|  |
| --- |
| ***Cours de Bioinformatique  (Les banques de séquences biologiques)*** |

**'**[**introduction générale du cours**](http://www.dsi.univ-paris5.fr/bio2/biocours/intro.html#INTRO)

I. [**LES BANQUES DE SEQUENCES BIOLOGIQUES**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#INTRO)

1. [**HISTORIQUE**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#HIST)
2. [**LES BANQUES GENERALISTES**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#BQGE)

2.1 [**La qualité des données**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#QUAL)

1. [**LES BANQUES SPECIALISEES**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#BQSP)

3.1 [**Les bases de motifs**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#MOTI)

3.1.1 [**Les bases de motifs nucléiques**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#MONU)

3.1.2 [**Les bases de motifs protéiques**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#MOPR)

1. [**LA DIFFUSION ET L'UTILISATION DES BANQUES DE DONNEES**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#UTIL)

4.1 [**La diffusion**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#DIFF)

4.2 [**L'interrogation**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#INTE)

4.2.1 [**Les systèmes dédiés**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#DEDI)

4.2.2 [**Les systèmes de Gestion de Bases de Données**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#SGBD)

1. [**CONCLUSION**](http://www.dsi.univ-paris5.fr/bio2/biocours/chapitreI.html#CONC)

**LES BANQUES DE SEQUENCES BIOLOGIQUES** 

           Il existe un grand nombre de bases de données d'intérêt biologique. Nous nous limiterons ici à une présentation des principales banques de données publiques, basées sur la structure primaire des séquences, qui sont largement utilisées dans l'analyse informatique des séquences. Nous distinguerons deux types de banques, celles qui correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations et celles qui correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt suscité par un groupe d'individus. En biologie, il est fréquent d'appeler les premières "banques de données" et les secondes "bases de données", mais cette distinction n'est pas universelle en dehors du domaine biologique. Aussi, pour éviter toute confusion sémantique nous parlerons ici de banques de données ou bases de données *généralistes* (pour les premières) et *spécialisées* (pour les secondes).

**1. HISTORIQUE**

           C'est au début des années 80 que les premières banques de séquences sont apparues sous l'initiative de quelques équipes comme celle du Professeur Grantham à Lyon (Gautier et al., [1981](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#43)). Très rapidement avec les évolutions techniques du séquençage, la collecte et la gestion des données ont nécessité une organisation plus conséquente. Ainsi, plusieurs organismes ont pris en charge la production de telles bases de données. En Europe, financée par l'EMBO (European Moleculary Biology Organisation), une équipe s'est constituée pour développer une banque de séquences nucléiques ([**EMBL**](http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html) data library) et en assurer la diffusion (Hamm et Cameron, [1986](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#53)). Cette équipe travaille au sein du Laboratoire Européen de Biologie Moléculaire qui est longtemps resté à Heidelberg et qui se trouve actuellement près de Cambridge au sein de [**l'EBI (European Bioinformatics Institute).**](http://www.ebi.ac.uk/ebi_home.html) Du coté américain, soutenue par le NIH (National Institute of Health) une banque nucléique nommée **[GenBank](http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html)** a été créée à Los Alamos (Bilofsky et al., [1986](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#11)). Cette base de données était distribuée par la société IntelliGenetics et est difusée maintenant par le [**NCBI (National Center for Biotechnology Information).**](http://www.ncbi.nlm.nih.gov/) La collaboration entre ces deux banques a commencé relativement tôt. Elle s'est étendue en 1987 avec la participation de la DDBJ (Dna Data Bank) du Japon pour donner naissance finalement en 1990 à un format unique dans la description des caractéristiques biologiques qui accompagnent les séquences dans les banques de données nucléiques [**(The DDBJ/EMBL/GenBank feature table : Definitions, 1999).**](http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html)

           Parallèlement, pour les protéines, deux banques principales ont été créées. La première, sous l'influence du National Biomedical Research Foundation (NBRF) à Washington, produit maintenant une association de données issues du MIPS (Martinsried Institute for Protein Sequences), de la base Japonnaise JIPID (Japan International Protein Information Database) et des données propres de la NBRF. Elle se nomme la Protein Identification Ressource ([**PIR-NBRF**](http://www.bis.med.jhmi.edu/)) (George et al., [1986](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#44)). La deuxième, **[Swissprot](http://www.expasy.ch/sprot/sprot-top.html)** a été constituée à l'Université de Genève à partir de 1986 et regroupe entre autres des séquences annotées de la PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL (Bairoch et Boeckmann, [1993](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#6) et [1994](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#8)).

