REVIEW ARTICLE
UDC: 616-006:616-056.7:577.11

Arch Oncol 2005;13(2):69-75.
DOI: 10.2298/AOO0502069S

Bioinformatic tools for cancer geneticists

Karmen Stankov

ABSTRACT

Early detection is essential for the control and prevention of many diseases, particularly cancer, which is the reason why the need for new disease markers with improved sensitivity and specificity continues to grow. Utilization of sophisticated bioinformatic tools enables the increased specificity and a relatively large quantity of high quality assays for any gene of interest. Understanding the molecular characteristics of diseases, such as cancer and the detection of mutations or changes in gene expression patterns that occur as a result of the disease, will bring researchers one step closer to achieving the predictive power needed for the development of new therapies, the design of clinical trials, and specific patient treatment planning. Genetic screening is one of the fastest moving areas of medical science, particularly in oncology, and as more genes are cloned, and more disease-associated mutations discovered, the workload is set to increase considerably with the utilization of bioinformatics tools used in integration and analysis of genomic, proteomic and metabolomic profiles of cancer.

KEY WORDS: *Computational Biology; Medical Oncology; Genomics; Cloning, Molecular; Proteome*

Department of Biochemistry, Medical Faculty Novi Sad, University of Novi Sad, Serbia & Montenegro; Address correspondence to: Doc. Dr. Karmen Stankov, MD, PhD, Department of Biochemistry, Medical Faculty Novi Sad, University of Novi Sad, Hajduk Veljkova 3, 21000 Novi Sad, Serbia & Montenegro,

E-mail: stankovkarmen@yahoo.com; The manuscript was received: 18.03.2005, Provisionally acepted: 25.03.2005, Accepted for publication: 18.04.2005

New horizons in oncogenomics

The International Agency for Research on Cancer (IARC) has recently published its predictions for worldwide cancer incidence (1). They estimate that there will be almost 16 million new cases in the year 2020, an increase of 5.6 million (55%) over the figure for 2000, and 70% of these cases will be in developing countries.

In the past 30 years of cancer research, substantial progress has been made in basic and strategic research, as evidenced by our improved understanding of the nature of cancer; but less progress has been made in clinical and population research into cancer management, as exemplified by the failure to significantly improve overall cancer control. This has led to a burgeoning of translational research, to exploit experimental findings for the good of humanity, as a whole and of cancer patient in particular. Most human genes can be simultaneously screened for potential disease associations by high throughput expression profiling (e.g., by mRNA or SNP chips), as well as high throughput in silico analyses of databases that contain genetic as well as clinical and functional information (2).

In the last few decades, advances in molecular biology and the equipment available for research in this field have allowed the increasingly rapid sequencing of large portions of the genomes of several species. In fact, to date, several bacterial genomes, as well as those of some eukaryotes have been sequenced in full. The Human Genome Project, designed to sequence all 24 of the human chromosomes, is also completed. Popular sequence databases, such as GenBank and EMBL, have been growing at exponential rates (3,4). This deluge of information has necessitated the careful storage, organization and indexing of sequence information. Information science has been applied to biology to produce the field called bioinformatics. Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. Bioinformatics may be defined as the integration of biological principles into the development of algorithms involved in characterization of genomes and proteomes, and in identification of disease-related genes (5,6).

In practice, the general term of bioinformatics covers two distinct areas: the science-based aspect and the data management part. Bioinformatics refers to the access, handling, and analysis of banks of scientific data that are available for scientists to feed into their own research. Data management is rather more an extension of the traditional laboratory notebook and allows researchers to seamlessly track, search, and archive large amounts of experimental data between themselves and the other scientists working on the project. The distinction between bioinformatics and data management, however, is becoming less defined and the two functions have now effectively merged under the generic banner of bioinformatics.

The main aim of bioinformatics is to provide systems to manage not only structured data, such as documents and tables, but also unstructured data (rich data) like mass spectrometric and gene expression data. Every researcher and every laboratory is unique and the situation is far more complicated than it may at first seem, because one is dealing not only with enormous amounts of data, but also with a variety of types of data from many different sources and different types and ages of technology.

The increasing availability of high-quality biological data from projects such as the sequencing and annotation of the human and other genomes, coupled with technological advances in instrumentation, reagents, and software are transforming the way scientific discovery is performed. As biomedicine and related disciplines begin to complement traditional laboratory research with information-based science, new opportunities exist for technology providers to improve, expedite, and lower the cost of life-science discovery and development programs. For example, the ability to combine genomic, proteomic, metabolic and other biological information, with validated assays that address the full continuum of the drug-discovery and development processes will help speed the development of safer, more effective and better-targeted treatments for disease (8).

