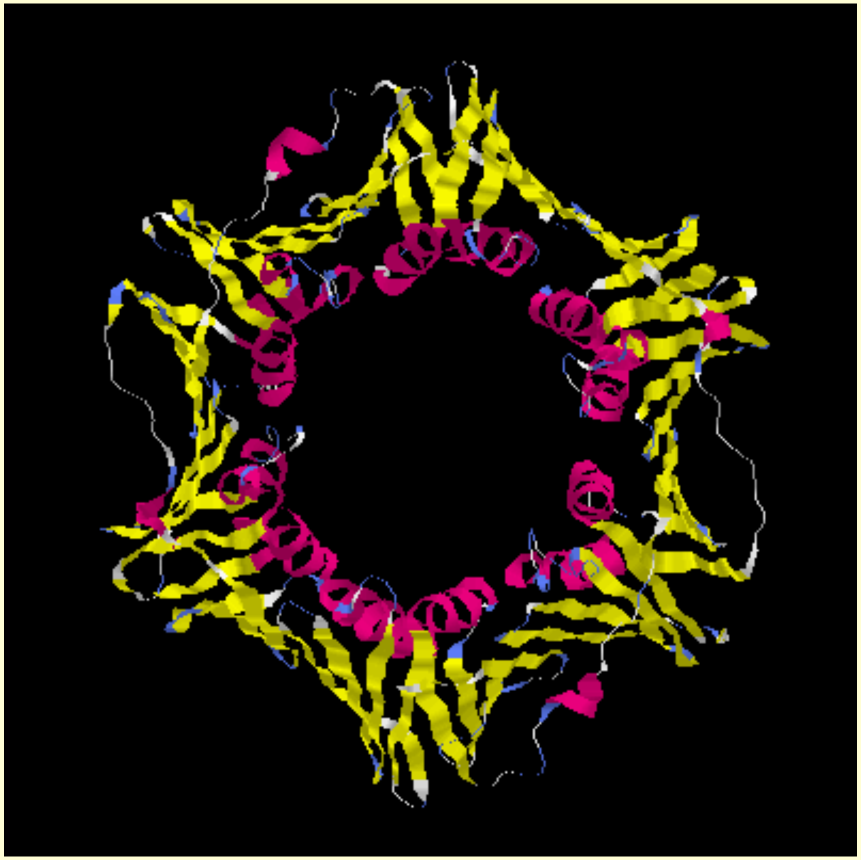
Computational Molecular Biology

Biochem 218 – BioMedical Informatics 231

http://biochem218.stanford.edu/

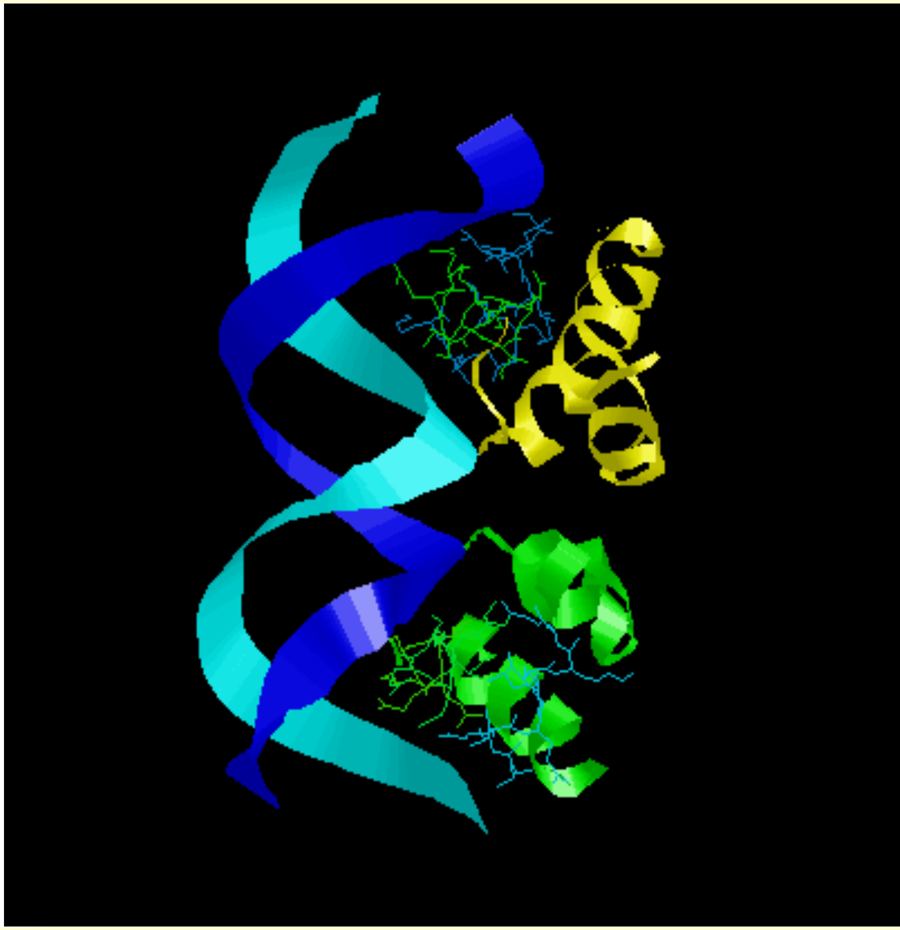
Sequence Alignment



Doug Brutlag
Professor Emeritus
Biochemistry & Medicine (by courtesy)

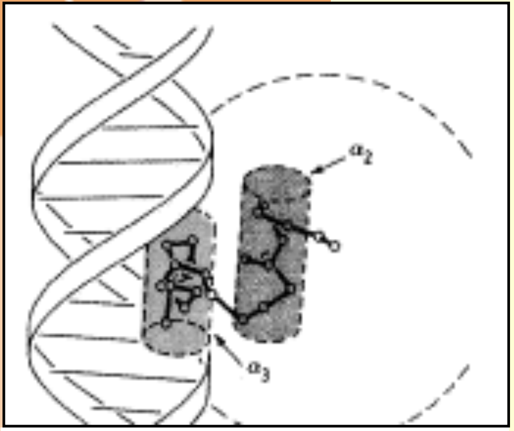
Position-Specific Scoring Matrix for Prokaryotic Helix-Turn-Helix Motifs

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



Doug Brutlag 2010

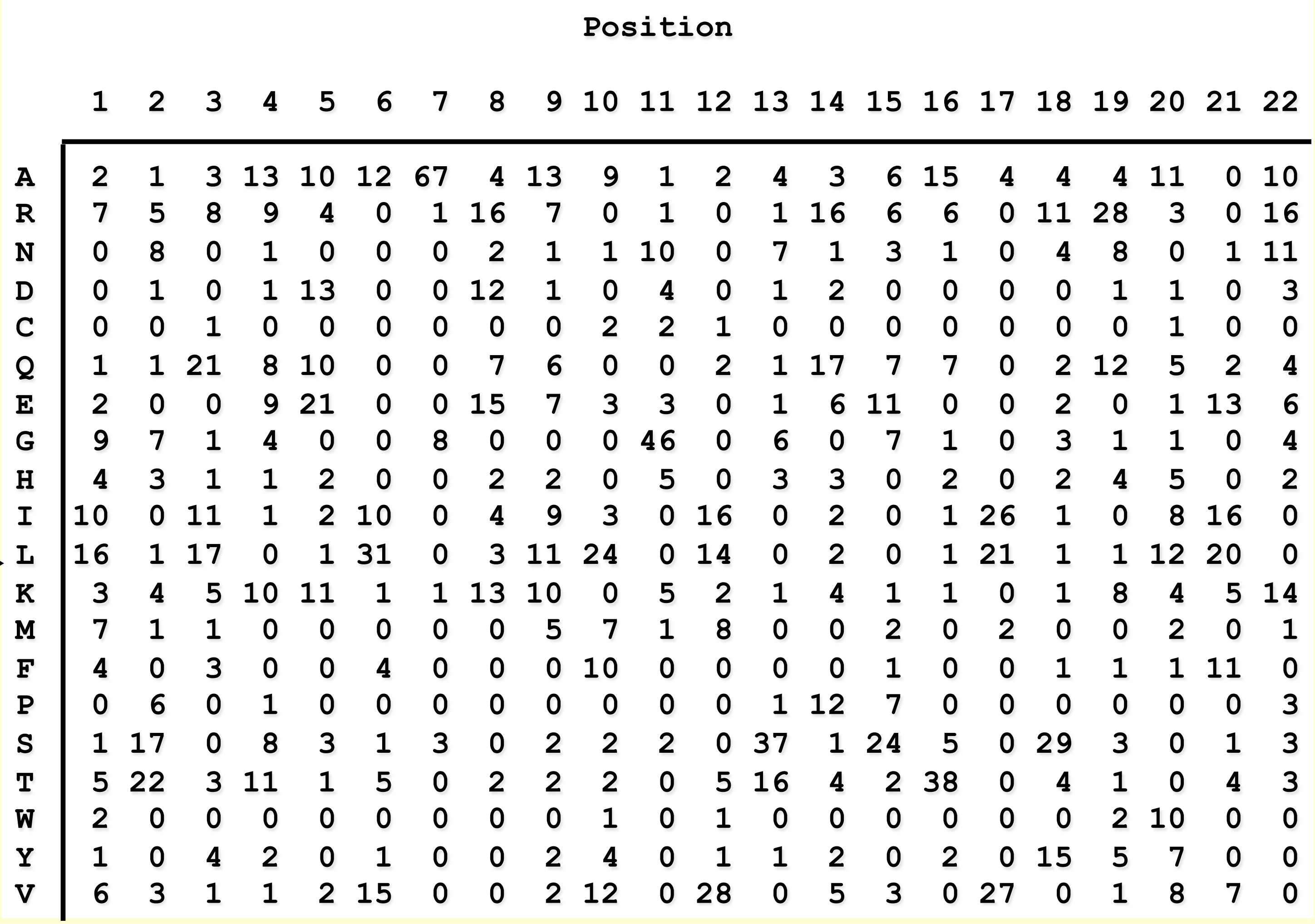
Position Specific Scoring Matrix for Prokaryotic Helix-Turn-Helix Motifs



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

Doug Brutlag 2010

Helix-Turn-Helix Weight Matrix

	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

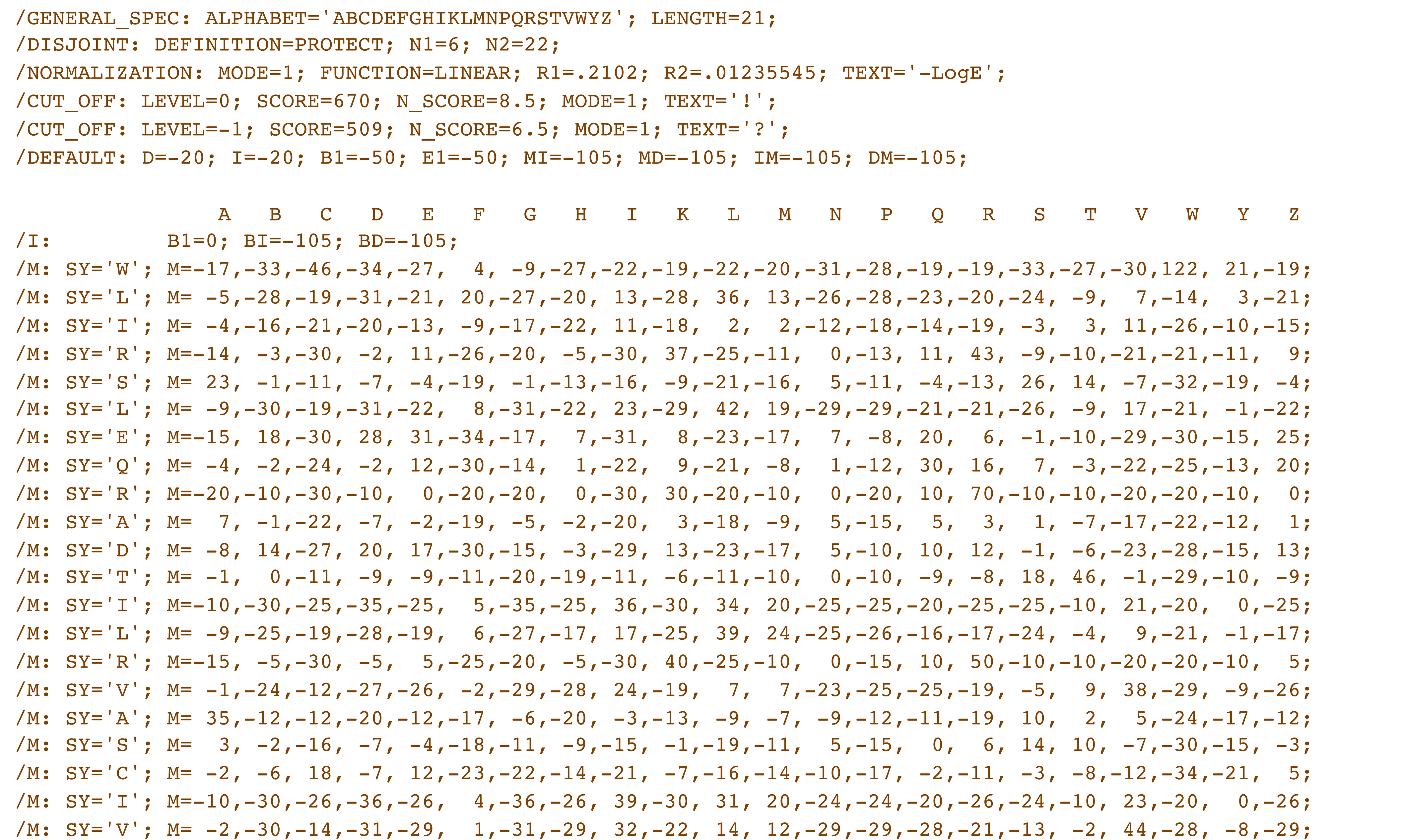
$$W_{ij} = \frac{\;\frac{N_{ij}}{N}\;}{f_i} \quad \text{where} \quad \begin{bmatrix} N_{ij} = & \text{number of amino acid of type i at position j} \\ N = & \text{number of sequences in training set, and} \\ f_i = & \text{frequency of amino acids of type } i \text{ in database} \end{bmatrix}$$

$$\text{Weight Matrix score for query of length L} = \sum_{j=1}^{L} \log W_{ij} = \sum \log\left(\frac{N_{ij}}{f_i}\right) - LN$$

