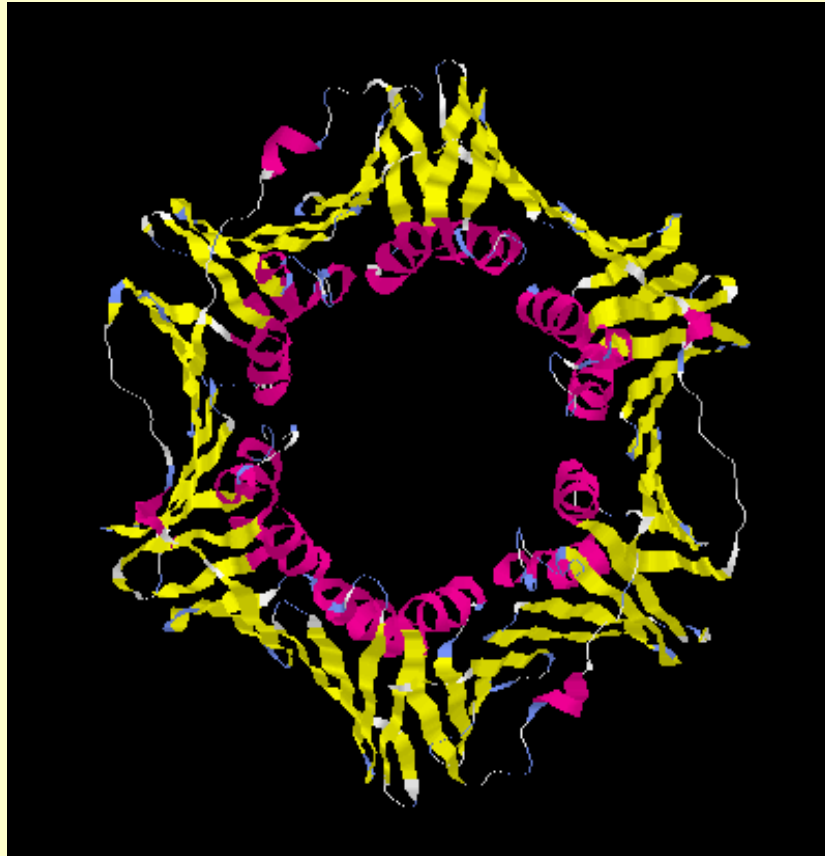


## Genomics and Bioinformatics



Doug Brutlag  
Professor Emeritus  
Biochemistry & Medicine (by courtesy)

# Faculty, TAs and Staff

---

Doug Brutlag



Lee Kozar



Maeve O'Huallachain



Dan Davison



# Course and Video Availability

---

- Always M114
  - Tuesdays & Thursdays 2:15-3:30 PM
- Course Web Site
  - <http://biochem218.stanford.edu/>
- Stanford Center for Professional Development
  - <http://scpd.stanford.edu/>
- Videos available 24 hours/day, 7 days/week
- Course offered Autumn, Winter and Spring quarters

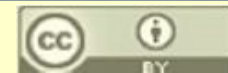
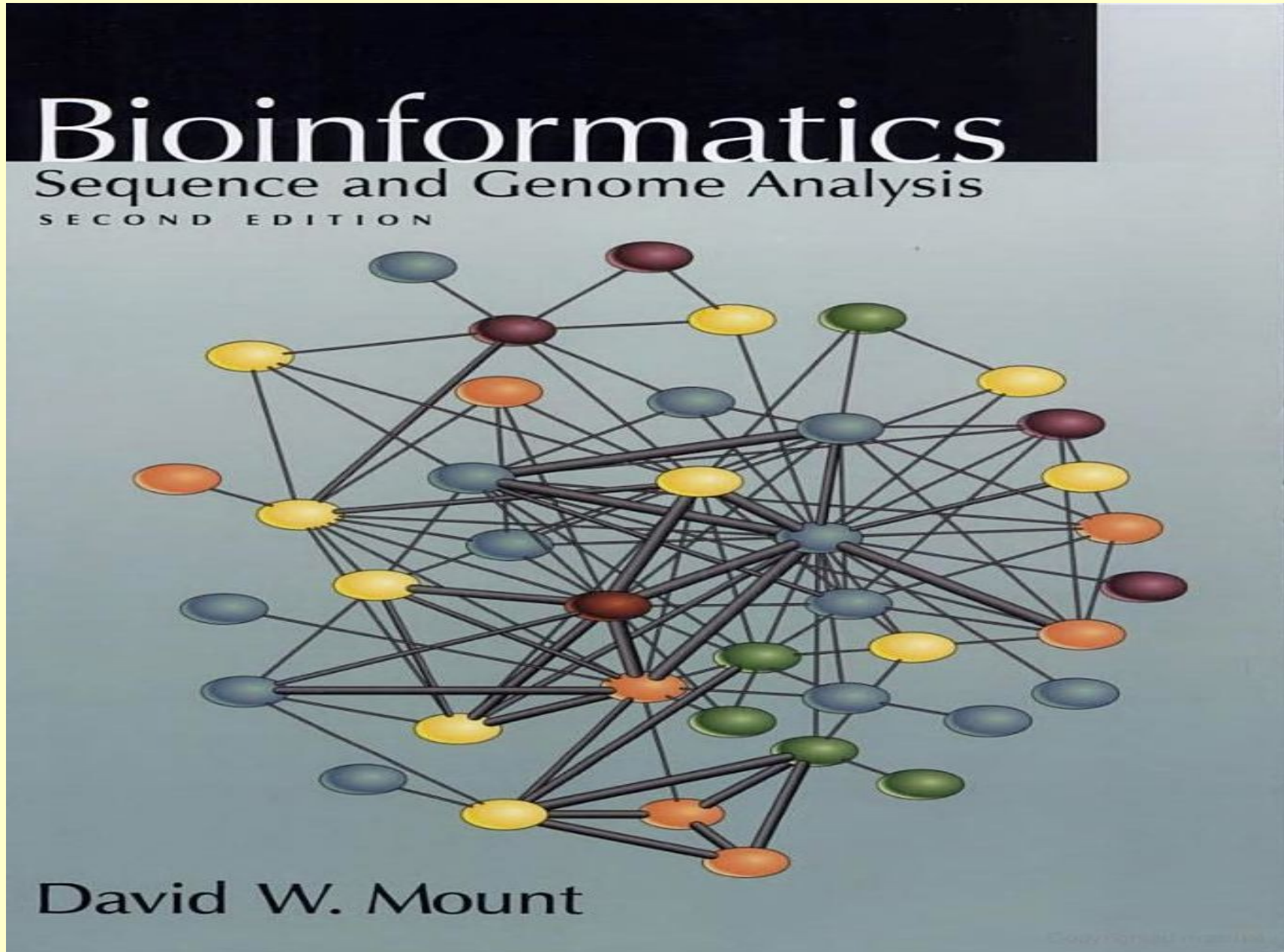
# Course Requirements

---

- Lectures
  - Theoretical background of current methods
  - Strengths and weaknesses of current approaches
  - Future directions for improvements
- Demonstrations
  - Applications (Mac, PC, Unix, Web)
  - Web applications
  - Illustrate homework
- All homework and questions must be submitted by email to [homework218@cmgm.stanford.edu](mailto:homework218@cmgm.stanford.edu)
- Several homework assignments (35%)
  - Due one week after assigned
- Final project (Due March 12th)
  - A critical or comparative review of computational approaches to any problem in computational molecular biology
  - Propose new approach
  - Implement a new approach
  - Examples of previous projects for the class can be found at <http://biochem218.stanford.edu/Projects.html>

David Mount

Bioinformatics: Sequence and Genome Analysis 2<sup>nd</sup> Edition

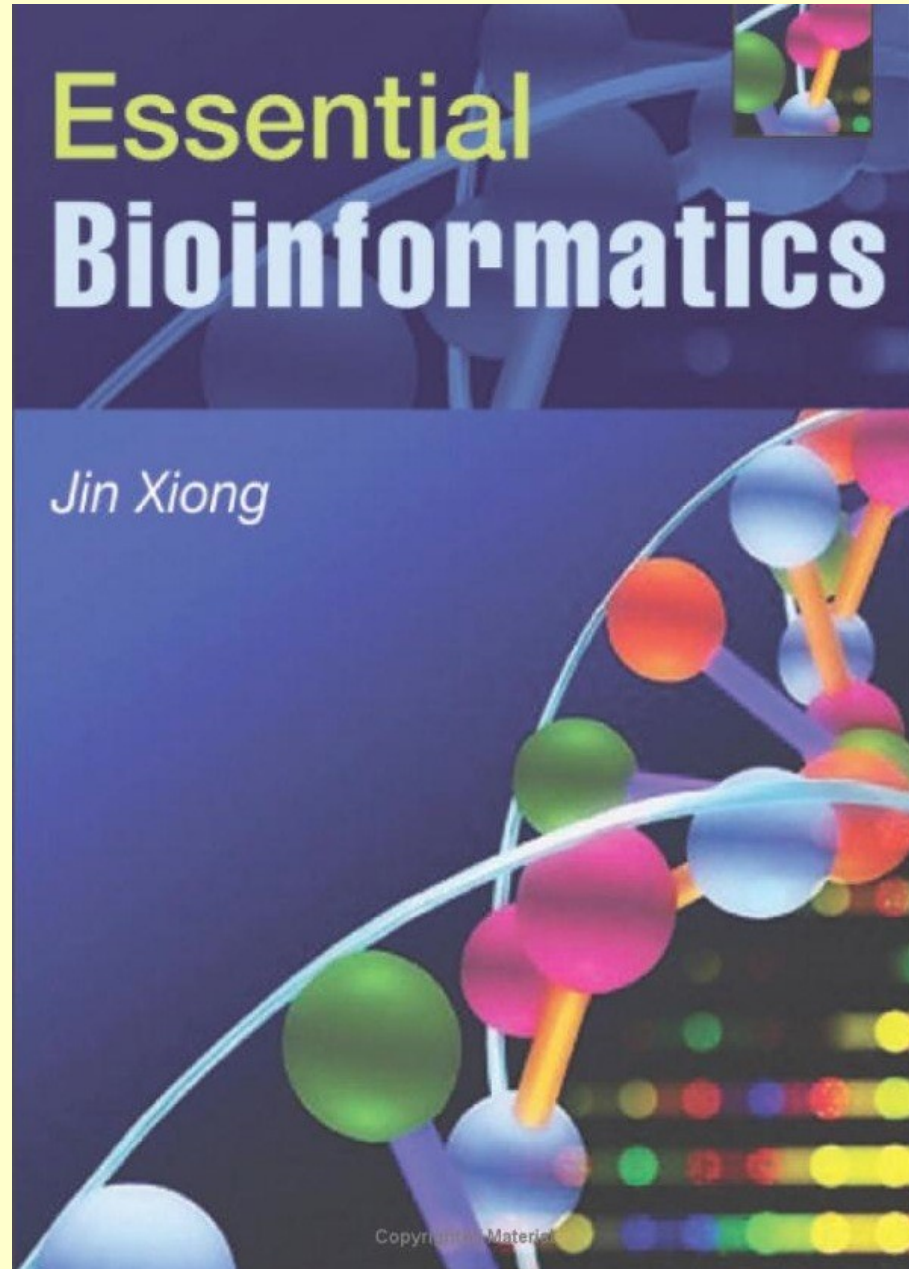


Doug Brutlag 2010

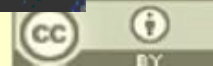
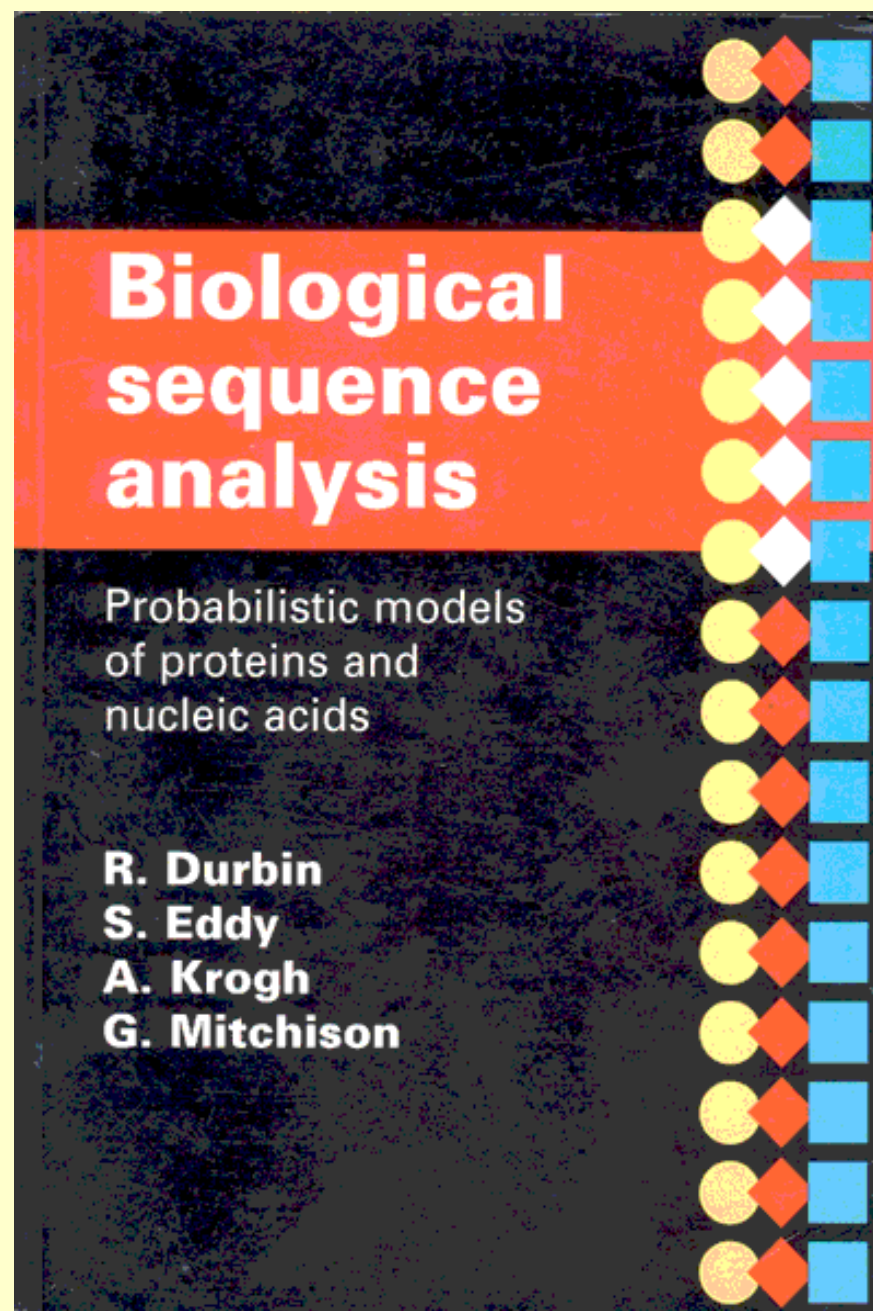
# Jin Xiong

