Bioinformatics Quick Start

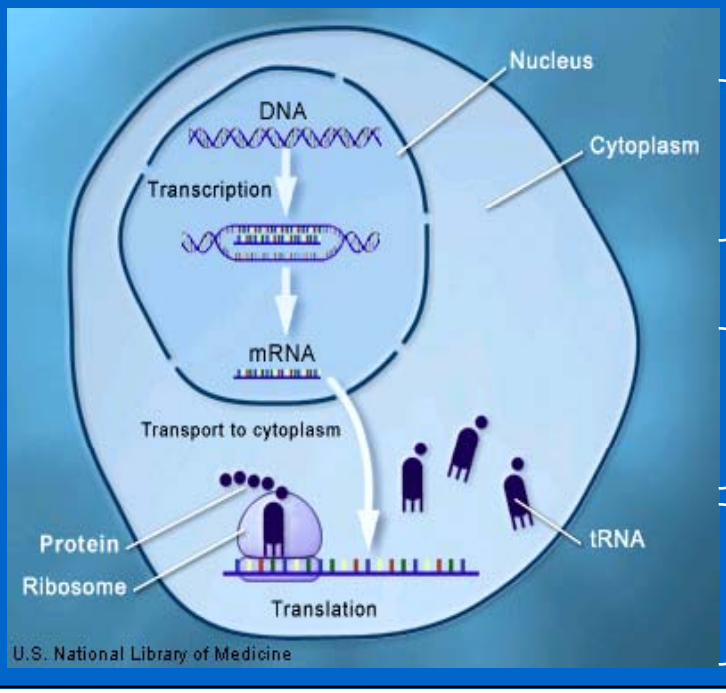
Medha Bhagwat
National Center for Biotechnology Information
National Institutes of Health

NCBI

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

NCBI



Information storage

Information carrier

Protein - active machinery

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

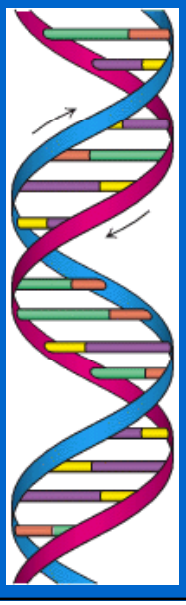
```
..AGCTGCATGCTAGCTGACGTCA....
  | | | | | | | | | | | | | | | | | | | | | |
..TCGACGTACGATCGACTGCAGT....
```

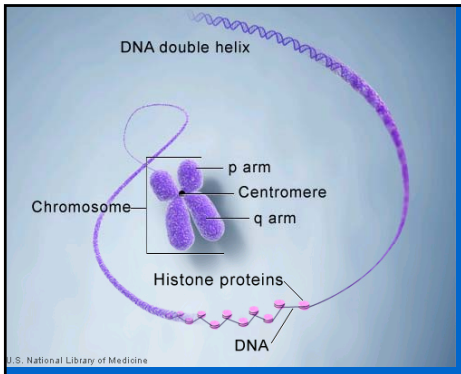
"Words" (genes) to encode proteins, RNA etc.

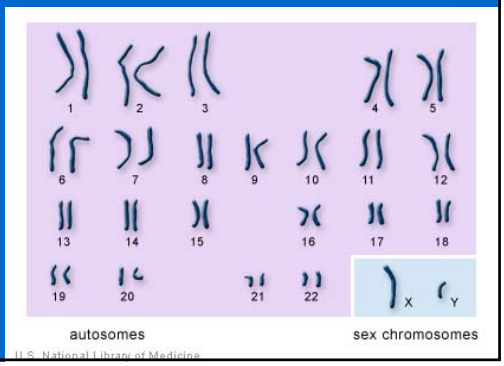
Double helical

The double-helical structure of DNA

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







Protein

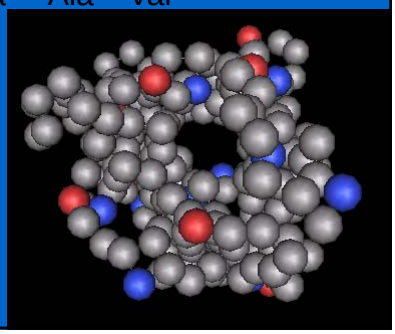
Alphabet: amino acids

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

ATG TGCAGCCTAGCTGCCGTC

Met—Cys—Ser—Leu —Ala — Ala —Val

Water channel protein

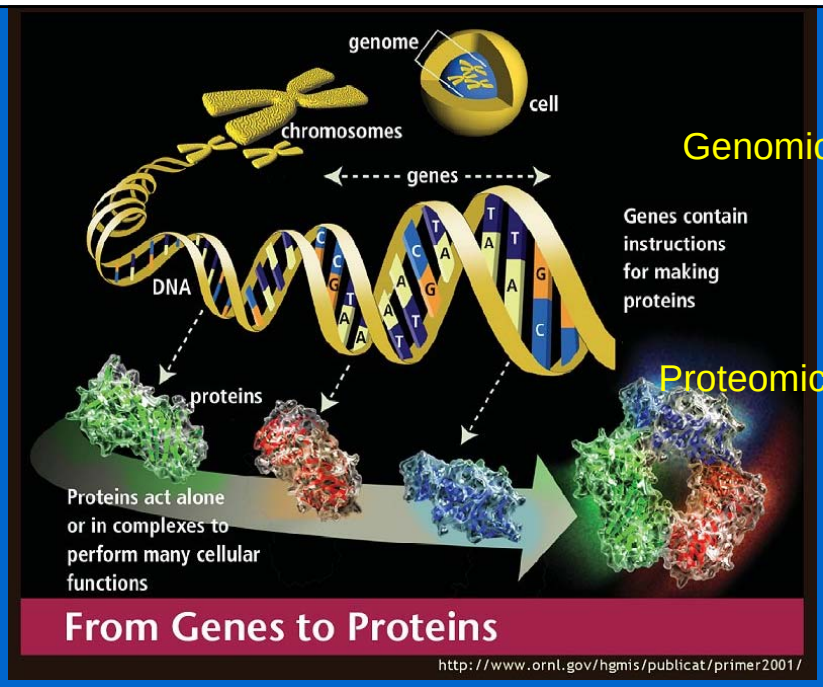


Genome (DNA)

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.



Genomics

Proteomics

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

NCBI

Additional Information

Genetics Home Reference

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

A Science Primer

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

Bookshelf

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.

