

Biological sequence databases

Frédérique Galisson

20th June 2001

Over the last two decades, while the progress of biotechnology led to an extraordinary accumulation of biological results and new data, the development of computer science enabled the appearance of powerful systems for information storage, management and processing. The automation of many molecular biology experiments, notably those related to DNA sequencing, made the productivity of biologists increase exponentially. Consequently, one of the most important aspects of bioinformatics is the development of numerous biological data banks (“data banks” will be called “databases” in the rest of the text, even if formally their two meanings are different) ([1]). There are several hundred of them (see the first issue of *Nucleic Acids Research* which each year is dedicated to biological databases); they contain various kinds of information (images, metabolic pathways, chemical structures, physical or genetics maps, gene expression levels, etc), most of them concerning sequence data ([2], [3]). The main criteria used to qualify and classify the biological sequence databases are whether or not the data they deal with are primary, a vocation to be general or on the contrary their specialization, and whether or not they link and integrate sequence data with other biological information resources. Primary data are defined as sequences obtained directly from experiments, not including those derived (from them) by computation. One should notice that a strict application of this definition would exclude all the protein sequences that are conceptual translations of nucleic acid sequences (the enormous majority of those that fill the banks), as well as genomic DNA sequences obtained by shotgun sequencing and which are consensus computed from multiple alignments of overlapping fragments. Nevertheless the general DNA and protein databases are considered to be primary data banks, aside from those containing derived sequence data (alignments, motifs...), and the specialized ones which focus on a particular organism or biological topic, often integrating information other than the sequences themselves (some are sometimes called “knowledge databases”) ([4], [5], [1]).

This article will focus on the big public repositories of nucleic acid and protein sequences, only mentioning some of the numerous specialized or derived databases. It will mainly concern the contents and organization of these resources and leaving out the computer science aspects of the subject.

1 Nucleic acid sequence databases

1.1 The EMBL/GenBank/DDBJ international collaboration

There are three major data repositories working in tight collaboration, whose missions are to collect, maintain and publicly release the primary information consisting in all the known DNA sequences:

- EMBL ([6]) was created in 1980 by the EMBL (European Molecular Biology Laboratory); it is maintained at the EBI (European Bioinformatics Institute, [7]), which depends on the EMBL.
- GenBank was created in 1982 at the Los Alamos National Laboratory, and is maintained at the NCBI (National Center for Biotechnology information, [8]) which depends on the American NIH (National Institute of Health) ([9], [4]).
- DDBJ (DNA Data Bank of Japan, [10]) was created in 1986 and is maintained at the Center for Biology Information of the National Institute of Genetics ([11]) in Japan.

EMBL and Genbank began working in close collaboration as early as the beginning of the eighties and were joined by DDBJ when it was created. The terms of this collaboration were officially formalized in 1988. They imply notably a daily exchange of collected data, the application of common rules regarding the wordlists and formats that are used to describe the data, and the recognition that an entry may be modified only by the center to which it had been initially submitted (in order to avoid possible update conflicts) ([4], [12]).

1.2 The data

These three banks deal with all kinds of primary nucleic acid sequences, wherever they come from, that is whatever the molecular species (DNA or RNA of all kinds, including synthetic genes), as well as the experimental methods that are used to obtain the sequence. Nevertheless, the information that enables to answer questions such as: “what does this sequence

represent?” and “to which molecule does it correspond?” are supposed to come with the sequence.

A first generation of sequences (those of the eighties) mainly correspond to genes that were individually isolated and cloned using various experimental approaches, in many cases from the study of a protein, a function or a phenotype. The beginning of the nineties was marked by the first massive productions of EST (Expressed Sequence Tags, [13]), followed by the first complete chromosome ([14]) and genome ([15], [16]) sequences. These systematic sequencing projects are responsible for the exponential growth rate of the databases, the doubling time of which is now less than one year. This acceleration is due in particular to the decision among the public sequencing centers to release their sequences in the public databases whatever their production status (one distinguishes four steps that are characterized by the finishing status of the sequences). Unfinished sequences are gathered under the generic term HTG (High Throughput Genomic sequences). Their quantitative importance (more than one third of the nucleotides contained in Genbank/EMBL/DDBJ) is illustrated in diagrams that show the evolution of the database contents ([17] and [18], more detailed), as well as the relative importance of the different kinds of data (the EBI also proposes a follow-up of genomic data: [19]).

1.3 Data collection and distribution

The sequences that are collected by each of the three centers may come from individual submissions by researchers, big sequencing centers, patent offices releasing sequences, or from one of the two other databases. Because they exchange their data on a daily basis, each one is comprehensive, and thus by a few hours or so their contents are equivalent. The only difference concerns the formats that are used to store the information. At regular intervals (two to three months), each center releases a new version of its database, to which are associated a version number and the date. Between two releases, each center maintains (and daily releases) an update file containing all the new sequences that have been collected since the last database release, as well as those that have been subjected to an update (for example, a correction about the sequence or a modification of its annotation). One should notice that in the case of an update, a sequence will thus be present simultaneously in the update file (the new version) and in the current release (the old version). A search in the comprehensive set “release plus updates” will then find this sequence two times (they both will have the same accession number, but different version numbers. See paragraph 1.4). When the next

release is built, the cumulative updates of the few months separating the two distributions are added to the previous release; in the case of sequences that occur in both sets, only the most recent version is kept. This operation (eliminating old versions) is performed daily by some WEB sites which propose a “*GenBank/EMBL/DDBJ non-redundant International Nucleotide Database*”: this is a non redundant merging (where the redundance is defined as the accession number identity) of the current release and the cumulative update file, from one of the three centers.

