# Bioinformatics and the developing world

**Sándor Pongor** and
International Center for Genetic Engineering and Biotechnology (ICGEB), Trieste Component, AREA Science Park, Padriciano 99, 34012 Trieste, Italy. Phone (+39) 040 3757 300; Fax (+39) 040 226 555

**David Landsman**
National Center for Biotechnology Information, NLM, NIH. Computational Biology Branch, Building 38A Room 5N507, Bethesda, MD 20894, USA. Phone (+1) 301 496 2475

Sándor Pongor: pongor@icgeb.trieste.it; David Landsman: landsman@ncbi.nlm.nih.gov

Bioinformatics is the science of managing and analyzing biological information. Whereas all branches of modern biology make some use of computers, molecular biology and especially genetic engineering could not even exist without them. Given a certain research and computer infrastructure, developing countries may have relatively easy access to the products of bioinformatics. However, their future use of this technology hinges on the availability of bioinformatics knowledge in the public domain.

Molecular biology studies the structure of genes and proteins, which are macromolecules consisting of several thousand atoms. Since it is not possible to draw such complex pictures with pencil and paper, computers are needed to gain insight into their structure to plan experiments. In recent years, biotechnology has made outstanding progress, allowing scientists to modify genetic structure in order to develop, for example, new drugs or pest-resistant plants.

Some of the underlying tools for this progress are provided by bioinformatics, which is a collective name used for computer approaches in fields such as molecular biology, biotechnology, medicine and agriculture. In the words of leading biologist Lee Hood, "*Biotechnology is the industrial use of biological information*".

The *Organisation for Economic Co-operation and Development* (OECD) has called bioinformatics a 'megascience', that is, a strategic discipline that forms the background of the entire biomedical field. This is because, *firstly*, bioinformatics is universally applicable to molecular biology (a discipline that underlies many new areas of biological research), and, *secondly*, bioinformatics has successfully integrated experimental and bibliographic databanks, thereby producing an exemplary scientific infrastructure which is equally applicable to biomedical and to agricultural research. For these two main reasons bioinformatics can be considered a switchboard between the various scientific fields since it allows easy translation and application of findings from one technical area into another.

According to market analysts the commercial bioinformatics market totalled US$ 500 million spent by the pharmaceutical and biotechnology industries on bioinformatics staff, data and software and hardware resources in 1997 and is estimated to have doubled in 1999.

## What are the products of bioinformatics?

Bioinformatics produces databanks such as collections of protein sequences for biology and biotechnology, as well as computer software for analysis. Users such as biotechnologists and academic biologists may choose to install these components on a local computer system, or access them over the internet, using the publicly available databanks.

Bioinformatics currently deals with several main types of biological data:

- á   *Sequences and structure of genes and proteins*. Sequences are the simplest way to represent a macromolecule. The structure of genes that code for the sequence of amino-acids in proteins is produced in this form by genome sequencing projects. Protein sequences are usually obtained via computer-based translation of genomic data.

- á   *3-D molecular structures*. These are obtained by physical measurement (X-ray, Nuclear Magnetic Resonance) combined with computer modelling.

- á   *Genome structures and functions*. The genome of an organisms is composed by the entire genetic material. Information on genome structure and function is a basic description that is continuously updated with new information including links to other databases.

- á   *Bibliographic data, such as abstracts of scientific articles*. The amount of data has increased exponentially, especially since the onset of genome projects, such as the human genome sequencing programmes. The data are currently organized into a small number of large public databanks available through the internet.

*Data management*, including data processing and database maintenance, is the first and most fundamental task of bioinformatics. A considerable share of the data are produced by publicly funded research and are deposited in public databanks. Annotation of the raw data, which means to add functional and other descriptions, is a significant and time-consuming part of this work, and is largely covered by the governmental research centres described below. Another large sub-field, biomathematics or biocomputing, is concerned with developing specialized algorithms for accessing and analysing these data. The most frequent research tasks in this sub-field are sequence similarity searching, to find a protein or gene similar to a novel sequence, and database retrieval.

