Contents

A Bioinformatics: The Information Technology of Living Things

An Introduction by Prof. Sugata Mitra

One of the important properties of living things is their ability to reproduce. Living things are able to make copies of themselves: not perfect copies, but close enough. The subtle differences between parents and offspring are the reason for evolution and biodiversity. If we could understand this copying process, we would know a lot about ourselves, perhaps even about the purpose of our existence.

The process of copying anything always involves a transfer of information. For example, in a photocopier, a paper document is scanned by a beam of light and the patterns of black and white in the reflected light are stored. This is the information that is then used to direct ink to the right places on a blank piece of paper so that a copy of the original document is produced. However, even in the best photocopiers, there are always subtle differences between the original document and the copy, all caused by tiny random errors of movement or ink application. Transformation of information from one form to another always has some error associated with it, just as the transformation of energy from one form to another will always have some waste. Indeed, energy and information are connected through these 'errors' and the concept of entropy in thermodynamics ties the two together. It is this connection between information and energy that makes life possible on earth. Life, as we know it on earth, is dependent on processes that convert information into chemical energy and eventually into the chemistry of macromolecules. This was unknown until the discovery of deoxyribonucleic acid (DNA).

DNA is a single, huge molecule, sometimes up to two metres long. It consists of two long strings of smaller molecules, called nucleotide, wound up against each other in a structure, now famous, called the *double helix*. The two strings of the double helix are connected to each other, periodically, by a set of four molecules or bases—adenine, guanine, cytosine, and thymine. The structure of DNA is like a ladder, twisted around to form a helix shown in Fig. A.1. The steps of this ladder are formed of pairs of molecules-adenine with thymine, or guanine with cytosine (AT or GC in short) (see Fig. A.2). There are no other connections allowed. Of course, an AT connection can also be seen as a TA connection, depending on which way it is looked at. The same holds for GC and CG. DNA molecules exist inside every cell nucleus in a living organism. A two-metre long molecule of DNA is coiled tightly and packed into the nucleus of every human cell.

If all the AT and GC pairs (called base pairs) in a section of DNA were to be opened, the unmatched base pairs could be thought of as a string of characters, for example, AAAGTTCTTCTTCAATTA. This, of course, would be matched on the other string by its complements, that is, TTTCAAGAAGAAGTTAAT. Such strings of 'alphabets' contain the procedure for assembling a certain sets of proteins from the necessary amino acids after several intermediate processes. These proteins eventually make up most of the organism. In other words, the sequences of base pairs in DNA are the 'raw materials' as well as 'instructions' for building and maintaining the organism. Sequences of base pairs, often millions of alphabets in number, are grouped together into genes, much as the English alphabet are grouped together into words, sentences, and paragraphs in order to make sense. When a gene inside a DNA molecule needs to be activated, the DNA molecule in a cell nucleus uncoils and unfurls to just the right extent to expose that gene. At this time, another molecule, RNA, is formed.

Fig. A.1 DNA double helix

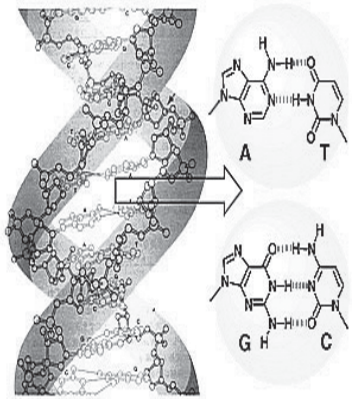

Fig. A.2 Zoomed view of double helixed DNA with side chains

Ribonucleic acid, or RNA, is thought to be even more fundamental to life than is DNA. A simple understanding of the structure of RNA would be to imagine a DNA double helix

that has been partly separated into a section with the broken ends of the base pairs sticking out of them. Free peptides and the A, T, G, C molecules floating in the nucleus are attracted to this open section of the DNA. Here they form an RNA molecule. For example, if the open section of the DNA contained the sequence CCG, the corresponding section of the RNA that is formed would contain the sequence GGC. In effect, the RNA that is formed would be a 'print' of the open DNA section. Such RNA is called messenger RNA or mRNA. However, RNA and DNA differ in an important respect. RNA does not contain the base thymine (T). Instead, it contains another base called uracil, coded as U. This is because uracil, though similar to thymine, takes less energy to bind and is, therefore, a more efficient form of copying, provided the copy does not need to last for long. This is indeed the case, as the mRNA has a very short-lived and specific purpose.