1.4 Structure of the entries

To each sequence corresponds an entry in the database. Each entry is composed of the sequence itself, as well as information enabling one to identify it and to associate it with biological features (and which make up the annotations). Whatever the computational choices made by each of the centers in terms of data storage technology and data structure (aspects that will not be addressed here even if very important with respect to the running and future of these resources), the three centers provide one with flat files (text files, in a human language), each gathering a great number of entries. Those are individualized using separators (“//”) and structured according to the database’s own format (there also exists a common format used for the data exchanges between the three centers). These formats are fully described in the *release notes* which come with each distribution and are available on the three databases WEB sites ([20], [21], [22]). The respect of strict format rules is required for enabling computer programs to be written that extract (and process) part of the stored information according to various criteria.

The sequence is separated from the annotations, among which one may distinguish two distinct parts:

- The format of the first part is not very constrained and may vary from one database to another. One finds there, among other information, the creation (and possibly modification) date, the accession number (it is unique and shared by the three databases), bibliographical references, the biological origin of the sequenced molecule, as well as possibly some other identifiers: in addition to the accession number, Genbank associates each sequence to one or several other numbers called “gi” (*geninfo identifier*). When a sequence is modified, the new version is given a new “gi” number *gi number* », while the accession number remains the same. Since February 1999, the international collaboration between the three centers has decided to follow a common and more

consistent terminology for describing these modifications: in addition to the accession number, each sequence is given a version number of the form *ac.version* (where “*ac*” is the accession number), beginning with *ac.1* and increasing *version* by one if a modification in the sequence arises (the accession number remaining stable). Therefore, the Genbank’s “gi” numbers are going to disappear (as did the “nid” and “pid” that were linked to it), even if there are still kept at the moment, mainly for compatibility reasons with computer programs using them or with other databases referring to them ([9], [23]).

- The second part concerns the properties or features (“Features” lines for GenBank and DDBJ, “FT” for EMBL) associated with part or all of the sequence. One of the important aspects of the collaboration between the three databases is the agreement on common rules and terminology with respect to the format of this part of the annotations. The *Feature Table Definition* that describes this format with details is available at the EBI WEB site: [24]. This part is structured into subparts, corresponding to the different features of the sequence. The type of a feature is indicated with a keyword chosen among a controlled vocabulary. Some examples of keys are “source” (all the information with respect to the molecular origin of the sequence) , or “CDS” (“coding sequence”). For a given feature are also indicated its location (the positions in the sequence of the sub-sequence to which the feature applies) as well as some qualifiers. There exists a list of possible qualifiers for each key. For a CDS feature, examples of qualifiers are “translation” (the amino acids sequence, inferred from the nucleotide sequence), “db_xref” (indicating cross references with other databases like protein sequence databases), or “protein_id” (followed by an identification number for the virtual corresponding protein). The last one is of the form “ac.version” and is aimed at replacing Genbank’s “pid”.

1.5 Data organization

Each database organizes the entries by clustering them into different sections. Most of these sections reflect a division according to taxonomic criteria, while others (more recent) correspond to special kinds of data, like the EST (“Expressed Sequence Tags”), STS (“Sequence tagged Sites”), GSS (“Genome Survey Sequences”), or HTG (“High Throughput Genomic Sequences”) sections ([25]). Each of the three databases has its own section organization which is documented in the “releases notes”. While overall their organizations are

very similar (identical for the most recent sections), there are a few differences. For example, in the case of mammalian sequences, while Genbank distinguishes the primates from the rodents (two distinct sections), EMBL gives the human species its own section. This organization allows restricting a search to only a subset of the sections.

1.6 The exhaustivity and its consequences

The Genbank/EMBL/DDBJ databases thus make up a comprehensive repository of all the primary nucleic acid sequences publicly known at the moment. It is important that these public primary data resources exist and that they aim at being exhaustive with respect to the sequences. Nevertheless, being comprehensive, particularly when dealing with primary data, has some corollaries that are difficult to avoid:

- Redundancy: the mission of these databases requires the integration of all new primary sequences that are produced, even if they are already present in the database (same or different molecular species, but identical sequence). This internal redundancy is neither eliminated nor even identified or documented, and it has many bad consequences when searching information in the database.
- Heterogeneity: this concerns several aspects of the information present in the databases:
 - the kind of molecule that has been sequenced.
 - the quality of the sequence itself, which is different depending on the kind of project the data come from.
 - the nature and the quality of the annotations. Except for the identifying numbers and for possible cross-links with other databases, all the information that comes with a sequence, as well as its updates are under the sole responsibility of the sequence's authors. The amount of known information is not the same in the case of a gene whose product(s) or function(s) have been studied by many biologists prior to its sequencing and in the case of a systematic genome sequencing. Thus, the amount and quality of these informations are very heterogeneous. Moreover, in no case has this information to be submitted to critical peer-review prior to public release through the databases.

