

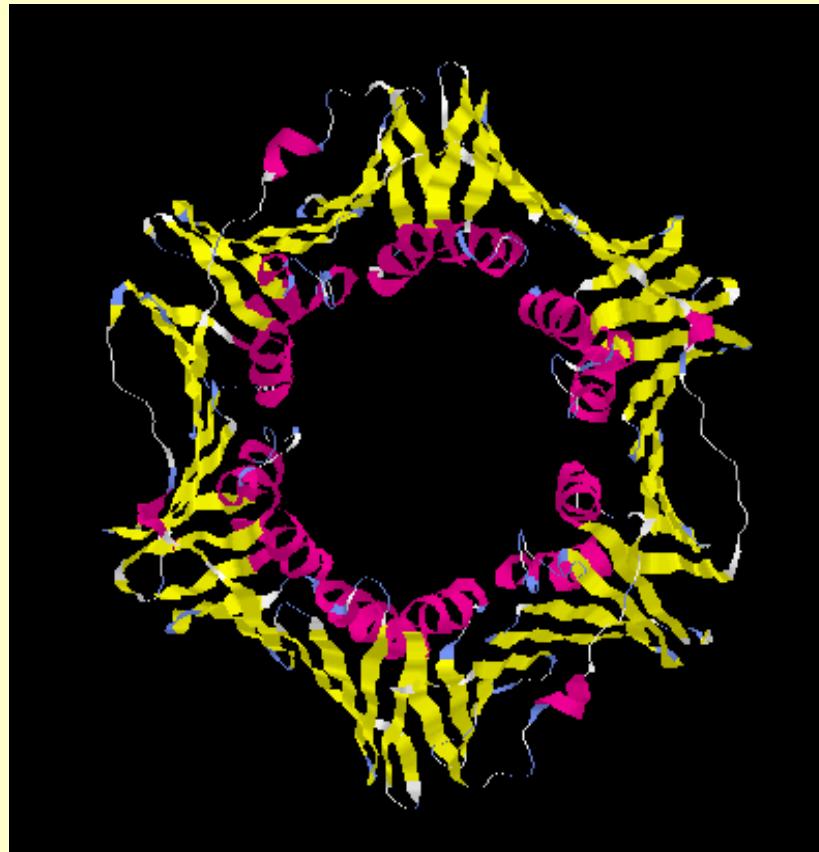
# Computational Molecular Biology

## Biochem 218 – BioMedical Informatics 231

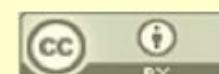
<http://biochem218.stanford.edu/>

---

## Genomics and Bioinformatics



Doug Brutlag  
Professor Emeritus  
Biochemistry & Medicine (by courtesy)



