Lecture 1 - Introduction to Structural Bioinformatics

Motivation and Basics of Protein Structure

Objectives of the course

- **Understanding protein function.**
- **Applications to Computer Aided Drug Design.**
- **Development of efficient algorithms to evaluate the above “in silico”.**
- **Emphasis on the “structure” related problems – Geometric Computing in Molecular Biology.**
- **Show relevance to other spatial “pattern discovery” tasks.**

Most of the Protein Structure slides – courtesy of Hadar Benyaminy.

Textbook

There is no single, double or triple textbook for this course.

Most of the material is based on journal articles and research done by the Wolfson-Nussinov Structural Bioinformatics group at TAU.

Nevertheless :

Recommended Literature (1):

- Setubal and Meidanis, Introduction to Computational Biology, (1997).
- A. Lesk, *Introduction to* Protein Architecture, 2'nd edition (2001).
- S.L. Salzberg, D.B.Searls, S. Kasif (editors), Computational Methods in Molecular Biology, (1998).

Recommended Literature (2):

- Branden and Tooze, Introduction to Protein Structure (2'nd edition).
- D. Gusfield, Algorithms on Strings, Trees and Sequences, (1997).
- Voet and Voet, Biochemistry (or, any other Biochemistry book in the Library).
- M. Waterman, Introduction to Computational Biology.

Strongly Recommended Literature (currently not in the library):

- Protein Bioinformatics.
- Structural Bioinformatics.

Recommended Web Sites:

- Enormous number of sites.
- Search using “google”.
- PDB site http://www.rcsb.org/pdb/
- Birbeck course on protein structure.

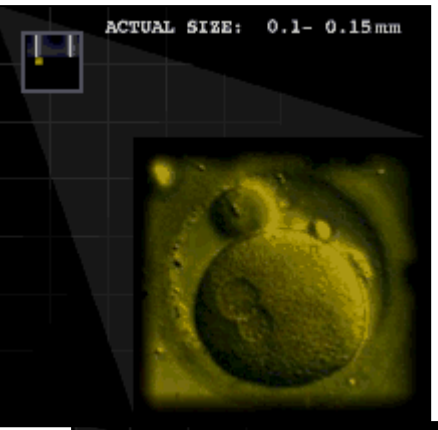
Journals :

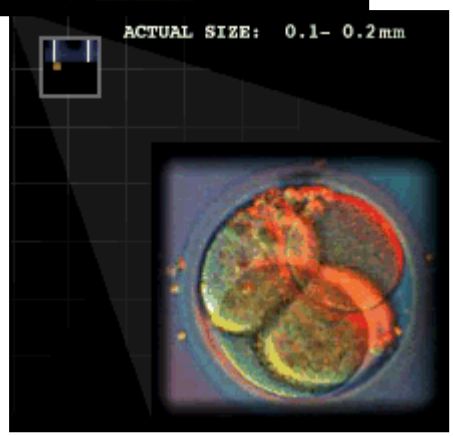
- Proteins : Structure, Function, bioinformatics.
- Journal of Computational Biology.
- Bioinformatics (former CABIOS).
- Journal of Molecular Biology.
- Journal of Computer Aided Molecular Design.
- Journal of Molecular Graphics and Modelling.
- Protein Engineering.

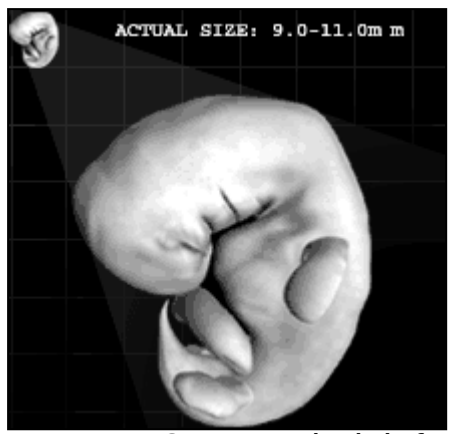
Computational Biology Conferences:

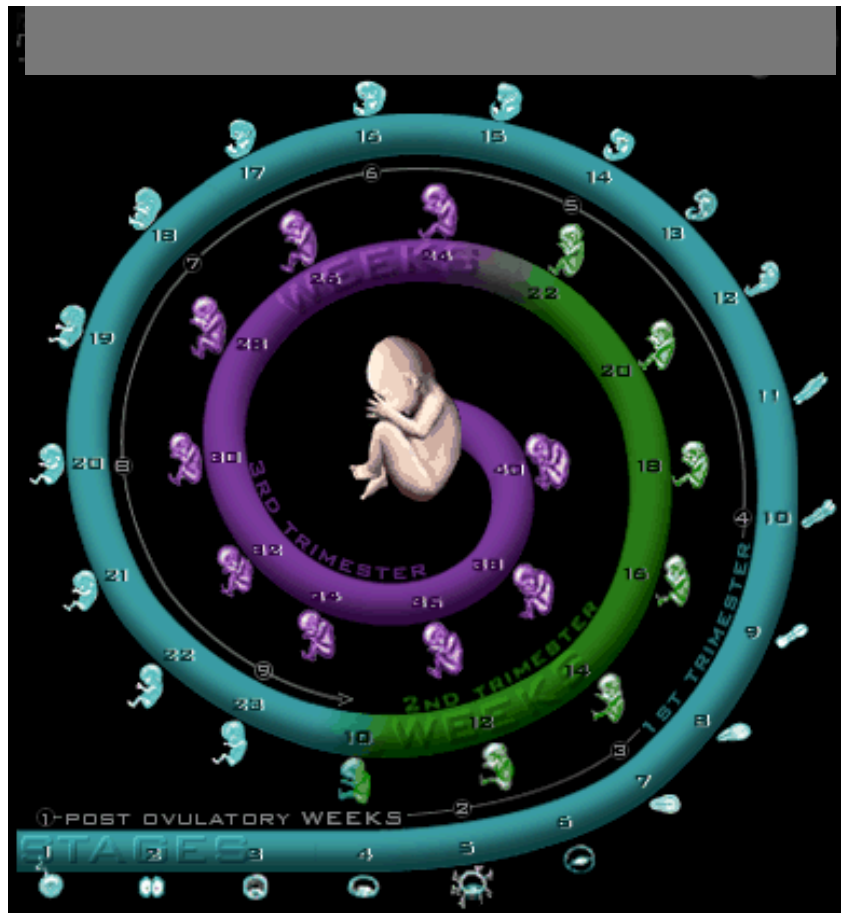
- **ISMB** - International Conference on Intelligent Systems in Molecular Biology.
- **RECOMB** - Int. Conference of Computational Molecular Biology.
- **ECCB** - European Conference on Computational Bio.
- **WABI** - Workshop of Algorithms in Bioinformatics .

Cell- the basic life unit

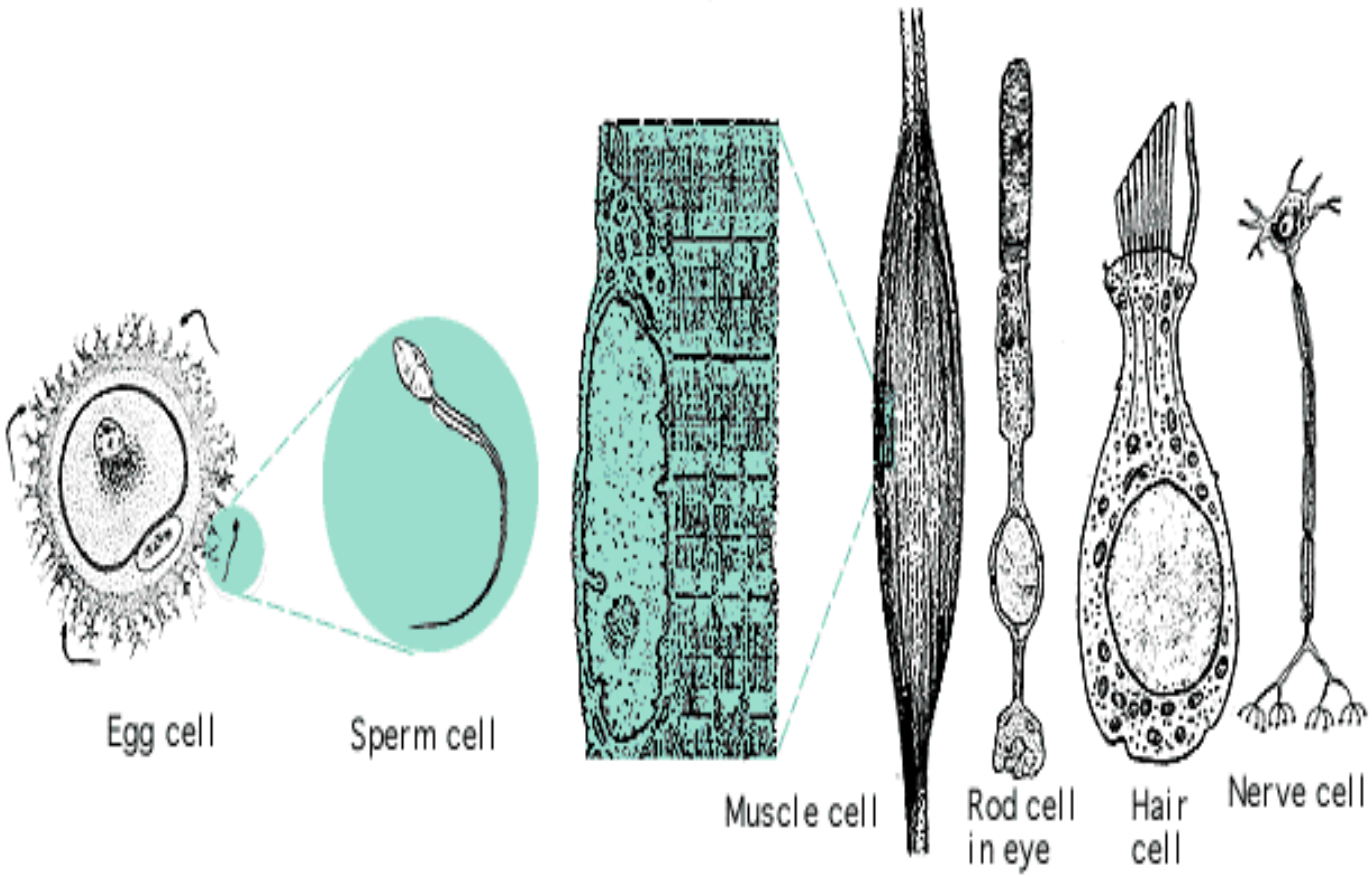








Different cell types



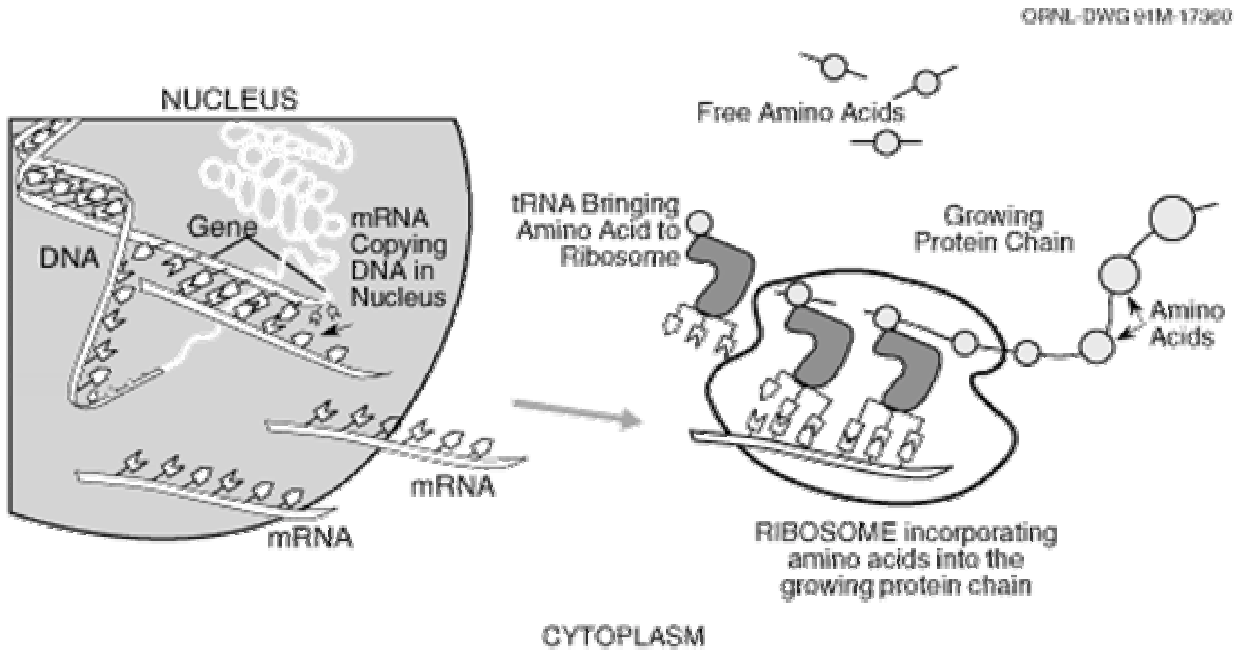
Size of protein molecules (diameter)

- cell (1x10^{-6} **m**) μ microns
- ribosome (1x10^{-9} **m**) nanometers
- protein (1x10^{-10} **m**) angstroms

The central dogma

- DNA ---> RNA ---> Protein

- {A,C,G,T} {A,C,G,U} {A,D,..Y}
- *4 letter alphabets* *20 letter alphabet*
- Sequence of nucleic acids seq of amino acids



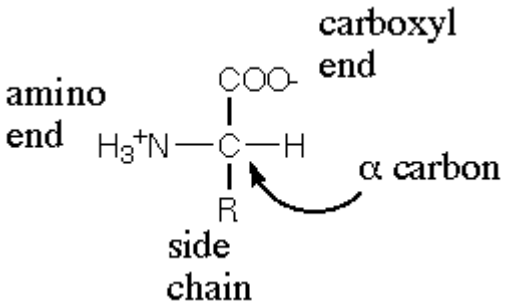
When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule of messenger RNA in a process similar to DNA replication The mRNA molecules then leave the cell nucleus and enter the cytoplasm, where triplets of (codons) forming the genetic code specify the particular amino acids that make up an) bases individual protein.

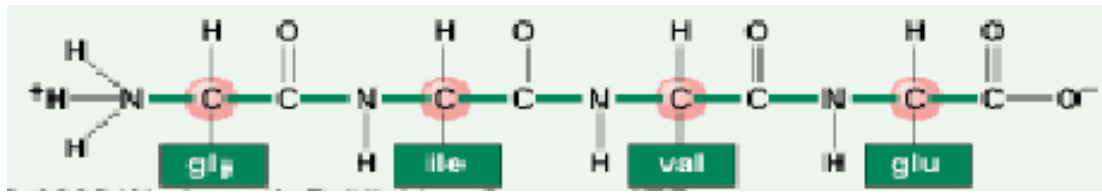
This process, called translation, is accomplished by ribosomes (cellular components composed of proteins and another class of RNA) that read the genetic code from the mRNA, and transfer RNAs (tRNAs) that transport amino acids to the ribosomes for attachment to the growing protein. (From www.ornl.gov/hgmis/publicat/primer/)

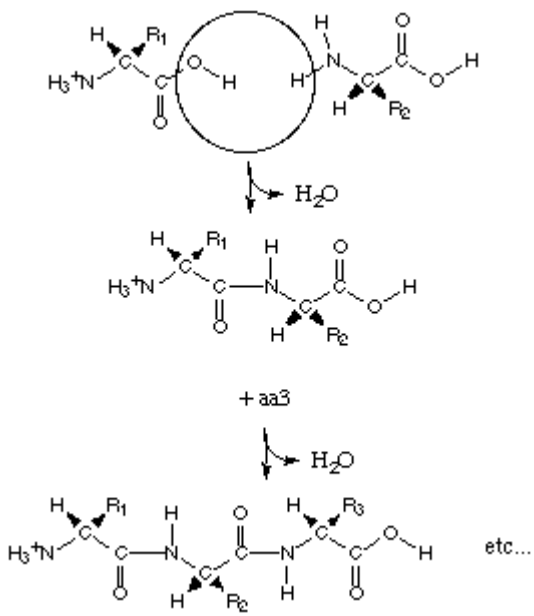
Proteins – our molecular machines (samples of protein tasks)

- Catalysis (enzymes).
- Signal propagation.
- Transport.
- Storage.
- Receptors (e.g. antibodies – immune system).
- Structural proteins (hair, skin, nails).

Amino acids and the peptide bond

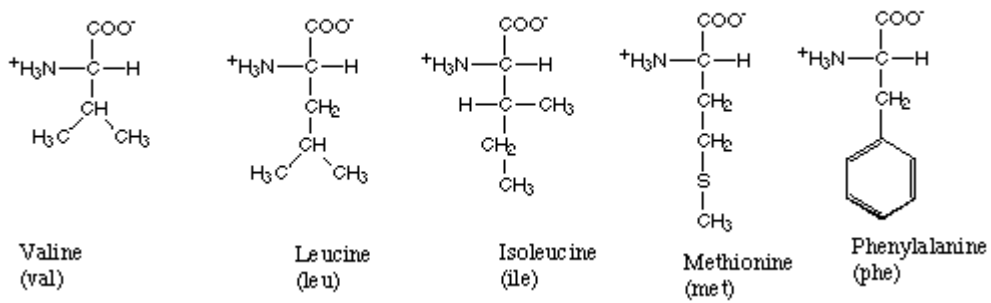






C_β – first side chain carbon (except for glycine).

Amino acids with hydrophobic side groups



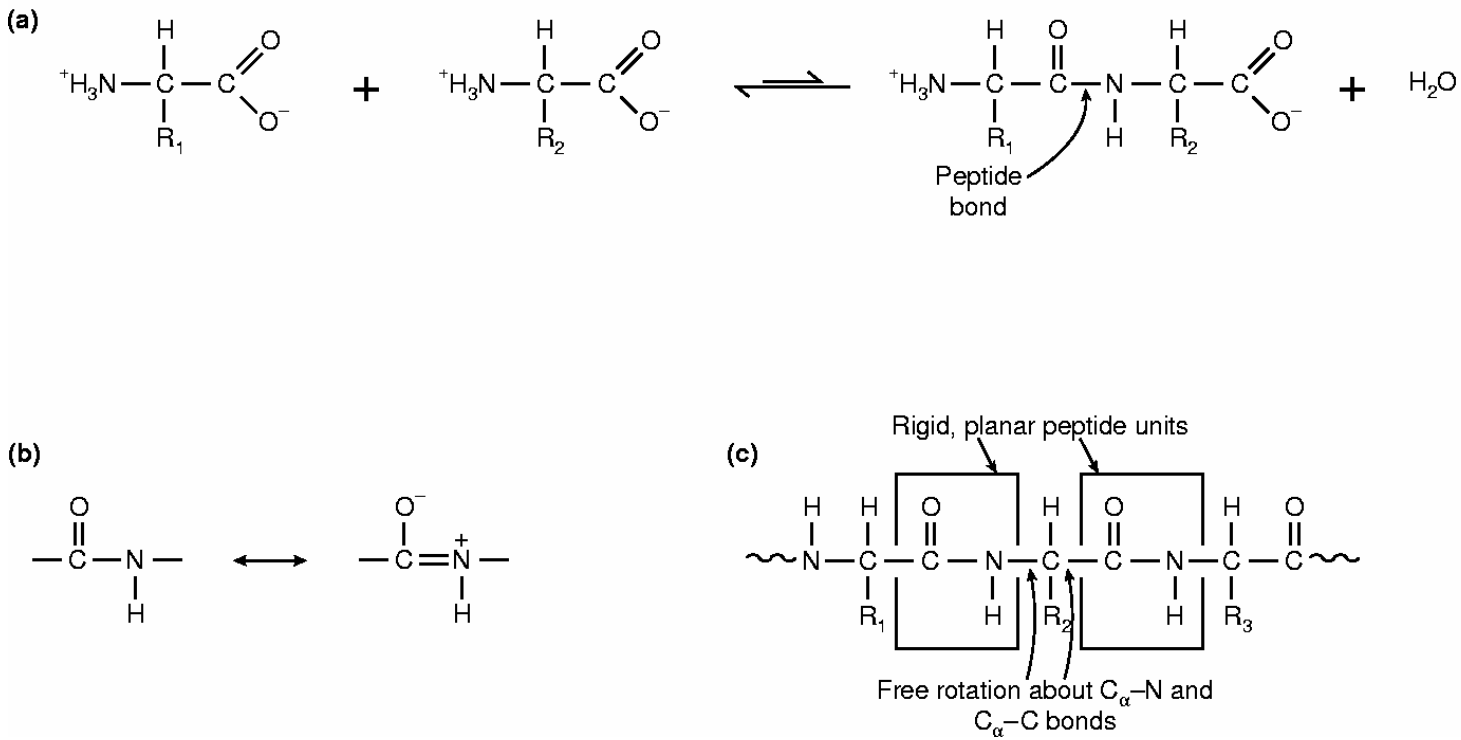
Primary through Quaternary structure

- Primary structure: The order of the amino acids composing the protein.

- AASGDXSLVEVHXXVFIVPPXIL…..

Folding of the Protein Backbone

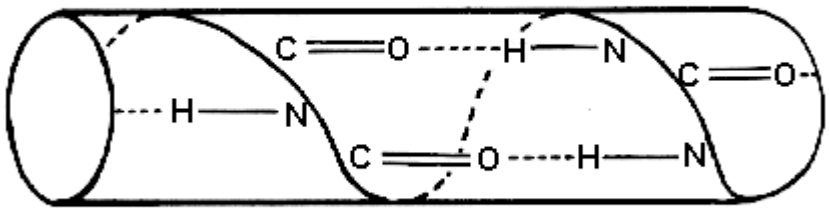
B2.1



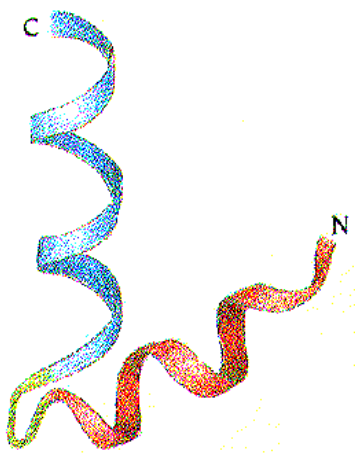
The Holy Grail - Protein Folding

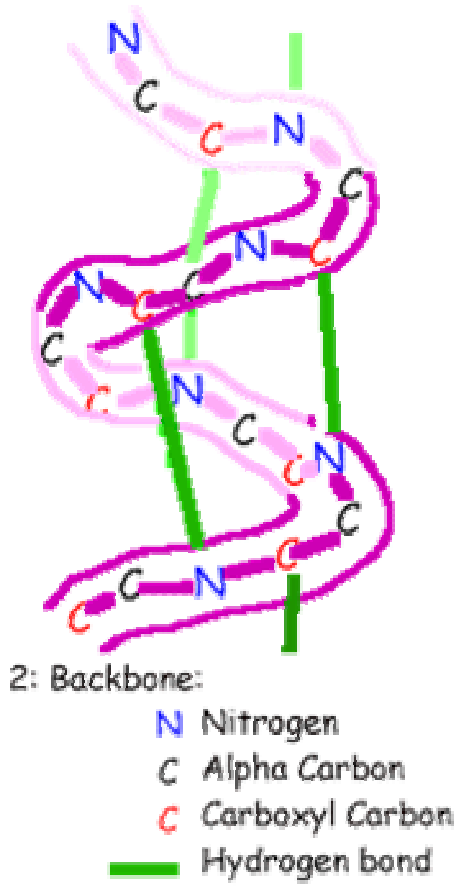
- How does a protein “know” its 3-D structure ?
- How does it compute it so fast ?
- Relatively primitive computational folding models have proved to be NP complete even in the 2-D case.

Secondary structure

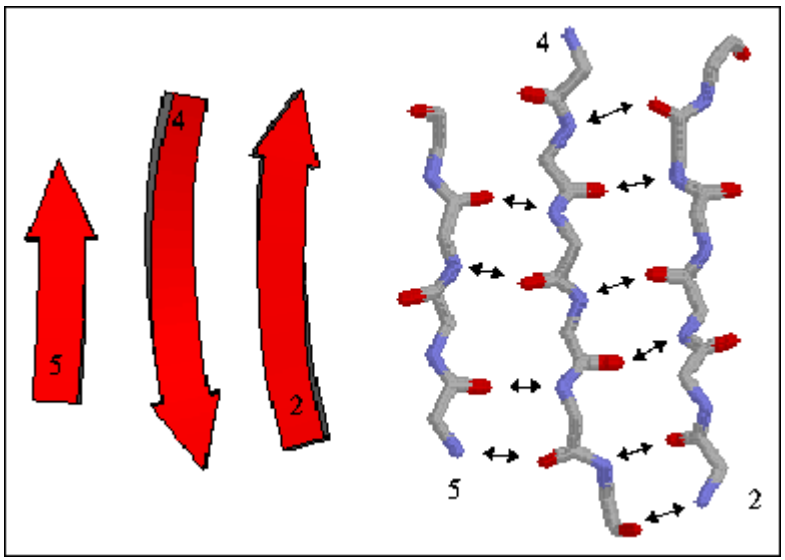


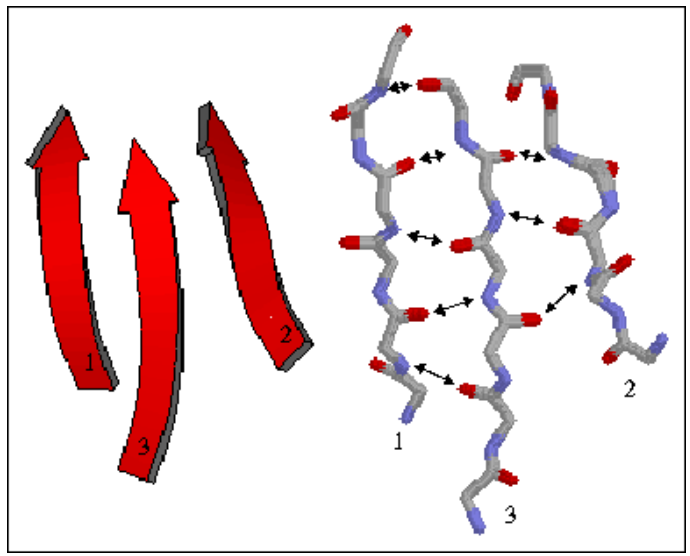
3.6 residues/turn (5.4 A dist.)



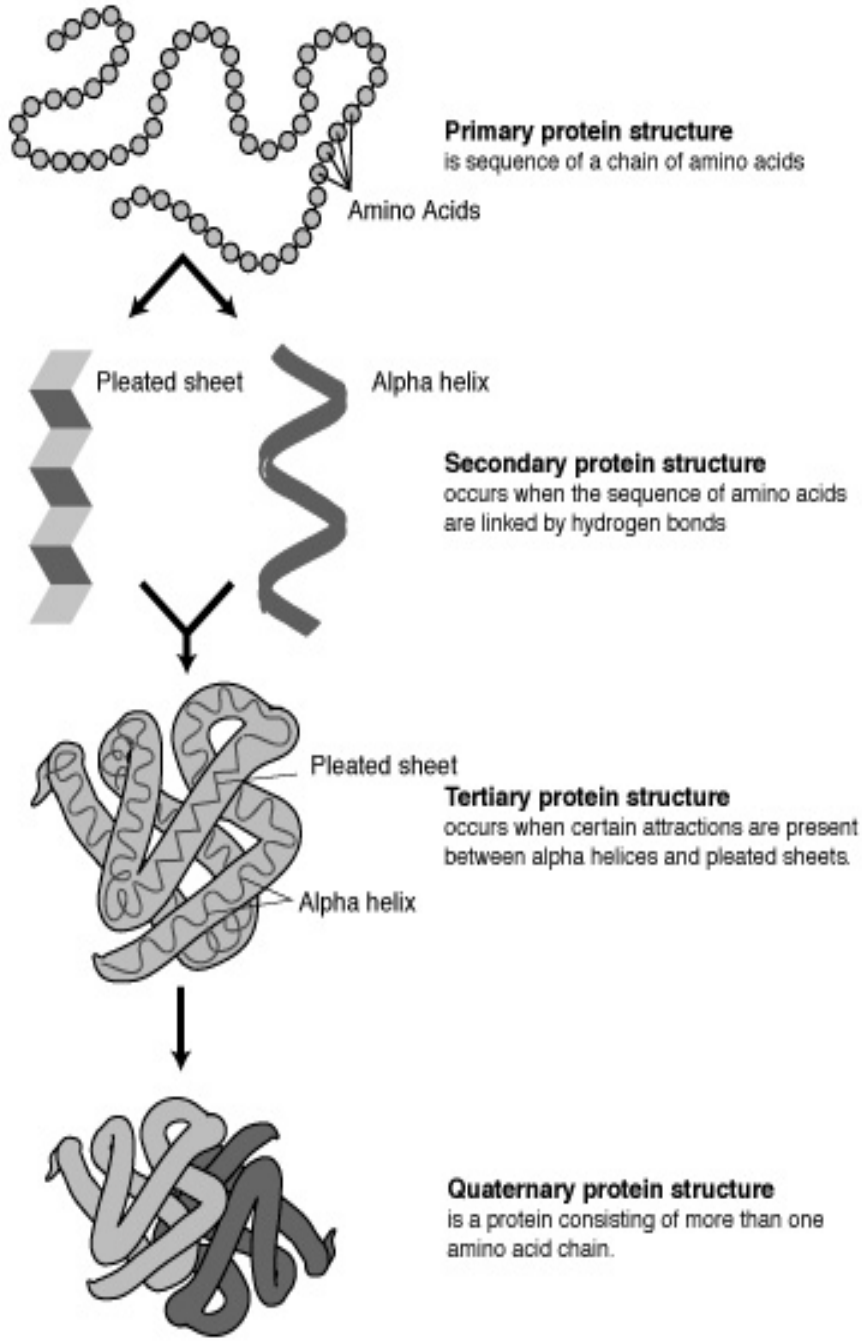


β strands and sheets

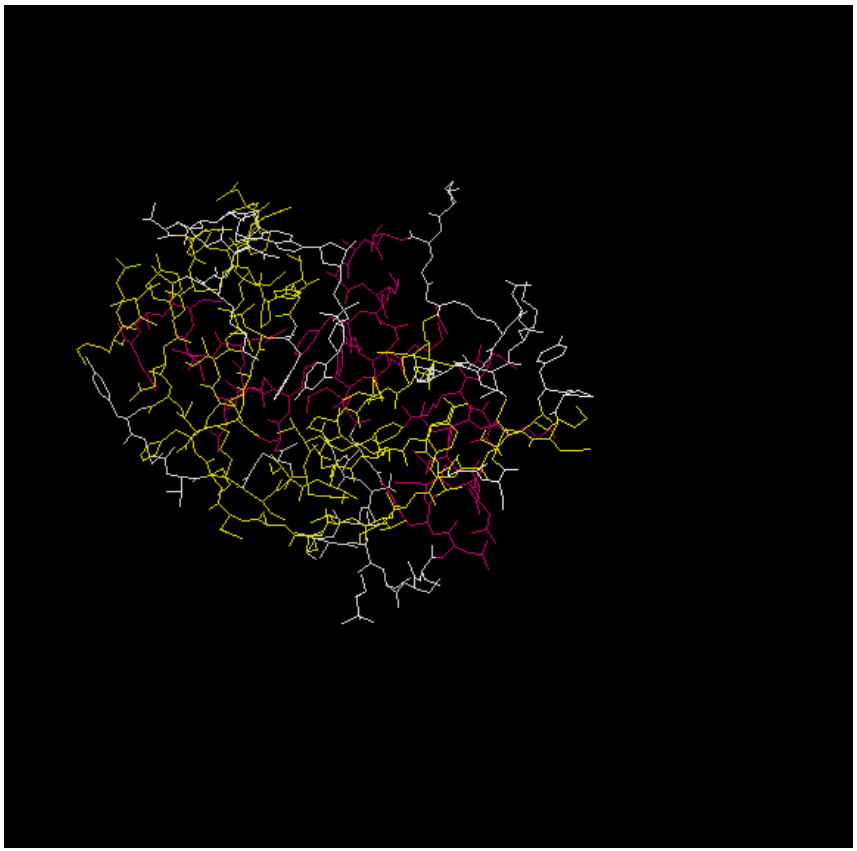




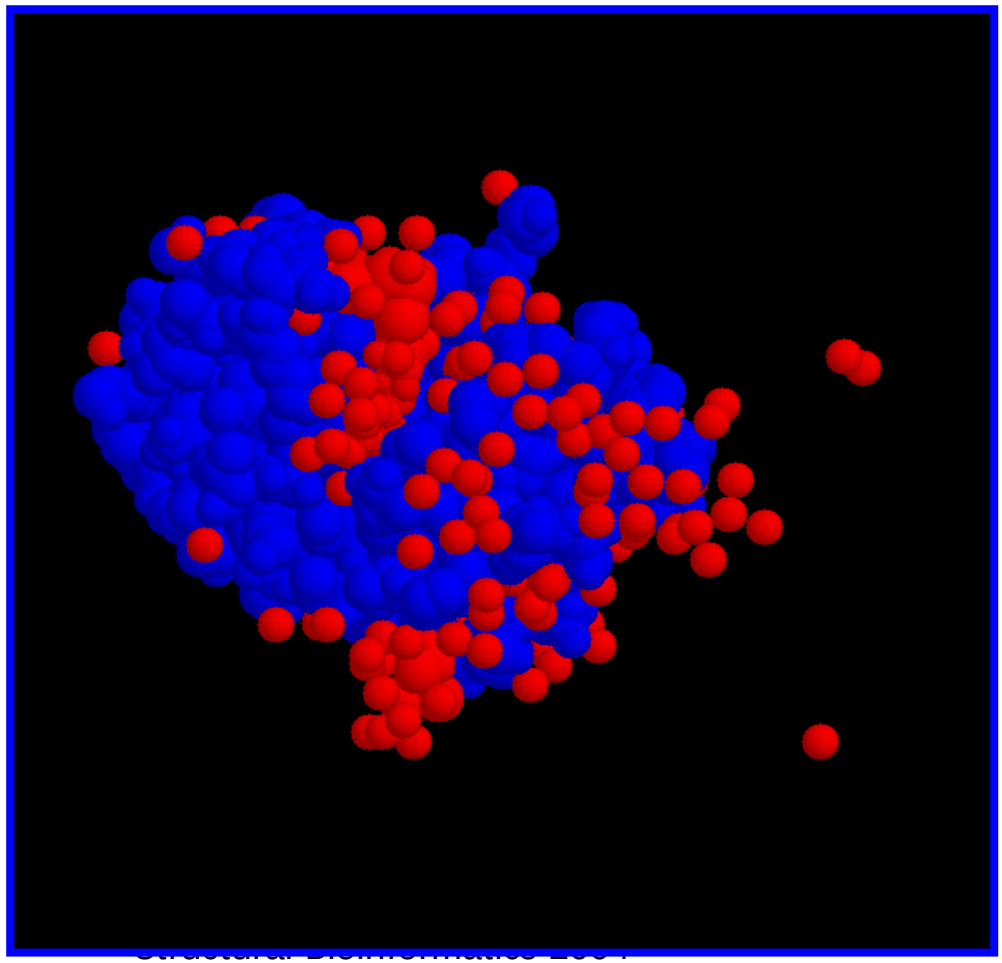
Bond. Hydrogen bond.


Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids
are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present
between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one
amino acid chain.

Wire-frame or ribbons display

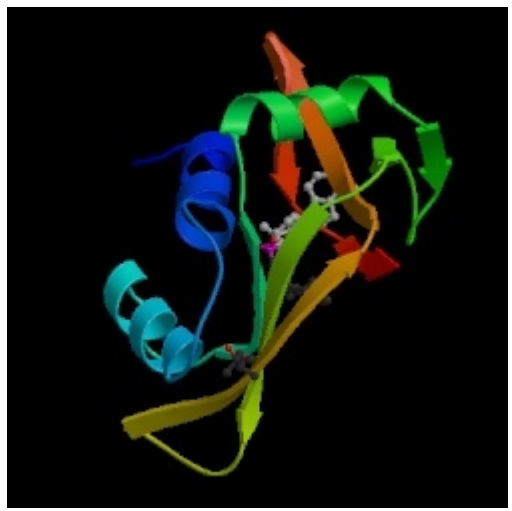


Space-fill display



Tertiary structure: full 3D folded structure of the polypeptide chain

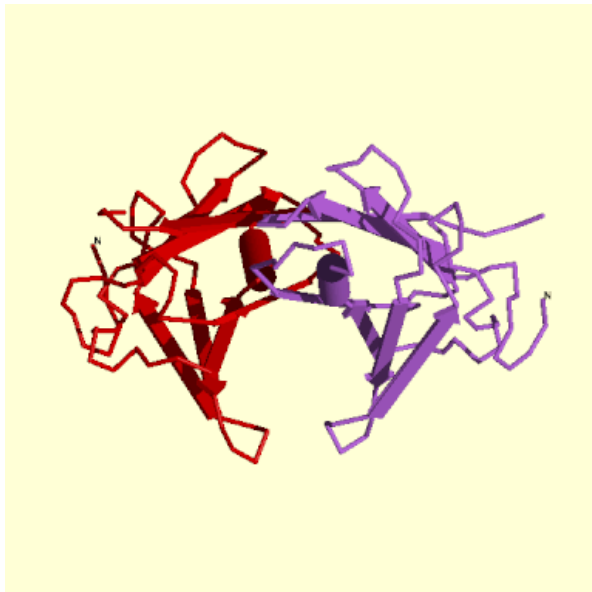
Ribonuclease - PDB code 1rpg



Quaternary structure

The interconnections and organization of more than one polypeptide chain.

