CS174: Other Topics in Bioinformatics

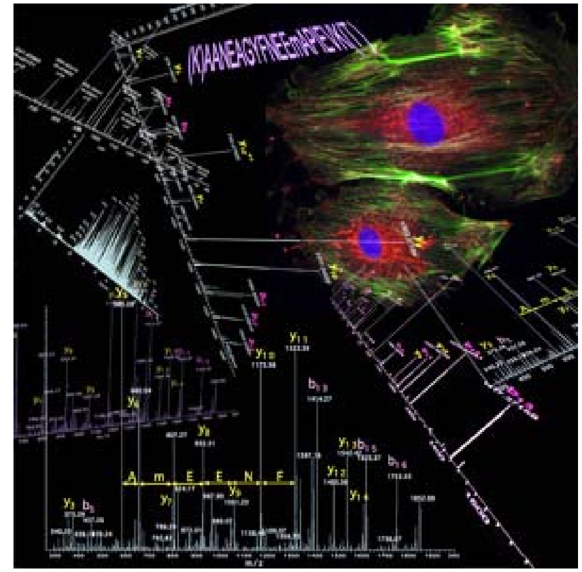
Topics covered so far

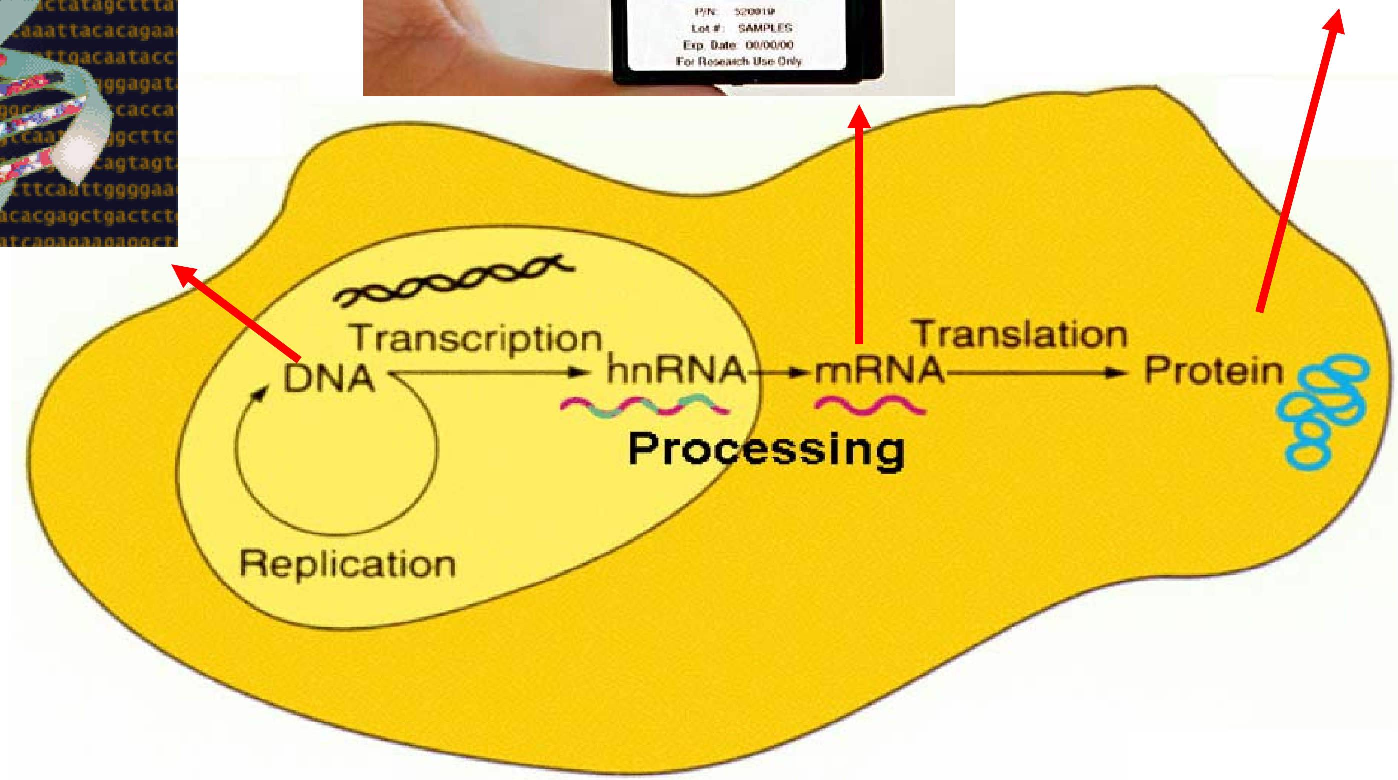
- Gene (ORF) discovery
- Regulatory motif discovery
- Sequence alignments
- Genome assembly
- Hidden Markov model

Other Topics:

- Comparative Genomics
- Protein structure prediction
- Systems biology: clustering, reverse-engineering approaches
- Evolutionary theory
- Population dynamics

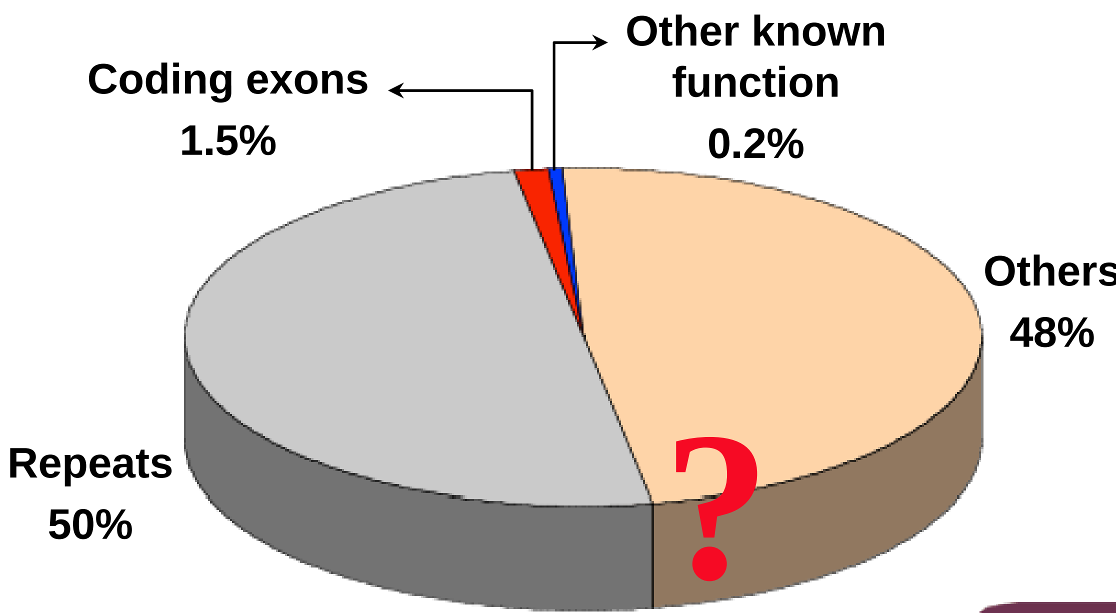
Readout from the genome



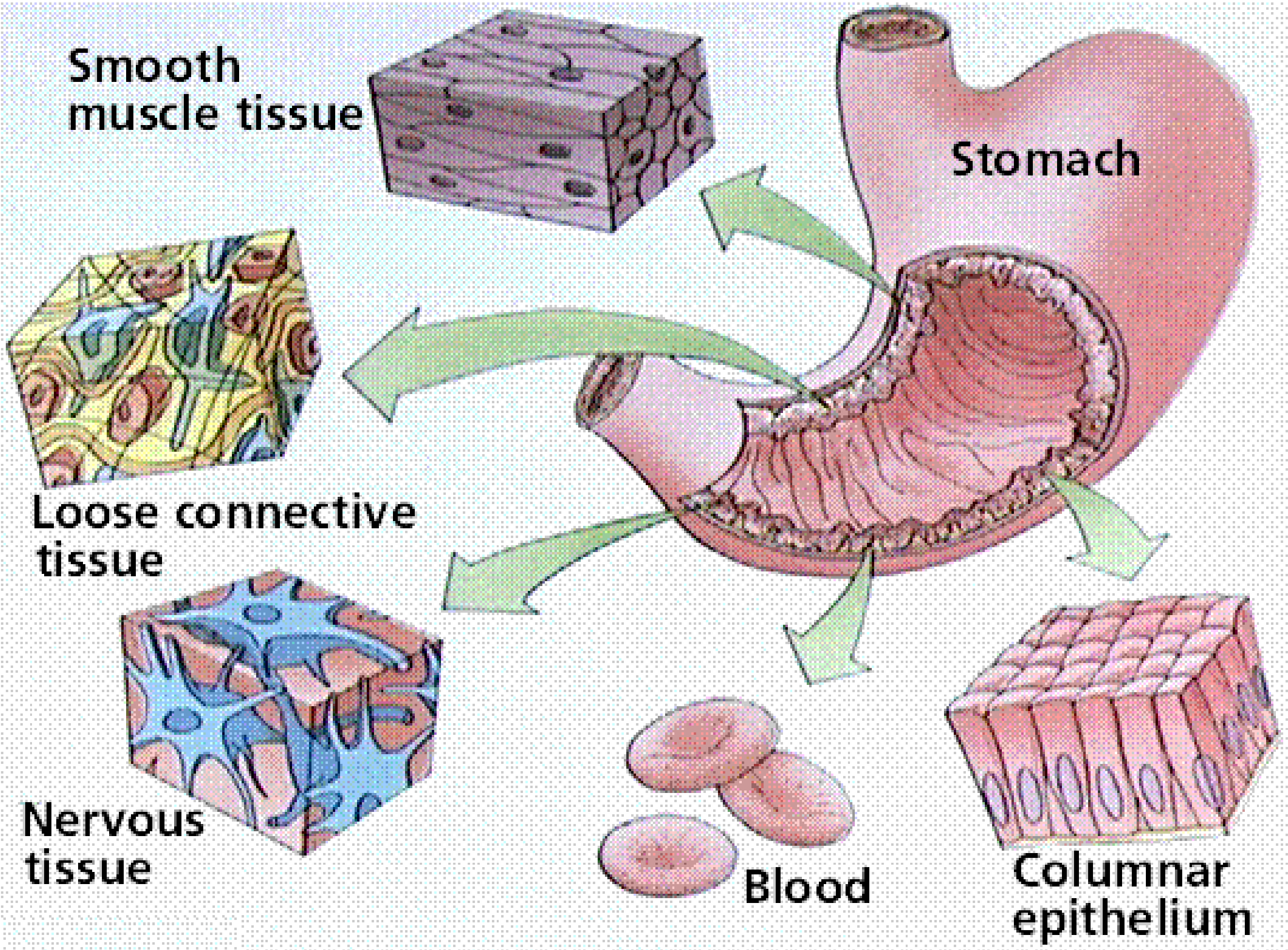


98% of the human genome unknown

Human
Genome
~3Gb



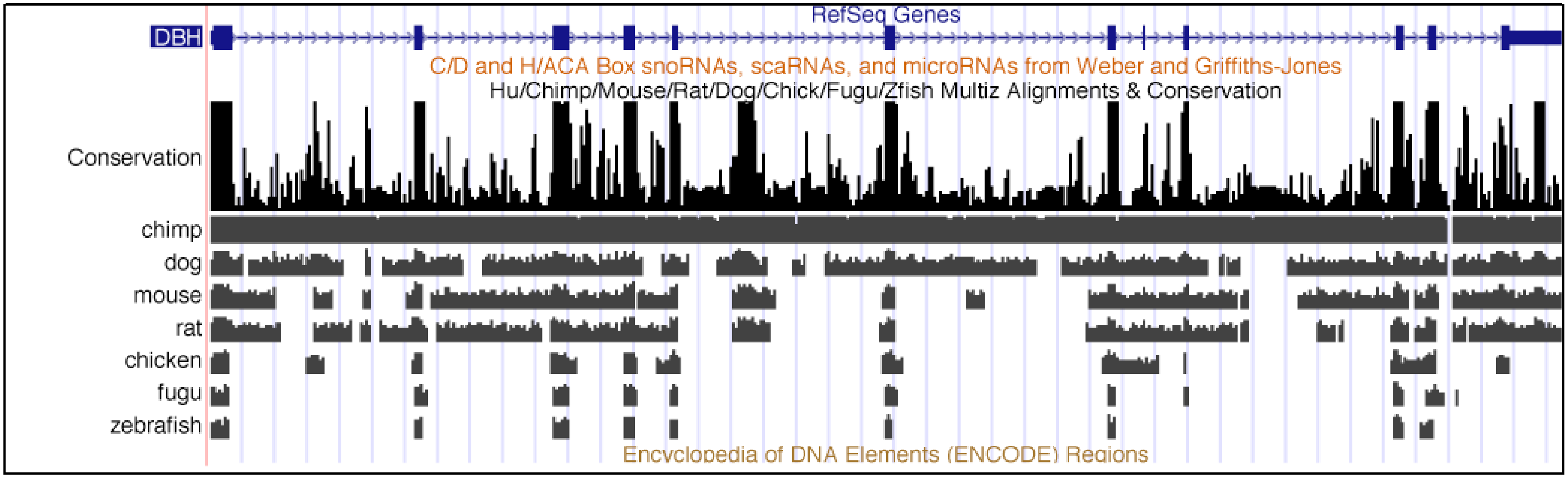
Example: Tissues in Stomach



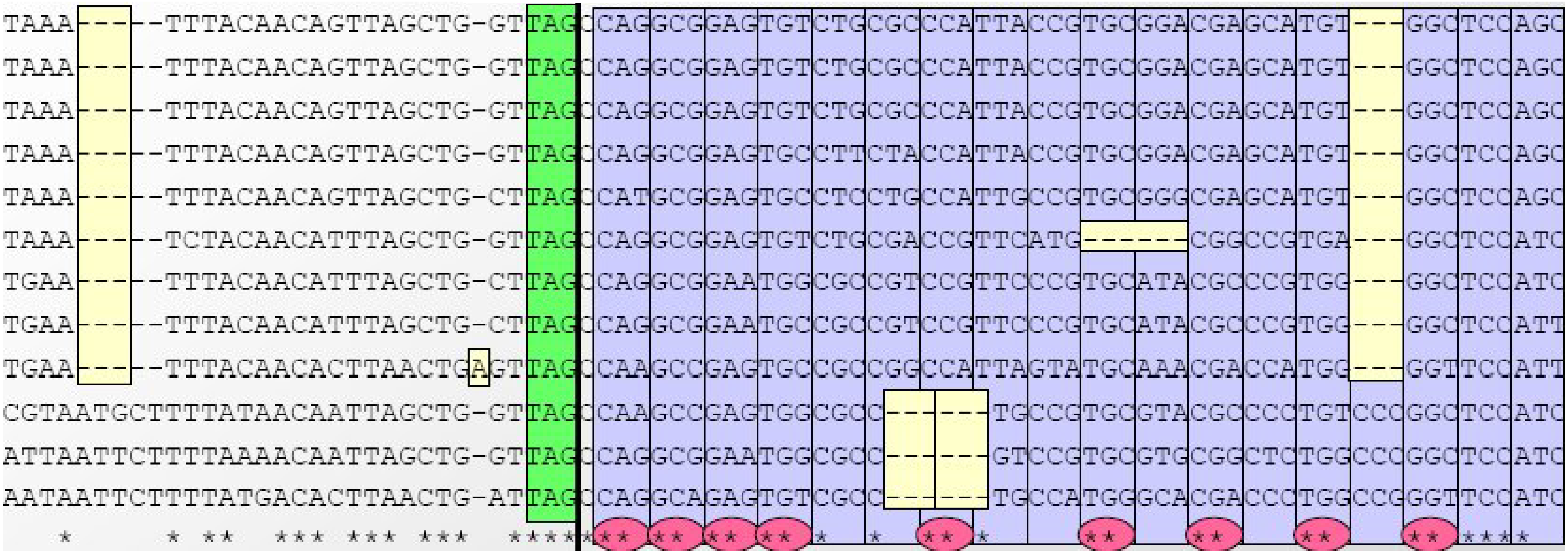
How is this variety encoded and expressed ?

