Bioinformatics in Post Genomic Era
• What is Bioinformatics?
• Availability of information about the human genome and other genomes
• Human health related databases
• Bioinformatics and Drug development
• Ethical, Legal and Social Issues (ELSI)
What is Bioinformatics?

- One idea for a definition:
  - (Molecular) **Bio-informatics** =
  - is conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied math, CS, and statistics) to understand and organize the information associated with these molecules, on a large-scale.
Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. There are three important sub-disciplines within bioinformatics:
• the development of new algorithms and statistics with which to assess relationships among members of large data sets;
• the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures;
• the development and implementation of tools that enable efficient access and management of different types of information.
Biological Data + Computer Calculations → Bioinformatics
# The Bioinformatics Spectrum

<table>
<thead>
<tr>
<th>Breadth: Homologs, Large-scale Surveys, Informatics</th>
<th>Pairwise comparison, sequence &amp; structure alignment</th>
<th>Multiple alignment, patterns, templates, trees</th>
<th>Databases, scoring schemes, censuses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3-100</td>
<td>100+</td>
</tr>
</tbody>
</table>

## Depth: Rational Drug Design (physics)

- **Structure Docking**
- **Force Field**
- **Protein Surface**
- **Geometry Calculation**
- **Structure Prediction**
- **Protein Sequence**
- **Gene Finding**
- **Genome Sequence**

- **Ligand Complex**
- **Molecular Simulation**
- **Protein Structure**
- **Depth: Rational Drug Design (physics)**
- **Structure Prediction**
- **Protein Sequence**
- **Gene Finding**
- **Genome Sequence**
What is the Human Genome?

• The entire genetic makeup of the human cell nucleus.

• Genes carry the information for making all of the proteins required by the body for growth and maintenance.

• The genome also encodes rRNA and tRNA which are involved in protein synthesis.
• Made up of ~35,000-50,000 genes which code for functional proteins in the body

• Includes non-coding sequences located between genes, which makes up the vast majority of the DNA in the genome (~95%)

• The particular order of nucleotide bases (As, Gs, Cs, and Ts) determines the amino acid composition of proteins
• Information about DNA variations (polymorphisms) among individuals can lend insight into new technologies for diagnosing, treating, and preventing diseases that affect humankind.
What Goals Were Established for the Human Genome Project When it Began in 1990?

- Identify all of the genes in human DNA.
- Determine the sequence of the 3 billion chemical nucleotide bases that make up human DNA.
- Store this information in data bases.
- Develop faster, more efficient sequencing technologies.
- Develop tools for data analysis.
- Address the ethical, legal, and social issues (ELSI) that arise from the project.
Two Different Groups Worked to Obtain the DNA Sequence of the Human Genome

- The HGP is a multinational consortium established by government research agencies and funded publicly.
- Celera Genomics is a private company whose former CEO, J. Craig Venter, ran an independent sequencing project.
- Differences arose regarding who should receive the credit for this scientific milestone.
- June 6, 2000, the HGP and Celera Genomics held a joint press conference to announce that TOGETHER they had completed ~97% of the human genome.
Banking on Genome data

- Britain is about embark on the world’s largest genome data project focussed on middle aged people which may shed light on the interaction between genes, health and the environment.

- Studies of families affected by genetic disease have proven useful for genetic linkage analyses (e.g. Huntington’s disease, neurofibromatosis, cystic fibrosis, Duchenne’s muscular dystrophy).
<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size (basepairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epstein-Barr virus</td>
<td>0.172 *10^6</td>
</tr>
<tr>
<td>Bacterium (E.coli)</td>
<td>4.6 *10^6</td>
</tr>
<tr>
<td>Yeast (S.cerevisiae)</td>
<td>12.1 * 10^6</td>
</tr>
<tr>
<td>Nematode worm (C.elegans)</td>
<td>95.5 * 10^6</td>
</tr>
<tr>
<td>Thale cress (A.thaliana)</td>
<td>117 * 10^6</td>
</tr>
<tr>
<td>Fruit fly (D.melanogaster)</td>
<td>180 * 10^6</td>
</tr>
<tr>
<td>Human (H.sapiens)</td>
<td>3200 * 10^6</td>
</tr>
</tbody>
</table>
Gene Sequence         Protein Sequences