The most obvious current task of bioinformatics is to provide the essential connection for the data generated by genomics and proteomics technologies. Handling massive amount of data requires powerful integrated systems. Issues related to database inter-operability, information presentation, data classification, and automatic extraction of information from

poorly organized sources (i.e., biological literature), are being addressed by different initiatives and large consortium. Even if complete solutions are still not at hand, very interesting developments are taking place in some of these areas, including the development of creation of interchange standards in genomics, proteomics, protein interactions and metabolic networks (9). The examples of utilization of the bioinformatic tools may be the studies of protein folding and structure, structural genomics, functional analyses of regulatory genomic regions, the pharmaceutical genomics, and associated studies.

The simplest tasks used in bioinformatics concern the creation and maintenance of databases of biological information. Nucleic acid sequences (and the protein sequences derived from them) comprise the majority of such databases. While the storage and organization of millions of nucleotides is far from trivial, designing a database and developing an interface whereby researchers can both access existing information and submit new entries is only the beginning.

The most pressing tasks in bioinformatics involve the analysis of sequence information. Computational biology is the name given to this process, and it involves the following:

- Finding the genes in the DNA sequences of various organisms
- Developing methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences.
- Clustering protein sequences into families of related sequences and the development of protein models.
- Aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships.

The process of evolution has produced DNA sequences that encode proteins with very specific functions. It is possible to predict the three-dimensional structure of a protein using algorithms that have been derived from our knowledge of physics, chemistry and most importantly, from the analysis of other proteins with similar amino acid sequences. Figure 1 summarizes the process by which DNA sequences are used to model protein structure.

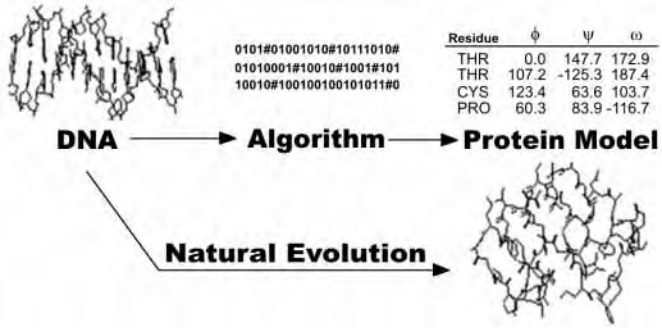

Figure 1. Modelling of protein structure

Searching for genes

The collecting, organizing, and indexing of sequence information into a database, a challenging task in it, provides the scientist with a wealth of information, albeit of limited use. The power of a database comes not from the collection of information, but in its analysis. A sequence of DNA does not necessarily constitute a gene. It may constitute only a fragment of a gene or alternatively, it may contain several genes. Luckily, in agreement with evolutionary principles, scientific research to date has shown that all genes share common elements. For many genetic elements, it has been possible to construct consensus sequences, those sequences best representing the norm for a given class of organisms (e.g., bacteria, eukaryotes). Common genetic elements include promoters, enhancers, polyadenylation signal sequences, and protein binding sites. These elements have also been further characterized into subelements. Genetic elements share common sequences, and it is this fact that allows mathematical algorithms to be applied to the analysis of sequence data. A computer program for finding genes should contain at least the elements shown in Table 1.

Table 1. The elements of a gene-seeking computer program

Algorithms for pattern recognition	Probability formulae are used to determine if two sequences are statistically similar
Data tables	These tables contain information on consensus sequences for various genetic elements. More information enables a better analysis
Taxonomic differences	Consensus sequences vary between different taxonomic classes of organisms. Inclusion of these differences in an analysis speeds processing and minimizes error
Analysis rules	These programming instructions define how algorithms are applied. They define the degree of similarity accepted and whether entire sequences and/or fragments thereof will be considered in the analysis. A good program design enables users to adjust these variables

