Bio-Informatics :Introduction - I

Dr. Vivek Upasani
Microbiology Department,
M. G. Science Institute,
Navrangpura, Ahmedabad 380009.

History

Year	Protein/Nucleic Acid	Scientist(s) / Remarks
1865	Laws of hereditary Birth of Genetics	G. Mendel
1950s	Edman degradation-dansylation method	Edman
1953	Double helix Structure of DNA	Watson J.D. & Crick F.H.C.
1955	First complete peptide sequence - Insulin	Ryle *et al.*
1960s	tRNA sequenced (74-95 nucleotides)	Holley R. *et al.*
1960	First complete enzyme sequence - Ribonuclease	Hirs *et al.*

History continued…….

Year	Protein/Nucleic Acid	Scientist(s) / Remarks
1965	Atlas of Protein Sequences and Structures	Margaret Dayhoff
1967	Automated protein Sequencers	Edman & Begg
1977	Nucleic Acid Sequencing Methods	Maxam & Gilbert Sanger & Coulson
1977	ØX174 genome sequenced	Sanger *et al.*
1990	Human Genome Project (HGP) launched	DoE, NIH & public funds
1995	First Bacterial Genome to be sequenced *Haemophilus influenzae*	
Feb. 2001	First Draft of Human Genome	HGP & Celera Genomics Nature (2001) 409:860-921
2003	Complete Sequence Human Genome	Estimated to be completed in 2005

What is ‘bioinformatics’?

- Who invented bioinformatics?
- Term Bioinformatics was invented by **Paulien Hogeweg** and Ben Hesper in 1970 as "the study of informatic processes in biotic systems".

- The term was first coined in 1988 by

  Dr. Hwa Lim
- The original definition was :

  “a collective term for data compilation, organisation, analysis and dissemination”

That means….

- Using information technology (IT) to help solve biological problems by designing novel and incisive algorithms and methods of analyses

- It also serves to establish innovative software and create new/maintain existing databases of information, allowing open access to the records held within them.

Other definitions of Bioinformatics-

- Bringing biological themes to computers
- BISTIC Bioinformatics Definition –

  Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data
- BISTIC Computational Biology Definition –

  Computational Biology: the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.
- http://grants2.nih.gov/grants/bistic/bistic.cfm

It's a huge subject

- 'Bioinformatics' - the new 'buzz word' in the scientific community
- It is an umbrella term for a world of "omics" namely genomics, proteomics, cellomics and evolution, and computer science
- It is now necessary for scientists to be inter-disciplinary

Why?

- The data is collected from a variety of sources
- The terminology is specific to its particular branch of science
- To allow the effortless transfer of information gathered and the interrogation of databases across the global interface

  i.e. to make the data easily and universally interpretable by scientists.

- It is a discipline vital in the era of post-genomics.

