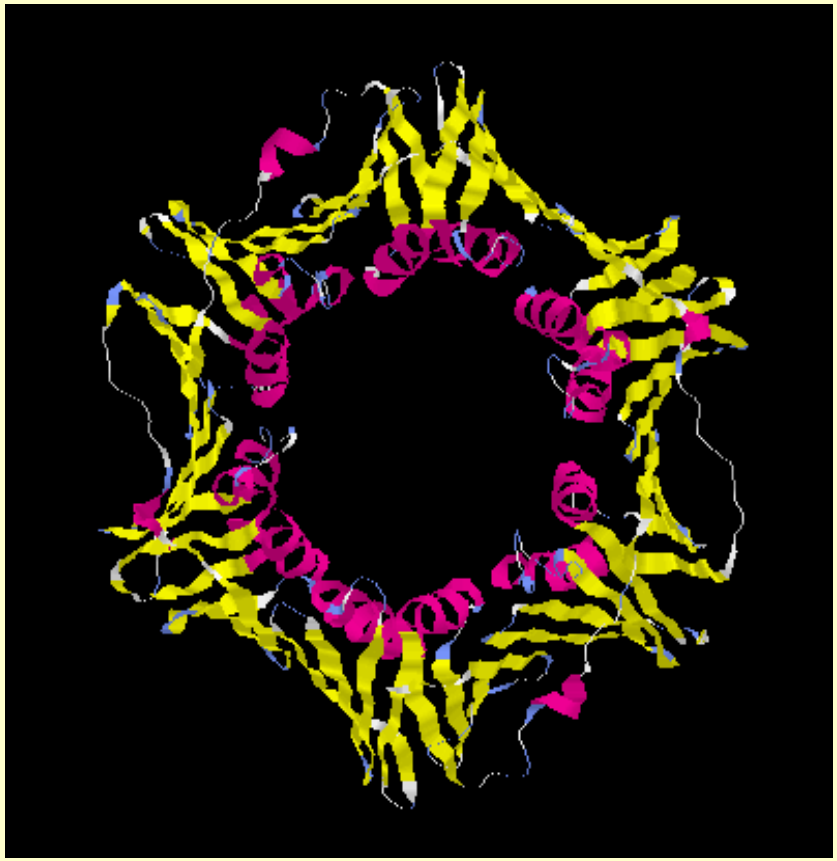
Computational Molecular Biology

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

Doug Brutlag 2010

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

Doug Brutlag 2010

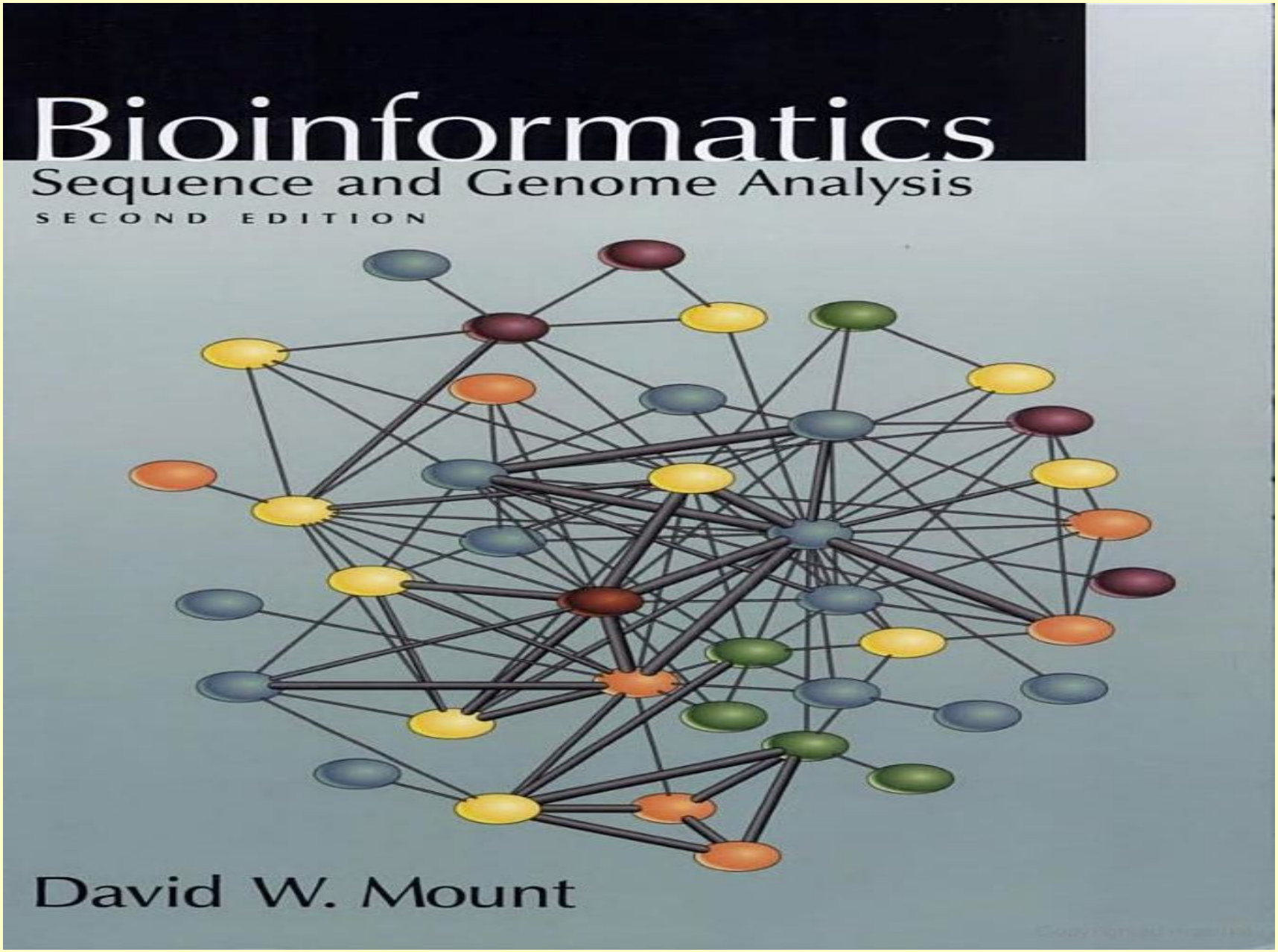
Course Requirements

- 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
- 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

