



\* **Biotechnology For You**

**Welcome to Biotechnology For You!**

BFY or Biotechnology for you is a magazine dedicated to all of them working in the field of life sciences. BFY will update you with the current news, research, jobs, software, games and many more exciting articles from Biotechnology sector. We request you to enjoy this and let us know your feedback.

May - 2006

***In This Issue..***

- \* **Human Genome Project - A Tutorial**
- \* **Book review**
- \* **Software Tutorial**

**Biotechnology For You**

# Human Genome Project – A Tutorial

## Introduction to the Human Genome Project and Molecular Biology

### Introduction to the human hieroglyphic

On April 15, 2003, the International Human Genome Sequencing Consortium (IHGSC) – the association of laboratories from around the world which had jointly undertaken the Human Genome Project (HGP) formally announced the completion of the project and the colossal task that lay at its core: the sequencing and assembly of the 3.2 billion bases that comprise the human genome. This is a truly landmark achievement for science and medicine. In the words of Elbert Branscom, Founding Director of the Joint Genome Institute (JGI), "We will see everything before this like the dark ages of biology".



### History of the HGP

Who were the thought leaders behind the project? What were the sequence of events that eventually led to the Human Genome Project? The answers to these questions takes us back almost hundred years into the history of the evolution of modern biology from the times of Thomas Hunt Morgan (1866–1945) whose laboratory first produced genetic maps of the fruitfly (*Drosophila melanogaster*) in 1911. From that first step when individual researchers painstakingly teased out small bits of information from short stretches of DNA to the present times when such work is routinely performed in an automated fashion at the rate of millions of bases a day, there have been many seminal discoveries that have spurred such dramatic advances. These include the elucidation of the double-helix structure of DNA through the work of Rosalind Franklin, James Watson, Francis Crick and Maurice Wilkins in 1953 (), the development of DNA sequencing methods by Frederick Sanger (Sanger et al., 1977) which allowed researchers to precisely determine the order of individual bases in DNA, the Polymerase chain reaction (Saiki et al., 1985, 1988) which provided a means to amplify DNA from minuscule amounts of starting material, to the advent of modern day capillary sequencing technology that transformed laboratories into high-throughput sequencing factories overnight.

We will catch the thread from the 1980's onwards when the first suggestion to sequence the human genome was made. The proposal came from Robert Sinsheimer in 1984, then the Chancellor of the University of California at Santa Cruz (UCSC). Sinsheimer was the former chairman of the Division of Biology at the California Institute of Technology which was founded by none other than Thomas Hunt Morgan more than 50 years ago. Sinsheimer was an accomplished Molecular Biologist and his ideas struck a chord with his contemporaries both within and outside UCSC. A meeting was called in May 1985 to discuss the proposal at what is known as The Santa Cruz Workshop. The workshop included leading researchers like David Botstein Leroy Hood, Robert Ludwig, Kivie Moldave, Harry Noller, John Sulston and Walter Gilbert among others and each of

Biotechnology For You

them had already made important contributions to the emerging field of Genomics.

Walter Gilbert and his graduate student Allan Maxam had earlier devised an alternate method for sequencing DNA in 1977, which as opposed to Sanger's enzymatic mechanism, was based on chemical degradation of double-stranded DNA. Gilbert, Sanger (and Paul Berg) shared the 1980 Nobel Prize in Chemistry for their fundamental studies and contributions towards the biochemistry of nucleic acids and their sequence determination. David Botstein had devised strategies for the construction of a genetic linkage map of the human genome which would enable the detection of polymorphisms and for studying genetic traits in humans by using a linkage map of restriction fragment length polymorphisms (RFLPs). John Sulston was an expert on the biology and genetics of the nematode worm, *Caenorhabditis elegans*, which would later be completely sequenced as an important model organism.

*The Santa Cruz Workshop set the idea of sequencing the entire human genome in motion and was followed up by efforts by other leading researchers in the field. A major contributing factor why the idea immediately gained momentum was the clear recognition among many scientists of the benefits of human genome sequencing. In 1986, Nobel Laureate, Renato Delbuco, who had done seminal work on oncogenic viruses and their effects on malignant transformation of the cells they infected, expressed his strong support for the program in an article in Science stating that the information would lead to an understanding of the causes of cancer. There was also an opposing view forwarded by many scientists borne out of a concern about the ethical issues surrounding the use of genetic information. However, the overwhelming advantages of the project in terms of betterment of human health and medicine ultimately triumphed over the moral issues and provided a strong impetus to take the project beyond the conceptual stage. The next major step in the process was taken by Charles DeLisi who was then the Associate Director for Health and Environmental Research at the US Department of Energy (DOE). DeLisi convened a meeting in Santa Fe in 1986 to outline the role of the DOE in sequencing the human genome and actually undertook the task of assessing the cost of the project. In 1987, as a result of his efforts, the DOE allocated 5.5 million dollars to the program.*