Once mRNA is formed, the open section of the DNA molecule closes, the gene is no longer exposed, and its copy, in the form of the mRNA, travels out of the cell nucleus into the cytoplasm. Here it encounters the ribosome, a relatively large structure made mostly of rRNA (another form of RNA) and some proteins to hold the mass together. Ribosome is capable of a form of biological computation. It 'reads' the base sequences on an mRNA molecule, and for every three bases it produces an amino acid from the cytoplasm (for example, for three bases CGC, it produces arginine).

Note

Amino acids are compounds that have an amine group at one end and a carboxyl group at other end. Proteins are formed by the polymerization of amino acids where the carboxyl and amine groups of adjacent amino acids bond together, releasing a molecule of water. There are 20 standard amino acids that are active in human beings. Each of these amino acids is related to one or more codons. You can refer Table 1 for a list of 20 amino acids and their related information.

A set of three consecutive bases in RNA is called a *codon*. In effect the ribosome is a process control computer that takes codons as the input and produces amino acids as the output. The amino acids are brought to the ribosome by another form of RNA, called transfer RNA or tRNA. These are relatively short RNA molecules that float freely in the cytoplasm with an amino acid attached to one end. When a ribosome detects a codon in an mRNA, it attracts that tRNA molecule which is 'carrying' that amino acid that the codon is referring to. Once close to the ribosome, the tRNA releases the amino acid. As a result, a sequence of codons results in a sequence of amino acids, which join together to form a polypeptide chain or a protein. Eventually, masses of intricately folded proteins form the gross structures such as skin and hair of living organisms. While it is usual to consider the human brain as a powerful central computer that controls all our actions and functions, it is interesting to note that our life processes are far more dependent, at a fundamental level, on the information processing that takes place in hundreds of millions of ribosomes. Life, in this context, appears as an immense distributed computing system.

Table 1 List of 20 amino acids and their related information

Amino acid	Abbreviation	Side chain	Hydrophobic	Polar	Charged	Small	Tiny	Aromatic or aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
										GCU	
Alanine	(Ala, A)	$-CH_3$	X	-	-	X	X	-	67	GCC GCA GCG	7.8
Cysteine	(Cys, C)	$-CH_2SH$	X	-	-	X	-	-	86	UGU UGC	1.9
Aspartate	(Asp, D)	$-CH_2COOH$	-	X	Negative	X	-	-	91	GAU GAC	5.3
Glutamate	(Glu, E)	$-CH_2CH_2COOH$	-	X	Negative	-	-	-	109	GAA GAG	6.3
Phenylalanine	(Phe, F)	$-CH_2C_6H_5$	X	-	-	-	-	Aromatic	135	UUU UUC	3.9
										GGU	
Glycine	(Gly, G)	–H	X	-	-	X	X	-	48	GGC GGA GGG	7.2
Histidine	(His, H)	$-CH_2-C_3H_3N_2$	-	X	Positive	-	-	Aromatic	118	CAU CAC	2.3
										AUU	
Isoleucine	(Ile, I)	$-CH(CH_3)CH_2CH_3$	X	-	-	-	-	Aliphatic	124	AUC AUA	5.3
Lysine	(Lys, K)	$-(CH_2)_4NH_2$	-	X	Positive	-	-	-	135	AAA AAG	5.9
										UUA	
										UUG	
										CUU	
Leucine	(Leu, L)	$-CH_2CH(CH_3)_2$	X	-	-	-	-	Aliphatic	124	CUC CUA CUG	9.1