## Bioinformatics in the laboratory

The use of computers for biology starts in the laboratory, for instance, to plan how a DNA molecule will be cut and tailored with the several hundreds of enzyme reagents available. In order to carry out the relatively simple task of cutting out a precise fragment of a DNA piece, it is necessary to find one or two enzymes that cut somewhere near the ends of the desired piece, but will not cut the fragment itself. One such enzyme may cut a piece of DNA into a few, or into several hundred fragments, depending on the sequence of the DNA piece. A computer can enumerate all the possible fragments that can be obtained, and suggest enzyme combinations, and a protocol for the experiment.

A more sophisticated task is the characterization of a gene sequence that is obtained from an experiment. To this end, the biologist performs a database search on several of the publicly accessible and frequently updated sequence databases available on the internet. The gene sequence is compared with the sequences in the DNA database, resulting in a ranked list of the 'hits' to the most similar sequences found in the database. Just a few sufficiently similar sequences are usually enough to predict the properties and hence the natural function of the new gene or protein with considerable probability. If no obviously similar sequences are found in the databank then more sophisticated tools, such as pattern searching could provide unknown characteristics in order to predict predict properties of the gene or protein in question. The majority of current molecular biology research relies on these techniques.

## Key players and proprietary issues

Bioinformatics is widespread among universities, public non-profit research institutions, and industry. The management and daily updating of public databanks are carried out by a number of institutions working in collaboration: the *European Bioinformatics Institute* (EBI), Cambridge, UK; the *National Centre for Biotechnology Information* (NCBI) of the *National Institutes of Health* (NIH), Bethesda, MD, USA; and the *DNA Data Bank of Japan* (DDBJ). In addition, several academic groups such as the Swiss *Institute of Bioinformatics,* the *Munnich Information Centre for Protein Sequences* (MIPS, Germany) and the *Sanger Centre* in Hinxton (UK), do not only produce data but also contribute to data annotation.

The NCBI, for example, has established a complex data access system, which allows both similarity search and data retrieval. The data are cross-referenced with other databases, so that users world-wide can freely navigate between related sequences, structures, and bibliographic data over the internet. The NCBI system is perhaps the best example of a complex scientific infrastructure that can be directly used by researchers not trained in bioinformatics. The search programmes and many of the database tools were originally developed by NCBI, which is also the curator of the public sequence databanks in the USA. The bibliographic database (MEDLINE) comes from the NCBI's parent institution, the *National Library of Medicine*. The result is an integrated system that allows for the investigation of complex questions by browsing through a network of cross-referenced databases.

Many academic groups maintain bioinformatics internet sites (see box 2 on page xx), and thereby the majority of new methods developed by biomathematics groups become freely available for public use also in developing countries. However, the advent of the genomics era prompted many large pharmaceutical and agri-biotech firms to establish biotechnology departments that maintain proprietary bioinformatics systems in closed computing environments. This is not only expensive but requires highly trained informatics staff. The main motivation is to extract every ounce of useful information out of the sequence data before it becomes public and the competing firms can apply it.

**Box 2**

**Bioinformatics internet sites**

| Institutions and Programmes | Internet address |
|---|---|
| National Center for Biotechnology Information (NCBI) in Bethesda, USA | http://www.ncbi.nlm.nih.gov/ |
| European Bioinformatics Institute (EBI) in Hinxton, UK | http://www.ebi.ac.uk/ |
| European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany | http://www.embl-heidelberg.de/ |
| Swiss Institute of Bioinformatics (SIB) in Geneva, Switzerland | http://www.isb-sib.ch/ |
| Munich Information Centre for Protein Sequences (MIPR) in Munich, Germany | http://www.mips.embnet.org/ |
| International Centre for Genetic Engineering and Biotechnology (ICGEB) in Trieste, Italy | http://www.icgeb.trieste.it/ |
| European Molecular Biology Network (EMBnet) | http://www.embnet.org/ |
| BINAS, the UNIDO site for biosafety regulations | http://binas.unido.org/binas/ |
| Genome Database (GDB), in support of the Human Genome Project | http://www.gdb.org/ |
| South African Bioinformatics Institute (SANBI), University of the Western Cape, South Africa | http://www.sanbi.ac.za/ |
| Bioinformatics Centre, University of Pune, India | http://bioinfo.ernet.in/ |
| **Companies** | |
| Human Genome Sciences, Inc. (HGSI), USA | http://www.hgsi.com/ |
| Celera Genomics, USA | http://www.celera.com/ |
| Incyte Pharmaceuticals, Inc., USA | http://www.incyte.com/ |
| Institute for Genomic Research (TIGR), USA | http://www.tigr.org/ |
| Genetics Computer Group, (GCG), USA | http://www.gcg.com/ |
| Molecular Simulations, Inc. (MSI), USA | http://www.msi.com/ |
| Lion Bioscience AG, Heidelberg, Germany | http://www.lion-ag.de/ |