Genome deciphered

Early in 2001, the duelling International Human Genome Sequencing Consortium (IHGSC) (public) and Celera Corporation (private) groups published papers in *Nature* and *Science* (10,11), describing the completion of so-called draft sequences. These sequences have revolutionized molecular biology by largely eliminating the need to clone and sequence genes involved in human health and disease. Instead of going to the bench, biologists now go to the web to look up gene sequences in public online databases. But despite their immediate usefulness, the draft sequences were far from perfect. Both drafts were missing some 10% of the so-called euchromatin, the gene-rich portion of the genome, and some 30% of the genome as a whole (which includes the gene poor regions of heterochromatin). The drafts contained hundreds of thousands of gaps, and had misassembled regions where portions of the genome were flipped or misplaced. As a result, any large-scale analyses of the genome, such as studies of the mechanisms of gene evolution or the long-range structure of the genome, had to contend with numerous uncertainties and artefacts. For example, studies of pseudogenes, the dying remnants of genes that have accumulated mutations that render them non-functional, had to contend with the possibility that any apparent pseudogene was instead the result of a sequencing error. Since the publication of the drafts, the IHGSC sequencing centres have quietly undertaken a laborious "finishing" process, in which each gap in the draft was individually examined and subjected to a battery of steps involving cloning and re-sequencing stretches of DNA. The high-quality reference sequence was completed in April 2003, marking the end of the Human Genome Project - 2 years ahead of the original schedule. Coincidentally, this was also the 50th anniversary of Watson and Crick's publication of DNA structure that launched the era of molecular biology. Available to researchers worldwide, the human genome reference sequence provides a magnificent and unprecedented biological resource that will serve throughout the century as a basis for research and discovery and, ultimately, myriad practical applications. The sequence is already having an impact on finding genes associated with human disease. The sequence announced today has just 341 gaps remaining, and consists of contiguous runs of sequence averaging 38 million base pairs. The authors estimate that the finished sequence covers 99% of the euchromatic portion of the genome and that the overall error rate is less than 1 error per 100,000 base pairs. This substantially exceeds the original goals for the project. The finishing procedure roughly doubled the total time and cost of the project. Does it contribute anything new to our understanding of the genome? It does indeed, and to prove the point the authors of the recent paper (12) describe several large-scale analyses of the genome that would have been difficult to perform on the draft sequence. One analysis studied the processes of gene birth and death. The authors find 1,183 human genes that show evidence of having been recently "born" by a process of gene duplication and divergence. They also find 37 genes that seem to have recently "died" by acquiring a mutation that rendered the gene non-functional. The resulting pseudogene then slowly degrades and disappears. In a second analysis, the authors use the finished sequence to map out segmental duplications, large regions of the genome that have duplicated in recent evolution. They find that 5% of the genome is involved in seg-

mental duplications, and that the distribution of these regions varies widely across the chromosomes. Knowing the nature and extent of such duplications is important for understanding the evolution of the human genome, and for studying the many medically relevant disorders that are involved in segmental duplications, such as DiGeorge syndrome and Charcot-Marie-Tooth syndrome (13).

Principles in identifying disease genes

Few areas have moved as fast as human disease gene identification. Before 1980, very few human genes had been identified as disease loci (most of these were for diseases with a known biochemical basis, where it was possible to purify the gene product). In the 1980s, advances in recombinant DNA technology allowed positional cloning and a consequent increase in the number of disease genes identified. With the advent of PCR for linkage studies and mutation screening, the pace accelerated. Now, the human and other genome projects have made available a vast range of resources (maps, clones, sequences, expression data, and phenotypic data), such as identifying novel genes is occurring on a weekly basis. Figure 2 summarizes the possible routes to identify human disease genes, and shows that there is significant interplay between clinical work, laboratory work, and computer analysis. It should be noted that there is no standard or single pathway to success, but the key step is to arrive at a plausible candidate gene, which can subsequently be tested for mutations in affected people.

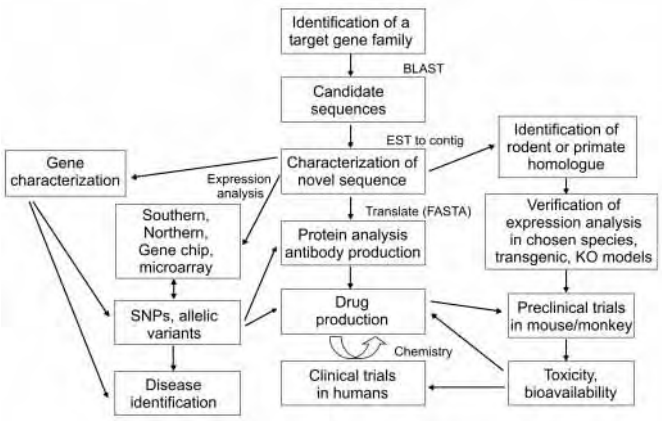

Figure 2. Target gene identification and validation is a multi-step process

Individual techniques or strategies for identifying disease genes may be classified in principle as:

- Purely position-independent, using methods based on sequence homology or functional complementation;
- Positional cloning, in which the identification of a disease gene is based solely on information regarding its approximate chromosomal location;

In reality, most disease genes have being identified by a positional candidate strategy, using a combination of positional and non-positional information. A purely position-independent approach will rarely succeed because molecular pathology is too complicated, such that predictions of the biochemical function of an unknown disease gene are often imprecise. A purely positional approach is inefficient because candidate regions identified by positional cloning usually contain dozens of genes.

Positional cloning may be defined as the identification and cloning of a specific human gene, with chromosomal location as the only available information about the gene. Many genetic disorders result from mutations in genes whose protein product and/or function is unknown. Positional cloning is the only approach to the isolation of genes when there is no biochemical or other data to suggest a likely candidate. The primary goal of a positional-cloning study is to identify and characterize the mutation(s), variants, or polymorphisms that give rise to specific disease phenotypes. The generally applied positional cloning strategies provide a guide to the procedure, yet it is impossible to predict the exact flow for any specific study. Many variables, such as the complexity of the disease phenotype (e.g., cancer), its frequency in the general population, its underlying genetic complexity, and the physical nature of the DNA surrounding the disease gene may require repeating certain steps many times.

