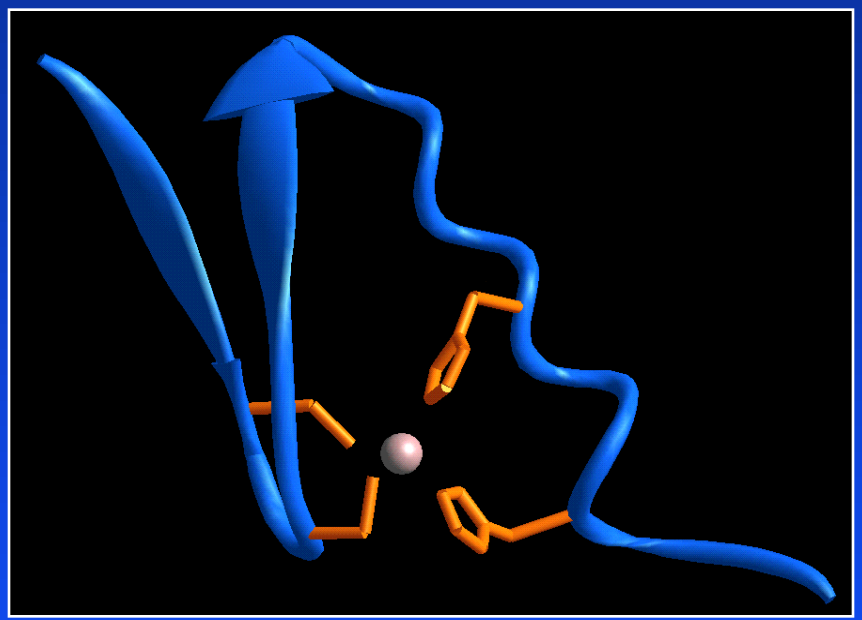
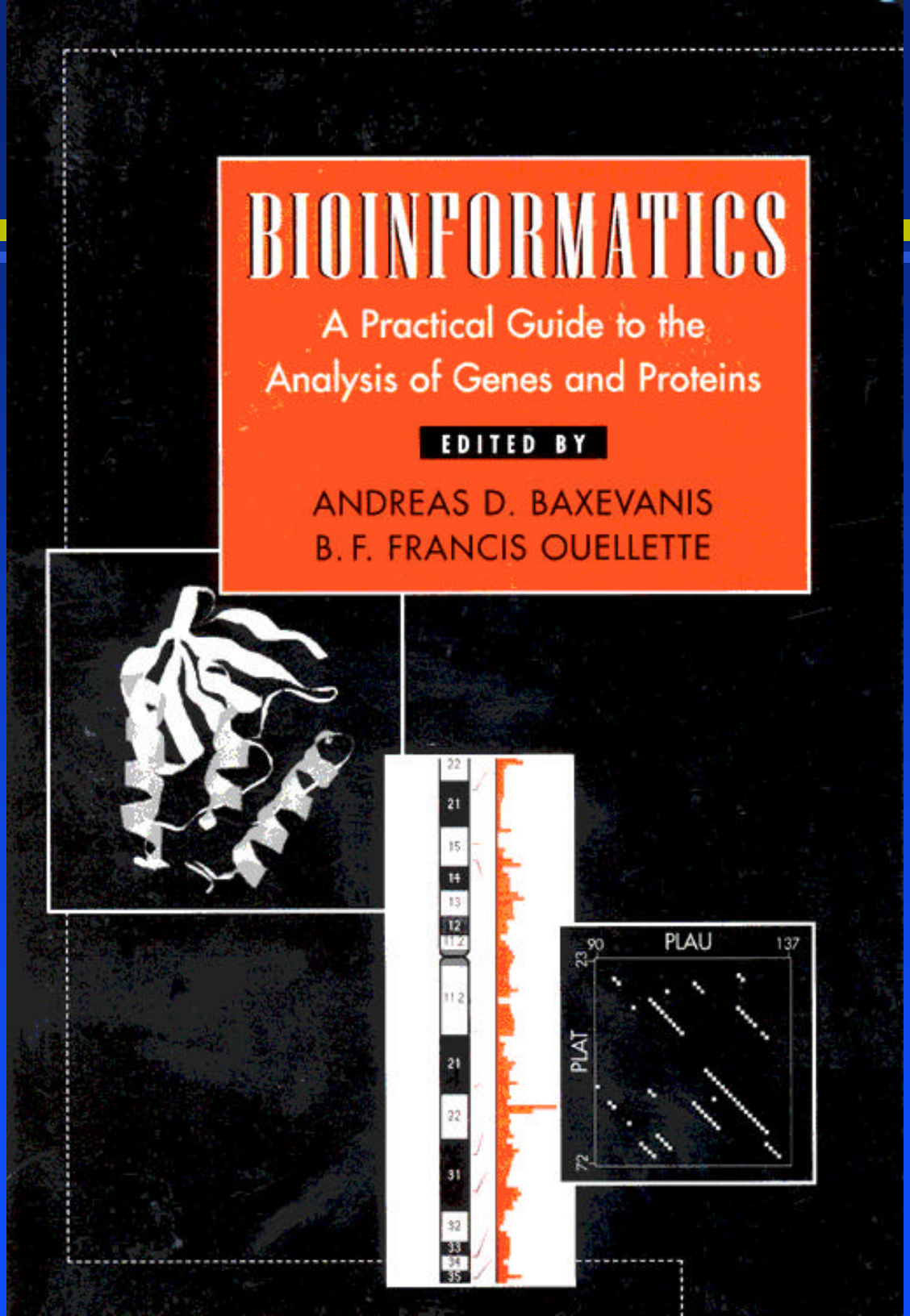
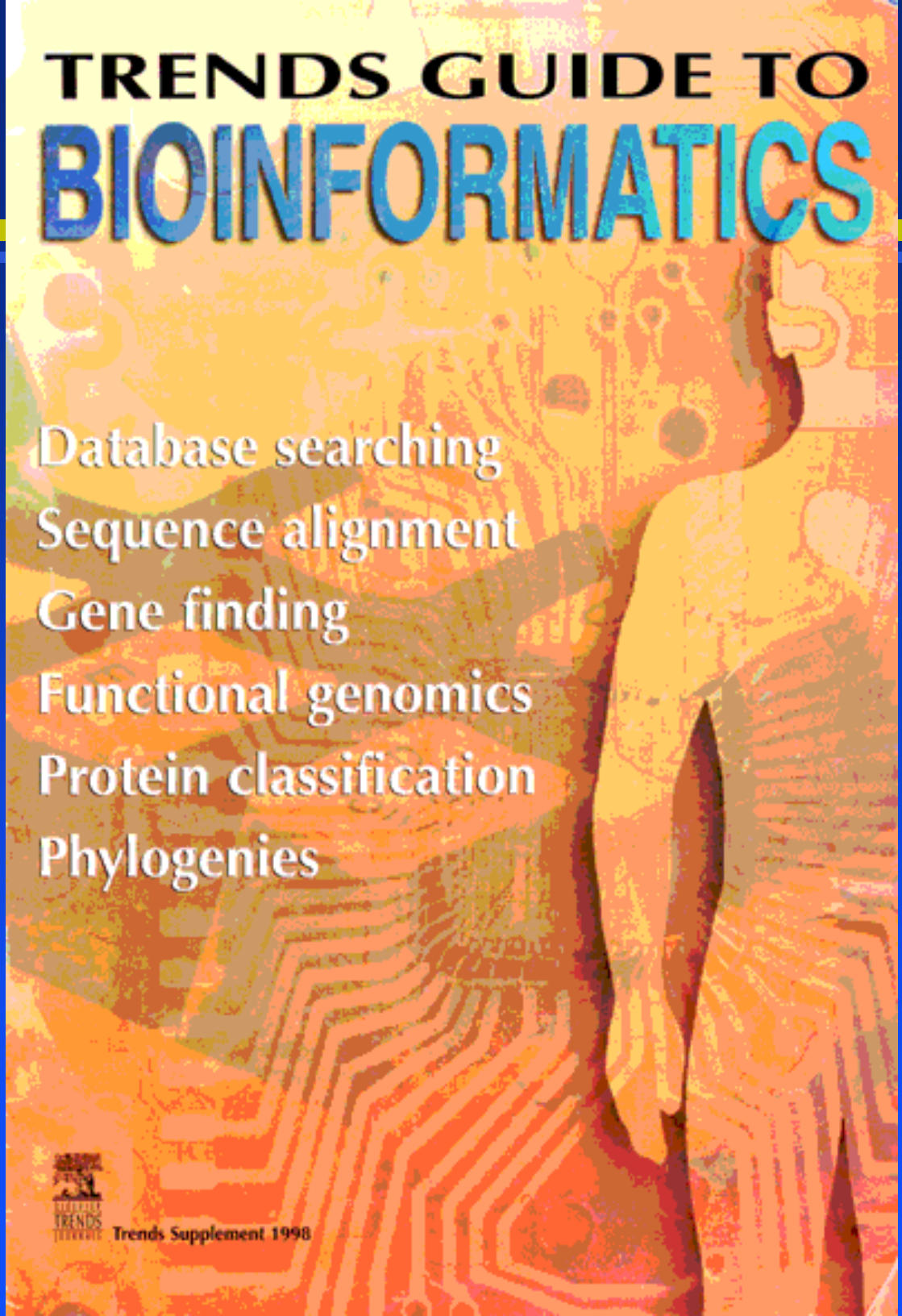
Biochemistry 201
Advanced Molecular Biology
http://cmgm.stanford.edu/biochem201/

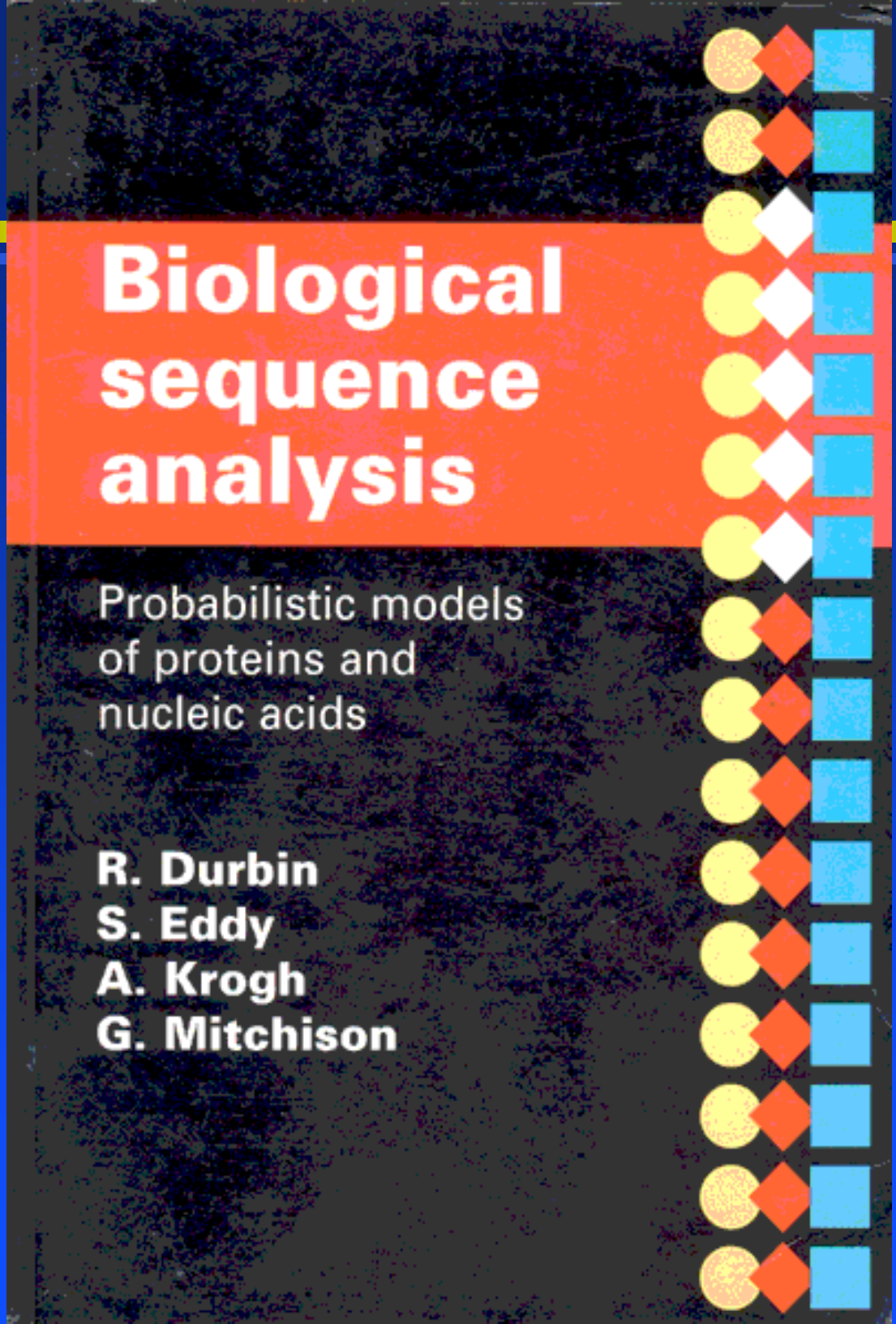
Bioinformatics: Discovering Function from Sequence



Doug Brutlag
Departments of Biochemistry
March 6, 2000







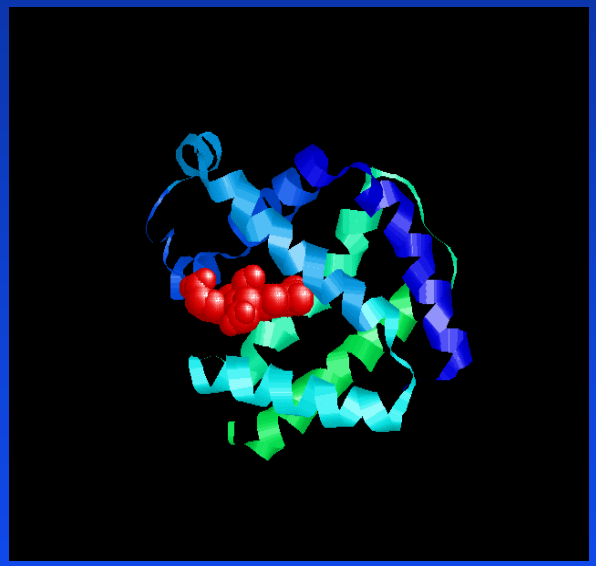
Central Paradigm of Molecular Biology

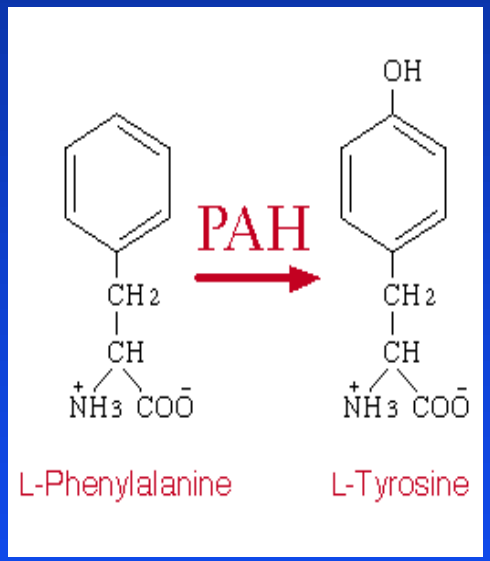
- **Molecules**
  - Structure
  - Function
- **Processes**
  - Mechanism
  - Specificity
  - Regulation

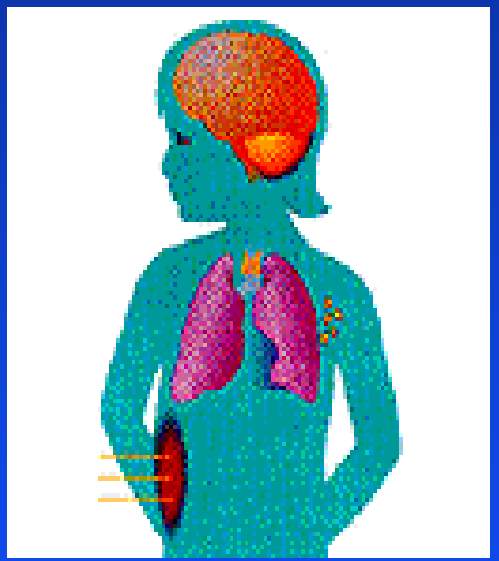
Central Paradigm of Bioinformatics

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

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



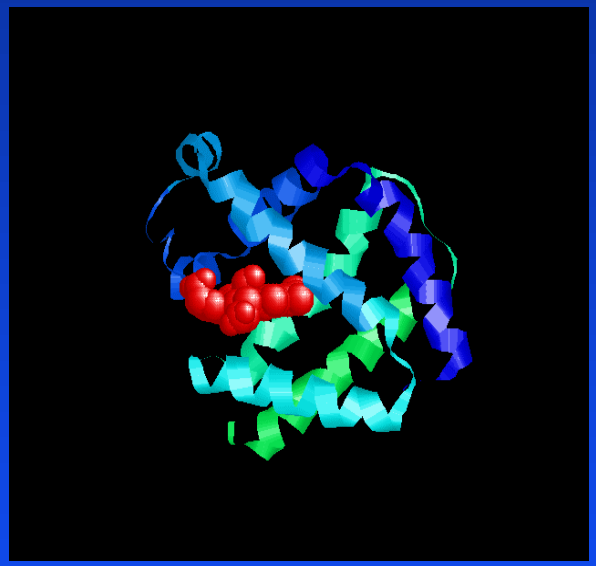


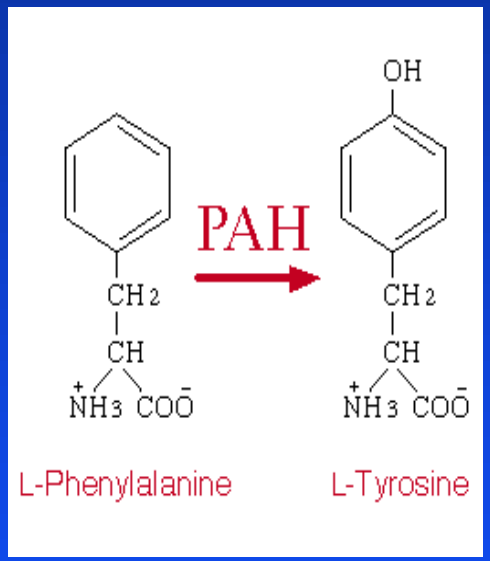


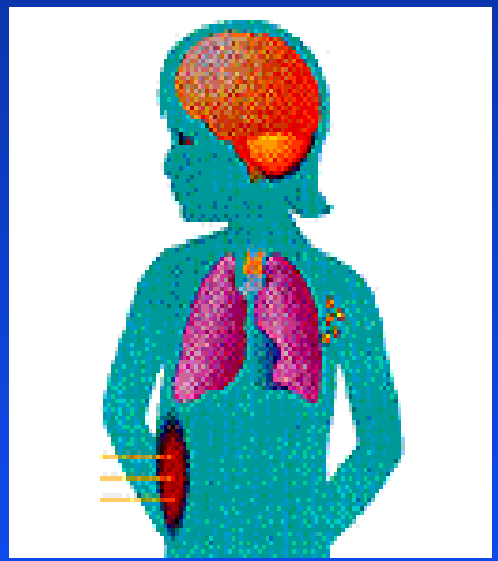
Central Paradigm of Bioinformatics

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

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







Discovering Function from Protein Sequence

Sequences of Common
Structure or Function

Sequence Alignments

```
          10        20        30        40              50
1   VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
     |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
2   HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
           10          20        30        40        50
```

Discovering Function from Protein Sequence

Consensus Sequences

```
Zinc Finger  (C2H2 type)
C.{2,4} C.{12} H.{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
```

Discovering Function from Protein Sequence

BLOCK, Weight Matrix or Position Specific Scoring Matrix

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

```
Zinc Finger  (C2H2 type)
C.{2,4} C.{12} H.{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
```

Discovering Function from Protein Sequence

BLOCK, Weight Matrix or Position Specific Scoring Matrix

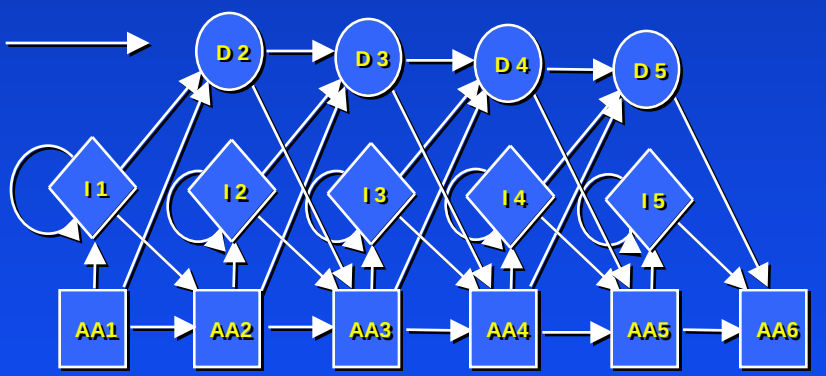
	Position											
	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	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

