Bases de données en biologie

A. VIGNAL

INRA, Laboratoire de Génétique Cellulaire, BP 27, 31326 Castanet-Tolosan cedex

e-mail : Alain.Vignal@toulouse.inra.fr

Résumé. *Les bases de données en biologie doivent permettre le stockage d'une quantité croissante d'informations dont la nature est hétérogène. Des entités parfois très différentes sont à représenter, ainsi que leurs relations. De plus, les techniques variées utilisées pour générer les données doivent être prises en compte. De ce fait, il existe un foisonnement de bases de données spécialisées, dont nous décrivons ici les principales, ainsi que leurs interrelations.*

Le problème des bases de données en biologie réside dans la complexité et la richesse des données qu'elles doivent contenir. En effet, une base de données doit pouvoir permettre tout d'abord un stockage cohérent et structuré des données, mais aussi un accès simplifié à une grande masse d'information, tout en préservant autant que possible sa richesse et sa diversité. Pour des raisons scientifiques, mais aussi pour des raisons historiques, les données en biologie sont disséminées dans plusieurs centaines de bases indépendantes et spécialisées.

La nature des données à représenter en biologie est extrêmement variée, ainsi que les techniques ayant été utilisées pour les générer et l'utilisation qui en sera faite. Rien qu'en se restreignant aux activités du domaine de la génétique moléculaire, sont produites et utilisées des données : de séquences nucléotidiques et protéiques, avec toute la variété d'annotations qui en découlent, en relation avec la structure, l'évolution, la fonction et la régulation des gènes ; de cartographie génique à diverses résolutions et utilisant des techniques différentes ; de profils d'expression ; de mutations avec leurs effets ; de voies métaboliques, etc. Le plus souvent, il est nécessaire de réaliser une combinaison de ces différents types de données, afin de progresser dans la résolution d'un problème donné. La mise en relation des données présentes dans les différents types de bases est parfois rendue difficile pour des raisons d'hétérogénéité des nomenclatures et du type des données.

1 / Les bases de données de séquences nucléotidiques

Description

Les trois principales bases de données de séquences nucléotidiques sont GenBank (http://www.ncbi.nlm.nih.gov/), EMBL (European Molecular Biology Laboratory ; http://www.ebi.ac.uk/embl.html) et DDBJ (DNA Data Bank of Japan ; http://www.ddbj.nig.ac.jp). Ces trois bases de données publiques contiennent toutes les séquences nucléiques et protéiques connues, avec annotations bibliographiques et biologiques. Des échanges journaliers permettent d'assurer que les trois bases sont à jour. Jusqu'à récemment, la source principale de données de ces trois bases a été les soumissions directes de séquences par les chercheurs et les échanges journaliers entre bases. Maintenant, la plus grande part de données provient des centres de séquençage, avec une part croissante de génomes complets et d'EST (Expressed Sequence Tags). Après avoir doublé tous les 18 mois, le nombre de séquences double actuellement tous les 15 mois. En août 1998, les 2,5 millions d'entrées représentaient 1,8 milliard de bases. Deux génomes complets avaient été ajoutés en 1996, 6 en 1997 et 10 en 1998, dont le génome de *Caenorhabditis elegans*, dont la taille est de 100 Mégabases. Environ 20 microorganismes sont actuellement en cours de séquençage et la plupart des résultats sont attendus pour 1999. Plus de 40 000 espèces différentes sont représentées et environ 900 sont ajoutées par mois. Les séquences humaines représentent 54 % des entrées.

Le type d'analyse le plus fréquent effectué sur les bases de séquences nucléotidiques est la recherche de séquences similaires à une séquence donnée. La recherche des meilleurs alignements est réalisée à l'aide de la famille de programmes BLAST. Chaque alignement BLAST est accompagné d'un score et d'une indication de la valeur statistique.

Traditionnellement, les séquences soumises aux bases de données étaient des fragments étudiés par des chercheurs en raison de leur intérêt direct en relation avec un sujet biologique donné : structure de gènes et de familles de gènes, étude de leurs régions régulatrices, phylogénie, étude des séquences répétées. Les soumissions correspondaient à des séquences lues sur les deux brins d'ADN, avec annotations précises par les auteurs.

Actuellement, une large part des entrées proviennent de projets systématiques par lesquels une grande quantité de séquences est produite, dont l'analyse du contenu biologique est réalisée automatiquement (annotation automatique), ou pour lesquelles l'information de séquence ne sert que d'outil pour la construction de cartes. Différentes catégories de séquences peuvent être ainsi distinguées, séparées en divisions (tableau 1).