Doug Brutlag 2010

# Faculty, TAs and Staff

---

Doug Brutlag



Lee Kozar



Maeve O'Huallachain



Dan Davison



# Course and Video Availability

---

- Alway 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

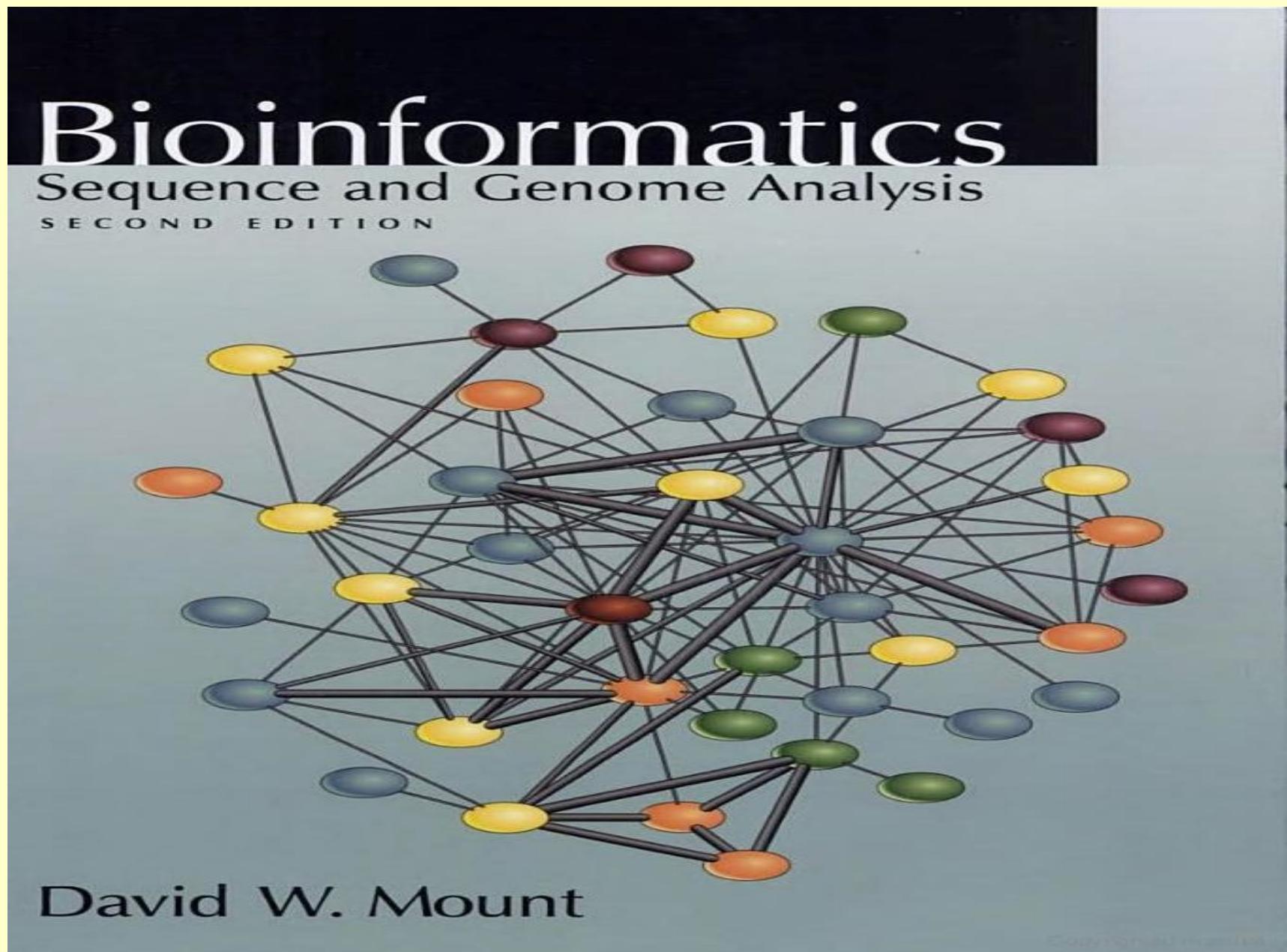
---

- **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>



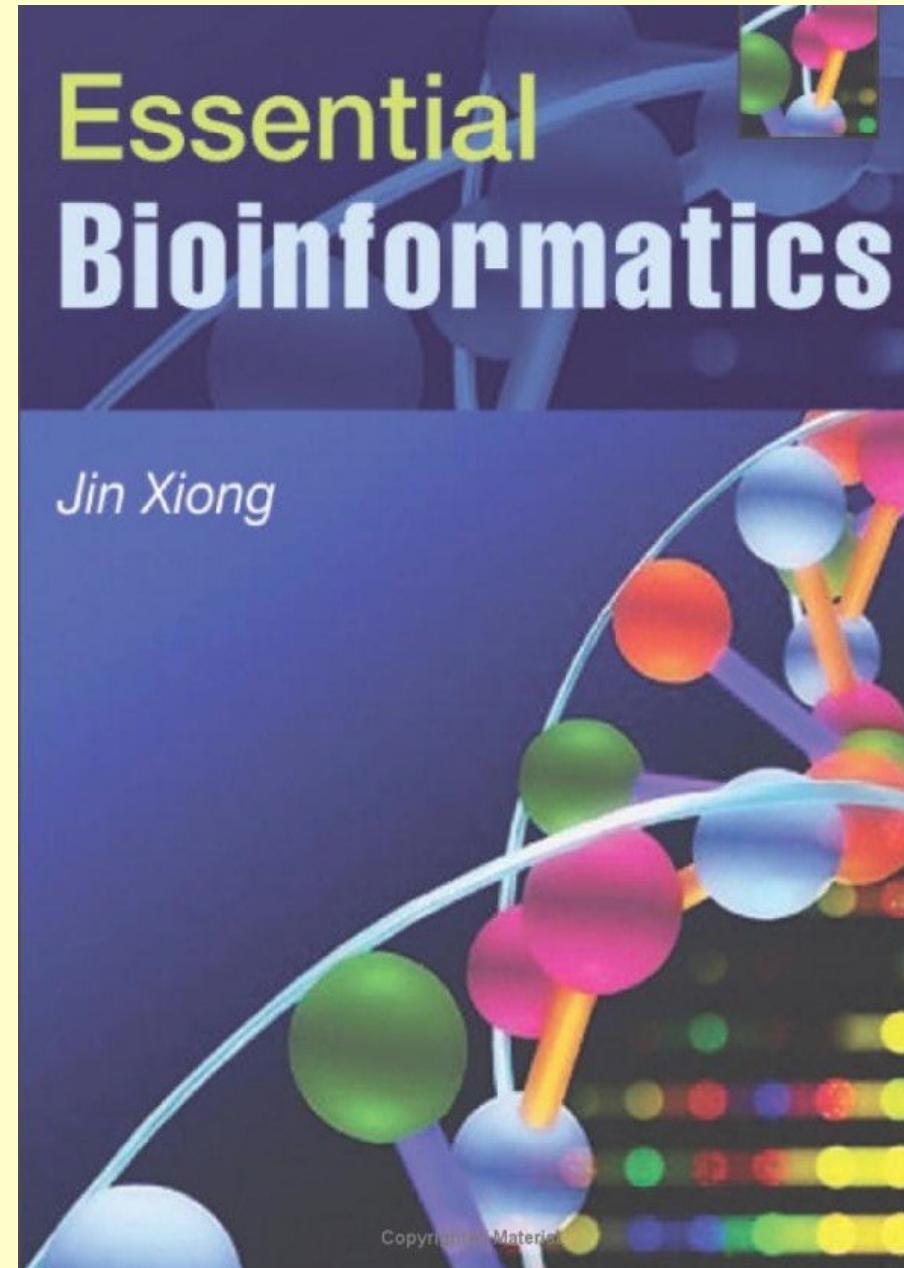
David Mount

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



# Jin Xiong

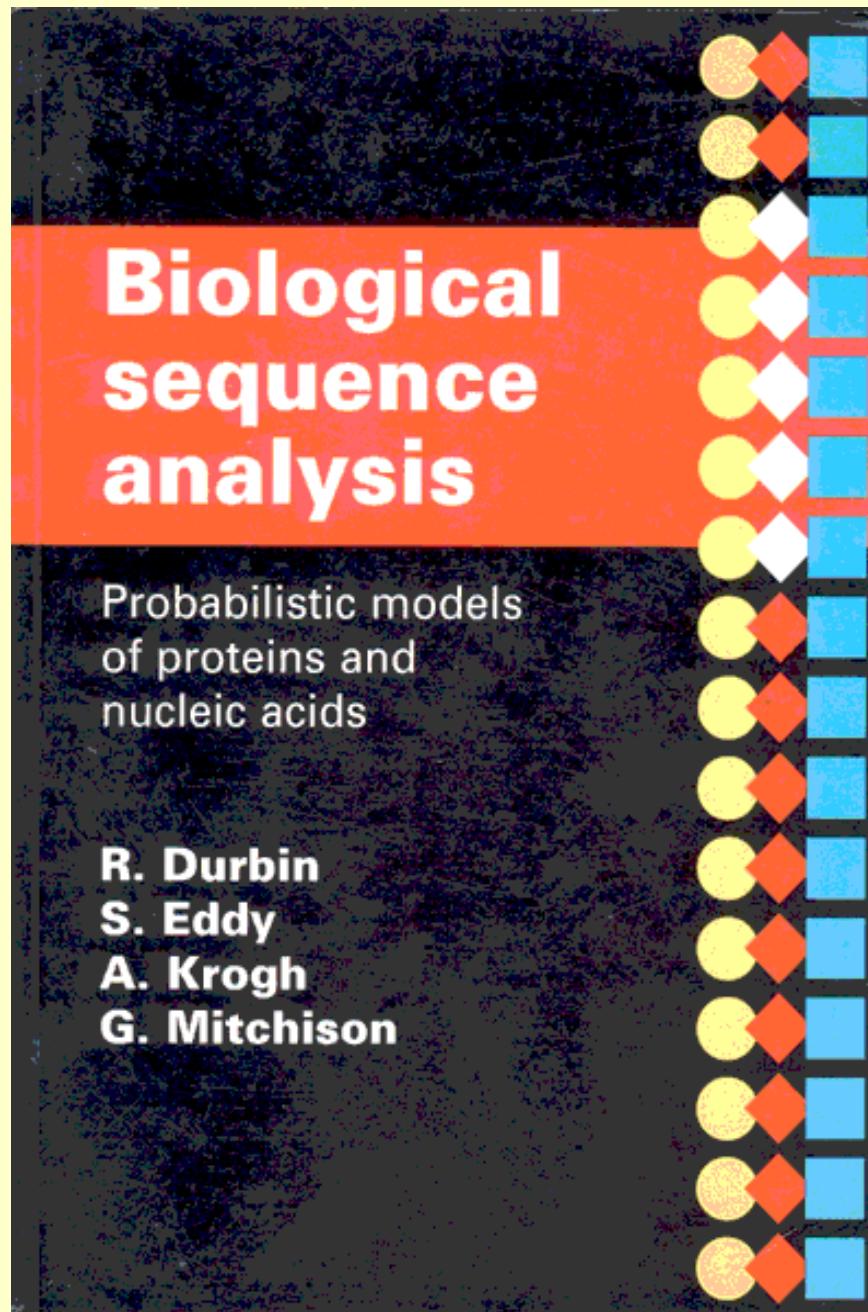
## Essential Bioinformatics



Doug Brutlag 2010

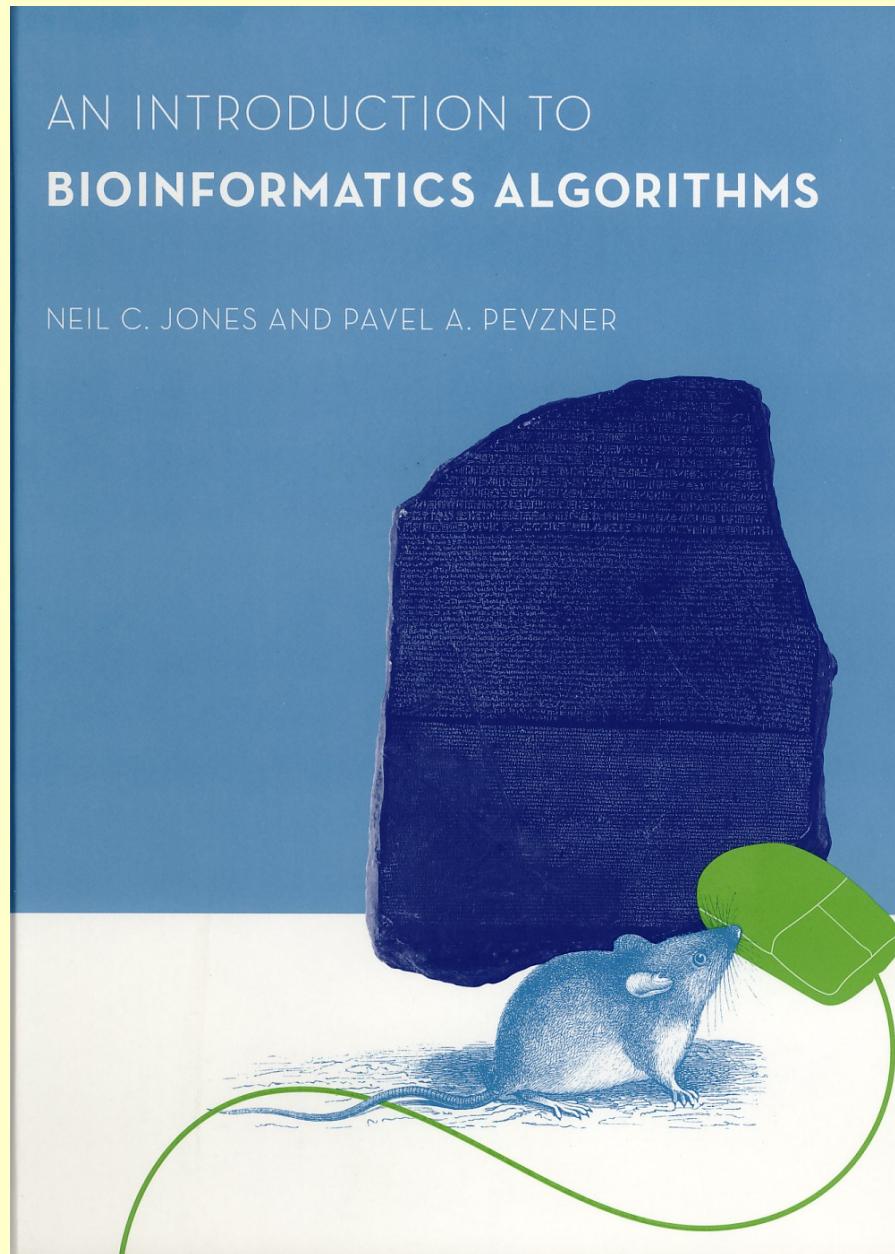
# Richard Durbin *et al.*

## Biological Sequence Analysis



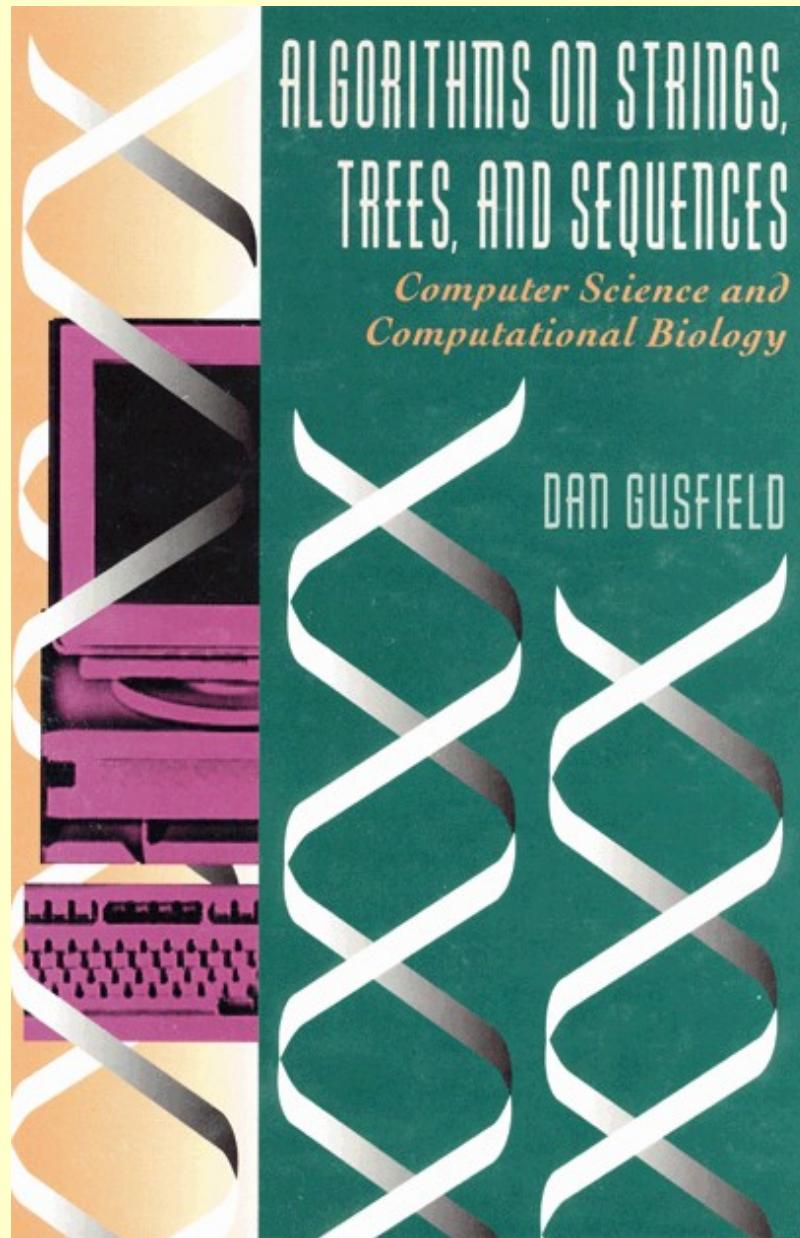
# Jones & Pevzner

## Bioinformatics Algorithms



# Dan Gusfield

## Algorithms on Strings, Trees & Sequences



# Baldi & Brunak

## Bioinformatics: The Machine Learning Approach



# Higgins & Taylor

## Bioinformatics: Sequence, Structure & Databanks



### Bioinformatics

*Sequence, structure  
and databases*

*Edited by*

Des Higgins  
Willie Taylor

PRACTICAL  
APPROACH

Copyrighted Material



Doug Brutlag 2010

# NCBI Handbook

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

The screenshot shows the homepage of the NCBI Handbook. The header features the NCBI logo (a stylized white 'S' on a blue background) and the text "The NCBI Handbook" and "The National Library of Medicine". Below the header is a search bar with the placeholder "Search for" and dropdown options "Within" and "This book", "All books", and "PubMed". The breadcrumb navigation shows the path "NCBI » Bookshelf » The NCBI Handbook ». The main content area includes a section titled "Book Information" with four expandable sections: "Part 1 The Databases", "Part 2 Data Flow and Processing", "Part 3 Querying and Linking the Data", and "Part 4 User Support". Below this is a link to a "Glossary". At the bottom of the page are links for "2002-2009 Copyright and disclaimer", "Bookshelf | NCBI | NLM | NIH", and "Help | Contact Bookshelf". In the bottom right corner are buttons for "Expand All" and "Collapse All". The footer contains a Creative Commons Attribution (CC BY) license logo and the text "Doug Brutlag 2010".

# NCBI Handbook

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



The screenshot shows the homepage of "The NCBI Handbook". The header features the NCBI logo, a blue banner with icons for search, databases, and other services, and the text "The NCBI Handbook" and "The National Library of Medicine". Below the banner is a search bar with fields for "Search for", "Within", and "This book", along with a radio button for "This book". The breadcrumb navigation shows the path: 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/>



EMBL-EBI  All Databases   Reset Advanced Search

Databases Tools EBI Groups Training Industry About Us Help Site Index  

**Data Resources & Tools**

- **EMBL-BANK**
- **UniProt**
- **ArrayExpress**
- **Ensembl**
- **InterPro**
- **PDBe**
- Genomes
- Nucleotide Sequences
- Protein Sequences
- Macromolecular Structures
- Small Molecules
- Gene Expression
- Molecular Interactions
- Reactions & Pathways
- Protein Families
- Enzymes
- Literature
- Taxonomy
- Ontologies
- Patent Resources
- Sequence Similarity & Analysis
- Pattern & Motif Searches
- Structure Analysis
- Text Mining
- Downloads
- Web Services



**European Bioinformatics Institute**

**About the EBI**

- **Research**
- **PhD Studies**
- **Training**
- **Industry Support**
- **Group & Team Leaders**
- **EBI Funders**
- User Support
- EBI Mission
- People
- Events at the EBI
- Genome Campus Events
- How to Find us

**EBI Hosted Project Websites**

- 1000 Genomes
- BioCatalogue
- BioSapiens
- E-MeP
- EGA
- ELIXIR
- EMBRACE
- EMERALD
- ENFIN
- FELICS
- IMPACT
- INSDC
- LRG
- SPINE
- SYMBIOmatics

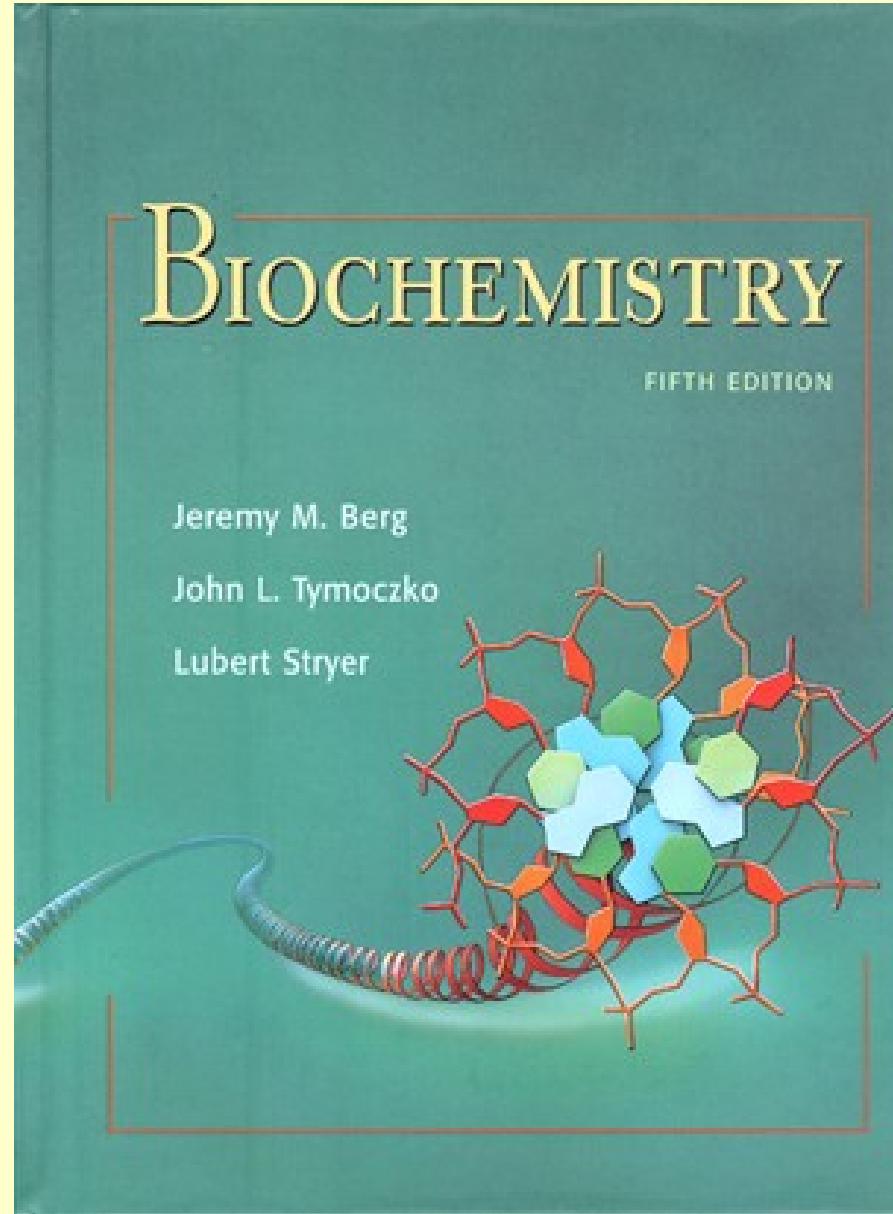
**Latest News** 

- **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](#)
- **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](#)