*The effort was also spurred on admirably by parallel advancements in sequencing technology. One of the most important developments was the automated sequencer invented by Leroy Hood. Hood founded Applied Biosystems Incorporated and (together with Lloyd Smith, Michael Hunkapiller and Tim Hunkapiller) fundamentally improved Sanger's method by replacing radioactive labeling of DNA with fluorescent dye tagging and by automating the process of base calling by integrating laser optics and computer technology. The automated sequencer became available commercially in 1986 and revolutionized Genomics by allowing the rapid sequencing of DNA. It was not much later, on October 1, 1990 that the Human Genome Project officially began.*

What is the International Human Genome Sequencing Consortium?

The consortium consists of 20 groups and represents countries like the United

States, the United Kingdom, Japan, France, Germany and China. The institutions are Whitehead Institute for Biomedical Research, Center for Genome Research (USA), the National Institutes of Health (NIH, USA), which created the National Human Genome Research Institute (NHGRI), US Department of Energy

Joint Genome Institute (USA), The Wellcome Trust Sanger Centre (UK), Washington University Genome Sequencing Center (USA), Baylor College of Medicine Human Genome Sequencing Center (USA), RIKEN Genomic Sciences Center (Japan), Genoscope (France), GTC Sequencing Center (USA), Department of Genome Analysis, Institute of Molecular Biotechnology, Beijing Genomics Institute/Human Genome Center (China), Multimegabase Sequencing Center, The Institute for Systems Biology (USA), Stanford Genome Technology Center (USA), Stanford Human Genome Center (USA), University of Washington Genome Center (USA), Department of Molecular Biology, Keio University School of Medicine (Japan), University of Texas Southwestern Medical Center at Dallas (USA), University of Oklahoma's Advanced Center for Genome Technology (USA), Max Planck Institute for Molecular Genetics (Germany), Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome center (USA) and GBF-German Research Centre for Biotechnology (Germany).

The IHGSC finished the sequencing project in time for the fiftieth anniversary of the discovery of the structure of DNA, more than two years ahead of schedule. The first draft of the sequence when the project was eighty-five percent complete was reported in 2000. Today, scientists have mapped 2.9 billion base pairs of human DNA to an accuracy of 99.99 percent and have estimated that human life is programmed by 30,000 odd individual genes that control all its molecular, cellular and organismal machinery.

**All those who were instrumental in bringing the HGP from concept to reality received recognition for their efforts by national and international bodies. UCSC named the building that houses its Molecular, Cellular and Developmental Biology research as the Sinsheimer Labs after Robert Sinsheimer. Charles DeLisi was recognized as the first government scientist to conceive and outline the feasibility, goals, and parameters of the Human Genome Project and received the Presidential Citizens Medal in 2001. Leroy Hood, currently President and Director of the Institute for Systems Biology was awarded the 2002 Kyoto Prize for Advanced Technology for his contributions to the rapid genome sequencing technology.**

Where are we going with the genome projects?

Much before the human genome was completely sequenced and assembled, scientists had realized the value of understanding the molecular basis of disease. For example, Doolittle et al., reported in 1983 of the similarity between the oncogene (a gene that contributes to cancer when mutated or expressed at abnormally-high levels), v-sis, of Simian sarcoma virus (an RNA tumor virus) and the gene encoding human platelet-derived growth factor (PDGF). The v-sis gene was the first oncogene to be identified as having homology to a known cellular gene. This discovery

provided the early insight into the critical role that growth factor signaling plays in the process of malignant transformation.

Biotechnology For You

Another example is the discovery that the defective gene that causes cystic fibrosis formed a protein that had similarity to a family of proteins that were involved in the transport of hydrophilic molecules across the cytoplasmic membrane. Cystic fibrosis

is the most common inherited disease in the Caucasian population and affects the respiratory, digestive and reproductive systems. It is now known that mutations in the cystic fibrosis gene leads to loss of chloride transport across the cell membrane, which is the underlying cause of the disease.