NCBI

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

NCBI

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

NCBI Databases

Primary Databases

Genomes (DNA)
↓
mRNA
↓
Protein

Organisms

Expression

Structures

Small compounds

Conserved Domains

Publications
Books

Derived Databases

Gene

Homologene

Unigene

RefSeq

NCBI

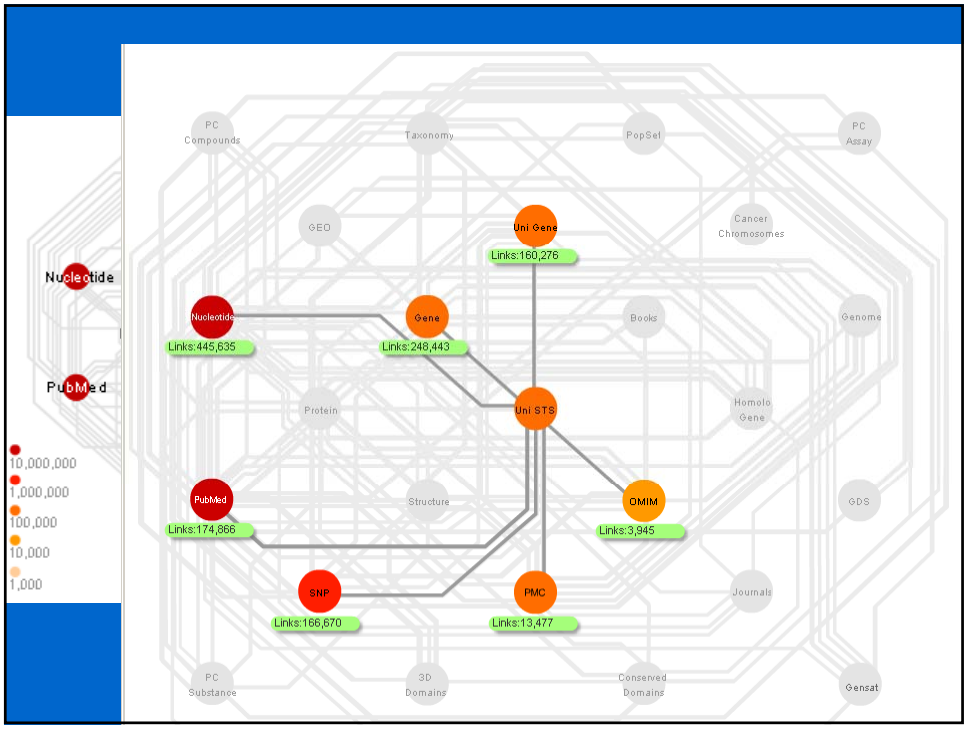
NCBI Databases

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

NCBI

NCBI Databases

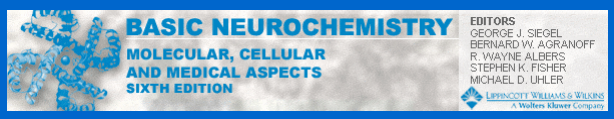
http://www.ncbi.nlm.nih.gov/sites/gquery

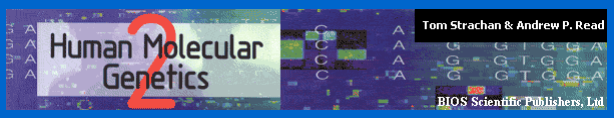


NCBI Databases

Bookshelf

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

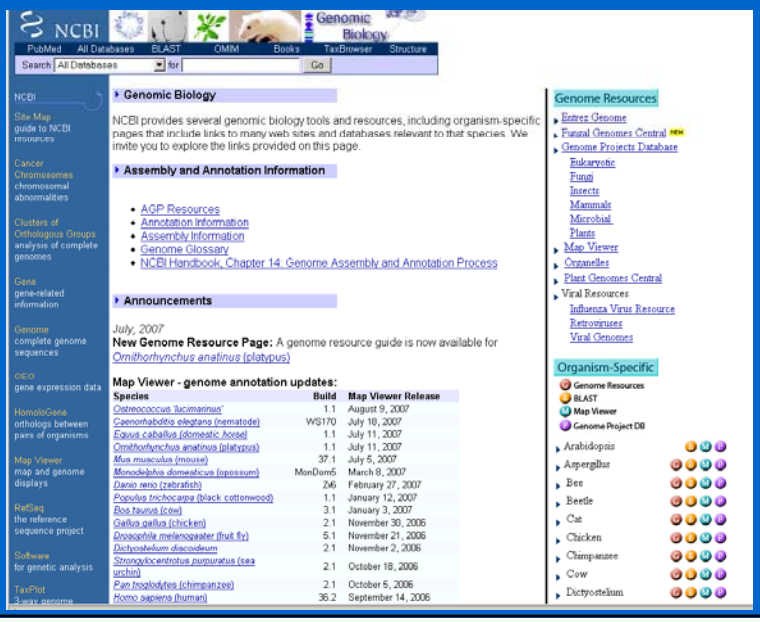




NCBI

Genome Sequence Data and Analysis Tools at NCBI

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



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 (*E. coli*)	4.6 million
Human immunodeficiency virus	9700

NCBI

Tools for Data Mining

BLAST | OMIM | Books | TaxBrowser | Structure

for Go

Protein Sequence Analysis | Structures | Genome Analysis | Gene Expression

Tools - Nucleotide Sequence Analysis

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.

electronic PCR 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 - 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 - 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.

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.

TaxPlot - a tool for 3-way comparisons of genomes on the basis of the protein sequences they encode. To use TaxPlot, 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.

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

Tools - Structures

Cn3D - Cn3D is a helper application for your web browser that allows you to view 3-dimensional structures from NCBI's Entrez retrieval service. Cn3D runs on Windows, Macintosh, and Unix.

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

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

Tools - Genome Analysis

Entrez Genomes - Whole genomes of over 1000 organisms. The 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, phages, 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 corresponds to an ancient conserved domain.

Map Viewer - shows integrated views of chromosome maps for many organisms, including human and numerous other vertebrates, 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/M-FISH & CGH Database - The NCI and NCBI SKY/M-FISH and CGH 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/M-FISH permits the simultaneous visualization of each human or mouse chromosome in a different color, facilitating the identification of chromosomal aberrations; CGH 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...)

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 eUtils 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. Interpreation/global analysis of data
   - Photoreceptors

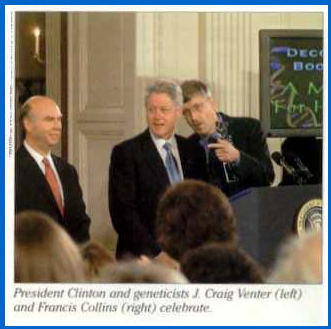
NCBI

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”

NCBI

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.

NCBI

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

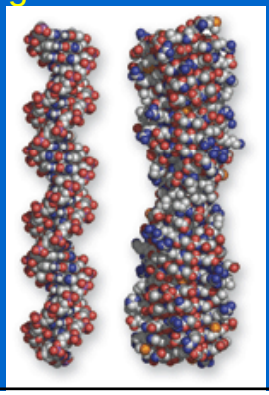
Understand Bacterial Resistance

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

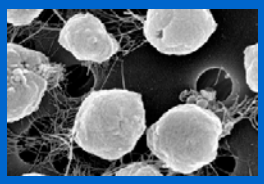
DNA



mfpA protein

Science (2005) 308, 1393

Methanocaldococcus jannaschii



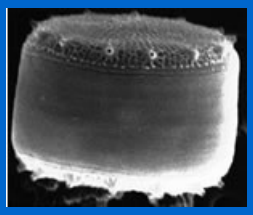
Methane-producing thermophilic archeon

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

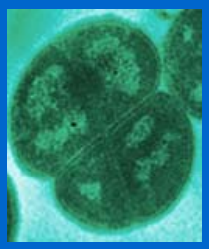


Ocean diatom, a major participant in biological pumping of carbon to ocean depths

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

NCBI

What's Next?????

Nature, Volume 449 Number 7164
18 October 2007

NCBI

What's Next?????

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

COMT genotyping could be applied to identify likely responders to burpoprion 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

