Bioinformatics Quick Start

Medha Bhagwat
National Center for Biotechnology Information
National Institutes of Health

Outline

1. What is a genome?
2. What is genomics?
3. What is bioinformatics?
4. Applications of genomics/bioinformatics
5. Future implications
6. A practical example
Proteins are Body's Worker Molecules

Hemoglobin carries oxygen to every part of the body

Ion channel proteins control brain signaling by allowing small molecules into and out of nerve cells

Enzymes in saliva, the stomach and the small intestine are proteins that help you digest food

Muscle proteins called actin and myosin enable all muscular movement

Antibodies are proteins that help defend your body against foreign invaders such as bacteria and viruses
DNA

Basic Unit (alphabet): Nucleotide (base)
Only 4: A, T, G, and C
Double-stranded

..AGCTGCATGCTAGCTGACGTCA….
    |||||||||||||||||||||||||
..TCGACGTACGATCGACTGCAGT….

“Words” (genes) to encode proteins, RNA etc.
Double helical

The double-helical structure of DNA
Protein

Alphabet: amino acids

There are 20 amino acids
Encoded by codons (triplets of nucleotides)

ATGTGCAGCCTAGCTGCCGTC

Met—Cys—Ser—Leu—Ala—Ala—Val

Water channel protein
Genome (DNA)

Exact spelling of a word is necessary

<table>
<thead>
<tr>
<th>CAT</th>
<th>DAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAT</td>
<td>GAT</td>
</tr>
<tr>
<td>MAT</td>
<td>KAT</td>
</tr>
<tr>
<td>FAT</td>
<td>CAT</td>
</tr>
<tr>
<td>BAT</td>
<td>CCT</td>
</tr>
<tr>
<td>EAT</td>
<td>CAD</td>
</tr>
<tr>
<td>HAT</td>
<td>CAC</td>
</tr>
</tbody>
</table>

Some changes in amino acids lead to diseases and some indicate normal differences among humans.

Genomics

Proteomics

From Genes to Proteins

http://genomics.energy.gov/
Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   - How to access the genome data?
   - How to access the analysis tools?
4. Applications of genomics/bioinformatics
   - Analysis of human and other genomes
5. Future implications
6. Interpretation/global analysis of data
   - Photoreceptors

Additional Information


Talking Glossary of Genetic Terms
http://www.genome.gov/10002096
Bioinformatics

Variety of definitions
By Luscombe et al Method Inform Med 2001; 40:346-58

Bioinformatics is conceptualizing biology in terms of molecules (in the sense of Physical chemistry) and applying "informatics techniques" (derived from disciplines such as applied math, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale.

Bioinformatics is a management information system for molecular biology and has many practical applications.

---

Bioinformatics

I. Organize data in databases
   researchers can access current data
   submit new data
II. Develop tools and resources to analyze data
III. Interpret data in a biologically useful manner
global analysis of data to uncover common principles that apply across many systems
National Center for Biotechnology Information

Created as a part of NLM in 1988
- To establish public databases
  U.S. National DNA Sequence Database
- To perform research in computational biology
- To develop software tools for sequence analysis
- To disseminate biomedical information

NCBI Databases

Primary Databases
- Genomes (DNA)
- mRNA
- Protein
- Organisms

Derived Databases
- Expression
- Structures
- Small compounds
- Conserved Domains
- Publications
- Books
- Gene
- Homologene
- Unigene
- RefSeq
## NCBI Databases

<table>
<thead>
<tr>
<th>Primary</th>
<th>Derived</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archival/repository</td>
<td>Curated</td>
</tr>
<tr>
<td>Redundant</td>
<td>Non-redundant</td>
</tr>
<tr>
<td>Submitter owner</td>
<td>NCBI owner</td>
</tr>
<tr>
<td>Sequenced</td>
<td>Combined/edited</td>
</tr>
</tbody>
</table>