Doug Brutlag 2010

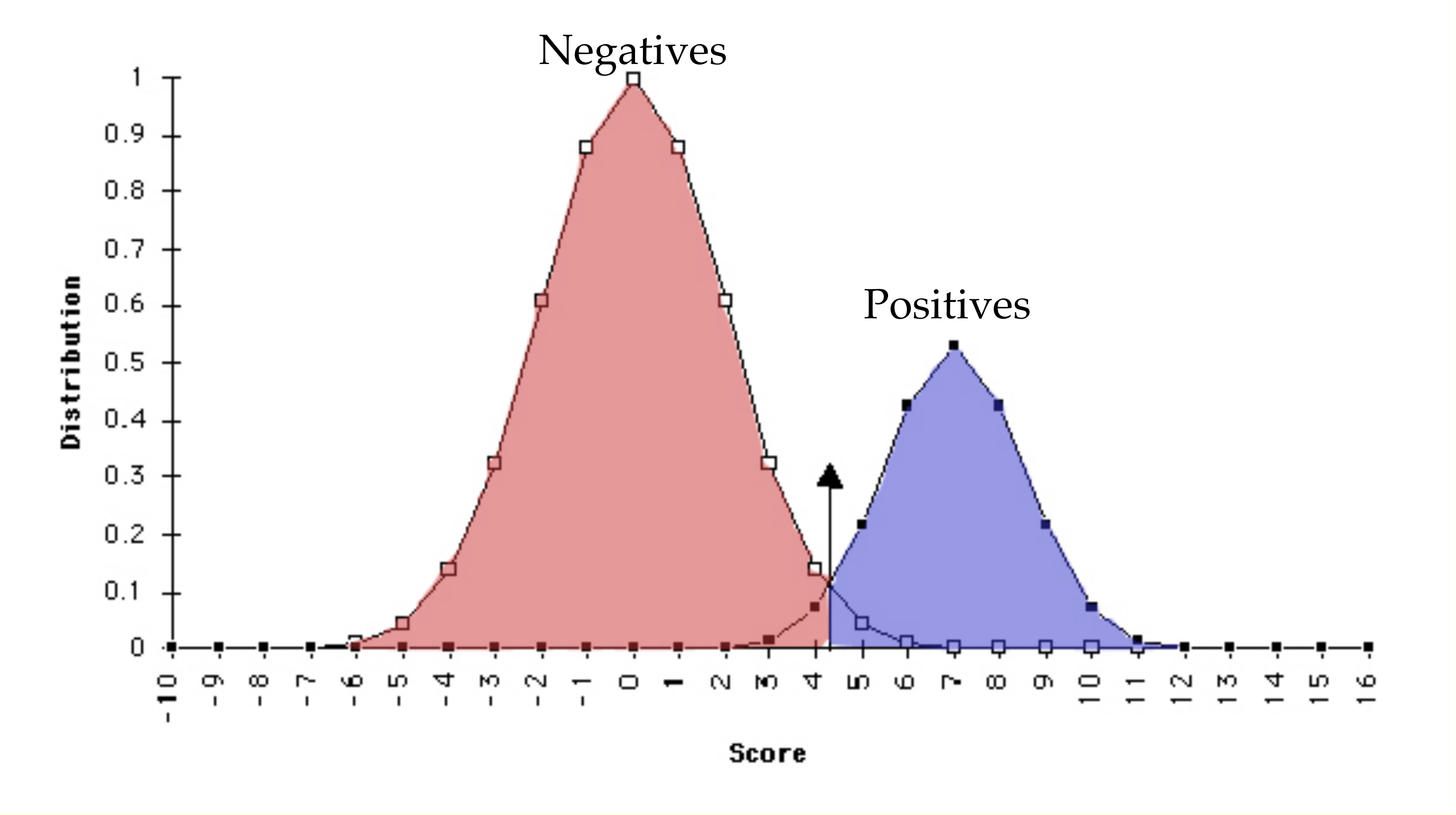
PSSM as a Scoring Matrix

http://ca.expasy.org/prosite/PS50044

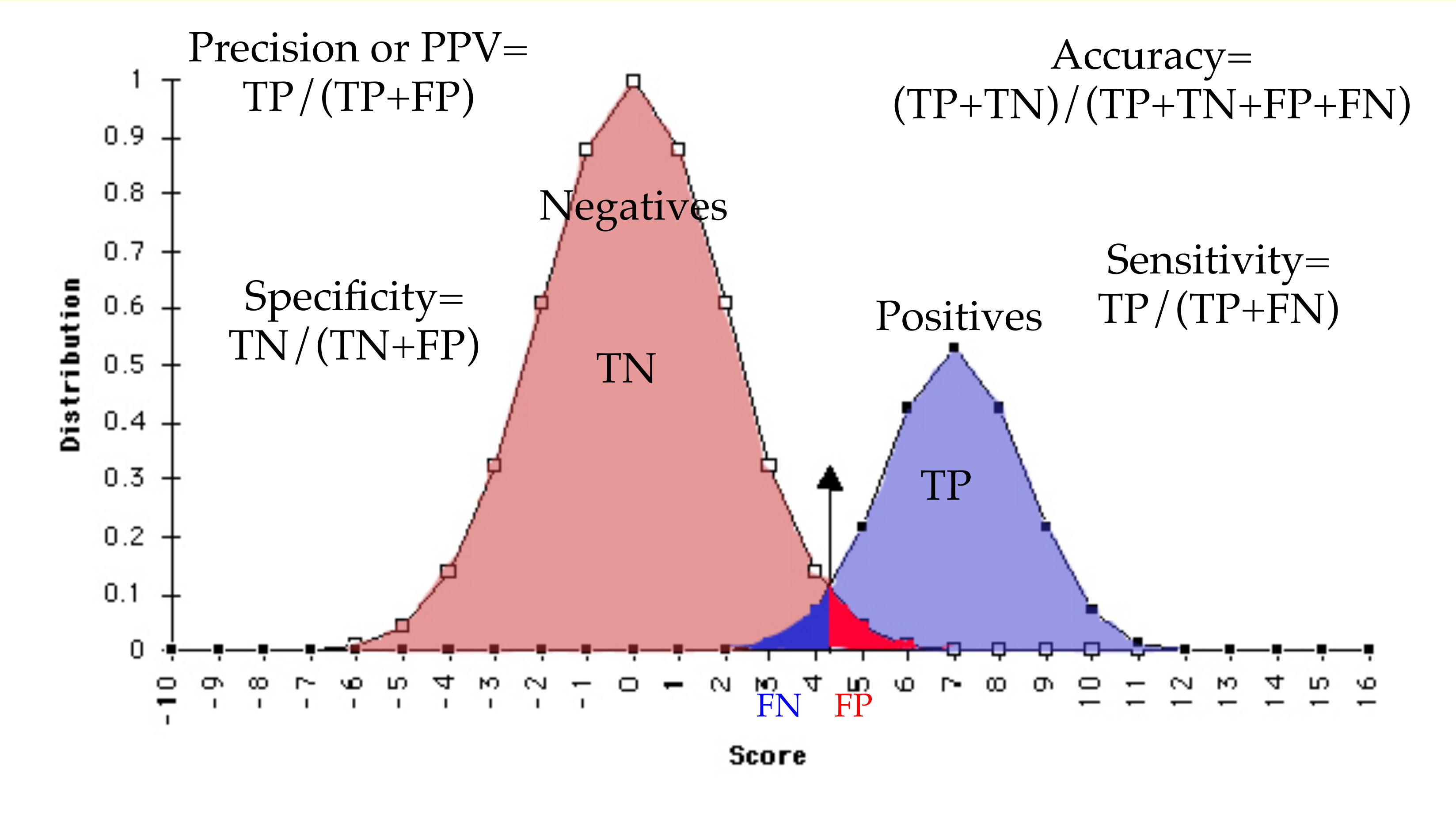


```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=21;
/DISJOINT: DEFINITION=PROTECT; N1=6; N2=22;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=.2102; R2=.01235545; TEXT='-LogE';
/CUT_OFF: LEVEL=0; SCORE=670; N_SCORE=8.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=509; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: D=-20; I=-20; B1=-50; E1=-50; MI=-105; MD=-105; IM=-105; DM=-105;

                 A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:           B1=0; BI=-105; BD=-105;
/M: SY='W'; M=-17,-33,-46,-34,-27,  4, -9,-27,-22,-19,-22,-20,-31,-28,-19,-19,-33,-27,-30,122, 21,-19;
/M: SY='L'; M= -5,-28,-19,-31,-21, 20,-27,-20, 13,-28, 36, 13,-26,-28,-23,-20,-24, -9,  7,-14,  3,-21;
/M: SY='I'; M= -4,-16,-21,-20,-13, -9,-17,-22, 11,-18,  2,  2,-12,-18,-14,-19, -3,  3, 11,-26,-10,-15;
/M: SY='R'; M=-14, -3,-30, -2, 11,-26,-20, -5,-30, 37,-25,-11,  0,-13, 11, 43, -9,-10,-21,-21,-11,  9;
/M: SY='S'; M= 23, -1,-11, -7, -4,-19, -1,-13,-16, -9,-21,-16,  5,-11, -4,-13, 26, 14, -7,-32,-19, -4;
/M: SY='L'; M= -9,-30,-19,-31,-22,  8,-31,-22, 23,-29, 42, 19,-29,-29,-21,-21,-26, -9, 17,-21, -1,-22;
/M: SY='E'; M=-15, 18,-30, 28, 31,-34,-17,  7,-31,  8,-23,-17,  7, -8, 20,  6, -1,-10,-29,-30,-15, 25;
/M: SY='Q'; M= -4, -2,-24, -2, 12,-30,-14,  1,-22,  9,-21, -8,  1,-12, 30, 16,  7, -3,-22,-25,-13, 20;
/M: SY='R'; M=-20,-10,-30,-10,  0,-20,-20,  0,-30, 30,-20,-10,  0,-20, 10, 70,-10,-10,-20,-20,-10,  0;
/M: SY='A'; M=  7, -1,-22, -7, -2,-19, -5, -2,-20,  3,-18, -9,  5,-15,  5,  3,  1, -7,-17,-22,-12,  1;
/M: SY='D'; M= -8, 14,-27, 20, 17,-30,-15, -3,-29, 13,-23,-17,  5,-10, 10, 12, -1, -6,-23,-28,-15, 13;
/M: SY='T'; M= -1,  0,-11, -9, -9,-11,-20,-19,-11, -6,-11,-10,  0,-10, -9, -8, 18, 46, -1,-29,-10, -9;
/M: SY='I'; M=-10,-30,-25,-35,-25,  5,-35,-25, 36,-30, 34, 20,-25,-25,-20,-25,-25,-10, 21,-20,  0,-25;
/M: SY='L'; M= -9,-25,-19,-28,-19,  6,-27,-17, 17,-25, 39, 24,-25,-26,-16,-17,-24, -4,  9,-21, -1,-17;
/M: SY='R'; M=-15, -5,-30, -5,  5,-25,-20, -5,-30, 40,-25,-10,  0,-15, 10, 50,-10,-10,-20,-20,-10,  5;
/M: SY='V'; M= -1,-24,-12,-27,-26, -2,-29,-28, 24,-19,  7,  7,-23,-25,-25,-19, -5,  9, 38,-29, -9,-26;
/M: SY='A'; M= 35,-12,-12,-20,-12,-17, -6,-20, -3,-13, -9, -7, -9,-12,-11,-19, 10,  2,  5,-24,-17,-12;
/M: SY='S'; M=  3, -2,-16, -7, -4,-18,-11, -9,-15, -1,-19,-11,  5,-15,  0,  6, 14, 10, -7,-30,-15, -3;
/M: SY='C'; M= -2, -6, 18, -7, 12,-23,-22,-14,-21, -7,-16,-14,-10,-17, -2,-11, -3, -8,-12,-34,-21,  5;
/M: SY='I'; M=-10,-30,-26,-36,-26,  4,-36,-26, 39,-30, 31, 20,-24,-24,-20,-26,-24,-10, 23,-20,  0,-26;
/M: SY='V'; M= -2,-30,-14,-31,-29,  1,-31,-29, 32,-22, 14, 12,-29,-29,-28,-21,-13, -2, 44,-28, -8,-29;
```

Evaluation of Classifiers



Evaluation of Classifiers



Precision or PPV=
TP/(TP+FP)

Accuracy=
(TP+TN)/(TP+TN+FP+FN)

Specificity=
TN/(TN+FP)

Sensitivity=
TP/(TP+FN)