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

Sequences of Common Structure or Function

Profiles, PSI-BLAST Hidden Markov Models



Sequence Alignments

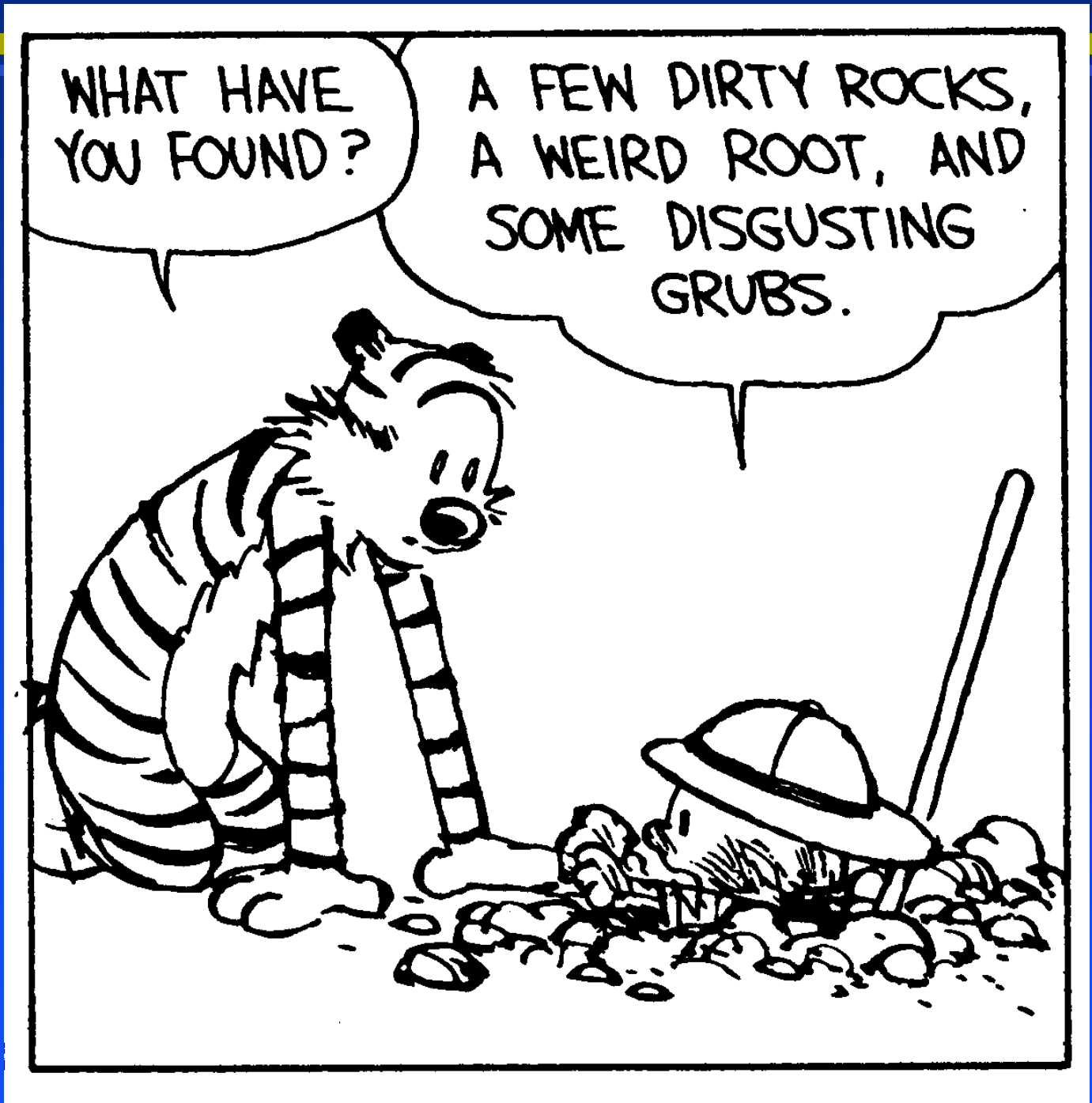
```
                10        20        30        40                50
1     VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
       |:| :|: | |:||||     | |:||| |: : :|:|  :|  |      |:  |
2     HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                10        20        30        40        50
```

Data Mining

http://www.calvinandhobbes.com/

Data Mining

http://www.calvinandhobbes.com/



Data Mining

http://www.calvinandhobbes.com/

Prosite Consensus Patterns

http://www.expasy.ch/sprot/prosite.htm

- Active site of trypsin-like serine proteases

G D S G G

- Zinc Finger (C_2H_2 type)

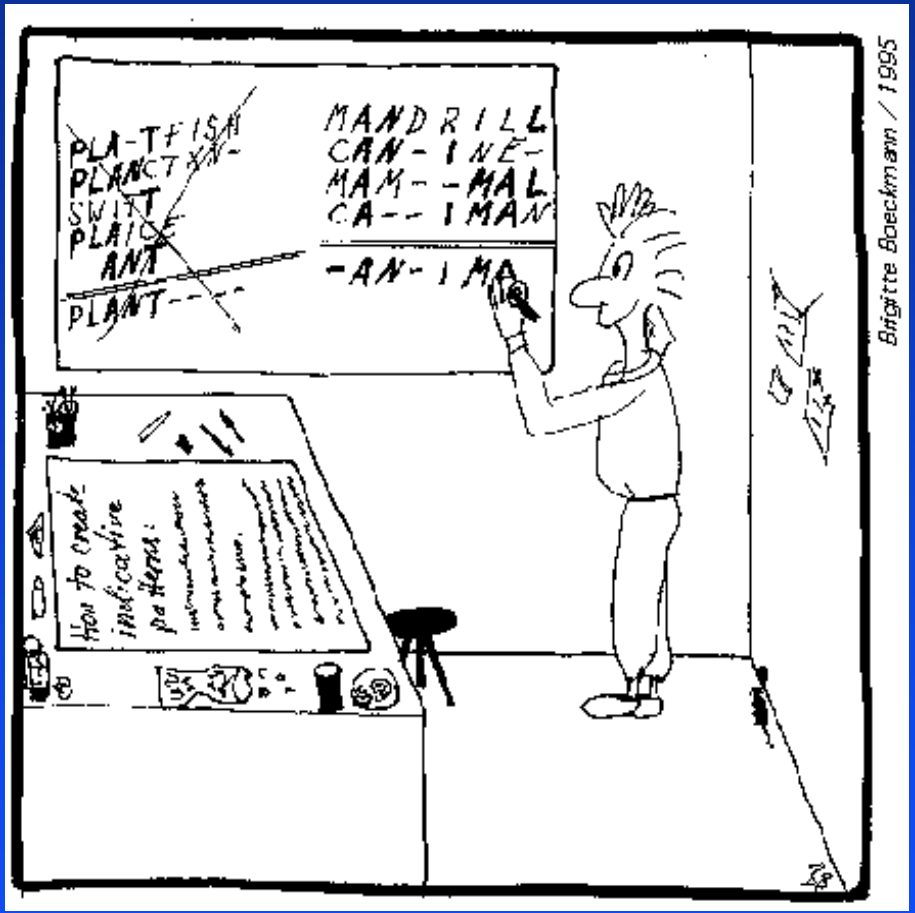
C .{2,4} C .{12} H .{3,5} H

- Homeobox Domain Signature

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

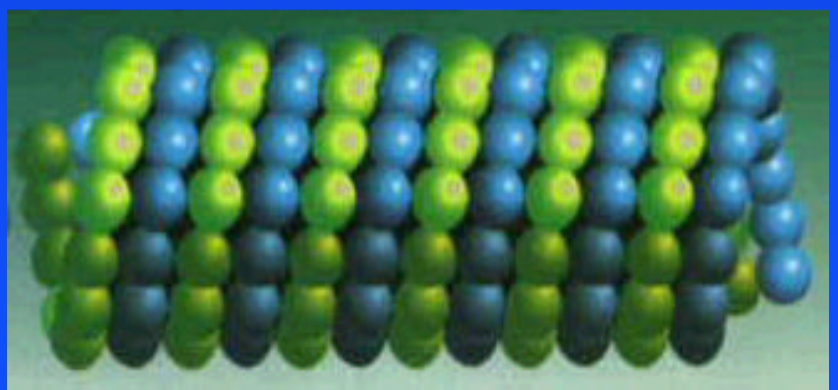
The Optimal Way to Develop Patterns

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



Generating Motifs from Aligned Protein Sequences

```
TEAESNMNDPVAEYQQY
TDARQDLYELEVDYANL
TEARENIAVLERDFEEV
TEAESNMNDLVSEYQQY
TEVRANMNDLVAEYQQY
SEAESNMNDLVSEYQQY
TEAREDLAALEKDYEEV
TEAREDLAALERDYIEV
SEAREDLAALEKDYEEV
AEAREDLAALEKDYIEV
SEAREDLAALEKDYEEV
SEAREDLAALERDYEEV
```



Generating Motifs from Aligned Protein Sequences

```
TEAESNMNDPVAEYQQY
TDARQDLYELEVDYANL
TEARENIAVLERDFEEV
TEAESNMNDLVSEYQQY
TEVRANMNDLVAEYQQY
SEAESNMNDLVSEYQQY
TEAREDLAALEKDYEEV
TEAREDLAALERDYIEV
SEAREDLAALEKDYEEV
AEAREDLAALEKDYIEV
SEAREDLAALEKDYEEV
SEAREDLAALERDYEEV
```

```
TEARENIAVLERDFEEV
SDVESDNNDPVAEYIQL
A   A LYE  V  ANY
    Q   A  S  Q
           K
```

Generating Motifs from Aligned Protein Sequences

```
TEAESNMNDPVAEYQQY
TDARQDLYELEVDYANL
TEARENIAVLERDFEEV
TEAESNMNDLVSEYQQY
TEVRANMNDLVAEYQQY
SEAESNMNDLVSEYQQY
TEAREDLAALEKDYEEV
TEAREDLAALERDYIEV
SEAREDLAALEKDYEEV
AEAREDLAALEKDYIEV
SEAREDLAALEKDYEEV
SEAREDLAALERDYEEV
```

```
TEAREDLAALERDYEEV
S          K  I
A
```

Generating Motifs from Aligned Protein Sequences

```
TEAESNMNDPVAEYQQY
TDARQDLYELEVDYANL
TEARENIAVLERDFEEV
TEAESNMNDLVSEYQQY
TEVRANMNDLVAEYQQY
SEAESNMNDLVSEYQQY
TEAREDLAALEKDYEEV
TEAREDLAALERDYIEV
SEAREDLAALEKDYEEV
AEAREDLAALEKDYIEV
SEAREDLAALEKDYEEV
SEAREDLAALERDYEEV
```

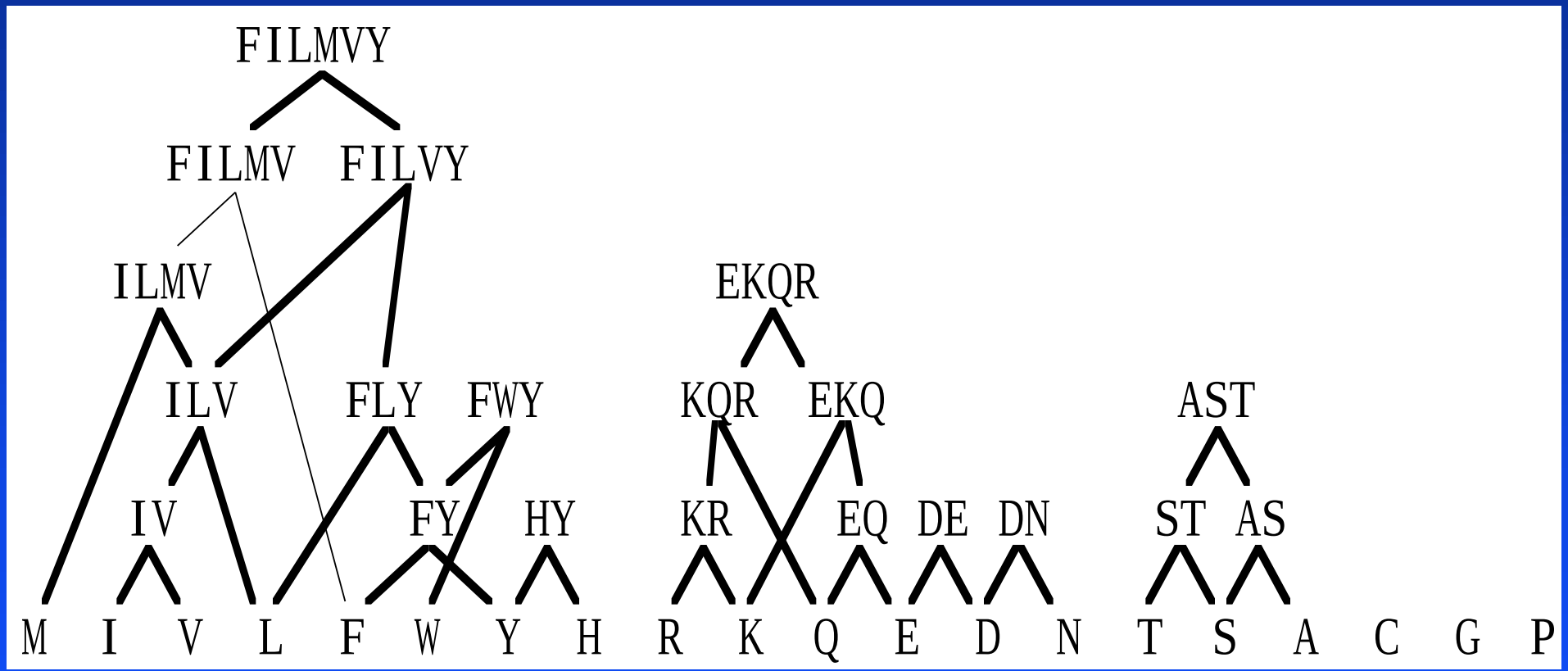
```
TEARENIAVLERDFEEV
SDVESDNNDPVAEYIQL
A   A LYE  V  ANY
    Q  A  S  Q
          K
```

Amino Acid Substitution Groups Based on Physical Properties

Only allow groups of amino acids sharing some chemical or physical property

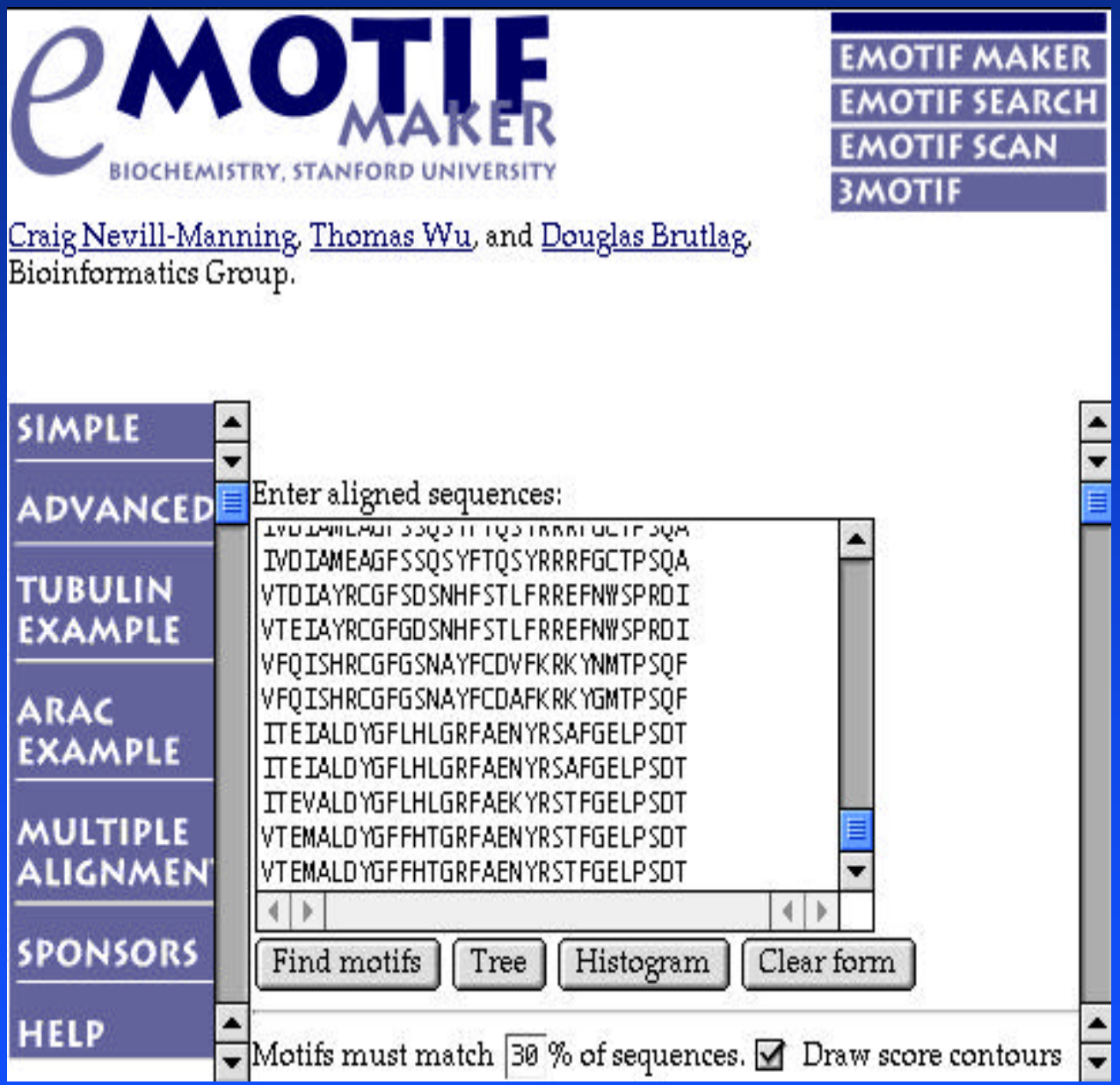
Group	Property
AG	Tiny
ST	Hydroxyl
PAGST	Small
QN	Glutamine/Glutamate
QNED	Acidic/Polar
KR	Strongly Basic
VLI	Small hydrophobic
VLIM	Small hydrophobic
FYW	Aromatic
KRH	Basic
DE	Acidic

