Bioinformatics

A Tramontano, *Sapienza University, Rome, Italy*

Based in part on the previous version of this Encyclopedia of Life Sciences (ELS) article, Bioinformatics by David A Adler and Darrell Conklin.

Bioinformatics is a discipline at the intersection of biology, computer science, information technology and mathematics. It aims at integrating and analysing a wealth of biological data with the aim of identifying and assigning a function to each of the parts list of a living organism and understanding the incredibly complex processes that define life at a systems level. It is applied, for example, in the construction of genetic and physical maps of genomes, gene discovery, the inference of the molecular function and three-dimensional structure of their products, the interpretation of the effect of gene variations on the phenotype, the reconstruction of interaction and signal transduction pathways and the simulation of biological systems. Bioinformatics is an essential part of modern biology and a key player in the quest for a complete systems-level understanding of a living cell and of an organism.

Advanced article

Article Contents

Online posting date: 15th September 2009

Introduction

Life is a complex process, brought about by the action and interaction of biomolecules within cells. The information for the synthesis, abundance localization and time of expression of each biomolecule is encoded in the genetic material, generally represented by deoxyribonucleic acid (DNA), although a slightly different molecule, ribonucleic acid (RNA), is used by some viruses. A few decades ago, Watson and Crick published a model for the structure and replication of DNA (Watson and Crick, 1953), building on the experimental work of Rosalind Franklin (Franklin and Gosling, 1953). This led to the codification of the central dogma of molecular biology (DNA is transcribed into RNA which in turn is translated into protein), although some important exceptions are found. Molecular biology has opened the road to the reliable and fast determination of the sequence of the information carrier DNA molecules and many technological breakthroughs have led to a data explosion. The first draft of the human sequence was published in 2001 (Lander *et al*., 2001; Venter *et al*., 2001) followed by the sequencing of the mouse and rat genomes. There are now hundreds of genome sequences in the public domain and many more are planned. This explosion in genomic information has been achieved in a remarkably short period of time, and new emerging technologies are expected to very substantially increase the flood of new sequence data, which is likely to continue for the next decades. However, DNA sequence is merely a string of letters; it must be interpreted in terms of the RNA and proteins that it encodes and regulatory regions that control transcription and translation.

Delving from the level of amino acid sequence encoded in a gene, to protein structure and its associated function and interactions touches on central questions in biology. Computational approaches are now making important contributions to our understanding of the relationship of the structure and function of biomolecules and their roles in biological processes. This article will introduce some of the concepts, methods and tools of biocomputing and describe a few examples of areas of inquiry in bioinformatics.

ELS subject area: Genetics and Molecular Biology

How to cite:
Tramontano, A (September 2009) Bioinformatics. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0001900.pub2

Scope of Bioinformatics

The need to retrieve, organize and link the content of very large databases requires the development of computational tools for data analysis. This is the role of Bioinformatics, a discipline at the intersection of biology, computer science, information technology and mathematics sometimes referred to as biocomputing or computational biology, the choice of term depending on the focus of activity. The recent appearance of extremely effective high-throughput techniques for sequencing and collecting data related to the composition of a cell has resulted in an exponential growth of data and has driven the rapid growth and maturation of the discipline of bioinformatics.

Bioinformatics is an essential part of modern biology and a key player in the quest for a complete systems-level understanding of a living cell or of an organism. Its main challenge resides in the need of analysing and especially correlating a huge amount of data of different nature and

complexity, from genome maps to gene and protein sequences, from three-dimensional (3D) structures of biomolecules to their interactions, from quantification of their abundance to the temporal patterns of their presence in a cell.

A primary goal of this data analysis is directed towards unravelling the information content of biomolecules and understanding how 'bioinformation' directs the development and function of living organisms. The analysis of nucleic acid sequences, protein structure/function relationships, genome organization, regulation of gene expression, interaction of proteins and mechanisms of physiological functions, can all benefit from a bioinformatics approach and cannot be achieved without it. Nucleic acid and protein sequence data from many different species and from population samplings provides a foundation for studies leading to new understandings of evolution and the natural history of humans.

Access to the published scientific literature is another important aspect in the realms of biotechnology and biomedicine in particular; this includes invention information contained in patent archives. Biological discoveries and data are mostly stored in biological and biomedical journals and one of the many challenges of bioinformatics is to devise tools for automatically extracting relevant information from text and linking them to the corresponding entries in biological databases (Krallinger *et al*., 2008).

Data handling and analysis clearly requires an infrastructure, that is, a communications medium with fast data transfer rates and high traffic capacity to provide almost simultaneous information access to thousands of people, and a set of efficient and fast software tools for the analyses.