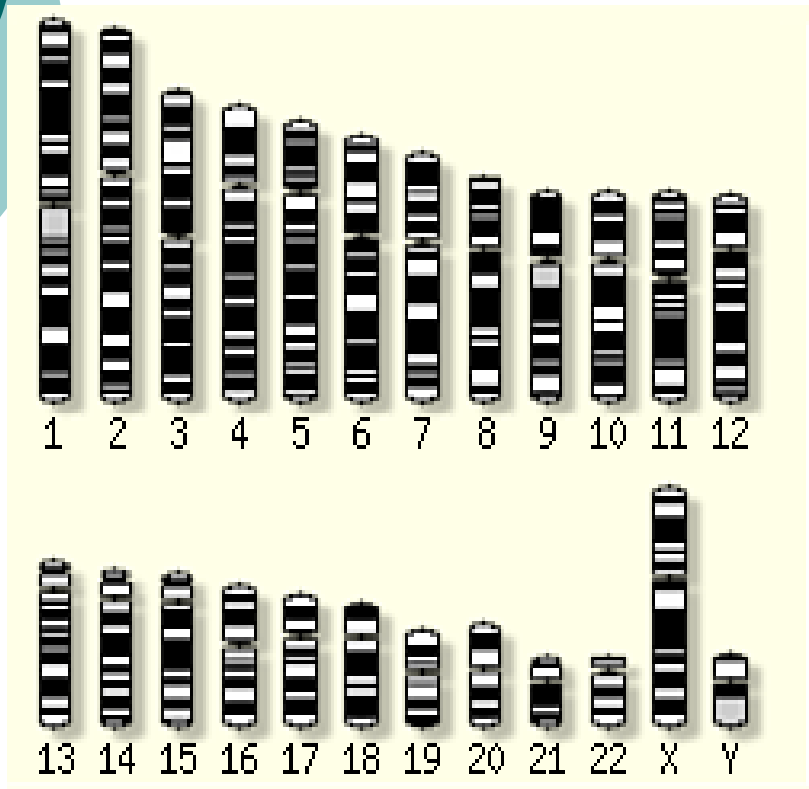
The Human Genome Project

- In 1990 the unveiling of the Human Genome Project (HGP) by the United States Department of Energy (DoE) and the National Institutes of Health
- Goals: to identify all chemical base pairs and all genes that make up the 23 chromosome pairs found in human DNA
- “To develop the next generation of methods for simulating cellular behaviour and pathways”
- Ultimately to devise means to apply IT to the modelling of cellular functions as specified by the enormous datasets

HGP continued…..

- Collaboration of 20 groups across the world
- The results would be free and data release would be rapid
- The completion date for which has been brought forward from the expected 15 years i.e. 2005 to 2003 due to technological advancement
- However, since the birth of the HGP biologists have already been inundated with at least 12 years worth of data
- The information is still accumulating at a great pace – the availability of the finished sequences has increased to an annual rate of 1Gb per year and this amount is on the rise

What does that mean?



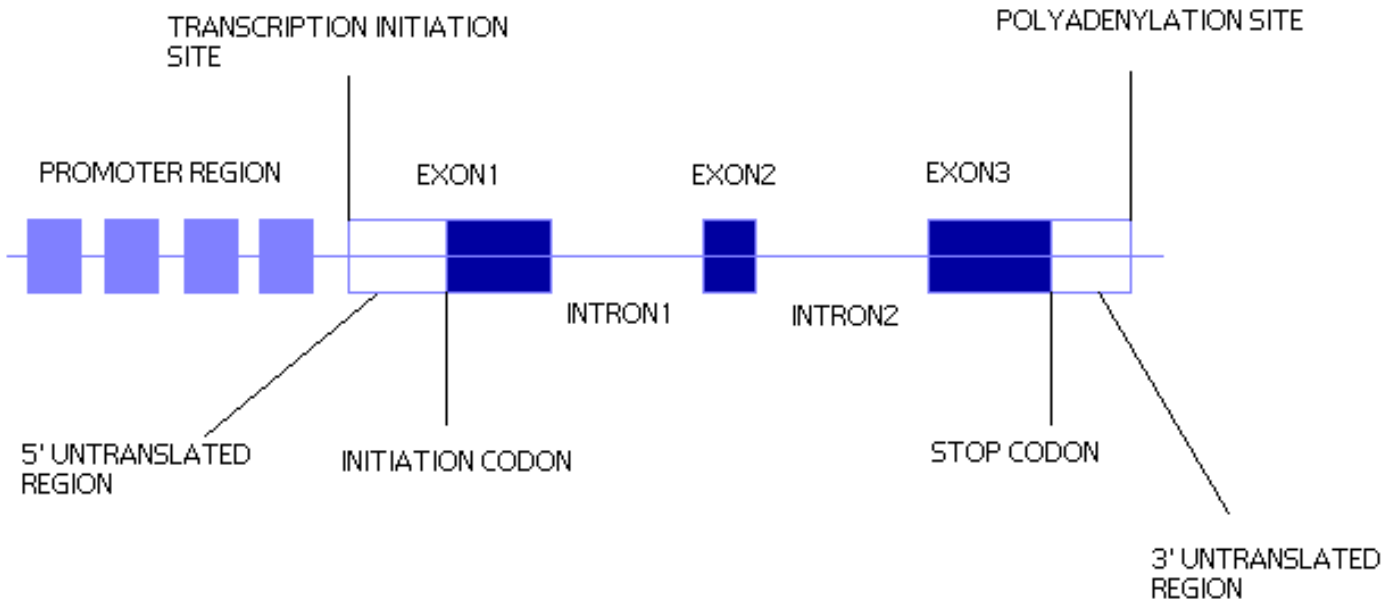
- To identify every base pair in the genome - there are 3×10^9
- To assign genes and what they code for or not
- RNA/proteins, regulatory genes
- Potentially revolutionise biomedical research