(*contd*)

Table 1 *(contd)*

Amino acid	Abbreviation	Side chain	Hydrophobic	Polar	Charged	Small	Tiny	Aromatic or aliphatic	van der Waals volume	Codon	Occurrence in proteins (%)
Methionine	(Met, M)	$-CH_2CH_2SCH_3$	X	-	-	-	-	-	124	AUG	2.3
Asparagine	(Asn, N)	$-CH_2CONH_2$	-	X	-	X	-	-	96	AAU AAC	4.3
Proline	(Pro, P)	$-CH_2CH_2CH_2-$	X	-	-	X	-	-	90	CCU CCC CCA CCG	5.2
Glutamine	(Gln, Q)	$-CH_2CH_2CONH_2$	-	X	-	-	-	-	114	CAA CAG	4.2
Arginine	(Arg, R)	$-(CH_2)3NH-C(NH)NH_2$	-	X	Positive	-	-	-	148	CGU CGC CGA CGG AGA AGG	5.1
Serine	(Ser, S)	$-CH_2OH$	-	X	-	X	X	-	73	UCU UCC UCA UCG AGU AGC	6.8
Threonine	(Thr, T)	$-CH(OH)CH_3$	X	X	-	X	-	-	93	ACU ACC ACA ACG	5.9
Valine	(Val, V)	$-CH(CH_3)_2$	X	-	-	X	-	Aliphatic	105	GUU GUC GUA GUG	6.6
Tryptophan	(Trp, W)	$-CH_2C_8H_6N$	X	-	-	-	-	Aromatic	163	UGG	1.4
Tyrosine	(Tyr, Y)	$-CH_2-C_6H_4OH$	X	X	-	-	-	Aromatic	141	UAU UAC	3.2

A somewhat lyrical (and not too accurate!) description of the process of life

You have a small cut on your finger. The skin is broken and so are some minor blood vessels. There is some bleeding. The blood clots quickly and the bleeding stops in a few minutes. Clotting is a physical property of blood useful for cuts, although it can be deadly when it happens internally. Your bleeding has stopped but the wound has to heal. The cells surrounding the cut begin to react. The DNA in their nuclei unfold and twist to expose those genes that have the instructions for building new skin. Floating molecules fit themselves to the open genes to form strands of mRNA that float out of the nuclei into the cytoplasm of the cells. Here they are scanned by floating ribosomes that process the codons in the mRNA and attract other floating tRNA from the cytoplasm. The tRNA bring with them the amino acids that will polymerize to form proteins. Wisps of proteins twist and fold to form structures that eventually join to form new cells—cells of skin. In a few days, the wound heals, as though it had never been there at all.

A.1 Bioinformatics: When and Why

It was in the 17th century that biologist started dealing with problems of information management. Early biologists were preoccupied with cataloguing and comparing species of living things. By the middle of the 17th century, John Ray introduced the concept of distinct species of animals and plants and developed guidelines based on anatomical features for distinguishing conclusively between species. In 1730, Carolus Linnaeus established the basis for the modern taxonomic naming system of kingdoms, classes, genera, and species. Taxonomy was the first informatics problem in biology. The University of Arizona's Tree of Life project and NCBI's taxonomy database are two examples of online taxonomy projects.

From Mark S. Boguski's article in *Trends Guide to Bioinformatics* Elsevier, Trends Supplement 1998, p1, '…. So bioinformatics has, in fact, been in existence for more than 30 years and is now middle-aged.'

Growth in laboratory technology facilitated collection of data at a rate that was faster than the rate of data interpretation. Biologists reached a similar information overload and started facing a lot of difficulties in the field of data analysis and interpretation. Collecting and cataloguing information about individual genes (approx. 30,000) in human DNA and determining the sequence of three billion chemical bases that made up the human DNA became the second informatics problem in biology.

In 1990, Human Genome Project was initiated as a prominent bioinformatics solution to the problem, and this labelled the 21st century as the era of genomes.

A.2 What is Bioinformatics?