The complexity of bioinformatics tasks has given rise to novel forms of *research and development* (R&D) collaborations. Highly trained bioinformaticians leaving genome projects have been known to found specialized bioinformatics companies. The new companies are usually based on the knowledge gained in academia, with venture capital used to buy the necessary hardware. They offer their services to large companies that have their own genomic data, but prefer not to establish a complete bioinformatics department. *Lion Bioscience* (Germany) is a typical example of the new-style companies, offering services in data processing as well as in data generation.

Bioinformatics knowledge itself has not been restricted by patenting. Most algorithms are public and most of the best software source codes are freely accessible. The specialized knowledge of small bioinformatics companies lies mostly in practical expertise, for instance, how to combine known algorithms and programmes into large systems suitable for massive data processing. Patentability and accessibility of sequence data are issues not directly related to bioinformatics.

While the majority of sequence data reaches the public databanks, there are also proprietary data, for instance some *Expressed Sequence Tag* (EST) databases. ESTs are short DNA sequences that are identified to be expressed and translated as proteins. Information from these EST databases generated by commercial companies are available for a fee..

## Is bioinformatics accessible to developing countries?

Public bioinformatics resources such as databanks and software tools that are crucial for biotechnology projects, are available via the internet. Scientists need only a computer and an internet connection to use them. With a good internet connection, the situation of a developing country biologist is no different than that of an academic biologist in an industrialized country.

Modern biomathematics research does not require more resources than any other field of computer science; almost all processes can be efficiently designed and modelled on a personal computer or workstation. If this basic infrastructure is provided, biomathematics can be recommended to developing country universities as an up-to-date and promising research subject that does not require excessive resources. However, the step from theoretical biomathematics to applied bioinformatics**,** intending to produce software from an algorithm, is not an easy one. It requires a supportive R&D climate that generates a local need for such research, and an appropriate computer science infrastructure. At present, bioinformatics has successfully been applied only in those developing countries where these requirements are met, such as Brazil, China, India, Mexico and South Africa. For instance, the *South African National Bioinformatics Institute* (SANBI) in Cape Town has developed STACK, a new database and querying system for expressed human genes, which is of high value to drug development and biotechnology companies. As a spin-off vompany, *Electric Genetics* (EG) has been created to commercialize the data system. With a grant from the South African government the product will be further developed and marketed. Profits derived from EG's product, which is already used by many bioinformatics groups around the world, are partly channelled back to non-profit SANBI in the form of equipment.

## Building bioinformatics capacity

Access to the results of bioinformatics, that is, teaching users to understand bioinformatics, is a general problem, and in this respect expertise is the bottleneck. The first challenge is to teach the fundamentals of bioinformatics to university students, which is even more complicated since senior professors and science managers are often not yet familiar enough with the methodology, not only in developing countries. The second source of difficulty is a problem of training. There are few university-level bioinformatics curricula worldwide, and very few in developing countries. Moreover, graduates are often absorbed by pharmaceutical or agri-biotech multinationals.

Furthermore, instead of maintaining bioinformatics research groups, many universities choose to support bioinformatics teaching in the general framework of biological curricula, using knowledgeable user-level teachers involved in other areas of biological research at the same university.

In the USA, this will change with new initiatives from the NIH and the *Howard Hughes Medical Institute* to establish 20 national bioinformatics programmes of excellence. In Europe pharmaceutical companies have founded an industrial collaboration centred around the *European Bioinformatics Institute* (EBI) in Hinxton, UK, that helps them train their own bioinformaticians. The UK *Research Councils* coordinate their efforts in creating new training and research facilities.