Ex: GenBank  
Ex: RefSeq

---

### NCBI Databases

NCBI Databases
Genome Sequence Data and Analysis Tools at NCBI

### Sizes of Different Genomes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aloe vera</td>
<td>16.0 billion</td>
</tr>
<tr>
<td>Rabbit</td>
<td>3.5 billion</td>
</tr>
<tr>
<td>Human</td>
<td>3.2 billion</td>
</tr>
<tr>
<td>Laboratory mouse</td>
<td>2.6 billion</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>137 million</td>
</tr>
<tr>
<td>Yeast</td>
<td>12.1 million</td>
</tr>
<tr>
<td>Bacterium (E. coli)</td>
<td>4.6 million</td>
</tr>
<tr>
<td>Human immunodeficiency virus</td>
<td>9700</td>
</tr>
</tbody>
</table>

### Tools for Data Mining

**BLAST**
- BLAST (Basic Local Alignment Search Tool) for comparing nucleotide and protein sequences against others in public databases, how worked in several ways including PSI-BLAST, PH-BLAST, and BLASTZ sequences. Specialized BLASTs are also available for human, microbial, malaria, and other pathogens, as well as for vector containment, immunoglobulins, and inherit human oncogenes.

**SAM/EMAP**
- SAM/EMAP - provides a tool for performing statistical tests designed specifically for differential gene expression of SAGE libraries. The data include SAGE libraries generated for individual labs as well as those generated by the Cancer Genome Anatomy Project (CGAP), which have been submitted to Gene Expression Omnibus (GEO). Gene expression profiles that compare the expression in different SAGE libraries are also available on the Entrez GEO Profiles page. It is possible to order a pair of sequence on the SAM/EMAP resource to determine what SAGE tags are in the sequence, then may be associated SAGEELD records and view the expression of those tags in different CGAP SAGE libraries.

**Entrez Gene**
- Entrez Gene record encapsulates a wide range for a given gene and organism. When possible, the information includes results of analyses that have been done on the sequence data. The amount and type of information presented depend on what is available for a particular gene and organism can include: (1) graphic summary of the genomic context, intron-exon structure, and flanking genes, (2) link to a graphic view of the mRNA sequence, which in turn shows biological features such as CDS, SNPs, etc., (3) links to gene ontology and phenotypic information, (4) links to corresponding protein sequence data and conserved domains, (5) links to mining resources, such as multigenic databases. Entrez Gene is a successor to LocusLink.

**Model Maker**
- Model Maker allows you to view the existing mRNA, ESTs, and gene interactions that was aligned to assembled genomic sequence to build a profile model and edit the model by creating or removing putative exons. You can then view the mRNA sequence and potential ORF for the edited model and save the mRNA sequence data for use in other programs.

**ORF Finder**
- ORF Finder identifies all possible ORF s in a DNA sequence for using the standard and alternative codon and start codons. The deduced amino acid sequences can then be used to BLAST against GenBank. ORF Finder is also packaged in the sequence submission software Seqin.

**Organism Specific Resources**
- Mouse, Rat, Chicken, Cow, etc.


- **SAEMAP** - a tool for computing ORF-to-Genomic alignments based on a variation of the Needleman-Wunsch algorithm combined with Blast for compartment detection and greater performance.
- **WebScreen** - a tool for identifying segments of a sequence and protein sequence for use in other programs.
- **WebQueening Tool** - a web-based program that identifies the putative intron exons for both intron-exon and non-intron-exon nucleotide sequences. It works by using BLAST to compare a query sequence to a set of reference sequences for known introns. Predictive reference genomes exist for three major virus pathogens: human immunodeficiency virus-1 (HIV-1), hepatitis C virus (HCV) and hepatitis B virus (HBV), as well as for bacteria. User-defined reference sequences can be used at the same time. The query sequence is broken into segments for comparison to reference sequence so that the mosaic organization of the sequence is revealed. The results are displayed graphically using color-coded genomics. Therefore, the genotype(s) of any portion of the query can be determined.
Genome Sequence Data and Analysis Tools at NCBI

Tools - BLAST

- BLAST: The Basic Local Alignment Search Tool (BLAST) is used for comparing gene and protein sequences against other genes or proteins in public databases, now offers several types including PSI-BLAST, Phx-BLAST, and BLAST 2.0. BLAST results are available for human, mouse, rat, and other genomes, as well as for vector contamination, immunoglobulins, and terminally-dependent consensus sequences.

- CD Search: Search the Conserved Domain Database with Reverse Position Specific BLAST.

- Distant BLAST: BLAST displays the results of all BLAST searches that have been made for every protein sequence in the Entrez Protein data.