           Devant la croissance quasi exponentielle des données [**(Figure 1)**](http://www.citi2.fr/bio2/biocours/Figure1.html) et l'hétérogénéité des séquences contenues dans les principales bases de séquences généralistes, d'autres bases spécialisées sont apparues. Elles se sont constituées autour de thématiques biologiques ou tout simplement en vue de réunir les séquences d'une même espèce et d'en enrichir les annotations pour diminuer, ou lever les ambiguités laissées par les grandes banques publiques. A titre d'exemple on peut citer la base sur les séquences nucléiques d'Eschérichia coli [**ECD**](http://bmc1.bmc.uu.se/srs/srsc?-info+ECD)(Kröger et al., [1991](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#67)), la base [**NRL3D**](http://www.bis.med.jhmi.edu/) des séquences protéiques dont la structure a été déterminée ou bien encore des bases de motifs nucléiques ou proteiques telles que TFD (Ghosh, [1993](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#46)) ou [**PROSITE**](http://www.expasy.ch/prosite/) (Bairoch et Bucher, [1994](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#9)). Le recensement et la description de toutes ces bases publiques généralistes et spécialisées liées à la structure primaire des séquences mais aussi liées à d'autres caractéristiques biologiques peuvent être consultées dans différentes revues (Bishop et al., [1987](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#12) ; Keil, [1990](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#66) ; Damerval et Dessen, [1992](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#28)), et plus récemment dans des numéros spéciaux de la revue Nucleic Acids Research consacrés aux bases de données (1993, 1994 et 1996). Il existe également des bases de données spécifiques qui donnent des listes actualisées des différentes banques d'intérêt biologique comme [**LIMB**](http://www.dna.affrc.go.jp/htdocs/LIMB/) (Listing of Molecular Biology Databases) (Keen et al, [1992](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#65)) ou [**DBCAT**](http://www.infobiogen.fr/services/dbcat/).

**2. LES BANQUES GENERALISTES**

           Les grandes banques de séquences généralistes telles que Genbank ou l'EMBL sont des projets internationaux et constituent des leaders dans le domaine. Elles sont maintenant devenues indispensables à la communauté scientifique car elles regroupent des données et des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique. Leur principale mission est de rendre publiques les séquences qui ont été déterminées, ainsi un des premiers intérêts de ces banques est la masse de séquences qu'elles contiennent. On y trouve également une bibliographie et une expertise biologique directement liées aux séquences traitées. Pour que l'utilisateur puisse s'y repérer, toutes ces informations sont mises à la disposition de la collectivité scientifique selon une organisation en rubriques ou en champs [**(Figure 2: informations d'une entrée de la banque EMBL)**](http://www.citi2.fr/bio2/biocours/Figure2.html). Un détail de cette organisation est donné dans la [**Figure 3**](http://www.citi2.fr/bio2/biocours/Figure3.html). 

