

Bioinformatics Quick Start



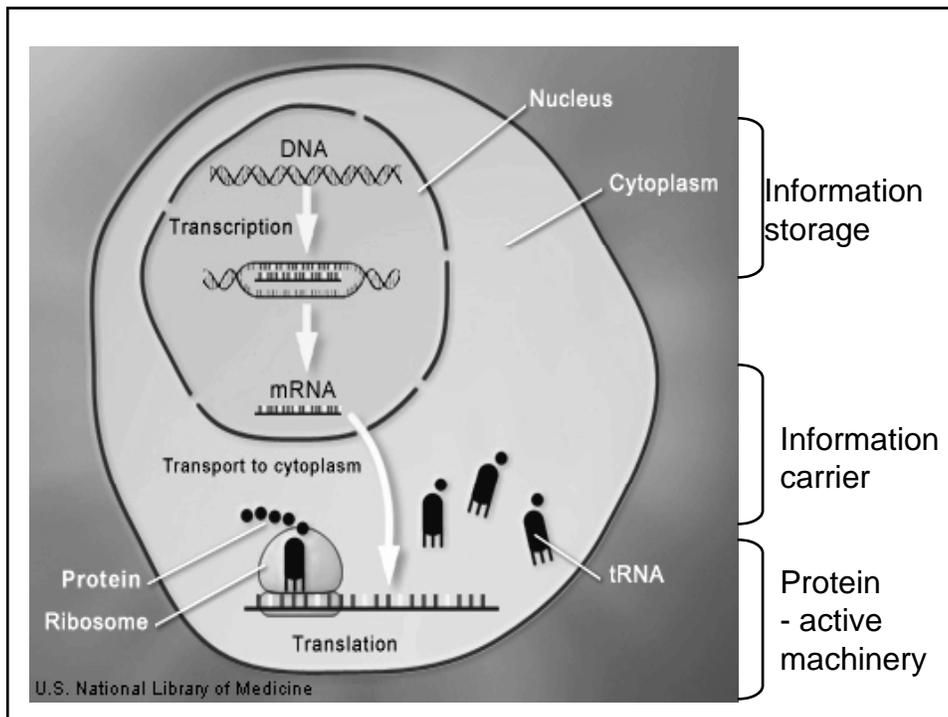
Medha Bhagwat
National Center for Biotechnology Information
National Institutes of Health



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.

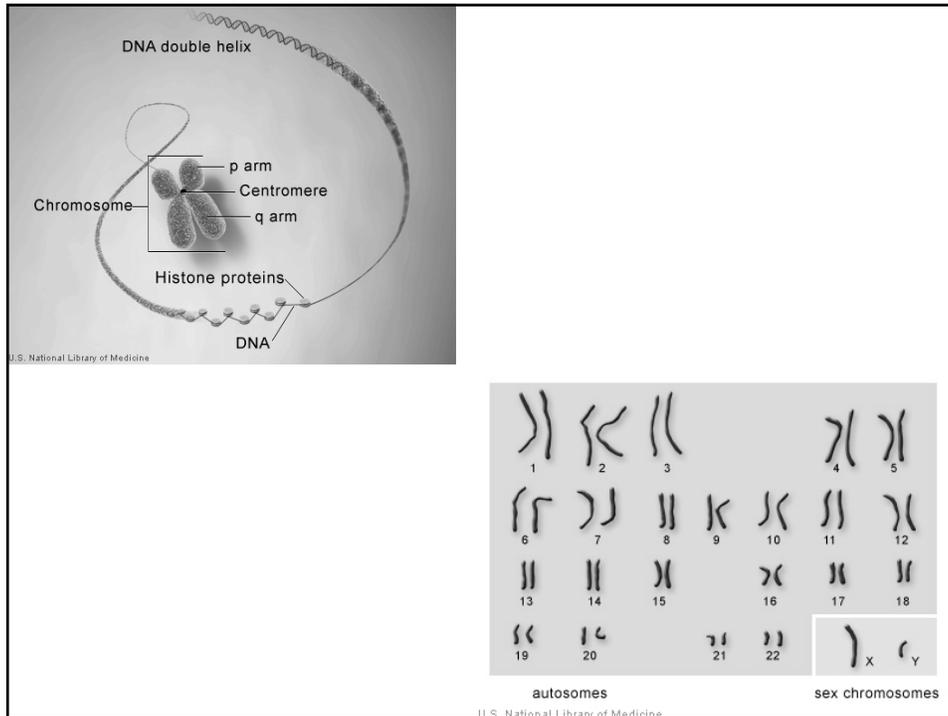
Double helical



The double-helical structure of DNA

<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=stryer.figgrp.146>





Protein

Alphabet: amino acids

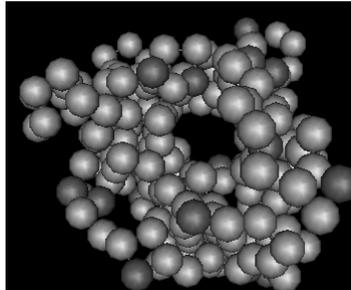
There are 20 amino acids

Encoded by codons (triplets of nucleotides)

ATG TGCAGCCTAGCTGCCGTC

Met—Cys—Ser—Leu —Ala — Ala —Val

Water channel protein

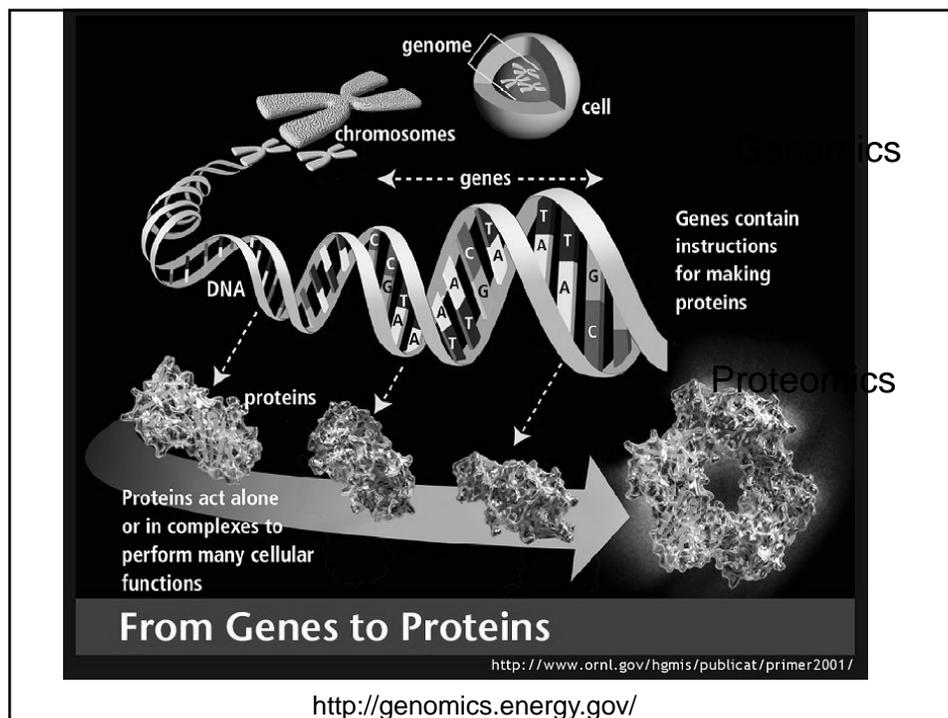


Genome (DNA)

Exact spelling of a word is necessary

CAT	DAT
RAT	GAT
MAT	KAT
FAT	CBT
BAT	CCT
EAT	CAQ
HAT	CAC

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



<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



<http://ghr.nlm.nih.gov/>



<http://www.ncbi.nlm.nih.gov/About/primer/>



<http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books>

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

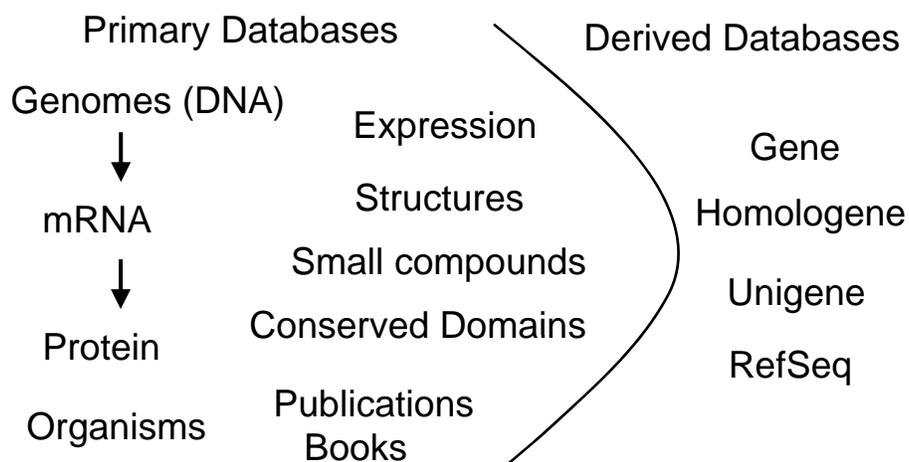
<http://www.ncbi.nlm.nih.gov>

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



NCBI Databases

Primary	Derived
Archival/repository	Curated
Redundant	Non-redundant
Submitter owner	NCBI owner
Sequenced	Combined/edited
Ex: GenBank	Ex: RefSeq