NCBI

More Information

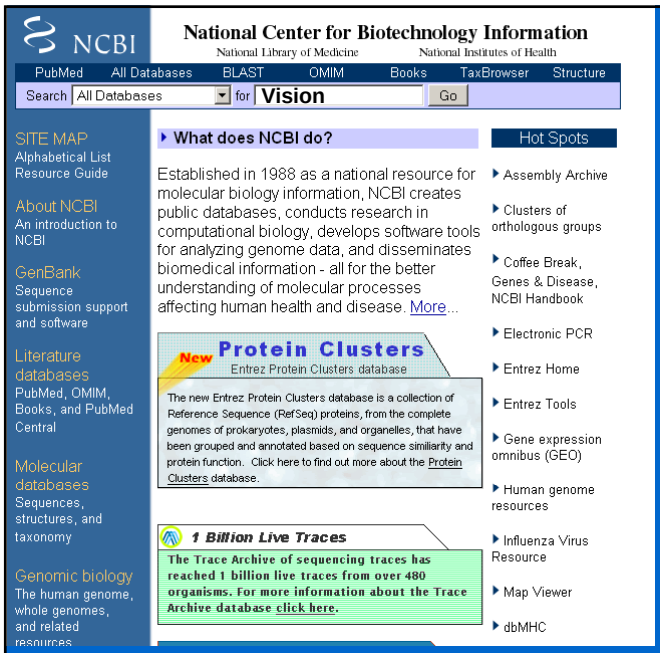
Volume 291, Issue 5507, Pages 1145-1434

Volume 409, Issue 6822, Pages 745-964

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

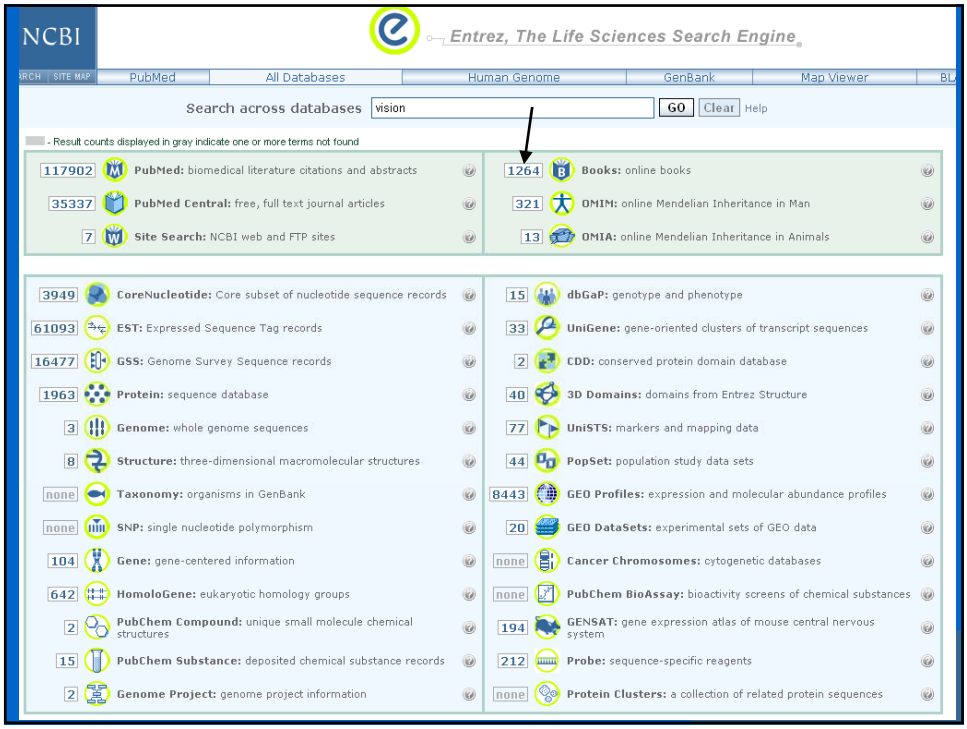
Human Genome Project Information
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

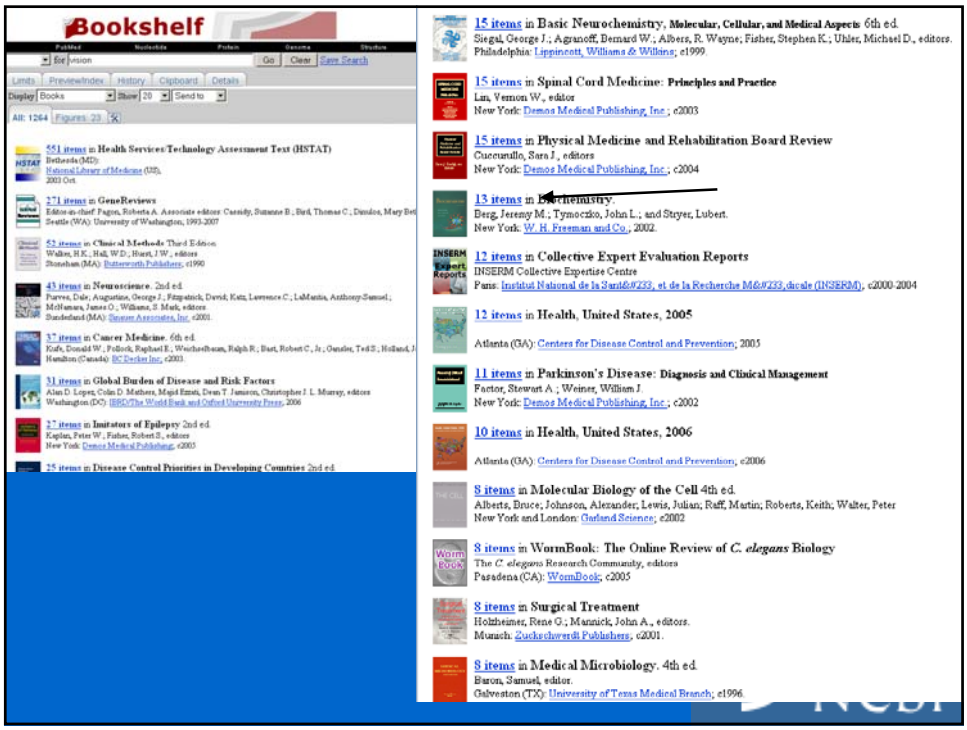
NHGRI Fact sheets
http://www.genome.gov/10000202



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

NCBI


NCBI
Entrez, The Life Sciences Search Engine
PubMed
All Databases
Human Genome
GenBank
Map Viewer
Search across databases
vision
GO
Clear
Help
- Result counts displayed in gray indicate one or more terms not found
117902 PubMed: biomedical literature citations and abstracts
35337 PubMed Central: free, full text journal articles
7 Site Search: NCBI web and FTP sites
1264 Books: online books
321 OMIM: online Mendelian Inheritance in Man
13 OMIA: online Mendelian Inheritance in Animals
3949 CoreNucleotide: Core subset of nucleotide sequence records
61093 EST: Expressed Sequence Tag records
16477 GSS: Genome Survey Sequence records
1963 Protein: sequence database
3 Genome: whole genome sequences
8 Structure: three-dimensional macromolecular structures
none Taxonomy: organisms in GenBank
none SNP: single nucleotide polymorphism
104 Gene: gene-centered information
642 HomoloGene: eukaryotic homology groups
2 PubChem Compound: unique small molecule chemical structures
15 PubChem Substance: deposited chemical substance records
2 Genome Project: genome project information
15 dbGaP: genotype and phenotype
33 UniGene: gene-oriented clusters of transcript sequences
2 CDD: conserved protein domain database
40 3D Domains: domains from Entrez Structure
77 UniSTS: markers and mapping data
44 PopSet: population study data sets
8443 GEO Profiles: expression and molecular abundance profiles
20 GEO DataSets: experimental sets of GEO data
none Cancer Chromosomes: cytogenetic databases
none PubChem BioAssay: bioactivity screens of chemical substances
194 GENSAT: gene expression atlas of mouse central nervous system
212 Probe: sequence-specific reagents
none Protein Clusters: a collection of related protein sequences


Bookshelf
15 items in Basic Neurochemistry, Molecular, Cellular, and Medical Aspects 6th ed.
Siegal, George J.; Agranoff, Bernard W.; Albers, R. Wayne; Fisher, Stephen K.; Uhler, Michael D., editors.
Philadelphia: Lippincott, Williams & Wilkins; c1999.
15 items in Spinal Cord Medicine: Principles and Practice
Lin, Vernon W., editor
New York: Demos Medical Publishing, Inc.; c2003
15 items in Physical Medicine and Rehabilitation Board Review
Cuccurullo, Sara J., editors
New York: Demos Medical Publishing, Inc.; c2004
13 items in Biochemistry.
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 Sant&#233, et de la Recherche M&#233,dicale (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, Samuel, 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 PreviewIndex History Clipboard Details

Limits: Current Records

Display Summary Show 20

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

Items 1 - 8 of 8

1: OPN1SW — Blue
Official Symbol OPN1SW and Name: opsin 1 (cone pigm
Other Aliases: BCP, BOP, CBT
Other Designations: blue cone photoreceptor pigment, blu
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: ECPN
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: OPN1MW2
Official Symbol OPN1MW2 and Name: opsin 1 (cone pigments), medium-wave-sensitive 2 [*Homo sapiens*]
Chromosome: X; Location: Xq28
Annotation: Chromosome X, NC_000023.9 (153178461..153151862)
GeneID: 728458