## Essential Bioinformatics

---



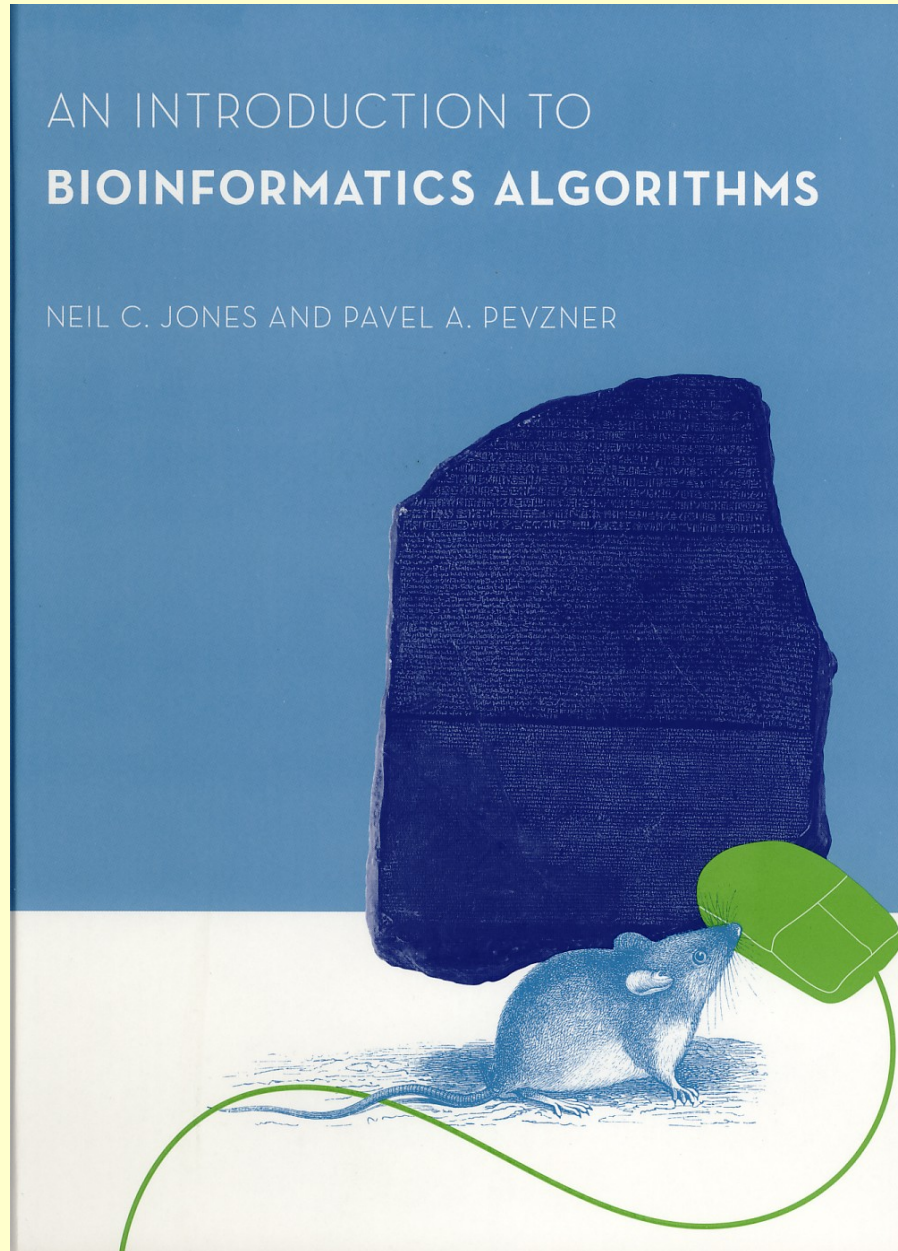
# Richard Durbin *et al.* Biological Sequence Analysis



# Jones & Pevzner

## Bioinformatics Algorithms

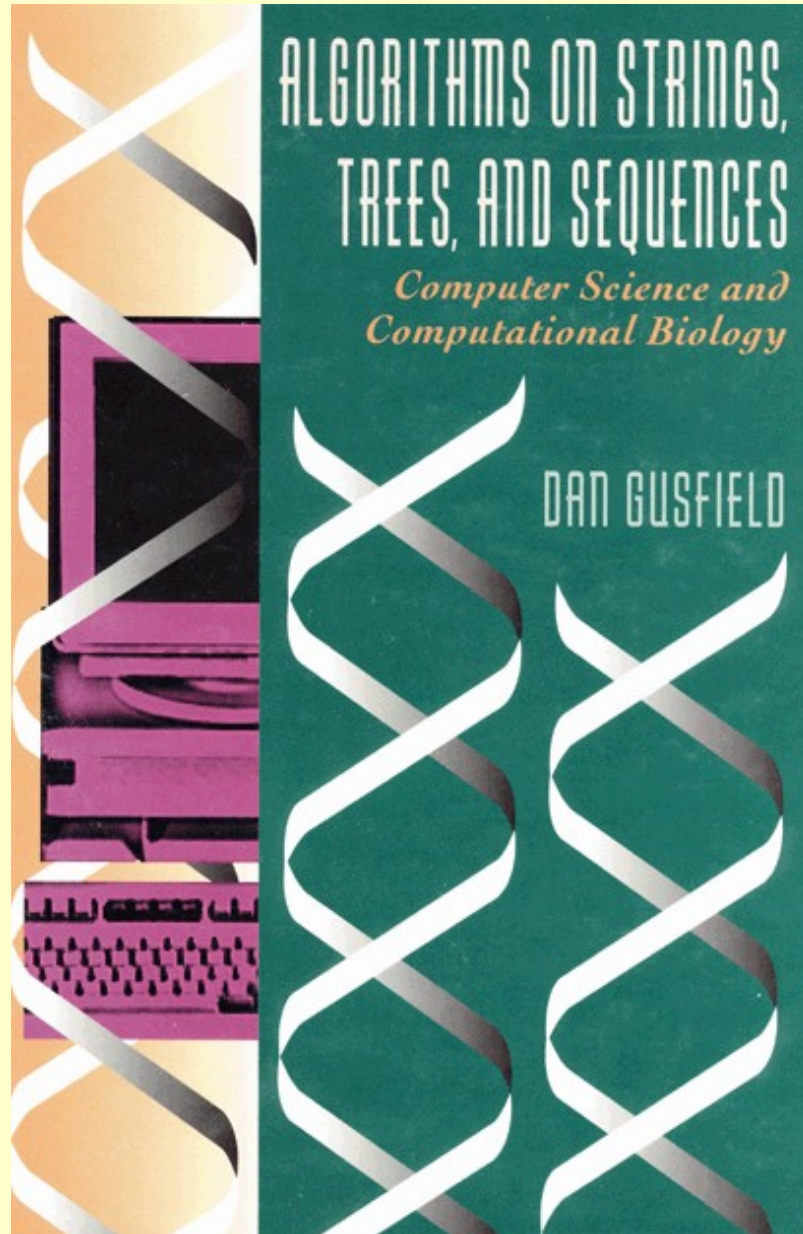
---





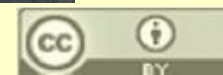
# Dan Gusfield

## Algorithms on Strings, Trees & Sequences



# Baldi & Brunak

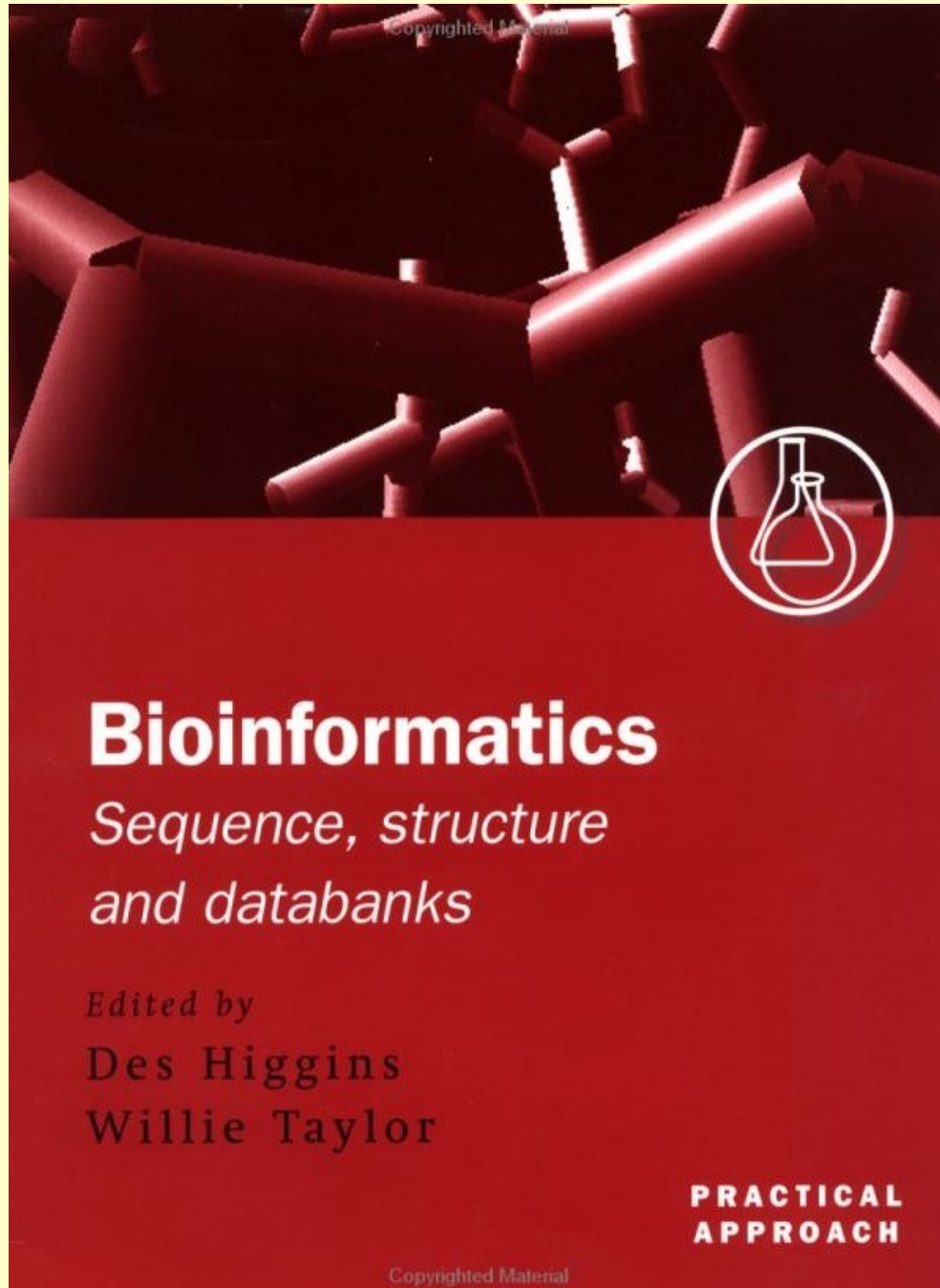
## Bioinformatics: The Machine Learning Approach



# Higgins & Taylor

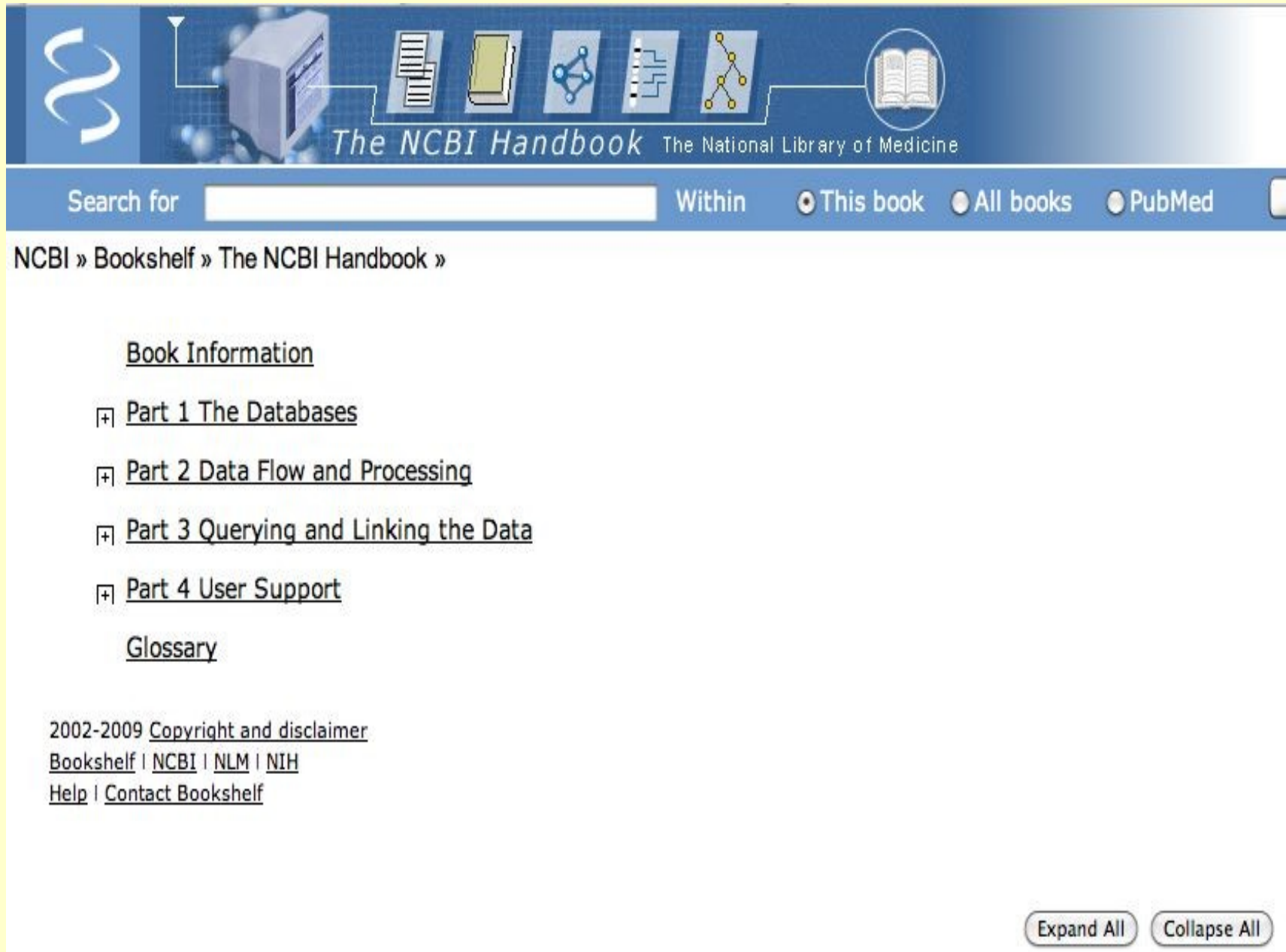

## Bioinformatics: Sequence, Structure & Databanks

---



# NCBI Handbook

<http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook>



**The NCBI Handbook** The National Library of Medicine

Search for  Within  This book  All books  PubMed

NCBI » Bookshelf » The NCBI Handbook »

Book Information

- Part 1 The Databases
- Part 2 Data Flow and Processing
- Part 3 Querying and Linking the Data
- Part 4 User Support

Glossary


2002-2009 Copyright and disclaimer  
Bookshelf | NCBI | NLM | NIH  
Help | Contact Bookshelf

Expand All Collapse All



# NCBI Handbook

<http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook>



The NCBI Handbook The National Library of Medicine

Search for  Within  This book

NCBI » Bookshelf » The NCBI Handbook » The Databases

## The Databases

---

[Chapter 1 GenBank: The Nucleotide Sequence Database](#)

[Chapter 2 PubMed: The Bibliographic Database](#)

[Chapter 3 Macromolecular Structure Databases](#)

[Chapter 4 The Taxonomy Project](#)

[Chapter 5 The Single Nucleotide Polymorphism Database \(dbSNP\) of Nucleotide Sequence Variation](#)

[Chapter 6 The Gene Expression Omnibus \(GEO\): A Gene Expression and Hybridization Repository](#)

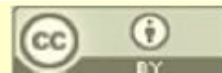
[Chapter 7 Online Mendelian Inheritance in Man \(OMIM\): A Directory of Human Genes and Genetic Disorders](#)

[Chapter 8 The NCBI BookShelf: Searchable Biomedical Books](#)

[Chapter 9 PubMed Central \(PMC\): An Archive for Literature from Life Sciences Journals](#)


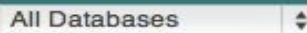


[Chapter 10 The SKY/CGH Database for Spectral Karyotyping and Comparative Genomic Hybridization Data](#)

[Chapter 11 The Major Histocompatibility Complex Database, dbMHC](#)





# EMBL-EBI Home Page

<http://www.ebi.ac.uk/>




---

[Databases](#) [Tools](#) [EBI Groups](#) [Training](#) [Industry](#) [About Us](#) [Help](#) [Site Index](#)  

### Data Resources & Tools

<ul style="list-style-type: none"><li>EMBL-BANK</li><li>UniProt</li><li>ArrayExpress</li><li>Ensembl</li><li>InterPro</li><li>PDBe</li></ul>	<ul style="list-style-type: none"><li>Genomes</li><li>Nucleotide Sequences</li><li>Protein Sequences</li><li>Macromolecular Structures</li><li>Small Molecules</li></ul>	<ul style="list-style-type: none"><li>Gene Expression</li><li>Molecular Interactions</li><li>Reactions &amp; Pathways</li><li>Protein Families</li><li>Enzymes</li></ul>	<ul style="list-style-type: none"><li>Literature</li><li>Taxonomy</li><li>Ontologies</li><li>Patent Resources</li></ul>	<ul style="list-style-type: none"><li>Sequence Similarity &amp; Analysis</li><li>Pattern &amp; Motif Searches</li><li>Structure Analysis</li><li>Text Mining</li><li>Downloads</li><li>Web Services</li></ul>
--	--	--	---	---

---



## European Bioinformatics Institute

---


### About the EBI

<ul style="list-style-type: none"><li>Research</li><li>PhD Studies</li><li>Training</li><li>Industry Support</li><li>Group &amp; Team Leaders</li><li>EBI Funders</li></ul>	<ul style="list-style-type: none"><li>User Support</li><li>EBI Mission</li><li>People</li><li>Events at the EBI</li><li>Genome Campus Events</li><li>How to Find us</li></ul>
---	---