Comparative Genomics

Comparative genomics and evolutionary signatures



- ✓ **Comparing genomes can reveal functional elements**
- ✓ **Can we also pinpoint specific functions of each elements? Yes!**
  - Patterns of change distinguish different types of functional elements
  - Specific function ⟷ Selective pressures ⟷ Patterns of mutations/indels
- ✓ **Develop evolutionary signatures characteristic of each function**



Signatures specific to protein genes:

- Indels are multiples of three
- Mutations are largely 3-periodic
- Conservation boundaries are sharp (splicing signals)

Evolutionary Signatures: Genes

```
hg18.chr15   AGACTAGTGATTTTGTTGTTGTCTTAC--TGC-GCTCAACAACAAATCCCAGTCT
panTro1      AGACTAGTGATTTTGTTGTTGTCTTAC--TGC-GCTCAACAACAAATCCCAGTCT
rheMac2      AGACTAGTGATTTTGTTGTTGTCTTAC--TGC-GCTCAACAACAAATCCCAGTCT
mm8          AGACTAGTGATTTTGTTGTTGTGTCTC--TGT-ATCCAACAACAAGTCCCAGTCT
oryCun1      AGACTAGTGATTTTGTTGTTGTCTCGC--TGT-GCTCAACAACAAGTCCCAGTCT
rn4          AGACTAGTGATTTTGTTGTTGTG--TC--TGT-GTCCAACAACAAGTCCCAGTCT
bosTau2      AGACTAGTGATTTTGTTGTTGTCTC-C--TGC-GCTCAACAACAAGTCCCAGTCT
canFam2      AGACTAGTGATTTTGTTGTTGTCTCAC--TGC-ATCCAACAACAAGTCCCAGTCT
dasNov1      AGACTAGTGATTTTGTTGTTGTCTCACTGTGC-GCTCAACAACAAGTCCCAGTCT
loxAfr1      AGACTAGTGATTTTGTTGTTGTCTCAT--TACCGTTCAACAACAAGTCCCAGTCT
echTel1      AGACTAGTGATTTTGTTGTTGTCTCGC--TACTGCTCAACAAC............
monDom4      AGACTAGTGATTTTGTTGTTTCTCTAACGTAAA-GATTGACAACAAATCCCAGTCT
```

has-mir-7

Signatures specific to RNA genes:

- Stem conservation >> loop conservation
- Compensatory changes for paired bases
- Gaps allowed

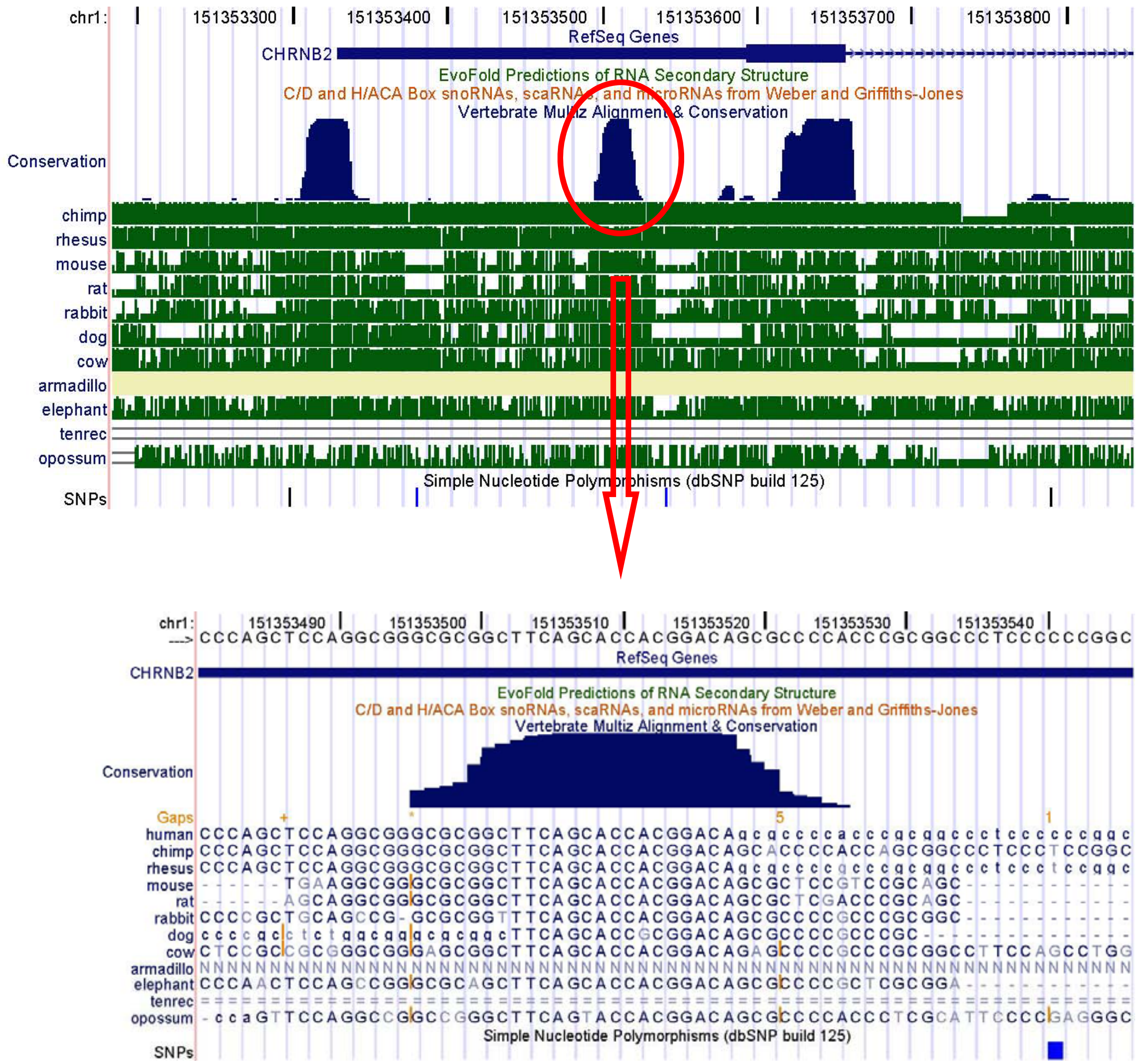
Evolutionary Signatures: RNA genes

Known motifs are preferentially conserved

```
human     CTCTTAATGGTACACGTTCTGCCT----AAGTAGCCTAGACGCTCCCGTGCGCCC-GGGG
dog       CTCTTA-CGGGGCACATTCTGCTTTCAACAGTGGGGCAGACGGTCCCGCGCGCCCCAAGG
mouse     GTCTTAGGAGGCT-CGATCGCC---------------------GCCTGCATTATT-----
rat       GTCTTAGTTGGCCACGACCTGC---------------------TCATGCATAATT-----
           *****   *    *   *  *                      *  *
                                             Errα
human     CGGGTAGGCCTGGCCGAAAATCTCTCCCGCGCGCCTGACCTTGGGTTGCCCCAGCCAGGC
dog       CAGGC---CCGGGCTGCAGACCTGCCCTGAGGGAATGACCTTGGGCGGCCGCAGCGGGGC
mouse     --------------CACAAGCCTGTGGCGCGC-CGTGACCTTGGGCTGCCCCAGGCGGGC
rat       --------------CACAAGTTTCTC---TGC-CCTGACCTTGGGTTGCCCCAGGCGAG-
                           *    *       *    **********  *** ***    *

human     TGCGGGCCCGAGACCCCCG-------------------GGCCTCCCTGCCCCCCGCGCCG
dog       CGCGGGCCCAGGCCCCCCTCCCTCCCTCCCTCCCTCCCTCCCTCCCTGCCCCCCGGACCG
mouse     TGCAGGCTCACCACCCCGTCTTTTCT---------------------GCTTTTCGAGTCG
rat       -GCATACACCCCGCCTTTTTTTTTTTTTT---------TTTTTTTTTGCCGTTCAAG-AG
           **   * *    **                                **    *     *
```

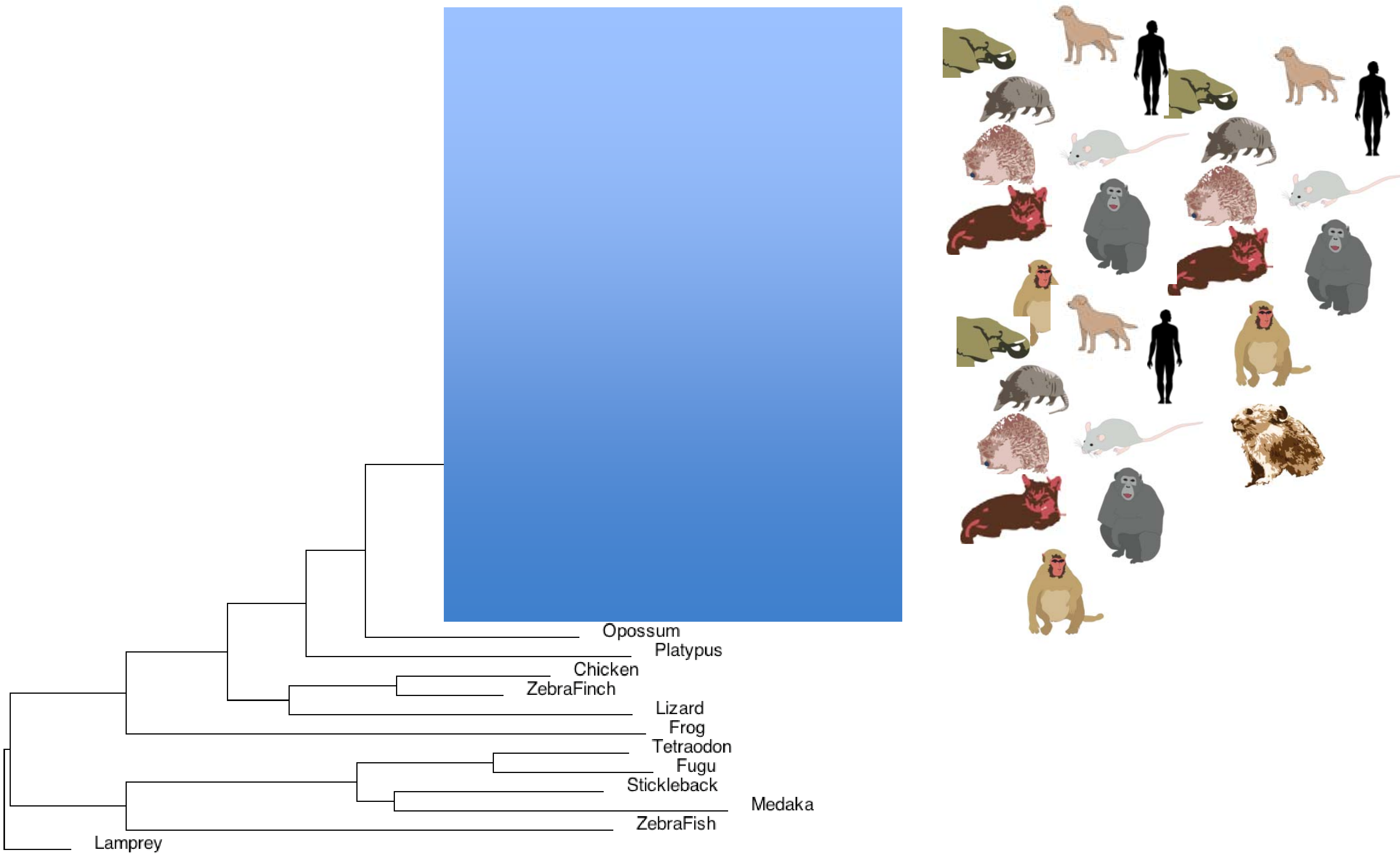
Gabpa

Evolutionary Signatures: Regulatory Motifs

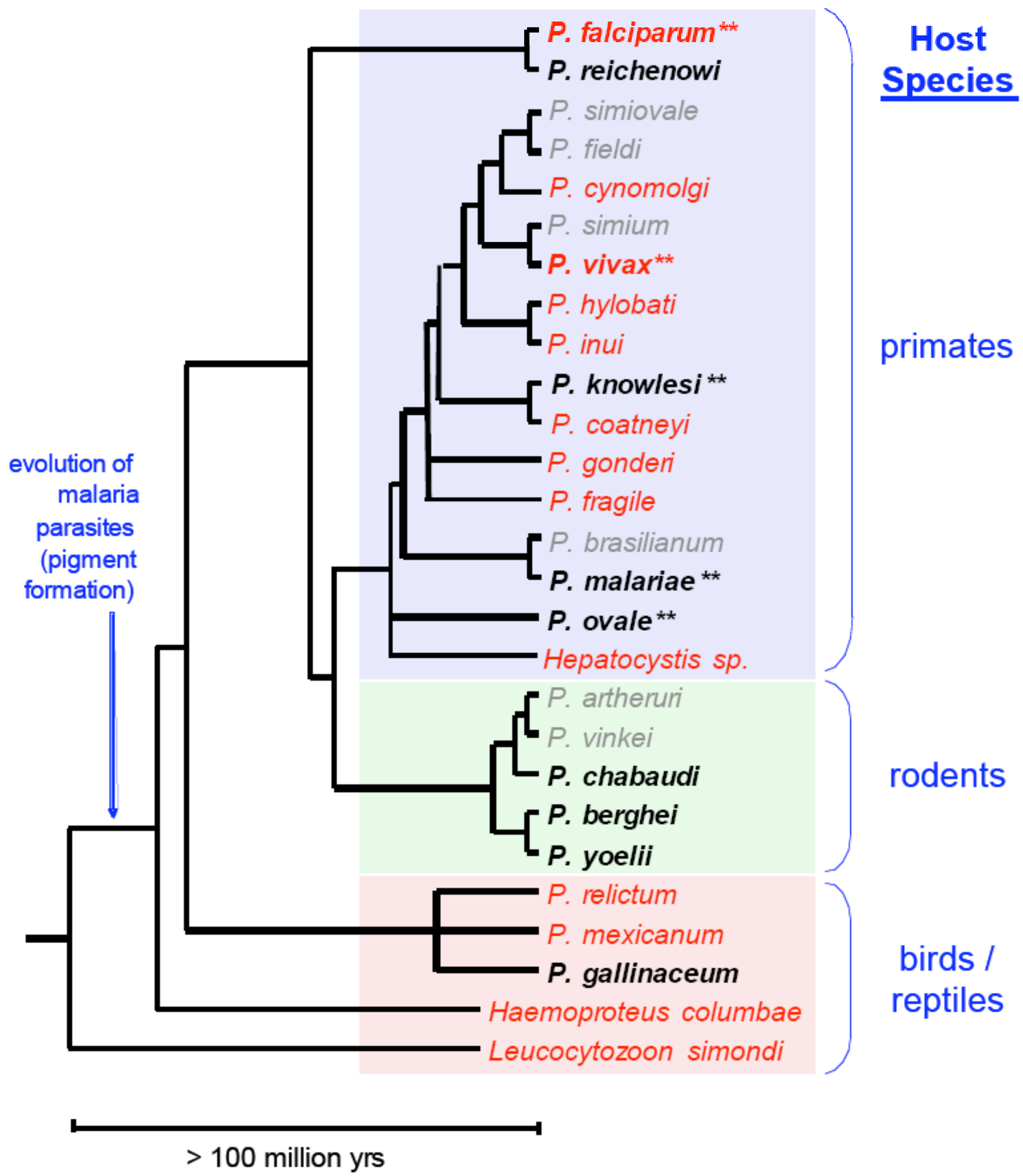


cholinergic receptor, nicotinic, beta 2 (neuronal)