Bioinformatics researchers, computer scientists and information specialists are also working on the conceptual foundations for the next generation of knowledge navigators. New hardware and software developments are being devised to facilitate data access and distribution. The technological innovations in bioinformatics are accompanied by ethical concerns, particularly pertaining to data repositories of identifiable genetic information. The ethical implications of advances in bioinformatics necessitate investigation and efforts on the part of scientists, lawmakers and the public are required to ensure the privacy of individuals. **See also**: DNA Sequence Analysis; Evolution of Protein Domains; Genome, Proteome, and the Quest for a Full Structure–Function Description of an Organism; Genome Sequence Analysis

Hardware

Bioinformatics requires an extensive physical computational infrastructure. All the biological data resources are growing very rapidly – with some doubling in size every year. During the past years, the advent of new generation sequencing technologies and of metagenomics and human variation projects threatens to overwhelm the current sequence databases by its sheer size. The advent of new data types and associated new databases (such as proteomic data) also increases the need for more computer hardware. Search algorithms, which are essential for accessing and exploiting the data, often scale as the square of the data size, and so rapidly consume ever-increasing computer resources. In the last decade, bioinformatics has explored and exploited both supercomputer class machines, providing high performance by harnessing multiple central processing units (CPUs) as well as novel technologies such as GRID-based computing, a type of decentralized parallel computing that takes advantage of a large number of computers, not necessarily very powerful, connected via a network. It is very likely that in the next few years, more innovative and coordinated solutions will be developed to face the enormous challenge of keeping up with the data deluge. **See also**: Genome Databases; Nucleotide Sequence Databases; Protein Databases

Software

Software for bioinformatics is task-driven. Publicly available and commercial packages are available for the analysis of genetic and physical-mapping data, for drawing pedigrees and evolutionary trees, assembling sequences, pattern or string searching, restriction analysis, motif identification, base or amino acid composition analysis, protein characterization, structure inference and, most importantly, sequence comparisons.

A fundamental issue in bioinformatics is that essential tools for solving specific tasks are many, often created and maintained by individual research groups and can become obsolete quite rapidly. This is confusing for the end-user and sometimes can lead to incorrect usage of the tools. Furthermore, today's trend towards systems biology demands very often the sequential application of several different tools in cascade (e.g. find the gene, translate it into the proteins sequence, predict or retrieve its 3D structure, analyse data about its abundance in specific cell types or disease state and about its interactions). A much better connection between different domains of knowledge is required, and the weaknesses in information integration are becoming a hindrance to scientific discoveries. Tool integration is at the same time essential and far from trivial.

The web medium remains the easiest means of accessing data from remote networked databases (**Table 1**), and the present trend for tools in bioinformatics is towards web services. A web service is a software system designed for interoperability among different computers on the same network. It offers a software interface described in a computer readable format (e.g. the Web Services Description Language) that other systems can use to 'activate' its operations by sending 'messages' and retrieving the answer. Ideally, a curated catalogue of available updated web services will reduce redundancy and ensure a more effective usage of the available resources.

Table 1 Bioinformatics resources on the web

Databases	
Nucleotide sequences	EMBL Nucleotide Database: http://www.ebi.ac.uk/embl/index.html
	Genbank: http://www.ncbi.nlm.nih.gov/Genbank/
	DDBJ: http://www.ddbj.nig.ac.jp/
Genomes	Ensembl: http://www.ensembl.org/index.html
	UCSC GB: http://genome.ucsc.edu/
	Genome reviews: http://www.ebi.ac.uk/GenomeReviews/
	Entrez Genome: http://www.ncbi.nlm.nih.gov/sites/entrez?db = genome
Amino acid sequences	UniProt: http://www.ebi.ac.uk/uniprot/index.html
Protein structures	PDB: http://www.wwpdb.org/
Genome variations	OMIM: http://www.ncbi.nlm.nih.gov/omim/
	HapMap: http://www.hapmap.org/
Expression data	ArrayExpress: http://www.ebi.ac.uk/microarray-as/ae/
	GEO: http://www.ncbi.nlm.nih.gov/geo/
Literature	PUBMED: http://www.ncbi.nlm.nih.gov/pubmed/
Gene ontology vocabulary	GO: http://www.geneontology.org/
Tools	
Biological database navigators	Entrez: http://www.ncbi.nlm.nih.gov/sites/gquery
	EB-eye: http://www.ebi.ac.uk/eb-eye/
BLAST series/PSI-BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
FASTA	http://www.ebi.ac.uk/Tools/fasta/
Multiple sequence alignment	Clustal: http://www.ebi.ac.uk/Tools/clustalw2/
	T-coffee: http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi
Comparative modeling	Modeller: http://salilab.org/modeller/modeller.html
	Swiss-model: http://swissmodel.expasy.org/SWISS-MODEL.html
Fold recognition	Threader: http://bioinf.cs.ucl.ac.uk/threader/
	Pcons: http://www.bioinfo.se/pcons/
Fragment based methods	Fragfold: http://bioinf.cs.ucl.ac.uk/
	Rosetta, Rosetta@home, FoldIt!: http://depts.washington.edu/bakerpg/
Useful starting points for exploring bioinformatics on the web	EBI: http://www.ebi.ac.uk
	NCBI: http://www.ncbi.nlm.nih.gov/
	EXPASY: http://www.expasy.org/
	CBS: http://www.cbs.dtu.dk/