These comprehensive and public resources, to which the other databases and the literature refer, are primordial, but their redundancy and heterogeneity make the existence of other resources necessary. This is why many specialized databases have been developed, dedicated to one organism or to one particular biological problem or topic, etc. The sequence itself may just be one type of data among others in some of these databases, which focus on other aspects of the biological information (for example, the Ecocyc database is focused on *E. coli* metabolism: [26]). Otherwise, while some of them are built around primary data, others contain derived data (obtained by computational sequence analysis of primary sequences).

Among all these databases, two categories are of general interest in the genomic context and will be briefly described in the two next paragraphs.

1.7 Organism specific databases

Presently, there exists at least one database for each sequenced (or currently under sequencing) organism. One of their goals is to provide one with an integration of the current knowledge about one organism, centered on the sequence. They are generally maintained by multidisciplinary teams, including database developers and biologists who are experts on the organism. In some cases, they also aim at enabling scientists who are not the authors of the sequence to take part in its annotation (this is not possible in the case of the primary DNA databases where only the sequence authors are involved in the annotations). Therefore, in general one may expect to find there homogeneous, non redundant, better documented and updated information, compared to that offered by the primary resources, to which the specialized databases refer. Many of them are presented as an article in the 1st of January issue of *Nucleic Acids Research* each year. The *Genome Research* and *Bioinformatics* journals also often publish articles about specialized biological databases.

1.8 The EST databases

EST make up a special category of nucleic acid sequences. They are short, one-pass sequences obtained from large scale cDNA sequencing projects where only the extremities of the clones are sequenced [13]. They are an essential resource for biologists, used for *in silico* gene cloning, and represent the main way of inferring exons boundaries in higher eucaryotic genomes (the programs that aim at predicting gene structures in these genomes perform poorly in terms of sensitivity, selectivity, and accuracy, due to the ambiguity

of our definitions of these structures, notably the splice sites. Since EST correspond to expressed sequences, they can be used to map genes on genomic DNA).

Today EST correspond to one fourth of the Genbank/EMBL/DDBJ nucleotides (only recently outnumbered by the HTG), and more than half of the entries (they are short sequences which explains this difference). The NCBI and the EBI both produce an EST-specific database. There also exist several databases that are specialized in the processing and analysis of EST data, which offer different levels of analysis and interpretation of this information (from a simple classification of the sequences up to an assembly of similar sequences enabling one to reconstruct exonic sequences, and even to discover alternative splicing patterns): UniGene ([27], [28]) developed at the NCBI, TGI ([29], [30]), developed at TIGR (*The Institute for Genomics Research*), STACK ([31], [32]), developed at the SANBI (*South-African National Bioinformatics Institute*), and EGI, the most recent, developed at the EBI ([33], [34]). Two recent reviews describe and compare these resources ([35], [36]).

2 Protein sequence databases

2.1 The data

Almost all the protein sequences currently filling the databases are conceptual translations inferred from experimentally determined DNA sequences. When the latter correspond to experimentally isolated and cloned genes, the transcription and translation start sites (as well as the splice sites) may have also been determined. In this case, it may be possible to define the coding sequence and its precise boundaries. The deducted protein sequence is then likely to correspond to an actual biological molecule, at least in some experimental conditions. It is very different in the case of the huge amounts of genomic sequences which interpretation and exploitation first require the identification of putative coding sequences. Prediction methods exist, grounded on the statistical biases imposed by the coding constraints for identifying “coding contents”, searching signals (like initiation and termination codons or ribosome binding sites), and implemented in computer programs. There are also validation methods, exploiting the sequence similarity with known proteins or with EST. The high gene density and the absence of introns in procaryotic genomes make these predictions easier and more accurate than in the case of eucaryotic genomes, but in any case they are only hypotheses, waiting for experimental testing.

Nevertheless, these are the sequences that currently constitute most of the protein sequence databases. Therefore a very important criteria for the choice of a protein database should be the care given to the annotations, indicating which methods and criteria have been used to infer the CDS features and the functional properties of the putative proteins (deduced from their sequences). The other important criteria for evaluating the relevance of the information one can get from a database search are its exhaustivity and its internal redundancy.

2.2 Several categories of protein sequence databases

At least historically, the motivation that led to the development of general protein sequence databases was not only (as in the case of nucleic acid databases) to establish a comprehensive collection of available sequences, but also to put together a knowledge resource focusing on proteins and their biological properties. The quality criteria they aim at applying are notably the non-redundancy, the homogeneity of information, and the scientific value of the annotations. In addition to collecting, storing and releasing the data, the authors perform a huge curating, sorting, and documentation work: the redundancy is defined, identified and eliminated, and the annotations (informations such as the cellular localization, post-translational modifications, functional features, or references to other databases) that are either extracted from the literature or stemmed from computational sequence analysis are added by the database authors.

