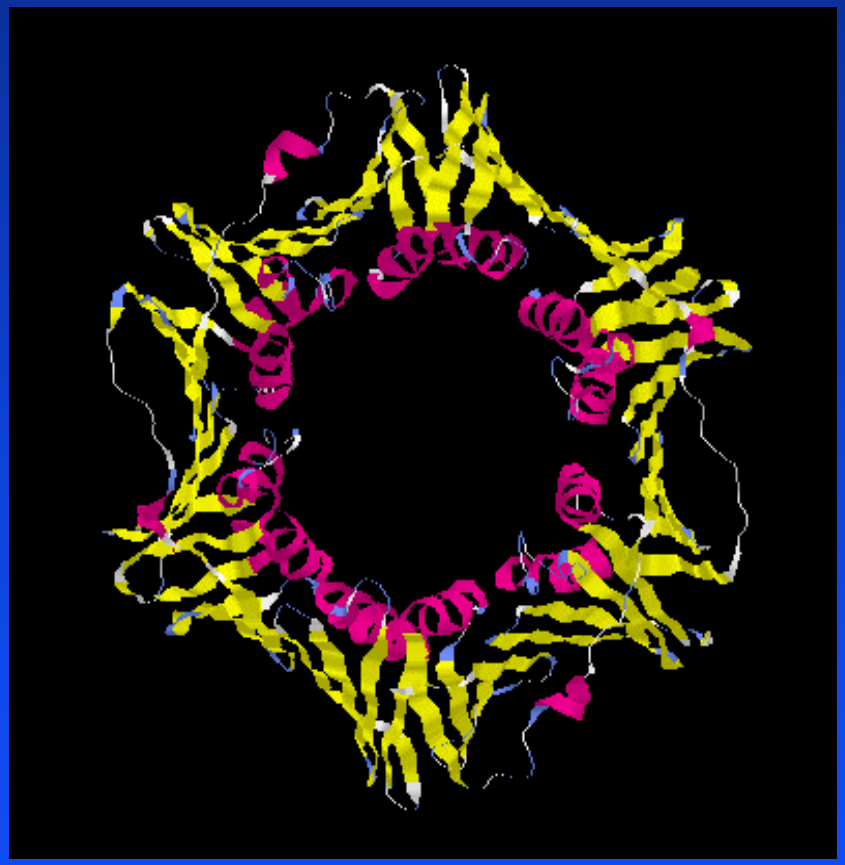
Computational Molecular Biology

(http://cmgm.stanford.edu/biochem218/)

Biochemistry 218



Dr. Doug Brutlag

Professor of Biochemistry & Medicine (by courtesy)

Lecture Syllabus

(http://cmgm.stanford.edu/biochem218/)

Date	Topic	Lecturer
Sept. 22	Representations of Sequences and Structures	Doug Brutlag
Sept 27	PubMed & Full Text Journal Access	Janet Morrison
Sept 29	Molecular Biology Databases on the Web	Doug Brutlag
Oct 4	Molecular Databases II	Doug Brutlag
Oct 6	Pattern Matching with Consensus Sequences	Doug Brutlag
Oct. 11	Quantitative and Probabilistic Pattern Matching	Doug Brutlag
Oct. 13	Sequence Alignment	Doug Brutlag
Oct. 18	Rapid Sequence Similarity Search I	Doug Brutlag
Oct. 20	Rapid Sequence Similarity Search II	Doug Brutlag
Oct. 25	Near-Optimal Sequence Alignments	Doug Brutlag
Oct. 27	Multiple Sequence Alignment	Doug Brutlag
Nov 1	Sequence Based Phylogenies	Doug Brutlag
Nov. 3	Sequence Blocks and Profiles	Doug Brutlag
Nov. 8	Discrete Protein Sequence Motifs	Doug Brutlag
Nov. 10	Protein Microenvironments	Russ Altman
Nov. 15	Probabilistic Protein Motifs	Tom Wu
Nov. 17	Motif Discovery Using Gibb's Sampling	Scott Schmidler
Nov. 22	Issues in Predicting Protein Secondary Structure	Scott Schmidler
Nov. 24	Protein Folds and Protein Structure Superposition	Amit P. Singh
Nov. 30	Protein Ligand Docking	Amit P. Singh

Course Availability

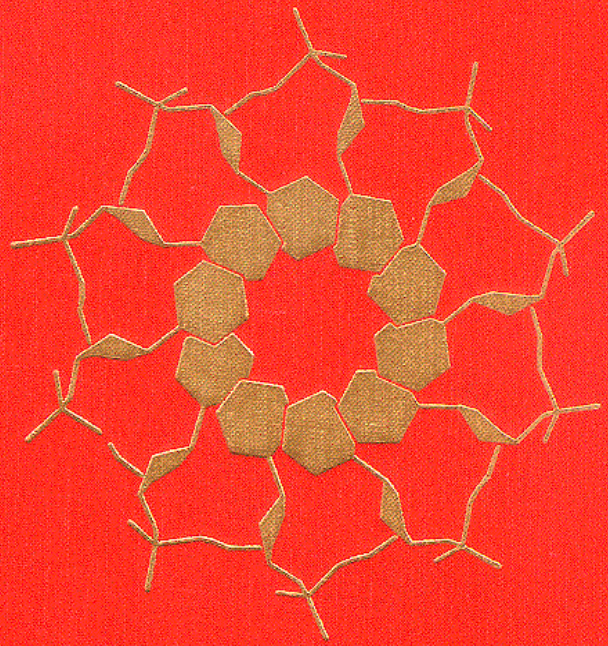
- Gates B03
  - Monday and Wednesday 2:15-3:45 PM
- Stanford Center for Professional Development
  - http://scpd.stanford.edu/
  - Live on SITN Channel E1
- Stanford Online
  - http://stanford-online.stanford.edu/
- Course available 24 hours/day, 7 days/week
- Students may register in any quarter

Course Requirements

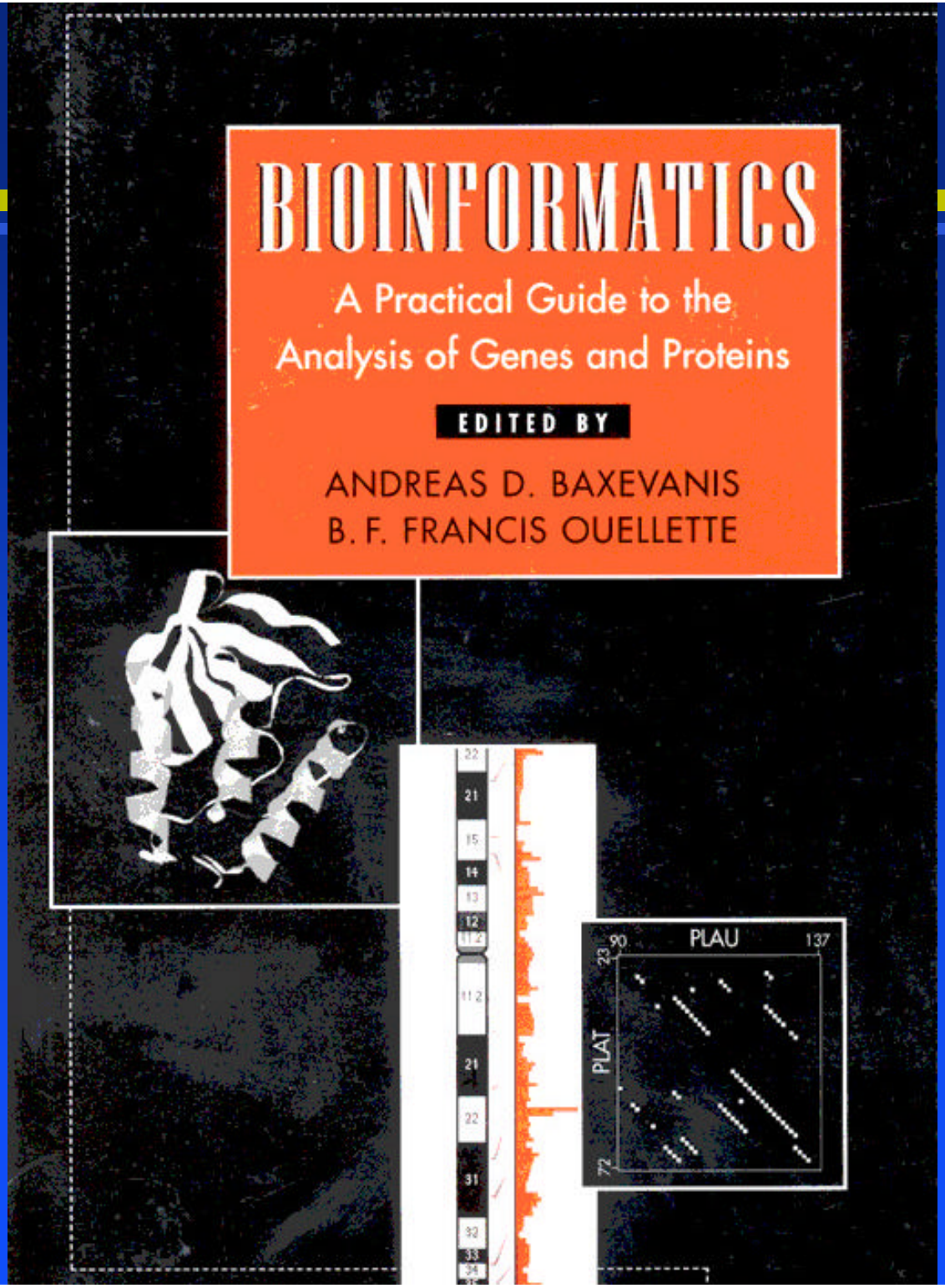
- Lectures
  - Theoretical background of current methods
  - Strengths and weaknesses of current approaches
  - Future directions for improvements
- Demonstrations
  - Implementations (Mac, PC, Unix, Web)
  - Illustrate homework
- Six to eight homework assignments
  - All homework submitted electronically as email attachments
  - Due one week after assigned
- Final project (DUE NOVEMBER 30TH)
  - Critically review an area
  - Critically analyze your own data sets
  - Propose new approach
  - Implement a new approach