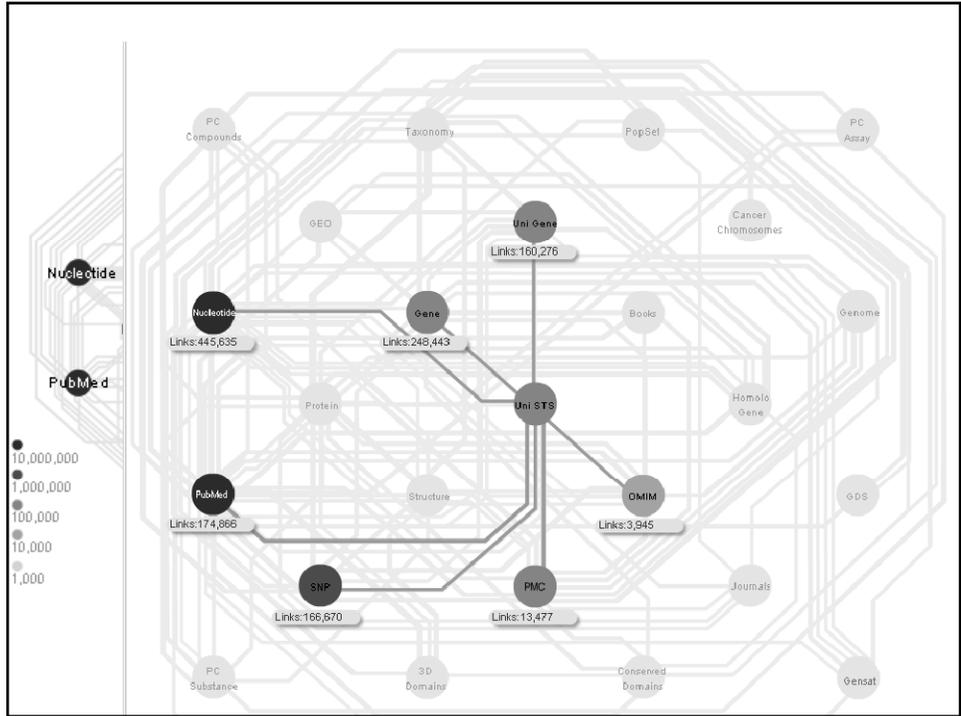


NCBI Databases

<http://www.ncbi.nlm.nih.gov/sites/gquery>

The screenshot displays the NCBI Entrez search engine interface. At the top, there is a search bar with the text "Search across databases" and buttons for "GO", "Clear", and "Help". Below the search bar, a grid of database categories is shown, each with a small icon and a brief description. The categories include:

- PubMed: biomedical literature citations and abstracts
- PubMed Central: free, full text journal articles
- Site Search: NCBI web and FTP sites
- Books: online books
- OMIM: online Mendelian Inheritance in Man
- OMIA: online Mendelian Inheritance in Animals
- CoreNucleotide: Core subset of nucleotide sequence records
- EST: Expressed Sequence Tag records
- GSS: Genome Survey Sequence records
- Protein: sequence database
- Genome: whole genome sequences
- Structure: three-dimensional macromolecular structures
- Taxonomy: organisms in GenBank
- SNP: single nucleotide polymorphism
- Gene: gene-centered information
- HemolaGenes: eukaryotic homology groups
- PubChem Compound: unique small molecule chemical structures
- PubChem Substance: deposited chemical substance records
- Genome Project: genome project information
- dbGaP: genotype and phenotype
- UniGene: gene-oriented clusters of transcript sequences
- CDD: conserved protein domain database
- 3D Domains: domains from Entrez Structure
- UniSTS: markers and mapping data
- PopSet: population study data sets
- GEO Profiles: expression and molecular abundance profiles
- GEO DataSets: experimental sets of GEO data
- Cancer Chromosomes: cytogenetic databases
- PubChem BioAssay: bioactivity screens of chemical substances
- GENSAT: gene expression atlas of mouse central nervous system
- Probes: sequence-specific reagents
- Protein Clusters: a collection of related protein sequences
- Journals: detailed information about the journals indexed in PubMed and other Entrez databases
- NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections
- MeSH: detailed information about NLM's controlled vocabulary



NCBI Databases

NCBI Entrez, The Life Sciences Search Engine

Search across databases [all filter] [GO] [Clear] [Help]

Result counts displayed in gray indicate one or more terms not found

17450582	PubMed: biomedical literature citations and abstracts	191460	Books: online books
1131060	PubMed Central: free, full text journal articles	18997	OMIM: online Mendelian Inheritance in Man
3855	Site Search: NCBI web and FTP sites	2484	OMIA: online Mendelian Inheritance in Animals
41888768	CoreNucleotide: Core subset of nucleotide sequence records	16987	dbGAP: genotype and phenotype
46394888	EST: Expressed Sequence Tag records	3025946	UniGene: gene-oriented clusters of transcript sequences
21105714	GS: Genome Survey Sequence records	23523	CDD: conserved protein domain database
18192257	Protein: sequence database	285813	3D Domains: domains from Entrez Structure
7423	Genome: whole genome sequences	504468	UniSTS: markers and mapping data
45213	Structure: three-dimensional macromolecular structures	66805	PopSet: population study data sets
387300	Taxonomy: organisms in GenBank	39319956	GEO Profiles: expression and molecular abundance profiles
35827062	SNPs: single nucleotide polymorphism	12338	GEO DataSets: experimental sets of GEO data
3723404	Gene: gene-centered information	128042	Cancer Chromosomes: cytogenetic databases
103771	HomoloGene: eukaryotic homology groups	631	PubChem BioAssay: bioactivity screens of chemical substances
10950023	PubChem Compound: unique small molecule chemical structures	64926	GENSAT: gene expression atlas of mouse central nervous system
19615099	PubChem Substance: deposited chemical substance records	8792499	Probes: sequence-specific reagents
2978	Genome Project: genome project information	222499	Protein Clusters: a collection of related protein sequences
21621	Journals: detailed information about the journals indexed in PubMed and other Entrez databases	197817	MeSH: detailed information about NLM's controlled vocabulary
1780789	NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections		

Bookshelf

<http://www.ncbi.nlm.nih.gov/sites/entrez?db=books>



Genome Sequence Data and Analysis Tools at NCBI

<http://www.ncbi.nlm.nih.gov/Genomes/>

NCBI provides several genomic biology tools and resources, including organism-specific pages that include links to many web sites and databases relevant to that species. We invite you to explore the links provided on this page.

Genomic Biology

Assembly and Annotation Information

- AGP Resources
- Annotation Information
- Assembly Information
- Genome Glossary
- NCBI Handbook, Chapter 14: Genome Assembly and Annotation Process

Announcements

July, 2007
New Genome Resource Page: A genome resource guide is now available for *Ornithorhynchus anatinus* (platypus).

Map Viewer - genome annotation updates:

Species	Build	Map Viewer Release
<i>Caenorhabditis elegans</i> (nematode)	WS170	July 18, 2007
<i>Equus caballus</i> (domestic horse)	1.1	July 11, 2007
<i>Ornithorhynchus anatinus</i> (platypus)	1.1	July 11, 2007
<i>Manus musculus</i> (muskrat)	37.1	July 5, 2007
<i>Monodelphis domestica</i> (opossum)	MonDom5	March 8, 2007
<i>Danio rerio</i> (zebrafish)	2.6	February 27, 2007
<i>Populus trichocarpa</i> (black cottonwood)	1.1	January 12, 2007
<i>Bos taurus</i> (cow)	3.1	January 3, 2007
<i>Gallus gallus</i> (chicken)	2.1	November 30, 2006
<i>Drosophila melanogaster</i> (fruit fly)	5.1	November 21, 2006
<i>Dictyostelium discoideum</i>	2.1	November 2, 2006
<i>Strongylocentrotus purpuratus</i> (sea urchin)	2.1	October 18, 2006
<i>Pan troglodytes</i> (chimpanzee)	2.1	October 5, 2006
<i>Homo sapiens</i> (human)	36.2	September 14, 2006

Genome Resources

- Entrez Genome
- Fungal Genomes Central
- Genome Projects Database
- Bacterial
- Fungi
- Insects
- Mammals
- Microbial
- Plants
- Map Viewer
- Organelles
- Plant Genomes Central
- Viral Resources
- Influenza Virus Resource
- Retroviruses
- Viral Genomes

Organism-Specific

- Genome Resources
- BLAST
- Map Viewer
- Genome Project DB
- Arabidopsis
- Aspergillus
- Bee
- Beetle
- Cat
- Chicken
- Chimpanzee
- Cow
- Dictyostelium

Sizes of Different Genomes

Aloe vera	16.0 billion
Rabbit	3.5 billion
Human	3.2 billion
Laboratory mouse	2.6 billion
Fruit fly	137 million
Yeast	12.1 million
Bacterium (<i>E. coli</i>)	4.6 million
Human immunodeficiency virus	9700



Tools for Data Mining

[BLAST](#) [OMIM](#) [Books](#) [TaxBrowser](#) [Structure](#)
 for
[Analysis](#) | [Protein Sequence Analysis](#) | [Structures](#) | [Genome Analysis](#) | [Gene Expression](#)

Tools: Nucleotide Sequence Analysis

BLAST - The Basic Local Alignment Search Tool (BLAST) for comparing a gene and protein sequences against others in public databases, now comes in several types including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, malaria, and other genomes, as well as for vector contamination, immunoglobulins, and tentative human consensus sequences.

Electronic PCR - allows you to search your DNA sequence for sequence tagged sites (STSs) that have been used as landmarks in various types of genomic maps. It compares the query sequence against data in NCBI's UniSTS, a unified, non-redundant view of STSs from a wide range of sources.

Entrez Gene - each Entrez Gene record encapsulates a wide range of information 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 and 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 related resources, such as mutation databases. Entrez Gene is a successor to LocusLink.

Model Maker - allows you to view the evidence (mRNAs, ESTs, and gene predictions) that was aligned to assembled genomic sequence to build a gene model and to edit the model by selecting or removing putative exons. You can then view the mRNA sequence and potential ORFs for the edited model and save the mRNA sequence data for use in other programs. Model Maker is accessible from sequence maps that were analyzed at NCBI and displayed in Map Viewer.

ORF Finder - identifies all possible ORFs in a DNA sequence by locating the standard and alternative stop 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 Sequin.

Organism Specific Resources - Bee, Cat, Chicken, Cow, etc.

