., Brunaud V., Gas S., Bessière P., “Normes et standards dans le processus de traitement du document numérique en biologie moléculaire ”, *Solaris*, nº6, 2000. [En ligne]. http://www.info.unicaen.fr/bnum/jelec/Solaris/d06/6gallezo.html.

La figure 4 détaille les dispositifs informationnels utilisés par les chercheurs en insistant, pour la collecte, le traitement et la diffusion de l'information scientifique et technique (IST), sur les pratiques informatives qui font intervenir des technologies de l'information et de la communication.

*Figure 4 : le cycle de l'IST et les dispositifs informationnels*




*Légende figure 4 : La partie gauche du schéma représente les sources et les entrepôts qui alimentent ou sont alimentés par le processus informationnel scientifique représenté par la collecte, le traitement et la diffusion de l'information scientifique et technique essentiellement médiatisés par des réseaux et des formats de données standardisés très utilisés. Les données issues de la paillasse sont volontairement distinguées de cette dernière partie pour signifier leur statut primaire et endogène aux laboratoires. La partie droite du schéma présente les dispositifs de traitement de l'information en distinguant ce qui relève du seul acte intellectuel du chercheur, puis des logiciels et des techniques qui l'aident dans cette action.*

Au delà de l'appropriation des outils de bureautique et de messagerie électronique, finalement commune à la plupart des activités professionnelles, la marque de l'influence des technologies de l'information et de la communication sur les pratiques informatives des chercheurs en génomique réside essentiellement dans le processus de traitement des données factuelles : les résultats issus de l'expérimentation (séquençage) sont sous forme numérique (stockés sur disque dur), puis ils font l'objet de traitements et d'analyses à l'aide de logiciels spécifiques en vue de leur annotation. Les

résultats annotés sont publiés (dépôts) dans les banques de données génomiques. Ces dépôts se réalisent par courrier électronique ou par une interface de saisie sur le site web des banques génomiques. Les banques génomiques diffusent l'ensemble des résultats collectés sous forme de notices standardisées *via* un serveur web (une ou quelques de notices à la fois) ou un serveur FTP (un ensemble de notices concernant un type d'organisme, les bactéries par exemple). Les collections de notices servent l'analyse de nouveaux résultats expérimentaux à travers leur traitement par des logiciels *ad hoc*. Les collections de notices peuvent être aussi intégrées dans des systèmes d'informations (*cf.* infra, cycle de l'information scientifique et technique en génomique), permettant ainsi une autre analyse, une autre présentation, et une autre "lecture" des résultats expérimentaux.

Pour la littérature en génomique l'influence des technologies de l'information et de la communication est marquée par les références bibliographiques (enrichies des *abtracts*) disponibles (en partie) sur le web (*cf.* Pubmed), par le texte intégral des articles souvent accessible en ligne sur le site des éditeurs et par de nombreux "portails" qui proposent des info-services (cours, forum, présentations de techniques...). Néanmoins, les revues "papier" existent toujours et continuent d'être une source prépondérante de collecte de la littérature.

Comment expliquer l'importance de l'influence des technologies de l'information et de la communication sur les données factuelles par rapport à la littérature scientifique ? Une première explication a trait à la gratuité des données factuelles. Une deuxième explication est relative aux entrepôts de données centralisés (et internationaux) que sont les banques génomiques. Cette situation a favorisé l'informatisation et l'échange des données factuelles, l'automatisation des tâches diminuant leur coût. Celle-ci a aussi appelé la réalisation de projets bioinformatiques, qu'il s'agisse de création de programmes informatiques ou de systèmes d'information. *A contrario,* la non gratuité du texte intégral et la multiplicité des revues[32] ne permettent pas encore de grands développements de traitement de l'information[33]. Néanmoins, la disponibilité des notices Pubmed au format XML ou les réflexions entamées à travers des projets comme Open Archive Initiative (OAI), CrossRef[34] présente des perspectives intéressantes en matière de gestion des connaissances.

Si cette analyse met en lumière l'aide des dispositifs informationnels apportée à l'activité cognitive du chercheur en génomique, elle laisse de côté l'autre activité de ces derniers qui consiste justement à construire ces dispositifs en usant des technologies de l'information et de la communication.