**2.1 La qualité des données**

           Il faut avoir conscience que l'information contenue dans ces bases présente un certain nombre de lacunes. Une des principales est le manque de vérifications des données soumises ou saisies surtout pour les séquences anciennes. Les auteurs des séquences ont parfois du mal à restituer les connaissances qu'ils détiennent à propos de leurs données ou bien n'ont pas fait un certain nombre de vérifications de base sur leurs séquences. Il arrive par exemple, que l'on retrouve des segments de vecteurs de clonage dans certaines séquences ou des incohérences dans les caractéristiques biologiques (parties codantes, définition des espèces ou des mots clés...) ou bien encore des informations biologiques incomplètes, voire erronnées. De ce point de vue l'établissement d'un thésaurus précis pour les mots clés faciliterait la vérification comme cela a été permis avec la définition d'arbres des espèces utilisés par plusieurs banques de données. Les organismes responsables de la maintenance de ces banques ont pris conscience de ces problèmes et maintenant de nombreuses vérifications sont faites systématiquement dès la soumission de la séquence. Ceci n'élimine pas la totalité des imprécisions comme par exemple l'existence de doublons car il s'agit là de séquences extrêmement similaires qui correspondent à des entrées différentes dans la banque et dont il est souvent difficile de savoir s'il s'agit de polymorphisme, de gènes dupliqués ou tout simplement d'erreurs établies lors de la détermination des séquences. Il existe d'ailleurs des boites aux lettres électroniques (e- mail) pour informer les gestionnaires des banques d'éventuelles erreurs ou rectifications que chacun pourrait déceler ou proposer. Un autre problème important est le retard de l'insertion d'une nouvelle séquence dans une banque, lié souvent au volume des séquences à traiter qui engendre des priorités ou des choix. Ainsi, il peut y avoir une dizaine de mois de décalage entre la détermination expérimentale d'une séquence et l'introduction de celle-ci dans une banque.

           Malgré cela, il faut souligner l'énorme richesse que représentent ces banques de données, en particulier dans le cadre de l'analyse des séquences. Tout d'abord, le fait que la majorité des séquences connues soit réunie en un seul ensemble est un élément fondamental pour la recherche de similitudes avec une nouvelle séquence. D'autre part, la grande diversité d'organismes qui y est représentée permet d'aborder des analyses de type évolutif. Par exemple, on peut extraire les séquences d'un même gène issu de plusieurs espèces. Un autre intérêt de ces bases réside dans l'information qui accompagne les séquences (annotations, expertise, bibliographie), même si celles-ci sont souvent de qualité inégale. Ces dernières peuvent parfois constituer les rares annotations disponibles sur certaines séquences. Enfin la présence de références à d'autres bases permet d'avoir accès à d'autres informations non répertoriées. Ainsi on peut connaitre l'entrée dans une base protéique de la protéine qui correspond au gène que l'on a repéré dans une base nucléique. La banque SWISSPROT particulièrement riche en références croisées avec d'autres banques et en annotations (par exemple, la notion de "prouvé ou pas expérimentalement" a été récemment introduite dans la table des caractéristiques biologiques) est un exemple de la qualité des données que l'on peut retrouver dans les différentes banques de séquences généralistes de ces dernières années.

**3. LES BANQUES SPECIALISEES**

           Pour des besoins spécifiques liés à l'activité d'un groupe de personnes, ou encore par compilations bibliographiques, de nombreuses bases de données spécifiques ont été créées au sein des laboratoires. Certaines ont continué d'être développées, d'autres n'ont pas été mises à jour et ont disparu car elles correspondaient à un besoin ponctuel. D'autres enfin sont inconnues ou mal connues et attendent qu'on les exploite davantage. Toutes ces bases de données spécialisées sont d'intérêt très divers et la masse des données qu'elles représentent peut varier considérablement d'une base à une autre. Elles ont pour but de recenser des familles de séquences autour de caractéristiques biologiques précises comme les signaux de régulation, les promoteurs de gènes, les signatures peptidiques ou les gènes identiques issus d'espèces différentes. Elle peuvent aussi regrouper des classes spécifiques de séquences comme les vecteurs de clonage, les enzymes de restriction, ou toutes les séquences d'un même génome. En fait très souvent ces bases correspondent à des améliorations ou à des regroupements par rapport aux données issues des bases généralistes. Pour illustrer ce type de banque, nous parlerons ici de bases spécialisées liées aux motifs qui sont particulièrement utilisées dans l'analyse des séquences. 

**3.1 Les bases de motifs**

           On sait que certains segments d'ADN ou de protéines sont déterminants dans l'analyse des séquences car ils correspondent à des sites précis d'activité biologique comme par exemple les éléments de régulation des gènes ou les signatures peptidiques. C'est pourquoi des bases spécialisées se sont naturellement constituées autour de ces séquences. 