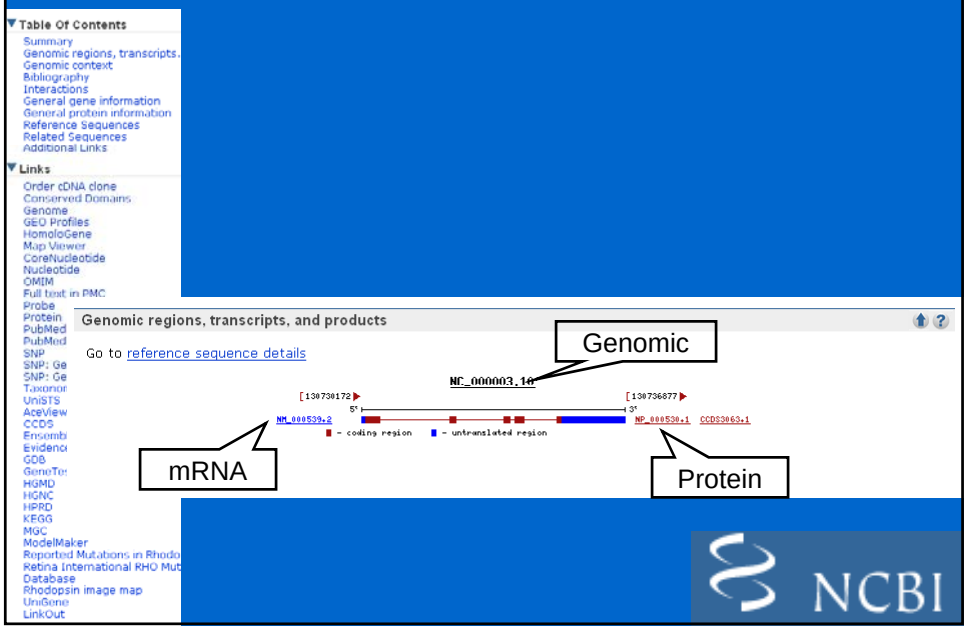
5: RHO — Rhodopsin
Official Symbol RHO and Name: rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant) [*Homo sapiens*]
Other Aliases: MGC138309, MGC138311, OPN2, RP4
Other Designations: rhodopsin
Chromosome: 3; Location: 3q21-q24
Annotation: Chromosome 3, NC_000003.10 (130730172..130736877)
MIM: 180380
GeneID: 6010

6: OPN1LW — Red
Official Symbol OPN1LW and Name: opsin 1 (cone pigments), long-wave-sensitive (color blindness, protan) [*Homo sapiens*]
Other Aliases: CBBM, 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; neuropsin; transm
Chromosome: 6; Location: 6p12.3
Annotation: Chromosome 6, NC_000006.10 (47857757..47902076)
MIM: 609042
GeneID: 221391

8: OPN1MW — Green
Official Symbol OPN1MW and Name: opsin 1 (cone pigments), medium-wave-sensitive (color blindness, deutan) [*Homo sapie*
Other Aliases: CBBM, CBD, GCP, OPN1MW1
Other Designations: green cone pigment; photopigment apoprotein
Chromosome: X; Location: Xq28
Annotation: Chromosome X, NC_000023.9 (153101361..153114725)
MIM: 303800
GeneID: 2652

