

# Bioinformatics – a definition

Michael Nilges and Jens P. Linge, Unité de Bio-Informatique Structurale, Institut Pasteur, 25–28 rue du Docteur Roux, F–75015 Paris, France

## **Synonyms**

Related: Computational Biology, Computational Molecular Biology, Biocomputing

## **Definition**

Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature. Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine.

## **Description**

The history of computing in biology goes back to the 1920s when scientists were already thinking of establishing biological laws solely from data analysis by induction (e.g. A.J. Lotka, *Elements of Physical Biology*, 1925). However, only the development of powerful computers, and the availability of experimental data that can be readily treated by computation (for example, DNA or amino acid sequences and three-dimensional structures of proteins) launched bioinformatics as an independent field. Today, practical applications of bioinformatics are readily available through the world wide web, and are widely used in biological and medical research. As the field is rapidly evolving, the very definition of bioinformatics is still the matter of some debate.

The relationship between computer science and biology is a natural one for several reasons. First, the phenomenal rate of biological data being produced provides challenges: massive amounts of data have to be stored, analysed, and made accessible. Second, the nature of the data is often such that a statistical method, and hence computation, is necessary. This applies in particular to the information on the building plans of proteins and of the temporal and spatial organisation of their expression in the cell encoded by the DNA. Third, there is a strong analogy between the DNA sequence and a computer program (it can be shown that the DNA represents a Turing Machine).

Analyses in bioinformatics focus on three types of datasets: genome sequences, macromolecular structures, and functional genomics experiments (e.g. expression data, yeast two-hybrid screens). But bioinformatic analysis is also applied to various other data, e.g. taxonomy trees, relationship data from metabolic pathways, the text of scientific papers, and patient statistics. A large range of techniques are used, including primary sequence alignment, protein 3D structure alignment, phylogenetic tree construction, prediction and classification of protein structure, prediction of RNA structure, prediction of protein function, and expression data clustering. Algorithmic development is an important part of bioinformatics, and techniques and algorithms were specifically developed for the analysis of biological data (e.g., the dynamic programming algorithm for sequence alignment).

Bioinformatics has a large impact on biological research. Giant research projects such as the human genome project [4] would be meaningless without the bioinformatics component. The goal of sequencing projects, for example, is not to corroborate or refute a hypothesis, but to provide raw data for later analysis. Once the raw data are available, hypotheses may be formulated and tested *in silico*. In this manner, computer experiments may answer biological questions which cannot be tackled by traditional approaches. This has led to the founding of dedicated bioinformatics research groups as well as to a different work practice in the average bioscience laboratory where the computer has become an essential research tool.

Three key areas are the organisation of knowledge in databases, sequence analysis, and structural bioinformatics.

## Organizing biological knowledge in databases

Biological raw data are stored in public databanks (such as Genbank or EMBL for primary DNA sequences). The data can be submitted and accessed via the world wide web. Protein sequence databanks like trEMBL provide the most likely translation of all coding sequences in the EMBL databank. Sequence data are prominent, but also other data are stored, e. g. yeast two–hybrid screens, expression arrays, systematic gene–knock–out experiments, and metabolic pathways.

The stored data need to be accessed in a meaningful way, and often contents of several databanks or databases have to be accessed simultaneously and correlated with each other. Special languages have been developed to facilitate this task (such as the Sequence Retrieval System (SRS) and the Entrez system). An unsolved problem is the optimal design of inter–operating database systems. Databases provide additional functionality such as access to sequence homology searches and links to other databases and analysis results. For example, SWISSPROT [1] contains verified protein sequences and more annotations describing the function of a protein. Protein 3D structures are stored in specific databases (for example, the Protein Data Bank [2], now primarily curated and developed by the Research Collaboratory for Structural Bioinformatics). Organism specific databases have been developed (such as ACEDB, the A C. Elegans DataBase for the *C. elegans* genome, FLYBASE for *D. melanogaster* etc). A major problem are errors in databanks and databases (mostly errors in annotation), in particular since errors propagate easily through links.

Also databases of scientific literature (such as PUBMED, MEDLINE) provide additional functionality, e.g. they can search for similar articles based on word–usage analysis. Text recognition systems are being developed that extract automatically knowledge about protein function from the abstracts of scientific articles, notably on protein–protein interactions.

## Analysing sequence data

The primary data of sequencing projects are DNA sequences. These become only really valuable through their annotation. Several layers of analysis with bioinformatics tools are necessary to arrive from a raw DNA sequence at an annotated protein sequences:

- establish the correct order of sequence contigs to obtain one continuous sequence;
- find the translation and transcription initiation sites, find promoter sites, define open reading frames (ORF);
- find splice sites, introns, exons;
- translate the DNA sequence into a protein sequence, searching all six frames;
- compare the DNA sequence to known protein sequences in order to verify exons etc with homologous sequences.