**3.1.1 Les bases de motifs nucléiques**

           La plupart de ces bases consiste à recenser dans des catalogues les séquences des différents motifs pour lesquels une activité biologique a été identifiée. Certains motifs sont simples et non ambigus comme ceux que l'on trouve dans la base Rebase (Roberts et Macelis, [1994](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#81)) qui donne une liste actualisée des sites de coupure des enzymes de restriction. D'autres correspondent à des activités biologiques plus complexes et engendrent donc des séquences moins précises comme par exemple les sites d'épissage ou les signaux de transcription. Pour ces derniers types de motifs, des compilations ont été établies pour donner des listes annotées de motifs qui peuvent être communs à plusieurs séquences. Par exemple, la revue Gnomic (Trifonov et Brendel, [1986](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#92)) a permis ainsi de recenser un grand nombre de signatures nucléiques. Maintenant, ce genre de compilation réunit essentiellement les informations liées à la transcription. Ainsi aujourd'hui, il existe principalement deux bases de motifs nucléiques qui sont régulièrement actualisées et qui correspondent à un travail de synthèse bibliographique : il s'agit des bases de facteurs de transcription [**TFD**](gopher://gopher.nih.gov/00/molbio/other/atfd) (Ghosh, [1993](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#46)), et [**TRANSFAC**](http://www.gene-regulation.com/pub/databases.html#transfac) (Knüppel et al., [1994](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#68)). Malgré tout, certaines redondances existent dans ces bases et certains facteurs peuvent être donnés sous des formes différentes. De plus, une autre approche s'est constituée pour représenter des sites nucléiques spécifiques impliqués dans des processus biologiques. Celle-ci permet de définir les sites, non plus sous forme de chaines de caractères avec d'éventuelles ambiguités à certaines positions, mais sous forme de matrices reflétant une probabilité d'apparition de chaque base à chaque position du site (c.f. la définition des motifs dans les programmes de recherche de motifs). Ces développements ont donc conduit à la constitution de bases de matrices représentant des sites spécifiques. Pour illustrer ces deux types de représentation de l'information, nous présentons ici une des bases de facteurs de transcription les plus utilisées TFD (Ghosh, [1993](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#46)), ainsi que la base IMD (Chen et al., [1995](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#18)) qui regroupe un grand nombre de matrices caractéristiques d'éléments de transcription. 

**La base de facteurs de transcription TFD**   
   
          TFD est une base dédiée aux facteurs de transcription eucaryotes. Une partie des données a été extraite de GenBank et une autre partie provient de synthèses bibliographiques réalisées à partir de publications traitant de différents aspects de la transcription. La base est organisée en plusieurs fichiers permettant de regrouper différentes classes d'information que l'on connait au niveau de la transcription. Ainsi la base renferme non seulement des données nucléiques mais aussi des informations sur les séquences protéiques directement impliquées dans la transcription comme les domaines protéiques interagissant avec l'ADN ou les cofacteurs de transcription. L'information la plus importante est bien sûr la liste des sites ou motifs nucléiques ([**Figure 4**](http://www.citi2.fr/bio2/biocours/Figure4.html)). Un effort particulier a été réalisé pour spécifier à partir de quel promoteur ou de quel gène sont issus ces sites et pour établir s'il s'agit d'une séquence consensus ou d'une séquence actuellement unique, et pour donner la localisation relative du site d'initiation de la transcription (Ghosh et al., 1990 et Ghosh, [1993](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#46)). 

**La base de motifs IMD (Information Matrix Database)**   
   
           Cette base est construite à partir de sites de facteurs de transcription trouvés dans les bases TFD (Ghosh, [1993](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#46)) et TRANSFAC (Knüppel et al., [1994](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#68)) ou à partir de données issues directement de publications. Durant la construction de cette base, les auteurs (Chen et al., [1995](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#18)) ont pris un soin particulier pour identifier les multiples représentations d'un même site de fixation protéique et pour regrouper tous les motifs correspondant au même facteur de transcription. Lorsque pour un site, un nombre suffisant de représentants est connu, les motifs sont utilisés pour établir une matrice de fréquences ou de pondération (weight matrix) qui donne la chance d'apparition pour chaque nucléotide de se trouver à une position déterminée. Cette base contient actuellement 532 matrices regroupant sept classes d'organismes différents auxquelles sont associées les références bibliographiques correspondantes. Un exemple de l'information contenue dans cette base est donnée dans la [**Figure 5**](http://www.citi2.fr/bio2/biocours/Figure5.html). 