<http://www.ncbi.nlm.nih.gov/Tools/>

SAGEmap - provides a tool for performing statistical tests designed specifically for differential-type analyses of SAGE (Serial Analysis of Gene Expression) data. The data include SAGE libraries generated by 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 pages. It is possible to enter a query sequence in the SAGEmap resource to determine what SAGE tags are in the sequence, then map to associated SAGETag records and view the expression of those tags in different CGAP SAGE libraries.

Spidey - Spidey - aligns one or more mRNA sequences to a single genomic sequence. Spidey will try to determine the exon/intron structure, returning one or more models of the genomic structure, including the genomic/mRNA alignments for each exon.

Splign - Splign - is a utility for computing cDNA-to-Genomic alignments based on a variation of the Needleman-Wunsch algorithm combined with Blast for compartment detection and greater performance.

VecScreen - a tool for identifying segments of a nucleic acid sequence that may be of vector, linker, or adaptor origin prior to sequence analysis or submission. VecScreen was developed to combat the problem of vector contamination in public sequence databases.

Viral Genotyping Tool - a web-based program that identifies the genotype (or subtype) of recombinant or non-recombinant viral nucleotide sequences. It works by using BLAST to compare a query sequence to a set of reference sequences for known genotypes. Predefined reference genotypes exist for three major viral pathogens: human immunodeficiency virus 1 (HIV-1), hepatitis C virus (HCV) and hepatitis B virus (HBV), as well as for poliovirus. User-defined reference sequences can be used at the same time. The query sequence is broken into segments for comparison to the reference so that the mosaic organization of recombinant sequences is revealed. The results are displayed graphically using color-coded genotypes. Therefore, the genotype(s) of any portion of the query can quickly be determined.

Genome Sequence Data and Analysis Tools at NCBI

Tools - Protein Sequence Analysis and Proteomics

BLAST - The Basic Local Alignment Search Tool (BLAST) for comparing gene and protein sequences against others in public databases, now comes in several types including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, malaria, and other genomes, as well as for vector contamination, immunoglobulins, and tentative human consensus sequences.

BLINK - ("BLAST Link") displays the results of BLAST searches that have been done for every protein sequence in the Entrez Proteins data domain.

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

CDART - when given a protein query sequence, CDART displays the functional domains that make up the protein and lists proteins with similar domain architectures.

OMSSA - The Open Mass Spectrometry Search Algorithm (OMSSA) - The OMSSA search service allows proteomics researchers to submit the mass spectra of peptides and proteins for identification. OMSSA then compares these mass spectra to theoretical ions generated from data libraries of known protein sequences and ranks the results using a score derived from classical hypothesis testing.

TopPlot - a tool for 3-way comparisons of genomes on the basis of the protein sequences they encode. To use TopPlot, one selects a reference genome to which two other genomes are compared. Pre-computed BLAST results are then used to plot a point for each predicted protein in the reference genome, based on the best alignment with proteins in each of the two genomes being compared.

Tools - Structures

3D Search - Use a help or application for your web browser that allows you to view 3-dimensional structures from NCBI's Entrez retrieval service. 3D Search runs on Windows, Macintosh, and Unix.

VAST Search - VAST Search is NCBI's structure-structure similarity search service. It compares 3-D coordinates of a newly determined protein structure to those in the MMDB/PPD database.

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

Tools - Genome Analysis

Entrez Genomes - Complete genomes of over 1000 organisms. The Entrez Genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life - bacteria, archaea, and eukaryota - are represented, as well as many viruses, plasmids, viroids, plasmids, and organelles. Entrez Genomes provides graphical overviews of complete genomes/chromosomes and the ability to explore regions of interest in progressively greater detail.

COGs - Clusters of Orthologous Groups - a natural system of gene families from complete genomes. Clusters of Orthologous Groups (COGs) were delineated by comparing protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus constitutes an ancient conserved domain.

Map Viewer - Shows integrated views of chromosome maps for many organisms, including human and numerous other mammals, invertebrates, fungi, protozoa, and plants. Map Viewer is used to view assembled genomes (either draft or complete) and is a valuable tool for the identification and localization of genes and other biological features. Multiple map displays are aligned based on shared marker and gene names when available, and sequence map displays are based on a common sequence coordinate system. Sequence data for chromosome regions of interest can be downloaded. Biological annotations can be viewed in graphical format and/or downloaded in tabular format, and gene models can be manipulated in the associated ModelMaker tool.

SKY-FISH & COH Database - The NCI and NCBI SKY-FISH and COH Database is a repository of publicly submitted data from Spectral Karyotyping (SKY), Multiplex Fluorescence In Situ Hybridization (M-FISH), and Comparative Genomic Hybridization (CGH), which are complementary fluorescent molecular cytogenetic techniques. SKY-FISH permits the simultaneous visualization of each human or mouse chromosome in a different color, facilitating the identification of chromosomal aberrations; COH can be used to generate a map of DNA copy number changes in tumor genomes. Collaborative project with the National Cancer Institute. (data submission instructions...)

<http://www.ncbi.nlm.nih.gov/Tools/>

Genome Sequence Data and Analysis Tools at NCBI

Tools - Gene Expression

GEO Gene Expression Omnibus - The Gene Expression Omnibus (GEO) provides several tools to assist with the visualization and exploration of GEO data. Datasets may be viewed as hierarchical cluster heat maps, providing insight into the relationships between samples and co-regulated genes. Individual gene expression profiles showing significant differences between experimental subsets may be located using average subset rank value comparisons. Related gene expression profiles may be identified on the basis of sequence similarity, profile similarity, or homology. Indicators of dataset normalization quality are provided as distribution graphs, and by flagging outliers. Links to other NCBI sequence, mapping and publication database resources are provided where possible.

SAGEmap - provides a tool for performing statistical tests designed specifically for differential-type analyses of SAGE (Serial Analysis of Gene Expression) data. The data include SAGE libraries generated by 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 pages. It is possible to enter a query sequence in the SAGEmap resource to determine what SAGE tags are in the sequence, then map to associated SAGETag records and view the expression of those tags in different CGAP SAGE libraries.

The Cancer Genome Anatomy Project (CGAP) - aims to decipher the molecular anatomy of cancer cells. CGAP develops profiles of cancer cells by comparing gene expression in normal, precancerous, and malignant cells from a wide variety of tissues.

UniGene DDD - Digital Differential Display - an online tool to compare computed gene expression profiles between selected cDNA libraries. Using a statistical test, genes whose expression levels differ significantly from one tissue to the next are identified and shown to the user. Additional information about UniGene is above, including a list of organisms represented.

<http://www.ncbi.nlm.nih.gov/Tools/>

Tools for Programmers

Entrez Programming Utilities - E-Utilities are a set of programs that provide a stable interface into the Entrez retrieval system. The eUtilities 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 23 Entrez databases.

Information Engineering Branch - IEB 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



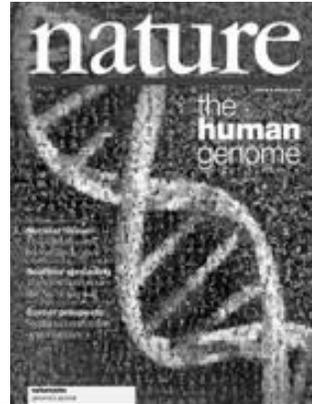
President Clinton and geneticists J. Craig Venter (left) and Francis Collins (right) celebrate.

www.sciencenews.org

June 23, 2000



The Human Genome Project



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

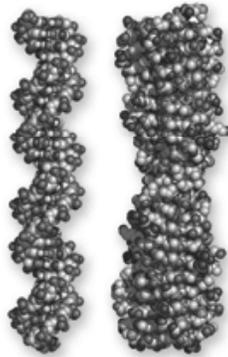
3-D structure of hemoglobin

3-D structure of mutant hemoglobin

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

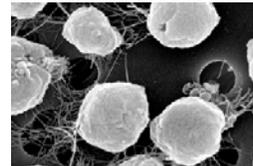


mfpA protein

Science (2005) 308, 1393

Methanocaldococcus jannaschii

Methane-producing thermophilic archeon

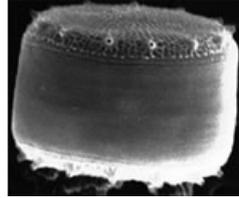


Produces methane, an important energy source

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

Photo: © UC Berkeley Electron Microscope Lab (GNN)

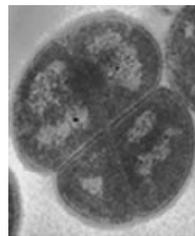
Thalassiosira pseudonana



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



What's Next?????



Nature, Volume 449 Number 7164
18 October 2007



What's Next?????

Catechol-O-Methyltransferase (COMT) gene variants predict response to bupropion therapy for tobacco dependence.

COMT genotyping could be applied to identify likely responders to bupropion treatment for smoking cessation.