 Doug Brutlag 2010

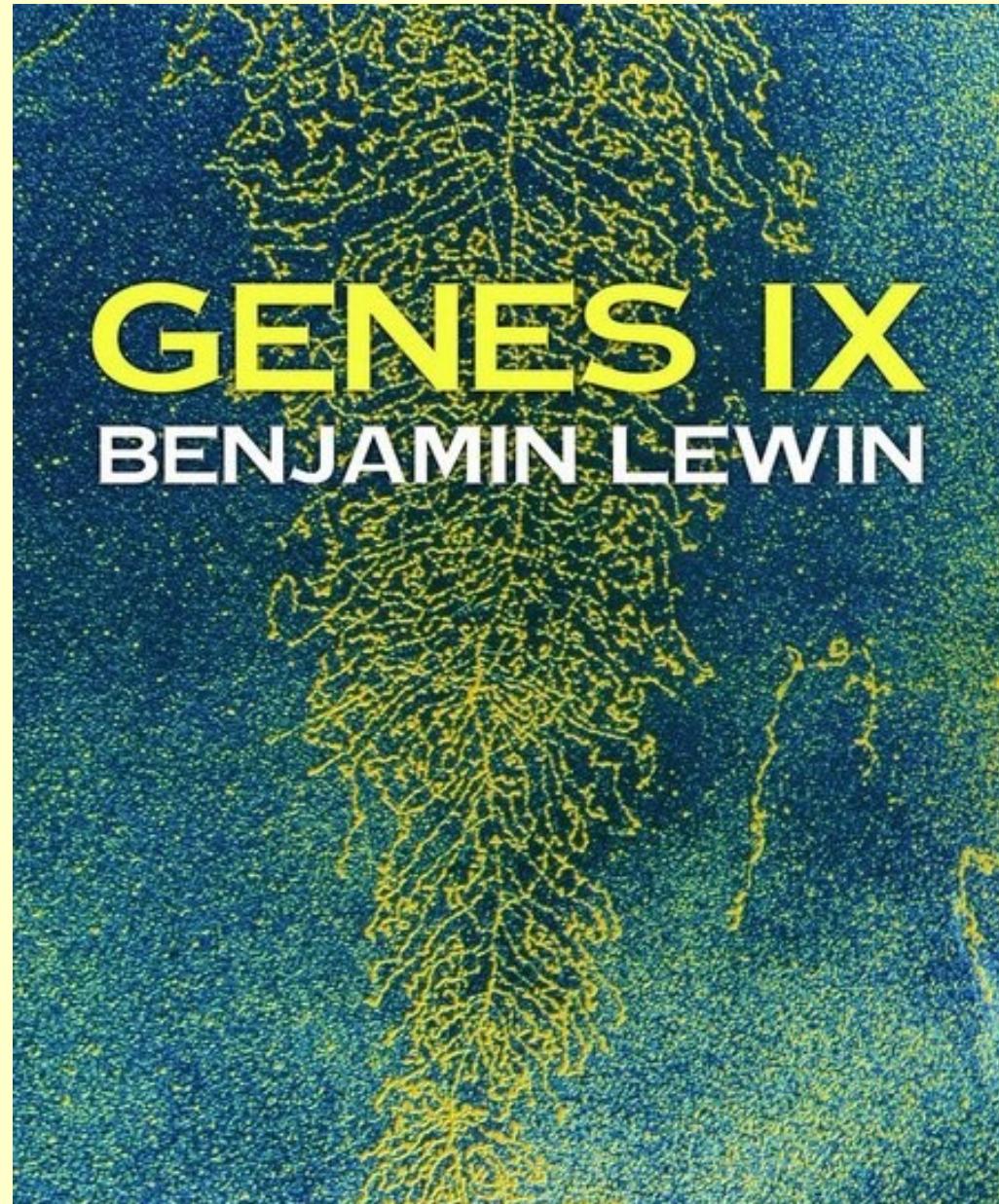
# Berg, Tymoczko & Stryer

## Biochemistry, Fifth Edition



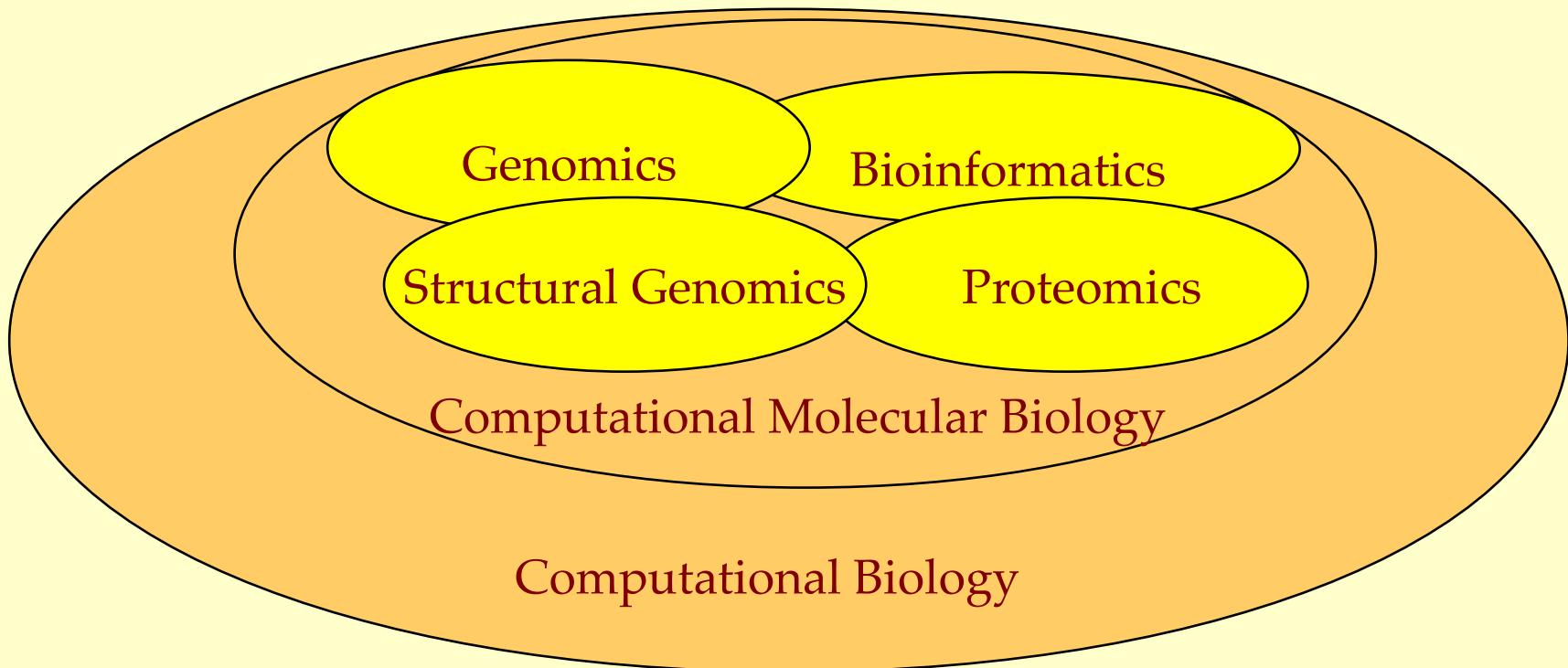
# Benjamin Lewin

## Genes IX



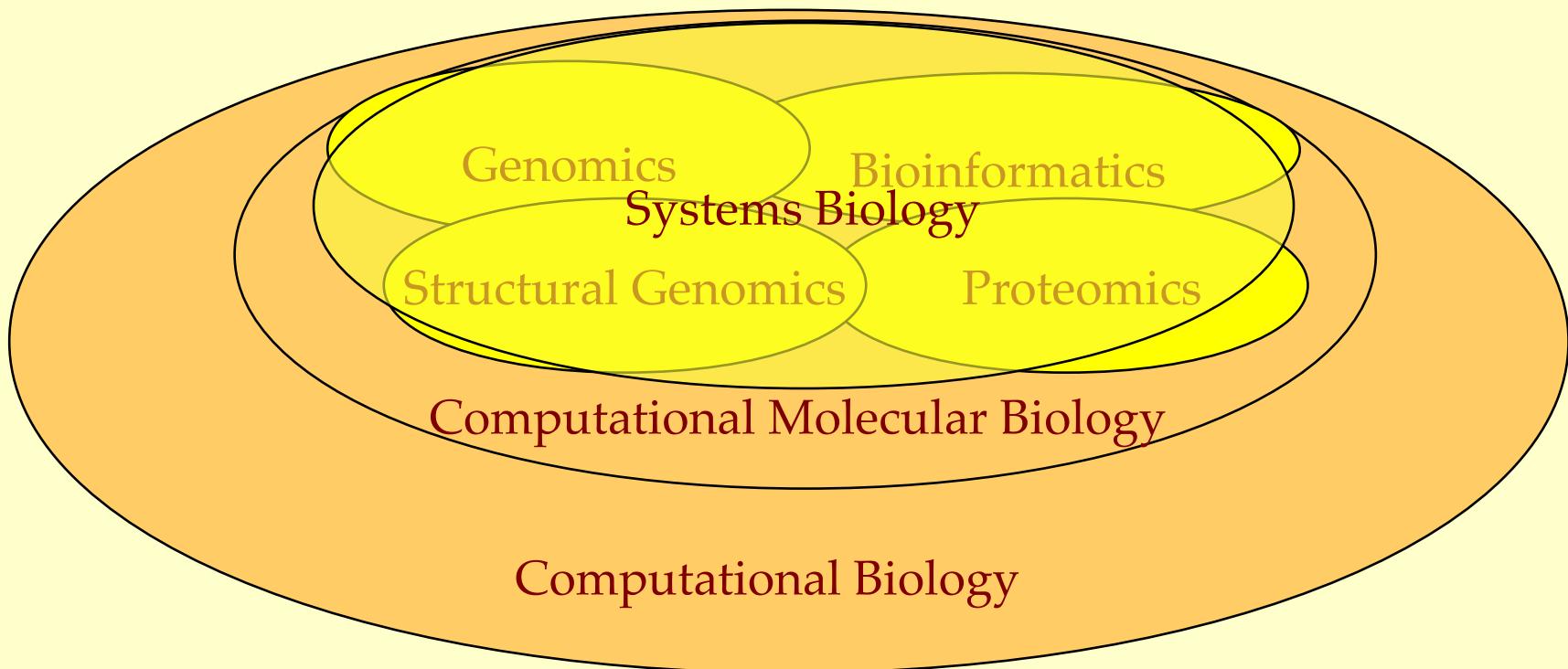
# Genomics, Bioinformatics & Computational Biology

---

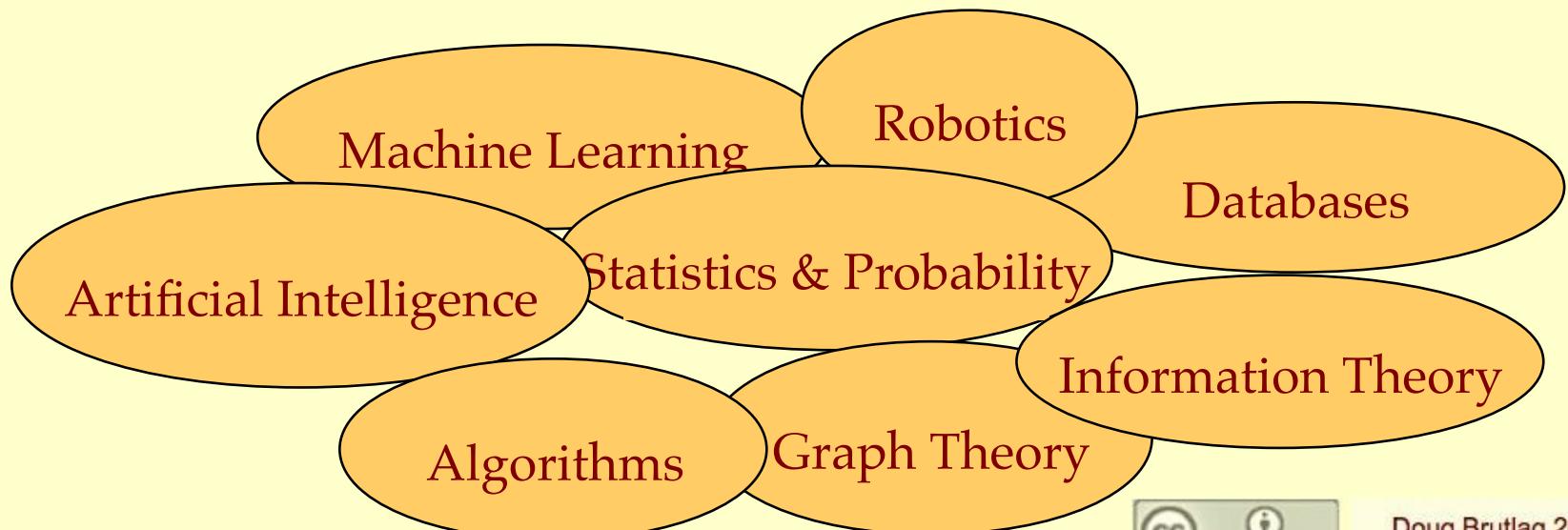
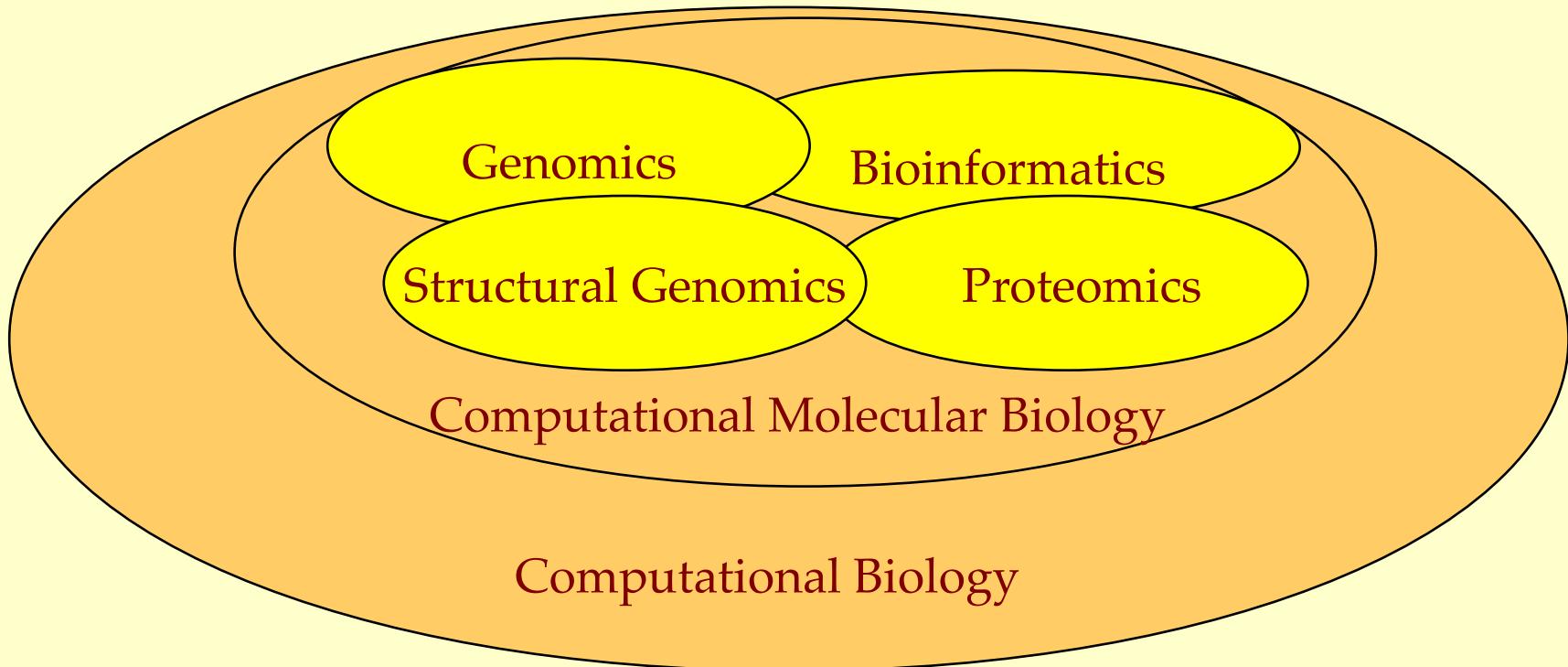