**3.1.2 Les bases de motifs protéiques**

           Il existe principalement deux types de bases de motifs qui permettent de recenser des signatures protéiques liées à des activités biologiques. Celles qui regroupent des motifs consensus et celles qui donnent des régions actives sous forme d'alignements multiples. Nous présenterons ici deux bases couramment utilisées qui reflètent ces deux aspects. 

**La base de motifs protéiques PROSITE**  
   
           La base [**PROSITE**](http://expasy.hcuge.ch/sprot/prosite.html) peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique. Elle est établie en regroupant, quand cela est possible, les protéines contenues dans Swissprot par famille comme par exemple les kinases ou les protéases. On recherche ensuite, au sein de ces groupes, des motifs consensus susceptibles de les caractériser spécifiquement. La conception de la base repose sur quatre critères essentiels : 1) collecter le plus possible de motifs significatifs, 2) avoir des motifs hautement spécifiques pour caractériser au mieux une famille de protéines, 3) donner une documentation complète sur chacun des motifs répertoriés, et 4) faire une révision périodique des motifs pour s'assurer de leur validité par rapport aux dernières expérimentations. L'essentiel de l'expertise est basé sur un réseau de correspondants spécialistes des sujets traités. La base est organisée en deux parties. La première contient l'identification et la description de chaque motif. La deuxième contient l'information qui documente chaque motif (Bairoch, [1993](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#7) ; Bairoch et Bucher, [1994](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#9)). Un exemple de ces deux formes d'information est donné dans la [**Figure 6**](http://www.citi2.fr/bio2/biocours/Figure6.html). 

**La base de motifs protéiques BLOCK**  
   
           La base [**BLOCK**](http://www.blocks.fhcrc.org/) est également basée sur un système qui détecte et assemble les régions conservées de protéines apparentées. La détection consiste en des alignements multiples à partir desquels des blocs de séquences sont engendrés. Un bloc est la superposition de segments protéiques très similaires sans insertion-délétion. L'ensemble de tous ces blocs forme la base. C'est ainsi que Henikoff et Henikoff ([1991](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#54)) ont défini 1764 blocs à partir des 437 groupes de protéines recensés durant l'établissement de PROSITE. Les motifs représentés par la base BLOCK sont généralement plus courts que ceux donnés par la base PROSITE mais les différences fondamentales entre ces bases résident dans la représentation des données. Les motifs de PROSITE sont définis sous forme de chaînes de caractères prenant en compte des insertions et des ambiguités sur les acides aminés conservés alors que les motifs de la base BLOCK sont représentés par des suites d'acides aminés donnés sous forme d'alignements multiples.

           L'utilisation de ces bases de motifs est devenue un outil essentiel dans l'analyse des séquences pour tenter de déterminer la fonction de protéines inconnues ou savoir à quelle famille appartient une séquence non encore caractérisée. En général, les bases comme TFD ou IMD sont employées sur les séquences promotrices des gènes tandis que celles comme PROSITE ou BLOCKS sont utilisées sur des protéines inconnues ou bien des séquences protéiques traduites à partir de cDNA ou de séquences génomiques. Pour détecter une fonctionalité sur une séquence, il suffit d'exécuter un programme qui s'appliquera à repérer la présence de certains motifs recensés dans ces bases et ainsi à prédire l'appartenance de la séquence testée à un groupe de séquences ayant une signature commune. Nous verrons dans le chapitre II (cf. les programmes de recherche de motifs) les avantages de chacune des définitions de motifs (motif consensus, matrice ou alignement multiple) ainsi que les différents algorithmes que l'on peut appliquer en fonction de ces définitions.