- Supposed to be raw data.
- One has to add layers of information to the sequence data
- Annotation of the data becomes very important
- Annotation : Theoretical methods
  Experimental methods
- Bioinformatics / Statistics / Mathematics
Complete Genome Sequences From Several Organisms Are Known

- Comparative Genomics
- Structural Genomics
- Functional Genomics
- Cellular Genomics
- Network Genomics
- Ethical Genomics
- Moral Genomics
Other Completed Genomes

- *Haemophilus influenzae*
- *Escherichia coli*
- *Bacillus subtilis*
- *Helicobacter pylori*
- *Borrelia burgdorferi*
- *Streptococcus pneumoniae*
- *Saccharomyces cerevisiae*
• Caenorhabditis elegans
• Arabidopsis thaliana
• Archaeoglobus fulgidus
• Methanobacterium thermoautotrophicum
• Methanococcus jannaschii
• Mycoplasma pneumoniae
• Mycoplasm genitaliu
• Rickettsia prowazekii
• Mycobacterium tuberculosis
• *Treponema pallidum*
• *Staphylococcus aureus*
• *And more!*
Completed Plant Genomes

- Arabidopsis thaliana

Completed Insect Genomes

- Drosophila melanogaster

Completed Rodent Genomes

- Mus musculus
Which Branches of Biology will Benefit from this Knowledge?

• Medicine
• Pharmacogenomics
• Biotechnology
• Bioinformatics
• Proteomics
Diagnosis of disease and disease risk

(a) when a patient presents with symptoms

(b) in advance of appearance of symptoms

[eg] Huntington disease (an inherited neurodegenerative disorder)

- symptoms: uncontrollable dance-like (choreatic) movements, mental disturbance, personality changes and intellectual impairment

- repeats of the trinucleotide CAG, corresponding to polyglutamine blocks in the corresponding protein, huntingtin
• 11-28 CAG repeats --> normal
• 29-34 CAG repeats --> likely to develop disease
• 35-41 CAG repeats develop mild symptoms
• more than 41 CAG repeats suffer full huntington disease

(c) for in utero diagnosis of potential abnormalities such as cystic fibrosis, asthma etc.

(d) for genetic counselling of couples contemplating having children
Online databases of disease-associated mutations
Online database of Mendelian Inheritance in Man (OMIM)

Welcome to OMIM(TM), Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere, and developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The database contains textual information, pictures, and reference information. It also contains copious links to NCBI's Entrez database of MEDLINE articles and sequence information.

The OMIM Morbid Map, a catalog of genetic diseases and their cytogenetic map locations arranged alphabetically by disease, is now available.

Browsing OMIM

- Search the OMIM Database
- Search the OMIM Gene Map
- Search the OMIM Morbid Map
- The OMIM numbering system
- View the OMIM Update Log
Welcome to the Human Gene Mutation Database at the Institute of Medical Genetics in Cardiff.

Copyright © University of Wales College of Medicine 2003. All Rights Reserved.

HGMD Search

- HGMD publications
- Newly added genes
- Locus-specific databases
- Electronic mutation submission
- Comment form
- Supplementary information
- Other useful links
- Meetings & news
- Mutation nomenclature
- C. publications

Statistics

What's new

HGMD Background

This database is maintained by D.N. Cooper, E.V. Ball, P.D. Stenson, A.D. Phillips, M.E. Mort and J.A. Shiel with the assistance of N.S.T. Thomas & S. Acyusgene. We gratefully acknowledge the support of:

Pfizer

MACMILLAN

SUN LIFE

HGMD enters into collaboration with Celera Genomics. Please see our What's New page for the full text of the announcement.

IARC p53 database

WORKSHOP: Functional consequences of TP53 mutations
Characterization of common and rare p53 mutants and relevance to
human cancer
Monday 30 June - Thursday 3 July 2003
Information and report here
Haemophilia B Mutation Database

Version 12

2003

A database of point mutations and short additions and deletions in the factor IX gene

- Introduction and background to the haemophilia B database
- Complete Reference List to the mutations listed in the database
- Polymorphisms and rare variants
- Gross deletions and insertions

View Mutation Table:
Von Willebrand factor database

The University of Sheffield

ISTH SSC VWF Database

Informations Pages

Up-to-date databases of point mutations, insertions, deletions, and polymorphisms found in the gene for human von Willebrand Factor.