Mapping and Linkage Analysis

The 3 billion base pairs comprising the human genome are distributed among 24 different linear DNA molecules which in turn are packaged into individual chromosomes (autosomes 1–22 and the sex chromosomes, X and Y). It is predicted, very approximately, that there are 20–25 000 genes and each has a distinct location on one of the human chromosomes. The process of assigning genes and DNA fragments to locations on particular chromosomes is called mapping.

Gene maps are of two primary types, genetic and physical. Genetic maps, determined by family studies in humans and defined crosses of laboratory organisms such as mice, provide the chromosomal assignment of a gene and its position relative to other genetic markers. Synteny is the term for two genes being on the same chromosome and they are 'linked' if they are sufficiently close, on the chromosome, that they appear to be inherited together. The determination of genetic map distance between genes is referred to as linkage analysis. Linkage can only be determined for polymorphic genes, those that have two or more distinguishable alleles in a population. Genetic map distances are calculated by assessing the frequency of recombination between two polymorphic loci on a chromosome and are usually expressed in units called morgans (0.01 M, or 1 cM, is equal to 1.0% recombination). A general principle of gene mapping is that the closer two loci are, the less likely a recombinational event, or breakage in the case of several physical-mapping methods, will occur between them. The significance of linkage data is typically reported as a prediction probability and expressed as a logarithmic odds ratio or LOD ('logarithm of differences') score. As a practical and general rule, a LOD score of three or greater between two genetic markers is considered significant evidence for linkage. **See also**: Genetic and Physical Map Correlation; Genetic Linkage Mapping; Genetic Variation: Polymorphisms and Mutations; Linkage Analysis

It is obviously not practical to control matings in human populations, so human genetic maps can only be elucidated by following the segregation of traits, or genetic markers, in family studies. In the laboratory mouse, with a long history of genetic analysis, genetic maps have been constructed over the years by following the segregation of alleles in experimental matings among well-characterized inbred strains. **See also**: Animal Models

Physical maps exploit techniques of molecular and cellular biology to localize genes and other markers without the need for family studies or genetic crosses, and do not require polymorphic genes. Methods of somatic cell genetics, DNA hybridization and the polymerase chain reaction (PCR) have provided a streamlined approach to human gene mapping. Cytogenetic techniques combined with nucleic acid hybridization provide direct means of localizing genes on chromosomes. Nucleic acid probes, labelled with fluorescent dyes, are allowed to bind to their complementary sequences on spread chromosomes and are detected by fluorescence microscopy. The incidence of random breakage events between markers and the occurrence of concordant cloning of two genes are both used to estimate physical distance between loci. **See also**: Fluorescence *In Situ* Hybridization (FISH) Techniques; Genome Mapping; Polymerase Chain Reaction (PCR)

Reconciliation and integration of maps derived by different methods, particularly the combining of physical and genetic maps, contribute to increasing the accuracy and resolution of mapping data. The correspondence of different map units can also be approximated from integrated maps, for example, 1 cM is roughly equivalent to 1–2 Mb of DNA and similar in size to a small cytogenetic band. Comparisons of the gene maps of different species have proven valuable in evolutionary studies as well as identification of human disease genes. The functional significance of the conservation of genome arrangement as evidenced by, for example, the homeotic genes maintained evolutionarily as clusters, a conserved linkage from fruitflies to humans, remains unclear. **See also**: Genetic and Physical Map Correlation