Examples of what they found?

- Human DNA is 25 times the size of any genome sequenced so far.
- The mutation rate is about twice as high in male as in female meiosis, which means that most mutation occurs in males.
- Information about evolution dating back 500million years - for example that some of the human DNA is there following the horizontal transfer from bacterial at some point along the vertebrate lineage.
- The initial estimates were that the human genome comprised some 100,000+ genes - now we know there are only 30-40,000 i.e. only twice as many as found in a worm or a fly.

Open Reading Frames

A schematic diagram of a typical eukaryotic protein coding gene



The complex splicing techniques of higher organisms means each protein-coding gene generates between 3 and 6 proteins = 50,000 to 500,000 proteins per individual

What this means…

- The idea that 1 gene = 1 protein is clearly wrong
- The gene structure and the components that regulate its expression must be much more complex than previously thought
- Poses the question whether identical proteins serve different functions depending on where in the organism they are found
- Still have roughly 100,000 genes of microbes, plants and animals whose functions are still to be revealed

The ‘omic’ revolution

- Bioinformatics has been split into various subjects:
  - Genomics – the sequencing and annotation of genomes
  - Functional and structural genomics – the comparison and characterisation of genomes of different species
  - Proteomics – the description of the complete set of proteins a particular genome codes for
- The data thus far obtained is substantial

Comparative Genomics

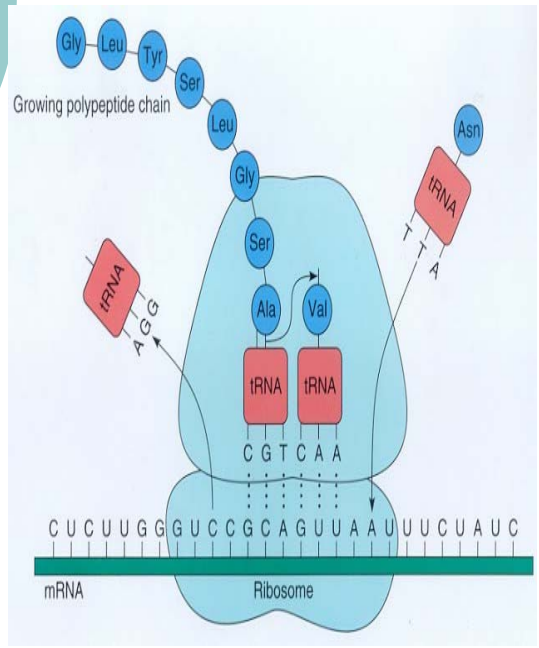
- Complete genomic sequences are now available for many organisms/bacteria/viruses/organelles including the following 'model' organisms:

- *Escherichia coli* (the bacteria)
- *Saccharomyces cerevisiae* (the yeast)
- *Caenorhabditis elegans* (the worm)
- *Drosophila melongaster* (the fruit fly)
- *Danio verio* (the zebrafish)
- *Aradopsis thaliana* (the plant)
- *Mus musculus* (the mouse)