Allowable Amino Acid Substitution Groups



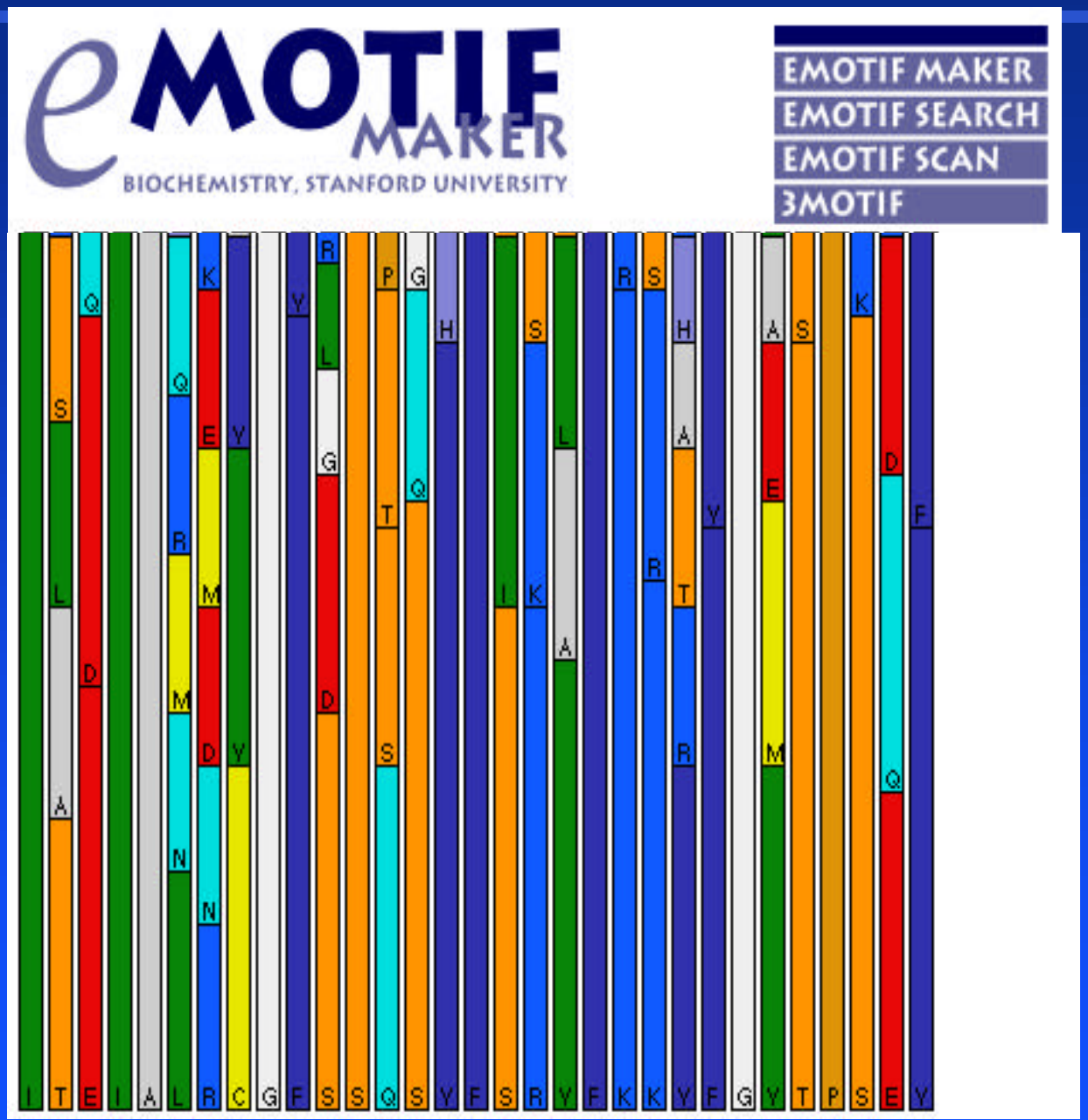
Finding *e*MOTIFs

http://motif.stanford.edu/emotif-maker/



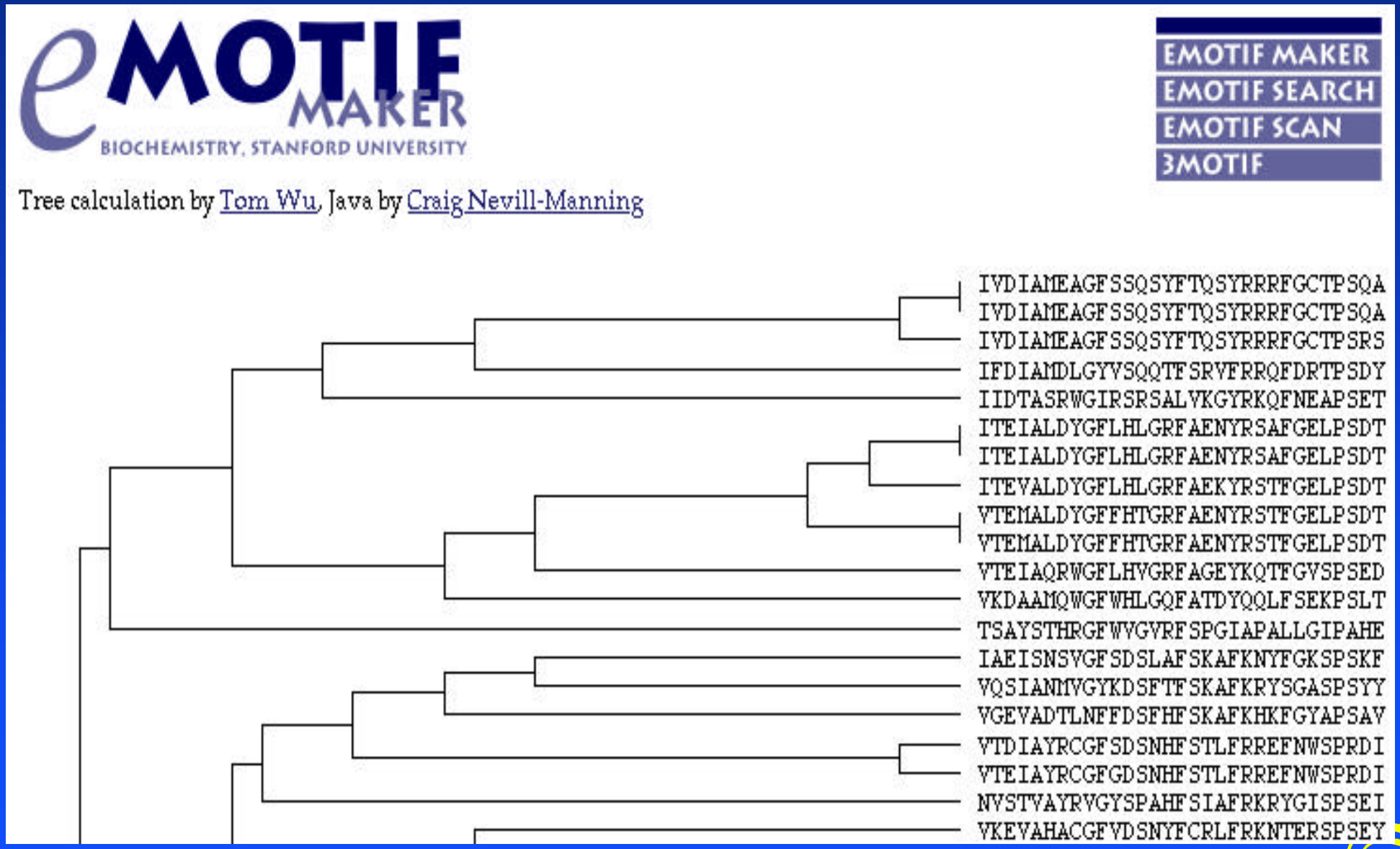
Histogram of Amino Acid Frequencies

http://motif.stanford.edu/emotif-maker/

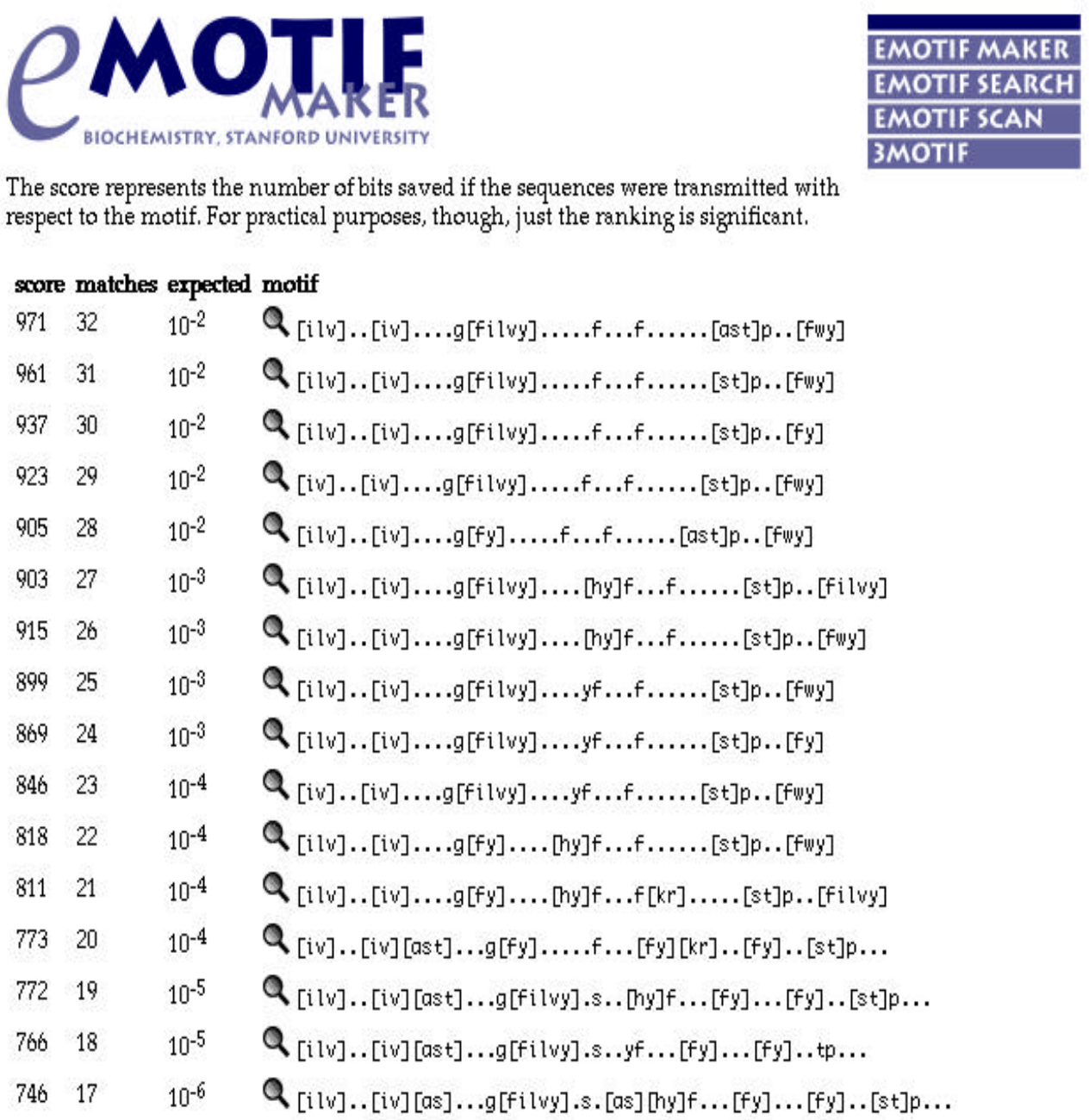


Motif Dendrogram

http://motif.stanford.edu/emotif-maker/



*e*MOTIFs Generated from the Helix-turn-helix Region of the LysR Family



eMOTIF MAKER
BIOCHEMISTRY, STANFORD UNIVERSITY

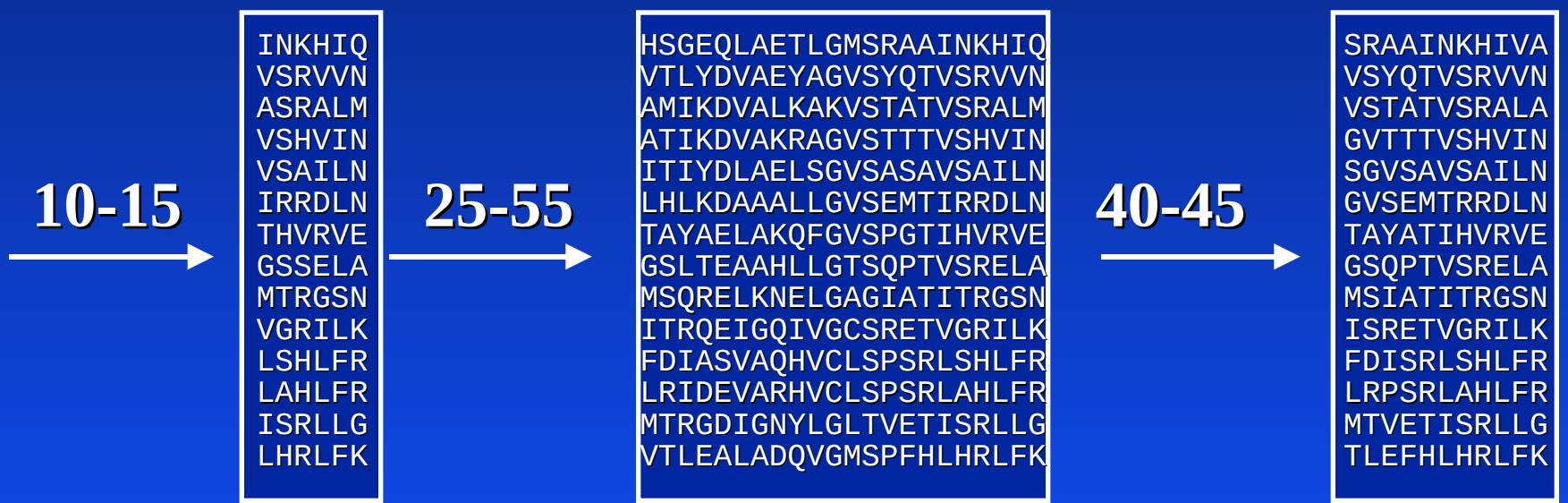
EMOTIF MAKER
EMOTIF SEARCH
EMOTIF SCAN
3MOTIF

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...

PRINTS & Block Signatures

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

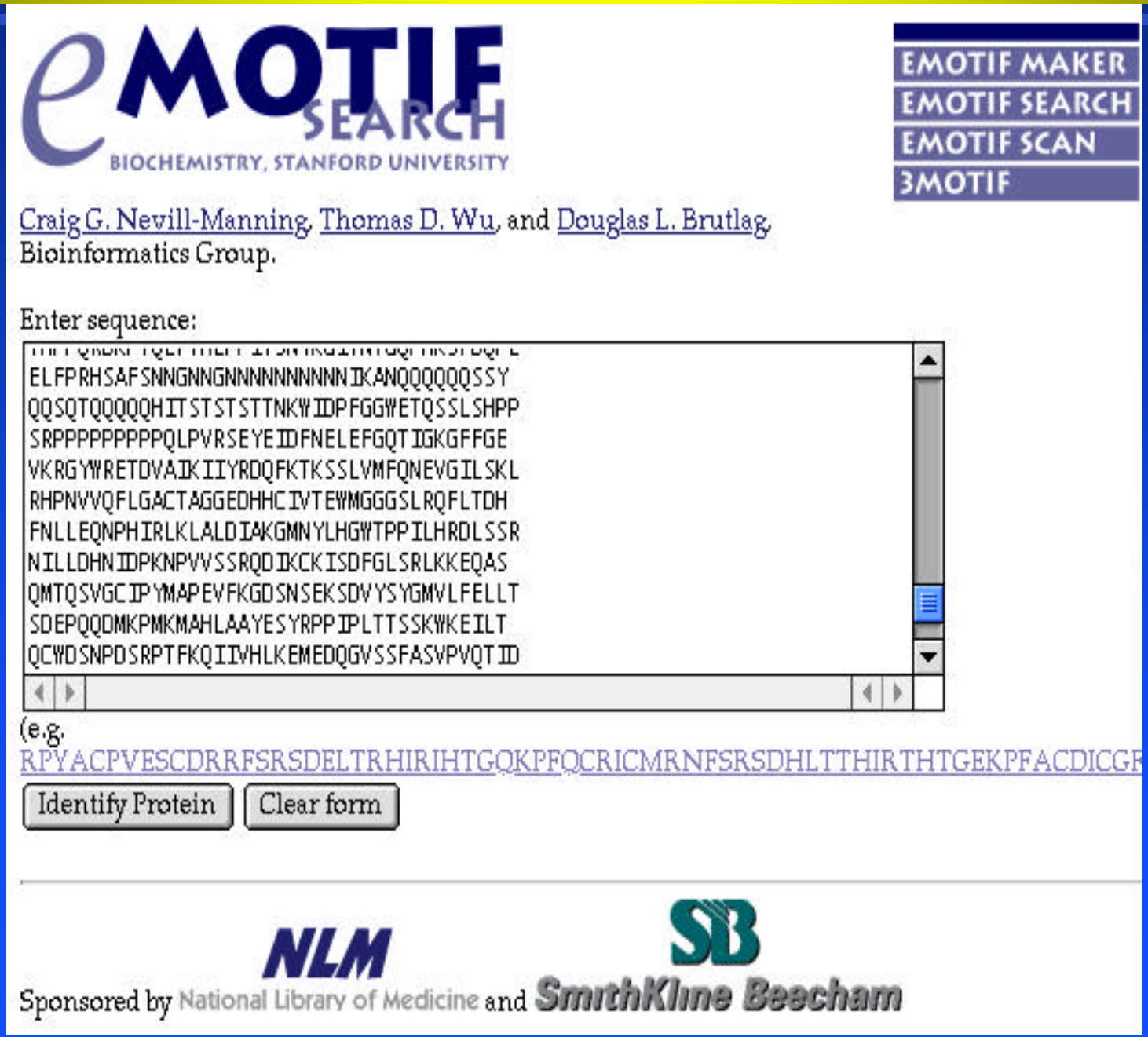


Attwood, T. K., et al., (2000).
PRINTS-S: the database formerly known as PRINTS.
Nucleic Acids Research, **28**(1), 225-227.

Henikoff, J. G., Greene, E. A., Pietrokovski, S. and Henikoff, S. (2000).
Increased coverage of protein families with the blocks database servers.
Nucleic Acids Research, **28**(1), 228-230.

Identifying Protein Function with *e*MOTIF

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



Identifying Protein Function with *e*MOTIF Search

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

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	...RPPIPLTTSSKWKEILTQCWDSNPDSRPTFKQIIVHLKEMEDQGV...	

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	...QQQHITSTSTSTTNKWIDPFGGWETQSSLSHPPSRPPP...	

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

Mapping Sequence Motifs to Structural Motifs

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

3MOTIF
MOTIFS IN THE 3RD DIMENSION
BIOCHEMISTRY, STANFORD UNIVERSITY

EMOTIF
IDENTIFY
SCAN
DECYPHER
CGNM
ALION
HOME

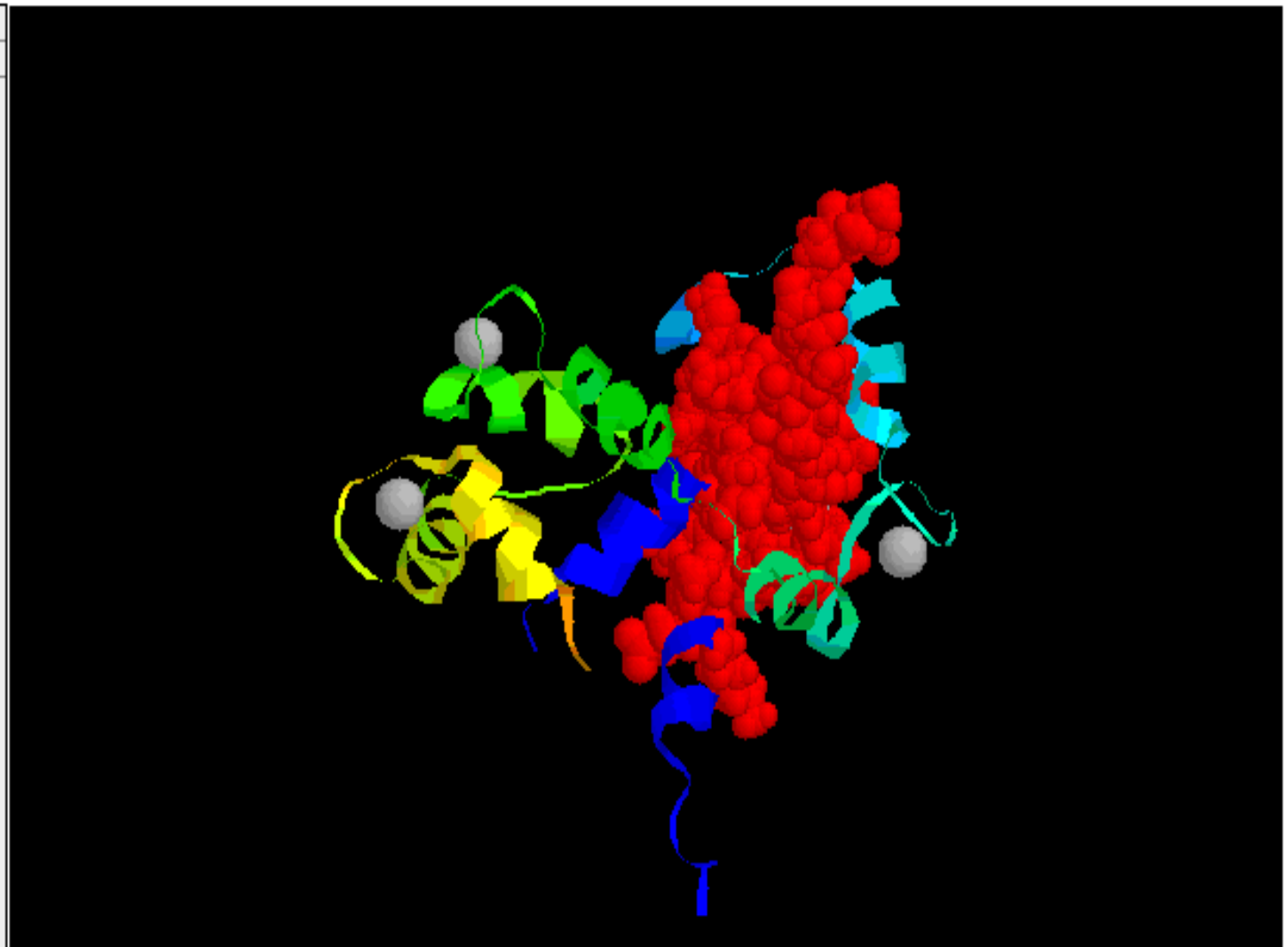
Motif:
[KR]..F.[ILMV][FILMVY]D.[DN]..C

Select
- block : 13-33
- conserved
- emotif
- all

Color
- rasmol 'amino' color code
- red
- gray
- all gray
- chain
- ligands
- custom coloring