Biol Psychiatry 2007, 61:111-118

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>

NCBI National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search All Databases for **Vision** Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to
NCBI

GenBank
Sequence
submission support
and software

**Literature
databases**
PubMed, OMIM,
Books, and PubMed
Central

**Molecular
databases**
Sequences,
structures, and
taxonomy

Genomic biology
The human genome,
whole genomes,
and related
resources

What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More ...](#)

Hot Spots

- ▶ Assembly Archive
- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home
- ▶ Entrez Tools
- ▶ Gene expression omnibus (GEO)
- ▶ Human genome resources
- ▶ Influenza Virus Resource
- ▶ Map Viewer
- ▶ dbMHC

New Protein Clusters
Entrez Protein Clusters database

The new Entrez Protein Clusters database is a collection of Reference Sequence (RefSeq) proteins, from the complete genomes of prokaryotes, plasmids, and organelles, that have been grouped and annotated based on sequence similarity and protein function. [Click here to find out more about the Protein Clusters database.](#)

1 Billion Live Traces
The Trace Archive of sequencing traces has reached 1 billion live traces from over 480 organisms. For more information about the Trace Archive database [click here.](#)

<http://www.ncbi.nlm.nih.gov/>

NCBI

NCBI  Entrez, The Life Sciences Search Engine

SEARCH | SITE MAP | PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST

Search across databases

Result counts displayed in gray indicate one or more terms not found

117902	PubMed: biomedical literature citations and abstracts	1264	Books: online books
35337	PubMed Central: free, full text journal articles	321	OMIM: online Mendelian Inheritance in Man
7	Site Search: NCBI web and FTP sites	13	OMIA: online Mendelian Inheritance in Animals

3949	CoreNucleotide: Core subset of nucleotide sequence records	15	dbGAP: genotype and phenotype
61093	EST: Expressed Sequence Tag records	33	UniGene: gene-oriented clusters of transcript sequences
16477	GSS: Genome Survey Sequence records	2	CDD: conserved protein domain database
1963	Protein: sequence database	40	3D Domains: domains from Entrez Structure
3	Genome: whole genome sequences	77	UniSTS: markers and mapping data
8	Structure: three-dimensional macromolecular structures	44	PopSet: population study data sets
none	Taxonomy: organisms in GenBank	8443	GEO Profiles: expression and molecular abundance profiles
none	SNP: single nucleotide polymorphism	20	GEO DataSets: experimental sets of GEO data
104	Gene: gene-centered information	none	Cancer Chromosomes: cytogenetic databases
642	HomoloGene: eukaryotic homology groups	none	PubChem BioAssay: bioactivity screens of chemical substances
2	PubChem Compound: unique small molecule chemical structures	194	GENSAT: gene expression atlas of mouse central nervous system
15	PubChem Substance: deposited chemical substance records	212	Probe: sequence-specific reagents
2	Genome Project: genome project information	none	Protein Clusters: a collection of related protein sequences

Bookshelf

Published | Multimedia | PubMed | Gateway | Structure

Limits | Preview/Index | History | Clipboard | Details

Display Books | Show 20 | Send to

All 1264 Figures: 23

- 551 items in **Health Services Technology Assessment Text (HSTAT)**
Bethesda (MD): National Library of Medicine (NLM), 2003 Cit.
- 271 items in **GeneReviews**
Editors-in-Chief: Fagan, Roberta A. Associate editors: Cassidy, Suzanne B., Bird, Thomas C., Dondos, Mary Beth Seattle (WA): University of Washington, 1993-2007
- 52 items in **Clinical Methods Third Edition**
Walker, H.K., Hall, W.D., Bart, J.W., editors
Shelton (MA): Butterworth Publishers, c1990
- 43 items in **Neuroscience**, 2nd ed.
Parvez, Dair, Augustine, George J., Fitzpatrick, David, Katz, Lawrence C., LeMaitre, Anthony-Samuel, McIlwain, James O., Williams, R. Mark, editors
Bundelwood (GA): Smart Associates, Inc., c2001
- 37 items in **Cancer Medicine**, 6th ed.
Kufe, Donald W., Folz, Robert E., Waisborthman, Ralph R., Bart, Robert C., Jr., Ossola, Ted S., Holland, J. Hamilton (Cincinnati, OH): Taylor & Francis, c2003
- 31 items in **Global Burden of Disease and Risk Factors**
Alan D. Lopez, Colin D. Mathers, Majid Ezzati, David T. Jamison, Christopher L. Murray, editors
Washington (DC): WHO/The World Bank and Oxford University Press, 2006
- 27 items in **Insulators of Epilepsy**, 2nd ed.
Explos, Peter W., Fisher, Robert S., editors
New York: Demos Medical Publishing, c2003
- 25 items in **Disease Control Priorities in Developing Countries**, 2nd ed.
- 15 items in **Basic Neurochemistry, Molecular, Cellular, and Medical Aspects**, 6th ed.
Singh, George J., Agrasoff, Donald W., Albers, R. Wayne, Fisher, Stephen K., Uhlir, Michael D., editors.
Philadelphia: Lippincott, Williams & Wilkins, c1999.
- 15 items in **Spinal Cord Medicine: Principles and Practice**
Liu, Vernon W., editor
New York: Demos Medical Publishing, Inc., c2003
- 15 items in **Physical Medicine and Rehabilitation Board Review**
Cuccunillo, Sara J., editors
New York: Demos Medical Publishing, Inc., c2004
- 13 items in **Psychiatry**
Berg, Jeremy M., Tymoczko, John L., and Stryer, Lubert.
New York: W. H. Freeman and Co., 2002.
- 12 items in **Collective Expert Evaluation Reports**
INSERM Collective Expertise Centre
Paris: Institut National de la Santé et de la Recherche Médicale (INSERM), c2000-2004
- 12 items in **Health, United States, 2005**
Atlanta (GA): Centers for Disease Control and Prevention, 2005
- 11 items in **Parkinson's Disease: Diagnosis and Clinical Management**
Factor, Stewart A.; Weiner, William J.
New York: Demos Medical Publishing, Inc., c2002
- 10 items in **Health, United States, 2006**
Atlanta (GA): Centers for Disease Control and Prevention, c2006
- 8 items in **Molecular Biology of the Cell**, 4th ed.
Alberts, Bruce, Johnson, Alexander, Lewis, Julian, Raff, Martin, Roberts, Keith, Walter, Peter
New York and London: Garland Science, c2002
- 8 items in **WormBook: The Online Review of C. elegans Biology**
The C. elegans Research Community, editors
Pasadena (CA): WormBook, c2005
- 8 items in **Surgical Treatment**
Holzheimer, Rene G.; Mannick, John A., editors.
Munich: Zuckschwerdt Publishers, c2001.
- 8 items in **Medical Microbiology**, 4th ed.
Baron, Sami, editor.
Galveston (TX): University of Texas Medical Branch, c1996.

Items 1 - 13 of 13

- 1: [Smell, Taste, Vision, Hearing, and Touch Are Based on Signal-Transduction Pathways Activated by Signals from the Environment](#)
Biochemistry -> Responding to Environmental Changes -> Sensory Systems -> Summary
- 2: [Sensory Systems](#)
Biochemistry -> Responding to Environmental Changes
- 3: [Seven-Transmembrane-Helix Receptors Change Conformation in Response to Ligand Binding and Activate G Proteins](#)
Biochemistry -> Transducing and Storing Energy -> Signal-Transduction Pathways: An Introduction to Information Metabolism
- 4: [Fat-Soluble Vitamins Participate in Diverse Processes Such as Blood Clotting and Vision](#)
Biochemistry -> The Molecular Design of Life -> Enzymes: Basic Concepts and Kinetics -> Vitamins Are Often Precursors to Coenzymes
- 5: [Color Vision Is Mediated by Three Cone Receptors That Are Homologs of Rhodopsin](#)
Biochemistry -> Responding to Environmental Changes -> Sensory Systems -> Photoreceptor Molecules in the Eye Detect Visible Light
- 6: [Color Perception](#)
Biochemistry -> Responding to Environmental Changes -> Sensory Systems
- 7: [Photoreceptor Molecules in the Eye Detect Visible Light](#)
Biochemistry -> Responding to Environmental Changes -> Sensory Systems
- 8: [Photoreceptor Molecules in the Eye Detect Visible Light](#)
Biochemistry -> Responding to Environmental Changes -> Sensory Systems -> Summary
- 9: [Chapter Integration Problem](#)
Biochemistry -> Responding to Environmental Changes -> Sensory Systems -> Problems
- 10: [Hearing Depends on the Speedy Detection of Mechanical Stimuli](#)
Biochemistry -> Responding to Environmental Changes -> Sensory Systems
- 11: [Cone-Pigment Absorption Spectra](#)
Biochemistry -> Responding to Environmental Changes -> Sensory Systems -> Photoreceptor Molecules in the Eye Detect Visible Light
- 12: [Biological functions mediated by 7TM receptors](#)
Biochemistry -> Transducing and Storing Energy -> Signal-Transduction Pathways: An Introduction to Information Metabolism -> Seven-Transmembrane-Helix Receptors Change Conformation in Response to Ligand Binding and Activate G Proteins
- 13: [Fat-soluble vitamins](#)
Biochemistry -> The Molecular Design of Life -> Enzymes: Basic Concepts and Kinetics -> Vitamins Are Often Precursors to Coenzymes

Obtain more information about opsin genes and proteins

Entrez Gene

PubMed Nucleotide Protein

for [opsin [Gene/Protein name] AND human[orgn]]