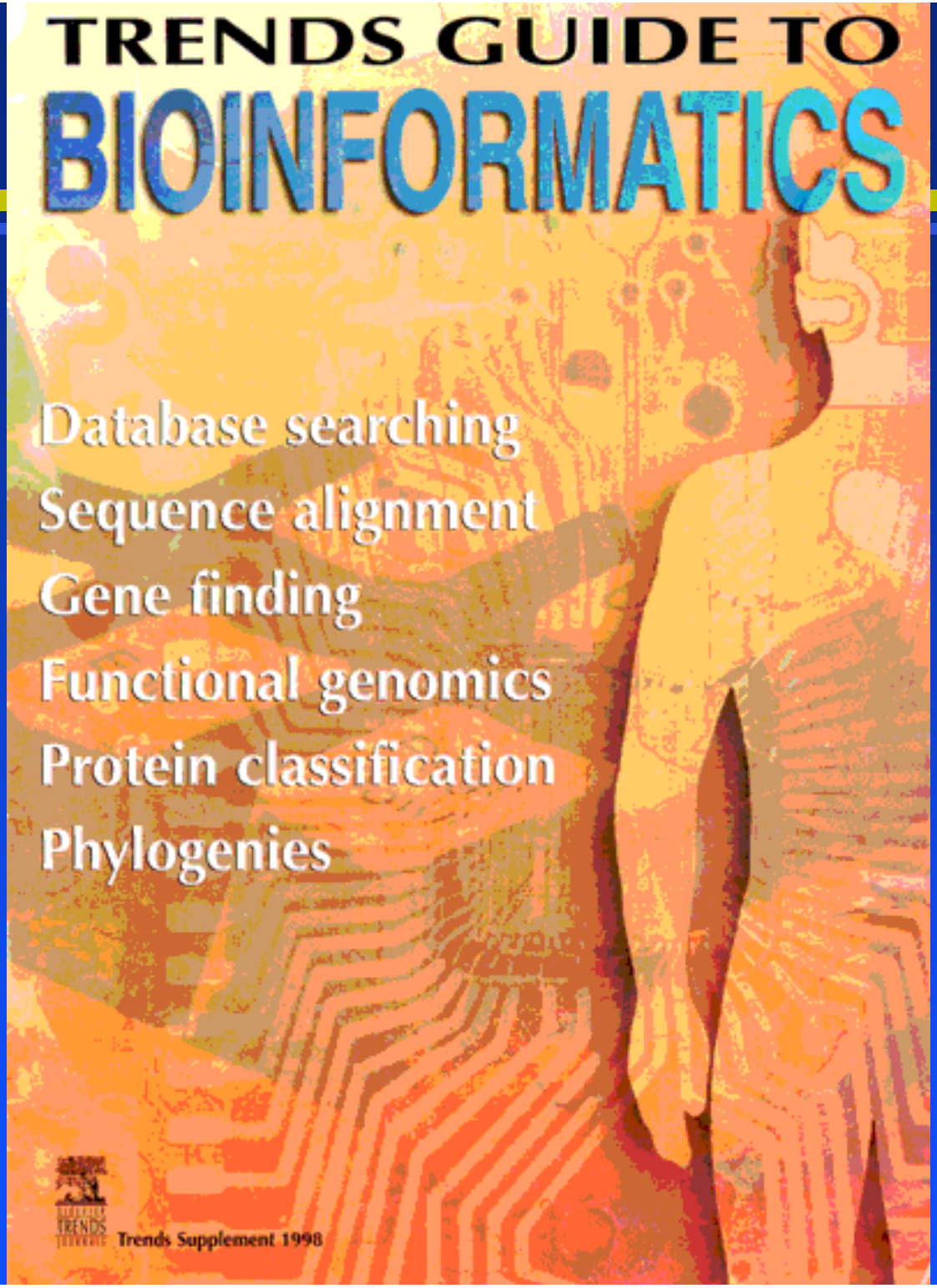
Biochemistry

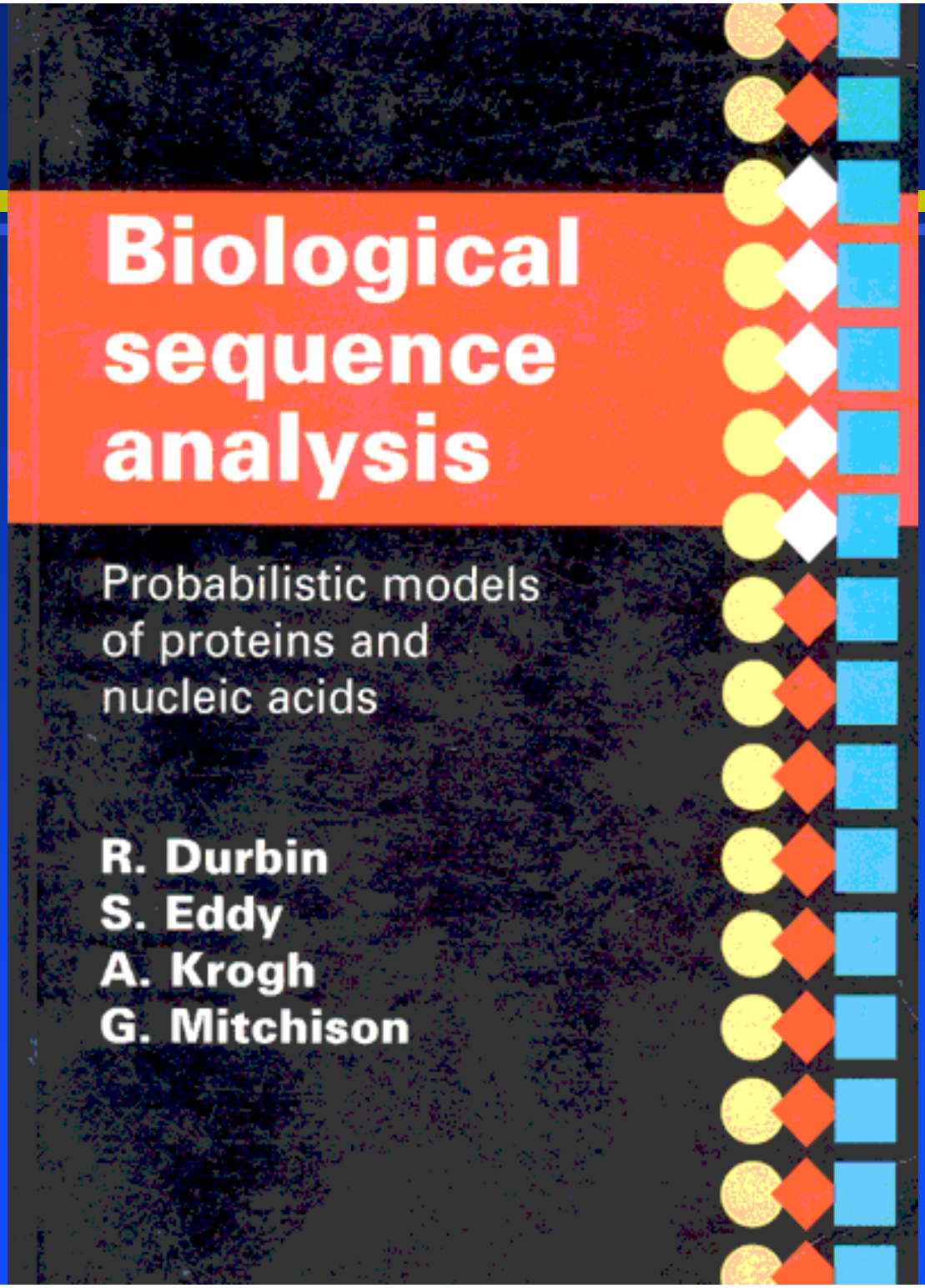
FOURTH EDITION

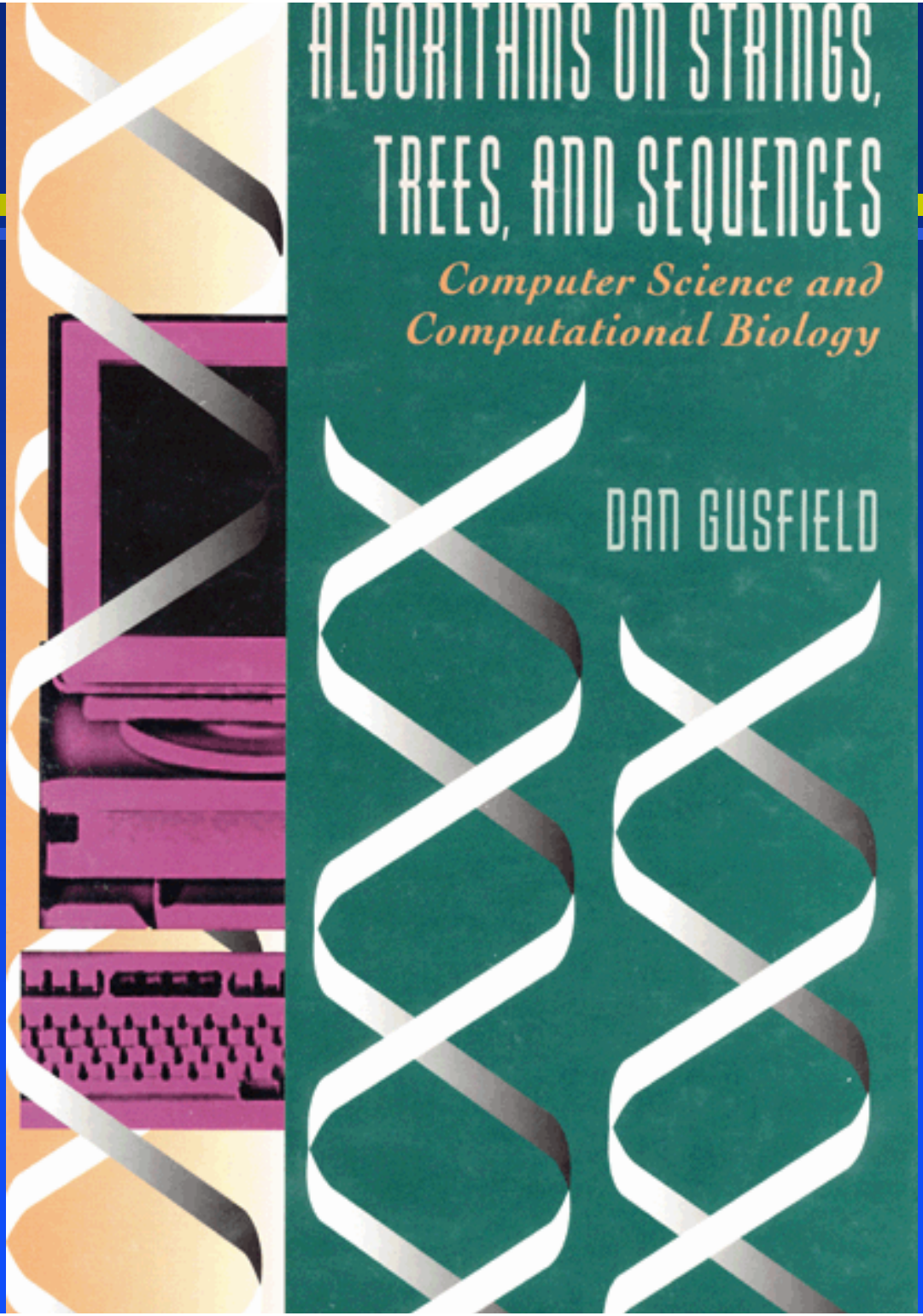


Lubert Stryer

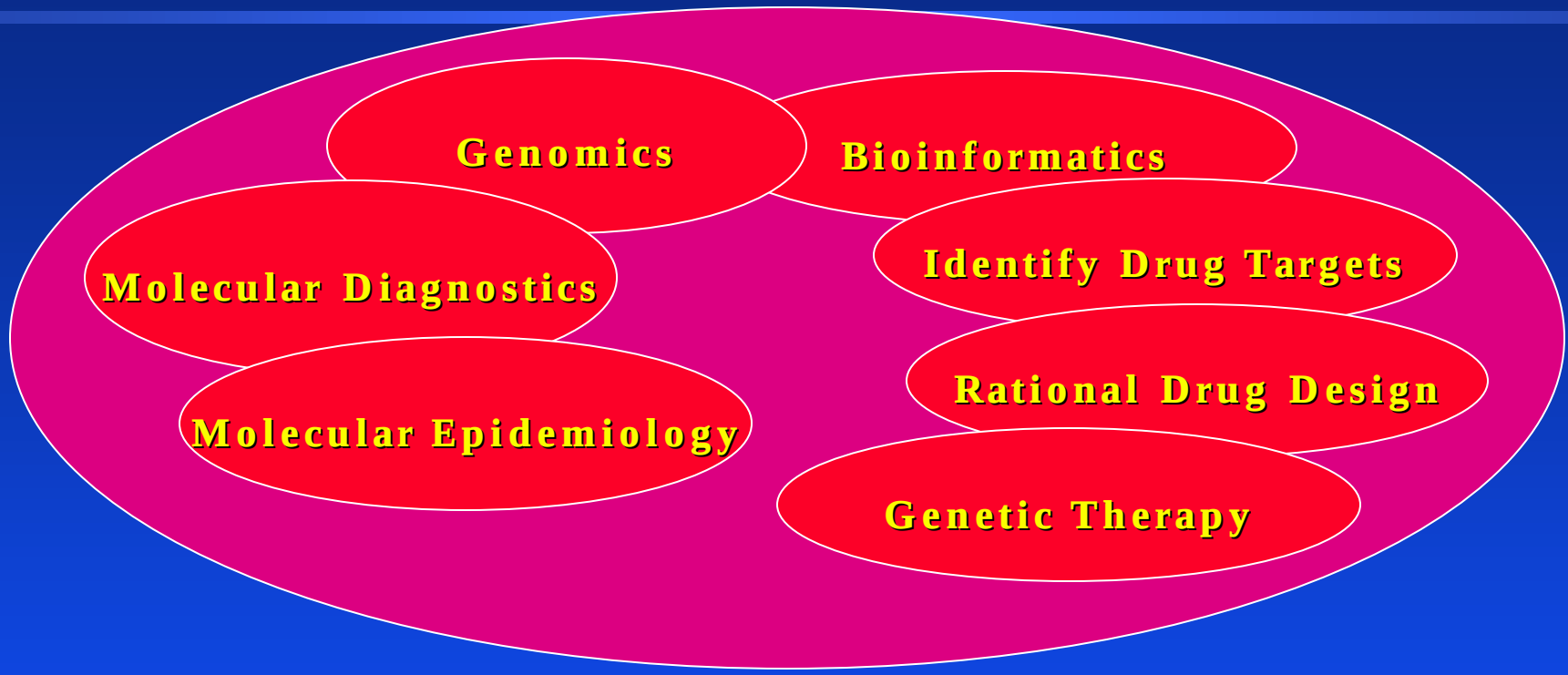

BIOINFORMATICS
A Practical Guide to the Analysis of Genes and Proteins
EDITED BY
ANDREAS D. BAXEVANIS
B. F. FRANCIS OUELLETTE
PLAU
PLAT
90
137
72


TRENDS GUIDE TO
BIOINFORMATICS
Database searching
Sequence alignment
Gene finding
Functional genomics
Protein classification
Phylogenies
Trends Supplement 1998