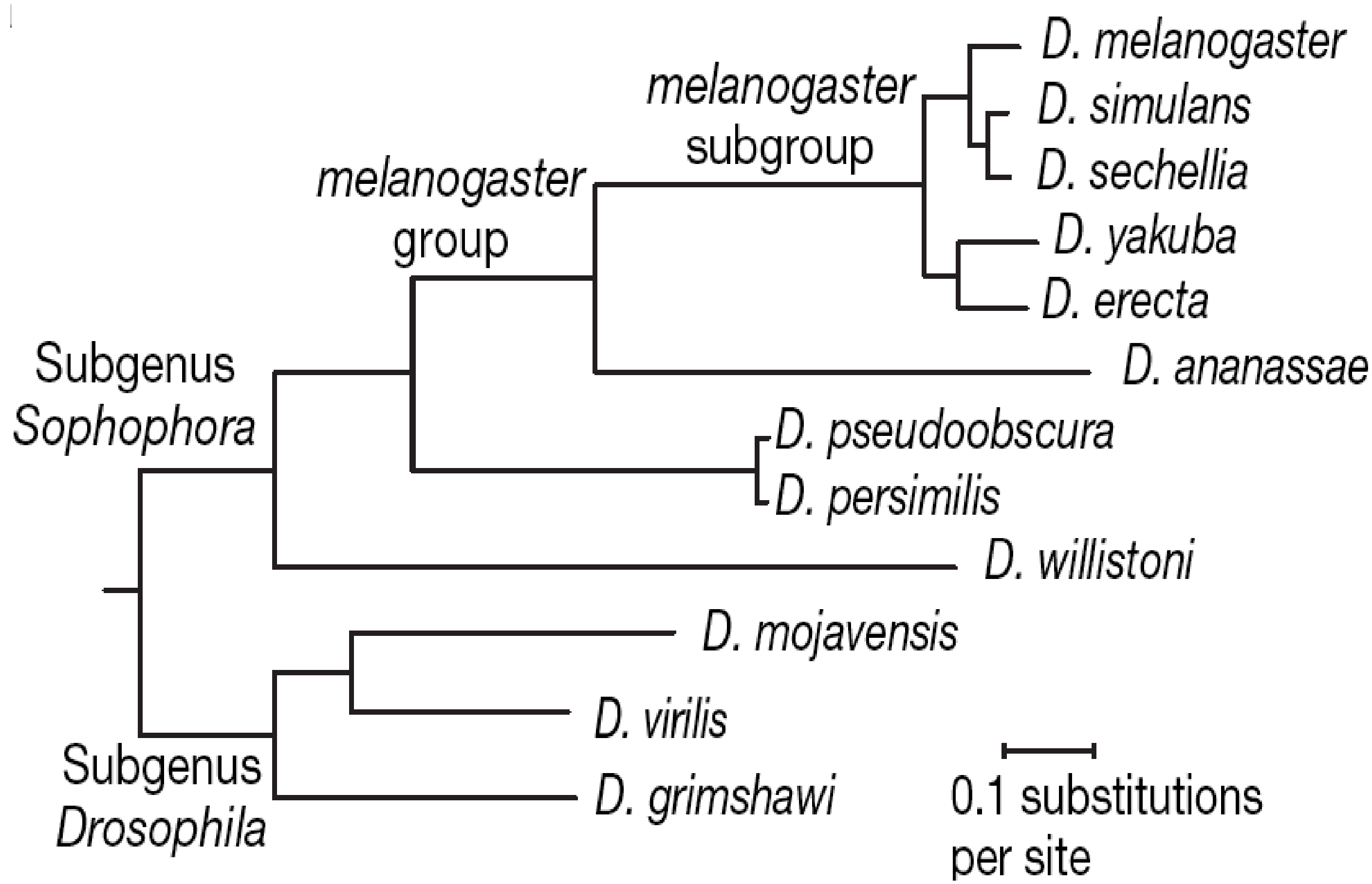
2009: 29 of the 44 vertebrates sequenced are eutherian mammals



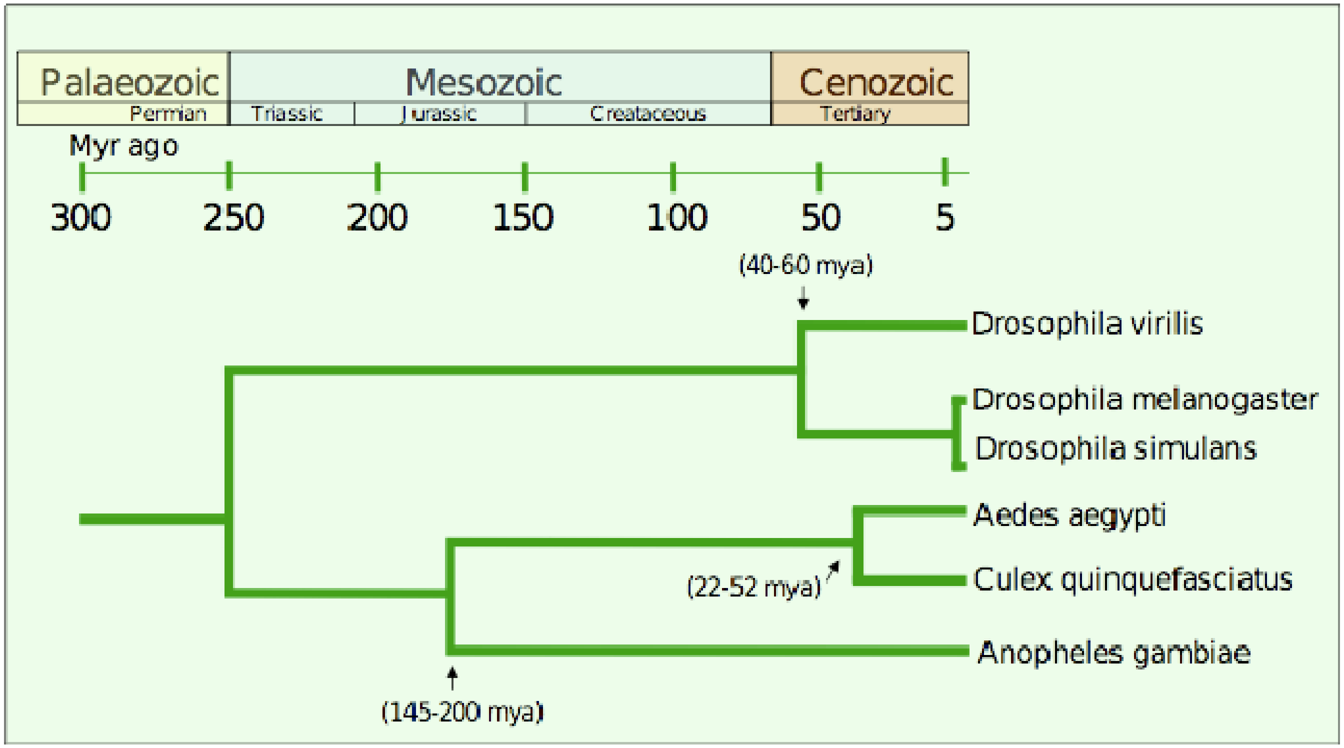
Plasmodium genomes



12 fly genomes



Mosquito genomes



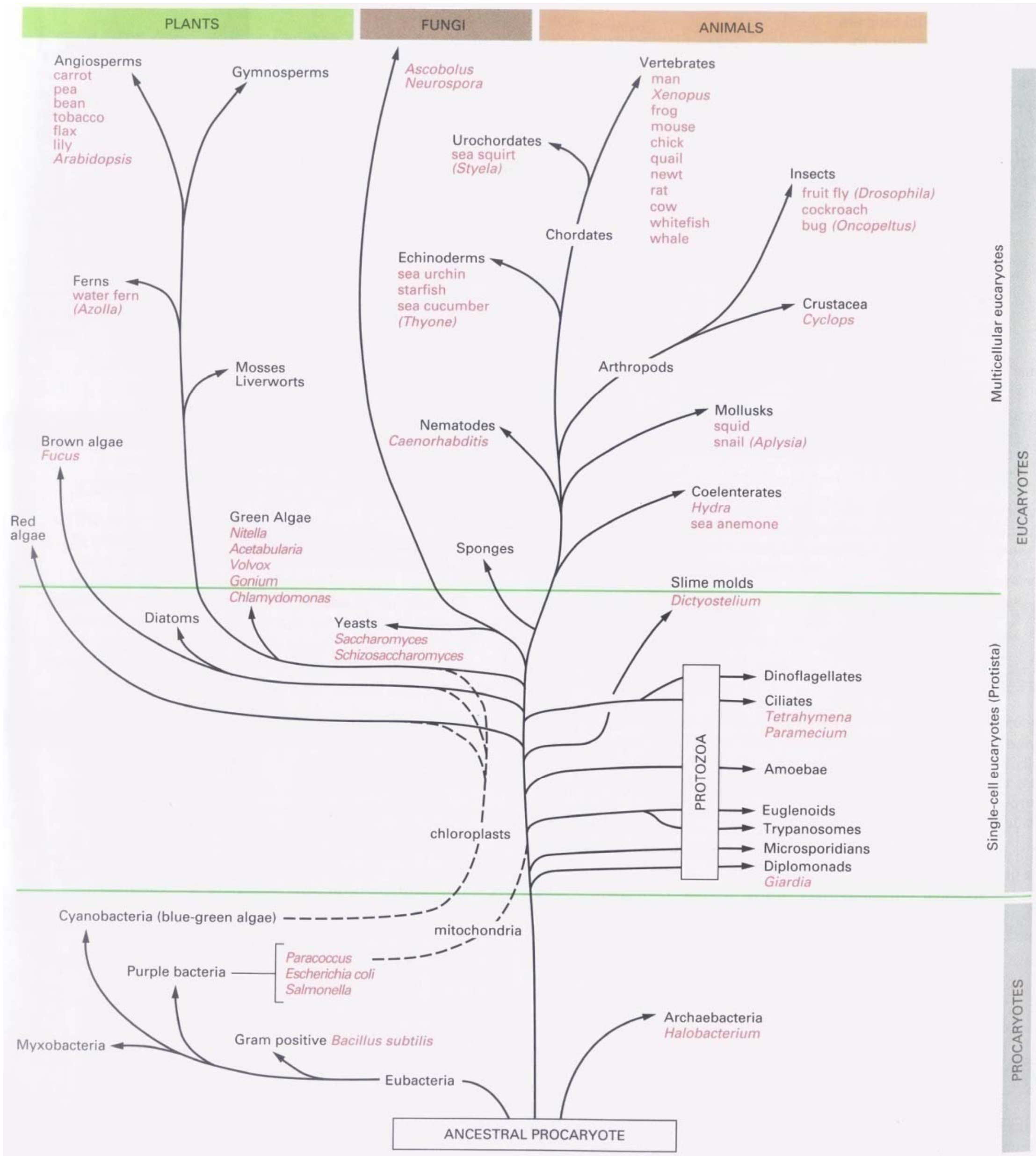
Anthony James Lab at UCI

Sieglaff et al., PNAS, 200

Evolution

- Related organisms have similar DNA
  - Similarity in sequences of proteins
  - Similarity in organization of genes along the chromosomes
- Evolution plays a major role in biology
  - Many mechanisms are shared across a wide range of organisms
  - During the course of evolution existing components are adapted for new functions