Doug Brutlag 2010

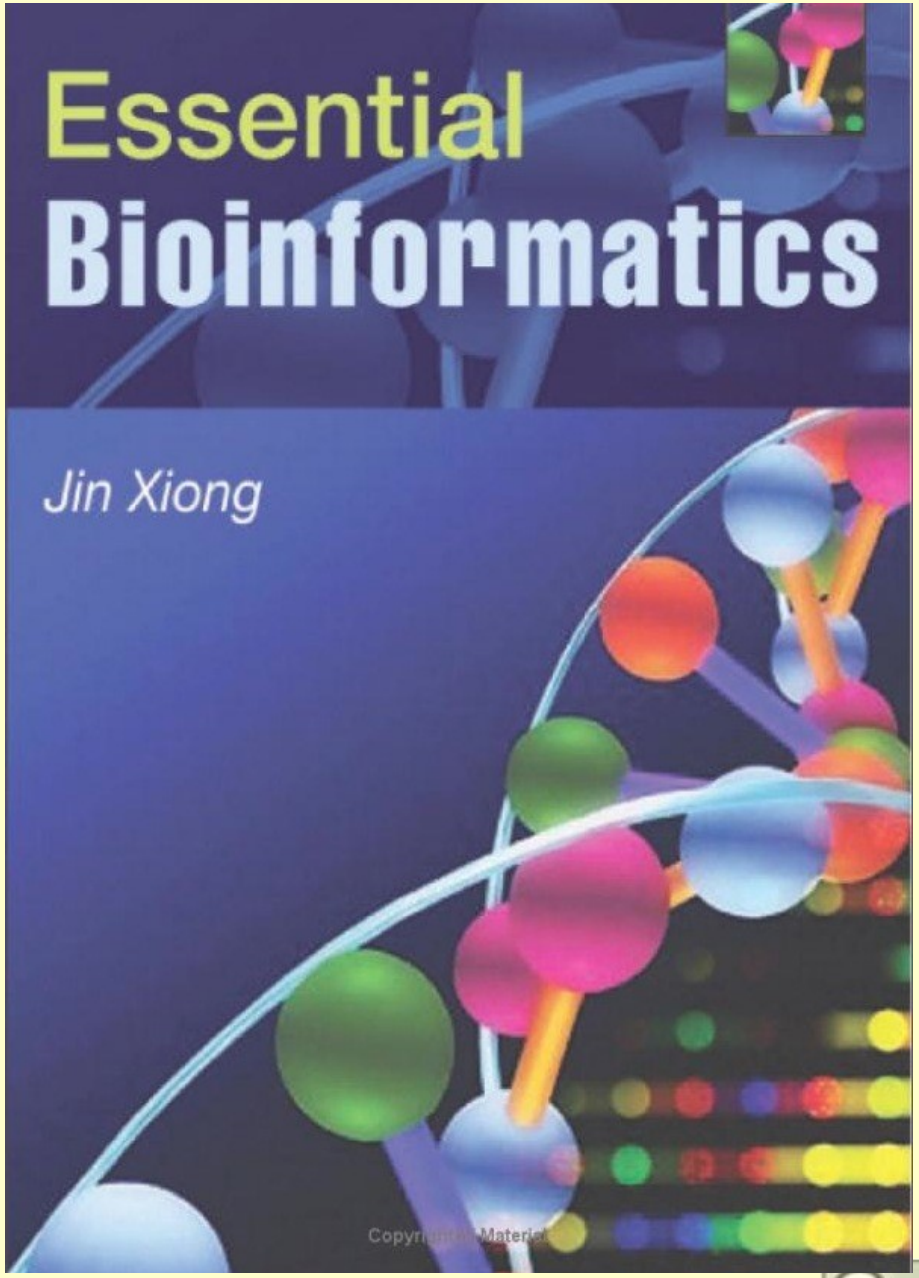
David Mount

Bioinformatics: Sequence and Genome Analysis 2nd Edition



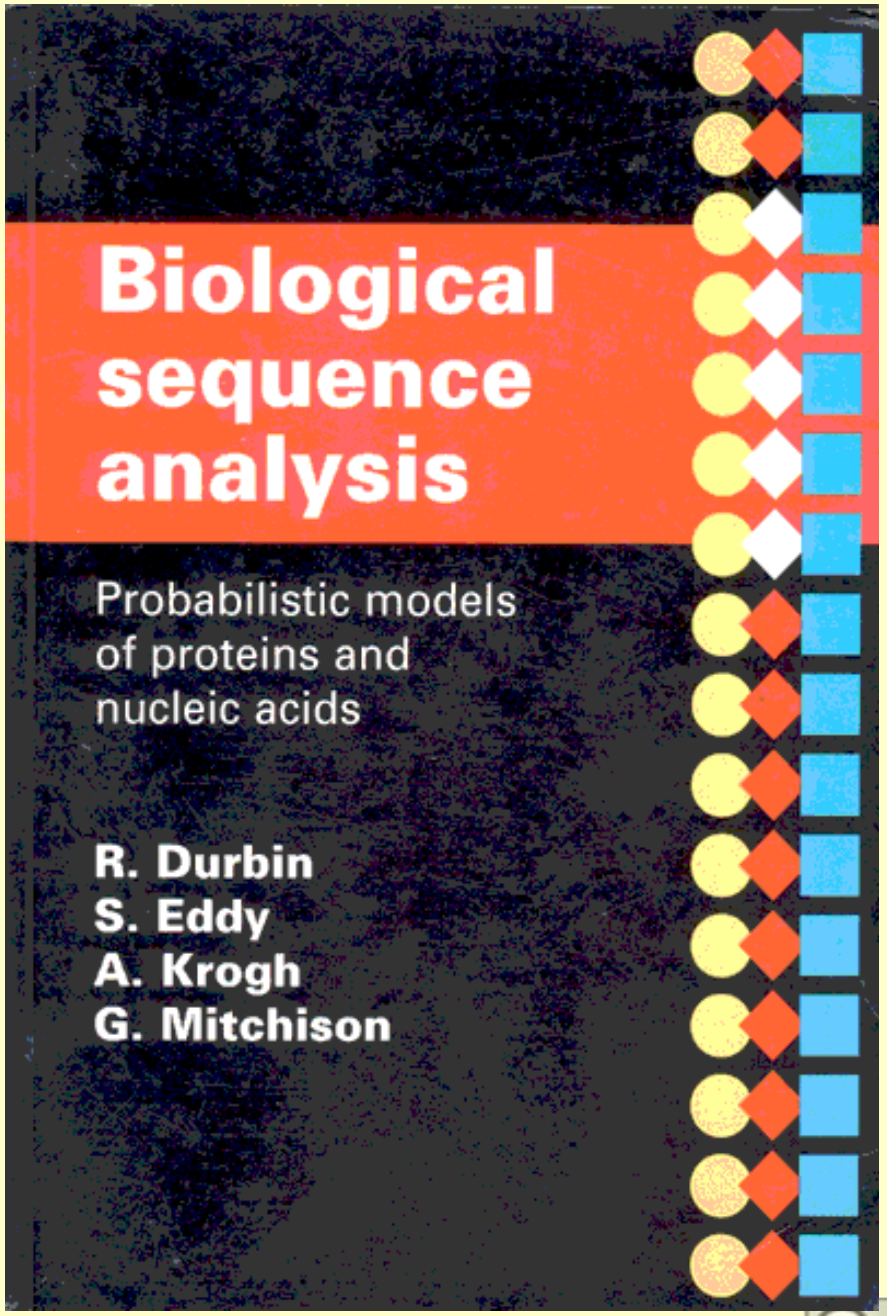
Doug Brutlag 2010

Jin Xiong
Essential Bioinformatics



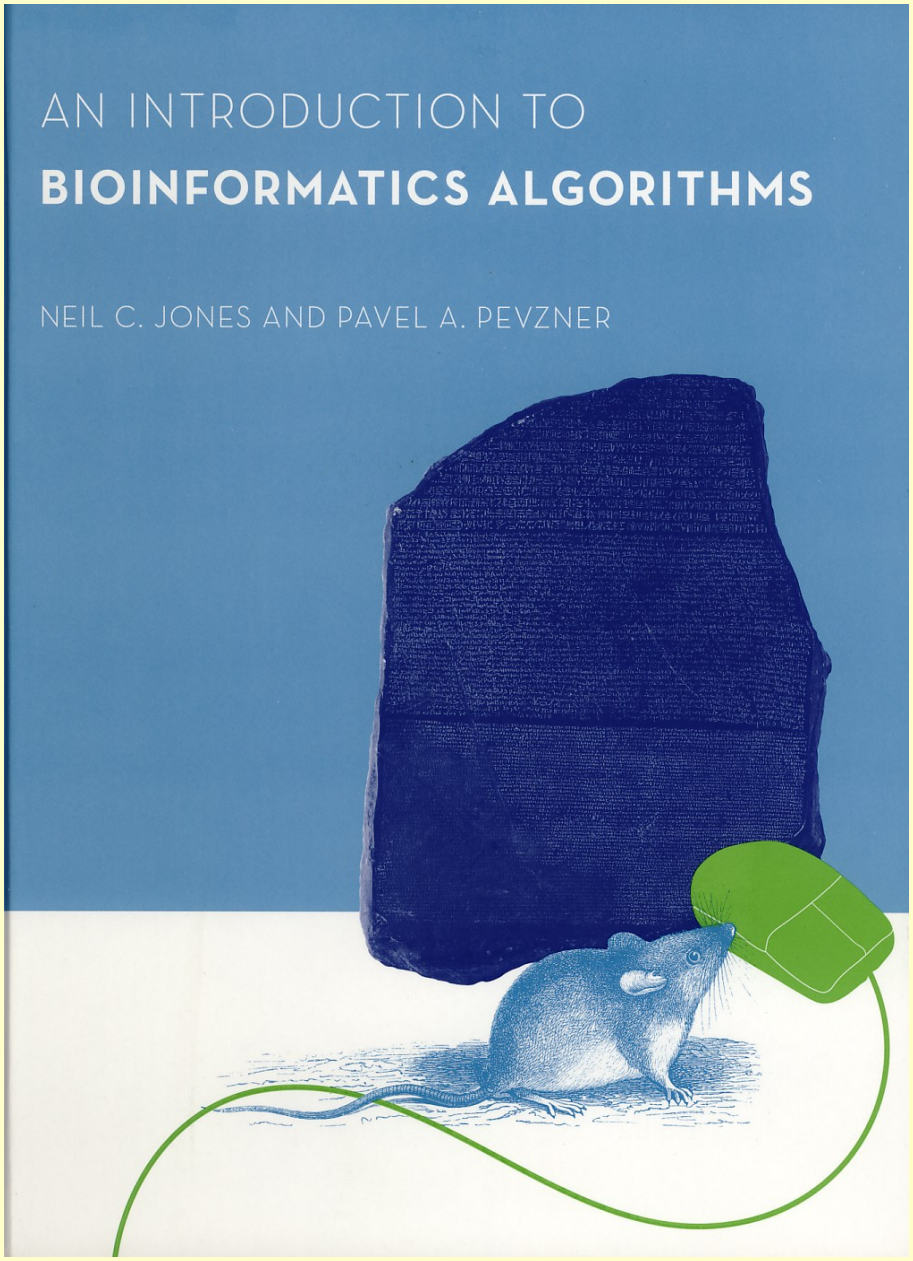
Doug Brutlag 2010

Richard Durbin *et al.*
Biological Sequence Analysis



Doug Brutlag 2010

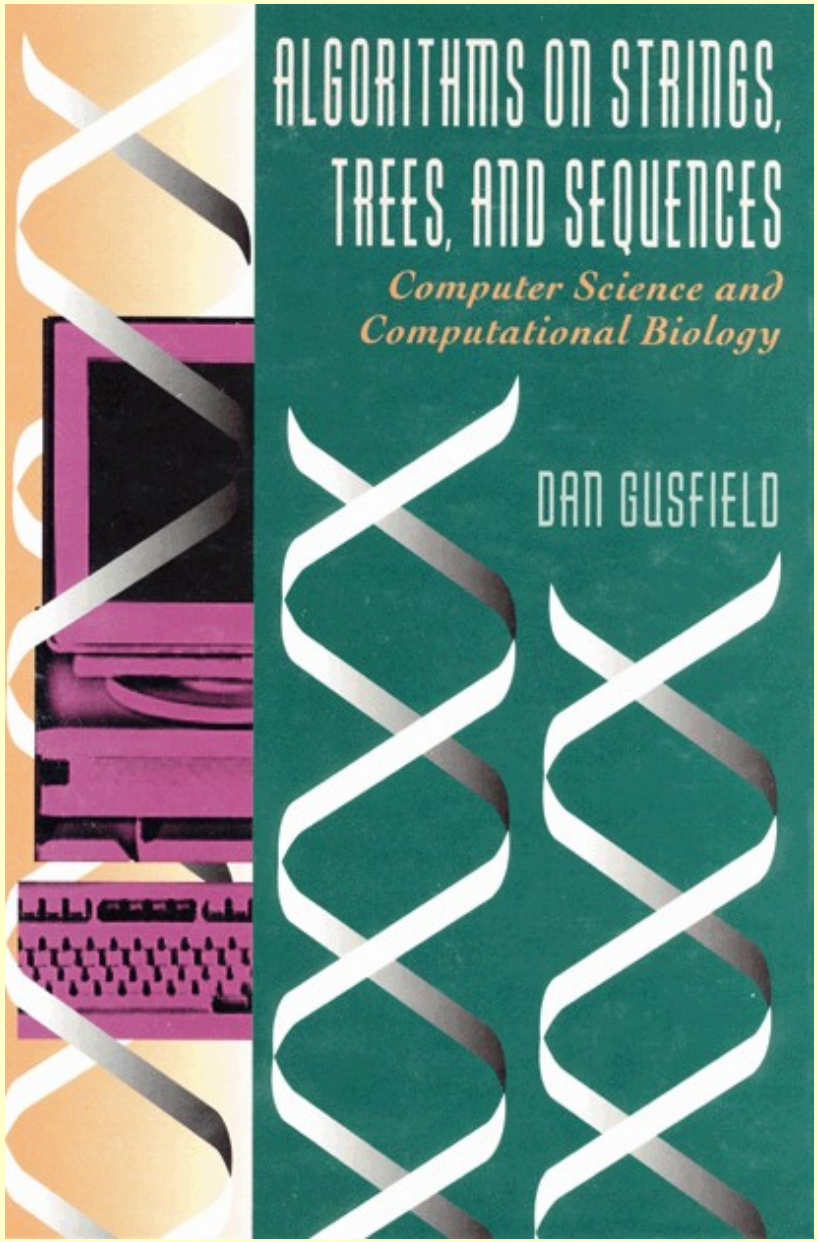
Jones & Pevzner
Bioinformatics Algorithms



Doug Brutlag 2010

Dan Gusfield

Algorithms on Strings, Trees & Sequences



Doug Brutlag 2010

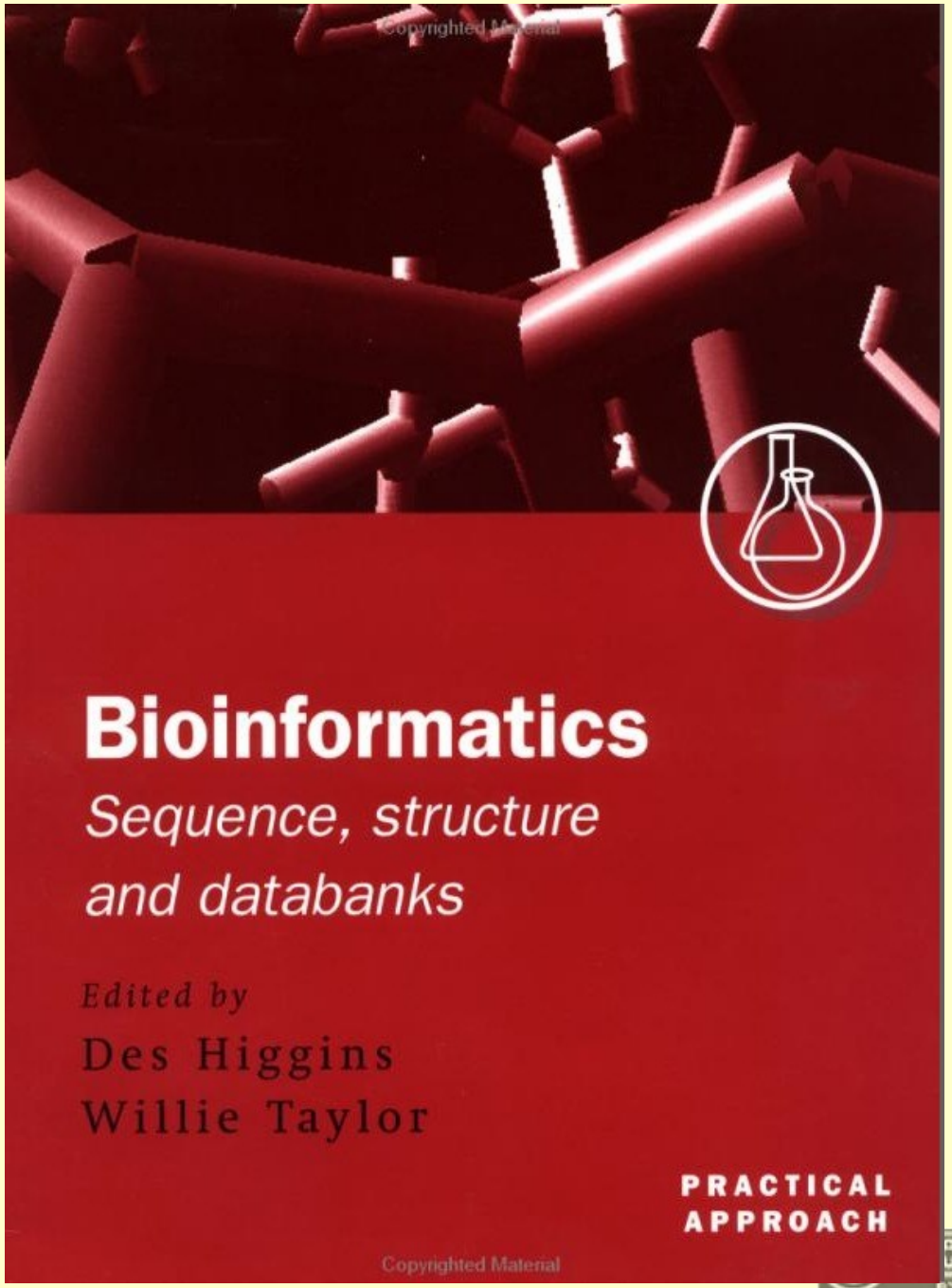
Baldi & Brunak

Bioinformatics: The Machine Learning Approach

Doug Brutlag 2010

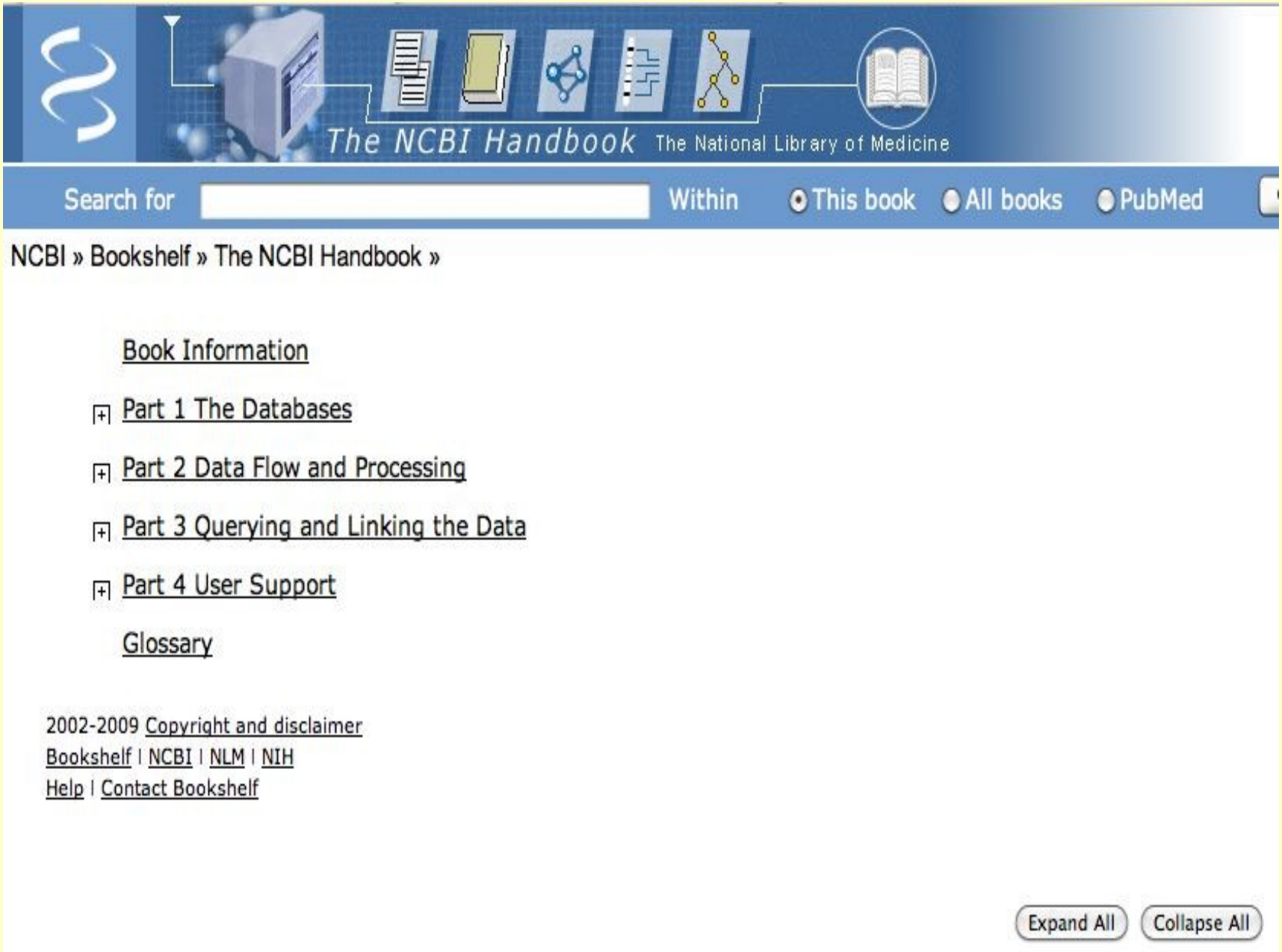
Higgins & Taylor

Bioinformatics: Sequence, Structure & Databanks



NCBI Handbook

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



Doug Brutlag 2010

NCBI Handbook

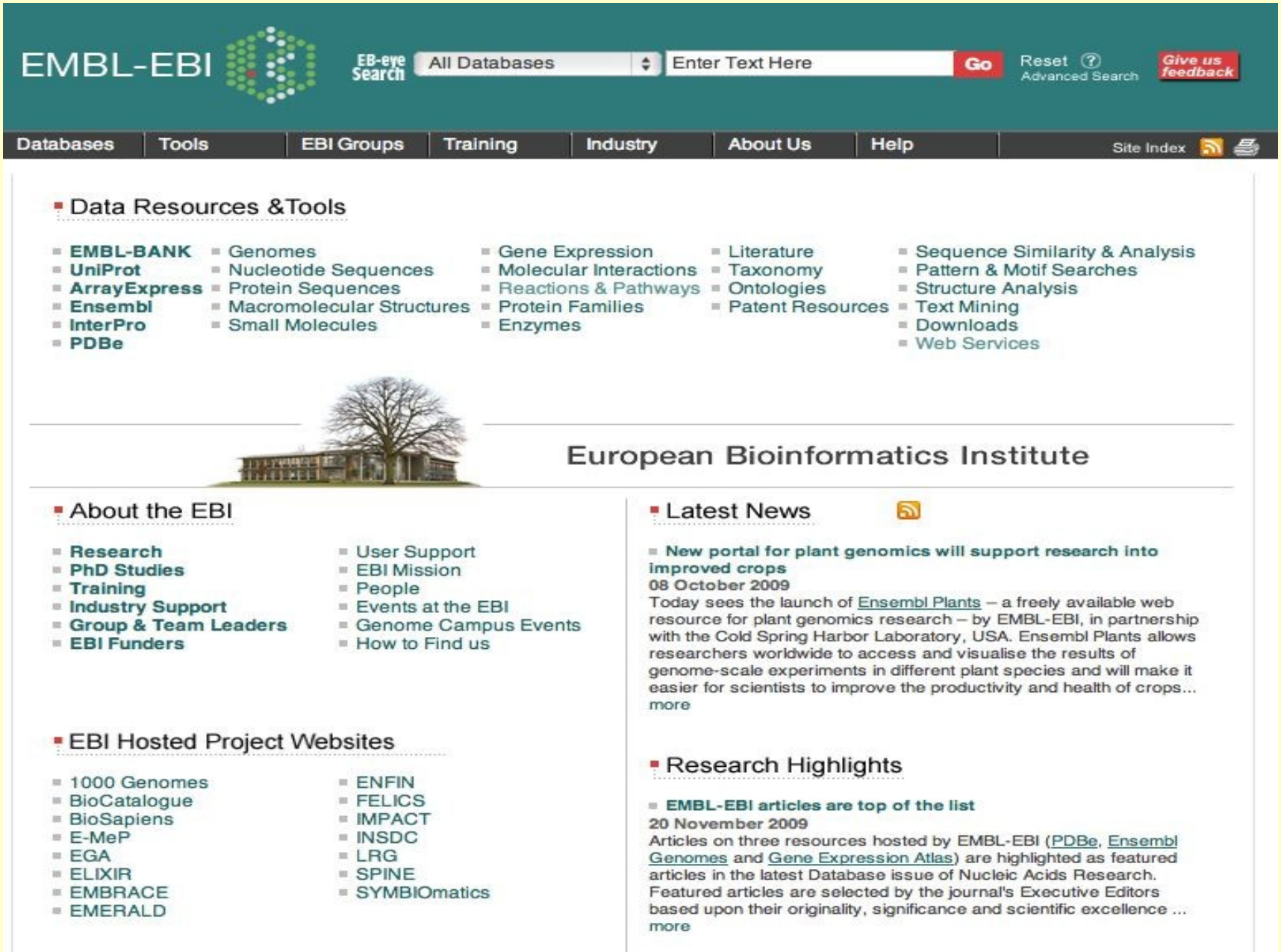
http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook