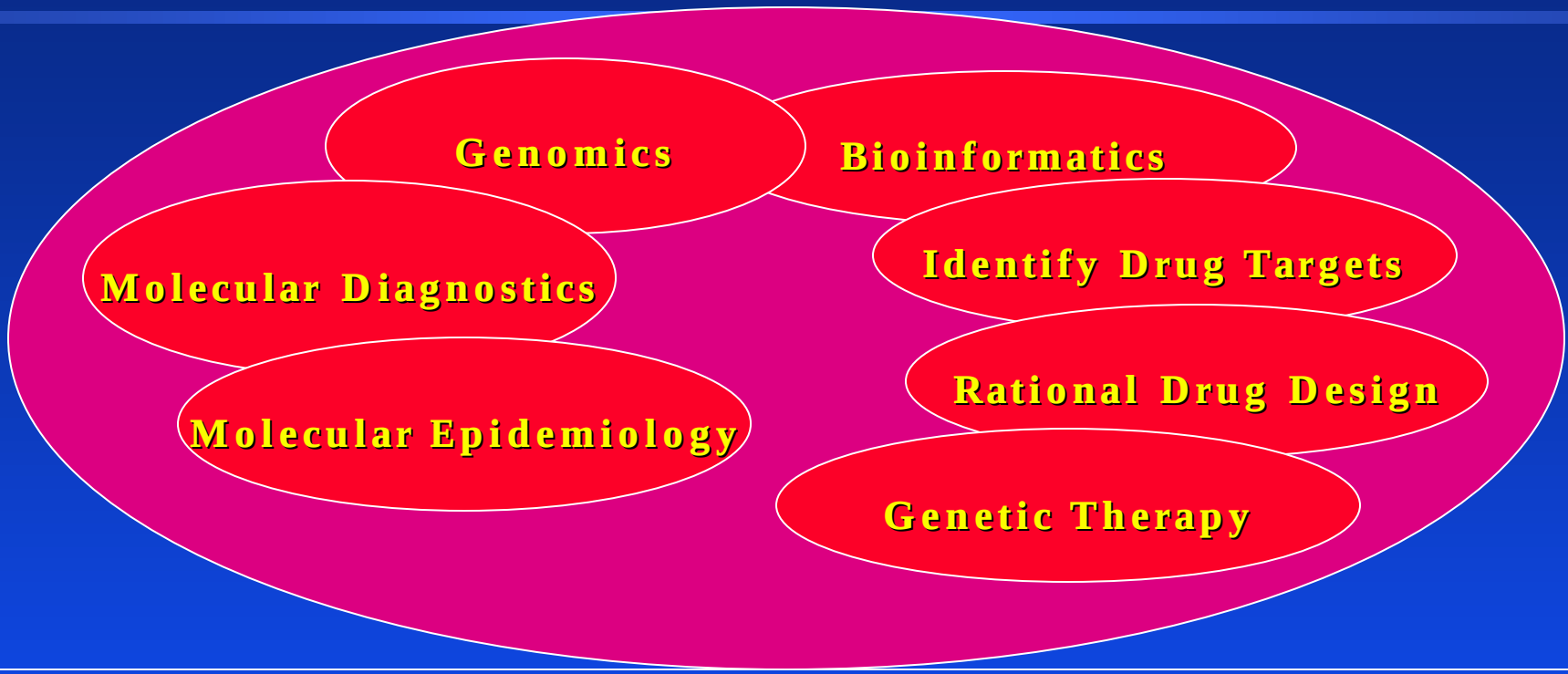

Biological sequence analysis
Probabilistic models of proteins and nucleic acids
R. Durbin
S. Eddy
A. Krogh
G. Mitchison

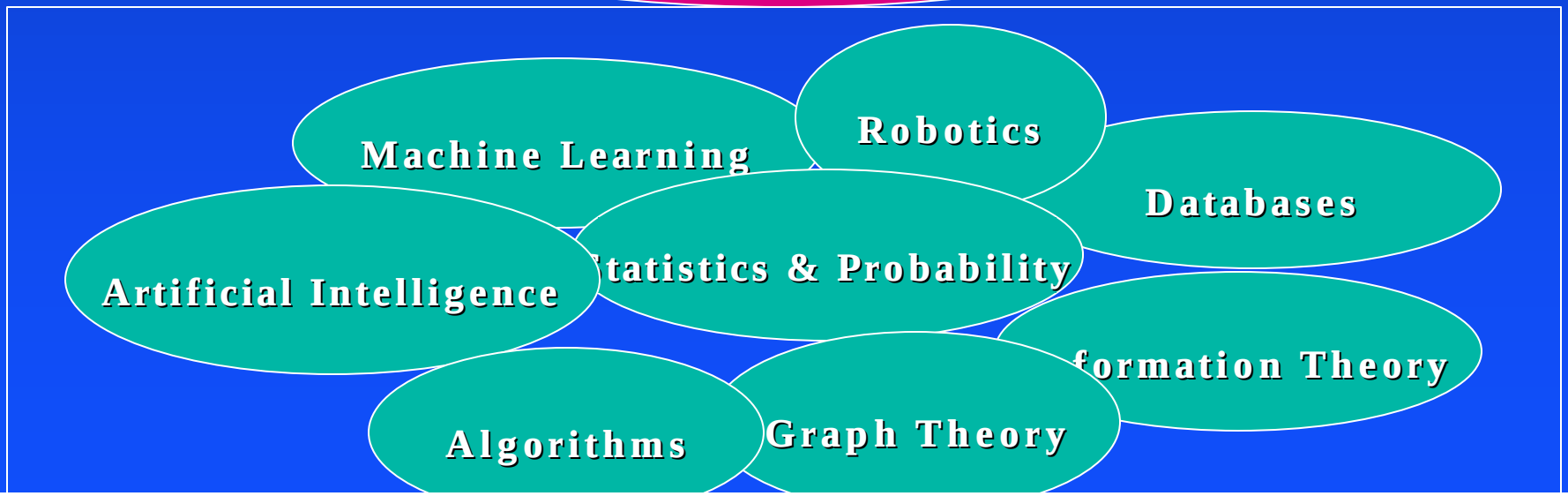

ALGORITHMS ON STRINGS, TREES, AND SEQUENCES
Computer Science and Computational Biology
DAN GUSFIELD

Genomics, Bioinformatics & Medicine



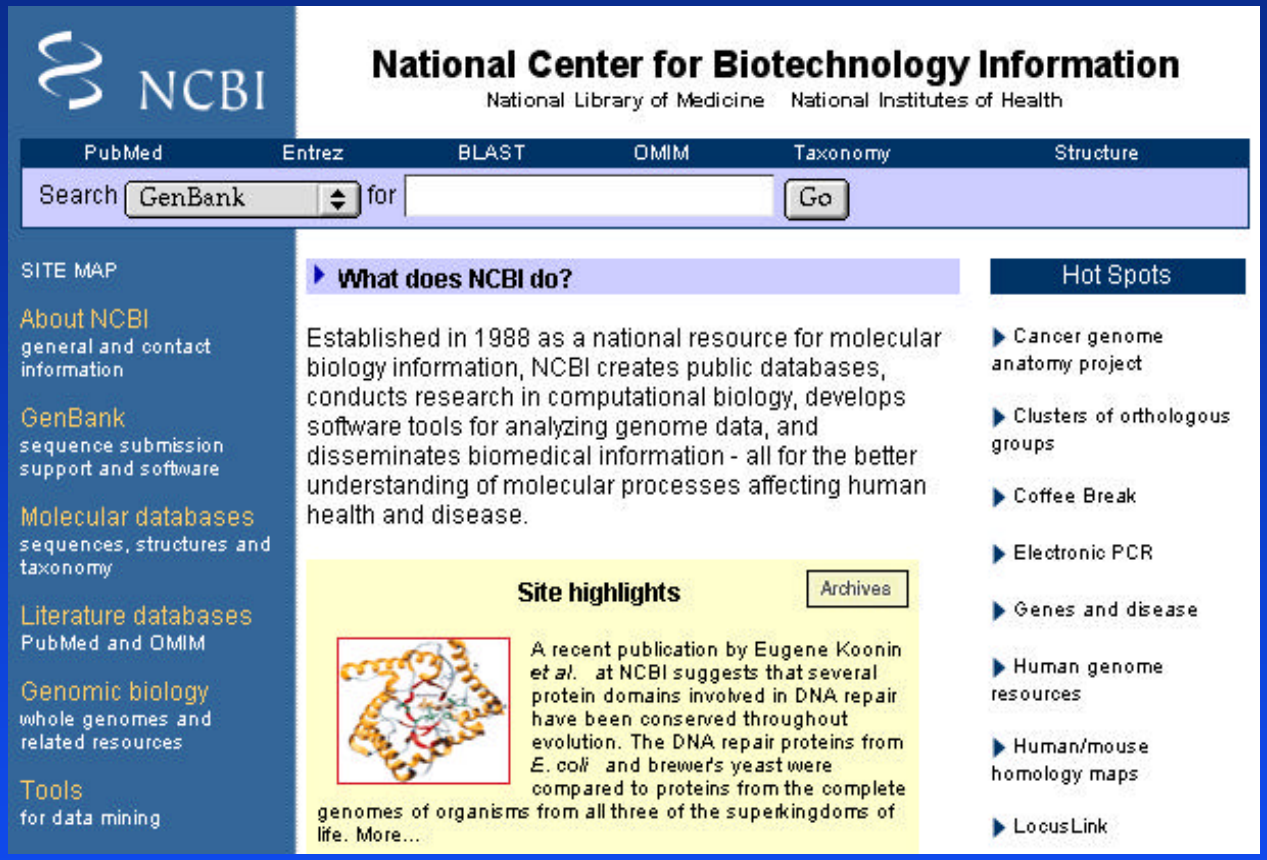
Genomics, Bioinformatics & Medicine





National Center for Biotechnology Information

(http://www.ncbi.nlm.nih.gov)



Human Genome Resources

(http://www.ncbi.nlm.nih.gov/genome/guide/)

NCBI

Genomes Guide

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

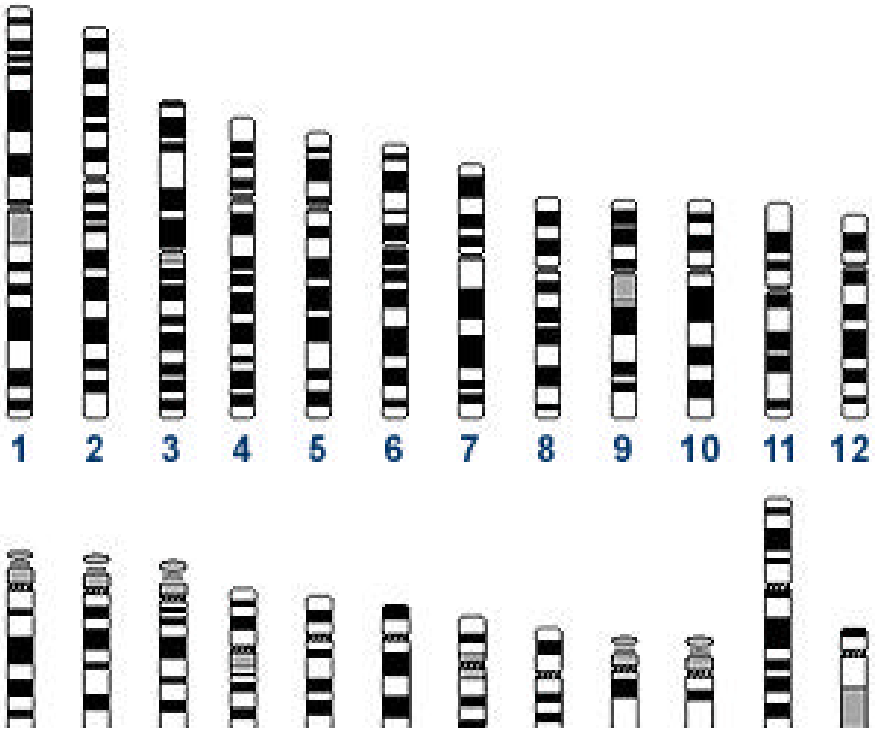
Search LocusLink for Go

Genomic Biology

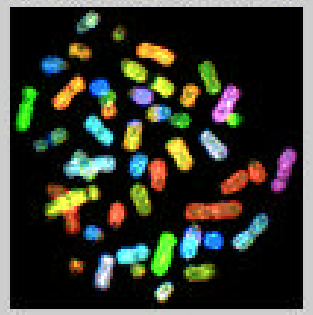
Human Genome Resources

- Human Genes
  - LocusLink
  - OMIM
- Sequences
  - Human genome sequencing
  - Reference mRNA sequences
  - UniGene
  - dbEST
- Maps
  - GeneMap'99