Information about Rhodopsin Gene RHO



Human Vision
Absorption of light by photoreceptor cells in eye

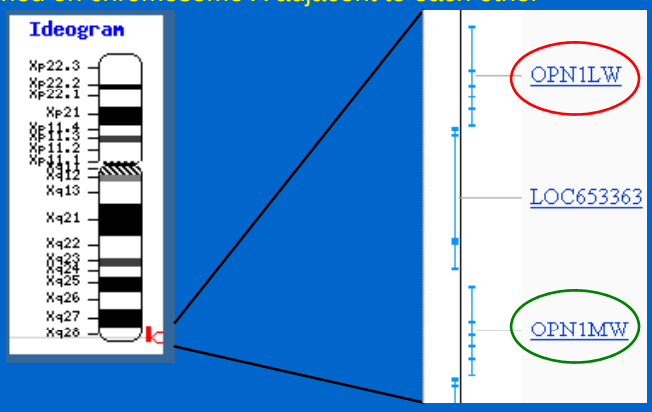
	Rods	Cones		
	Noncolor vision in dim light	color vision in bright light		
	pigment opsin (chromophore retinal)			
	rhodopsin	blue	green	red
Absorbs at	500 nm	426nm	530nm	560nm
gene	RHO	OPN1SW	OPN1MW	OPN1LW
Chromosome	3	7	X	X
Protein length	348	348	362	362
	Rest 40% identity		95% identity	

```
Red     MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM  60
Green   MAQQWSLQRLAGRHPQDSYEDSTQSSIFTYTNSNSTRGPFEGPNYHIAPRWVYHLTSVWM  60

Red     IFVVTASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISIVNQVSGYFV  120
Green   IFVVIASVFTNGLVLAATMKFKKLRHPLNWILVNLAVADLAETVIASTISVVNQVYGYFV  120

Red     LGHPMCVLEGYTVSLCGITGLWSLAIISWERWLVVCKPFGNVRFDAKLAIVGIAFSWIVS  180
Green   LGHPMCVLEGYTVSLCGITGLWSLAIISWERWMVVCKPFGNVRFDAKLAIVGIAFSWIVA  180

Red     AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCIIPLAIIMLCYL  240
Green   AVWTAPPIFGWSRYWPHGLKTSCGPDVFSGSSYPGVQSYMIVLMVTCCITPLSIIVLCYL  240

Red     QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMIFAYCVCWGPYTFFACFAAANPGYAFH  300
Green   QVWLAIRAVAKQQKESESTQKAEKEVTRMVVVMVLAFCFCWGPYAFFACFAAANPGYPFH  300

Red     PLMAALPAYFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS  360
Green   PLMAALPAFFAKSATIYNPVIYVFMNRQFRNCILQLFGKKVDDGSELSSASKTEVSSVSS  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

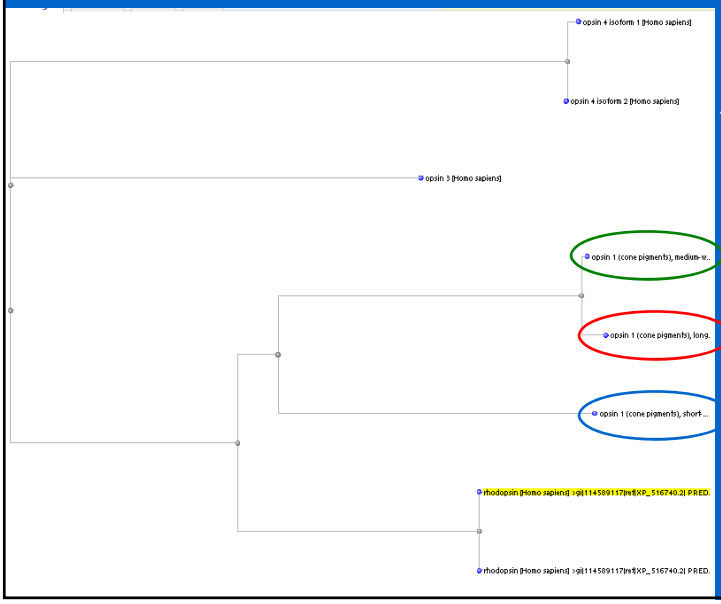
High level of identity between them
Positioned on chromosome X adjacent to each other



- 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

NCBI

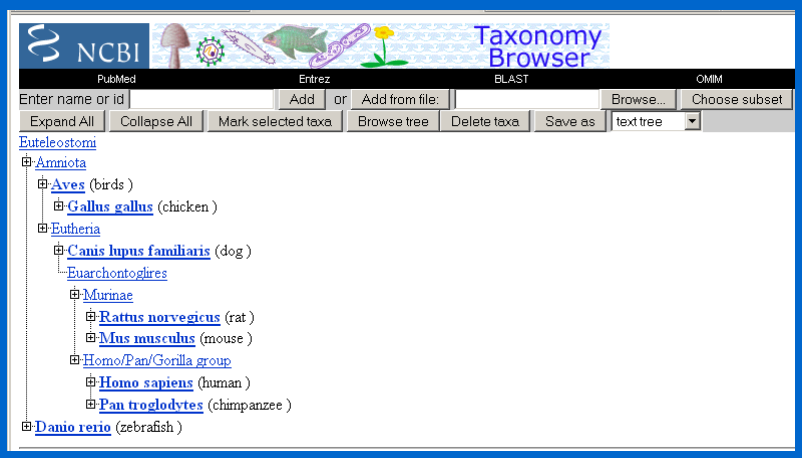
Distance Tree Generated from the BLAST Results of Rhodopsin Protein against Human RefSeq Proteins



Humans have three cone pigments red, green and blue.

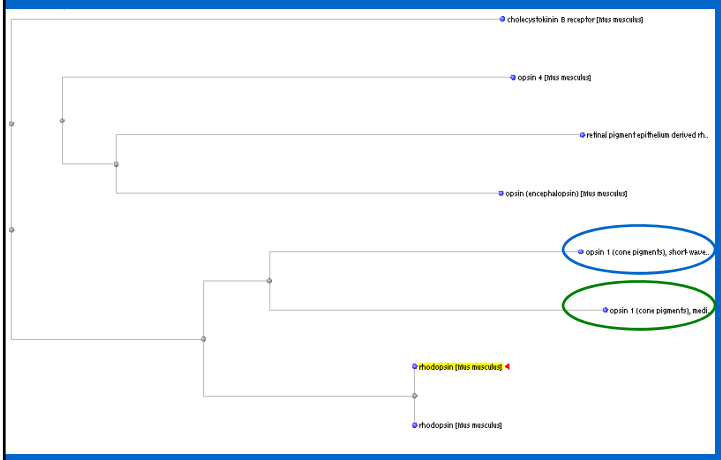
Common Lineage Tree Generated from the Taxonomy Browser

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



NCBI

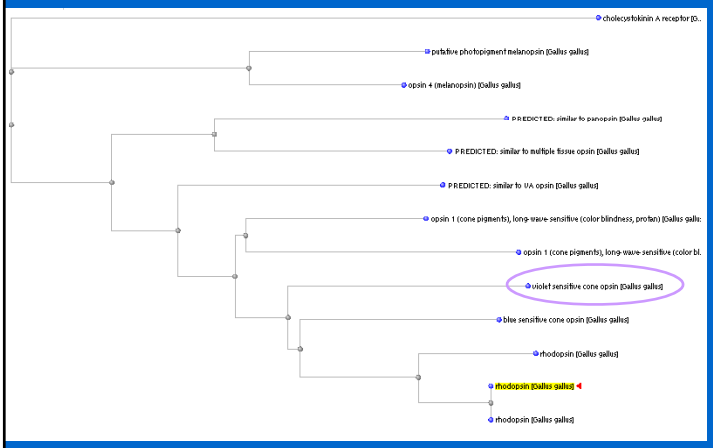
Distance tree Generated from the BLAST Results of Mouse Rhodopsin Protein against Mouse RefSeq Proteins



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.

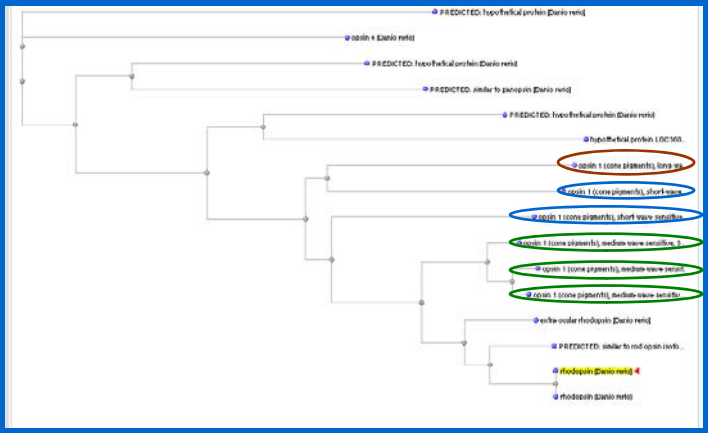
Distance Tree Generated from the BLAST Results of Chicken Rhodopsin Protein against Chicken RefSeq Proteins



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.

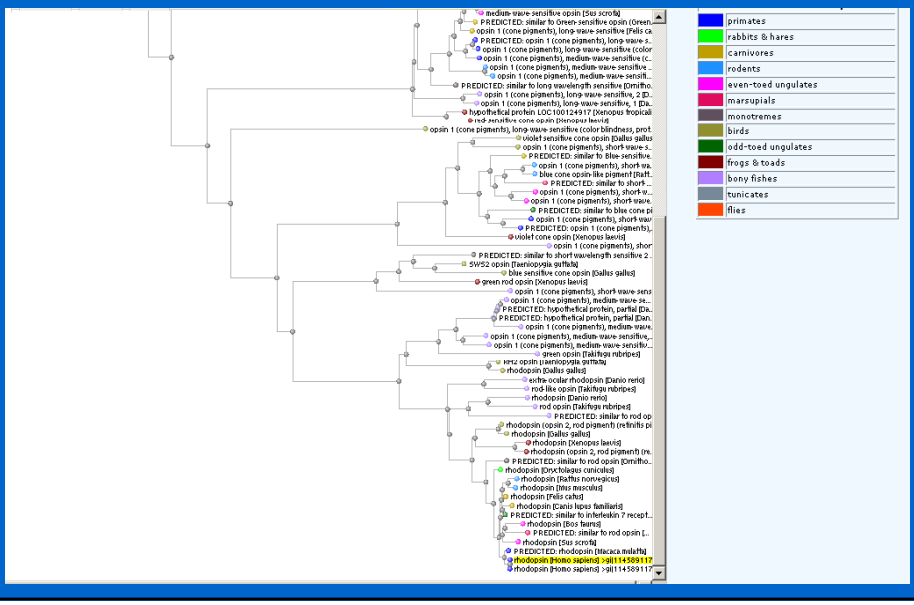
Distance Tree Generated from the BLAST Results of Zebrafish Rhodopsin Protein against Zebrafish RefSeq Proteins



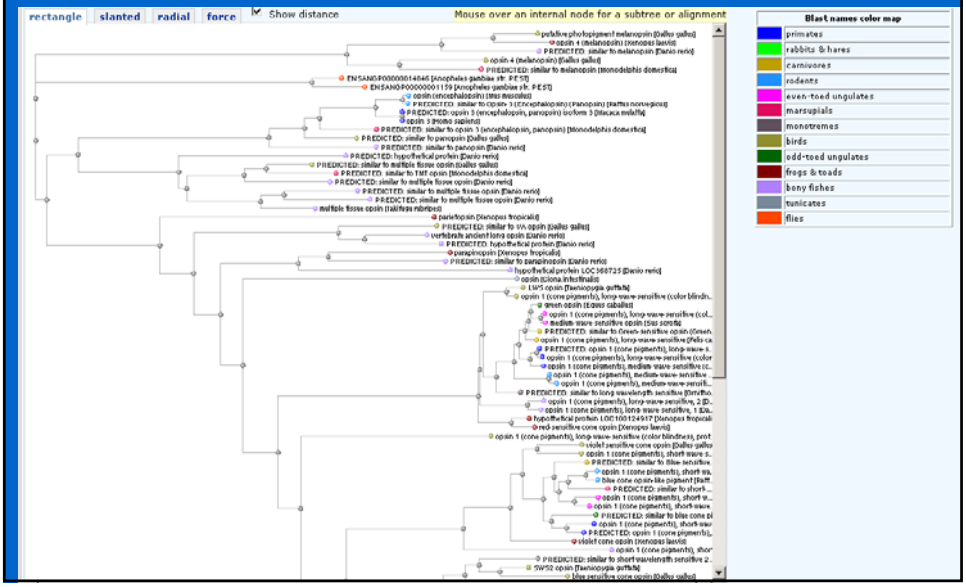
Fish, for example, Zebrafish have multiple copies of 4 pigments.

Fishes have special receptor requirement because of the variation in the amount of light in water.

Blast Distance Tree of Animal Proteins Similar to Human Rhodopsin



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

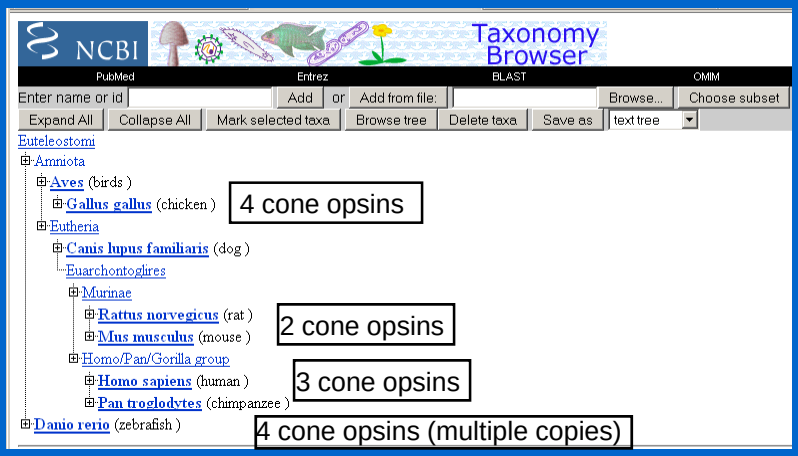


Tax BLAST Report of the Previous BLAST Search

Lineage Report