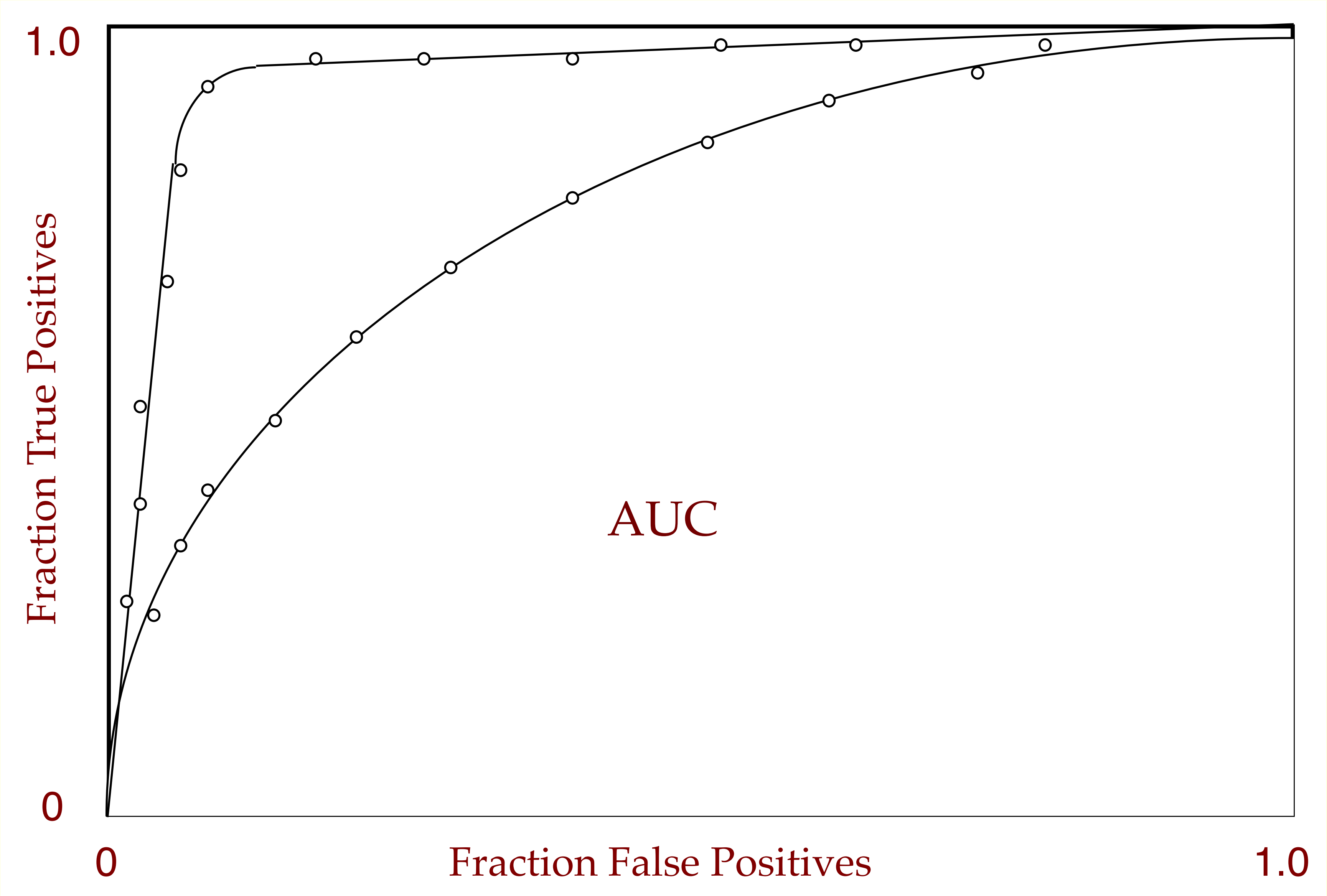
Doug Brutlag 2010

Criteria Used to Select Threshold

- Minimize the False Negatives
- Minimize False Positives
- Minimize Total Misclassified Cases
- Maximize Specific Utility Function
- Optimize Arbitrary Objective Function

Receiver-Operator Characteristic Shows Sensitivity versus Specificity with Threshold

ROC Curve



Homework Assignment 3

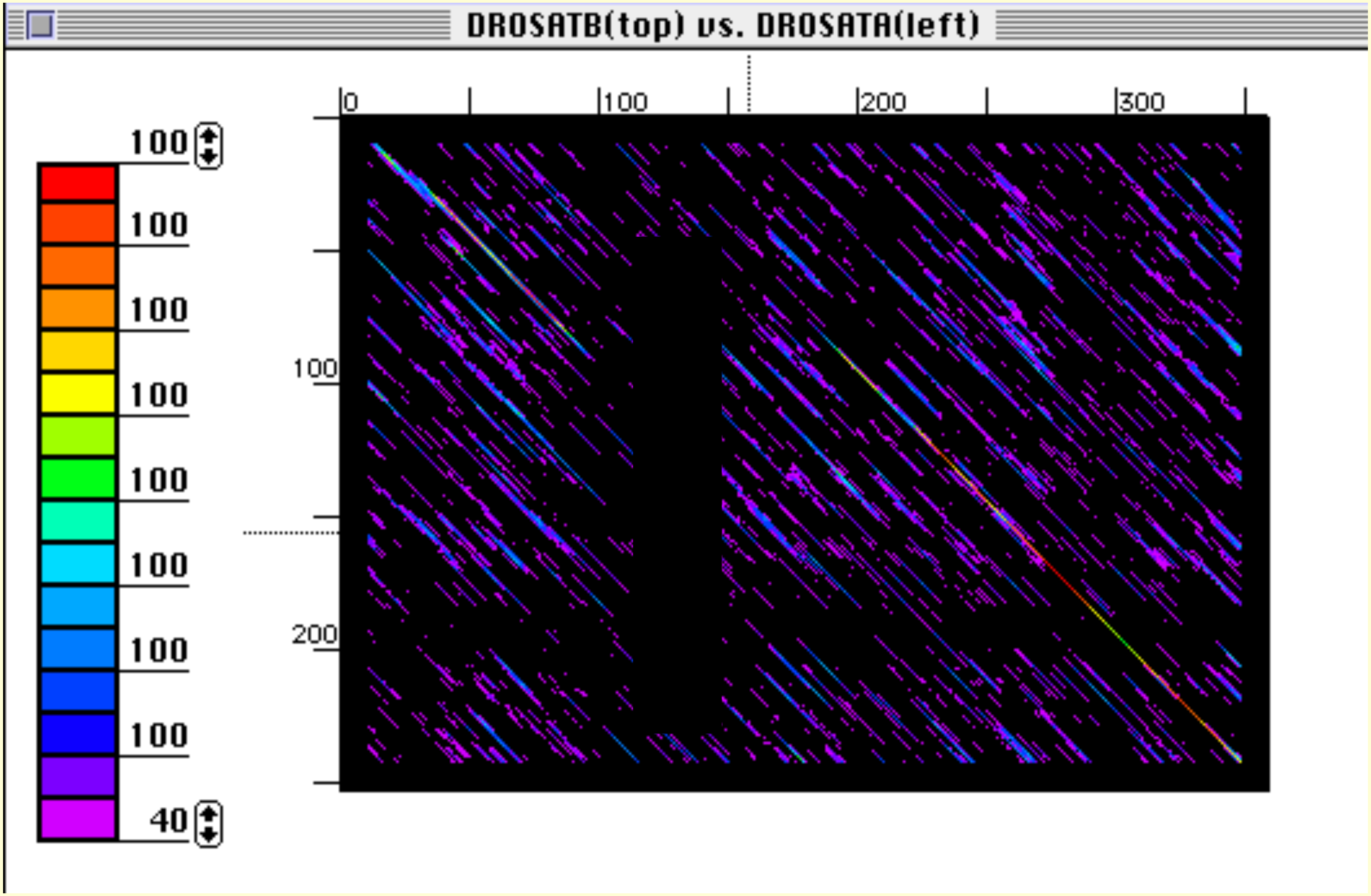
http://biochem218.stanford.edu/03Homework.pdf

1. Select a protein of interest to you from UniProt/SwissProt database whose function is well known and well characterized. Obtain the FASTA format of the protein and the Gene Ontology terms associated with your protein.
2. Search your protein for similar sequences using the BLAST method on the UniProt site. Please report two or three hits which are both statistically and biologically significant. Also report two or three hits which you think are neither statistically nor biologically significant. If your protein family is very large, you may have to ask BLAST to return more hits to find statistically insignificant hits.
3. Search your protein for motifs with the MyHits Motif Scan Query. Be sure to Include Protsite Patterns, Prosite Frequent Patterns, Prosite Profiles, Prefiles, Pfam HMMSs (local Models) in your search. Please send the MyHits you think are biologically significant and at least 1 or 2 hits which you think are not statistically or biologically significant. Please note that only the Profiles have expectation values. The patterns do not have a measure of statistical significance.
4. Search your protein for motifs using the InterPro database. Please report a few of the InterPro domains hits you think are significant and any hits which you think are not statistically or biologically significant. Please note that the default graphic output of InterPro does not list expectation values. You must switch to the Tabular view to obtain the statistical significance.
5. Are the results from these functional searches compatible with the gene ontology terms associated with your protein? Did you discover any statistically significant functional similarities or motifs not represented by the known gene ontology terms?

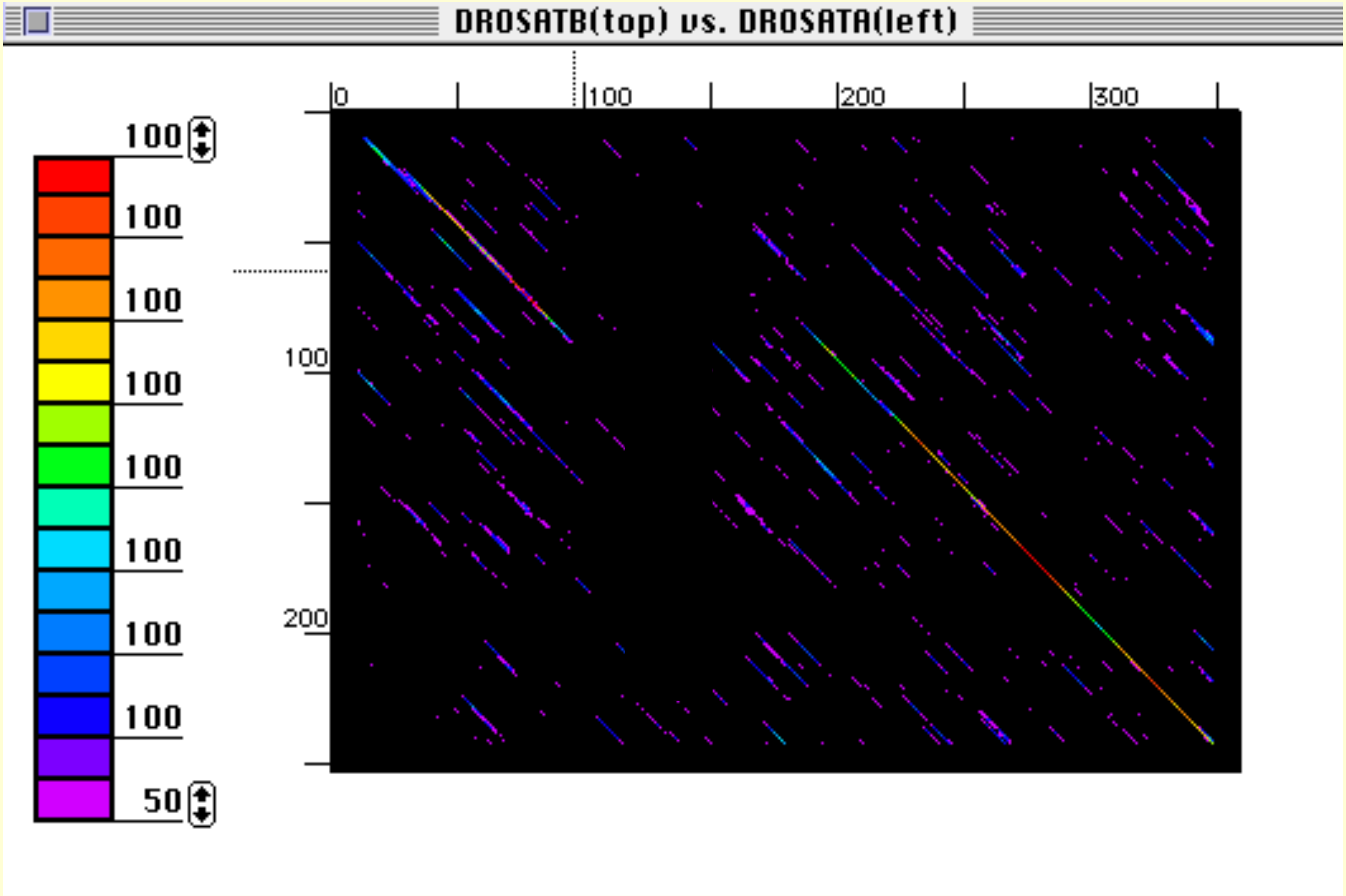
Biological vs. Statistical Significance

- Statistically significant results always have biological significance.
- Statistically insignificant similarities or motifs may still be biologically significant, especially those at the borderline of statistical significance.
- Biologically significant results that are not statistically significant can often be detected by multiple observations.
- Biological significance can have multiple hierarchical interpretation or meaning.
- Algorithms that miss biologically significant results should be improved to more accurately reflect the biology.