This work, which produces the “added value” compared to primary sequence databases, is mostly human and much more time-consuming than the sequence production itself. At the moment, the choice of a protein sequence database inevitably involves a trade-off between quality and exhaustivity: a database like SwissProt which is renowned for the quality of its information is not complete (it does not contain all the conceptual translations of all the CDS that are present in the DNA databases) and it is now supplemented by TrEMBL (see below, paragraph 2.4). In front of the massive production of new genomic sequences, new protein sequence databases are born (Genpept, NCBI-nrprot, OWL) which display exhaustivity as a goal, and for reaching it, sacrifice almost all the other quality criteria mentioned above.

2.3 PIR

The first efforts of gathering and archiving protein sequences had been initiated at the NBRF (*National Biomedical Research Foundation*) by Margaret

Dayhoff in the beginning of the sixties. She created the first protein sequence collection, which was not computerized, but published in the *Atlas of Protein Sequence and Structure* ([37]) and made up of sequences that were obtained by direct protein sequencing. The purpose behind its creation was to study the evolutionary relationships between sequences that were classified into families (this was from these data that she derived the PAM matrices, used in sequence comparison for quantifying the similarities between amino acids [38]). This collection was the predecessor of the PIR database which is developed and maintained by an international collaboration composed of the *Protein Information Resource* at the NBRF, the MIPS (*Martinsried Institute for Protein Sequences*), and the JIPID (*International Protein Information Database of Japan*) ([39]). Following the Dayhoff's tradition, the sequences are classified by families (the process leading to this classification is now mainly automated, based on sequence comparison...). PIR is organized in four sections depending on the information quality and the degree to which the data have been processed. The annotations are generally poor ([40]). The NBRF also develops the Nrl3D database which only contains sequences from proteins whose structure is represented in the PDB (*Protein DataBank*) structure database [41]).

2.4 SwissProt and TrEMBL

2.4.1 SwissProt

SwissProt was created in 1986 by Amos Bairoch at the University of Geneva. It is currently developed as a collaboration between the Swiss Institute of Bioinformatics (ISB-SIB, [42]) and the EBI ([7]). It is certainly the less redundant, better annotated, and most integrated (through cross-references) with other databases. The sequences are either directly submitted to SwissProt by their authors, or (in majority now) extracted from the CDS sequences from the EMBL database. The integration process of a new sequence into SwissProt consists of several steps (detailed below) that aim at checking its relevance, reducing the database internal redundancy (therefore a SwissProt entry may contain several sequences, identical or "almost": see paragraph 2.4.3), validating and enriching the associated biological information. Notably, great care is taken in indicating the experimental or computational nature of the functional information, as well as the confidence level one attaches to them ([43]).

Until the middle of the nineties, the small Bairoch's group in collaboration with the EMBL team was able to follow the growth rate of EMBL, ap-

plying this (essentially human) expertise to all the newly sequenced “genes”. In 1995-96, with the arrival of the first complete genomes (*Haemophilus Influenzae* by the TIGR [15], *Saccharomyces cerevisiae* [16]), came a new problem: the rate at which new putative CDS were entering EMBL became incompatible with the amount of work that was required in order to fulfill SwissProt’s quality criteria, given the human and financial resources of the group. The dilemma was either to respect these criteria, but to be only partially representative of the known coding sequence set, or to propose a comprehensive database, but at the cost of a decrease and an heterogeneity in the information’s quality. The solution they chose was to create the TrEMBL database (*TR*anslation of *EMBL*) as a supplement to SwissProt, and to develop informatics tools ([44]) in order to automate some of the steps of the path leading a putative coding sequence from EMBL, through TrEMBL, to SwissProt ([45]).

2.4.2 TrEMBL

TrEMBL, which is built at the EBI, is made up of the protein counterparts to all the CDS of EMBL, which are not already present in Swissprot (based on the existence of a link to SwissProt in the EMBL entry). The search for CDS features in the EMBL annotations and the extraction of the corresponding protein sequences are performed by a program that also automatically extracts some information from the EMBL annotations. All the EMBL CDS leading to a correct translation give rise to TrEMBL entries, written in a format very similar to the one of SwissProt entries. They are given an identifier that corresponds to the *protein_id* of EMBL entries. Then the new entries are sorted and pre-annotated, notably:

- all the CDS which EMBL annotation contains a cross-reference to either Swissprot or TrEMBL are removed, since it means that the sequences are already present in the protein databases (nucleic acid databases are highly redundant). A first level of redundancy is thus eliminated, and TrEMBL is defined as the set of all the known protein sequences that are not already present into SwissProt.
- TrEMBL is split into two parts, Rem-TrEMBL and Sp-TrEMBL. All the CDS belonging to the following categories enter Rem-TrEMBL and will not be concerned by the next steps leading towards SwissProt: immunoglobulin, T cell receptor and MHC sequences, which diversity is due to somatic recombination and which are over-represented in the DNA databases (they are the focus of specialized databases

like IMGT [46]); sequences coming from patents; sequences less than eight amino acids long; those corresponding to artificial, synthetic or chimeric genes (with no known biological reality); those corresponding to pseudo-genes. These sequences make up Rem-TrEMBL and are not meant to enter SwissProt. All the remaining sequences make up Sp-TrEMBL, and will then be subjected to the long (here is the bottleneck...) integration process into SwissProt.