## 3.2 - Le cycle de l'information scientifique et technique médiatisé par un artefact informationnel

A travers l'exemple d'un système d'information nommé MICADO[35], nous avons analysé l'appropriation de l'information numérique par des bioinformaticiens (*cf.* figure 5)

---

32 Multiplicité des supports, multiplicité des politiques d'accès aux contenus, multiplicité des formats.

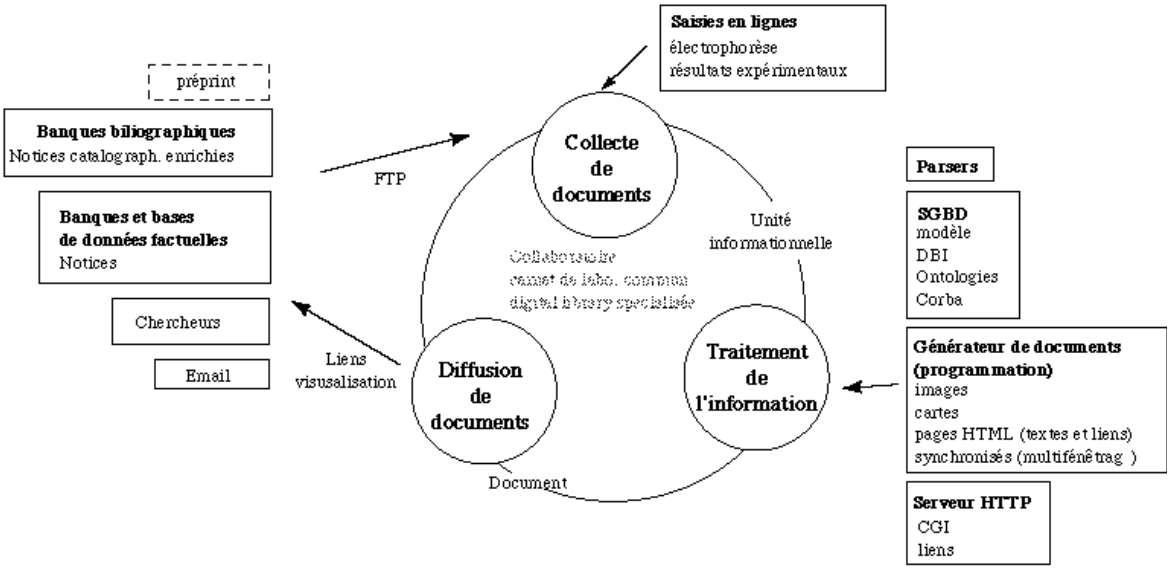
33 Toutefois les projets Pubmed Central ou E-biosci, permettront peut-être ce type de projets. Gallezot G., *op. cit.*, pp137-144. L'Open Archive Initiative participe à cette évolution avec notamment la mise en place de la "Santa Fe convention", http://www.openarchives.org.

34 Buttler D., "The future of electronic scientific literature", *Nature* 413, 2001, p.1-3

35 MICrobial Advanced Database Organisation, système d'information maintenu par l'équipe MIG (Mathématique, Informatique et Génome) de l'INRA

L'appellation *"*base de donnée intégrative" attribuée à ce système d'information par les bioinformaticiens dans les premiers temps, résume bien le processus d'intégration physique des données en un seul site. Elle dénote la volonté de collecter l'information scientifique et technique présente dans de nombreuses banques de données sur Internet. Elle reflète l'idée de posséder de manière exhaustive l'information relative à des organismes spécifiques (ex : la bactérie *Bacillus Subtilis*) et de l'organiser pour faciliter le traitement de cette somme d'informations. Mais cette appellation ne prend pas en considération l'ensemble du processus de collecte de traitement et de diffusion de l'information scientifique et technique. En effet l'organisation de l'information, sa manipulation, puis sa restitution sous forme de documents disponibles à travers une interface web relèvent d'une suite de processus techniques qui confèrent à cette "base de données intégrative" le statut de système d'information[36]. Les entrées sont caractérisées par les notices des banques et les sorties par des documents hypermédia. La combinatoire des techniques composant ce système[37] opère donc une médiation informationnelle. Elle permet de transmettre l'information des producteurs vers les usagers. Mais, le seul fait de transmettre de l'information ne suffit pas à parler de "médiation". Il doit y avoir une action de "négociation". Ce système "négocie" entre l'information (brute) des producteurs et la demande d'information des chercheurs, en transformant des données génomiques en documents "prêts à lire". Il régule ainsi le flux informationnel des données génomiques. A l'instar de l'homéostasie des organismes qu'ils étudient, ce système permet aux chercheurs de maintenir dans un état de stabilité leurs connaissances. Si ce système d'information manipule essentiellement les données factuelles, des travaux concernant l'exploitation de la littérature (données textuelles) sont en cours. Déjà commencé avec le traitement de co-citations d'objets biologiques et la création d'outils sémantiques qui s'attachent à détecter leurs relations dans des résumés, l'exploitation automatisée du texte intégral viendra, sous peu, compléter certainement le dispositif informationnel des chercheurs.