The Genome at a Glance



Genes & Disease



Disease gene profiles for students and the public

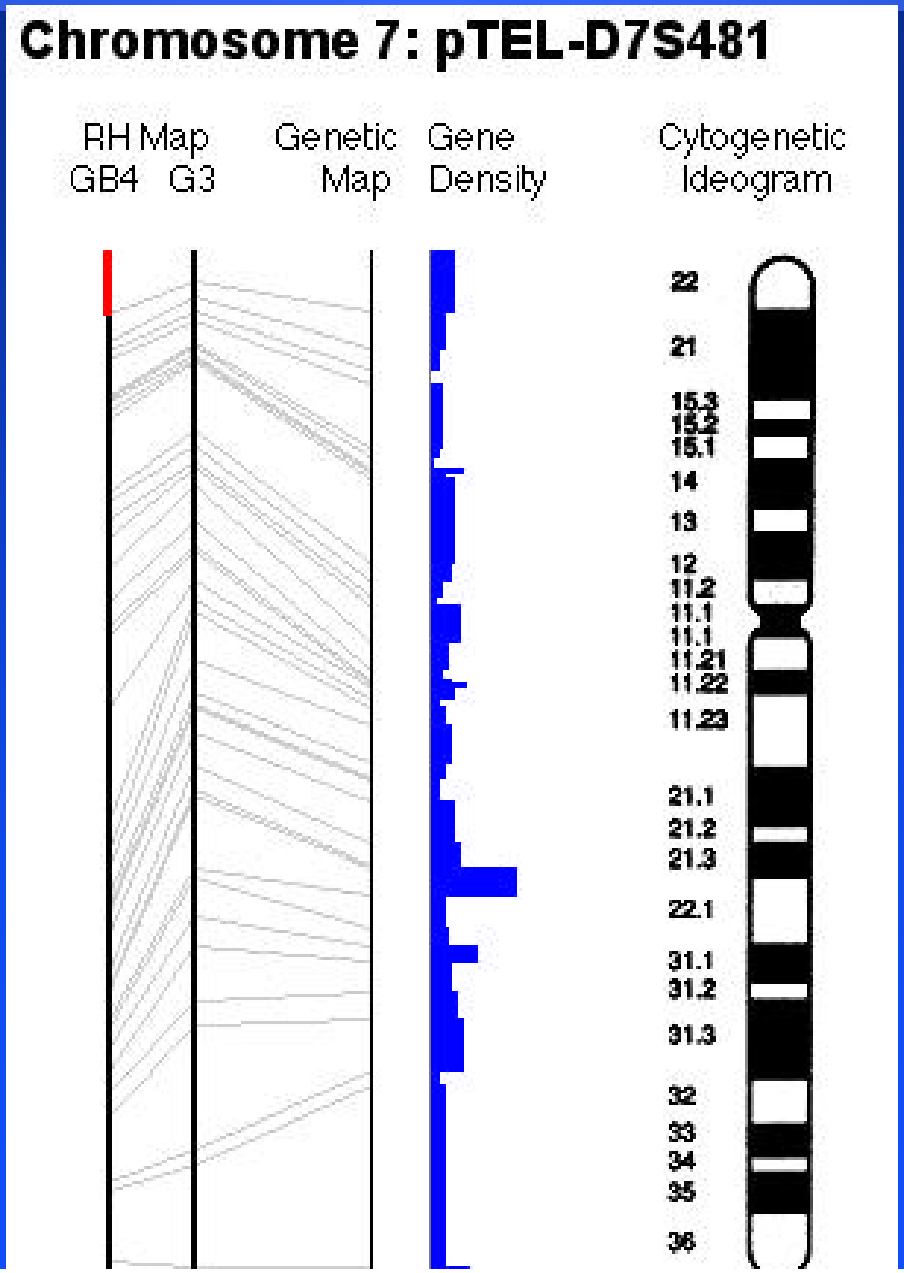
NHGRI

National Human Genome Research Institute

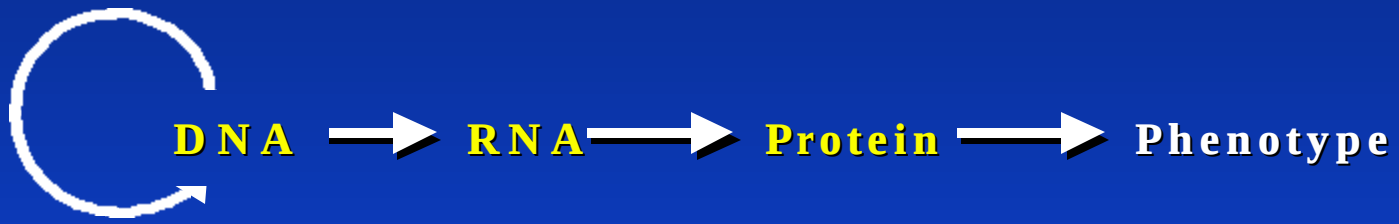
Human genome project

Genes on Chromosome 7

(http://www.ncbi.nlm.nih.gov/genemap/map.cgi?CHR=7)



Central Paradigm of Molecular Biology



- **Molecules**
  - Structure
  - Function
- **Processes**
  - Mechanism
  - Specificity
  - Regulation

Central Paradigm of Bioinformatics

Genetic Information

SRAAINKHIV
A
VSYQTVSRVV
N
VSTATVSRAL
A
GVTTTVSHVI
N
SGVSAVSAIL
N
GVSEMTRRDL
N
TAYATIHVRV
E
GSQPTVSREL
A
MSIATITRGS
N
ISRETVGRILK
FDISRLSHLFR
LRPSRLAHLF
R
MTVETISRLL
G
TLEFHLHRLF
K

Central Paradigm of Bioinformatics

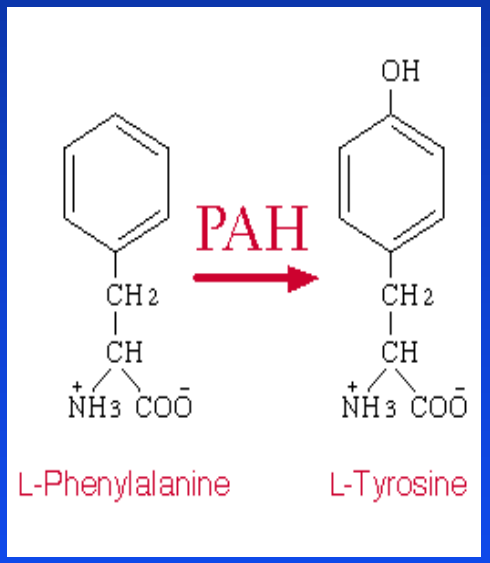
Genetic Information → **Molecular Structure**

SRAAINKHIV
A
VSYQTVSRVV
N
VSTATVSRAL
A
GVTTTVSHVI
N
SGVSAVSAIL
N
GVSEMTRRDL
N
TAYATIHVRV
E
GSQPTVSREL
A
MSIATITRGS
N
ISRETVGRILK
FDISRLSHLFR
LRPSRLAHLF
R
MTVETISRLL
G
TLEFHLHRLF
K

Central Paradigm of Bioinformatics

Genetic Information → **Molecular Structure** → **Biochemical Function**

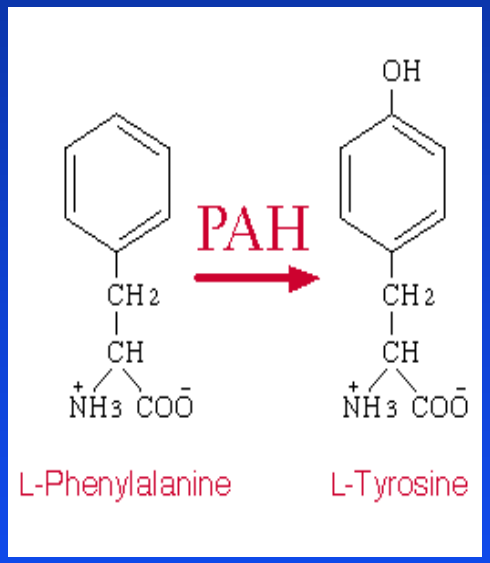
SRAAINKHIV
A
VSYQTVSRVV
N
VSTATVSRAL
A
GVTTTVSHVI
N
SGVSAVSAIL
N
GVSEMTRRDL
N
TAYATIHVRV
E
GSQPTVSREL
A
MSIATITRGS
N
ISRETVGRILK
FDISRLSHLFR
LRPSRLAHLF
R
MTVETISRLL
G
TLEFHLHRLF
K

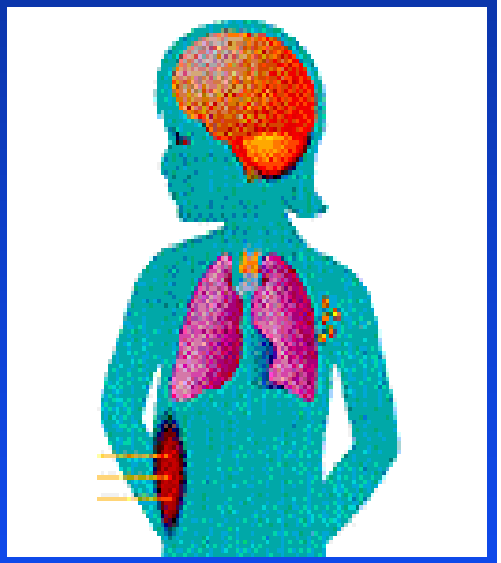


Central Paradigm of Bioinformatics

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

SRAAINKHIV
A
VSYQTVSRVV
N
VSTATVSRAL
A
GVTTTVSHVI
N
SGVSAVSAIL
N
GVSEMTRRDL
N
TAYATIHVRV
E
GSQPTVSREL
A
MSIATITRGS
N
ISRETVGRILK
FDISRLSHLFR
LRPSRLAHLF
R
MTVETISRLL
G
TLEFHLHRLF
K

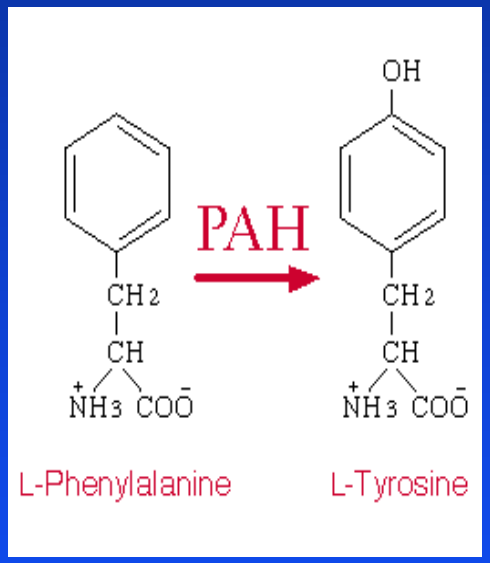


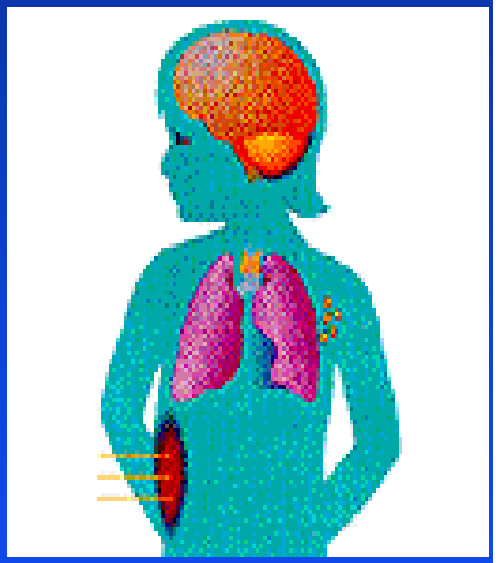


Central Paradigm of Bioinformatics

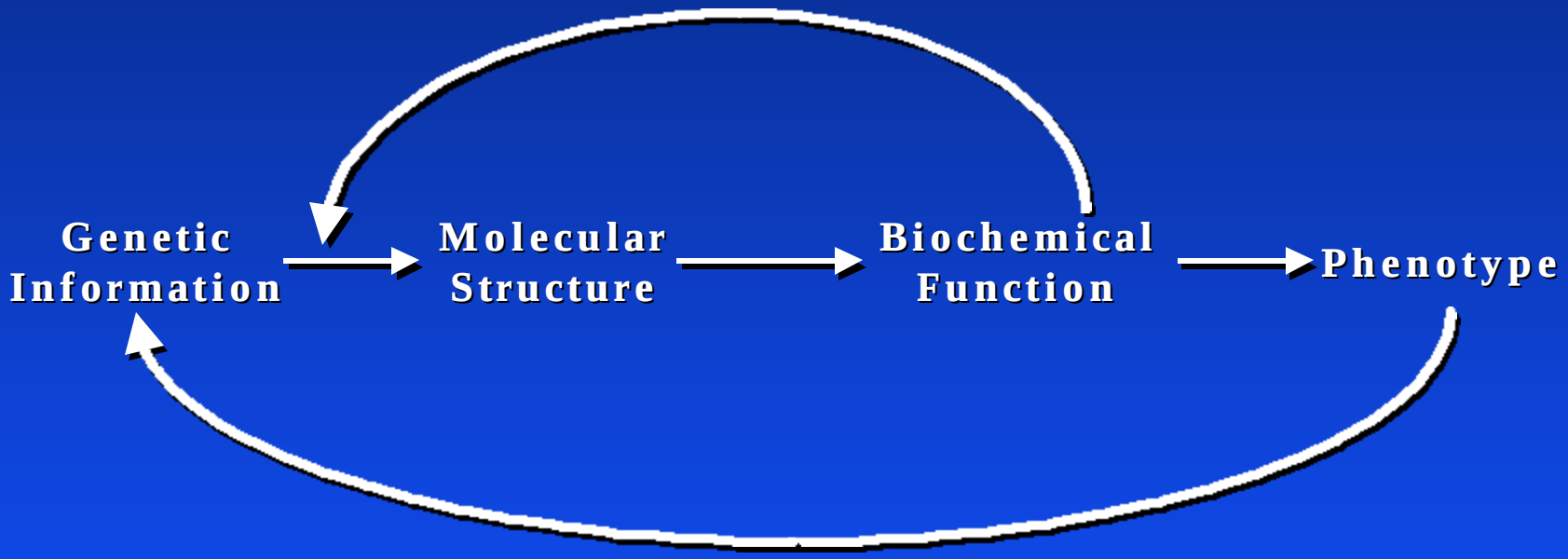
Genetic Information → **Molecular Structure** ⟷ **Biochemical Function** → **Symptoms (Phenotype)**

SRAAINKHIVA
VSYQTVSRVVN
VSTATVSRALA
GVTTTVSHVIN
SGVSAVSAILN
GVSEMTRRDLN
TAYATIHVRVE
GSQPTVSRELA
MSIATITRGSN
ISRETVGRILK
FDISRLSHLFR
LRPSRLAHLFR
MTVETISRLLG
TLEFHLHRLFK

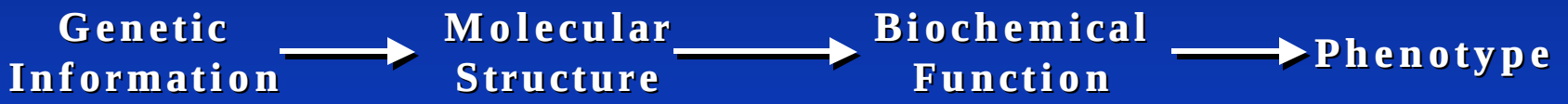




Central Paradigm of Bioinformatics



Challenges Understanding Genetic Information



- Genetic information is redundant
- Structural information is redundant
- 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
- Strain variation
- Sequencing errors