Example :Transthyretin dimer (**1tta)**



Determination of protein structures

- X-ray Crystallography
- NMR (Nuclear Magnetic Resonance)
- EM (Electron microscopy)
- Nano – sensors (?)

X-ray Crystallography

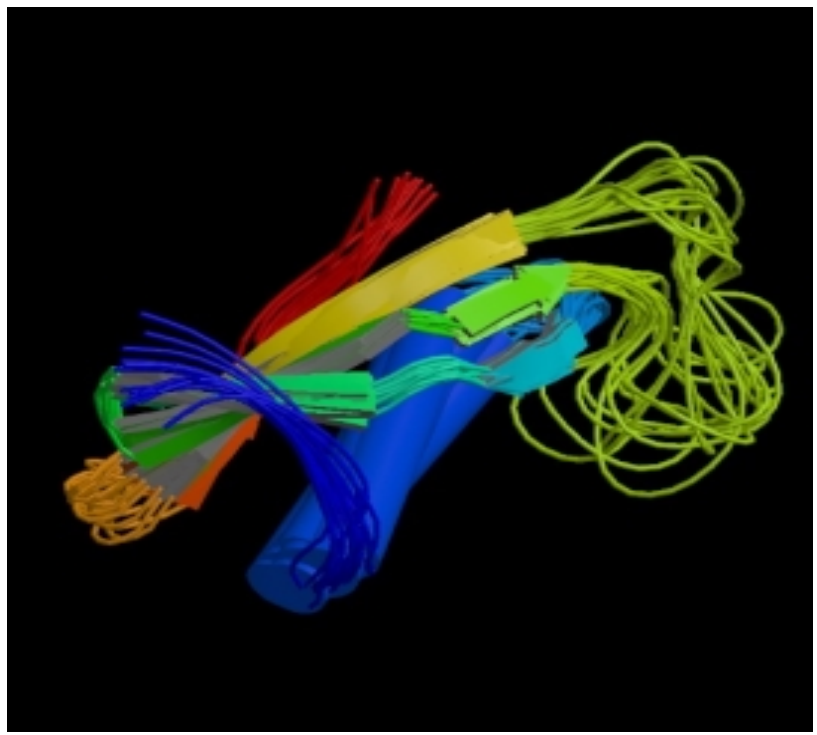
- Crystallization
- Each protein has a unique **X-ray pattern diffraction.**
- The **electron density map** is used to build a model of the protein.

Nuclear Magnetic Resonance

- Performed in an **aqueous solution**.
- NMR analysis gives **a set of estimates of distances between specific pairs of protons (H – atoms).**
- Solved by Distance Geometry methods.
- The result is an **ensemble of models** rather than a single structure.

An NMR result is an ensemble of models

Cystatin (**1a67)**



The Protein Data Bank (PDB)

- International repository of 3D molecular data.
- Contains x-y-z coordinates of all atoms of the molecule and additional data.

PDB™
PROTEIN DATA BANK
Research Collaboratory for Structural Bioinformatics

RCSB Home Contact Us

RCSB Sites: SDSC * Rutgers * NIST
Mirror Sites

Welcome to the PDB, the single international repository for the processing and distribution of 3-D macromolecular structure data primarily determined experimentally by X-ray crystallography and NMR.

DEPOSIT **Contribute structure data**

STATUS **Find entries awaiting release**

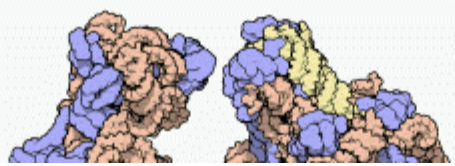
DOWNLOAD **Retrieve structure files (FTP)**

LINKS **Browse related information**

PREVIEW **Beta-test new features**

About the PDB

General Information
WWW User Guides
Get Educated



Current Holdings

13505 Structures
Last Update: 24-Oct-2000
PDB Statistics

Search

Enter a PDB ID: Explore
SearchLite: simple keyword search
SearchFields: advanced search

News

Complete News
PDB Newsletter
Subscribe
Browse Mailing List

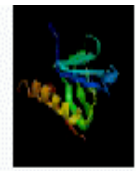
24-Oct-2000

Issue 7 of the PDB Newsletter Now

Feb. 2003 – about 20,000 structures.

PDB
PROTEIN DATA BANK

Structure Explorer - 1IRS



Summary Information

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details

Explore
SearchLite SearchFields

Compound: **Mol_Id: 1; Molecule: Irs-1; Chain: A; Fragment: Ptb Domain; Synonym: Insulin Receptor Substrate 1; Engineered: Yes**
Mol_Id: 2; Molecule: Il-4 Receptor Phosphopeptide; Chain: B; Engineered: Yes

Authors: **M.-M. Zhou, B. Huang, E. T. Olejniczak, R. P. Meadows, S. B. Shuker, M. Miyazaki, T. Trub, S. E. Shoelson, S. W. Feisk**

Exp. Method: **NMR, Minimized Average Structure**

Classification: **Complex (Signal Transduction/Peptide)**

Source: **Homo Sapiens**

Primary Citation: **Zhou, M. M., Huang, B., Olejniczak, E. T., Meadows, R. P., Shuker, S. B., Miyazaki, M., Trub, T., Shoelson, S. E., Fesik, S. W.: Structural basis for IL-4 receptor phosphopeptide recognition by the IRS-1 PTB domain.** *Nat Struct Biol* **3** *pp.* **388 (1996)**
[Medline]

Deposition Date: **22-Mar-1996** *Release Date:* **15-May-1997**

Polymer Chains: **A, B** *Residues:* **123**

Atoms: **971**

HET groups:

ID	*Name*	*Formula*
PTR	PHOSPHOTYROSINE	$C_9H_{12}N_1O_6P_1$

Classification of 3D structures

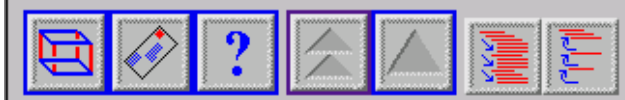
SCOP

- Provides a description of the structural and evolutionary relationships between all proteins whose structure is known.
- Created largely by manual inspection.

- J. Mol. Biol. 247, 536-540, 1995

SCOP

Structural Classification of Proteins



Protein: Hemoglobin, alpha-chain from Human (*Homo sapiens*)

Lineage:

1. Root: scop
2. Class: All alpha proteins
3. Fold: Globin-like
   core: 6 helices; folded leaf, partly opened
4. Superfamily: Globin-like
5. Family: Globins
   Heme-binding protein
6. Protein: Hemoglobin, alpha-chain
7. Species: Human (*Homo sapiens*)

PDB Entry Domains:

1. 1bab
   complexed with hem, so4; mutant
   1. chain a
   2. chain c
2. 1bz0

CATH - Protein Structure Classification
http://www.biochem.ucl.ac.uk/bsm/cath/

CATH

Protein Structure Classification

Version 1.6 : Released June 1999

Welcome to the **CATH** protein classification home page
Biomolecular Structure and Modelling Unit,
University College London.

Dr. Frances M.G. Pearl, Mr. James Bray, Ms. Annabel E. Todd,
Dr. David Lee, Dr. Adrian J. Shepherd, Dr. Andrew Harrison, Prof. Janet Thornton
Dr. Christine A. Orengo

Available options:

- Browse or search classification
- Lexicon
- Glossary

CATH

- Class: derived from secondary structure content.
- Architecture: gross orientation of secondary structures, independent of connectivities.
- Topology: clusters according to topological connections and numbers of secondary structures.

- Homology: clusters according to structure and function.

HOME PAGE | TOP | UP | HIDE | LEX | GLOS | HELP

Class Architecture Topology Homologous superfamily

1.10.490

C Mainly Alpha

A Non-Bundle

T Globin-like

[H 10] 1hlm (12 S's)

[S 1] 1hlm

1hlm GET DATA...

PDB header: Oxygen Transport
PDB comp: Hemoglobin (Cyano-Met) (Sea Cucumber)
PDB source: Sea Cucumber (Caudina (Molpadia) Arenicola)

1hlb GET DATA...

PDB header: Oxygen Transport
PDB comp: Hemoglobin (Sea Cucumber)
PDB source: Sea Cucumber (Caudina (Molpadia) Arenicola)

[S 2] 1hbg

1hbg GET DATA...

PDB header: Oxygen Transport
PDB comp: Hemoglobin (Carbon Monoxy)
PDB source: Marine Bloodworm (Glycera Dibranchiata)

2hbg GET DATA...

PDB header: Oxygen Transport
PDB comp: Hemoglobin (Deoxy)

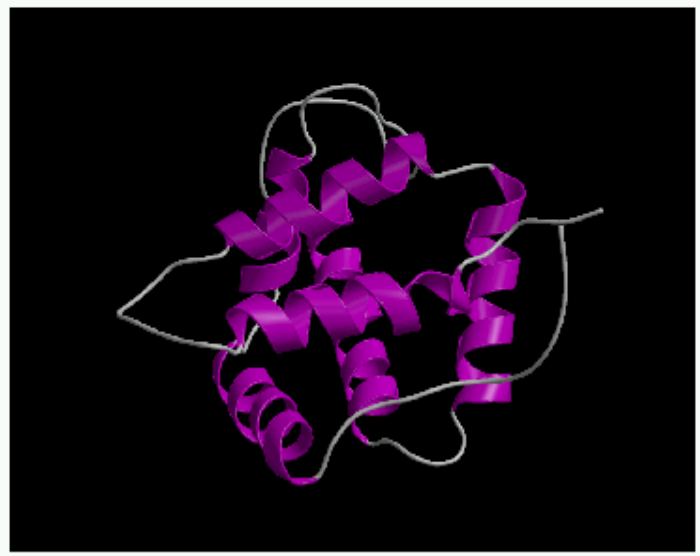
CATH LEXICON

Mainly Alpha

These proteins consist predominantly of alpha helix secondary structures, although many also contain a small percentage of beta sheet on their peripheries. Mainly alpha proteins have been assigned using a cutoff of >60% alpha and <5% beta secondary structure assignment. In addition these proteins must have >50% alpha-alpha and <5% beta-beta secondary structure contacts (Michie *et al.*,1996)

Non-Bundle

The non-bundle architecture is a general

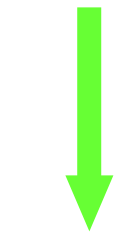


Representative for CATH code **1.10.490.10.1**
Mainly Alpha : Non-Bundle : Globin-like : 1hlm
PDB: **1hlm**

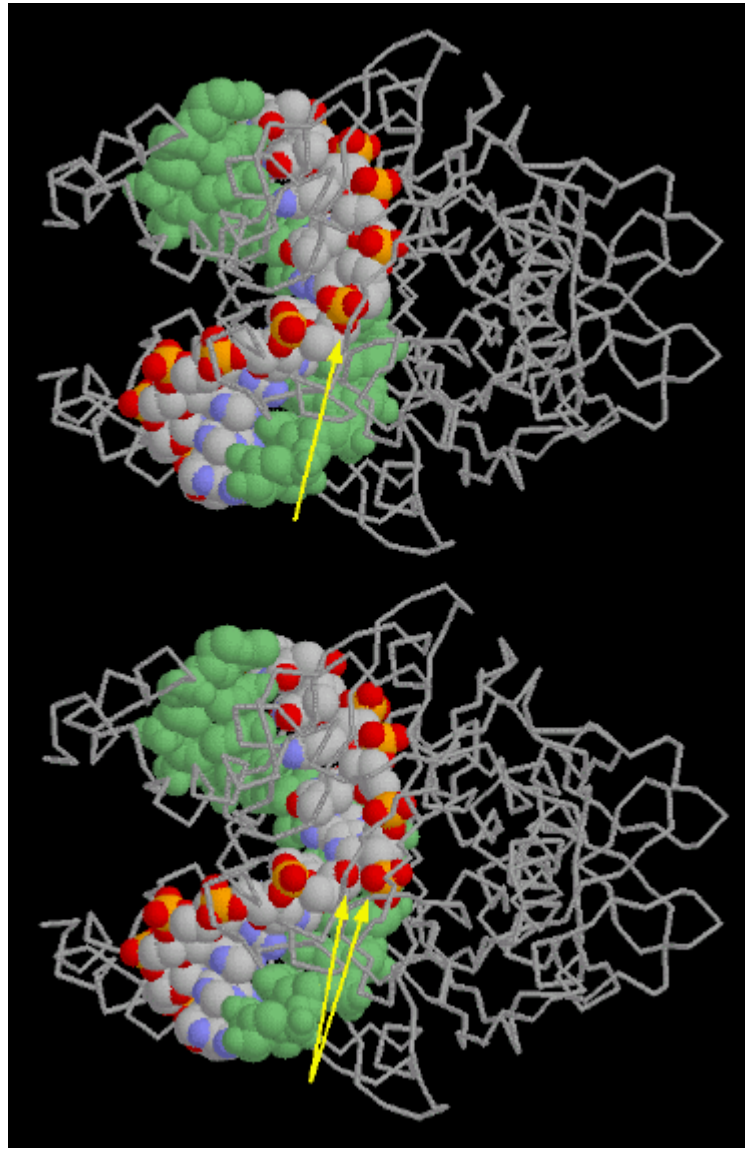
- **PDB http://pdb.tau.ac.il**
- **PDB http://www.rcsb.org/pdb/**
- **CATH**
  http://www.biochem.ucl.ac.uk/bsm/cath/
- **SCOP http://scop.mrc-lmb.cam.ac.uk/scop/**

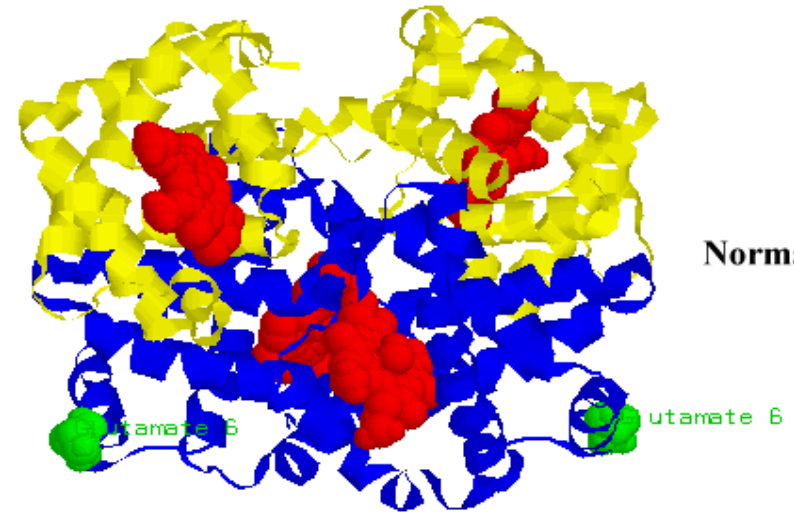
Restriction enzymes

```
---GAATTC---
||||||||||||
---CTTAAG---
```

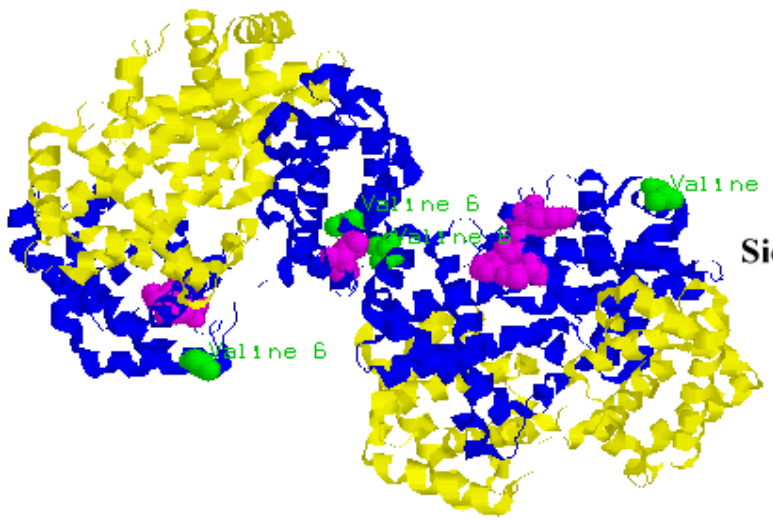


```
---G     AATTC---
||||         ||||
---CTTAA     G---
```





Normal Hemoglobin



Sickle Hemoglobin

Note: The Sickle hemoglobin image is drawn at 50% of the size of the Normal hemoglobin

The Structural Genomics Pipeline (X-ray Crystallography)

Basic Steps

Target Selection

Crystallomics
- Isolation,
- Expression,
- Purification,
- Crystallization

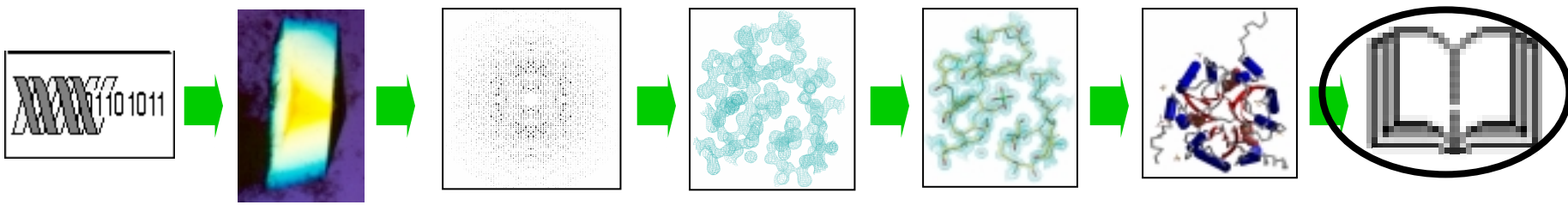
Data Collection

Structure Solution

Structure Refinement

Functional Annotation

Publish



Bioinformatics
- Distant homologs
- Domain recognition

Automation
Bioinformatics
- Empirical rules

Automation
Better sources

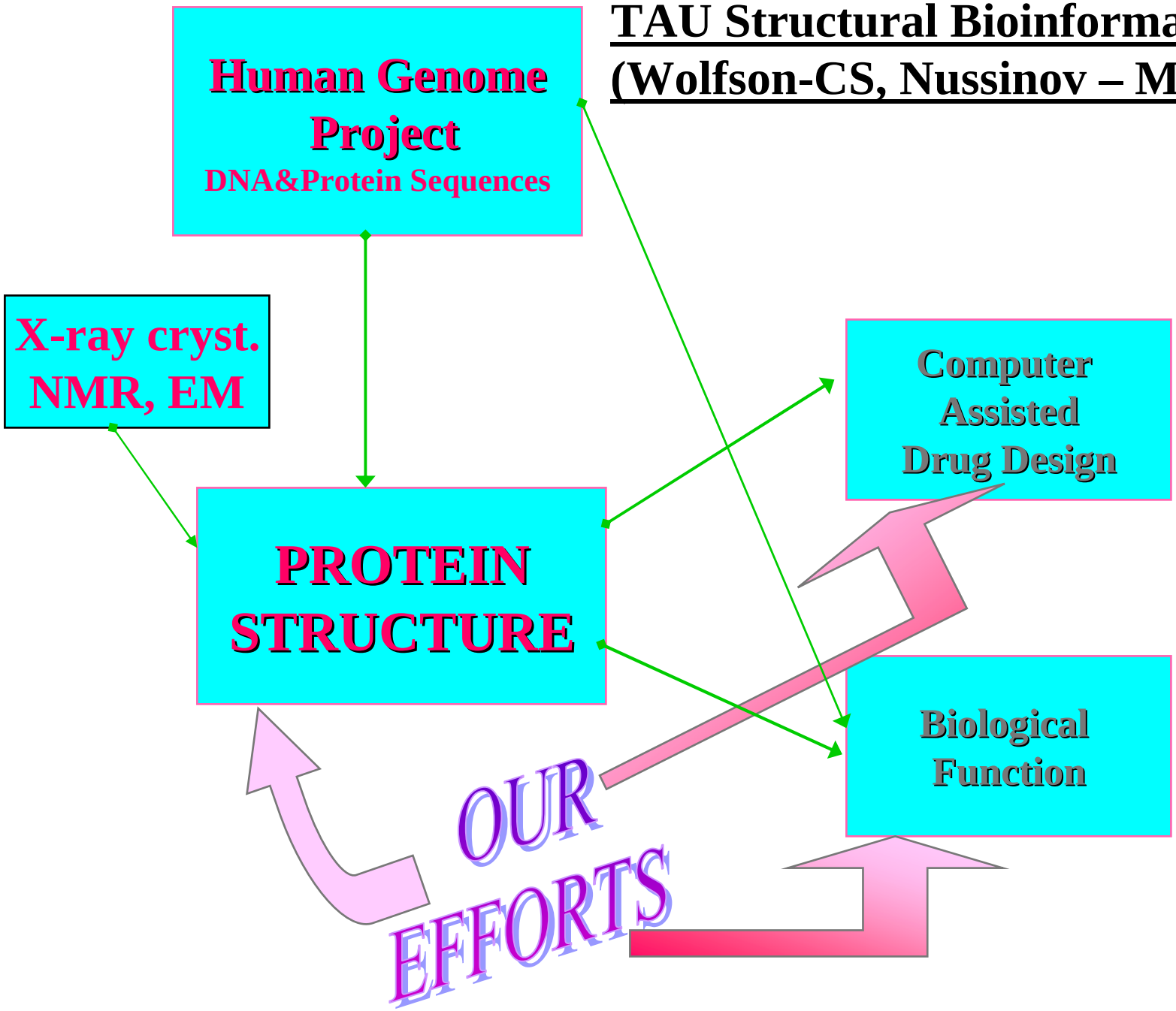
Software integration
Decision Support
MAD Phasing
Automated fitting

Bioinformatics
- Alignments
- Protein-protein interactions
- Protein-ligand interactions
- Motif recognition

No?

Borrowed from Bourne's (UCSD) lecture on CADD

TAU Structural Bioinformatics Lab (Wolfson-CS, Nussinov – MB)



Structural Bioinformatics Lab Goals

Development of *state of the art* algorithmic methods to tackle major computational tasks in protein structure analysis, biomolecular recognition, and *Computer Assisted Drug Design*.

Establish truly *interdisciplinary* collaboration between Life and Computer Sciences.

Bioinformatics and Genomics - Economic Impact

- **Medicine and public health.**
- **Pharmaceutics.**
- **Agriculture.**
- **Food industry.**
- **Biological Computers (?).**

Bioinformatics and Genomics - the Computational Viewpoint

- **Molecular Biology is becoming a Computational Science.**
- **The emergence of large databases of DNA, proteins, small molecules and drugs requires computational techniques to analyze the data.**
- **Efficient CPU and memory intensive algorithms are being developed.**
- **Many of the computational tasks have analogs in other well established fields of Computer Science allowing cross-fertilization of ideas.**

Bioinformatics - Computational Genomics

- DNA mapping.
- Protein or DNA sequence comparisons , *primary structure*.
- Exploration of huge textual databases.
- In essence one- dimensional methods and intuition.
- Graph - theoretic methods.

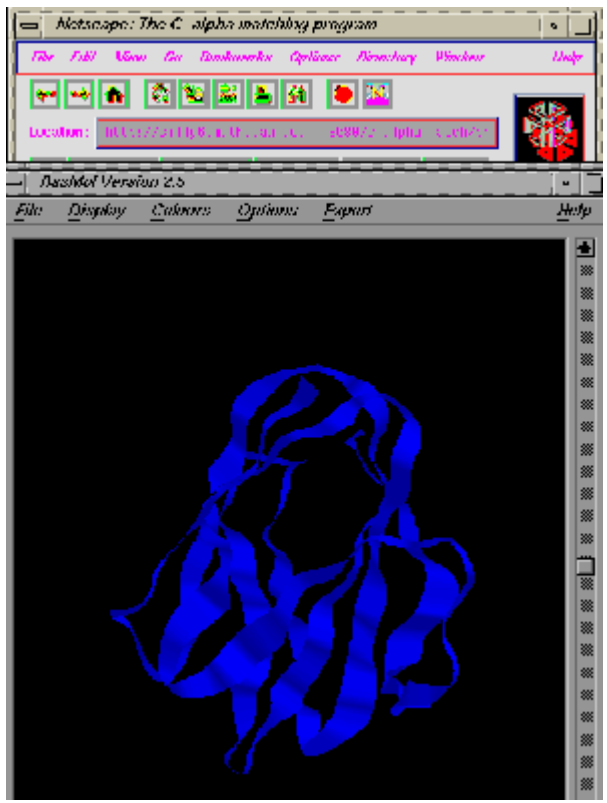
Structural Bioinformatics - Structural Genomics

- **Elucidation of the 3D structures of biomolecules.**
- **Analysis and comparison of biomolecular structures.**
- **Prediction of biomolecular recognition.**
- **Handles three-dimensional (3-D) structures.**
- ***Geometric Computing*.**

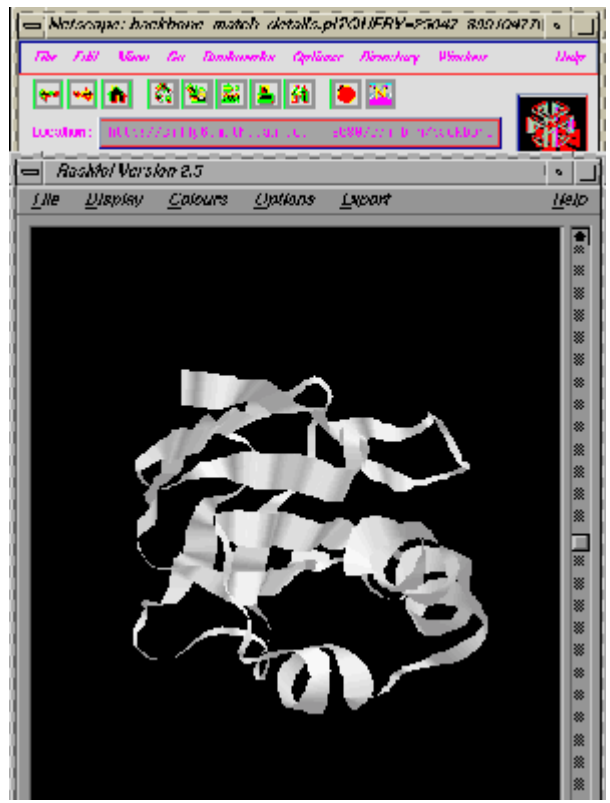
Why bother with structures when we have sequences ?

- **In evolutionary related proteins structure is much better preserved than sequence.**
- **Structural motifs may predict similar biological function .**
- **Getting insight into protein folding. Recovering the limited (?) number of protein folds.**

Case in Point : Protein Structural Comparison



ApoAmicyanin - 1aaj



Pseudoazurin - 1pmy

Geometric Task :

Given two configurations of points in the three dimensional space,

find those rotations and translations of one of the point sets which produce "large" superimpositions of corresponding 3-D points.

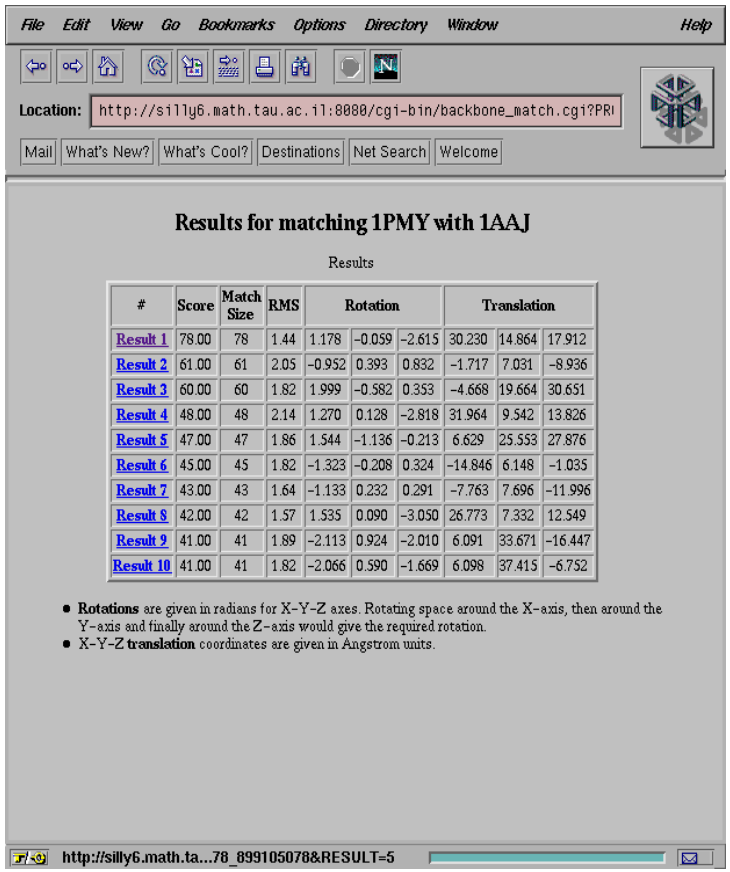
Remarks :

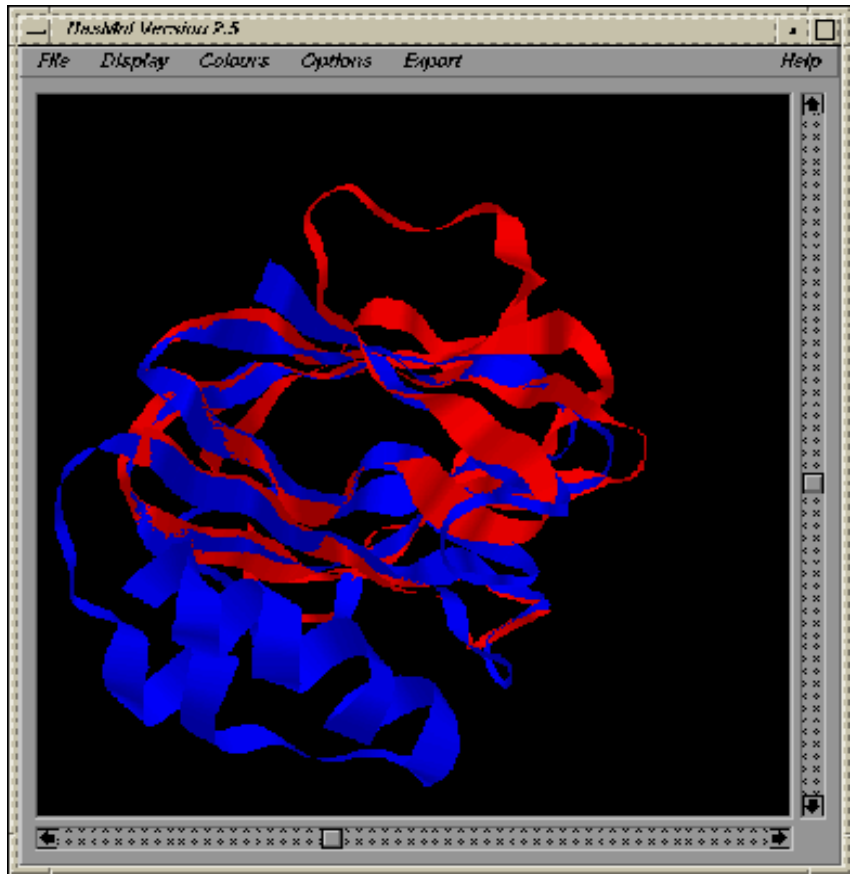
The superimposition pattern is not known a-priori – *pattern discovery* .

The matching recovered can be *inexact*.

We are looking not necessarily for the largest superimposition, since other matchings may have *biological meaning*.

Algorithmic Solution



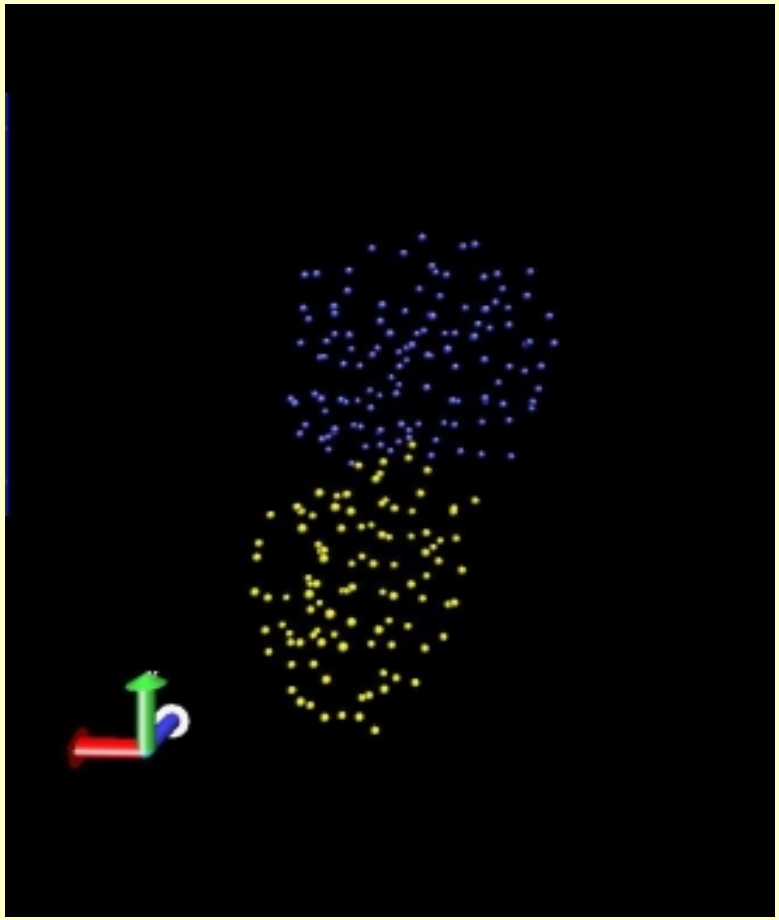


About 1 sec. Fischer, Nussinov, Wolfson ~ 1990.