*Figure 5: l'activité scientifique médiatisée par un système d'information*




36 Chaumier J., *Systèmes d'information*, Paris, Les Editions ESF, 1986. D'autres qualificatifs peuvent être proposés pour compléter l'analyse de Micado, notamment ceux de *Collaboratoire* et *Bibliothèque numérique [Cf.* Gallezot G., *op. cit.*, pp145-154]

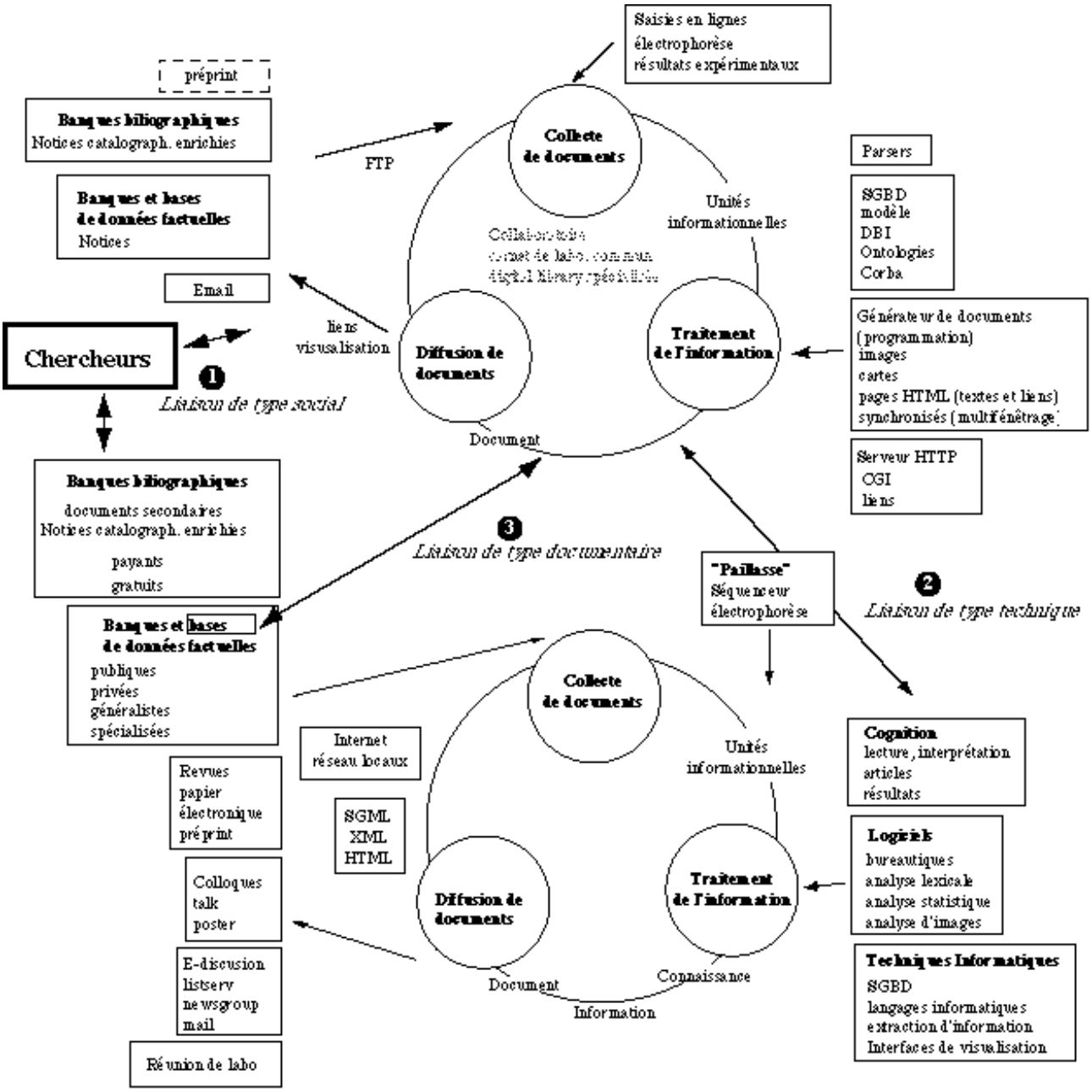
37 Gallezot G., *op. cit.*, pp101-119

*Légende figure 5 : ce schéma présente un processus informationnel guidé par des technologies de l'information et de la communication. des chercheurs à l'aide d'outils de traitement disponibles (cf. figure 4) construisent un artefact qui médiatise en partie leur activité scientifique. Ainsi la partie gauche représente les entrées et les sorties de l'artefact, la partie droite le dispositif technique nécessaire à son fonctionnement et à son développement.*