DNA Dot Matrix

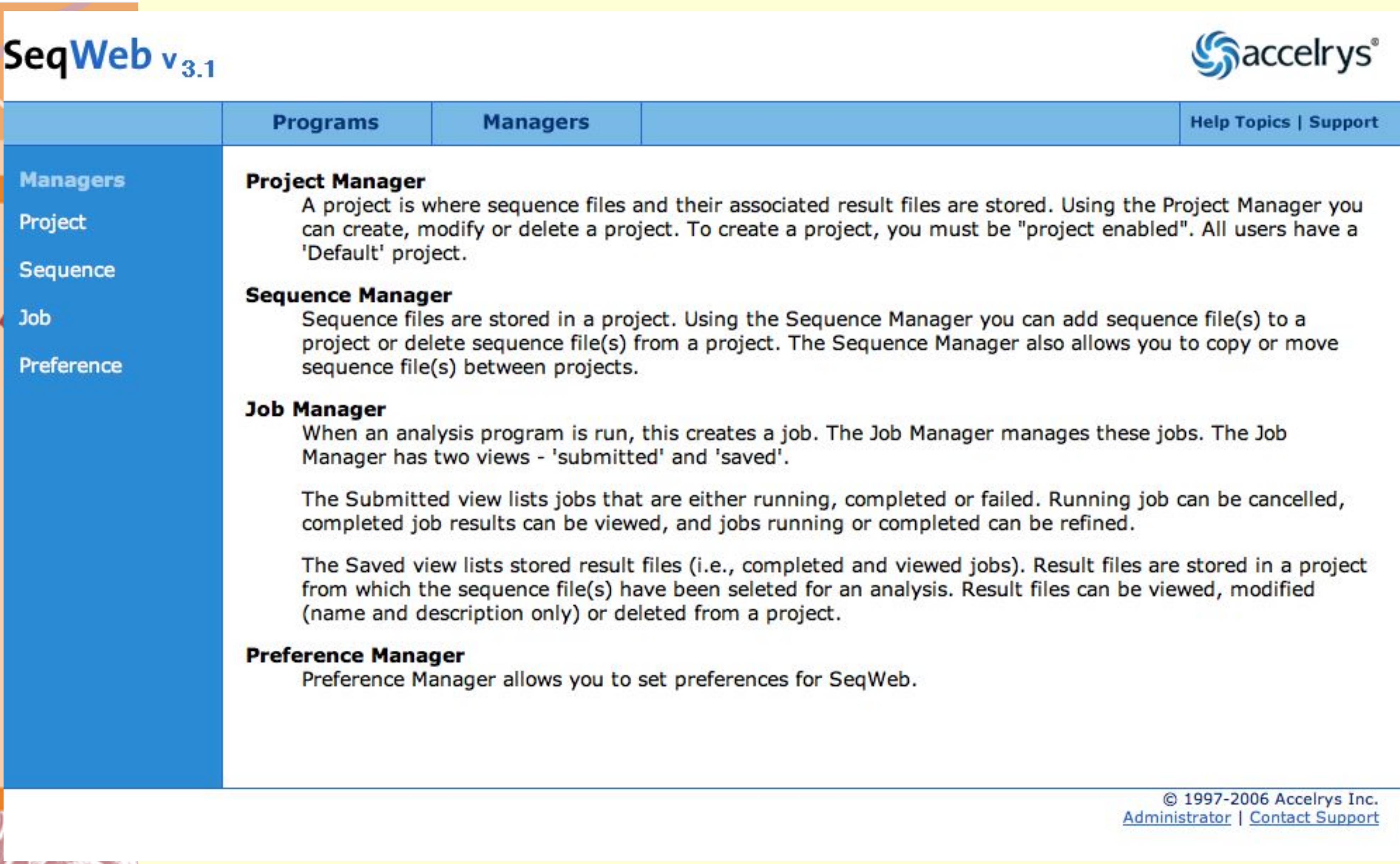


DNA Dot Matrix (2)



SeqWeb's Compare DotPlot

http://seqweb.stanford.edu:81/



Project Manager

A project is where sequence files and their associated result files are stored. Using the Project Manager you can create, modify or delete a project. To create a project, you must be "project enabled". All users have a 'Default' project.

Sequence Manager

Sequence files are stored in a project. Using the Sequence Manager you can add sequence file(s) to a project or delete sequence file(s) from a project. The Sequence Manager also allows you to copy or move sequence file(s) between projects.

Job Manager

When an analysis program is run, this creates a job. The Job Manager manages these jobs. The Job Manager has two views - 'submitted' and 'saved'.

The Submitted view lists jobs that are either running, completed or failed. Running job can be cancelled, completed job results can be viewed, and jobs running or completed can be refined.

The Saved view lists stored result files (i.e., completed and viewed jobs). Result files are stored in a project from which the sequence file(s) have been seleted for an analysis. Result files can be viewed, modified (name and description only) or deleted from a project.

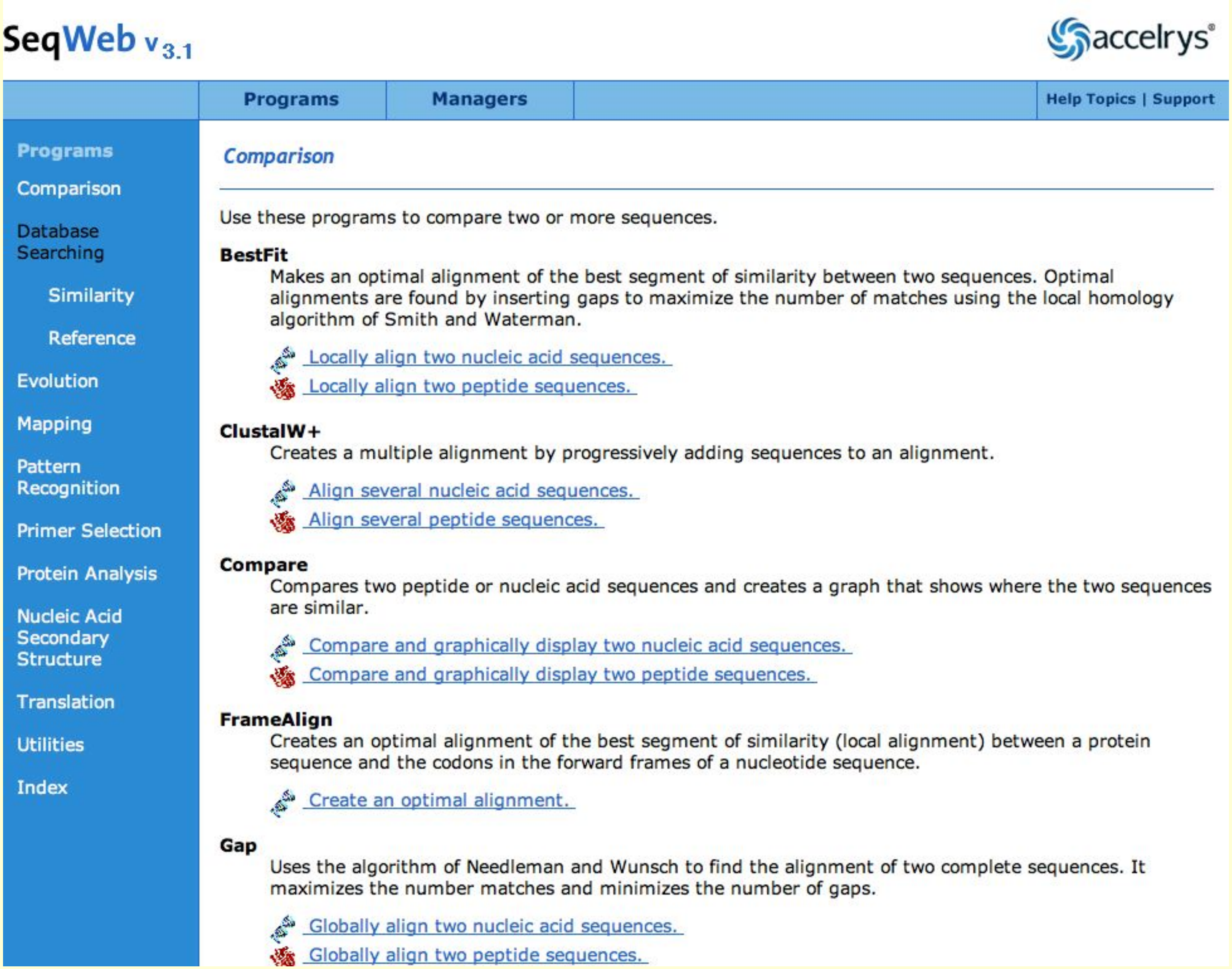
Preference Manager

Preference Manager allows you to set preferences for SeqWeb.

Administrator | Contact Support

SeqWeb's Comparison Programs

http://seqweb.stanford.edu:81/gcg-bin/programs.cgi?name=comparison



SeqWeb v 3.1

Comparison

Use these programs to compare two or more sequences.

BestFit

Makes an optimal alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maximize the number of matches using the local homology algorithm of Smith and Waterman.

- Locally align two nucleic acid sequences.
- Locally align two peptide sequences.

ClustalW+

Creates a multiple alignment by progressively adding sequences to an alignment.

- Align several nucleic acid sequences.
- Align several peptide sequences.

Compare

Compares two peptide or nucleic acid sequences and creates a graph that shows where the two sequences are similar.

- Compare and graphically display two nucleic acid sequences.
- Compare and graphically display two peptide sequences.

FrameAlign

Creates an optimal alignment of the best segment of similarity (local alignment) between a protein sequence and the codons in the forward frames of a nucleotide sequence.

- Create an optimal alignment.

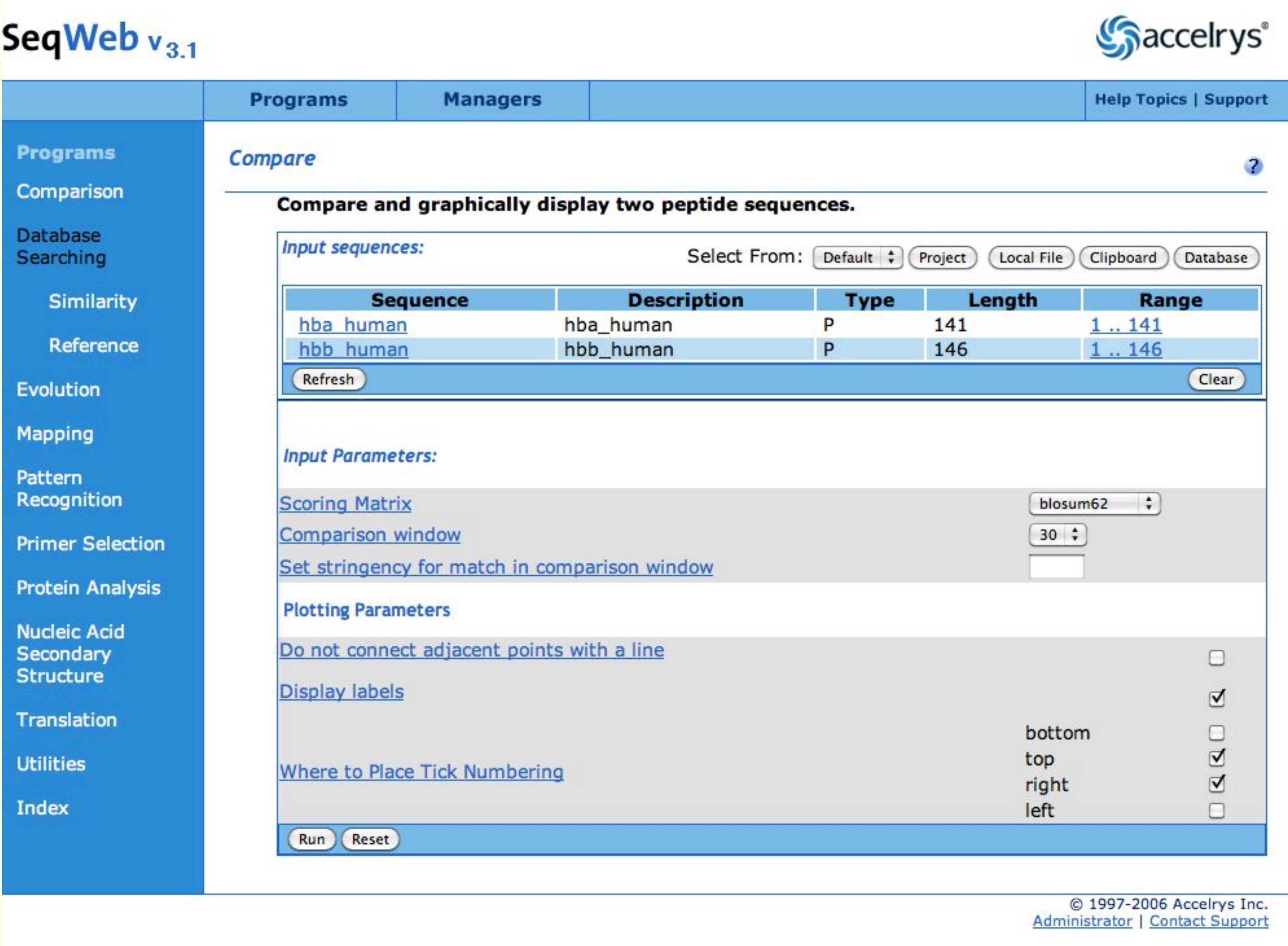
Gap

Uses the algorithm of Needleman and Wunsch to find the alignment of two complete sequences. It maximizes the number matches and minimizes the number of gaps.

- Globally align two nucleic acid sequences.
- Globally align two peptide sequences.

SeqWeb's Compare Peptide Sequences

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=compdot-prot



Compare

Compare and graphically display two peptide sequences.