2.4.3 From Sp-TrEMBL to SwissProt

Reduction in the redundancy. First, the internal redundancy of Sp-TrEMBL is examined: all the entries corresponding to identical sequences, as well as all the fragments belonging to the same protein in the same species, are merged into one common Sp-TrEMBL entry. Next all the Sp-TrEMBL sequences are compared to those of SwissProt: when two sequences are identical, the one from TrEMBL is integrated into the corresponding SwissProt entry, thus minimizing the redundancy introduced into SwissProt; the entries which contain fragments of sequences from other entries or polymorphic variations of the same protein, are also merged ([47]).

All the steps of constructing TrEMBL entries, splitting into Rem- and Sp-TrEMBL, and eliminating the main redundancy, are now automated.

Increasing the information. The purpose here is to enrich the annotation of the Sp-TrEMBL entries that have been selected through the previous steps. Some information that was present in the EMBL entry is automatically extracted from there (for example, cross-references to other databases that give information regarding the correct terminology of the gene, or the accession number in a database like *Enzyme* [48]), these information being then recorded in the Sp-TrEMBL entry. Furthermore, in order to generate information from the sequence's analysis, an automatic annotation system has been developed. It enables one to call several programs performing different analysis, to chain the analysis in a selective and dynamic manner (it is possible to specify some conditions governing which program has to be called and when, as a function of the sequence features or of the results given by other programs), and to generate annotations inferred from the analysis results ([49]). As an example, the sequences are compared to the

Prosite database of protein motifs ([50]) and rules for automatic pre-annotation from the result of this search have been implemented ([51]).

At that stage, the pre-annotated entries are still not SwissProt entries. They are Sp-TrEMBL entries (with a SpTrEMBL accession number which become their SwissProt accession number), ready to be examined by human experts, the swissprot annotators, who will judge the biological relevance of the automatic annotations, and depending on the cases, will modify, complete or invalidate them. During this last step (the longer one), biology researchers from around the world also participate. The ultimate validation of a new entry (or an old one to which new sequences from Sp-TrEMBL have been added) is made by one responsible scientist from the SwissProt team (in principle, they all are validated by Amos Bairoch). Then, it is never finished since any SwissProt entry may be modified or updated...

Thus at one moment one can find two kinds of entries in Sp-TrEMBL:

- those which have followed the automated annotation process, but which have not yet been examined by human beings. They contain two parts: one which is visible by everybody (the “core” data; that is the sequence, the accession number, the information directly extracted from the EMBL entries), and the other one which is hidden (the results of the automatic pre-annotations) except to the annotators.
- those which have been processed by the human annotators and are thus very similar to SwissProt entries, but have not yet been granted the ultimate validation.

2.4.4 SwissProt + TrEMBL + TrEMBL_new = Sptrnrdb

What does it happen between two releases of SwissProt and TrEMBL? Ideally, immediately after a new SwissProt distribution, Sp-TrEMBL should be empty (if all the sequences were integrated into SwissProt). However, it is far from being the case, and many sequences are transiently stored into Sp-TrEMBL. For example, at the end of 2000, SwissProt contained around 91000 sequences while there were 371000 stored in Sp-TrEMBL. Six months ago, these numbers were respectively 85000 and less than 30000... If one considers the situation at a given time, one has got three databases: SwissProt, TrEMBL and TrEMBL_new. Just after a TrEMBL release, TrEMBL_new, which contains the translations of the new EMBL CDS, is empty, since its data have just been incorporated into the new Rem- and Sp-TrEMBL releases. Between two TrEMBL releases, TrEMBL_new gets bigger, while

Sp-TrEMBL gets smaller, some of its entries being integrated into SwissProt. It gets bigger again at the next release when TrEMBL_new contents are added to it and to Rem-TrEMBL. Swissprot gets only bigger and bigger.

The collaboration between the SIB and the EBI proposes a complete and - almost - non redundant protein sequence database to be updated weekly, called Sptrnrdb, standing for *SwissProt*, *TREMBL*, *Non-Redundant DataBase* ([52]), and composed of three files:

- sprot.dat: SwissProt, last released version plus the updates since this last distribution.
- trembl.dat: Sp-TrEMBL, last released version minus the entries that have been integrated into SwissProt since then.
- trembl_new.dat: all the new TrEMBL entries since the last release (that is an update file).

SwissProt is renowned as being the protein sequence database that is the best curated, annotated, and integrated to other sequence databases. Without sacrificing the quality which makes SwissProt valuable, the creation of TrEMBL as its supplement and the gathering of the weekly versions of SwissProt and Sp-TrEMBL give rise to a database that is complete and as least redundant as possible.

2.5 The other comprehensive databases

There are other databases that aim at providing one with a complete collection of all the known protein sequences:

- Genpept, which is produced at the *Frederick Biomedical Supercomputing Center* is built by simply extracting all the CDS features in the Genbank annotations ([53]). This is the same as what is performed at the first step of building TrEMBL, and therefore, since Genbank and EMBL contents are equivalent, those of Genpept and “raw TrEMBL” (before the next curating steps) are equivalent too. In the case of Genpept, no sorting, no reduction of redundancy nor annotation work is performed. Thus it is exactly the translated version of Genbank coding sequences and like Genbank, it is complete, but highly redundant and poorly annotated (the annotations are just the result of automatic extractions from those of Genbank).
- NRDB, also called nrprot ([54]) and produced at the NCBI, and OWL ([55]) are composite databases made up by gathering several databases

(such as SwissProt, Nrl-3D, PIR, Genpept) and keeping only one exemplar of exactly identical sequences. OWL establishes a hierarchy among its source databases, giving the highest priority to entries coming from SwissProt. With the NCBI nrdb/nrprot database, no priority is considered and the result is something probably comprehensive, but redundant and very heterogeneous.