Ce système d'informations globalise au mieux l'ensemble de l'information d'un domaine et fait l'objet de développements techniques incessants. Avec d'autres systèmes similaires, il est le reflet de l'influence des technologies de l'information et de la communication sur l'activité des chercheurs en génomique et contribue aux changements de leurs pratiques informatives. Il est un instrument évolutif et polymorphe. Evolutif parce que l'organisation *in silico* du vivant peut être adaptée à un grand nombre d'organismes. Polymorphe, parce que son agencement peut agréger de nouvelles techniques ou de nouvelles applications qui permettront de nouvelles collectes, de nouveaux traitements, d'autres interfaces de visualisation et d'autres documents à diffuser. C'est un système modulaire et modulable qui peut être valorisé dans d'autres activités scientifiques

## 3.3 - Le cycle de l'information scientifique et technique en génomique

***Figure 6 : Le cycle de l'****information scientifique et technique* ***en génomique***




*<u>Légende figure 6</u> : Ce schéma à pour objectif de globaliser l'activité scientifique des génomistes à travers des dispositifs informationnels génériques (partie inférieure) et médiatisé par des artefacts informationnels dédiés (partie supérieure).*

L'adjonction des figures 4 et 5 exposées précédemment, permet de présenter l'ensemble des actants qui interviennent dans le processus de construction des connaissances en génomique. Les chercheurs sont au coeur de ce schéma puisqu'il s'agit de leur activité scientifique. La mise en évidence d'une "liaison de type social" [**1**] permet d'insister sur ce fait, le(s) chercheur(s) doit(vent) disposer et se servir des ressources informationnelles et des techniques disponibles (partie inférieure du schéma) pour construire des artefacts informationnels (partie supérieure du schéma) capables de les aider dans leurs pratiques informatives. Dans cette lecture nous retrouvons les deux catégories de chercheurs évoquées plus haut, ceux qui usent des outils construits pour eux et ceux qui usent des outils construits par eux. Dans les laboratoires où ces deux catégories de chercheurs se retrouvent cette distinction ne tient pas. De leur interaction continuelle émerge une nouvelle catégorie : "les usagers-constructeurs".

La mise en évidence d'une "liaison de type technique" [**2**] souligne le fait que la disponibilité et la maîtrise de techniques d'information permettent une technologie qui débouche sur la construction d'artefacts informationnels *ad hoc.* La mise en évidence d'une "liaison de type documentaire" [**3**] indique que la réalisation d'artefacts informationnels *ad hoc*, par et pour un groupe de chercheurs, permet aussi la mise à disposition d'un ensemble de documents pour une communauté scientifique plus large.

# IV - Illustration de l'influence des dispositifs informationnels sur la production des connaissances

La productivité mesure le rapport la quantité de biens produits aux facteurs nécessaires pour cette production, c'est-à-dire ici le rapport de la quantité d'information produite aux facteurs techniques d'information et de temps[38]

Ainsi la productivité dépend directement de la façon dont se déroule le cycle de l'information scientifique et technique en génomique : de la façon dont les documents sont collectés, de la manière dont est extraite l'information, des moyens de traitement de cette information, de la capacité à produire des connaissances et de les "traduire" en information, des possibilités d'édition et des canaux de communication utilisés. La corrélation forte entre technologies de l'information et de la communication et production de connaissance est mise en évidence par les courbes présentées ci-dessous. Nous pensons toutefois que ce n'est pas le seul facteur explicatif. D'autres facteurs (budgets des laboratoires, budgets de la recherche, politiques de recherche, nombre de chercheurs travaillant sur le sujet...) devraient également être mis en correspondance.

Pour vérifier cette corrélation, une étude quantitative des données factuelles et bibliographiques contenues dans Micado (*cf.* supra) nous a permis de mesurer la productivité des chercheurs. Elle a permis, de suivre pour une même communauté scientifique, les chercheurs qui travaillent sur *Bacillus Subtilis,* l'évolution des productions d'information scientifique et technique dans le temps (hors conférences, posters, etc.). Cette évolution a été couplée avec l'évolution diachronique des techniques présentées (*cf. Part I, importance des techniques*). On met ainsi en lumière différentes périodes qui rendent compte de la relation entre l'introduction de techniques / la généralisation des techniques / le volume de production d'information scientifique et technique.