### EBI Hosted Project Websites

<ul style="list-style-type: none"><li>1000 Genomes</li><li>BioCatalogue</li><li>BioSapiens</li><li>E-MeP</li><li>EGA</li><li>ELIXIR</li><li>EMBRACE</li><li>EMERALD</li></ul>	<ul style="list-style-type: none"><li>ENFIN</li><li>FELICS</li><li>IMPACT</li><li>INSDC</li><li>LRG</li><li>SPINE</li><li>SYMBIOmatics</li></ul>
---	--

### Latest News



- New portal for plant genomics will support research into improved crops**  
08 October 2009  
Today sees the launch of [Ensembl Plants](#) – a freely available web resource for plant genomics research – by EMBL-EBI, in partnership with the Cold Spring Harbor Laboratory, USA. Ensembl Plants allows researchers worldwide to access and visualise the results of genome-scale experiments in different plant species and will make it easier for scientists to improve the productivity and health of crops... more

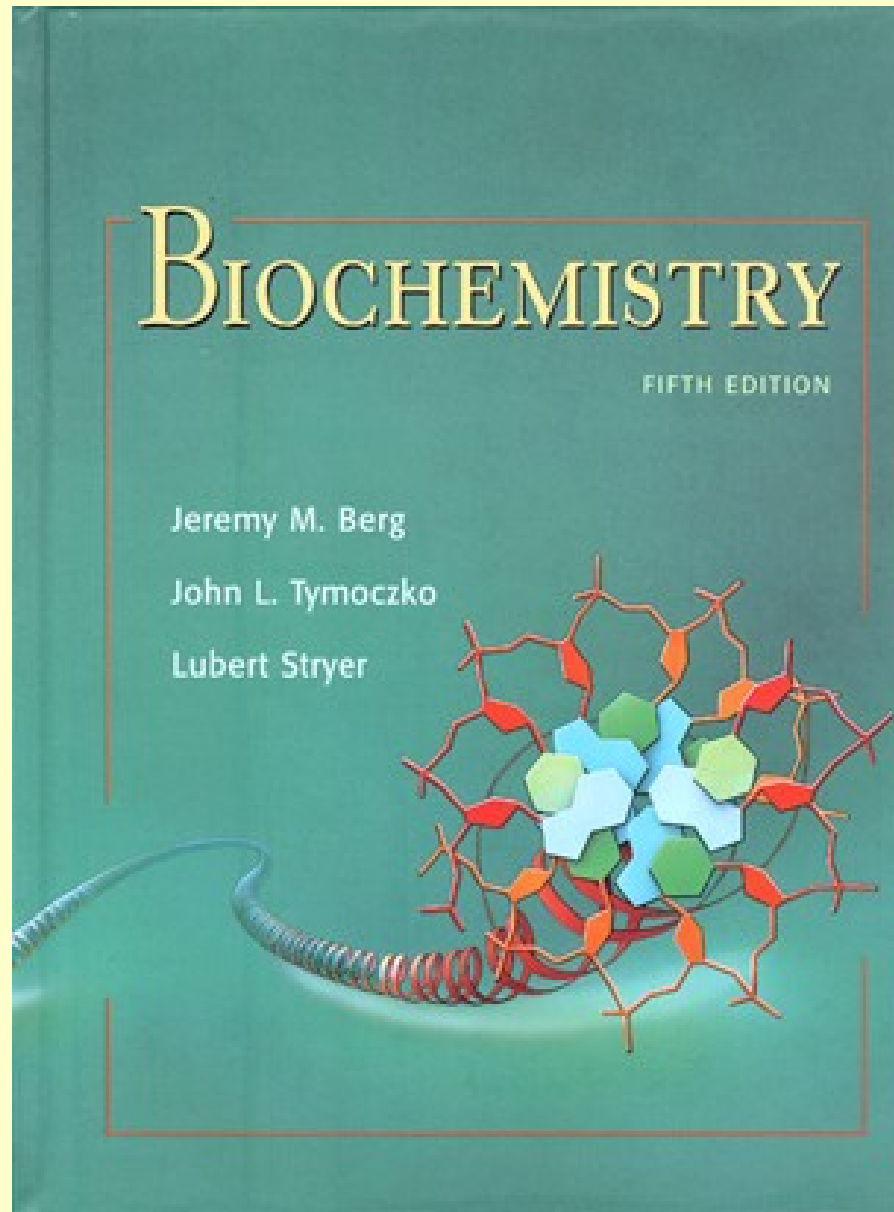
### Research Highlights

- EMBL-EBI articles are top of the list**  
20 November 2009  
Articles on three resources hosted by EMBL-EBI ([PDBe](#), [Ensembl Genomes](#) and [Gene Expression Atlas](#)) are highlighted as featured articles in the latest Database issue of Nucleic Acids Research. Featured articles are selected by the journal's Executive Editors based upon their originality, significance and scientific excellence ... more



# Berg, Tymoczko & Stryer Biochemistry, Fifth Edition

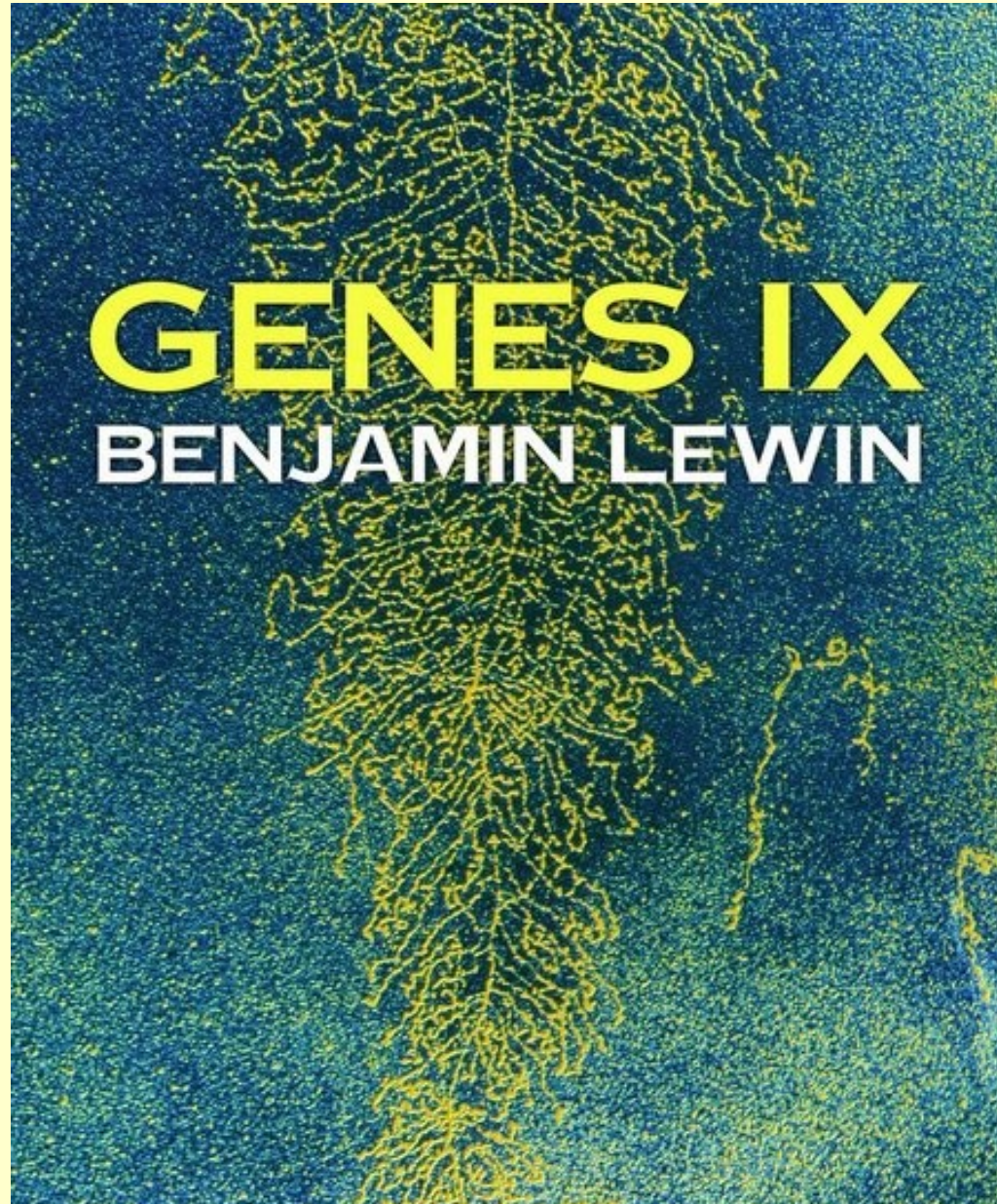
---



# Benjamin Lewin

## Genes IX

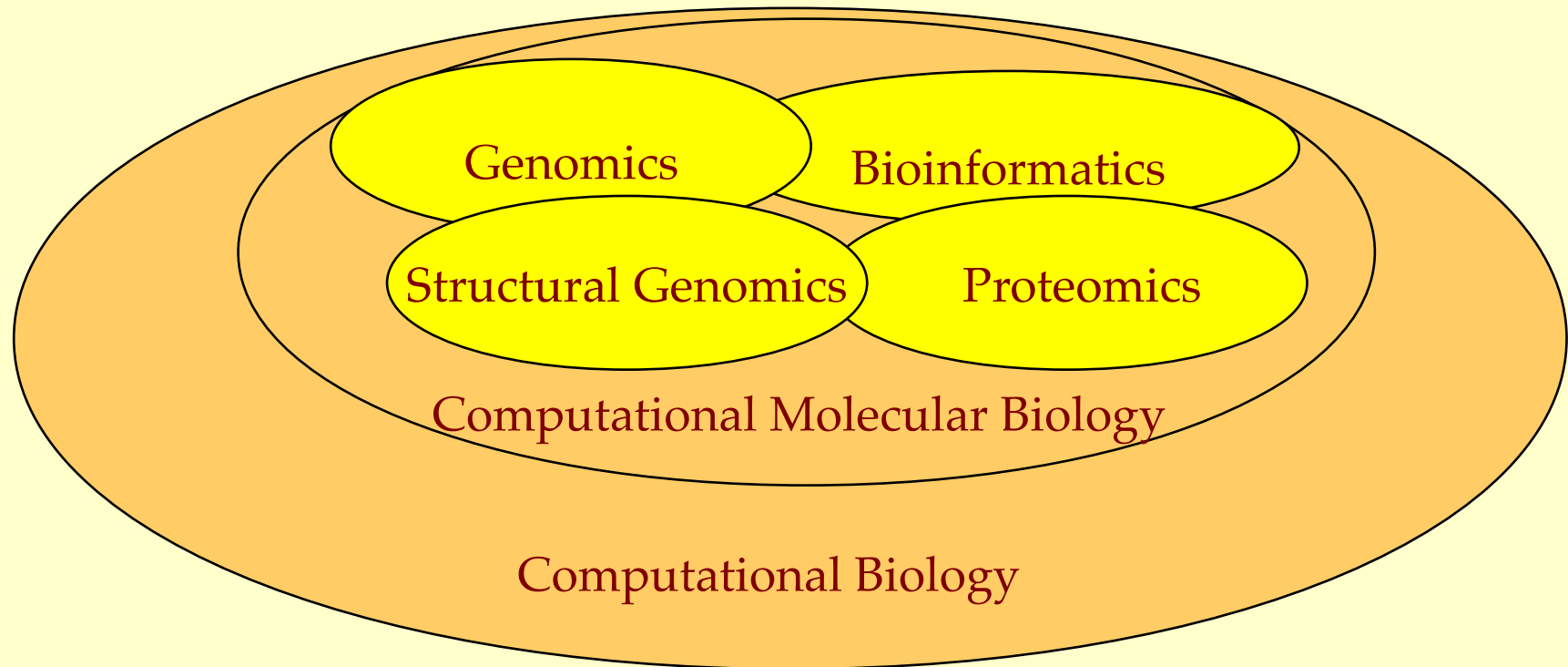
---





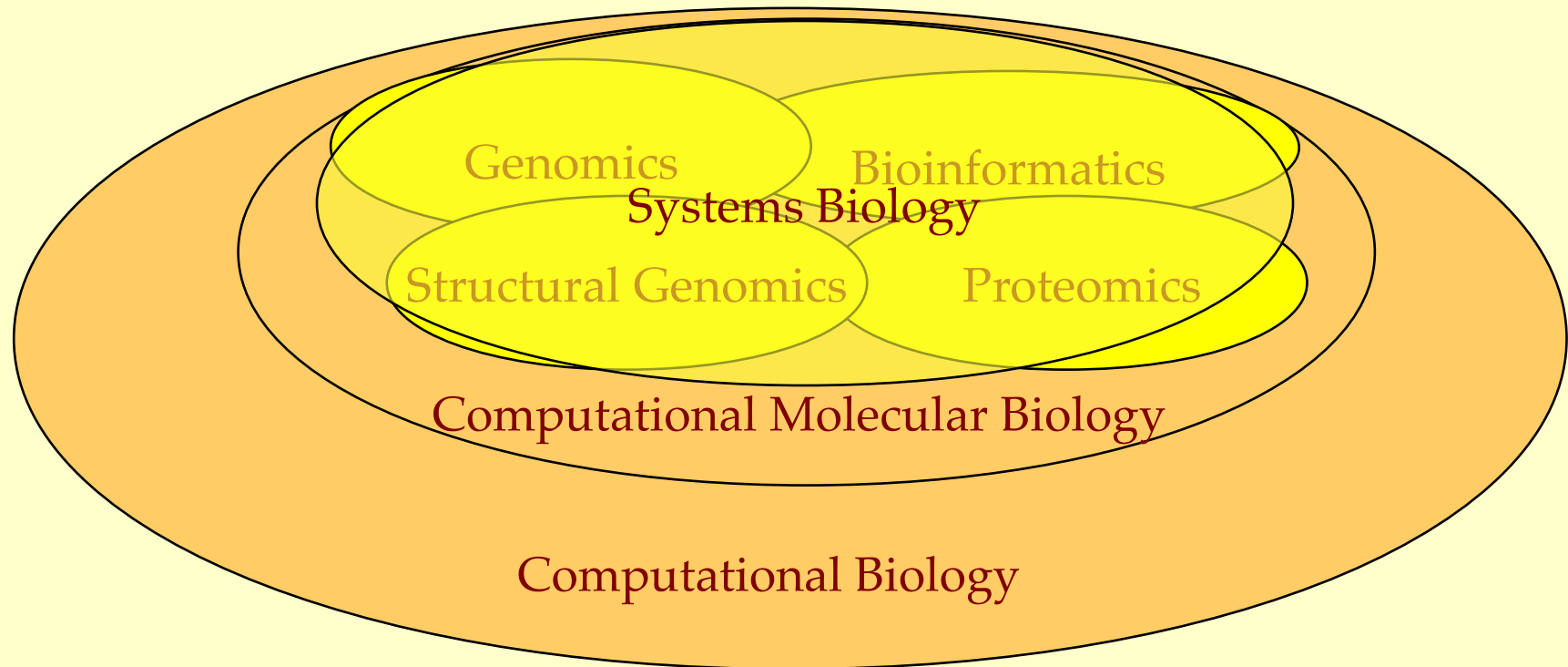
# Genomics, Bioinformatics & Computational Biology