Some completely automated annotation systems have been developed (e.g., GENEQUIZ), which use a multitude of different programs and methods.

The protein sequences are further analysed to predict function. The function can often be inferred if a sequence of a homologous protein with known function can be found. Homology searches are the predominant bioinformatics application, and very efficient search methods have been developed [3]. The often difficult distinction between orthologous sequences and paralogous sequences facilitates the functional annotation in the comparison of whole genomes. Several methods detect glycosylation, myristylation and other sites, and the prediction of signal peptides in the amino acid sequence give valuable information about the subcellular location of a protein.

The ultimate goal of sequence annotation is to arrive at a complete functional description of all genes of an organism. However, function is an ill-defined concept. Thus, the simplified idea of “one gene – one protein – one structure – one function” cannot take into account proteins that have multiple functions depending on context (e.g., subcellular location and the presence of cofactors). Well-known cases of “moonlighting” proteins are lens crystalline and phosphoglucose isomerase. Currently, work on ontologies is under way to explicitly define a vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing.

Families of similar sequences contain information on sequence evolution in the form of specific conservation patterns at all sequence positions. Multiple sequence alignments are useful for

- building sequence profiles or Hidden Markov Models to perform more sensitive homology searches. A sequence profile contains information about the variability of every sequence position. Improving structure prediction methods (secondary structure prediction). Sequence profile searches have become readily available through the introduction of PsiBLAST [3];
- studying evolutionary aspects, by the construction of phylogenetic trees from the pairwise differences between sequences: for example, the classification with 70S, 30S RNAs established the separate kingdom of archaea;
- determining active site residues, and residues specific for subfamilies;

- predicting protein–protein interactions;
- analysing single nucleotide polymorphisms to hunt for genetic sources of diseases.

Many complete genomes of microorganisms and a few of eukaryotes are available [4]. By analysis of entire genome sequences a wealth of additional information can be obtained. The complete genomic sequence contains not only all protein sequences but also sequences regulating gene expression. A comparison of the genomes of genetically close organisms reveals genes responsible for specific properties of the organisms (e.g., infectivity). Protein interactions can be predicted from conservation of gene order or operon organisation in different genomes. Also the detection of gene fusion and gene fission (i.e, one protein is split into two in another genome) events helps to deduce protein interactions.

## Structural bioinformatics

This branch of bioinformatics is concerned with computational approaches to predict and analyse the spatial structure of proteins and nucleic acids. Whereas in many cases the primary sequence uniquely specifies the three–dimensional (3D) structure, the specific rules are not well understood, and the protein folding problem remains largely unsolved. Some aspects of protein structure can already be predicted from amino acid content. Secondary structure can be deduced from the primary sequence with statistics or neural networks. When using a multiple sequence alignment, secondary structure can be predicted with an accuracy above 70 %.

3D models can be obtained most easily if the 3D structure of a homologous protein is known (homology modelling, comparative modelling). A homology model can only be as good as the sequence alignment: whereas protein relationships can be detected at the 20% identity level and below, a correct sequence alignment becomes very difficult, and the homology model will be doubtful. From 40 to 50% identity the models are usually mostly correct; however, it is possible to have 50% identity between two carefully designed protein sequences with different topology (the so–called JANUS protein). Remote relationships that are undetectable by sequence comparisons may be detected by sequence–to–structure–fitness (or threading) approaches: the search sequence is systematically compared to all known protein structures. Ab initio predictions of protein 3D structure remains the major challenge; some progress has been made recently by combining statistical with force–field based approaches.

Membrane proteins are interesting drug targets. It is estimated that membrane receptors form 50 % of all drug targets in pharmacological research. However, membrane proteins are underrepresented in the PDB structure database. Since membrane proteins are usually excluded from structural genomics initiatives due to technical problems, the prediction of transmembrane helices and solvent accessibility is very important. Modern methods can predict transmembrane helices with a reliability greater than 70 %.

Understanding the 3D structure of a macromolecule is crucial for understanding its function. Many properties of the 3D structure cannot be deduced directly from the primary sequence. Obtaining better understanding of protein function is the driving force behind structural genomics efforts, which can be thus understood as part of functional genomics. Similar structure can imply similar function. General structure–to–function relationships can be obtained by statistical approaches, for example, by relating secondary structure to known protein function or surface properties to cell location.