The Tree of Life

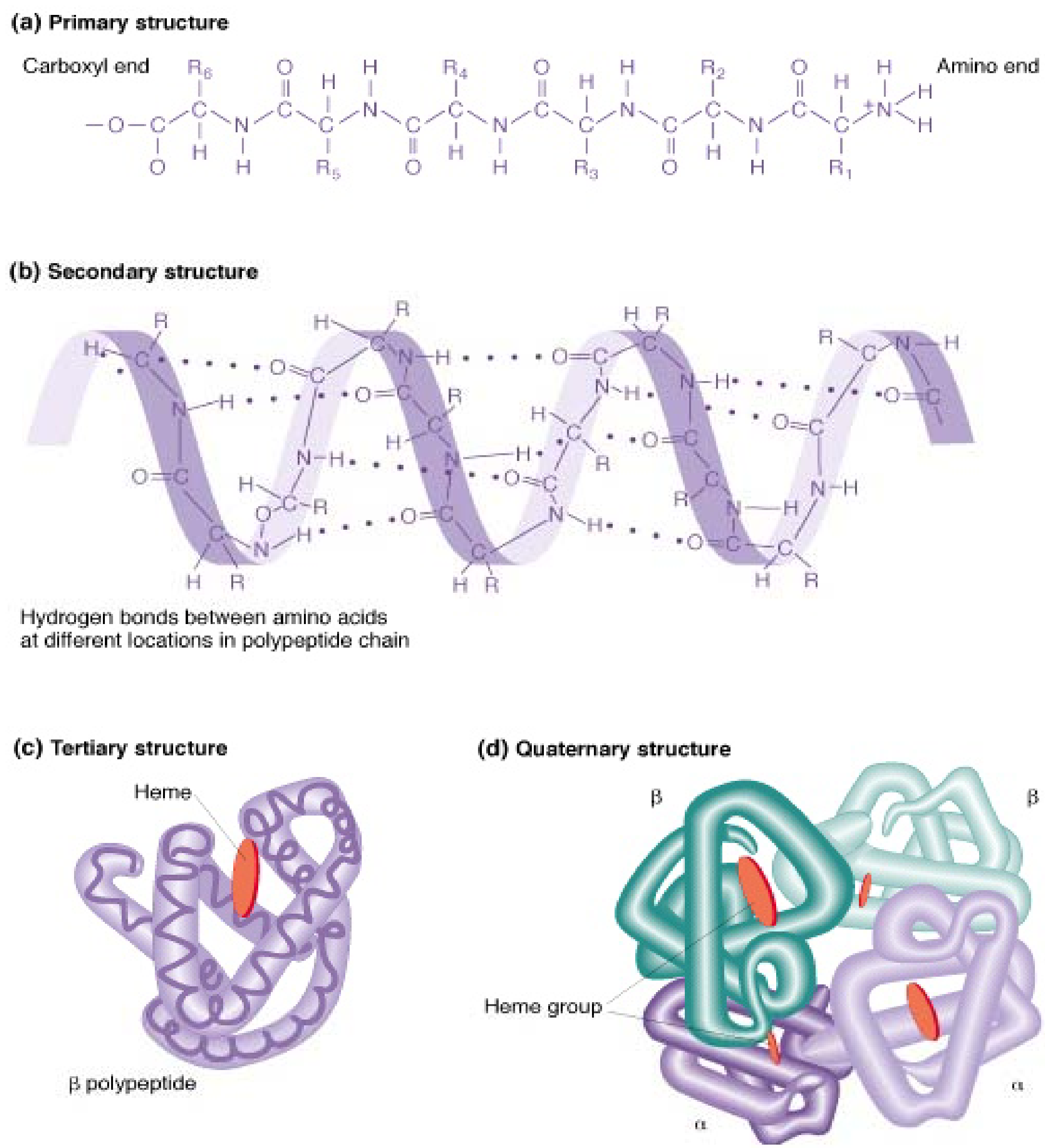


Source: Alberts et al

Protein Structure Prediction

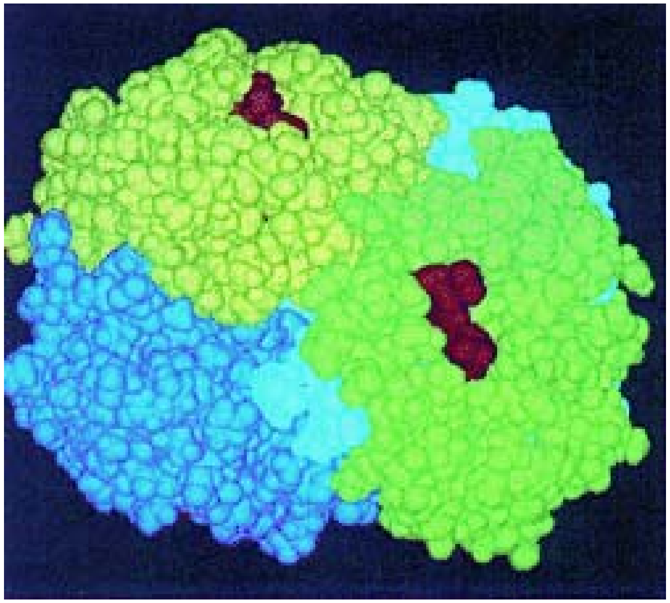
Protein Structure

- Proteins are poly-peptides of 70-3000 amino-acids
- This structure is (mostly) determined by the sequence of amino-acids that make up the protein



Hemoglobin

- protein built from 4 polypeptides
- responsible for carrying oxygen in red blood cells



Protein structures

Why do we need protein structures?

Protein structure is the key to understand:
• Sequence-function relationships
• Evolution (a history of changes at the level of sequence and function)

Based on the knowledge of protein structure we can attempt to:
• Design new drugs
• Design proteins with new functions (e.g. industrial enzymes)

Protein structure determination / prediction

Protein model accuracy vs time required for its construction

Experimental protein structure solution provides high resolution models,
BUT:
…it is costly and time-consuming,
…fails for many proteins

Drug design requires hyper-accurate protein structures.

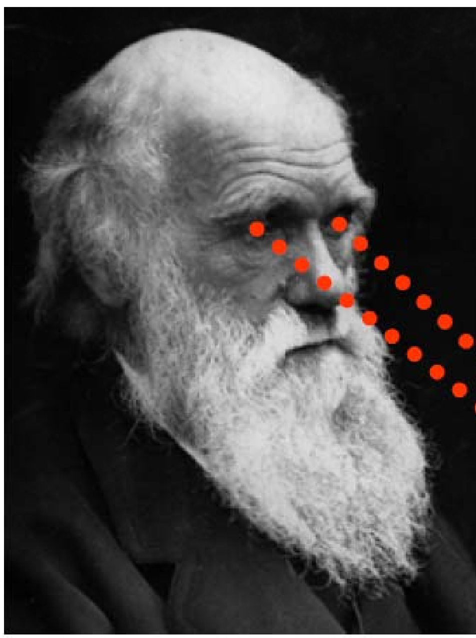
NONETHELESS:
to understand the protein function or to engineer proteins by genetic methods it is often sufficient to know only an approximate structure (at the level of amino acid residues, not necessarily all atoms)

Theoretical modeling can quickly provide approximate structures

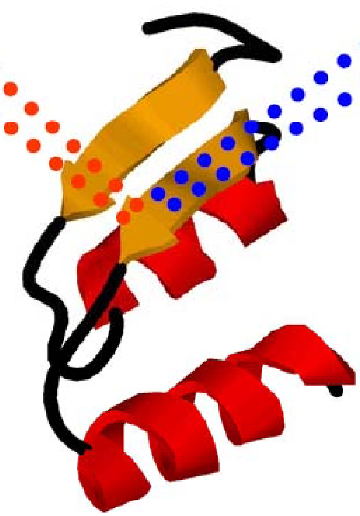
TWO DIFFERENT PERSPECTIVES ON PROTEIN STRUCTURE PREDICTION

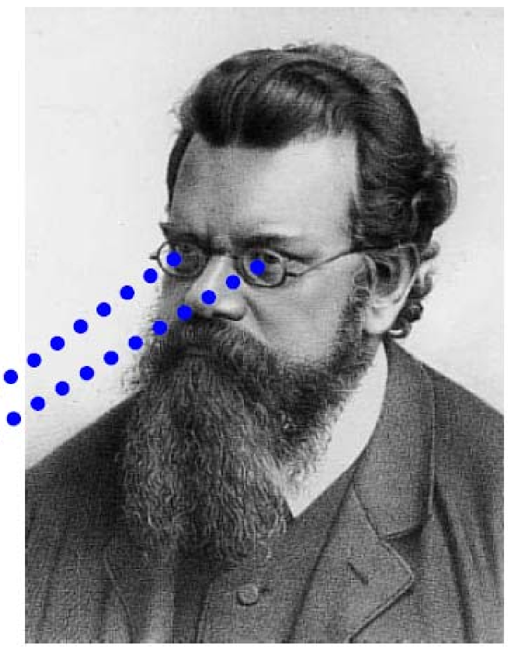
EVOLUTIONARY BIOLOGY

STATISTICAL THERMODYNAMICS

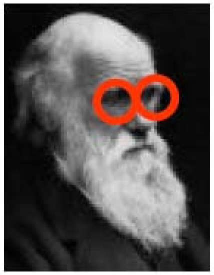


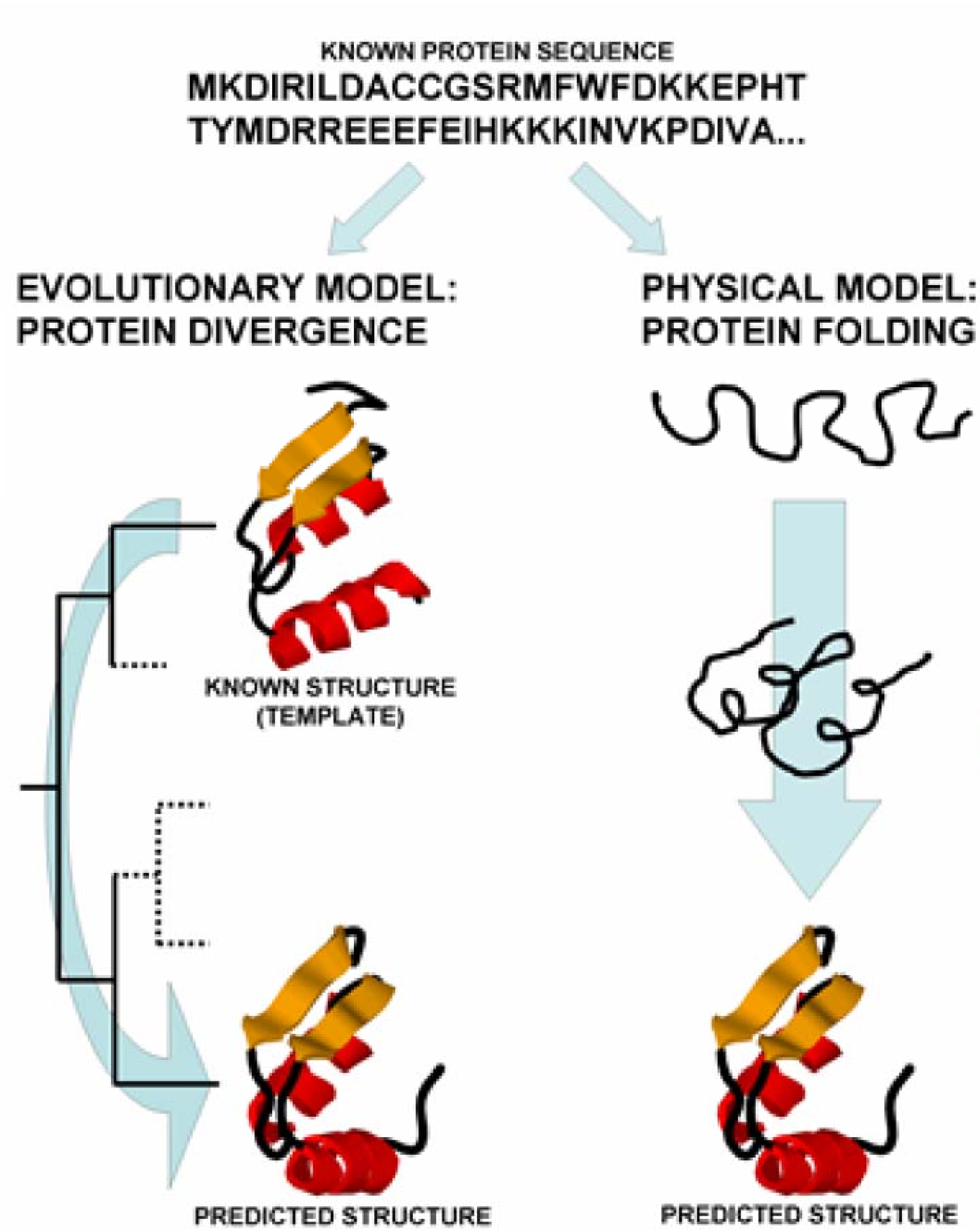
Charles
Darwin
(1809-1882)





Ludwig Edward
Boltzmann
(1844-1906)





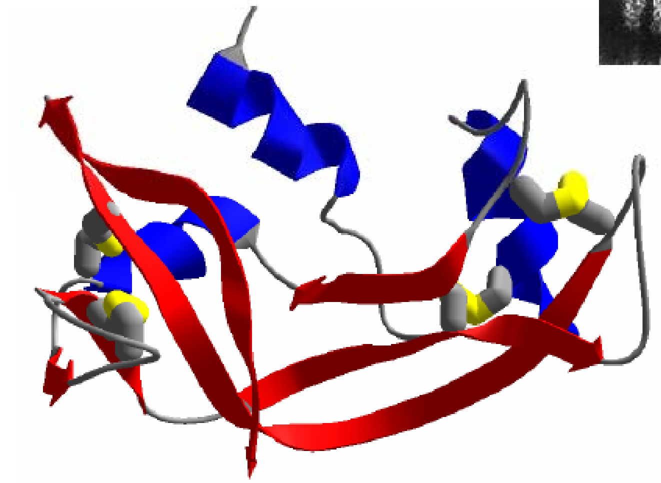
billions
of years

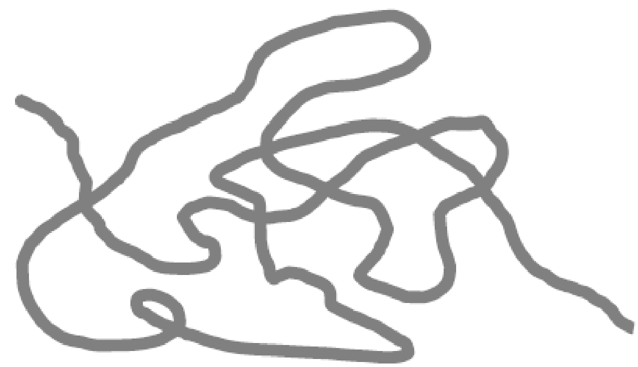
miliseconds
/ seconds

Protein refolding:

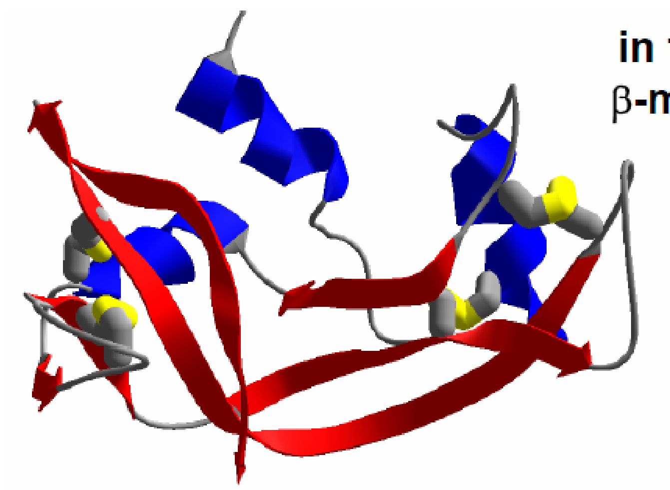
```
        10        20        30        40        50
....|....|....|....|....|....|....|....|....|....|
KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHES

        60        70        80        90       100
....|....|....|....|....|....|....|....|....|....|
LADVQAVCSQKNVACKNGQTNCYQSYSTMSITDCRETGSSKYPNCAYKTT

       110       120
....|....|....|....|....
QANKHIIVACEGNPYVPVHFDASV
```

RNase A





renaturation in the presence of β-mercaptoethanol



Anfinsen et al, *PNAS* 47, 1309 (1961)

Physics of protein folding

… Anfinsen's thermodynamic hypothesis:

"The three-dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, etc.) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment."