Fredij Tekaia at the Institute Pasteur offers the definition of bioinformatics as 'The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.'

From *Introduction to Bioinformatics* by T.K. Attwood and D.J. Parry-Smith (published by Prentice Hall in 1999): 'The term bioinformatics is used to encompass almost all computer applications in biological sciences…'

As per Cynthia Gibas, Bioinformatics is the intersection of information technology and biology. It was in 1730 that Carolus Linnaeus established the basis for the modern taxonomic naming system of kingdoms, classes, genera, and species. Taxonomy was the first informatics problem in biology.

Bioinformatics is a highly interdisciplinary field of biology, relying on basic principles from computer science, biology, physics, chemistry, and mathematics (statistics). The subject facilitates biological information to be handled by the use of computers. Or to be more precise, bioinformatics is the subject that involves the use of techniques from all these subjects to deal with problems of biology, understand biological processes, and find methods and solutions with information technology to solve biological problems.

Harold Morowitz believes that 'computers are to biology, what mathematics is to physics.'

The following four statements sum up the use of IT in bioinformatics:

1. Database management is used to store, retrieve, analyse, and/or predict the huge biological data.
2. Software development is used for implementing algorithms and developing applications and tools for insilico experiments.
3. CAD 4 multimedia are used for developing user interfaces, static/dynamic web pages, and graphic representations of data for prediction and visualization.
4. Next generation operating systems, networking, and software can facilitate seamless data transfer and execution of tools.

A.3 In This Book

As mentioned before, bioinformatics deals with databases. Among many varieties of databases, sequence databases are records of the sequence of bases found in DNA, sequence of amino acids, and protein sequences. While such databases are essential for storing the vast information collected from labs, their real use lies in the interpretation of the stored data. Searching a base sequence database for genes forms an important task in bioinformatics. A secion of DNA may be a part of a gene, or it may contain many genes. Software applications are used to determine where genes lie in the sequence databases of DNA with searching techniques. Once a gene is identified, the amino acid and protein sequence that it generates can be visualized and modelled. How proteins fold is an important, unsolved problem. In order to do all of this efficiently, new algorithms need to be developed and new software or tools written. The creation of tools is an important and emerging area in bioinformatics. While consolidating the learnings of a new subject such as bioinformatics what emerged as broad sections (called parts in this book) are databases, tools, and algorithms. Each part begins with a part-introducting chapter, followed by chapters comprising that part. The content and coverage of these parts is as follows.

Part A: Introduction This is an introductory part of the whole book and does not have its own introduction chapter or related chapters unlike other parts.

Part B: Databases This part talks about the characteristics and categories of Bioinformatics databases wherein the technical design aspect is given more importance than other database categories. This will not only help to understand any new database you come across, it will also help you to design your own mirror database in case you need to create one, with a different structure of an already existing public database. You may also create a subset of an existing database and extend that as your own private database by adding data from your own experiments. This part also explains how you can navigate in a database and retrieve information that you need. This part has three main chapters after part introduction, namely, Sequence Databases, Structure Databases, and Other Databases.

Part C: Tools The need for tools, their role in knowledge discovery, and various types of bioinformatics tools are discussed in this part. All these tools are software applications developed specifically for the databases to perform specific tasks. The chapters in this part are Data Submission Tools, Data Analysis Tools, Prediction Tools, and Modelling Tools. Each chapter discusses systematically one or two samples of each type of tools. The purpose of the tool, the set attributes, desired input, and expected output with a broad analysis is also discussed in the chapters. Customized tools can be developed to suit specific needs of research using specialized programming languages such as BioCORBA, BioJava, BioPerl, BioPython, BioXML, CellML, GEML (Gene Expression Markup Language), SBML (Systems Biology Markup Language). However it is a separate area of study altogether and is not covered in this book.