The development, refinement and application of all these mapping technologies have produced dense maps of entire genomes. Human gene maps of individual chromosomes that could once be reported in graphic form on a single sheet of paper can now only be displayed with computer technology due to the exponential increase in the number of localized markers. The availability of dense gene maps also greatly facilitates positional cloning of disease loci. Positional cloning refers to a commonly used strategy that starts from knowing only the approximate location of a gene and progressively narrowing the critical region until mutations in a single gene are shown to be associated with the phenotype. There are many examples of the cloning of human inherited diseases using this approach, including Huntington disease, Duchenne muscular dystrophy and cystic fibrosis. Dense maps also provide the foundation for the realization of the ultimate map, the complete genome sequence. To ensure the value and accessibility of mapping data it is essential to maintain authoritative repositories. The ability to search and display this information is essential and is thus another important aspect of bioinformatics. **See also**: Susceptibility to Human Infectious Diseases, Genetics of

Genome Variation and Mutation Data

Improved sequencing technologies are generating a flood of data on human variation and phenotypes (in terms of mutations and disease associations). Recent human population studies have revealed correlations between certain genetic variants and disease occurrence. Furthermore, an international research consortium has been formed to sequence the genomes of at least a thousand people from around the world. This will lead to a new map of the human genome including a view of biomedically relevant DNA variations at a resolution unmatched by current resources.

These data are being collected supplementing well-known databases such as the 'Online Mendelian Inheritance in Man' (OMIM) (Amberger *et al.*, 2008), a catalogue of human genes and genetic disorders, with links to literature references, sequence records, maps and related databases and HapMap (The International HapMap Consortium, 2007), an effort to identify and catalogue genetic similarities and differences in human beings.

Using the information in these resources, researchers will be able to find genes that affect health, disease and individual responses to medications and environmental factors, but understanding how a given variation or mutation produces a given phenotype is still an open problem and bioinformatics is called on to combine available experimental data and develop new methods to help scientists interpreting the data.

Available techniques include methods for analysing the evolutionary conservation of the site affected by the mutation (which is an indication of its functional role), for predicting its effect on a protein's stability and interactions (often based on modelling its 3D structure and/or on machine-learning techniques). We will certainly witness many new developments in bioinformatics in this area that are bound to have a major scientific impact for biologists and will impact human health by increasing our understanding of how variations in genotype can determine susceptibility to different diseases and infections. **See also**: Human Variation Databases; Identifying Regions of the Human Genome that Exhibit Evidence for Positive Selection; Single Nucleotide Polymorphism (SNP)

Biosequence Analysis

The determination of the linear sequence of amino acids in proteins and the nucleotides in DNA and RNA leads to the requisite for compiling and analysing sequence data. Sequence analysis is the process of investigating the

information content of linear raw nucleic and protein sequence data.

Nucleic Acid Sequence Analysis

The bulk of genomic DNA does not code for proteins, and the protein-coding regions of human genes are not contiguous but are arranged with exons interspersed with introns. Therefore, an important question for computational biology is how to detect protein-coding regions within genomic DNA. Other common tasks include translating DNA into protein, assembling partially overlapping fragments, analysing sequences, comparing sequences and specific DNA or amino acid motif discovery and recognition. Current DNA sequencing technologies are not capable of generating complete sequence for long nucleic acid molecules in a single sequencing run and so it is necessary to utilize computational methods to assemble contiguous sequences from individual short-sequence determinations. If a large DNA molecule is randomly broken into smaller pieces for the actual sequence determinations then a contiguous linear sequence can be reconstructed by aligning the overlapping portions from different random fragments. However, the recent development of massively parallel platforms for DNA sequencing (Gupta, 2008) is having a striking impact not only on genetics, but also on the computational infrastructure required to reassemble the short reads from many of these instruments into genomic scaffolds or exons. It should be mentioned that these new generation sequencers can provide gigabytes up to terabytes of data in a single 2-h experiment. The problem of interpreting the experimental data is somehow tractable when the aim of the study is resequencing of a known genome to characterize its variations; much more complex is the task of assembling a new genome starting from fragments that can be a few nucleotides long. New generations of programs are being developed and have been applied to small genomes, but innovative computational tools are required.