Limits: **Current Records**

Display: Summary show 20

All: 8 Current Only: 8 Genes Genomes: 8 SNP GeneView

Items 1 - 8 of 8

- 1: **OPN1LW** **Blue**
Official Symbol OPN1LW and Name: opsin 1 (cone pigment)
Other Aliases: BCF, EOP, CBT
Other Designations: blue cone photoreceptor pigment, bla
Chromosome: 7, Location: 7q31.3-q32
Annotation: Chromosome 7, NC_000007.12 (128199783)
MIM: 190900
GeneID: 611
- 2: **OPN3**
Official Symbol OPN3 and Name: opsin 3 (encephalopsin)
Other Aliases: ECFN
Other Designations: opsin 3
Chromosome: 1, Location: 1q43
Annotation: Chromosome 1, NC_000001.9 (239823075)
MIM: 606695
GeneID: 23596
- 3: **OPN4**
Official Symbol OPN4 and Name: opsin 4 (melanopsin)
Other Aliases: MGC142118, MOP
Other Designations: melanopsin, opsin 4
Chromosome: 10, Location: 10q22
Annotation: Chromosome 10, NC_000010.9 (88404354)
MIM: 606665
GeneID: 94233
- 4: **OPN1M2**
Official Symbol OPN1M2 and Name: opsin 1 (cone pigments), medium-wave-sensitive 2 [*Homo sapiens*]
Chromosome: X, Location: Xq28
Annotation: Chromosome X, NC_000003.9 (333328161..333351953)
GeneID: 728458 **Rhodopsin**
- 5: **RHO**
Official Symbol RHO and Name: rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant) [*Homo sapiens*]
Other Aliases: MGC138309, MGC138311, OPN2, RP4
Other Designations: rhodopsin
Chromosome: 3, Location: 3q21-q24
Annotation: Chromosome 3, NC_000003.10 (130730172..130736877)
MIM: 180380
GeneID: 6010 **Red**
- 6: **OPN1LW**
Official Symbol OPN1LW and Name: opsin 1 (cone pigments), long-wave-sensitive (color blindness, protan) [*Homo sapiens*]
Other Aliases: CEBM, CBP, RCP
Other Designations: red cone pigment
Chromosome: X, Location: Xq28
Annotation: Chromosome X, NC_000023.9 (153062934..153077705)
MIM: 303900
GeneID: 5956
- 7: **OPN5**
Official Symbol OPN5 and Name: opsin 5 [*Homo sapiens*]
Other Aliases: GPR136, PGR12, TMEM13
Other Designations: G protein-coupled receptor 136, OTTHUMP00000016565, OTTHUMP00000039910, neuropilin, transmembrane protein 136
Chromosome: 6, Location: 6p12.3
Annotation: Chromosome 6, NC_000006.10 (47857757..47902076)
MIM: 609042
GeneID: 221391 **Green**
- 8: **OPN1MW**
Official Symbol OPN1MW and Name: opsin 1 (cone pigments), medium-wave-sensitive (color blindness, deutan) [*Homo sapiens*]
Other Aliases: CEBM, CBD, GCP, OPN1MW1
Other Designations: green cone pigment, photopigment ap-protein
Chromosome: X, Location: Xq28
Annotation: Chromosome X, NC_000023.9 (153101361..153114725)
MIM: 303800
GeneID: 2652

Red	MAQQWSLQRLACRHPQDSYEDSTQSSIFTYTNENST RCP FEGPNYHIAPRUVYHLSVWM	60
Green	MAQQWSLQRLACRHPQDSYEDSTQSSIFTYTNENST RCP FEGPNYHIAPRUVYHLSVWM	60
Red	IFVVVASVF TNGLVLAATMKFRKL RHPLNWI LWNLA VAD LAETVIASTI SIVNQVSGYFV	120
Green	IFVVLASVF TNGLVLAATMKFRKL RHPLNWI LWNLA VAD LAETVIASTI SIVNQVGYFV	120
Red	LGHPMCYLECYTWS LC GIT GLWSLAI ISWERMLVCKPFGVRFDAKLAIVGLAFSWINIS	180
Green	LGHPMCYLECYTWS LC GIT GLWSLAI ISWERMLVCKPFGVRFDAKLAIVGLAFSWINIA	180
Red	AVWTAPP IFCWS RYWPHELKTS CGPDVFS GSSYPGVQSYMI VLMVT CCIIP LAIIMLCYL	240
Green	AVWTAPP IFCWS RYWPHELKTS CGPDVFS GSSYPGVQSYMI VLMVT CCIIP LSIIVLCYL	240
Red	QVWLAI RAVAKQKES EST QKAEKEV TRMVVVM I FAYCV CGPPTT FACFAAANPCYAFH	300
Green	QVWLAI RAVAKQKES EST QKAEKEV TRMVVVM I FCF CGPPTT FACFAAANPCYPPH	300
Red	PLMAALPAYFAKSATTIYNPVIYVFMNRQF RNCILQLFGKRWDDGSELSSASKTEVSVSS	360
Green	PLMAALPAF FAKSATTIYNPVIYVFMNRQF RNCILQLFGKRWDDGSELSSASKTEVSVSS	360
Red	VSPA	364
Green	VSPA	364

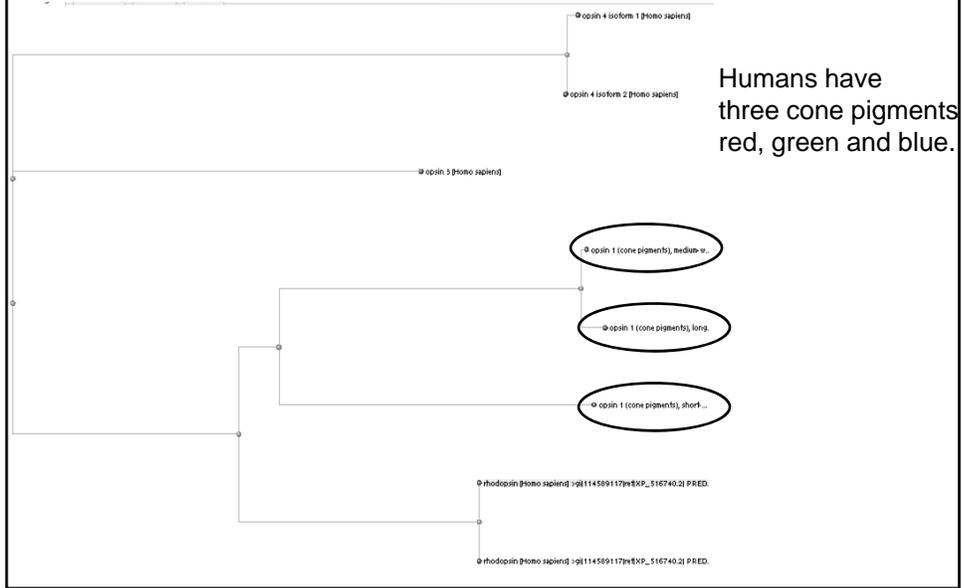
- Only 15 amino acids different
- 3 residues determine the wavelength of absorption
 - At 180 (serine/alanine), 277 (tyrosine/phenyl alanine)
 - 285 (threonine/alanine)
- Hydroxyl containing amino acids in the red pigment interact with the photo excited state of retinal and lower its energy, leading 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

Distance Tree Generated from the BLAST Results of Rhodopsin Protein against Human RefSeq Proteins



Common Lineage Tree Generated from the Taxonomy Browser

<http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi>

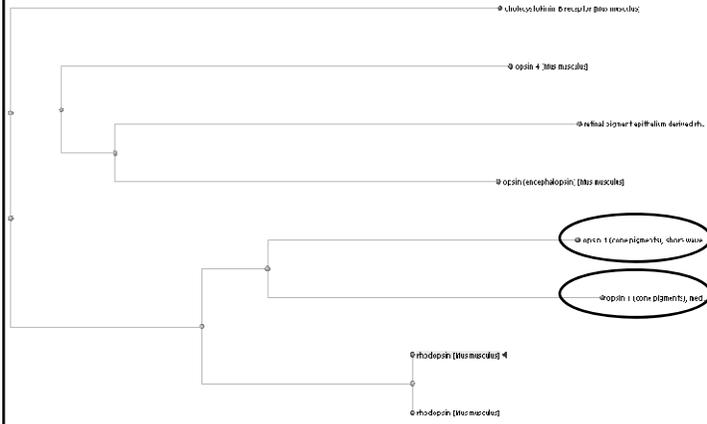
The screenshot shows the NCBI Taxonomy Browser interface. At the top, there are logos for NCBI, PubMed, Entrez, BLAST, and OMM. Below the logos is a search bar with the text "Enter name or id" and buttons for "Add", "or", "Add from file:", "Browse...", and "Choose subset". There are also buttons for "Expand All", "Collapse All", "Mark selected taxa", "Browse tree", "Delete taxa", and "Save as" with a dropdown menu set to "text tree".

The main content area displays a common lineage tree starting with "Euteleostomi". The tree is expanded to show the following taxa:

- Amniota
 - Aves (birds)
 - Gallus gallus (chicken)
 - Euthera
 - Canis lupus familiaris (dog)
 - Euarchontoglires
 - Murinae
 - Rattus norvegicus (rat)
 - Mus musculus (mouse)
 - Homo/Pan/Gorilla group
 - Homo sapiens (human)
 - Pan troglodytes (chimpanzee)
- Danio rerio (zebrafish)



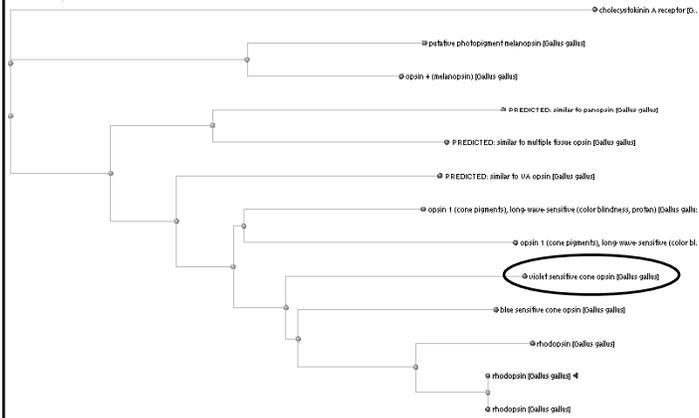
Distance tree Generated from the BLAST Results of Mouse Rhodopsin Protein against Mouse RefSeq Proteins



Mice have only two cone pigments, blue and green.

Mice are not sensitive to light as far toward the infrared region and they do not discriminate colors well.