Part D: Algorithms This part discusses major algorithms that are used in bioinformatics tools for various bioinformatics tasks. These algorithms are categorized based on the methods of algorithm design as well as on their implementation. Understanding these algorithms will help one to know the method adopted to perform a task as well as help decide which tool to select to get the specific input. The chapters included in this part are Data Analysis Algorithms and Prediction Algorithms. You may learn Part D at your own pace and you do not need Internet connectivity to study and understand the content in this part.

Each chapter starts with chapter objectives and comprises general content, analogies, tasks, notes, examples, and exercises. While all of these are self-explanatory, tasks are step-by-step procedures that you can follow to explore a database or tool to get their respective outputs. This feature actually strengthens ones confidence by giving step-by-step route to explore or solve a problem without leaving any chance for getting defocused. The respective output ensures that you are going in the right direction. Thus, it is recommended that while you go through a chapter, read the concepts first and then sit at your computer (with internet connectivity) to explore the database and tools.

A.4 Bioinformatics: Applications and Research

With adequate data and right tools, it is possible to explore a number of new areas in biology. Chief among these are given below.

Biodiversity In addition to species and physical measures, bioinformatics provides a genetic measure of biodiversity. Eventually, it is the genetic diversity in a biosphere that will provide the correct measure of its extent.

Analysis of gene expression A physical or chemical change in a living system is not caused by a single gene but by the combined effect of many genes. Understanding the action of many genes on a single condition will, one day, provide a genetic basis for disease and change control.

Analysis of gene regulation Regulation is the chain of events, starting with an extracellular event (such as temperature change) and leading to a change in the activity of proteins. The analysis of what promotes and regulates the activity of genes and proteins forms a part of this study.

Comparative genomics By comparing the genes of different organisms, it is possible to trace evolutionary pathways by which one organism could have evolved into another. Such studies can not only throw new light on evolution, but also provide evidence for the migration of species, thereby bringing new evidence to historical and anthopological studies.

Molecular medicine More emphasis needs to be given on tracing the fundamental cause of diseases rather than treating symptoms. Specific diagnostic tests and faster generation of test reports can improve the situation. It is believed that gene testing and thereby through gene therapy may even make replacement of defective genes feasible.

Microbial genomics The genomes of bacteria can help throw light on energy sources, environmental monitoring to detect pollutants, find disease-producing properties of genes, and improve industrial efficiency.

Risk assessment More research on human genome can help to assess individual risk on exposure to toxic elements as resistance to external agents varies from person to person. It can also help to reduce the likelihood of heritable mutations.

Bioarchaeology, anthropology, evolution, and human migration Understanding human and other genomes will help to understand evolution, inheritance, traits, and disease carriers. The study of the genome comparison across organisms can help to understand similar genes with associated disease.

DNA forensics (identification) DNA profile of an individual, called DNA fingerprints, can help in identifying criminals, establishing family relationships, protecting rare wildlife species, and matching organ donors.

Agriculture, livestock breeding, and bioprocessing Genome research on plants can provide nutritious, disease-resistant, pesticide-free crops. Even alternate use of crops can be found, e.g., tobacco has been found to produce bacterial enzymes that break down explosives such as TNT and dinitroglycerin.

Bioinformatics research is a multidisciplinary and vast area. It ranges from drawing a mathematical or physical model of a biological system to implementing data analysis algorithms to develop databases and web tools to access them. The large number of informatics tools and data resources already available or still being developed need to be fully integrated and coordinated. Drug design is one of the major areas of research that is gaining importance in academics as well as pharmaceutical companies. In this direction, a major challenge is to integrate genome and genome-related databases to design common interfaces, implementing 'minimonolithic' databases containing subsets of relevant data extracted from a set of larger public databases. Major scope lies in the improvement of laboratory-systems integration and information-management systems to promote large-scale genomics and other biology programs in academia and industry.

To sum up, the sky is the limit, and bioinformaticians have ample scope for research in specialized fields within bioinformatics, e.g., informatics, cheminformatics, proteomics, genomics, etc. It is not only biology that stands to benefit from bioinformatics. The genetic code is a language. It is the first language of non-human origin that we have encountered; it is a truly alien system. Ironically, it has come to us not from outer space, but from within ourselves. Understanding the structure and function of this language is vital to our understanding of this language, of communication and, eventually, of sentience.