Christian B. Anfinsen (1916-1995)

Structure prediction based on physics

1. Assume Anfinsen's hypothesis is correct (native protein structure is thermodynamically stable and located at the global free energy minimum)

2. SAMPLING: Within a reasonable time, effectively sample the conformational space available to the polypeptide chain of the target protein to generate a large number of „decoys", of which at least one should be sufficiently similar to the native structure to be in the same energy minimum.

3. ENERGY: Use an accurate energy function to rank the decoys and identify the native structure as one of the lowest energy.

WHAT IS POSSIBLE

So, is it possible to predict the protein structure based solely on the principles of physics?

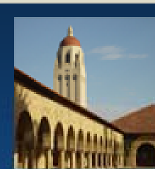
Not yet. But hope remains… ;-)

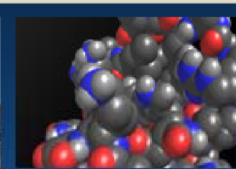
It is not yet possible to sample all conformations within reasonable time to guarantee that one of them will be sufficiently similar to the native structure. It is also not yet possible to guarantee that the native structure (or the corresponding closest decoy) would have the lowest energy.

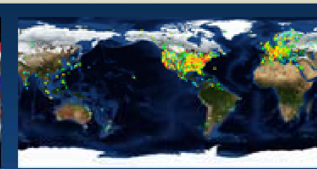
http://folding.stanford.edu/

Latest Headlines PubMed ICS Yahoo New York Times UCSC Java

Folding@home
distributed computing







Home Download Guides FAQ Stats Science Results Awards About Us

Main
News
Forum (Help)
Donate

العربية
中文
Dansk
Deutsch
English
Español
Français
Italiano
Ελληνικά
日本語
연구말
Lietuvių
Magyar
Nederlands
Norsk
Occitan
فارسی
Polski
Português
Русский
Suomeksi
Svenska
Turkish
Tiếng Việt

Our goal: to understand protein folding, misfolding, and related diseases

What is protein folding and how is folding linked to disease?
Proteins are biology's workhorses -- its "nanomachines." Before proteins can carry out these important functions, they assemble themselves, or "fold." The process of protein folding, while critical and fundamental to virtually all of biology, in many ways remains a mystery.

Download Folding@home

Moreover, when proteins do not fold correctly (i.e. "misfold"), there can be serious consequences, including many well known diseases, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, Huntington's, Parkinson's disease, and many Cancers and cancer-related syndromes.

You can help by simply running a piece of software.
Folding@home is a distributed computing project -- people from throughout the world download and run software to band together to make one of the largest supercomputers in the world. Every computer takes the project closer to our goals. Folding@home uses novel computational methods coupled to distributed computing, to simulate problems millions of times more challenging than previously achieved.

What have we done so far?
We have had several successes. You can read about them on our Science page, on our Awards page, or go directly to our Results page.

Want to learn more?
Click on the links on the left for downloads or more information. You can also download our Executive Summary, which is a PDF suitable for distribution. Also, you can learn more by watching recent seminars (Stanford BMI ; Xerox PARC). One can also help by donating funds to the project, via Stanford University.

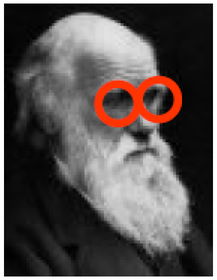
Make a gift *now!*
Click here to make a gift ›

National Science Foundation — National Institutes of Health — Google — Dell — Apple — intel — ATI — simbios — nVIDIA — Stanford School of Medicine

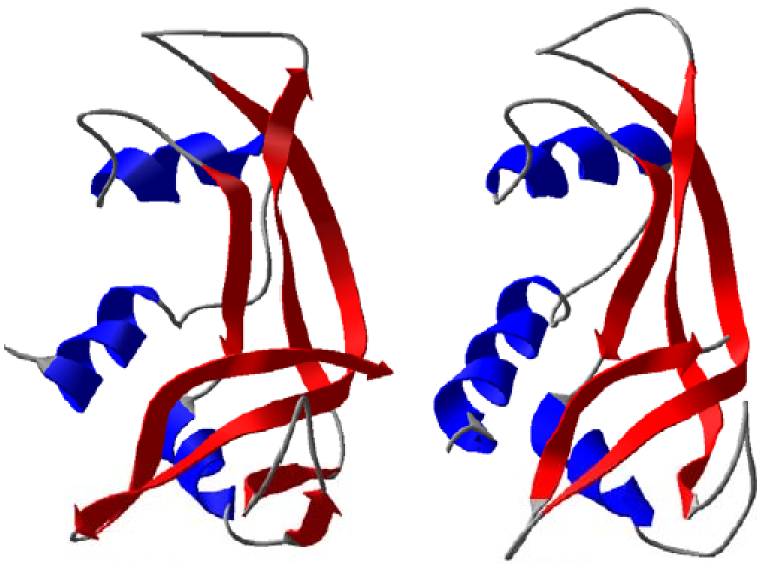
THE EVOLUTIONARY SCHOOL OF PROTEIN STRUCTURE PREDICTION

Theodosius Dobzhansky (1900-1975)

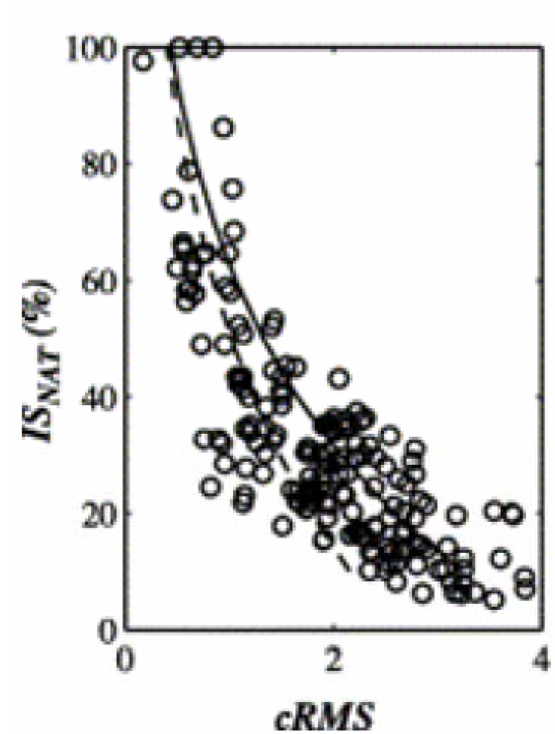
nothing in biology makes sense except in the light of evolution



Evolution of proteins: sequence-structure-function

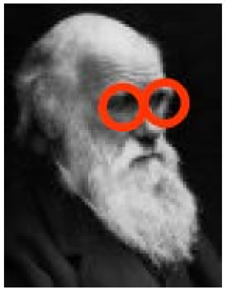


RNase A RNase 4




```
RNase A KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVC
RNase 4 QDGMYQRFLRQHVHPEE-TGGSDRYCNLMMQRRKMTLYHCKRFNTFIHEDIWNIRSIC

SQKNVACKNGQTNCYQSYSTMSITDCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHFDASV
STTNIQCKNGKMNCHEG--VVKVTDCRDTGSSRAPNCRYRAIASTRRVVIACEGNPQVPVHFDG--
```

Evolutionarily related proteins retain very similar tertiary structure despite the accumulation of amino acid substitutions, deletions and insertions



Structure prediction based on evolution

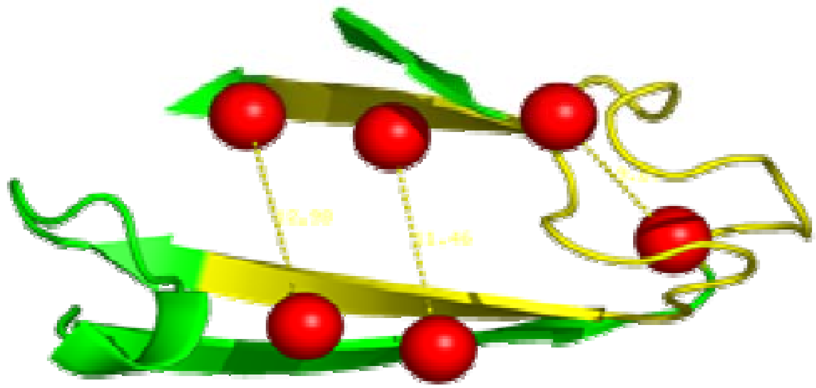
1. Assume that extant proteins evolved from common ancestors

2. Assume that evolutionarily related (homologous) proteins are structurally similar

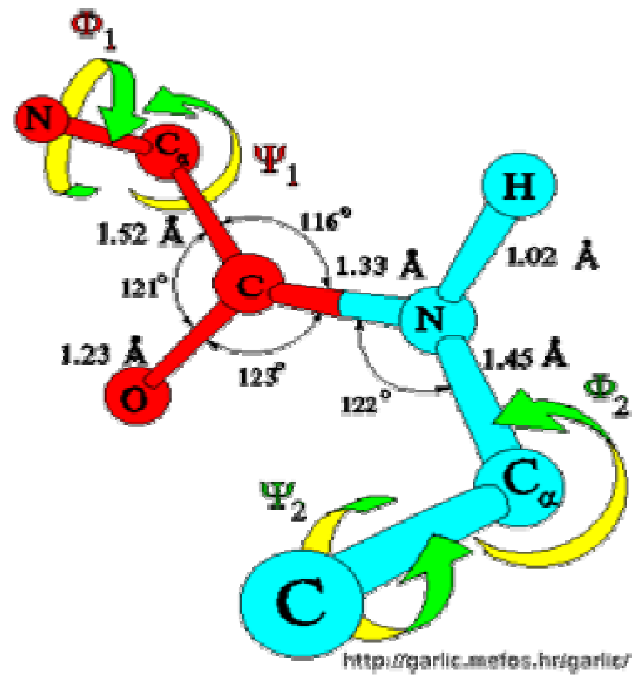
3. Find a structure of a protein homologous to the target protein and model the evolutionary changes that occurred since these two proteins diverged from their common ancestor (provided that an accurate model of divergent evolution of protein sequences and structures is available)

Template-based modeling

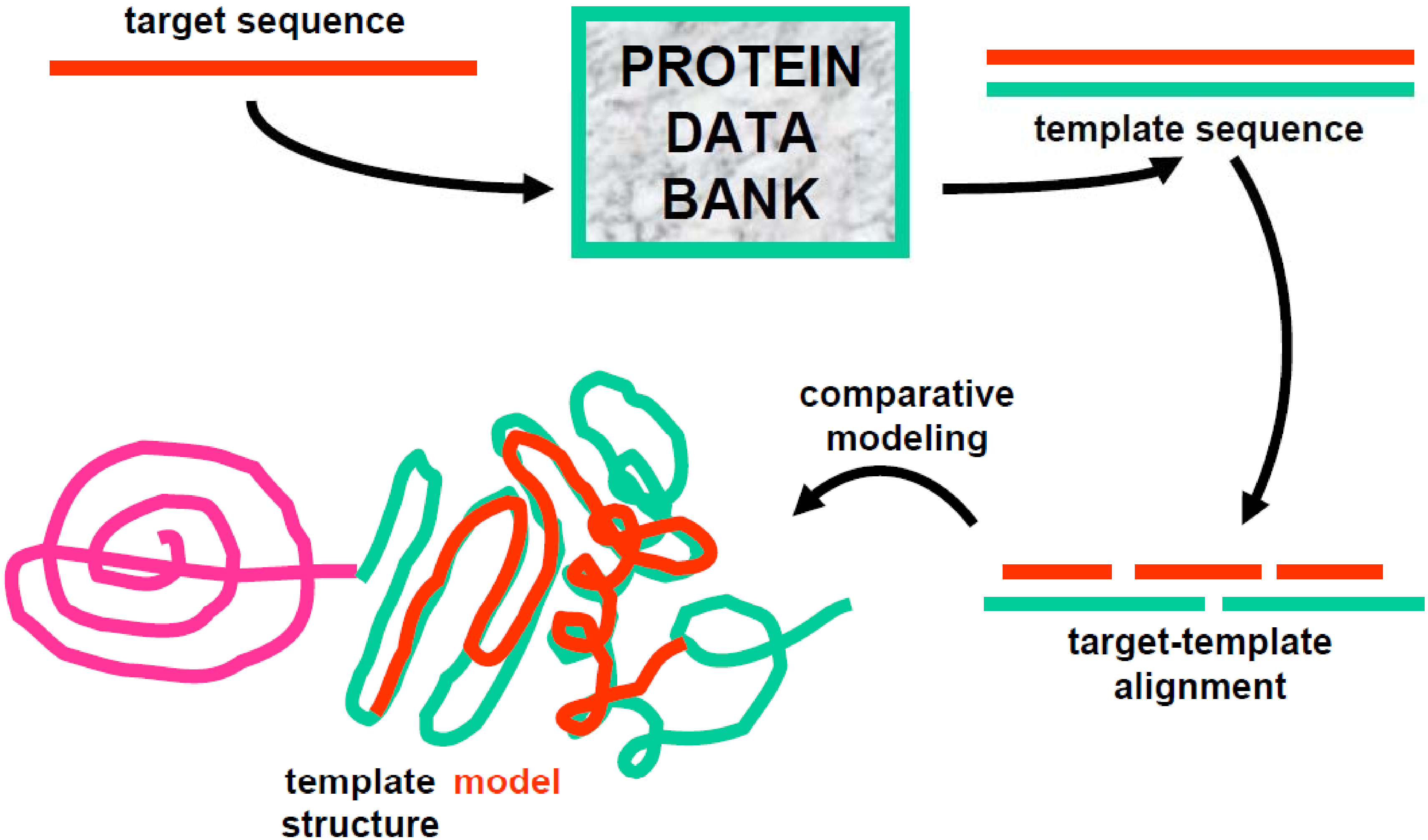
1. For all residues in the target protein assign restraints derived from the template structure:
   - bond lenths, angles
   - interactomic distances



2. Minimization of the target function usually comprising:
   - a term describing the agreememnt of the model with the restraints
   - a term describing the ideal values of various structural parameters