```
Coelomata        [animals]
. Chordata         [chordates]
. . Euteleostomi     [vertebrates]
. . . Tetrapoda        [vertebrates]
. . . . Amniota          [vertebrates]
. . . . . Mammalia         [mammals]
. . . . . . Theria           [mammals]
. . . . . . . Eutheria         [placentals]
. . . . . . . . Euarchontoglires [placentals]
. . . . . . . . . Catarrhini       [primates]
. . . . . . . . . . Macaca mulatta (rhesus macaque) ---------  663  3 hits [primates]
. . . . . . . . . . Homo sapiens (man) ......................  652  6 hits [primates]
. . . . . . . . . . Pan troglodytes .........................  652  4 hits [primates]
. . . . . . . . . . Oryctolagus cuniculus (domestic rabbit) ---  660  1 hit  [rabbits & hares]
. . . . . . . . . . Rattus norvegicus (brown rat) ............  651  4 hits [rodents]
. . . . . . . . . . Mus musculus (mouse) ....................  651  4 hits [rodents]
. . . . . . . . . . Felis catus (cat) -------------------------  660  3 hits [carnivores]
. . . . . . . . . . Bos taurus (cow) .........................  641  4 hits [even-toed ungulates]
. . . . . . . . . . Sus scrofa (wild boar) ...................  638  3 hits [even-toed ungulates]
. . . . . . . . . . Canis lupus familiaris (dogs) ............  629  5 hits [carnivores]
. . . . . . . . . . Equus caballus (equine) ..................  586  3 hits [odd-toed ungulates]
. . . . . . . . . Monodelphis domestica -------------------------  640  5 hits [marsupials]
. . . . . . . . . Ornithorhynchus anatinus (duck-billed platypus) -  629  3 hits [monotremes]
. . . . . . . . . Taeniopygia guttata (zebra finch) -------------  597  5 hits [birds]
. . . . . . . . . Gallus gallus (bantam) ...........
. . . . . Xenopus laevis (clawed frog) ------
. . . . . Xenopus tropicalis ...............
. . . Danio rerio (leopard danio) ---------
. . . Takifugu rubripes (torafugu) ........
. . Ciona intestinalis ----------------------
. Anopheles gambiae str. PEST -------------
```

```
Homo sapiens (man) [primates] taxid 9606
ref|NP_000530.1| rhodopsin [Homo sapiens]                      652   0.0
ref|NP_001699.1| opsin 1 (cone pigments), short-wave-sensi...  304   1e-81
ref|NP_000504.1| opsin 1 (cone pigments), medium-wave-sens...  278   6e-74
ref|NP_001041646.1| 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|NP_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...  296   2e-79
ref|NP_446000.1| opsin 1 (cone pigments), medium-wave-sens...  274   1e-72