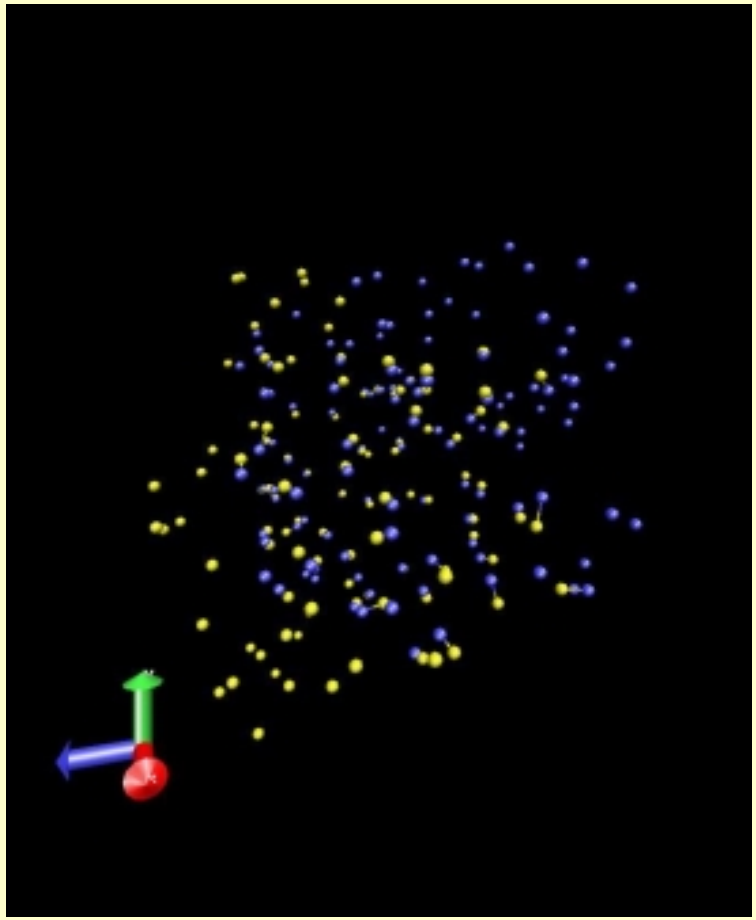
Applications

- **Classification of protein databases by structure.**
- **Search of partial and disconnected structural patterns in large databases.**
- **Detection of structural pharmacophores in an ensemble of drugs.**
- **Comparison and detection of drug receptor *active sites*.**

Geometric Matching task = Geometric Pattern Discovery

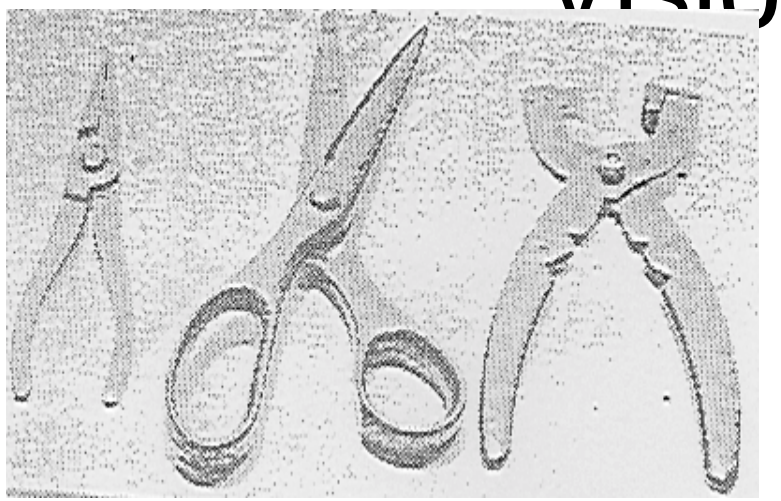


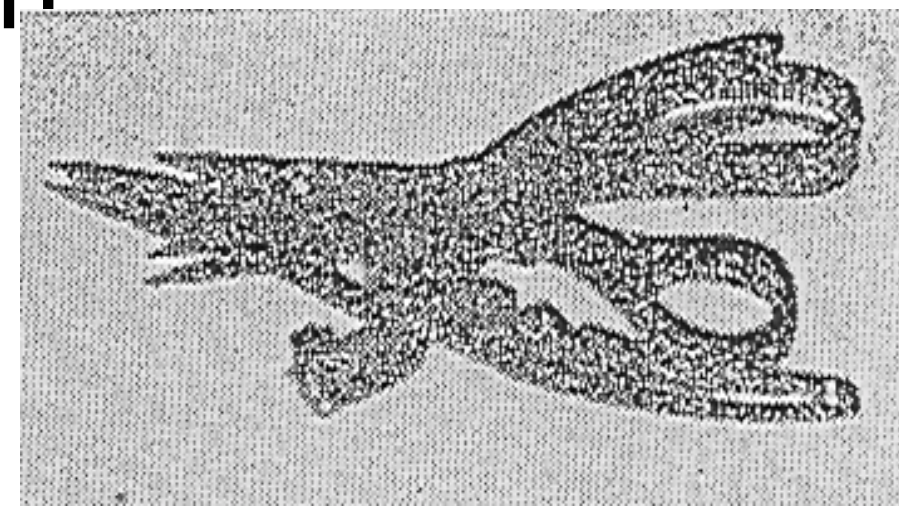
C_α constellations - before

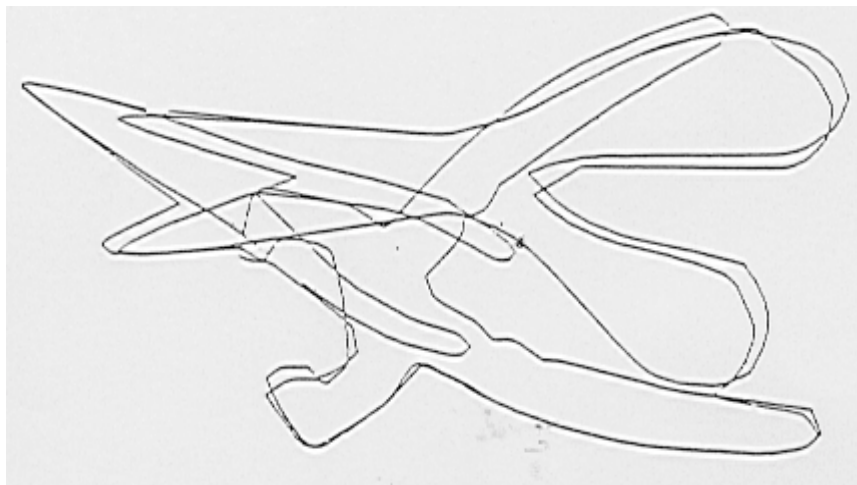


Superimposed constellations

Analogy with Object Recognition in Computer Vision

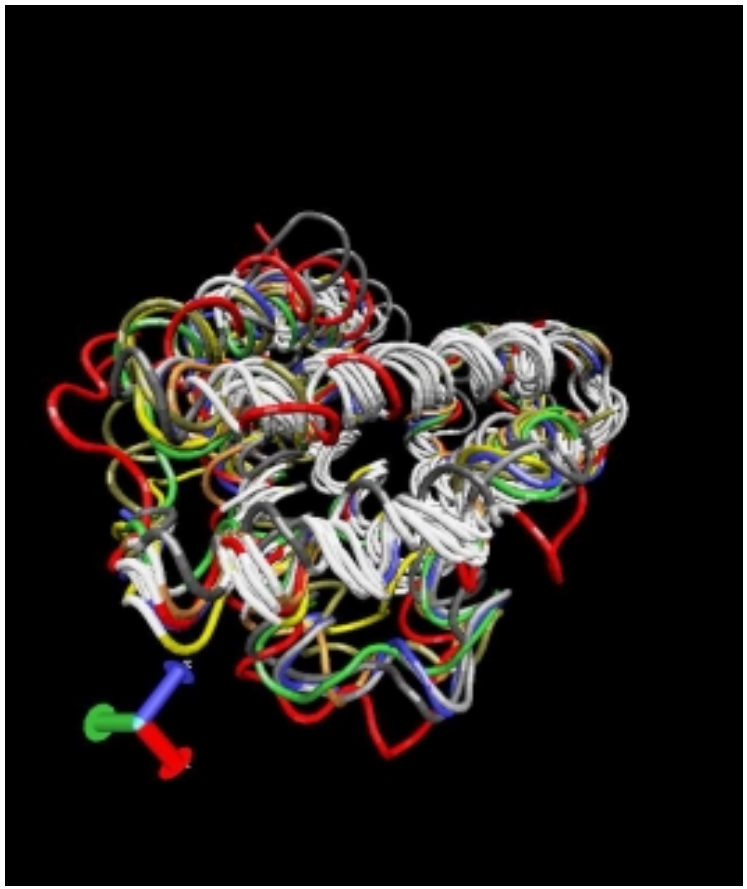


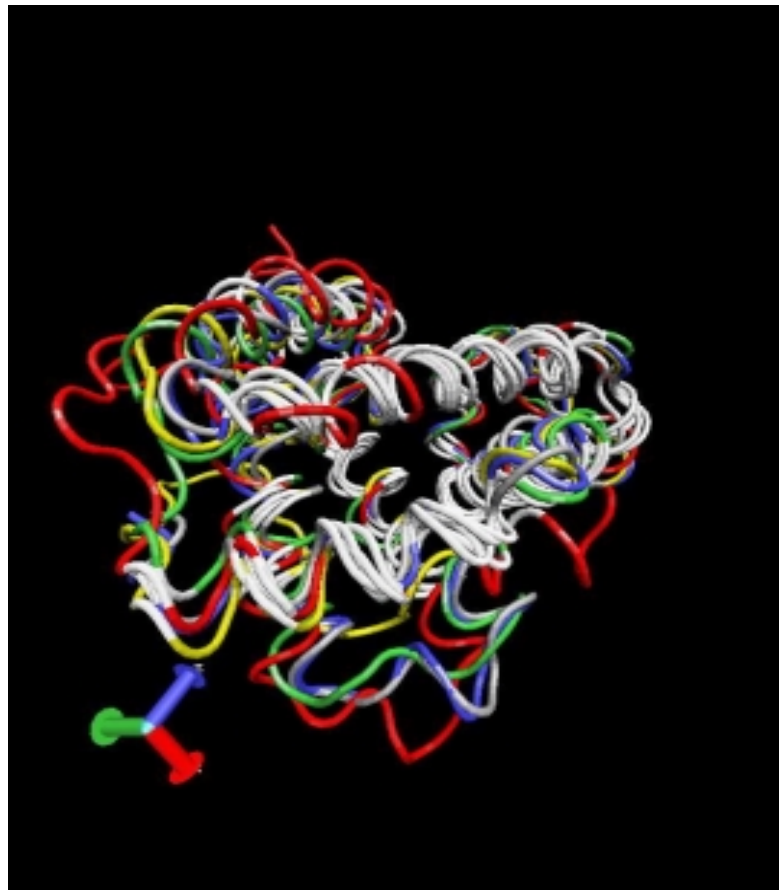




Wolfson, “Curve Matching”,1987.

Multiple Structural Alignment (Globin example)





Leibowitz, Fligelman, Nussinov, Wolfson, - ISMB'99 – Heidelberg.

Biomolecular Recognition - docking

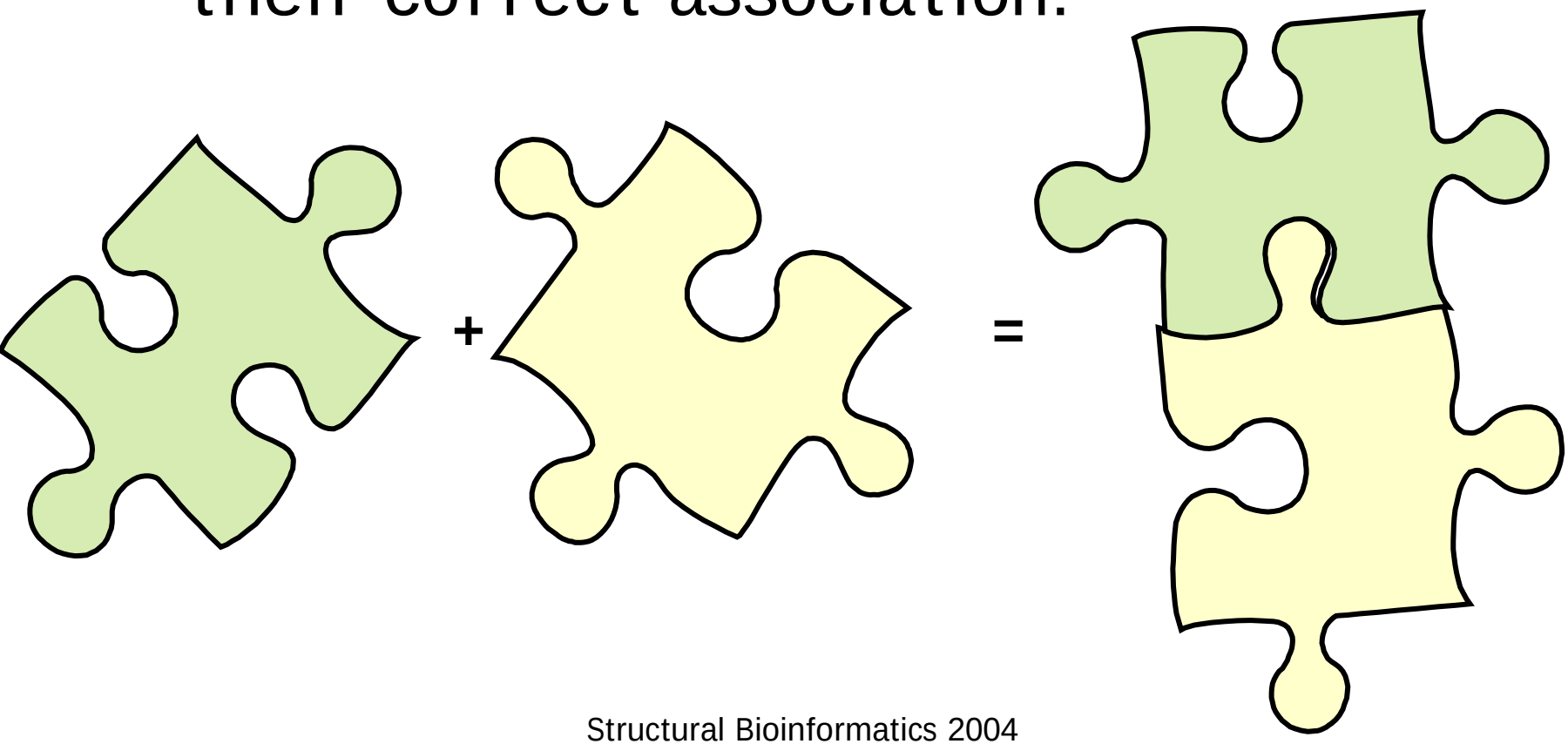
- **Predict association of protein molecules.**
- **Predict binding of a protein molecule with a potential drug.**
- **Scan libraries of drugs to detect a suitable inhibitor for a target molecule.**

Docking Algorithms

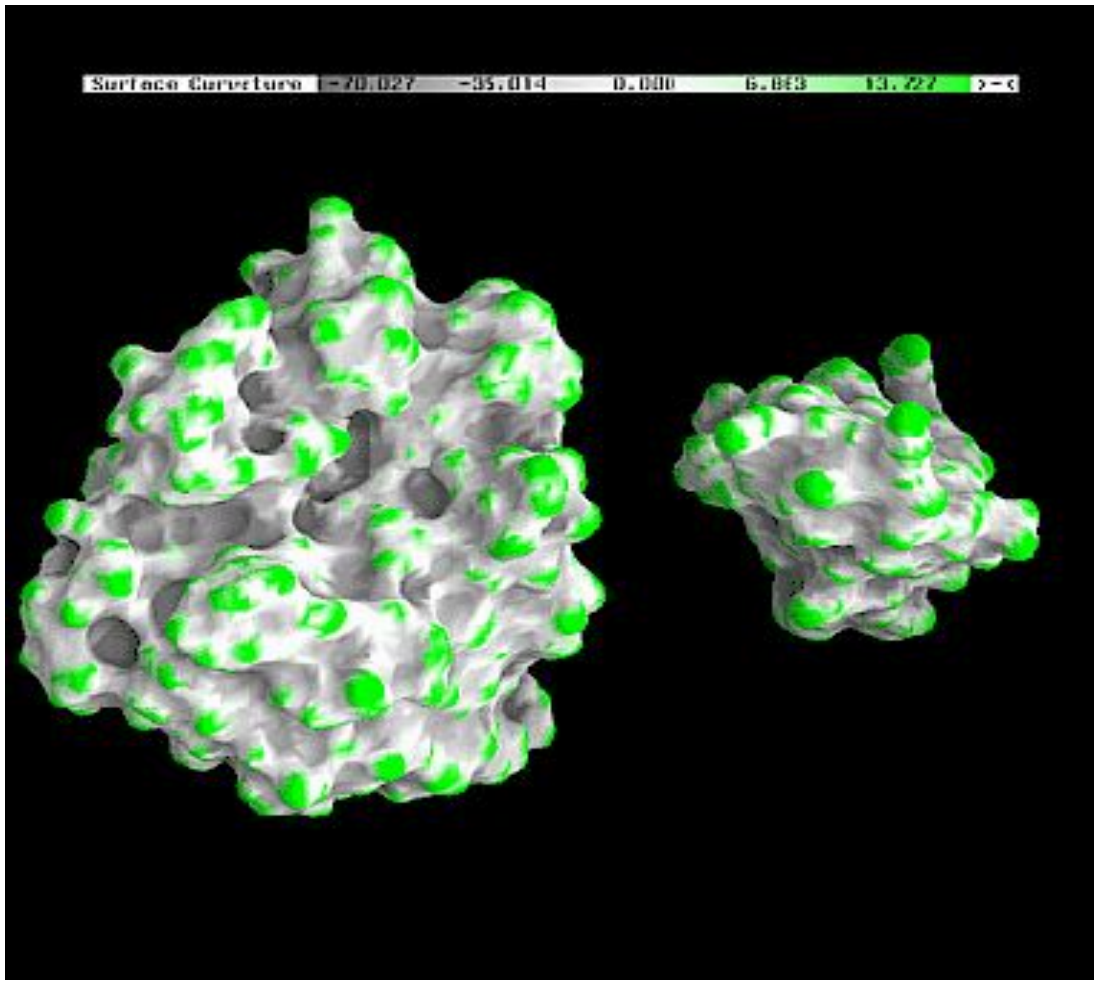
- **Rigid receptor-ligand and protein-protein docking**.
- **Flexible receptor-ligand docking allowing a small number of hinges either in the ligand or the** receptor.

Docking - Problem Definition

- Given a pair of molecules find their correct association:



Docking - Trypsin and BPTI



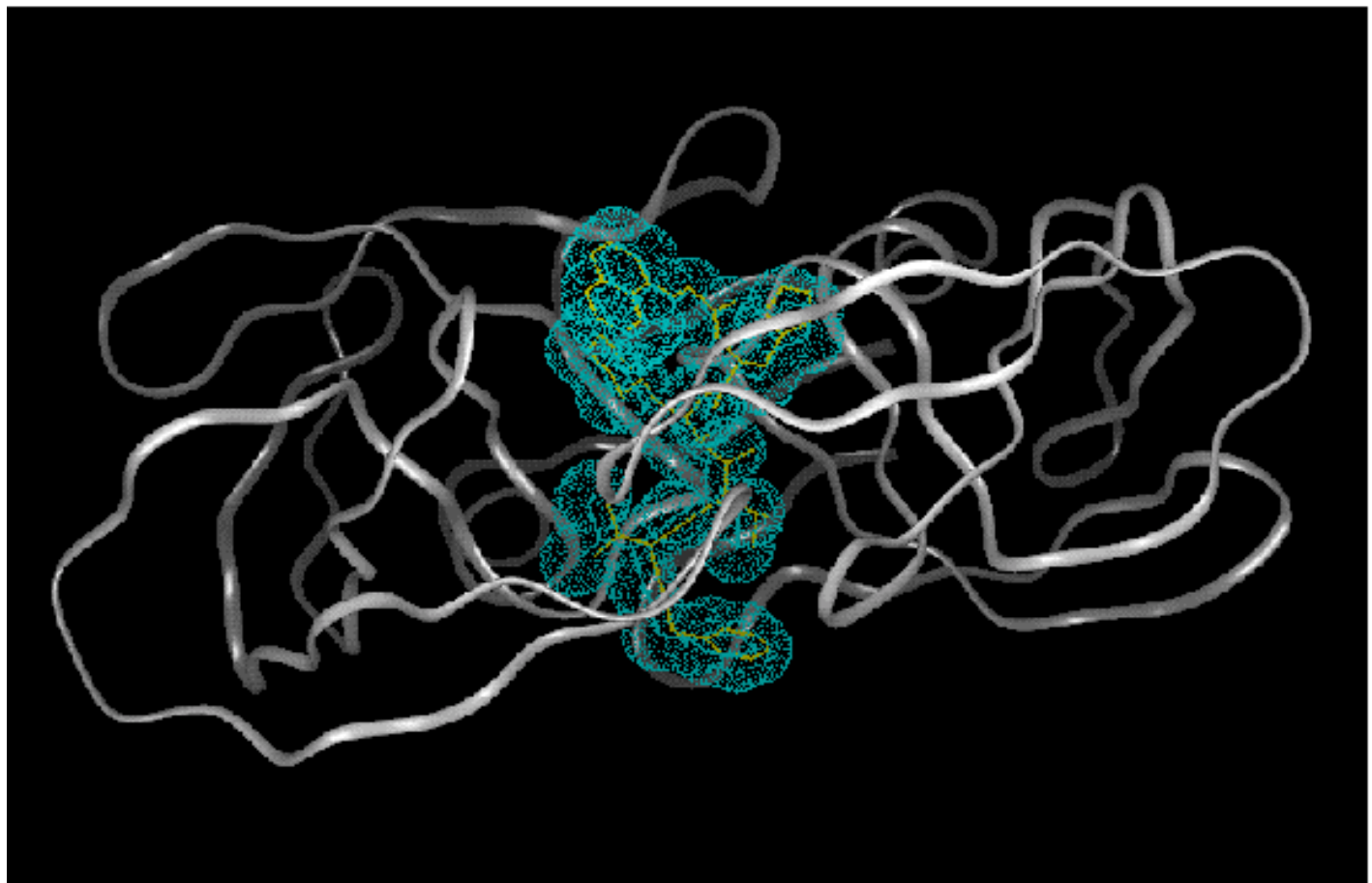
Docking - Relevance

- Computer aided drug design – a new drug should fit the active site of a specific receptor.
- Understanding of the biochemical pathways - many reactions in the cell occur through interactions between the molecules.
- Crystallizing large complexes and finding their structure is difficult.

Flexible Docking
Calmodulin with M13 ligand

Sandak, Nussinov, Wolfson - JCB 1998.

Flexible Docking
HIV Protease Inhibitor



Sandak, Nussinov, Wolfson - CABIOS 1995.

Software Infrastructure

- **Development of a software infrastructure for Geometric Computing in Molecular Biology.**
- **Object oriented, C++ library.**
- **Speed up development of new and re-usability of old software.**
- **Development of building blocks for fast testing of new ideas.**

Cross - fertilization 1

- **Analogous tasks appear in Computer Vision, Medical Imaging, Structural Bioinformatics, Target Recognition.**
- **Similar software and hardware can handle all of these *Geometric Computing* tasks - *method based cross fertilization*.**

Cross - fertilization 2

- **Bioinformatics brings together Computer Scientists, Molecular Biologists, Chemists etc. to tackle major problems in Computational Biology and Computer Assisted Drug Design - *task based cross-fertilization*.**

Conclusions 1

- **Molecular Biology and Biotechnology have entered a stage in which advanced algorithmic methods make the difference between theory and practice**.
- **Only true interdisciplinary collaboration among Computer and Life scientists can deliver biologically relevant computational techniques**.

Conclusions 2

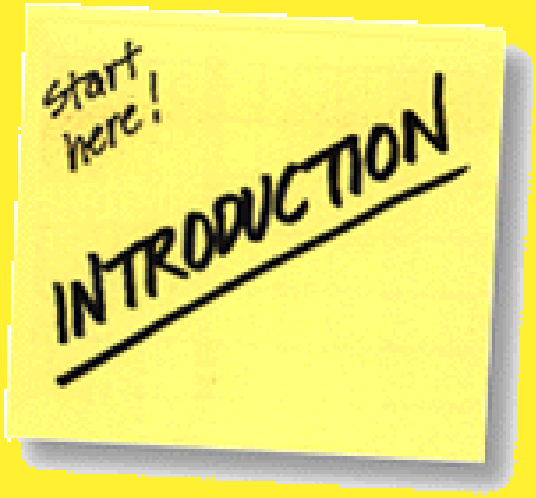
- **The b.c. (before Computer Science) algorithms in Computational Biology/Biotechnology, which have been mostly developed by chemists and physicists, are analogous to the first generation CS algorithms. The current state-of-the-art of CS (~fifth generation) provides a quantum leap.**

Sample of Topics to be covered

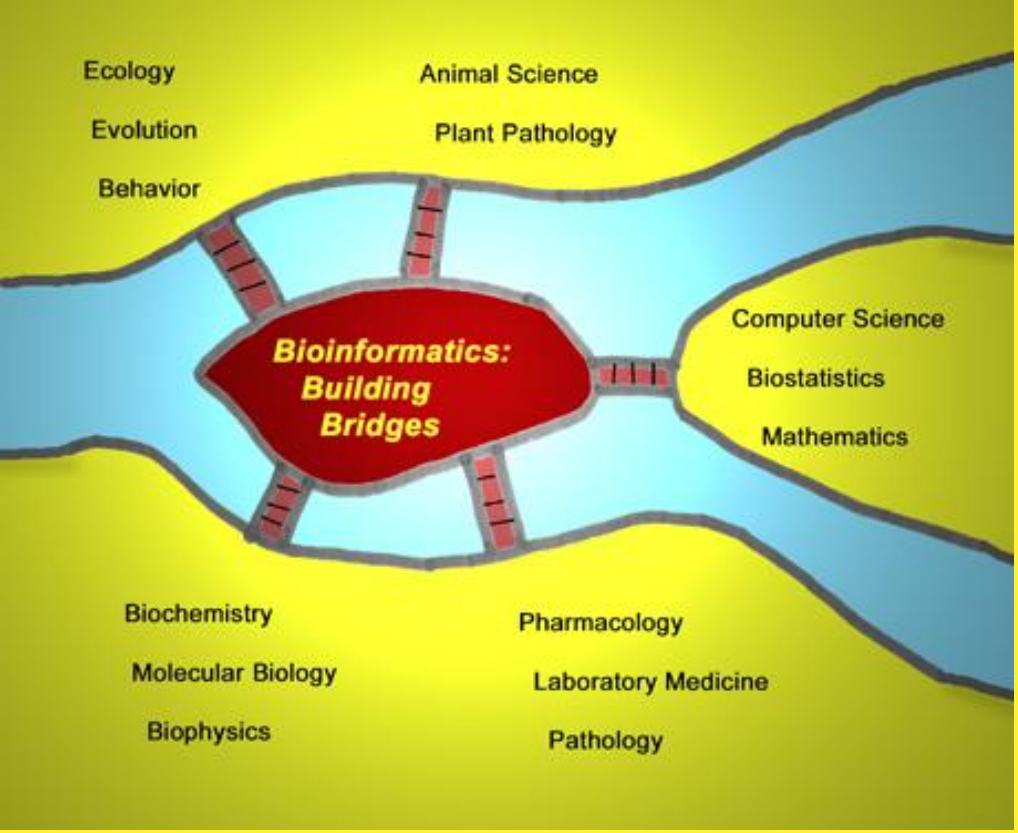
- Protein and DNA sequence alignment.
- Protein structural alignment and classification.
- Biomolecular recognition prediction – docking.
- Folding (homology modelling, threading, ab-initio).
- Distance Geometry for structure calculation from NMR data (?)
- Computer Assisted Structural Drug Design.

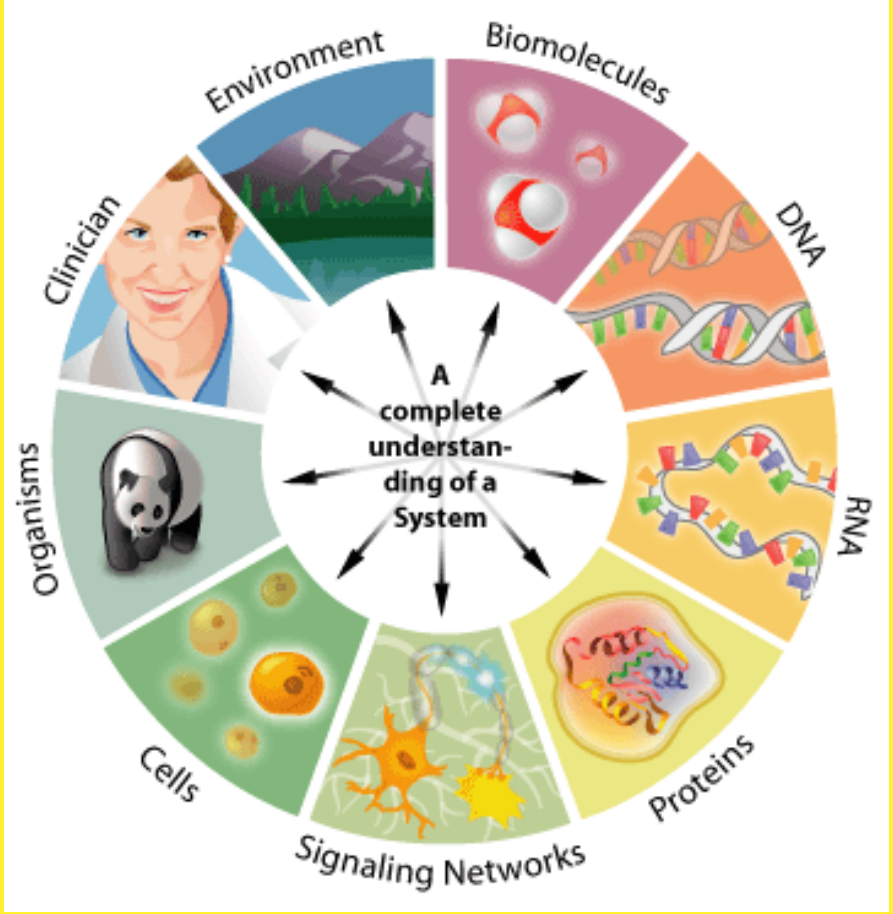
GRADING

- Exercises - 50%.
- Final (individual) Project, which involves heavy programming, based on the exercises – 50%.
- Most likely, all the students will get the same project assignment.
- The exact grading details will be supplied by the TA, Maxim Shatsky.



Introduction to Bioinformatics





What is Bioinformatics?

NIH – definitions

What is Bioinformatics? - Research, development, and application of computational tools and approaches for expanding the use of biological, medical, behavioral, and health data, including the means to acquire, store, organize, archive, analyze, or visualize such data.

on molecular

What is Computational Biology? The development and application of analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social data.

NSF – introduction

Large databases that can be accessed and analyzed with sophisticated tools have become central to biological research and education. The information content in the genomes of organisms, in the molecular dynamics of proteins, and in population dynamics, to name but a few areas, is enormous. Biologists are increasingly finding that the management of complex data sets is becoming a bottleneck for scientific advances. Therefore, **bioinformatics** is rapidly become a key technology in all fields of biology.

NSF – mission statement

The present bottlenecks in **bioinformatics** include the education of biologists in the use of advanced computing tools, the recruitment of computer scientists into this evolving field, the limited availability of developed databases of biological information, and the need for more efficient and intelligent search engines for complex databases.

NSF – mission statement

The present bottlenecks in **bioinformatics** include ***the education of biologists in the use of advanced computing tools***, the recruitment of computer scientists into this evolving field, the limited availability of developed databases of biological information, and the need for more efficient and intelligent search engines for complex databases.

Molecular Bioinformatics

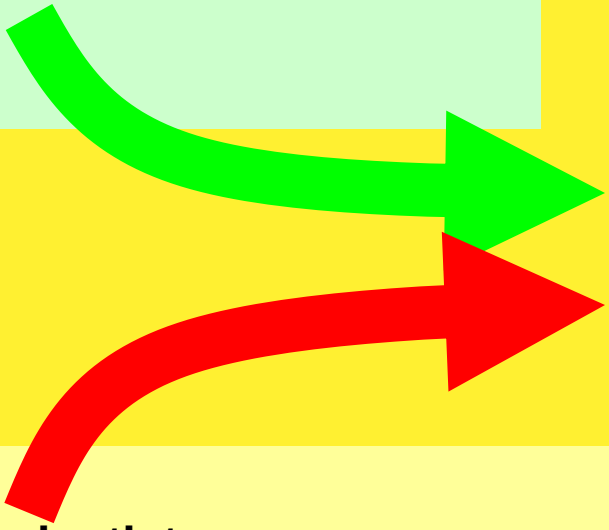
Molecular Bioinformatics involves the use of computational tools to discover new information in complex data sets (from the one-dimensional information of DNA through the two-dimensional information of RNA and the three-dimensional information of proteins, to the four-dimensional information of evolving living systems).

Bioinformatics (Oxford English Dictionary):

The branch of science concerned with information and information flow in biological systems, esp. the use of computational methods in genetics and genomics.

The field of science in which **biology**, **computer science** and **information technology** merge into a single discipline

Biologists
collect molecular data:
DNA & Protein sequences,
gene expression, etc.



Bioinformaticians
Study biological questions by analyzing molecular data

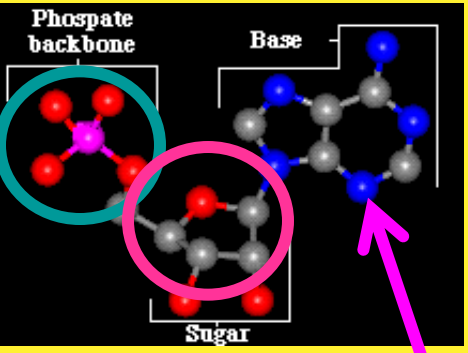
Computer scientists
(+Mathematicians, Statisticians, etc.)
Develop tools, softwares, algorithms
to store and analyze the data.

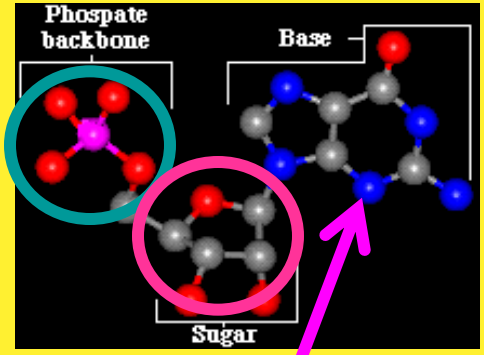
Some biological background….

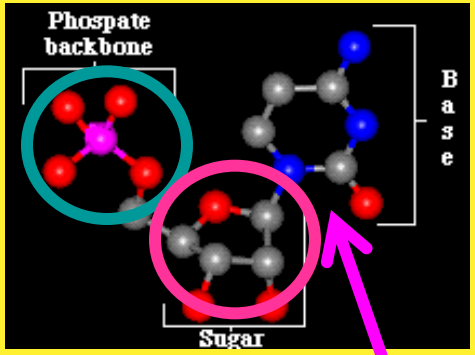
A biologist

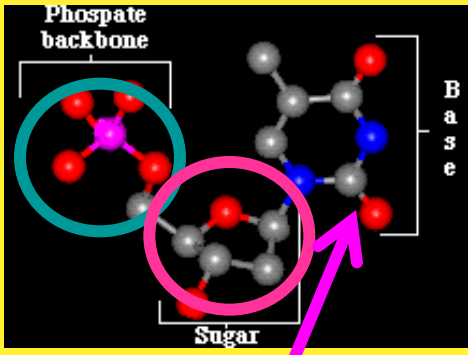
The hereditary information of all living organisms, with the exception of some viruses, is carried by **deoxyribonucleic acid (DNA)** molecules.

2 purines: 2 pyrimidines:









adenine (A) guanine (G) cytosine (C) thymine (T)

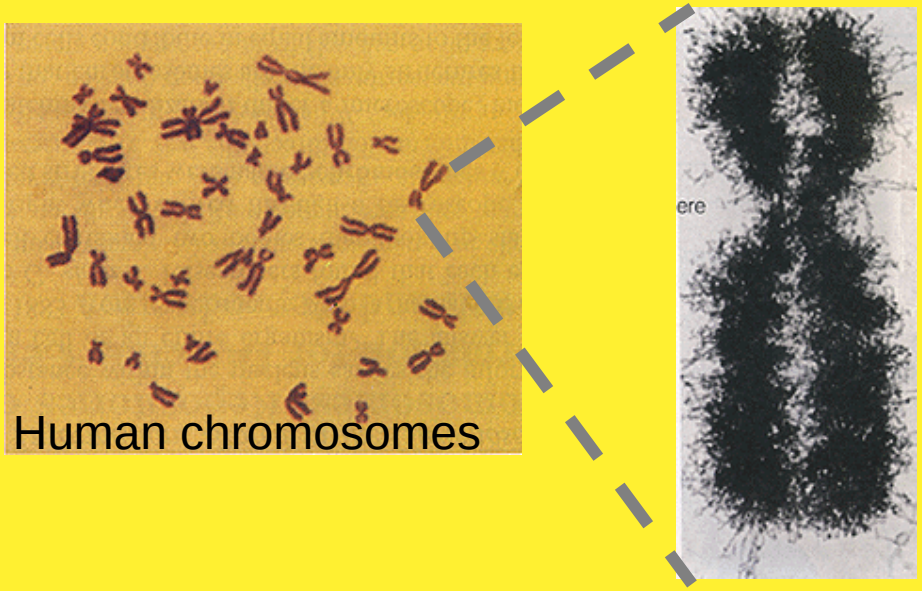
two rings

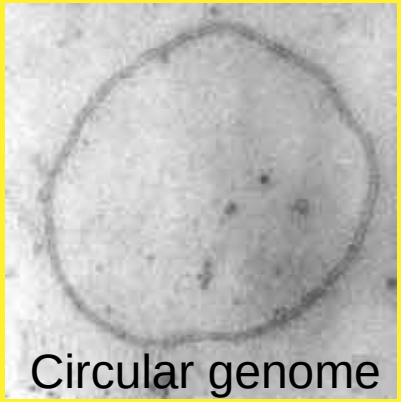
one ring

The entire complement of genetic material carried by an individual is called the **genome**.