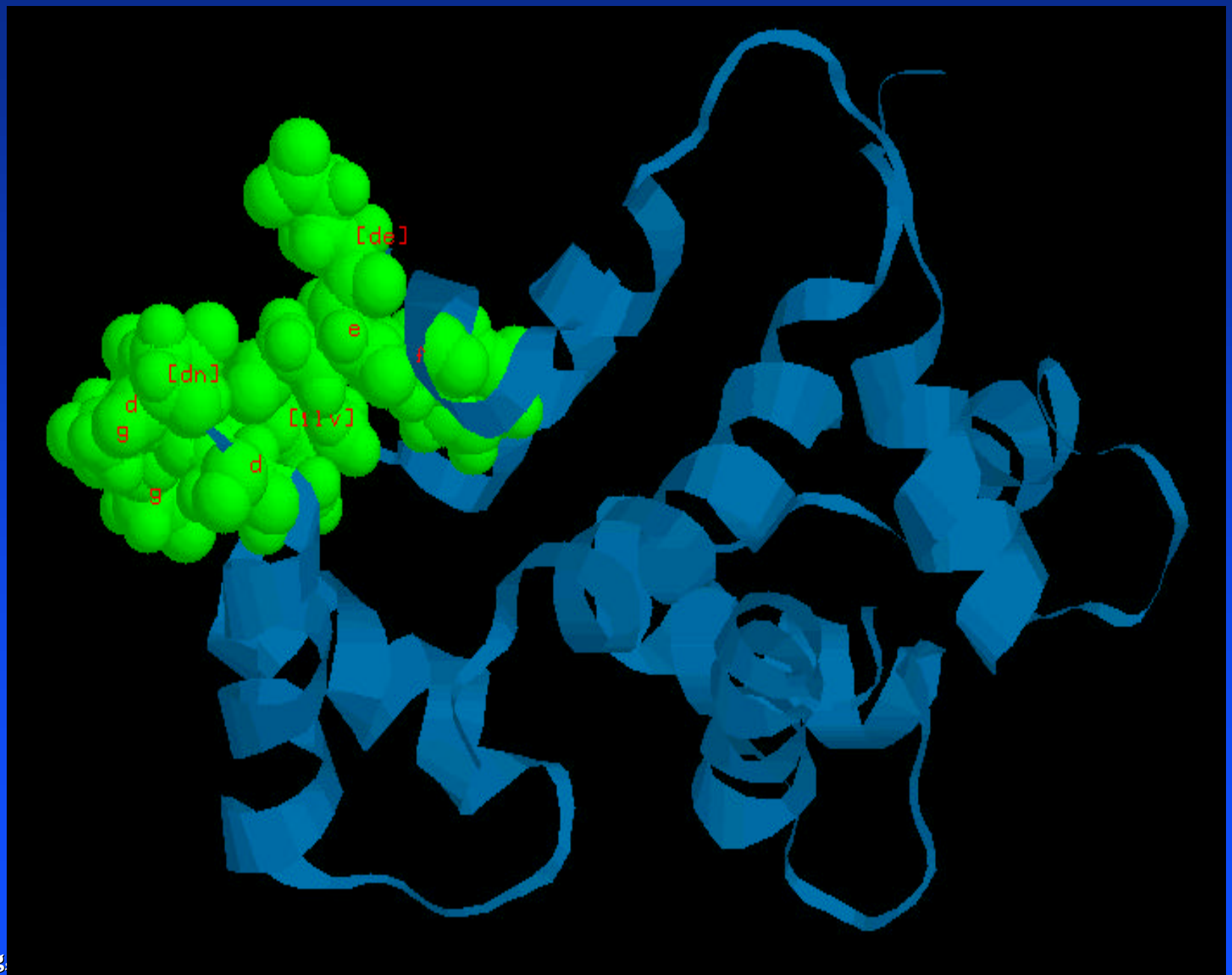
Shape
- space-filling
- ribbon

- reset to default



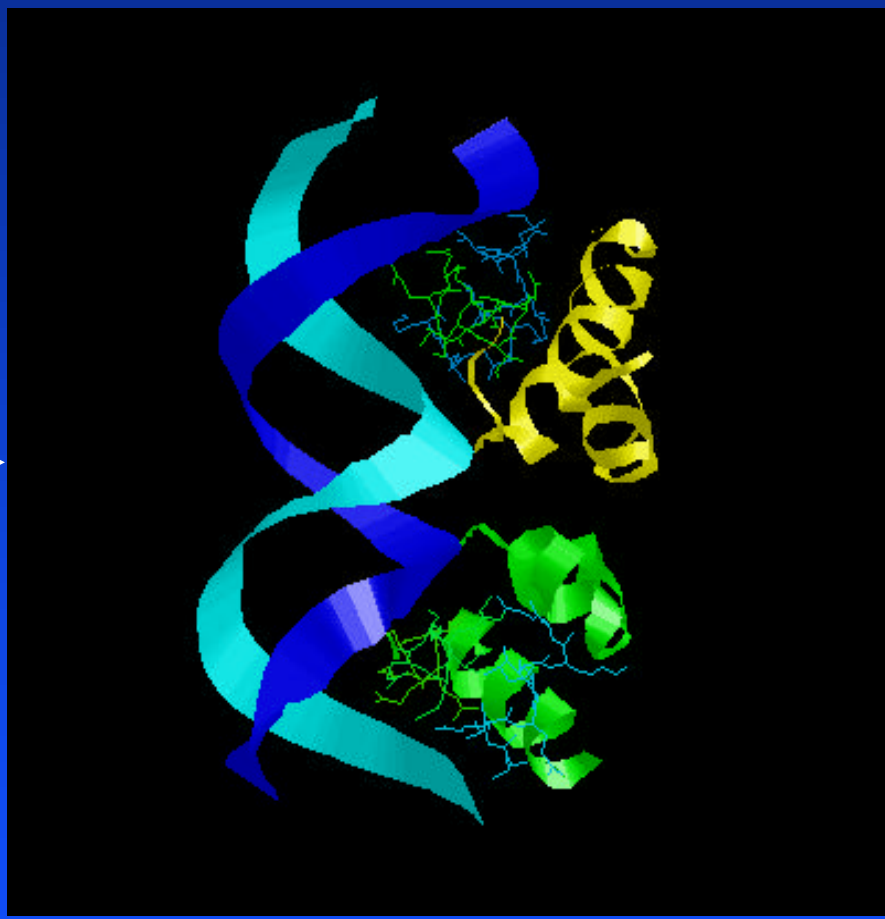
3MOTIF Labeling

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



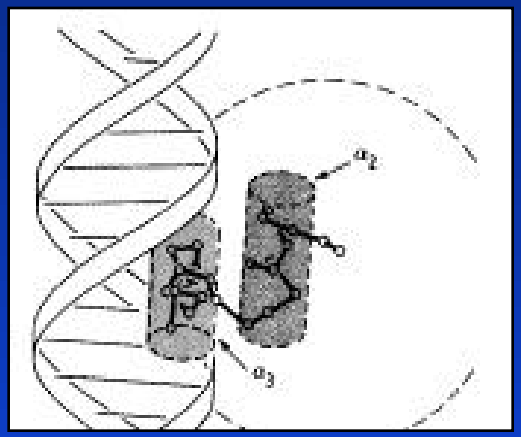
Sequence Redundancy in Prokaryotic Helix-Turn-Helix Motifs

Sequence						Helix						Turn				Helix					
RCRO$LAMBD	F	G	Q	T	K	T	**A**	K	D	L	**G**	V	Y	Q	S	A	**I**	N	K	A	I	H
RCRO$BP434	M	T	Q	T	E	L	**A**	T	K	A	**G**	V	K	Q	Q	S	**I**	Q	L	I	E	A
RCRO$BPP22	G	T	Q	R	A	V	**A**	K	A	L	**G**	I	S	D	A	A	**V**	S	Q	W	K	E
RPC1$LAMBD	L	S	Q	E	S	V	**A**	D	K	M	**G**	M	G	Q	S	G	**V**	G	A	L	F	N
RPC1$BP434	L	N	Q	A	E	L	**A**	Q	K	V	**G**	T	T	Q	Q	S	**I**	E	Q	L	E	N
RPC1$BPP22	I	R	Q	A	A	L	**G**	K	M	V	**G**	V	S	N	V	A	**I**	S	Q	W	E	R
RPC2$LAMBD	L	G	T	E	K	T	**A**	E	A	V	**G**	V	D	K	S	Q	**I**	S	R	W	K	R
LACR$ECOLI	V	T	L	Y	D	V	**A**	E	Y	A	**G**	V	S	Y	Q	T	**V**	S	R	V	V	N
CRP$ECOLI	I	T	Q	Q	E	I	**G**	Q	I	V	**G**	C	S	R	E	T	**V**	G	R	I	L	K
TRPR$ECOLI	M	S	Q	R	E	L	**K**	N	E	L	**G**	A	G	I	A	T	**I**	T	R	G	S	N
RPC1$CPP22	R	G	Q	R	K	V	**A**	D	A	L	**G**	I	N	E	S	Q	**I**	S	R	W	K	G
GALR$ECOLI	A	T	I	K	D	V	**A**	R	L	A	**G**	V	S	V	A	T	**V**	S	R	V	I	N
Y77$BPT7	L	S	H	R	S	L	**G**	E	L	Y	**G**	V	S	Q	S	T	**I**	T	R	I	L	Q
TER3$ECOLI	L	T	T	R	K	L	**A**	Q	K	L	**G**	V	E	Q	P	T	**L**	Y	W	H	V	K
VIVB$BPT7	D	Y	Q	A	I	F	**A**	Q	Q	L	**G**	G	T	Q	S	A	**A**	S	Q	I	D	E
DEOR$ECOLI	L	H	L	K	D	A	**A**	A	L	L	**G**	V	S	E	M	T	**I**	R	R	D	L	N
RP32$BACSU	R	T	L	E	E	V	**G**	K	V	F	**G**	V	T	R	E	R	**I**	R	Q	I	E	A
Y28$BPT7	E	S	N	V	S	L	**A**	R	T	Y	**G**	V	S	Q	Q	T	**I**	C	D	I	R	K
IMMRE$BPPH	S	T	L	E	A	V	**A**	G	A	L	**G**	I	Q	V	S	A	**I**	V	G	E	E	T



Position-Specific Scoring Matrices

http://expasy.hcuge.ch/www/tools.html



Structural or functional motif

⇓

Examples of motif

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

.

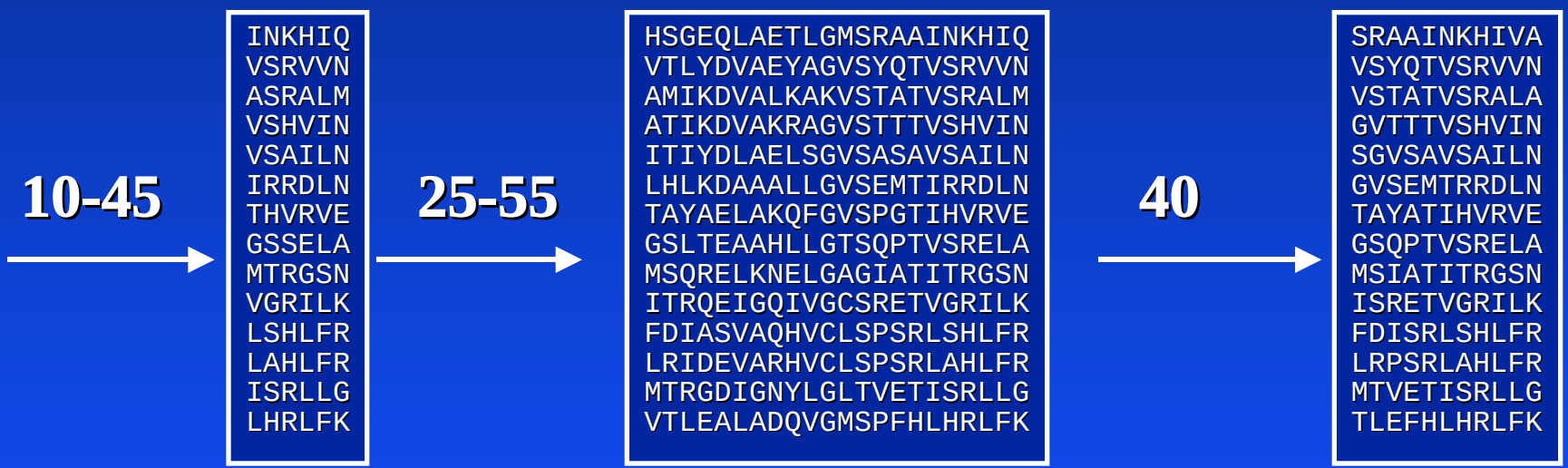
Position

	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

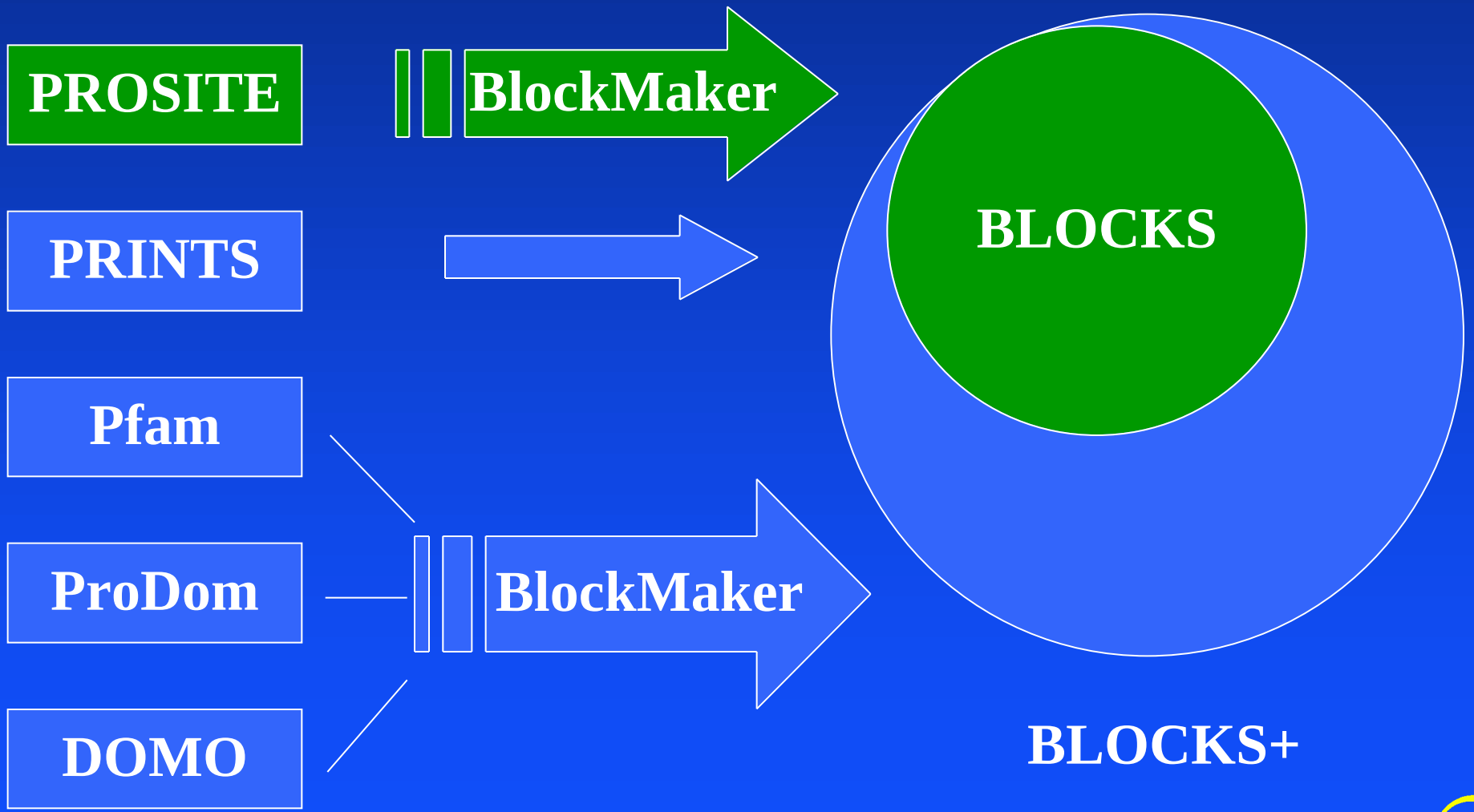
Block Signatures for a Protein Family

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

(After Henikoff and Henikoff)