---

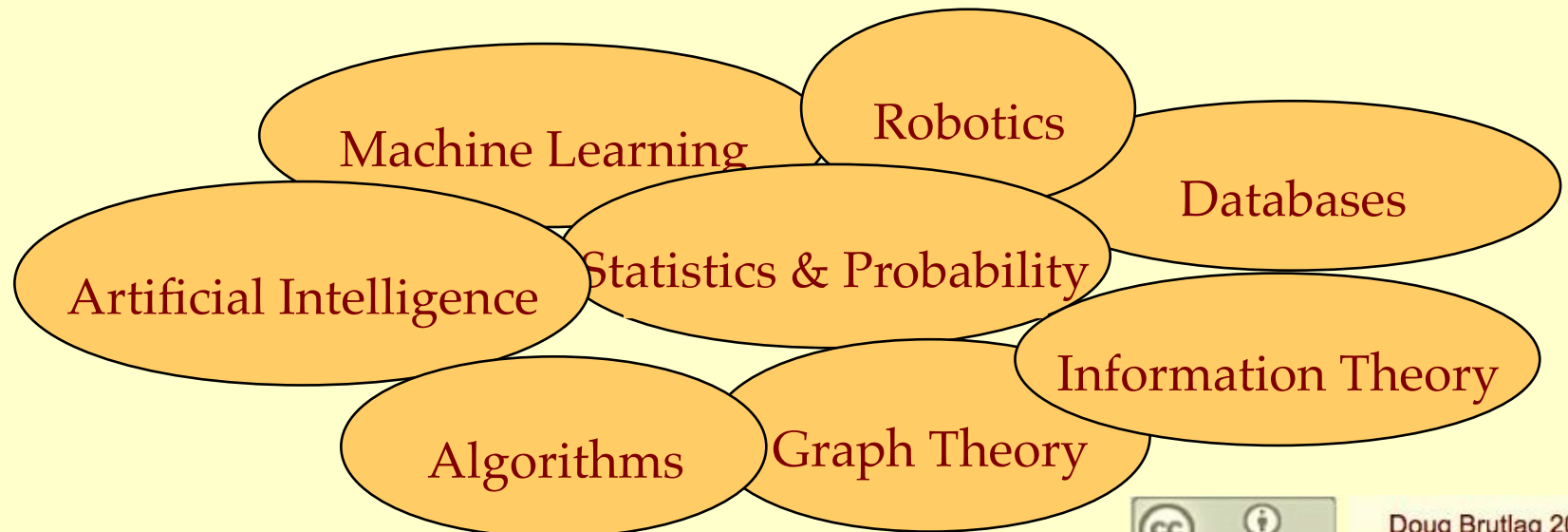
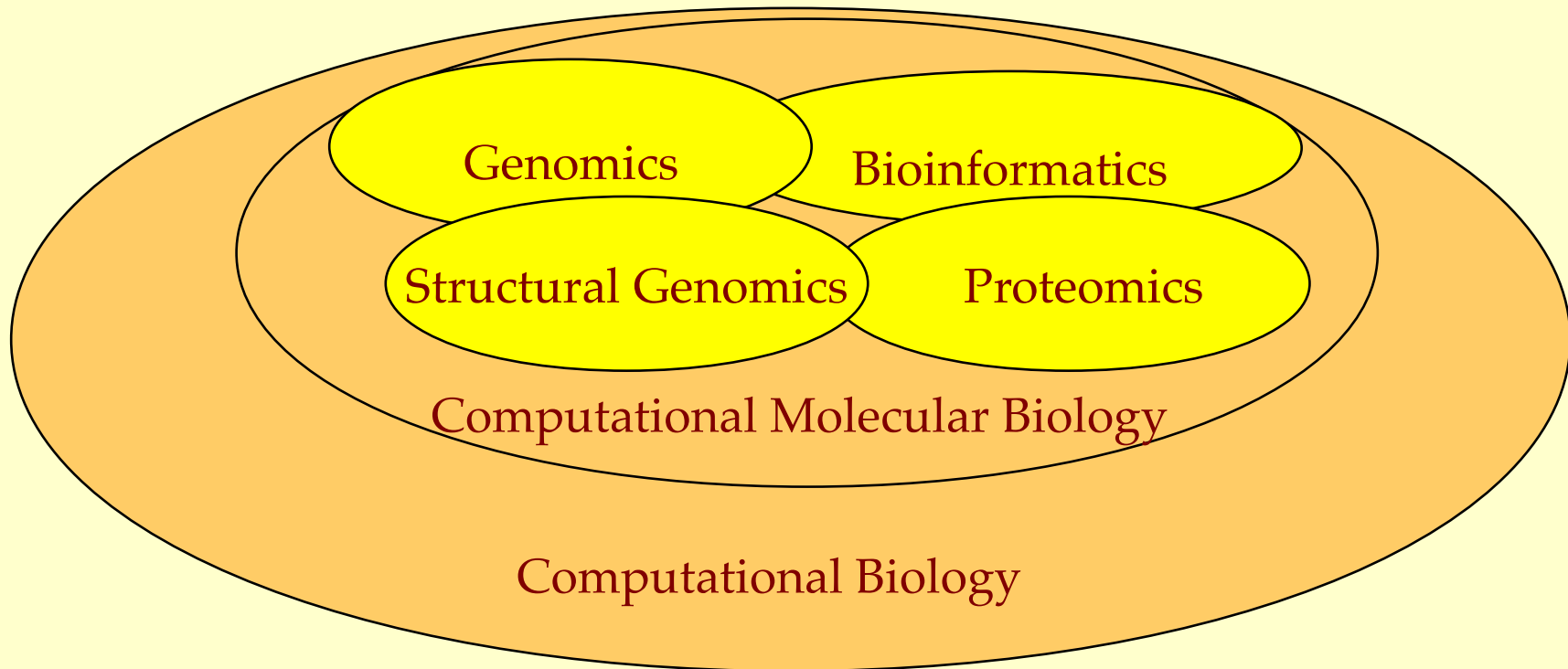


# Genomics, Bioinformatics & Computational Biology

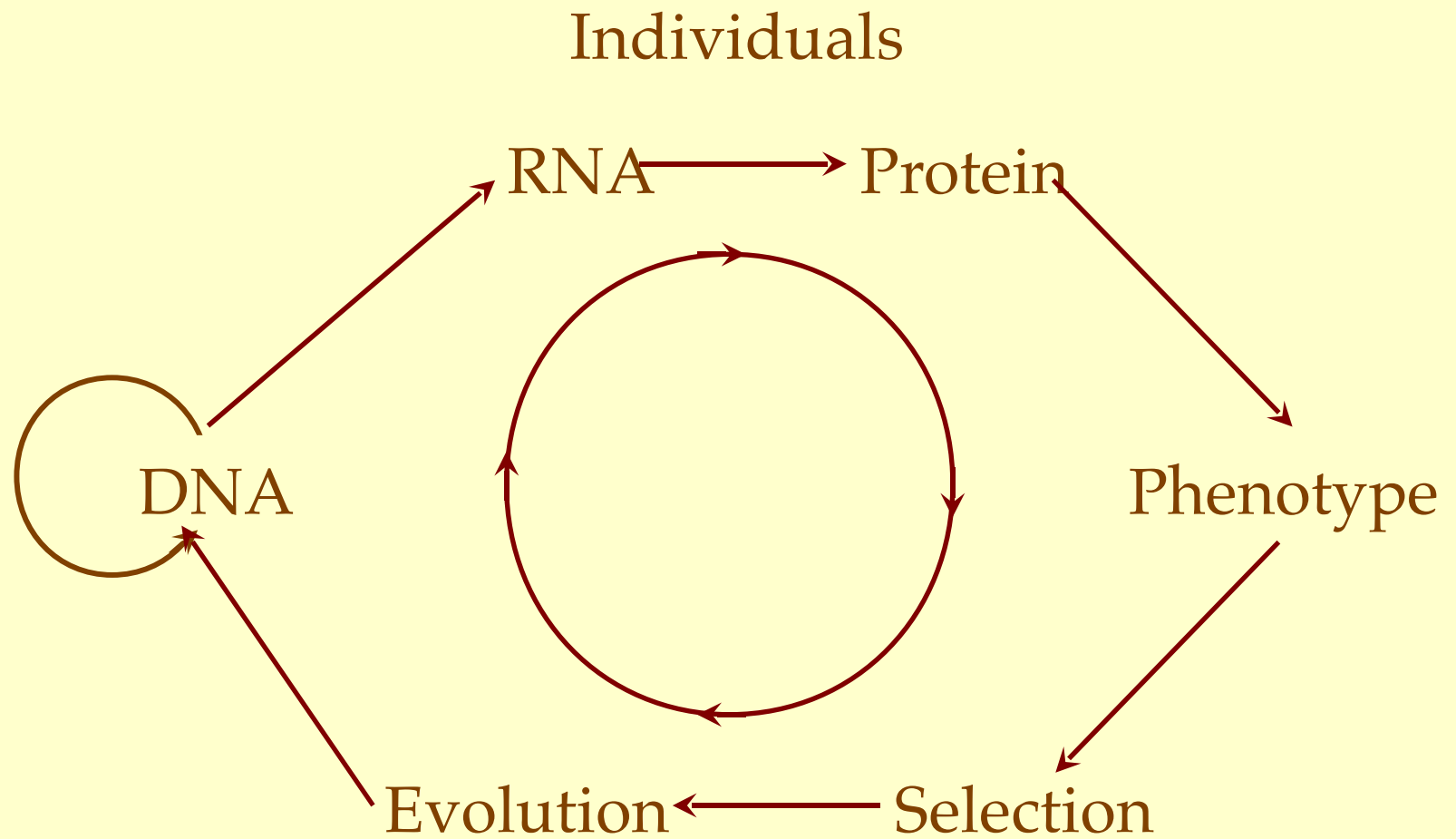
---



# Genomics, Bioinformatics & Computational Biology

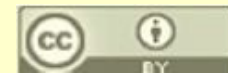


# What is Bioinformatics?



Populations

Biological Information



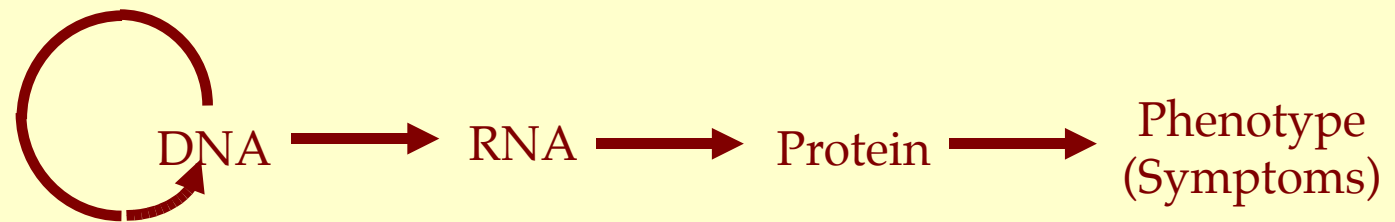
# Computational Goals of Bioinformatics

---

- Learn & Generalize: Discover conserved patterns (models) of sequences, structures, interactions, metabolism & chemistries from well-studied examples.
- Prediction: Infer function or structure of newly sequenced genes, genomes, proteins or proteomes from these generalizations.
- Organize & Integrate: Develop a systematic and genomic approach to molecular interactions, metabolism, cell signaling, gene expression...
- Simulate: Model gene expression, gene regulation, protein folding, protein-protein interaction, protein-ligand binding, catalytic function, metabolism...
- Engineer: Construct novel organisms or novel functions or novel regulation of genes and proteins.
- Gene Therapy: Target specific genes, or mutations, RNAi to change a disease phenotype.

# Central Paradigm of Molecular Biology

---



# Molecular Biology of the Gene 1965

MOLECULAR BIOLOGY  
OF THE GENE



JAMES D. WATSON

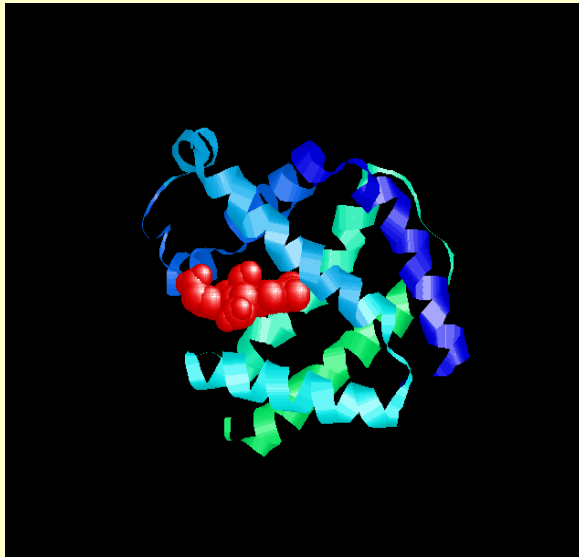


# Central Paradigm of Bioinformatics

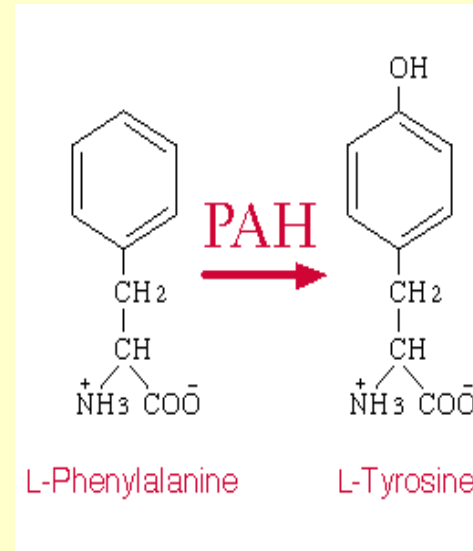
Genetic Information

MVHLTPEEKT  
AVNALWGKVN  
VDAVGGEALG  
RLLVVYPWTQ  
RFFESFGDLS  
SPDAVMGNPK  
VKAHGKKVLG  
AFSDGLAHL  
NLKGTFSQLS  
ELHCDKLHVD  
PENFRLLGNV  
LVCVLARNFG  
KEFTPQMCAA  
YQKVVAGVAN  
ALAHKYH

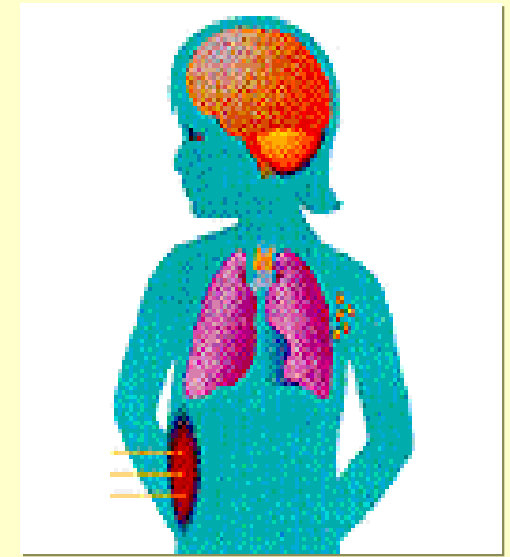
Molecular Structure



Biochemical Function



Phenotype (Symptoms)

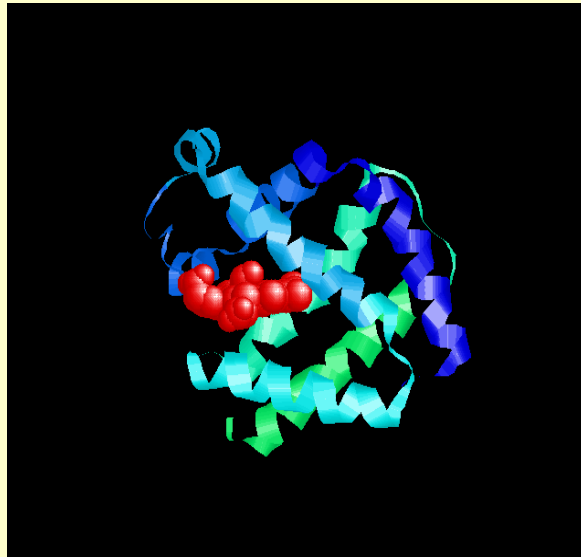




# Central Paradigm of Bioinformatics

Genetic Information

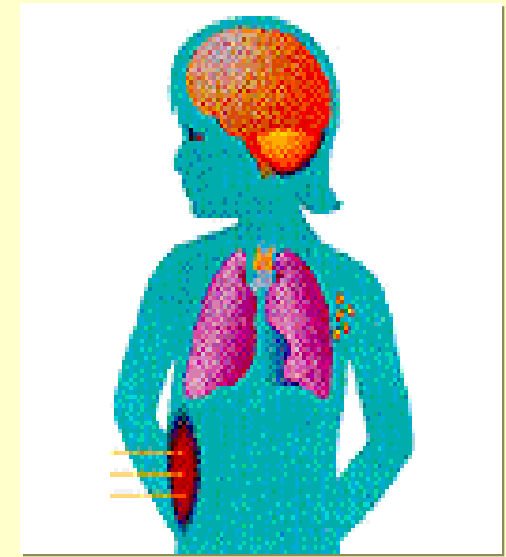
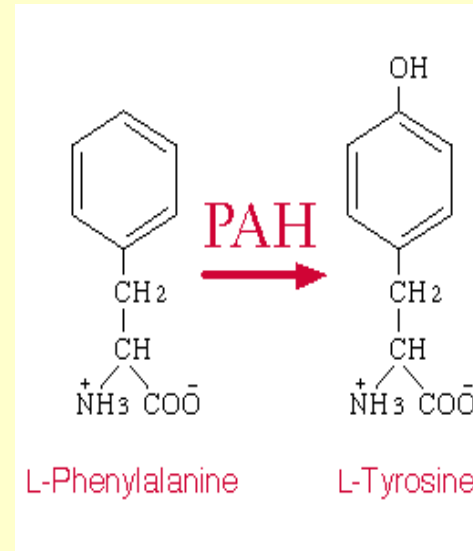
MVHLTPEEKT  
AVNALWGKVN  
VDAVGGEALG  
RLLVVYPWTQ  
RFFESFGDLS  
SPDAVMGNPK  
VKAHGKKVLG  
AFSDGLAHLA  
NLKGTFSQLS  
ELHCDKLHVD  
PENFRLLGNV  
LVCVLARNFG  
KEFTPQMCAA  
YQKVVAGVAN  
ALAHKYH