Tableau 1. *Les divisions dans les banques de séquences nucléotidiques (d'après : http://www.ebi.ac.uk /embl/Documentation/User_manual/database_div.html).*

Division	Code
ESTs	EST
Bacteriophage	PHG
Fungi	FUN
Genome survey	GSS
High Throughput Genome	HTG
Human	HUM
Invertebrates	INV
Organelles	ORG
Other Mammals	MAM
Other Vertebrates	VRT
Plants	PLN
Prokaryotes	PRO
Rodents	ROD
STSs	STS
Synthetic	SYN
Unclassified	UNC
Viruses	VRL

EST (Expressed Sequence Tags)
Fin 1998, il y avait environ 1,8 million d'entrées pour 130 organismes. L'organisation utile des ESTs est réalisée par la création de la collection UniGene (http://www.ncbi.nlm.nih.gov/UniGene/index.html), dans laquelle les entrées correspondant à un même organisme sont groupées par séquences ayant des 3' non codant identiques. De cette manière, plus d'un million d'ESTs humains ont été regroupés en 52 000 clusters, chacun pouvant être considéré comme représentant un gène.

STS (Sequence-Tagged Site)
Plus de 60 000 séquences y sont répertoriées, avec les conditions de PCR. Ces étiquettes servent à construire des cartes physiques.

GSS (Genome Survey Sequence)
Plus de 220 000 entrées, correspondant à des séquences d'extrémités de BACs, générées lors des projets de séquençage de génomes complets. Ces séquences sont utilisées avec des STSs, pour ordonner les clones BAC en contigs.

HTG (High Throughput Genomic)
Ces séquences proviennent de projets de grande ampleur et sont en cours de finition. Une fois terminées, elles seront placées dans la division correspondant à leur organisme.

Le problème des annotations et les limites des bases généralistes

Les données soumises initialement par des projets de séquençage comportent des annotations préliminaires, basées sur les résultats obtenus en utilisant des programmes de prédiction. Par ailleurs, une séquence donnée peut avoir présenté des similarités avec un grand nombre de séquences cibles, la liste de ces séquences cibles pouvant varier en fonction des algorithmes et des paramètres utilisés, ainsi que de l'époque où l'analyse a été faite. Les groupes de séquençage peuvent ne pas maintenir les annotations après qu'un projet soit terminé. Par la suite, elles seront maintenues par des bases de données spécialisées. Il existe plusieurs catégories de bases spécialisées, dans lesquelles les séquences sont regroupées en fonction d'un sujet défini et associées à des informations complémentaires : base de données par espèce, de motifs nucléotidiques, de mutations, etc.

2 / Les bases de données de séquences protéiques

Les catégories principales sont les bases de séquences traduites à partir de séquences nucléotidiques ; les bases de séquences comportant des annotations, pouvant parfois être de très haut niveau sur la fonction, les domaines, les modification post-traductionelles, les variants ; les bases de séquences de domaines protéiques, tenant compte de la modularité des structures protéiques.

3 / Les bases de données en cartographie

Il existe un grand nombre de types de cartes géniques produites par des techniques variées, utilisées en fonction de la résolution qu'elles permettent d'obtenir et de leur utilité pour la localisation de gènes, de phénotypes ou pour le séquençage. Les principales cartes sont de type cytogénétique, génétique, d'hybrides irradiés ou de contigs de clones (tableau 2).

Les données ayant servi à réaliser les différentes cartes géniques figurent dans des bases spécialisées en fonction du type de carte et de l'espèce. Il existe par exemple plusieurs bases de données contenant des données brutes pour des cartes humaines d'hybrides irradiés, en fonction de la collection d'hybrides utilisée. Les différentes cartes géniques humaines sont regroupées dans la base GDB (Genome Database : http://www.gdb.org/), qui permet de visualiser les différentes cartes du génome, de rechercher les marqueurs polymorphes, les réactifs tels que des clones ou des oligonucléotides, ainsi que de rechercher la localisation de gènes et de marqueurs. Des liens permettent parfois l'accès aux bases contenant les données brutes, telles que RHdb (Radiation Hybrid database : http://corba.ebi.ac.uk/ RHdb/), permettant la récupération des données de typage sur panel d'hybrides pour un marqueur. Les données de typage sur plusieurs marqueurs peuvent être récupérées dans des RHdb, afin de réaliser des calculs. De même, pour les cartes génétiques, les données de typages sont accessibles dans la base de données du CEPH (Centre d'Etudes du Polymorphisme Humain : http://landru.cephb.fr/cephdb/). Le