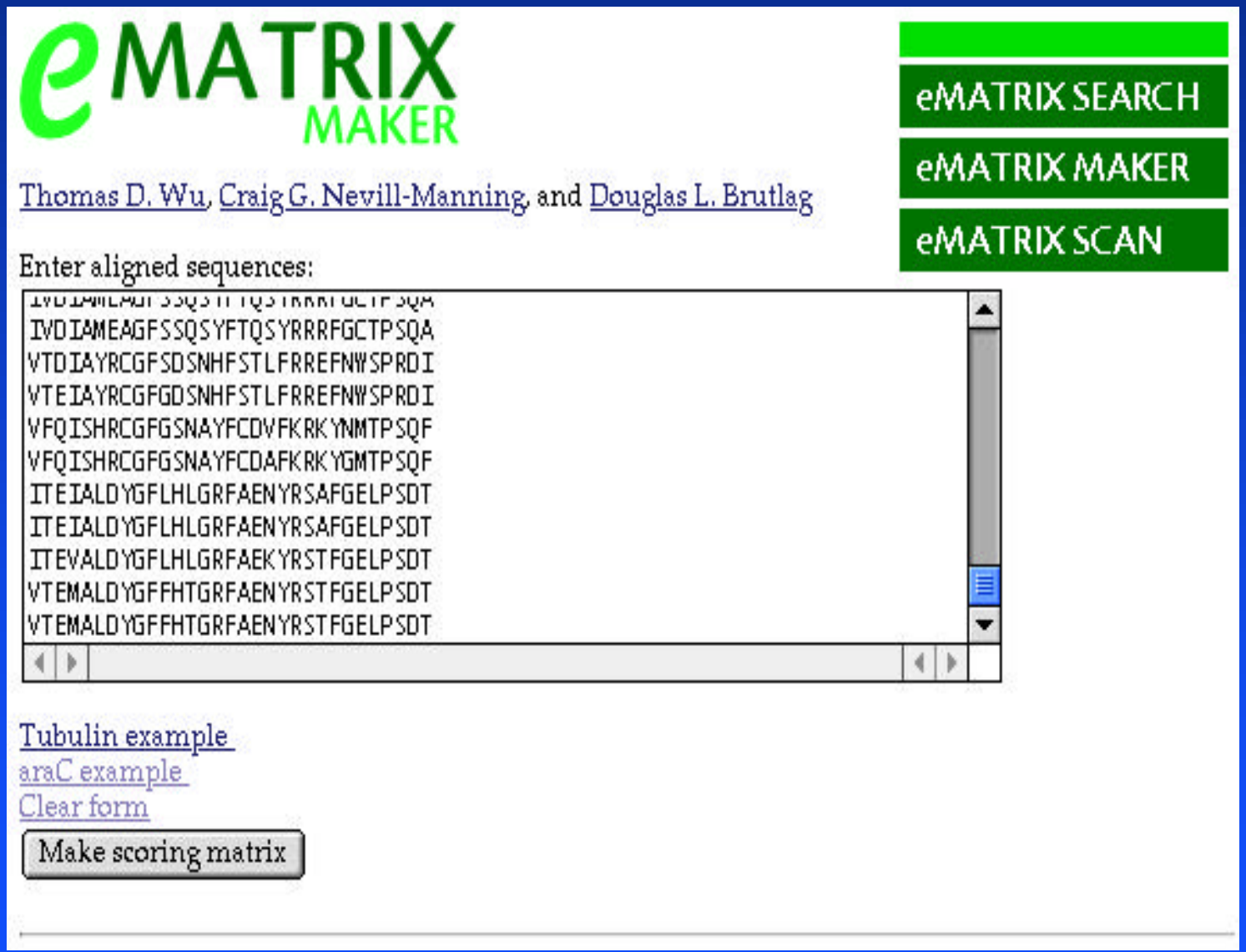


BLOCKS+ Is Based On Several Protein Family Databases



eMATRIX Maker

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



eMATRIX SEARCH

To paste this matrix into eMATRIX-SCAN, click here.

```
ID
AC
DE
MA
  A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   X   Y   Z   *   -
-51 -36 -46 -60 -63 -46 -63 -59  38 -59  -8 -40  -4 -64 -61 -60 -58  12  21 -61 -36 -54 -62   0   0
  6  -9 -17 -14 -23  -2   0  -4   2   8   6 -15  -2 -28 -10 -19   7  15   2 -28  -4 -22 -15   0   0
  4   8 -33  24  22 -37 -32 -26 -36 -25 -38  -9 -12 -37  19  -5   8  -4 -20 -40 -13  -7  20   0   0
  3 -59 -43 -59 -60  -3 -58 -53  37 -56 -10  -4 -58 -61 -58 -57 -54   1  16 -53 -29   9 -59   0   0
 35 -57 -48 -58 -58 -59  -1 -56 -53 -55  -7 -53 -56 -60 -57 -58  19 -16  -9 -64 -31 -60 -57   0   0
  5  -1  -1  -7  -2  -7 -21  17  -5 -15   5  20   7 -27   9   4  -7   2  -4 -23  -3   3   4   0   0
 -7   9   5   8   3 -17 -19  13  -9   5  -6  14  10 -24  13  10  -3  -4 -11 -23  -3   1   8   0   0
  4 -21  37 -33 -32 -18 -31   1   3 -26  -3 -20  -6 -37  -6   8 -17 -14  14  27 -10  21 -18   0   0
-60 -42 -67 -68 -73 -15  42 -68 -69 -68 -74 -70 -10 -77  -2 -17 -66 -70   7 -74 -46 -74 -34   0   0
 -4 -59 -48 -60 -60  42 -10 -40  10 -57 -41 -43 -57 -64 -58 -56 -56 -54 -43 -36 -36  38 -59   0   0
 -6   7 -23  13  -6   1  -2  -3 -23 -12  -7  12  -1  -6  -4   5  14 -14 -11  30  -3   2  -5   0   0
-53   8 -58  21 -57 -67 -59  26 -64 -58 -68 -64 -10 -66 -14 -61  36 -54  -3 -72 -40 -63 -33   0   0
 -1 -18 -24 -26  -2 -31  -2 -24 -11 -22  -3 -23  -8  17  22  -8   9  13  -1 -35  -6 -31  11   0   0
  0 -21 -37 -40 -40  18   7 -37 -39 -38  -9 -38   2 -11  19 -40  26 -36  -2 -44 -14  -7  -7   0   0
 -4 -33 -45 -48 -45 -29 -49  38 -46 -40 -25 -43 -14 -55  -5  17 -11   6 -45 -36 -24  42 -23   0   0
-84 -91 -78 -92 -93  52 -89 -77 -68 -89   6 -68 -90 -96 -92 -87 -89 -85 -73 -67 -70 -56 -92   0   0
  7 -42  23 -44 -45 -41 -42  16  20 -43 -40 -39 -40 -49 -44 -45  27  11 -12 -48 -22   7 -45   0   0
-28  -5  -2 -10  -2 -35 -18 -25  -7  16 -33 -30   1   2  -1  28   4   7  -8 -39  -8 -11  -1   0   0
 10  -5 -18  -7   3 -22   3 -24   1 -10   7 -15  -2 -31 -23  -5  -9   4  17 -31  -4 -26 -12   0   0
-69 -76 -64 -77 -77  49 -74 -58  11 -74 -54 -56 -74 -81 -76 -72 -74 -70 -57 -51 -57  28 -77   0   0
 -5 -27 -55 -50 -46 -60 -22 -46 -58  37 -58 -52   2 -58  -4  27 -20 -21 -28 -62 -28 -56 -23   0   0
  2  -3 -33  -8   0 -38 -32   3 -38  27 -38 -33   5   2  10  17   0  -8 -35 -40 -11  -7   6   0   0
  2 -18   8 -24  -2  -2 -26  13  -4   3  -4 -18 -12 -31   4  12  -9   8  -4 -20  -5  26   1   0   0
-40 -46 -36 -48 -48  36 -47 -32 -11 -45   5   8 -44 -52 -46 -44 -21  17 -10 -29 -23  32 -47   0   0
-20   4 -53  -3 -10 -59  38 -51 -59 -17  -9 -55  12 -61  -5 -54  -7 -54 -58 -60 -26 -61  -7   0   0
 -7 -33  18 -33   7   4  -3 -28   6 -19  -6  31 -34 -40   0 -10 -33 -11  18  10  -8   4   3   0   0
 -1 -47 -43 -49 -50 -52 -49 -50   0 -11  -8 -46 -45  -2 -50 -49  23  34 -43 -58 -24 -53 -50   0   0
-14 -88 -88 -87 -87 -97 -89 -88 -91 -84 -94 -91 -90  51 -88 -15 -86 -87 -90 -99 -73 -95 -88   0   0
  5 -45 -45 -47 -46 -54 -17   8 -53  18 -21 -49 -43 -53  -3  10  31  -9 -50 -58 -22 -52 -22   0   0
  2   3 -27  13  18 -27  -8  20 -28   7  -6 -24 -11 -31  19  -2  -4 -24 -27 -31  -7   2  19   0   0
-18 -29 -36 -18  -3  27 -45 -28   1 -41  -7 -32 -42 -50 -42 -42 -24  10 -12  30 -18  39 -24   0   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
IIVHLKEMEDQGVSSFASVPVQTIDTGVYA

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.

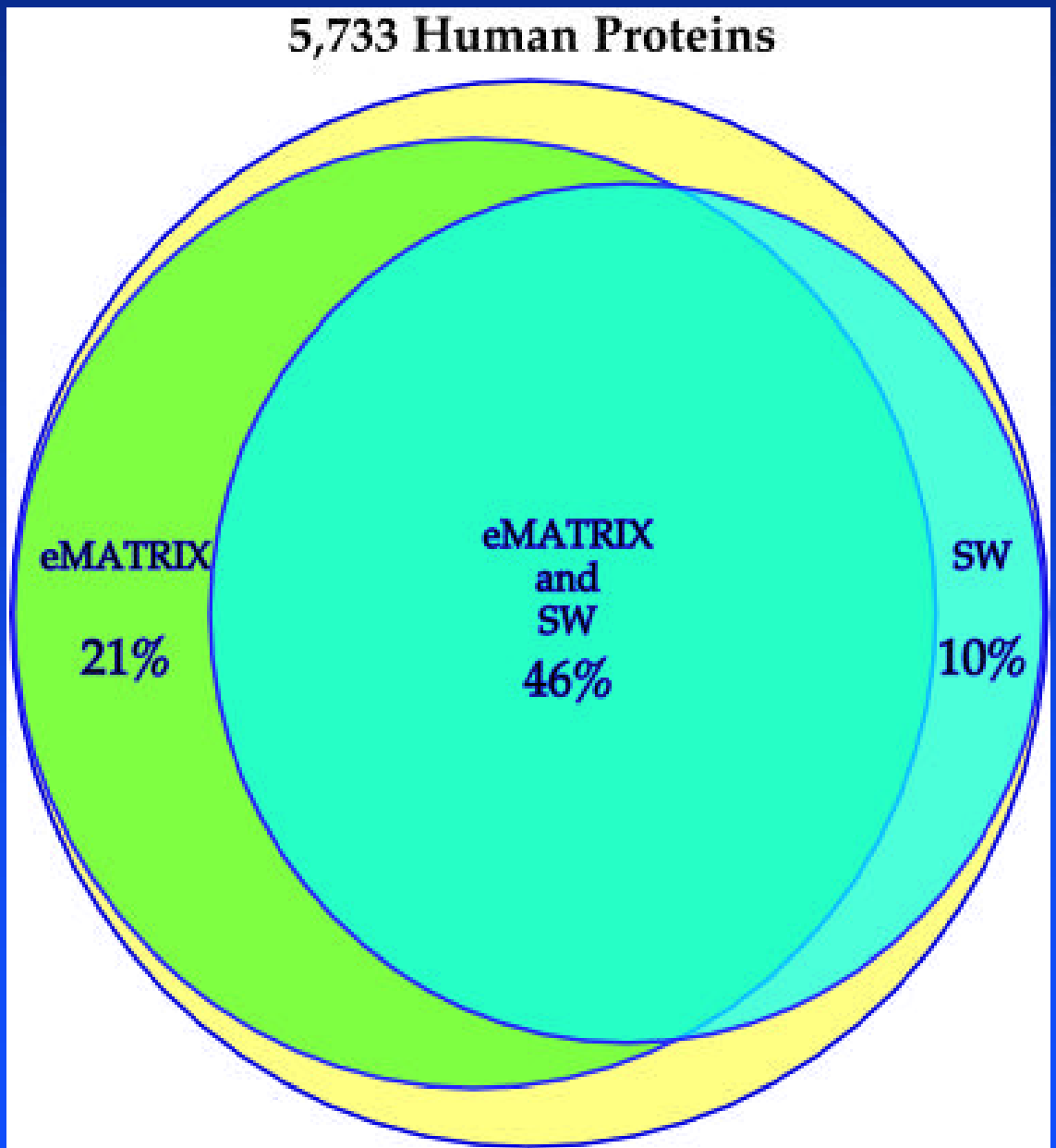
eMATRIX Search Results

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

eMATRIX SEARCH

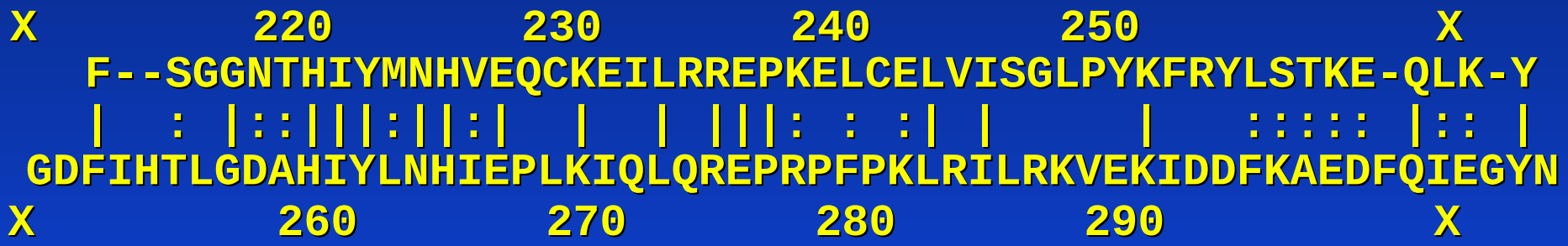
Rank	Prob.	Profile and Matching Segment
1.	5.765e-12	PR00109D TYROSINE KINASE CATALYTIC DOMAIN SIGNATURE 72 SDVYSYGMVLFELLTSDEPQQDM 95
2.	3.876e-11	BL00239E Receptor tyrosine kinase class II proteins. 43 EQASQMTQSVGCIPYMAPEVFKGDSNSEKSDVYSYGMVLFELLTSDEPQQ 93
3.	7.253e-11	BL00240G Receptor tyrosine kinase 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

eMATRIX & Similarity Search Identify 77% of Human Proteins



Sequence Alignment

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



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

where:

Penalty = Gap-penalty + Size-of-gap x Gap-size-penalty

Smith-Waterman Similarity Search

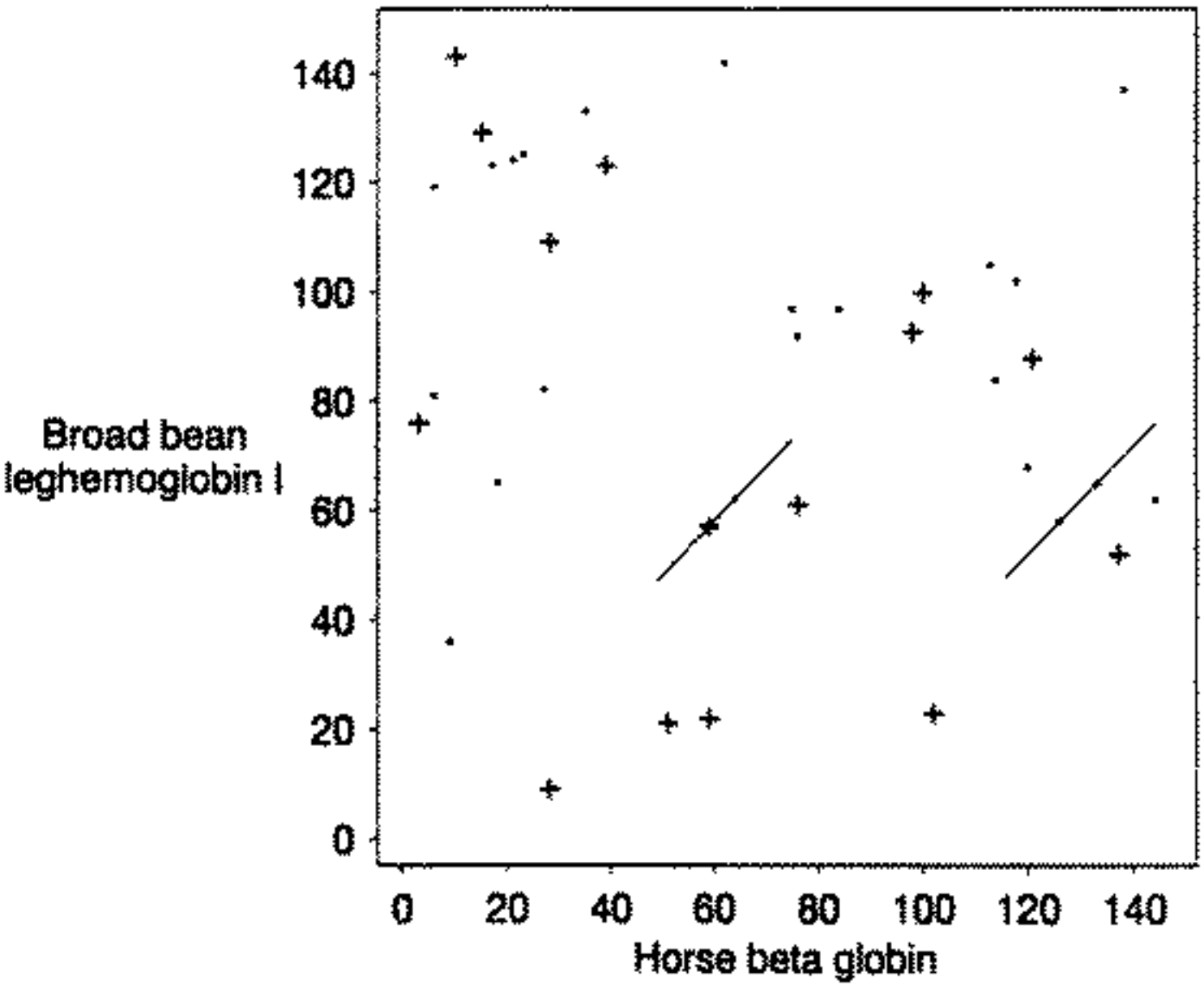
Query: HU-NS1 Maximal Score: 452
PAM Matrix: 200 Gap Penalty: 5 Gap Extension: 0.5

No.	Score	Match	Length	DB	ID	Description	Expectation
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
24	76	16.8	477	2	GLGA_ECOLI	GLYCOGEN SYNTHASE (EC	3.80e-01

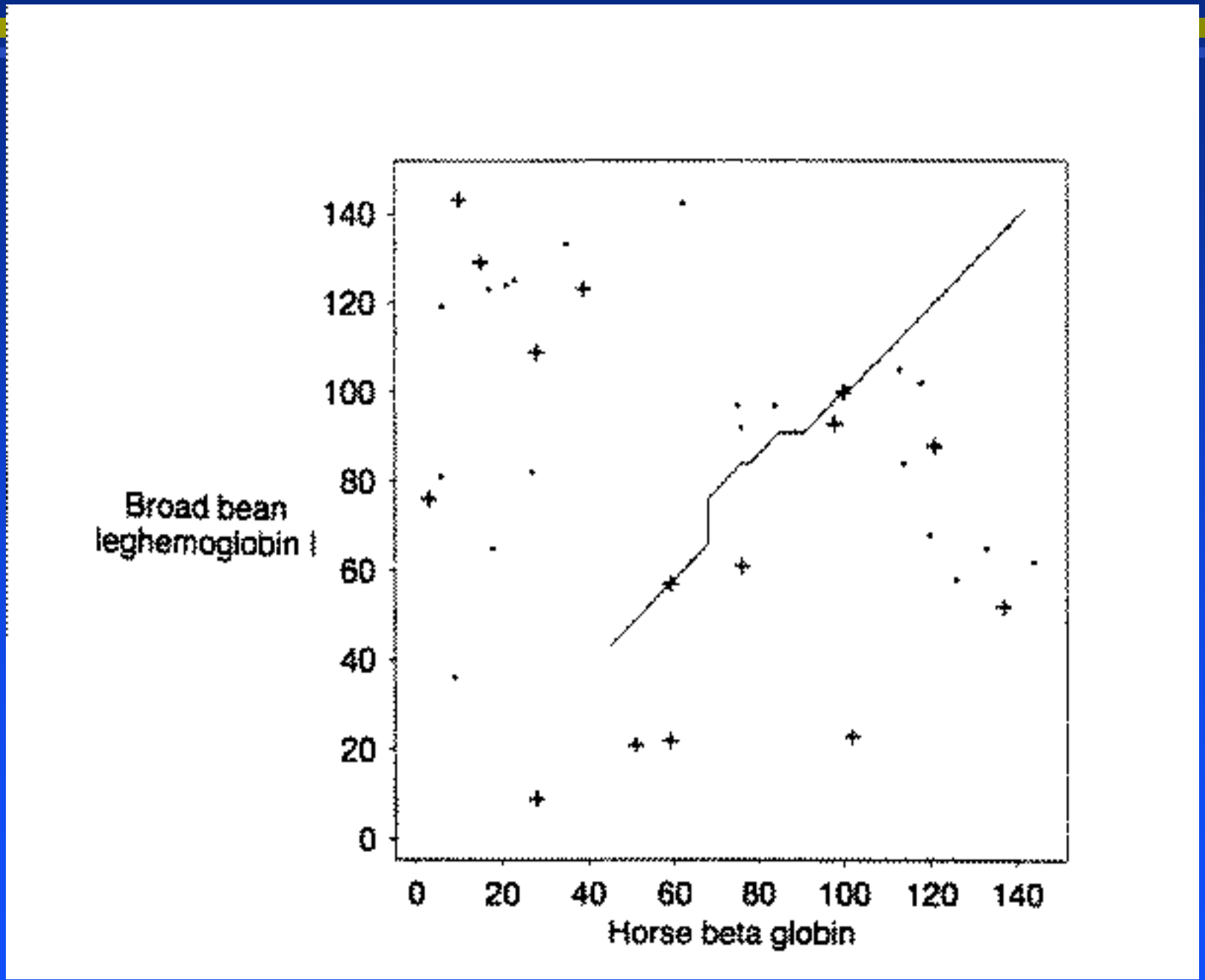
Original BLAST Algorithm

- Indexes database
- Starts with all overlapping words from query
- Calculates “neighborhood” of each word using BLOSUM matrix and probability threshold
- Looks up all words and neighbors from query in database index
- Extends High-Scoring Segment Pairs (HSPs) left and right to maximal length
- Finds Maximal Segment Pairs (MSPs) between query and database
- Does not permit gaps in alignments

GAPPED BLAST Starts with a Two Hit Approach



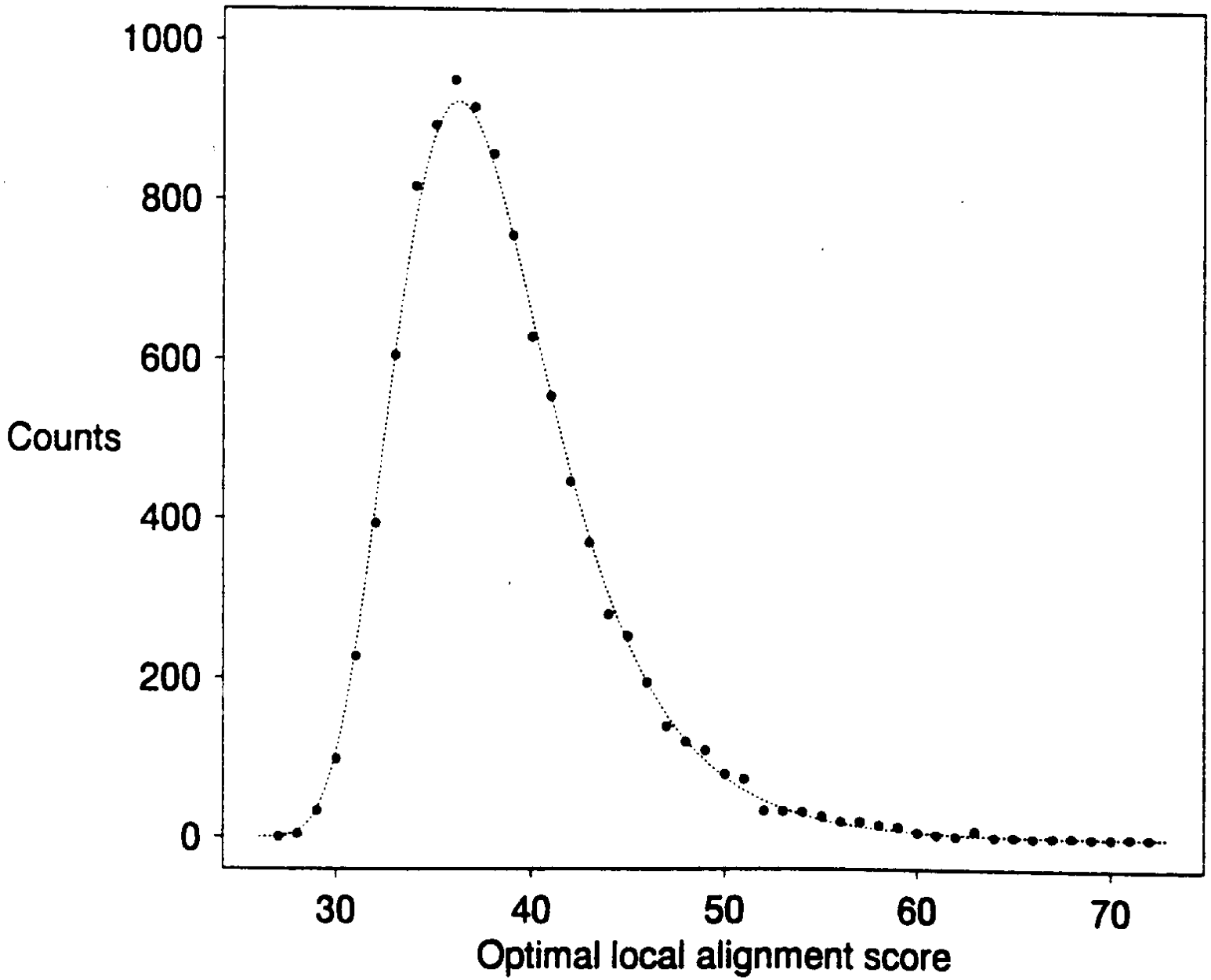
GAPPED BLAST Extension of Two Hit HSP



GAPPED BLAST Alignment

```
Leghemoglobin   43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
                   F  L +    V+ +PK+ AH +KV           L + GE V   LD    G+
Beta globin     45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE  90

Leghemoglobin   91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                   +H  K   +DP +F ++    L+  +     G   ++ EL A+++     G+A A+
Beta globin     91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

Extreme Value Distribution of Scores



Expectation of High Scoring Segment Pairs (HSPs)

$$\Pr ob(S > X) \approx 1 - \exp\left\{-Ke^{-\lambda X}\right\}$$

where λ is the root of the equation:

$$\sum_{i=1}^{r}\sum_{j=1}^{r} p_i p_j \exp\left\{\lambda s_{ij}\right\} = 1$$

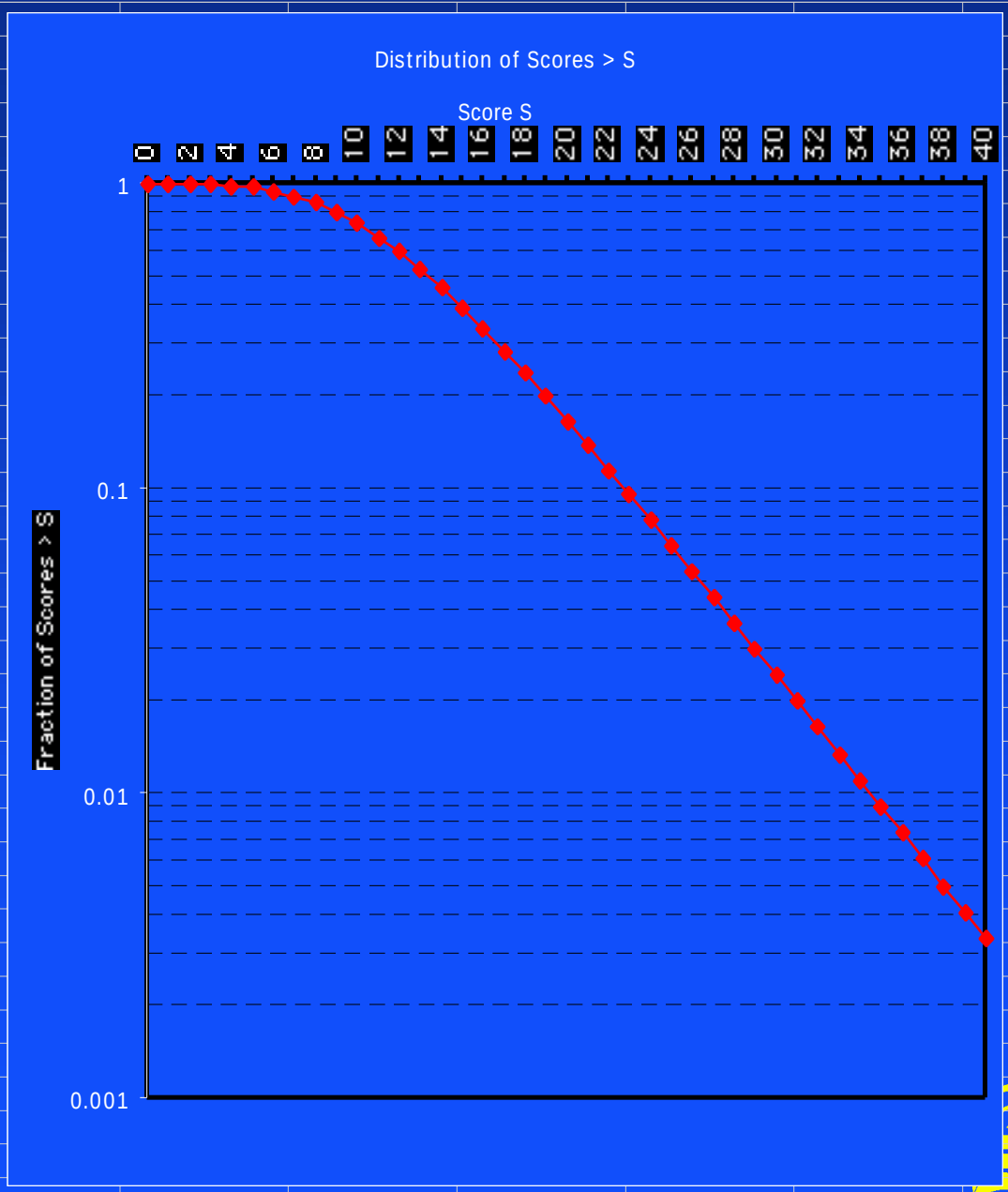
p_i and p_j are the probabilities of each residue in each sequence,

s_{ij} are the similarity scores of two residues.