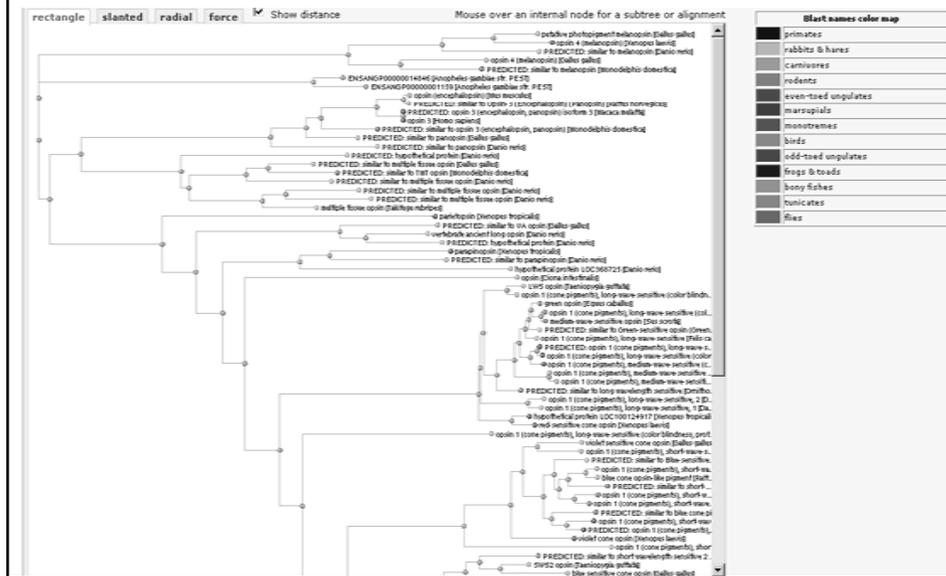
Distance Tree Generated from the BLAST Results of Chicken Rhodopsin Protein against Chicken RefSeq Proteins



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.

Blast Distance Tree of Animal Proteins Similar to Human Rhodopsin (contd)



Tax BLAST Report of the Previous BLAST Search

Lineage Report

Lineage	Count	Hit	Species	Accession	Score	E-value
Coelomata [animals]						
Chordata [chordates]						
Mammalia [mammals]						
Primates [primates]						
Homo sapiens (man)	663	3 hits	[primates]		652	0.0
Pan troglodytes	652	4 hits	[primates]		304	1e-81
Oryctolagus cuniculus (domestic rabbit)	660	1 hit	[rabbits & hares]		278	6e-74
Rattus norvegicus (brown rat)	651	4 hits	[rodents]		278	6e-74
Mus musculus (mouse)	661	4 hits	[rodents]		276	2e-73
Felis catus (cat)	660	3 hits	[carnivores]		286	2e-76
Bos taurus (cow)	641	4 hits	[even-toed ungulates]		236	1e-79
Sus scrofa (wild boar)	638	3 hits	[even-toed ungulates]		274	1e-72
Canis lupus familiaris (dog)	629	5 hits	[carnivores]		274	1e-72
Equus caballus (equine)	586	3 hits	[odd-toed ungulates]		274	1e-72
Monodelphis domestica	640	5 hits	[marsupials]		274	1e-72
Ornithorhynchus anatinus (duck-billed platypus)	629	3 hits	[monotremes]		274	1e-72
Taeniopygia guttata (zebra finch)	597	5 hits	[birds]		274	1e-72
Gallus gallus (bantam)						
Xenopus laevis (clawed frog)						
Xenopus tropicalis						
Panthera leo (leopard)						
Takinops rimbriipes (rorafugu)						
Clona intestinalis						
Anopheles gambiae str. PEST						
Tetrapoda [vertebrates]						
Amniota [vertebrates]						
Mammalia [mammals]						
Theria [mammals]						
Eutheria [placentals]						
Euarchontoglires [placentals]						
Catarrhini [primates]						
Macaca mulatta (rhesus macaque)	663	3 hits	[primates]		652	0.0
Homo sapiens (man)	663	3 hits	[primates]		304	1e-81
Pan troglodytes	652	4 hits	[primates]		278	6e-74
Oryctolagus cuniculus (domestic rabbit)	660	1 hit	[rabbits & hares]		278	6e-74
Rattus norvegicus (brown rat)	651	4 hits	[rodents]		276	2e-73
Mus musculus (mouse)	661	4 hits	[rodents]		286	2e-76
Felis catus (cat)	660	3 hits	[carnivores]		236	1e-79
Bos taurus (cow)	641	4 hits	[even-toed ungulates]		274	1e-72
Sus scrofa (wild boar)	638	3 hits	[even-toed ungulates]		274	1e-72
Canis lupus familiaris (dog)	629	5 hits	[carnivores]		274	1e-72
Equus caballus (equine)	586	3 hits	[odd-toed ungulates]		274	1e-72
Monodelphis domestica	640	5 hits	[marsupials]		274	1e-72
Ornithorhynchus anatinus (duck-billed platypus)	629	3 hits	[monotremes]		274	1e-72
Taeniopygia guttata (zebra finch)	597	5 hits	[birds]		274	1e-72
Gallus gallus (bantam)						
Xenopus laevis (clawed frog)						
Xenopus tropicalis						
Panthera leo (leopard)						
Takinops rimbriipes (rorafugu)						
Clona intestinalis						
Anopheles gambiae str. PEST						
Marsupialia [marsupials]						
Didymelasma [marsupials]						
Pan troglodytes [primates] taxid 9598						
ref NP_000301.1 rhodopsin [Homo sapiens]					652	0.0
ref NP_001699.1 opsin 1 (cone pigments), short-wave-sens...					304	1e-81
ref NP_000504.1 opsin 1 (cone pigments), medium-wave-sens...					278	6e-74
ref NP_00104646.1 opsin 1 (cone pigments), medium-wave-s...					278	6e-74
ref NP_064445.1 opsin 1 (cone pigments), long-wave-sensit...					276	2e-73
ref NP_055137.2 opsin 3 [Homo sapiens]					157	1e-37
Pan troglodytes [primates] taxid 9598						
ref XP_516740.2 PREDICTED: rhodopsin [Pan troglodytes]					652	0.0
ref XP_001009127.1 opsin 1 (cone pigments), short-wave-se...					304	1e-81
ref XP_001144896.1 PREDICTED: opsin 1 (cone pigments), lo...					286	2e-76
ref XP_514302.2 PREDICTED: opsin 3 [Pan troglodytes]					157	1e-37
Rattus norvegicus (brown rat, ...) [rodents] taxid 10116						
ref NP_254276.1 Rhodopsin [Rattus norvegicus]					651	0.0
ref NP_112277.1 blue cone opsin-like pigment [Rattus norv...					236	1e-79
ref NP_446006.1 opsin 1 (cone pigments), medium-wave-sens...					274	1e-72
Mus musculus (mouse) [rodents] taxid 10090						
ref NP_053335.1 rhodopsin [Mus musculus]					651	0.0
ref NP_031564.1 opsin 1 (cone pigments), short-wave-sens...					305	3e-82
ref NP_032132.1 opsin 1 (cone pigments), medium-wave-sens...					272	4e-72
ref NP_034228.1 opsin (encephalopsin) [Mus musculus]					158	1e-37

The screenshot shows the NCBI Taxonomy Browser interface. The taxonomic tree is expanded to show the following species and their associated cone opsin counts:

- Gallus gallus** (chicken): 4 cone opsins
- Rattus norvegicus** (rat): 2 cone opsins
- Mus musculus** (mouse): 2 cone opsins
- Homo sapiens** (human): 3 cone opsins
- Pan troglodytes** (chimpanzee): 3 cone opsins
- Danio rerio** (zebrafish): 4 cone opsins (multiple copies)

The NCBI logo is visible in the bottom right corner of the screenshot.

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



Automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes

1: HomoloGene:68068. Gene conserved in Verteostomi

Download, Links

Genes

Genes identified as putative homologs of one another during the construction of HomoloGene.

- H.sapiens RHO rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)
- P.troglodytes RHO rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)
- C.lupus RHO_2 rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)
- M.musculus Rho rhodopsin
- R.norvegicus Rho rhodopsin
- G.gallus RHO rhodopsin (opsin 2, rod pigment)
- D.rerio rho rhodopsin

Proteins

Proteins used in sequence comparisons and their conserved domain architectures.

- human** NP_000530.1 348 aa 
- chimp** XP_516740.2 348 aa 
- dog** XP_655608.1 358 aa 
- mouse** NP_663358.1 348 aa 
- rat** NP_254276.1 348 aa 
- chicken** NP_990821.1 355 aa 
- zebrafish** NP_571159.1 354 aa 

Expression profile suggested by analysis of EST counts.

Hs.247565: RHO: Rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)

See Legend
 Note: Please mouseover the Tissue criterion to view complete details

Breakdown by Tissue

Tissue	Count	Percentage
adipose tissue	0	0 / 12911
adrenal gland	0	0 / 33281
ascites	0	0 / 40099
bladder	0	0 / 29919
blood	0	0 / 122703
bone	0	0 / 71743
bone marrow	0	0 / 48088
brain	0	0 / 943418
canxix	0	0 / 47887
cochlea	0	0 / 16285
colon	0	0 / 182454
connective tissue	0	0 / 148193
cranial nerve	547	10 / 18286
embryonic tissue	0	0 / 158205
esophagus	0	0 / 19857
eye	2085	439 / 210537
heart	0	0 / 89096
kidney	0	0 / 211808
larynx	0	0 / 24401
liver	0	0 / 207525
lung	0	0 / 337300
lymph	0	0 / 44398
lymph node	0	0 / 91709
mammary gland	0	0 / 152721
mouth	0	0 / 67428
muscle	371	48 / 107737
nerve	0	0 / 15760
ovary	0	0 / 102196
pancreas	0	0 / 215169
parathyroid	0	0 / 20801
pharynx	0	0 / 41999
pituitary gland	0	0 / 16549
placenta	0	0 / 283807
prostate	0	0 / 190459
salivary gland	0	0 / 20249
skin	0	0 / 197275

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

Homo sapiens gene RHO, encoding rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant).

Table of Contents/open all paragraphs

BIOLOGICAL AND FUNCTIONAL ANNOTATION

In the spirit of systems biology, this chapter provides links to all genes with similar annotations.

► Diseases (New!) ↑

▼ Pathways, biological processes, molecular function and cellular localization (GO) ↑

This section summarizes the functional aspects: pathways, processes, molecular function, enzymatic activity, or localization of the protein(s) to cell compartments. Some annotations are documented in PubMed, some are inferred. The lists of related genes with the same process or function GO annotation are reported in the last column only, if they are supported by a PubMed publication.