Tools - Gene Expression

- GEO: Gene Expression Omnibus (GEO) provides several tools to assist with the visualization and analysis of GEO data. GEO allows the browsing of microarray cluster heat maps, providing insight into the relationships between samples and regulated genes. Individual gene expression profiles showing significant differences between experimental subsets may be located using average subset rank value profiles. Plotted gene expression profiles may be identified by the basis of sequence similarity, profile similarity, or homology. Indicators of dataset normalization quality are provided as distribution graphs, and binned scatter plots, useful for assessing sequence, mapping, and publication database resources are provided where possible.

- RNASeq: RNA-Seq is a technology for sequencing RNA to identify expressed genes. The National Cancer Institute (NCI) is conducting an RNA-seq pilot project. RNA-Seq provides a measure of the relative abundance of RNA transcripts in a sample. RNA-Seq data is generated using high-throughput technologies, such as deep sequencing and next-generation sequencing, which can produce millions of reads from a single sample. These reads are then aligned to the reference genome, and the abundance of each transcript is quantified. RNA-Seq is a powerful tool for identifying novel transcripts, splicing variants, and regulatory elements. RNA-Seq data can be used to study gene expression across different cell types, tissues, and conditions, and it is widely used in cancer research to identify potential therapeutic targets and biomarkers.


Tools for Programmers

- Entrez Programming Utilities (E-Utilities) are a set of programs that provide a simple interface for the Entrez retrieval system. E-Utilities use a fixed URL syntax that translates a standard set of input parameters into values necessary for various NCBI software components to search for and retrieve data from Entrez databases.

- Information Engineering Branch - IE is responsible for developing NCBI's resources and databases. Access is provided to documentation, access to NCBI software tools and libraries, and announcements.
Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   - How to access the genome data?
   - How to access the analysis tools?
4. Applications of genomics/bioinformatics
   - Analysis of human and other genomes
5. Future implications
6. Interpretation/global analysis of data
   - Photoreceptors

Applications of Genomics and Proteomics

1. Understand basic biology
2. Diagnosis and treatment of diseases
3. Rationale for drug design
4. Protect plant life
5. Understand bacterial resistance
6. Solve environmental problems
7. Develop new energy sources
8. Improve industrial processes
9. Study evolutionary changes among organisms
Human Genome Sequenced

Celera

The Human Genome Project

www.sciencenews.org

June 23, 2000

The Human Genome

23 pairs of chromosomes
3.2 billion base pairs
Estimated number of genes about 30,000
Only 2% of the human genome “codes”
Average gene size 4000 base pairs
Largest gene dystrophin 2.4 million base pairs
More than 50% in repeat elements or so called “junk DNA”
Analysis of the Human Genome

The DNA sequence of any two people is 99.9 percent identical.

Sites in the DNA sequence where individuals differ at a single DNA base are called single nucleotide polymorphisms (SNPs).

The SNPs may greatly affect an individual's disease risk.

Sickle Cell Anemia

- Sickled red blood cells
- Mutation in the HBB gene that codes for hemoglobin
- one nucleotide change in the 7th codon GAG to GTG
- changing glutamic acid to valine
- interaction between valine and the complementary regions on adjacent molecules results in the formation of polymers that aggregate and distort the shape of the red blood cells

3-D structure of hemoglobin

3-D structure of mutant hemoglobin
What are the genetic changes that make us human?

Only 1.2% difference between human and chimpanzee:
   Every 100<sup>th</sup> nucleotide different
   Could affect thousands of genes
Many are probably the consequence of 6 million years
   of genetic drift, with little effect on body or behavior
Other small changes--perhaps in regulatory, noncoding
   sequences--may have dramatic consequences

What are the genetic changes that make us human?

Study clinical mutations in the genes that impair key traits
   and trace the genes' evolution.
For example, FOXP2, the first ever gene associated with
   human speech.
FOXP2 knock-out mice are squeak less.

A complete understanding of uniquely human traits will,
   however, include more than DNA such as nurture and
   nature.
Understand Bacterial Resistance

Fluoroquinolone antibiotics kill Tuberculosis bacteria by binding to DNA-DNA gyrase complex. Tuberculosis bacterium encodes a novel protein mfpA resembling DNA. mfpA competes with DNA for binding to Fluoroquinolone antibiotics thus making bacteria resistant to the antibiotic.