Input sequences:

Sequence	Description	Type	Length	Range
hba_human	hba_human	P	141	1 .. 141
hbb_human	hbb_human	P	146	1 .. 146

Input Parameters:

Scoring Matrix: blosum62

Comparison window: 30

Set stringency for match in comparison window:

Plotting Parameters

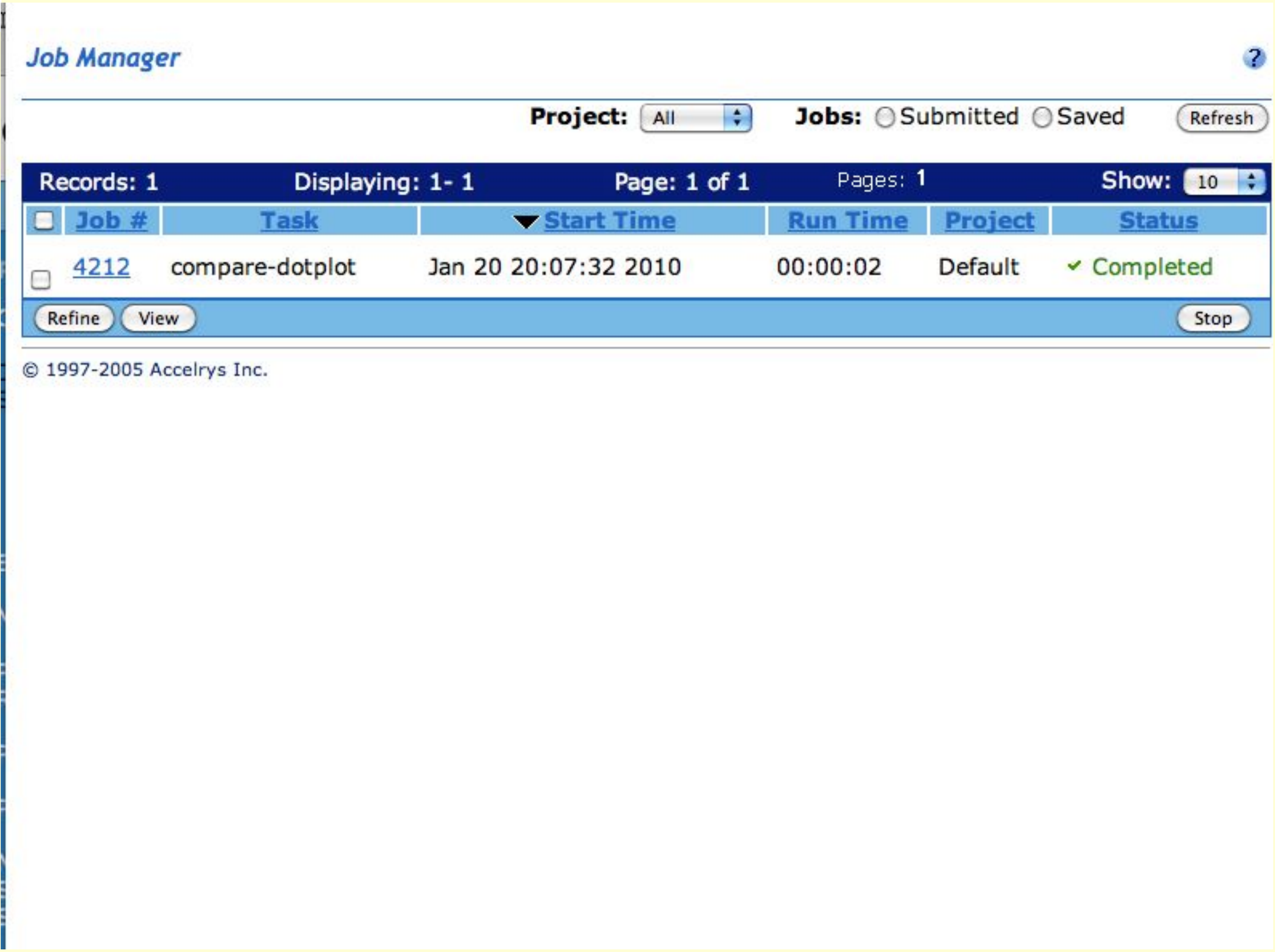
Do not connect adjacent points with a line □

Display labels ☑

Where to Place Tick Numbering
bottom □
top ☑
right ☑
left □

SeqWeb's Job Manger

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=compdot-prot



Job Manager

Project: All **Jobs:** ○ Submitted ○ Saved

Records: 1 Displaying: 1- 1 Page: 1 of 1 Pages: 1 Show: 10

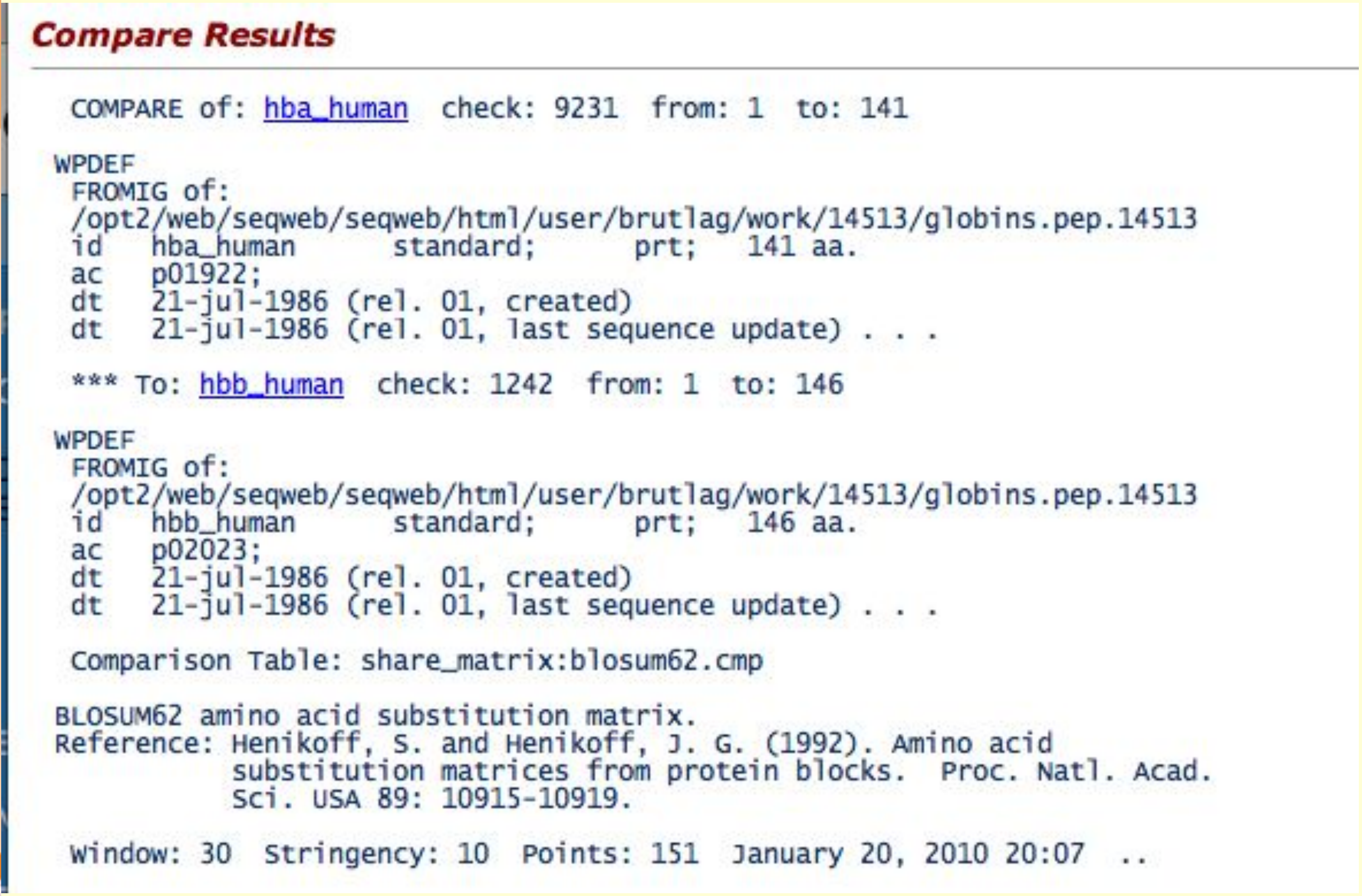
□	Job #	Task	▼ Start Time	Run Time	Project	Status
□	4212	compare-dotplot	Jan 20 20:07:32 2010	00:00:02	Default	✓ Completed

© 1997-2005 Accelrys Inc.

Doug Brutlag 2010

SeqWeb’s Compare Results

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=compdot-prot

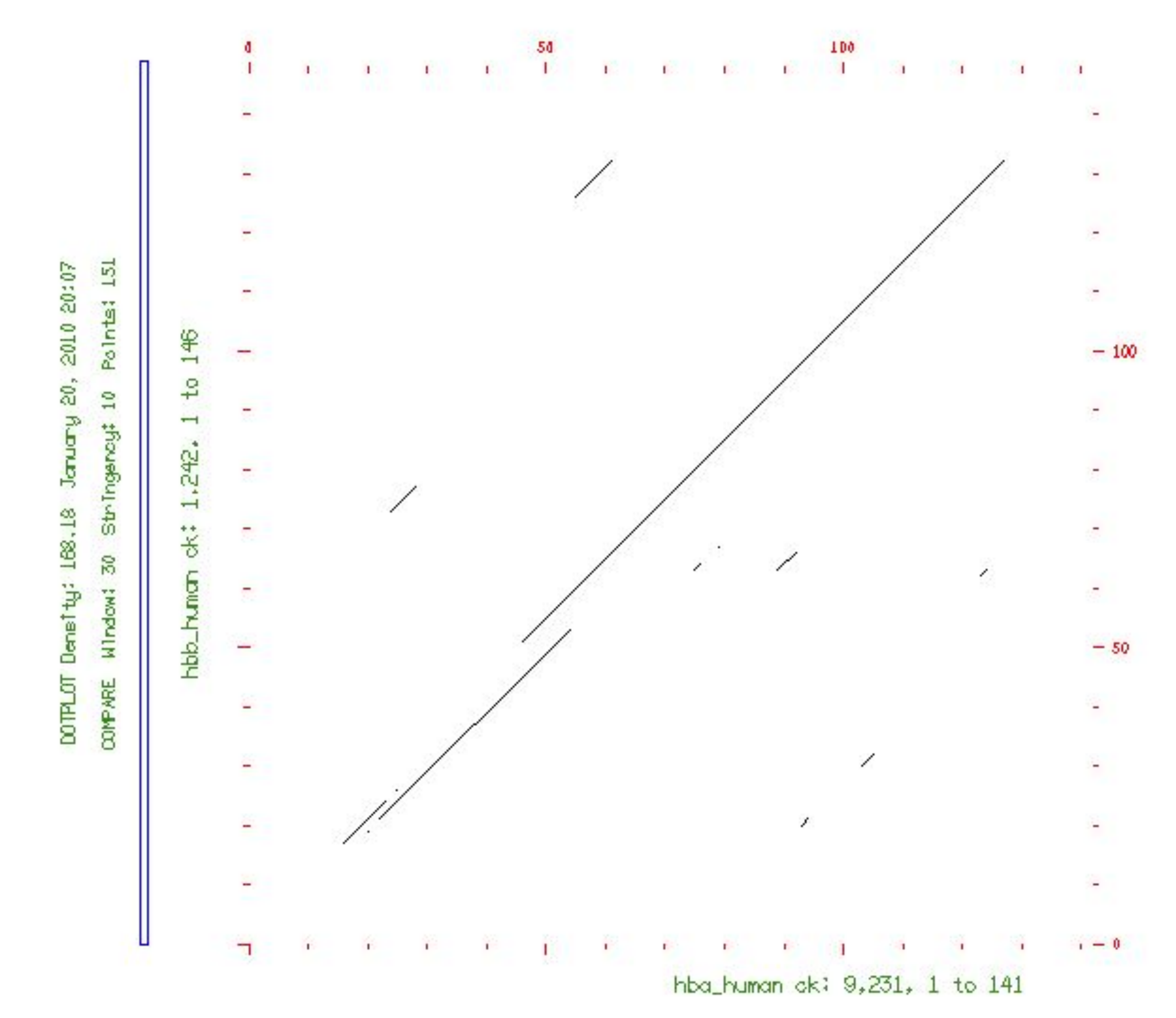


Compare Results

```
 COMPARE of: hba_human  check: 9231  from: 1  to: 141

WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hba_human      standard;      prt;   141 aa.
 ac    p01922;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .

 *** To: hbb_human  check: 1242  from: 1  to: 146

WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hbb_human      standard;      prt;   146 aa.
 ac    p02023;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .

 Comparison Table: share_matrix:blosum62.cmp

BLOSUM62 amino acid substitution matrix.
Reference: Henikoff, S. and Henikoff, J. G. (1992). Amino acid
           substitution matrices from protein blocks.  Proc. Natl. Acad.
           Sci. USA 89: 10915-10919.

 Window: 30  Stringency: 10  Points: 151  January 20, 2010 20:07  ..
```

SeqWeb's Compare Results

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=compdot-prot



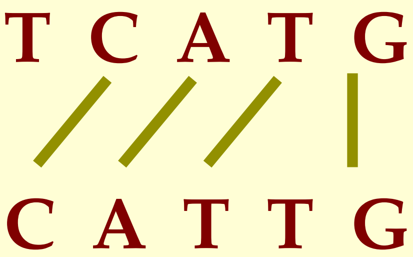
```
DOTPLOT Density: 168.18 January 20, 2010 20:07
COMPARE Window: 30 Stringency: 10 Points: 151
```