2.6 Derived databases

Because the proteins of the thousands of organisms that are represented in the databases may be clustered in evolutionary families, because the functional properties of a protein may reside in some short motifs inside the sequence, because they may be made up with domains behaving like independent (structural, functional, evolutionary...) modules, for all these reasons and others, databases of families, alignments, motifs, patterns, domains, etc, have been developed. A good review about these derived or secondary databases is given in [56]. Some of them will be presented during the course about "Multiple alignments and motifs".

3 Conclusions and future prospects

At the beginning of year 2001, the Genbank/EMBL/DDBJ nucleic acid sequence database contains more than 11 billions of nucleotides (it was less than 5 at the beginning of last year) and no decrease in the mondial production of sequences is expected in the near future. Apart from the informatics aspects concerning the efficient management of such amounts of data, many questions arise concerning the future development of these resources. These questions relate to scientific as well as epistemologic, political or economical issues...

The quality of a sequence database depends on criteria such as the scientific value of its annotations, the wealth of its references, the absence of internal redundancy, and the homogeneity of the information it contains. Obviously, homogeneity and non-redundancy are difficult if not impossible to reconcile with exhaustivity (as illustrated by the current state of the big comprehensive databases). The solution that seems to be adopted for the short-term future is the co-existence of complete repositories that aim at being comprehensive but from which it would not be wise to expect more, with some "secondary" databases (like Swissprot and TrEMBL) and many derived or specialized databases. The latter, because of their size and their restricted focus on an organism or a particular topic have got good chances

of reaching high quality levels. Such a solution nonetheless raises at least three problems:

- the quality level is under the sole responsibility and scientific demand of the database authors. Indeed in the case of specialized databases as well as in the case of the big comprehensive repositories, the information that is publicly released by that way is not submitted for review and validation to peers. The annotation or update work made by scientists is not renowned and valued like classical scientific publications are. The marginal position of the scientific activities involved in biological database development is one of the reasons for the difficulty they encounter in striving to be high quality resources ([57]).
- the development of all these resources is driven by needs and motivations but without the status of the databases and their authors being clearly defined. Therefore, there is much fuzziness regarding the scientific fields and institutions to which the scientists who develop them should belong or refer (biology *versus* computer science, pure scientific research *versus* technological development). This leads to (among other consequences) the fact that funding is a problem([58]). One striking example is the case of SwissProt, which was at the point of disappearing in 1996, when their short-term funding sources stopped ([59]). The solution that was chosen was to make industry pay for SwissProt, and to re-invest one part of the money earned that way in the running of SwissProt ([60]). The *Geneva BioInformatics* ([61]) company was created in 1997 in order to deal with the commercial and financial aspects of SwissProt (and other databases developed at the SIB) in addition to its own development activities. The uncertainty of most databases funding sources make their future relationship with public research questionable.
- Another, more scientific, aspect of having comprehensive primary data resources on one side and many secondary, specialized databases on the other side, is the problem of their communication and information exchange ([62], [5], [63]). From the viewpoint of the data, this interoperability requires the adoption of common languages for describing the information they contain, that is some common definitions or ontologies for specifying the biological objects and concepts they deal with ([64]). Other concerns are the informatics aspects of the question and the use of modern information technologies such as CORBA ([65],

[66]). In both cases, the biological, semantic side, and the technological, computer science side, these are active research fields.

The model mentioned above relies on the existence of common, primary, comprehensive, public resources. What will happen in the future? Will Genbank, EMBL and DDBJ remain public and continue to collaborate? One condition for that is that they be funded by public money. While in the US steps have been taken in this direction ([67]), the situation was until recently more uncertain in Europe: the EBI which develops many high quality resources had to face serious funding problems ([68], [69], [70], [71])

Finally, do not forget that besides the sequence data, other data types coming from “functional genomics” (expression, molecular interaction data...) are beginning to be massively produced also, and the same questions occur with these new data and databases ([72]).

References

- [1] William M. Gelbart. Databases in genomic research. *Science*, 282:659–661, 1998.
- [2] Mark S. Boguski. Bioinformatics - a new era. *Trends Guide in Bioinformatics*, pages 1–3, 1998.
- [3] Mark S. Boguski. Biosequence exegesis. *Science*, 286:453–455, 1999.
- [4] Andreas D. Baxevanis and B. F. Francis Ouellette, editors. *Bioinformatics - A Practical Guide to the Analysis of Genes and Proteins*, volume 39 of *Methods in Biochemical Analysis*, chapter 2, pages 16–45. John Wiley & Sons, 1998.
- [5] Patricia G. Baker and Andy Bass. Recent developments in biological databases. *Current Opinion in Biotechnology*, 9:54–58, 1998.
- [6] Guenter Stoesser, Wendy Baker, Alexandra van den Broek, Evelyn Camon, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Vincent Lombard, Rodrigo Lopez, Helen Parkinson, Nicole Redaschi, Peter Sterk, Peter Stoehr, and Mary Ann Tuli. The embl nucleotide sequence database. *Nucleic Acids Res.*, 29:17–21, 2001.
- [7] <http://www.ebi.ac.uk>.
- [8] <http://www.ncbi.nlm.nih.gov>.