Fast link:

Please select one of the following links:

- **Mutations** Mutations in the human VWF gene sorted by nucleotide number and by association with type of von Willebrand disease (VWD)
- **Polymorphisms** Polymorphisms in the human VWF gene
- **Nomenclature** Current nomenclature recommended by the ISTH VWF SSC
- **VWF Sequences** Amino acid, cDNA and genomic DNA links with Numerous genomic and pseudogene sequences

Supported by Aventis
Amyotrophic lateral sclerosis database

Welcome to alsod.org

alsod.org is the online database for ALS genetic (SOD1, ALS and other) mutations. alsod.org is designed to provide both the scientific community and wider public with up-to-date information on SOD1 associated ALS.

We have performed a major upgrade of the database, and alsod.org can now store data on multiple genes including SOD1 and ALS2. Read more about the changes.

Reports
Summaries of data by gene or mutation

Search
Find specific mutations or subjects

Add data
If you wish to add some of your own data, you must first register.

Public access
Although not all the data submitted to alsod.org is publicly available, alot of it is. Feel free to browse the ‘reports’ or ‘search’ sections.

What is ALS/MND?
Amyotrophic lateral sclerosis (ALS) or Motor Neurone Disease (MND) is a progressive, fatal neurological disease (get more information on ALS/MND)
Bioinformatics and Drug development
<table>
<thead>
<tr>
<th>Compound</th>
<th>Target enzyme</th>
<th>Clinical use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetazolamide</td>
<td>Carbonic anhydrase</td>
<td>Glaucoma</td>
</tr>
<tr>
<td>Aspirin</td>
<td>Cylooxygenases</td>
<td>Inflammation</td>
</tr>
<tr>
<td>Amoxicillin</td>
<td>Pencillin binding proteins</td>
<td>Bacterial infections</td>
</tr>
<tr>
<td>Digoxin</td>
<td>Sodium, potassium ATPase</td>
<td>Heart disease</td>
</tr>
<tr>
<td>Omeprazole</td>
<td>H^+ ,K^+-ATPase</td>
<td>Peptic ulcers</td>
</tr>
<tr>
<td>Sorbinol</td>
<td>Aldose reductase</td>
<td>Cancer</td>
</tr>
<tr>
<td>VIAGRA</td>
<td>Phosphodiesterase</td>
<td>Erectile Dysfunction</td>
</tr>
</tbody>
</table>
RECEPTORS

- G-protein coupled receptors
- Ligand-gated ion channels
- Tyrosine kinase receptors
- Nuclear receptors
Drug Discovery Pipeline

DNA sequences and maps → Genomic database → Gene expression analysis → Target Genes → Protein data → Disease selection → Empirical medicine

Genomic database → Receptor identification

DNA sequences and maps → Disease model → Disease biology

Disease selection → In vivo cell Biology database

Protein structure prediction analysis → Lead identification and optimization → Predictive ADME

Predictive ADME → Preclinical trials → Clinical trials → Preclinical and experimental data

Preclinical trials → Clinical trials → ACME Toxicity

Molecular diversity chemical descriptions → Physical chemistry

HTS → Rational drug design → Medicinal chemistry
Workflow of a virtual screening run against a specific target

Starting Database

Prepared Database

Starting Target

Prepared Target

Molecular Docking

Post Process

Select Components for assay
Genetics of responses to therapy-customized treatment

- sequence analysis permits selecting drugs and dosages optimal for individual patients, a fast-growing field called pharmacogenomics [eg] 6-mercaptopurine used in the treatment of childhood leukaemia
Identification of drug targets

(a) drug design process
(b) drugs act on targets such as receptors, enzymes, hormones and some unknown targets
(c) differential genomics [eg] tumour cells

Gene therapy

(a) direct supply of proteins [eg] insulin
(b) antisense therapy [eg] crohn disease
Eliminating side effects

Developing revolutionary new drugs and treatments for illness that previously couldn't be treated/preventing or avoiding serious diseases.