Le génome de *Bacillus subtilis* représente 4100 gènes pour quelques 4.2 millions de paires de bases. Il a été séquencé par 35 équipes (28 européennes et 7 japonaises) en 7 ans[39]. Le séquençage a démarré en 1990 et s'est terminé en juillet 1997 dans les laboratoires, pour être publié dans les banques de séquences et faire l'objet d'un article en novembre 1997[40] (indexation intégrée dans

---

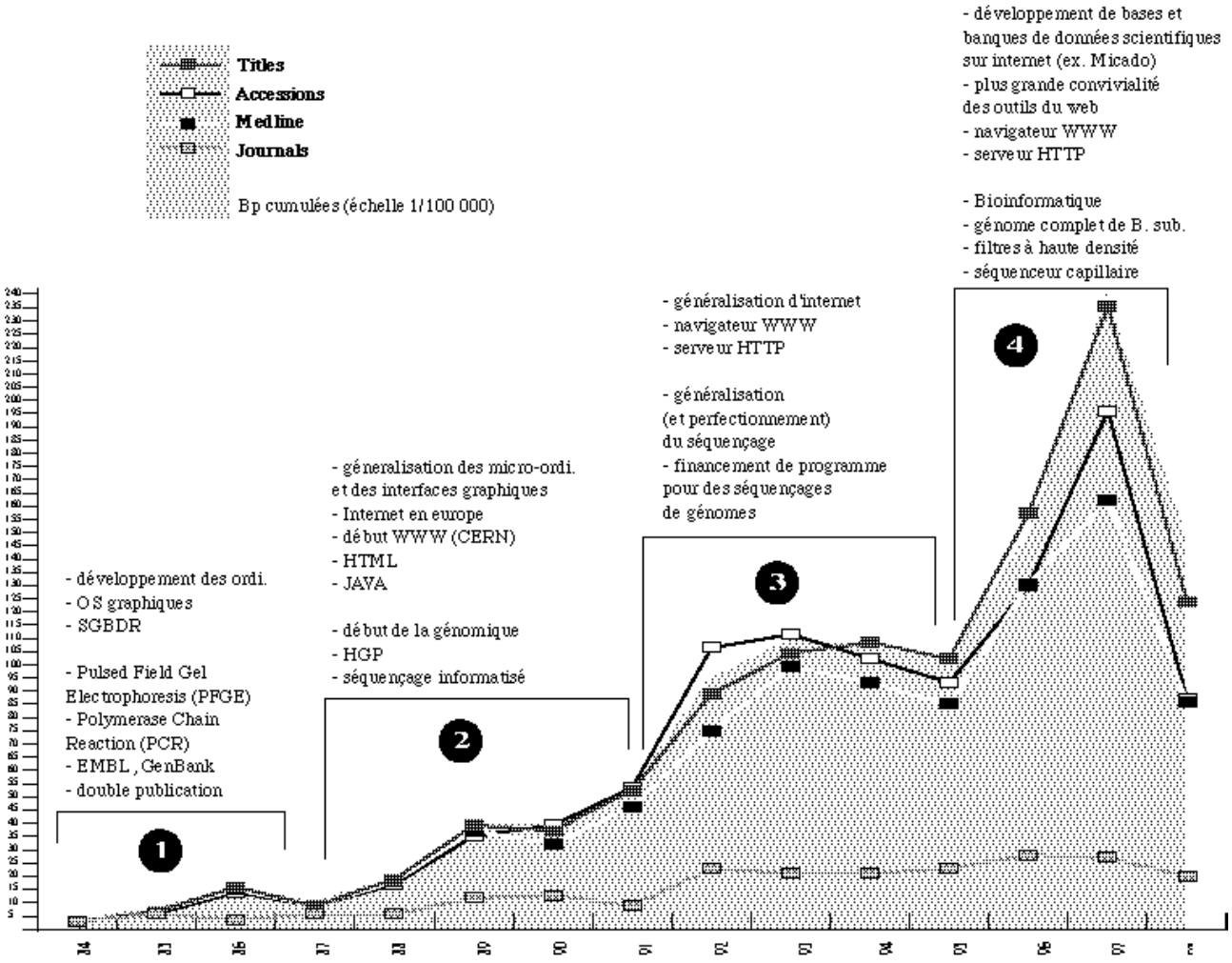
38 Nous rappelons ici que nous définissons l'information comme la mise en forme des connaissances (in-formation). La description d'une séquence génétique est un niveau de connaissance sur un génome. Sa mise en forme à travers une notice d'une banque constitue une information. Ainsi en quantifiant les informations des notices des banques nous quantifions un niveau connaissance (la composition d'un génome)..

39 Danchin A. et al., " *Bacillus subtilis* dévoile ses gènes ", *Biofutur*, n°174, janvier 1998, pp.14-17.

40 Kunst F. et al., " The Complete Genome Sequence of the Gram-positive Bacterium *Bacillus subtilis*", *Nature*, n°390, 1997, p. 249-256.

MedLine en décembre 1997). Aujourd'hui les séquenceurs permettent un séquençage complet d'un génome de même importance en moins de deux ans par une seule équipe. Le temps de recherche s'est considérablement raccourci, la publication des données factuelles est quasi immédiate. Seule la publication papier d'un article présente encore un délai qui varie entre 6 mois et un an. Pour comprendre cette évolution, la figure ci-dessous[41] montre l'évolution de la production scientifique liée aux séquences de *Bacillus subtilis* et les dispositifs informationnels prépondérants qui l'ont influencé.