The increased speed of structure determination necessary for the structural genomics projects make an independent validation of the structures (by comparison to expected properties) particularly important. Structure validation helps to correct obvious errors (e.g., in the covalent structure) and leads to a more standardized representation of structural data, e.g., by agreeing on a common atom name nomenclature. The knowledge of the structure quality is a prerequisite for further use of the structure, e.g. in molecular modelling or drug design.

In order to make as much data on the structure and its determination available in the databases, approaches for automated data harvesting are being developed. Structure classification schemes, as implemented for example in the SCOP, CATH, and FSSP databases, elucidate the relationship between protein folds and function and shed light on the evolution of protein domains.

Combined analysis of structural and genomic data will certainly get more important in the near future. Protein folds can be analysed for whole genomes. Protein–protein interactions predicted on the sequence level, can be studied in more detail on the structure level. Single Nucleotide Polymorphisms can be mapped on 3D structures of proteins in order to elucidate specific structural causes of disease.

More detailed aspects of protein function can be obtained also by force–field based approaches. Whereas protein function requires protein dynamics, no experimental technique can observe it directly on an atomic scale, and motions have to be simulated by molecular dynamics (MD) simulations. Also free energy differences (for example between binding energies of different protein ligands) can be characterized by MD simulations. Molecular mechanics or molecular dynamics based approaches are also necessary for homology modelling and for structure refinement in X–ray crystallography and NMR structure determination.

Drug design exploits the knowledge of the 3D structure of the binding site (or the structure of the complex with a ligand) to construct potential drugs, for example inhibitors of viral proteins or RNA. In addition to the 3D structure, a force field is necessary to evaluate the interaction between the protein and a ligand (to predict binding energies). In virtual screening, a library of molecules is tested on the computer for their capacities to bind to the macromolecule.

## **Pharmacological Relevance**

Many aspects of bioinformatics are relevant for pharmacology. Drug targets in infectious organisms can be revealed by whole genome comparisons of infectious and non–infectious organisms. The analysis of single nucleotide polymorphisms reveals genes potentially responsible for genetic diseases. Prediction and analysis of protein 3D structure is used to develop drugs and understand drug resistance.

Patient databases with genetic profiles, e.g. for cardiovascular diseases, diabetes, cancer, etc. may play an important role in the future for individual health care, by integrating personal genetic profile into diagnosis, despite obvious ethical problems. The goal is to analyse a patient's individual genetic profile and compare it with a collection of reference profiles and other related information. This may improve individual diagnosis, prophylaxis, and therapy.

## **References**

1. Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28:45–48
  2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res.* 28:235–42
  3. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402
- Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132:185–219
4. The Genome International Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- JC Venter et al. (2001) The sequence of the human genome. *Science* 291:1304–1351
- R.D. Fleischmann et al. (1995) Whole-genome random sequencing and assembly of *haemophilus-influenzae*. *Science* 269:496–512

## **Glossary**

### **Keyword: Analogous Proteins**

#### **Definition:**

Two proteins with related folds but unrelated sequences are called analogous. During evolution, analogous proteins independently developed the same fold.

### **Keyword: Databank**

#### **Definition:**

In the biosciences, a databank (or data bank) is a structured set of raw data, most notably DNA sequences from sequencing projects (e.g. the EMBL and GenBank databases).

### **Keyword: Database**

#### **Definition:**

A database (or data base) is a collection of data that is organized so that its contents can easily be accessed, managed, and modified by a computer. The most prevalent type of database is the relational database which organizes the data in tables; multiple relations can be mathematically defined between the rows and columns of each table to yield the desired information. An object-oriented database stores data in the form of objects which are organized in hierarchical classes that may inherit properties from classes higher in the tree structure.

In the biosciences, a database is a curated repository of raw data containing annotations, further analysis, links to other databases. Examples of databases are the SWISSPROT database for annotated protein sequences or the FlyBase database of genetic and molecular data for *Drosophila melanogaster*.

### **Keyword: Dynamic Programming**

#### **Definition:**

In general, dynamic programming is an algorithmic scheme for solving discrete optimization problems that have overlapping subproblems. In a dynamic programming algorithm, the definition of the function that is optimized is extended as the computation proceeds. The solution is constructed by progressing from simpler to more complex cases, thereby solving each subproblem before it is needed by any other subproblem.

In particular, the algorithm for finding optimal alignments is an example of dynamic programming.

**Keyword: Force-field****Definition:**

In molecular dynamics and molecular mechanics calculations, the intra- and intermolecular interactions of a molecule are calculated from a simplified empirical parametrization called a force field. These include atom masses, charges, dihedral angles, improper angles, van-der-Waals and electrostatic interactions, etc.