Eukaryotes may have up to 3 subcellular genomes:
1. Nuclear
2. Mitochondrial
3. Plastid

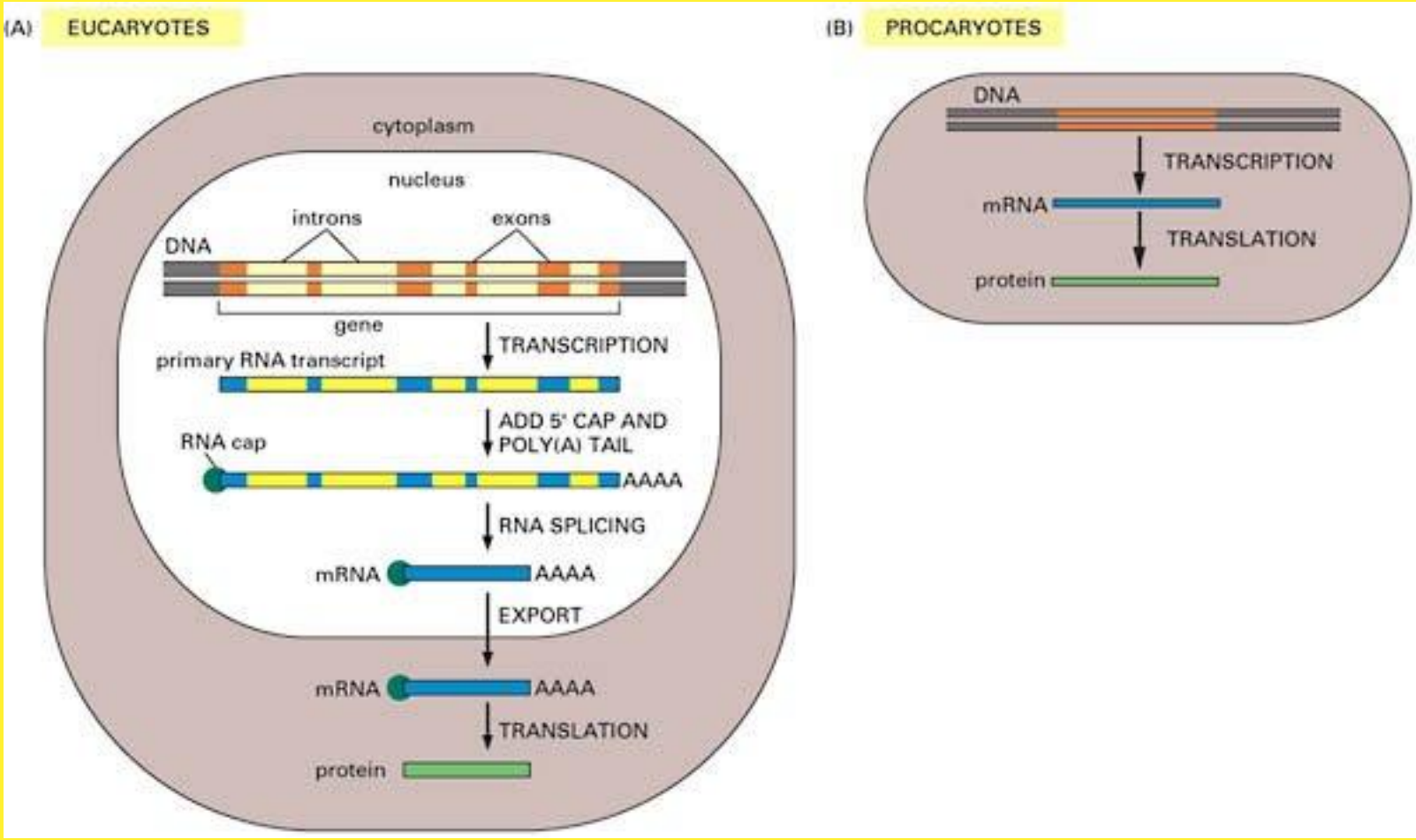
Bacteria have either circular or linear genomes and may also carry plasmids


Human chromosomes

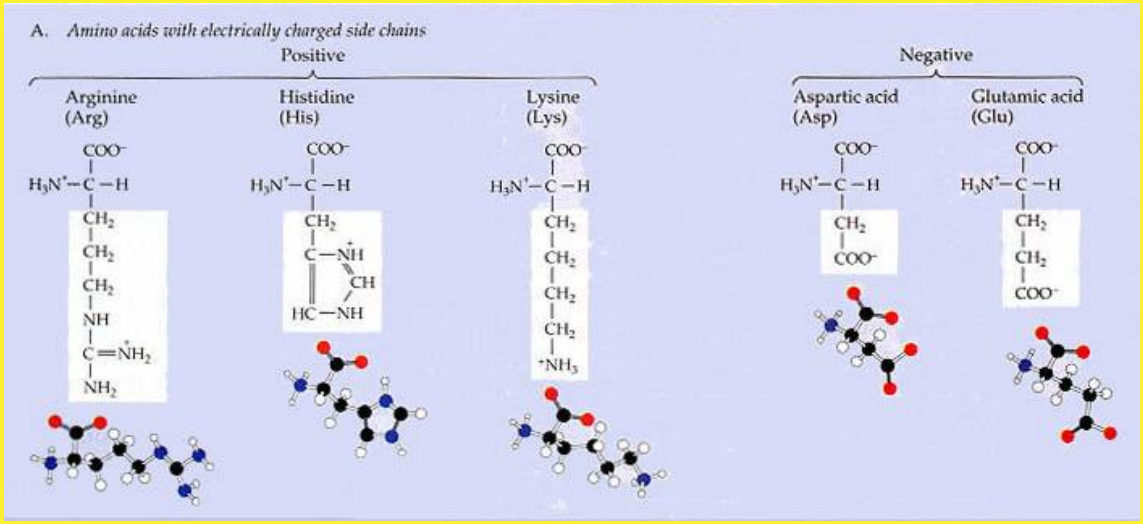

Circular genome

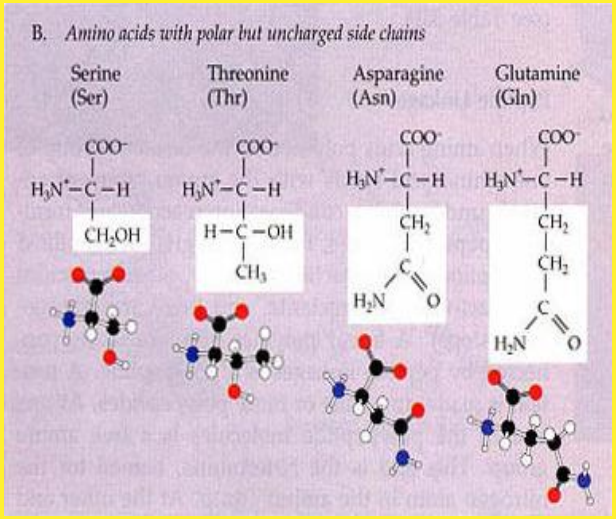
Central dogma: DNA makes RNA makes Protein

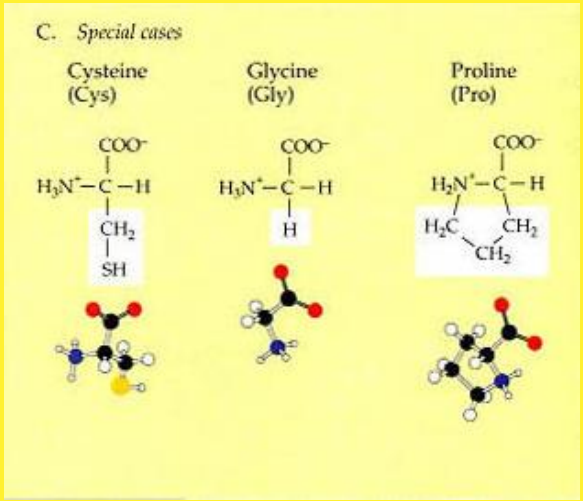
Modified dogma: DNA makes DNA and RNA, RNA makes DNA, RNA an Protein

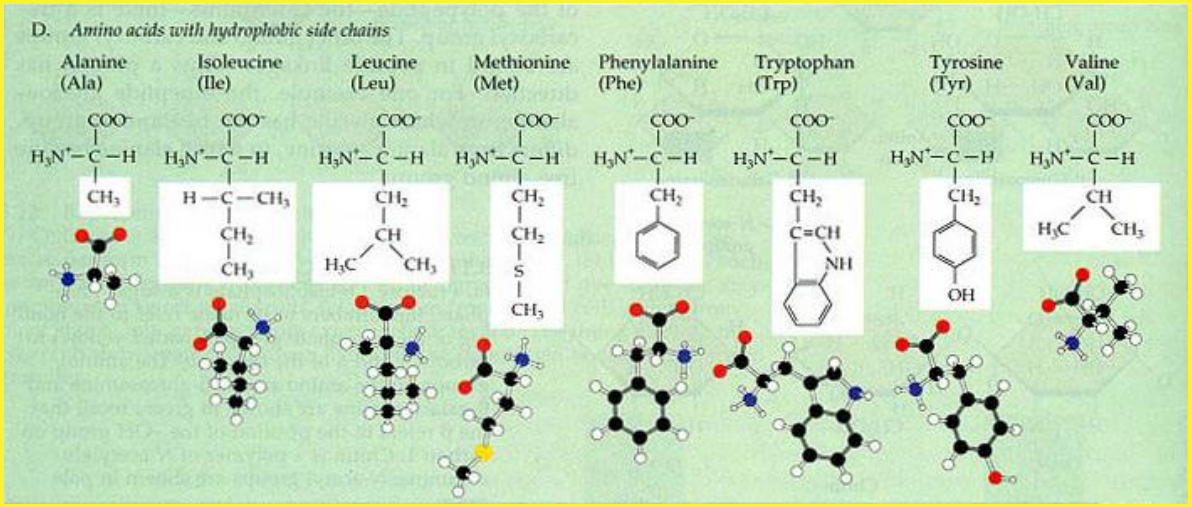


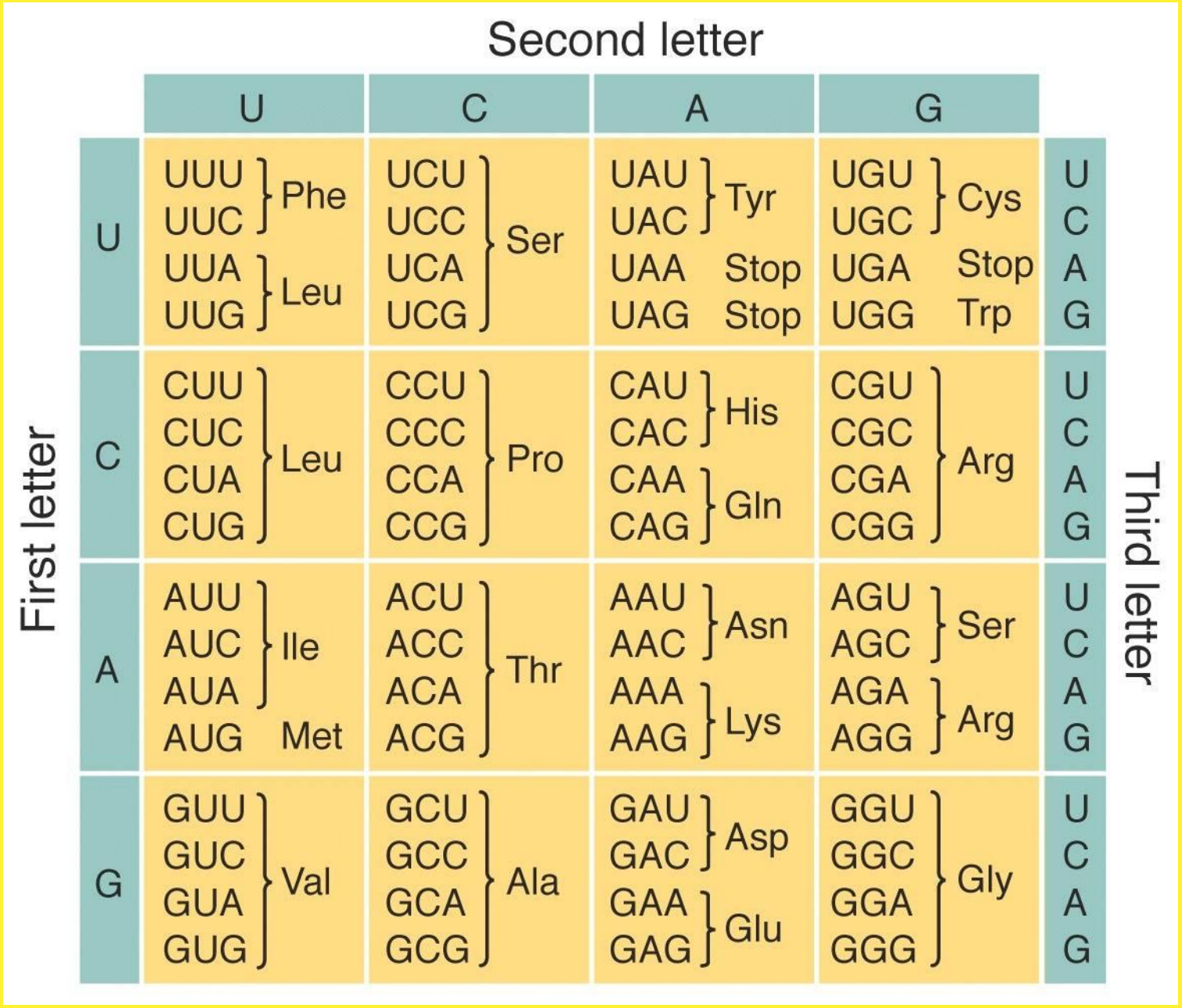
Amino acids - The protein building blocks



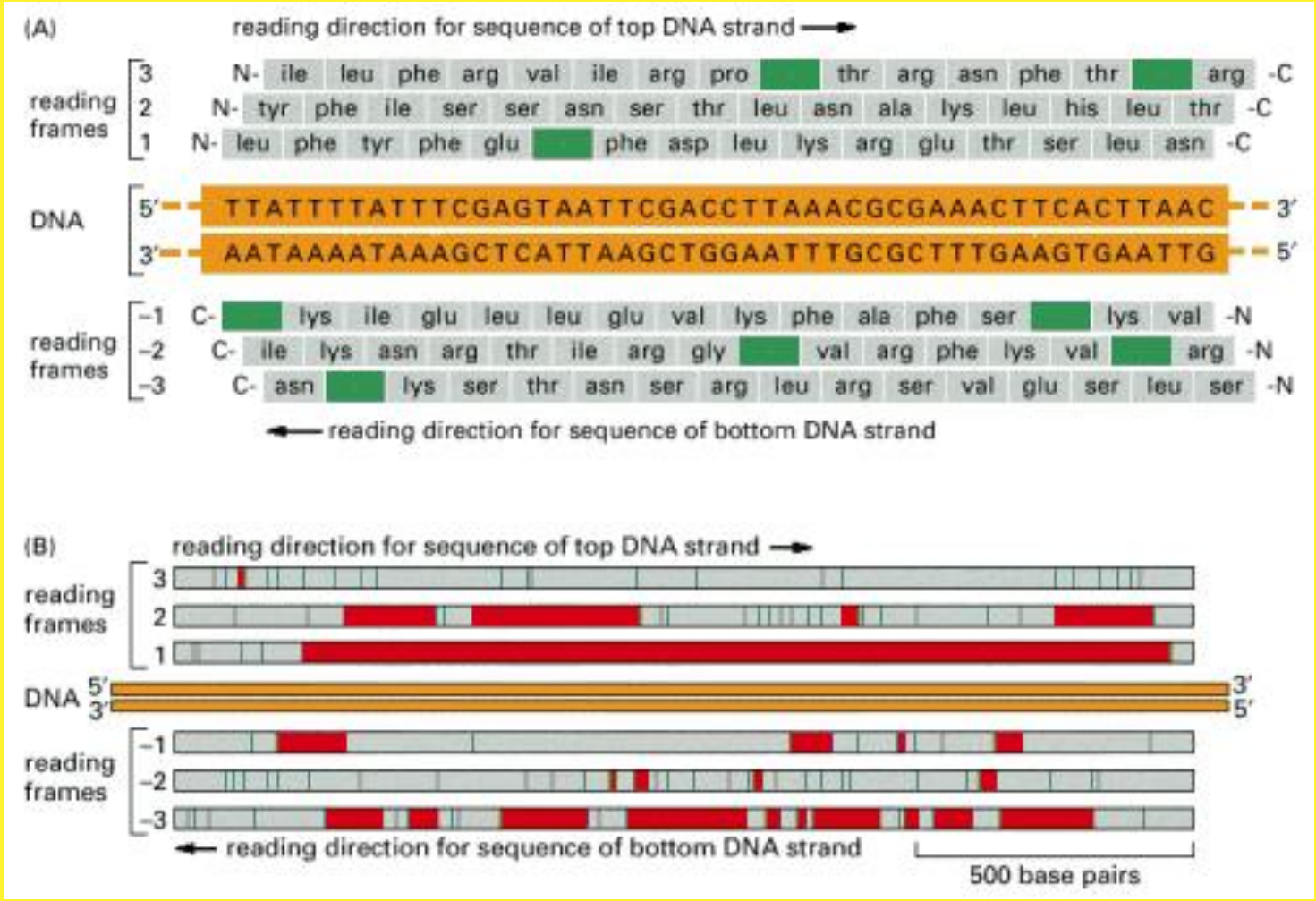







Second letter
U
C
A
G
First letter
Third letter
U
UUU Phe
UUC
UUA Leu
UUG
UCU Ser
UCC
UCA
UCG
UAU Tyr
UAC
UAA Stop
UAG Stop
UGU Cys
UGC
UGA Stop
UGG Trp
C
CUU Leu
CUC
CUA
CUG
CCU Pro
CCC
CCA
CCG
CAU His
CAC
CAA Gln
CAG
CGU Arg
CGC
CGA
CGG
A
AUU Ile
AUC
AUA
AUG Met
ACU Thr
ACC
ACA
ACG
AAU Asn
AAC
AAA Lys
AAG
AGU Ser
AGC
AGA Arg
AGG
G
GUU Val
GUC
GUA
GUG
GCU Ala
GCC
GCA
GCG
GAU Asp
GAC
GAA Glu
GAG
GGU Gly
GGC
GGA
GGG

Any region of the DNA sequence can, in principle, code for six different amino acid sequences, because any one of three different reading frames can be used to interpret each of the two strands.



Protein folding

A human Hemoglobin:

How does it all looks like on a computer monitor?

A cDNA sequence

>gi|14456711|ref|NM_000558.3| **Homo sapiens hemoglobin, alpha** 1 (HBA1), mRNA
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAA
CGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAG
AGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCT
GCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGG
ACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCG
GTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAA
TACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTC
CCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC

A cDNA sequence (reading frame)

>gi|14456711|ref|NM_000558.3| **Homo sapiens hemoglobin, alpha** 1 (HBA1), mRNA

ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCA
ACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGA
GAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTC
TGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTG
GACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCC
GGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTT
CACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAA
ATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCC
TCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC

A protein sequence

>gi|4504347|ref|NP_000549.1| **alpha 1 globin [Homo sapiens]**

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVA
DALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLA
SVSTVLTSKYR

And, a whole genome…

ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTC
AAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTC
CTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGG
GCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCG
CTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCC
ACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGG
ACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATG
CTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAAT
AAAGTCTGAGTGGGCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCT
GCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAG
GCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACG
GCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTG
GACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTC
AACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTG
CGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCT
GGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGT
ACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACC
CACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGC
TGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCG
CACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACC
AACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAG
CTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCC
CCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGAC
CTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCC
TCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGCGCCGTGGCGCACGTG
GACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTC
AACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTG
CGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCT
GGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGT

How big are whole genomes?

E. coli 4.6 x 10^6 nucleotides

- Approx. 4,000 genes

Yeast 15 x 10^6 nucleotides

- Approx. 6,000 genes

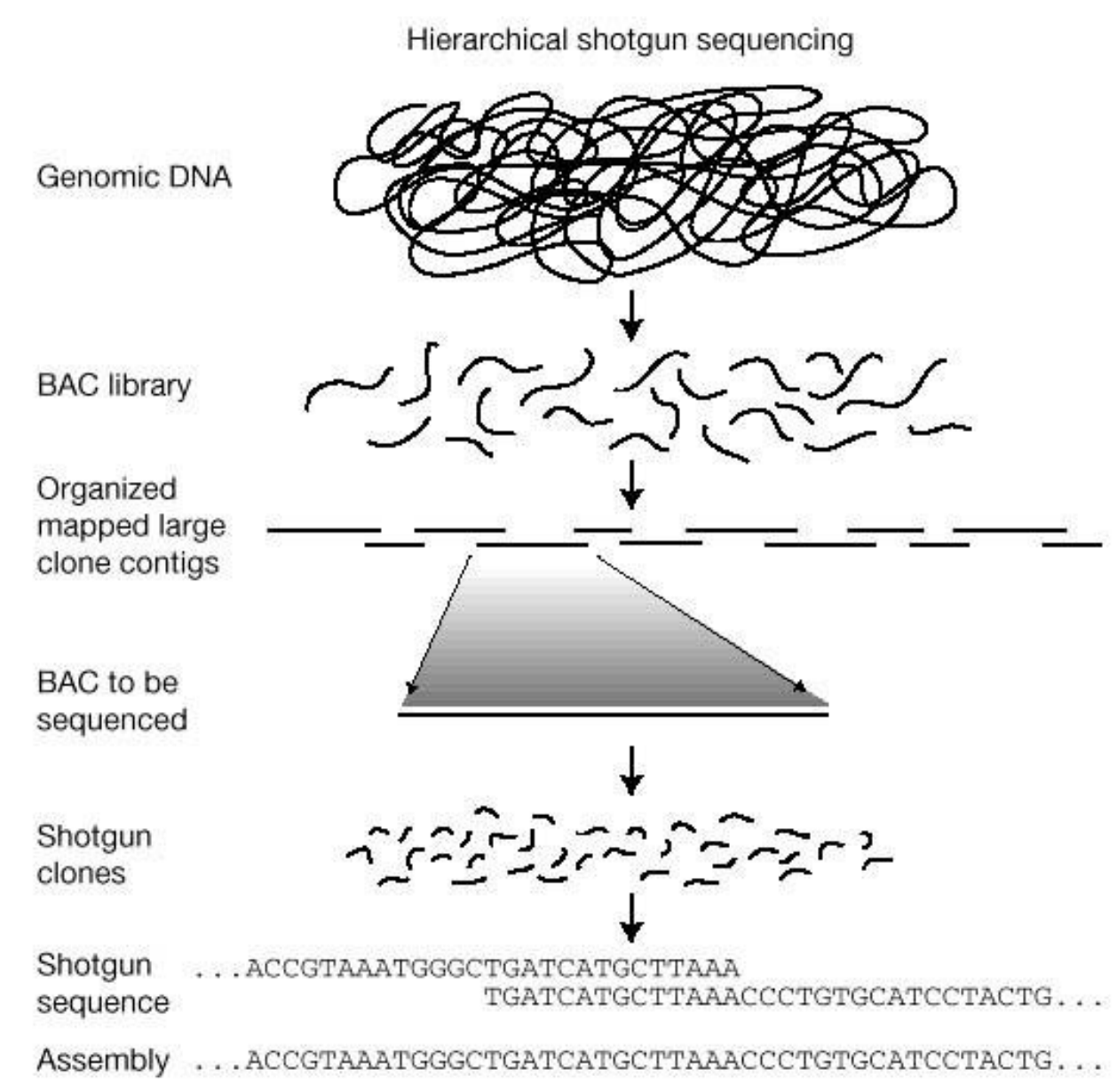
Human 3 x 10^9 nucleotides

- Approx. 30,000 genes

Smallest human chromosome 50 x 10^6 nucleotides

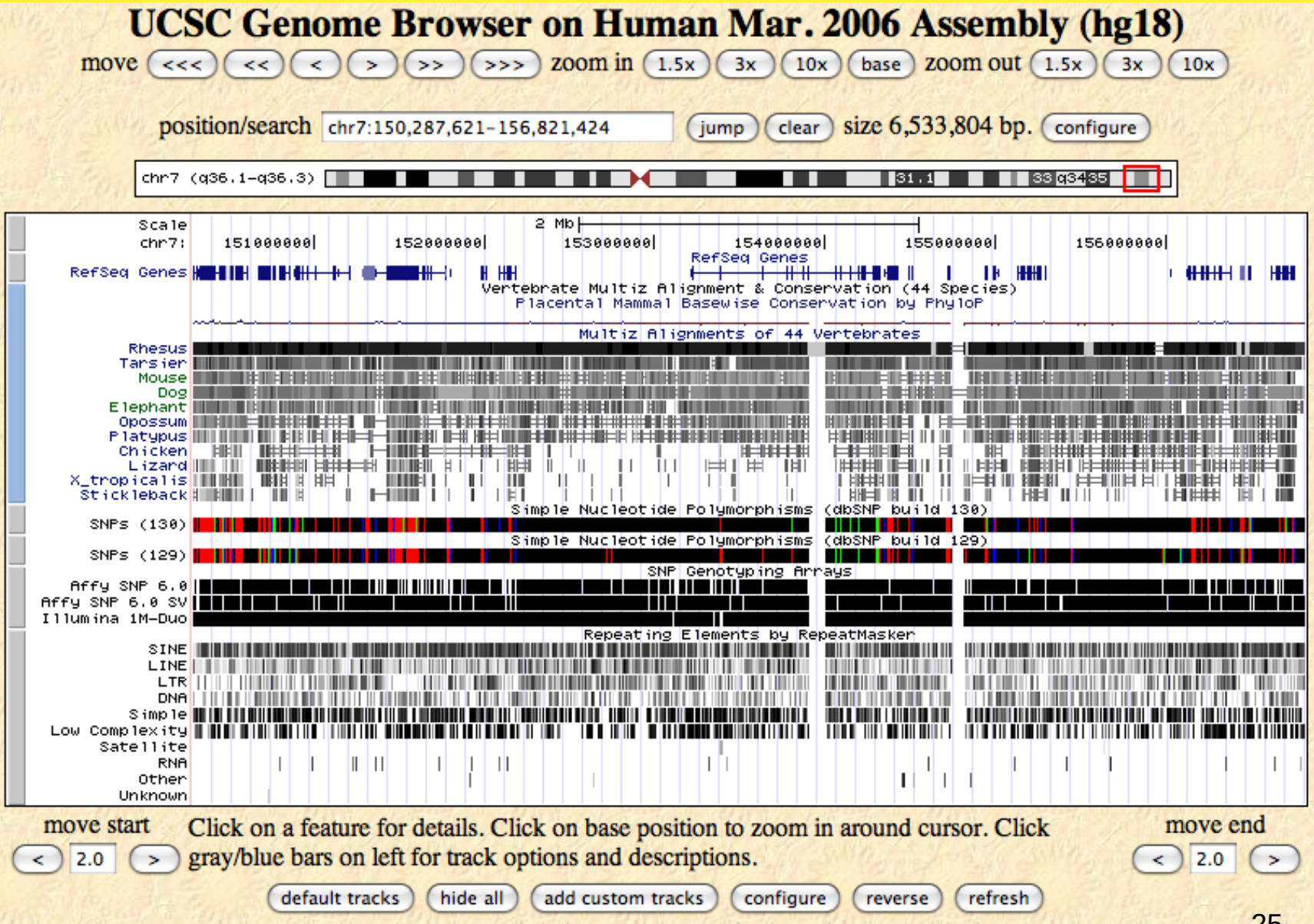
What do we actually do with bioinformatics?

Sequence assembly



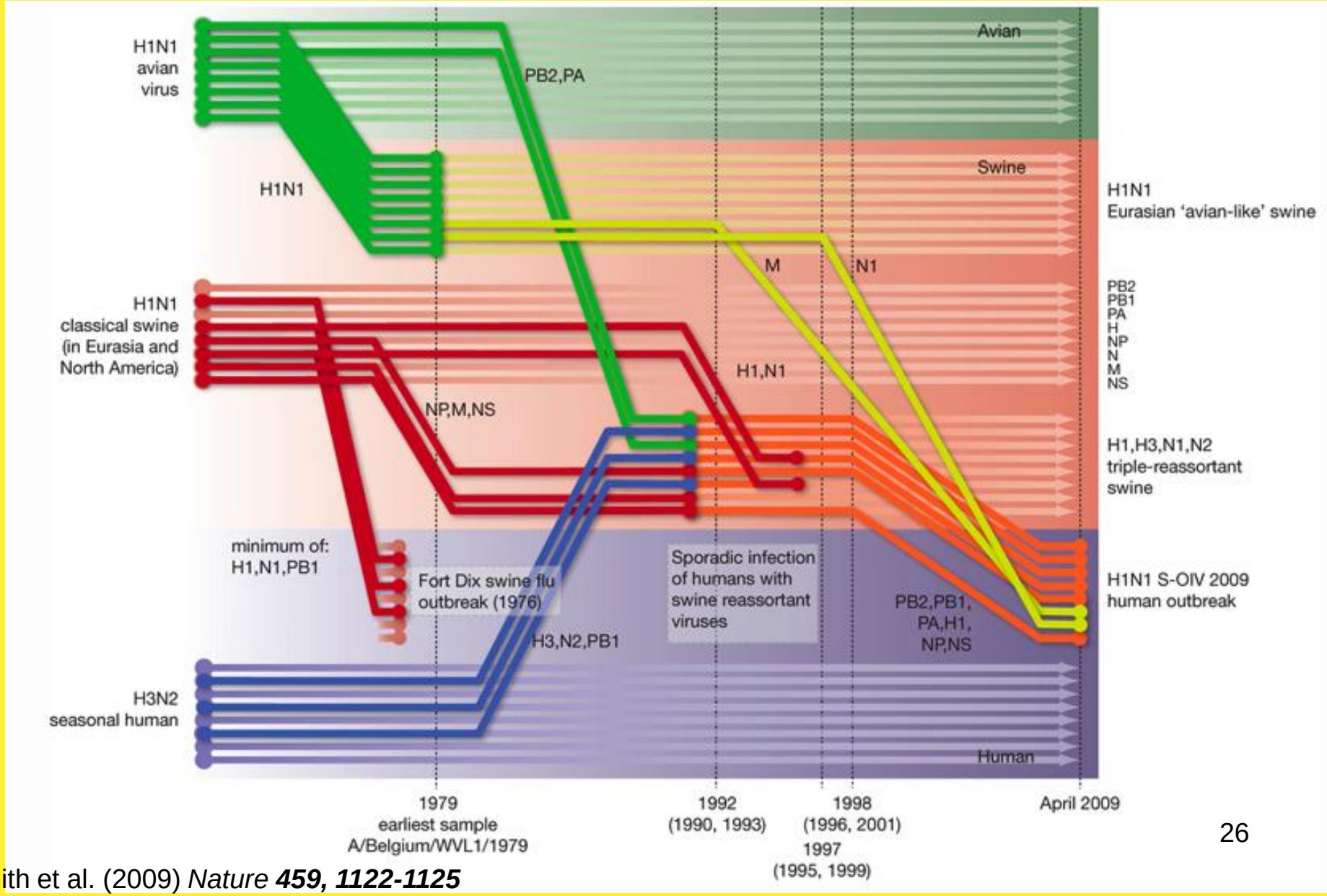
(next generation sequencing)

Genome annotation



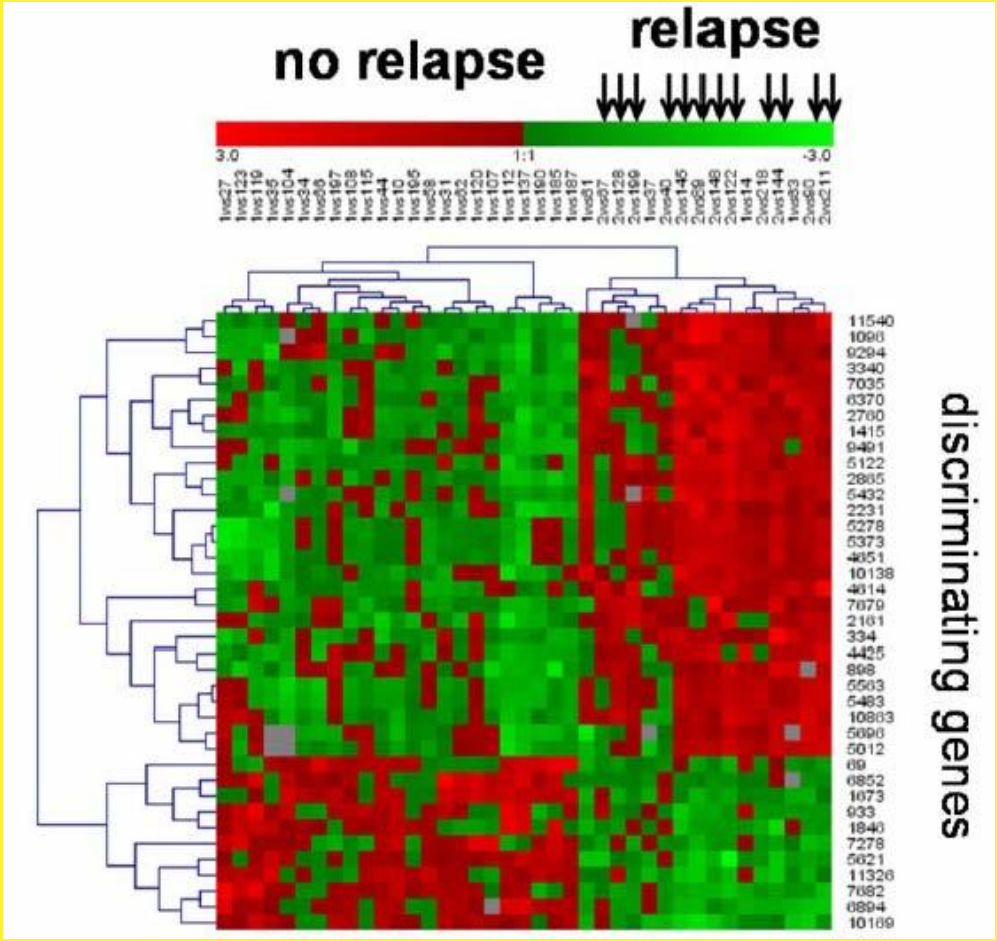
Molecular evolution

Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic



Smith et al. (2009) *Nature* ***459, 1122-1125***

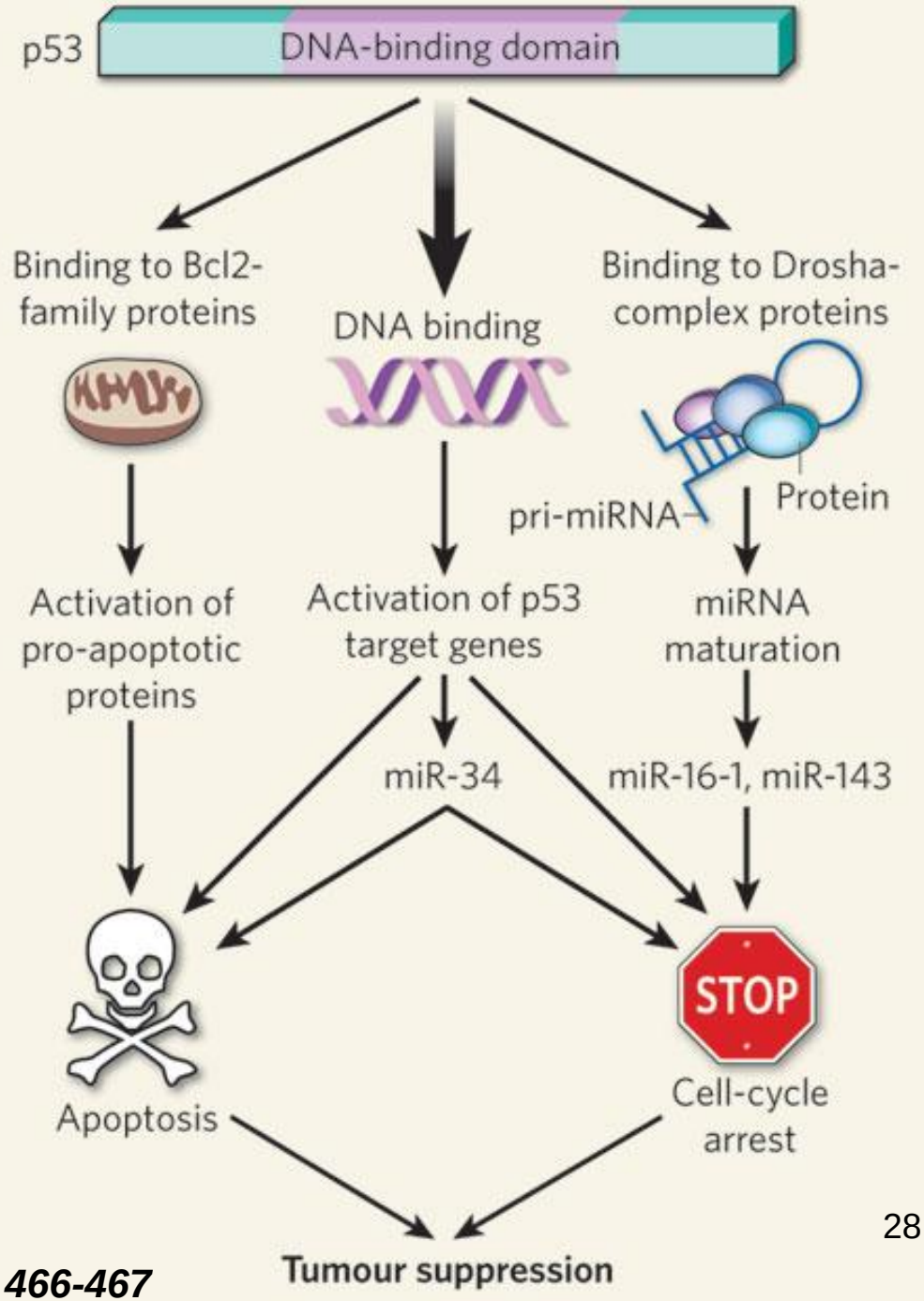
Analysis of gene expression



Gene expression profile of relapsing versus non-relapsing Wilms tumors. A set of 39 genes discriminates between the two classes of tumors.