Molecular Structure

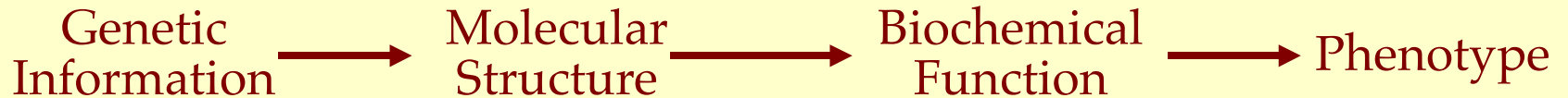
Biochemical Function

Phenotype (Symptoms)



# Challenges Understanding Genetic Information

---



- Genetic information is redundant
- Structural information is redundant
- Genes and proteins are meta-stable
- Single genes have multiple functions
- Genes are one dimensional but function depends on three-dimensional structure

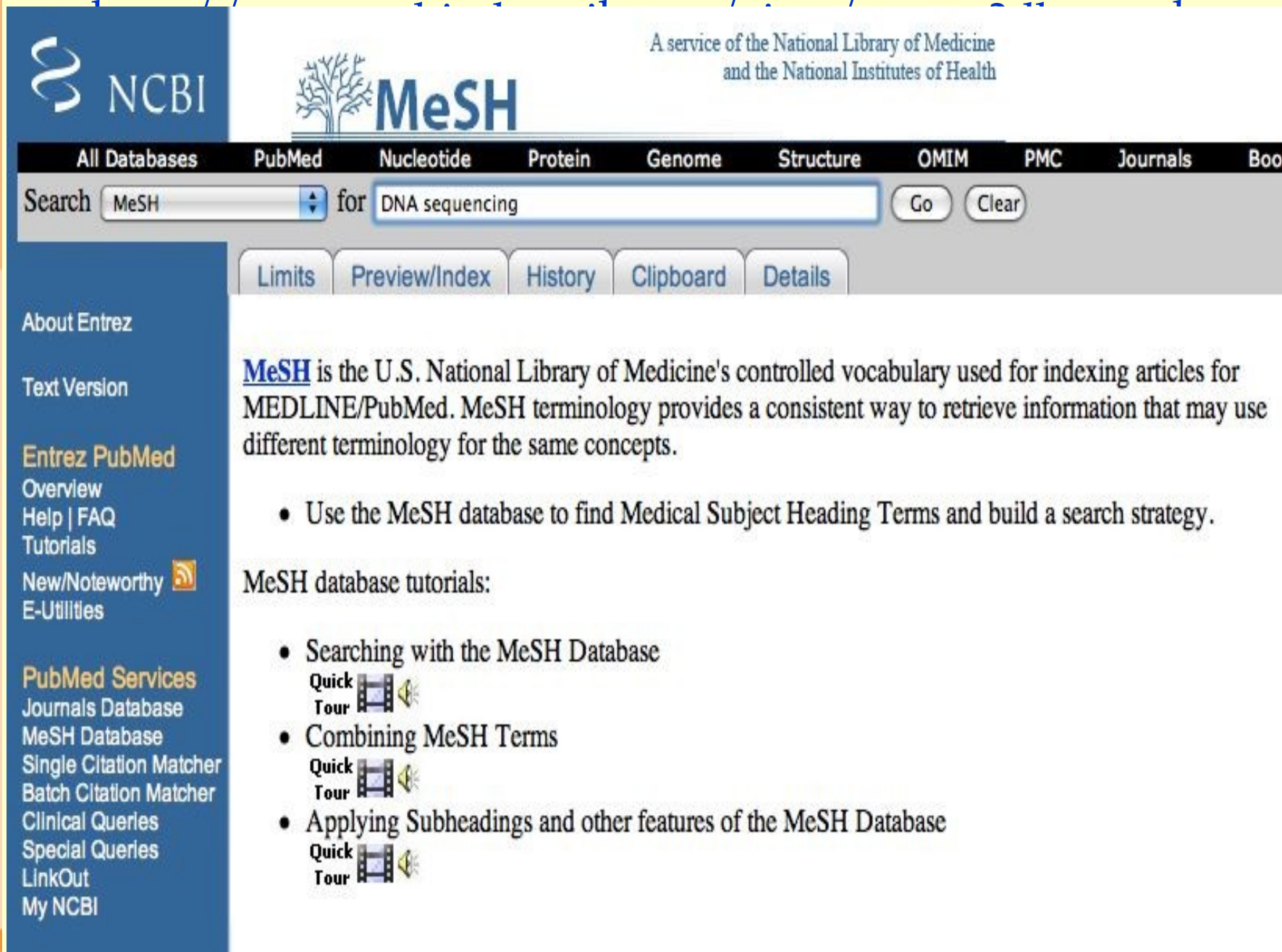
# Redundancy in Genomic & Protein Sequences

---

- DNA is double-stranded
- Genetic code
- Acceptable amino-acid replacements
- Intron-exon variation
- Alternative splicing
- Strain variations (SNPs)
- Sequencing errors



# Using A Controlled Vocabulary for Literature Search



NCBI

A service of the National Library of Medicine  
and the National Institutes of Health

MeSH

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Boo

Search MeSH for DNA sequencing Go Clear


Limits Preview/Index History Clipboard Details

About Entrez

Text Version

**Entrez PubMed**

Overview  
Help | FAQ  
Tutorials

New/Noteworthy   
E-Utilities




**PubMed Services**

Journals Database  
MeSH Database  
Single Citation Matcher  
Batch Citation Matcher  
Clinical Queries  
Special Queries  
LinkOut  
My NCBI

**MeSH** is the U.S. National Library of Medicine's controlled vocabulary used for indexing articles for MEDLINE/PubMed. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts.

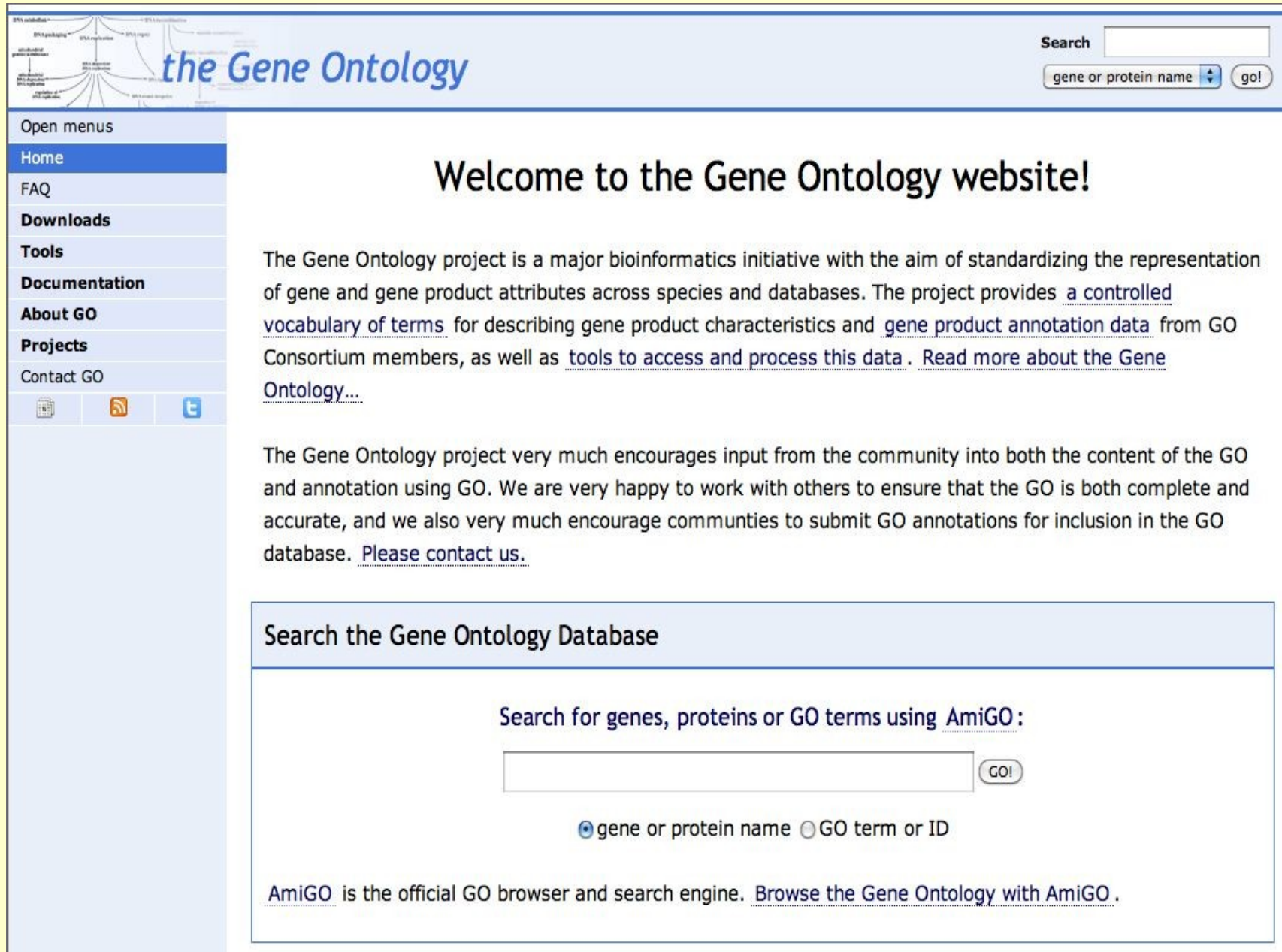

- Use the MeSH database to find Medical Subject Heading Terms and build a search strategy.

MeSH database tutorials:

- Searching with the MeSH Database  
Quick Tour 
- Combining MeSH Terms  
Quick Tour 
- Applying Subheadings and other features of the MeSH Database  
Quick Tour 

# Gene Ontology Database

<http://www.geneontology.org/>





**the Gene Ontology**

Search   
gene or protein name

Open menus

- Home
- FAQ
- Downloads
- Tools
- Documentation
- About GO
- Projects
- Contact GO

## Welcome to the Gene Ontology website!

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides [a controlled vocabulary of terms](#) for describing gene product characteristics and [gene product annotation data](#) from GO Consortium members, as well as [tools to access and process this data](#). [Read more about the Gene Ontology...](#)

The Gene Ontology project very much encourages input from the community into both the content of the GO and annotation using GO. We are very happy to work with others to ensure that the GO is both complete and accurate, and we also very much encourage communities to submit GO annotations for inclusion in the GO database. [Please contact us.](#)

### Search the Gene Ontology Database

Search for genes, proteins or GO terms using [AmiGO](#) :

gene or protein name  GO term or ID

[AmiGO](#) is the official GO browser and search engine. [Browse the Gene Ontology with AmiGO.](#)



# UCSC Genome Browser

<http://genome.ucsc.edu/>

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS

### UCSC Genome Browser on Human Mar. 2006 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr20:5,043,600-5,055,268 jump clear size 11,669 bp. configure

chr20 (p12.3) 20p13.1 20p12.1 q12 q13.2

chr20: 5045000 | 5050000 | 5055000

STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps

Gap Locations

UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA

RefSeq Genes

Non-Human RefSeq Genes

Bos MGC128125  
Rattus Pcna  
Mus Pcna  
Xenopus pcna  
Danio pcna  
Drosoph mus209  
Arabid PCNA1  
Arabid PCNA2  
Drosoph mus209  
Bombyx Pcna  
Gallus PCNA  
Hyza Os02g0805200  
Drosoph CG10262

AceView Gene Models With Alt-Splicing (lifted from hg17)

Human mRNAs from GenBank

CR617091  
CR619580  
M15796  
BC000491  
BC062439  
CR617056  
G43531  
CR536501  
DQ894581  
CR541799

Spliced ESTs

Human ESTs That Have Been Spliced

UniGene Alignments

Hs.147433  
Hs.601337  
Hs.603585

Vertebrate Multiz Alignment & Conservation (17 Species)

Conservation

mouse  
rat  
dog  
opossum  
chicken  
x\_tropicalis  
tetraodon

Simple Nucleotide Polymorphisms (dbSNP build 126)

SNPs

Repeating Elements by RepeatMasker



RepeatMasker

move start Click on a feature for details. Click on base position to zoom in move end  
< 2.0 > around cursor. Click on left mini-buttons for track-specific options. < 2.0 >



# ExPASy Proteomics Server

<http://www.expasy.ch/doc.html>



Search  for

## ExPASy Proteomics Server

Databases Tools Services Mirrors About Contact

You are here: [ExPASy CH](#) > [About](#)

### Complete table of available documents

#### ExPASy

- [What's New on ExPASy](#)
- [Quick Guide to ExPASy](#)
- [SWISS-FLASH](#) electronic bulletins
- [How to create HTML links to services on ExPASy](#)
- [Disclaimer](#)
- [ExPASy: the proteomics server for in-depth protein knowledge and analysis](#)

Databases	Tools and Software Packages
<ul style="list-style-type: none"><li>• <b>Swiss-Prot and TrEMBL, the UniProt Knowledgebase</b> - Protein knowledgebase<ul style="list-style-type: none"><li>◦ <a href="#">User manual &amp; release notes</a></li><li>◦ <a href="#">General documents</a></li><li>◦ <a href="#">Nomenclature documents</a></li><li>◦ <a href="#">Species specific documents</a></li><li>◦ <a href="#">Other documents</a></li><li>◦ <a href="#">Printable Quick Guide to UniProtKB</a></li></ul></li><li>• <b>PROSITE</b> - Protein families and domains<ul style="list-style-type: none"><li>◦ <a href="#">User manual &amp; release notes</a></li><li>◦ <a href="#">List of documentation entries</a></li><li>◦ <a href="#">Syntax of PROSITE patterns</a></li><li>◦ <a href="#">Generalised profile syntax</a></li></ul></li><li>• <b>SWISS-2DPAGE</b> - Two-dimensional polyacrylamide gel electrophoresis<ul style="list-style-type: none"><li>◦ <a href="#">User manual &amp; release notes</a></li><li>◦ <a href="#">FAQ</a> - (Frequently Asked Questions about SWISS-2DPAGE)</li><li>◦ <a href="#">Protocols</a></li></ul></li><li>• <b>ENZYME</b> - Enzyme nomenclature<ul style="list-style-type: none"><li>◦ <a href="#">User manual &amp; release notes</a></li></ul></li></ul>	<ul style="list-style-type: none"><li>• <a href="#">UniProt web site</a></li><li>• <a href="#">Swiss-Shop</a>: automatically obtain (by email) new sequence entries relevant to your field(s) of interest</li><li>• <a href="#">Protein identification and characterization</a><ul style="list-style-type: none"><li>◦ <a href="#">AAComplident</a> - Identify proteins with amino acid composition</li><li>◦ <a href="#">AACompSim</a> - Compare the amino acid composition of a Swiss-Prot entry with all other entries</li><li>◦ <a href="#">Multident</a> - Identify proteins with <i>pI</i>, <i>Mw</i>, amino acid composition, sequence tag and peptide mass fingerprinting data</li><li>◦ <a href="#">TagIdent</a> - Identify proteins with <i>pI</i>, <i>Mw</i> and sequence tag, or generate a list of proteins close to a given <i>pI</i> and <i>Mw</i></li><li>◦ <a href="#">Aldente</a> - Identify proteins with peptide mass fingerprinting data, <i>pI</i> and <i>Mw</i></li><li>◦ <a href="#">FindMod</a> - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides</li><li>◦ <a href="#">GlycoMod</a> - Predict possible oligosaccharide structures that occur on proteins from their experimentally determined masses</li><li>◦ <a href="#">FindPept</a> - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications,</li></ul></li></ul>



# Inferring Biological Function from Protein Sequence

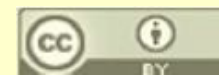
Consensus Sequences  
or Sequence Motifs