Sequence Alignment Problem

T C A T G

C A T T G

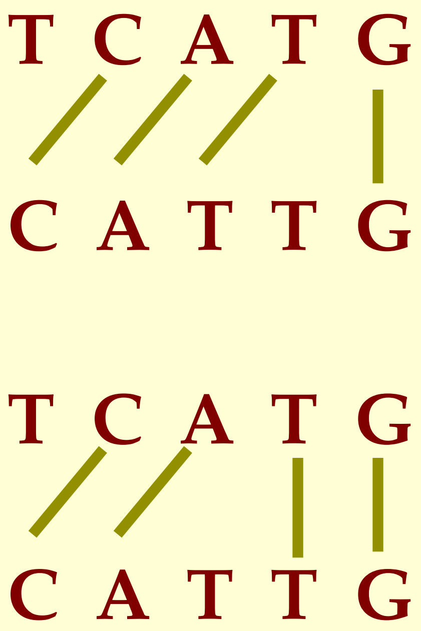
Sequence Alignment Problem



T C A T G

C A T T G

Sequence Alignment Problem



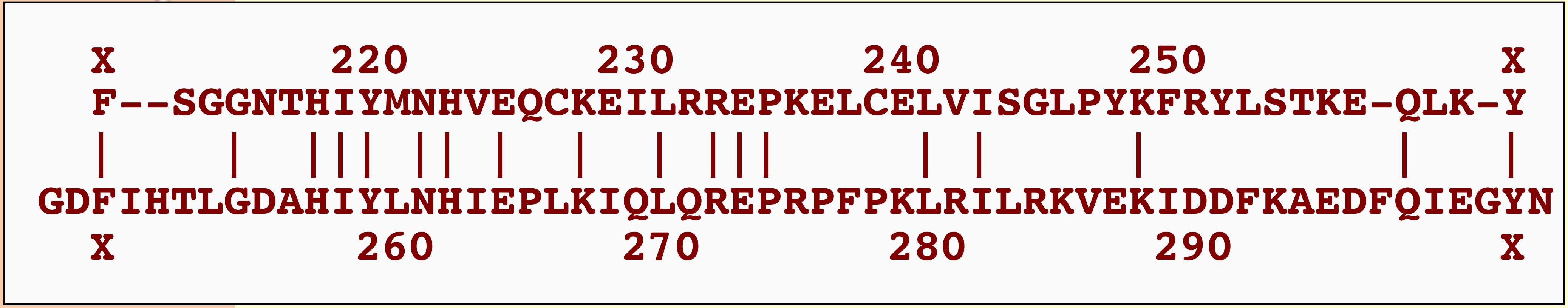
T C A T G

C A T T G

T C A T G

C A T T G

Sequence Alignment
Exact Matches Only

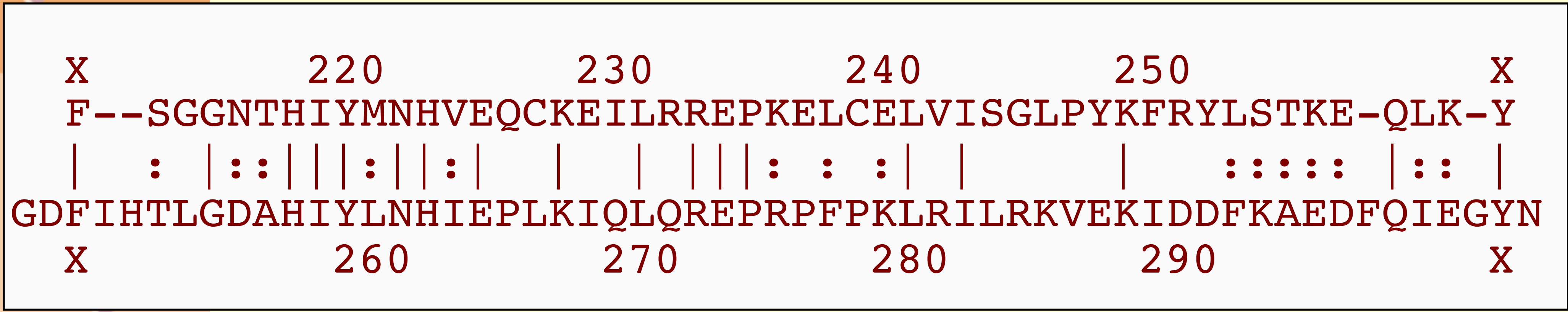


```
  X        220       230       240       250           X
  F--SGGNTHIYMNHVEQCKEILRREPKELCELVISGLPYKFRYLSTKE-QLK-Y
  |    |  ||| || |  |  | |||     | |     |         |   |
GDFIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKAEDFQIEGYN
  X         260       270       280       290          X
```

Sequence Alignment
Amino Acid Similarity

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

Sequence Alignment and Typical Objective Function



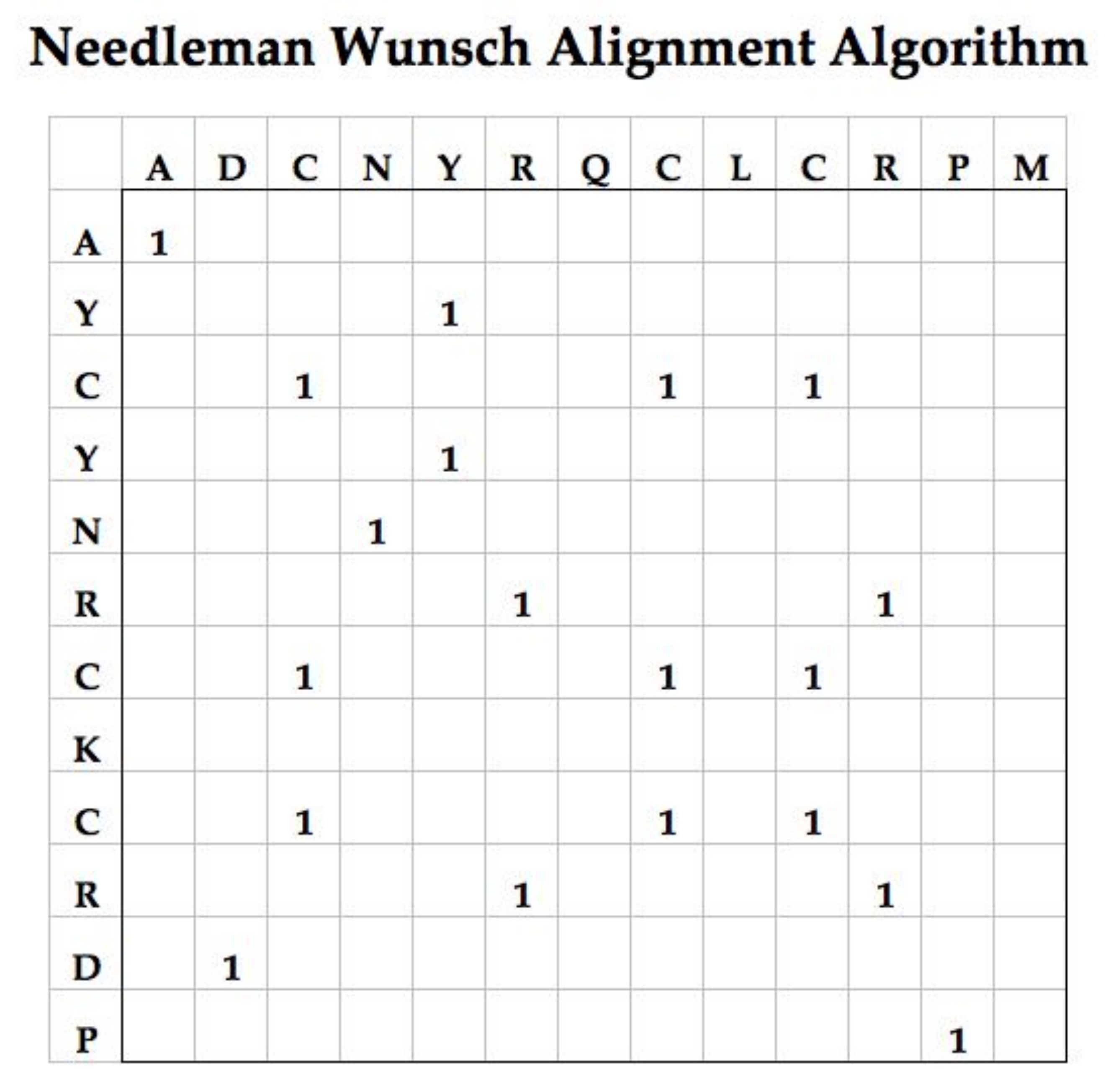
```
  x        220       230       240       250           x
  F--SGGNTHIYMNHVEQCKEILRREPKELCELVISGLPYKFRYLSTKE-QLK-Y
  |  : |::|||:||:|  |  | |||: : :| |     |   ::::: |:: |
GDFIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKAEDFQIEGYN
  x         260       270       280       290          x
```

$$Score = \sum_{Re\,gion\_Start}^{Re\,gion\_End} Similarity\_Weights - \sum_{Re\,gion\_start}^{Re\,gion\_End} Gap\_Penalties$$

$$where:$$

$$Gap\_Penalty = Gap\_Start\_Penalty + (Gap\_Size - 1) * Gap\_Size\_Penalty$$

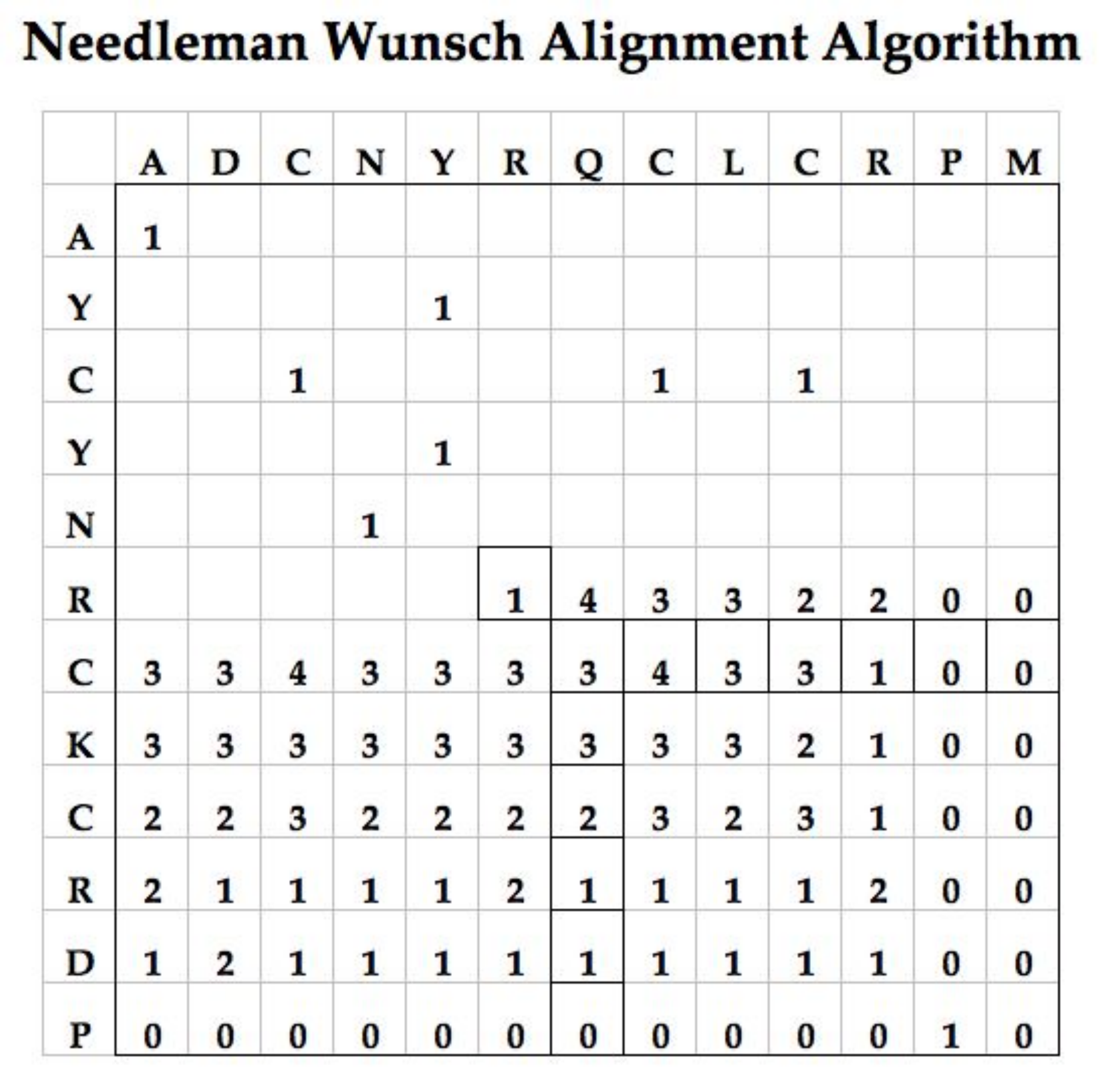
Needleman-Wunsch Alignment Algorithm Matches and Mismatches



Needleman Wunsch Alignment Algorithm

	A	D	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
D		1											
P												1	