NCBI » Bookshelf » The NCBI Handbook » The Databases

The Databases

Doug Brutlag 2010

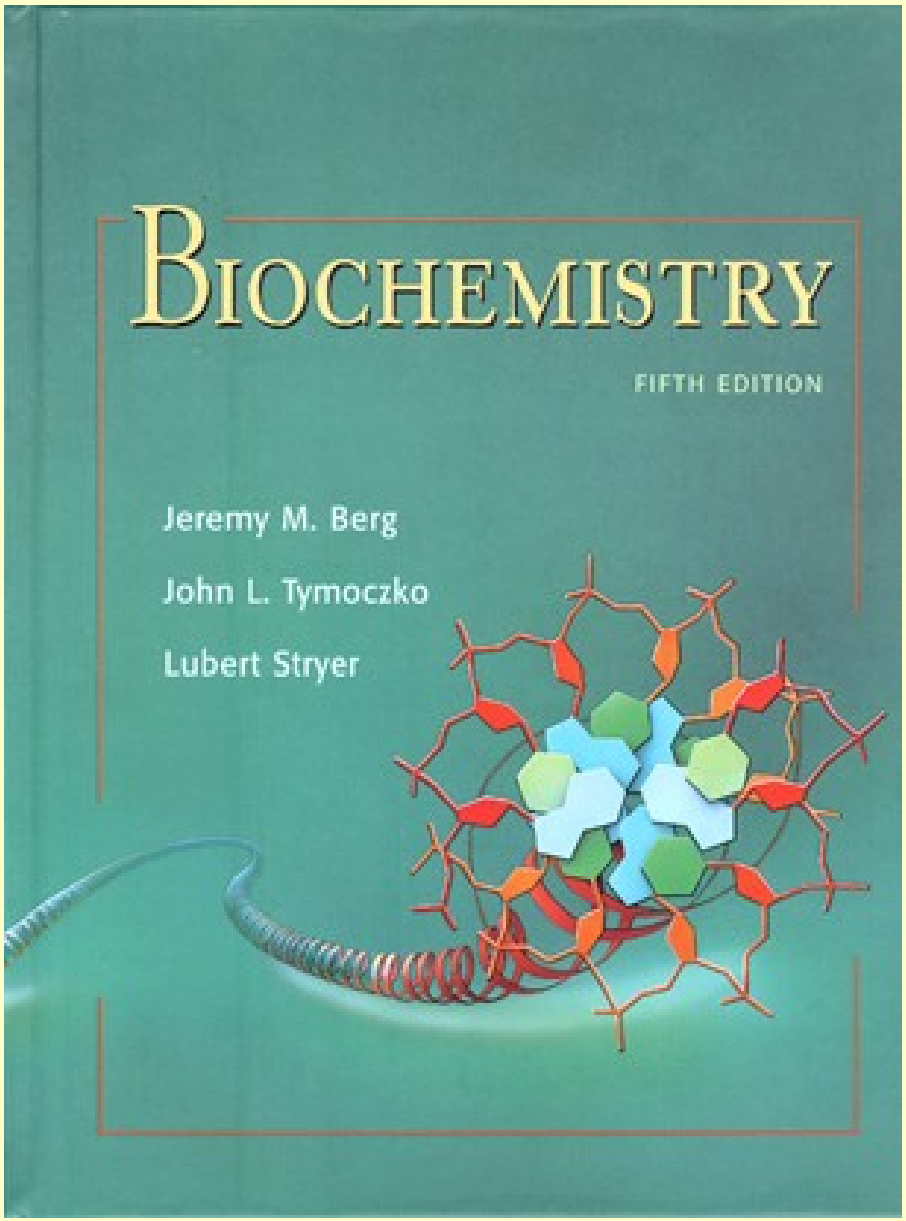
EMBL-EBI Home Page
http://www.ebi.ac.uk/



Doug Brutlag 2010

Berg, Tymoczko & Stryer

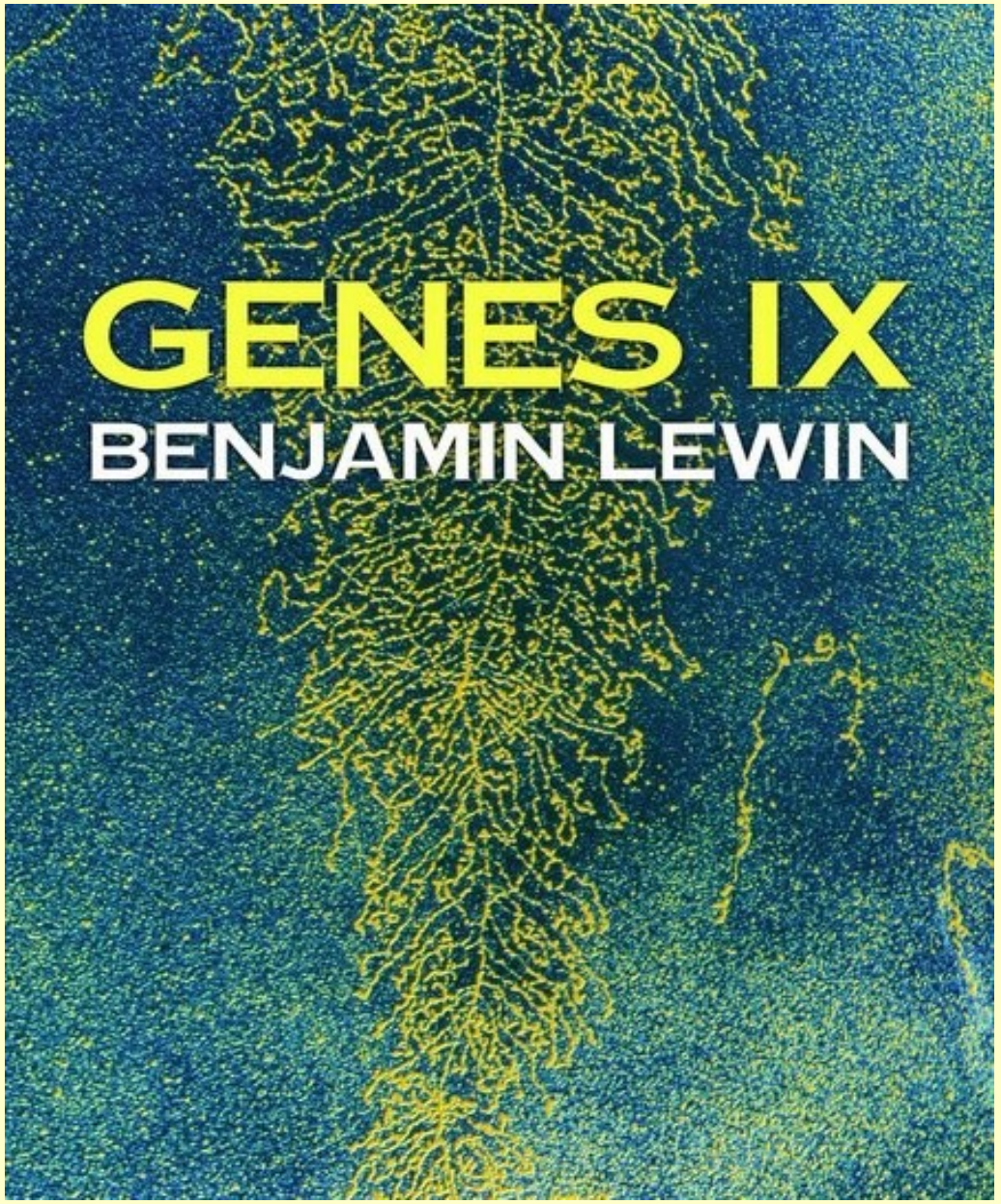
Biochemistry, Fifth Edition



Doug Brutlag 2010

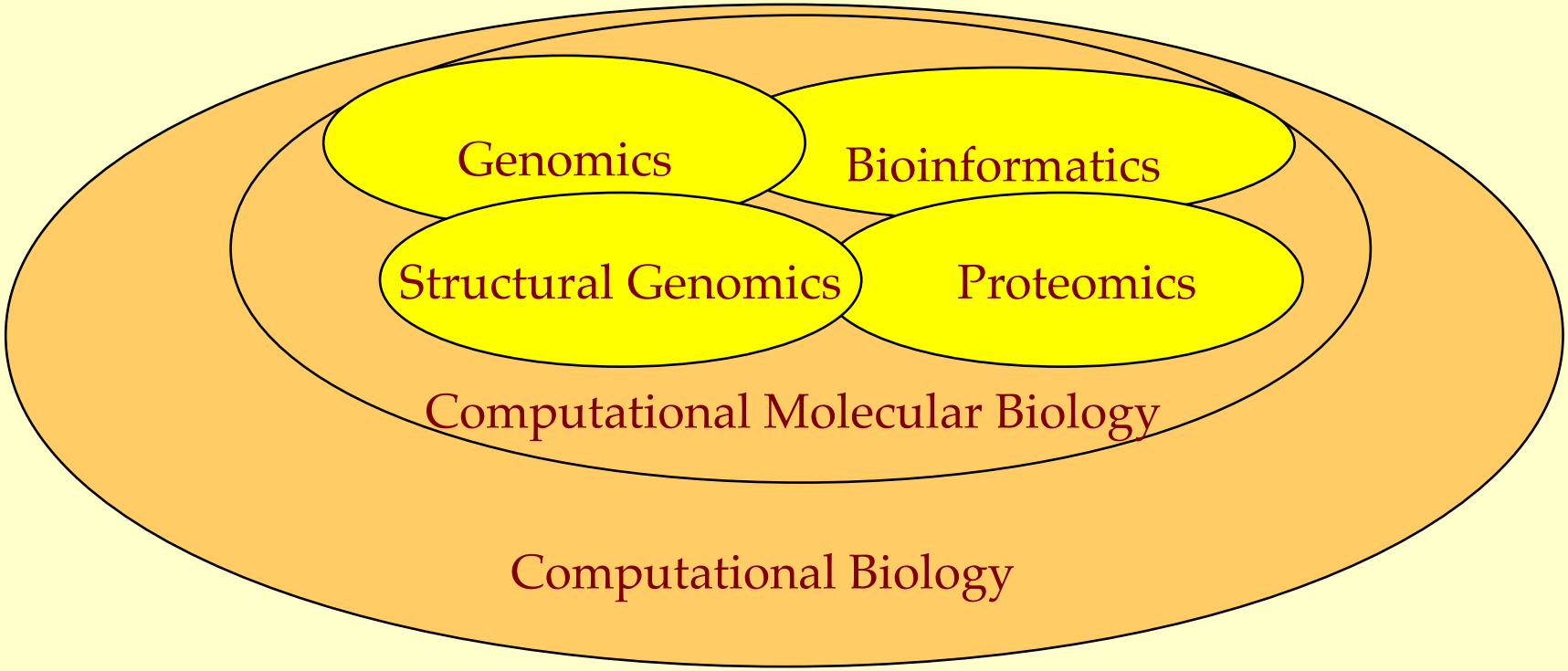
Benjamin Lewin

Genes IX



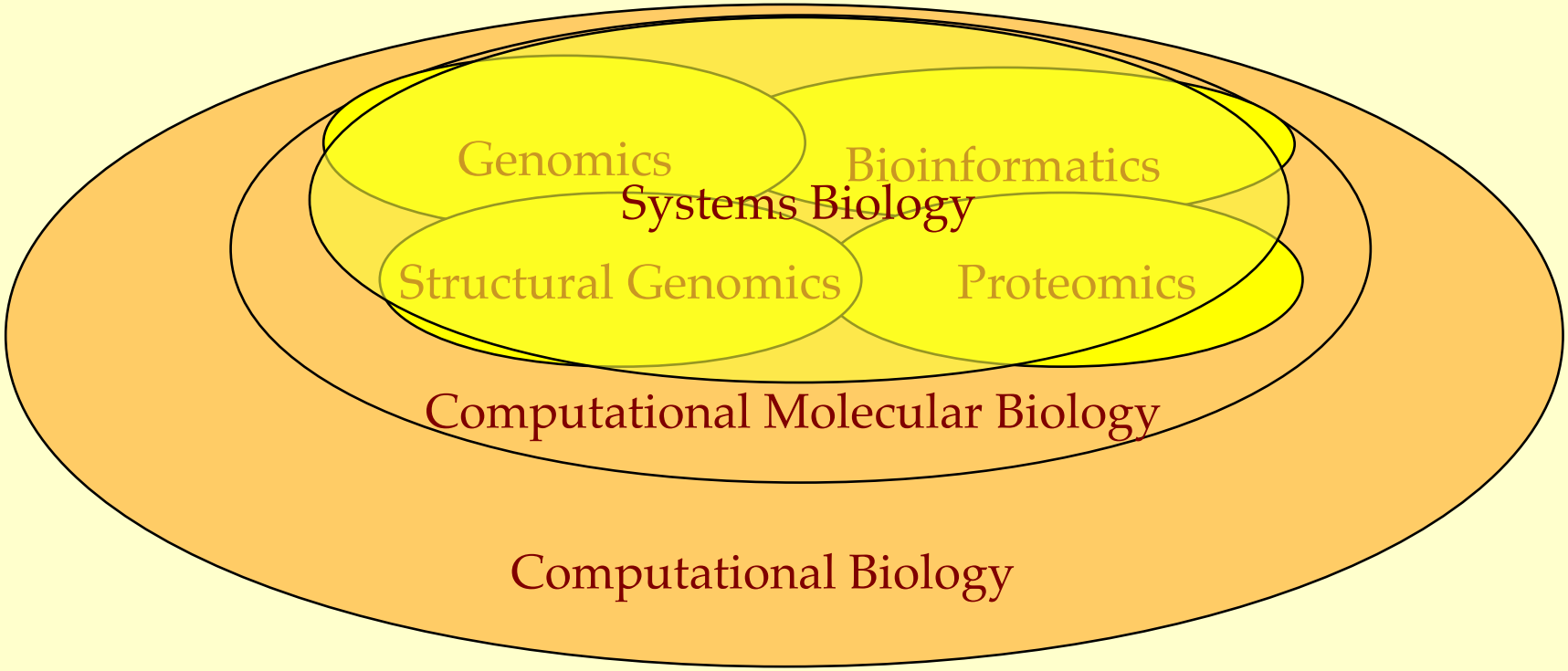
Doug Brutlag 2010

Genomics, Bioinformatics & Computational Biology



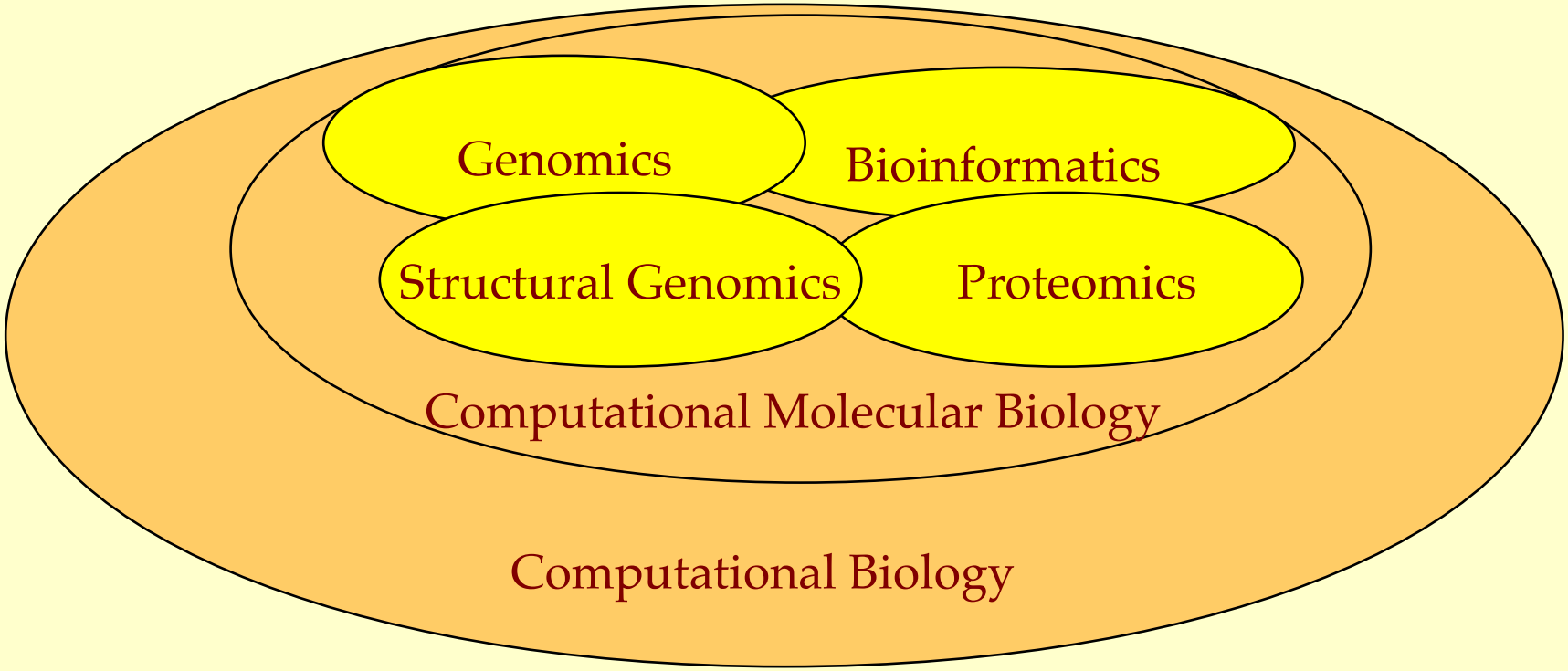
Doug Brutlag 2010

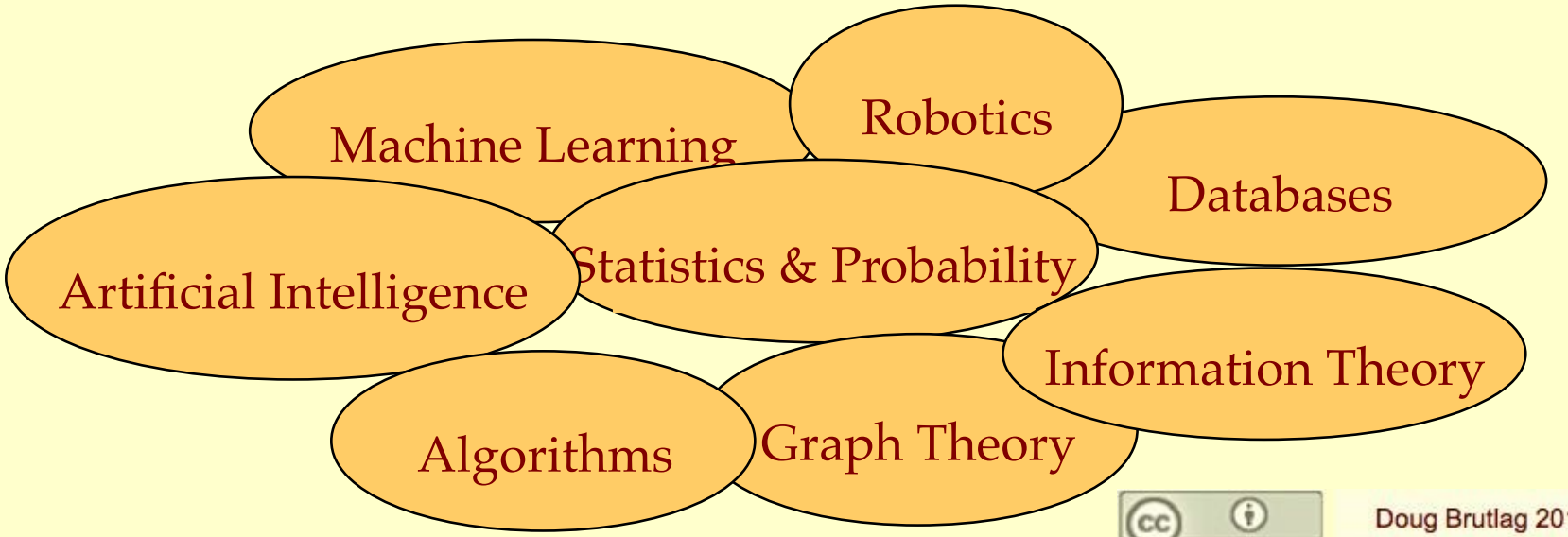
Genomics, Bioinformatics & Computational Biology