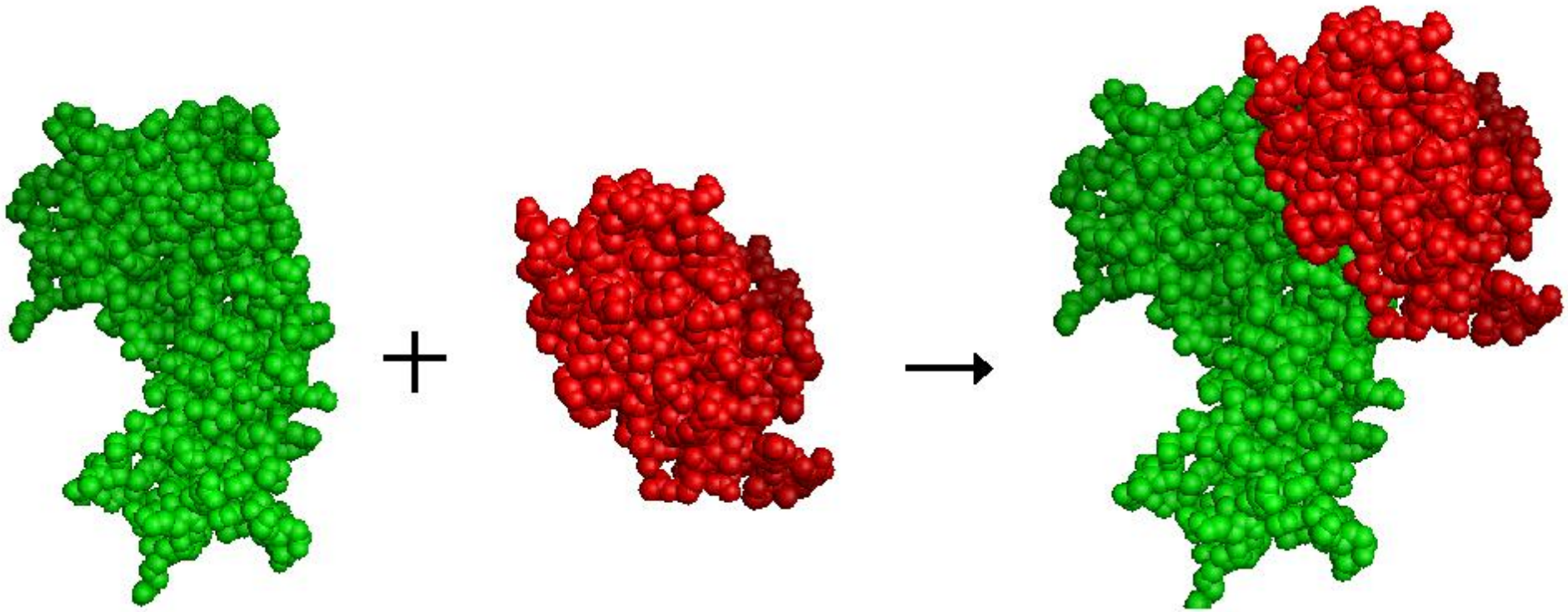
(http://www.biozentrum2.uni-wuerzburg.de/, Prof. Gessler)

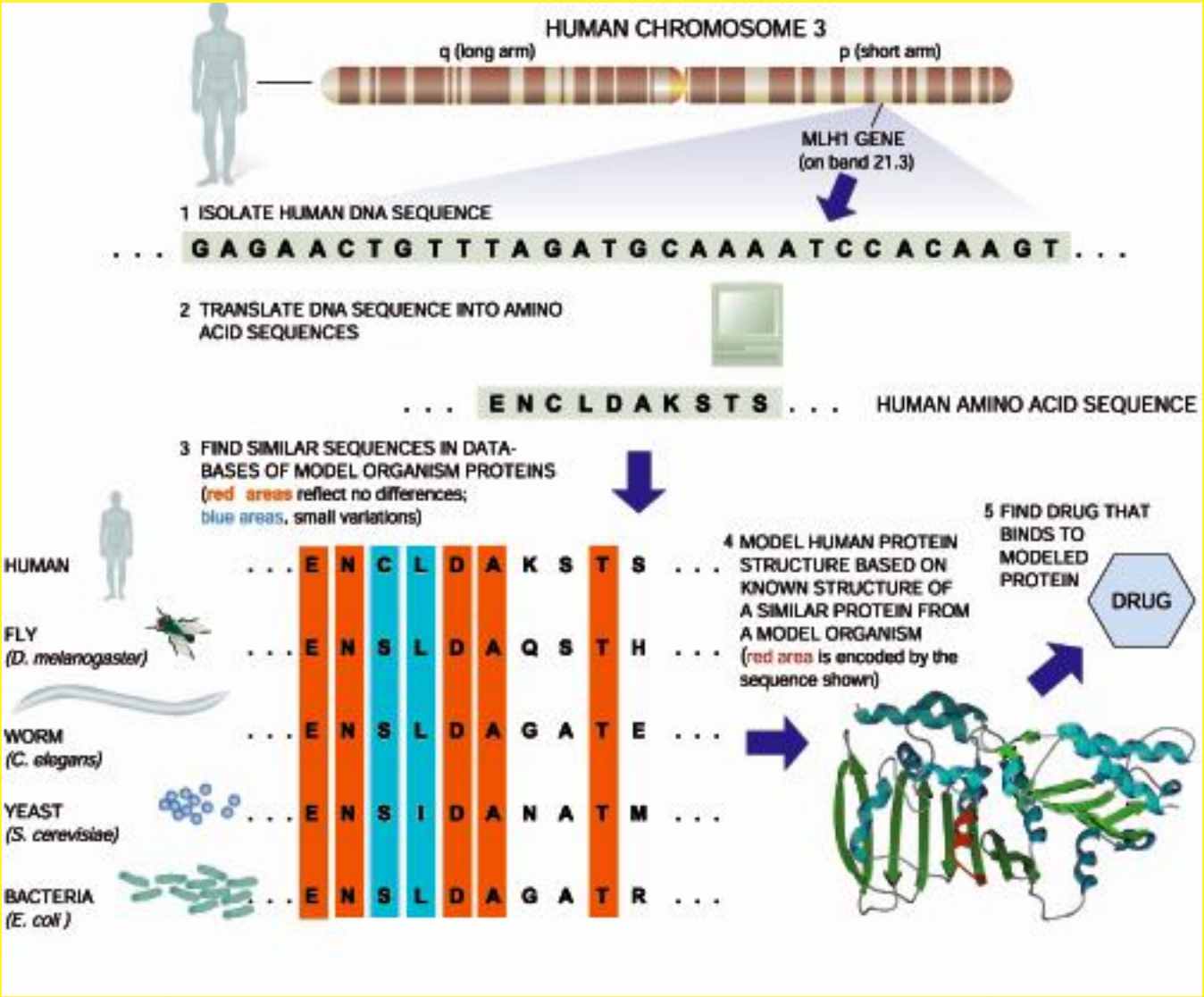
Analysis of regulation



Toledo and Bardot (2009) *Nature* ***460, 466-467***

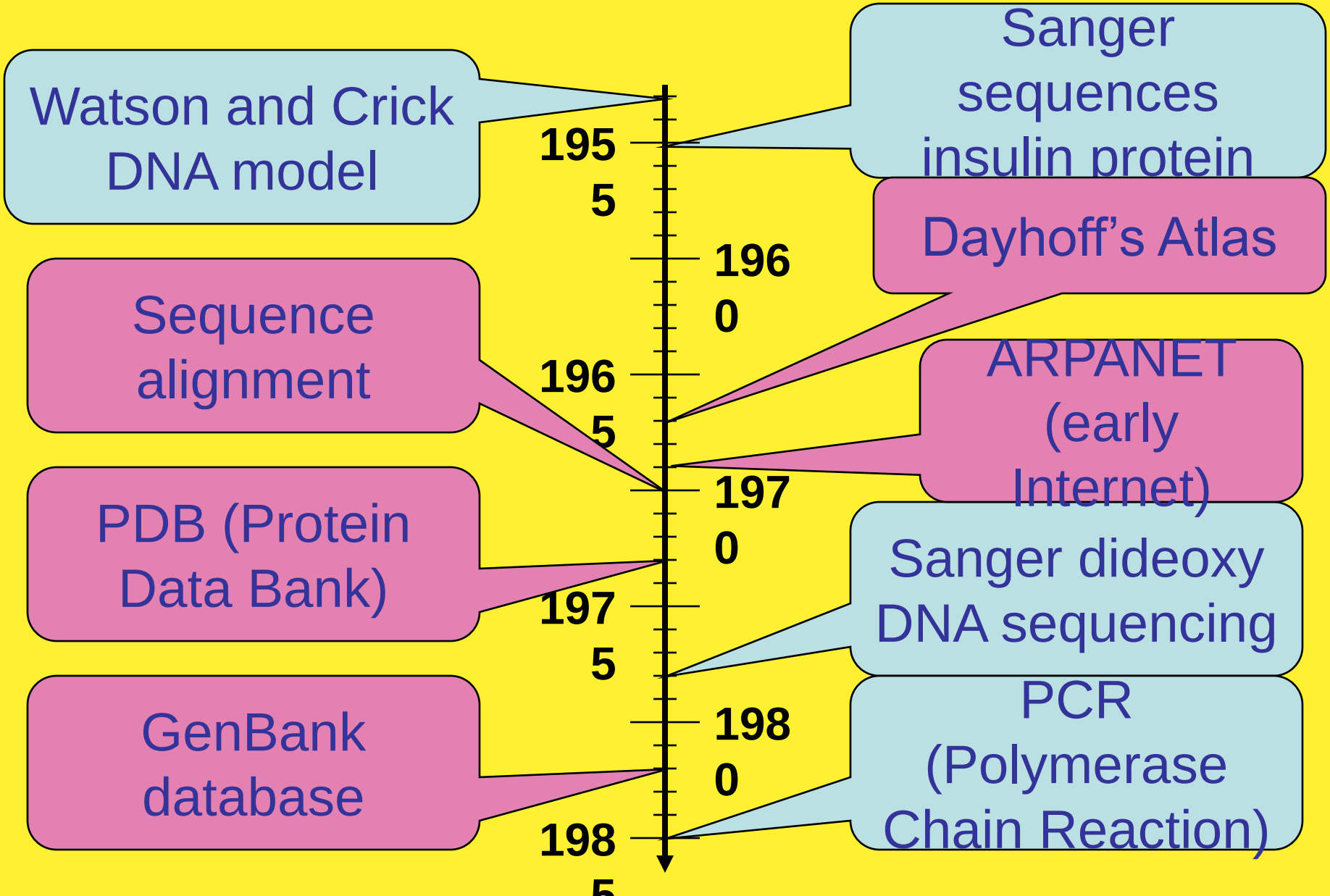
Protein structure prediction
Protein docking

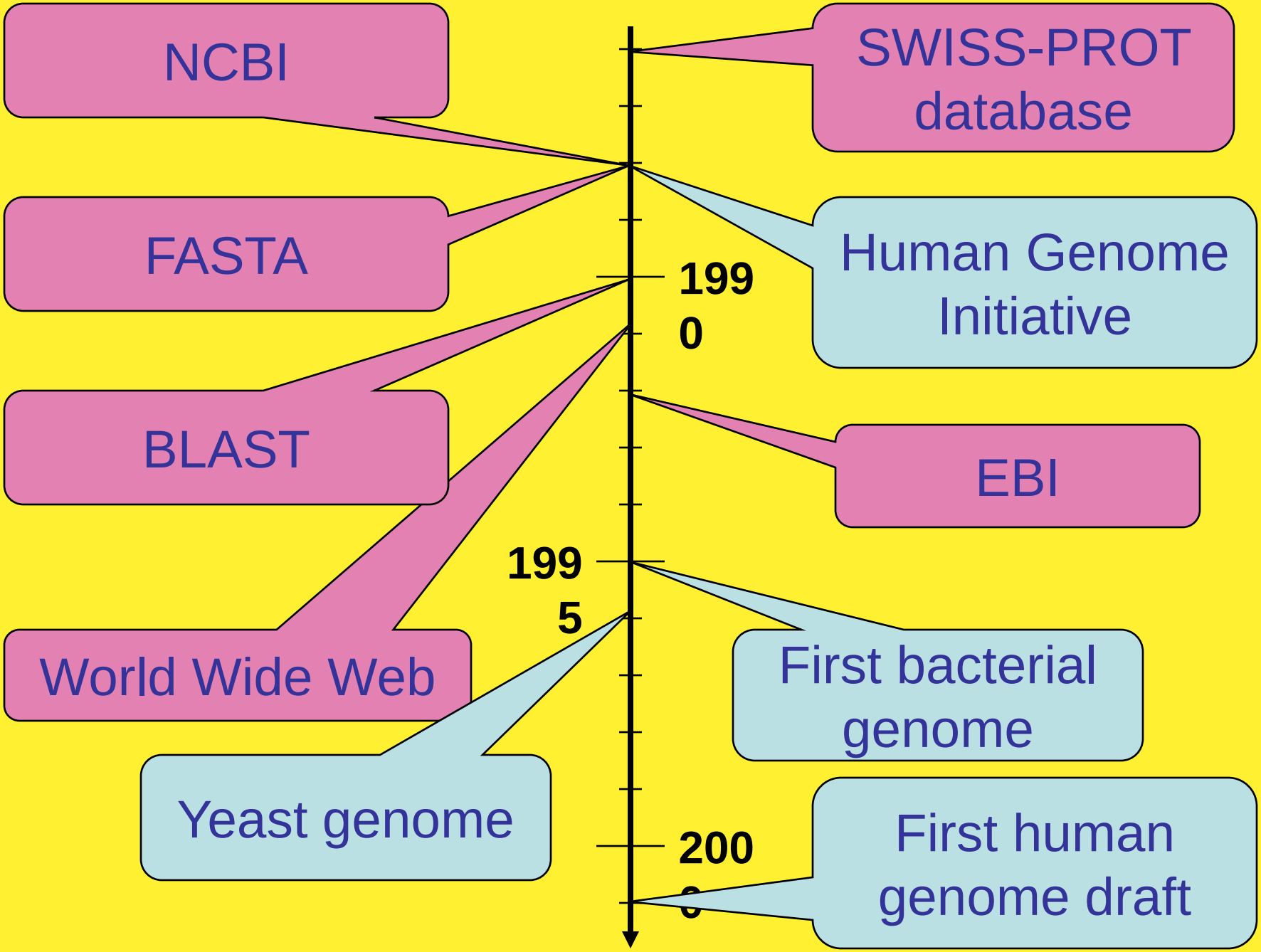




Luscombe, Greenbaum, Gerstein (2001)

From DNA to Genome




NCBI
SWISS-PROT database
FASTA
Human Genome Initiative
1990
BLAST
EBI
1995
World Wide Web
First bacterial genome
Yeast genome
2000
First human genome draft

Origin of bioinformatics and biological databases:

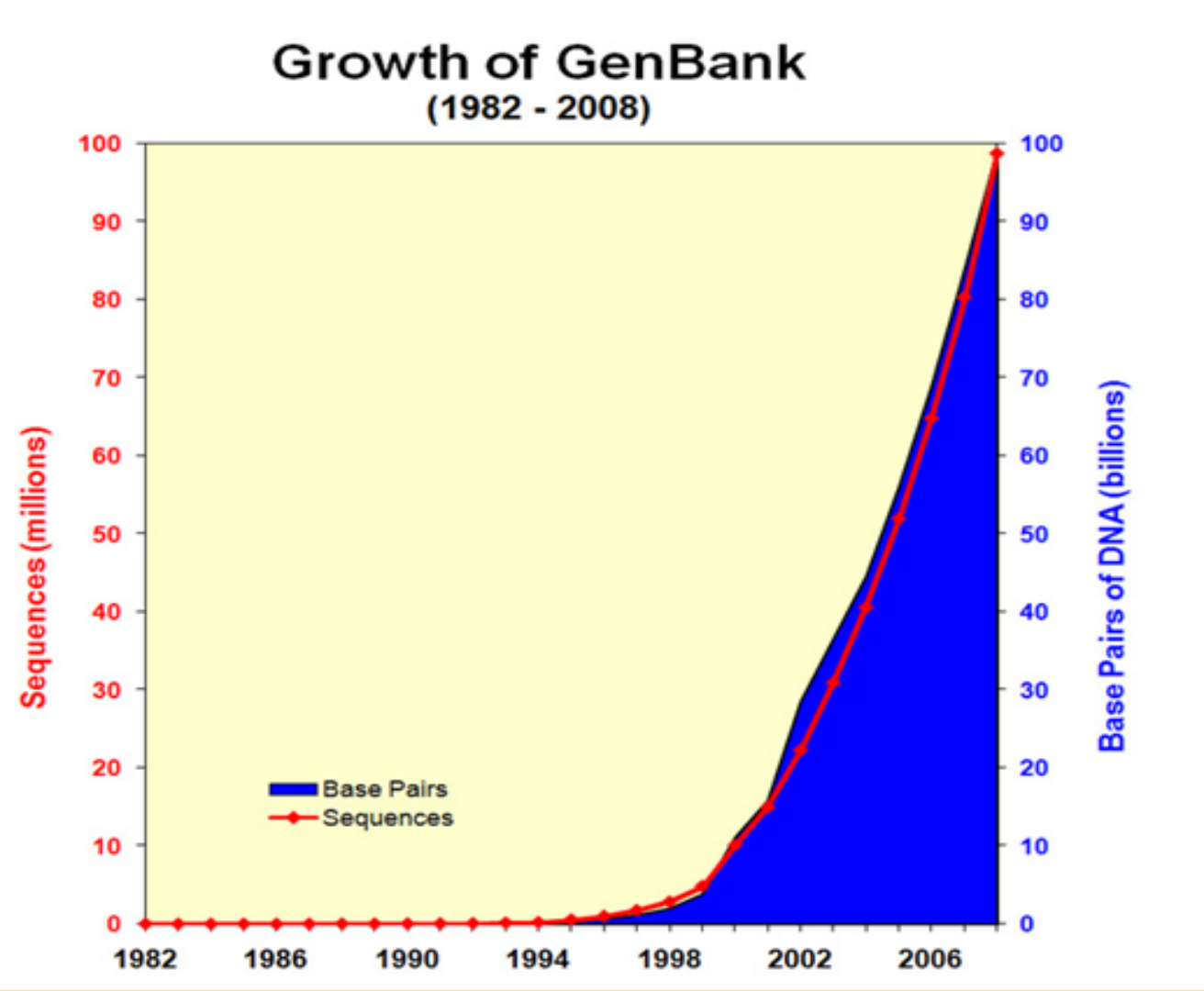
The first protein sequence reported was that of bovine insulin in **1956**, consisting of 51 residues.

Nearly a decade later, the first nucleic acid sequence was reported, that of yeast $tRNA^{alanine}$ with 77 bases.

In 1965, **Dayhoff** gathered all the available sequence data to create the first bioinformatic database (*Atlas of Protein Sequence and Structure*).

The Protein DataBank followed in 1972 with a collection of ten X-ray crystallographic protein structures. The SWISSPROT protein sequence database began in 1987.

Nucleotides

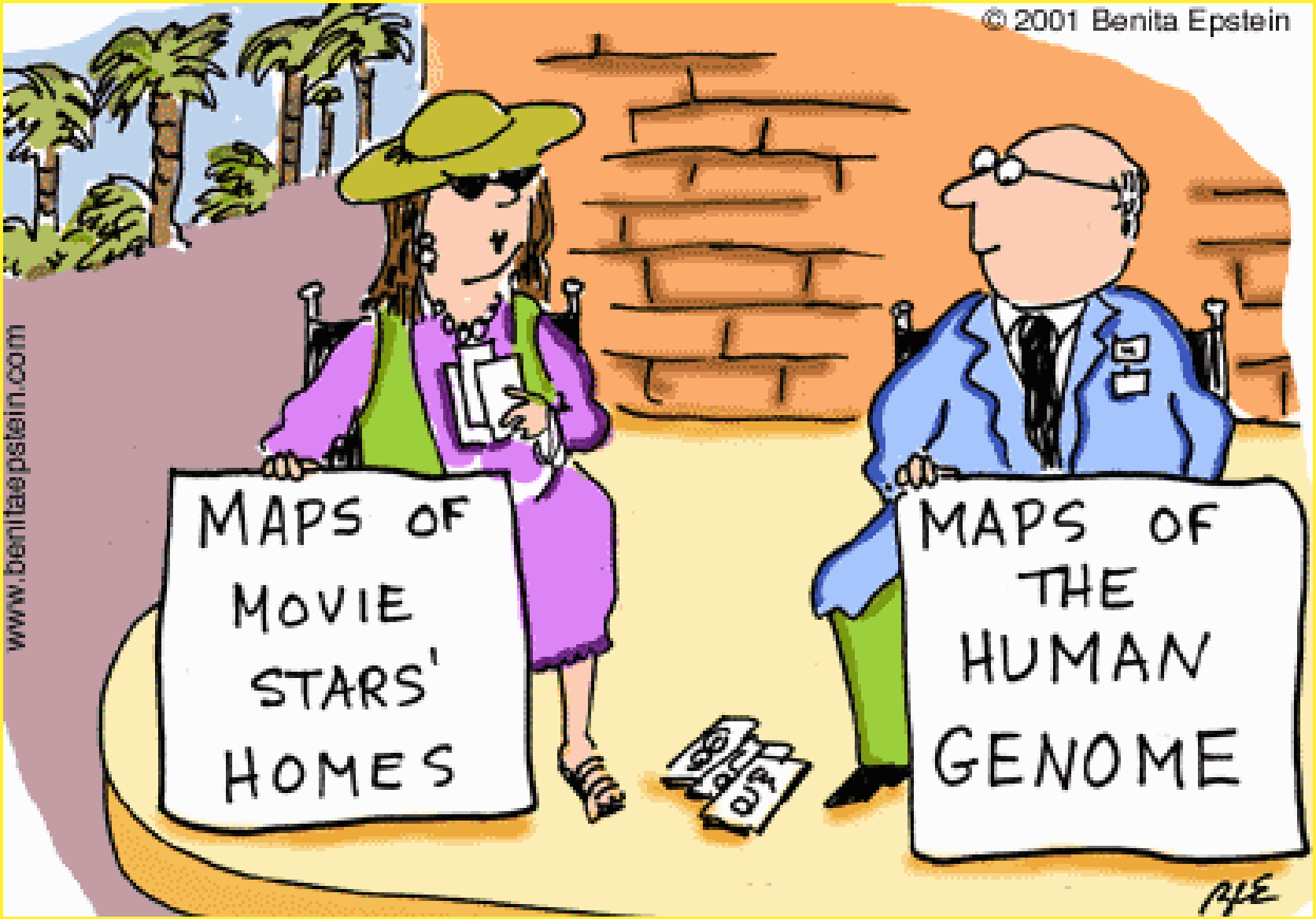


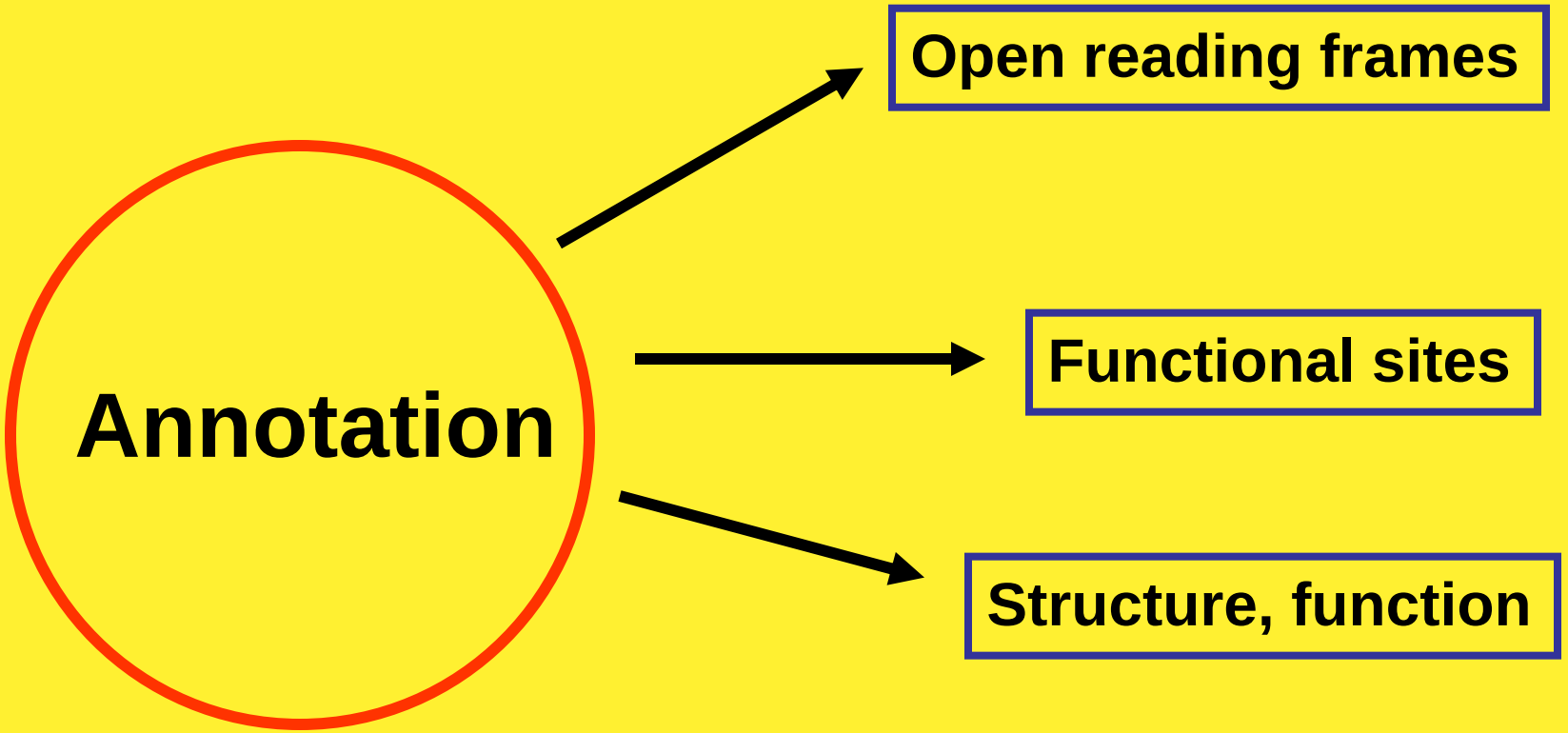
Complete Genomes

as of August 2011:

Eukaryotes	**37**
Prokaryotes	**1708**
Total	**1745**

What can we do with sequences and other type of molecular information?




Annotation
Open reading frames
Functional sites
Structure, function

CCTGACAAATTCGACGTGCGGCATTGCATGCAGACGTGCATG

CGTGCAAATAATCAATGTGGACTTTTCTGCGATTATGGAAGAA

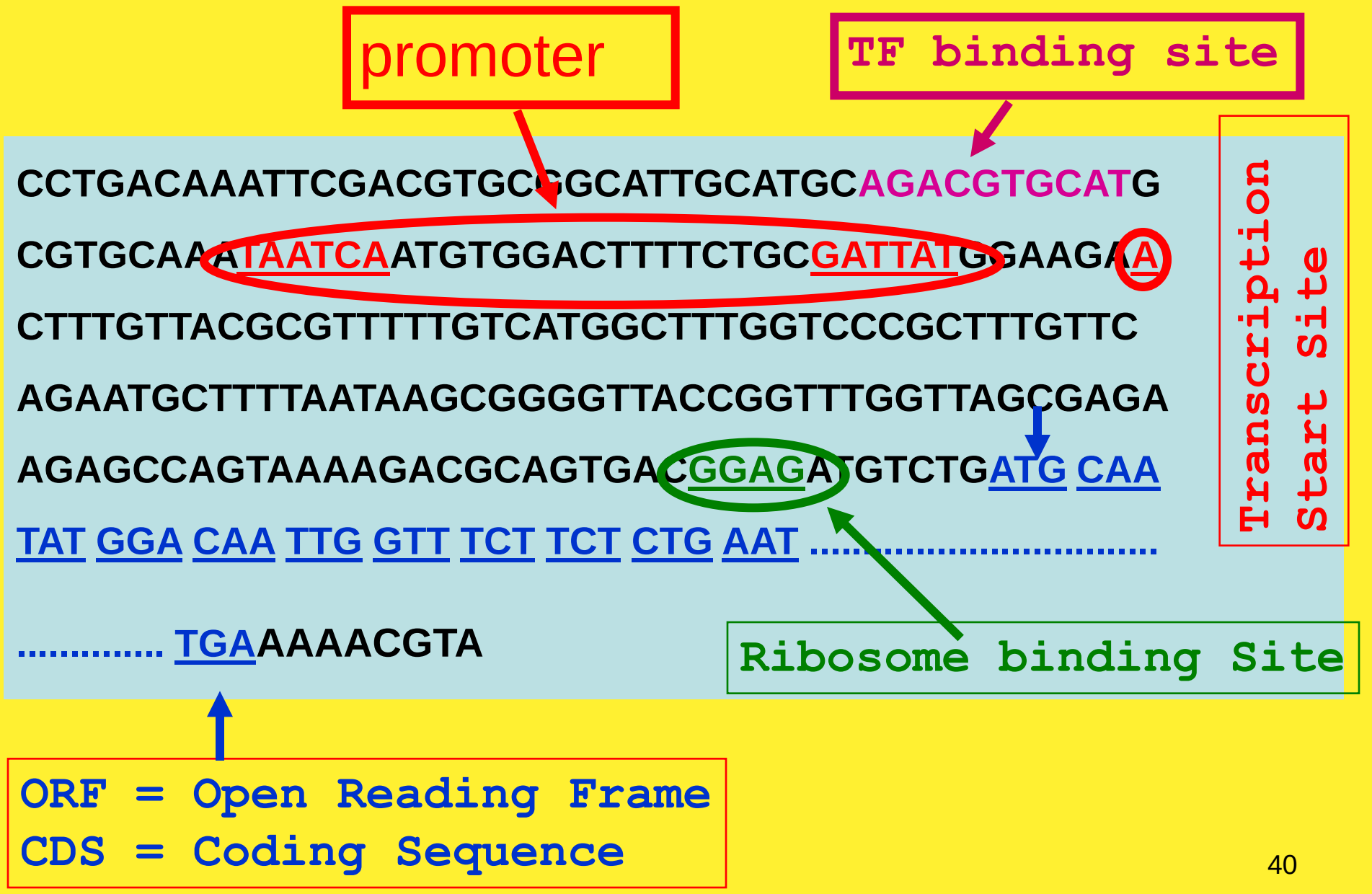
CTTTGTTACGCGTTTTTGTCATGGCTTTGGTCCCGCTTTGTTC

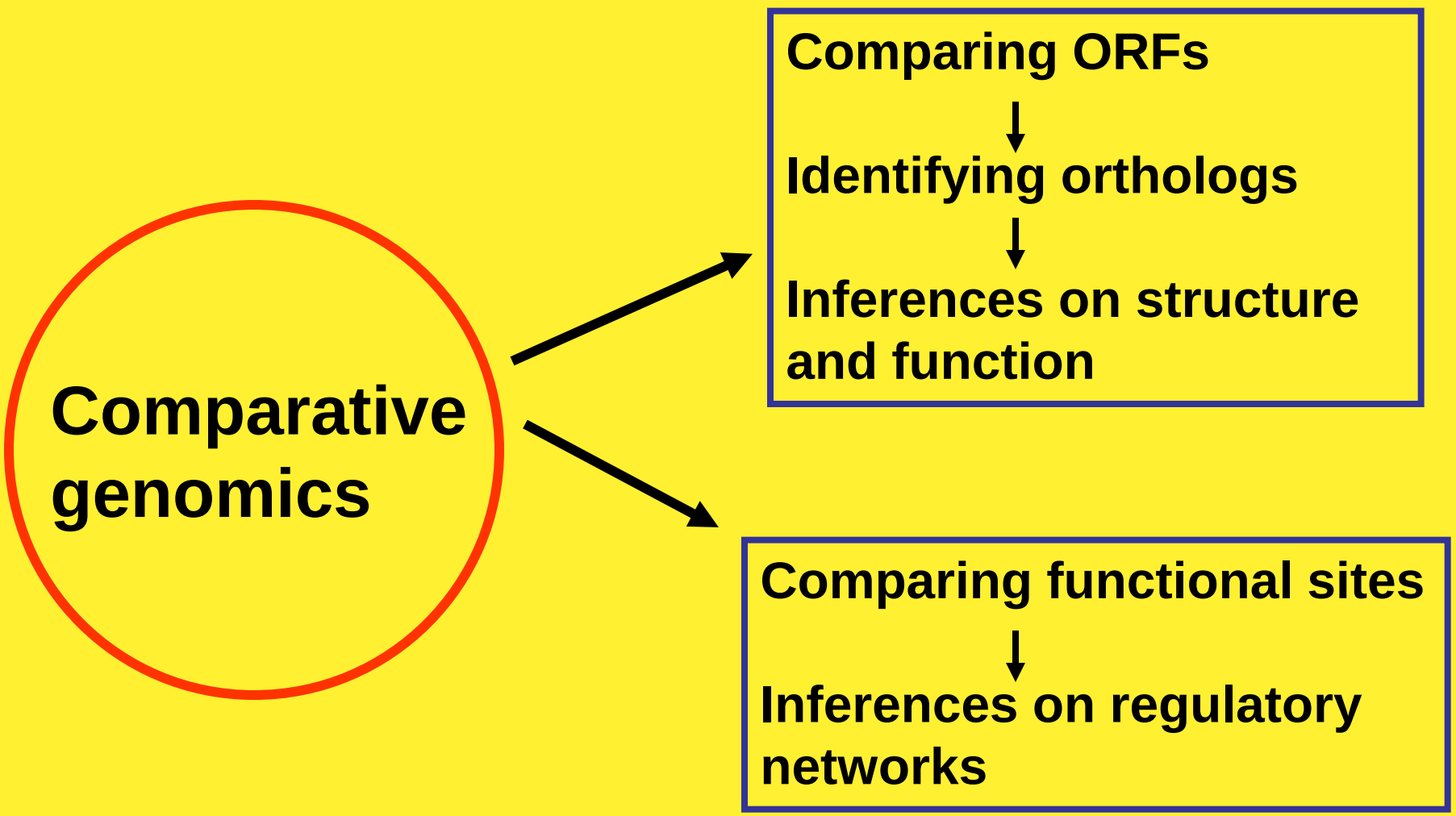
AGAATGCTTTTAATAAGCGGGGTTACCGGTTTGGTTAGCGAGA

AGAGCCAGTAAAAGACGCAGTGACGGAGATGTCTGATG CAA

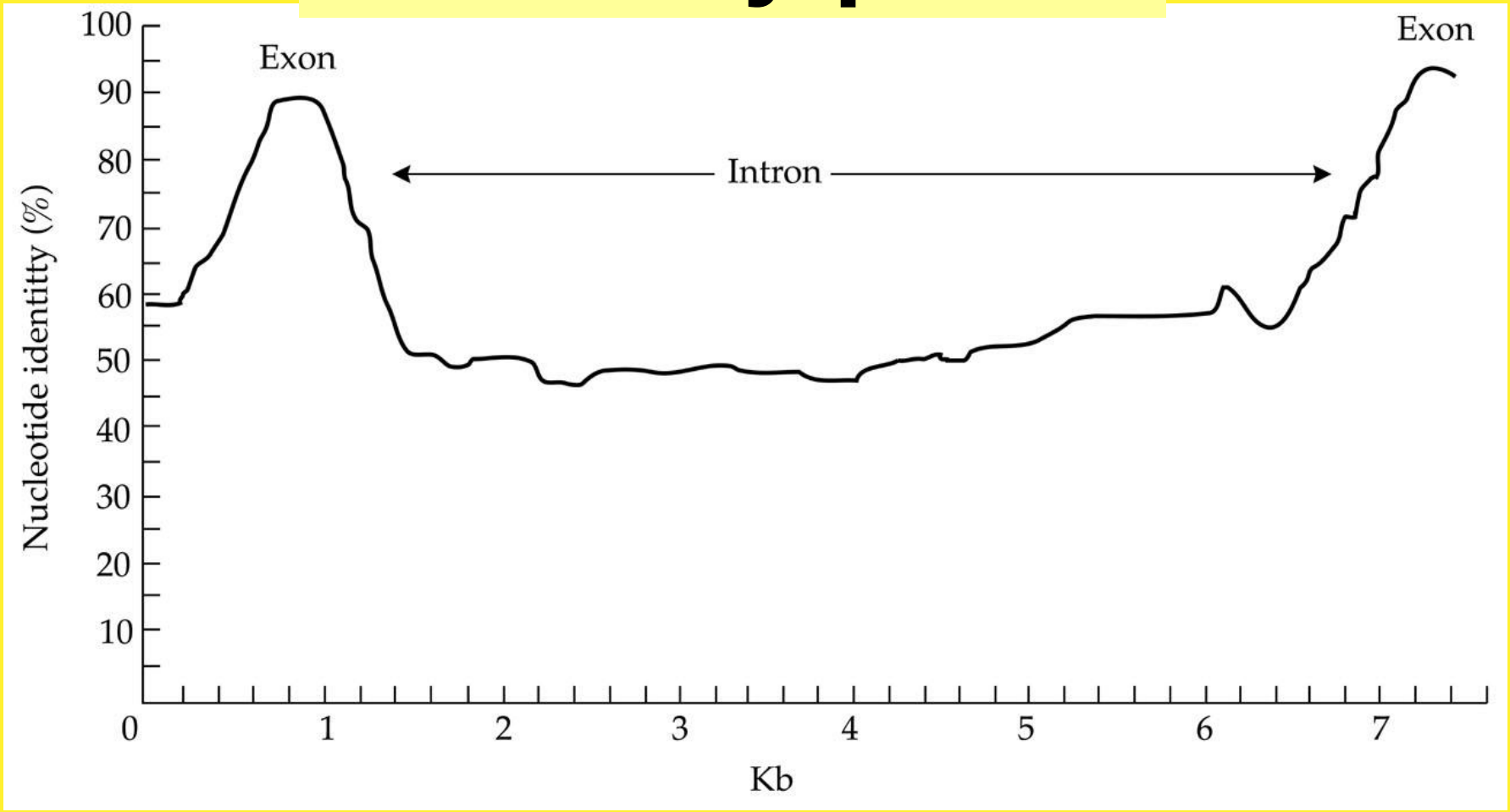
TAT GGA CAA TTG GTT TCT TCT CTG AAT .....