# Genomics, Bioinformatics & Computational Biology

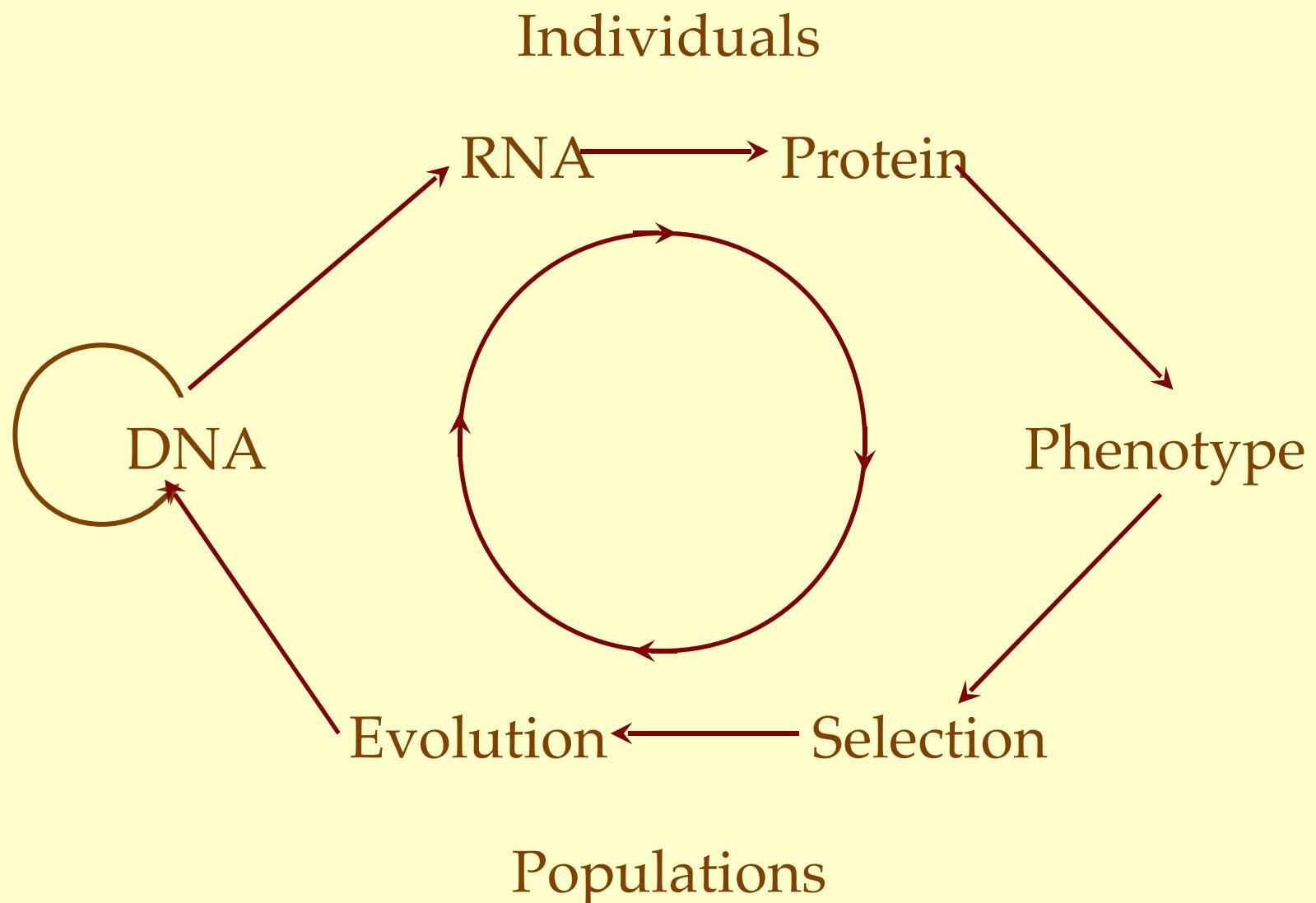
---



# Genomics, Bioinformatics & Computational Biology



# What is Bioinformatics?





# Computational Goals of Bioinformatics

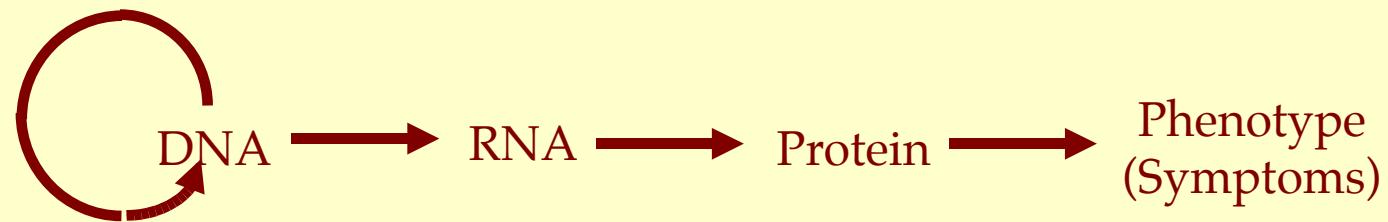
---

- 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.



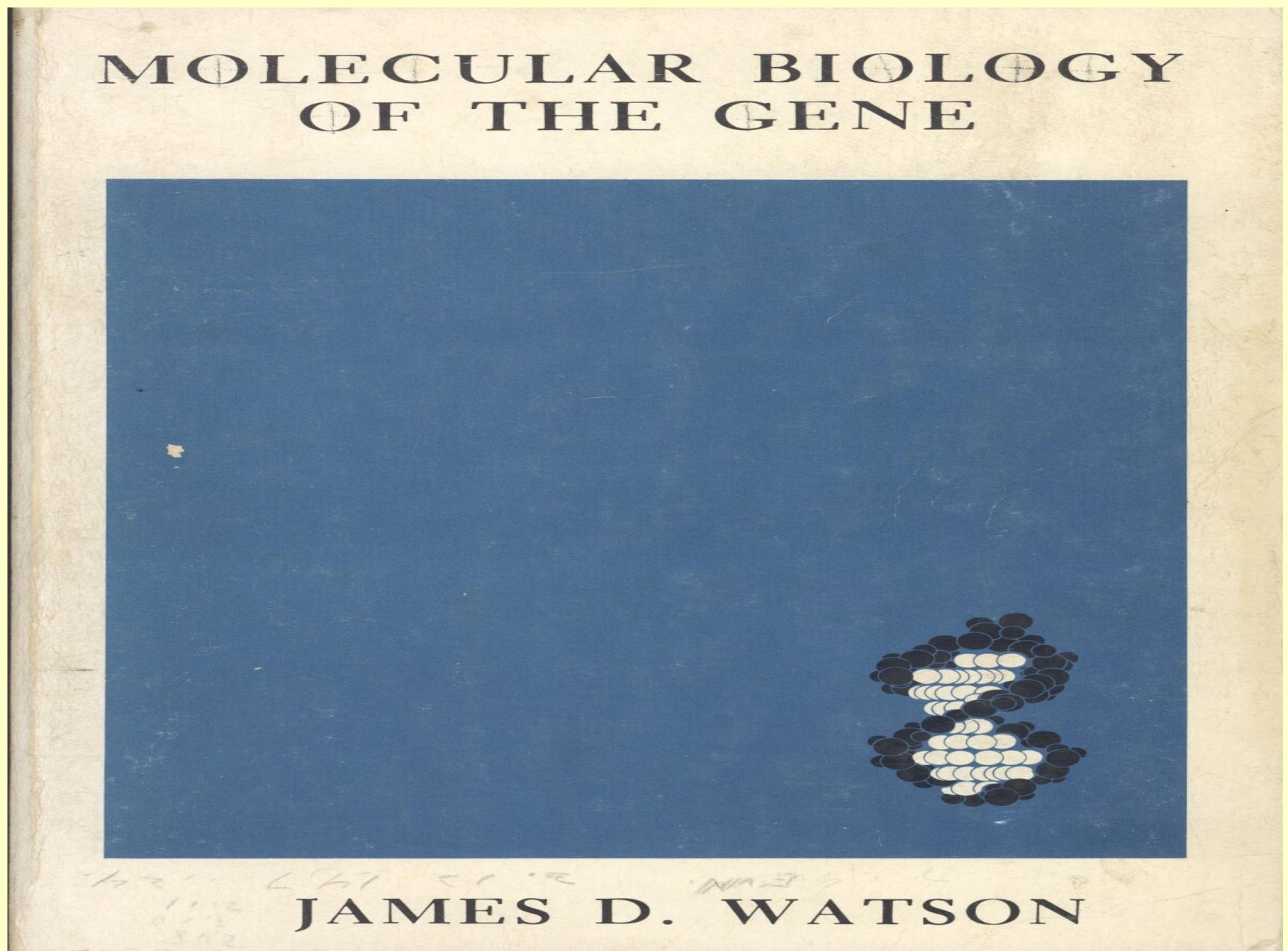
# Central Paradigm of Molecular Biology

---



# Molecular Biology of the Gene 1965

---

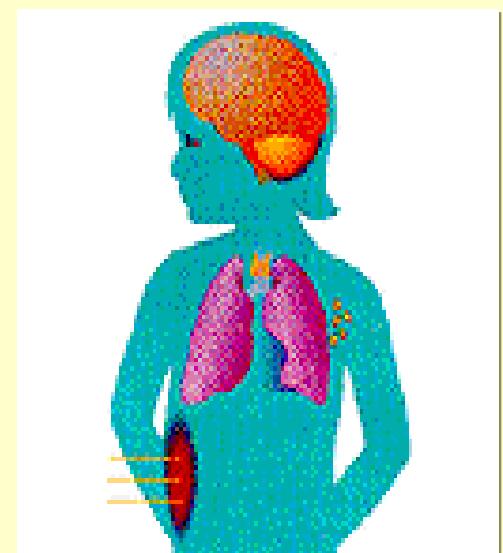
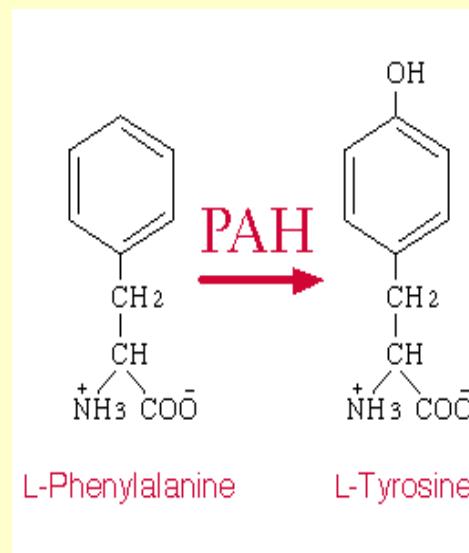
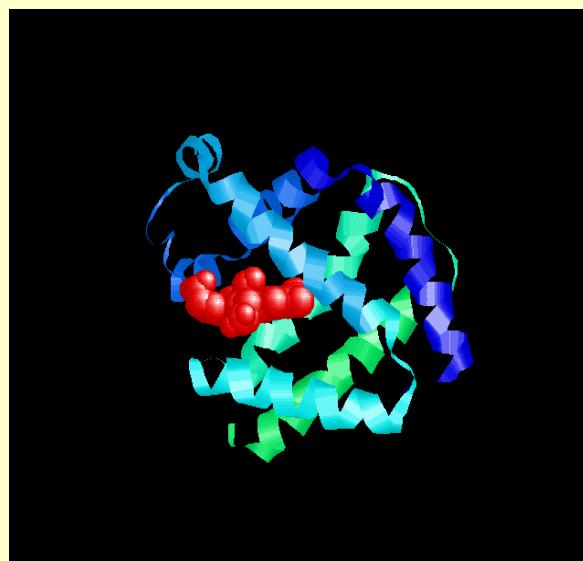


# Central Paradigm of Bioinformatics

Genetic Information

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

→ Molecular Structure → Biochemical Function → Phenotype (Symptoms)