There are four main strategies for identifying human genes:

- Functional cloning: information about the function of an unidentified gene is used to isolate the gene - either a gene product or a functional assay is required;
- Candidate gene approach: requires sufficient information about the molecular basis of pathogenesis or the existence of a suitable animal or human model where the gene is already known to be able to make an educated guess;
- Positional cloning: isolation of the gene knowing only its chromosomal location, which is typically identified by linkage analysis. This approach includes the construction of physical and genetic map of candidate region, identification of the genes within the region and investigation of each candidate gene until the disease gene is identified;
- Positional candidate approach: combines the positional and candidate gene approaches. Candidate region may be already identified, usually by linkage, and genes known to map to this region are then considered as candidates.

There is a considerable overlap between the four approaches. Ultimately, all four methods are used to identify a candidate gene, which then has to be tested for association with the disease.

All of the steps in positional cloning strategy involve the utilization of molecular biology methods (such as cDNA library screening, cDNA selection, CpG island identification, exon trapping, sequence analysis, STSs, ESTs, linkage, loss of heterozygosity, etc.) and bioinformatics methods. Our ability to find the genes involved in genetic susceptibility to many diseases, including the cancer is increasing rapidly. The utilization of bioinformatic methods in cancer research already became a routine, owing to powerful analytical tools and the completed human genome sequence information (14-16).

An essential feature of the positional candidate approach (and of successful disease gene identification) is to prioritise candidate genes. This requires position-independent information about their pattern of expression, likely function, or homology to genes implicated in relevant mutants in model organisms. Any or all of following approaches can be used:

a) Expression pattern and function

From the list of genes that map to a candidate region (from databases such as Genome Browser) (17), one would look for a gene that shows appropriate expression and/or function (or one that has homology to a human or non-human gene that displays these characteristics). A good candidate gene should have an expression pattern consistent with the disease phenotype. Although expression need not be restricted to the affected tissue, the gene should at least be expressed at the time and in the tissue where the pathology is seen. Expression patterns can be tested by RT-PCR or Northern blotting, but the best method for revealing exact expression patterns is in situ hybridisation against mRNA tissue sections. Studying the pathology of a genetic disease often allows inference of possible function(s) of the causative gene, which may allow good positional candidates to be selected. Candidate genes may also be suggested on the basis of a close functional relationship to a gene known to be involved in a similar disease, for example the genes encoding a receptor and its ligand.

b) Homology to a relevant human gene or EST

Preliminary identification of transcripts often comes from analysis of genomic sequence generated from the candidate region. If a sequence matches an unmapped EST (expressed sequence tag) in the databases, the presence of an exon is suggested. Sometimes, a sequence may show similarity to a known disease gene, and if the diseases are similar then the new gene becomes a compelling candidate. However, attempting to prioritise candidate genes on the basis of homology to other human genes may not always be successful, since many diseases show extensive locus heterogeneity, i.e., the different genes involved in a disease or group of diseases may not always be related in any obvious way, either struc-

turally or functionally.

c) Homology to a relevant gene in a model organism

Since structural and functional homologies are known to extend across even very distantly related species, it may often be more fruitful to select candidate genes by homology to genes in model organisms. Not only do gene sequences show significant homology between species, but also the pathways are often highly conserved. It should be noted that mammals might have several pathways corresponding to a single path in lower organisms. A very powerful method of selecting good candidate genes is therefore to search databases for evidence of homologous genes in other model organisms. Additionally, in the mouse, though not in non-mammalian species, the likely chromosomal location of the human ortholog can often be predicted from mouse mapping data, which allows prediction of as yet uncharacterised positional candidates.

The challenge of protein modelling

There are a myriad of steps following the location of a gene locus to the realization of a three-dimensional model of the protein that it encodes (9).

The first step is the location of transcriptional start and stop. A proper analysis to locate a genetic locus will usually have already pinpointed at least the approximate sites of the transcriptional start and stop. Such an analysis is usually sufficient in determining protein structure. It is the start and end codons for translation that must be determined with accuracy. The second phase is the identification of the position of the translational start and stop. The first codon in a messenger RNA sequence is almost always AUG. While this reduces the number of candidate codons, the reading frame of the sequence must also be taken into consideration.

There are six reading frames possible for a given DNA sequence, three on each strand, which must be considered, unless further information is available. Since genes are usually transcribed away from their promoters, the definitive location of this element can reduce the number of possible frames to three. There is not a strong consensus between different species surrounding translation start codons. Therefore, location of the appropriate start codon will include a frame in which they are not apparent abrupt stop codons. Knowledge of a protein's predicted molecular mass can assist this analysis. Incorrect reading frames usually predict relatively short peptide sequences. Therefore, it might seem deceptively simple to ascertain the correct frame in bacteria such is frequently the case. However, eukaryotes add a new obstacle to this process: detection of intron/exon splice sites.