Doug Brutlag 2010

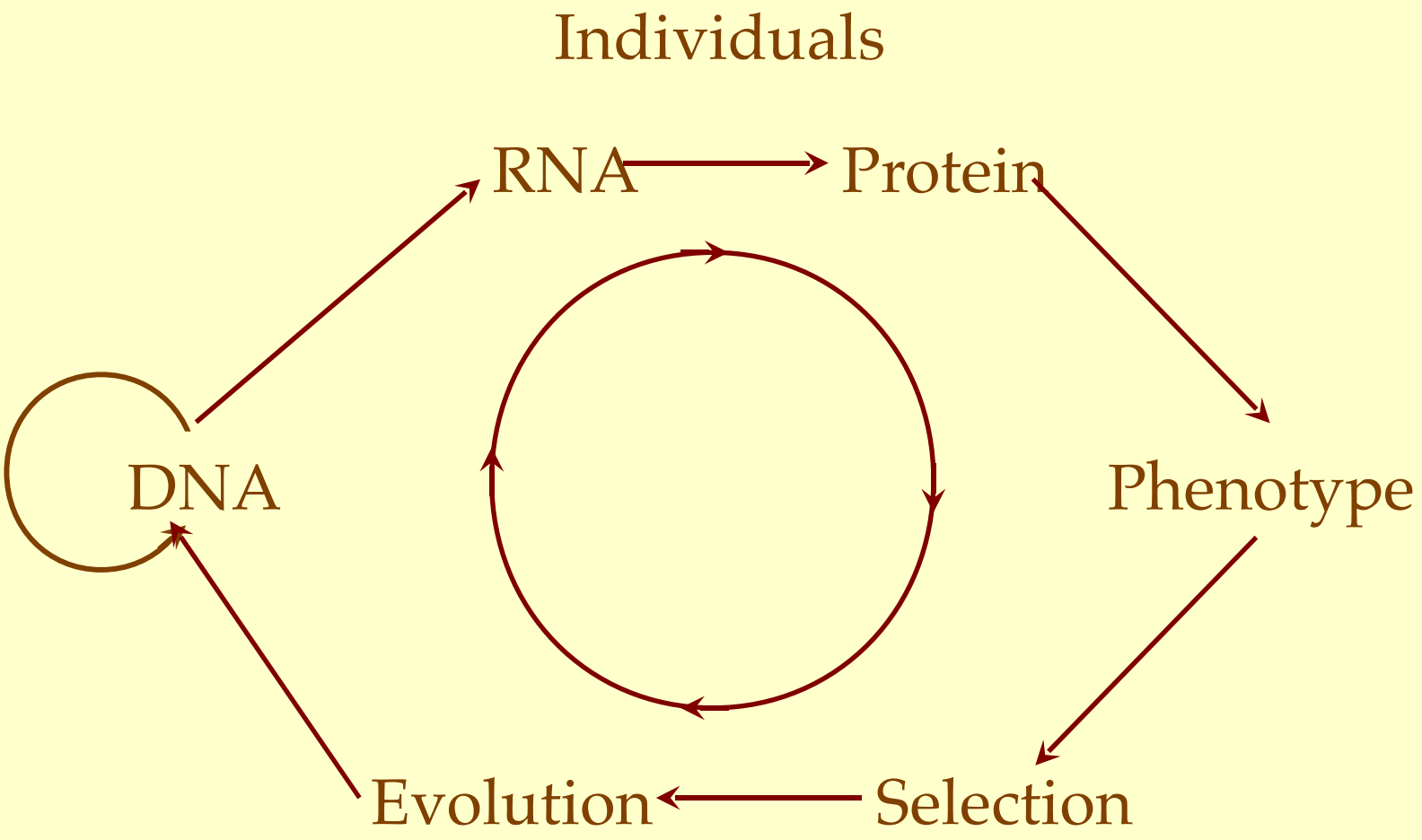
Genomics, Bioinformatics & Computational Biology





Doug Brutlag 2010

What is Bioinformatics?



Biological Information

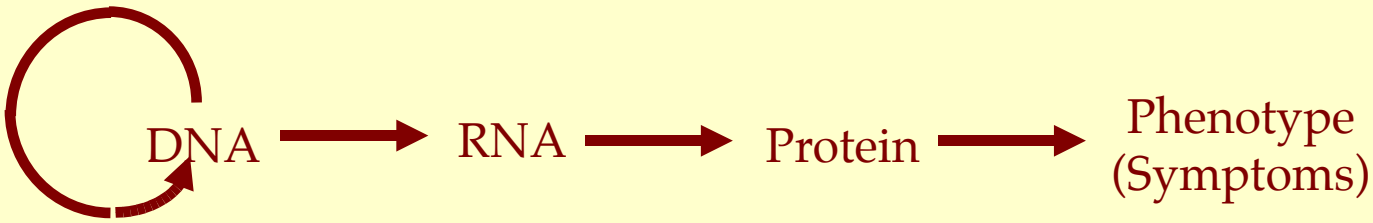
Doug Brutlag 2010

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.

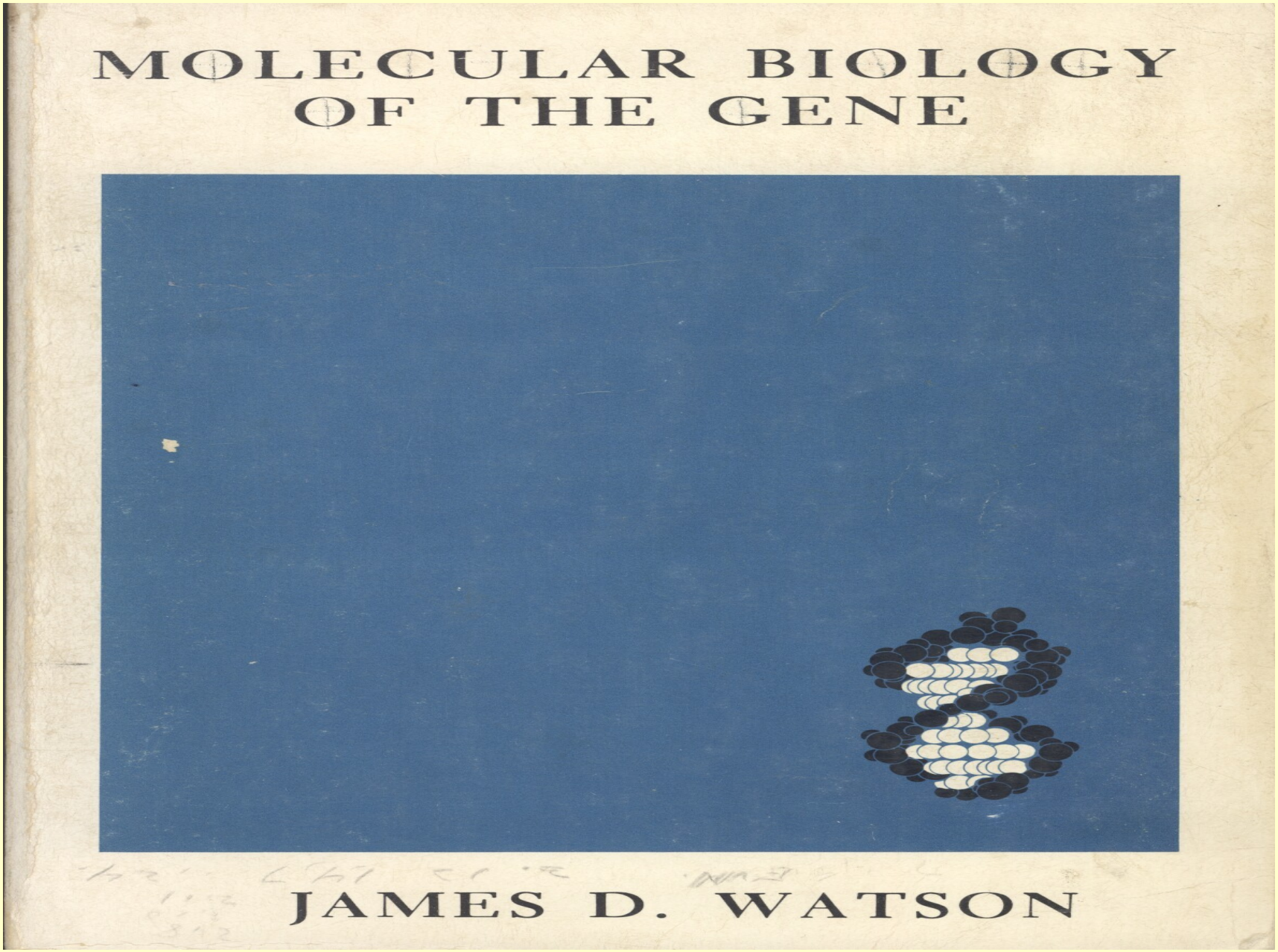
Doug Brutlag 2010

Central Paradigm of Molecular Biology



Doug Brutlag 2010

Molecular Biology of the Gene 1965

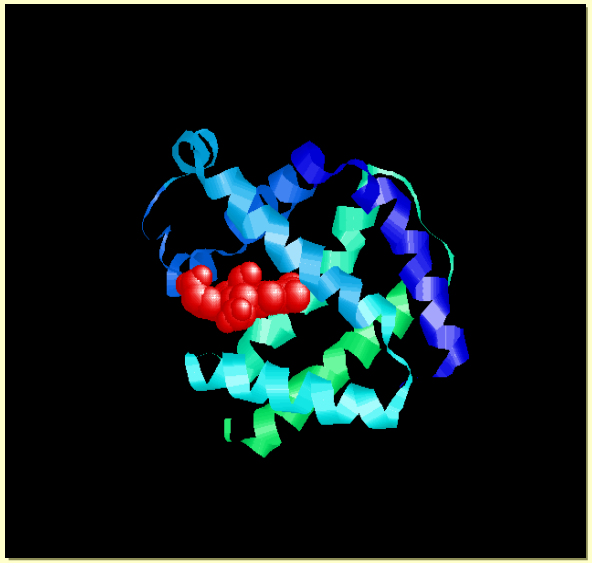


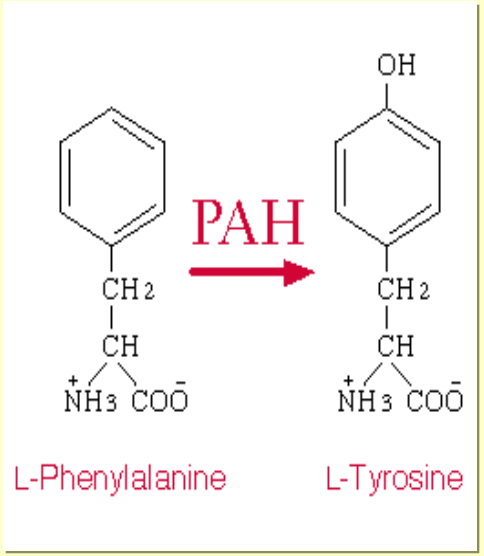
Doug Brutlag 2010

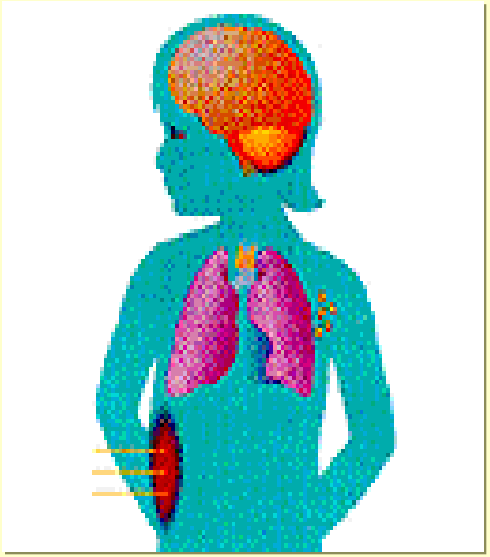
Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Phenotype (Symptoms)

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





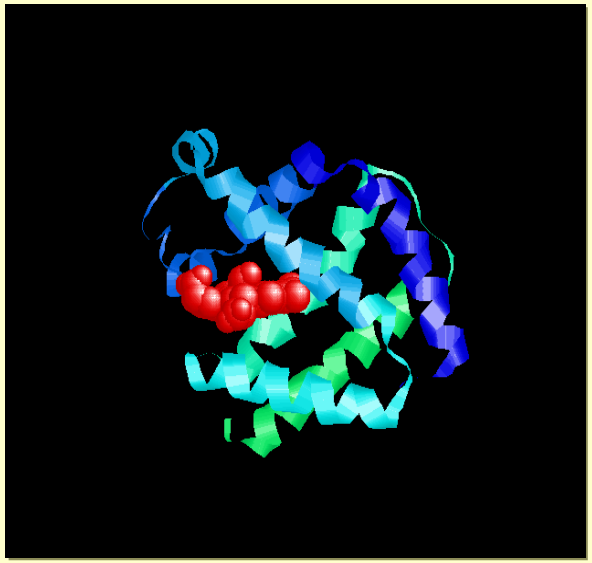


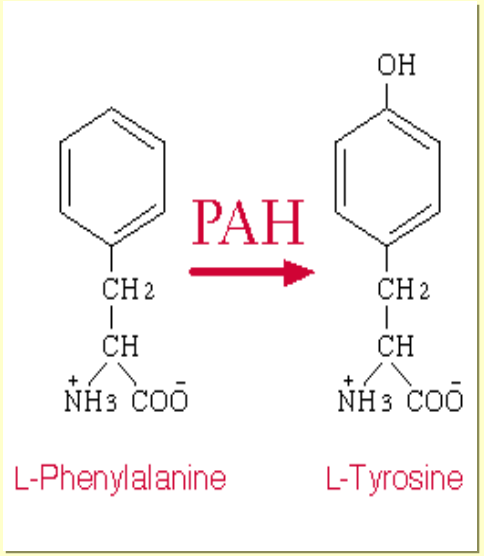
Doug Brutlag 2010

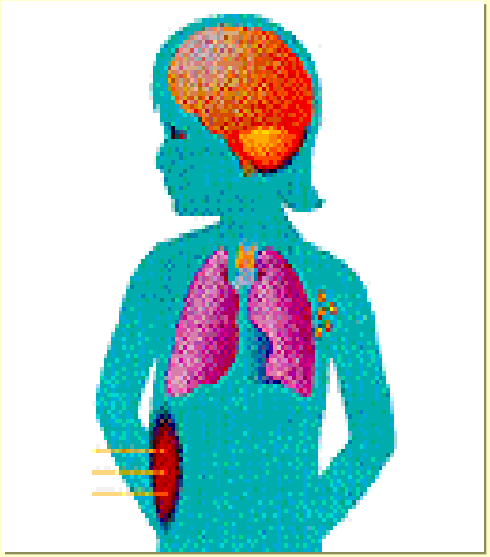
Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Phenotype (Symptoms)