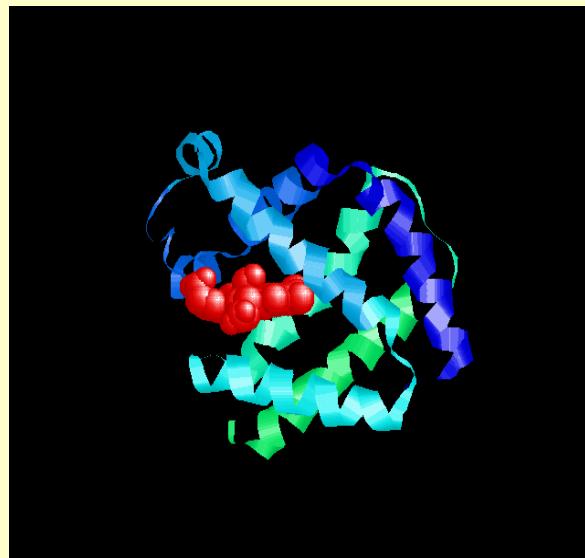


# Central Paradigm of Bioinformatics

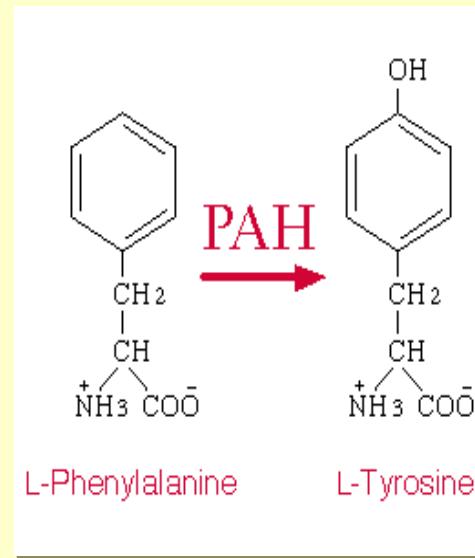
Genetic Information

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

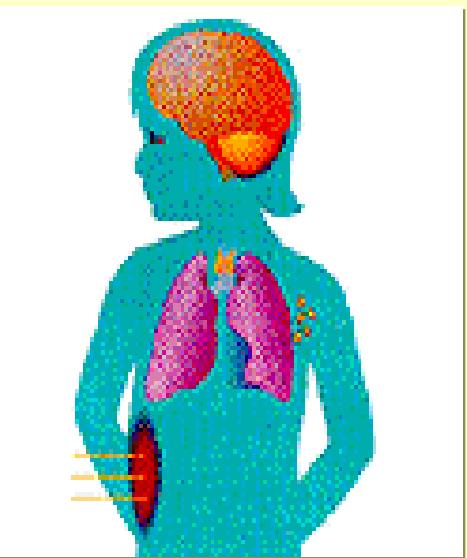
Molecular Structure



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



The image shows a screenshot of the NCBI MeSH website. The header features the NCBI logo and the text "A service of the National Library of Medicine and the National Institutes of Health". Below the header is a navigation bar with links to All Databases, PubMed, Nucleotide, Protein, Genome, Structure, OMIM, PMC, Journals, and Books. A search bar contains the text "Search MeSH for DNA sequencing". Below the search bar are buttons for Limits, Preview/Index, History, Clipboard, and Details. To the left of the main content area is a sidebar with links to About Entrez, Text Version, Entrez PubMed, Overview, Help | FAQ, Tutorials, New/Noteworthy (with a RSS icon), and E-Utilities. The main content area contains a paragraph about MeSH and a bulleted list of tutorials.

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

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

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

New/Noteworthy

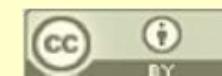
PubMed Services Journals Database MeSH Database Single Citation Matcher Batch Citation Matcher Clinical Queries Special Queries LinkOut My NCBI

**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 screenshot shows the Gene Ontology website homepage. At the top, there is a navigation bar with a search field and a menu for "gene or protein name". Below the header, a diagram illustrates the life cycle of a cell, showing various processes like DNA replication, RNA processing, and protein synthesis. The main content area features a large "Welcome to the Gene Ontology website!" heading. A detailed paragraph explains the project's goal of standardizing gene and gene product attributes across species and databases, mentioning a controlled vocabulary of terms, gene product annotation data from Consortium members, and tools for access and processing. Another section encourages community input and submission of GO annotations. The footer contains a search form for AmiGO, a note about AmiGO being the official GO browser, and a Creative Commons Attribution license logo.

the Gene Ontology

Open menus

Home

FAQ

Downloads

Tools

Documentation

About GO

Projects

Contact GO

Search

gene or protein name

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...

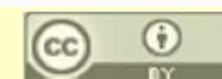
The Gene Ontology project very much encourages input from the community into both the content of the GO and annotation using GO. We are very happy to work with others to ensure that the GO is both complete and accurate, and we also very much encourage communities to submit GO annotations for inclusion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

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

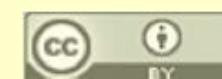
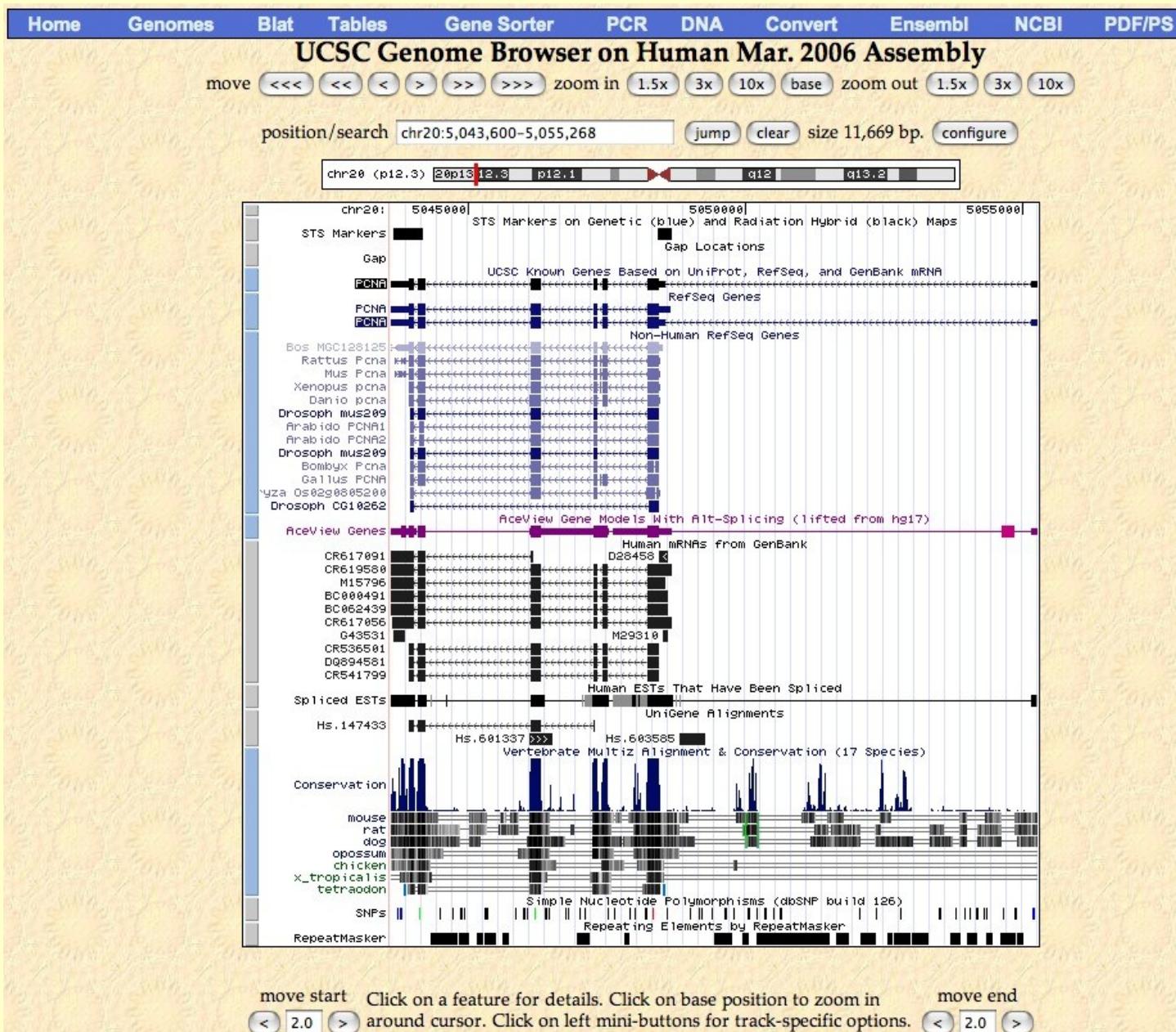
gene or protein name  GO term or ID

AmiGO is the official GO browser and search engine. Browse the Gene Ontology with AmiGO.



# UCSC Genome Browser

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



# ExPASy Proteomics Server

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



The ExPASy Proteomics Server website interface. At the top, there are logos for SIB (Swiss Institute of Bioinformatics) and ExPASy. A search bar is located at the top right. Below the header, a navigation menu includes links for Databases, Tools, Services, Mirrors, About, and Contact. The main content area features a "Complete table of available documents" table with three columns: "ExPASy", "Databases", and "Tools and Software Packages".

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



Doug Brutlag 2010

# Inferring Biological Function from Protein Sequence

Consensus Sequences  
or Sequence Motifs

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

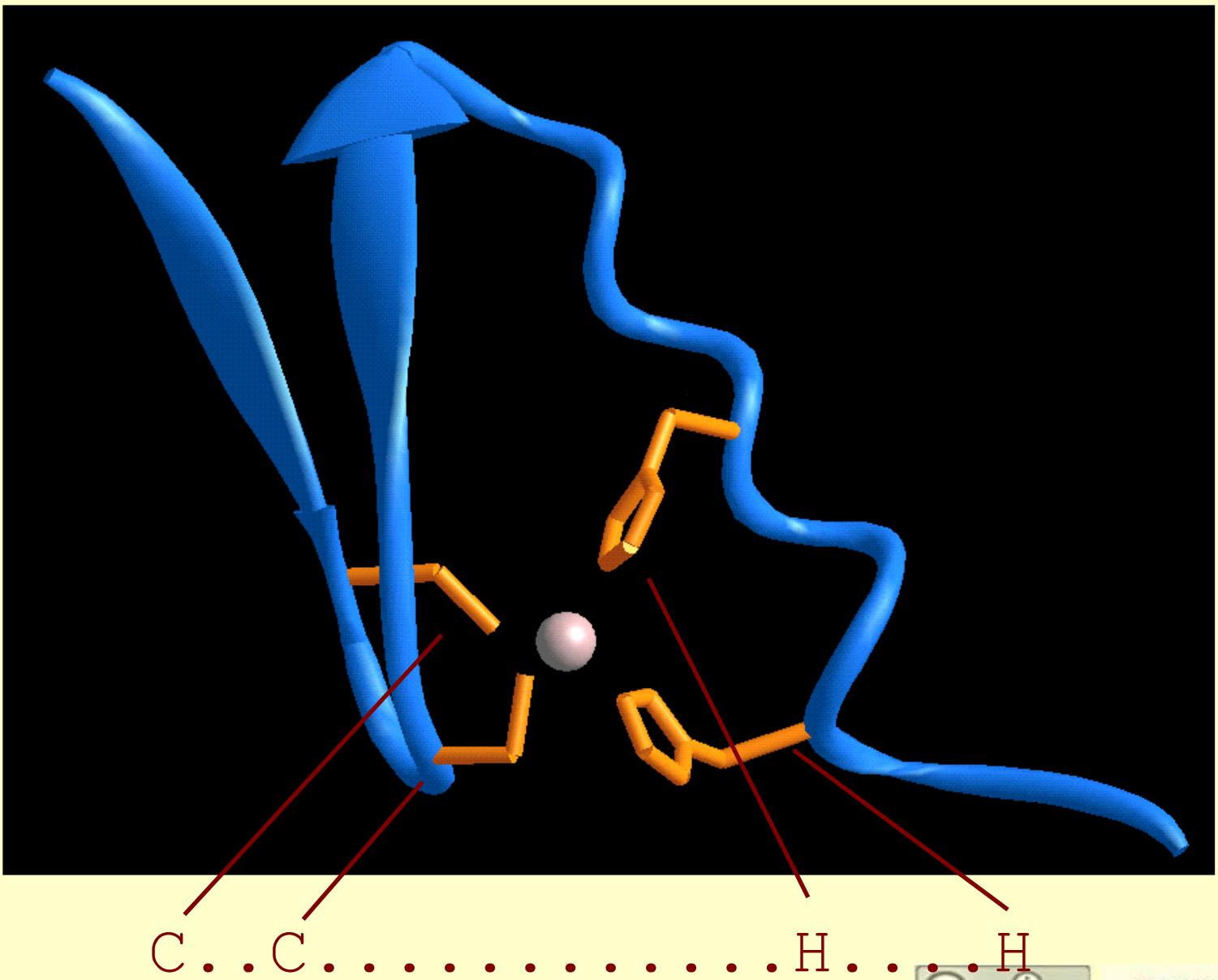
Sequences of Common  
Structure or Function

Sequence Similarity

Query VLSPADKTNVKAAGKVGKAHAGEVGAEALERMFLSFPTTKTYFPHF-----DLSHGS  
Match HLTPEEKSAVTALWGKV--NVDEYGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN  
10 20 30 40 50  
10 20 30 40 50