![DNA and mfpA protein](image)

Science (2005) 308, 1393

Methanocaldococcus jannaschii

Methane-producing thermophilic archeon

Produces methane, an important energy source

encodes enzymes that withstand high temperatures and pressures possibly useful for industrial processes

Photo: © UC Berkeley Electron Microscope Lab (GNN)
**Thalassiosira pseudonana**

Ocean diatom, a major participant in biological pumping of carbon to ocean depths

has potential for mitigating global climate change

Photo: courtesy of DOE-Genomes to Life

---

**Deinococcus radiodurans**

Survives extremely high levels of radiation

has high potential for radioactive waste cleanup

Photo: DOE Joint Genome Institute
What’s Next?????

1. HapMap: Genetic variation mapping project
   Discovery of genes related to diseases
   Gene Testing
   Gene Therapy
2. Pharmacogenomics: Pharmacology and genomics
   Custom effective drugs based on genetic profile
   Reduce adverse reactions
3. ENCODE: Encyclopedia of functional elements
   Study expression of genes

Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   How to access the genome data?
   How to access the analysis tools?
4. Applications of genomics/bioinformatics
   Analysis of human and other genomes
5. Future implications
6. Interpretation/global analysis of data
   Photoreceptors
More Information

Volume 291, Issue 5507, Pages 1145-1434

Volume 409, Issue 6822, Pages 745-964

Nature Supplement on “Human Genome”
http://www.nature.com/nature/supplements/collections/humangenome/index.html

Human Genome Project Information
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

NHGRI Fact sheets
http://www.genome.gov/10000202

Obtain more information about opsin genes and proteins

**Rhodopsin**
- **Blue**: Chromosome 1, Exon 3, Gene: GJB3
- **Red**: Chromosome 1, Exon 8, Gene: GJB3
- **Green**: Chromosome 1, Exon 8, Gene: GJB3
Information about Rhodopsin Gene RHO

Genomic regions, transcripts, and products

Go to reference sequence details

mRNA

Protein

Human Vision
Absorption of light by photoreceptor cells in eye

Rods

Noncolor vision in dim light

pigment opsin (chromophore retinal)

rhodopsin

Absorbs at

500 nm

Cones

Color vision in bright light

blue

426nm

green

530nm

red

560nm

gene

RHO

OPN1SW

OPN1MW

OPN1LW

Chromosome

3

7

X

X

Protein length

348

348

362

362

Rest 40% identity

95% identity
- Only 15 amino acids different
- 3 residues determine the wavelength of absorption
  - At 180 (serine/alanine), 277 (tyrosine/phenyl alanine), 285 (threonine/alanine)
- Hydroxyl containing amino acids in the red pigment interact with the photo excited state of retinal and lower its energy, leading to a shift toward the lower-energy (red) region of the spectrum
- Mutations in these amino acids lead to color “blindness”

Red and Green Genes Susceptible to Unequal Homologous Recombination

- High level of identity between them
- Positioned on chromosome X adjacent to each other
- Leading to different number of individual genes or hybrid genes and thus color “blindness”
- Trouble distinguishing red and green color
- Approximately 5% of males have only the red gene
Humans have three cone pigments: red, green and blue.
Mice are not sensitive to light as far toward the infrared region and they do not discriminate colors well.

Birds, for example, chickens have 4 cone pigments red, green, blue similar to humans and an additional one, violet.

Birds have highly acute color perception.
Fishes have special receptor requirement because of the variation in the amount of light in water.
Green and red photoreceptors are products of a recent evolutionary event.

The green and red pigments appear to have diverged in the primate lineage approximately 35 million years ago.

Mammals, such as dogs and mice, that diverged from primates earlier have only two cone photoreceptors, blue and green, an event believed to have resulted from the nocturnal life.

In contrast, birds such as chickens have a total of six pigments: rhodopsin, four cone pigments, and a pineal visual pigment called pinopsin. Birds have highly acute color perception.

Aquatic environment offers a single system to study evolution of color vision because of the variations in underwater light.