*figure 7* :Évolution de la production scientifique et des dispositifs informationnels.
*Le cas de Bacillus Subtilis*




Les courbes présentent, année par année, la somme de données factuelles et des publications relatives aux séquences de *Bacillus subtilis*. Bien que des données sur d'autres organismes bactériens soient disponibles dans Micado, le choix de cette bactérie s'est imposé, d'une part, parce que le laboratoire étudié travaille plus particulièrement sur cet organisme, d'autre part parce que son génome fut l'un des premiers à faire l'objet d'un programme de séquençage suivi et réunissant des laboratoires européens et japonais, et enfin parce qu'elle fut l'une des premières bactéries à révéler son génome.

41 Le graphique est réalisé à partir d'extraction de données sur Micado (mise à jour du 1er mars 1999).

“Titles” représente le nombre d’articles publiés, “Accessions” le nombre de séquences (de notices), “Medline” le nombre d’articles référencés dans cette banque, “Journals” le nombre de revues dans lesquelles les articles sont publiés et “Bp cumulées” le nombre de paires de bases cumulées.

Les périodes indiquées en **[1]**, **[2]**, **[3]** et **[4]** ont été déterminées par paliers d’accroissement des données. En effet, pour déterminer l’influence des technologies de l'information et de la communication une étude année par année n’est pas signifiante, il faut prendre en compte l’émergence d’une technique, ses premières appropriations par la communauté étudiée, puis sa généralisation. L’échelle débute en 1984 avec les premières séquences (3 notices) enregistrées dans GenBank et se termine en 1998 (86 notices) afin d’avoir les données complètes pour chaque année (l’année 1999 est donc exclue).

Les dispositifs informationnels indiqués au-dessus des périodes portent sur les périodes de synchronie, entre les événements informatiques et les événements en biologie moléculaire, intitulées “adolescence” et “une certaine maturité”.

On remarque en premier lieu l’évolution concomitante de l’ensemble des données. Ce fait s’explique par le principe de double publication qui impose aux chercheurs la publication de la séquence dans les banques de séquences pour pouvoir publier un article sur une séquence. Le pic de 1997 est relatif aux séquençages complet de *Bacillus subtilis*, la chute de 1998 s’explique par ce même phénomène. D’ailleurs on peut supposer que les séquences encore publiées en 1998 sont relatives à une autre souche de *Bacillus subtilis* ou à des séquençages de segments spécifiques de la molécule d’ADN.

La production de données factuelles ou non factuelles (publications) est en constante progression sauf pour les années 1987 et 1994-1995 qui représentent des périodes de faible productivité. Les trois sauts quantitatifs en 1989, 1992 et 1996 représentent l’appropriation de nouvelles techniques et certainement l’apport de moyens financiers non étudiés en détail dans ce travail.

L’accroissement des données entre 1984 et 1986 **[1]** marque le début du séquençage de masse avec les techniques de PFGE et de PCR et l’apparition des premiers entrepôts qui servent à l’enregistrement des résultats issus du séquençage : les banques de séquences.

La relative chute de production de données en 1987 peut s’expliquer par l’interrogation sur l’utilité du séquençage et sur les moyens techniques, financiers et humains à mettre en œuvre pour décrypter des génomes entiers.

Le saut quantitatif de 1989 est une réponse à cette interrogation. Le lancement du *Human Genome Project* a affirmé l’utilité de décrypter l’ADN du vivant et a impulsé d’autres programmes de séquençage. La volonté d’organismes essentiellement publics de financer ce type de programme a eu un effet d’électrochoc sur les techniques *ad hoc*. Les séquenceurs informatisés ont fait leur apparition permettant d’accroître le nombre de données factuelles pour un nombre quasi identique de publications. La période **[2]** voit la naissance du programme de séquençage de *Bacillus subtilis*, la forte évolution des données en 1992 illustre ce fait, avec près de 130 000 paires de bases[42] en plus entre 1989[43] et 1992.

Pendant la période **[3]** on distingue une différence entre les données factuelles et les publications. Les séquences sont plus nombreuses que les articles en 1992 et les articles sont plus nombreux que

---

42 Somme des paires de bases de 1990 et 1991

43 Lancement du HGP

les séquences en 1995. Ce phénomène s'explique par la production massive de séquences en début de période[44], pour laquelle un certain nombre n'ont pas fait l'objet d'une publication (*direct submission*), ou qu'une publication faisait référence à plusieurs séquences à la fois. Les raisons que l'on peut invoquer sont liées aux techniques de traitement de représentation numérique des séquences. Les outils d'analyse de ces documents n'étaient pas généralisés ou peu utilisés en 1992. La généralisation d'Internet, du web et des micro-ordinateurs dans les laboratoires a permis une diffusion plus large de ces outils, voir leur interfaçage sur des serveurs web. Un autre point à relever en 1992 est la différence entre le nombre d'articles et le nombre de notices MedLine. Une première explication se trouve dans l'accroissement du nombre de revues[45] : des articles faisaient l'objet d'une publication dans de nouvelles revues ou des revues considérées comme mineures, non référencées par MedLine. Une seconde explication peut être avancée : la publication de texte dans des ouvrages ou actes de colloques.