If the expected value of the scores for random sequences is

$$< 0, \text{ i. e. } \left(\sum_{i=1}^{r}\sum_{j=1}^{r} p_i p_j s_{ij} < 0\right)$$

then there are two solutions for λ, zero and one other positive root.



Decypher Database Search Engine

http://decypher.stanford.edu/

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	Synopsis
Blastn		DNA	DNA	Nucleic comparison.
Blastx		DNA	Protein	Translated search in protein space.
Tblastx		DNA	DNA	Translated search in all vs. all frames in protein space.
Blastp	Example	Protein	Protein	Protein comparison.
Tblastn		Protein	DNA	Search of the translated database.
PSI-Blast		Protein	Protein	BLASTP with position-specific iterative refinement.
All Methods Form - BLASTALL		DNA or Protein	DNA or Protein	Combination form supporting all heuristic methods.

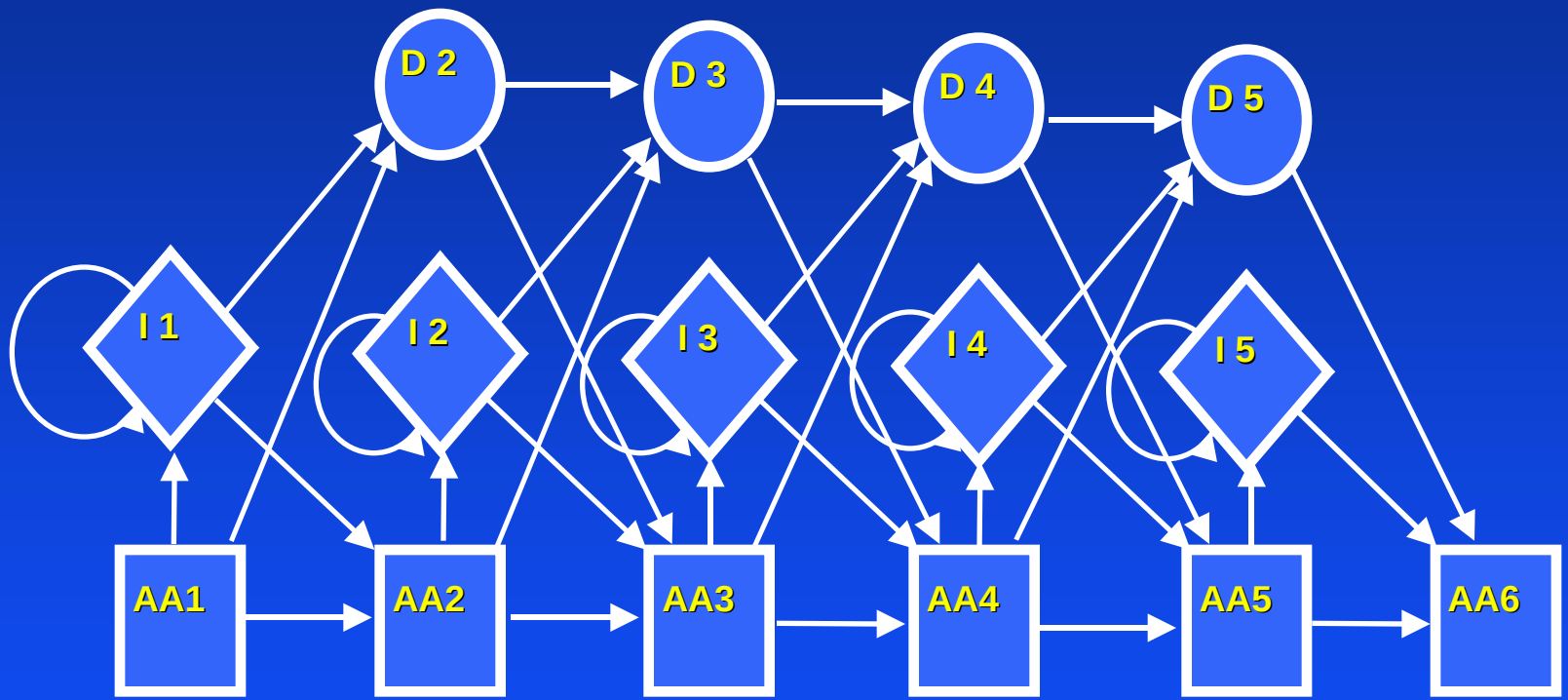
Decypher Database Search Engine

http://decypher.stanford.edu/

DYNAMIC METHODS		Your Query	Database	Synopsis
Hidden Markov Model	Example	DNA	Protein HMM	Translated search vs. protein HMM database.
	Example	Protein	Protein HMM	Protein sequence vs. HMM search.
	Example	Protein HMM	Protein	Compares protein HMM model to protein sequence d
Smith-Waterman	Example	DNA	DNA	Gapped, affine nucleic search.
		DNA	Protein	Translated gapped affine search.
		Protein	Protein	Gapped affine search in protein space.
FrameSearch	Example	DNA	Protein	Gapped, affine translated search in protein space with spanning codon frameshifts.
		Protein	DNA	
SFI™ Symmetric Frame Independent™	Example	DNA	DNA	Gapped affine protein space search with frameshift spanning in both query and database.
ProfileSearch	Example	Protein Profile	Protein	Position-specific gapped, affine search of profile to se
ProfileFrameSearch	Example	Protein Profile	DNA	Translated ProfileSearch with frameshift spanning.
ProfileScan		Protein	Protein Profile	Compares a protein sequence to profile database.
ClustalW	Example	DNA		Multiple sequence alignment.
		Protein		

Profiles & Hidden Markov Models

http://pfam.wustl.edu/



Discovering Function from Protein Sequence

BLOCK, Weight Matrix or Position Specific Scoring Matrix

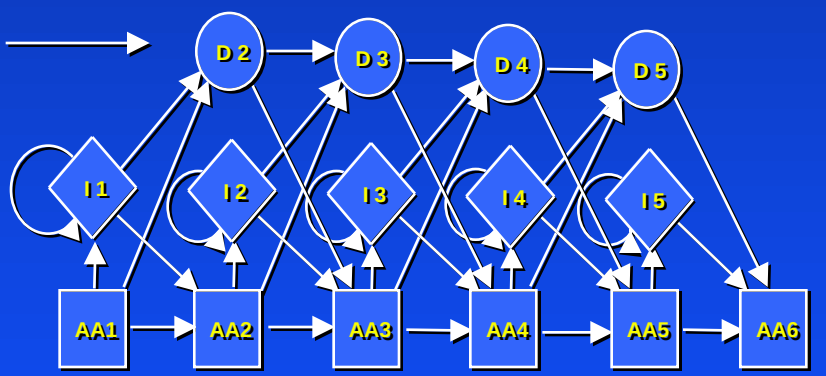
	Position											
	1	2	3	4	5	6	7	8	9	10	11	12
A	2	1	3	13	10	12	67	4	13	9	1	2
R	7	5	8	9	4	0	1	16	7	0	1	0
N	0	8	0	1	0	0	0	2	1	1	10	0
D	0	1	0	1	13	0	0	12	1	0	4	0
C	0	0	1	0	0	0	0	0	0	2	2	1
Q	1	1	21	8	10	0	0	7	6	0	0	2
E	2	0	0	9	21	0	0	15	7	3	3	0
G	9	7	1	4	0	0	8	0	0	0	46	0
H	4	3	1	1	2	0	0	2	2	0	5	0
I	10	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

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

Sequences of Common Structure or Function

Profiles, PSI-BLAST Hidden Markov Models



Sequence Alignments

```
                10        20        30        40                50
1     VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGS
       |:| :|: | |:||||     | |:||| |: : :|:|  :|  |       |:  |
2     HLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
                10        20        30        40        50
```

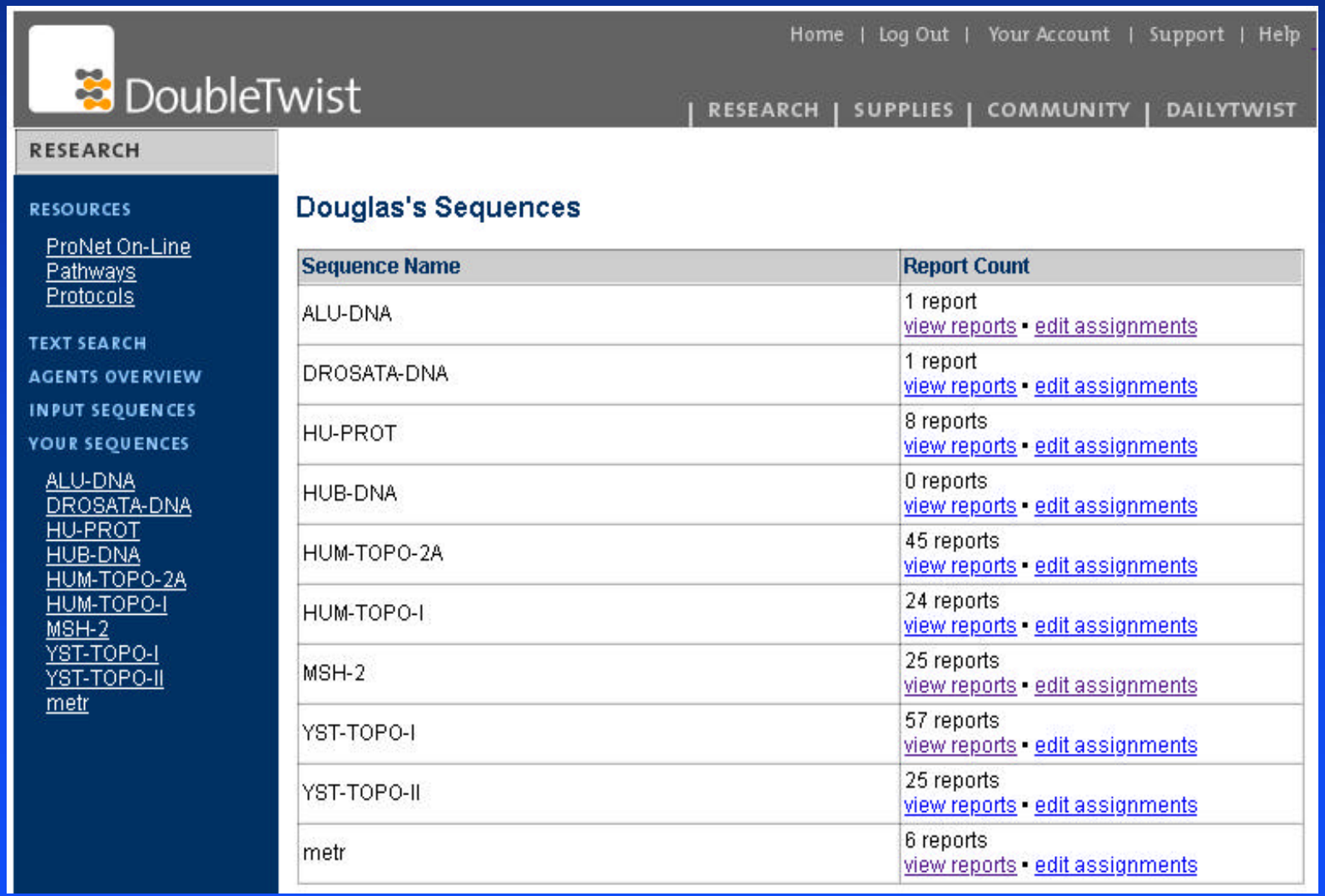
DoubleTwist's Home Page

http://doubletwist.com/

What is *DoubleTwist.com?*

- DoubleTwist is a web portal for life sciences
- DoubleTwist accepts sequence & text queries
- DoubleTwist provides multiple analyses
  - DNA & protein sequence databases
  - Completed genome database
  - Clustered EST databases
  - Motif & domain databases
- DoubleTwist provides research reports
- DoubleTwist provides evaluation of results
- DoubleTwist links to reagents, clones & protocols
- DoubleTwist is agent driven
- DoubleTwist runs daily

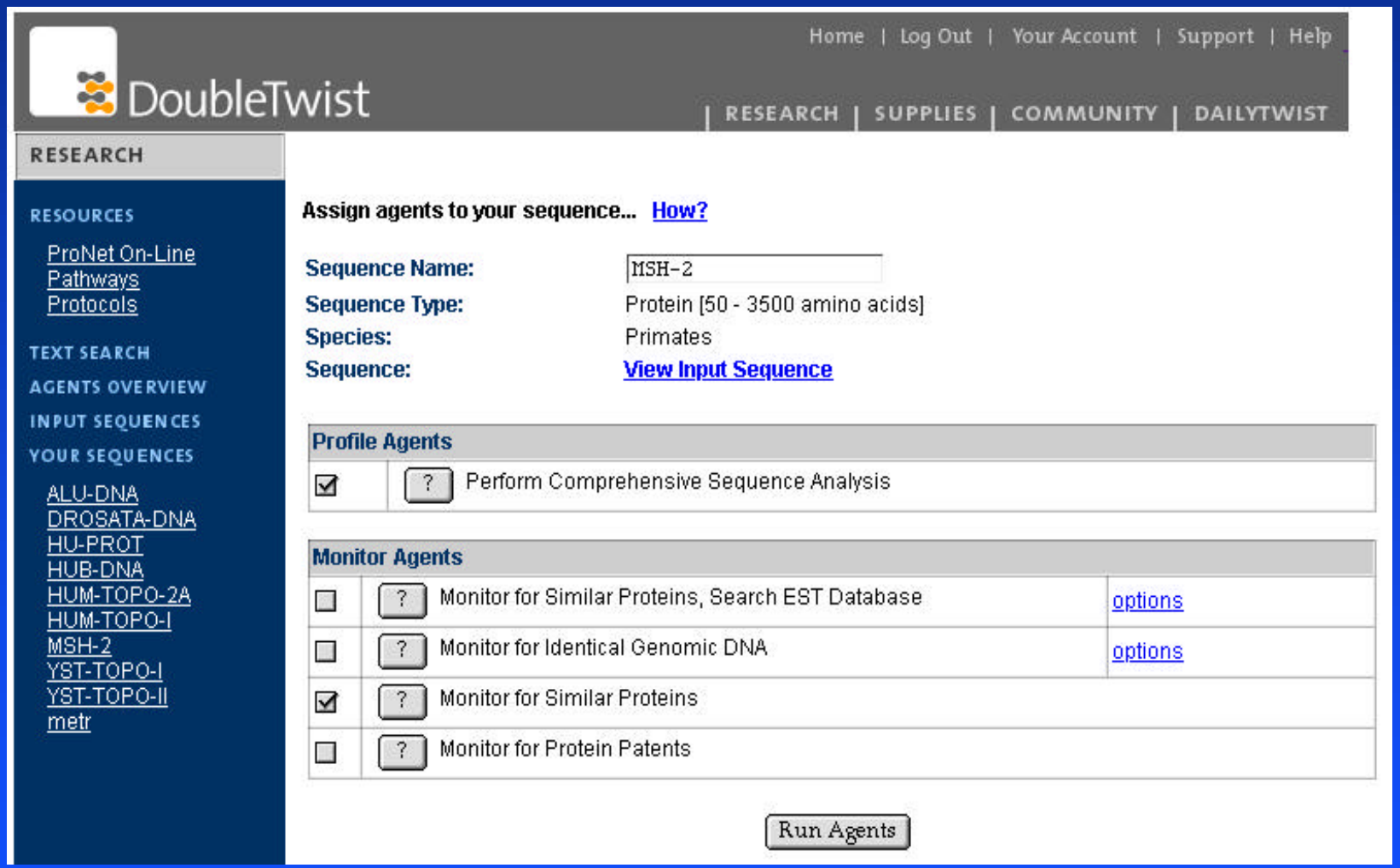
Query Sequences



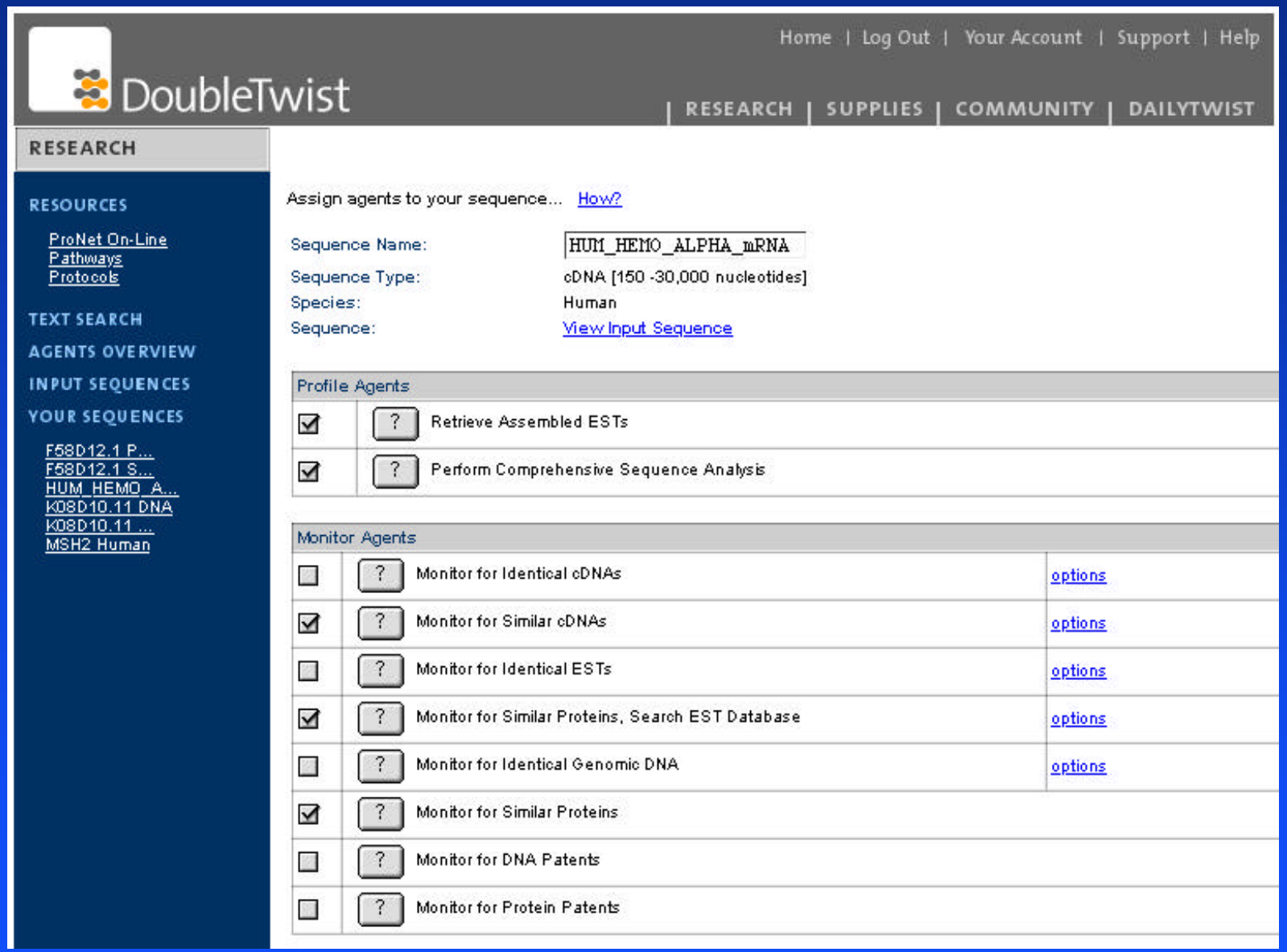
DoubleTwist Research Agents

- Given an EST, cDNA, genomic DNA or protein sequence DoubleTwist will:
  - Find identical cDNA, EST or genomic sequences
  - Find similar cDNA, EST or genomic sequences
  - Find identical protein coding regions
  - Find coding regions for similar proteins
  - Find functional motifs in coding regions
  - Find similar proteins whose structure is known
  - Find similar sequences that have been patented

Assigning DoubleTwist Agents to Queries



Assigning DoubleTwist Agents to Queries



How one DoubleTwist Agent Works

Agent name: Find Similar ESTs

- Initial steps
  - 1. Remove trailing poly(A) and poly(T)
  - 2. Mask vector sequences
  - 3. Mask repetitive / low complexity regions
- Every night
  - Run BLAST2N against GenBank nightly updates
  - Return hits that fall within acceptable parameters
    - determine if hit is an EST from specified tissue / organism
  - Notify user via E-mail (Web Link points to Report)
    - flag hits with confidence values (high / medium / low)

DoubleTwist Protein Query versus EST Similarity Agent

EST Identities and Similarities
Finds any similar published protein sequences using TBLAST2N (gapped BLAST) to search GenBank's EST (dbEST) database.

The top three matches, determined by our "Basis for a Match," are reported for this section.

Note: All "tied" matches (separate records with identical E Value scores) are included in your report.