..... TGAAAAACGTA


promoter
TF binding site
CCTGACAAATTCGACGTGCGGCATTGCATGCAGACGTGCATG
CGTGCAAATAATCAATGTGGACTTTTCTGCGATTATGGAAGAA
CTTTGTTACGCGTTTTTGTCATGGCTTTGGTCCCGCTTTGTTC
AGAATGCTTTTAATAAGCGGGGTTACCGGTTTGGTTAGCGAGA
AGAGCCAGTAAAAGACGCAGTGACGGAGATGTCTGATG CAA
TAT GGA CAA TTG GTT TCT TCT CTG AAT .....
..... TGAAAAACGTA
Transcription Start Site
Ribosome binding Site
ORF = Open Reading Frame
CDS = Coding Sequence

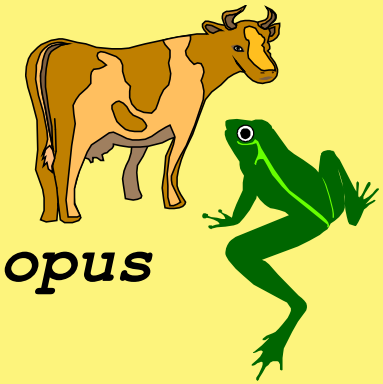

Comparative genomics
Comparing ORFs
Identifying orthologs
Inferences on structure and function
Comparing functional sites
Inferences on regulatory networks

Similarity profiles



Researchers can learned a great deal about the structure and function of human genes by examining their counterparts in model organisms.

Alignment preproinsulin

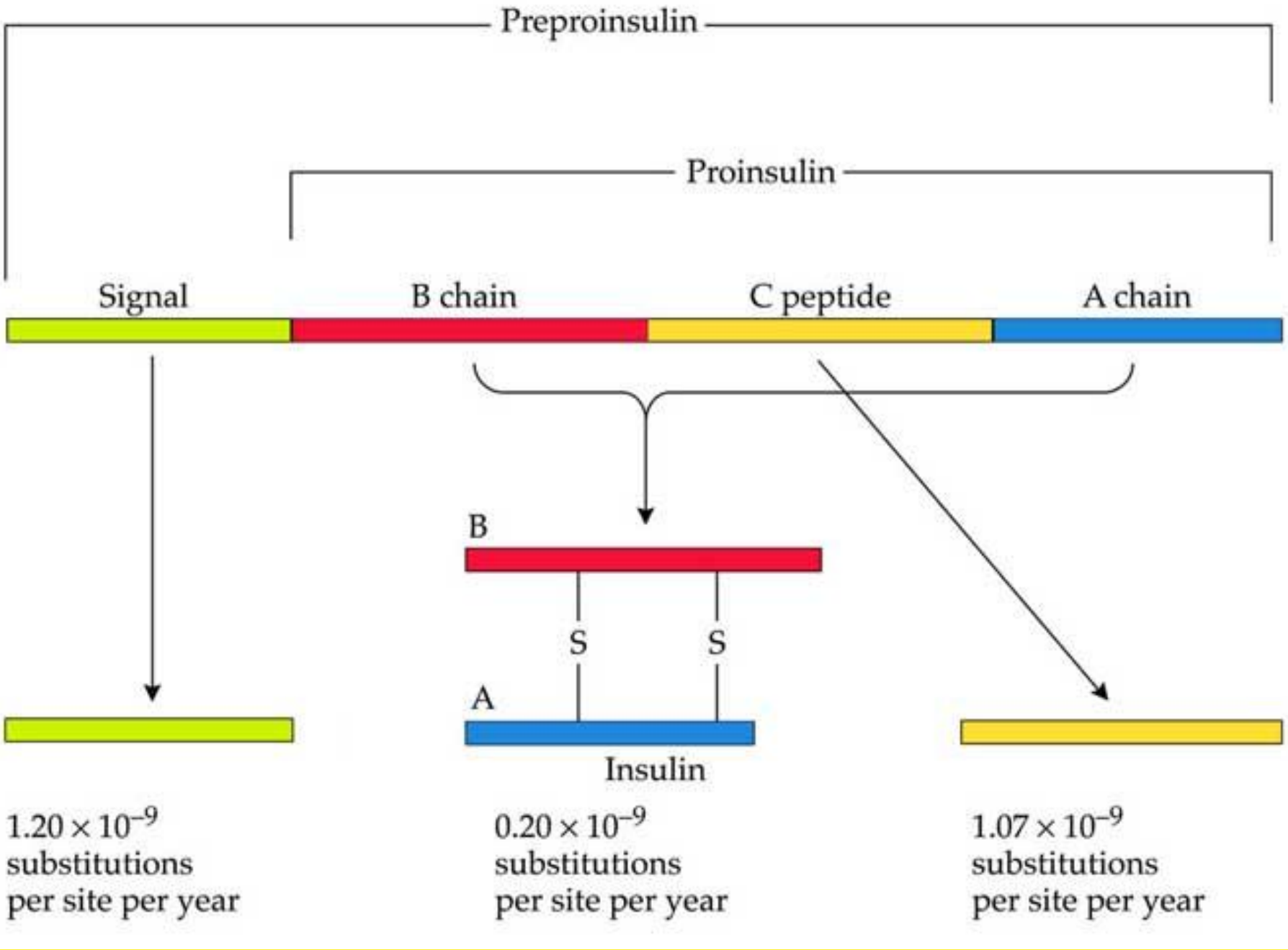


```
Xenopus   MALWMQCLP-LVLVLLFSTPNTEALANQHL
Bos       MALWTRLRPLLALLALWPPPPARAFVNQHL
          **** :   * *.*: *:..* :. *:****

Xenopus   CGSHLVEALYLVCGDRGFFYYPKIKRDIEQ
Bos       CGSHLVEALYLVCGERGFFYTPKARREVEG
          ****************:***** ** :*::*

Xenopus   AQVNGPQDNELDG-MQFQPQEYQKMKRGIV
Bos       PQVG---ALELAGGPGAGGLEGPPQKRGIV
          .**.      ** *       *     *****

Xenopus   EQCCHSTCSLFQLENYCN
Bos       EQCCASVCSLYQLENYCN
          **** *.***:*******
```


Preproinsulin
Proinsulin
Signal
B chain
C peptide
A chain
B
S
S
A
Insulin
1.20 × 10−9
substitutions
per site per year
0.20 × 10−9
substitutions
per site per year
1.07 × 10−9
substitutions
per site per year

Ultraconserved Elements in the Human Genome

Gill Bejerano, Michael Pheasant, Igor Makunin, Stuart Stephen, W. James Kent, John S. Mattick, & David Haussler (*Science* 2004. 304:1321-1325)

There are 481 segments longer than 200 base pairs (bp) that are absolutely conserved (100% identity with no insertions or deletions) between orthologous regions of the human, rat, and mouse genomes. Nearly all of these segments are also conserved in the chicken and dog genomes, with an average of 95 and 99% identity, respectively. Many are also significantly conserved in fish. These ultraconserved elements of the human genome are most often located either overlapping exons in genes involved in RNA processing or in introns or nearby genes involved in the regulation of transcription and development.

There are 156 intergenic, untranscribed, ultraconserved segments

Junk:
Supporting evidence

Megabase deletions of gene deserts result in viable mice

Marcelo A. Nóbrega*, Yiwen Zhu*, Ingrid Plajzer-Frick, Veena Afzal & Edward M. Rubin

...e Institute Walnut Creek, California 94598, USA, and ...n Lawrence Berkeley National Laboratory Berkeley, California

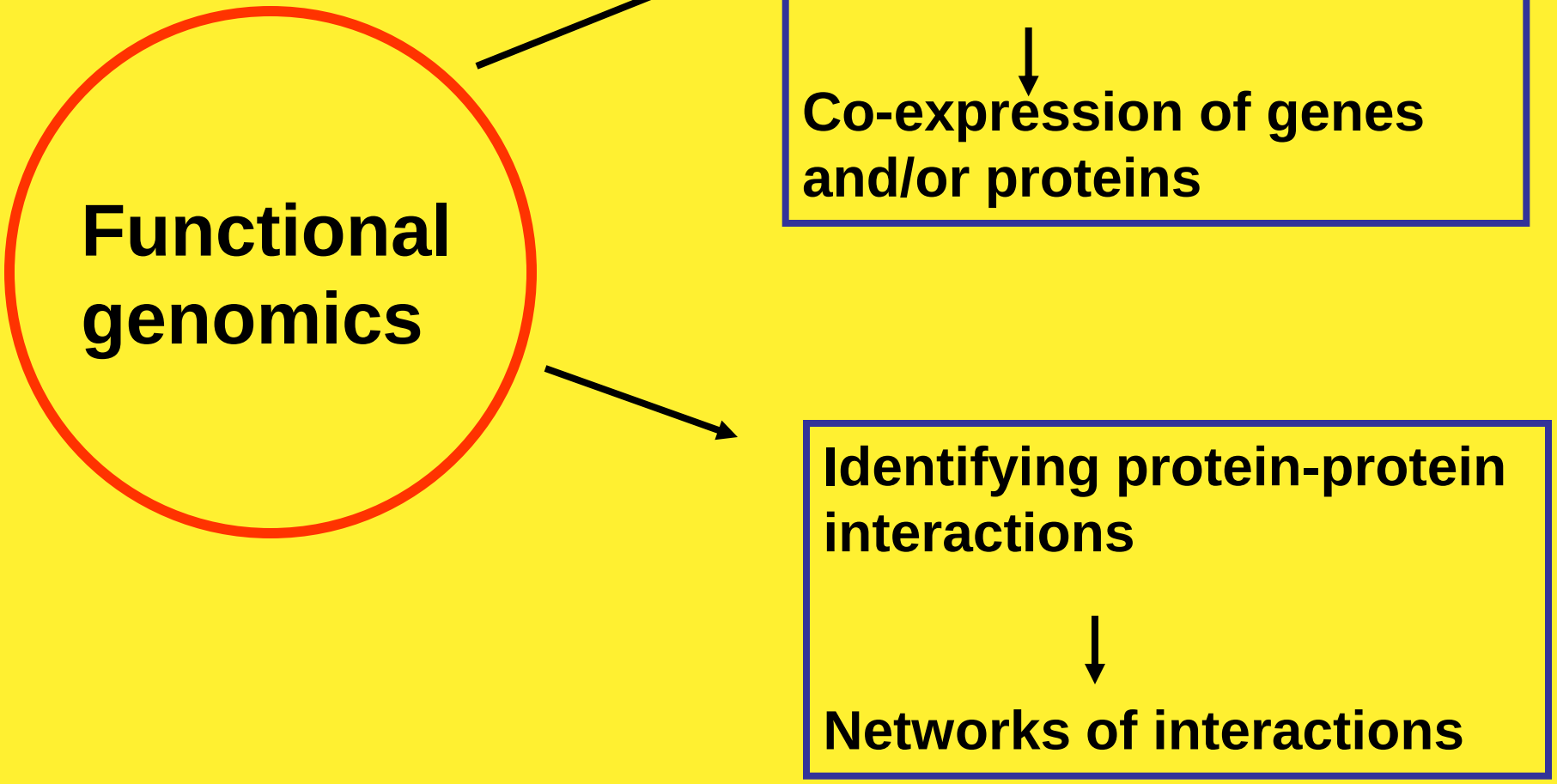
...buted equally to this work

...l importance of the roughly 98% of mammalian genomes not corresponding to protein coding sequences remains largely undetermined[1]. Here we show that some large-scale deletions of the non-coding DNA referred to as gene deserts[2–4] can be well tolerated by an organism. We deleted two large non-coding intervals, 1,511 kilobases and 845 kilobases in length, from the mouse genome. Viable mice homozygous for the deletions were generated and were indistinguishable from wild-type littermates with regard to morphology, reproductive fitness, growth, longevity and a va... homeostasis. Further det... multiple genes bracketing... expression differences in ... mice. Together, the two deleted segments harbour 1,243 non-coding sequences conserved between humans and rodents (more than 100 base pairs, 70% identity). Some of the deleted sequences might encode for functions unidentified in our screen; none-...r support the existence of potentially ...omes of mammals.

The genome of an organism is frequently referred to as its 'book

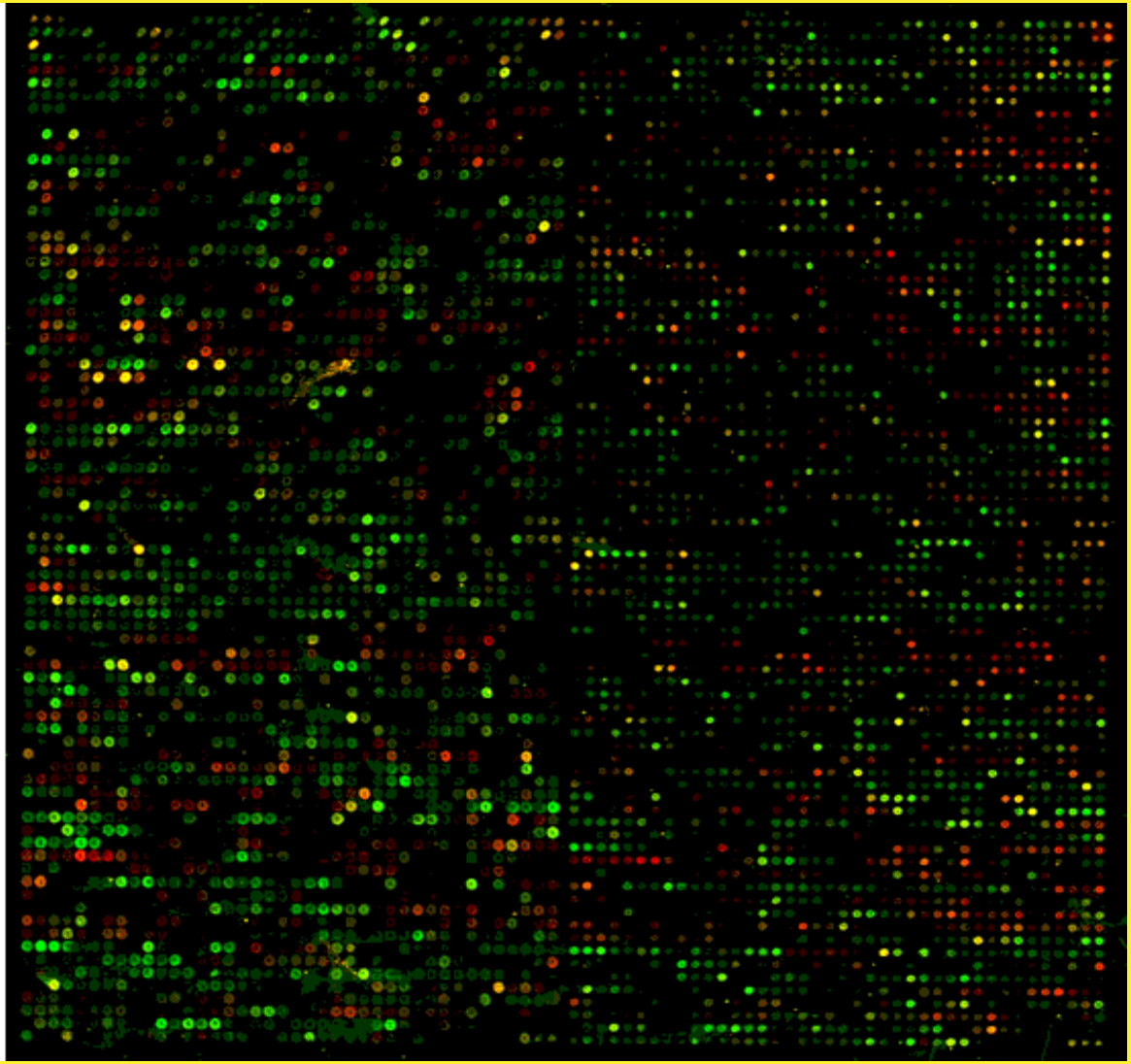
Junk is real!

Nature **431**, 988-993 (21 October 2004)


Functional genomics
Genome-wide profiling of:
• mRNA levels
• Protein levels
Co-expression of genes and/or proteins
Identifying protein-protein interactions
Networks of interactions

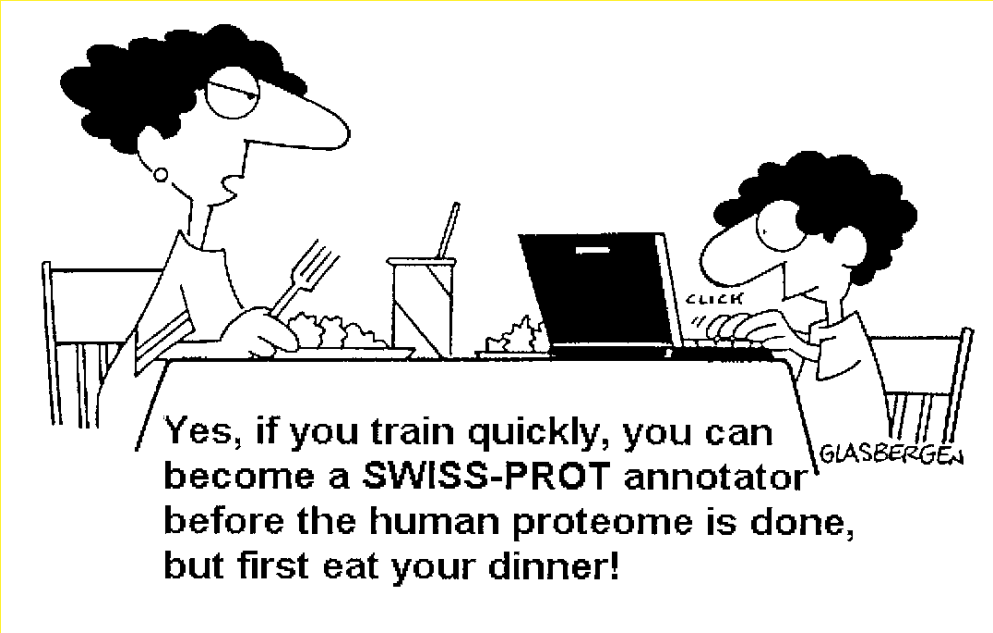
Understanding the function of genes and other parts of the genome

The complete S. cerevisiae genome on a microarray chip hybridised to RNA from cultures in anaerobic and aerobic stationary phase



Structural genomics

Assign structure to all proteins encoded in a genome


CLICK
Yes, if you train quickly, you can become a SWISS-PROT annotator before the human proteome is done, but first eat your dinner!
GLASBERGEN

databases

Database or databank?

Initially

- Databank (in UK)
- Database (in the USA)

Solution

- The abbreviation *db*

What is a Database?

A **structured collection** of data held in computer storage; *esp.* one that incorporates software to make it accessible in a variety of ways; *transf.*, any **large collection** of information.

database management: the organization and manipulation of data in a database.

database management system (DBMS): a software package that provides all the functions required for database management.

database system: a database together with a database management system.

Oxford Dictionary

What is a database?

- A collection of data
  - structured
  - searchable (index) -> table of contents
  - updated periodically (release) -> new edition
  - cross-referenced (hyperlinks) -> links with other db

- Includes also associated tools (software) necessary for access, updating, information insertion, information deletion….

- Data storage management: flat files, relational databases…

Database: a « flat file » example

Flat-file database (« flat file, 3 entries »):

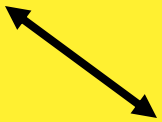
Accession number: 1
First Name: Amos
Last Name: Bairoch
Course: Pottery 2000; Pottery 2001;
//
Accession number: 2
First Name: Dan
Last name: Graur
Course: Pottery 2000, Pottery 2001; Ballet 2001, Ballet 2002
//
Accession number 3:
First Name: John
Last name: Travolta
Course: Ballet 2001; Ballet 2002;
//

- Easy to manage: all the entries are visible at the same time !

Database: a « relational » example

Relational database (« table file »):

Teacher	Accession number	Education
Amos	1	Biochemistry
Dan	2	Genetics
John	3	Scientology



Course	Year	Involved teachers
Advanced Pottery	2000; 2001	1; 2
Ballet for Fat People	2001; 2002	2; 3

Why biological databases?

- Exponential growth in biological data.
- Data (genomic sequences, 3D structures, 2D gel analysis, MS analysis, Microarrays….) are no longer published in a conventional manner, but directly submitted to databases.
- Essential tools for biological research. The only way to publish massive amounts of data without using all the paper in the world.

Distribution of sequences

- Books, articles 1968 -> 1985
- Computer tapes 1982 -> 1992
- Floppy disks 1984 -> 1990
- CD-ROM 1989 ->
- FTP 1989 ->
- On-line services 1982 -> 1994
- WWW 1993 ->
- DVD 2001 ->

Some statistics

- More than 1000 different ‘biological’ databases
- Variable size: <100Kb to >20Gb
  - DNA: > 20 Gb
  - Protein: 1 Gb
  - 3D structure: 5 Gb
  - Other: smaller
- Update frequency: **daily** to **annually** to **seldom** to **forget about it**.
- Usually accessible through the web (some free, some not)

Some databases in the field of molecular biology…

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genline, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene,Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, YEPD, YPD, YPM, etc ..... !!!!

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolic networks
- Regulatory networks
- Bibliography
- Expression (Microarrays,…)
- Specialized

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

EBI:
http://www.ebi.ac.uk/

DDBJ:
http://www.ddbj.nig.ac.jp/

Literature Databases:

Bookshelf: A collection of searchable biomedical books linked to PubMed.

PubMed: Allows searching by author names, journal titles, and a new Preview/Index option. PubMed database provides access to over 12 million MEDLINE citations back to the mid-1960's. It includes History and Clipboard options which may enhance your search session.

PubMed Central: The U.S. National Library of Medicine digital archive of life science journal literature.

OMIM: Online Mendelian Inheritance in Man is a database of human genes and genetic disorders (also OMIA).

NCBI
National Center for Biotechnology Information

Search All Databases

Search Clear

NCBI Home
Site Map (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

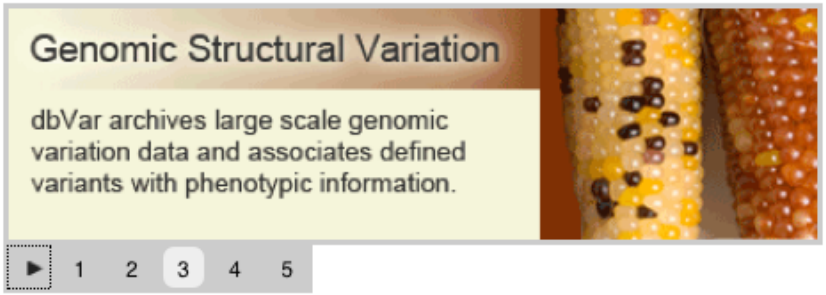
Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases



Popular Resources

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed
- PubMed Central
- SNP

NCBI News

Preliminary genomic assemblies from two isolates from the European E. coli outbreak now available

07 Jun 2011

Preliminary genomic assemblies of two isolates are in the Nucleotide

New version of Cn3D (v.4.3) now available

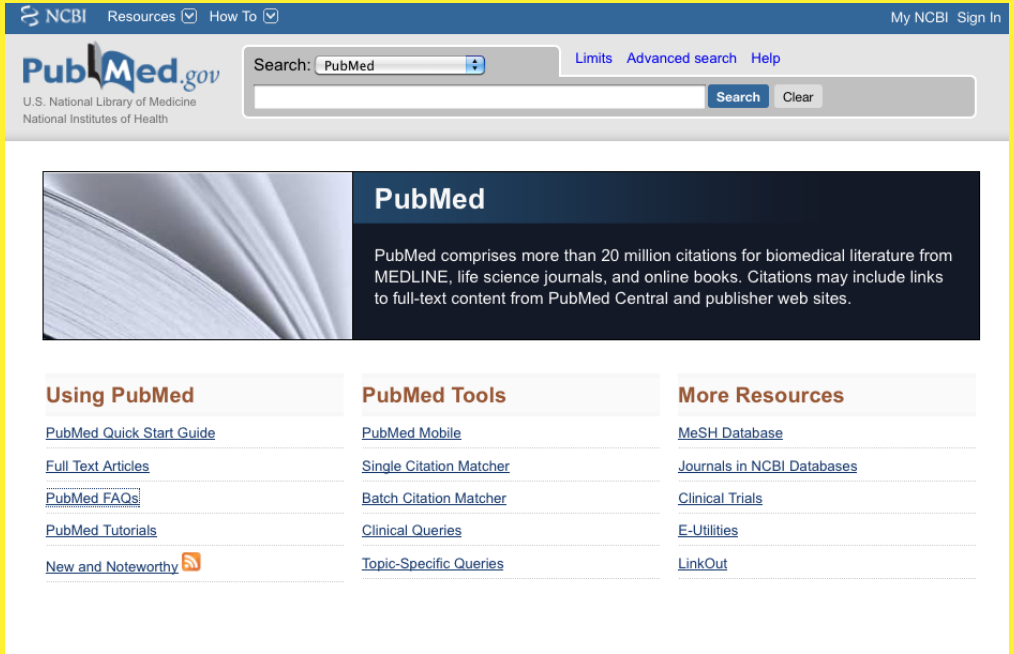
07 Jun 2011

A new version of this popular 3D molecular visualization program

More...

PubMed (Medline)

- MEDLINE covers the fields of medicine, nursing, dentistry, veterinary medicine, public health, and **preclinical sciences**
- Contains citations from approximately 5,200 worldwide journals in 37 languages; 60 languages for older journals.
- Contains over 20 million citations since 1948
- Contains links to biological db and to some journals
- New records are added to PreMEDLINE daily!



- Alerting services
  - http://www.pubcrawler.ie/
  - http://www.biomail.org

A search by subject: “mitochondrion evolution”

Biol Direct. 2011 Jun 30;6(1):35. [Epub ahead of print]

Energetics and genetics across the prokaryote-eukaryote divide.

Lane N.

Abstract

ABSTRACT: BACKGROUND: All complex life on Earth is eukaryotic. All eukaryotic cells share a common ancestor that arose just once in 4 billion years of **evolution**. Prokaryotes show no tendency to evolve greater morphological complexity, despite their metabolic virtuosity. Here I argue that the eukaryotic cell originated in a unique prokaryotic endosymbiosis, a singular event that transformed the selection pressures acting on both host and endosymbiont. Results: The reductive **evolution** and specialisation of endosymbionts to **mitochondria** resulted in an extreme genomic asymmetry, in which the residual mitochondrial genomes enabled the expansion of bioenergetic membranes over several orders of magnitude, overcoming the energetic constraints on prokaryotic genome size, and permitting the host cell genome to expand (in principle) over 200,000-fold. This energetic transformation was permissive, not prescriptive; I suggest that the actual increase in early eukaryotic genome size was driven by a heavy early bombardment of genes and introns from the endosymbiont to the host cell, producing a high mutation rate. Unlike prokaryotes, with lower mutation rates and heavy selection pressure to lose genes, early eukaryotes without genome-size limitations could mask mutations by cell fusion and genome duplication, as in allopolyploidy, giving rise to a proto-sexual cell cycle. The side effect was that a large number of shared eukaryotic basal traits accumulated in the same population, a sexual eukaryotic common ancestor, radically different to any known prokaryote. Conclusions: The combination of massive bioenergetic expansion, release from genome-size constraints, and high mutation rate favoured a protosexual cell cycle and the accumulation of eukaryotic traits. These factors explain the unique origin of eukaryotes, the absence of true evolutionary intermediates, and the **evolution** of sex in eukaryotes but not prokaryotes. Reviewers: This article was reviewed by: Eugene Koonin, William Martin, Ford Doolittle and Mark van der Giezen. For complete reports see the Reviewers' Comments section.

PMID: 21714941 [PubMed - as supplied by publisher] Free full text

LinkOut - more resources

Related citations

The origin of introns and their role in eukaryogenesis: a compromise [Biol Direct. 2006]

The energetics of genome complexity. [Nature. 2010]

On the need for widespread horizontal gene transfers under genome size cc [Biol Direct. 2009]

Review The neomuran origin of archaebacteria, the negibacterial ı [Int J Syst Evol Microbiol. 2002]

Review The phagotrophic origin of eukaryotes and phylogenetic [Int J Syst Evol Microbiol. 2002]

See reviews...

See all...

Recent activity

Turn Off Clear

Energetics and genetics across the prokaryote-eukaryote divide. PubMed

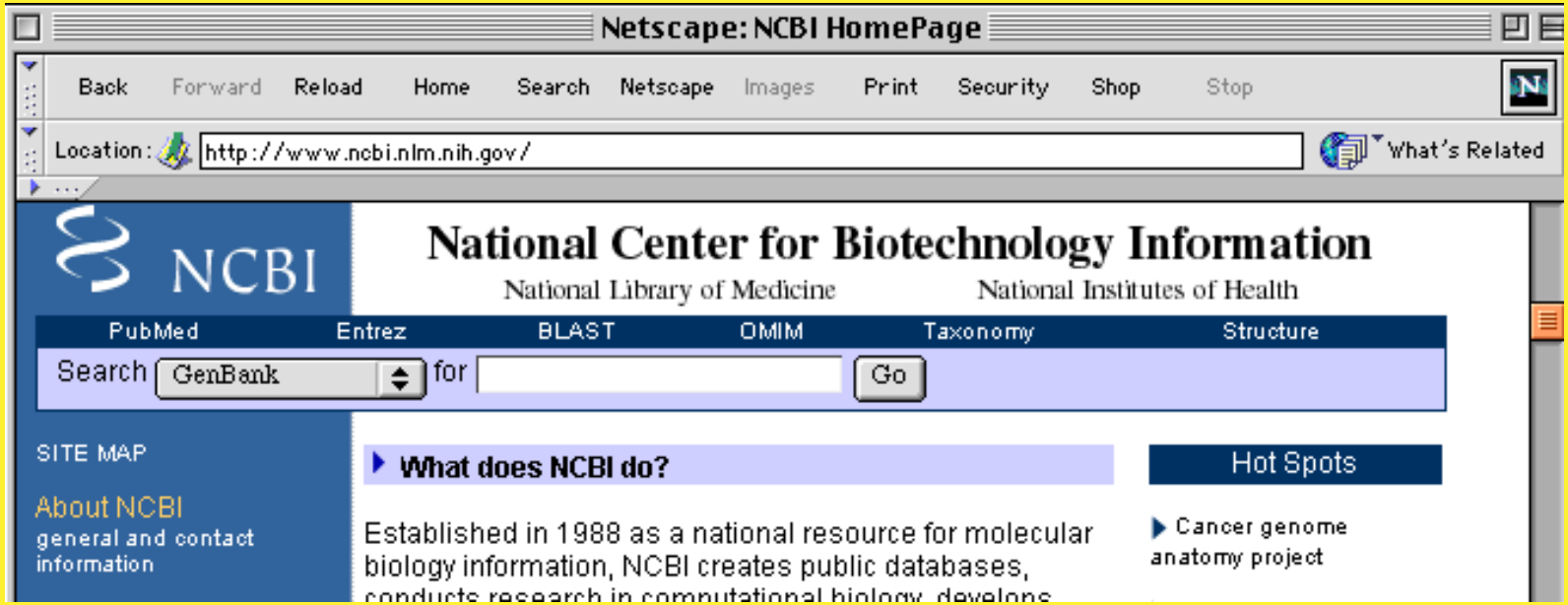
mitochondrion evolution (3990) PubMed

PubMed Help - PubMed Help Bookshelf

See more...

Type in a Query term

- Enter your search words in the query box and hit the "**Go**" button



http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html#Searching

The Syntax …

1. **Boolean operators:** AND, OR, NOT must be entered in UPPERCASE (e.g., promoters OR response elements). The default is AND.

2. Entrez processes all Boolean operators in a left-to-right sequence. The order in which Entrez processes a search statement can be changed by enclosing individual concepts in parentheses. The terms inside the parentheses are processed first. For example, the search statement: g1p3 OR (response AND element AND promoter).

3. Quotation marks: The term inside the quotation marks is read as one phrase (e.g. “public health” is different than public health, which will also include articles on public latrines and their effect on health workers).

4. Asterisk: Extends the search to all terms that start with the letters before the asterisk. For example, dia* will include such terms as diaphragm, dial, and diameter.

Refine the Query

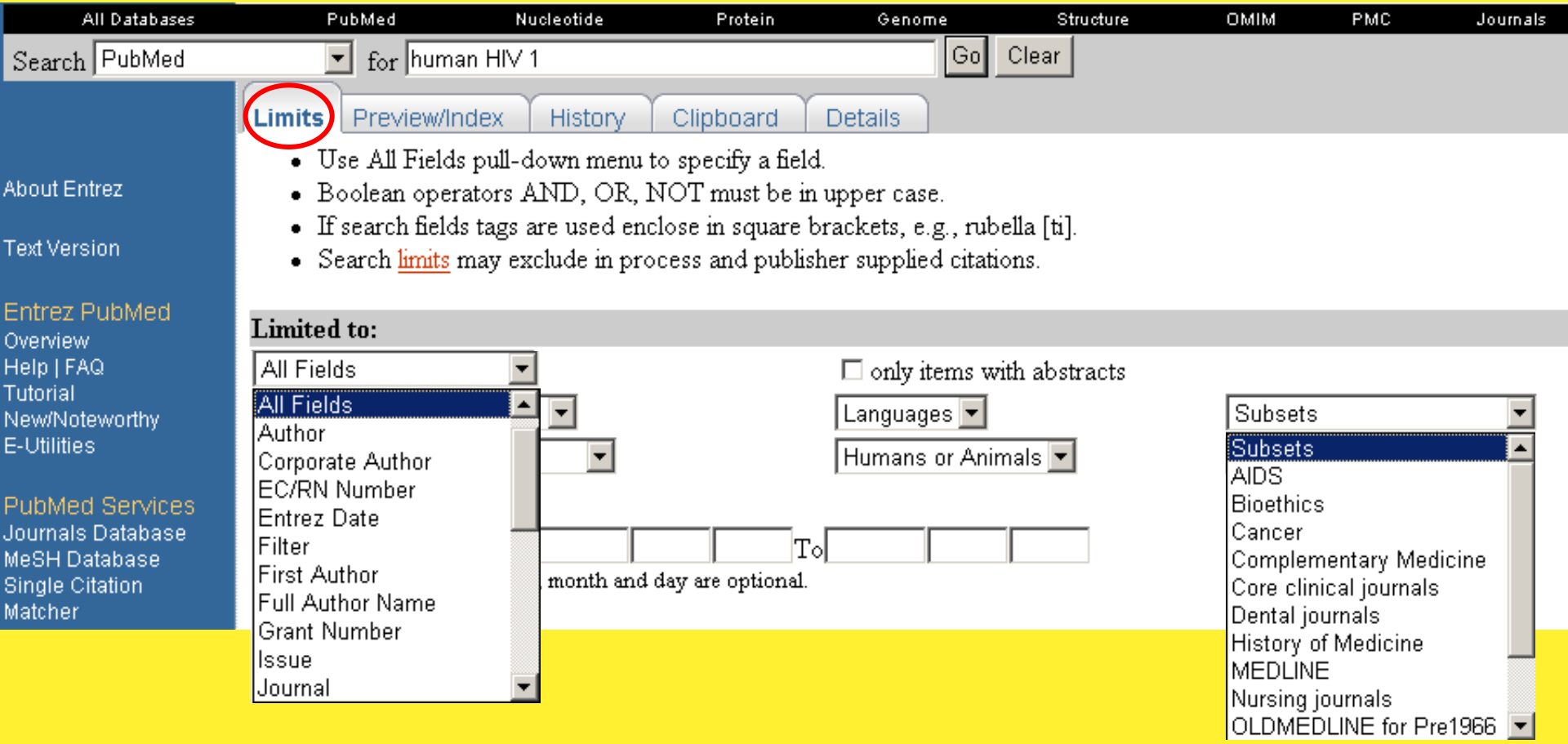
- Often a search finds too many (or too few) sequences, so you can go back and try again with more (or fewer) keywords in your query
- The “History” feature allows you to combine any of your past queries.
- The “Limits” feature allows you to limit a query to specific organisms, sequences submitted during a specific period of time, etc.
- [Many other features are designed to search for literature in MEDLINE]

Related Items

You can search for a text term in sequence annotations or in **MEDLINE** abstracts, and find all articles, DNA, and protein sequences that mention that term.

Then from any article or sequence, you can move to "related articles" or "related sequences".

- Relationships between sequences are computed with BLAST
- Relationships between articles are computed with "MESH" terms (shared keywords)
- Relationships between DNA and protein sequences rely on accession numbers
- Relationships between sequences and **MEDLINE** articles rely on both shared keywords and the mention of accession numbers in the articles.


All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals
Search PubMed for human HIV 1 Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation
Matcher
Use All Fields pull-down menu to specify a field.
Boolean operators AND, OR, NOT must be in upper case.
If search fields tags are used enclose in square brackets, e.g., rubella [ti].
Search limits may exclude in process and publisher supplied citations.
Limited to:
All Fields
All Fields
Author
Corporate Author
EC/RN Number
Entrez Date
Filter
First Author
Full Author Name
Grant Number
Issue
Journal
only items with abstracts
Languages
Humans or Animals
To
month and day are optional.
Subsets
Subsets
AIDS
Bioethics
Cancer
Complementary Medicine
Core clinical journals
Dental journals
History of Medicine
MEDLINE
Nursing journals
OLDMEDLINE for Pre1966

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals

Search PubMed for human HIV 1 Preview Go Clear

Limits | Preview/Index | History | Clipboard | Details

- Enter terms and click Preview to see only the number of search results.
- To combine searches use # before search number, e.g., (#2 OR #3) AND asthma.
- Click on query # to add to strategy

Search	Most Recent Queries
#2	Search **human HIV 1**

Add Term(s) to Query or View Index:

- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within

All Fields Preview Index

add a term to the query box.