Type	Description	Evidence	Source	Related genes (* if published support)
Process	G-protein coupled receptor protein signaling pathway	1 article	GOA/TAS	178 genes *
	phototransduction, visible light	1 article	GOA/TAS	4 genes *
	rhodopsin mediated signaling	1 article	GOA/TAS	3 genes *
	protein-chromophore linkage		GOA/EA	
	response to stimulus		GOA/EA	
	visual perception		GOA/EA	108 genes *
Function	rhodopsin-like receptor activity	Pfam	AcqView	one gene: GPR139 *
	Genes most related through pathways, process or function (with published evidence)			GRI1 RPI, PDE6B, ABCA4 SAG OPN1MW, RGR1, RRH, GNAT2, OPN1DCHNL
Localization	plasma membrane	1 article	GOA/IDA	1262 genes
	integral to plasma membrane	1 article	GOA/TAC	617 genes
	integral to membrane	Pfam	AcqView	447 genes
	photoreceptor outer segment		GOA/EA	one gene: MYO7A
	Different localizations may apply to different protein isoforms.			
	Cumulated literature	5 articles		

▼ Protein domains and motifs ↑

Protein domains or motifs

The rhodopsin-like GPCR superfamily domain is found in 2 isoforms: a, b.

[InterPro annotation] G-protein-coupled receptors, GPCRs, constitute a vast protein family that encompasses a wide range of functions (including various autocrine, paracrine and endocrine processes). They show considerable diversity at the sequence level, on the basis of which they can be separated into distinct groups. We use the term clan to describe the GPCRs, as they embrace a group of families for which there are indicators of evolutionary relationship, but between which there is no statistically significant similarity in sequence. The currently known clan members include the rhodopsin-like GPCRs, the secretin-like GPCRs, the cAMP receptors, the fungal mating pheromone receptors, and the metabotropic glutamate receptor family. There is a specialized database for GPCRs: <http://www.gpcr.org/7tm/>. The rhodopsin-like GPCRs themselves represent a widespread protein family that includes hormone, neurotransmitter and light receptors, all of which transduce extracellular signals through interaction with guanine nucleotide-binding (G) proteins. Although their activating ligands vary widely in structure and character, the amino acid sequences of the receptors are very similar and are believed to adopt a common structural framework.

721 genes: 7tm_1.1, 7tm_1.3, 7tm_1.4, 7tm_1.5, 7tm_1.6, 7tm_1.8, 7tm_1.9, 7tm_1.11, 7tm_1.12, 7tm_1.13.

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



Information about Rhodopsin Gene RHO

NCBI Single Nucleotide Polymorphism

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez [SNP] for [] Go

BUILD 127 SNP linked to Gene RHO(geneID:6010) Via Contig Annotation

Color Legend

Region	Contig position	mRNA pos	dbSNP rs#	Heterozygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
exon_1	35742723	96						start codon				1
exon_1	35742880	253	rs29233395	N.D.		Yes		nonsynonymous	G	Arg [R]	2	53
				N.D.		Yes	Yes	contig reference	C	Pro [P]	2	53
	35742895	268	rs29233394	N.D.		Yes		nonsynonymous	G	Arg [R]	2	58
				N.D.		Yes	Yes	contig reference	C	Ile [I]	2	58
exon_3	35746341	727	rs29233393	N.D.		Yes		nonsynonymous	C	Pro [P]	2	211
				N.D.		Yes	Yes	contig reference	A	His [H]	2	211
exon_4	35746711	981	rs29001653	N.D.		Yes		nonsynonymous	G	Glu [E]	1	296
				N.D.		Yes	Yes	contig reference	C	Lys [K]	1	296
exon_5	35747699	1134	rs29001637	N.D.		Yes		nonsynonymous	T	Val [V]	1	347
				N.D.		Yes	Yes	contig reference	C	Pro [P]	1	347
	35747706	1135	rs29001566	N.D.		Yes		nonsynonymous	G	Arg [R]	2	347
				N.D.		Yes		nonsynonymous	T	Leu [L]	2	347
				N.D.		Yes	Yes	contig reference	C	Pro [P]	2	347

Information about Rhodopsin Gene RHO

NCBI OMIM Online Mendelian Inheritance in Man

All Databases PubMed **RHO**

Search OMIM **ALLELIC VARIANTS** (selected examples)

Limits Preview/Index History

Display Detailed

+180380
RHO
RHODOPSIN; RHO

Alternative titles: symbols
OPSN 2; OPN2
RETINITIS PIGMENTOSA 4, I
RETINITIS PIGMENTOSA, R-

- 0001 RETINITIS PIGMENTOSA 4 [RHO, PRO23HIS]
- 0002 RETINITIS PIGMENTOSA 4 [RHO, PRO347LEU] dbSNP
- 0003 RETINITIS PIGMENTOSA 4 [RHO, PRO347SER] dbSNP
- 0004 RETINITIS PIGMENTOSA 4 [RHO, THR56ARG] dbSNP
- 0005 RETINITIS PIGMENTOSA 4 [RHO, 3-BP DEL]
- 0006 RETINITIS PIGMENTOSA 4 [RHO, THR17MET]
- 0007 RETINITIS PIGMENTOSA 4 [RHO, PHE45LEU]
- 0008 RETINITIS PIGMENTOSA 4 [RHO, VAL87ASP]
- 0009 RETINITIS PIGMENTOSA 4 [RHO, GLY89ASP]
- 0010 RETINITIS PIGMENTOSA 4 [RHO, GLY106TRP]
- 0011 RETINITIS PIGMENTOSA 4 [RHO, ARG135LEU]
- 0012 RETINITIS PIGMENTOSA 4 [RHO, ARG135TRP]
- RETINITIS PUNCTATA ALBESCENS, INCLUDED
- 0013 RETINITIS PIGMENTOSA 4 [RHO, TYR178CVS]
- 0014 RETINITIS PIGMENTOSA 4 [RHO, ASP190GLY]
- 0015 RETINITIS PIGMENTOSA 4 [RHO, GLN247TER]
- 0016 RETINITIS PIGMENTOSA 4 [RHO, LYS296GLU] dbSNP**
- 0017 RETINITIS PIGMENTOSA 4 [RHO, ASP190TYR] dbSNP
- 0018 RETINITIS PIGMENTOSA 4 [RHO, HIS211PRO] dbSNP
- 0019 RETINITIS PIGMENTOSA 4 [RHO, 12-BP DEL, EX1]
- 0020 RETINITIS PIGMENTOSA 4 [RHO, PRO347ARG] dbSNP
- 0021 RETINITIS PIGMENTOSA 4 [RHO, GLY182SER]
- 0022 RETINITIS PIGMENTOSA 4 [RHO, PRO267LEU]
- 0023 RETINITIS PIGMENTOSA 4 [RHO, GLU249TER]
- 0024 RETINITIS PIGMENTOSA 4 [RHO, PRO53ARG] dbSNP
- 0025 RETINITIS PIGMENTOSA 4 [RHO, GLY106ARG] dbSNP
- 0026 RETINITIS PIGMENTOSA, AUTOSOMAL RECESSIVE [RHO, IVS4, G-T, *1]
- 0027 RETINITIS PIGMENTOSA 4 [RHO, ASP190TYR] dbSNP
- 0028 RETINITIS PIGMENTOSA 4 [RHO, ARG207MET] dbSNP
- 0029 RETINITIS PIGMENTOSA 4 [RHO, ASN165SER]
- 0030 RETINITIS PIGMENTOSA 4 [RHO, MET207ARG] dbSNP
- 0031 NIGHT BLINDNESS, CONGENITAL STATIONARY, AUTOSOMAL DOMINANT 1 [RHO, ALA292GLU]
- 0032 NIGHT BLINDNESS, CONGENITAL STATIONARY, AUTOSOMAL DOMINANT 1 [RHO, GLY90ASP]
- 0033 RETINITIS PIGMENTOSA, AUTOSOMAL RECESSIVE [RHO, GLU150LYS]

Blink: Precalculated top 200 protein BLAST hits

Query: gi4506527 rhodopsin [Homo sapiens]
 Matching gi: 31873264, 129207, 114589117, 28798805, 117644328, 10069598, 108752084, 21928611, 85567017, 85567192, 1236137, 119599650

Show identical | All hits | Common Tree | Taxonomy Report | 3D structures | CDD-Search | G list | Run BLAST

200 BLAST hits to 154 unique species Sort by taxonomy proximity

Archaea 0 Bacteria 199 Metazoa 0 Fungi 0 Plants 0 Viruses 0 Other Eukaryotes

Keep only [] Cut-Off 100 Select Reset New search by GI 4506527 Go

SCORE	P	ACCESSION	GI	N	ORGANISM
Conserved Domain Database hits					
1821	26	XP_001...	109098032	-	<i>Macaca mulatta</i>
1816	26	Q09096	2024200	-	<i>Macaca fascicularis</i>
1803	21	NP_001...	57163783	-	<i>Felis catus</i>
1800	22	P49912	1709478	-	<i>Oryzotagus cuniculus</i>
1794	21	CAAS0502	588566	-	<i>Canis lupus familiaris</i>
1791	21	Q62796	6093621	-	<i>Trichechus manatus</i>
1786	21	Q62794	6093620	-	<i>Phoca vitulina</i>
1786	21	AAB86808	2636722	-	<i>Sus scrofa</i>
1783	22	AAH31766	21594395	-	<i>Mus musculus</i>
1792	22	CAA97091	603975	-	<i>Rattus norvegicus</i>
1791	21	AAC12765	2025842	-	<i>Phoca groenlandica</i>
1781	22	AB032113	133854420	-	<i>Cavia porcellus</i>
1777	22	P28681	125206	-	<i>Cricetulus griseus</i>
1772	23	BA024008	20524400	-	<i>Otolemus crassicaudatus</i>
1767	22	AAQ25707	10954020	-	<i>Mannosipalax ehrenbergi</i>
1760	21	P02700	129212	-	<i>Ovis aries</i>
1757	1	AAL24791	16588405	-	synthetic construct
1754	20	XP_001...	126336211	-	<i>Monodelphis domestica</i>
1752	20	Q693E1	75071958	-	<i>Caluromys philander</i>
1752	21	Q69J47	75071458	-	<i>Loxodonta africana</i>
1751	21	Q62791	6093616	-	<i>Delphinus delphis</i>
1751	21	AAF13021	30314002	-	<i>Micounga angustirostris</i>
1750	21	P02499	129204	-	<i>Bos taurus</i>
1743	21	Q62793	6093618	-	<i>Mesoplecton bidens</i>
1731	21	AAC12940	2037082	-	<i>Tasayoa truncatus</i>
1724	19	AAU90201	21448820	-	<i>Canis lupus familiaris</i>

Query: gi4506527 rhodopsin [Homo sapiens]
 Matching gi: 31873264, 129207, 114589117, 28798805, 117644328, 10069598, 108752084, 21928611, 85567017, 85567192, 1236137, 119599650

Get CaJD New!