It is believed that we are approaching a new era of ‘personalized medicines’ medicine that understands as individual patient at the genetic level and offers the optimum treatment.
Rationales for Drug Design

- Tuberculosis is a global threat affecting 1/3 of world population with latent infections. 50% of HIV patients develop TB.
- TB cases are on the rise and approximately 2 million people each year die from the infection.
- The spread of HIV/AIDS and the emergence of multidrug-resistant TB are contributing to the worsening impact of this disease.
- It is estimated that between now and 2020, approximately 1000 million people will be newly infected, over 150 million people will get sick, and 36 million will die of TB - if control is not further strengthened.
Realistic Design Cycle

**Lead Generation:**
- natural ligands
- random screening
- literature leads

**Chemical Library Preparation:**
- synthesis of compounds

**Biological Testing:**
- Biological testing of compounds

**Structure Analysis and Compound Design:**
- Rationalise SAR, predict better analogs, de-novo leads

**3D structure of ligand:**
- QSAR, receptor mapping, 3D QSAR

**3D structure of target-ligand complex:**
- Homology modeling

**3D structure of target-ligand complex:**
- X-ray & NMR structures

**Physical Properties and 3D ligand databases:**

**Pre-clinical studies:**

**Molecular Biology:**
- Cloning/Expression
- ²H/N/²C labeling, purification, characterisation

**3D protein structure sequence databases:**
Blockbuster Drugs

HIV drugs

In 1998 in the US, NRTIs accounted for $865 million in sales, PIs $865 million and NNRTIs for $100 million.

The market in the rest of the world is about $2 billion (1998).
Cartoon representation of TA xylanase along with the active site Glu 131 and Glu 237, the salt bridge (Arg 124 - Glu 232) and disulphide bridge.
The “salad bowl” view showing the substrate binding cleft. The Active site is at the C-terminus of the β barrel and the salt bridge is at the N-terminus of the β barrel.
Figure shows an example for the competition for polar atoms by water molecules is more at low temperature.
A Water dimer formed by Wat 533 (W1) and Wat 511 (W2) and its interactions. Conserved residues are labeled in red. Interactions involving water molecules appear to contribute to the stability of residues in the active site region. β-strands 1 and 8 are not shown.
The HIV-1 protease-pepstatin complex

aspartyl residues of the active site

pepstatin
(HIV protease dimer complexed with protease inhibitor(red), GIF generated using RasMol)
HIV protease & inhibitor (red)
Biotechnology

- Production of useful protein products for use in medicine, agriculture, bioremediation and pharmaceutical industries.
  - Antibiotics
  - Protein replacement (factor VIII, TPA, streptokinase, insulin, interferon…)
  - BT insecticide toxin (from *Bacillus thuringiensis*)
  - Herbicide resistance (glyphosate resistance)
• Bioengineered foods [e.g. Flavr Savr tomato (antisense – polygalacturonase) to delay rotting]
• “Pharm” animals
Proteomics

- Investigates patterns and levels of gene expression in diseased cells that can be analyzed to build databases of expression profiles.
Developmental Biology

- Regulation of embryonic development.
- Regulation of the aging process.
Because DNA mutates at a constant rate, comparisons of DNA between different organisms can provide evolutionary histories.
Ethical, Legal and Social Issues (ELSI)

- Privacy legislation
- Gene testing
- Patenting
- Forensics
- Behavioral Genetics
- Genetics in the Courtroom
Philosophical Implications

Human responsibility
Free will versus genetic determinism
Psychological Impact and igmaization

- Affects on the individual
- Affects on society’s perceptions and expectations of the individual
Clinical Issues

- Growing demand to educate health care workers to accurately evaluate genetic tests.
- Public needs to gain scientific literacy and understand the capabilities, limitations and risks.
- Standards need to be established including quality controls to ensure accuracy and reliability.
- Federal regulation?
Genetic Counseling

- Informed consent for complex procedures
- Counseling about the risks, limitations and reliability of genetic screening techniques
- Reproductive decision making based on genetic information
- Reproductive rights
Multifactorial Diseases and Environmental Factors

- Genetic predispositions do not mandate disease development
- Caution must be exercised when correlating genetic tests with predictions
Summary

• The significance of the completion of the human genome project cannot be overstated.
• With the dictionary of the genome available, the molecular mechanisms of human health and disease will be resolved.
• Armed with this knowledge a transformation in medical diagnostics and therapy is underway and will continue into the next few decades.
• The application of this knowledge needs to be regulated and restricted to practices deemed ethically sound.
In nature’s infinite book of secrecy
A little I can read
Reference:
Prof. S. Ramakumar,
Bioinformatics Center,
IISc,
Bangalore-12.
THANK YOU FOR YOUR KIND ATTENTION