# A Typical Motif: Zinc Finger DNA Binding Motif



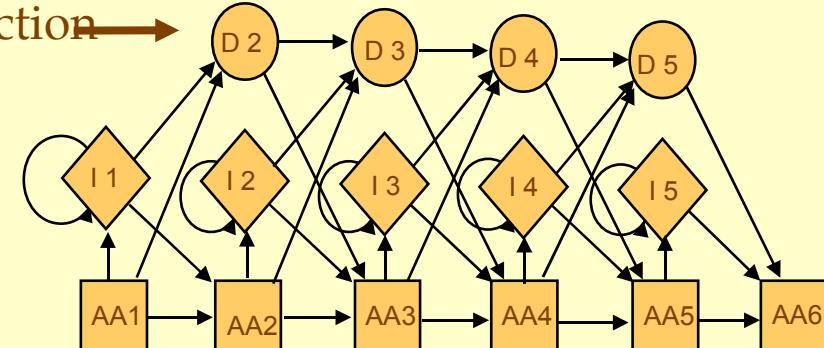
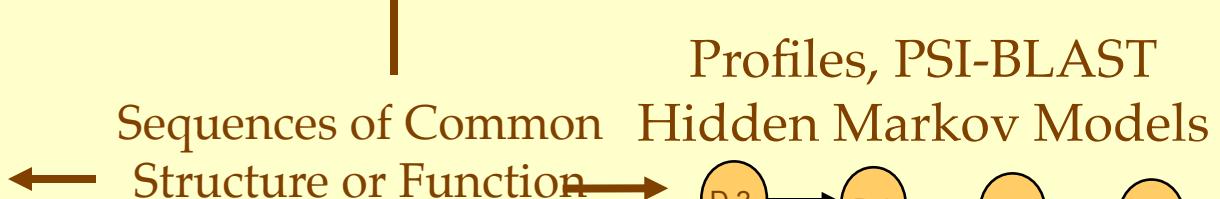
# Inferring Biological Function from Protein Sequence

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	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

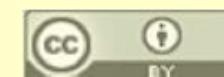
Consensus Sequences  
or Sequence Motifs

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



Sequence Similarity

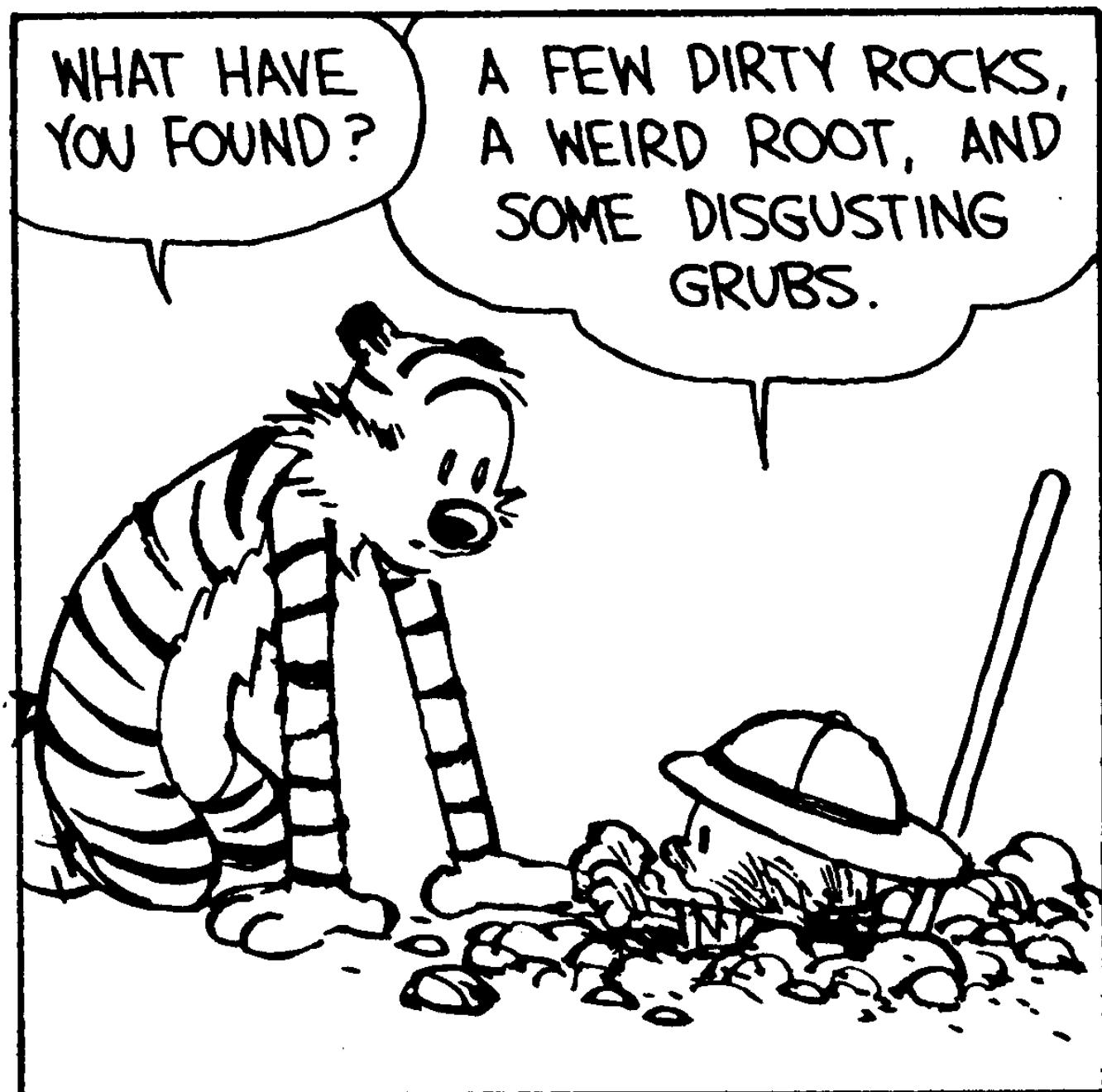
Query	VLSPADKTNVKAAGKVGKVAHAGEVGAEALERMFLSPTTKTYFPHF-----DLSHGS
	:   :   :     :
Match	HLTPEEKSAVTALWGKV--NVDEYGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
	10 20 30 40 50
	10 20 30 40 50



# 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		E	F	G	H	
1	VHLTPEEKSAVTALWGKVNV	EVGGEALGRLLVVY	PWTQRFFESFGDLSTPDAVMGNPK		VKAHGKKVLGAFSDG	DAHLDNLKGTFAT	LSELHCDKLHVDPENFRLLGNVLVCVLAH	FKEFTPPVQAA	YQKVVAGVANALAHKYH
2	VQLSGEEKAAVLALWDKVNEE	EVGGEALGRLLVVY	PWTQRFFDSFGDLSNPGAVMGNPK		VKAHGKKVLHSFGE	VHLDNLKGTFAA	LSELHCDKLHVDPENFRLLGNVLVVVLARH	FKGDFTEQ	SYQKVVAGVANALAHKYH
3	VLSPADKTNVKAAWGKVGAHAGEYGAELERMFLSFPTTKTYFPHF	DLSH	GSAQ		VKGHGKKVADALTNA	VAHVDDMPNALSA	LSDLHAKLRVDPVNFKLLSHCLLVTLAAH	LPAEFTPAVHASLDKF	FLASVSTVLTSKYR
4	VLSAADKTNVKAAWSKVGGHAGEYGAELERMFLGFPPTKTYFPHF	DLSH	GSAQ		VKAHGKKVGDA	VGHLDLPGALSN	LSDLHAKLRVDPVNFKLLSHCLLSTLAVH	LPNDFTP	AVHASLDKF
5	PIVDTGSVAPLSAAEKT	KIRSAWAPVYS	YSDVETSGVDILVKFFTSTPAAEEFFPKFKGLT	ADELKKSAD	VKAHGKKVDA	VASMDDTEK	LSDLHAKLRVDPVNFKLLSHCLLSTLAVH	LPNDFTP	SSVSTVLTSKYR
6	VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSH	PETLEK	FDRFKHLKTEAEMKASED		LLGAI	MSSMKDLSGKHAKSFEVDPEYFKVLA	LPNDFTP	AGFEKLLRMICILLRSAY	
7	GALTESQAALVKSSWEEFN	PKH	THRF		LOAHAGKVF	VIYEAAI	QLEVTVGVVAS	QGAMNKALELFRKDIAAKYKELGYQG	
	*	*	*		*	*	*	*	

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

# Consensus Sequence From a Multiple Sequence Alignment

ClustalW Insulin Alignments

	10	20	30
IPGP			
IPDK			
IPDG	M <b>A L W M R</b> L L P <b>L L</b> A <b>L L</b> A L W A P A P T R A F V <b>N Q H</b>		F V S <b>R H</b>
IPCH	M <b>A L W I R</b> S L P <b>L L</b> A <b>L L</b> V F S G P G - T S Y <b>A A N Q H</b>		<b>A A N Q H</b>
IPCA	M <b>A V W I Q</b> A G A <b>L L</b> F <b>L L</b> A V S S V N A N A G <b>A P - Q H</b>		<b>A P - Q H</b>
IPBO			F V <b>N Q H</b>
IPAF	M A <b>A L W L Q</b> S F S <b>L L</b> V <b>L L</b> V V S W P G S Q A V <b>A P A Q H</b>		<b>A P A Q H</b>
	A . W . .	L L L L	A N Q H
	40	50	60
IPGP	<b>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</b>		
IPDK	<b>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</b>		
IPDG	<b>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</b>		
IPCH	<b>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</b>		
IPCA	<b>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</b>		
IPBO	<b>L C G S H L V D A L Y L V C G D R G F F Y T P K A R R E V E</b>		
IPAF	<b>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</b>		
	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 <b>P Q V</b> E Q T E <b>L G M G</b> - - - - L <b>G A G G L Q</b> P - - L Q G		
IPDK	Q <b>P - L</b> V N G P <b>L H G E</b> - - - - V <b>G E L P F Q</b> - - - H <b>E</b>		
IPDG	D L Q <b>V R D V E L A G A</b> - - - - P <b>G E G G L Q</b> P L A L E G		
IPCH	Q <b>P - L</b> V S S P <b>L R G E</b> - - - - A <b>G V L P F Q</b> - - - Q <b>E</b>		
IPCA	L G F <b>L P P K S</b> - - - - A Q E T E V A D <b>F A F K D H A E</b>		
IPBO	G <b>P Q V G A L E L A G G</b> - - - - P <b>G A G G L E</b> - - - G		
IPAF	L G F <b>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</b>		
	P L L G	G F Q	E
	100	110	120
IPGP	A <b>L Q X X</b> - - <b>G I V D Q C C</b> T G T <b>C T R H Q L Q S Y C N</b>		
IPDK	E Y <b>Q X X</b> - - <b>G I V E Q C C</b> E N P <b>C S L Y Q L E N Y C N</b>		
IPDG	A <b>L Q K R</b> - - <b>G I V E Q C C</b> T S I <b>C S L Y Q L E N Y C N</b>		
IPCH	E Y <b>E K V K R</b> <b>G I V E Q C C</b> H N T <b>C S L Y Q L E N Y C N</b>		
IPCA	V <b>I R K R</b> - - <b>G I V E Q C C</b> H K P <b>C S I F E L Q N Y C N</b>		
IPBO	P <b>P Q K R</b> - - <b>G I V E Q C C</b> A S V <b>C S L Y Q L E N Y C N</b>		
IPAF	M <b>M V K R</b> - - <b>G I V E Q C C</b> H R P <b>C N I F D L Q N Y C N</b>		
	. 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](#)

---

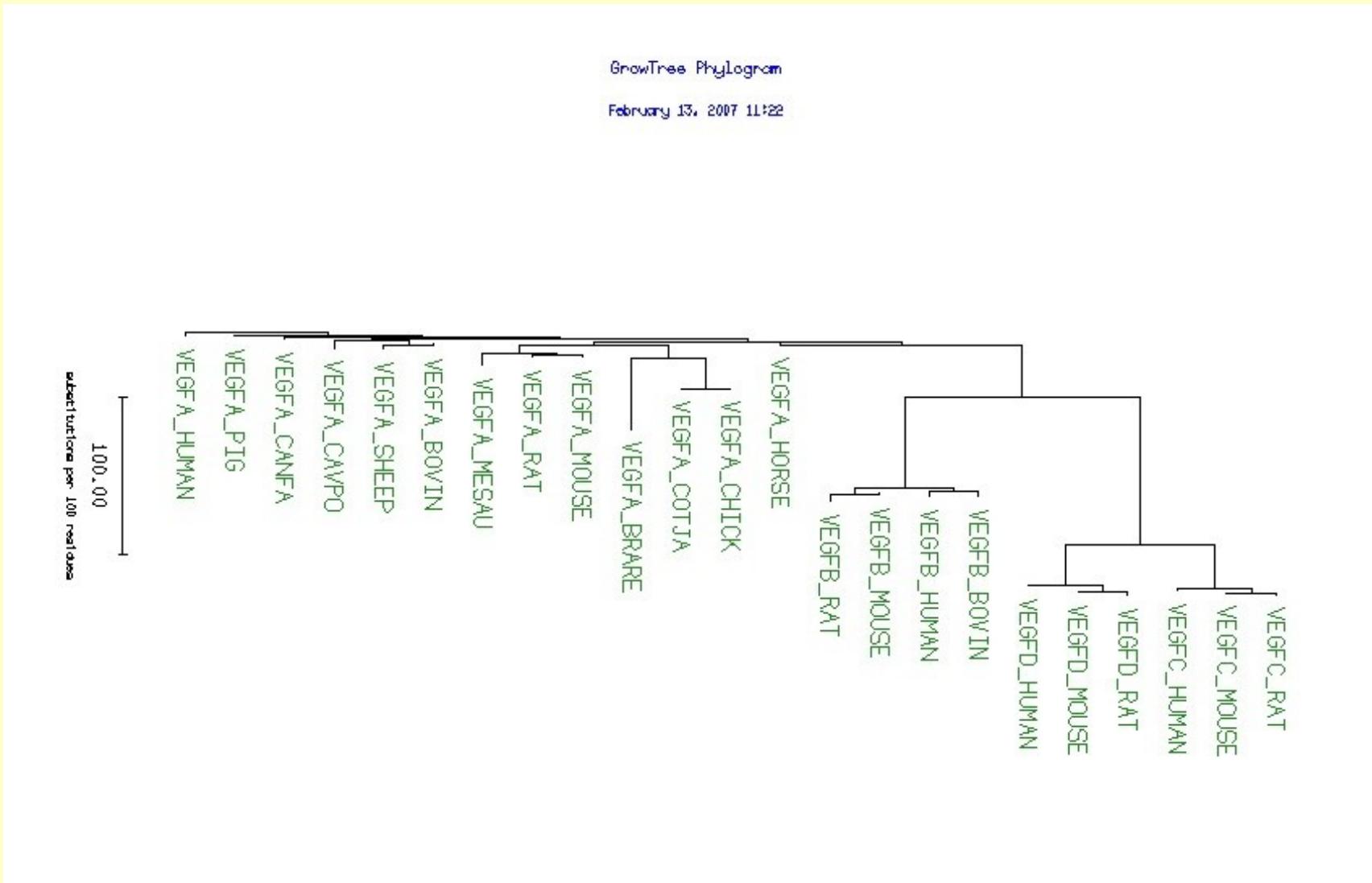
```

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
DESC
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 * *