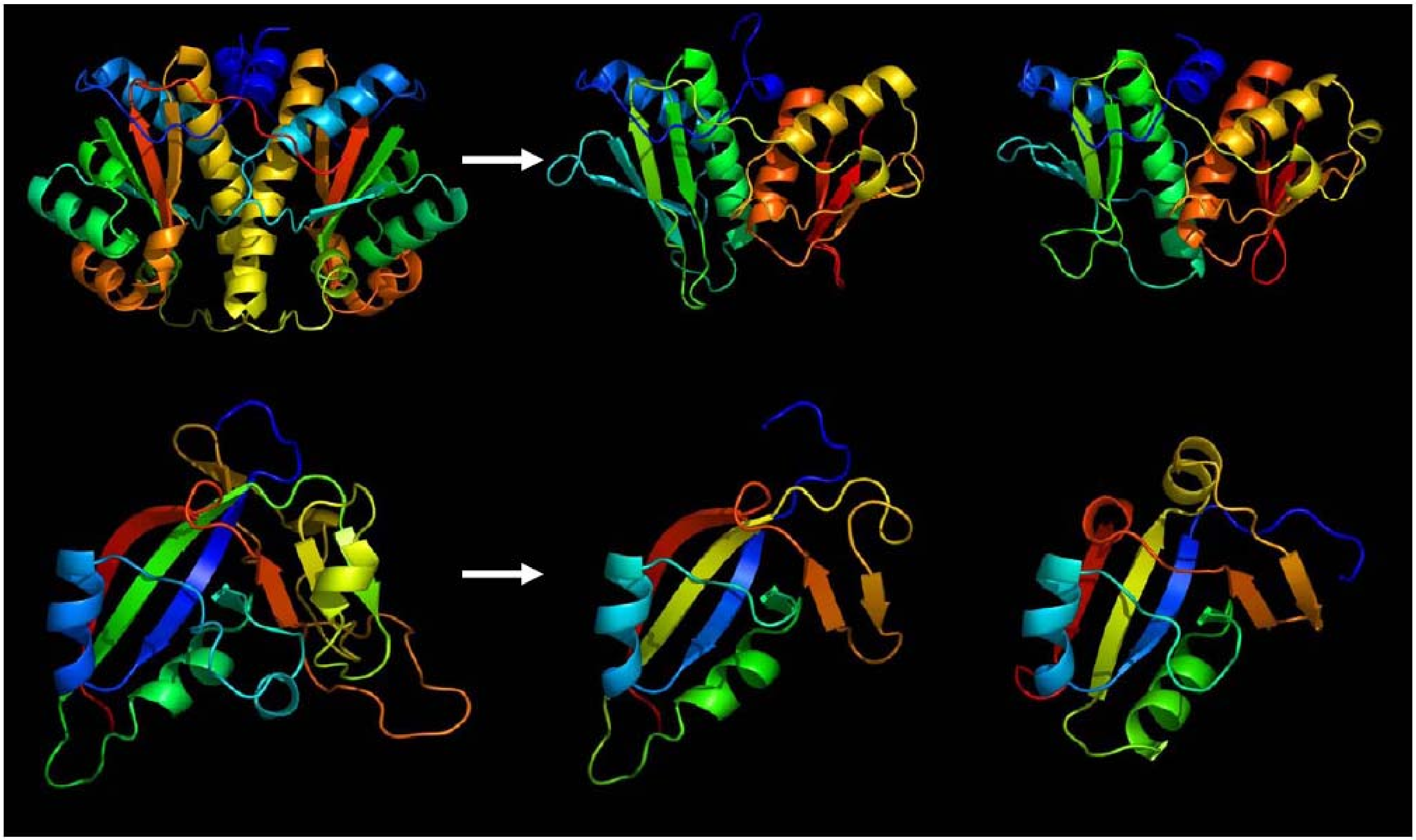


And get a model derived from the template (with some modifications)



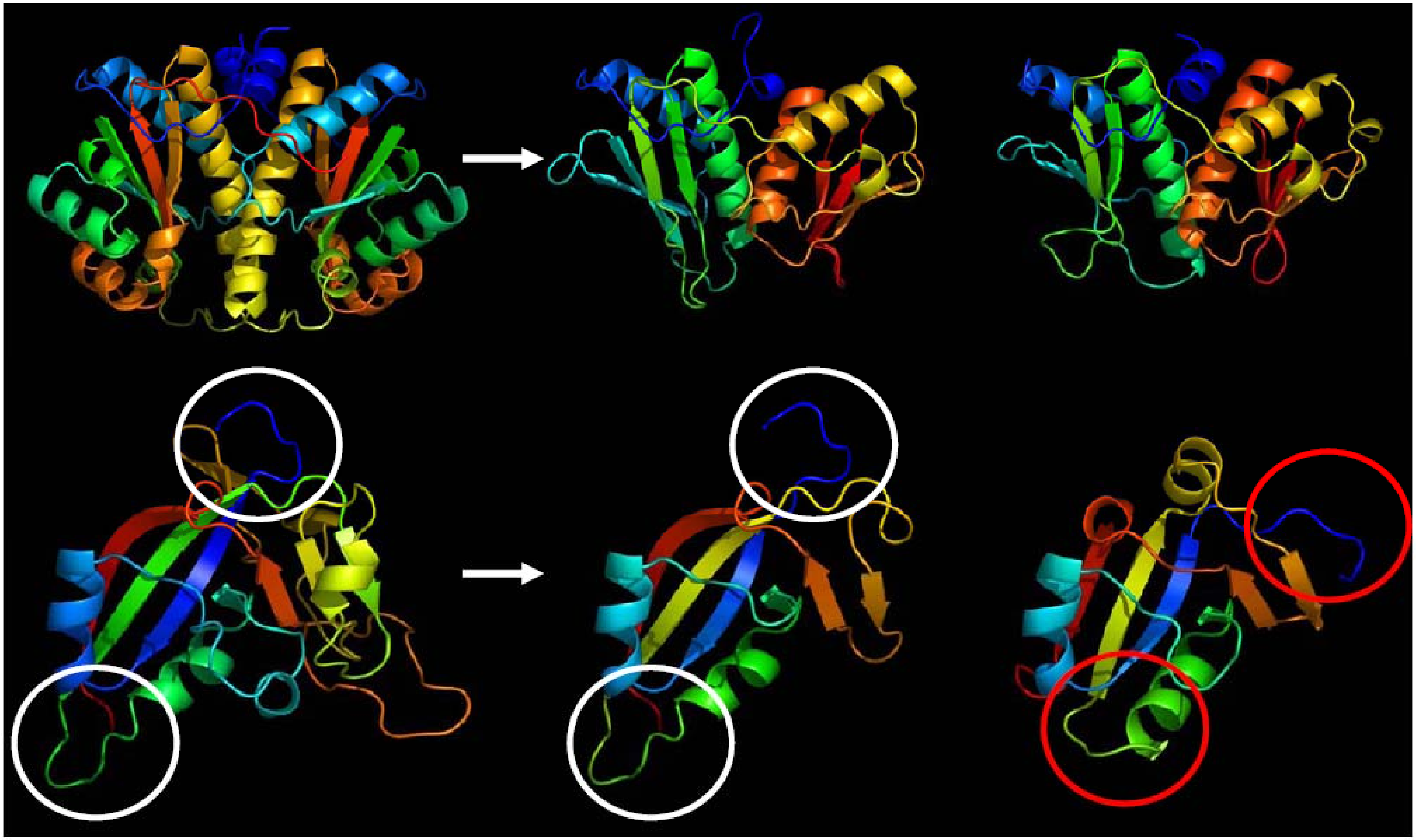
Template-based models can be quite good

template structure | **modeled target structure** | **native terget structure**



Template-based models can be quite good
…but devil is in the details

template structure	modeled target structure	native terget structure



SO, WHAT IS POSSIBLE?

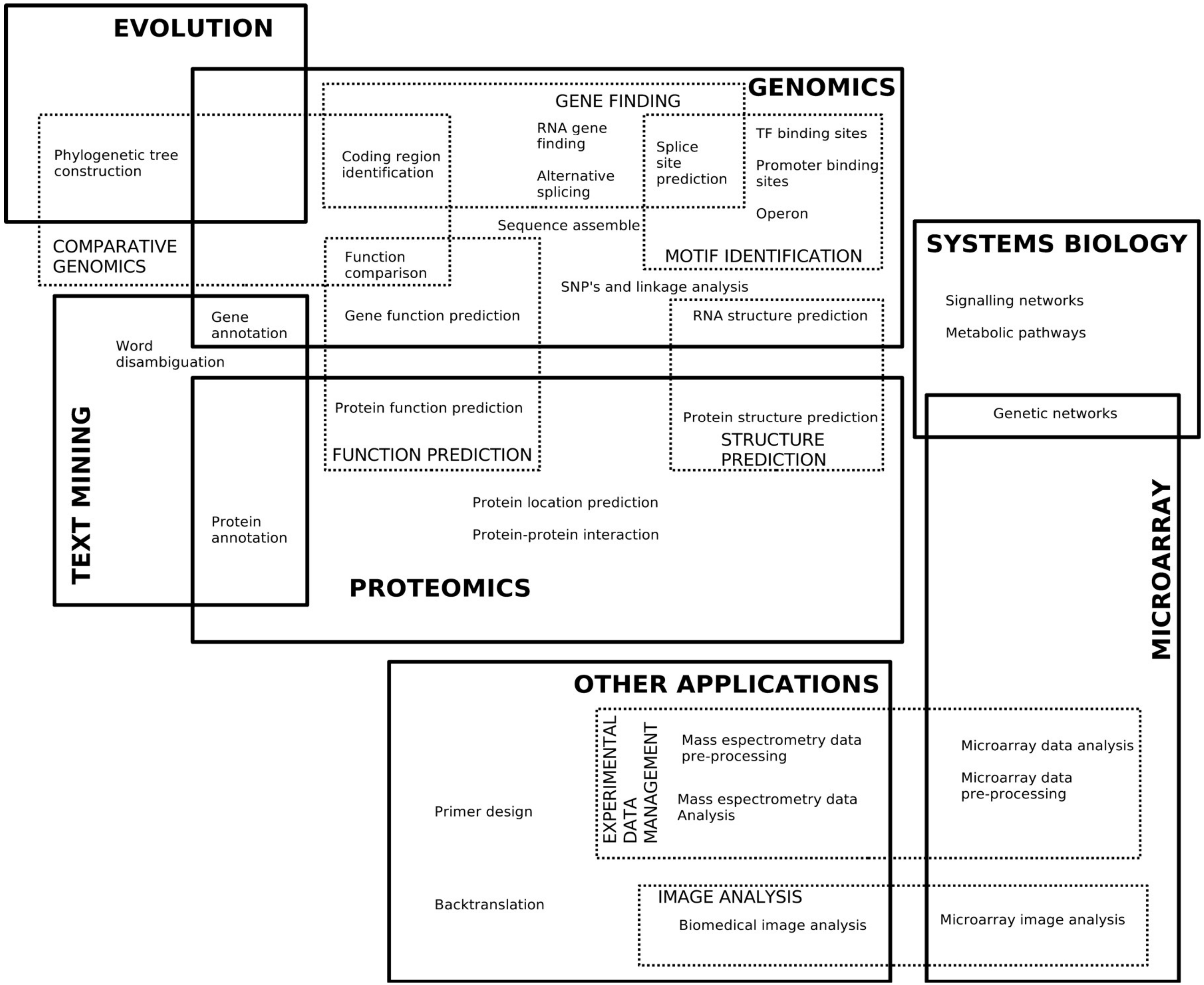
Is it possible to predict the protein structure based solely on the principles of physics?

Not yet. But hope remains… ;-)
It is not yet possible to sample all conformations within reasonable time to guarantee that one of them will be sufficiently similar to the native structure.
It is also not yet possible to guarantee that the native structure (or the corrsponing closest decoy) would have the lowest energy.

Is it possible to predict the protein structure based solely on the principles of evolution?
Yes…But only if 1) there is a homologous protein with known structure,
2) if we can correctly identify it among all proteins with known structures
and 3) if we can approximate the evolutionary changes at the level of sequences (alignment) and structures (3D coordinates).