Zinc Finger (C2H2 type)  
C x {2,4} C x {12} H x {3,5} H

Sequences of Common  
Structure or Function

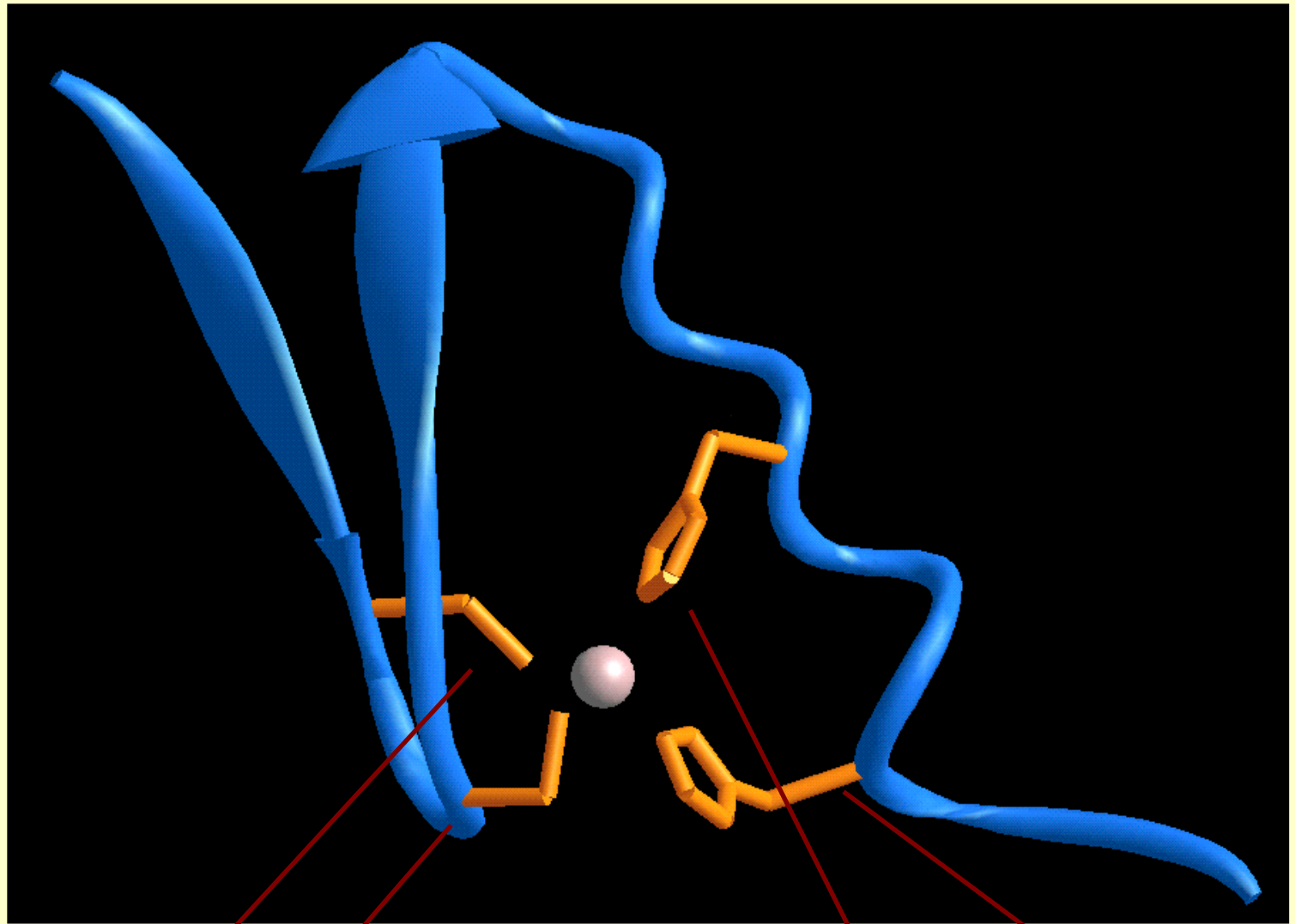
Sequence Similarity

	10	20	30	40	50
Query	VLSPADKTNVKA	AWGKVG	AHAGEVGA	EALERMFLSF	PTTKTYFPHF-----DLSHGS
	:  : :    :	:	: : : :	:	:
Match	HLTPEEKSAVT	ALWGKV--	NVDEYGG	EALGRLLV	VYPWTQRFFESFGDLSTPDAVMGN
	10	20	30	40	50





# A Typical Motif: Zinc Finger DNA Binding Motif



C . . C . . . . . H . . . . H



# Inferring Biological Function from Protein Sequence

## Weight Matrices or Position-Specific Scoring Matrices

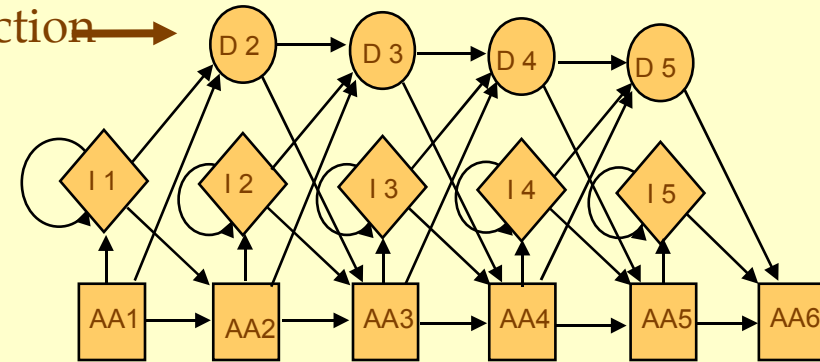
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	0	11	1	2	10	0	4	9	3	0	16
L	16	1	17	0	1	31	0	3	11	24	0	14
K	3	4	5	10	11	1	1	13	10	0	5	2
M	7	1	1	0	0	0	0	0	5	7	1	8
F	4	0	3	0	0	4	0	0	0	10	0	0
P	0	6	0	1	0	0	0	0	0	0	0	0
S	1	17	0	8	3	1	3	0	2	2	2	0
T	5	22	3	11	1	5	0	2	2	2	0	5
W	2	0	0	0	0	0	0	0	0	1	0	1
Y	1	0	4	2	0	1	0	0	2	4	0	1
V	6	3	1	1	2	15	0	0	2	12	0	28

## Consensus Sequences or Sequence Motifs

Zinc Finger (C2H2 type)  
 $C \times \{2,4\} C \times \{12\} H \times \{3,5\} H$

## Profiles, PSI-BLAST Sequences of Common Hidden Markov Models

Structure or Function



## Sequence Similarity

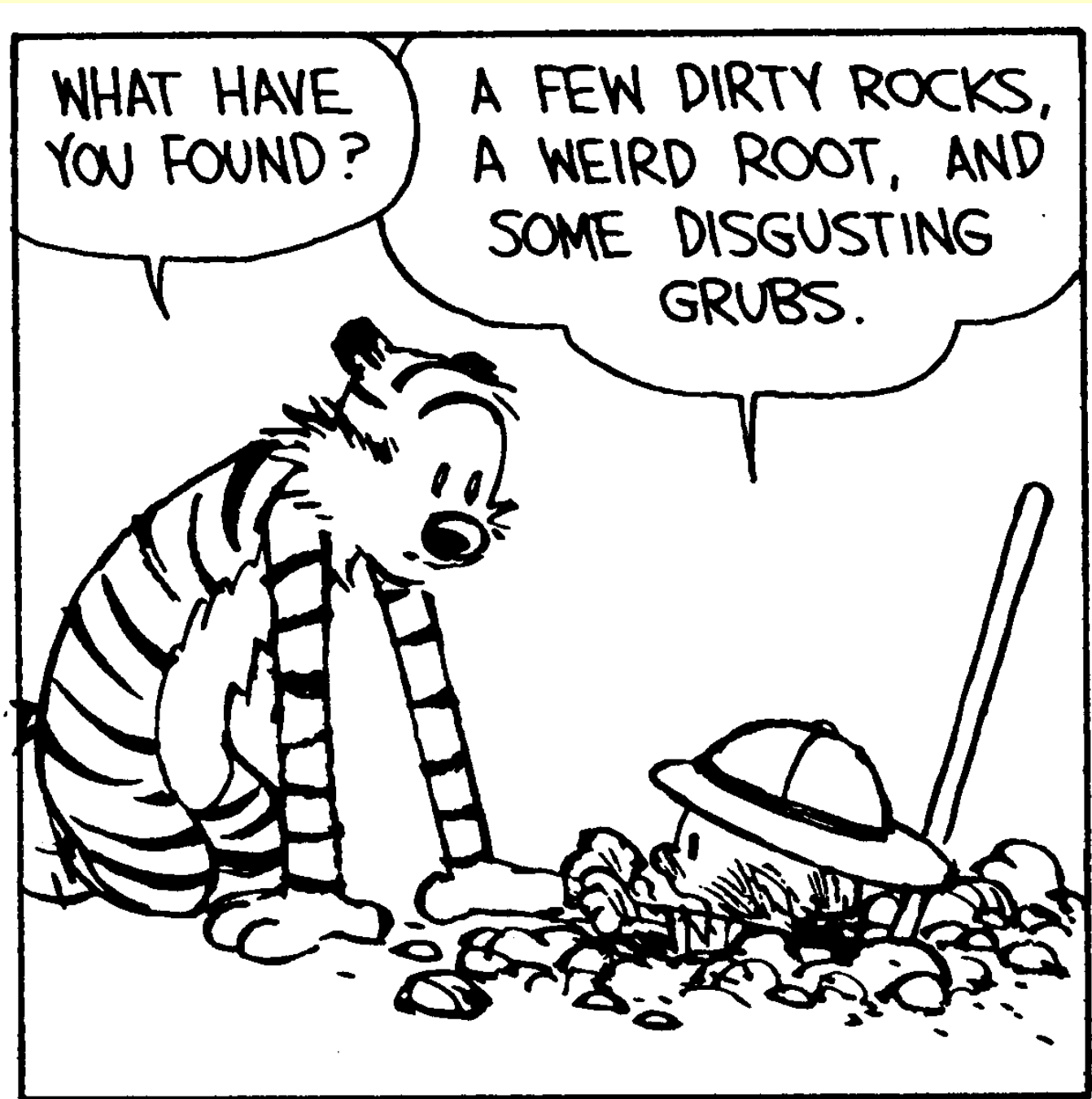
	10	20	30	40	50													
Query	VLSPADKTNVKA	AWGKVG	AHAGEVGA	EALERMFLSF	PTTKTYFPHF-----	DL	SHGS											
	:	: :	: :	: :	: :	: :	:	:										
Match	HLTPEEKSA	V	TALW	GKV--	NVDEY	GGEAL	GRLLV	VYPWT	Q	RFFES	F	GDL	STP	DA	V	M	G	N
	10	20	30	40	50													



# Buried Treasure



# Buried Treasure



# Buried Treasure



# Clustal Globin Alignment

- 1 human beta globin
- 2 horse beta globin
- 3 human alpha globin
- 4 horse alpha globin
- 5 cyanohaemoglobin
- 6 whale myoglobin
- 7 leghaemoglobin

```

          A          B          C
VHLTPEEKSAVTALWGKVNVND  EVGGEALGRLLVYPWVWQRFESFGDLSTPDAVMGNPK
VQLSAGEEKAAVLALWDKVNNEE  EVGGEALGRLLVYPWVWQRFDSFGDLSNPGAVMGNPK
VLSPADKTNVKAAWGKVGAAHAGEYGAELERMFLSFPTTKTYFPHF  DLSH  GSAQ
VLSAADKTNVKAAWSKVGGHAGEYGAELERMFLGFPTTKTYFPHF  DLSH  GSAQ
PIVDTGTSVAPLSAAEKTIRSAWAPVYSYDYESGVDILVKFFTSTFAAEFFPKFKGLTTADELKKKSAD
VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDKFKHLKTEAEMKASED
GALTESQAALVKSSWEFFNANIPKHTHRFFILVLEIAPAAKDLFSSFLKGGTSEVPQNNPE
          *   .   .   *           .   .   *   *   *
    
```

```

          E          F          G          H
VKAHGKKVLGAFSDG  LAHLDNLRKGTFFAT  LSELHCDKLHVDPENFRLLGNVLCVLAHHEGKEFTTPPVQAAAYQKVVAGVANALAHKYH
VKAHGKKVLHSEFGE  VHLDNLRKGTFAA  LSELHCDKLHVDPENFRLLGNVLVVVLARHEGKDFTPPELQASVYQKVVAGVANALAHKYH
VKGHGKKVADALTNA  VAHVDDMPNALS  LSDLHAHAKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR
VKAHGKKVGDALTNA  VGHLDDLPGALS  LSDLHAHAKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTISKYR
VRWHAERIIDAVIDA  VASMDDEKMSMKDLSGKHAKSFEVDPEYFKVLA AVIADTVAAGD  AGFEKLLRMICILLRSAY
LKKHGVTVLTAIGAI  LKKKGHEAELKP  LAQSHATKHKIPKYLEFISEAIIHVLHSRHPGDFGADAQCAMNKALELFRKDIAAKYKELGYQG
LQAHAGKVFKLVEAA  IQLVETGVVAS  DATLKNLGSVHVS  KGVVA DAHFPVVK EAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMDDAA
          .   *   .           *   *           .   .   .   .   .   .   .
    
```

Fig. 3. CLUSTAL-produced multiple alignment of seven globin sequences taken from Lesk and Chothia (1980) (see RESULTS, section 4)

# Consensus Sequence From a Multiple Sequence Alignment

## ClustalW Insulin Alignments

	10	20	30
IPGP			F V S R H
IPDK			A A N Q H
IPDG	M A L W M R	L L P L L A L L A L W A P A P T R A	F V N Q H
IPCH	M A L W I R	S L P L L A L L V F S G P G - T S Y	A A N Q H
IPCA	M A V W I Q	A G A L L F L L A V S S V N A N A G	A P - Q H
IPBO			F V N Q H
IPAF	M A A L W L Q	S F S L L V L L V V S W P G S Q A V	A P A Q H
	A . W . .	L L L L	A N Q H

	40	50	60
IPGP	L C G S N L V E T L Y S V C Q D D G F F Y I P K	D X X E L E	
IPDK	L C G S H L V E A L Y L V C G E R G F F Y S P K T	X X D V E	
IPDG	L C G S H L V E A L Y L V C G E R G F F Y T P K A	R R E V E	
IPCH	L C G S H L V E A L Y L V C G E R G F F Y S P K A	R R D V E	
IPCA	L C G S H L V D A L Y L V C G P T G F F Y N P K	R D V D P P	
IPBO	L C G S H L V E A L Y L V C G E R G F F Y T P K A	R R E V E	
IPAF	L C G S H L V D A L Y L V C G D R G F F Y N P K	R D V D Q L	
	L C G S H L V E A L Y L V C G E R G F F Y . P K .		D V E

	70	80	90
IPGP	D P Q V E Q T E L G M G - - - - L G A G G L Q P - - L Q G		
IPDK	Q P - L V N G P L H G E - - - - V G E L P F Q - - - H E		
IPDG	D L Q V R D V E L A G A - - - - P G E G G L Q P L A L E G		
IPCH	Q P - L V S S P L R G E - - - - A G V L P F Q - - - Q E		
IPCA	L G F L P P K S - - - - A Q E T E V A D F A F K D H A E		
IPBO	G P Q V G A L E L A G G - - - - P G A G G L E - - - - G		
IPAF	L G F L P P K S G G A A A A G A D N E V A E F A F K D Q M E		
	P L L G	G F Q	E

	100	110	120
IPGP	A L Q X X - - G I V D Q C C T G T C T R H Q L Q S Y C N		
IPDK	E Y Q X X - - G I V E Q C C E N P C S L Y Q L E N Y C N		
IPDG	A L Q K R - - G I V E Q C C T S I C S L Y Q L E N Y C N		
IPCH	E Y E K V K R G I V E Q C C H N T C S L Y Q L E N Y C N		
IPCA	V I R K R - - G I V E Q C C H K P C S I F E L Q N Y C N		
IPBO	P P Q K R - - G I V E Q C C A S V C S L Y Q L E N Y C N		
IPAF	M M V K R - - G I V E Q C C H R P C N I F D L Q N Y C N		
	. Q K R	G I V E Q C C	C S L Y Q L E N Y C N

# HMM Model of Hemoglobins

<http://decypher.stanford.edu/>

ClustalW Search on DeCypher... Results for Job CGI\_Temp7772...

[Home Page](#)

HMM Model created from: ClustalW Search on DeCypher Protein Sequences

[Download](#) (For Internet Explorer Browsers, Right Click on Hyperlink and Select "Save Target As...")