Mus musculus (mouse) [rodents] taxid 10090
ref|NP_663358.1| rhodopsin [Mus musculus]                      651   0.0
ref|NP_031564.1| opsin 1 (cone pigments), short-wave-sensi...  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
```



NCBI

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 form the nocturnal life.

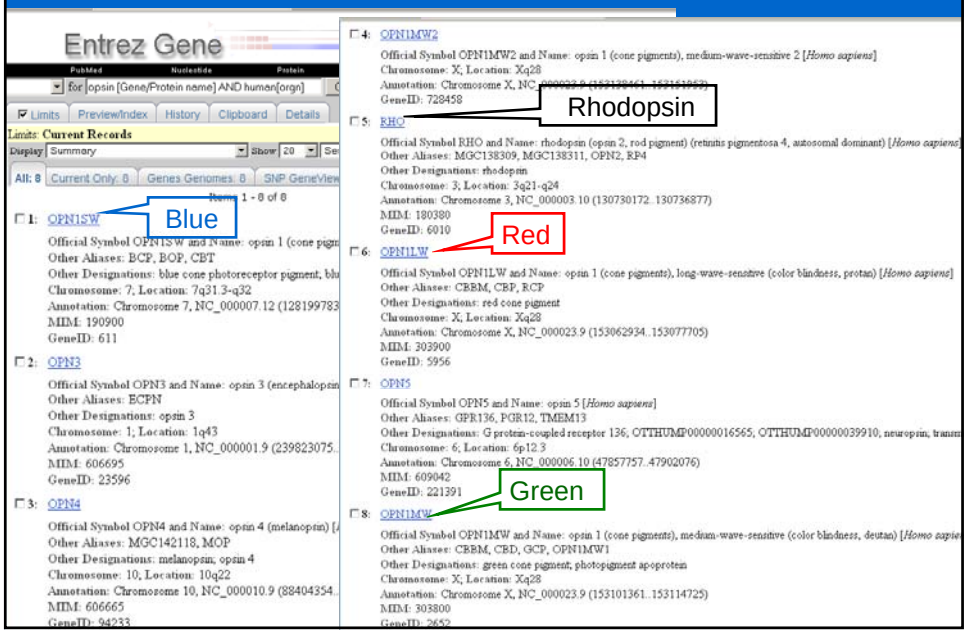
In contrast, birds such as chickens have a total of six pigments: rhodopsin, four cone pigments, and a pineal visual pigment called *pinopsin*. Birds have highly acute color perception.

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

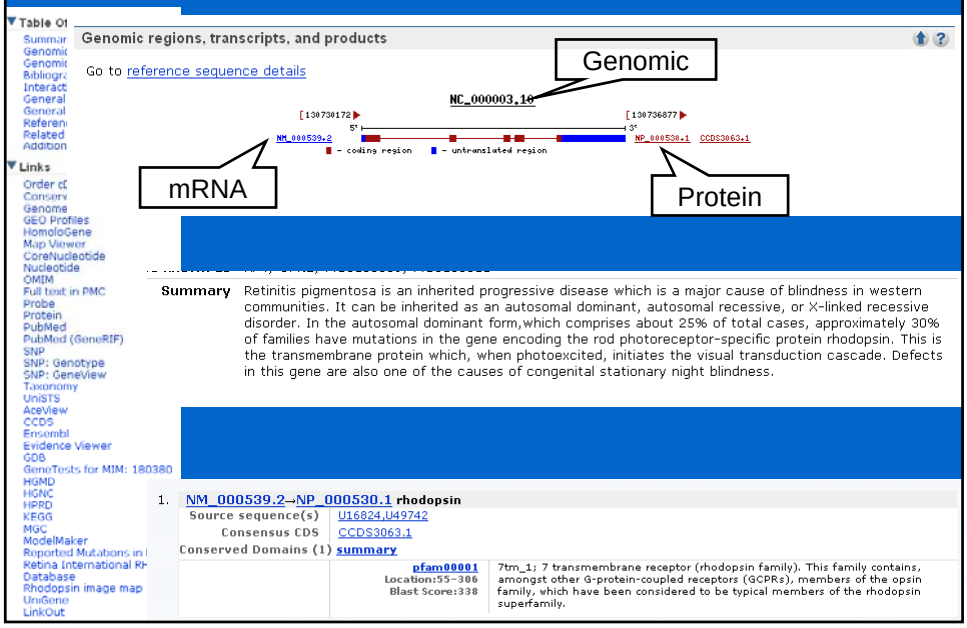
Review articles:
Bowmaker and Hunt Current Biology vol16, R484
Hunt et al. CMLS, Cell.Mol.Lifr Sci. 58, 1583

NCBI

Obtain More Information about Opsin Genes and Proteins



Information about Rhodopsin Gene RHO



HomoloGene

Discover Homologs

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

1: HomoloGene:68068. Gene conserved in Euteleostomi Download, Links

Genes		Proteins
Genes identified as putative homologs of one another during the construction of HomoloGene.		*Proteins used in sequence comparisons and their conserved domain architectures.*
H.sapiens RHO rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)	human	NP_000530.1 348 aa
P.troglodytes RHO rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)	chimp	XP_516740.2 348 aa
C.lupus RHO_2 rhodopsin (opsin 2, rod pigment) (retinitis pigmentosa 4, autosomal dominant)	dog	XP_855608.1 358 aa
M.musculus Rho rhodopsin	mouse	NP_663358.1 348 aa
R.norvegicus Rho rhodopsin	rat	NP_254276.1 348 aa
G.gallus RHO rhodopsin (opsin 2, rod pigment)	chicken	NP_990821.1 355 aa
D.rerio rho rhodopsin	zebrafish	NP_571159.1 354 aa

UniGene

ORGANIZED VIEW OF THE TRANSCRIPTOME

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

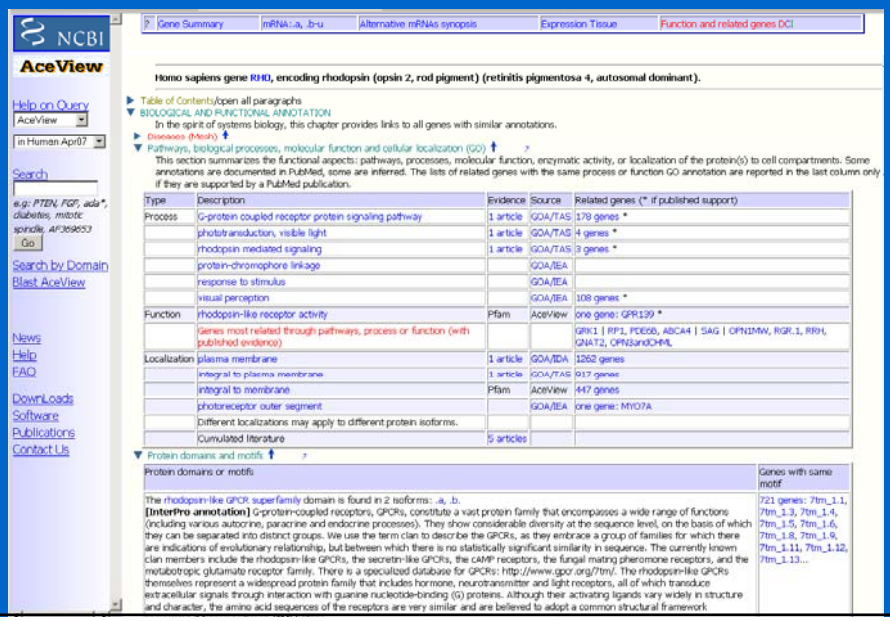
	Hs.247565	
adipose tissue	0	0 / 12911
adrenal gland	0	0 / 33281
ascites	0	0 / 40059
bladder	0	0 / 29919
blood	0	0 / 122703
bone	0	0 / 71743
bone marrow	0	0 / 48988
brain	0	0 / 943418
cervix	0	0 / 47897
cochlea	0	0 / 16285
colon	0	0 / 193454
connective tissue	0	0 / 149193
cranial nerve	547	10 / 18285
embryonic tissue	0	0 / 198205
esophagus	0	0 / 19857
eye	2085	439 / 210537
heart	0	0 / 89696
kidney	0	0 / 211908
larynx	0	0 / 24401
liver	0	0 / 207525
lung	0	0 / 337300
lymph	0	0 / 44399
lymph node	0	0 / 91709
mammary gland	0	0 / 153721
mouth	0	0 / 67429
muscle	371	40 / 107737
nerve	0	0 / 15760
ovary	0	0 / 102196
pancreas	0	0 / 215169
parathyroid	0	0 / 20601
pharynx	0	0 / 41999
pituitary gland	0	0 / 16649
placenta	0	0 / 283807
prostate	0	0 / 190459
salivary gland	0	0 / 20249
skin	0	0 / 197225

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

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

NCBI

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

Have a question

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

in gene region cSNP has frequency double hit refresh