In eukaryotes, the reading frame is discontinuous at the level of the DNA because of the presence of introns. Unless one is working with a cDNA sequence in analysis, these introns must be spliced out and the exons joined to give the sequence that actually codes for the protein.

Intron/exon splice sites can be predicted on the basis of their common features. Most introns begin with the nucleotides GT and end with the nucleotides AG. There is a branch sequence near the downstream end of each intron involved in the splicing event. There is a moderate consensus around this branch site.

With the completed primary amino acid sequence in hand, the challenge of modelling the three-dimensional structure of the protein awaits as the fourth step. This process uses a wide range of data and CPU-intensive computer analysis. Most often, one is only able to obtain a rough model of the protein, and several conformations of the protein may exist that are equally probable. The best analyses will utilize data from all the sources shown in Table 2.

Table 2. The sources of data used for protein analyses

Pattern comparison	Alignment to known homologues whose conformation is more secure
X-ray diffraction data	Most ideal when some data is available on the protein of interest. However, diffraction data from homologous proteins is also very valuable
Physical forces/energy states	Biophysical data and analyses of an amino acid sequence can be used to predict how it will fold in space

All of this information is used to determine the most probable locations of the atoms of the protein in space and bond angles. Graphical programs can then use this data to depict a three-dimensional model of the protein on the two-dimensional computer screen.

Cancer target discovery using proteomics

With the development of exceptionally sensitive instruments every few years, progress in the field of proteomics has been rapid although there are still a few areas where current technology is either not sensitive or fast enough to be useful (18). Identifying non-abundant proteins, determining millisecond physiologic responses and elucidating the expression of proteins as dynamic transitions occur in cells, organs and individuals are three such areas where there are more questions than answers.

For many diseases, the pancreatic and breast cancer are amongst the most striking examples, a diagnosis tends to occur in the advanced stages of the disease and the prognosis is very poor. Determining the molecular profiles to stratify subtypes of patients' diseases would prove the opportunity for earlier intervention and targeted treatment.

The development of mass spectrometers capable of providing data for peptide sequencing has enabled the acquisition of protein data in high throughput, giving rise to industrialized proteomics. Proteomics on such a scale offers a rapid route of discovery of new targets for pharmaceutical intervention (19,20).

Introduction to molecular biology databases

Recent years have seen an explosive growth in biological data, which is often not published anymore in a conventional sense, but deposited in a database. Sequence data from mega-sequencing projects may not even be linked to a conventional publication. This trend and the need for computational analyses of the data made databases essential tools for biological research. Under the URL of Swiss Institute of Bioinformatics, it is possible to find a comprehensive web document that lists many databases and many other information sources for molecular biologists (21).

A biological database is a large, organized body of persistent data, usually associated with computerised software designed to update, query, and retrieve components of the data stored within the system. A simple database might be a single file containing many records, each of which includes the same set of information.

Services that abstract the scientific literature began to make their data available in machine-readable form in the early 1960s. However, up to date, none of the abstracting services has a complete coverage. The best known is MEDLINE and now PUBMED, abstracting mainly the medical literature. MEDLINE/ PUBMED is best accessible through NCBI's ENTREZ (22, 23). EMBASE is a commercial product for the medical literature. Most of the bibliographical databases are only available through commercial database vendors.

Taxonomic databases are rather controversial since the soundness of the taxonomic classifications done by one taxonomist will be directly questioned by the next taxonomist. Various efforts are going on to create a taxonomy source (e.g. "The Tree of Life" project; "Species 2000"; Integrated Taxonomic Information System, etc.) (24-26). The most generally useful taxonomic database is that maintained by NCBI (27). That hierarchical taxonomy is used by the Nucleotide Sequence Databases, SWISS-PROT and TrEMBL, and is created by an informal group of experts.

The International Nucleotide Sequence Database Collaboration is a joint production of the nucleotide sequence database by the EBI (European Bioinformatics Institute), DDBJ (DNA Data Bank of Japan), and NCBI (National Centre for Biotechnology Information) (28-30). In Europe, the vast majority of the nucleotide sequence data produced is collected, organized and distributed by the EMBL Nucleotide Sequence Database located at the EBI, Cambridge, UK, an Outstation of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany (31). The nucleotide sequence databases are data repositories, accepting nucleic acid sequence data from the community and making it freely available. The databases strive for completeness, with the aim of recording every publicly known nucleic acid sequence. These data are heterogeneous; they vary with respect to the source of the mate-

rial (e.g. genomic versus cDNA), the intended quality (e.g. finished versus single pass sequences), the extent of sequence annotation and the intended completeness of the sequence relative to its biological target (e.g., complete versus partial coverage of a gene or a genome). EMBL, NCBI, and DDBJ automatically update each other every 24 hours with the new sequences they collected or updated. The result is that they contain the same information, except for sequences that have been added in the last 24 hours. Each entry in the database must have a unique identifier that is a string of letters and/or numbers that only that record has. This unique identifier, which is known as the accession number, can be quoted in the scientific literature, as it will never change. As the accession number must always remain the same, another code is used to indicate the different versions due to sequence corrections. Therefore, it should be always taken care to quote both the unique identifier and the version number, when referring to records in a nucleotide sequence database.