As discussed RNA protects itself by replacing the uracil base with thymine and twisting itself into the DNA molecule. It then copies itself (mRNA), reads itself (rRNA), and assembles itself (tRNA) to perpetuate a cycle of self-referential, self-organized, reproduction. The evolution of life on earth, looked at from this perspective, appears as an incredibly efficient attempt by one molecule, RNA, to survive. The key to this immense process of survival lies in the connections between information science and biology. You will learn more about the present bioinformatics scenario in India in the following section.

A.5 Present Bioinformatics Scenario in India

The following subsections discuss the present bioinformatics scenario in India as also the role of various organization in this field.

A.5.1 Government Organizations

In 1986, the Department of Biotechnology (DBT) launched the Biotechnology Information System (BTIS), during the 7th five-year plan. It is a nationwide network with ten distributed information centres (DICs) and 48 distributed information subcentres (sub-DICs). Its mission was to establish India as a leader in bioinformatics. In the 8th five-year plan, BTISnet was established—a distributed database and network infrastructure. It comprised the above as well as six national facilities for high-end interactive graphics and molecular modelling. All these centres were connected through satellites and terrestrial links under two major network service providers—NICNET and ERNET. They also use X.25 links of DoT/VSNL for international access.

These DICs and sub-DICs have been given the task of providing discipline-oriented information to all institutions interested in relevant fields. The information catered to the major areas of research. A list of various disciplines with respective DICs responsible for the same is given below:

1. Genetic engineering: Jawaharlal Nehru University, New Delhi; Indian Institute of Science, Bangalore; Madurai Kamraj University, Madurai; Bose Institute, Calcutta.
2. Protein modelling and protein engineering: Institute of Microbial Technology, Chandigarh.
3. Plants tissue culture, photosynthesis, and plant molecular biology: Indian Agricultural Research Institute, New Delhi.
4. Animal cell culture and virology: University of Pune, Pune.
5. Oncogenes, reproduction physiology, cell transformation, nucleic acid and protein sequences: Centre for Cellular and Molecular Biology, Hyderabad.
6. Immunology: National Institute of Immunology, New Delhi.
7. Neuro-informatics: National Brain Research Centre, Delhi.
8. Terra flops supercomputing facility: Supercomputing facility for Bioinformatics and Computational Biology, at IIT Delhi.
9. Biotechnology and other related fields: Department of Biotechnology, Government of India, New Delhi.

A.5.2 Private Organizations

There are a lot of international biotech companies such as Strand Genomics, Jubilant Biosys, etc., with their offices in India, with plans to harness Indian manpower. There are also Indian based companies such as Biocon India, Informatics Pvt. Ltd, Advance Biochemical Lab, working as contract research organizations for companies abroad as well as companies involved in customized software development work for India. There are pure Information Technology (IT) companies heading towards bioinformatics by collaborations with government organizations. Few such companies are TCS, Nicholas Piramal, Satyam, with collaborative ventures with the Centre for DNA Fingerprinting and Diagnostics (CDFD), Centre for Biochemical Technology (CBT), and Center for Cellular and Molecular Biology (CCMB), respectively. Spectramind eServices is one of the IT enabled service (ITES) companies where professionals are given to analyse biological literature and create databases. GE call centre also has plans on similar lines. Other companies such as Avestha Gengraine, Mahindra-BT, and DSQ software also ventured in the field with a people-intensive business model for the bioinformatics service sector.