gene model (contig mRNA transcript):	Contig Label	Contig	mrna	protein	mrna orientation	transcript	snp count
	reference	NT_005612.15	NM_000539.2	NP_000530.1	forward	plus strand	6, coding

Color Legend

Region	Contig position	mRNA pos	dbSNP rs# cluster id	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
exon_1	35742723	96						start codon				1
exon_1	35742880	253	rs28933395	N.D.		Yes		nonsynonymous	G	Arg [R]	2	53
				N.D.		Yes	Yes	contig reference	C	Pro [P]	2	53
	35742895	268	rs28933394	N.D.		Yes		nonsynonymous	G	Arg [R]	2	58
				N.D.		Yes	Yes	contig reference	C	Thr [T]	2	58
exon_3	35746341	727	rs28933393	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	A	Lys [K]	1	296
exon_5	35747699	1134	rs29001637	N.D.		Yes		nonsynonymous	T	Ser [S]	1	347
				N.D.		Yes	Yes	contig reference	C	Pro [P]	1	347
	35747700	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 Inheritan Johns Hopkins

All Databases PubMed

Search OMIM

Limits Preview/Index Histo

Display Detailed

MIM +180380
Cloning
Gene Function
Mapping
Biochemical Features
Molecular Genetics
Genotype/Phenotype
Correlations
Clinical Management
Animal Model
History
Allelic Variants
View List
See Also
References
Contributors
Creation Date

+180380
RHODOPSIN; RHO

Alternative titles; symbols

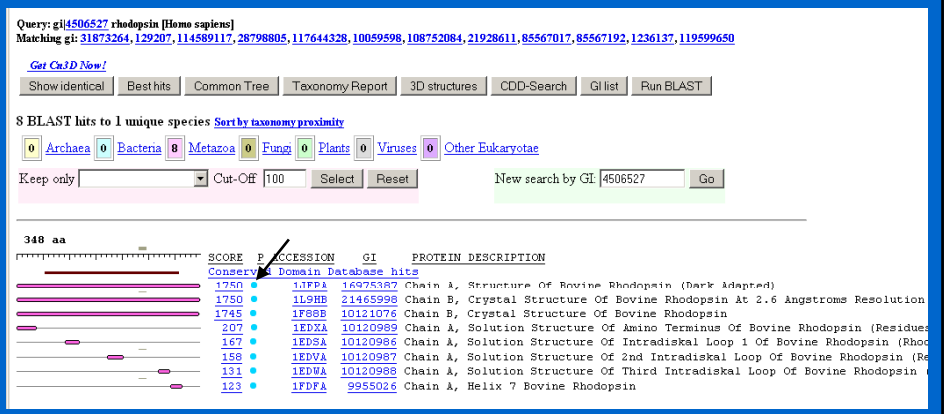
OPSIN 2; OPN2
RETINITIS PIGMENTOSA 4, I
RETINITIS PIGMENTOSA, RH

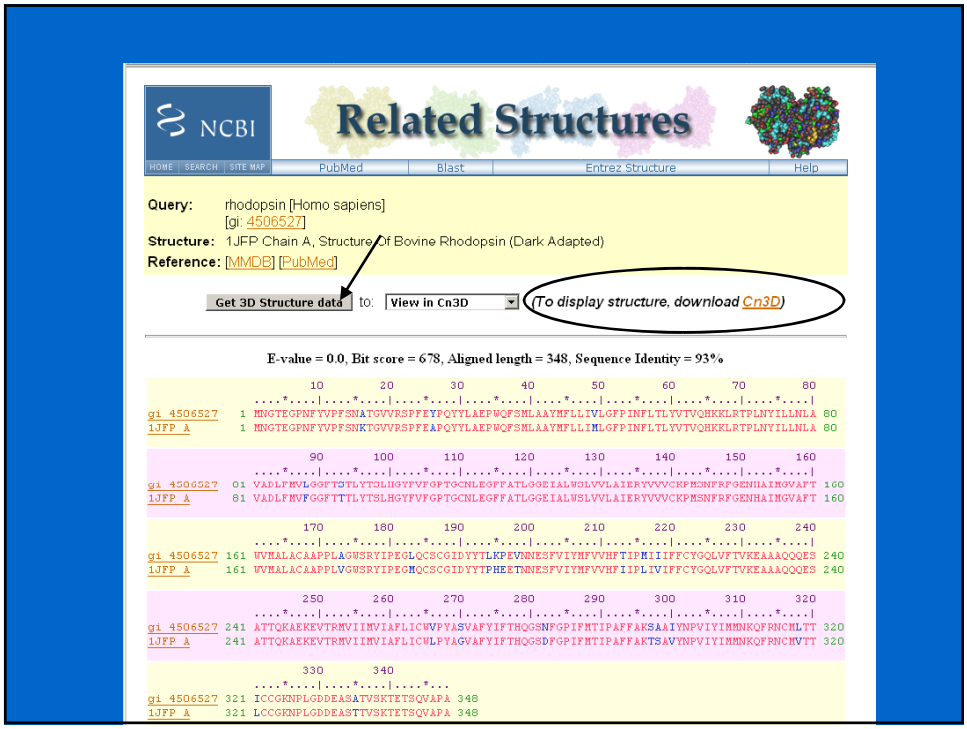
+180380
RHODOPSIN; RHO

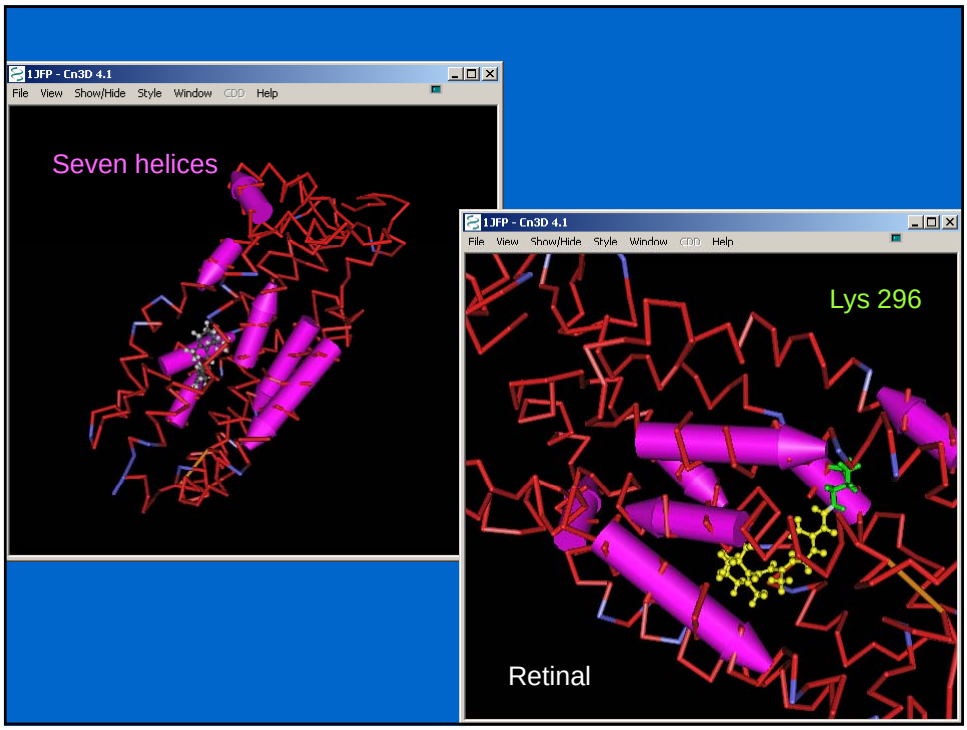
ALLELIC VARIANTS
(selected examples)

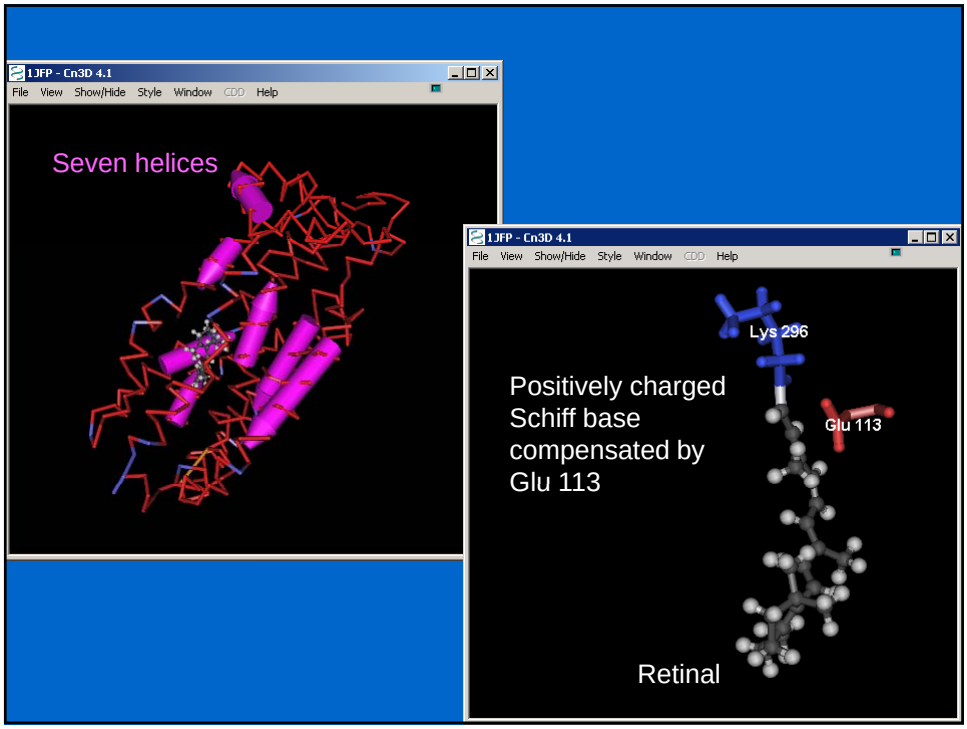
- 0001 RETINITIS PIGMENTOSA 4 [RHO, PRO23HIS]
- 0002 RETINITIS PIGMENTOSA 4 [RHO, PRO347LEU] dbSNP
- 0003 RETINITIS PIGMENTOSA 4 [RHO, PRO347SER] dbSNP
- 0004 RETINITIS PIGMENTOSA 4 [RHO, THR58ARG] dbSNP
- 0005 RETINITIS PIGMENTOSA 4 [RHO, 3-BP DEL]
- 0006 RETINITIS PIGMENTOSA 4 [RHO, THR17MET]
- 0007 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, TYR178CYS]
- 0014 RETINITIS PIGMENTOSA 4 [RHO, ASP190GLY]
- 0015 RETINITIS PIGMENTOSA 4 [RHO, GLN344TER]
- 0016 RETINITIS PIGMENTOSA 4 [RHO, LYS296GLU] dbSNP
- 0017 RETINITIS PIGMENTOSA 4 [RHO, ASP190ASN] 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, ASN15SER]
- 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




NCBI
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]
Get 3D Structure data
to:
View in Cn3D
(To display structure, download Cn3D)
E-value = 0.0, Bit score = 678, Aligned length = 348, Sequence Identity = 93%


1JFP - Cn3D 4.1
File View Show/Hide Style Window CDD Help
Seven helices
1JFP - Cn3D 4.1
File View Show/Hide Style Window CDD Help
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

NCBI

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

NCBI

Vision for Bioinformatics

Databases

Entrez global search
Genomes
Books
Gene
Nucleotide
RefSeq
Protein
Taxonomy
Homologene
UniGene
AceView
dbSNP
OMIM
Structure

Interpretation

Photoreceptors
cones and rods
Sequence similarity
Phylogeny
Homology
Expression
Structure
Function

Tools

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

NCBI