For organisms of major interest to geneticists, there is a long history of conventionally published catalogues of genes or mutations. In the past few years, most of these have been made available in an electronic form, and a variety of new databases have been developed. There are several databases for genomes of Escherichia coli (E. coli Genetic Stock Centre), Saccharomyces (MIPS yeast database), Drosophila melanogaster (FlyBase), Caenorhabditis elegans, or mouse (Mouse Genome Database), used in the functional study of house keeping and other evolutionary conserved genes in humans and human diseases (32-37). Two major databases for human genes and genomics are in existence. McKusick's Mendelian Inheritance in Man (MIM) is a catalogue of human genes and genetic disorders and is available in an online form (OMIM) from the NCBI (38). The Genome Database (GDB) is the major human genome database including both molecular and mapping data (39). Both OMIM and GDB include information on genetic variation in humans, but there is also the Sequence Variation Database project at the EBI, with links to many single sequence variation databases at the EBI, and to the SRS (Sequence Retrieval System) interface to many human mutation databases (40). The GeneCards resource at the Weizmann Institute integrates information about human genes from a variety of databases, including GDB, OMIM, SWISS-PROT and the nucleotide sequence databases (41). GENATLAS also provides a database of human genes, with links to diseases and maps (42). The goal of the National Cancer Institute's Cancer Genome Anatomy Project (CGAP) is to determine the gene expression profiles of normal, precancerous, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient. By collaborating with scientists worldwide, such as the Ludwig Institute for Cancer Research and Lund University, CGAP seeks to increase its scientific expertise and expand its databases for the benefit of all cancer researchers (43).

A relatively new database has been created by EnsEMBL, a joint project between EMBL-EBI and the Sanger Centre that strives to develop a software system, which produces and maintains automatic annotation on eukaryotic genomes. Human data are available now, with worm and mouse added later (44).

The protein sequence databases are the most comprehensive source of information on proteins. It is necessary to distinguish between universal databases covering proteins from all species and specialized data collections storing information about specific families or groups of proteins, or about the proteins of a specific organism. Two categories of universal protein databases can be discerned: simple archives of sequence data and annotated databases where additional information has been added to the sequence record.

The Protein Information Resource (PIR) was established in 1984 by the National Biomedical Research Foundation (NBRF) as a successor of the original NBRF Protein Sequence Database, developed over a 20-year period (45). Since 1988, the database has been maintained by PIR-International, collaboration between the NBRF, The Munich Information Centre for Protein Sequences (MIPS) and the Japan International Protein Information Database (JIPID) (46). The PIR database is portioned into four sections, according to the degree of classification of protein entries, with section four that includes sequences identified as not naturally occurring or expressed. PIR provides also some degree of cross-referencing to other biomolecular databases by linking to the DDBJ/EMBL/GenBank nucleotide sequence databases, GDB, FlyBase, and OMIM.

SWISS-PROT is an annotated protein sequence database established in 1986 and maintained collaboratively by the Swiss Institute of Bioinformatics (SIB) and the EMBL Outstation, The European Bioinformatic Institute (EBI) (47-49). It strives to provide a high level of annotation, a minimal level of redundancy, a high level of integration with other molecular databases as well as extensive external documentation. Each entry in SWISS-PROT gets thoroughly analysed and annotated by biologists ensuring a high standard of annotation and maintaining the quality of the database (50). SWISS-PROT contains data that originate from a wide variety of organisms; the latest release number 46 from February 1, 2005 contains 168297 annotated sequence entries from more than 8826 different species. SWISS-PROT includes in addition to citations about sequencing work also references to other scientific work like 3-D structure determination, mutagenesis, and detection of post-translational modifications and variants.