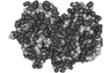
Show identical | Best hits | Common Tree | Taxonomy Report | 3D structures | CDD-Search | G list | Run BLAST

8 BLAST hits to 1 unique species Sort by taxonomy proximity

Archaea 0 Bacteria 8 Metazoa 0 Fungi 0 Plants 0 Viruses 0 Other Eukaryotes

Keep only [] Cut-Off 100 Select Reset New search by GI 4506527 Go

SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
Conserved Domain Database hits				
1750	*	1FFP1	16075387	Chain A, Structure Of Bovine Rhodopsin (Dark Adapted)
1750	*	1L9HB	21465998	Chain B, Crystal Structure Of Bovine Rhodopsin At 2.6 Angstroms Resolution
1745	*	1F88B	10121076	Chain B, Crystal Structure Of Bovine Rhodopsin
207	*	1EDXA	10120989	Chain A, Solution Structure Of Amino Terminus Of Bovine Rhodopsin (Residues
167	*	1EDSA	10120986	Chain A, Solution Structure Of Intradiskal Loop 1 Of Bovine Rhodopsin (Rho
158	*	1EDVA	10120987	Chain A, Solution Structure Of 2nd Intradiskal Loop Of Bovine Rhodopsin (Re
131	*	1EDWA	10120988	Chain A, Solution Structure Of Third Intradiskal Loop Of Bovine Rhodopsin (R
123	*	1FDFA	9955026	Chain A, Helix 7 Bovine Rhodopsin


Related Structures


[HOME](#) | [SEARCH](#) | [SITE MAP](#) | [PubMed](#) | [Blast](#) | [Entrez Structure](#) | [Help](#)

Query: rhodopsin [Homo sapiens]
[gi: 4506527]

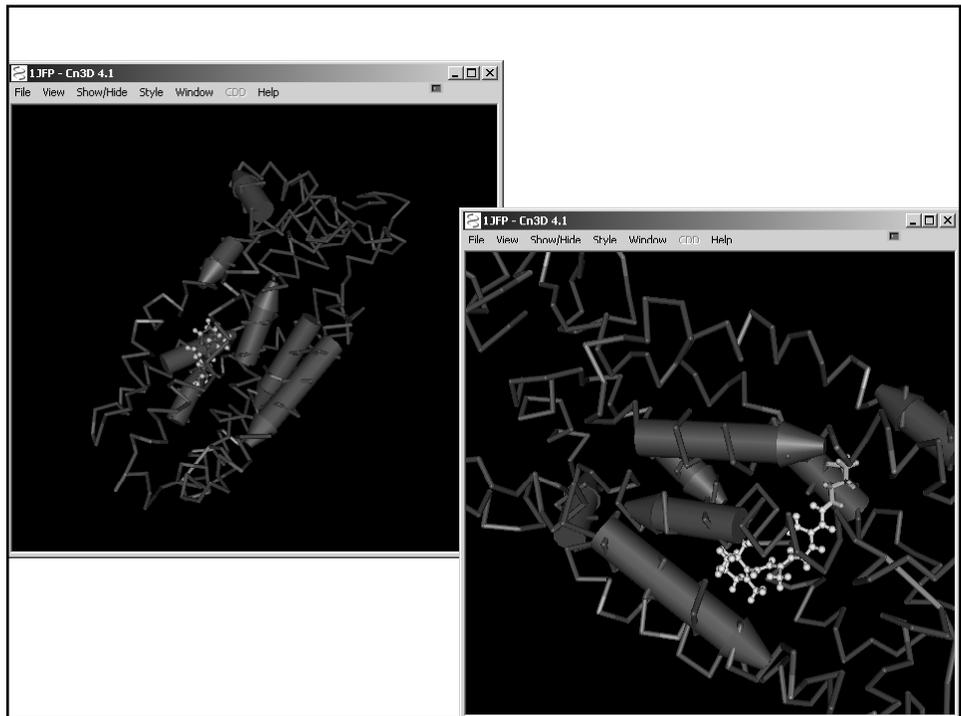
Structure: 1JFP Chain A, Structure of Bovine Rhodopsin (Dark Adapted)

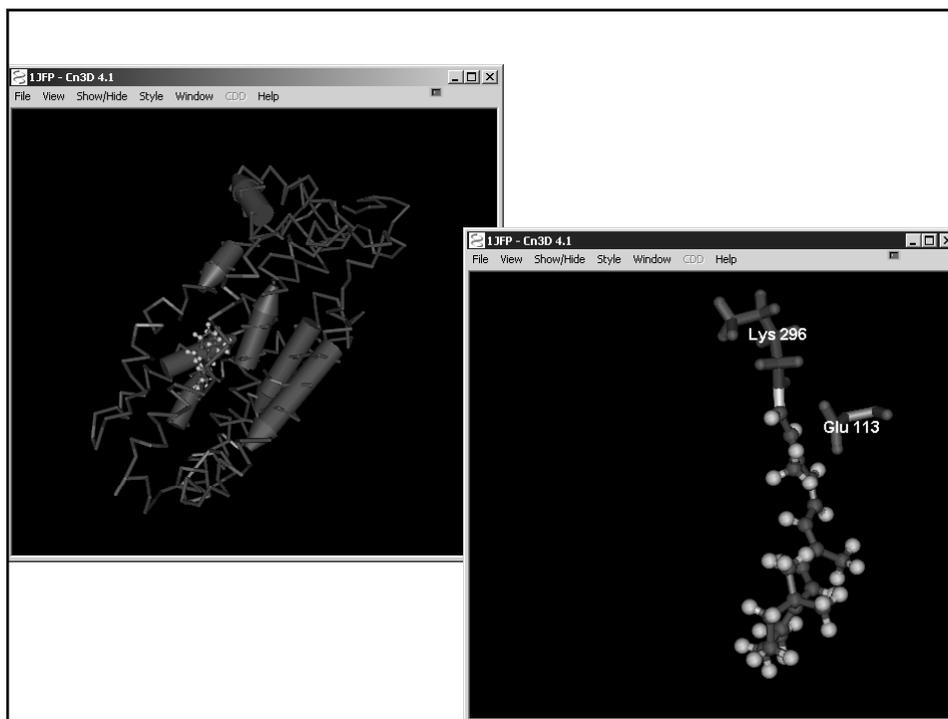
Reference: [MMDB] [PubMed]

to: (To display structure, download Cn3D)

E-value = 0.0, Bit score = 678, Aligned length = 348, Sequence Identity = 93%

	10	20	30	40	50	60	70	80
gi 4506527	1	MNGTEGPNFYVFFSNATGVVRSFF	YFPQYLLAEPQFSLAAYMFL	LLVGLGFFINFLTYVTVQHKLR	PLNLYILLNLA	80		
1JFP_A	1	MNGTEGPNFYVFFSNKTVVRSFF	EAPQYLLAEPQFSLAAYMFL	LLMLGFFINFLTYVTVQHKLR	PLNLYILLNLA	80		
.....								
	90	100	110	120	130	140	150	160
gi 4506527	81	VADLFVLOGFTSLYTDLHGFFV	FGPTGCHLEGFATLGGELAG	WGLVLAIERVVVCRPMSNFR	FGENHAIKGVAF	160		
1JFP_A	81	VADLFVFGGFITLLYTSLHGFF	VFGPTGCHLEGFATLGGELAG	WGLVLAIERVVVCRPMSNFR	FGENHAIKGVAF	160		
.....								
	170	180	190	200	210	220	230	240
gi 4506527	161	WVMALACAAPPLAGWSRYIPE	GLQCSGIDYTLKPEVNNE	SFVIYHFVVHFTIPMII	IFFCYGQLVFTVKEAA	QQQES	240	
1JFP_A	161	WVMALACAAPPLVGSRYIPE	GHQCSGIDYTPHEETNNE	SFVIYHFVVHFTIPLI	LVIFFCYGQLVFTVKE	AAQQQES	240	
.....								
	250	260	270	280	290	300	310	320
gi 4506527	241	ATTQKAEREVTRMVIINVI	AFILICVVPPLASVAFYI	FTHQGSDFGPIFHTIP	AFFAKSAATINPVYI	IIMNKQFENCHLIT	320	
1JFP_A	241	ATTQKAEREVTRMVIINVI	AFILICVLPYAGVAFYI	FTHQGSDFGPIFHTIP	AFFAKSAATINPVYI	IIMNKQFENCHVIT	320	
.....								
	330	340						
gi 4506527	321	ICCGKNPLGDDEASATVSK	TETSQVAPA	348				
1JFP_A	321	LCCGKNPLGDDEASTVSK	TETSQVAPA	348				





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

Databases	Interpretation	Tools
Entrez global search Genomes Books Gene Nucleotide RefSeq Protein Taxonomy Homologene UniGene AceView dbSNP OMIM Structure	Photoreceptors cones and rods Sequence similarity Phylogeny Homology Expression Structure Function	BLAST Blastp Blast2 sequences Distance Tree Tax Blast MapViewer UniGene DDD Cn3D Taxonomy Common Tree Blink Related Structures



Questions about NCBI Resources?

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