A common question arising when new genes are cloned and sequenced is whether the sequence is already known or does not occur in current databases. Answering this question requires comparing the newly obtained sequence to every sequence in the database. The algorithm of choice for this task is the extremely rapid BLASTN algorithm (Altschul *et al*., 1990). A list of all W-mers (contiguous fragments of length W, which is typically set between 11 and 16), in the query sequence, is first compiled and then every sequence in the database is in turn checked against this list. This can be done rapidly and serves to rule out most sequences from consideration. These regions are then extended in either direction, using less stringent matching, to form HSPs (high-scoring segment pairs). The expectation value of the HSP (derived from the probability that an HSP of a similar score will occur between two random sequences) is computed and all database sequences having significant HSPs are reported. Overall database access time by BLASTN is minimized by using a compressed form of the nucleotide data and by using a memory-mapped file. It is an algorithm highly amenable to parallelism and can be compiled to run on multiprocessor hardware. **See also**: BLAST Algorithm; FASTA Algorithm; Genome Sequence Analysis; Similarity Search

Amino Acid Sequence Analysis

Linear chains of amino acids, proteins, the product of gene translation, are normally found in cells folded into functionally active structures. It is established that the primary sequence of the protein, that is, its amino acid sequence, determines the ultimate conformation of the protein and therefore its biological function (Sela *et al*., 1957). However, the flexibility of long-chain polypeptides can generate an almost infinite number of shapes (Levinthal, 1968), and the computational task of predicting correct structures is beyond the reach of current knowledge. Predicting the shape of a protein from its linear amino acid sequence is one of the holy grails of computational biology. Solving the protein-folding problem holds the promise of spawning major advances in assigning the function to a gene product. **See also**: Protein Structure Prediction; Protein Structure Prediction and Databases

An indispensable resource for the bioinformatics scientist is the Protein Data Bank (PDB) (Berman *et al*., 2000), a repository of solved protein structures, that is, mappings of each atom in a protein onto 3D space. This is done using X-ray crystallography or nuclear magnetic resonance. These techniques are time consuming and not necessarily applicable to every protein. For this reason, the number of known protein sequences vastly exceeds the number of sequences with solved structures in the PDB. **See also**: Disordered Proteins; Primary Protein and Nucleic Acid Three-dimensional Structure Databases

Comparative Modelling

Through the ages, the human genome has been the target of major evolutionary processes such as gene duplication, gene fusion, gene rearrangement and gene deletion. The individual gene has been subjected to the more subtle process of base mutations that often change the protein sequence of the gene product. Genes have evolved substantially while still preserving the 3D structure of their protein. This is because mutations that substantially alter a protein fold will destroy the normal function of the protein, and will not persist through generations. Therefore, the first step in predicting the fold of a new protein is to determine whether it is evolutionarily related to some sequence in the PDB. **See also**: Evolution of Protein Domains

One technique for testing two protein sequences for an evolutionary relationship is pair-wise alignment using a dynamic programming algorithm (Needleman and Wunsch, 1970; Smith and Waterman, 1981). A pair-wise alignment is the correspondence between the amino acids

of one protein and those of the other deemed to reflect their evolutionary history. The alignment algorithms output the alignment that maximizes the similarity between the two amino acid chains in the assumption that this is the most likely description of the evolutionary events. The next important question is which threshold of similarity between two proteins is sufficient to guarantee that they are homologous, that is, that they share a common ancestor and are thereby evolutionary related. Statistical evaluation of the likelihood that the observed similarity is observed by chance result can provide the answer. Sander and Schneider (1991) analysed the relationship between percentage identity, alignment length and structural similarity by studying proteins of known structure in the PDB. Chothia and Lesk (1986) quantified the relationship between sequence identity and structural divergence expressed in terms of root mean square deviation of the corresponding main chain atoms in the core of two structures. Roughly stated, when two protein sequences have at least 50% identity over at least 80 amino acids, their structures are expected to be similar (approximately 1Å root mean square deviation). It must be stressed that the converse of this implication is not true: due to evolutionary divergence, two related sequences may not have a statistically significant similarity and detecting these evolutionary relationships is a very active area of research in bioinformatics. **See also**: Protein Structure Prediction; Similarity Search; Substitution Matrices

Dynamic programming algorithms, unless implemented in massively parallel hardware, are too slow for interactive application to very large databases. This is because they require, for every database sequence, the computation of every cell in a score matrix, the total number of cells being equal to the product of the query and subject sequence lengths. Several clever algorithms have been devised to avoid the computation of the full score matrix; two popular methods are FAST-All (FASTA) (Pearson and Lipman, 1988) and BLASTP (Basic Local Alignment Search Tool Protein) (Altschul *et al*., 1990). FASTA initially computes a hash table containing all k-tuples (peptide of length k) in the query sequence. A target sequence can be tested very rapidly for the presence of these k-tuples. The second step of the FASTA algorithm performs a limited computation of the full score matrix, only in regions which join selected k-tuples. The BLASTP method initially compiles a finite state machine, employing an extremely rapid technique from computer science for finding common substrings in sequences. Using an amino acid comparison matrix all W-mer peptides (W is typically set at 2 or 3) that could possibly attain a score greater than a threshold score to any W-mer in the query are placed into the machine. The value of this threshold depends on the comparison matrix and on other parameters of the algorithm. Each region of a database sequence that reaches the final state of the machine is then extended in either direction to form HSPs. Sequences with statistically significant HSPs are reported to the scientist.