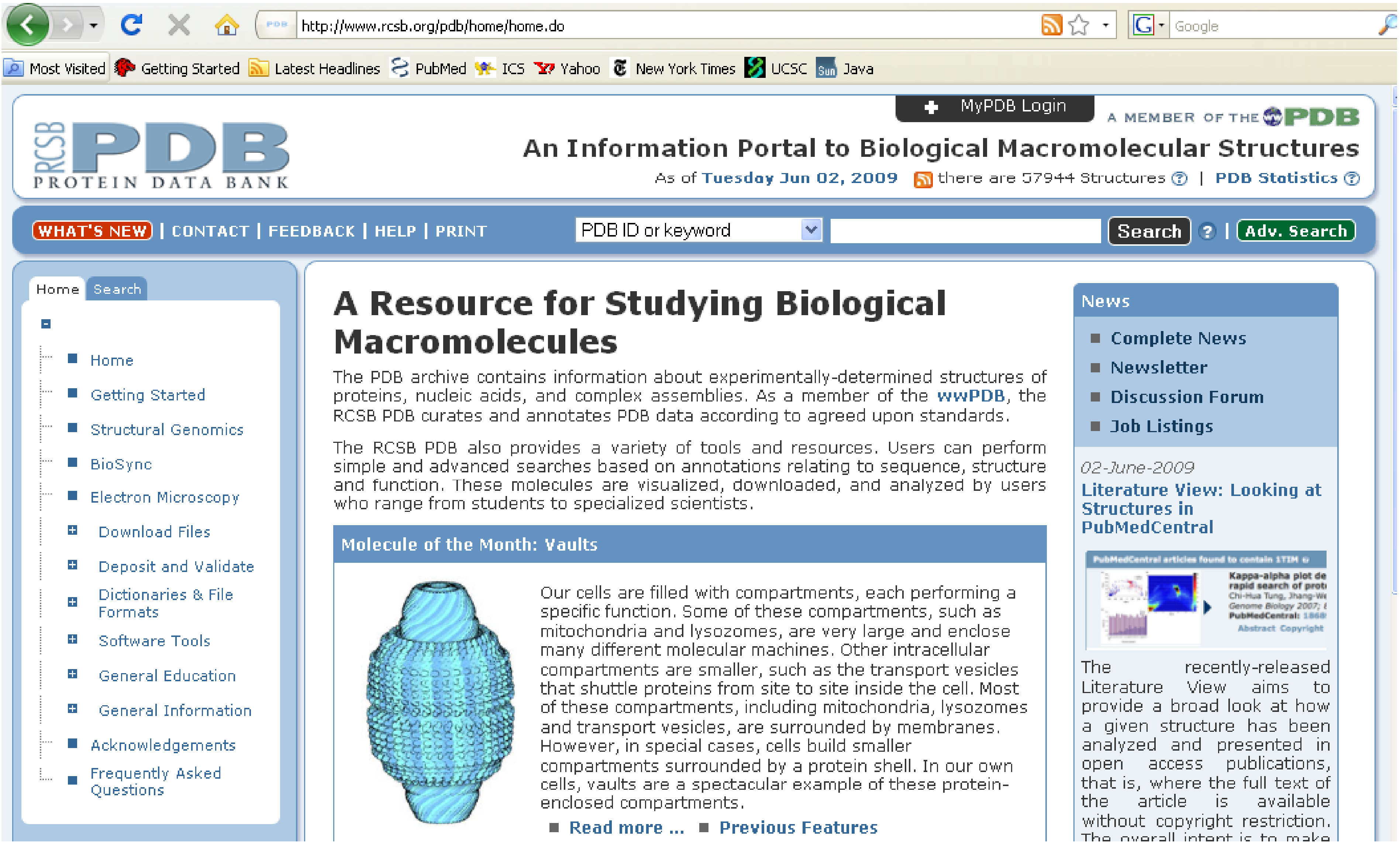
Protein data bank

Machine Learning in Bioinformatics

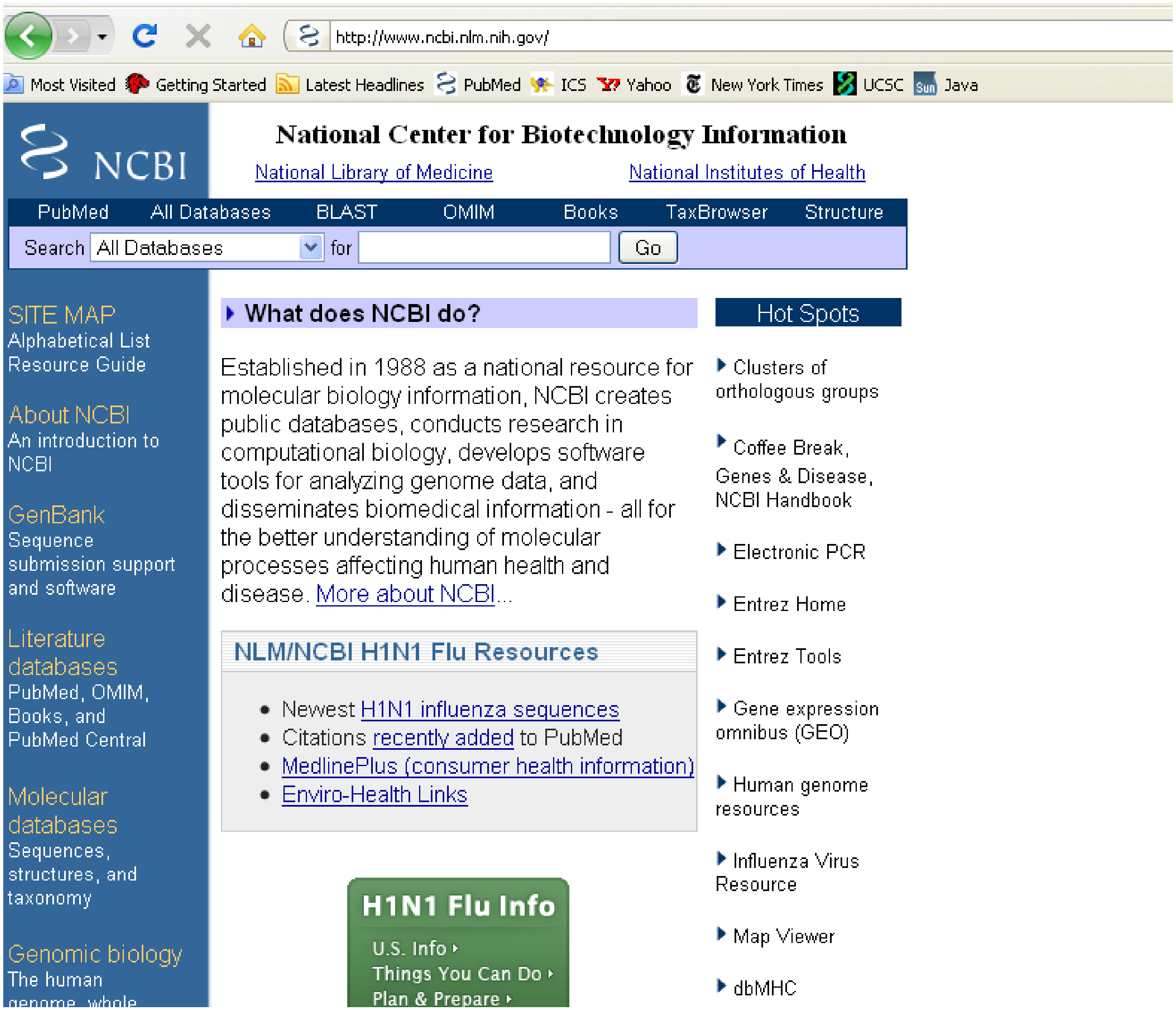

EVOLUTION
GENOMICS
GENE FINDING
Phylogenetic tree construction
Coding region identification
RNA gene finding
Alternative splicing
Splice site prediction
TF binding sites
Promoter binding sites
Operon
Sequence assemble
COMPARATIVE GENOMICS
Function comparison
MOTIF IDENTIFICATION
SNP's and linkage analysis
SYSTEMS BIOLOGY
Signalling networks
Metabolic pathways
Gene annotation
Gene function prediction
RNA structure prediction
Word disambiguation
Genetic networks
Protein function prediction
Protein structure prediction
STRUCTURE PREDICTION
FUNCTION PREDICTION
TEXT MINING
Protein location prediction
Protein-protein interaction
Protein annotation
PROTEOMICS
MICROARRAY
OTHER APPLICATIONS
EXPERIMENTAL DATA MANAGEMENT
Mass espectrometry data pre-processing
Mass espectrometry data Analysis
Microarray data analysis
Microarray data pre-processing
Primer design
Backtranslation
IMAGE ANALYSIS
Biomedical image analysis
Microarray image analysis

Online Resources

Protein Data Bank



NCBI

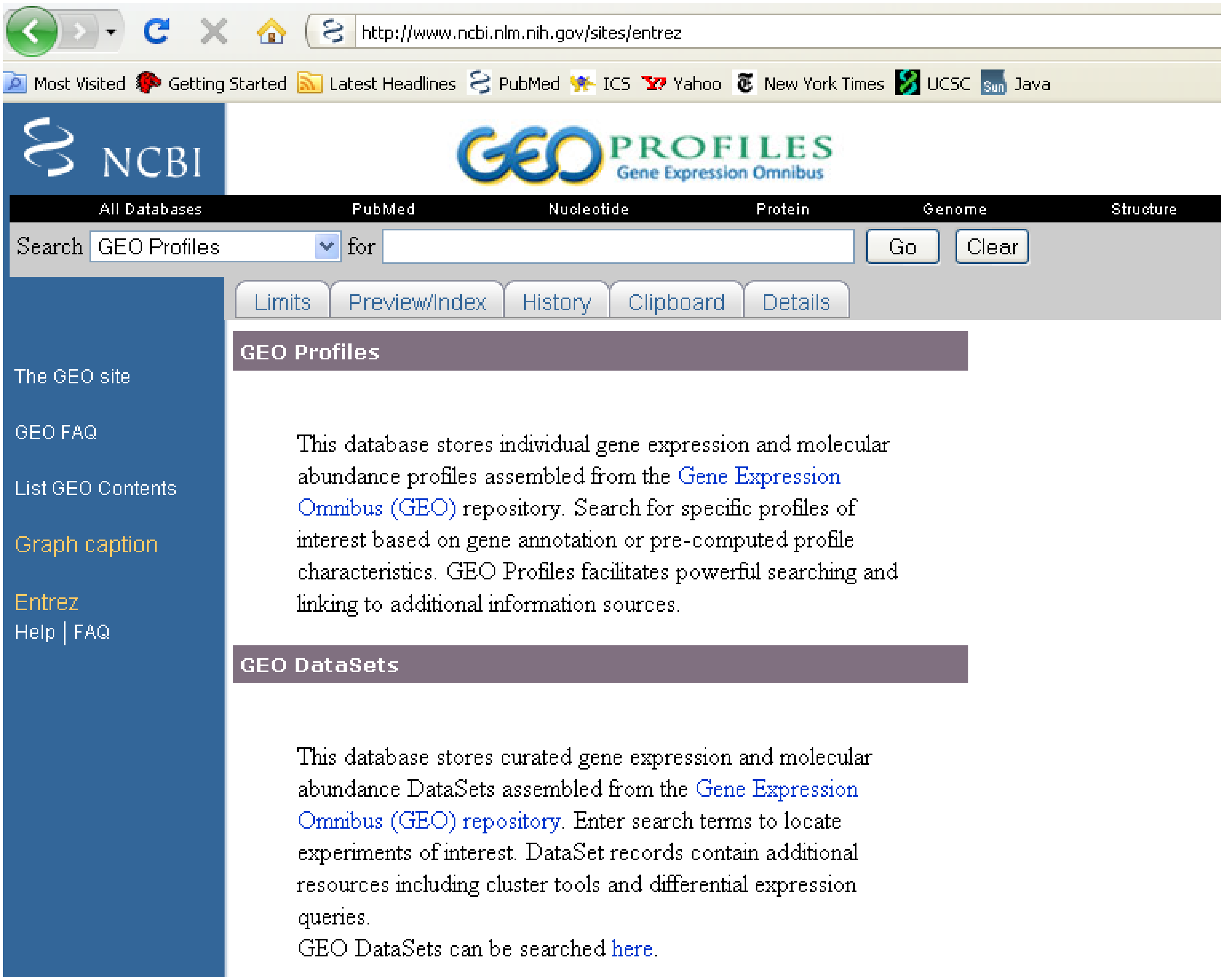


http://www.ncbi.nlm.nih.gov/
Most Visited
Getting Started
Latest Headlines
PubMed
ICS
Yahoo
New York Times
UCSC
Java
NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed
All Databases
BLAST
OMIM
Books
TaxBrowser
Structure
Search
All Databases
for
Go
SITE MAP
Alphabetical List
Resource Guide
About NCBI
An introduction to NCBI
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
Genomic biology
The human
What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More about NCBI...
NLM/NCBI H1N1 Flu Resources
Newest H1N1 influenza sequences
Citations recently added to PubMed
MedlinePlus (consumer health information)
Enviro-Health Links
H1N1 Flu Info
U.S. Info
Things You Can Do
Plan & Prepare
Hot Spots
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
Influenza Virus Resource
Map Viewer
dbMHC

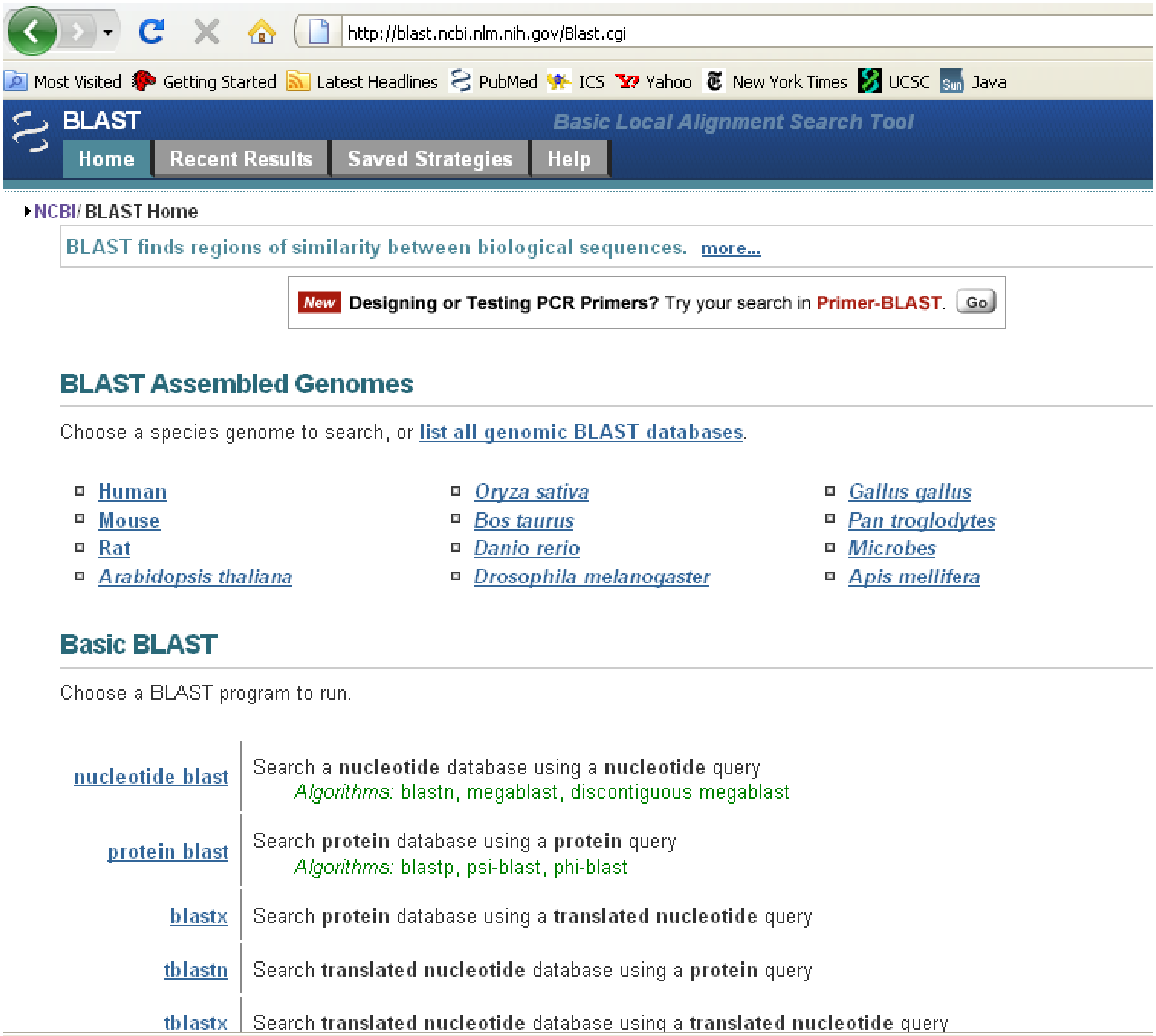
PubMed


http://www.ncbi.nlm.nih.gov/sites/entrez
pubmed
Most Visited
Getting Started
Latest Headlines
PubMed
ICS
Yahoo
New York Times
UCSC
Java
NCBI
PubMed
www.pubmed.gov
A service of the U.S. National Library of Medicine
and the National Institutes of Health
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
OMIM
PMC
Journals
Books
Search
PubMed
for
hidden Markov model
Go
Clear
Advanced Search
Save Search
Limits
Preview/Index
History
Clipboard
Details
Display
Summary
Show
20
Sort By
Send to
All: 985
Review: 30
Items 1 - 20 of 985
Page
1
of 50
Next
1: Fast statistical alignment.
Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L.
PLoS Comput Biol. 2009 May;5(5):e1000392. Epub 2009 May 29.
PMID: 19478997 [PubMed - in process]
Related Articles
Free article at journal site
2: Identifying novel constrained elements by exploiting biased substitution patterns.
Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X.
Bioinformatics. 2009 Jun 15;25(12):i54-62.
PMID: 19478016 [PubMed - in process]
Related Articles
Free article at journal site
3: Model-based clustering of array CGH data.
Shah SP, Cheung KJ Jr, Johnson NA, Alain G, Gascoyne RD, Horsman DE, Ng RT, Murphy KP.
Bioinformatics. 2009 Jun 15;25(12):i30-8.
PMID: 19478003 [PubMed - in process]
Related Articles
Free article at journal site
4: Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data.
Yin J, Jordan MI, Song YS.
Also try:
hidden markov model review
profile hidden markov model
Recent Activity
Turn Off
Clear
hidden Markov model (985)
PubMed
Machine learning in bioinformatics.