```

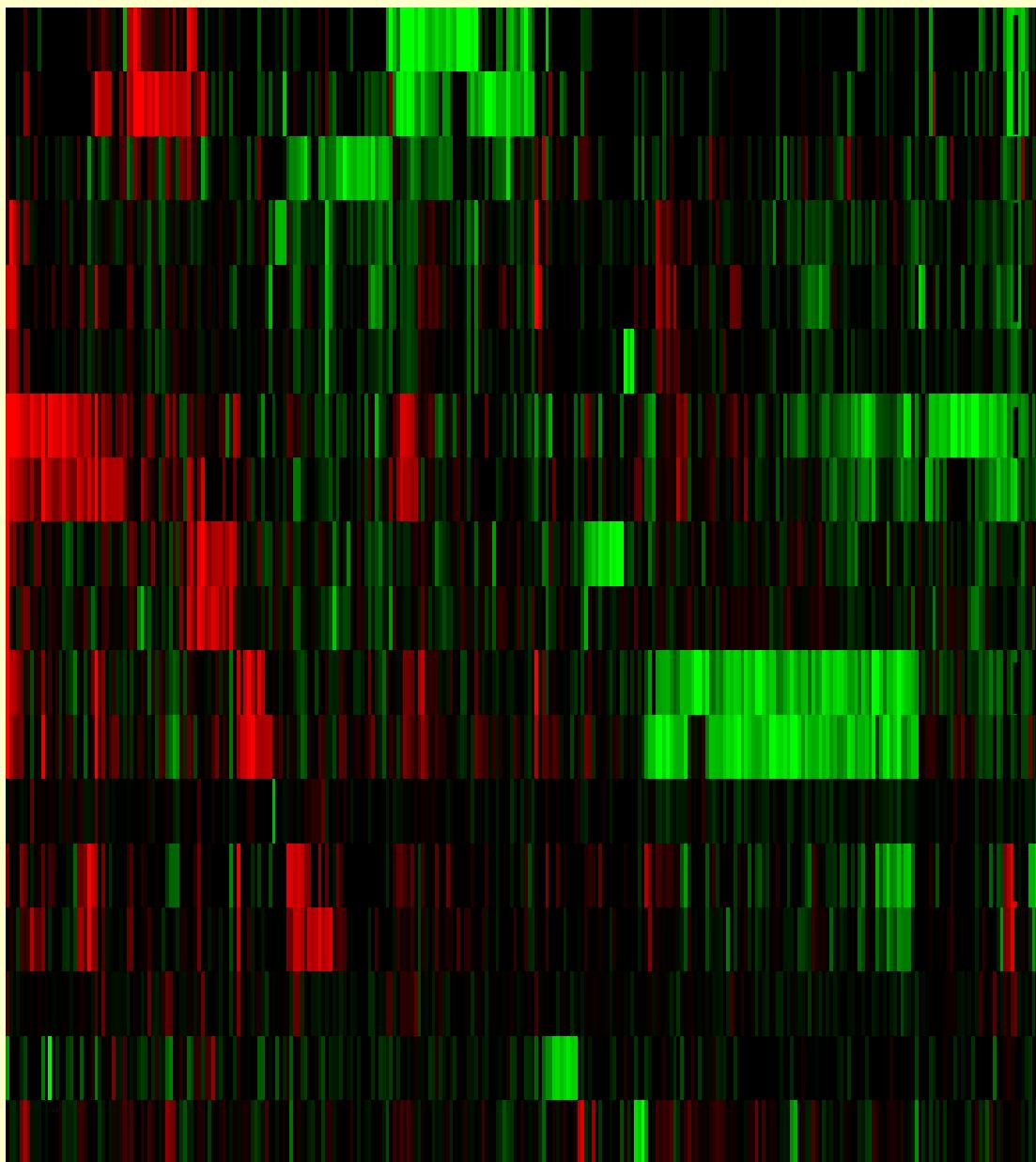


# GrowTree VegF Neighbor Joining Tree



Doug Brutlag 2010

# Human Gene Expression Signatures



T Cells Signaling

DNA Damage

Fibroblast Stimulation

B Cells Signaling

CMV Infection

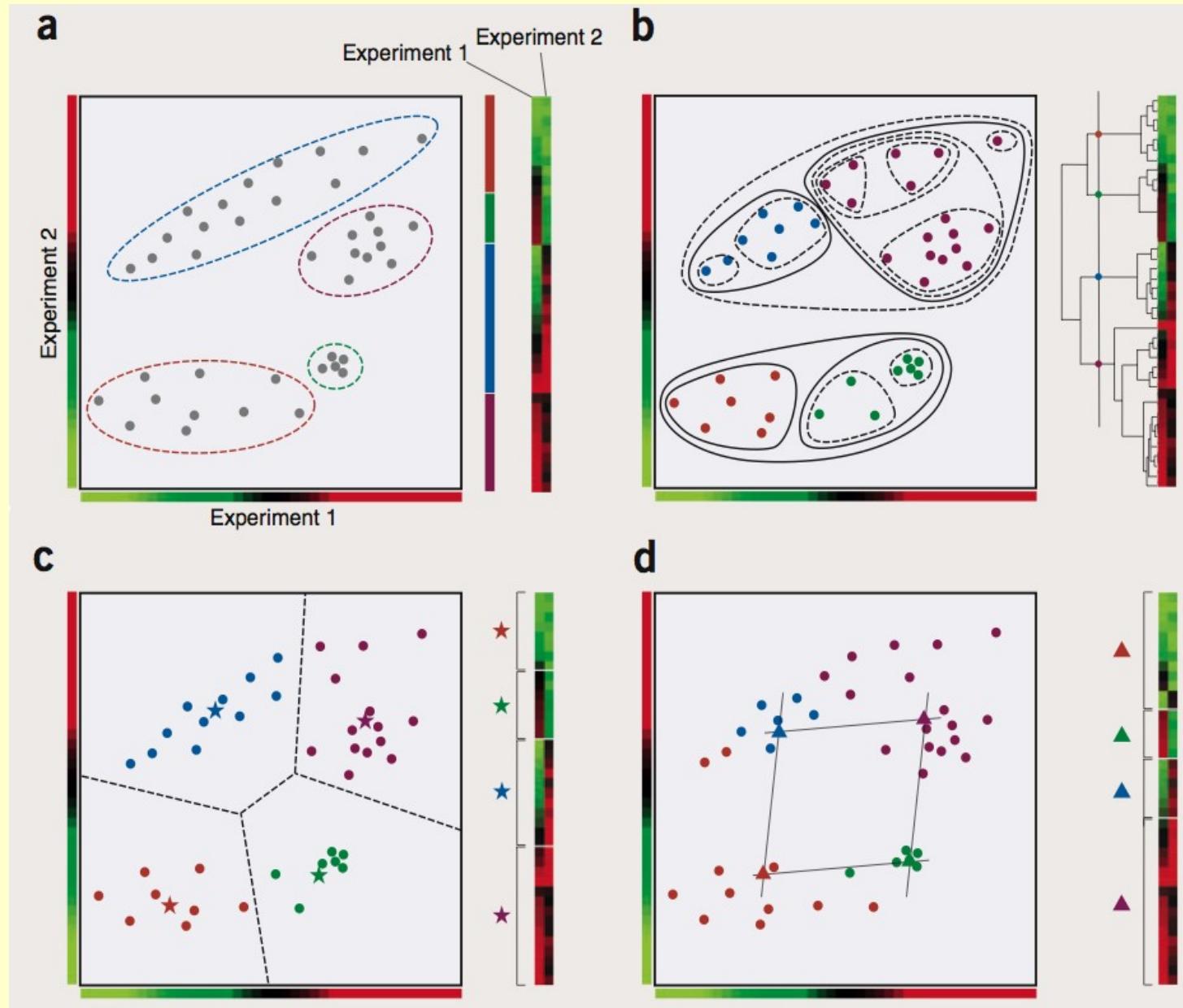
Anoxia

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*

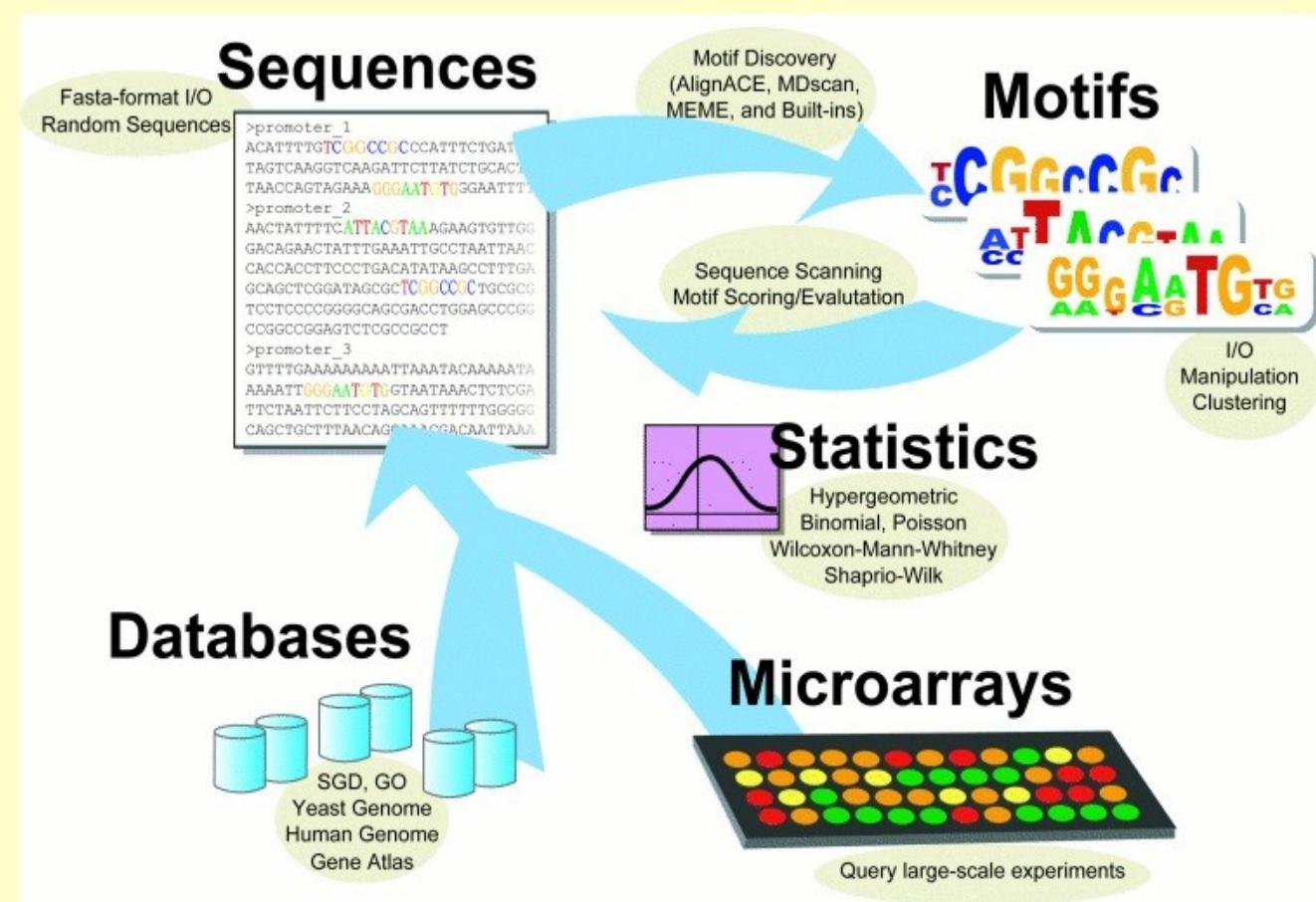
[Home](#)  
[People](#)  
[Publications](#)  
**Data / Download**

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-

Cores

GATGGCTGCACCACGTGTATGC...ACG  
CACATCGCATCACGTGACCAAGT...GAC  
GCCTCGCACGTGGTGGTACAGT...AAC  
TCTCGTTAGGACCATCACGTGA...ACA  
CGCTAGCCCACGTGGATCTTGA...AGA 

Pho 5

Pho 8

Pho 81

Pho 84

Pho ...