Tableau 2. *Les types de cartes géniques.*

Type de carte	Méthode	Avantage	Inconvénient
Cytogénétique	Hybridation sur chromosomes en méthaphase	Localisation rapide de gènes ou marqueurs anonymes	Faible résolution Nécessite des fragments d'ADN de grande taille.
Génétique	Estimation des fréquences de recombinaison	Localisation de locus contrôlant des phénotypes	Localisation dépendante de l'existence de polymorphisme
Hybrides irradiés	PCR sur hybrides cellulaires ayant retenu des fragments d'ADN cassés par irradiation	Haute résolution Rapidité	Nécessite l'obtention d'une collection d'hybrides. Artefacts PCR
Contigs physiques	Alignement de clones d'ADN génomique	Cartographie à très haute résolution	Uniquement dans des régions ciblées

fonctionnement des bases de données de génétique animale est similaire (adresses Internet des principaux sites : tableau 3).

Tableau 3. *Bases de données en génétique animale.*

Bovins http://locus.jouy.inra.fr/cgi/bovmap/intro.pl http://probe.nalusda.gov:8000/animal/aboutbovgbase.html
Porcins http://probe.nalusda.gov:8000/animal/aboutpigbase.html http://sol.marc.usda.gov/genome/swine/swine.html
Ovins http://www.ri.bbsrc.ac.uk/sheepmap/
Poule http://www.ri.bbsrc.ac.uk/chickmap/

4 / Autres bases de données

Les gènes peuvent se caractériser par leur structure et par leur position sur le génome, mais il est essentiel de pouvoir aussi décrire les aspects fonctionnels de leurs implications dans le métabolisme au niveau de la cellule ou de l'organisme, ainsi que de pouvoir retracer des liens phylogénétiques. Pour cela, il existe un grand nombre de bases de données spécialisées, contenant des données sur les voies métaboliques et les régulations, l'expression, l'anatomie, la taxonomie, les structure tridimensionnelles de protéines. Certaines bases sont spécialisées pour un organisme, une fonction, un tissu, ou pour une famille de protéines, voire même pour une seule protéine, avec un répertoire des mutations et de leurs effets phénotypiques. Des résultats expérimentaux tels que ceux de gels d'électrophorèse 2D de protéines peuvent être comparés dans des bases spécialisées. Des outils, tels que les enzymes de restriction, des vecteurs et des souris transgéniques pouvant servir de modèles animaux, sont également répertoriés dans d'autres bases de données. Dbcat (DataBase catalogue : http://www.infobiogen.fr/services/dbcat/), un catalogue de bases de données comprenant 416 entrées, a été réalisé par le groupe Infobiogen.

5 / Le problème des liens

Les recherches en génétique moléculaire nécessitent de pouvoir recouper des informations de nature différente parfois sur un grand nombres de gènes. En génétique animale, les données de cartographie comparée entre espèces sont beaucoup utilisées. L'utilisation de plusieurs bases de données présentant des structures parfois différentes est donc nécessaire. De plus, les problèmes techniques d'accès aux données peut se retrouver compliqué par l'hétérogénéité de la nomenclature des gènes, qui peut varier en fonction des espèces, mais aussi des disciplines de la biologie : un même gène peut être nommé selon la structure de la protéine pour laquelle il code, sa fonction, son implication dans une pathologie ou selon la description d'un phénotype mutant. Les travaux actuels portent donc sur les problèmes de nomenclature et sur des outils communs permettant d'améliorer les connexions entre bases et les requêtes. L'outil SRS (Sequence Retrievial System : http://srs.ebi.ac.uk:/), a été développé dans un premier temps pour l'extraction de séquences nucléiques et protéiques, et son application s'est aujourd'hui étendue aux requêtes dans des bases de données plus variées : voies métaboliques, structure tri-dimensionnelle des protéines, cartographie, mutations, taxonomie, etc. Les liens et les échanges entre bases différentes sont en cours d'amélioration grâce à l'utilisation de CORBA (Common Object Request Broker Architecture), développé pour répondre aux besoins d'homogénéisation du développement d'applications.

Référence

Roberts R.J., Gait M.J. (eds), 2000. Nucleic Acids Research, Database issue. Oxford University Press.