Almost all Indian pharmaceutical companies are into bioinformatics research and Development. Some of them are Ranbaxy, Dr. Reddy's Lab, Dabur, Smith Klein Beecham, Pfizer, Cipla, Zydus Cadila, Wockhardt, Astra Zeneca, and East India Pharmaceuticals, Pharma and pharma-related companies in India have a unique and exciting opportunity of

growing their annual revenue from $5.5 bn in 2000 to $25 bn in 2010 see Table 2. Of this, $1.5–2.0 bn can accrue from IT-related new horizon areas including bioinformatics, genomics/proteomics, data management for contract research, and remote sales and marketing. And $20 bn from the growth in the current core business and the rest $3 bn from emergence of a new horizon of non-IT innovation. However, success in IT-related fields (especially informatics) will be difficult and will require a focus on biological knowledge and innovation and not just investment.

Table 2 Projected IT-related revenue for year 2010 from pharma and pharma-related companies

Total IT-related revenue	$1.5–2.0 bn
Informatics research	~$0.7 bn
Development (data management)	~$0.5–0.6 bn
Sales and marketing	~$0.3–0.7 bn

Trends that can create new and exciting opportunities for India, especially through the proliferation of IT-related technologies, are the following.

Dramatic technologies Global bio-pharmacos are keen to foman alliance with Indian companies, especially for genomics, proteomics, and bioinformatics. Companies in India can provide contract research services in bioinformatics and cheminformatics.

Global harmonization There is a need to reduce costs across the business system, globally. To meet this need, outsourced IT services are likely to be better supported, provided significant quality and confidentiality can be ensured.

Cost containment Global pharmacos are finding it acceptable to source from non-domestic markets. Their growing demand for low-cost R&D, especially IT-related work can be off-shored from India.

India, thus, becomes the preferred destination for pharma and bioinformatics companies for their network with global academic and biotech community. Thus annual profit and market capital of IT-related Indian pharma companies have the opportunity to grow five times, i.e., from $1 bn to $5 bn and ten times, i.e., $15–20 bn to $150–200 bn, respectively. As pharmaceutical research is in the midst of a paradigm shift towards gene-based drug discovery where proteins are tailor made, companies are entering into alliances to access emerging research technologies.

A.5.3 Informatics Research

In research, India can focus on several options with the main focus on informatics. The strengths on which India can compete is skill, infrastructure, availability of research sample, and cost competitiveness. As said by a director of a premier research institution in India, 'Biotech is one of the areas where India can compete…To create a biotech hub in India

is a dream but it is one worth pursuing'. Different research areas in the descending order of Indian strengths are informatics, synthetic and medicinal chemistry, structure-based drug design, positional cloning, and expression profiling and proteomics. Of these, informatics and structure-based drug design can give a higher profitability.

Looking from the IT perspective, research in informatics has broadly three categories.

Research enterprise application service providers (ASPs) They provide a user-friendly interface and access to bioinformatics and cheminformatics data (public and proprietary) as well as analytical solutions. Such companies require data acquisition assets along with bioinformatics capabilities. In India, bioinformatics skills are available at (1/3) rd cost. Other than US and Western Europe, Ireland is one of the potential competitors of India. India's global share in this technology is about 15%.

Software application developer and service providers They provide analytical solutions to discrete stages of the discovery process. They require extensive IT and bioinformatics capabilities. India's superiority in mathematics and statistics allows informatician to think at different levels. India's global share compared to the emerging competitors such as Hungary, other than US and Western Europe, is 15%.

Database content providers These companies provide access to proprietary and public data on genomics, proteomics, and combinatorial chemistry. They require extensive IT capabilities. The genetic diversity in India is of tremendous advantage for the generation of genomic databases. With the US and Western Europe as dominating players, India's other emerging competitors are Iceland, Ireland, and China, with only 5% global share. India can capture $400–500 million IT-related data management opportunity for clinical trials.

Table 3 shows the likely Indian market capital of informatics for year 2010.