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

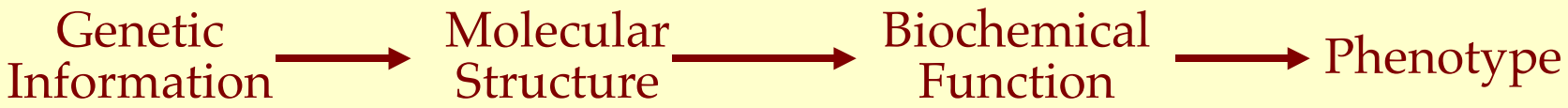






Doug Brutlag 2010

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

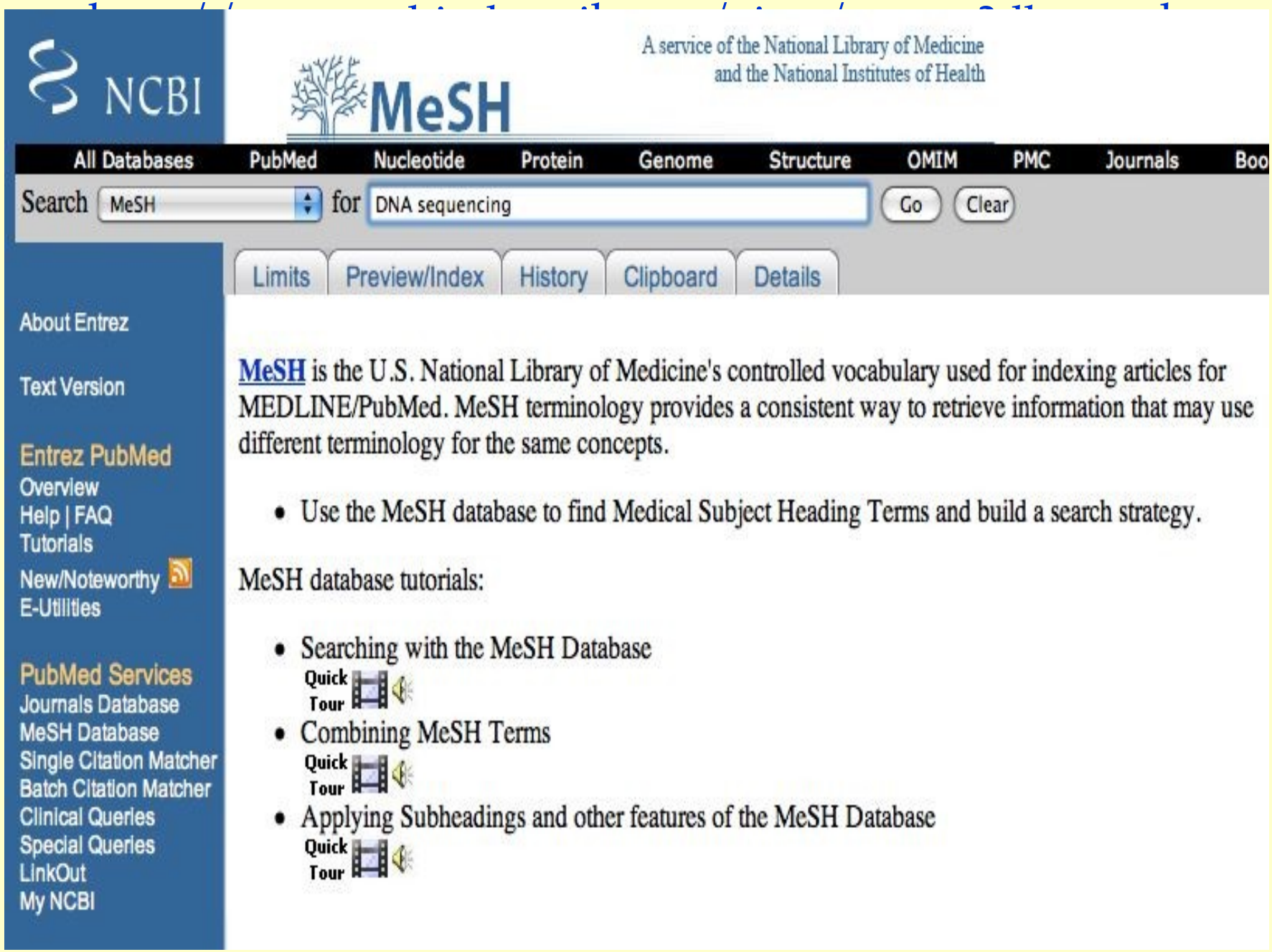
Doug Brutlag 2010

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

Doug Brutlag 2010

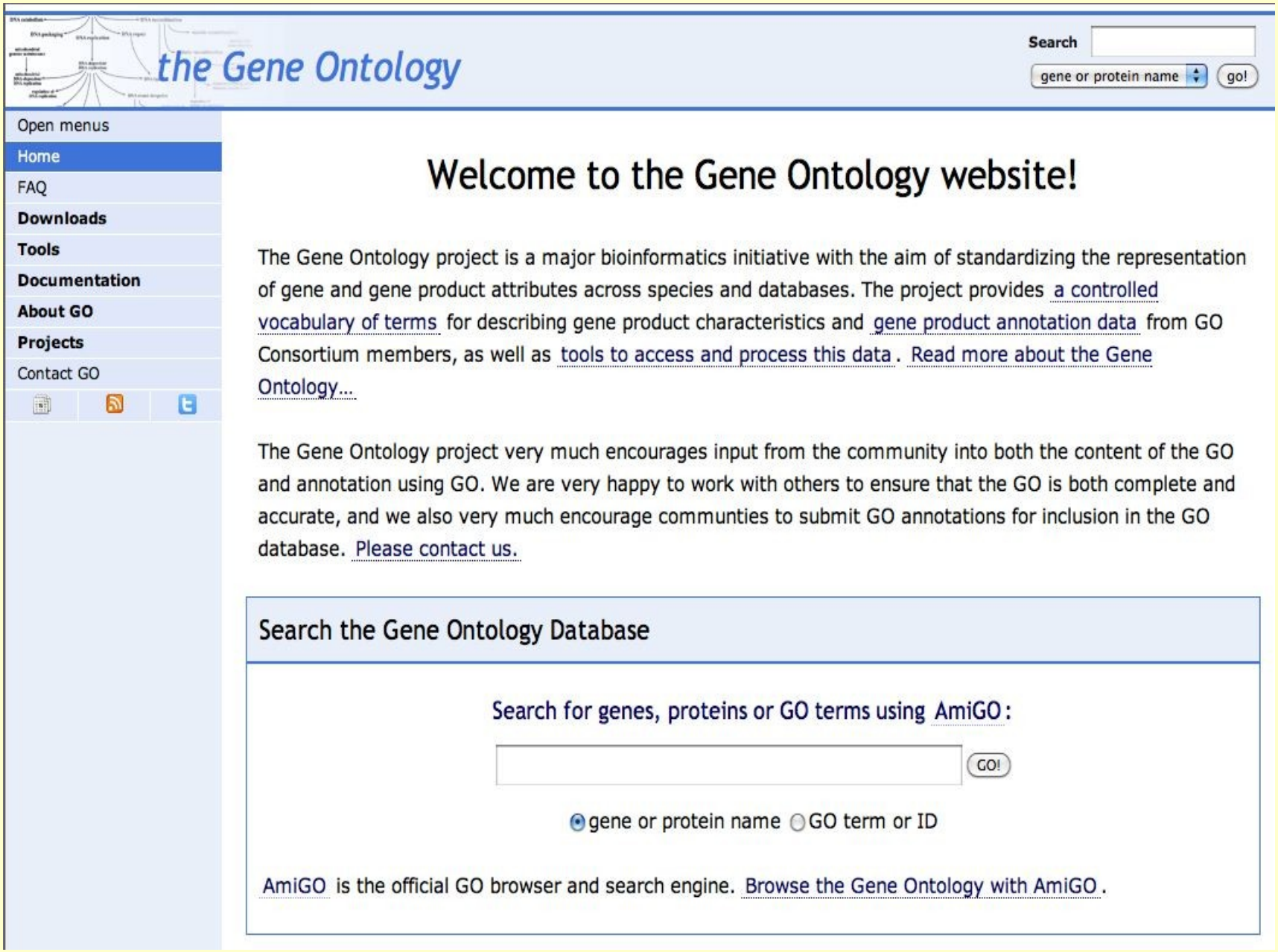
Using A Controlled Vocabulary for Literature Search



Doug Brutlag 2010

Gene Ontology Database

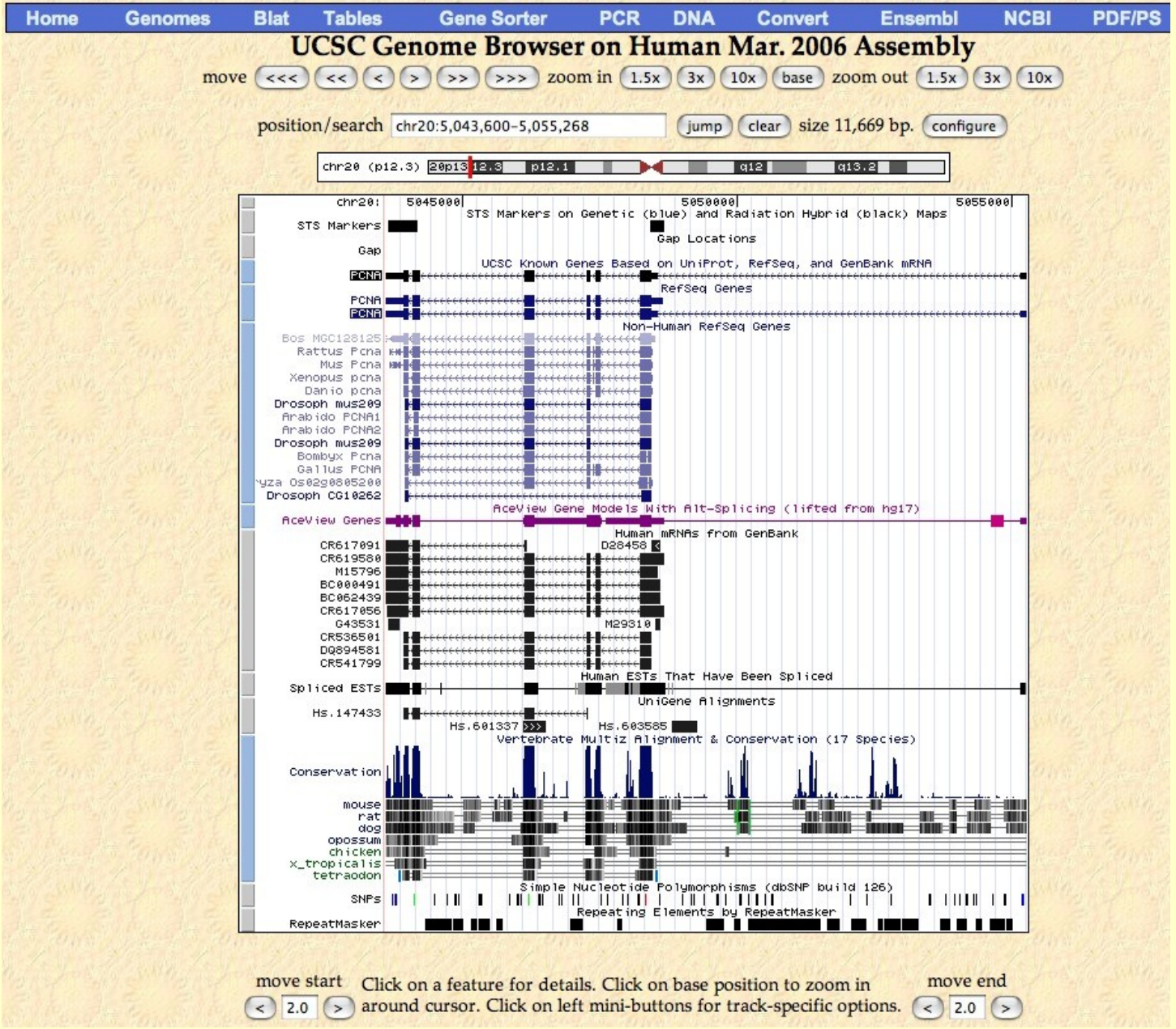
http://www.geneontology.org/



Doug Brutlag 2010

UCSC Genome Browser

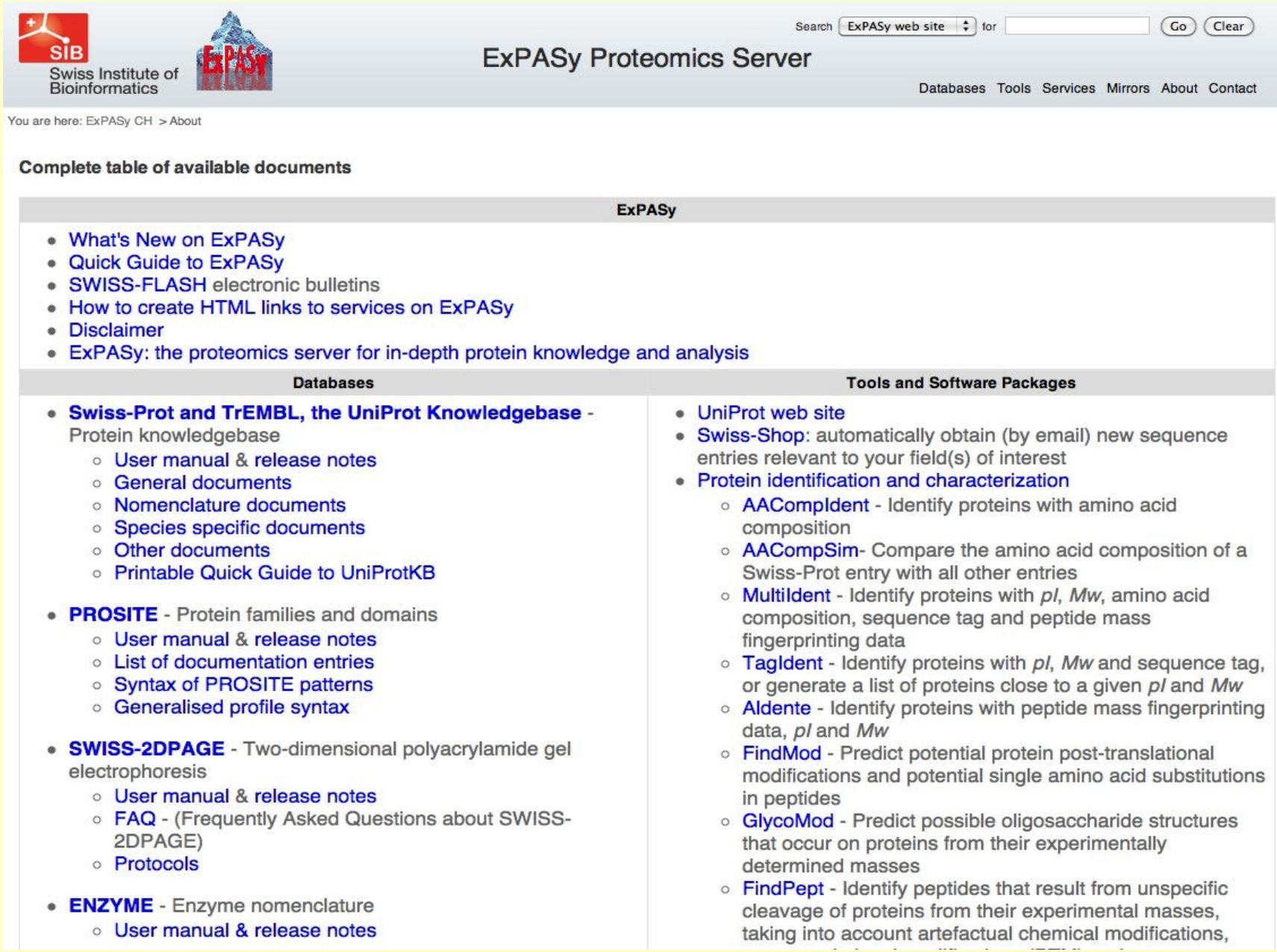
http://genome.ucsc.edu/



Doug Brutlag 2010

ExPASy Proteomics Server

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



Search ExPASy web site for Go Clear

Swiss Institute of Bioinformatics

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

- **Swiss-Prot and TrEMBL, the UniProt Knowledgebase** - Protein knowledgebase
  - User manual & release notes
  - General documents
  - Nomenclature documents
  - Species specific documents
  - Other documents
  - Printable Quick Guide to UniProtKB
- **PROSITE** - Protein families and domains
  - User manual & release notes
  - List of documentation entries
  - Syntax of PROSITE patterns
  - Generalised profile syntax
- **SWISS-2DPAGE** - Two-dimensional polyacrylamide gel electrophoresis
  - User manual & release notes
  - FAQ - (Frequently Asked Questions about SWISS-2DPAGE)
  - Protocols
- **ENZYME** - Enzyme nomenclature
  - User manual & release notes

Tools and Software Packages

- UniProt web site
- Swiss-Shop: automatically obtain (by email) new sequence entries relevant to your field(s) of interest
- Protein identification and characterization
  - AACompIdent - Identify proteins with amino acid composition
  - AACompSim- Compare the amino acid composition of a Swiss-Prot entry with all other entries
  - MultiIdent - Identify proteins with *pI*, *Mw*, amino acid composition, sequence tag and peptide mass fingerprinting data
  - TagIdent - Identify proteins with *pI*, *Mw* and sequence tag, or generate a list of proteins close to a given *pI* and *Mw*
  - Aldente - Identify proteins with peptide mass fingerprinting data, *pI* and *Mw*
  - FindMod - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides
  - GlycoMod - Predict possible oligosaccharide structures that occur on proteins from their experimentally determined masses
  - FindPept - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications,