# Finding Transcription Factor Binding Sites

---



Upstream Regions

Co-expressed  
Genes

GATGGCTGCAC**CACGTG**TATGC...ACG**ATGTCTCGC**  
CACATCGCAT**CACGTG**ACCAGT...GAC**ATGGACGGC**  
GCCTCG**CACGTG**GTGGTACAGT...AAC**ATGACTAAA**  
TCTCGTTAGGACCAT**CACGTG**A...ACA**ATGAGAGCG**  
CGCTAGCC**CACGTG**GATCTTGT...AGA**ATGGCCTAT**

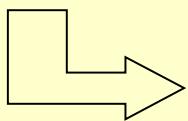
# Finding Transcription Factor Binding Sites



Upstream Regions

Co-expressed Genes

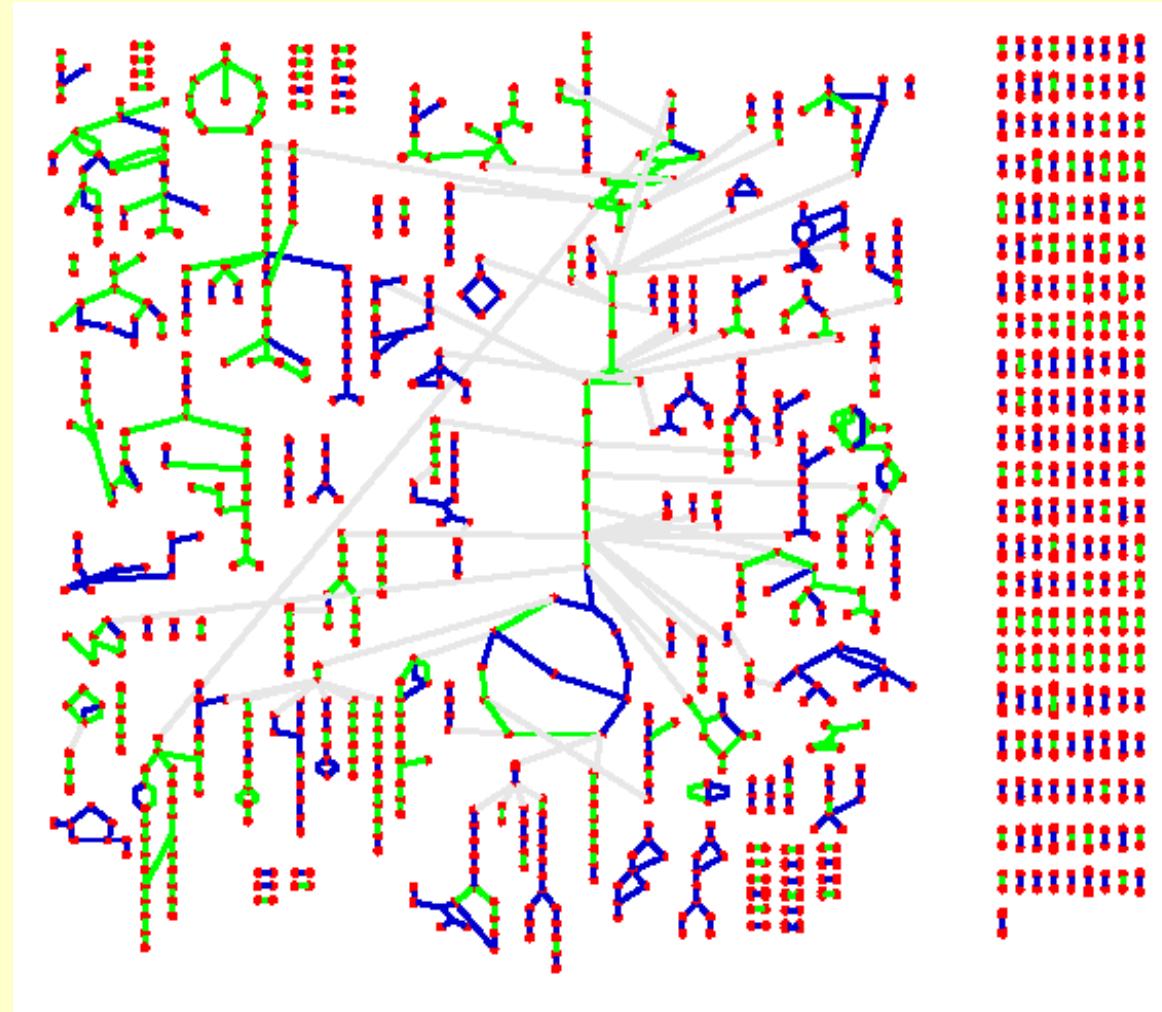
ATGGCTGCAC	CACGTT	TATGC...ACG	ATGTCTCGC
CACATCGCAT	CACGTG	ACCAGT...GAC	ATGGACGGC
GCCTCG	CACGTG	GTGGTACAGT...AAC	ATGACTA
TTAGGACCAT	CACGTG	A...ACA	ATGAGAGCG
CGCTAGCC	CACGTI	GATCTTGT...AGA	ATGGCCTAT



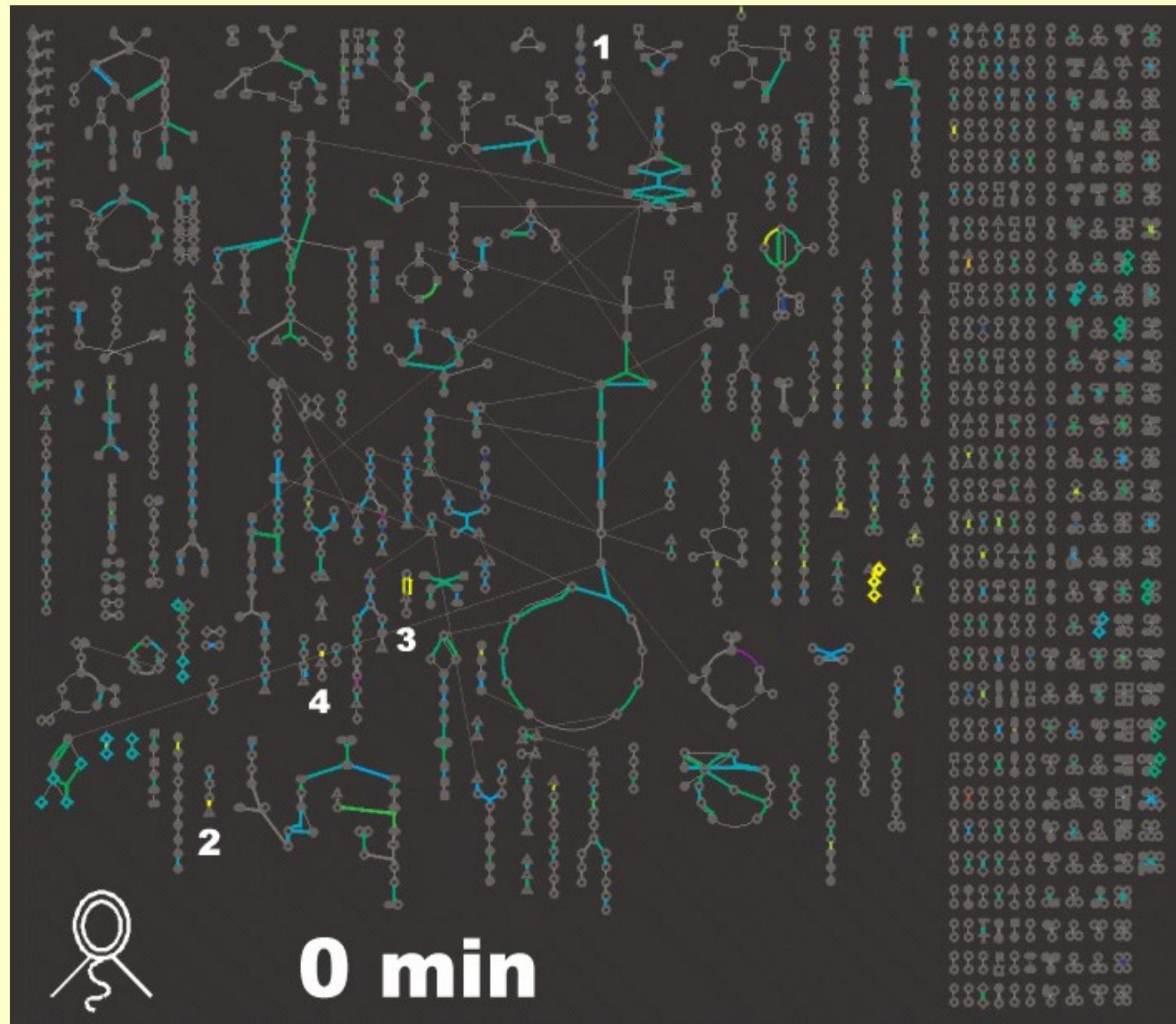
Pho4 binding

# Metabolic Networks: BioCyc

<http://biocyc.org/>

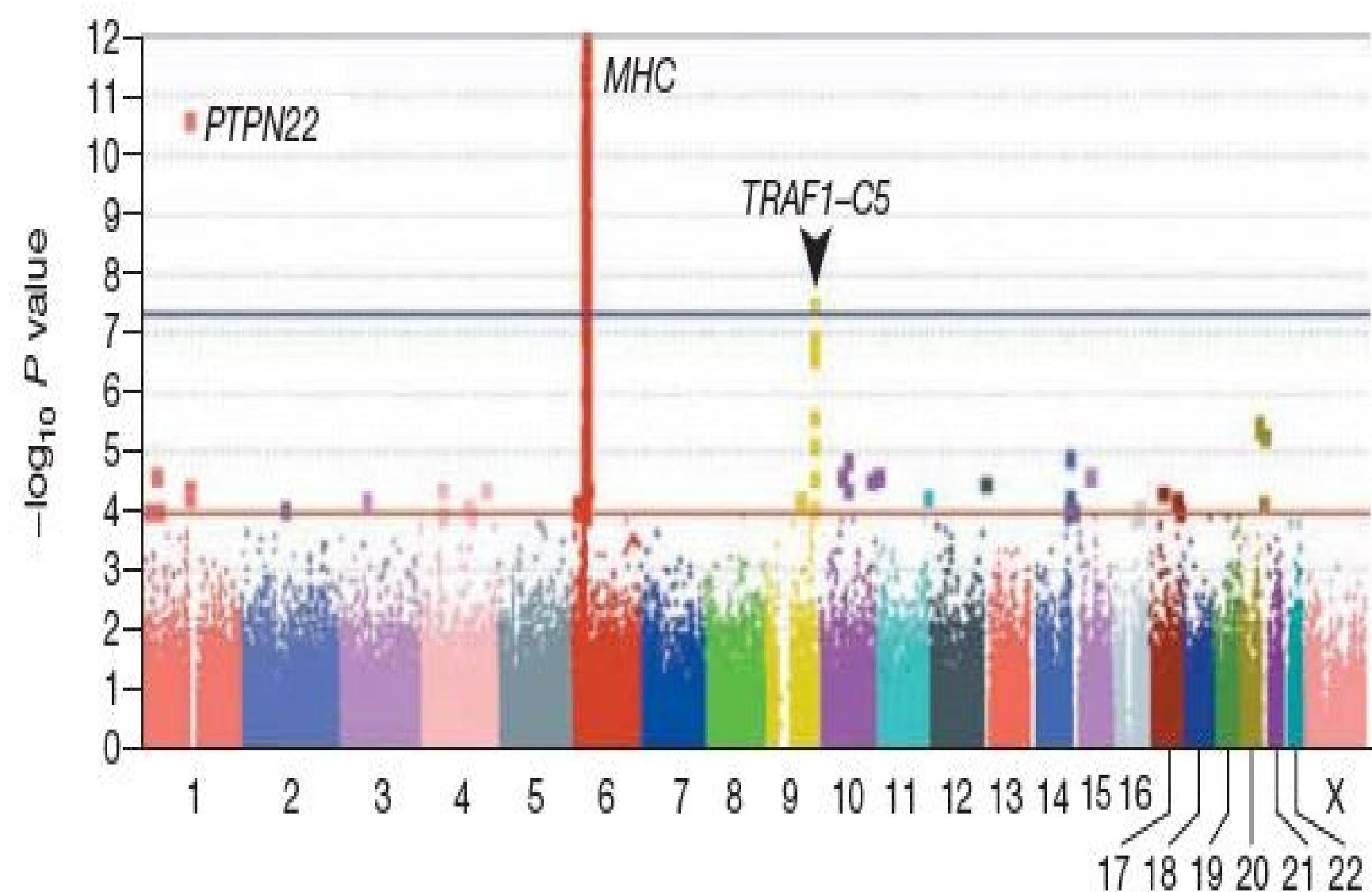


# *C. crescentus* Cell Cycle Gene Expression



# Genome Wide Associations in

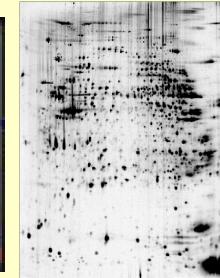
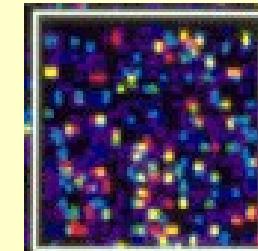
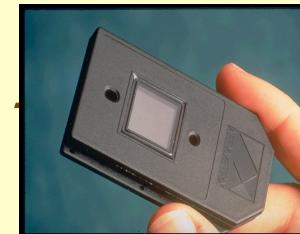
**Figure 3.** Genome-wide Association Findings in Rheumatoid Arthritis



# Leveraging Genomic Information in Medicine

Novel Diagnostics

Microchips & Microarrays

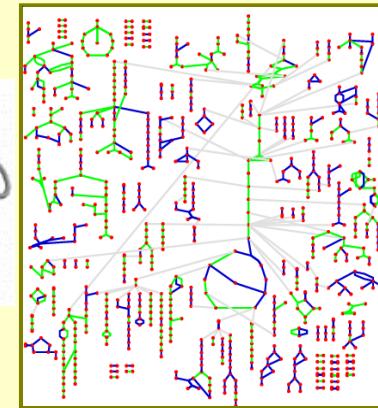
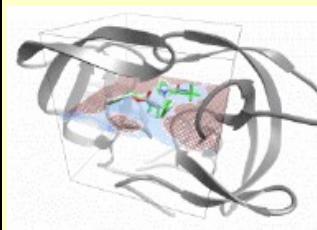
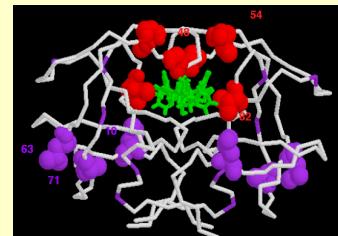


Gene Expression - RNA

Proteomics - Protein

Novel Therapeutics

Drug Target Discovery



Rational Drug Design

Molecular Docking

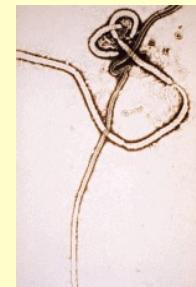
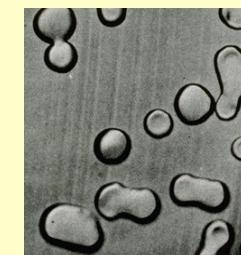
Gene Therapy

Stem Cell Therapy

Understanding Metabolism

Understanding Disease

Inherited Diseases - OMIM



Infectious Diseases

Pathogenic Bacteria

Viruses