Recent and effective methods for detecting evolutionary relationships are based on profiles and Hidden Markov models (HMMs). Protein sequences related by ancient evolutionary events of gene duplication are said to form a family of sequences and they can be aligned recursively to obtain a 'multiple sequence alignment'. The multiple sequence alignment can be used to derive the sequence profile of the family. A profile is an n-tuple of probability distributions of amino acids, derived from the group of related proteins, where n is the length of the multiple alignments of these proteins. It is represented as a 2-dimensional matrix of 20 rows and n columns, where each column is a probability distribution p over the 20 amino acids in one position in the multiple alignments. The profile can be used to compute the probability that a new sequence belongs to the family used to derive the profile. The very successful Position-Specific Iterative BLAST (PSI-BLAST) (tool; Altschul *et al*., 1997) is a modification of BLAST in which a profile is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search. The profile is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second (etc.)-BLAST search and the results of each 'iteration' used to refine the profile. This iterative searching strategy results in increased sensitivity.

An even recent extension of the concept of profile searching is the 'profile–profile' search method (Sadreyev and Grishin, 2003) where the profile derived from the multiple sequence alignment of the protein family under study is compared with pre-derived profiles of known protein families.

Multiple sequence alignments can also be used to derive a HMM, a probabilistic model in which the system being modelled is assumed to be a Markov process with unknown parameters (Krogh *et al*., 1994; Eddy, 1996). HMMs are used to transform the information contained within a multiple sequence alignment into a position-specific scoring system. Their application to sequence searching further increases sensitivity. Incidentally, HMMs are also largely used to detect protein-coding regions in genomes.

Detection of an evolutionary relationship to a protein whose structure has been experimentally determined provides the basis for comparative modelling. There are several flavours of the technique, the most used and traditional one consists in assigning the atomic coordinates of the backbone of the amino acids of the protein of known structure (template) to those of the protein under study (target) according to the correspondence defined by their sequence alignment. Next, regions where insertions and deletions have occurred during evolution are modelled usually grafting regions with the appropriate characteristics from proteins of known structure (Tramontano, 1998). The combinatorial problem of predicting the optimal position of the amino acid side chains is solved by only exploring conformations commonly observed in known protein structure, often using a dead-end elimination algorithm (Canutescu *et al*., 2003). Often the rough model is

subjected to energy minimization, although it is now clear that this step does not improve the final model (Cozzetto *et al*., 2008). **See also**: Hidden Markov Models; Profile Searching; Protein Tertiary Structures: Prediction from Amino Acid Sequences

Fold Recognition

Often a new protein sequence contains no recognizable motifs, nor can its structure be inferred by comparative modelling. In such cases, one can resort to fold recognition approaches. The task of fold recognition is easily defined but notoriously difficult to solve: for a given sequence, determine which, if any, structures in the PDB are compatible with the sequence.

This is justified by the observation that, despite the vast space of protein sequences explored by evolution over the ages, there seem to exist only several thousand unique protein topologies (Chothia, 1992; Tramontano and Pearson, 2007). As the PDB continues to expand with new solved protein structures, the chance that a new gene product folds like a known structure will continue to increase. **See also**: Structural Predictions and Modeling

Because of the improved sensitivity of sequence searching techniques, it is increasingly rare that a similarity between the topology of two proteins due to an even distant evolutionary relationship goes undetected. The remaining cases are clearly extremely difficult. One possible strategy is threading. Threading methods are based on the assumptions that protein structures are in a state of minimum free energy, and that this energy can be roughly computed for any given structure. The energy computation takes into account the compatibility of different amino acids at each position in the structure. This compatibility usually reflects the preference of hydrophobic amino acids in the core environment of the protein, and the potential energy created when two amino acids are spatially close to one another.

Given a function that can evaluate the compatibility of a sequence with a structural template whose native sequence has been removed, threading algorithms attempt to minimize this function by considering various possible sequences to structure alignments. The threading task is enormously complex since exponentially many (as a function of sequence and structure sizes) alignments are possible, and the presence of arbitrarily many pair-wise interactions in a protein structure precludes the use of dynamic programming alignment algorithms to produce optimal solutions. There are two interesting heuristic algorithms for obtaining at least a feasible solution in the face of this complexity. One is the approach of Jones *et al*. (1992) who uses a variant of the standard dynamic programming algorithm. Another is the statistical sampling approach of Madej *et al*. (1995), who iteratively modifies a working alignment until a local minima is reached. Both approaches have had some success in predicting the fold of unknown proteins.