[Search with this Model](#)

```

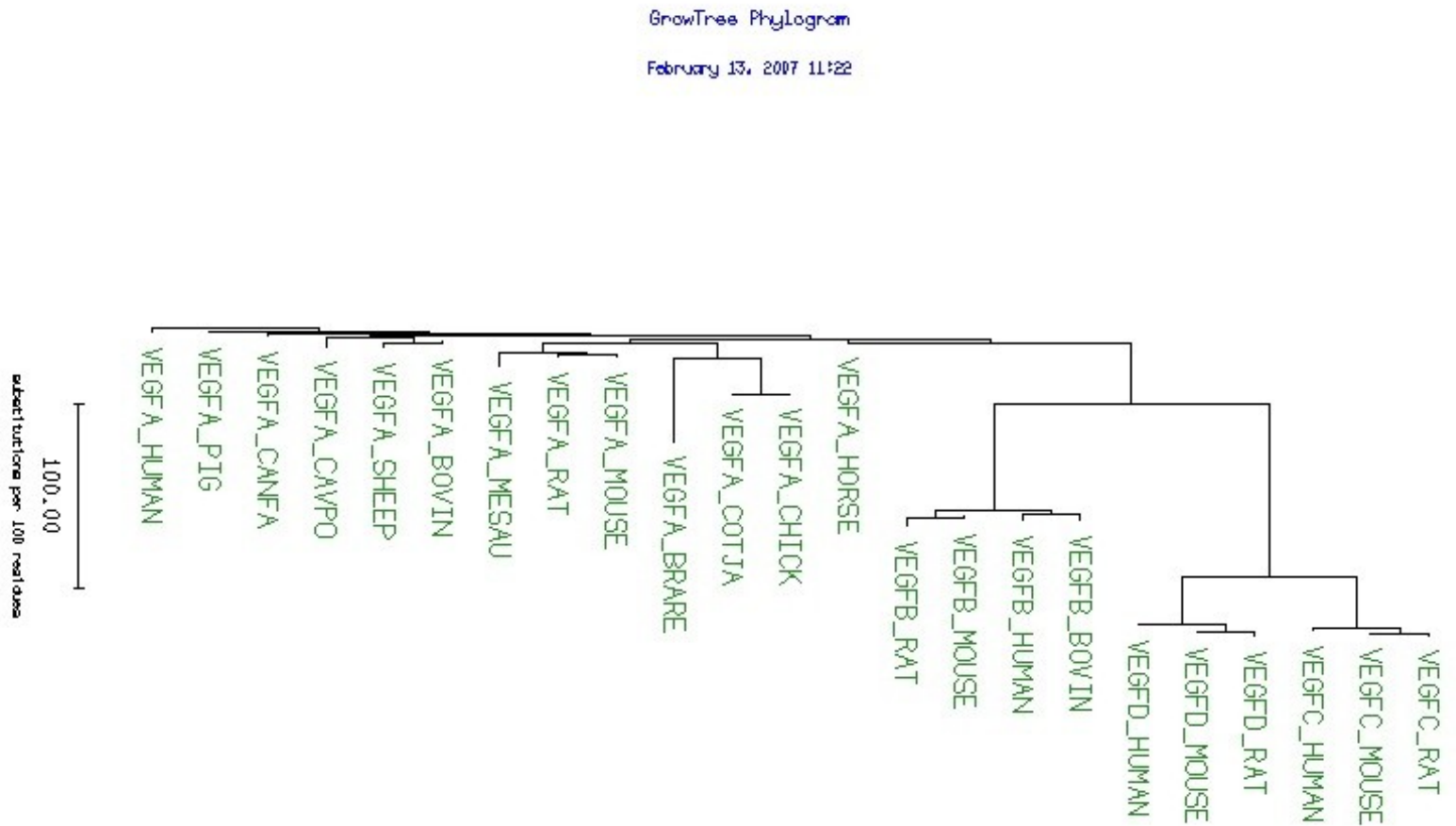
HMMER2.0
NAME d:\decypher\output\CGI_Temp77466aad91_H
DESC
LENG 161
ALPH Amino
RF no
CS no
MAP yes
COM D:\decypher\bin\hmmbuild.exe d:\decypher\output\CGI_Temp7772fbe351.out.tmp d:\decypher\output\CGI_Temp77466aad91_H.seq
NSEQ 7
DATE Thu Feb 15 10:29:18 2007
CKSUM 1944
XT -8455 -4 -1000 -1000 -8455 -4 -8455 -4
NULT -4 -8455
NULE 595 -1558 85 338 -294 453 -1158 197 249 902 -1085 -142 -21 -313 45 531 201
HMM A C D E F G H I K L M N P Q R S T
m->m m->i m->d i->m i->i d->m d->d b->m m->e
-606 * -1543
1 -820 -1123 -1229 -1315 -2065 -1220 -1331 -2014 -1345 -2133 -1685 -1232 3693 -1331 -1440 -1011 -1095
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -34 -5976 -7018 -894 -1115 -701 -1378 -606 *
2 -958 -760 -2548 -2272 -535 -2401 -1765 2979 -1942 375 395 -2092 -2523 -1850 -1979 -1799 -993
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -34 -5976 -7018 -894 -1115 -701 -1378 *
3 -641 -619 -2245 -2036 -813 -1897 -1637 1127 -1762 -54 76 -1764 -2218 -1697 -1826 -1292 -764
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -34 -5976 -7018 -894 -1115 -701 -1378 *
4 -1041 -1978 3376 310 -2404 -1091 -582 -2397 -748 -2472 -1917 -27 -1587 -337 -1332 -886 -1147
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -34 -5976 -7018 -894 -1115 -701 -1378 *
5 1 -504 -1094 -1024 -1493 -825 -961 -774 -860 -1268 -782 -771 -1401 -850 -1002 -231 2989
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -34 -5976 -7018 -894 -1115 -701 -1378 *
6 -654 -1010 -1195 -1358 -2342 3274 -1442 -2362 -1583 -2515 -1948 -1206 -1681 -1458 -1665 -851 -984
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -34 -5976 -7018 -894 -1115 -701 -1378 *
7 163 -452 -764 -783 -1673 -626 -870 -1534 -827 -1810 -1134 -553 -1252 -739 -1014 2714 -167
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -34 -5976 -7018 -894 -1115 -701 -1378 *
8 -641 -619 -2245 -2036 -813 -1897 -1637 1127 -1762 -54 76 -1764 -2218 -1697 -1826 -1292 -764
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -34 -5976 -7018 -894 -1115 -701 -1378 *
9 1686 -842 -2427 -2196 -2039 1808 -1802 -1205 -2040 -1858 -1181 -1585 -1986 -1795 -2118 -663 -682
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -15 -7120 -8162 -894 -1115 -701 -1378 *
10 963 -1368 -1105 -579 -1489 -1865 1547 -1083 -439 -1321 -579 -769 1032 1030 -821 -852 -691
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96 359 117
- -10 -7714 -8756 -894 -1115 -701 -1378 *

```

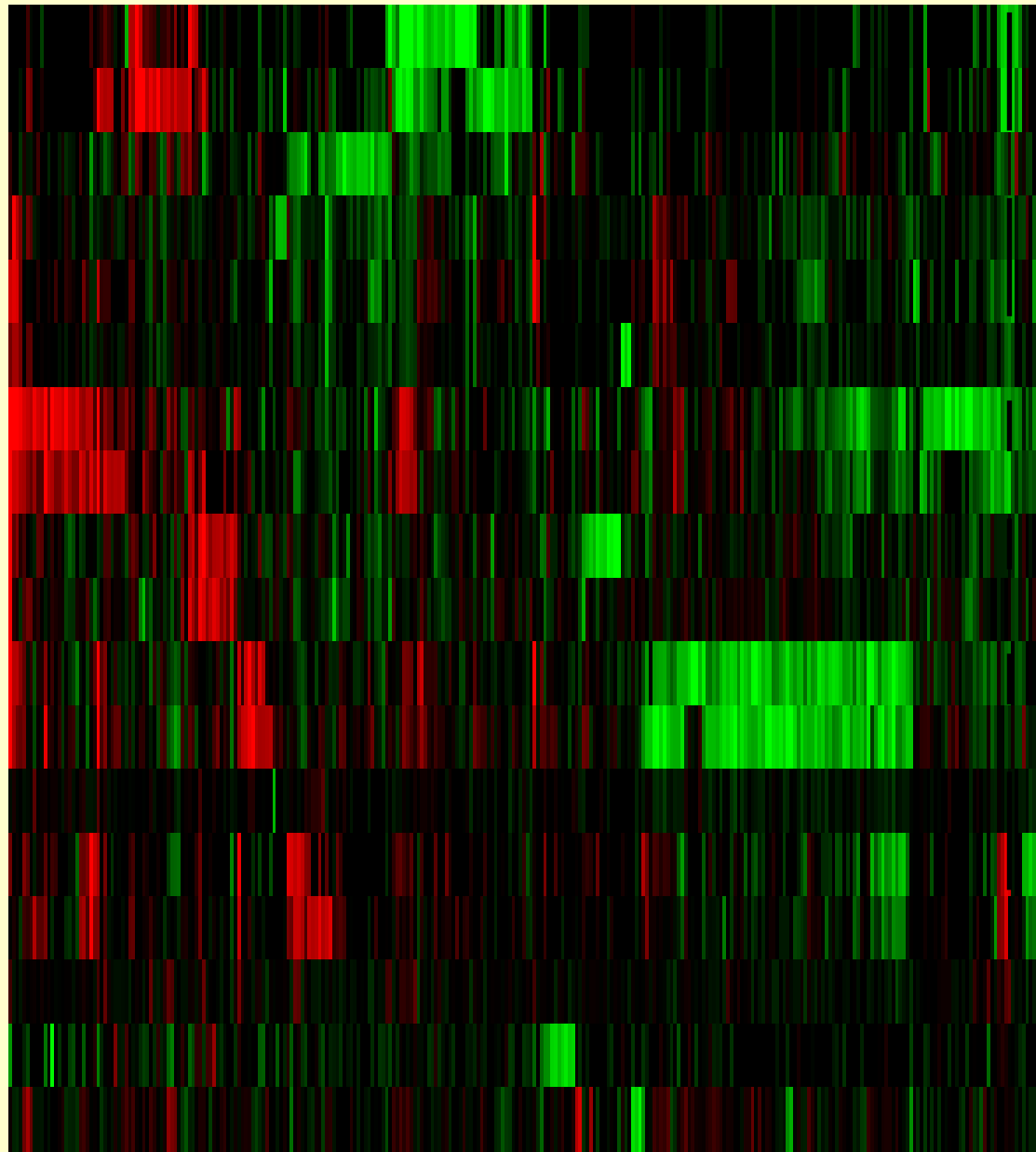




# GrowTree VegF Neighbor Joining Tree



# Human Gene Expression Signatures



T Cells Signaling

DNA Damage

Fibroblast Stimulation

B Cells Signaling

CMV Infection

Anoxia

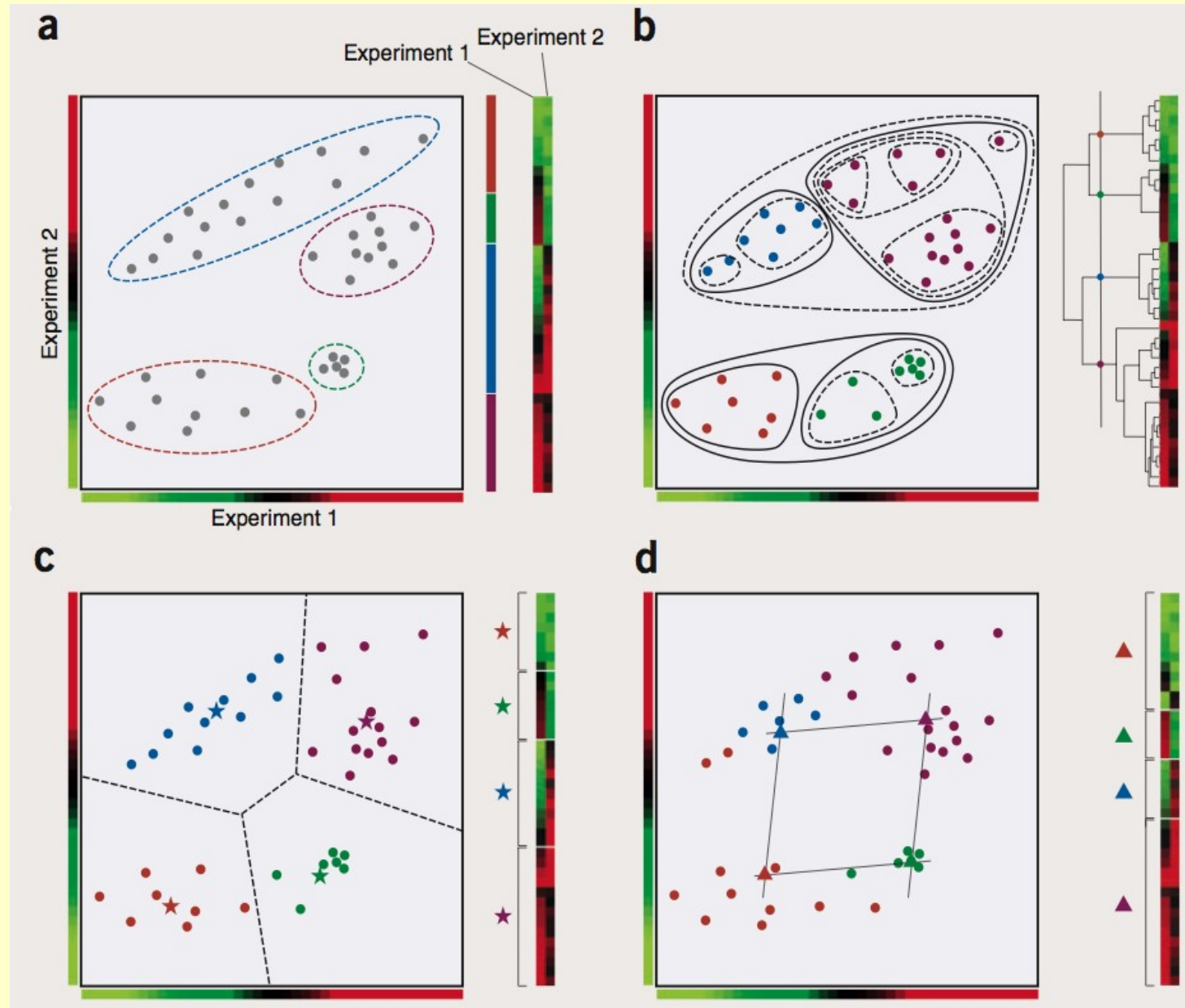
Polio Infection

Monocytes Signaling IL4

Hormone



# Clustering Gene Expression Profiles: Comparison of Methods





# The Fraenkel Lab TAMO -- Tools for the Analysis of Motifs

Download and Support for the TAMO package

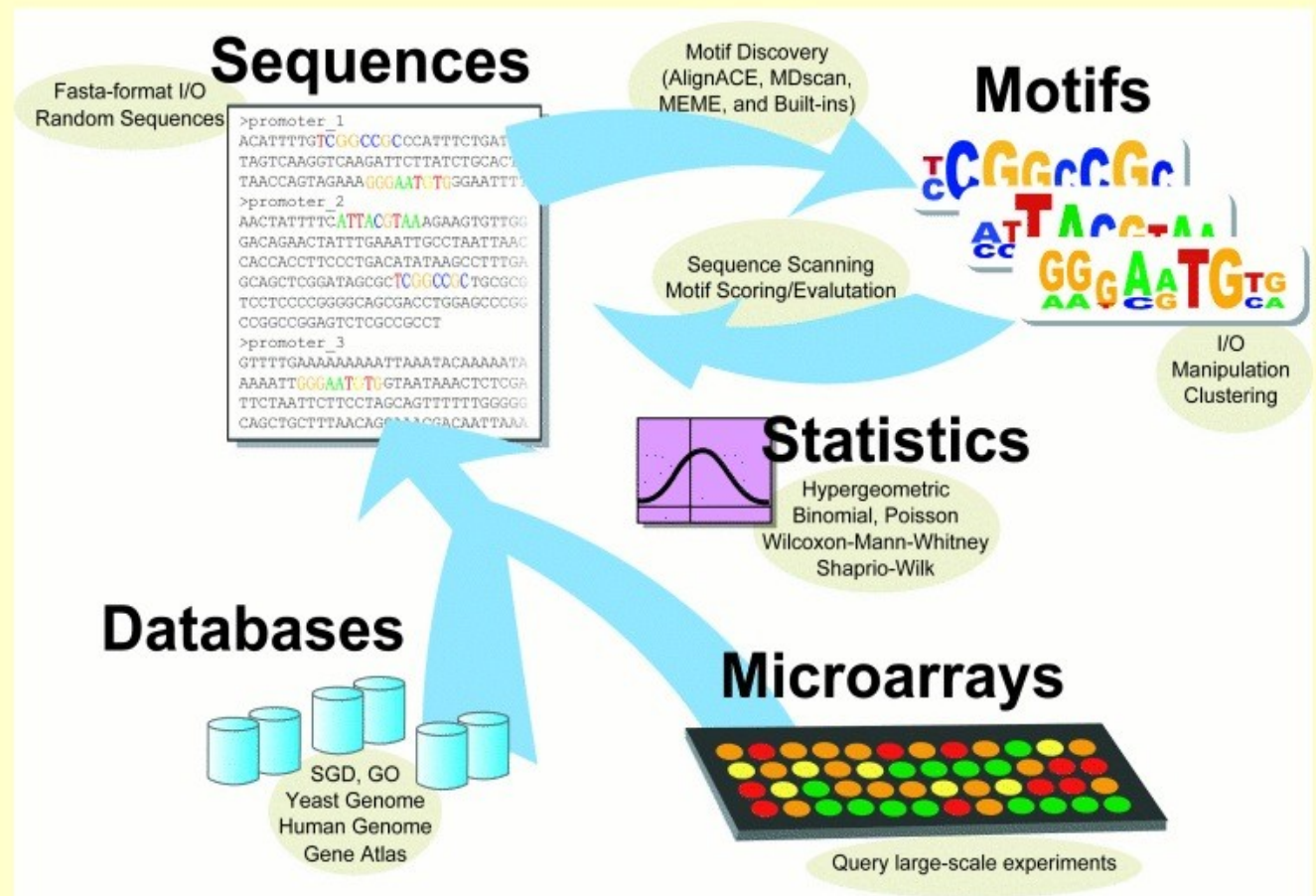
- Home
- People
- Publications
- Data / Download**

The TAMO package can be downloaded from this page. Support information, and possibly new modules will be distributed from this location.