- [9] Dennis A. Benson, Ilene Karsch-Mizrachi, James Ostell David J. Lipman, Barbara A. Rapp, and David L. Wheeler. Genbank. *Nucleic Acids Res.*, 28:15–18, 2000.
- [10] Yoshio Tateno, Satoru Miyazaki, Motonori Ota, Hideaki Sugawara, and Takashi Gojobori. Dna data bank of japan (ddbj) in collaboration with mass sequencing teams. *Nucleic Acids Res.*, 28:24–26, 2000.
- [11] <http://www.ddbj.nig.ac.jp/>.
- [12] <http://www.ncbi.nlm.nih.gov/collab/>.
- [13] Mark D. Adams, Jenny M. Kelley, Jeannine D. Gocayne, Mark Dubnick, Mihael H. Polymeropoulos, Hong Xiao, Carl R Merrill, Andrew Wu, Bjorn Olde, Ruben F. Moreno, Anthony R. Kerlavage, W. Richard McCombie, and J. Craig Venter. Complementary dna sequencing: Expressed sequence tags and human genome project. *Science*, 252:1651–1656, 1991.
- [14] Steve G. Oliver, Q.J. Van der Aart, M.L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J.P. Ballesta, P. Benit, and et al. The complete dna sequence of yeast chromosome iii. *Nature*, 357:38–46, 1992.
- [15] R. D. Fleischmann, Mark D. Adams, Owen White, R.A. Clayton, E.F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-François Tomb, B.A. Dougherty, J.M. Merrick, and et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269:496–512, 1995.
- [16] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274:546–547, 1996.
- [17] <http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html>.
- [18] <http://www3.ebi.ac.uk/Services/DBStats/>.
- [19] <http://www.ebi.ac.uk/~sterk/genome-MOT/>.
- [20] <http://www.ebi.ac.uk/embl/Documentation/>.
- [21] <http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>.

- [22] <http://www.ddbj.nig.ac.jp/fromddbj-e.html>.
- [23] <http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html>.
- [24] http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html.
- [25] Francis B.F. Ouellette and Mark S. Boguski. Database divisions and homology search files: a guide for the perplexed. *Genome Research*, 7:952–955, 1997.
- [26] Peter D. Karp, Monica Riley, Milton Saier, Ian T. Paulsen, Suzanne M. Paley, and Alida Pellegrini-Toole. The ecocyc and metacyc databases. *Nucleic Acids Res.*, 28:56–59, 2000.
- [27] MS Boguski and GD Schuler. Establishing a human transcript map. *Nat. Genet.*, 10:369–71, 1995.
- [28] <http://www.ncbi.nlm.nih.gov/UniGene/index.html>.
- [29] John Quackenbush, Feng Liang, Ingeborg Holt, Geo Pertea, and Jonathan Upton. The tigr gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.*, 28:141–145, 2000.
- [30] <http://www.tigr.org/tdb/tgi.shtml>.
- [31] R. T. Miller, AG. Christoffels, C. Gopalakrishnan, J. A. Burke, A. A. Ptitsyn, T. R. Broveak, and W. A. Hide. A comprehensive approach to clustering of expressed human gene sequence: The sequence tag alignment and consensus knowledgebase. *Genome Res.*, 28:141–145, 1999.
- [32] <http://www.sanbi.ac.za/Dbases.html>.
- [33] J.D. Parsons and Patricia Rodriguez-Tome. Jesam: Corba software components to create and publish est alignments and clusters. *Bioinformatics*, 16(4):313–325, 2000.
- [34] <http://corba.ebi.ac.uk/EST/egi.html>.
- [35] John Bouck, Wei Yu, Richard Gibbs, and Kim Worley. Comparison of gene indexing databases. *Trends in Genetics*, 15(4):159–162, 1999.
- [36] C. Victor Jongeneel. Searching the expressed sequence tag (est) databases. *Briefings in Bioinformatics*, 1(1):76–92, 2000.