Doug Brutlag 2010

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

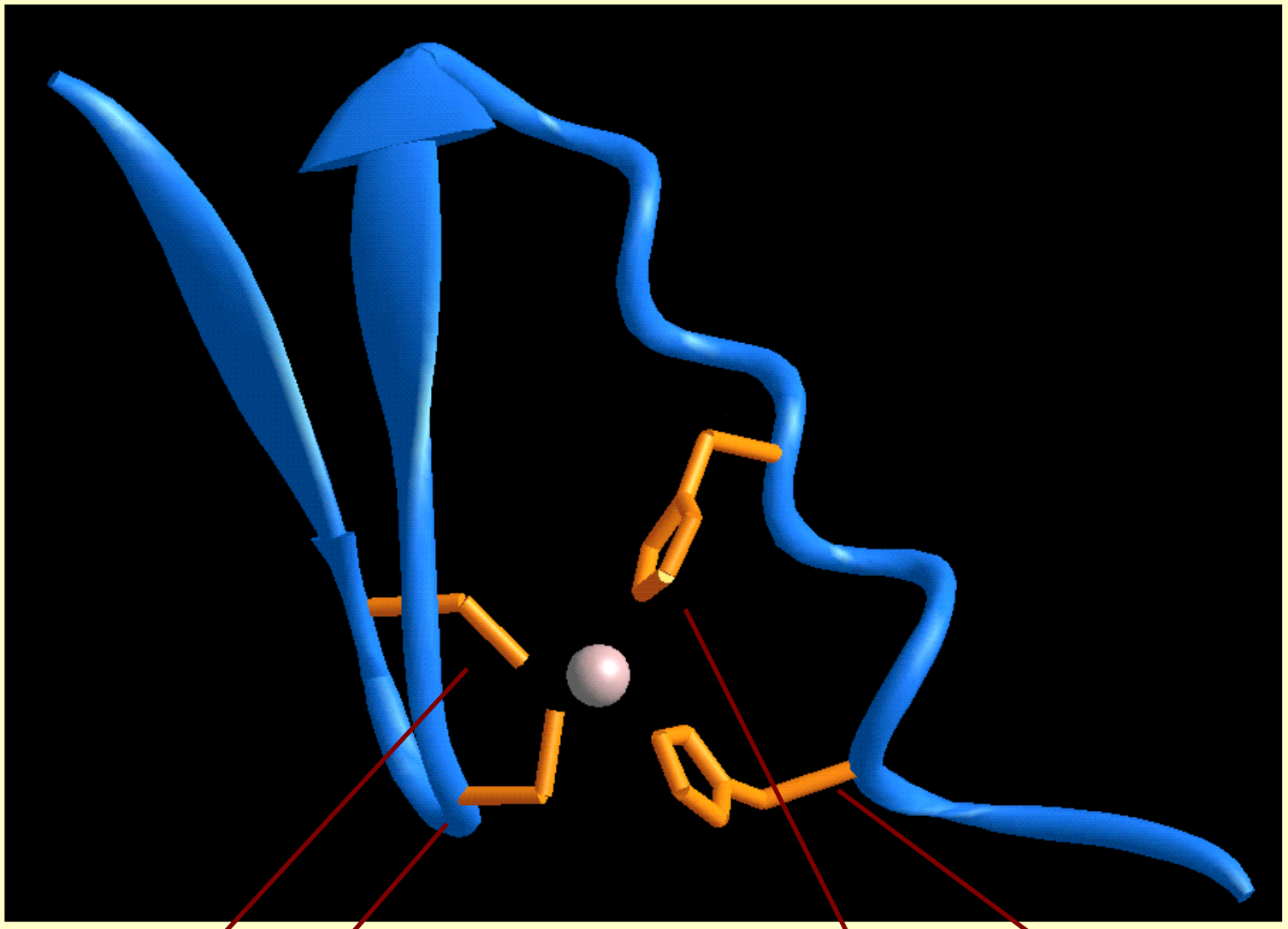
Sequences of Common Structure or Function

Sequence Similarity

```
                10        20        30        40                  50
Query VLSPADKTNVKAAWGKVGAHAGEVGAEALERMFLSFPTTKTYFPHF------DLSHGS
      |:| :|: | |:|||||     | |:||| |: : :|:|  :|  |       |:  |
Match HLTPEEKSAVTALWGKV--NVDEYGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                10        20        30        40        50
```

Doug Brutlag 2010

A Typical Motif:
Zinc Finger DNA Binding Motif



Doug Brutlag 2010

Inferring Biological Function from Protein Sequence

Weight Matrices or Position-Specific Scoring Matrices

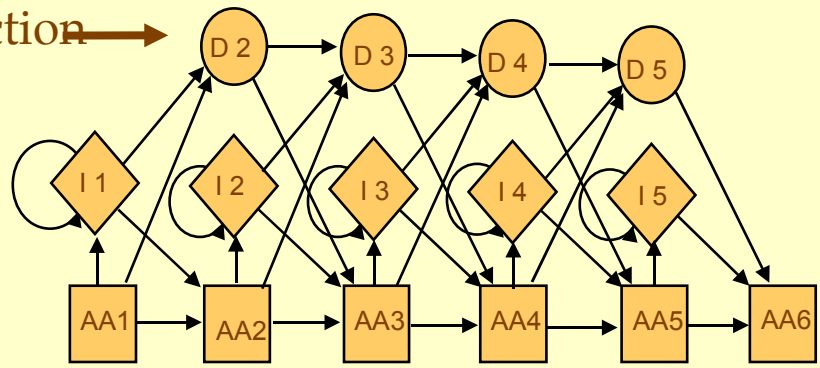
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	0	11	1	2	10	0	4	9	3	0	16
L	16	1	17	0	1	31	0	3	11	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 x {2,4} C x {12} H x {3,5} H

Sequences of Common Structure or Function

Profiles, PSI-BLAST Hidden Markov Models



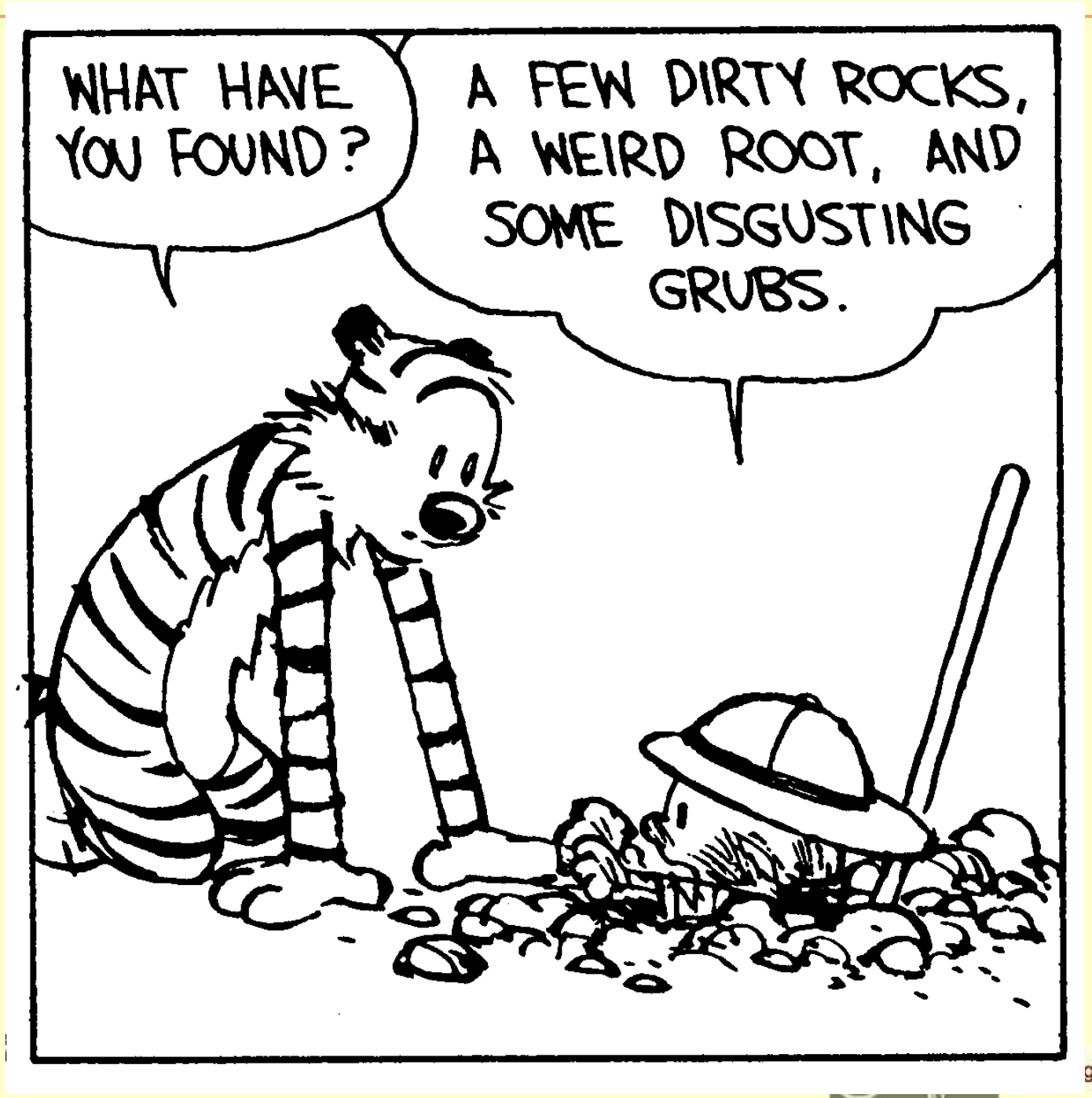
Sequence Similarity

```
                10        20        30        40                  50
Query VLSPADKTNVKAAWGKVGAHAGEVGAEALERMFLSFPTTKTYFPHF------DLSHGS
      |:| :|: | |:|||||     | |:||| |: : :|:|  :|  |       |:  |
Match HLTPEEKSAVTALWGKV--NVDEYGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                10        20        30        40        50
```

Doug Brutlag 2010

Buried Treasure

Buried Treasure



g Brutlag 2010

Buried Treasure

Brutlag 2010

Clustal Globin Alignment

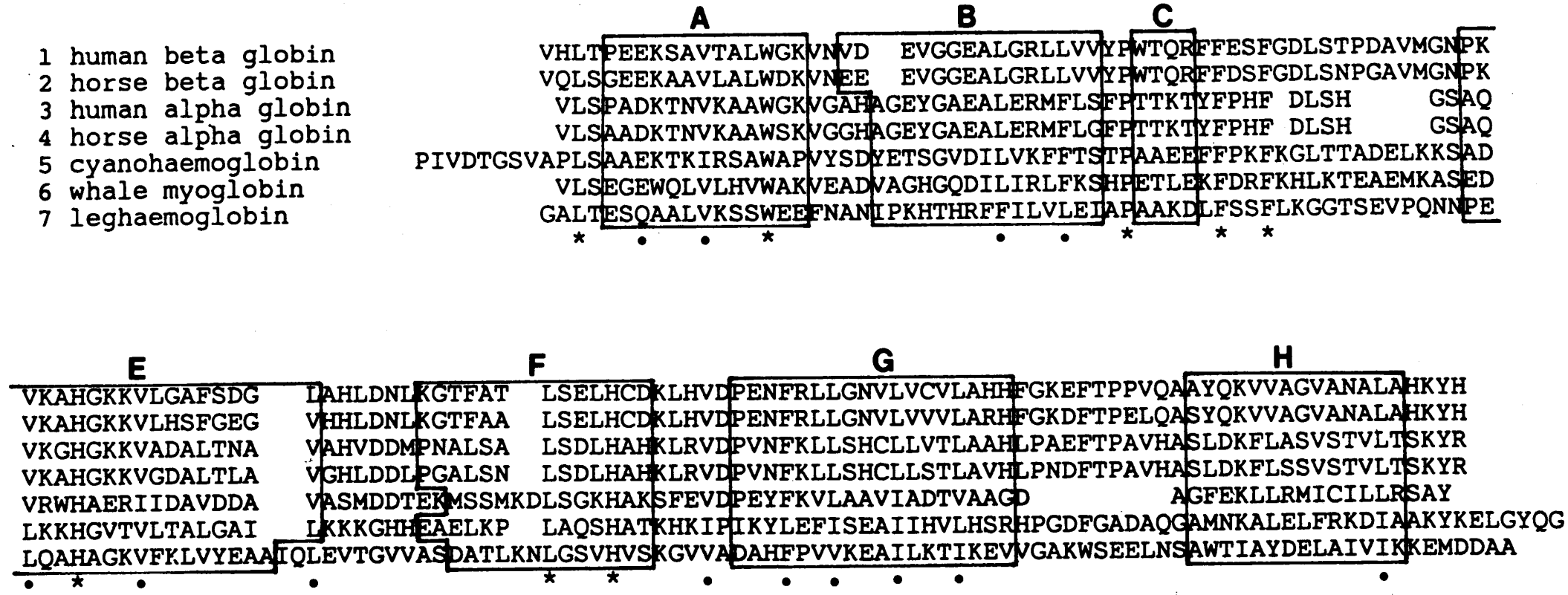
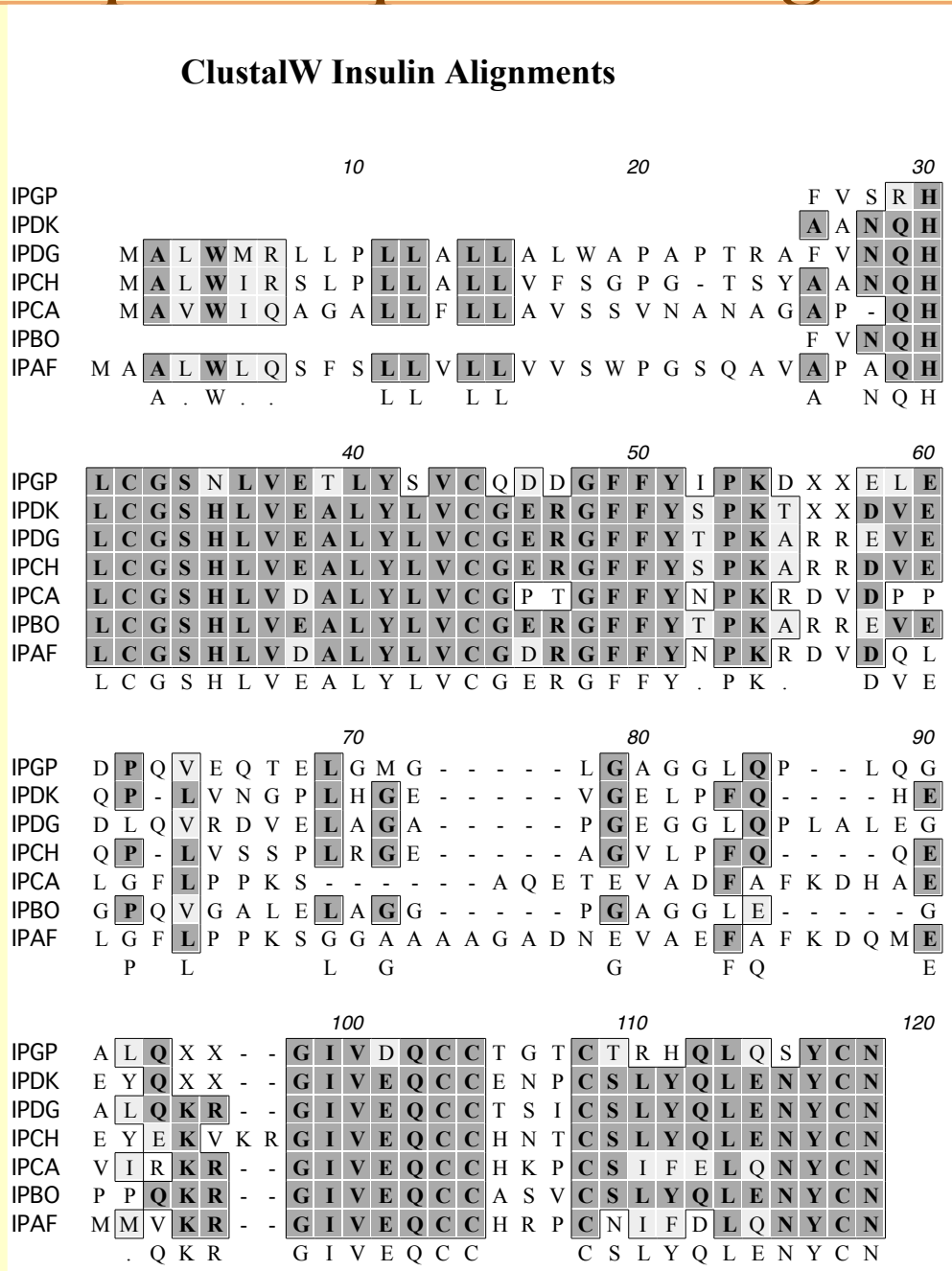


Fig. 3. CLUSTAL-produced multiple alignment of seven globin sequences taken from Lesk and Chothia (1980) (see RESULTS, section