Discovering Function from Protein Sequences

Discovering Function from Protein Sequences

Sequences of Common Structure or Function

Sequence Alignments

```
          10        20        30        40          50
1   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
     |:| :|: | |:||||    | |:||| |: : :|:|  :|  |     |:  |
2   HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
          10        20        30        40        50
Initial Score    =    63  Optimized Score  =    98  Significance =  5.51
Residue Identity =   14%  Matches          =    21  Mismatches   =   22
Gaps             =     2  Conservative Substitutions          =   11
```

Discovering Function from Protein Sequences

Consensus Sequences

Zinc Finger (C2H2 type)
CX{2,4}CX{12}HX{3,5}H

Sequences of Common Structure or Function

Sequence Alignments

```
          10        20        30        40          50
1   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
     |:| :|: | |:||||    | |:||| |: : :|:|  :|  |     |:  |
2   HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
          10        20        30        40        50
Initial Score    =    63  Optimized Score  =    98  Significance =  5.51
Residue Identity =   14%  Matches          =    21  Mismatches   =   22
Gaps             =     2  Conservative Substitutions          =   11
```

Discovering Function from Protein Sequences

locks, Profiles or Templates

Position

```
 1  2  3  4  5  6  7  8  9 10 11 12

 2  1  3 13 10 12 67  4 13  9  1  2
 7  5  8  9  4  0  1 16  7  0  1  0
 0  8  0  1  0  0  0  2  1  1 10  0
 0  1  0  1 13  0  0 12  1  0  4  0
 0  0  1  0  0  0  0  0  0  2  2  1
 1  1 21  8 10  0  0  7  6  0  0  2
 2  0  0  9 21  0  0 15  7  3  3  0
 9  7  1  4  0  0  8  0  0  0 46  0
 4  3  1  1  2  0  0  2  2  0  5  0
10  0 11  1  2 10  0  4  9  3  0 16
16  1 17  0  1 31  0  3 11 24  0 14
 3  4  5 10 11  1  1 13 10  0  5  2
 7  1  1  0  0  0  0  0  5  7  1  8
 4  0  3  0  0  4  0  0  0 10  0  0
 0  6  0  1  0  0  0  0  0  0  0  0
 1 17  0  8  3  1  3  0  2  2  2  0
 5 22  3 11  1  5  0  2  2  2  0  5
 2  0  0  0  0  0  0  0  0  1  0  1
 1  0  4  2  0  1  0  0  2  4  0  1
 6  3  1  1  2 15  0  0  2 12  0 28
```

Consensus Sequences

Zinc Finger (C2H2 type)
CX{2,4}CX{12}HX{3,5}H

Sequences of Common Structure or Function

Sequence Alignments

```
                10        20        30        40          50
      1   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
          |:| :|: | |:||||    | |:||| |: : :|:|  :|  |     |:  |
      2   HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                10        20        30        40        50
Initial Score    =    63  Optimized Score  =    98  Significance =  5.51
Residue Identity =   14%  Matches         =    21  Mismatches   =   22
Gaps             =     2  Conservative Substitutions          =   11
```

Discovering Function from Protein Sequences

Blocks, Profiles or Templates

Position

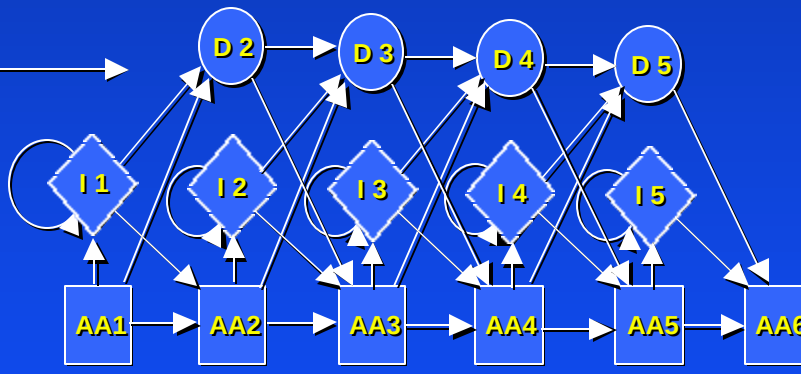
```
1  2  3  4  5  6  7  8  9 10 11 12
   2  1  3 13 10 12 67  4 13  9  1  2
   7  5  8  9  4  0  1 16  7  0  1  0
   0  8  0  1  0  0  0  2  1  1 10  0
   0  1  0  1 13  0  0 12  1  0  4  0
   0  0  1  0  0  0  0  0  0  2  2  1
   1  1 21  8 10  0  0  7  6  0  0  2
   2  0  0  9 21  0  0 15  7  3  3  0
   9  7  1  4  0  0  8  0  0  0 46  0
   4  3  1  1  2  0  0  2  2  0  5  0
  10  0 11  1  2 10  0  4  9  3  0 16
  16  1 17  0  1 31  0  3 11 24  0 14
   3  4  5 10 11  1  1 13 10  0  5  2
   7  1  1  0  0  0  0  0  5  7  1  8
   4  0  3  0  0  4  0  0  0 10  0  0
   0  6  0  1  0  0  0  0  0  0  0  0
   1 17  0  8  3  1  3  0  2  2  2  0
   5 22  3 11  1  5  0  2  2  2  0  5
   2  0  0  0  0  0  0  0  0  1  0  1
   1  0  4  2  0  1  0  0  2  4  0  1
   6  3  1  1  2 15  0  0  2 12  0 28
```

Consensus Sequences

Zinc Finger (C2H2 type)
CX{2,4}CX{12}HX{3,5}H

Sequences of Common Structure or Function

Hidden Markov Model



Sequence Alignments

```
                10        20        30        40          50
      1   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
          |:| :|: | |:||||    | |:||| |: : :|:|  :|  |     |:  |
      2   HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                10        20        30        40        50
Initial Score    =    63  Optimized Score  =    98  Significance =  5.51
Residue Identity =   14%  Matches         =    21  Mismatches   =   22
Gaps             =     2  Conservative Substitutions          =   11
```

Multiple Representations of Protein Structure

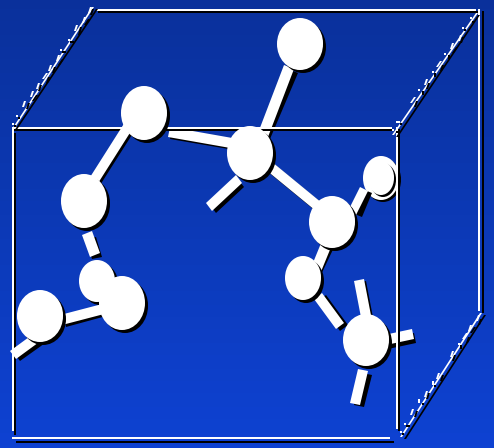
$$\sum_{\text{all bonds}} K_b (b_i - b_o)^2 +$$

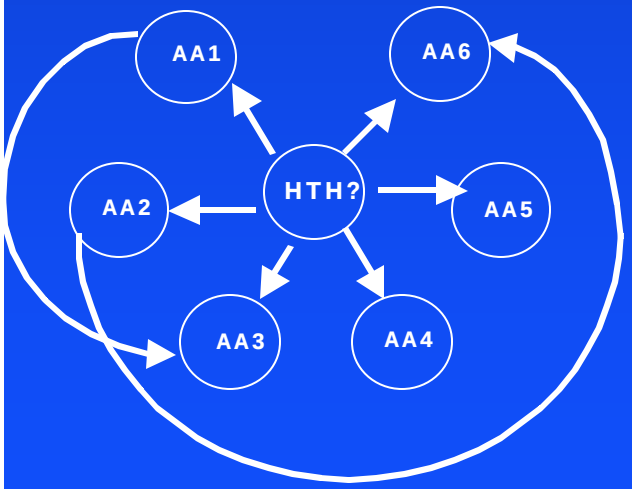
$$\sum_{\text{all angles}} K_q (q_i - q_o)^2 +$$

$$\sum_{\text{all torsion angles}} K_f [1 - \cos (n f_i + d)] +$$

$$\sum_{\text{all non-bonded pairs}} e \{(r_o / r_{ij})^{12} - 2 (r_o / r_{ij})^6\}$$

$$\sum_{\text{all partial charge pairs}} q_i q_j / r_{ij}$$





82% Hydrophobic

18% Hydrophilic

```
  40          50
FPTTKTYFPHF-DLS-----HGS
:| |: :|  | |||     |
YPWTQRFFESFGDLSTPDAVMGI
  40      50
```

Buried Treasure

Buried Treasure

Buried Treasure

Sequence Alignment

(http://motif.stanford.edu/alion/)

```
   X     220     230     240     250       X
F--SGGNTHIYMNHVEQCKEILRREPKELCELVISGLPYKFRYLSTKE-QLK-Y
          |  : |::|||:||:|  |  | |||: : :| |    |  ::::: |:: |
FIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKAEDFQIEGY
   X      260     270     280     290       X
```

Sequence Alignment

(http://motif.stanford.edu/alion/)

```
  X     220     230     240     250       X
F--SGGNTHIYMNHVEQCKEILRREPKELCELVISGLPYKFRYLSTKE-QLK-Y
          |  : |::|||:||:|  |  | |||: : :| |    |  ::::: |:: |
FIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKAEDFQIEGY
  X      260     270     280     290       X
```

$$\text{Score} = \sum_{\text{Region Start}}^{\text{Region End}} \text{Similarity-weights} - \sum_{\text{Region Start}}^{\text{Region End}} \text{Penalties}$$

where:

$$\text{Penalty} = \text{Gap-penalty} + \text{Size-of-gap} \times \text{Gap-size-penalty}$$

Smith-Waterman Similarity Search

Query: HU-NS1 Maximal Score: 452
PAM Matrix: 200 Gap Penalty: 5 Gap Extension: 0.5

No.	Score	Match	Length	DB	ID	Description	Pred. No.
1	452	100.0	90	2	DBHB_ECOLI	DNA-BINDING PROTEIN H	8.74e-86
2	451	99.8	90	2	DBHB_SALTY	DNA-BINDING PROTEIN H	1.54e-85
3	336	74.3	90	2	DBHA_ECOLI	DNA-BINDING PROTEIN H	1.64e-57
4	336	74.3	90	2	DBHA_SALTY	DNA-BINDING PROTEIN H	1.64e-57
5	328	72.6	90	2	DBH_BACST	DNA-BINDING PROTEIN I	1.35e-55
6	328	72.6	92	2	DBH_BACSU	DNA-BINDING PROTEIN I	1.35e-55
7	327	72.3	90	2	DBH_VIBPR	DNA-BINDING PROTEIN H	2.35e-55
8	302	66.8	90	2	DBH_PSEAE	DNA-BINDING PROTEIN H	2.14e-49
9	273	60.4	91	2	DBH1_RHILE	DNA-BINDING PROTEIN H	1.47e-42
10	272	60.2	91	2	DBH_CLOPA	DNA-BINDING PROTEIN H	2.52e-42
11	263	58.2	90	2	DBH_RHIME	DNA-BINDING PROTEIN H	3.18e-40
12	261	57.7	91	2	DBH5_RHILE	DNA-BINDING PROTEIN H	9.29e-40
13	250	55.3	94	2	DBH_ANASP	DNA-BINDING PROTEIN H	3.32e-37
14	233	51.5	93	2	DBH_CRYPH	DNA-BINDING PROTEIN H	2.70e-33
15	226	50.0	95	2	DBH_THETH	DNA-BINDING PROTEIN I	1.07e-31
16	210	46.5	99	3	IHFA_SERMA	INTEGRATION HOST FACT	4.46e-28
17	206	45.6	100	3	IHFA_RHOCA	INTEGRATION HOST FACT	3.52e-27
18	205	45.4	99	3	IHFA_SALTY	INTEGRATION HOST FACT	5.90e-27
19	204	45.1	99	3	IHFA_ECOLI	INTEGRATION HOST FACT	9.87e-27
20	200	44.2	94	3	IHFB_ECOLI	INTEGRATION HOST FACT	7.71e-26
21	200	44.2	94	3	IHFB_SERMA	INTEGRATION HOST FACT	7.71e-26
22	165	36.5	99	5	TF1_BPSP1	TRANSCRIPTION FACTOR	3.42e-18
23	147	32.5	90	2	DBH_THEAC	DNA-BINDING PROTEIN H	2.12e-14

Decypher Similarity Search

(http://decypher.stanford.edu/)

TimeLogic DeCypher **Bioinformatics Supercomputer**

Select a Sequence Analysis Method

Click on the hyperlink under **Your Query** next to the method you wish.