**4. LA DIFFUSION ET L'UTILISATION DES BANQUES DE DONNEES** 

**4.1 La diffusion**

           La plupart des bases de données sont mises à jour par la production de versions actualisées. Le rythme est de trois à six versions par an pour les bases de données les plus couramment utilisées (actuellement, quatre versions par an pour L'EMBL et six pour GenBank) et de seulement une version par an ou tous les deux ans pour la plupart des autres. Pendant longtemps, la principale distribution fut l'envoi postal de bandes magnétiques aux personnes ayant souscrit un abonnement. Progressivement le CD-ROM a remplacé ce support de stockage et a permis une plus grande diffusion des données. Depuis le début des années 90, avec l'installation massive des réseaux informatiques à hauts débits qui permettent d'atteindre une machine située à plusieurs milliers de kilomètres de son terminal, beaucoup de laboratoires rapatrient les bases de données via ces réseaux à partir de serveurs publics. Ces réseaux informatiques rapides et les services qui en découlent permettent une large diffusion des bases. Ainsi beaucoup de serveurs mettent gratuitement à disposition de nombreuses bases, dont les grandes banques de séquences généralistes comme l'EMBL avec une mise à jour quotidienne des données, mais également un grand nombre d'autres bases dont la diffusion était auparavant plus restreinte. De ce fait, il résulte une banalisation de l'accès à l'information. Il n'est même plus nécessaire d'avoir localement les bases de données ou de se connecter par des procédures complexes à un centre serveur privilégié pour pouvoir exploiter aisément le contenu de ces bases. C'est ainsi que l'on estime maintenant le nombre total d'utilisateurs de la banque EMBL à plus de 50 000. 

**4.2 L'interrogation**

           Pour que les bases de données soient plus facilement exploitables, et que les utilisateurs puissent extraire les sous-ensembles de séquences qui les intéressent, deux types de logiciels leur sont généralement offerts soit localement, soit sur les nombreux serveurs publics. Les premiers logiciels sont des systèmes d'interrogation dédiés qui sont programmés exclusivement pour la manipulation de séquences biologiques. Par contre, les deuxièmes sont des programmes établis à l'aide de systèmes de gestion de bases de données (SGBD) qui utilisent un langage de requête standard et un format de stockage des données indépendant de la nature de l'information contenue dans les bases. 

**4.2.1 Les systèmes dédiés**

           Il existe plusieurs systèmes dédiés dont la plupart permettent des interrogations multicritères simples. On en trouvera une liste assez complète dans une publication de Ginsburg ([1987](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#47)). Ils sont généralement adaptés aux principales bases de données (EMBL, GenBank etc...). On peut citer parmi eux le programme Stringsearch du logiciel GCG (Devereux et al., [1984](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#31)) qui permet une interrogation à deux critères. D'autres permettent des interrogations plus complexes comme les logiciels [**ACNUC**](http://pbil.univ-lyon1.fr/databases/acnuc.html) (Gouy et al., [1985](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#50)) ou [**SRS**](http://www.infobiogen.fr/srs/index.html) (Etzold et Argos, [1993](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#37)). Ces deux logiciels méritent que l'on s'y attarde un peu plus car le premier est l'un des plus performants pour l'interrogation des bases de biomolécules et le deuxième illustre assez bien l'orientation des nouveaux logiciels développés dans le domaine durant les dernières années.