Fragment-based Methods for Protein Structure Prediction

These rather recent methods for protein structure prediction are based on the idea of generating a set of candidate models likely to be highly populated with native-like members to increase the probability of selecting a near-native model. The generation of near-native structures is based on the assembly of fragments derived from known protein structures and selected according to their local sequence similarity with the corresponding fragment of the target protein. A number of fragments (usually in the hundreds) are selected to cover every stretch of residues in the target protein (based on the size of the fragment) and randomly assembled to construct a large set of putative models. These tools use optimization procedures (Monte Carlo, simulated annealing, genetic algorithms) to enrich the set of low-energy structures through an iterative procedure. The two pioneers of this technique have been David Jones with his Fragfold tool (Jones, 2001) and David Baker with the well-known Rosetta method (Das and Baker, 2008).

Some interesting developments of the latter are worth mentioning. The first is Rosetta@home (Das *et al*., 2007), a project that uses idle computer-processing resources from volunteers' computers to perform calculations on individual work units. Completed results, that is, computer models of the target protein, are sent to a central project server where they are validated and assimilated into project databases. An even recent development, FoldIt!, turns protein folding into a free, downloadable video game, hoping that the ability of human brain in visual processing, spatial reasoning and problem solving makes them more efficient than existing computer programs at recognizing the correct fold. Should this be the case, it might be possible to identify the successful human strategies and implement them in a computer program. **See also**: Protein Structure Prediction; Protein Tertiary Structures: Prediction from Amino Acid Sequences

Protein Function Prediction

The ultimate goal of the analysis of a gene or protein sequence is the prediction of its biological function. The first problem faced by any functional annotation process is the definition of function itself. If the protein is an enzyme, the Enzyme Commission (EC) numbering scheme can be used. In more general terms, however, the problem is 'multi-dimensional': a protein can have a molecular function, a cellular role and be part of a functional complex or a pathway, and these are the distinctions used in Gene Ontology, a controlled vocabulary to describe gene and gene product functional properties.

A protein function can be inferred based on its evolutionary relationship with proteins of known function. Orthologous proteins in different species most often share function, but paralogy (i.e. divergence following

duplication of the original gene) does not guarantee common function. The distinction between orthology and paralogy can be attempted on the basis of the observed sequence similarity patterns, by analysing the specific conservation pattern of residues responsible for function in the family, or on the basis of the protein structure (either experimentally determined or modelled). When the structure, or a model, of a protein is known, distant evolutionary relationships are easier to detect and it is possible to inspect the protein structure for conservation and spatial localization of residues known to be important for function.

Even if a protein family is divergent, it may be possible to identify short regions that appear to have conserved sequences and, therefore, locally conserved structure and perhaps biochemical function. Each region can be described using a motif that states, for each position, the allowed variation in possible amino acids using a distinct score for each. The computational techniques used to create motifs fall into four classes. The standard technique creates the motif using variation observed in columns of a multiple sequence alignment of the family. There are several ways to compute motifs from a multiple alignment (Gribskov *et al*., 1987; Tatusov *et al*., 1994). Other techniques are machine-learning algorithms that attempt to create motifs without requiring a multiple alignment of the family (Brazma *et al*., 1998). The hidden Markov model techniques (Krogh *et al*., 1994) try to fit available data to a sequence of probability distributions using a local optimization algorithm. Finally, some techniques are iterative algorithms, which generalize a motif by repeated searches of a sequence database (Tatusov *et al*., 1994) using the evolving motif. **See also**: Pattern Searches

Methods based on evolutionary relationships are the most effective for function prediction, and sophisticated tools to decipher the functional meanings of these relationships are constantly being developed. Nevertheless, protein function prediction remains a challenging area of research.

Conclusion

The ability to store, integrate and analyse complex data from multiple experimental sources using interdisciplinary tools is the core of Bioinformatics.

These data come from high-throughput technological platforms, such as genomics, transcriptomics (whole cell or tissue gene expression measurements), proteomics (the complete identification of proteins and protein expression patterns of a cell or tissue), metabolomics (the identification and measurement of all small-molecules metabolites within a cell or tissue), glycomic (identification of all carbohydrates in a cell or tissue), interactomics (the catalogue of protein–protein interaction within a cell), genome variation studies and scientific literature. Their effective combination is key to the process of transforming knowledge to understanding with the ultimate goal to be able to uncover how the biological whole changes over time, for example, during evolution, in response to a perturbation or at the onset of disease.