HEURISTIC METHODS		Your Query	Database	
Blastn		DNA	DNA	Nucleic comparison.
Blastx		DNA	Protein	Translated search in p
Tblastx		DNA	DNA	Translated search in a
Blastp	Example	Protein	Protein	Protein comparison.
Tblastn		Protein	DNA	Search of the translate
PSI-Blast		Protein	Protein	BLASTP with position
All Methods Form - BLASTALL		DNA or Protein	DNA or Protein	Combination form su

Prosite Consensus Patterns

(http://www.expasy.ch/prosite/)

- Active site of trypsin-like serine proteases

G D S G G

- Zinc Finger (C_2H_2 type)

C .{2,4} C .{12} H .{3,5} H

- N-Glycosylation Site

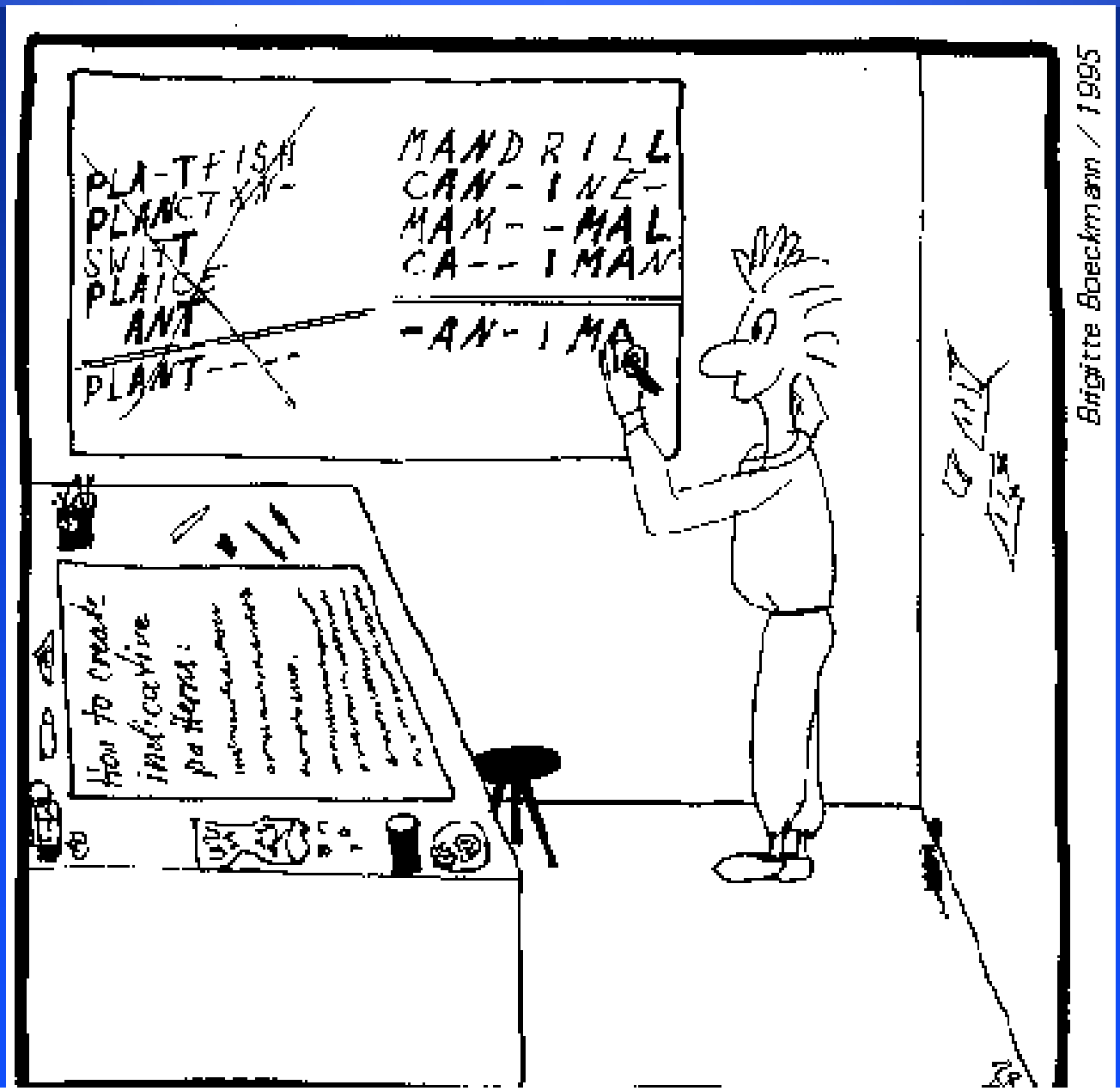
N [^P] [S T] [^P]

- Homeobox Domain Signature

[LIVMF] .{5} [LIVM] .{4} [IV] [RKQ] . W .{8} [RK]

The Optimal Way to Develop Patterns

(http://www.expasy.ch/images/cartoon/prosite.gif)



EMOTIF Pattern Discovery

(http://motif.stanford.edu/emotif/)

Craig G. Nevill-Manning, Thomas D. Wu, and Douglas L. Brutlag, Bioinformatics Group.

Sequences:

```
VIDIALKYQFDSQQSFAKRFKAYLGISPSLY
IGDIAICVGIANAPYFITLFKKKTGQTPARF
INVVAQKCGYNSTSYFICAFKDYYGVTPSHY
IANIGRVVGYDDQLYFSRVFRKRVGVSPSDF
LASIAHSVGYGSESALSVAFKRVLGMPPGDY
IIEISAKLFYDSQQTFTREFKKIFGYTPRQY
VTDIAFEAGYSSPSLFIKTFKKLTSFTPKSY
IGMIASLVGYTSVSYFIKTFKEYYGVTPKKF
IIDTASRWGIRSRSALVKGYRKQFNEAPSET
ITQLAVNHGYSSPSHFSSEIKELIGVSPRKL
```

Sponsored by

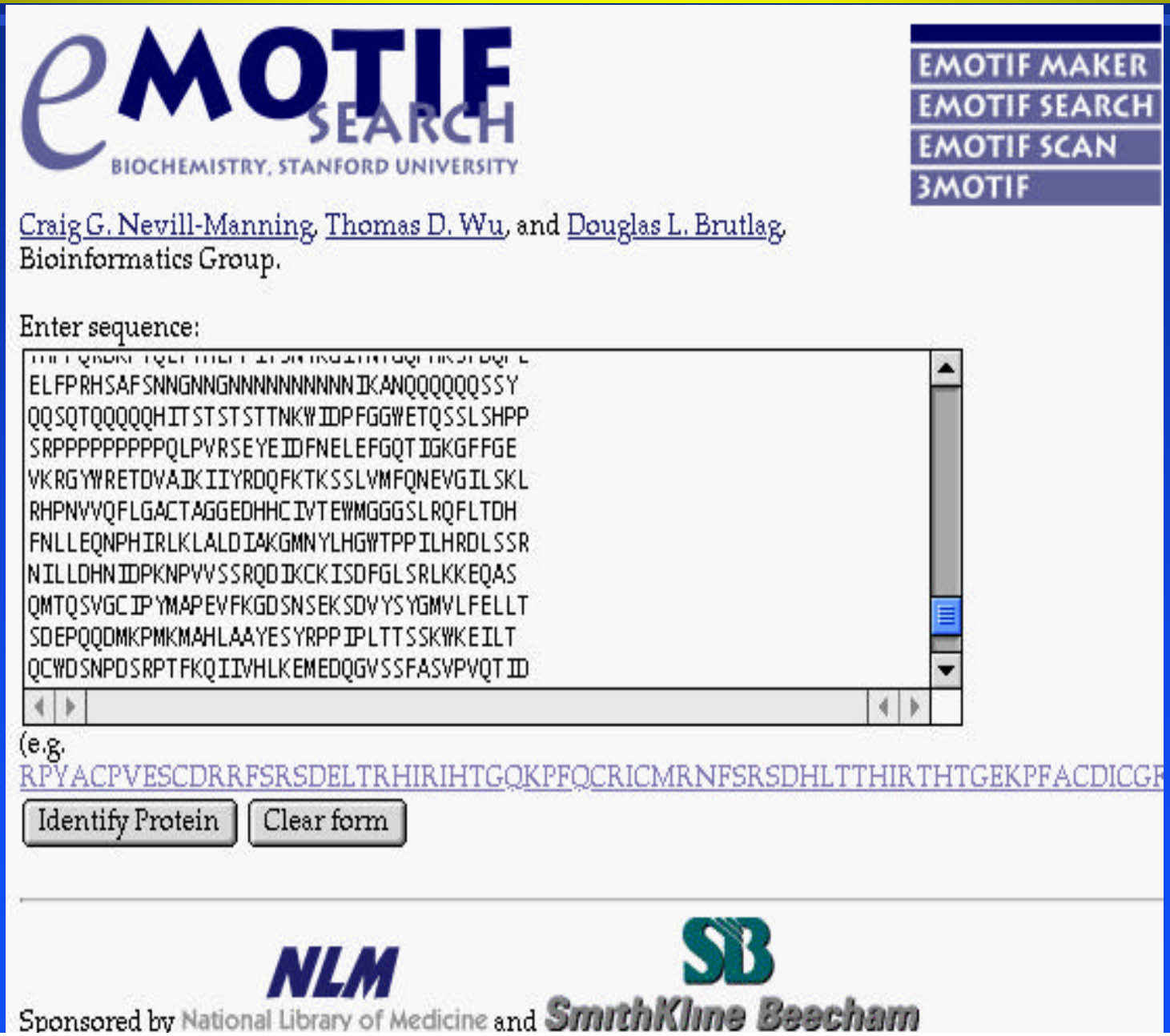
Find motifs | Clear form

Motifs must match 30 % of sequences. ☒ Draw score contours

Calculate specificity relative to distribution of ◉ SWISS-PROT ○ sequences supplied.

Identifying Protein Functions

(http://motif.stanford.edu/emotif-search)



Identifying Protein Function

eMOTIF SEARCH
BIOCHEMISTRY, STANFORD UNIVERSITY

EMOTIF MAKER
EMOTIF SEARCH
EMOTIF SCAN
3MOTIF

At a stringency of at least one in 10^{10} (no false positives expected) no matches.

At a stringency of at least one in 10^{9} (no false positives expected) no matches.

At a stringency of at least one in 10^{8} (no false positives expected)

Name	Description	Motif	Specificity
TYRKINASE	TYROSINE KINASE CATALYTIC DOMAIN SIGNATURE	[ilmv]..cw.....rp.f	$10^{-8.2}$
	positions 1537-1552	...RPPIPLTTSSKWKEILTQC▼DSNPDSRPTFKQIIVHLKEMEDQGV...	

At a stringency of at least one in 10^{7} (no false positives expected)

Name	Description	Motif	Specificity
PROTEIN_KINASE_ATP	Protein kinases ATP-binding region proteins.	[hy]rd[ilv]...n.[filmv][ilmv]	10^{-7}
	positions 1414-1425	...AKGMNYLHGWTPPILHRDLSSRNILLDHNIDPKNPVVSSRQ...	3D
		wi...ggw	$10^{-7.3}$
	positions 1245-1253	...QQQHITSTSTSTTNK▼IDPFGG▼ETQSSLSHPPSRPPP...	

At a stringency of at least one in 10^{6} (expect one false positive)

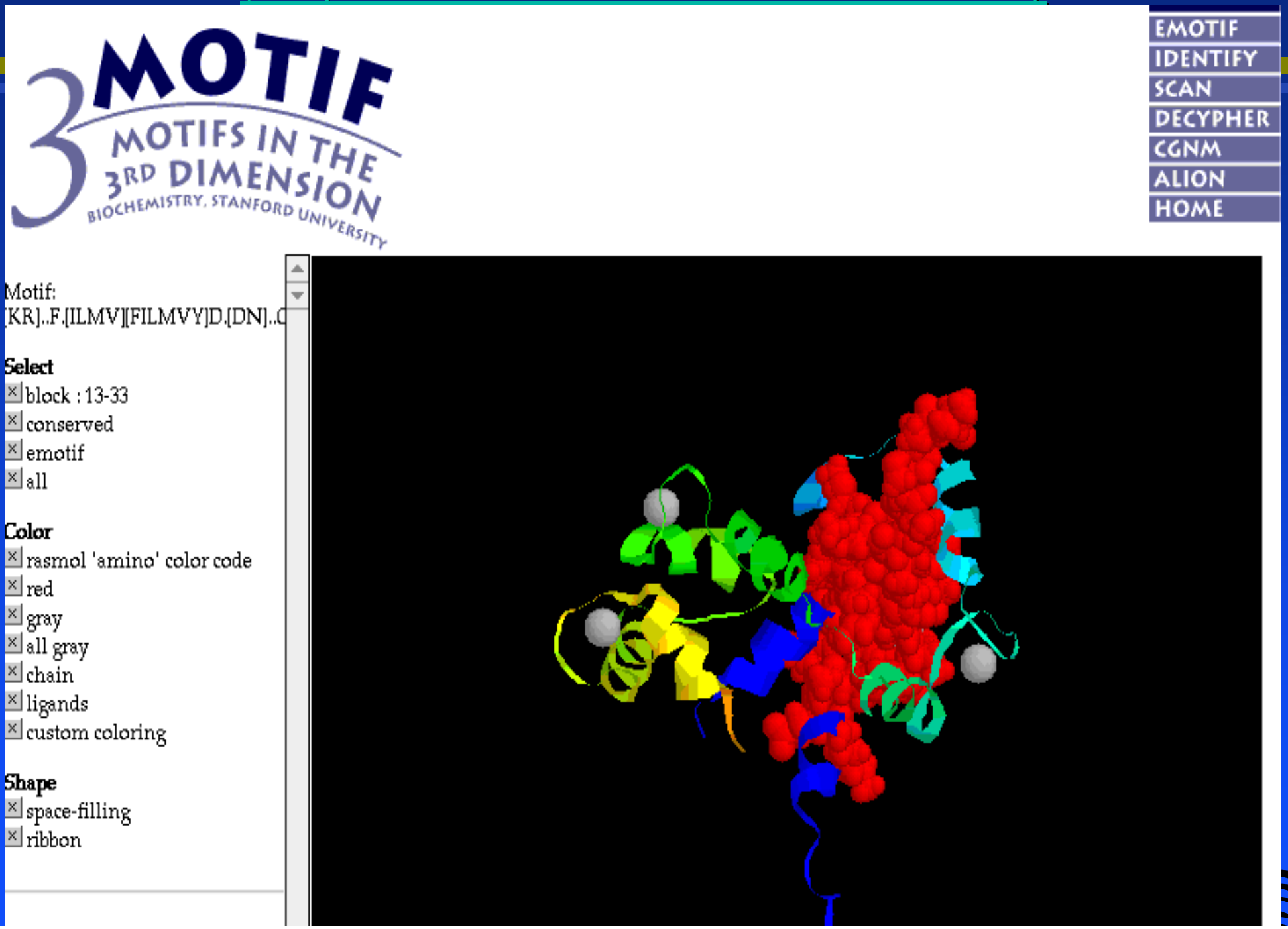
Name	Description	Motif	Specificity
PROTEIN_KINASE_ATP	Protein kinases ATP-binding region proteins.	[filmvy].[hy].d[filmv]...n.[filmv][filmvy]	10^{-6}
	positions 1412-1425	...DIAKGMNYLHGWTPPILHRDLSSRNILLDHNIDPKNPVVSSRQ...	3D