Review articles:
Bowmaker and Hunt Current Biology vol16, R484
Hunt et al. CMLS, Cell.Mol.Lifr Sci. 58, 1583
Obtain More Information about Opsin Genes and Proteins

Information about Rhodopsin Gene RHO

Summary: Retinitis pigmentosa is an inherited progressive disease which is a major cause of blindness in western communities. It can be inherited as an autosomal dominant, autosomal recessive, or X-linked recessive disorder. In the autosomal dominant form, which comprises about 25% of total cases, approximately 30% of families have mutations in the gene encoding the rod photoreceptor-specific protein rhodopsin. This is the transmembrane protein which, when photoactivated, initiates the visual transduction cascade. Defects in this gene are also one of the causes of congenital stationary night blindness.
Automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes

<table>
<thead>
<tr>
<th>Genes</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens RHO rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)</td>
<td>human (NP_000528.1) 340 aa</td>
</tr>
<tr>
<td>P. troglodytes RHO rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)</td>
<td>chimpanzee (XP_616740.2) 340 aa</td>
</tr>
<tr>
<td>C. lupus RHO rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)</td>
<td>dog (NP_853908.1) 356 aa</td>
</tr>
<tr>
<td>M. musculus Rho rhodopsin</td>
<td>mouse (NP_863368.1) 340 aa</td>
</tr>
<tr>
<td>P. troglodytes Rho rhodopsin</td>
<td>rat (NP_254276.1) 340 aa</td>
</tr>
<tr>
<td>G. gallus RHO rhodopsin (opsin 2, rod pigment)</td>
<td>chicken (NP_99082.1) 355 aa</td>
</tr>
<tr>
<td>D. rerio Rho rhodopsin</td>
<td>zebrafish (NP_671158.1) 354 aa</td>
</tr>
</tbody>
</table>

Cluster of transcript sequences that appear to come from the same gene/expressed pseudogene

Restricted Expression (contributing more than half of the EST frequency)

**Hs.247565**: Expression restricted to eye
Information about Rhodopsin Gene RHO

Human Vision

Opsins are 7 transmembrane helix receptors (7TM family)

Chromophore 11 cis-retinal covalently binds to lysine (296) to form positively charged schiff base

A positive schiff base is compensated by glutamate(113)

On absorption of light isomerizes to 11 trans-retinal

Leads to cascade of events that cause hyperpolarization of the membrane and neuronal signaling
Blink: Precalculated top 200 protein BLAST hits
Seven helices

Lys 296

Retinal
Outline

1. What is a genome?
2. What is genomics?
3. What is Bioinformatics?
   - How to access the genome data?
   - How to access the analysis tools?
4. Applications of genomics/bioinformatics
   - Analysis of human and other genomes
5. Future implications
6. Interpretation/global analysis of data
   - Photoreceptors
Bioinformatics

I. Organize data in databases
   researchers can access current data
   submit new data
II. Develop tools and resources to analyze data
III. Interpret data in a biologically useful manner
     global analysis of data to uncover common
     principles that apply across many systems

Vision for Bioinformatics

<table>
<thead>
<tr>
<th>Databases</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entrez global search</td>
<td>BLAST</td>
</tr>
<tr>
<td>Genomes</td>
<td>Blastp</td>
</tr>
<tr>
<td>Books</td>
<td>Blast2 sequences</td>
</tr>
<tr>
<td>Gene</td>
<td>Distance Tree</td>
</tr>
<tr>
<td>Nucleotide</td>
<td>Tax Blast</td>
</tr>
<tr>
<td>Ref/Seq</td>
<td>MapViewer</td>
</tr>
<tr>
<td>Protein</td>
<td>UniGene DDD</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Cn3D</td>
</tr>
<tr>
<td>Homologene</td>
<td>Taxonomy Common Tree</td>
</tr>
<tr>
<td>UniGene</td>
<td>Blink</td>
</tr>
<tr>
<td>AceView</td>
<td>Related Structures</td>
</tr>
<tr>
<td>dbSNP</td>
<td></td>
</tr>
<tr>
<td>OMIM</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td></td>
</tr>
</tbody>
</table>

Photoreceptors cones and rods
Sequence similarity
Phylogeny
Homology
Expression
Structure
Function
Questions about NCBI Resources?

E-mail
info@ncbi.nlm.nih.gov