- [37] M.O. Dayhoff, R.V. Eck, M.A. Chang, and M.R. Sochard. *Atlas of Protein Sequence and Structure*. Natl. Biomed. Res. Fnd., Silver Spring MD, 1965.
- [38] M.O. Dayhoff and R.V. Eck. A model of evolutionary change in proteins. In Natl. Biomed. Res. Fnd., editor, *Atlas of Protein Sequence and Structure*, chapter 4, pages 33–41. Silver Spring MD, 1967-68.
- [39] Winona C. Barker, John S. Garavelli, Hongzhan Huang, Peter B. McGarvey, Bruce C. Orcutt, Geetha Y. Srinivasarao, Chunlin Xiao, Lai-Su L. Yeh, Robert S. Ledley, Joseph F. Janda, Friedhelm Pfeiffer, Hans-Werner Mewes, Akira Tsugita, and Cathy Wu. The protein information resource (pir). *Nucleic Acids Res.*, 28:41–44, 2000.
- [40] Teresa K. Attwood. Protein information resources. In Teresa K. Attwood and David J. Parry-Smith, editors, *Introduction to Bioinformatics*, chapter 3, pages 35–67. Longman, 1999.
- [41] Helen M. Berman, John Westbrookd, Zukang Feng, T. N. Bhat Gary Gilliland, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.
- [42] <http://www.isb-sib.ch/>.
- [43] Vivien Junker, Rolf Apweiler, and Amos Bairoch. Representation of functional information in the swiss-prot data bank. *Bioinformatics*, 15(12):1066–1067, 1999.
- [44] Rolf Apweiler, Alain Gateau, Sergio Contrino, Maria Jesus Martin, Vivien Junker, Claire O’Donovan, Fiona Lang, Nicoletta Mitaritonna, Stephanie Kappus, and Amos Bairoch. Protein sequence annotation in the genome era: the annotation concept of swiss-prot + trembl. In *ISMB-97; Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, pages 33–43. AAAI Press, Menlo Park, 1997.
- [45] Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res.*, 28:45–48, 2000.
- [46] Manuel Ruiz, Véronique Giudicelli, Chantal Ginestoux, Peter Stoehr, James Robinson, Julia Bodmer, Steven G. E. Marsh, Ronald Bontrop, Marc Lemaitre, Gérard Lefranc, Denys Chaume, and Marie-Paule

- Lefranc. Imgt, the international immunogenetics database. *Nucleic Acids Res.*, 28:219–221, 2000.
- [47] Claire O’Donovan, Maria Jesus Martin, Eric Glemet, Jean-Jacques Codani, and Rolf Apweiler. Removing redundancy in swiss-prot and trembl. *Bioinformatics*, 15(3):258–259, 1999.
- [48] Amos Bairoch. The enzyme database in 2000. *Nucleic Acids Res.*, 28:304–305, 2000.
- [49] Steffen Moller, Ulf Leser, Wolfgang Fleischmann, and Rolf Apweiler. Edittotrembl: a distributed approach to high-quality automated protein sequence annotation. *Bioinformatics*, 15:219–227, 1999.
- [50] Kay Hofmann, Philip Bucher, L. Falquet, and Amos Bairoch. The prosite database, its status in 1999. *Nucleic Acids Res.*, 27:215–219, 1999.
- [51] Wolfgang Fleischmann, Steffen Moller, Alain Gateau, and Rolf Apweiler. A novel method for automatic functional annotation of proteins. *Bioinformatics*, 15(3):228–233, 1999.
- [52] ftp://ftp.expasy.ch/databases/sp_tr_nrdb/README.
- [53] <ftp://ftp.ncifcrf.gov/pub/genpept/>.
- [54] <ftp://ncbi.nlm.nih.gov/blast/db/README>.
- [55] <http://www.biochem.ucl.ac.uk/bsm/dbbrowser/OWL/OWL.html>.
- [56] Terri K. Attwood. The role of pattern databases in sequence analysis. *Briefings in Bioinformatics*, 1(1):45–59, 2000.
- [57] Elisabeth Pennisi. keeping genome databases clean and up to date. *Science*, 286:447–450, 1999.
- [58] Lynda B.M. Ellis and Doyle Kalumbi. Financing a future for public biological data. *Bioinformatics*, 15(9):717–722, 1999.
- [59] <http://www.expasy.ch/sprot/crisis96/>.
- [60] http://www.expasy.ch/announce/sp_98sum.html.
- [61] <http://www.genebio.com/>.

- [62] Nigel Williams. Bioinformatics: how to get databases talking the same language. *Science*, 275:301–302, 1997.
- [63] Dimitrij Frishman, Klaus Heumann, Arthur Lesk, and Hans werner Mewes. Comprehensive, comprehensible, distributed and intelligent databases: current status. *Bioinformatics*, 14(7):551–561, 1998.
- [64] Elisabeth Pennisi. Seeking common language in a tower of babel. *Science*, 286:448, 1999.
- [65] Robert Stevens and Crispin miller. Wrapping and interoperating bioinformatics resources using corba. *Briefings in Bioinformatics*, 1:9–21, 2000.
- [66] <http://industry.ebi.ac.uk:80/applab/>.
- [67] Sylvia D. spengler. Bioinformatics in the information age. *Science*, 18(287):1221–1223, 2000.
- [68] Declan Butler. Life science facilities in crisis as brussels switches off funding. *Nature*, 402:3, 1999.
- [69] Alison Abbott. Embl rescue package keeps bioinformatics centre running. *Nature*, 402:450, 1999.
- [70] Alison Abbott. European centres rebuffed in infrastructure funding bid. *Nature*, 405:723, 2000.
- [71] Alison Abbott. Europe boosts genome resource centres. *Nature*, 408:393, 2000.
- [72] Alvis Brazma, Alan Robinson, Graham Cameron, and Michael Ashburner. One-stop shop for microarray data. *Nature*, 403:699–700, 2000.