The first program specifically addressing the bioinformatics needs of developing countries has been initiated in 1990 by the *International Centre for Genetic Engineering and Biotechnology* (ICGEB [see box on page XX]). An estimated US$ 200,000 per year are spent in this program for establishing ICGEBnet, a high level computing environment freely available for developing country scientists, as well as for theoretical and practical courses, WWW-based tutorials and research grants. ICGEBnet has to date served over 1300 registered users and a comparable number of course participants from Asia, Africa, South America and (mostly Central and Eastern) Europe.

## Box 1

### ICGEB

The *International Centre for Biotechnology and Genetic Engineering* (ICGEB), was established in 1987 as an international research organization focusing on molecular biology and biotechnology. ICGEB is supported by 43 countries. The centre is hosted by the governments of Italy and India, and the member countries include most of Latin America, and many Asian, African and East-European countries, including China, India and Russia. Today, ICGEB has two main laboratories, in Trieste, Italy, and New Delhi, India, and a network of 32 *Affiliated Centres* selected from the national research institutes of the member countries. The core of its activities is a research programme carried out at the two main laboratories, but ICGEB also operates a grant system, and a system of postgraduate fellowships, as well as two PhD programmes. A series of ICGEB seminars and symposia are organized free of charge for member country scientists. In addition ICGEB provides a forum for the promotion of the safety and regulatory aspects of biotechnology worldwide.

A European initiative, the *European Molecular Biology Network* (EMBnet) was organized in 1989 to help the spread of bioinformatics data and techniques in the European Union. Although EMBnet is mainly active in Europe, it recently adopted some non-European countries including Argentina, China, India and South Africa as members, and has sponsored several courses, partly in collaboration with ICGEB, outside Europe. EMBnet can therefore be considered a worldwide network of centres of bioinformatics expertise that spreads bioinformatics knowledge through a local infrastructure and local training courses. Such centres can be established at university campuses; a practice that could be recommended to developing countries.

Other programmes include a recent initiative of the *International Council of Scientific Unions* (ICSU) for the establishment of an internet tutorial by NCBI, and a project of the

*United Nations Educational, Scientific and Cultural Organization* (UNESCO), in which an estimated US$ 40,000 was spent in 1997 to 98 on supporting an initiative at Israel's *Weizmann Institute* aimed at organizing courses in Middle Eastern countries.

There is a growing number of teaching materials on the internet itself, and one can expect that the availability of the internet will help to compensate the missing basic level training courses in many countries.

## Future perspectives of bioinformatics

Bioinformatics cannot be disregarded by any country intending to remain up-to-date in the biomedical, biotechnological and agricultural sectors. In addition to this general trend, developing countries may also want to manage their own specific data on indigenous biological species, on local epidemiology and biodiversity programmes. These tasks clearly require that statisticians and informatics experts become advanced users of bioinformatics software and develop a capability to solve problems locally. This process does not require large resources in itself but will allow developing countries further investigate their own biological resources. To facilitate this process biomathematics/biocomputing should be introduced to universities, and the establishment of small software groups and companies encouraged.

Whereas the 1990s have been characterized by genome projects which called for massive data processing solutions, the next step will be to interpret and understand the results. At present, advanced bioinformatics is concentrated in a few research centres and private companies around the world which have the capacity to employ personnel with highly specialized training. In spite of the fact that bioinformatics methods are freely accessible, there is clearly a gap between the developing and the industrialized world which must be consciously narrowed. Bioinformatics is indeed the enabling technology for several fields of biomedical and agricultural research. The use of bioinformatics spreads freely through the internet and it helps developing countries to catch up with industrialized countries. All this is based, however, on the principle that information resources world-wide remain freely accessible.

## Sources

1. Bradbury, EM.; Pongor, S. Structural Biology and Functional Genomics. Kluyver Academic Publishers; Dordrecht, The Netherlands: 1999. p. 370

2. Andrade MA, Sander C. Bioinformatics: from genome data to biological knowledge. Curr Opin Biotechnol. 8:675–683. [PubMed: 9425655]

3. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S. Structural genomics: beyond the human genome project. Nat Genet. 23:151–157. [PubMed: 10508510]

4. Boguski MS. Biosequence exegesis. Science. 1999; 286:453–4555. [PubMed: 10521334]