Table 3 Global versus Indian market capital of information for year 2010

Informatics research areas	Global market (in million dollars) Year 2010	Indian market share	
		In %	(in million dollars)
Research enterprise application service providers	500	15%	75
Software application developer and service providers	4100	15%	615
Database content providers	900	5%	45
Total	**5500**	**13.36%**	**735**

A.5.4 Data Management

Data management is rapidly evolving into an IT-driven paperless activity. Earlier data management was done using software installed on dedicated servers, which were batch processed. This resulted in delay in query processing and report generation. Now it involves browser-based programs which are machine and operating system independent. Also data processing is in real time, enabling receiving and resolving queries in real time. The key activities are purely IT, starting from the development of the database to data capturing, analysis, and monitoring. To support all these, the key technology enablers are electronic data capture, point-of-care data collection, point-of-entry queries, data warehousing, and 'real-time' data monitoring. India can leverage its IT skills and emerging track record to participate in this opportunity. Database service is an easier and quicker revenue opportunity. It involves annotation of gene and protein sequences for large database providers and pharmacos and curation of scientific literature on an outsourced basis.

A.5.5 Academic Scenario in India

India is gearing up to improve the academic scenario in the field of bioinformatics. DBT has started one-year Advance Diploma in Bioinformatics in five Indian universities—Jawaharlal Nehru University, Calcutta University, Pune University, Madurai Kamraj University, and Pondicherry University. The admission to the course is through All India Entrance Examination, conducted by the respective centres. The Union Government is planning to set NBI under the DBT in the 10th five year plan, which will be on the lines of the National Centre for Bioinformatics under the National Institute of Health (NIH) of USA. In the 10th plan, there is a substantial increase in the allocation for bioinformatics, i.e., nearly 1.2 billion rupees as against a total of about 300 million rupees during the 9th plan. The proposed NBI (the location of the centre is yet to be decided) is planned to regulate bioinformatics research in India. It would have separate wings for various bioinformatics activities such as software and database development, human resources, genomics, proteomics, and services. Indian Institute of Technology (IIT) Delhi conducted certificate courses in supercomputing facility and plans to offer M.Tech in bioinformatics.

Other than what we have from the DBT charter, there are other universities as well with their own courses and resources development planning. Bharathiar University, Coimbatore, offers a two-year MSc course in bioinformatics. Karnataka Government in association with ICICI has set up the Institute of Bioinformatics and Applied Biotechnology (IBAB) at Bangalore offering one-year diploma course. These courses are not DBT authorized.

Bioinformatics Institute of India (BII), the only institute providing a distance learning programme (DLP) on bioinformatics in the country, plans to invest 150 million rupees to open 500 centres by the year 2003. This is to fulfil the global biotech market share that is expected to cross US$ 500 billion by the year 2010.

A.5.6 IT Contribution: Skills

Presently institutes providing bioinformatics as primary degree course or offering it as an elective subject are providing students with basic knowledge of molecular biology, protein (bio) chemistry, and evolutionary biology for computer science students. On the contrary, biology students need to cover programming, optimization, cluster analysis, C, C++, Java, Perl, Fortran, CGI scripts, Linux, RDBMS such as Oracle/ Sybase, Maths including statistical techniques and calculus. There are a host of new languages tailored for bioinformatics need, e.g., BioCORBA, BioJava, BioPerl, BioPython, BioXML, Cell Modelling (CellML), gene expression markup language (GEML), systems biology markup language (SBML), etc.

The job profiles for non-biology bioinformatics science graduates are the same as that of IT sectors. They are system analyst, system engineer, software engineer, scientific programmer, application analyst, application programmer, database administrator, database designer, database programmer, database developer, database services, network administrator, technical support, marketing, etc. Bioinformatics tools to be handled by non-IT bioinformatics science graduates are production and data submission tools, data mining tools, research tools, analysing tools, annotation tools, map integration tools, visualizing tools, etc.

Conclusion

The discussions so far highlight the tremendous scope in the field of bioinformatics. Low cost Indian IT and biology workforce are two of the key success factors to achieve the target. Consultant scientists with a network of global scientific alliances and scientific advisory board is likely required. It is widely recognized that India has expertise in software and a good pool of qualified and experienced biologists. In this context, India needs to get its act together by way of helping create a workforce that is capable of integrating the expertise of software developers and biologists to achieve the objective.