Doug Brutlag 2010

Consensus Sequence From a Multiple Sequence Alignment



TimeLogic®
biocomputing solutions

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

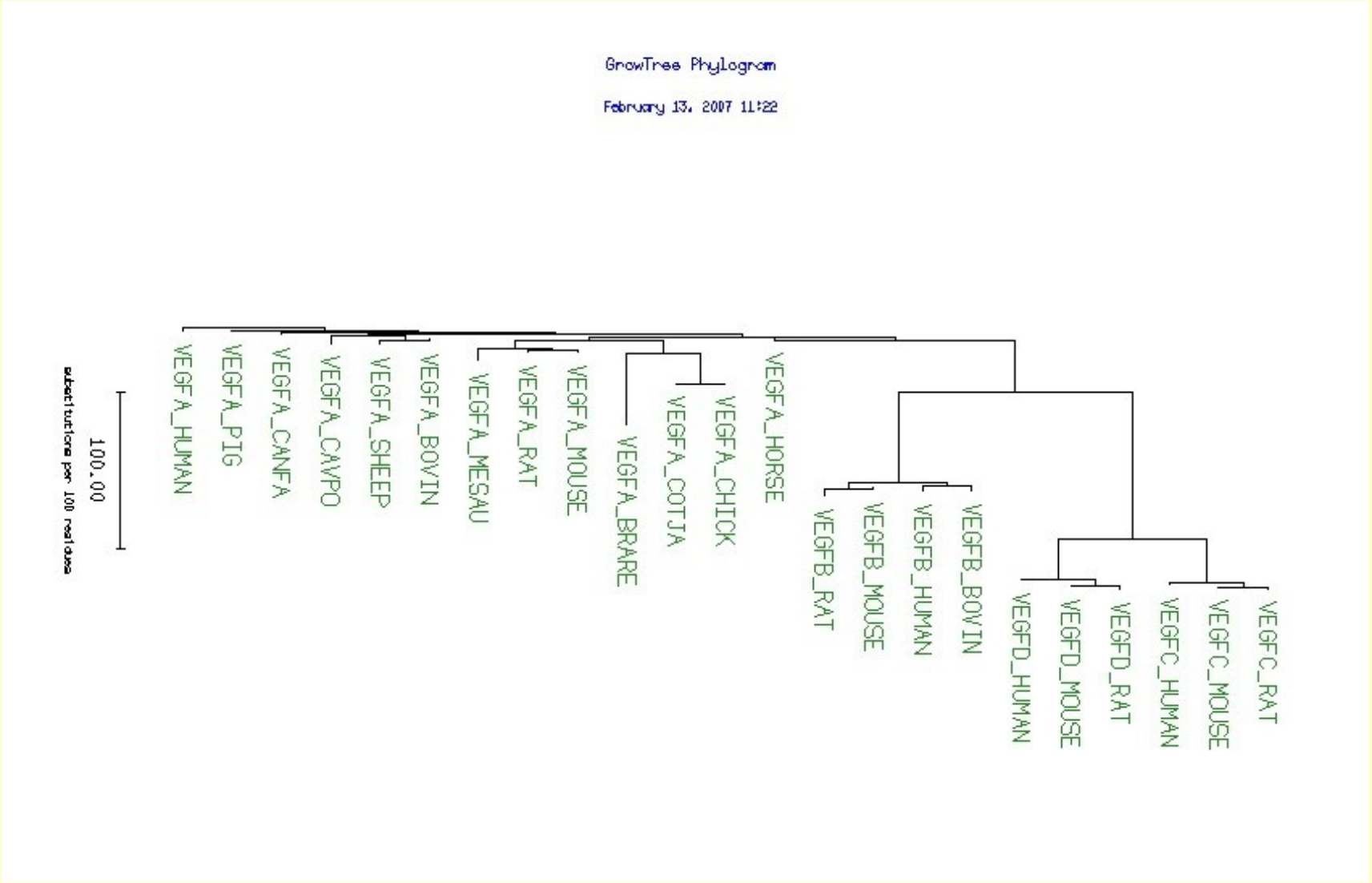
Download (For Internet Explorer Browsers, Right Click on Hyperlink and Select "Save Target As...")