The score represents the number of bits saved if the sequences were transmitted with respect to the motif. For practical purposes, though, just the ranking is significant.

score	matches	expected	motif
971	32	10^{-2}	[ilv]..[iv]....g[filvy].....f...f......[ast]p..[fwy]
961	31	10^{-2}	[ilv]..[iv]....g[filvy].....f...f......[st]p..[fwy]
937	30	10^{-2}	[ilv]..[iv]....g[filvy].....f...f......[st]p..[fy]
923	29	10^{-2}	[iv]..[iv]....g[filvy].....f...f......[st]p..[fwy]
905	28	10^{-2}	[ilv]..[iv]....g[fy].....f...f......[ast]p..[fwy]
903	27	10^{-3}	[ilv]..[iv]....g[filvy]....[hy]f...f......[st]p..[filvy]
915	26	10^{-3}	[ilv]..[iv]....g[filvy]....[hy]f...f......[st]p..[fwy]
899	25	10^{-3}	[ilv]..[iv]....g[filvy]....yf...f......[st]p..[fwy]
869	24	10^{-3}	[ilv]..[iv]....g[filvy]....yf...f......[st]p..[fy]
846	23	10^{-4}	[iv]..[iv]....g[filvy]....yf...f......[st]p..[fwy]
818	22	10^{-4}	[ilv]..[iv]....g[fy]....[hy]f...f......[st]p..[fwy]
811	21	10^{-4}	[ilv]..[iv]....g[fy]....[hy]f...f[kr].....[st]p..[filvy]
773	20	10^{-4}	[iv]..[iv][ast]...g[fy].....f...[fy][kr]..[fy]..[st]p...
772	19	10^{-5}	[ilv]..[iv][ast]...g[filvy].s..[hy]f...[fy]...[fy]..[st]p...
766	18	10^{-5}	[ilv]..[iv][ast]...g[filvy].s..yf...[fy]...[fy]..tp...
746	17	10^{-6}	[ilv]..[iv][as]...g[filvy].s.[as][hy]f...[fy]...[fy]..[st]p...

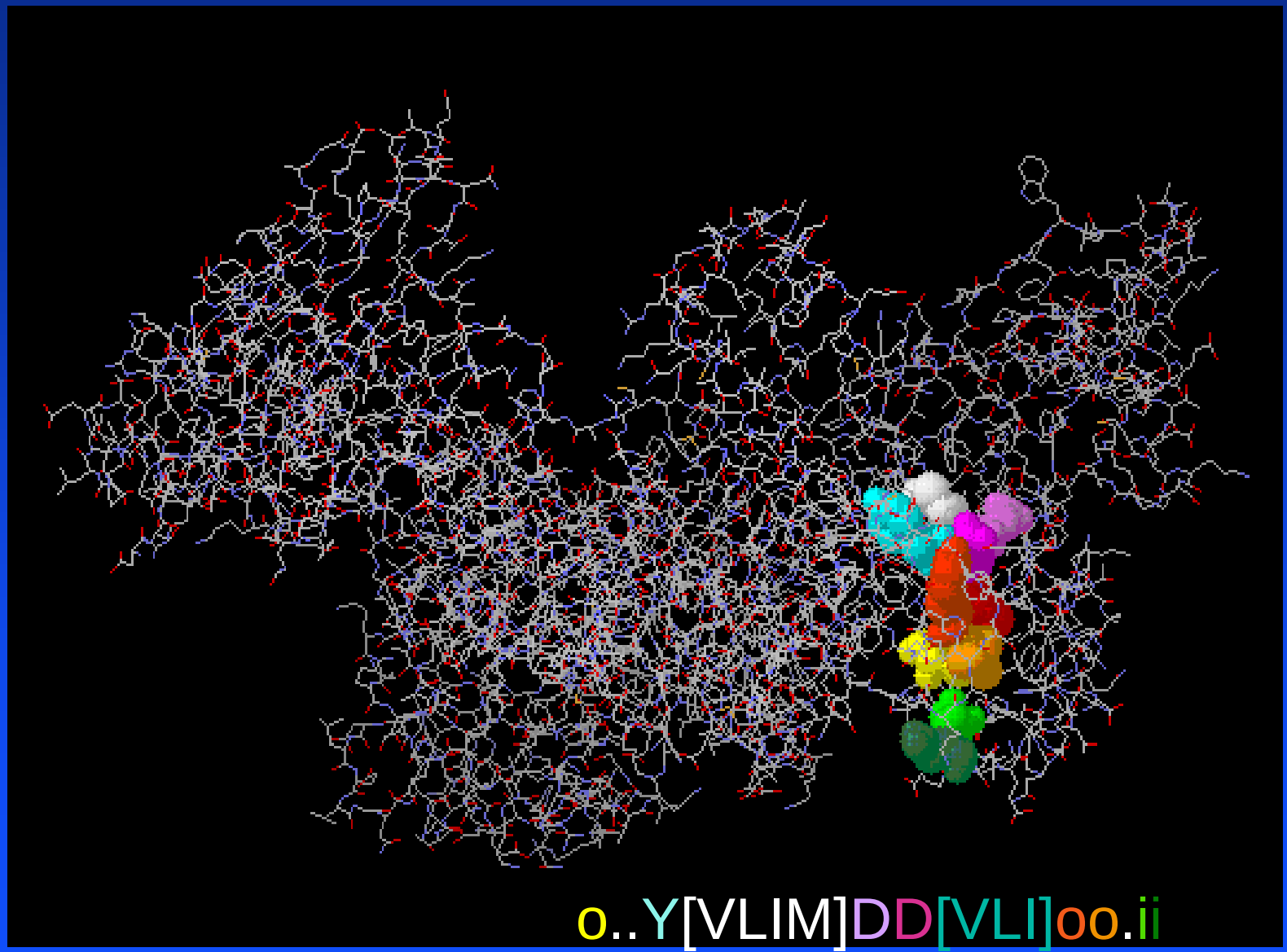
Mapping Sequence Motifs to Structural Motifs

(http://motif.stanford.edu/3motif/)



Motifs as Potential Drug Targets

HIV Reverse Transcriptase



Discovering Function from Protein Sequences

Blocks, Profiles or Templates

Position

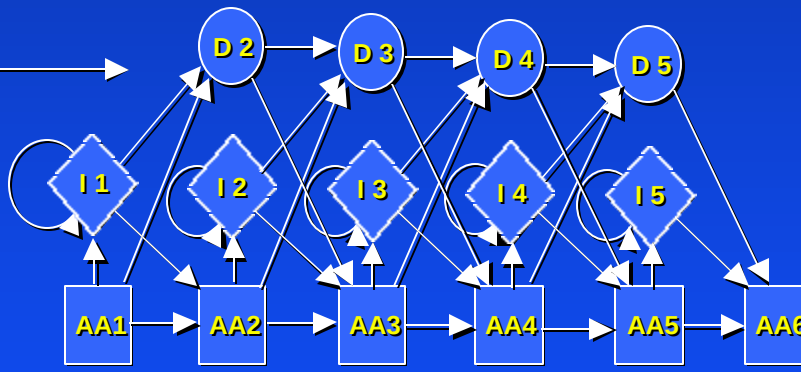
1	2	3	4	5	6	7	8	9	10	11	12
2	1	3	13	10	12	67	4	13	9	1	2
7	5	8	9	4	0	1	16	7	0	1	0
0	8	0	1	0	0	0	2	1	1	10	0
0	1	0	1	13	0	0	12	1	0	4	0
0	0	1	0	0	0	0	0	0	2	2	1
1	1	21	8	10	0	0	7	6	0	0	2
2	0	0	9	21	0	0	15	7	3	3	0
9	7	1	4	0	0	8	0	0	0	46	0
4	3	1	1	2	0	0	2	2	0	5	0
10	0	11	1	2	10	0	4	9	3	0	16
16	1	17	0	1	31	0	3	11	24	0	14
3	4	5	10	11	1	1	13	10	0	5	2
7	1	1	0	0	0	0	0	5	7	1	8
4	0	3	0	0	4	0	0	0	10	0	0
0	6	0	1	0	0	0	0	0	0	0	0
1	17	0	8	3	1	3	0	2	2	2	0
5	22	3	11	1	5	0	2	2	2	0	5
2	0	0	0	0	0	0	0	0	1	0	1
1	0	4	2	0	1	0	0	2	4	0	1
6	3	1	1	2	15	0	0	2	12	0	28

Consensus Sequences

Zinc Finger (C2H2 type)
CX{2,4}CX{12}HX{3,5}H

Sequences of Common Structure or Function

Hidden Markov Model

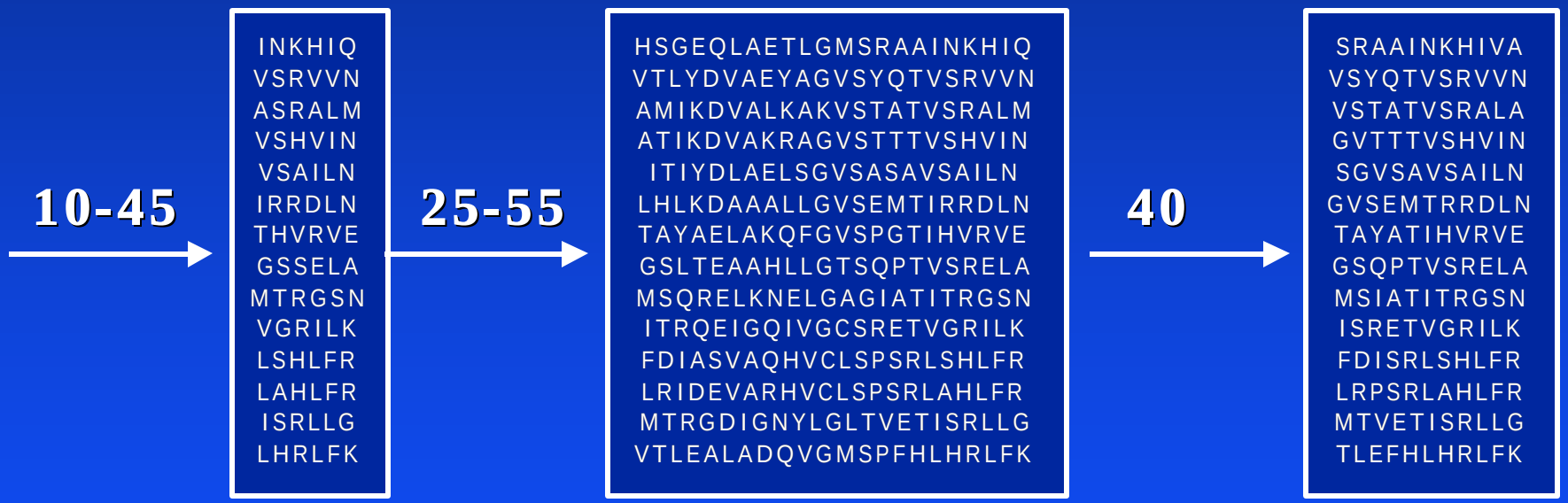


Sequence Alignments

```
          10        20        30        40          50
1   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
     |:| :|: | |:||||    | |:||| |: : :|:|  :|  |     |:  |
2   HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
          10        20        30        40        50
Initial Score    =    63  Optimized Score  =    98  Significance =  5.51
Residue Identity =   14%  Matches          =    21  Mismatches   =   22
Gaps             =     2  Conservative Substitutions          =   11
```

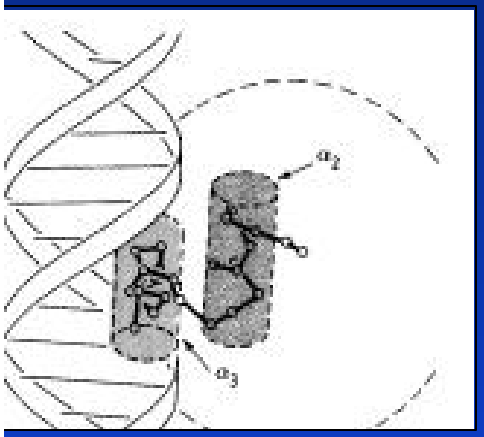
Block Signatures for a Protein Family

(http://www.blocks.fhcrc.org/)



eMATRIX: Position-Specific Scoring Matrices

(http://motif.stanford.edu/ematrix)