Today, our understanding of disease mechanisms is further enhanced by the availability of the complete sequences of several genomes. We have the sequences of almost 600 viruses and viroids, more than 30 bacteria, seven archaea, one fungus, two animals and two plants (rice and arabidopsis). The Human genome carries special significance for obvious reasons. Francis Collins, who followed James Watson as the director of the National Center for Human Genome Research (NHGRI), likened the Human genome to a book with multiple uses. In his words, "It's a history book - a narrative of the journey of our species through time. It's a shop manual, with an incredibly detailed blueprint for building every human cell. And it's a transformative textbook of medicine, with insights that will give health care providers immense new powers to treat, prevent and cure disease." Eric Lander, the Director of the Whitehead Institute/MIT Center for Genome Research called the Human genome the biological equivalent of the periodic table of the elements. Dr. Robert Waterston, head of Department of Genetics of the Washington University School of Medicine Genome Sequencing Center, said, "We have before us the instruction set that carries each of us from a one-cell egg through adulthood to the grave."

The HGP has already contributed significantly towards the goal it had set out to accomplish. An important discovery was reported on April 24 by scientists at the University of Michigan in Ann Arbor and the Clinical Research Institute of Montreal in Canada who simultaneously published the identification of a key gene (called Bmi-1) that regulates the growth of stem cells involved in hematopoiesis and the development of the immune system.

On April 16, the NHGRI announced the discovery of the genetic basis of progeria, an accelerated aging disease that causes children to age at 5 to 10 times the normal rate. NHGRI director Francis Collins commented: "The implications of our work may extend far beyond progeria to each and every human being. What we learn about the molecular basis of this model of premature aging may provide us with a better understanding of what occurs in the body as we all grow older."

To date, 1,400 disease genes have been identified; enabling the production of specifically targeted drugs and treatments. More than 350 new drugs derived from the HGP research are currently undergoing tests. Undoubtedly, the sequencing of the human and other genomes is just the beginning of the revolution that is unfolding right in front of our eyes. We are moving towards a paradigm shift in medicine, from just-in-time treatment that is given after the onset of symptoms to predictive and personalized treatment where the determination of the genetic factors predisposing an individual to disease is made right at birth and treatment started much before the onset of disease.

Biotechnology For You

## Book review

### 1- **Bioinformatics: Principles and Applications by Harshawardhan P Bal.—**

New Delhi: Tata McGraw-Hill  
Publishing, 2005. xiii, 217p.  
ISBN : 0-07-058320-x.

With the signing of the WTO by India, the Indian Pharma industry has to move its strategy from imitation to innovation. For India to be able to produce novel drugs, it has to spend more in R&D and vertically integrate its operations from basic research (laboratory based) to translation research (patient based). Bioinformatics has a key role to play in this transformation. To the millions of students and practioners of various domains that contribute to the research and drug development process, this book provides a solid foundation that enables them to combine existing skills and learn a new skills to perform truly high-throughput analysis fo datasets. This book is highly recommended as a practical companion book for people chemistry, biology, pharmacy, medicine, etc.

### 2- **BIOINFORMATICS (Principles and Applications) By- BAL, HARSHAWARDHAN, P.**

Product Details  
Publisher: Tata Mcgraw-Hill  
ISBN: 007058320X  
Edition: 01/e Paperback (Special Indian Edition)  
Publication Year: **2005**

#### Title Details

The International Human Genome Project (IHGP), accomplished in 2003, was undoubtedly a giant step for mankind in the sequencing and assembly of the entire human genome. This event has fundamentally altered the manner in which we analyze biological systems, our approach to the many unsolved problems, and indeed, our attempt to unravel the closely held mysteries of even the simplest of living organisms. However, deriving biological meaning from the raw stack of three billion bases of human DNA would not have been possible without the development of new algorithms for data mining and analysis. This book is about those fundamental tools and techniques that revolutionized biomedical research, and enable us today to perform biology in silico. The book uses an integrative approach to illustrate the use of these tools, and binds them together to create a coherent strategy to tackle the overwhelming problem of biological information overload. Divided into two parts, the book covers the core set of tools that have become indispensable to scientific discovery in the post-genome era, and also demonstrates how these tools can be integrated programmatically with BioPerl to be used in an enhanced, truly high throughput- biology on steroids- manner. With this coverage, the book will be useful to a diverse array of life science professionals, computational biologists, bioinformaticists, as well as students.

#### Table of Contents

##### Part I: Principles

1. Web-based Sequence Analysis: BLAST I
2. Web-based Sequence Analysis: BLAST II
3. Web-based Sequence Analysis: BLAST III

Biotechnology For You

## Book review

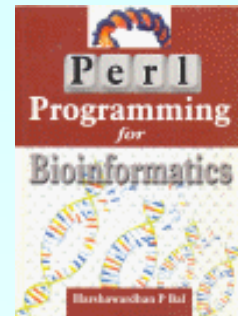
4. Web-based Sequence Analysis: Gene Prediction
5. Web-based Sequence Analysis: HMMER
6. PSI-BLAST