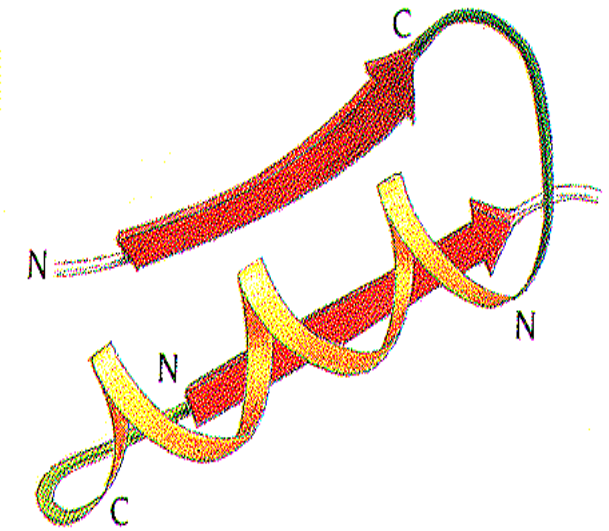
Why have model organisms?

- Cheap
- Plentiful
- Short generation times
- Easily manipulated
- Test novel drug candidates
- Illustrating which genes, and therefore which proteins, are responsible for which phenotype/disease
- 85% genetic similarity between the mouse and human genome

Next challenge is to figure out the proteome



- Proteins comprise of various combinations of 20 different amino acids
- The protein sequence assembled on the ribosome during protein synthesis is known as the primary structure
- This string of amino acids then folds into various elements known as the secondary structure
- These may be α-helices, β-sheets or 'random' coils



- Protein tertiary structure is the specific packing of these elements to give various ‘motifs’
- It is the 3D pattern of the protein – directly linked to its function
- This is an example of a β-α-β motif

Why are proteins important?

- Proteins may be simple (just amino acids)
- Or “conjugated” – have additional groups like metal ions for example iron ions in haemoglobin
- Proteins may be
  - Enzymes
  - Structural eg.keratin (hair) and collagen (tendons and cartilage)
  - Transport proteins like Haemoglobin, myoglobin
  - Antibodies
  - Hormones
  - Regulatory eg. in metabolism
  - First point of call for drug targeting

Where does computer science come into it?

The HGP has brought to light the limitations of traditional lab work – although mostly automated they are expensive and time consuming

- **We need to incorporate original techniques to allow greater understanding of protein function, protein-protein interactions and protein-DNA interactions and put it all in a cellular context**

- **Bioinformaticists act to bridge the gap between the data stored and its biological significance**

How is this accomplished in proteomics?

- **Sequence alignment – looking for homology**
  - **Homology is defined as the divergent evolution of two proteins from a common ancestor**
- **Solving protein structure – current methods are using NMR and X-ray crystallography**
- **Aims are to predict protein folding, protein structure and protein structure-function relationships**

Challenges in bioinformatics…

- To attempt to further our understanding of biochemical pathways via *in silico* protein docking experiments
- Automation of the design of drug compounds
- The better understanding (through modelling) of protein structure and function
- Site directed mutagenesis experiments – virtual manipulation of structure to make better/synthetic proteins
- Better understanding of evolutionary processes and pressures for example why is it that many proteins with completely different amino acid sequences fold to give similar structures?

Progress….

- Bioinformatics therefore plays a significant role as a tool to aid biological research rather than as an end in itself.

- Scientists implement bioinformatics to:
  - analyse
  - interpret
  - and apply the vast amounts of high resolution data, both archived and new, to further their understanding of how biological systems work

- Ultimately to assist structure based drug design and curing diseases.

However….

- The magnitude of data is ever increasing.
- The invention of computers, however, and especially their improvement in the last couple of decades, has allowed scientists to more easily collate information which can be added to, edited, organized, and maintained in a more efficient and less time consuming manner than back in the days when Europeans first started to travel and explore the world.

In a nutshell….

- Bioinformatics will also serve to advance medical research regarding the drug discovery process and therapeutic intervention.

- Implementing the information disclosed permits us to discern biological systems well enough and at such a level to build models reflecting how natural pathways/processes work, to predict their response and behavior, to manipulate them, as well as to identify defects in order to better understand and fight disorders and disease.

- **THANK YOU!!!!**