Structural or functional motif

ß

Examples of motif

HSGEQLAETLGMSRAAINKHIQ
VTLYDVAEYAGVSYQTVSRVVN
AMIKDVALKAKVSTATVSRALM
ATIKDVAKRAGVSTTTVSHVIN
ITIYDLAELSGVSASAVSAILN
LHLKDAAALLGVSEMTIRRDLN
TAYAELAKQFGVSPGTIHVRVE
GSLTEAAHLLGTSQPTVSRELA
MSQRELKNELGAGIATITRGSN
ITRQEIGQIVGCSRETVGRILK
FDIASVAQHVCLSPSRLSHLFR
LRIDEVARHVCLSPSRLAHLFR
MTRGDIGNYLGLTVETISRLLG
VTLEALADQVGMSPFHLHRLFK

Þ

Position

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	2	1	3	13	10	12	67	4	13	9	1	2	4	3	6	15	4	4	4	11	0	10
R	7	5	8	9	4	0	1	16	7	0	1	0	1	16	6	6	0	11	28	3	0	16
N	0	8	0	1	0	0	0	2	1	1	10	0	7	1	3	1	0	4	8	0	1	11
D	0	1	0	1	13	0	0	12	1	0	4	0	1	2	0	0	0	0	1	1	0	3
C	0	0	1	0	0	0	0	0	0	2	2	1	0	0	0	0	0	0	0	1	0	0
Q	1	1	21	8	10	0	0	7	6	0	0	2	1	17	7	7	0	2	12	5	2	4
E	2	0	0	9	21	0	0	15	7	3	3	0	1	6	11	0	0	2	0	1	13	6
G	9	7	1	4	0	0	8	0	0	0	46	0	6	0	7	1	0	3	1	1	0	4
H	4	3	1	1	2	0	0	2	2	0	5	0	3	3	0	2	0	2	4	5	0	2
I	10	0	11	1	2	10	0	4	9	3	0	16	0	2	0	1	26	1	0	8	16	0
L	16	1	17	0	1	31	0	3	11	24	0	14	0	2	0	1	21	1	1	12	20	0
K	3	4	5	10	11	1	1	13	10	0	5	2	1	4	1	1	0	1	8	4	5	14
M	7	1	1	0	0	0	0	0	5	7	1	8	0	0	2	0	2	0	0	2	0	1
F	4	0	3	0	0	4	0	0	0	10	0	0	0	0	1	0	0	1	1	1	11	0
P	0	6	0	1	0	0	0	0	0	0	0	0	1	12	7	0	0	0	0	0	0	3
S	1	17	0	8	3	1	3	0	2	2	2	0	37	1	24	5	0	29	3	0	1	3
T	5	22	3	11	1	5	0	2	2	2	0	5	16	4	2	38	0	4	1	0	4	3
W	2	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	2	10	0	0
Y	1	0	4	2	0	1	0	0	2	4	0	1	1	2	0	2	0	15	5	7	0	0
V	6	3	1	1	2	15	0	0	2	12	0	28	0	5	3	0	27	0	1	8	7	0

eMATRIX-Search

eMATRIX SEARCH

eMATRIX SEARCH
eMATRIX MAKER
eMATRIX SCAN

Thomas D. Wu, Craig G. Nevill-Manning, and Douglas L. Brutlag

Desired significance threshold: 10e -8

Threshold on information: 60.0

Enter sequence:

HRDLSSRNILLDHNIDPKNPVVSSRQDIKCKISDFGLSRLKKEQASQMTQSVGCIPYMAPEVFKGDSNSE
KSDVYSYGMVLFELLTSDEPQQDMKPMKMAHLAAYESYRPPIPLTTSSKWKEILTQCWDSNPDSRPTFKQ
IIVHLKEMEDQGVSSFASVPVQTIIDTGVYA

Fill in example
Clear form

Search for function

---

eMATRIX is based on minimal-risk scoring matrices, optimized for speed and accuracy. To cite this work, use:

Thomas D. Wu, Craig G. Nevill-Manning, and Douglas L. Brutlag, "Minimal-risk scoring matrices for sequence analysis", Journal of Computational Biology, 1999, in press.

eMATRIX Search Results

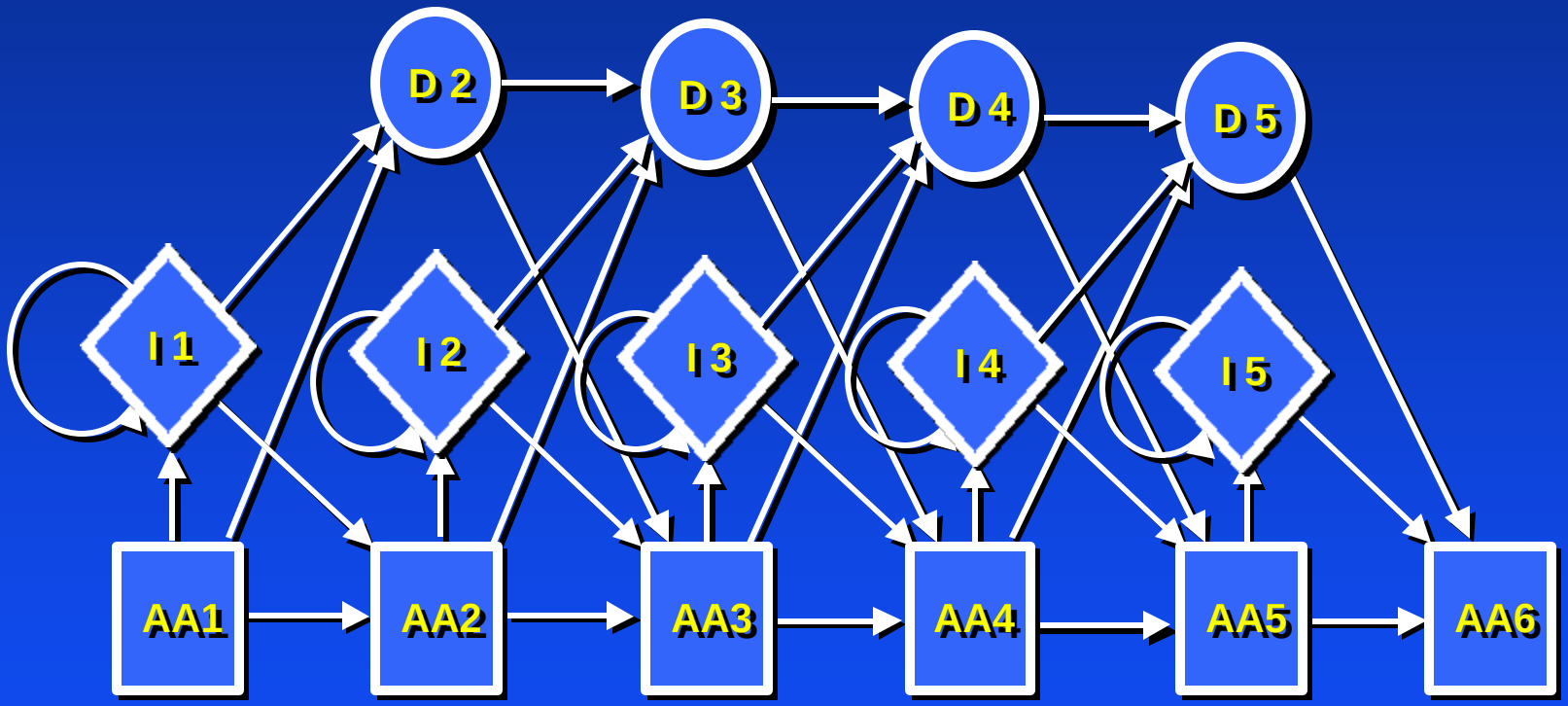
(http://motif.stanford.edu/ematrix-search)

eMATRIX SEARCH

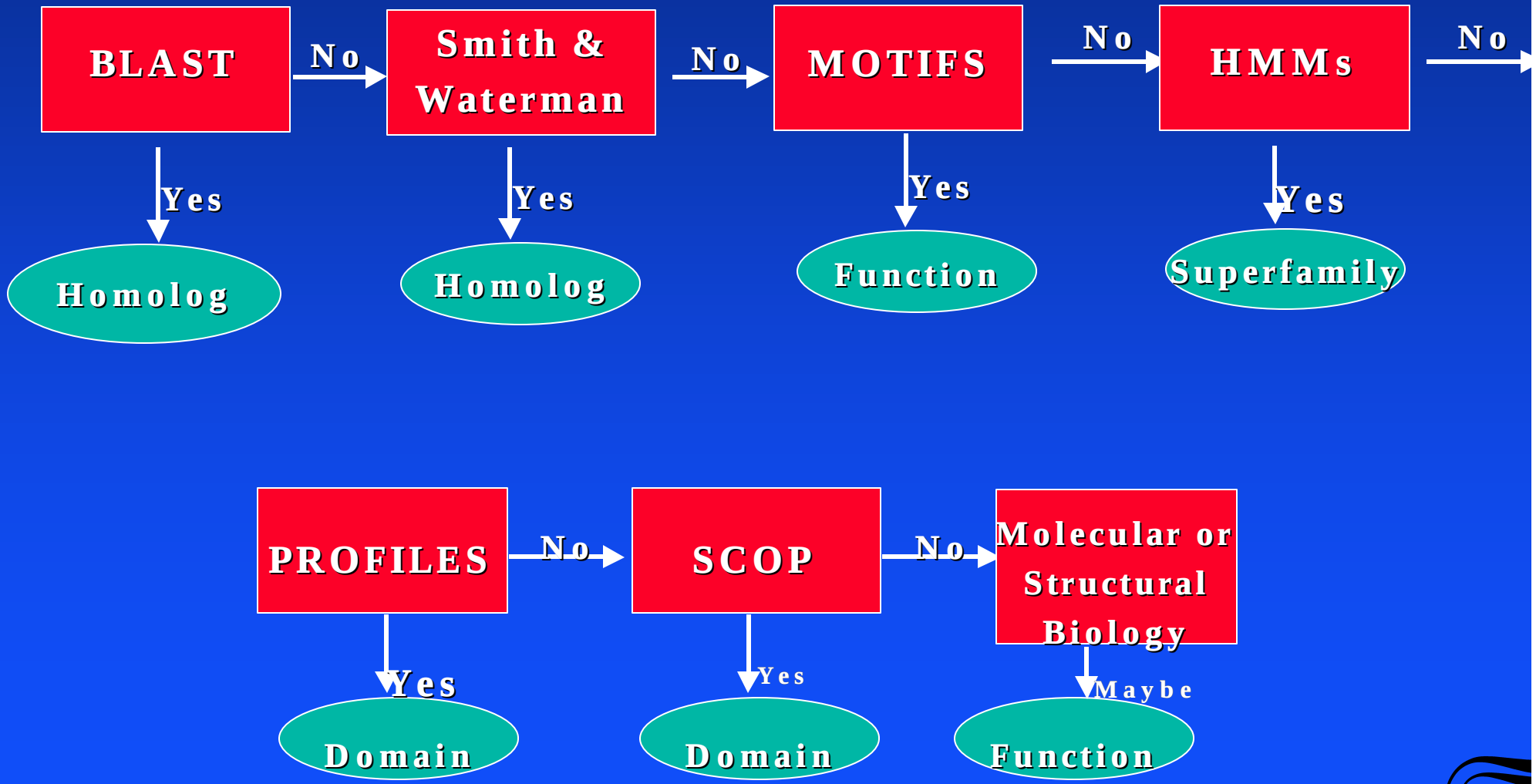
Rank	Prob.	Profile and Matching Segment
1.	5.765e-12	PR00109D TYROSINE KINASE CATALYTIC DOMAIN SIGNATURE 72 SDVYSYGMVLFELLTSDEPQQDM 95
2.	3.876e-11	BL00239E Receptor tyrosine kinase class II proteins. 43 EQASQMTQSVGCIPYMAPEVFKGDSNSEKSDVYSYGMVLFELLTSDEPQQ 93
3.	7.253e-11	BL00240G Receptor tyrosine kinase class III proteins. 89 EPQQDMKPMKMAHLAAYESYRPPIPLTTSSKWKEILTQCWDSNPDSRPTFKQI 142
4.	5.596e-10	PR00109E TYROSINE KINASE CATALYTIC DOMAIN SIGNATURE 116 TSSKWKEILTQCWDSNPDSRPTF 139
5.	4.757e-09	BL00790Q Receptor tyrosine kinase class V proteins. 108 YRPPIPLTTSSKWKEILTQCWDSNPDSRPTFKQIIVHLKEMEDQGVSSF 157
6.	6.824e-09	BL00107B Protein kinases ATP-binding region proteins. 71 KSDVYSYGMVLFELLT 87
7.	7.247e-09	BL00239F Receptor tyrosine kinase class II proteins. 97 MKMAHLAAYESYRPPIPLTTSSKWKEILTQCWDSNPDSRPTFKQI 142
8.	9.224e-09	BL00240F Receptor tyrosine kinase class III proteins. 42 KEQASQMTQSVGCIPYMAPEVFKGDSNSEKSDVYSYGMVLFELLTSDE 90

Hidden Markov Models

(after Haussler)



Protein Identification



Sequence Representations

- Consensus
- Alignment
- Blocks or Weight Matrices
- Templates or Profiles
- Bayesian Networks
- Hidden Markov Models

Deterministic

Probabilistic