### Part II: Applications

7. Accessing Sequence Information Using BioPerl
8. Bio::DB::GenBank
9. Accessing GenBank Data
10. BioPerl BLAST Modules
11. Parsing BLAST Output

### 3- PERL PROGRAMMING FOR BIOINFORMATICS By BAL, P. HARSHAWARDHAN

ISBN: 0070474478  
Publisher: Tata McGraw-Hill  
Language: English  
List Price: Rs. 595.00  
Year Of Publication: 2002



#### Description:

Why does one person respond to a certain medication while another does not? Why are some individuals more susceptible to an infectious disease than others? These are the kind of mysteries that biologists are now trying to unravel. The next few decades will be completely consumed in research that leads to answers to these issues. Bioinformatics is the solution. Recent developments in fledgling fields of Genomics and Proteomics brought about by the Human Genome Project have given birth to Bioinformatics a new scientific discipline at the crossroads of biology, medicine and information technology. The sequencing of whole genomes such as *Homo sapiens*, *Arabidopsis thaliana* (thale cress), *Caenorhabditis elegans* (worm) and *Drosophila melanogaster* (fruit fly), among many others is only the first step towards the discovery process that is aimed at devising new methods and technologies to address our unmet medical needs. This book introduces Bioinformatics and its tools and techniques to a new hybrid professional the biologist and the would-be informatician and encourages his/her to use this vast and powerful resource to address biological problems. The book has been designed keeping in mind Unix and Windows 95/98/2000/XP operating systems and does not assume prior specialized knowledge in either of the operating systems. Biology or computing and all teaching material is formulated from the ground-up so that professionals from diverse backgrounds can follow the content. Relevance for a biologist who traditionally used to analyzing his or her favorite sequence(s) manually, this book offers methods of performing such analysis in a truly high-throughput fashion, giving the ability to analyze hundreds and thousands of sequences at once using the power of Perl. For an IT professional, the book provides a vignette into the world of biology into the universe of genes and proteins, and the bewildering complexity therein that he has perhaps wondered about only from a safe distance. For professionals from a variety of backgrounds including medicine, chemistry, engineering, and business as Bioinformatics is a multi-disciplinary field of study.

Biotechnology For You

# Software Tutorial

## Bioinformatics Applications:

### *Why Java is used for developing Bioinformatics tools?*

- Bioinformatics challenges developers to create widely distributable tools that elucidate biological relationships.
- The principal value of these tools is measured in the contribution they provide to research biologists. Furthermore, as the amount and nature of available data increases, bioinformaticians are forced to provide rapidly evolving tools to their user communities.
- Java has allowed bioinformaticians to rapidly develop user-friendly, cross-platform applications that are accessible to users at all levels of computational ability. Physiome Sciences' computer-based biological simulation technologies and Bioinformatics Solutions' PatternHunter are two examples of the growing adoption of Java in bioinformatics.

### *Which Bioinformatics Applications Use Java?*

As mentioned, the main bioinformatics tasks are data retrieval and analysis. We have discussed several Java-based data retrieval methods for accessing genomic sequence and annotation. But what types of tools are being constructed to perform novel bioinformatics analysis? How are they delivering their analyses to users of varying computational abilities?

- TIGR MultiExperiment Viewer (MEV) integrates large numbers of experiments to facilitate researchers in analyzing patterns of gene expression. MEV is an important example of a bioinformatics application that aims at integrating bioinformatics experimentation. Furthermore, MEV is extensively documented and tested, and is freely available as open source.
- 
- J-Express is a commercial Java application similar to MEV. It has also leveraged its success as a bioinformatics application by combining suites of bioinformatics algorithms relevant to microarray/genechip analysis (a technique used to measure what genes are being transcribed/expressed).
- 
- WebMol analyzes molecular structure information. This program uses Java3D to allow researchers to visualize and manipulate complex protein structures. It is a good example of how Java applets can be used to deliver novel information. Figure 1 shows it in action.

The development of bioinformatics applications can be broken down into two main tasks: data management and biological analysis. Let's find out more about Data management in Bioinformatics.

#### Facts about Data Management in Bioinformatics

- The Data management generally involves aggregating sequences (strings representing DNA or proteins, the former being represented by non-random but complex patterns of As, Cs, Ts and Gs) and/or annotation information (the known properties of a given sequence; for instance, annotation consists of the locations of genes and other biologically relevant features).
- Genome browsers, like EnSEMBL and UCSC, actively aggregate and serve annotation for several genomes (the chemical sequences that make up an organism's DNA).
- Bioinformaticians wishing to exploit new data constantly need to know what data exists, how they can access this data, and when to provide it to their users.
- Genome browsers provide bioinformaticians with a central location for exploring the common types of annotation that are readily available.