- Click [here](#) for the package overview.
- Click [here](#) for short descriptions of each file.
- Click [here](#) for an introductory tutorial.
- [Download](#) the package.
- Browse the automatically generated [documentation](#) (via pydoc).
  - For each command-line program that can be executed from the unix shell, (e.g. [Sitemap.py](#), [AlignAce.py](#), [MotifMetrics.py](#), [UPGMA.py](#), etc...) documentation is obtained by executing the program without any arguments.
  - If you're looking for a place to start, the core data structure is the [Motif](#) object. This file also includes tools for constructing motifs from different data sources.
- Installation instructions are [here](#).

### License:

- The TAMO package is free for academic use. Please contact [Ernest Fraenkel](#) for commercial licensing.



# Finding Transcription Factor Binding Sites

Upstream Regions  
expressed

Co-

Genes

GATGGCTGCACCACGTGTATGC...ACG

Pho 5

CACATCGCATCACGTGACCAGT...GAC

Pho 8

GCCTCGCACGTGGTGGTACAGT...AAC

Pho 81

TCTCGTTAGGACCATCACGTGA...ACA

Pho 84

CGCTAGCCCACGTGGATCTTGA...AGA

Pho ...

ATGACTGGC

# Finding Transcription Factor Binding Sites

---

Upstream Regions

Co-expressed  
Genes

GATGGCTGCAC**CACGTG**TATGC . . . ACGATGTCTCGC  
CACATCGCAT**CACGTG**ACCAGT . . . GACATGGACGGC  
GCCTCG**CACGTG**GTGGTACAGT . . . AACATGACTAAA  
TCTCGTTAGGACCAT**CACGTG**A . . . ACAATGAGAGCG  
CGCTAGCC**CACGTG**GATCTTGT . . . AGAATGGCCTAT

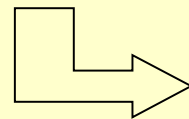
# Finding Transcription Factor Binding Sites

---

Upstream Regions

Co-expressed  
Genes

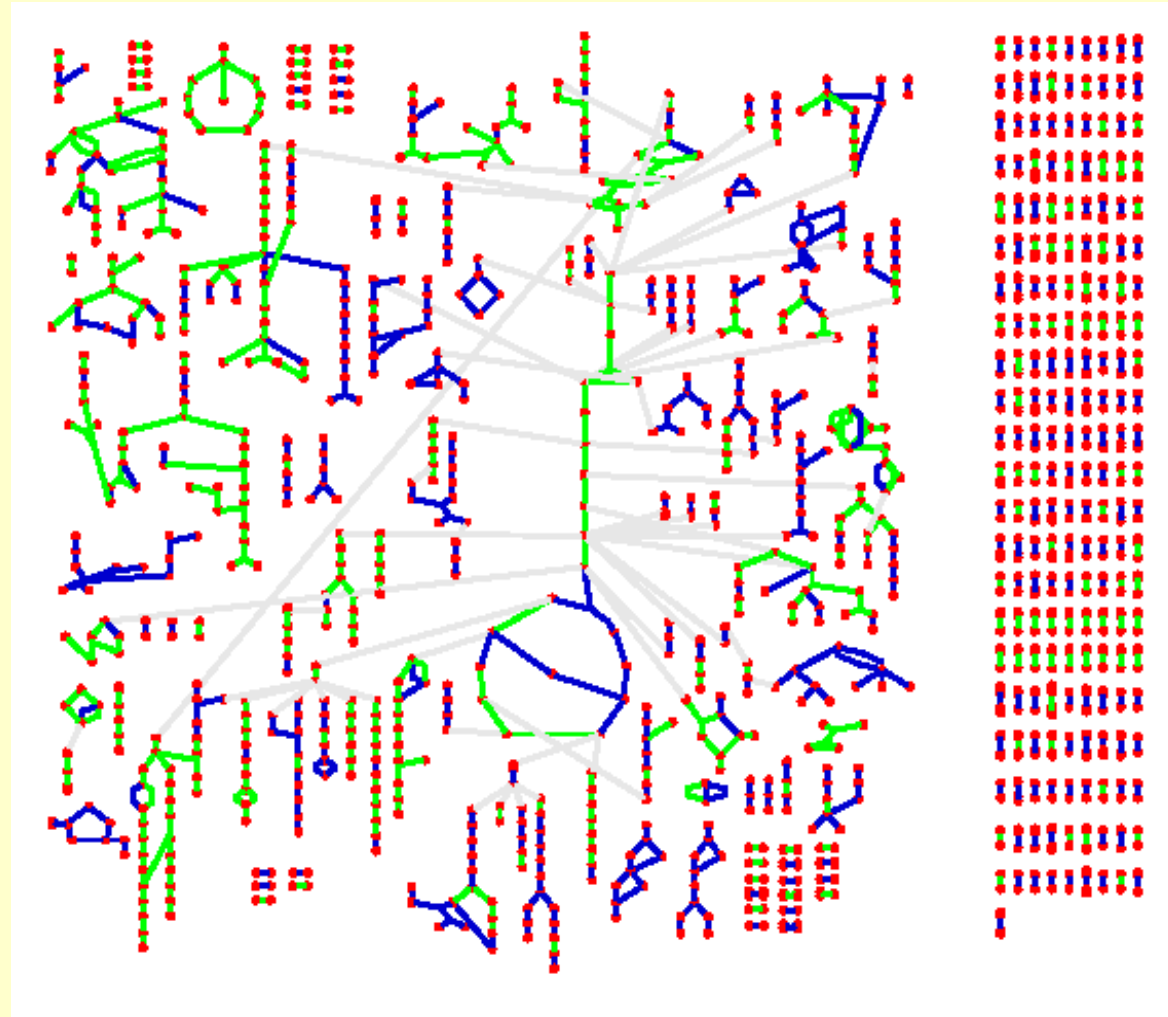
```
ATGGCTGCAC CACGTT TATGC . . . ACGATGTCTCGC
CACATCGCAT CACGTG ACCAGT . . . GACATGGACGGC
GCCTCG CACGTG GTGGTACAGT . . . AACATGACTA
TTAGGACCAT CACGTG A . . . ACAATGAGAGCG
CGCTAGCC CACGTT GATCTTGT . . . AGAATGGCCTA
```



Pho4 binding

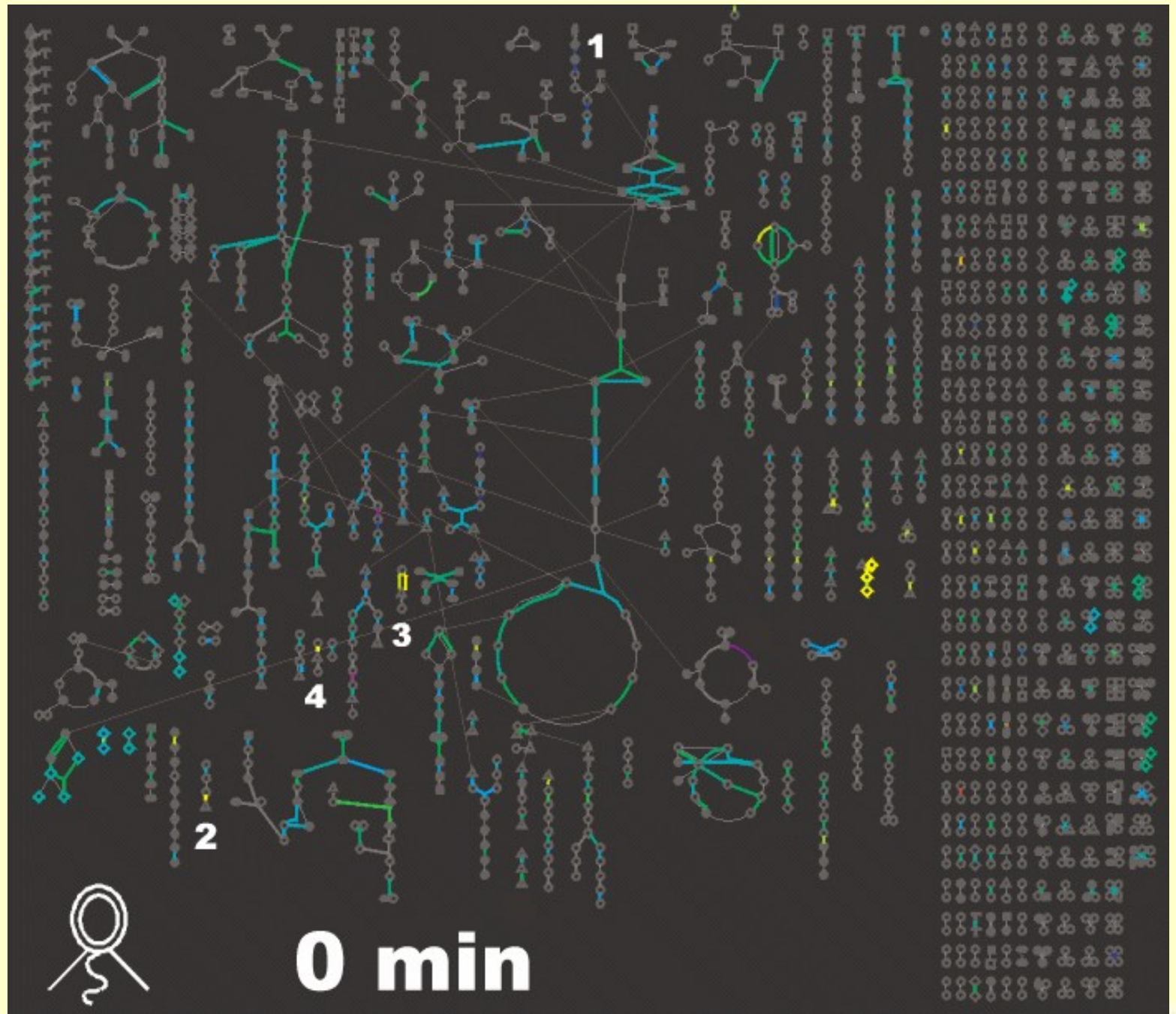
# Metabolic Networks: BioCyc

<http://biocyc.org/>



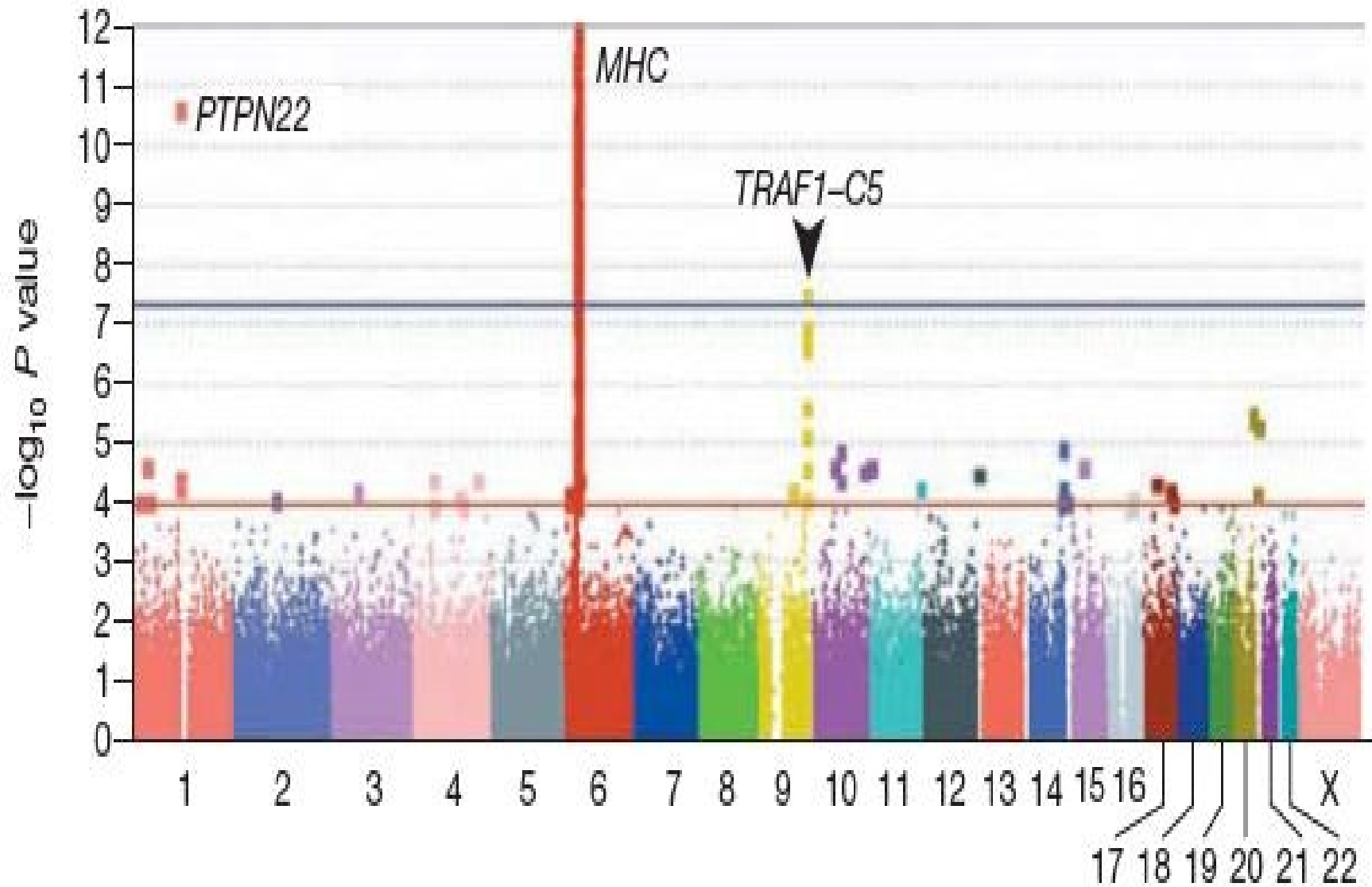


# *C. crescentus* Cell Cycle Gene Expression



# Genome Wide Associations in

**Figure 3.** Genome-wide Association Findings in Rheumatoid Arthritis



# Leveraging Genomic Information in Medicine

## Novel Diagnostics

Microchips & Microarrays

Gene Expression - RNA

Proteomics - Protein

## Novel Therapeutics

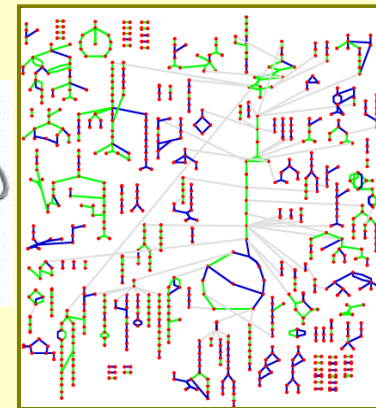
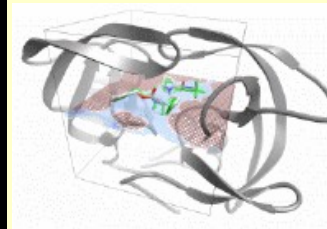
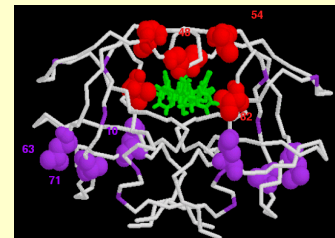
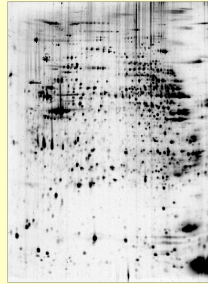
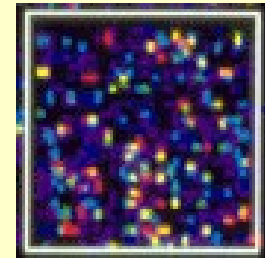
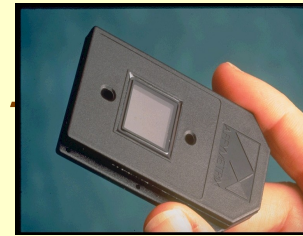
Drug Target Discovery

Rational Drug Design

Molecular Docking

Gene Therapy

Stem Cell Therapy



## Understanding Metabolism

## Understanding Disease

Inherited Diseases - OMIM

Infectious Diseases

Pathogenic Bacteria

Viruses