**Le logiciel ACNUC**  
   
           Développé à Lyon chez le Professeur Grantham, il a été un des premiers logiciels qui puisse répondre en peu de temps à des questions sophistiquées telles que : "je recherche dans l'EMBL les séquences codants pour des protéines de levure ou de souris qui ne soient pas mitochondriales et ayant une longueur supérieure à 500 paires de bases". Ceci a été réalisé grâce à la construction de fichiers index représentant des critères de sélection (mot-clé, auteurs, espèces, revues, type de molécule...) et une organisation des fichiers permettant d'effectuer des liens entre critères ainsi qu'un langage de requête basé sur les opérations logiques ET, OU et NON. On peut regrouper en quatre catégories les commandes possibles qui servent à la manipulation du logiciel :  
           1) sélection : Ces commandes permettent de constituer des listes de séquences qui correspondent à des sous-ensembles de la banque. Il est possible d'utiliser différents critères selon une syntaxe logique où les critères utilisés doivent être nommés explicitement (Exemple: SP pour species, KE pour keyword, AU pour author, M pour molécule...)  
           2) définition : Ces commandes permettent de définir avec plus de précision les critères de sélection autorisés. On peut ainsi visualiser tout ou partie de l'arbre des espèces ou des mot-clés qui sont utilisés dans la base que l'on consulte.  
           3) information : Ces commandes permettent l'édition de toute ou partie de l'information liée aux séquences sélectionnées.  
           4) gestion : Ces commandes permettent de faire des modifications, des extractions, ou des suppressions dans les listes de séquences déjà selectionnées.

           Ce logiciel, très répandu en France, est l'un des plus performants pour l'interrogation malgré le nombre croissant de séquences dans les banques de données. Il est principalement utilisé pour les grandes banques généralistes (EMBL, GenBank, PIR-NBRF) même si certaines bases thématiques l'utilisent comme HOVERGEN, banque de données sur les gènes homologues de vertébrés (Duret et al., [1994](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#36)) ou NRsub, base non redondante sur le génome de Bacillus subtilis (Perriere et al., [1994](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#78)). Le logiciel intègre une interface graphique sous X Window ainsi qu'une [**interface WWW**](http://pbil.univ-lyon1.fr/search/query.html) disponible sur le [**pole bioinformatique lyonnais**](http://pbil.univ-lyon1.fr/) (Université Lyon1).

**Le logiciel SRS (Sequence Retrieval system)**  
   
           C'est un système d'interrogation basé sur l'indexation des champs contenus dans les différents formats que proposent les banques de données. Il est construit à l'aide de deux types de programmes distincts qui sont en étroite interaction. Les premiers, classiques, écrits en langage de programmation standard, servent entre autre à exploiter les données indexées. Les deuxièmes utilisent un langage de programmation propre ODD (Object Design and Definition) qui permet de reconnaître le format des banques utilisées et de rendre celui-ci indépendant de l'exploitation des données. Ainsi le principal avantage de ce système est la possibilité de pouvoir indexer en même temps une grande quantité de banques sans se soucier de l'organisation de celles-ci et donc de pouvoir manipuler avec le même langage de requête les principales bases de séquences généralistes et beaucoup de bases spécialisées. Le langage de requête est similaire à celui d'ACNUC, c'est-à-dire que l'on peut utiliser plusieurs critères de sélection avec les opérateurs logiques ET, OU et NON. Ce système permet néanmoins une moins grande puissance d'interrogation qu'ACNUC sur une base particulière mais a l'avantage de pouvoir faire des interrogations sur plusieurs bases à la fois ou de pouvoir utiliser les références croisées entre celles-ci. Par exemple, on peut rechercher dans l'EMBL les séquences issues d'une interrogation de SwissProt. Ce logiciel dispose d'une [**interface WWW**](http://www.infobiogen.fr/srs/index.html) que la plupart des [**serveurs**](http://www.infobiogen.fr/srs/srs5list.html) impliqués dans la diffusion et la consultation de données biologiques mettent à la disposition des utilisateurs.

**4.2.2 Les Systèmes de Gestion de Bases de Données (SGBD)**

           L'information structurée d'une base de données, qu'elle soit biologique ou pas, peut être exploitée par un SGBD pour gérer ou retrouver rapidement les données. La construction de la base passe par une définition des données qui est indépendante des programmes d'exploitation de ces données. Durant la définition, qui se fait sous forme de tables, on peut établir des liens qui vont déterminer une organisation logique des données qui sera ensuite exploitée à l'interrogation. Il existe plusieurs types de SGBD, le modèle dit Relationnel étant maintenant le plus utilisé à travers des systèmes comme Oracle, Sybase ou 4ème Dimension. Le langage de requête qu'ils utilisent est standard et permet d'exploiter la définition des données que l'on a faite. Pour plus d'informations sur les SGBDs, on peut consulter des ouvrages comme celui de Carrez ([1990](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#17)).