There is a tremendous increase of sequence data due to technological advances (such as sequencing machines), the use of new biochemical methods (such as PCR and gene chip technology), as well as the implementation of projects to sequence complete genomes (51). These advances have brought along an enormous flood of sequence information. Maintaining the high quality of SWISS-PROT requires, for each entry, a time-consuming process that involves the extensive use of sequence analysis tools along with detailed creation steps by expert annotators, is the rate-limiting step in the production of the database. A supplement to SWISS-PROT was created in 1996, since it is vital to make new sequences available as quickly as possible without relaxing the high editorial standards of SWISS-PROT. This supplement, the TrEMBL (Translation of EMBL nucleotide sequence database), consists of computer-annotated entries derived from the translation of all coding sequences (CDS) in the EMBL nucleotide sequence database, except for those already included in SWISS-PROT. All EMBL nucleotide sequence database divisions are regularly scanned for new or updated CDS features. These are translated to TrEMBLnew entries, which are in SWISS-PROT format. This section is organized in six subsections:

1. Immunoglobulins and T-cell receptors (file name Immuno.dat)
2. Synthetic sequences (file name Synth.dat)
3. Small fragments (file name Small.dat), a subsection with protein fragments with less then eight amino acids.
4. Patent application sequences (file name Patent.dat). Coding sequences captured from patent applications.
5. CDS not coding for real proteins (file name Pseudo.dat).
6. Truncated proteins (file name Truncated.dat), which result from events like mutations introducing a stop codon leading to the truncation of protein product.

Amongst the specialized proteins databases, the ENZYME database presents an annotated extension of the Enzyme Commission's publication, linked to SWISS-PROT (52). There are also databases of enzyme properties - BRENDA, Ligand Chemical Database for Enzyme Reactions (LIGAND), and the Database of Enzymes and Metabolic Pathways (EMP) (53, 54). BRENDA, LIGAND and EMP are searchable via SRS at the EBI. LIGAND is linked to the metabolic pathways in Kyoto Encyclopaedia of Genes and Genomes (55, 56).

The number of known protein structures is increasing very rapidly and these are available through the Protein Data Bank (PDB) (57). The Nucleic Acid Database (NDB) is the database for structural information about nucleic acid molecules (58). There is also a database of structures of "small" molecules, of interest to biologists concerned with protein-ligand interactions, from the Cambridge Crystallographic Data Centre (59).

CONCLUSION

The rapidly emerging field of bioinformatics promises to lead to advances in understanding basic biological processes, and in turn, advances in the diagnosis, treatment and preven-

tion of many genetic diseases. Bioinformatics has transformed the biomedical disciplines from purely lab-based science to information sciences as well. Increasingly, biological studies begin with a scientist conducting a vast numbers of database and Web site searches to formulate specific hypotheses or design large-scale experiments. The implications behind this change, for both science and medicine are staggering.