Enfin, pendant la période **[4]**, le saut quantitatif en 1996 (séquences et publications) qui se poursuit jusqu'en 1997 (date du séquençage complet du génome) est lié à une appropriation affirmée des outils informatiques et du perfectionnement des outils pour l'analyse de l'ADN. Le décalage entre le nombre de publication et le nombre de séquences (débuté en 1994) montre que la production de données factuelles accumulées et gérées par des technologies de l'information et de la communication permet une production accrue des connaissances et marque l'importance de la bioinformatique.

La croissance de la productivité des périodes **[1]** et **[2]** résulte des techniques de biologie, des impulsions politiques et financières de programmes d'analyse des génomes, et de la mise en oeuvre du principe de double publication. La croissance de la productivité de la période **[3]** s'explique en grande partie par l'introduction de l'informatique dans les techniques en biologie et dans les laboratoires, permettant ainsi d'accélérer la production de données factuelles, mais aussi, du démarrage du programme d'analyse du génome de *Bacillus subtilis.* La croissance de la productivité de la période **[4]** s'explique par une utilisation accrue des outils informatiques permettant ainsi d'automatiser une partie de l'analyse du vivant, de favoriser les échanges entre chercheurs et d'opérer de nouveau traitement sur l'information scientifique et technique en génomique. Ainsi, les données factuelles sont produites plus rapidement, les documents circulent plus vite et la construction des connaissances s'accélère.

Les années de ralentissement de la productivité (1987 et 1994-1995 pour les données factuelles) résultent de phénomènes plus ponctuels, mais tout aussi conjoncturels. Ce sont pour 1987 des problèmes essentiellement liés aux financements de projets de séquençage et pour les années 1994-1995 vraisemblablement des problèmes liés à la rapidité de l'annotation des séquences.

Ainsi nous avons apporté quelques éléments de réponses concernant les corrélations entre les technologies de l'information et de la communication et la production de connaissances en génomique. Néanmoins ces corrélations ne doivent pas être détachées de leur contexte : des orientations politiques de recherche fortes autour des génomes, donnant lieu à des budgets de recherche conséquents.

---

44 Le décalage entre le nombre de paires de base cumulées et le nombre de notice de banques indique que les notices présentaient des séquences composées de peu de paires de bases.

45 9 revues en 1991 contre 23 en 1992

# Conclusion

Le travail scientifique en génomique et tout le processus éditorial des résultats de la recherche ont largement été reconfigurés par les technologies de l'information et de la communication.
Avec la production en masse de séquence d'ADN, le principe de double publication s'est imposé, indiquant la voie d'une publication gratuite des données factuelles. Néanmoins, toutes les données de l'analyse des génomes ne sont pas obligatoirement enregistrées dans les banques publiques (Genbank, EMBL, DDBJ...). Les résultats de recherches financées par des fonds privés peuvent échapper à cette règle.
La situation des références bibliographiques est plus simple: Medline qui indexe un nombre conséquent de revues[46], est consultable gratuitement sur le web (Pubmed) et satisfait largement les chercheurs. Il existe des liens croisés entre les notices catalographiques et les enregistrements des données factuelles. Des systèmes d'information dédiés (par exemple, Micado) ou plus généralistes (par exemple, Entrez) mettent en évidence ces liens.
Finalement, seuls les articles scientifiques relèvent encore essentiellement du secteur marchand. Pour quelles raisons ? Quelques éléments de réponses concernent certainement l'organisation de la fonction d'évaluation scientifique, les habitudes de lecture... Mais au regard de la pensée non-marchande qui s'est développée dans les deux précédents types de publication, il est probable que des projets comme Pubmed Central impliqueront certaines reconfigurations, notamment parce que l'analyse des données dites “textuelles” (les savoirs enregistrés dans les articles scientifiques) est une prochaine étape de la recherche *in silico*[47].

---

46 Le 22 mai 2001 Medline comptait quelques 4300 indexées. Source : http://www.nlm.nih.gov/pubs/factsheets/medline.html

47 En effet, les connaissances biologiques (les interactions géniques par exemple) ne sont pas uniquement décrites dans les notices des banques mais aussi dans les articles scientifiques.