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       *       *
```

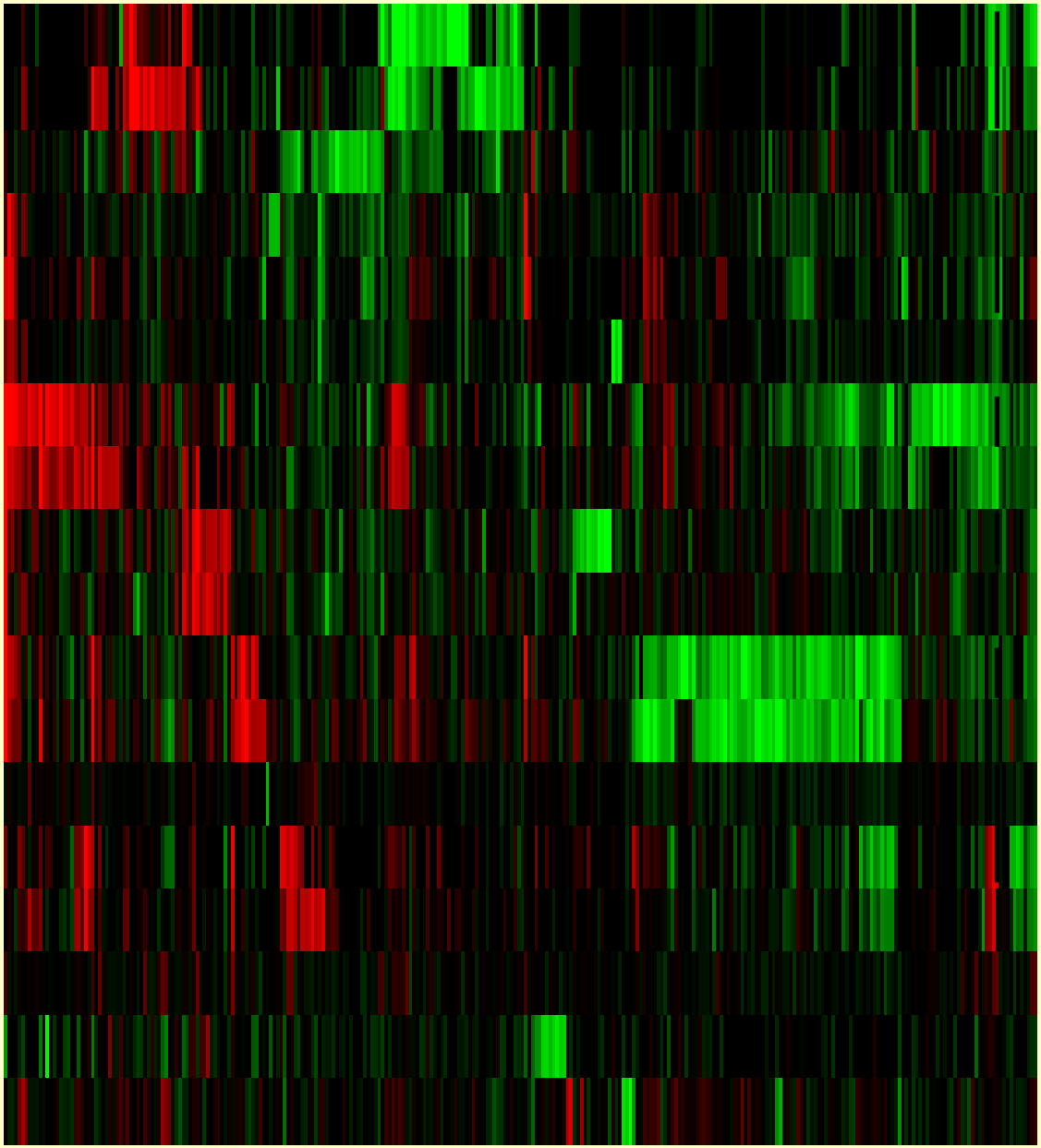
Doug Brutlag 2010

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

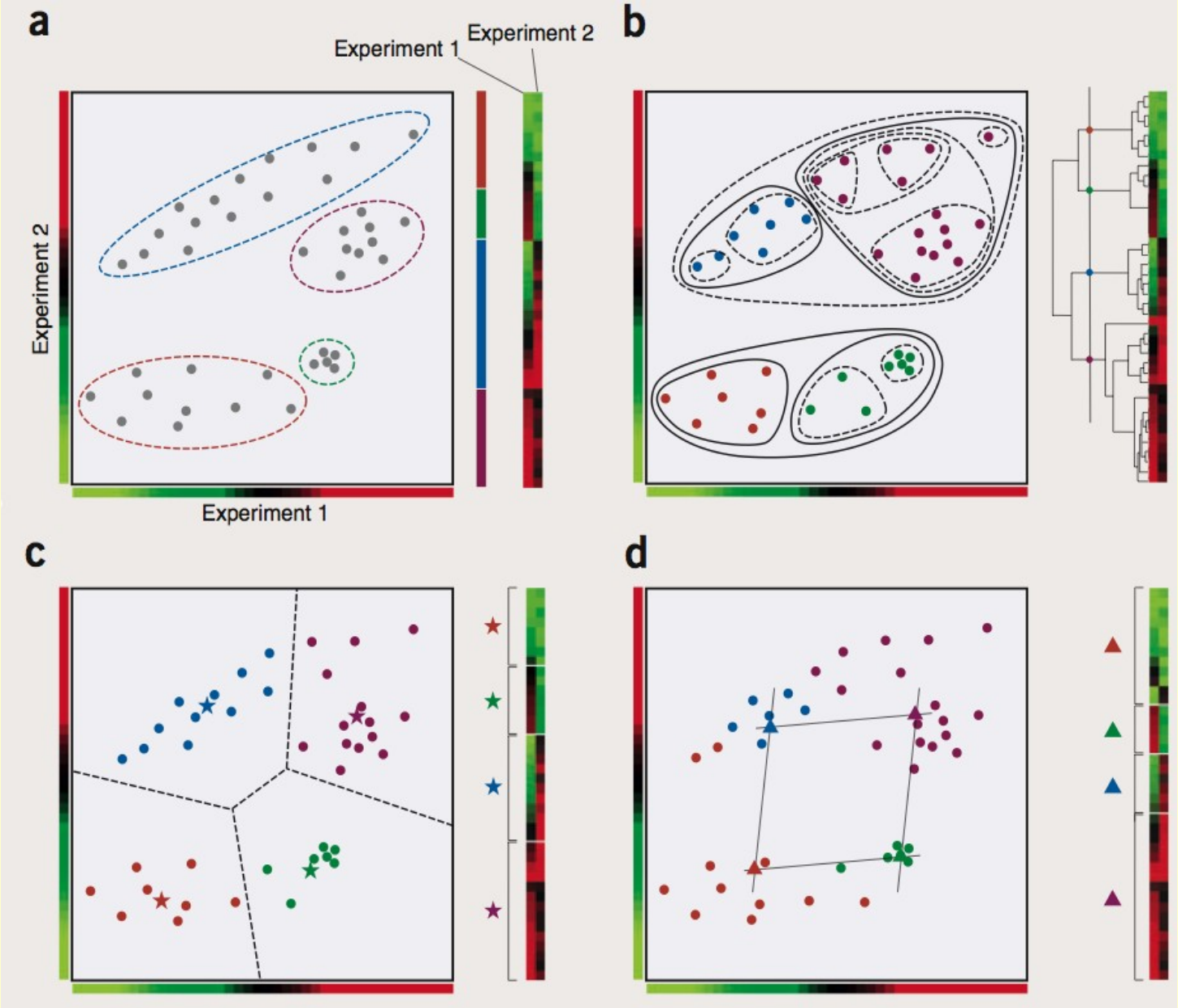
Polio Infection

Monocytes Signaling IL4

Hormone

Doug Brutlag 2010

Clustering Gene Expression Profiles: Comparison of Methods



D'haeseleer P (2005). *Nat Biotechnol.* 23,1499-501.

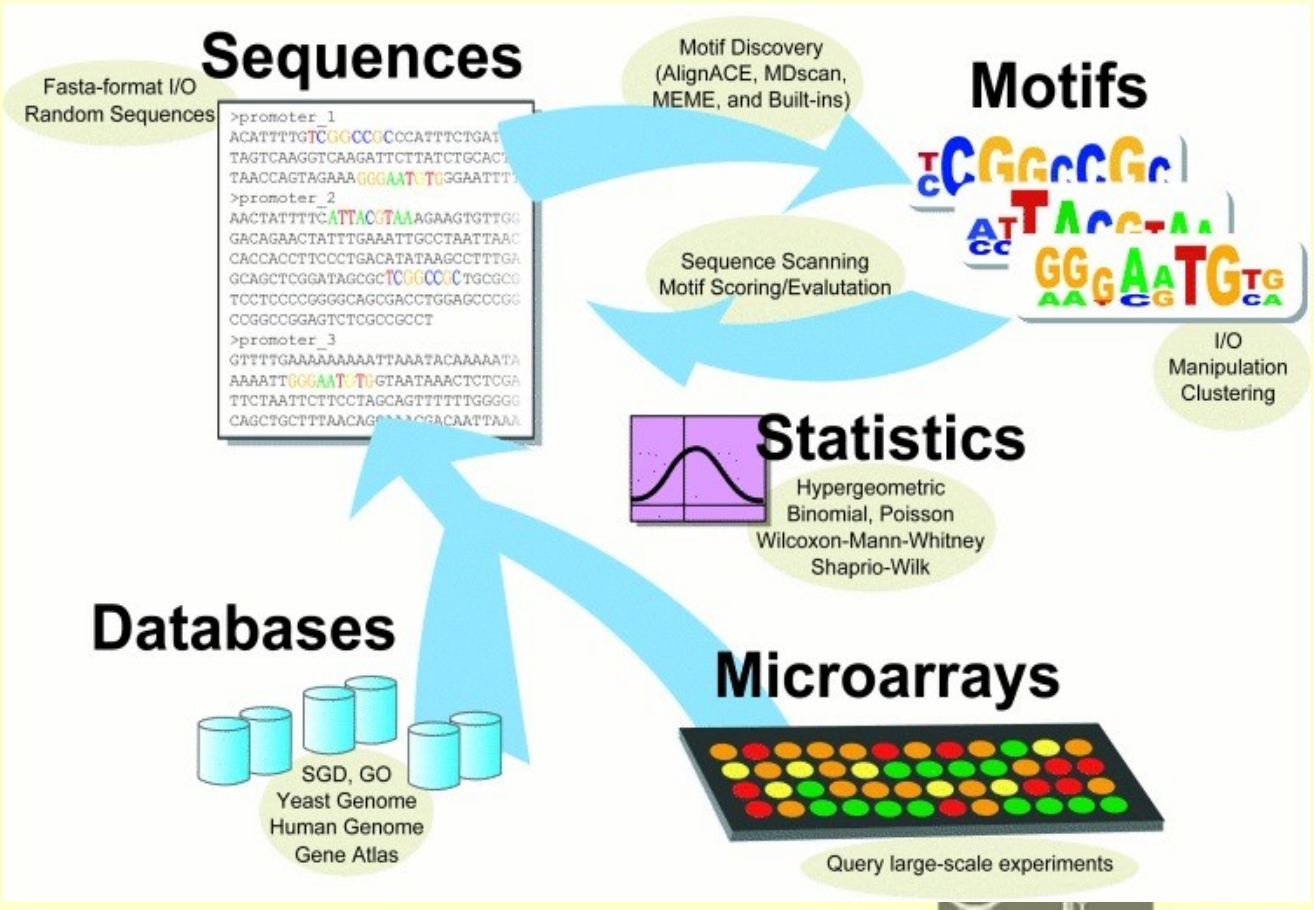
The Fraenkel Lab
TAMO -- Tools for the Analysis of MOtifs
Download and Support for the TAMO package

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 commerical licensing.



Finding Transcription Factor Binding Sites

Upstream Regions

Co-expressed

GATGGCTGCACCACGTGTATGC...ACG

CACATCGCATCACGTGACCAGT...GAC

GCCTCGCACGTGGTGGTACAGT...AAC

TCTCGTTAGGACCATCACGTGA...ACA

CGCTAGCCCACGTGGATCTTGA...AGAATGACTGGC

Pho 5

Pho 8

Pho 81

Pho 84

Pho …

Doug Brutlag 2010

Finding Transcription Factor Binding Sites

Upstream Regions — Co-expressed Genes

GATGGCTGCACCACGTGTATGC...ACGATGTCTCGC

CACATCGCATCACGTGACCAGT...GACATGGACGGC

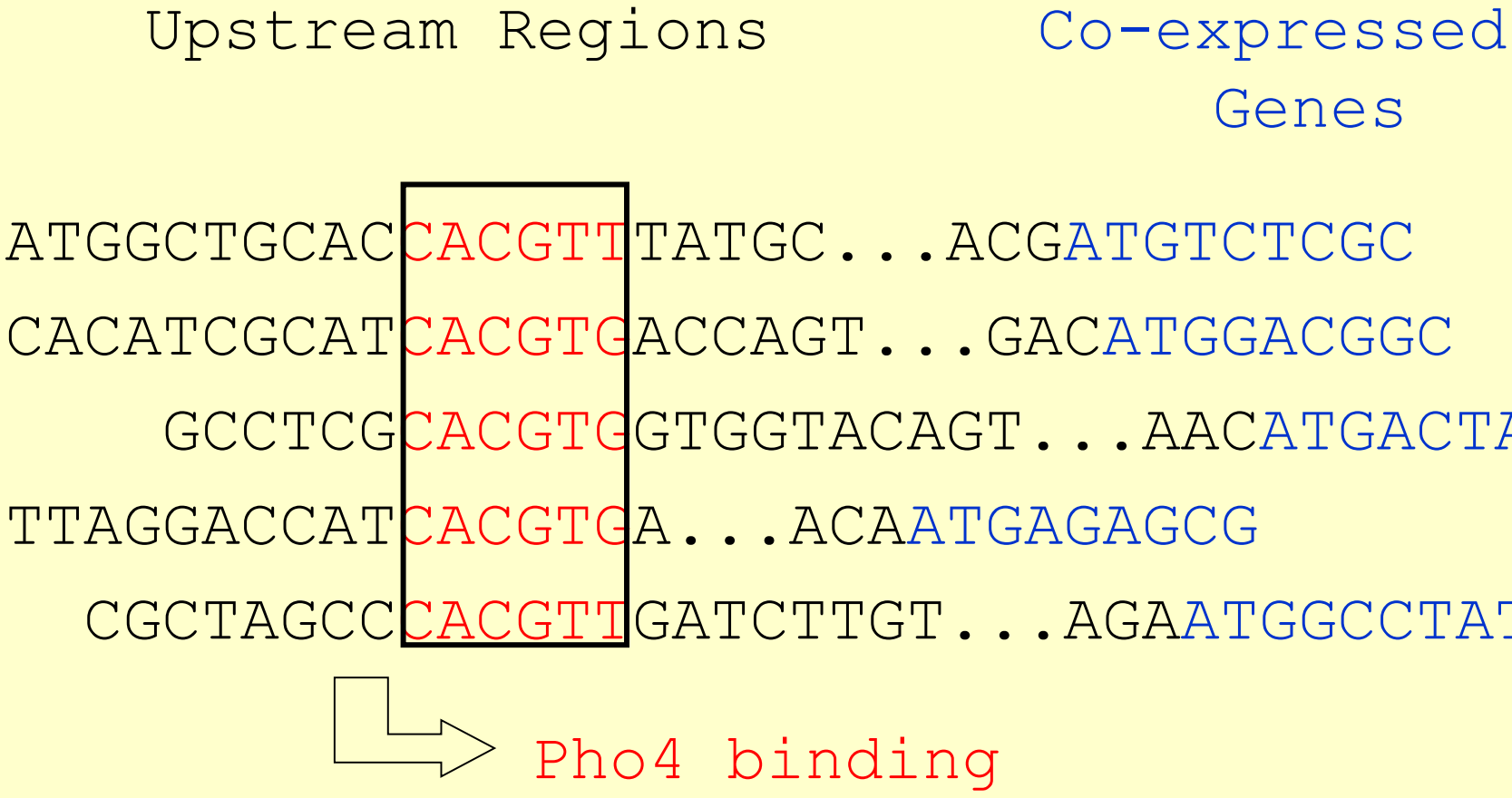
GCCTCGCACGTGGTGGTACAGT...AACATGACTAAA

TCTCGTTAGGACCATCACGTGA...ACAATGAGAGCG

CGCTAGCCCACGTGGATCTTGT...AGAATGGCCTAT

Doug Brutlag 2010

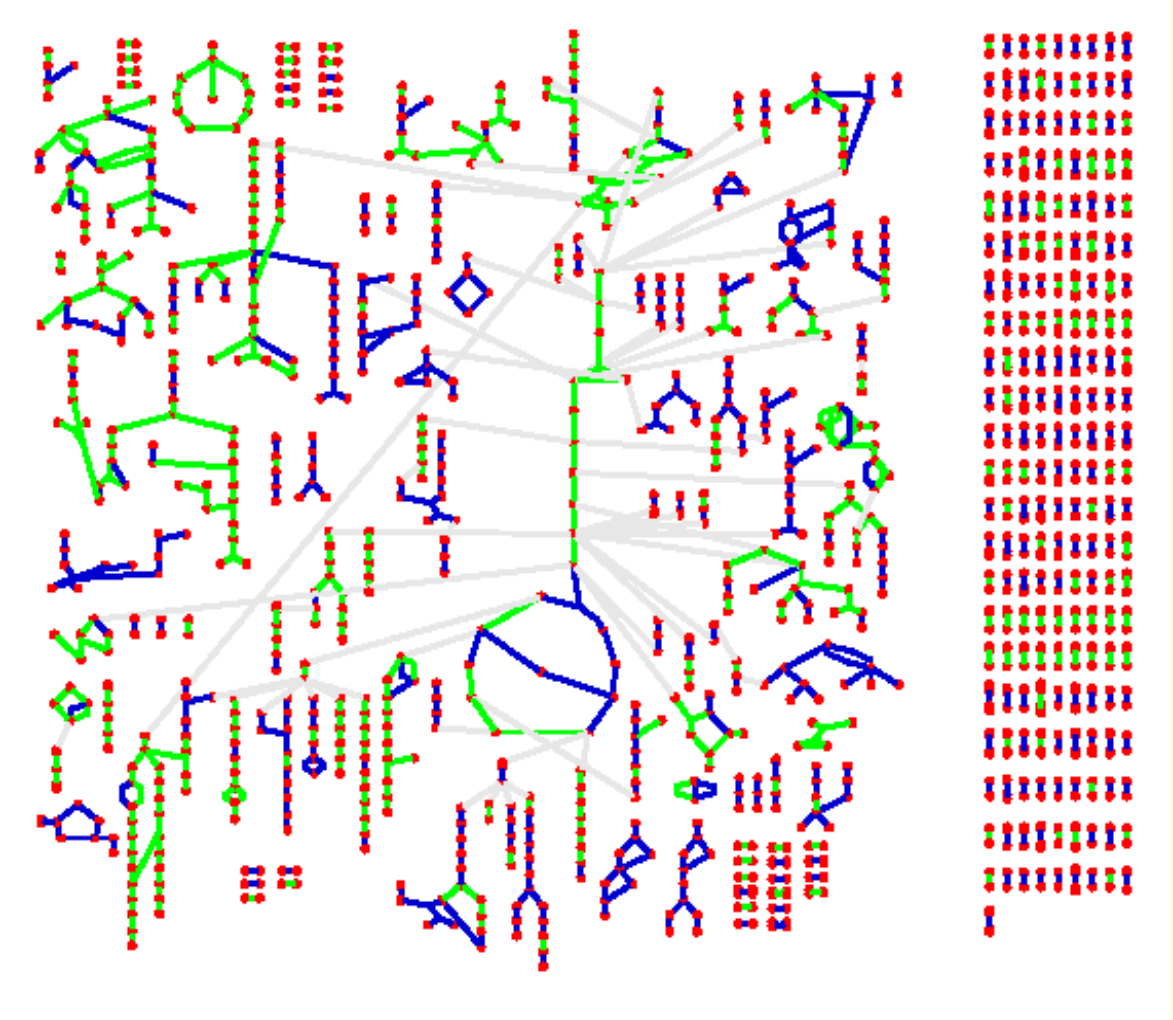
Finding Transcription Factor Binding Sites



Doug Brutlag 2010

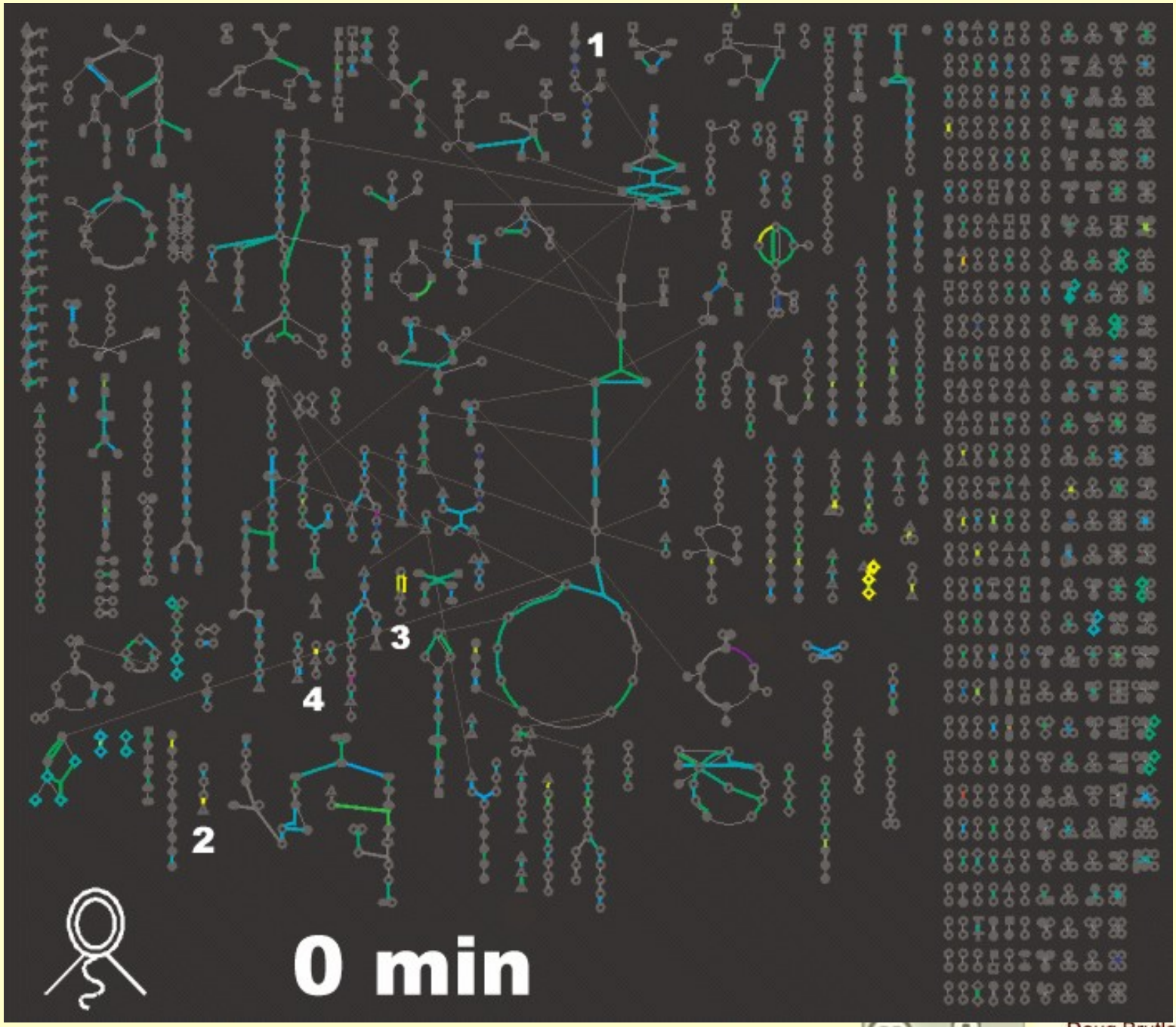
Metabolic Networks: BioCyc

http://biocyc.org/



Doug Brutlag 2010

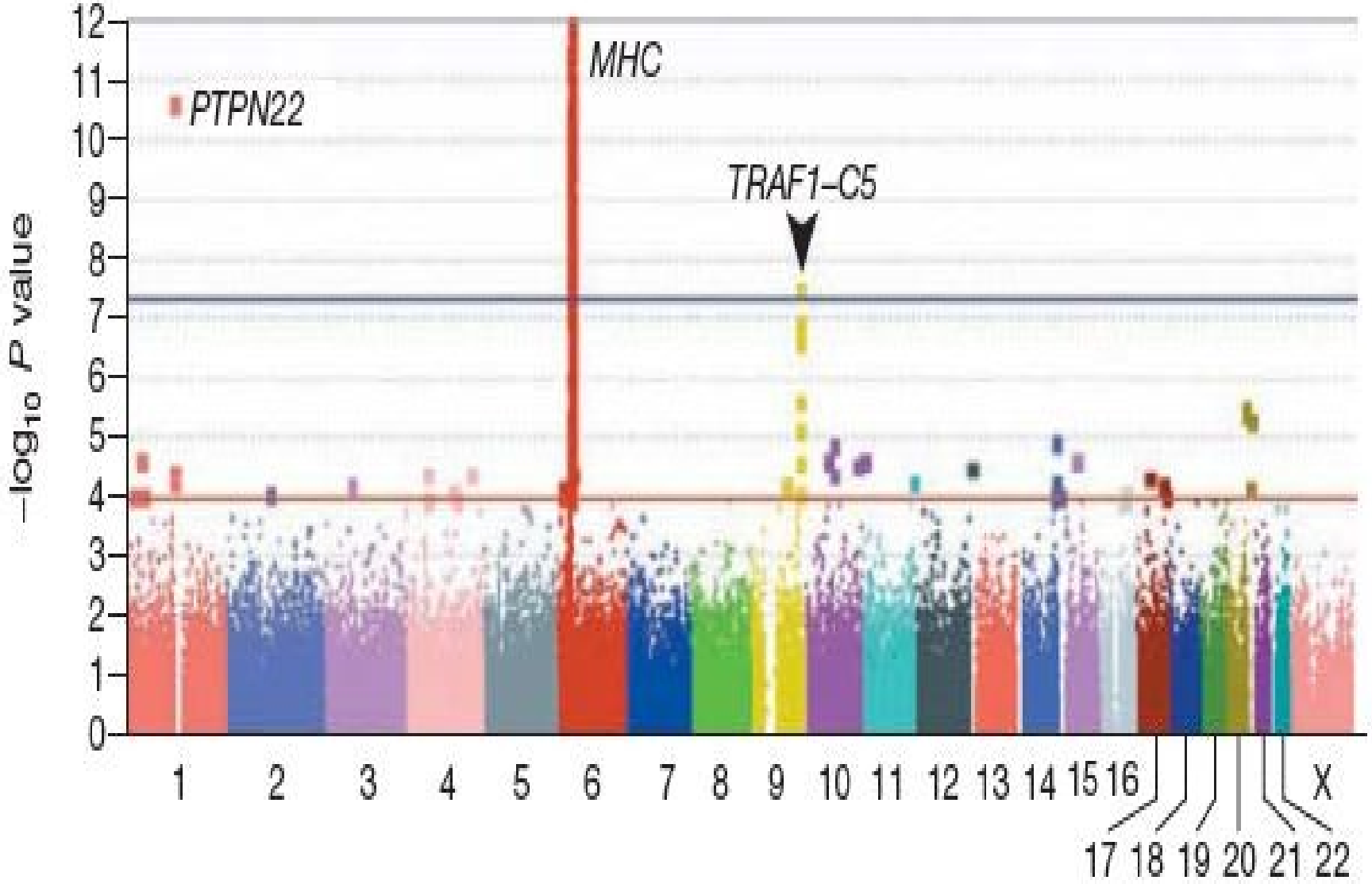
C. crescentus Cell Cycle Gene Expression



Doug Brutlag 2010

Genome Wide Associations in

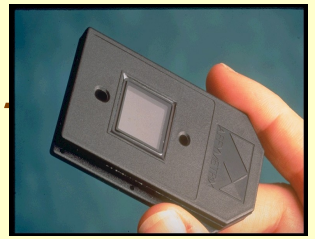
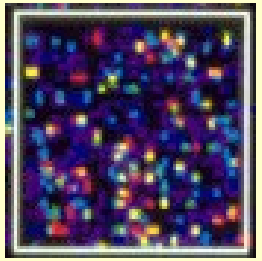
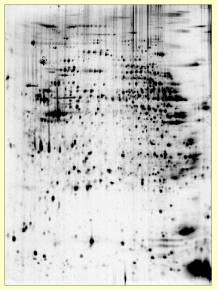
Figure 3. Genome-wide Association Findings in Rheumatoid Arthritis



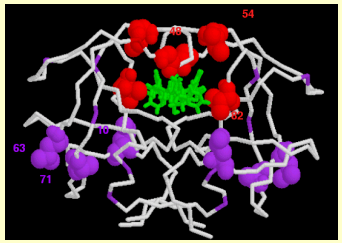
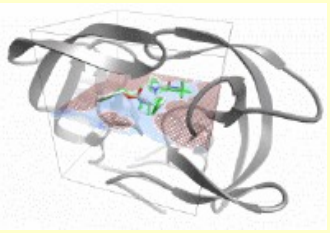
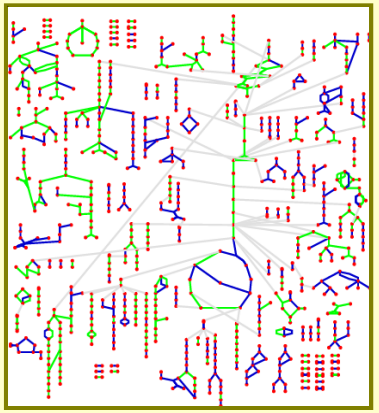
Pearson, T. A. et al. JAMA 2008;299:1335-1344

Leveraging Genomic Information in Medicine

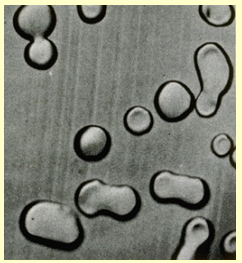
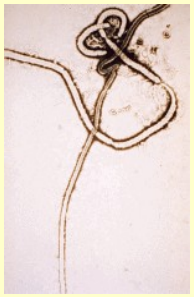
- Novel Diagnostics
  - Microchips & Microarrays
  - Gene Expression - RNA
  - Proteomics - Protein

- 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

Doug Brutlag 2010