           Un des avantages de ces systèmes est leur mise en oeuvre rapide car ils possèdent un certain nombre d'outils et de fonctionnalités de base qui facilitent certaines étapes de programmation comme la présentation des résultats ou les opérations d'ajout, de suppression, de recherche ou de tri d'information. Ils permettent également une évolution facile de la base par ajout de nouvelles entités et de nouveaux liens sans modification de la structure initiale. Cependant, indépendemment de leur coût financier, ils prennent souvent une part trop importante des ressources des machines sur lesquelles ils sont installés, limitant ainsi le nombre d'utilisateurs simultanés ou nécessitant l'exclusivité d'une machine. Dans le domaine de la biologie, ils sont généralement utilisés pour l'interrogation de bases qui ont un nombre limité d'entrées car les temps de réponse peuvent devenir importants lors d'interrogations qui nécessitent le croisement de plusieurs listes volumineuses, comme par exemple, la recherche de l'intersection entre toutes les séquences humaines et toutes les séquences codants pour des protéines dans une base comme l'EMBL. Par contre, ils sont souvent utiliseés dans la construction des bases, notamment par les organismes qui gèrent les grandes bases de séquences. 

**5. CONCLUSION**

           Le nombre de données dans le domaine de la biologie ne cesse d'augmenter en particulier avec le séquençage des génomes de différents organismes mais l'on assiste également à une grande diversification des informations produites (séquences primaires, structures moléculaires, cartographie, collection de souches ou de clones...). Toutes ces données sont regroupées dans des banques de données très variées dans leur volume et leur nature. On ne peut maintenant imaginer leur consultation sans l'apport de l'informatique. Cet apport est devenu considérable durant les dernières années, en particulier avec l'extension des réseaux à haut débits. Il permet aux scientifiques d'utiliser de nouveaux outils allant de la simple interrogation textuelle à la présentation graphique des données en passant par l'utilisation du multifenêtrage ou de documents sonores ou vidéo. Il est donc évident que l'organisation et l'interrogation des données vont en être profondément changées. Cette transformation est déjà apparente dans le développement de certains logiciels qui proposent de plus en plus d'interactions entre les bases de données, ceci en exploitant davantage les liens qui existent entre elles. Le logiciel [**SRS**](http://www.infobiogen.fr/srs/index.html), qui est installé sur de nombreux [**serveurs WWW**](http://www.infobiogen.fr/srs5/srs5list.html), est un exemple de cette évolution en proposant une consultation multibase avec la même interface graphique.

           Un des modèles de l'intégration des données qui montre ce que seront les outils de demain est le système de base de données [**ACEDB**](http://www.sanger.ac.uk/Software/Acedb/) qui a été développé au cours du projet génome de Caenorhabditis elegans par Richard Durbin et Jean Thierry-Mieg en 1992. Dans ce sytème, les données sont stockées comme objets appartenant à des classes. Ces objets (gènes, allèles, clones, cartes génétiques, cartes physiques, séquences, publications...) sont mémorisés sous formes d'arbres permettant une organisation hiérarchisée et des représentations différentes de ces objets. De plus, on trouve dans ce système une grande efficacité et une grande flexibilité dans la manipulation, la visualisation et l'interrogation des informations stockées. En fait, ce qui caractérise essentiellement la tendance actuelle, c'est l'intégration de plus en plus grande des données, liée à une souplesse accrue de l'utilisation. Ceci permet à l'utilisateur d'exploiter en peu de temps et avec peu de connaissances techniques des données intégrées et d'avoir une vision synthétique de son objet de recherche. Par exemple, à partir du nom d'un gène, on peut avoir accès à la séquence primaire de la partie codante et à la protéine qui en découle, à sa localisation physique et génétique, à la bibliographie associée etc... Ainsi, on s'éloigne de l'époque où il fallait que l'utilisateur connaisse suffisamment bien les bases de données et les logiciels associés pour en extraire toute l'information existante.