REFERENCES

1. International Agency for Research on Cancer. World cancer report. Lyon: IARC; 2003.
2. Koenig-Hoffmann K, Bonin-Debs AL, Boche I, Gawin B, Gnirke A, Hergersberg C, et al. High throughput functional genomics: identification of novel genes with tumor suppressor phenotypes. Int J Cancer 2005;113:434-9.
3. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: update. Nucleic Acids Research 2004;32:23-6.
4. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL Nucleotide Sequence Database. Nucleic Acids Research 2005;33:29-33.
5. Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. Science. 2002;298:2345-9.
6. Futreal PA, Kasprzyk A, Birney E, Mullikin JC, Wooster R, Stratton M. Cancer and genomics. Nature 2001;409:850-2.
7. Bardelli A, Velculescu VE. Mutational analysis of gene families in human cancer. Current Opinion in Genetics & Development 2005;15:5-12.
8. Brindle K. Metabolomics: Pandora's box or Aladin's cave. Biochemist 2003;25:15-7.
9. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Current Opinion in Structural Biology 2002;12:368-73.
10. International Human Genome Sequencing Consortium. Initial sequencing and analysis of human genome. Nature 2001;409:860-921.
11. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291:1304-51.
12. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931-45.
13. Stein LD. End of the beginning. Nature 2004;431:915-6.
14. Balmain A, Gray J, Ponder B. The genetics and genomic of cancer. Nature Genetics 2003;33:238-44.
15. Stankov K, Pastore A, Toschi L, Kraimps JL, Bonneau D, Gibelin H, et al. Allelic loss on chromosomes 2q21 and 19p13.2 in Hürthle thyroid tumors. Intl J Cancer 2004;111:463-7.
16. McKay JD, Thompson D, Lesueur F, Stankov K, Pastore A, Watfah C, et al. Evidence for interaction between the TCO and NMTC1 loci in familial non-medullary thyroid cancer. J Med Genet 2004;41:407-12.
17. Home page: Genome Browser [Internet]. Available from: http://www.genome.ucsc.edu/
18. Cooper JW, Wang Y, Lee CS. Recent advances in capillary separations for proteomics. Electrophoresis 2004;25:3913-26.
19. Adam PJ, Boyd R, Tyson KL, Fletcher GC, Stamps A, Hudson L, et al. Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer. J Biol Chem 2003;278:6482-9.
20. Fletcher GC, Patel S, Tyson K, Adam PJ, Schenker M, Loader JA, et al. hAG-2 and hAG-3, human homologues of genes involved in differentiation, are associated with oestrogen receptor-positive breast tumors and interact with metastasis gene C4.4a and dystroglycan. Br J Cancer 2003;88:579-85.
21. Home page: Swiss Institute of Bioinformatics [Internet]. Available from: http://www.expasy.ch/alinks.html
22. Home page: MEDLINE/ PUBMED [Internet]. Available from: http://www.ncbi.nlm.nih.gov/PubMed/
23. Home page: National Centre for Biotechnology Information's Entrez [Internet]. Available from: http://www.ncbi.nlm.nih.gov/Entrez/
24. Home page: "The Tree of Life" project [Internet]. Available from: http://phylogeny.arizona.edu/tree/life.html
25. Home page: "Species 2000" project [Internet]. Available from: http://www.sp2000.org/
26. Home page: Integrated Taxonomic Information System [Internet]. Available from: http://www.itis.usda.gov/itis
27. Home page: National Centre for Biotechnology Information's Taxonomy [Internet]. Available from: http://www.ncbi.nlm.nih.gov/Taxonomy/
28. Home page: European Bioinformatics Institute [Internet]. Available from: http://www.ebi.ac.uk
29. Home page: DNA Data Bank of Japan [Internet]. Available from: http://www.ddbj.nig.ac.jp
30. Home page: National Centre for Biotechnology Information [Internet]. Available from: http://www.ncbi.nlm.nih.gov
31. Home page: European Molecular Biology Laboratory Nucleotide Sequence Database [Internet]. Available from: http://www.ebi.ac.uk/embl.html
32. Home page: E. coli Genetic Stock Centre [Internet]. Available from: http://cgsc.biology.yale.edu/top.html
33. Home page: Munich Information Centre for Protein Sequences, Yeast Database [Internet]. Available from: http://www.mips.biochem.mpg.de/proj.yeast/
34. Home page: FlyBase [Internet]. Available from: http://flybase.bio.indiana.edu/
35. Home page: Sanger Institute [Internet]. Available from: http://www.sanger.ac.uk/Projects/C_elegans/
36. Home page: Mouse Genome Database [Internet]. Available from: http://www.informatics.jax.org
37. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 2002;420:520-62.
38. Home page: Online Mendelian Inheritance in Man [Internet]. Available from: http://www3.ncbi.nlm.nih.gov/Omim
39. Home page: The Genome Database [Internet]. Available from: http://gdb.org/
40. Home page: Sequence Variation Database project [Internet]. Available from: http://www.ebi.ac.uk/mutations/index.html
41. Home page: Weizmann Institute's GeneCards resource [Internet]. Available from: http://bioinfo.weizmann.ac.il/cards
42. Home page: Genatlas [Internet]. Available from: http://web.citi2.fr/GENATLAS/
43. Home page: National Cancer Institute's Cancer Genome Anatomy Project [Internet]. Available from: http://cgap.nci.nih.gov/
44. Home page: EMBL-EBI and Sanger Centre database [Internet]. Available from: http://www.ensembl.org
45. Barker WC, Garavelli JS, McGarvey PB, Marzec CR, Orcutt BC, Srinivasarao GY, -et al. The PIR-International Protein Sequence Database. Nucleic Acid Research. 1999;27:39-43
46. Home page: National Biomedical Research Foundation [Internet]. Available from: http://www-nbrf.georgetown.edu/
47. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acid Research. 1999;27:49-54
48. Home page: Swiss Institute of Bioinformatics [Internet]. Available from: http://www.expasy.ch/
49. Home page: European Bioinformatic Institute [Internet]. Available from: http://www.ebi.ac.uk/swissprot/
50. Apweiler R, Gateau A, Contrino S, Martin MJ, Junker V, O'Donovan C, et al. In: Gaasterland T, Karp P, Karplus C, Ouzounis C, Sander C, Valencia A, editors. Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB). Menlo Park: AAAI Press, 1997. p. 33-43.
51. Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, Kincaid R, et al. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. PNAS 2004;101:17765-70.
52. Home page: ENZYME database [Internet]. Available from: http://www.expasy.ch/enzyme/
53. Home page: Database of enzyme properties, BRENDA [Internet]. Available from: http://www.brenda.uni-koeln.de/brenda/

54. Home page: Ligand Chemical Database for Enzyme Reactions [Internet]. Available from: http://www.genome.ad.jp/dbget/ligand.html

55. Home page: Sequence Retrieval System [Internet]. Available from: http://srs.ebi.ac.uk/

56. Home page: Kyoto Encyclopaedia of Genes and Genomes [Internet]. Available from: http://www.genome.ad.jp/kegg/kegg.html

57. Home page: Protein Data Bank [Internet]. Available from: http://www.rcsb.org/pdb/

58. Home page: Nucleic Acid Database [Internet]. Available from: http://ndbserver.rutgers.edu/

59. Home page: Cambridge Crystallographic Data Centre [Internet]. Available from: http://www.ccdc.cam.ac.uk/