Gene Expression Omnibus

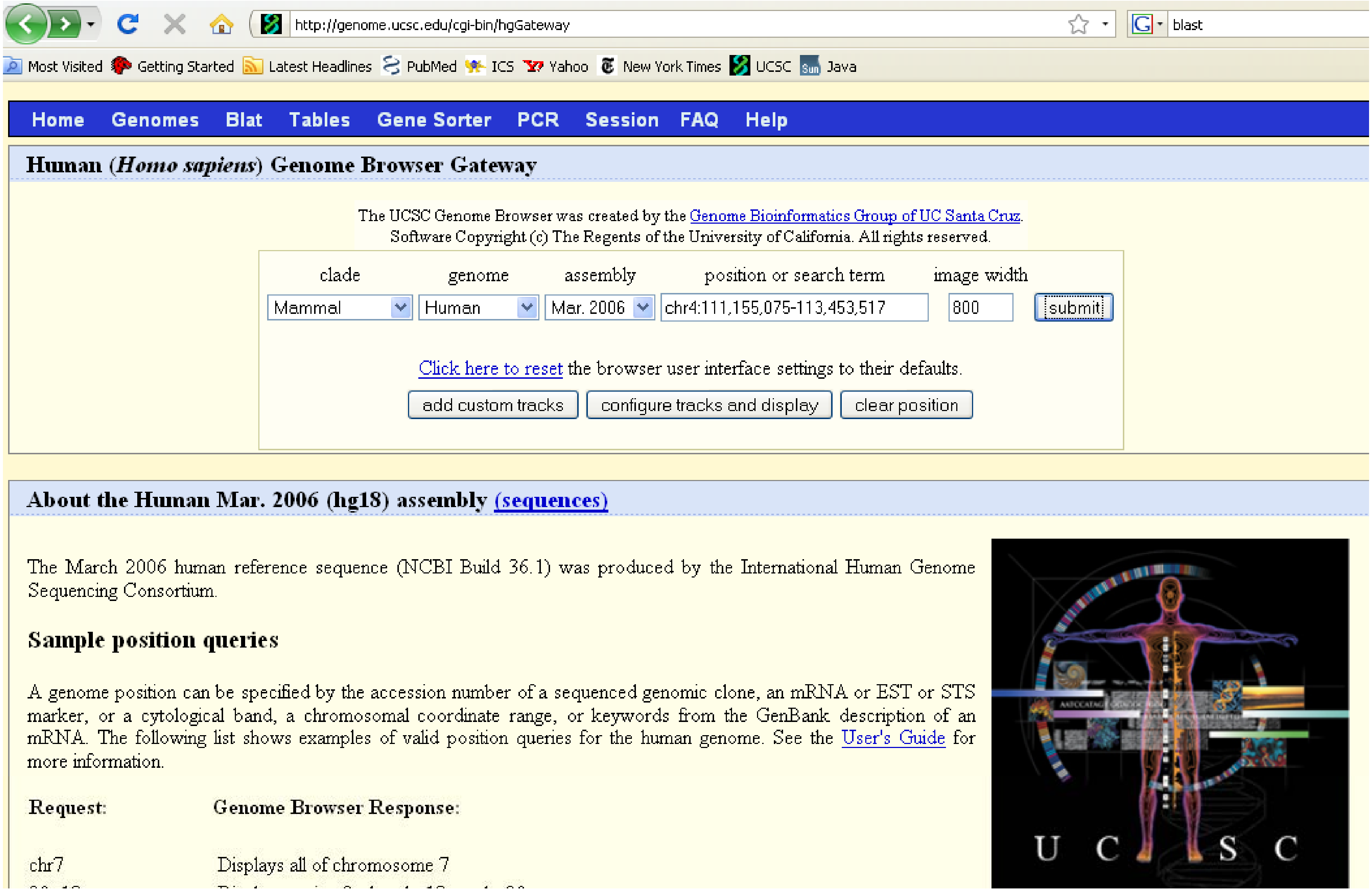


BLAST

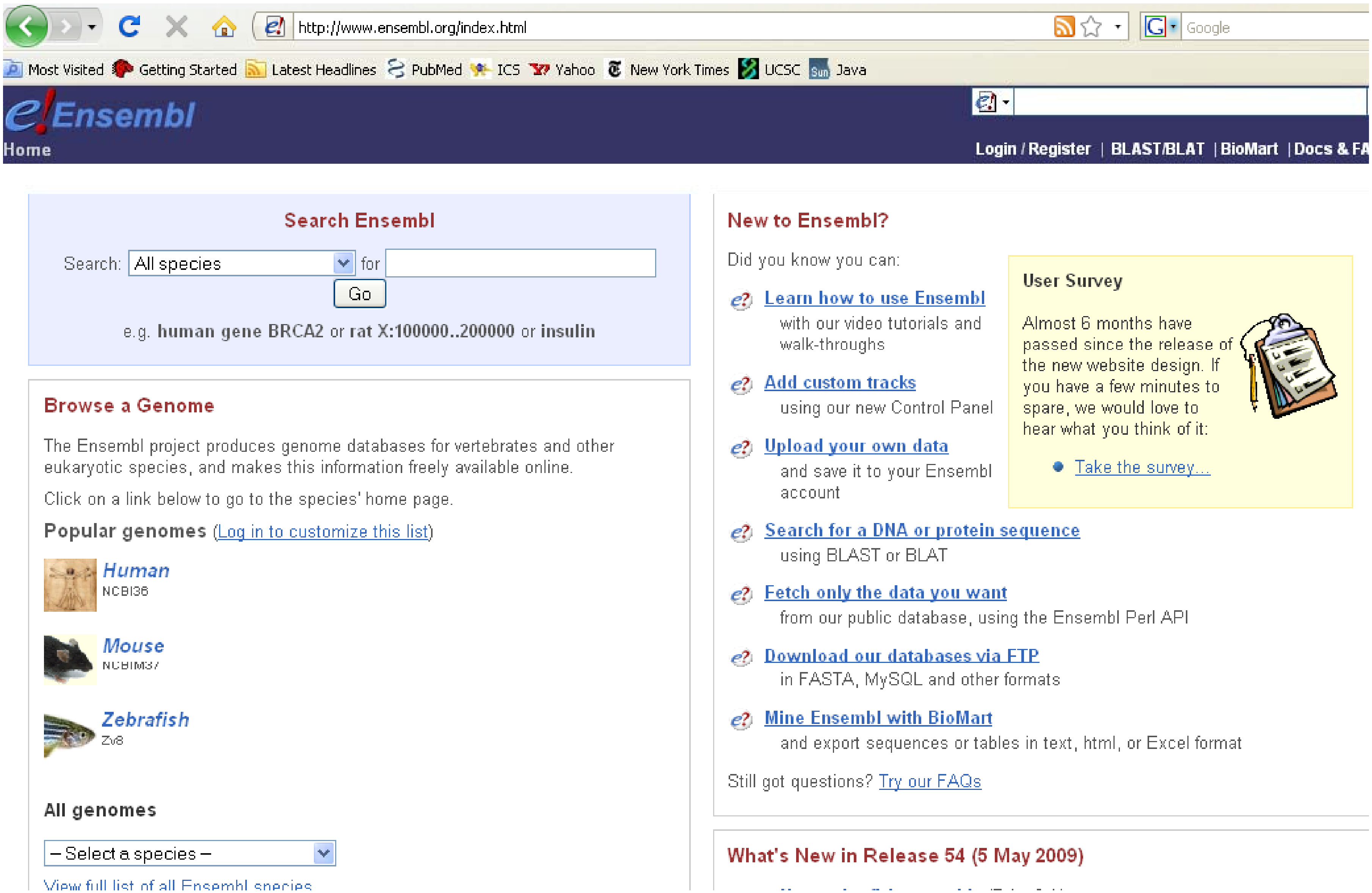


http://blast.ncbi.nlm.nih.gov/Blast.cgi
Most Visited
Getting Started
Latest Headlines
PubMed
ICS
Yahoo
New York Times
UCSC
Java
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
New Designing or Testing PCR Primers? Try your search in Primer-BLAST. Go
BLAST Assembled Genomes
Choose a species genome to search, or list all genomic BLAST databases.
Human
Mouse
Rat
Arabidopsis thaliana
Oryza sativa
Bos taurus
Danio rerio
Drosophila melanogaster
Gallus gallus
Pan troglodytes
Microbes
Apis mellifera
Basic BLAST
Choose a BLAST program to run.
nucleotide blast
Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast
protein blast
Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast
blastx
Search protein database using a translated nucleotide query
tblastn
Search translated nucleotide database using a protein query
tblastx
Search translated nucleotide database using a translated nucleotide query

UCSC Genome Browser



Ensembl Genome Browser



Educational Opportunities at UCI

Biomedical Computing Major

BIOMEDICAL COMPUTING

Degrees available: B.S.

(*open to freshmen in Fall 2009; open to transfer applicants in Fall 2011*)

What is Biomedical Computing?

Biomedical Computing (BMC) is the intersection of computer science and information technology with biology and medicine. Students will receive a firm quantitative grounding in mathematics, statistics, and computation; become familiar with the basic foundations of physics, chemistry, and biology; master a rigorous and demanding Biomedical Computing year-long sequence; study theory, algorithms, data mining, and machine learning; and carry out a year-long immersive capstone Senior Project. The immense growth and impact of biomedical information has led to a critical need for people who can understand the languages, tools, and techniques of both life sciences and computational sciences. The Biomedical Computing program aims to create a new generation of professionals with these complementary cross-disciplinary skills.

Program Educational Objectives:

SUGGESTED CURRICULUM for the Biomedical Computing major

Freshman

Introduction to Computer Science I and II
Fundamental Data Structures
Single Variable Calculus
Multivariable Calculus
Biotech Basics[2]
From DNA to Organisms
From Organisms to Ecosystems
General Education[1] (one course)

Sophomore

Introduction to Computer Organization
Introduction to Software Engineering
Genetics
Introduction to Computational Biology
Discrete Mathematics
Boolean Algebra and Logic
Linear Algebra
Introduction to Probability and Statistics for Computer Science
Classical Physics (two courses)
Classical Physics Laboratory (two courses)
General Education (one course)

Junior

Representations and Algorithms for Molecular Biology
Probabilistic Modeling of Biological Data
Computational Systems Biology
Design and Analysis of Algorithms
Machine Learning and Data Mining
General Chemistry (two courses)
General Chemistry Laboratory (one course)
Biology Elective[3]
General Education (two courses)

Senior

Biomedical Computing Project (three courses)
Quantitative Elective[4] (three courses)
Upper-division Writing
General Education (five courses)

BIT (Bioinformatics Training Program)

INSTITUTE FOR GENOMICS AND BIOINFORMATICS

University of California, Irvine

- About IGB
- People
- Education
- Research
- Tools & Databases
- Facilities & Services
- Technology Transfer
- Events
- Sponsorship Programs

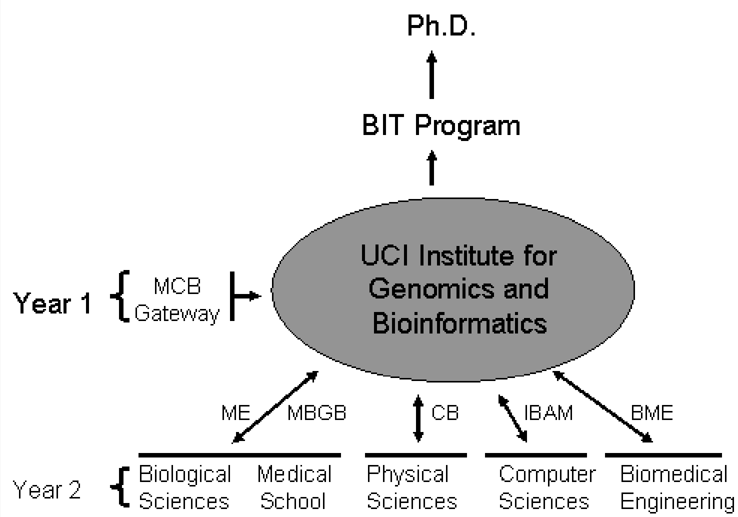
Predoctoral Program

Predoctoral :: Postdoctoral

Shortcuts: Nomination Process — Mandatory Courses — Laboratory Rotations — Choosing a BIT Faculty Thesis Advisor — Continuing BIT Program Requirements.

Admission Procedures

Recruitment of first-year students directly into the BIT Program through the MCSB Gateway will be performed in a manner similar to recruitment of first-year students into other campus graduate programs.



The Biomedical Informatics Training (BIT) Program. ME, Mol. Evolution track in Ecology and Evolution Dept.; MBGB, Combined Mol. Biol., Genet. and Biochem Program in School of Biol. Sci.; CB, Chemical Biol. Track in Chem Dept.; IBAM, Informatics in Biology and Medicine track in ICS; BME Biomedical Engineering Program. Existing graduate programs and BIT cooperate to jointly admit first year students directly into the MCSB Gateway. Alternatively, second-year students can join the BIT Program at the beginning of their second year as in the past.

MCB program

UCIrvine | Graduate Program in
Mathematical, Computational and Systems Biology

About MCSB | Admissions | Research & Faculty | Curriculum | Participating Departments | Contact MCSB | Links

Admissions & Application

Overview & Application
Application Checklist
Current Class

MCSB Home > Admissions and Applications > Overview & Application

Admissions and Application:
Students are admitted via the MCB gateway program, as described below. After successful completion of the MCB year, students in good standing who have been accepted into the group of a participating thesis advisor are automatically admitted to the department Ph.D. program of that advisor.

Prerequisites:
The MCB curriculum is designed to teach students at the beginning of their graduate careers the necessary mathematical, computational, and biological knowledge for successful research at the interface between these disciplines. The needs of students with a variety of backgrounds can be met provided that they have had mathematical training comparable to a standard one-year university-level calculus course and a lower-division university course in elementary differential equations and linear algebra. Exceptional students not meeting these prerequisites may be admitted to the program on the condition that they fulfill these requirements during the first fall quarter of their graduate study or the summer preceding, and pass with a grade of B or better.

Admissions Information:
The MCB Executive Committee and Director act as the admissions committee. Following an initial screening of applications, the Executive Committee usually invites potential trainees for an interview at UCI. The Director makes final offers of admissions to applicants, based on the recommendations of the interviewing faculty and the Executive Committee.