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) BLAST – basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Altschul S, Madden T, Schaffer A *et al*. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Amberger J, Bocchini CA, Scott AF and Hamosh A (2008) McKusick's online Mendelian inheritance in man (OMIM). *Nucleic Acids Research* **37**: D793–D796.

Berman HM, Westbrook J, Feng Z *et al*. (2000) The protein data bank. *Nucleic Acids Research* **28**: 235–242.

Brazma A, Jonassen I, Eidhammer I and Gilbert D (1998) Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology* **5**(2): 279–305.

Canutescu A, Shelenkov AA and Dunbrack RL Jr (2003) A graph theory algorithm for protein side-chain prediction. *Protein Science* **12**: 2001–2014.

Chothia C (1992) Proteins. One thousand families for the molecular biologist. *Nature* **357**: 543–544.

Chothia C and Lesk A (1986) The relation between the divergence of sequence and structure in proteins. *EMBO Journal* **5**: 823–826.

Cozzetto D, Giorgetti A, Raimondo D and Tramontano A (2008) The evaluation of protein structure prediction results. *Molecular Biotechnology* **39**: 1–8.

Das R and Baker D (2008) Macromolecular modeling with Rosetta. *Annual Review of Biochemistry* **77**: 363–382.

Das R, Qian B, Raman S *et al*. (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69**: 118–128.

Eddy SR (1996) Hidden Markov models. *Current Opinions in Structural Biology* **6**: 361–365.

Franklin RE and Gosling RG (1953) Molecular configuration of sodium thymonucleate. *Nature* **171**: 740–741.

Gribskov M, McLachlan AD and Eisenberg D (1987) Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the USA* **84**(13): 4355–4358.

Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology* **26**: 602–611.

Jones DT (2001) Predicting novel protein folds by using FRAGFOLD. *Proteins* **5**: 127–132.

Jones DT, Taylor WR and Thornton JM (1992) A new approach to protein fold recognition. *Nature* **358**: 86–89.

Krallinger M, Morgan A, Smith L *et al*. (2008) Evaluation of text-mining systems for biology: overview of the second biocreative community challenge. *Genome Biology* **9**. doi:10.1186/gb-2008-9-s2-s1.

Krogh A, Brown M, Mian IS, Sjolander K and Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology* **235**: 1501–1531.

Lander ES, Linton LM, Birren B *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Levinthal C (1968) Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique* **65**: 44–45.

Madej T, Gibrat JF and Bryant SH (1995) Threading a database of protein cores. *Proteins* **23**(3): 356–369.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**(3): 443–453.

Pearson W and Lipman D (1988) Improved tools for biological sequence analysis. *Proceedings of the National Academy of Sciences of the USA* **85**: 2444–2448.

Sadreyev R and Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *Journal of Molecular Biology* **326**: 317–336.

Sander C and Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9**: 56–68.

Sela M, White FH Jr and Anfinsen CB (1957) Reductive cleavage of disulfide bridges in ribonuclease. *Sciente* **125**: 691–692.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* **147**: 195–197.

Tatusov RL, Altschul SF and Koonin EV (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proceedings of the National Academy of Sciences of the USA* **91**: 12091–12095.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.

Tramontano A (1998) Homology modeling with low sequence identity. *Methods* **14**: 293–300.

Tramontano A and Pearson WR (2007) The completeness of biological space. *Current Opinions in Structural Biology* **17**: 334–336.

Venter *et al*. (2001) The sequence of the human genome. *Science* **291**: 1304–1351.

Watson JD and Crick FHC (1953) A structure for deoxyribose nucleic acid. *Nature* **171**: 737–738.

Further Reading

Lattman EE and Loll PJ (2008) *Protein Crystallography: A Concise Guide*. Baltimore, USA: Johns Hopkins University Press.

Lesk AM (2002) *Introduction to Bioinformatics*. Oxford, UK: Oxford University Press. ISBN/ISSN: 9780199251964.

Lesk AM (2007) *Introduction to Genomics*. Oxford, UK: Oxford University Press. ISBN/ISSN: 9780199296958.

Tramontano A (2006) *Protein Structure Prediction: Concepts and Applications*. Weinheim, D: Wiley. ISBN/ISSN: 978352731167-5.

Wüthrich K (1986) *NMR of Proteins and Nucleic Acids*. New York, USA: Wiley.