- Affiliation
- All Fields
- Author
- Corporate Author
- EC/RN Number
- Entrez Date
- Filter
- First Author
- Full Author Name
- Grant Number
- Issue

About Entrez

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher

Search Field Descriptions and Tags

Affiliation [AD]	Issue [IP]	Place of Publication [PL]
All Fields [ALL]	Journal Title [TA]	Publication Date [DP]
Author [AU]	Language [LA]	Publication Type [PT]
Comment Corrections	Last Author [LASTAU]	Publisher Identifier [AID]
Corporate Author [CN]	MeSH Date [MHDA]	Secondary Source ID [SI]
EC/RN Number [RN]	MeSH Major Topic [MAJR]	Subset [SB]
Entrez Date [EDAT]	MeSH Subheadings [SH]	Substance Name [NM]
Filter [FILTER]	MeSH Terms [MH]	Text Words [TW]
First Author Name [1AU]	NLM Unique ID [JID]	Title [TI]
Full Author Name [FAU]	Other Term [OT]	Title/Abstract [TIAB]
Full Investigator Name [FIR]	Owner	Transliterated Title [TT]
Grant Number [GR]	Pagination [PG]	UID [PMID]
Investigator [IR]	Personal Name as Subject [PS]	Volume [VI]
	Pharmacological Action MeSH Terms [PA]	

A search by authors:

"Esser" [au] AND "martin" [au]

NCBI Resources How To My NCBI Sign In

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed RSS Save search Limits Advanced search Help

"Esser C" [au] AND "Martin W" [au] Search Clear

Display Settings: Summary, Sorted by Recently Added Send to:

Results: 4

1. Supertrees and symbiosis in eukaryote genome evolution.
Esser C, **Martin W**.
Trends Microbiol. 2007 Oct;15(10):435-7. Epub 2007 Sep 19.
PMID: 17884500 [PubMed - indexed for MEDLINE]
Related citations

2. The origin of mitochondria in light of a fluid prokaryotic chromosome model.
Esser C, **Martin W**, Dagan T.
Biol Lett. 2007 Apr 22;3(2):180-4.
PMID: 17251118 [PubMed - indexed for MEDLINE] **Free PMC Article**
Free full text Related citations

3. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes.
Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, Bryant D, Steel MA, Lockhart PJ, Penny D, **Martin W**.
Mol Biol Evol. 2004 Sep;21(9):1643-60. Epub 2004 May 21.
PMID: 15155797 [PubMed - indexed for MEDLINE] **Free Article**
Related citations

4. Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a streamlined genome overrun by mobile genetic elements.
Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, Brownlie JC, McGraw EA, **Martin W**, **Esser C**, Ahmadinejad N, Wiegand C, Madupu R, Beanan MJ, Brinkac LM, Daugherty SC, Durkin AS, Kolonay JF, Nelson WC, Mohamoud Y, Lee P, Berry K, Young MB, Utterback T, Weidman J, Nierman WC, Paulsen IT, Nelson KE, Tettelin H, O'Neill SL, Eisen JA.
PLoS Biol. 2004 Mar;2(3):E69. Epub 2004 Mar 16.
PMID: 15024419 [PubMed - indexed for MEDLINE] **Free PMC Article**
Free full text Related citations

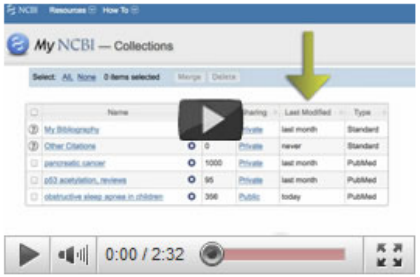
Display Settings: Summary, Sorted by Recently Added Send to:

Filter your results:

All (4)
Free Full Text (3)
Review (0)

Manage Filters

Save Results in Collections Tutorial



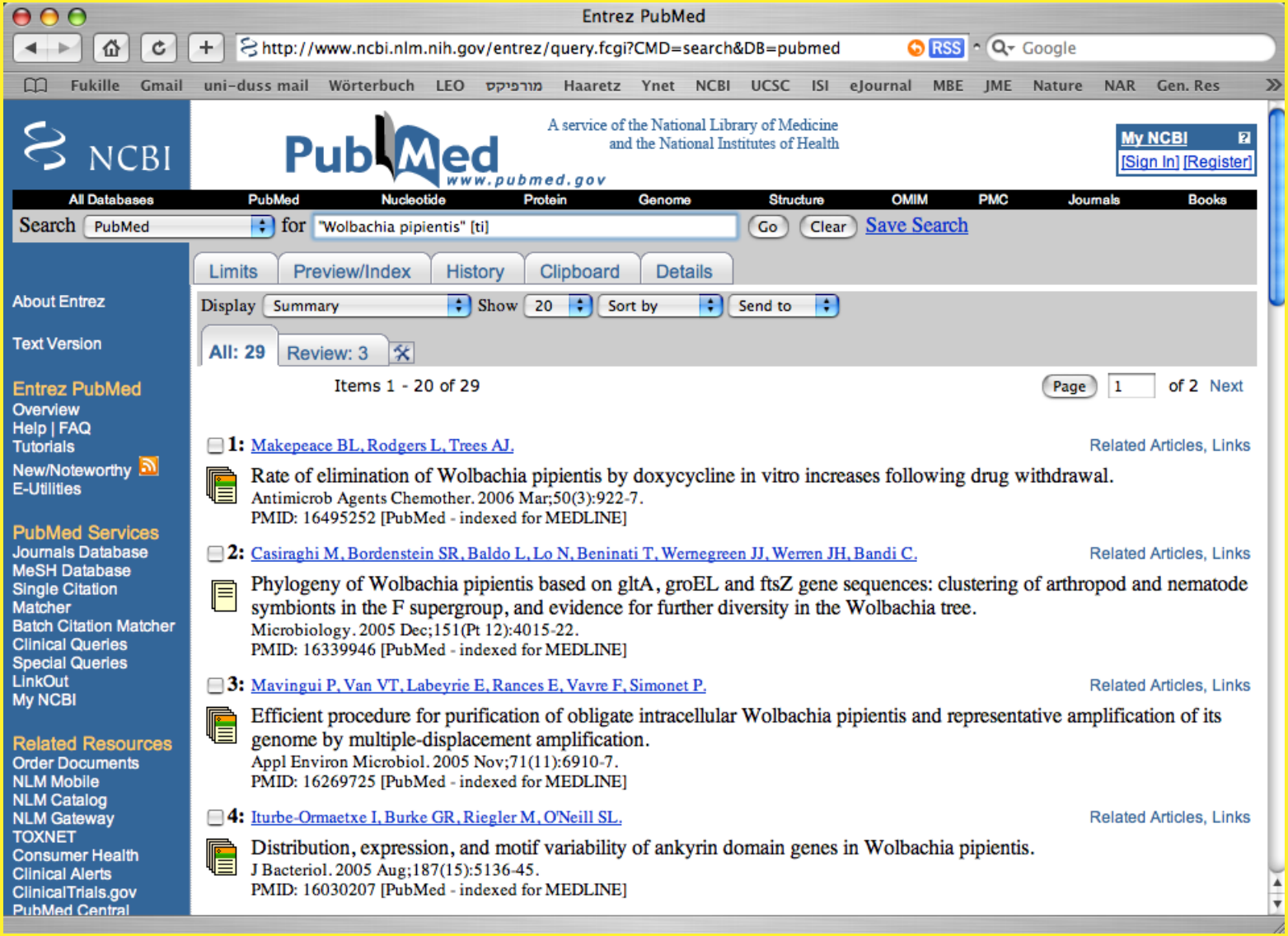
See larger video at YouTube
See all NCBI YouTube video channel videos

2 free full-text articles in PubMed Central

The origin of mitochondria in light of a fluid prokaryotic chromosome model. [Biol Lett. 2007]
Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a stre [PLoS Biol. 2004]
See all (2)...

A search by title word:

"Wolbachia pipientis" [ti]



Database Search Strategies

- General search principles - not limited to sequence (or to biology).
- Start with broad keywords and narrow the search using more specific terms.
- Try variants of spelling, numbers, etc.
- Search many databases.
- **Be persistent!!**

Searching PubMed

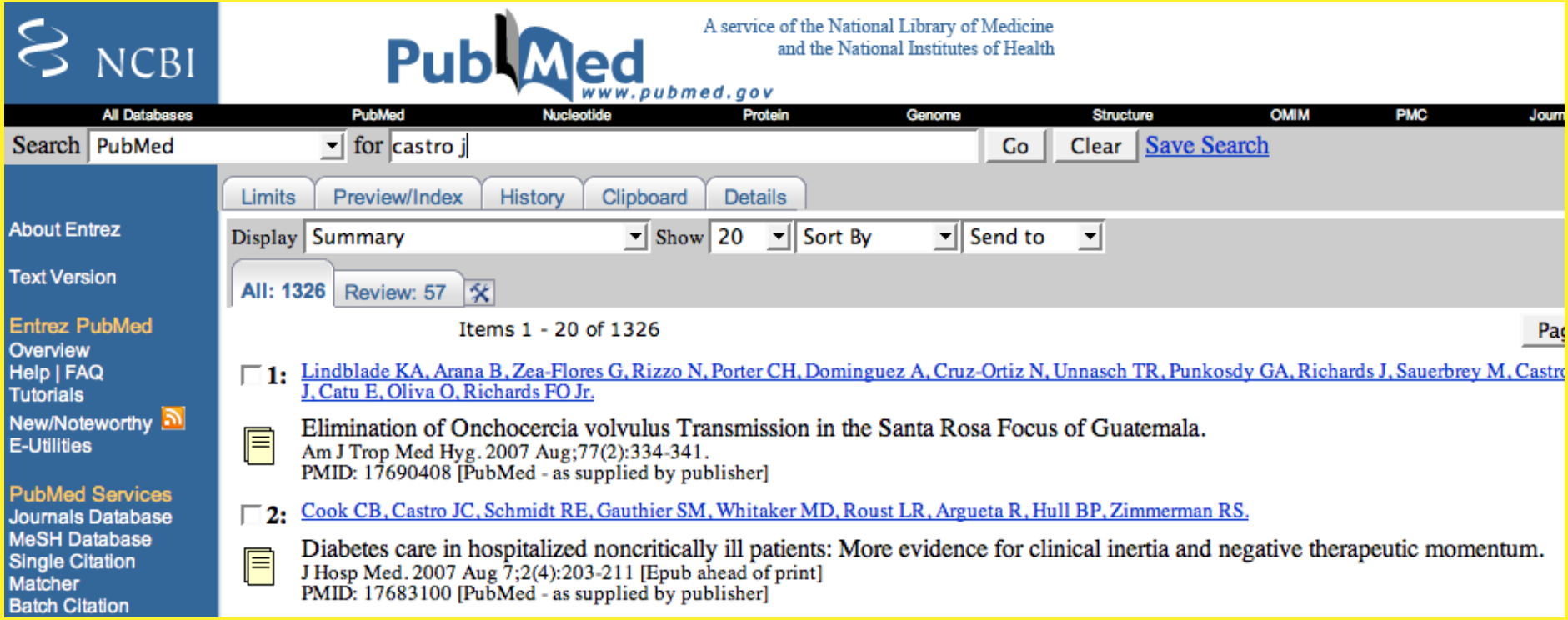
- Structureless searches
  - Automatic term mapping
- Structured searches
  - Tags, e.g. [au], [ta], [dp], [ti]
  - Boolean operators, e.g. AND, OR, NOT, ()
- Additional features
  - Subsets, limits
  - Clipboard, history

Start working:

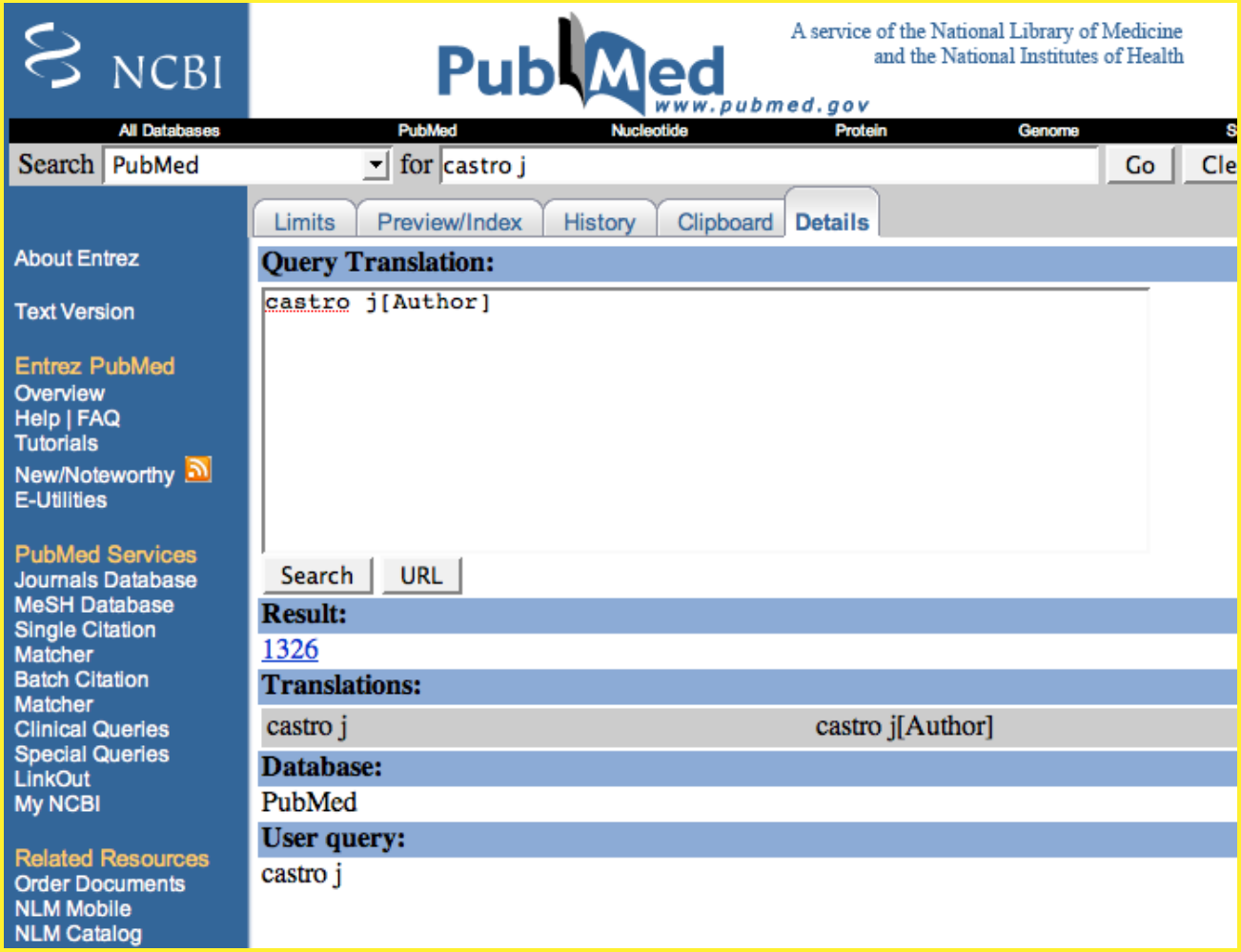
Search PubMed

1. cuban cigars
2. cuban OR cigars
3. “cuban cigars”
4. cuba* cigar*
5. (cuba* cigar*) NOT smok*
6. Fidel Castro
7. “fidel castro”
8. #6 NOT #7

“Details” and “History” in PubMed



“Details” and “History” in PubMed



The **OMIM** (**Online Mendelian Inheritance in Man**)

- Genes and genetic disorders
- Edited by team at Johns Hopkins
- Updated daily

MIM Number Prefixes

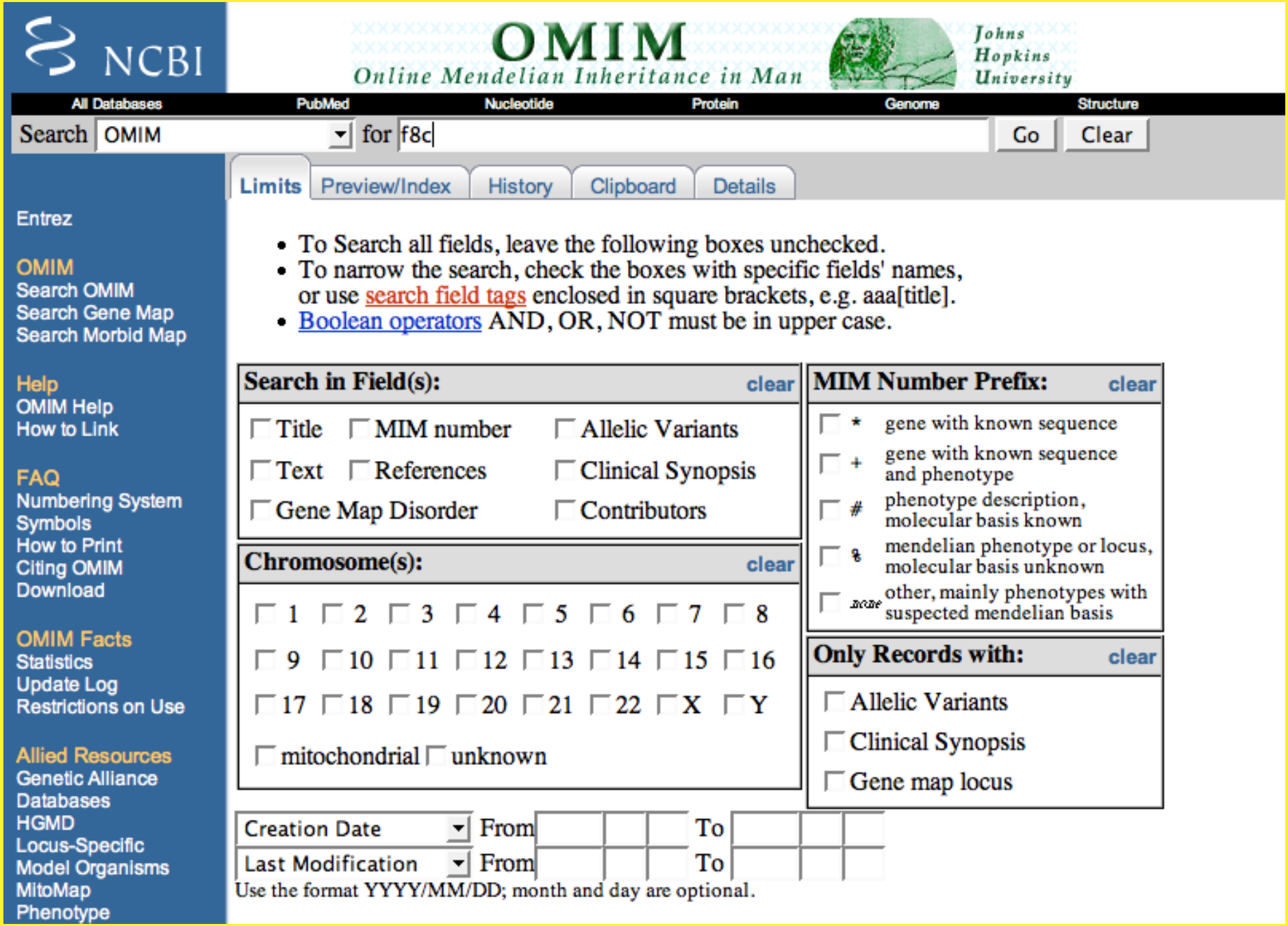
*	gene with known sequence
+	gene with known sequence and phenotype
#	phenotype description, molecular basis known
%	mendelian phenotype or locus, molecular basis unknown
no prefix	other, mainly phenotypes with suspected mendelian basis

Searching OMIM

- Search Fields
  - Name of trait, e.g., hypertension
  - Cytogenetic location, e.g., 1p31.6
  - Inheritance, e.g., autosomal dominant
  - Gene, e.g., coagulation factor VIII

OMIM search tags

All Fields	[ALL]
Allelic Variant	[AV] or [VAR]
Chromosome	[CH] or [CHR]
Clinical Synopsis	[CS] or [CLIN]
Gene Map	[GM] or [MAP]
Gene Name	[GN] or [GENE]
Reference	[RE] or [REF]


NCBI
OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University
All Databases PubMed Nucleotide Protein Genome Structure
Search OMIM for f8c Go Clear
Limits Preview/Index History Clipboard Details
Entrez
OMIM
Search OMIM
Search Gene Map
Search Morbid Map
Help
OMIM Help
How to Link
FAQ
Numbering System
Symbols
How to Print
Citing OMIM
Download
OMIM Facts
Statistics
Update Log
Restrictions on Use
Allied Resources
Genetic Alliance
Databases
HGMD
Locus-Specific
Model Organisms
MitoMap
Phenotype
To Search all fields, leave the following boxes unchecked.
To narrow the search, check the boxes with specific fields' names, or use search field tags enclosed in square brackets, e.g. aaa[title].
Boolean operators AND, OR, NOT must be in upper case.
Search in Field(s): clear
Title MIM number Allelic Variants
Text References Clinical Synopsis
Gene Map Disorder Contributors
Chromosome(s): clear
1 2 3 4 5 6 7 8
9 10 11 12 13 14 15 16
17 18 19 20 21 22 X Y
mitochondrial unknown
MIM Number Prefix: clear
* gene with known sequence
+ gene with known sequence and phenotype
# phenotype description, molecular basis known
% mendelian phenotype or locus, molecular basis unknown
other, mainly phenotypes with suspected mendelian basis
Only Records with: clear
Allelic Variants
Clinical Synopsis
Gene map locus
Creation Date From To
Last Modification From To
Use the format YYYY/MM/DD; month and day are optional.

Start working:

Search OMIM

How many types of hemophilia are there?

For how many is the affected gene known?

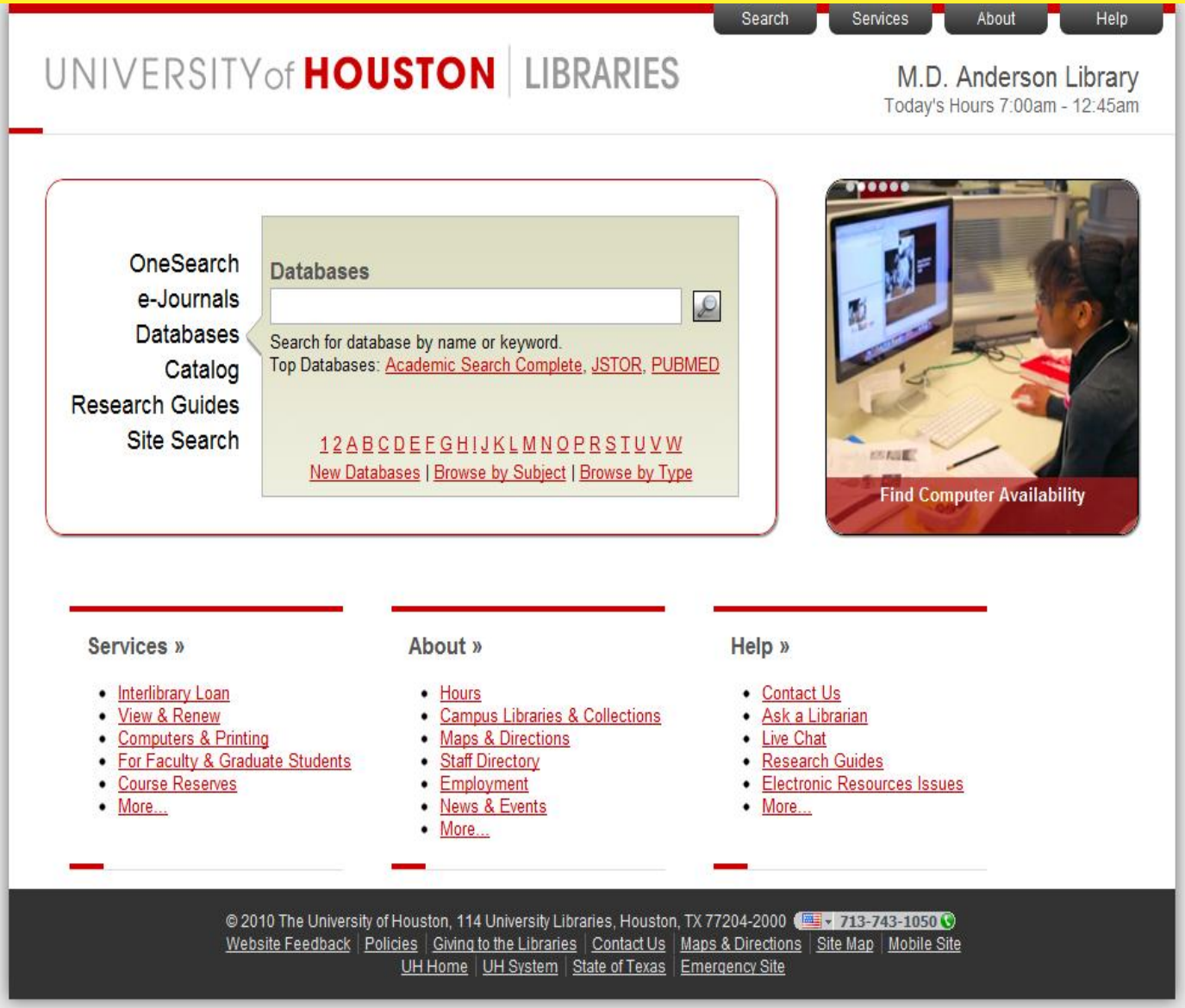
What are the genes involved in hemophilia A?

What are the mutations in hemophilia A?

Online Literature databases

1. How to use the UH online Library?
2. Online glossaries
3. Google Scholar
4. Google Books
5. Web of Science

How to use the online UH Library? http://info.lib.uh.edu/



UNIVERSITY of HOUSTON LIBRARIES

UH Library User Authentication

Access to this electronic resource is limited to current University of Houston main campus students, faculty an person to ensure that he or she uses these resources only for individual, noncommercial use without systematically downlo portions of the information provided. Furthermore, copying or tampering with the software that organizes and displays info agreements.

Misuse of these resources can result in the termination of the licensing agreement and the loss of use of the resource by the campus. Penalties for misuse can and will be imposed by the University of Houston. Please read the policy for Appropriat details: http://www.uh.edu/infotech/refguide/appropriate.html

Providers and the University of Houston Libraries offer these resources on an "as is basis." There are no warranties for co availability of storage and delivery devices. The University of Houston Libraries do not assume and hereby disclaim any lia resulting from the use of information in licensed resources.

By entering your login information, you agree to the above terms of use for the electronic resources provided by the UH L

Please login by entering your:

Last Name:

Your Library ID/Barcode:

What is my Library ID/Barcode?

Submit Query

Online Glossaries

Bioinformatics :
http://www.geocities.com/bioinformaticsweb/glossary.html
http://big.mcw.edu/

Genomics:
http://www.geocities.com/bioinformaticsweb/genomicglossary.html

Molecular Evolution:
http://workshop.molecularevolution.org/resources/glossary/

Biology dictionary:
http://www.biology-online.org/dictionary/satellite_cells

Other glossaries, e.g., the list of phobias:
http://www.phobialist.com/class.html

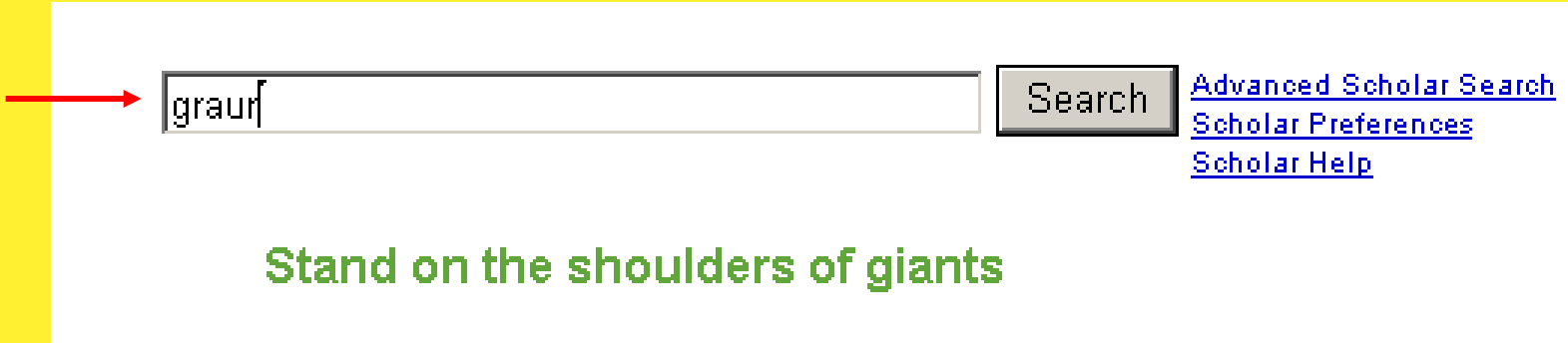
4. Google Scholar

http://www.scholar.google.com/

What is Google Scholar?

Enables you to search specifically for scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts and technical reports from all broad areas of research.

Use Google Scholar to find articles from a wide variety of academic publishers, professional societies, preprint repositories and universities, as well as scholarly articles available across the web.



Google Scholar orders your search results by how relevant they are to your query, so the most useful references should appear at the top of the page

This relevance ranking takes into account the: full text of each article. the article's author, the publication in which the article appeared and how often it has been cited in scholarly literature.

Scholar

[BOOK] Fundamentals of molecular evolution
WH Li, D **Graur** - 1991 - Sunderland, Mass.: Sinauer Associates
Cited by 372 - Web Search - Library Search

[BOOK] Fundamentals of molecular evolution
D **Graur**, WH Li - 2000 - Sunderland, Mass.: Sinauer Associates
Cited by 186 - Web Search - Library Search

Patterns of nucleotide substitution in pseudogenes and functional genes
T Gojobori, WH Li, D **Graur** - J. Mol. Evol, 1982 - ncbi.nlm.nih.gov
Patterns of nucleotide substitution in pseudogenes and functional genes. Gojobori T, Li WH, **Graur** D. MeSH Terms: Base Sequence; Codon; DNA/genetics*; Evolution* ...
Cited by 116 - Web Search

[CITATION] Extent of protein polymorphism and the neutral mutation theory
M Nei, D **Graur** - Evol. Biol, 1984
Cited by 86 - Web Search

Is the guinea-pig a rodent?
D **Graur**, WA Hide, WH Li - Nature, 1991 - ncbi.nlm.nih.gov
The guinea-pig (Cavia porcellus), traditionally classified as a New World hystricomorph rodent, often shows anomalous morphological and molecular ...
Cited by 86 - Web Search

Phylogenetic position of the order Lagomorpha(rabbits, hares and allies)
D **Graur**, L Duret, M Gouy - Nature, 1996 - ncbi.nlm.nih.gov
Ever since they have been classified as ruminants in the Old Testament (Leviticus 11:6, Deuteronomy 14:7) and equated with hyraxes in the vulgate ...
Cited by 78 - Web Search

What other DATA can we retrieve from the record?

Scholar

[BOOK] Fundamentals of molecular evolution
WH Li, D **Graur** - 1991 - Sunderland, Mass.: Sinauer Associates
Cited by 372 - Web Search - Library Search

Scholar Results **1** - **10** of about **359** citing **Li: Fundamentals of molecular evolution**. (**0.01** sec

Archaea and the prokaryote-to-eukaryote transition
JR Brown, WF Doolittle - Microbiology and Molecular Biology Reviews, 1997 - mmbr.asm.org
Page 1. M ICROBIOLOGY AND M OLECULAR B IOLOGY R EVIEWS , 1092-2172/97/$04.00
0 Dec. 1997, p. 456–502 Vol. 61, No. 4 Copyright © 1997 ...
Cited by 171 - Web Search - uprm.edu - plantbio.berkeley.edu - pubmedcentral.nih.gov - all 7 versions »

METALLOTHIONEIN: An Intracellular Protein to Protect Against Cadmium Toxicity
CD Klaassen, J Liu, S Choudhuri - Annual Review of Pharmacology and Toxicology, 1999 - pharmtox.annualreviews.org
Page 1. Annu. Rev. Pharmacol. Toxicol. 1999. 39:267–94 Copyright c 1999 by Annual
Reviews. All rights reserved METALLOTHIONEIN: An Intracellular ...
Cited by 142 - Web Search - unc.edu - ncbi.nlm.nih.gov - csa.com - all 6 versions »

Scholar

[BOOK] Fundamentals of molecular evolution
WH Li, D **Graur** - 1991 - Sunderland, Mass.: Sinauer Associates
Cited by 372 - Web Search - Library Search

Google

Web Images Groups News Froogle Local Scholar **more »**

"Li" "Fundamentals * molecular evolution" Search Advanced Search Preferences

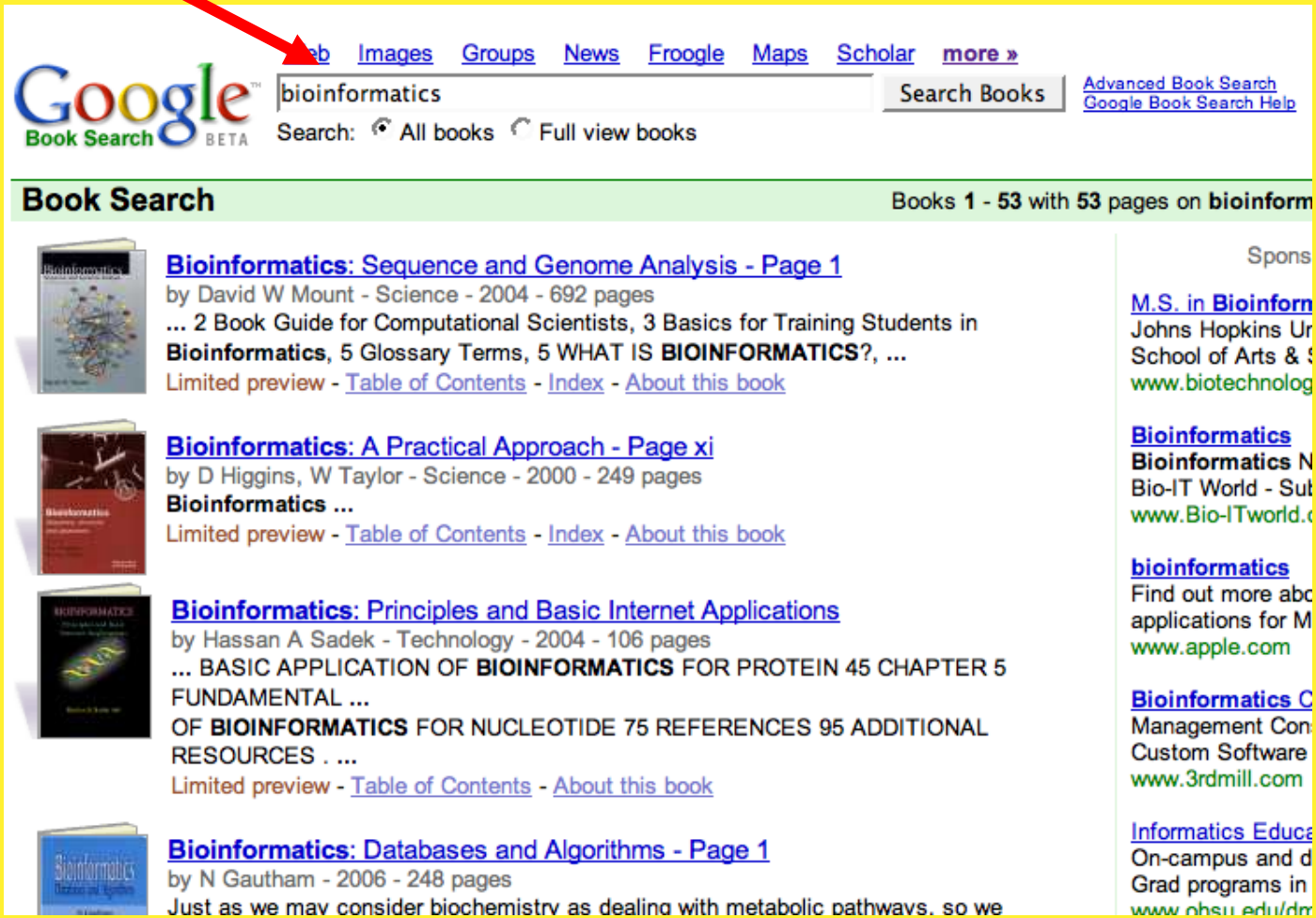
Web Results **1** - **7** of about **4,460** for **"Li" "Fundamentals * molecular evolution"**.