Basis for a Match	
Confidence Level	*E Value Range*
HIGH	$<1E^{-30}$
MEDIUM	$\leq 1E^{-8}$ and $\geq 1E^{-30}$
LOW	< 0.1 and $>1E^{-8}$
NONE	≥ 0.1

DoubleTwist's Databases

Sequence / motif databases

- SwissProt
- TREMBL
- dbEST
- PDB
- PIR
- BLOCKS+
- PRODOM
- PRINTS
- PFAM
- UNIGENE
- CGAP
- Patent Database
- Repbase
- GenBank non-redundant nucleotide database
- GenBank non-redundant protein database
- GenBank daily DNA database updates
- GenBank daily protein database updates

DoubleTwist curated databases

- Human EST cluster database
- Mouse EST cluster database
- Arabidopsis EST cluster database
- Completed Genomes Database
- Comprehensive Vector Database

Other databases in development

- AlphaGene clone database
- ClonTech clones & reagents
- ProNet™ (Myriad Genetics)
- Microbial metabolic database (Karp)
- Many others

DoubleTwist's Algorithms

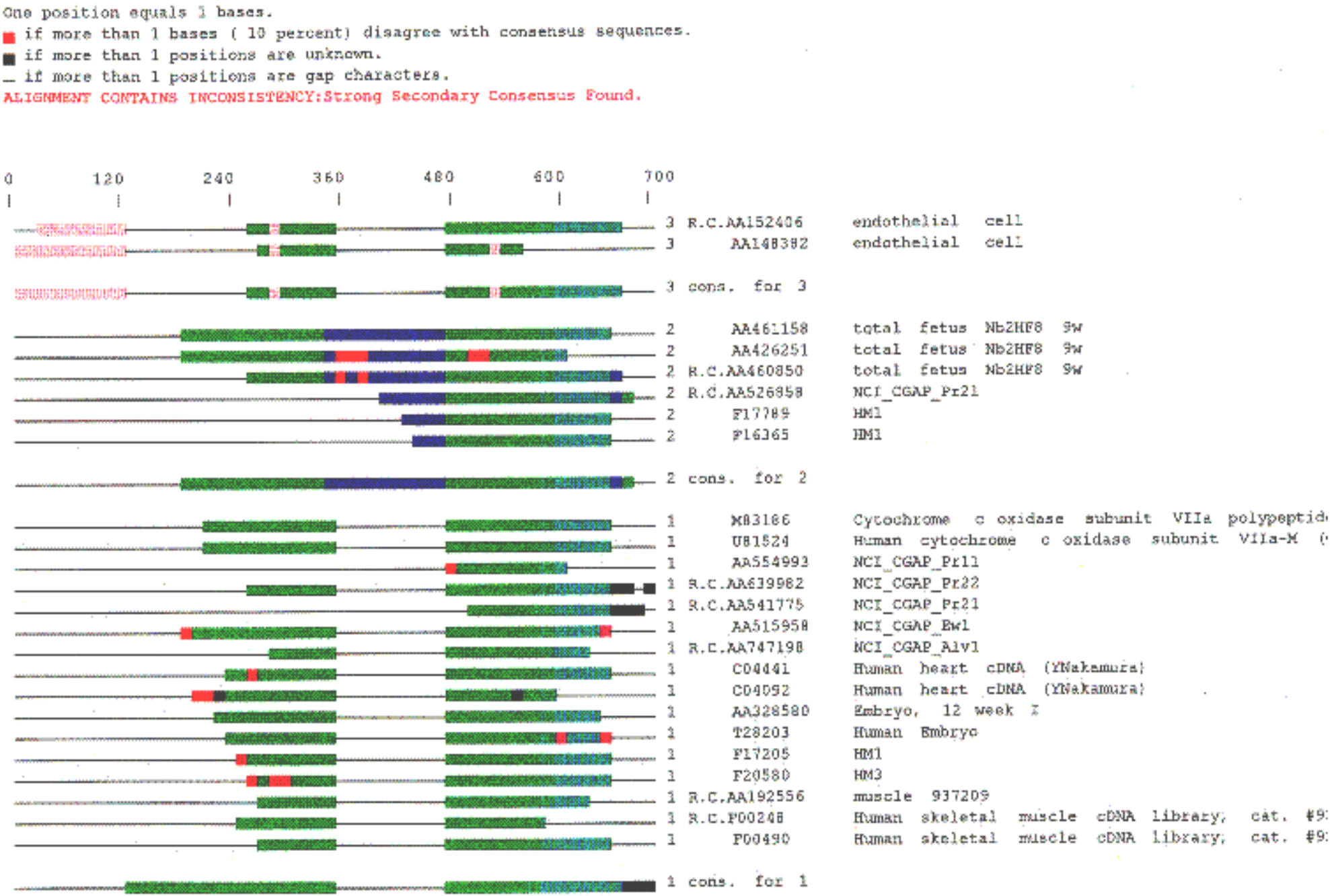
- BLAST2N (NCBI)
- BLAST2X (NCBI)
- TBLAST2N (NCBI)
- TBLAST2X (NCBI)
- MASK (vector masking, DoubleTwist Inc.)
- RepeatMasker (Smit & Green, U. Washington)
- Blocks Search (Henikoff & Henikoff, FHCRC)
- GRAIL (Uberbacher et al.,DOE)
- GenScan (Burge et al., Stanford & MIT)
- HalfWise (Birney et al., Sanger Center)
- HMMer (Eddy & Sonnhammer, Wash U. & MRC)
- SIM4 (Miller et al., Penn State)

Advantages of DoubleTwist's Clustered EST Databases

Clustering and alignment tool (CAT) detects alternative splicing and sequence polymorphisms

- CAT corrects cloning & sequencing errors
  - CAT removes insertion / deletion errors resulting in longer open reading frames (ORFs)
  - CAT distinguishes alternative splicing from sequencing artifacts (chimeras & lane tracking errors)
  - CAT distinguishes SNPs from sequencing errors
- Better drug target discovery
  - Increased sensitivity (target discovery)
  - Improved selectivity (target validation)

DoubleTwist Clusters

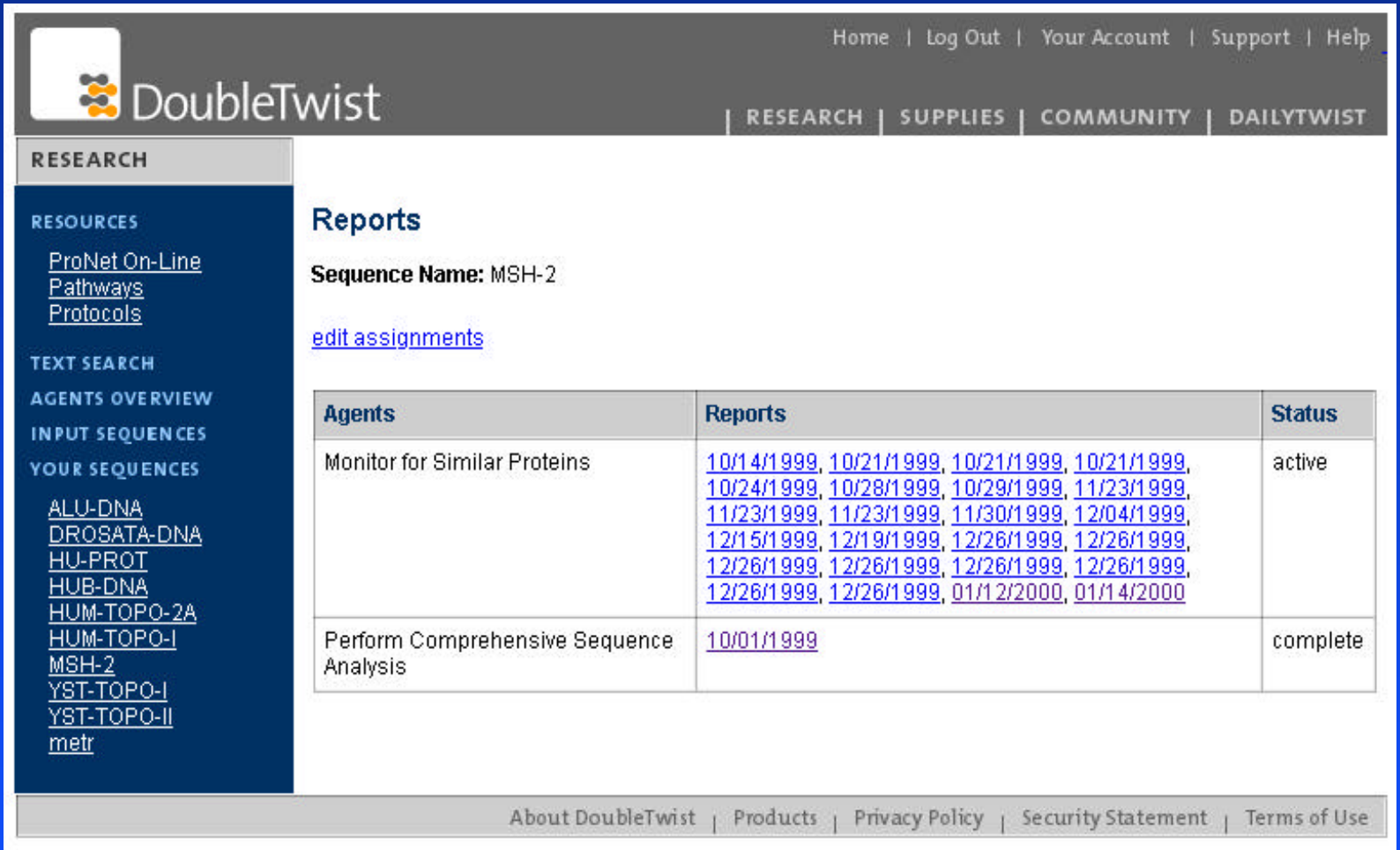


Single Nucleotide Polymorphisms: Resolving Adjacent Polymorphic Sites

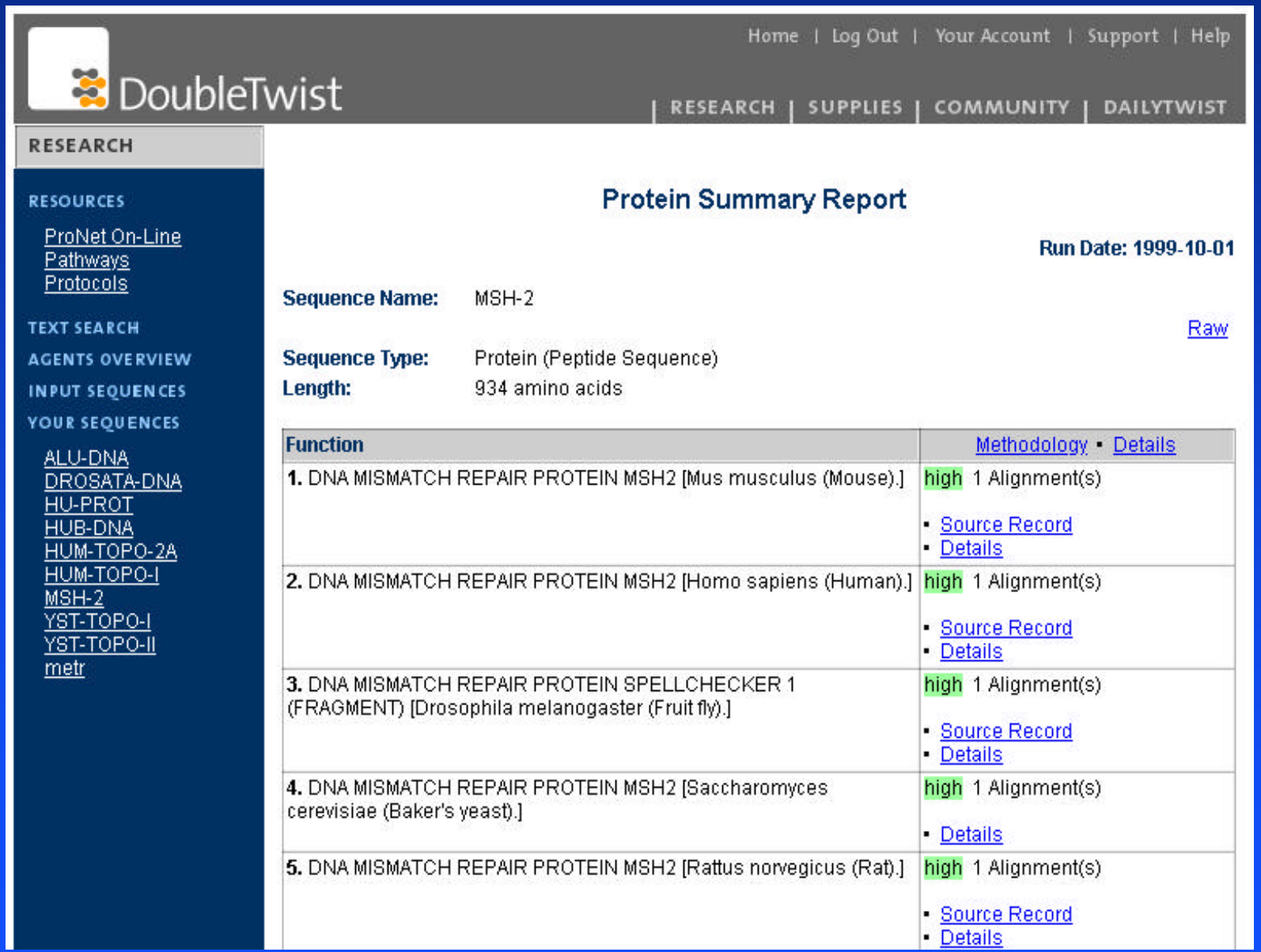
```
                                    A
WIAF-1302 GTCTTCAGCCTCCTCTACCCTACGAGATCTGGAGCAACAGCTAGGAAA
                                    ^
                                   [A/G]
                                    |
          GTTTTCAGCCTCCTCTACCCTACATGATCTGGAGCAA gene_gnl|UG|Hs#S576
          GTTTTCAGCCTCCTCTACCCTACATGATCTGGAGCAA AA484792
          GTTTTCAGCCTCCTCTACCCTACATGATCTGGAGCAA R.C.T40995
            |
          GTCTTCAGCCTCCTCTACCCTACATGATCTGGAGCAA gene_gnl|UG|Hs#S1790
          GTCTTCAGCCTCCTCTACCCTACATGATCTGGAGCAA gene_gnl|UG|Hs#S5761
                                    ||
          GTCTTCAGCCTCCTCTACCCTACGAGATCTGGAGCAA gene_gnl|UG|Hs#S4384
          GTCTTCAGCCTCCTCTACCCTACGAGATCTGGAGCAA genel|UG|Hs#S268846
          GTCTTCAGCCTCCTCTACCCTACGAGATCTGGAGCAA gene_gnl|UG|Hs#S1791
                                    |
          GTCTTCAGCCTCCTCTACCCTACAAGATCTGGAGCAA gene_gnl|UG|Hs#S269861
          GTCTTCAGCCTCCTCTACGCTACAAGATCTGGAGCAA AA233679
          GTCTTCAGCCTACTATACCCTACAAGATCTGGCGCAA R.C.AI000973
          GTCTTCAGCCTCCTCTACCCTACCAGATCTGGAGCAA AA736471
          GTCTTCAGCCTCCTCTACCCTACAAGATCTGGAGCAA D57354
          GTCTTCAGCCTCCTCTACCCTACAAGATCTGGAGCAA AA393215
            ^                       ^^
            [C/T]                   |[A/T]
                                    [A/G]
```


DoubleTwist's View of an EST Cluster

DoubleTwist Research Reports



A DoubleTwist Research Report: Protein Function



A DoubleTwist Research Report: Protein Functional Motifs

Protein-Protein Interactions (ProNet on-line)	Methodology • Details
None Found	
Motifs	Methodology • Details
1. DNA mismatch repair prot	high 1 Alignment(s) • Details
2. DNA mismatch repair prot	high 1 Alignment(s) • Details
3. DNA mismatch repair prot	high 1 Alignment(s) • Details
4. DNA mismatch repair prot	high 1 Alignment(s) • Details
5. DNA mismatch repair prot	medium 1 Alignment(s) • Details
6. DNA mismatch repair prot	medium 1 Alignment(s) • Details
7. Bacterial ring hydroxyla	low 1 Alignment(s) • Details
8. Eukaryotic RNA polymeras	low 1 Alignment(s) • Details

A DoubleTwist Research Report
Late Breaking Database Entries

"Late Breaking" Protein Similarities	Methodology • Details
1. (AJ245967) mismatch repair protein msh6-1 [Arabidopsis thaliana]	high 1 Alignment(s) • Details
2. (AJ245967) mismatch repair protein msh6-1 [Arabidopsis thaliana]	high 1 Alignment(s) • Details
3. (AJ245967) mismatch repair protein msh6-1 [Arabidopsis thaliana]	high 1 Alignment(s) • Details
4. (U58758) C. elegans HIM-5 (GB:AF178755); contains similarity to Pfam domain PF00488 (mutS), Score=204.6, E-value=4.9e-58, N=1 [Caenorhabditis elegans]	high 1 Alignment(s) • Details

Patent, Structure and EST Reports

Patents	Methodology ▪ Details
1. Sequence 2 from patent US 5591826	1 Alignment(s) ▪ Details
Structure Similarities	Methodology ▪ Details
None Found	
EST Similarities	Methodology ▪ Details
1. w68d09.r1 Stratagene mouse skin (#937313) Mus musculus cDNA clone 1227569 5' similar to gb:X81143 M.musculus msh2 mRNA (MOUSE);	medium 1 Alignment(s) ▪ Source Record ▪ Details
2. ah13d02.y5 Gessler Wilms tumor Homo sapiens cDNA clone IMAGE:1156515 5' similar to SW:MSH2_HUMAN P43246 DNA MISMATCH REPAIR PROTEIN MSH2. ;; mRNA sequence	high 1 Alignment(s) ▪ Source Record ▪ Details
3. ne42b03.s1 NCI_CGAP_Co3 Homo sapiens cDNA clone IMAGE:899981 similar to SW:MSH2_HUMAN P43246 DNA MISMATSCH REPAIR PROTEIN MSH2. ;	medium 1 Alignment(s) ▪ Source Record ▪ Details

Viewing Monitor Agent Reports

Notification of New Findings

Date: Friday, 4 Feb 2000 16:18:11 -0800 (PST)
From: report <report@DoubleTwist.com>
To: BRUTLAG <brutlag@stanford.edu>
Subject: Find Similar Proteins

This mail is auto-generated, and was triggered by NEW FINDINGS for the sequence you specified at:
http://www.DoubleTwist.com

Here is the outline:
Agent: Find Similar Proteins
Sequence: MSH-2
Details:
http://www.DoubleTwist.com/report.jsp?id=26074

The links above are secured by password, and all access to DoubleTwist is monitored to ensure privacy. For more information about DoubleTwist, security and privacy, please go to: http://www.DoubleTwist.com

Similar Protein Monitor Agent Report

Monitor for Similar Proteins Report

Agent Monitor for Similar Proteins report for 2000-02-04

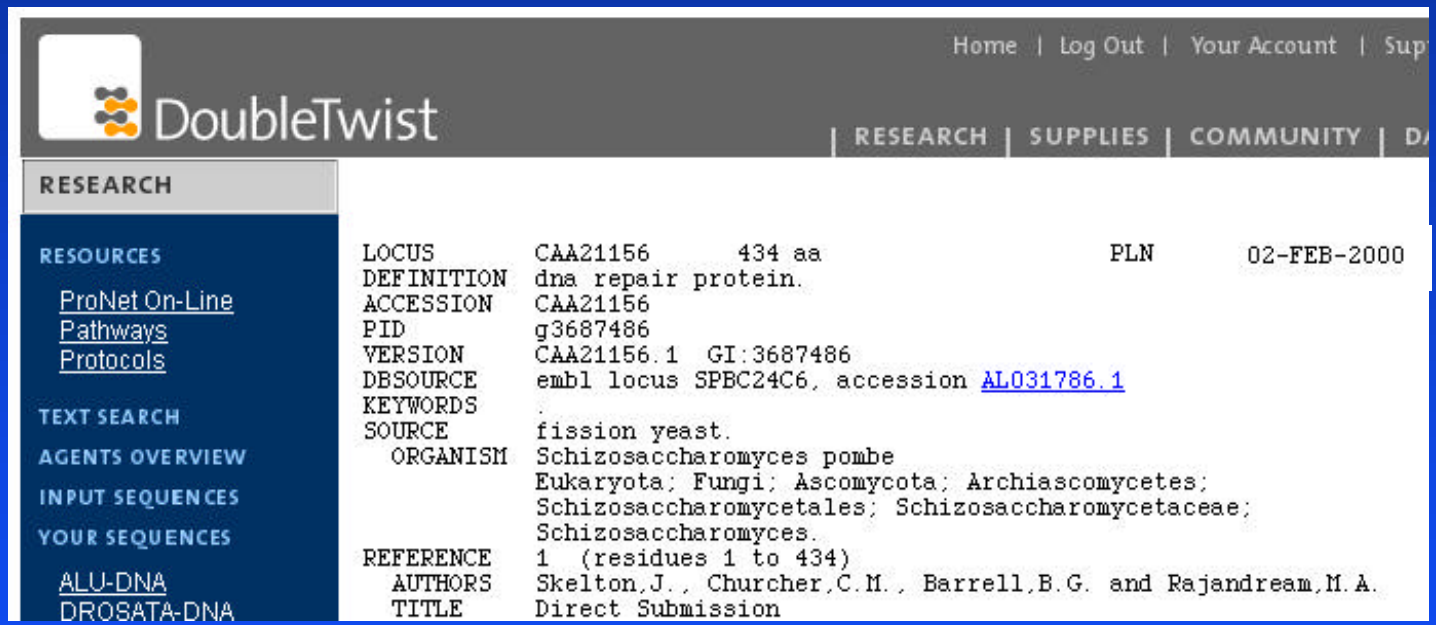
Query Sequence	
Name:	MSH-2
Input Sequence Type:	Protein [50 - 3500 amino acids]
Length:	934