**Keyword: Genome****Definition:**

The genome is the gene complement of an organism. A genome sequence comprises the information of the entire genetic material of an organism.

**Keyword: Genomics, Functional Genomics, Structural Genomics****Definition:**

The goal of Genomics is to determine the complete DNA sequence for all the genetic material contained in an organism's complete genome.

Functional genomics (sometimes referred to as functional proteomics) aims at determining the function of the proteome (the protein complement encoded by an organism's entire genome). It expands the scope of biological investigation from studying single genes or proteins to studying all genes or proteins at once in a systematic fashion, using large-scale experimental methodologies combined with statistical analysis of the results.

Structural Genomics is the systematic effort to gain a complete structural description of a defined set of molecules, ultimately for an organism's entire proteome. Structural genomics projects apply X-ray crystallography and NMR spectroscopy in a high-throughput manner.

**Keyword: Hidden Markov Model****Definition:**

A Hidden Markov Model (HMM) is a general probabilistic model for sequences of symbols. In a Markov chain, the probability of each symbol depends only on the preceding one. Hidden Markov models are widely used in bioinformatics, most notably to replace sequence profile in the calculation of sequence alignments

## **Keyword: Homologous Proteins**

### **Definition:**

Two proteins with related folds and related sequences are called homologous. Commonly, homologous proteins are further divided into orthologous and paralogous proteins. While orthologous proteins evolved from a common ancestral gene, paralogous proteins were created by gene duplication.

## **Keyword: Neural Network**

### **Definition:**

A neural network is a computer algorithm to solve non-linear optimisation problems. The algorithm was derived in analogy to the way the densely interconnected, parallel structure of the brain processes information.

## **Keyword: Ontology**

### **Definition:**

The word ontology has a long history in philosophy, in which it refers to the study of being as such. In information science, an ontology is an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships among them.

## **Keyword: Open Reading Frame (ORF)**

### **Definition:**

An opening frame contains a series of codons (base triplets) coding for amino acids without any termination codons. There are six potential reading frames of an unidentified sequence.

## **Keyword: Protein Folding Problem**

### **Definition:**

Proteins fold on a time scale from  $\mu s$  to s. Starting from a random coil conformation, proteins can find their stable fold quickly although the number of possible conformations is astronomically high. The Protein Folding Problem is to predict the folding and the final structure of a protein solely from its sequence.

The Protein Structure Prediction Problem refers to the combinatorial problem to calculate the three-dimensional structure of a protein from its sequence alone. It is one of the biggest challenges in structural bioinformatics.

**Keyword: Proteome****Definition:**

The Proteome is the protein complement expressed by a genome. While the genome is static, the proteome continually changes in response to external and internal events.

**Keyword: Proteomics****Definition:**

Proteomics aims at quantifying the expression levels of the complete protein complement (the proteome) in a cell at any given time. While proteomics research was initially focussed on two-dimensional gel electrophoresis for protein separation and identification, proteomics now refers to any procedure that characterizes the function of large sets of proteins. It is thus often used as a synonym for functional genomics.

**Keyword: Sequence Contig****Definition:**

A contig consists of a set of gel readings from a sequencing project that are related to one another by overlap of their sequences. The gel readings of a contig can be combined to form a contiguous consensus sequence whose length is called the length of the contig.

**Keyword: Sequence Profile****Definition:**

A sequence profile represents certain features in a set of aligned sequences. In particular, it gives position-dependent weights for all 20 amino acids and as for insertion and deletion events at any sequence position.

**Keyword: Single Nucleotide Polymorphism****Definition:**

Single Nucleotide Polymorphisms (SNPs) are single base pair positions in genomic DNA at which normal individuals in a given population show different sequence alternatives (alleles) with the least frequent allele having an abundance of 1 % or greater. SNPs occur once every 100 to 300 bases and are hence the most common genetic variations.

**Keyword: Threading****Definition:**

Threading techniques try to match a target sequence on a library of known three-dimensional structures by „threading“ the target sequence over the known coordinates. In this manner, threading tries to predict the three-dimensional structure starting from a given protein sequence. It is sometimes successful when comparisons based on sequences or sequence profiles alone fail due to a too low sequence similarity.

**Keyword: Turing Machine****Definition:**

The Turing machine is one of the key abstractions used in modern computability theory. It is a mathematical model of a device that changes its internal state and reads from, writes on, and moves a potentially infinite tape, all in accordance with its present state. The model of the Turing machine played an important role in the conception of the modern digital computer.