Li Lab -- Publications (Full List)
Li, W.-H. and D. Graur (1991) **Fundamentals of Molecular Evolution**, Sinauer Associates
... Graur, D. And W.-H. **Li** (1999) **Fundamentals of Molecular Evolution**, **...**
pondside.uchicago.edu/~**li**lab/pubs_full.htm - 65k - Cached - Similar pages

Scholar

[BOOK] Fundamentals of molecular evolution
WH Li, D **Graur** - 1991 - Sunderland, Mass.: Sinauer Associates
Cited by 372 - Web Search - Library Search

5. Google Book Search


Images Groups News Froogle Maps Scholar more »
Google Book Search BETA
bioinformatics
Search Books
Advanced Book Search
Google Book Search Help
Search: All books Full view books
Book Search
Books 1 - 53 with 53 pages on bioinform
Bioinformatics: Sequence and Genome Analysis - Page 1
by David W Mount - Science - 2004 - 692 pages
... 2 Book Guide for Computational Scientists, 3 Basics for Training Students in Bioinformatics, 5 Glossary Terms, 5 WHAT IS BIOINFORMATICS?, ...
Limited preview - Table of Contents - Index - About this book
Bioinformatics: A Practical Approach - Page xi
by D Higgins, W Taylor - Science - 2000 - 249 pages
Bioinformatics ...
Limited preview - Table of Contents - Index - About this book
Bioinformatics: Principles and Basic Internet Applications
by Hassan A Sadek - Technology - 2004 - 106 pages
... BASIC APPLICATION OF BIOINFORMATICS FOR PROTEIN 45 CHAPTER 5 FUNDAMENTAL ...
OF BIOINFORMATICS FOR NUCLEOTIDE 75 REFERENCES 95 ADDITIONAL RESOURCES . ...
Limited preview - Table of Contents - About this book
Bioinformatics: Databases and Algorithms - Page 1
by N Gautham - 2006 - 248 pages
Just as we may consider biochemistry as dealing with metabolic pathways, so we
Spons
M.S. in Bioinform
Johns Hopkins Ur
School of Arts &
www.biotechnolog
Bioinformatics
Bioinformatics N
Bio-IT World - Sub
www.Bio-ITworld.
bioinformatics
Find out more abo
applications for M
www.apple.com
Bioinformatics C
Management Con
Custom Software
www.3rdmill.com
Informatics Educa
On-campus and d
Grad programs in
www.obsu.edu/dm

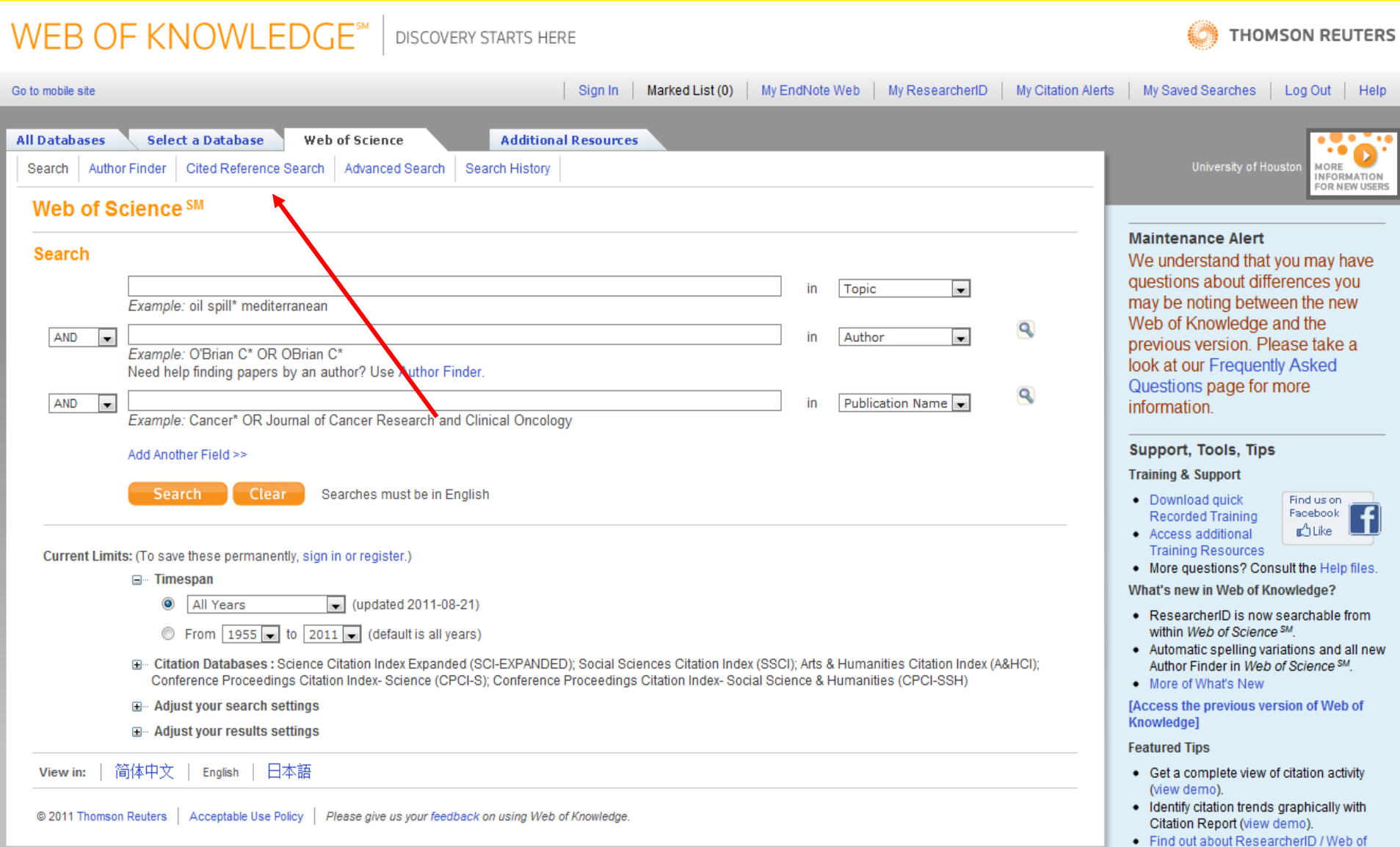
Start working:

Search Google Books

How many times is the tail of the giraffe mentioned in *On the Origin of Species* by Mr. Darwin?

6. Web of science

http://http://apps.webofknowledge.com.ezproxy.lib.uh.edu/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch&SID=4FB7LbbLgDMhG9fDiLh&preferencesSaved=



Web of Science℠

Cited Reference Search (Find the articles that cite a person's work)

View our Cited Reference Search tutorial.

Step 1: Enter information about the cited work. Fields are combined with the Boolean AND operator.

* Note: Entering the volume, issue, or page in combination with other fields may reduce the number of cited reference variants found.

graur d in Cited Author

Example: O'Brian C* OR OBrian C*

in Cited Work

Example: J Comp* Appl* Math* (journal abbreviation list)

in Cited Year(s)

Example: 1943 or 1943-1945

Add Another Field >>

Search | Clear | Searches must be in English

Current Limits: (To save these permanently, sign in or register.)

- **Timespan**
  - All Years (updated 2011-08-21)
  - From 1955 to 2011 (default is all years)
- **Citation Databases :** Science Citation Index Expanded (SCI-EXPANDED); Social Sciences Citation Index (SSCI); Arts & Humanities Citation Index (A&HCI); Conference Proceedings Citation Index- Science (CPCI-S); Conference Proceedings Citation Index- Social Science & Humanities (CPCI-SSH)
- **Adjust your search settings**
- **Adjust your results settings**

Web of Science SM

<< Back to previous page

Cited Reference Search (Find the articles that cite a person's work)

View our Cited Reference Search tutorial.

Step 2: Select cited references and click "Finish Search."

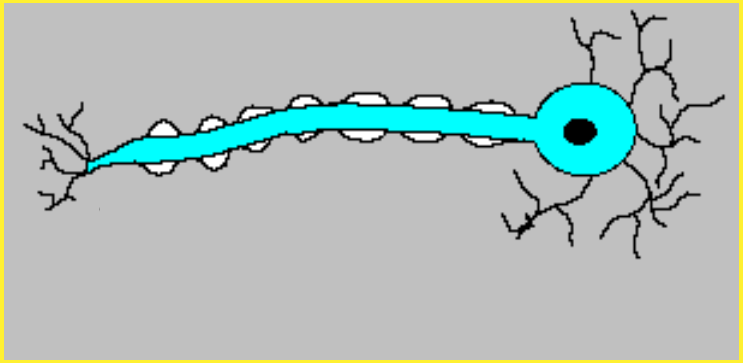
Hint: Look for cited reference variants (sometimes different pages of the same article are cited or papers are cited incorrectly).

CITED REFERENCE INDEX
References: **1 - 50 of 193**

Page 1 of 4 Go

Select Page | Select All* | Clear All | Finish Search

Select References	Cited Author	Cited Work [SHOW EXPANDED TITLES]	Year	Volume	Page	Article ID	Citing Articles **	View Record
☐	Abulafia, S...Graur, D	ISRAEL J PLANT SCI	1996	44	77		1	View Record
☐	Armon, A...Graur, D	J MOL BIOL	2001	307	447	10.1006/jmbi.2000.4474	198	View Record
☐	Arnason, U...Graur, D	J MOL EVOL	1996	43	41	10.1007/BF02352298	21	View Record
☐	Barki, Y...Graur, D	MAR BIOL	2000	136	37	10.1007/s002270050005	13	View Record
☐	Barki, Y...Graur, D	MAR ECOL-PROG SER	2002	231	91	10.3354/meps231091	17	View Record
☐	Ben-Dor, A...Graur, D	J COMPUT BIOL	1998	5	377	10.1089/cmb.1998.5.377	21	View Record
☐	BREIMAN, A...GRAUR, D	THEOR APPL GENET	1991	82	201		9	View Record
☐	BREIMAN, A...GRAUR, D	ISRAEL J PLANT SCI	1995	43	85		10	View Record
☐	Cohen, N...Graur, D	MOL BIOL EVOL	2005	22	1260	10.1093/molbev/msi115	37	View Record
☐	Crozier, R. H....Graur, D.	MOL PHYLOGENET EVOL	1995	4	20	10.1006/mpev.1995.1003	29	View Record
☐	Dagan, T...Graur, D	MOL BIOL EVOL	2006	23	310	10.1093/molbev/msj036	28	View Record
☐	Dagan, T...Graur, D	MOL BIOL EVOL	2005	22	496		9	View Record
☐	Dagan, T...Graur, D	MOL BIOL EVOL	2002	19	1022		29	View Record
☐	Dagan, T...Graur, D	NUCLEIC ACIDS RES	2004	32	D489	10.1093/nar/gkh132	36	View Record
☐	Elhaik, Eran...Graur, Dan	BIOL DIRECT	2010	5		ARTN 10	1	View Record
☐	Elhaik, Eran...Graur, Dan	COMPUT BIOL CHEM	2008	32	147	10.1016/j.compbiolchem.2007.11.003	4	View Record
☐	Elhaik, Eran...Graur, Dan	MOL BIOL EVOL	2010	27	1015	10.1093/molbev/msp307	2	View Record
☐	Elhaik, Eran...Graur, Dan	MOL BIOL EVOL	2009	26	1829	10.1093/molbev/msp100	6	View Record
☐	Elhaik, E...Graur, D	MOL BIOL EVOL	2006	23	1	10.1093/molbev/msj006	17	View Record



Databases

Sequence Databases

Bibliographic Databases

Clinical Databases

Integrated Databases

Structural Databases

Sequence Databases

Release/Up dates

Nucleotide Databases:

EMBL: European Molecular Biology Laboratory

Genbank

DDBJ: DNA Data Bank of Japan

International repository for all nucleotide sequences submitted by researchers

Current Release: 18,324,138 entries

Accession numbers are unique to each entry.

One alphabetical character is followed by five digits, or two alphabetical characters are followed by six digits.

The EMBL Nucleotide Sequence Database Stoesser G., Baker W., van den Broek A., Camon E., Garcia-Pastor M., Kanz C., Kulikova T., Lombard V., Lopez R., Parkinson H., Redaschi N., Sterk P., Stoehr P., Tuli MA

Sequence Databases

Nucleotide Databases:

RefSeq: Reference Sequence

Current Release: 93,285 entries

NC_123456

Complete Prokaryote Genome

Complete Eukaryote Chromosome

NG_123456

***Homo sapiens* Genomic Region**

A database of non-redundant reference sequences standards, including genomic DNAcontigs, mRNAs and proteins for known genes. Contributions are taken from the NCBI and collaborative sequencing efforts

NM_123456

mRNA of several organisms, including *Homo sapiens, Mus musculus, Rattus norvegicus*

Those accession numbers beginning with X indicate model entries produced as a result of the Genome Annotation process.

Sequence Databases

Protein Databases:

SwissProt: Swiss Protein

Current Release: 115,105 entries

Contains translated sequences from EMBL, adaptations from PIR, extracted from the literature and directly submitted by researchers. Annotation is high quality and the data is cross-referenced to other databases.

Entry names are often the name of the gene followed by the species.
Accession numbers are of the following format:

[O,P,Q] [0-9] [A-Z, 0-9] [A-Z, 0-9] [A-Z, 0-9] [0-9],

e.g. P26367 (PAX6_HUMAN)

Amos Bairoch and Rolf Apweiler "The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000", Nucleic Acids Res. 28:45-48(2000).

Sequence Databases

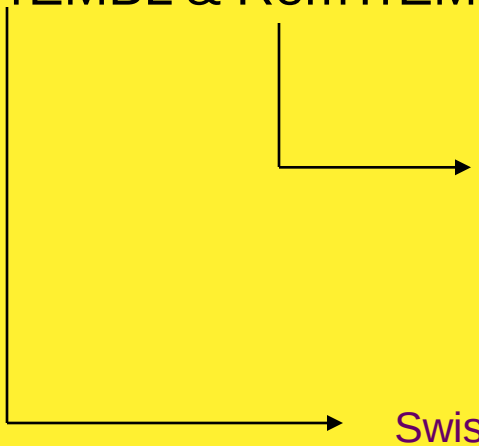
Protein Databases:

TrEMBL: Translated EMBL

Current Release: 632,013 entries

Acts as a supplement to SwissProt and contains translated EMBL sequences with automatic annotation. TrEMBL entries are manually annotated before being entered into SwissProt.

SpTrEMBL & RemTrEMBL



Remaining TrEMBL contains entries that will never be incorporated into SwissProt. These include: immunoglobulins; T-cell receptors; small fragments; synthetic sequences; CDS not coding for real proteins; patent application sequences

SwissProt TrEMBL contains entries which will eventually be integrated into the SwissProt database. SwissProt accession numbers have been assigned.

Sequence Databases

Protein Databases:

PIR: Protein Information Resource

Current Release: 283,175 entries

The PIR is a computer system offering both peptide an nucleotide sequences designed to aid protein identification.

Although much of the protein information in the PIR has been integrated into SwissProt, it may contain some unique sequences.

Sidman KE. George DG. Barker WC. Hunt LT (1988). The protein identification resource (PIR). Nucleic Acids Res 16:1869-71

Sequence Databases

Protein Databases:

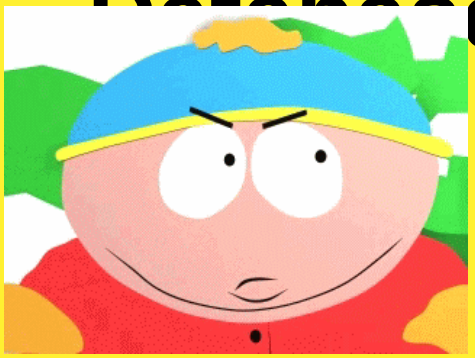
RefSeqP: Reference Sequence Proteins

Current Release: 402,006 entries

RefSeqP provides a protein reference standard for the central dogma. It is used, as is RefSeq, to provide a foundation for the functional annotation of the human genome

Accession numbers for all proteins are of the format: NP_123456

Sequence Databases



Searching for a sequence:

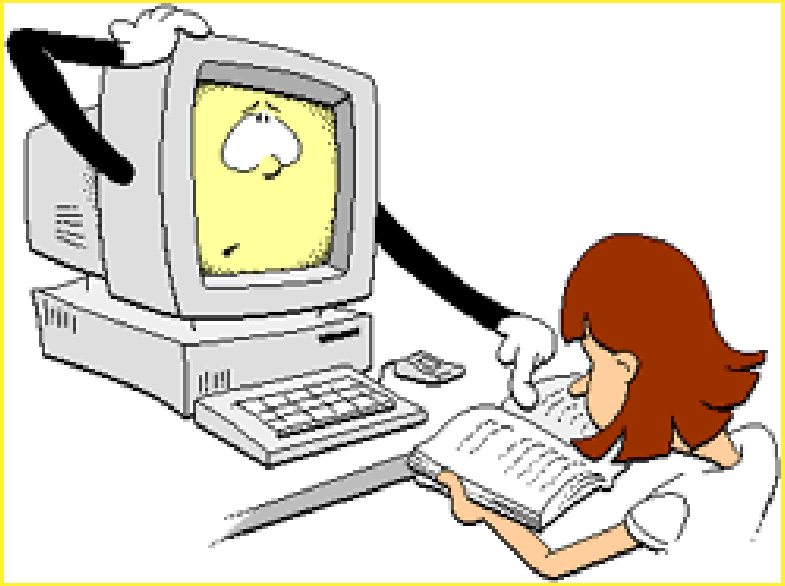
Text Search:

Use text with a boolean operator

BRCA1 & BRCA2 – searches for BRCA1 **AND** BRCA2

BRCA1 | BRCA2 – searches for one gene **OR** the other

BRCA1 ! BRCA2 – searches for BRCA1 **BUT NOT** BRCA2



Computers are

THICK!

Database entries often presented as **flatfiles**

Each piece of information is on a separate line, distinguished by a code. Computers index this code, so they can search for the relevant entry.

EMBL entry for a sequence fragment implicated in Human Breast Cancer

Identification	ID	AY144588 standard; DNA; HUM; 68 BP.
Accession	AC	AY144588;
Sequence Version	SV	AY144588.1
Date	DT	23-SEP-2002 (Rel. 73, Created)
	DT	23-SEP-2002 (Rel. 73, Last updated, Version 1)
Description	DE	Homo sapiens truncated breast and ovarian cancer susceptibility protein
	DE	(BRCA1) gene, partial cds.
Keyword	KW	.
Organism Source	OS	Homo sapiens (human)
Organism Classification	OC	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
	OC	Eutheria; Primates; Catarrhini; Hominidae; Homo.

Reference Number	RN	[1]
Reference Position	RP	1-68
Reference Author	RA	Rajkumar T., Soumittra N., Nirmala Nancy K., Shanta V.;
Reference Title	RT	"Novel 5bp deletion in BRCA1 gene in South Indian family";
Reference Location	RL	Unpublished.
	RN	[2]
	RP	1-68
	RA	Rajkumar T., Soumittra N., Nirmala Nancy K., Shanta V.;
	RT	;
	RL	Submitted (27-AUG-2002) to the EMBL/GenBank/DDBJ databases.
	RL	Molecular Oncology, Cancer Institute (WIA), Canal Bank Road, Adyar, RL Chennai, TN 600020, India

Feature Table Header

Feature Table Data

```
FH  Key   Location/Qualifiers
FH
FT          source 1..68
FT          /country="India: South India“
FT          /db_xref="taxon:9606"
FT          /note="identical sequence found in daughter with
breast
FT          cancer"
FT          /sex="female"
FT          /organism="Homo sapiens"
FT          /isolation_source="mother with breast cancer"
FT          /dev_stage="adult"
FT          /mRNA 68
FT          /gene="BRCA1"
FT          /product="truncated breast and ovarian cancer
FT          susceptibility protein"
```

FT CDS <1..68

FT /codon_start=3

FT /note="contains premature stop codon due to frameshift

FT caused by deletion"

FT /product="truncated breast and ovarian cancer

FT susceptibility protein"

FT /protein_id="AAN10167.1"

FT /translation="EAASGCESETSVSEDCSGLSE"

FT exon 1..68

FT /number=12

FT /gene="BRCA1"

FT misc_feature 61..62

FT /note="site of deletion"

FT /gene="BRCA1"

Sequence Header

SQ Sequence 68 BP; 19 A; 12 C; 23 G; 14 T; 0 other;

gtgaagcagc atctgggtgt gagagtgaaa caagcgtctc tgaagactgc tcagggctat 60

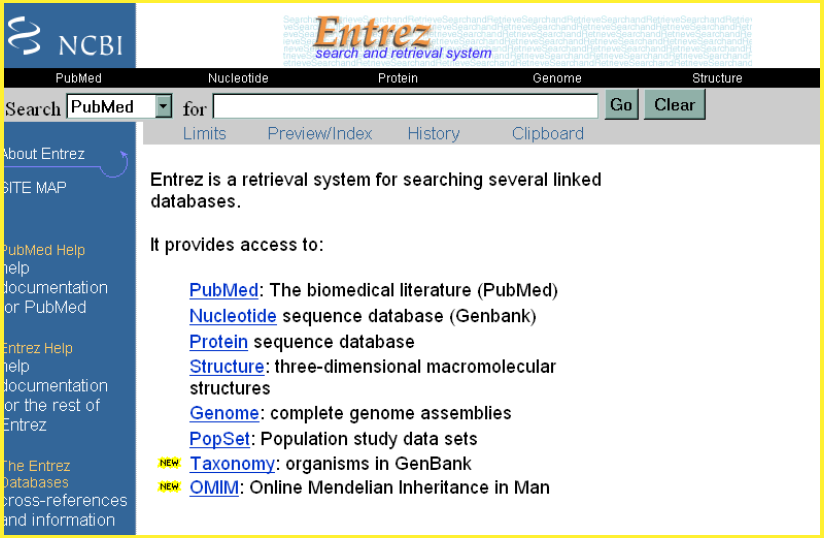
cagagtga 68

//

Searching the databases with a "search engine":

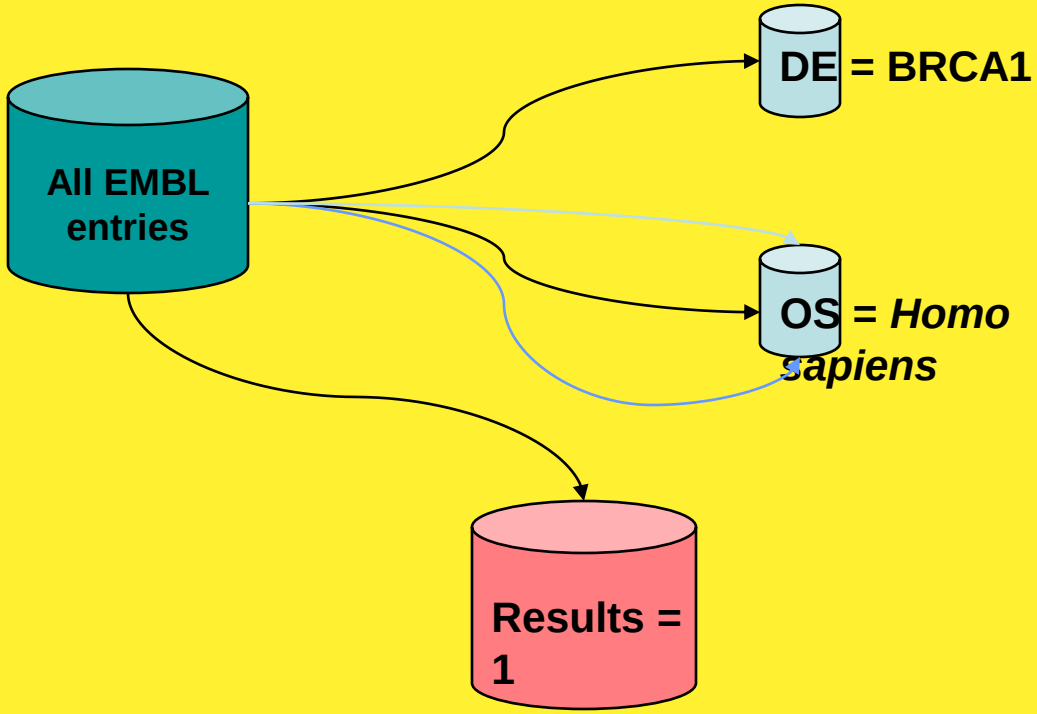
The Sequence Retrieval System (SRS) from LION Bioscience AG is a very common search tool

The NCBI in the USA has its own search engine called Entrez.



To search for the BRCA1 gene in Homo sapiens in the EMBL database:

BRCA1 [DE] & Human [OC]

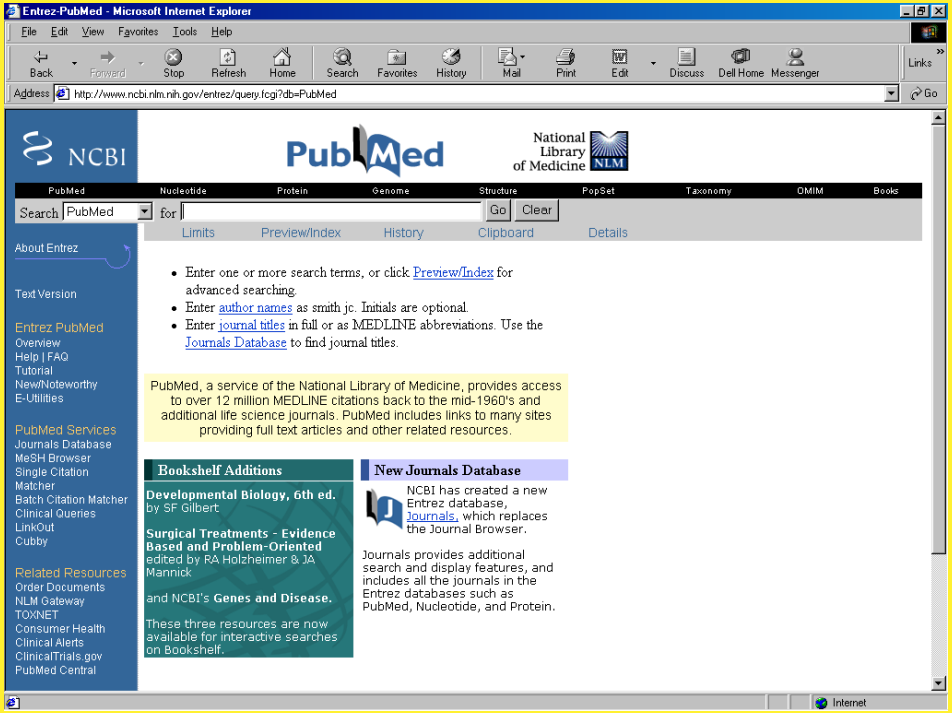


Bibliographic Databases

Used for searching for reference articles

For all (loosely) medically related papers, use PubMed from the NCBI

Currently holds over 12 million MEDLINE entries.



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

Bibliographic Databases

Other scientific databases may include:

Web of Science: http://wos.mimas.ac.uk

Free to academics, but requires username and password

PubCrawler: http://www.pucrawler.ie

Free to academics, will search journals and sequences daily, weekly or mor
and alert the user when results are found corresponding to their search

Clinical Databases

Generally contain information from the Human.

Human Gene Mutation Database, Cardiff, UK:
http://www.hgmd.org

Registers known mutations in the human genome and the diseases they cause.

dbSNP, Bethesda, USA:
http://ncbi.nlm.nih.gov/SNP/

The largest database for single nucleotide polymorphisms. Accession numbers used in dbSNP are not compatible with other SNP databases.

Integrated Databases

These contain overview information garnered from a variety of different databases, and then offer links to further information.

GeneCards: http://bioinformatics.weizmann.ac.il/cards

An extremely thorough overview of a particular gene, with links to various other integrated and clinical databases.

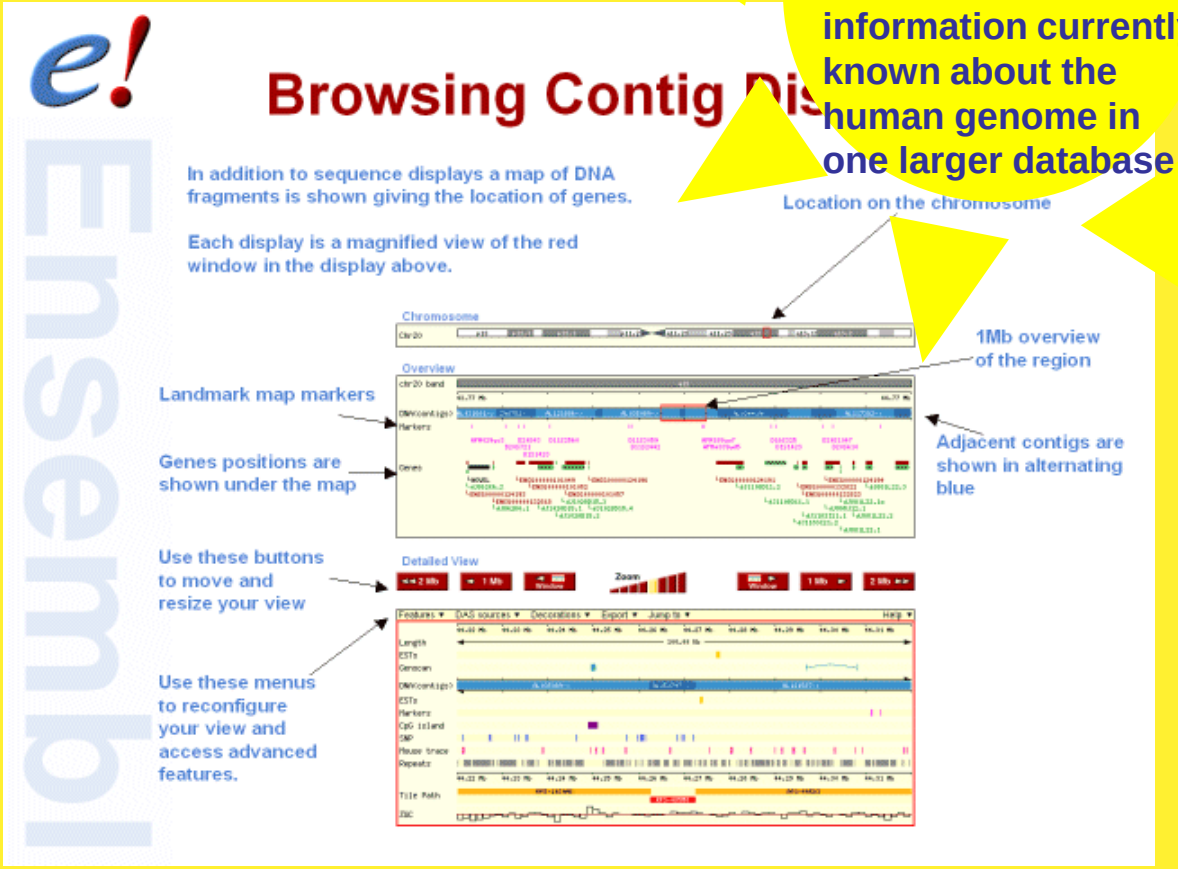
Interpro: http://www.ebi.ac.uk/interpro

Integration of individual protein resources PRINTS; PROSITE; SMART; ProDom; Pfam; TIGRfam into one database. A search will scan entries of each and output results.

Integrated Databases

Ensembl: http://www.ensembl.org

A joint project by EBI and Sanger to annotate all the information currently known about the human genome in one larger database



Structural Databases

Tertiary protein structure prediction is possibly the Holy Grail of bioinformatics.

PDB: Protein DataBank, New Jersey, USA

http://www.rcsb.org/

This houses a collection of 3D coordinates of each atom in a protein, allowing the structure to be displayed by viewing software. Protein structures are submitted by individual researchers and have been determined by x-ray diffraction, or NMR.

EMSD: EBI Macromolecular Structure Database

http://www.ebi.ac.uk/msd/index.html

Management and distribution of data on macromolecular structures in close collaboration with the PDB.

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, **28** pp. 235-242 (2000)

Structural Databases

SCOP: Structural Classification of Proteins

http://scop.mrc-lmb.cam.ac.uk/scop/

Current Release: 686 folds; 1073 Superfamilies; 1827 Familes representing 15,979 PDB entries

CATH: Classification, Architecture, Topology, Homology

http://www.biochem.ucl.ac.uk/bsm/cath_new/

Current Release: 36,480 Domains

Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540

Alignment III
PAM Matrices

PAM250 scoring matrix

Cys	12																			
Ser	0	2																		
Thr	-2	1	3																	
Pro	-1	1	0	6																
Ala	-2	1	1	1	2															
Gly	-3	1	0	-1	1	5														
Asn	-4	1	0	-1	0	0	2													
Asp	-5	0	0	-1	0	1	2	4												
Glu	-5	0	0	-1	0	0	1	3	4											
Gln	-5	-1	-1	0	0	-1	1	2	2	4										
His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Scoring Matrices

S = [s_{ij}] gives score of aligning character i with character j for every pair i, j.

	C	S	T	P	A
C	12				
S	0	2			
T	-2	1	3		
P	-3	1	0	6	
A	-2	1	1	1	2

STPP
CTCA

0 + 3 + (-3) + 1

= 1

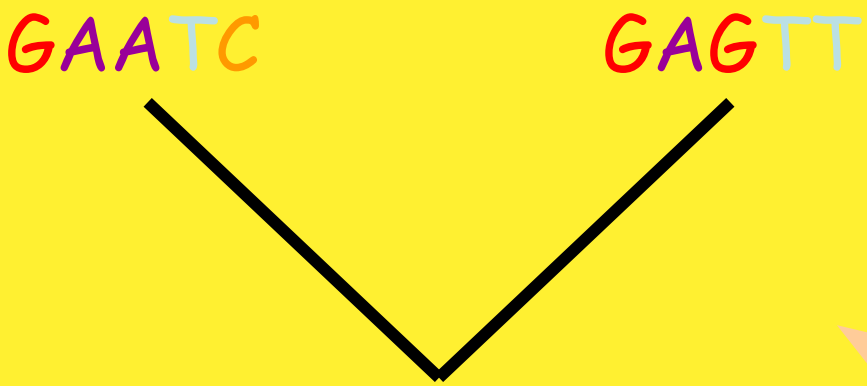
Scoring with a matrix

- Optimum alignment (global, local, end-gap free, etc.) can be found using dynamic programming
  - No new ideas are needed
- Scoring matrices can be used for any kind of sequence (DNA or amino acid)

Types of matrices

- PAM
- BLOSUM
- Gonnet
- JTT
- DNA matrices
- PAM, Gonnet, JTT, and DNA PAM matrices are based on an explicit evolutionary model; BLOSUM matrices are based on an implicit model

PAM matrices are based on a simple evolutionary model



Two changes

- Only mutations are allowed
- Sites evolve independently

Log-odds scoring

- What are the odds that this alignment is meaningful?

  $X_1 X_2 X_3 \ldots X_n$
  $Y_1 Y_2 Y_3 \ldots Y_n$

- **Random model:** We're observing a chance event. The probability is

$$\prod_i p_{X_i} \prod_i p_{Y_i}$$

  where p_X is the frequency of X

- **Alternative:** The two sequences derive from a common ancestor. The probability is

$$\prod_i q_{X_i Y_i}$$

  where q_{XY} is the joint probability that X and Y evolved from the same ancestor.

Log-odds scoring

- Odds ratio:

$$\frac{\prod_i q_{X_i Y_i}}{\prod_i p_{X_i} \prod_i p_{Y_i}} = \prod_i \frac{q_{X_i Y_i}}{p_{X_i} p_{Y_i}}$$

- Log-odds ratio (score):

$$S = \sum_i s(X_i, Y_i)$$

where

$$s(X, Y) = \log\left(\frac{q_{XY}}{p_X p_Y}\right)$$

is the **score** for X, Y. The $s(X,Y)$'s define a **scoring matrix**

PAM matrices: Assumptions

- Only mutations are allowed
- Sites evolve independently
- Evolution at each site occurs according to a simple ("first-order") Markov process
  - Next mutation depends only on current state and is independent of previous mutations
- Mutation probabilities are given by a **substitution matrix** $M = [m_{XY}]$, where m_{xy} = Prob(X $\rightarrow$ Y mutation) = Prob(Y|X)

PAM substitution matrices and PAM scoring matrices

- Recall that

$$s(X,Y) = \log\left(\frac{q_{XY}}{p_X p_Y}\right)$$

- Probability that X and Y are related by evolution:

$$q_{XY} = \text{Prob(X)} \cdot \text{Prob(Y|X)} = p_x \cdot m_{XY}$$

- Therefore:

$$s(X,Y) = \log\left(\frac{m_{XY}}{p_Y}\right)$$

Mutation probabilities depend on evolutionary distance

- Suppose *M* corresponds to one unit of evolutionary time.
- Let *f* be a frequency vector (f_i = frequency of a.a. *i* in sequence). Then
  - $M \cdot f$ = frequency vector after one unit of evolution.
  - If we start with just amino acid *i* (a probability vector with a 1 in position *i* and 0s in all others) column *i* of *M* is the probability vector after one unit of evolution.
  - After *k* units of evolution, expected frequencies are given by $M^k \cdot f$.

PAM matrices

- **P**ercent **A**ccepted **M**utation: Unit of evolutionary change for protein sequences [Dayhoff78].
- A PAM unit is the amount of evolution that will on average change 1% of the amino acids within a protein sequence.

PAM matrices

- Let *M* be a PAM 1 matrix. Then,

$$\sum_i p_i (1 - M_{ii}) = 0.01$$

- **Reason:** M_{ii}s are the probabilities that a given amino acid does not change, so (1-M_{ii}) is the probability of mutating away from *i*.

The PAM Family

Define a ***family*** of substitution matrices — PAM 1, PAM 2, etc. — where PAM n is used to compare sequences at distance n PAM.

PAM n = $(\text{PAM 1})^n$

Do not confuse with scoring matrices!

Scoring matrices are derived from PAM matrices to yield log-odds scores.

Generating PAM matrices

- **Idea:** Find amino acids substitution statistics by comparing evolutionarily close sequences that are highly similar
  - Easier than for distant sequences, since only few insertions and deletions took place.
- Computing PAM 1 (Dayhoff's approach):
  - Start with highly similar aligned sequences, with known evolutionary trees (71 trees total).
  - Collect substitution statistics (1572 exchanges total).
  - Let m_{ij} = observed frequency (= estimated probability) of amino acid A_i mutating into amino acid A_j during one PAM unit
  - Result: a 20× 20 real matrix where columns add up to 1.

Dayhoff's PAM matrix

	A	R	N	D	C
A	9867	2	9	10	3
R	1	9913	1	0	1
N	4	1	9822	36	0
D	6	0	42	9859	0
C	1	1	0	0	9973

All entries $\times 10^4$