Methodology

Matching Sequence	
Accession number:	CAA21156
Description:	>gi|3687486|emb|CAA21156| (AL031786) dna mismatch repair protein Msh2p [Schizosaccharomyces pombe]
Literature:	None Found
Sequence Alignment 1:	(Alignment Length: 424, E-Value: 3.0E-95, Alignment % Identity: 45.0)

```
Query: 536 NKNFSTVDIQKNGVKFTNSKLTSLXXXXXXXXXXXXXXAQDAIVKEIVNISSGYVEPMQTL 595
          + +++ +  QKNGV FT  +L SL               Q+ + +E++ I++ Y  P++ +
Sbjct: 5   SSHYTELSTQKNGVYFTTKRLHSLNNSYMDHQKSYRYHQNGLAREVIKIAATYGPPLEAI 64

Query: 596 NDVLAQLDAVVSFAHVSNGAPVPYVRPAILEK-------GQG------------------ 630
             V+A LD ++SFAH S  A +PYVRP I++         GQ
Sbjct: 65  GQVIAHLDVILSFAHASTVAVIPYVRPNIVDSSIAQEKHGQSSNILDIVSLEDTPNFEEI 124

Query: 631 ----------RIILKASRHACVEVQDEIAFIPNDVYFEKDKQMFHIITGPNMGGKSTYIR 680
                     R+ LK +RH C+E QD++ FIPNDV  E       IITGPNMGGKSTYIR
Sbjct: 125 RRTLENNHCARLYLKQARHPCLEAQDDVKFIPNDVNLEHGSSELLIITGPNMGGKSTYIR 184

Query: 681 QTGVIVLMAQIGCFVPCESAEVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSAT 740
           Q GVI +MAQIGC VPCE A++ I+D ILARVGA DSQLKG+STFMAEMLETA+ILR+AT
Sbjct: 185 QVGVITVMAQIGCPVPCEVADLDIIDAILARVGASDSQLKGISTFMAEMLETATILRAAT 244

Query: 741 KDSLIIIDELGRGTSTYDGFGLAWAISEYIATKIGAFCMFATHFHELTALANQIPTVNNL 800
             SLIIIDELGRGTST DGFGLAWAI+E+I T+IG FC+FATH+HE+T L+ +I TV NL
Sbjct: 245 PRSLIIIDELGRGTSTTDGFGLAWAITEHIVTQIGCFCLFATHYHEMTKLSEEITTVKNL 304
```

Submission Date for CAA21156



Similar Protein Monitor Agent Report

Monitor for Similar Proteins Report

Agent Monitor for Similar Proteins report for 2000-01-12

Query Sequence	
Name:	MSH-2
Input Sequence Type:	Protein [50 - 3500 amino acids]
Length:	934

Methodology

Matching Sequence	
Accession number:	BAA87137
Description:	>gi|6473463|dbj|BAA87137| (AB027833) Hypothetical protein [Schizosaccharomyces pombe]
Literature:	None Found
Sequence Alignment 1:	(Alignment Length: 172, E-Value: 2.0E-16, Alignment % Identity: 30.0)

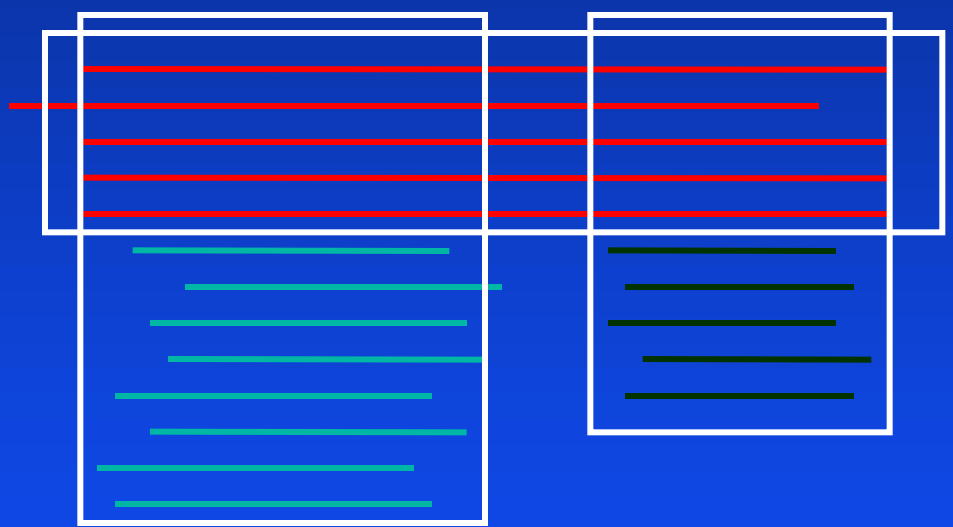
```
Query:  546 KNGVKFTNSKLTSLXXXXXXXXXXXXXXAQDAIVKEIVNISSGYVEPMQTLNDVLAQLDAV 605
           K+   F     TSL               +  ++K I +    + + +++L + L +LD
Sbjct:  3   KSTASFQLPGWTSLGMDLENTKLHIHQEEQRVLKSITDEIVSHHKTLRSLANALDELDIS 62

Query:  606 VSFAHVSNGAPVPYVRPAILEKGQGRIILKASRHACVEV---QDEIAFIPNDVYFEKDKQ 662
            S A ++       +VRP + +      +I    RH  VE        I F PND +
Sbjct:  63  TSLATLAQEQD--FVRPVVDDSHAHTVI--QGRHPIVEKGLSHKLIPFTPNDCFVGNGNV 118

Query:  663 MFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESAEVSIVDCILARVGA 714
               +ITGPNM GKST++RQ   +I ++AQIG FVP   +A + IVD I +R+G+
Sbjct:  119 NIWLITGPNMAGKSTFLRQNAIISILAQIGSFVPASNARIGIVDQIFSRIGS 170
```

DoubleTwist Genome Viewer

PSI-BLAST In Brief

1) Compare the query sequence to database
2) Construct profile from significant similarities
3) Compare the profile to database
4) Repeat step 2 and 3 until convergence

PSI-BLAST Results Contain Multiple Similar Regions



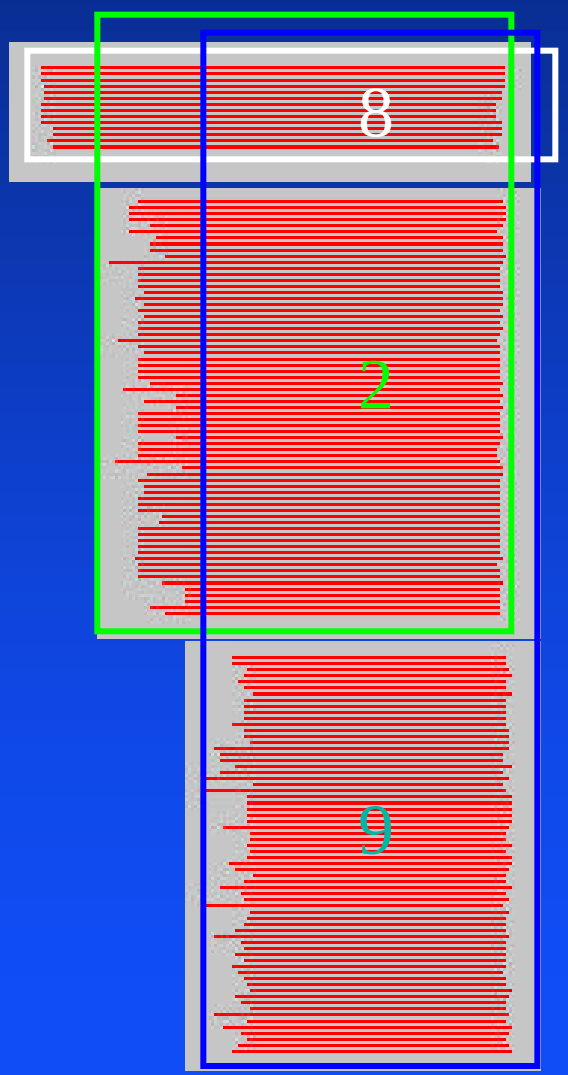
Major Steps:

1) Clustering

2) Alignment

3) Trimming

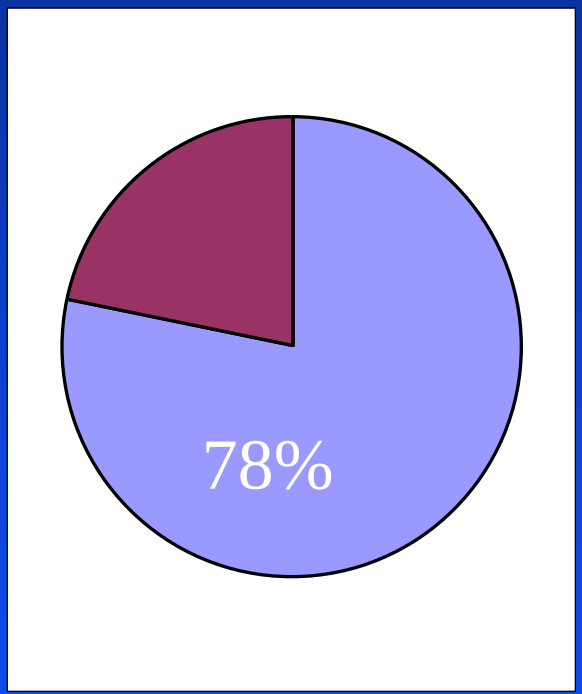
Clusters Are Organized Into Groups



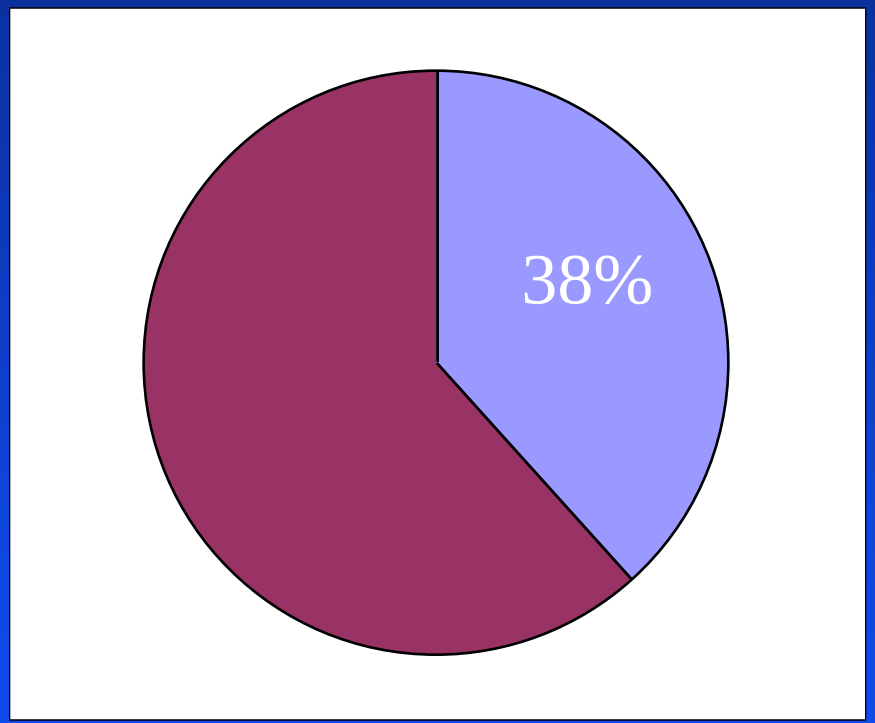
eBLOCKs Summary

- SWISS-PROT
  - 79,449 Sequences
- Filtered Target Set
  - 57,266 Sequences
- PSI-BLAST Searches
  - 17,415
- Final Number Of Groups
  - 16,658
- Final Number Of Blocks
  - 74,396

eBLOCKs Is More Comprehensive



BLOCKS+
(9,498)



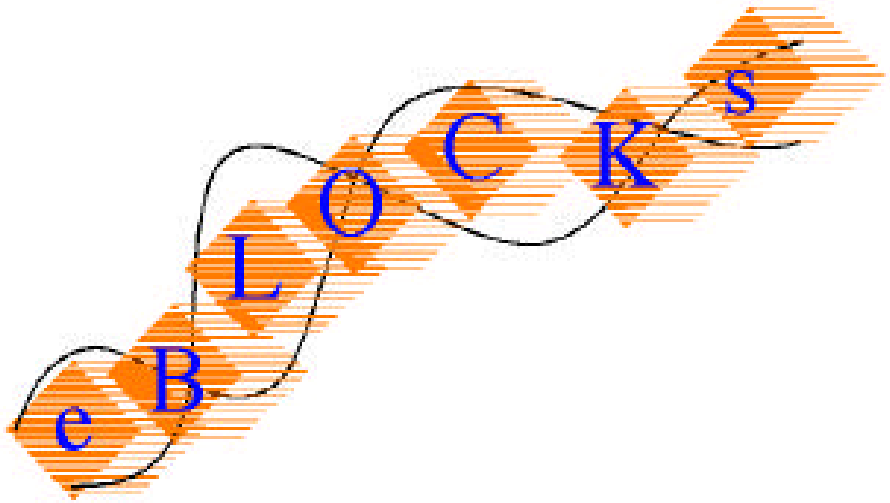
eBLOCKs
(74,396)

Web Access to eBLOCKS
(http://eblocks.stanford.edu/)

Netscape: eBLOCKs Home

Back Forward Reload Home Search Guide Images Print Security Stop

Netsite: http://eblocks/



Enumeration of Blocks From PSI-Blast Results

About eBLOCKs

Search By Accession

Search By Keyword

Search A Sequence

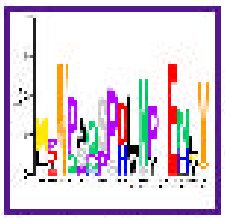
Qiaojuan Su and Douglas Brutlag
Brutlag Bioinformatics Group
Stanford University

An Entry From eBLOCKs

Netscape: eBLOCK P29358G1B1

eBLOCK P29358G1B1

- Sequence Logo



A PostScript File Will Be Generated. The Left Picture (GIF) Is Illustration Only.
You Could Use Ghostscript To View PostScript Files.
[About Sequence Logo]

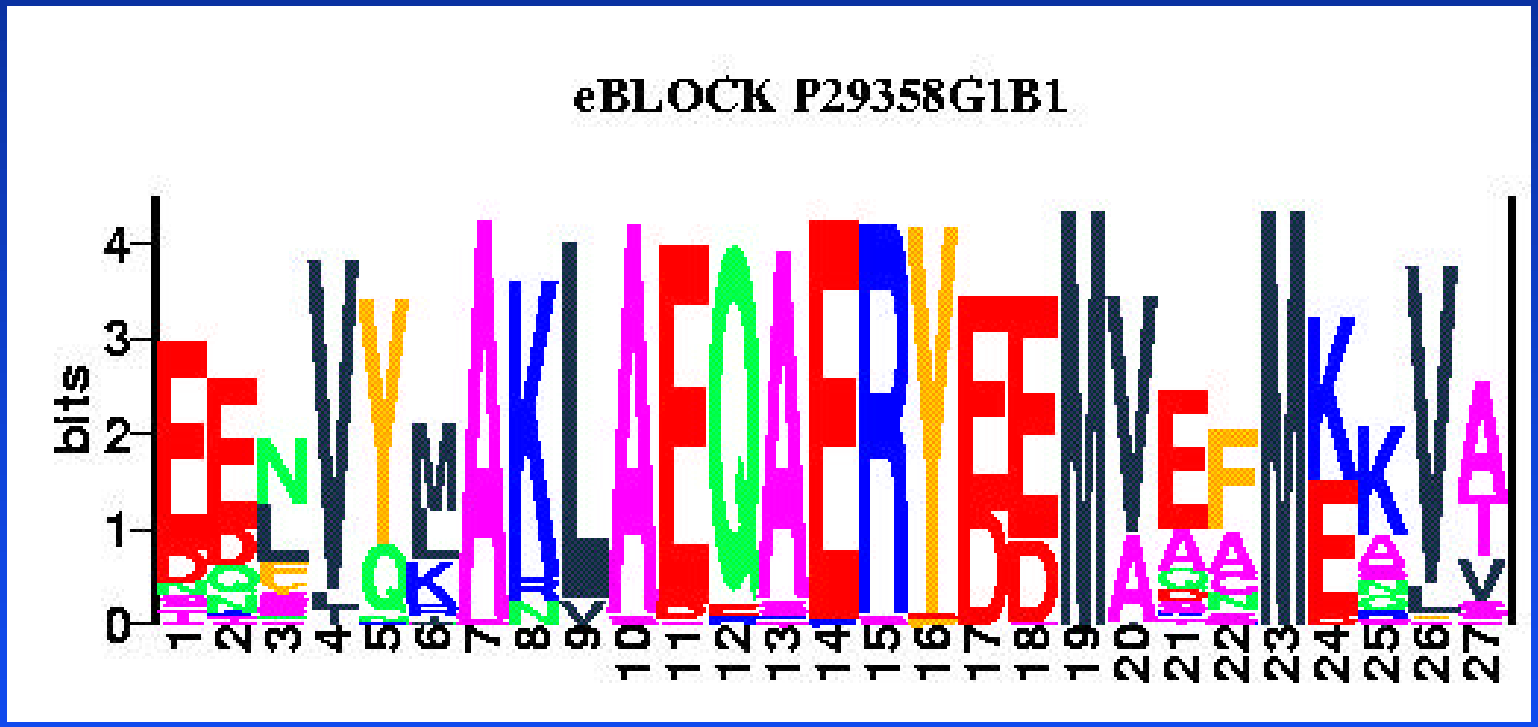
- Links To Blocks+ And Source Documents

  BLOCKS+ BL00796A: PROSITE PS00796

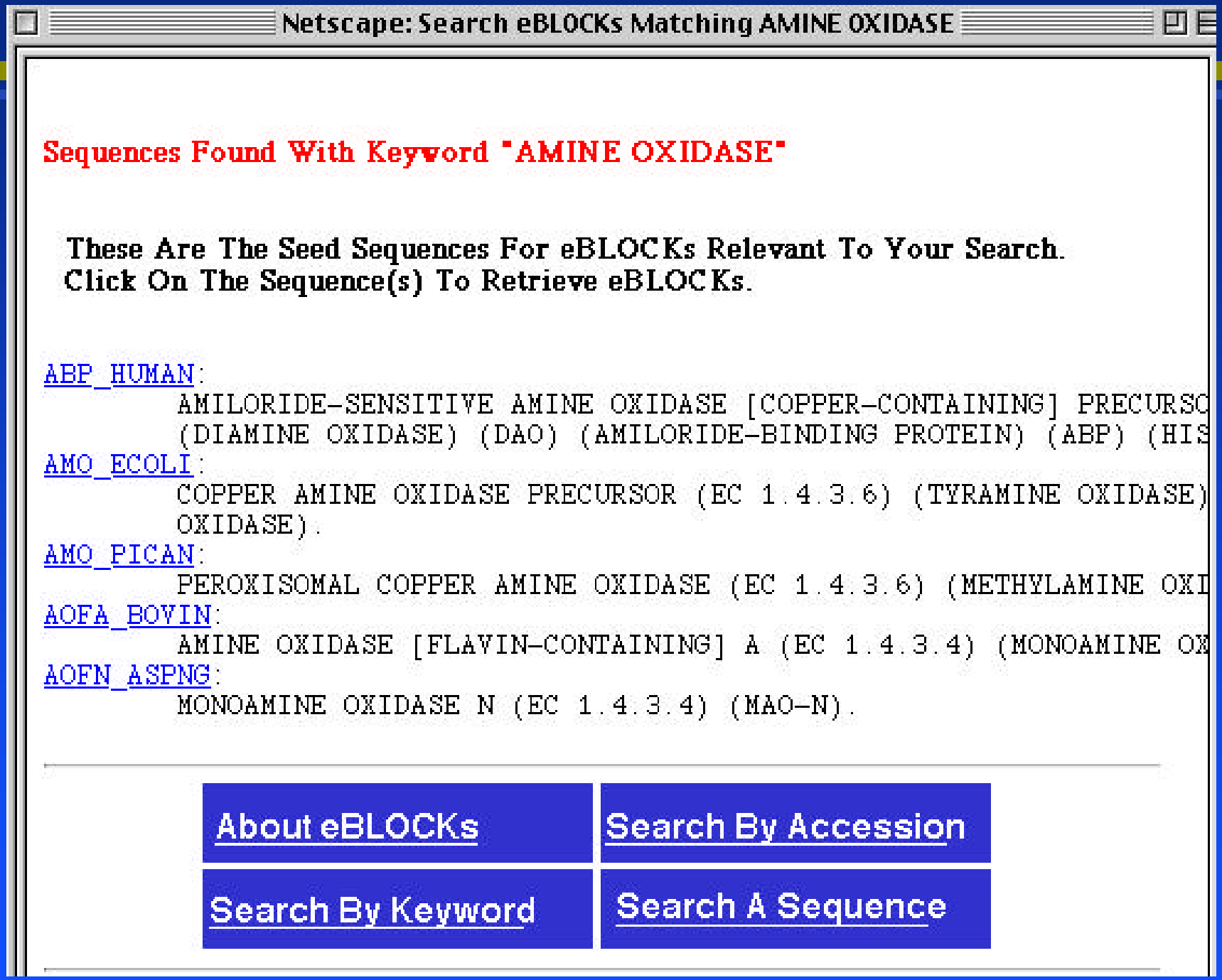
- **All eBLOCKs From The Same Seed Sequence** (143B_BOVIN)

```
ID   143B_BOVIN;
AC   P29358G1B1
DE   14-3-3 PROTEIN BETA/ALPHA (PROTEIN KINASE C INHIBITOR PROTEIN-1)
DE   (KCIP-1).
BL   width=27 seqs=72

143B_TOBAC (O49995) (    4) EENVYMAKLAEQAERYEEMVSFMEKVS  18
1435_SOLTU (P93784) (    6) EENVYMAKLAEQAERYEEMVEFMEKVV  13
1433_PEA   (P46266) (    9) EENVYMAKLAEQAERYEEMVEFMEKVS  15
1433_MESCR (P93259) (    8) EENVYMAKLAEQAERYEEMVEFMEKVA  13
143C_TOBAC (P93343) (    9) EENVYMAKLAEQAERYEEMVEFMEKVS  15
143O_ARATH (Q01525) (    6) EELVYMAKLAEQAERYEEMVEFMEKVS  15
```

An Entry From eBLOCKs



A Sample Keyword Search



Search A Sequence