# Références :

Benton D., " Bioinformatics -Principles and Potential of a New Multidiciplinary Tool ", *Trends in Biotechnology*, vol. 14, n° 8, 1996, p. 261-272.

Biaudet V., Samson F., Bessières P., "Micado: a Network Oriented Database for Microbial Genomes", *Computer Application in Biosciences*, vol. 13, n° 4, 1997, p. 431-438.

Blanquet M.F., S*cience de l'information et philosophie*, Paris, ADBS éditions, 1997.

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

Briet S. *Qu'est-ce que la documentation*. Paris : Editions Documentaires Industrielles et Techniques (EDIT), 1951.

Bucland M. K., "Information As Thing", *JASIS*, Vol. 42, n°5, 1991, p. 351-360.

Bucland M. K., "What Is a Document ?", *JASIS*, Vol. 48, n°9, 1997, p. 804-809.

Buttler D., "The future of electronic scientific literature", *Nature* 413, 2001, p.1-3

Buttler D., " US Biologists Propose Launch of Electronic Preprint Archive ", *Nature*, Vol. 397, jan. 1999, p.91.

Cacaly S. (sous la dir.), *Dictionnaire encyclopédique de l'information et de la documenation*, Paris, Nathan Université, 1997.

Cockerill M., " A Versatile Tool for Retrieving Molecular Sequences: Entrez ", *Trends in Biochemical Sciences*, vol 9, nº2, 1994, p. 94-96.

D.O.E., *To Know Ourselves*. [en ligne], [consultée en 02/98] URL http://www.ornl.gov/TechResources/Human_Genome/tko/

Danchin A. et al., " *Bacillus subtilis* dévoile ses gènes ", *Biofutur*, n°174, janvier 1998, pp.14-17.

Dujon B., " The Yeast Genome Project: What Did we Learn? ", *Trends in Genetics*, vol. 12, n° 7, 1996, p. 263-270.

Flichy P., *L'innovation technique : récents développements en sciences sociales. Vers une théorie de l'innovation*. Paris, Ed. la découverte, 1995.

Gallezot G., Sansom F., Brunaud V., Gas S., Bessière P., " Normes et standards dans le processus de traitement du document numérique en biologie moléculaire ", *Solaris*, nº6, 2000. [En ligne]. http://www.info.unicaen.fr/bnum/jelec/Solaris/d06/6gallezo.html.

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

Gibbons et al. *The New Production of Knowledge. The dynamic of Science and Research in Contemporary Societies*. London, Sage publications, 1994.

Guinchat C. Skouri Y., *Guide pratique des techniques documentaires*, Vanves, Edicef, 1996, vol.2.

Hieter P., Boguski M., " Functional genomics : it's all how you read it ", *Science*, Vol.278, 1997, p. 601-602.

Jordan B.,*Voyage autour du génome : le tour du monde en 80 labos*, Paris, INSERM John Libbey Eurotext, 1993.

Kunst F. et al., " The Complete Genome Sequence of the Gram-positive Bacterium *Bacillus subtilis*", *Nature*, n°390, 1997, p. 249-256.

Le Coadic Y.-F., *La science de l'information*, Paris, PUF, Coll. Que sais-je, 1994.

Letovsky S.," Beyond the information maze ", *J Comput Biol,* Vol.2, n°4, 1995 Winter, p. 539-546.

Lomme L., " L'information électronique en biologie moléculaire ", *Documentaliste-science de l'information*, Vol.35, n°3, 1998, p. 179-185.

Meadows A. J., *Communicating Research*, Academic Press, 1998.

Morange M., "Bréve histoire de la biologie moléculaire", *Biofutur*, 1995, n°142, p.15-18

Nedellec C., Ould Abdel Vetah M., Bessières P., " Sentence filtering for information extraction in genomics, a classification problem ", *PKDD'01 - Proceedings of the Conference on Practical Knowledge Discovery in Databases*, Springer Verlag, Freiburg, September 2001.

Schazt B.R., Chen H., "Digital Librairie: Technological Advances and Social Impacts",*Computer* , fev, 1999, pp. 45-50.

Swynghedauw B., *La biologie moléculaire*. Paris, Nathan-Université, 1994, coll. 128 – Sciences.

Varmus H. [Consulté en mai 1999]. *E-BIOMED : A Proposal for Electronic Publication in the Biomedical Sciences.* [En ligne] http://www.nih.gov/welcome/director/ebiomed/ebi.htm
Weller A. C. "The Human Genome Project" In : Crawford S.Y., Hurd J.M., Weller A.C., *From Print to Electronic, The Transformation of Scientific Communication*, Medford, Information today Inc (Assis monograph series), Washington, USA, 1996, p. 46-73.