Needleman-Wunsch Alignment Algorithm Recursion

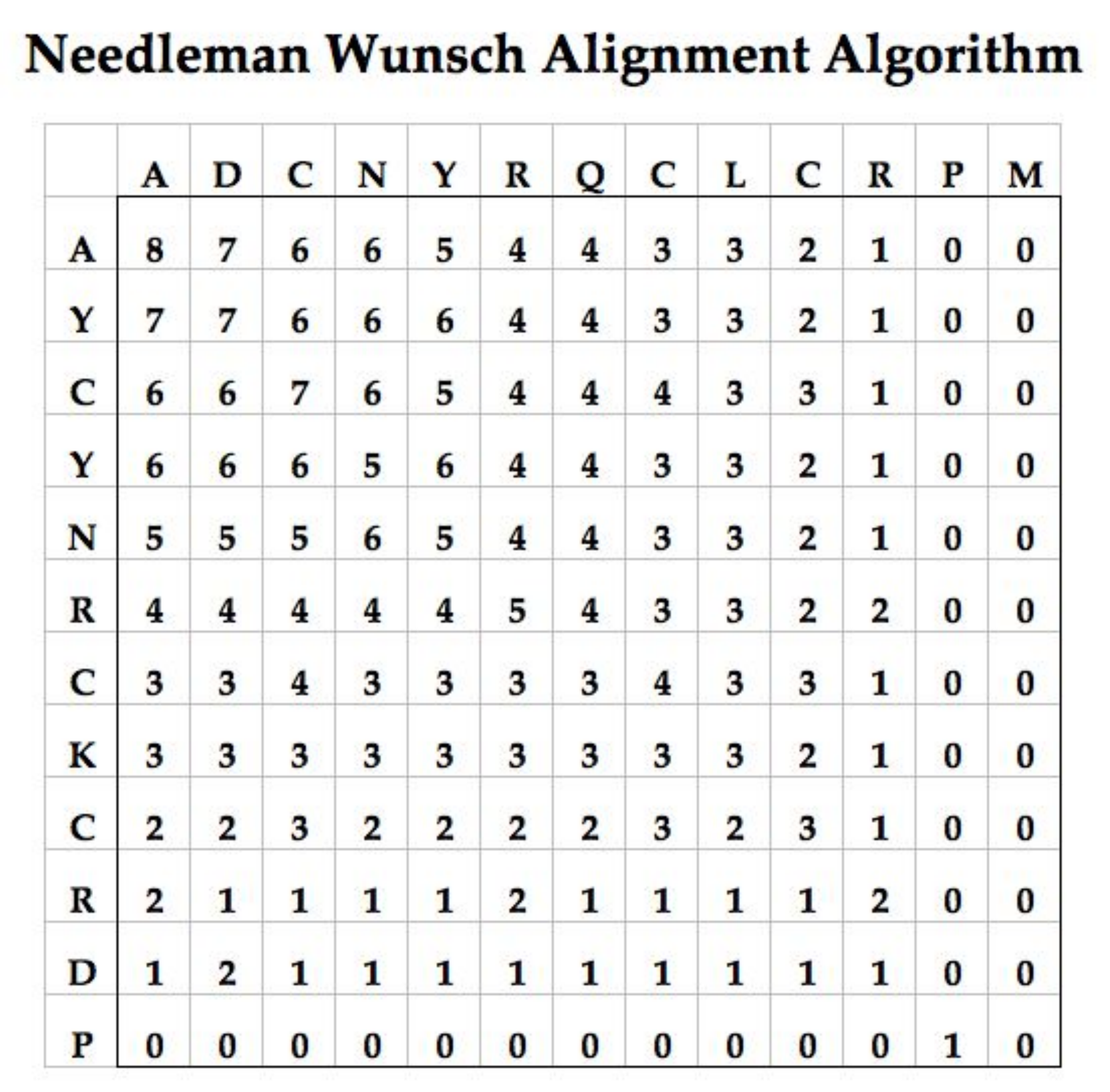


Needleman Wunsch Alignment Algorithm

	A	D	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
D	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Doug Brutlag 2010

Needleman-Wunsch Alignment Algorithm Maximal Scores

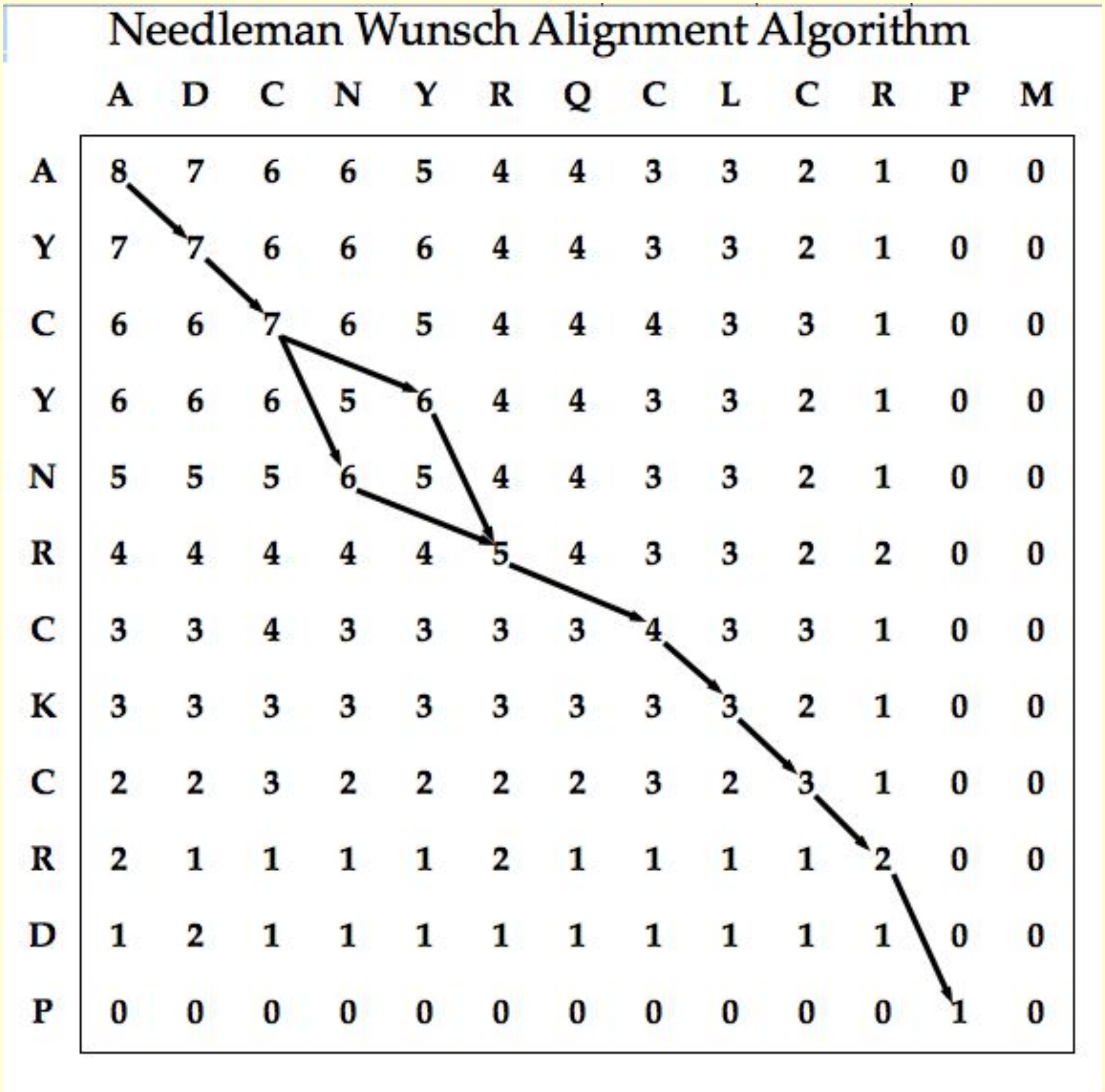


Needleman Wunsch Alignment Algorithm

	A	D	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
D	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Doug Brutlag 2010

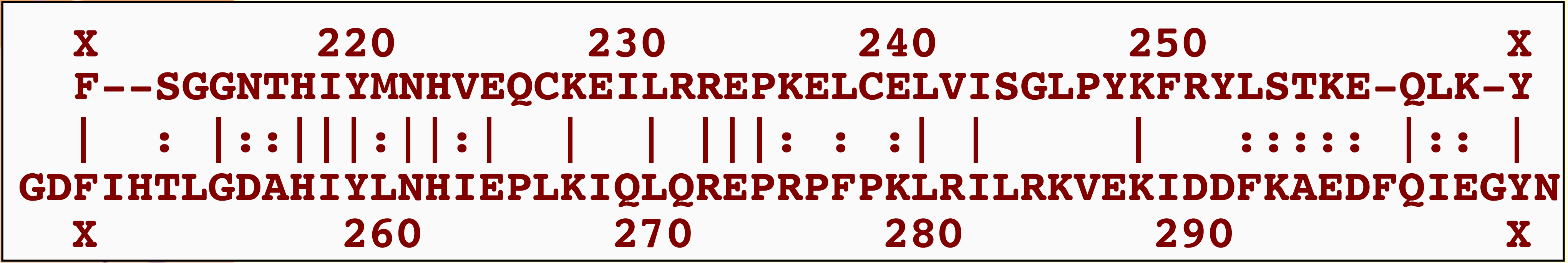
Needleman-Wunsch Alignment Algorithm Trace Back



Needleman Wunsch Alignment Algorithm

	A	D	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
D	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Sequence Alignment and Typical Objective Function



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

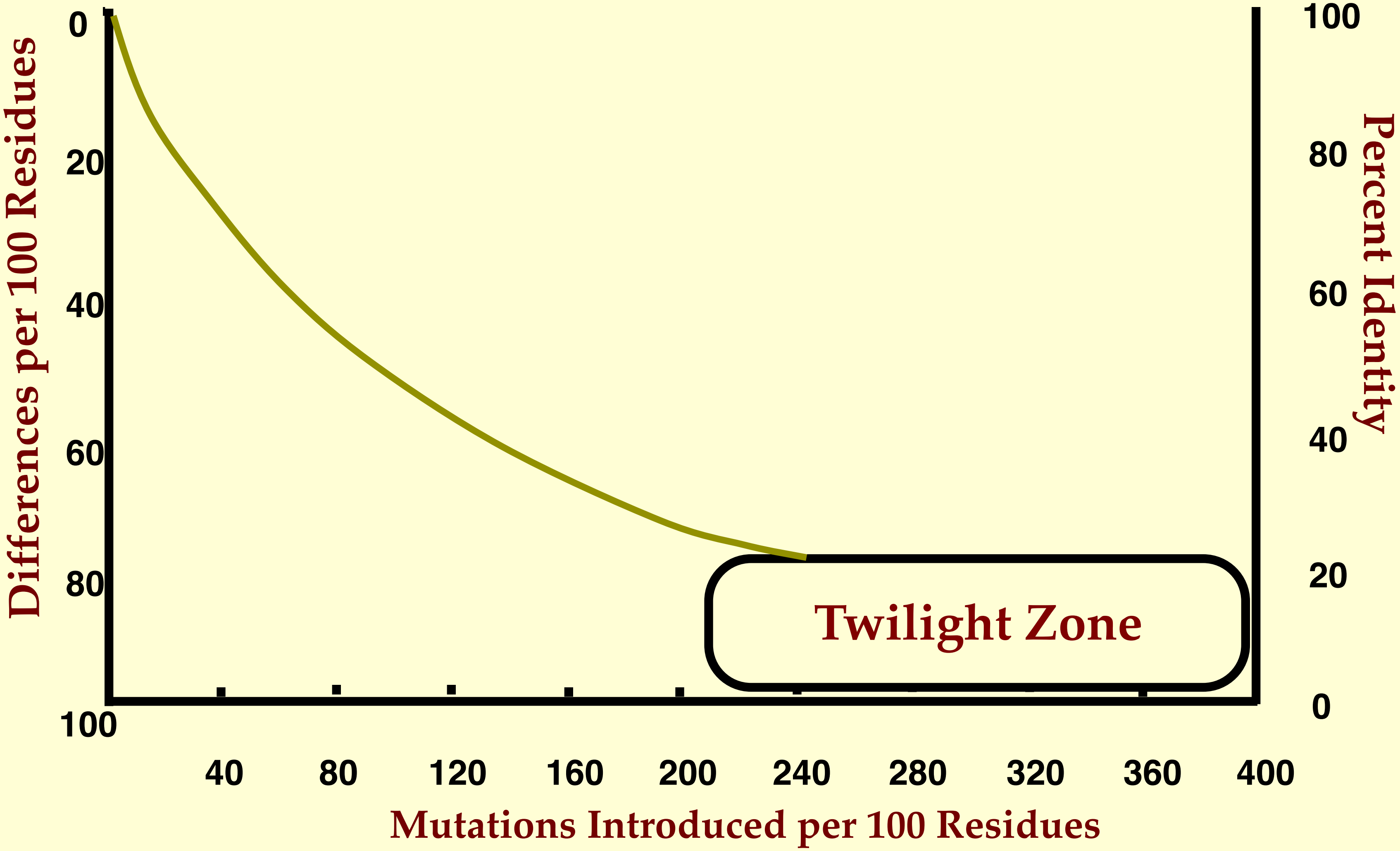
$$Score = \sum_{\mathrm{Re}\,gion\_Start}^{\mathrm{Re}\,gion\_End} Similarity\_Weights - \sum_{\mathrm{Re}\,gion\_start}^{\mathrm{Re}\,gion\_End} Gap\_Penalties$$

where:

$$Gap\_Penalty = Gap\_Start\_Penalty + Gap\_Size * Gap\_Size\_Penalty$$

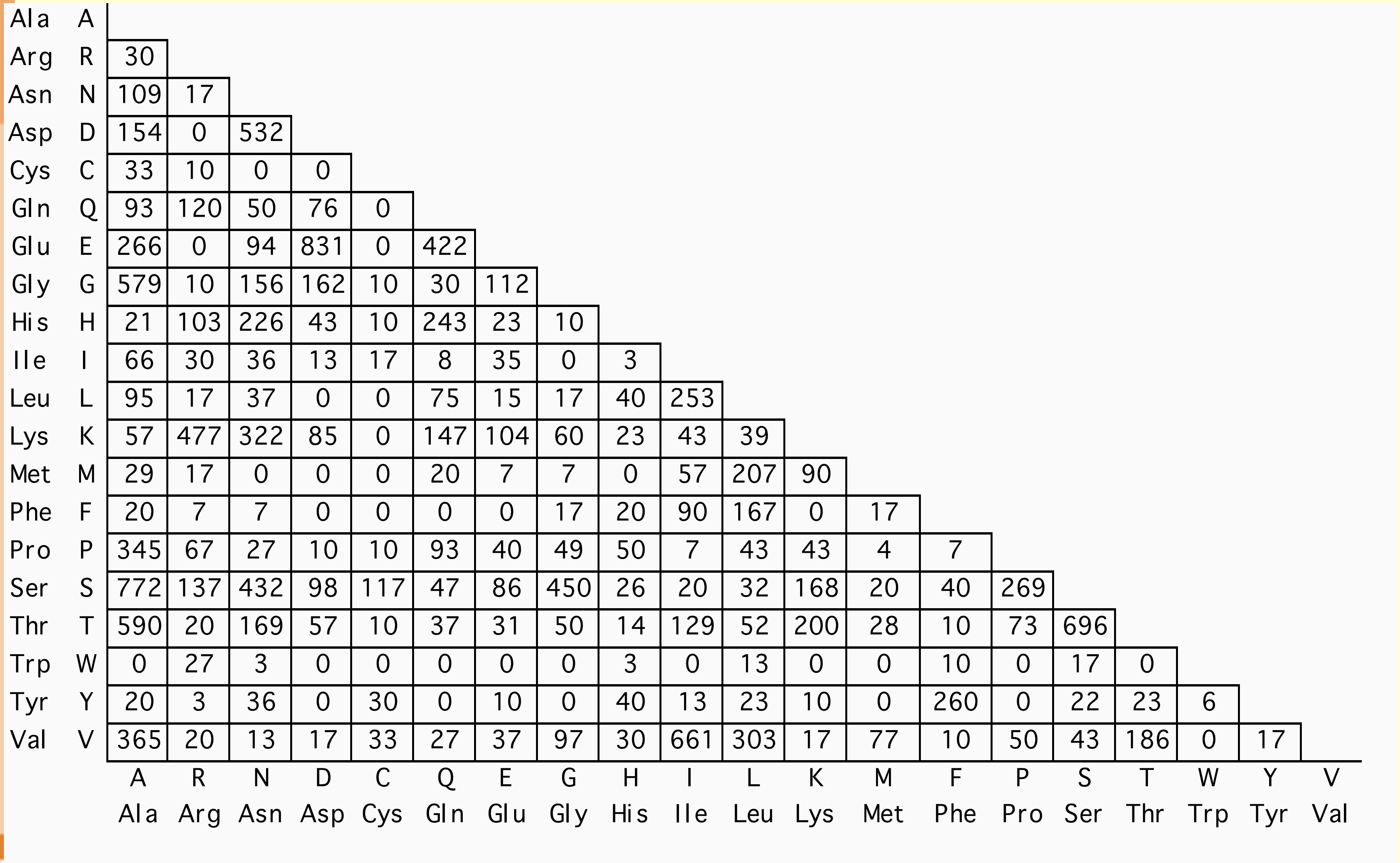
Doug Brutlag 2010

Sequence Similarity vs Evolutionary Distance



After Russ Doolittle

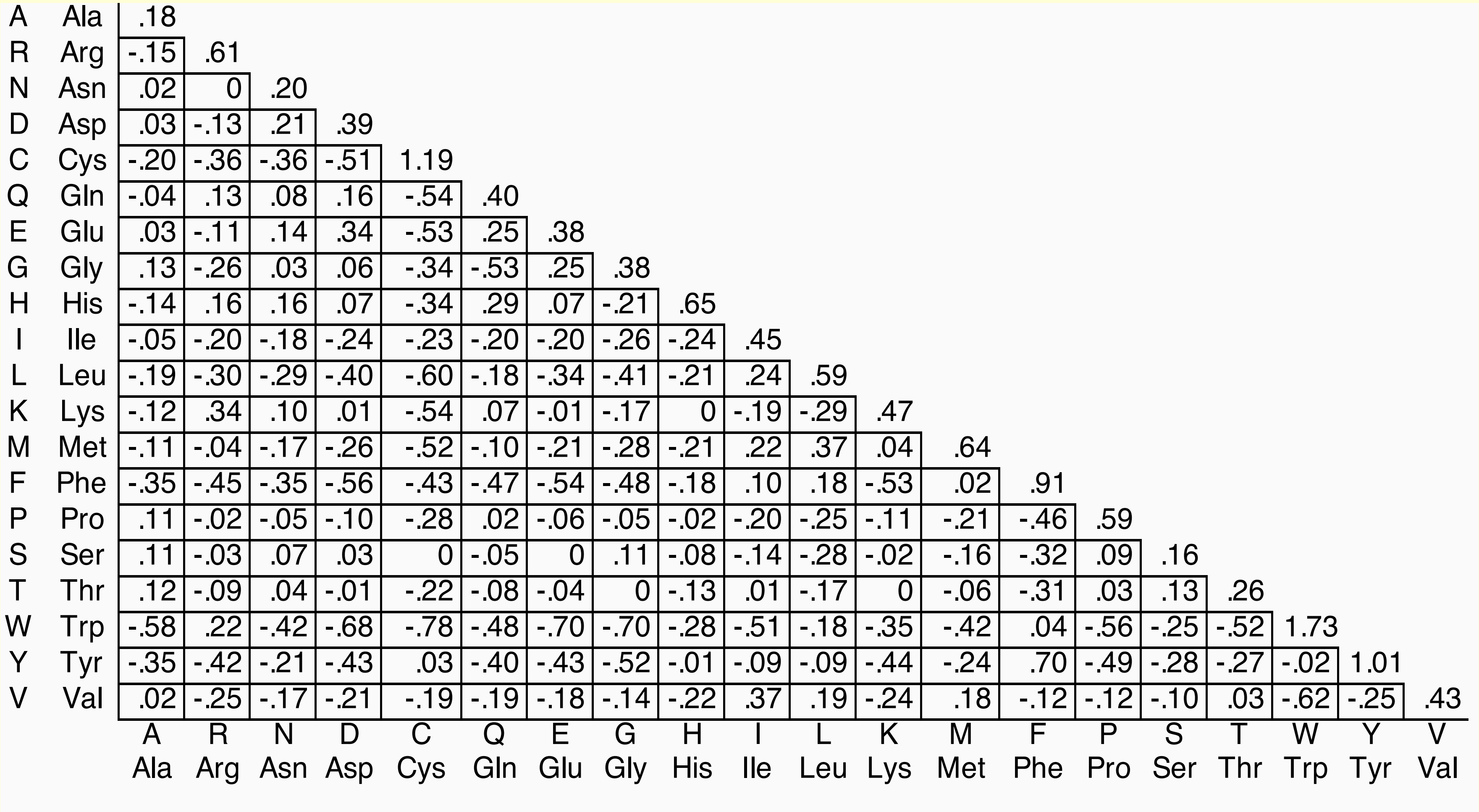
Dayhoff's Acceptable Point Mutations (PAMs)



Ala	A																				
Arg	R	30																			
Asn	N	109	17																		
Asp	D	154	0	532																	
Cys	C	33	10	0	0																
Gln	Q	93	120	50	76	0															
Glu	E	266	0	94	831	0	422														
Gly	G	579	10	156	162	10	30	112													
His	H	21	103	226	43	10	243	23	10												
Ile	I	66	30	36	13	17	8	35	0	3											
Leu	L	95	17	37	0	0	75	15	17	40	253										
Lys	K	57	477	322	85	0	147	104	60	23	43	39									
Met	M	29	17	0	0	0	20	7	7	0	57	207	90								
Phe	F	20	7	7	0	0	0	0	17	20	90	167	0	17							
Pro	P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
Ser	S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
Thr	T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
Trp	W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Tyr	Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
Val	V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

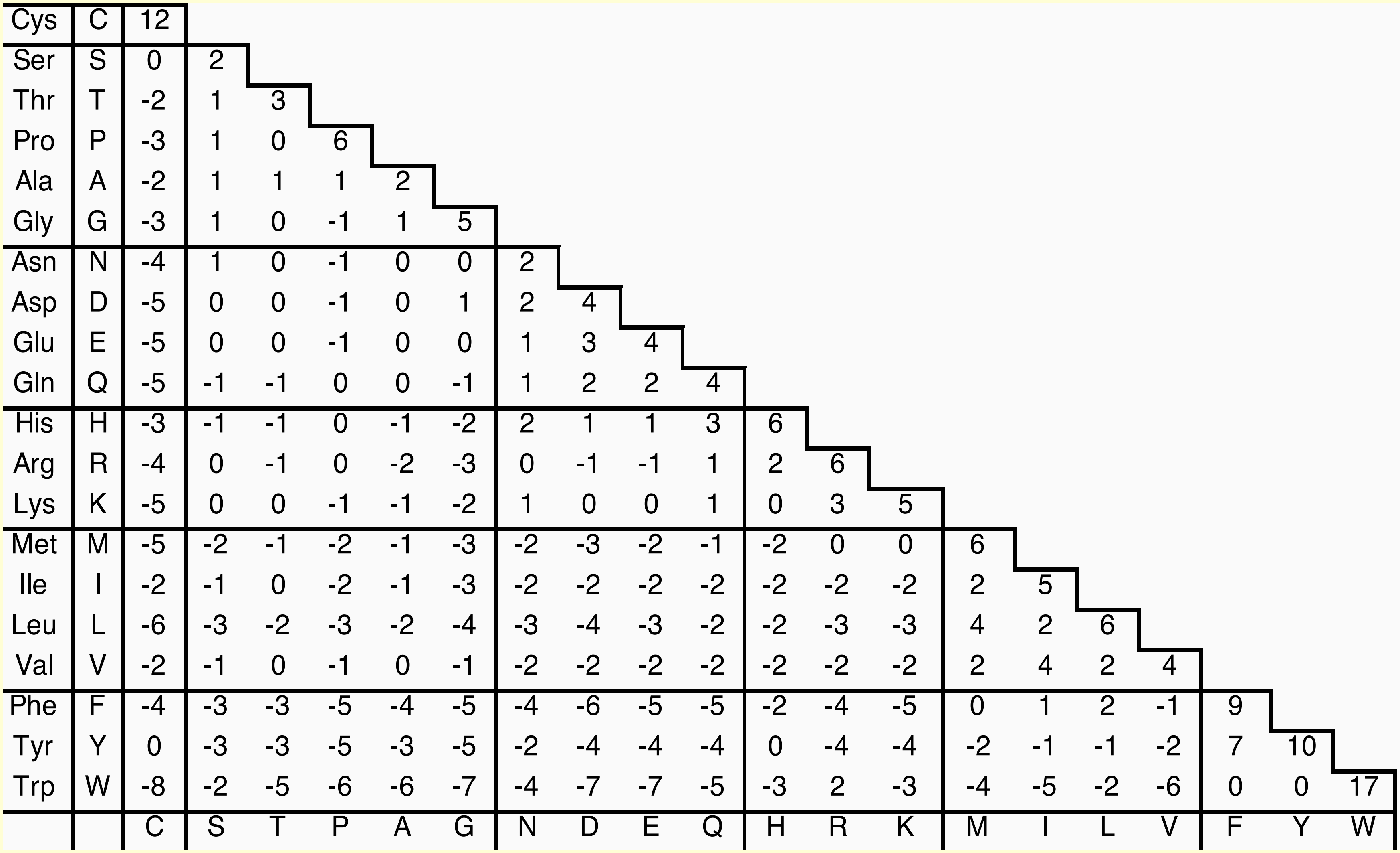
Doug Brutlag 2010

Dayhoff’s PAM 250 Matrix (Log-Odds Form)



A	Ala	.18																			
R	Arg	-.15	.61																		
N	Asn	.02	0	.20																	
D	Asp	.03	-.13	.21	.39																
C	Cys	-.20	-.36	-.36	-.51	1.19															
Q	Gln	-.04	.13	.08	.16	-.54	.40														
E	Glu	.03	-.11	.14	.34	-.53	.25	.38													
G	Gly	.13	-.26	.03	.06	-.34	-.53	.25	.38												
H	His	-.14	.16	.16	.07	-.34	.29	.07	-.21	.65											
I	Ile	-.05	-.20	-.18	-.24	-.23	-.20	-.20	-.26	-.24	.45										
L	Leu	-.19	-.30	-.29	-.40	-.60	-.18	-.34	-.41	-.21	.24	.59									
K	Lys	-.12	.34	.10	.01	-.54	.07	-.01	-.17	0	-.19	-.29	.47								
M	Met	-.11	-.04	-.17	-.26	-.52	-.10	-.21	-.28	-.21	.22	.37	.04	.64							
F	Phe	-.35	-.45	-.35	-.56	-.43	-.47	-.54	-.48	-.18	.10	.18	-.53	.02	.91						
P	Pro	.11	-.02	-.05	-.10	-.28	.02	-.06	-.05	-.02	-.20	-.25	-.11	-.21	-.46	.59					
S	Ser	.11	-.03	.07	.03	0	-.05	0	.11	-.08	-.14	-.28	-.02	-.16	-.32	.09	.16				
T	Thr	.12	-.09	.04	-.01	-.22	-.08	-.04	0	-.13	.01	-.17	0	-.06	-.31	.03	.13	.26			
W	Trp	-.58	.22	-.42	-.68	-.78	-.48	-.70	-.70	-.28	-.51	-.18	-.35	-.42	.04	-.56	-.25	-.52	1.73		
Y	Tyr	-.35	-.42	-.21	-.43	.03	-.40	-.43	-.52	-.01	-.09	-.09	-.44	-.24	.70	-.49	-.28	-.27	-.02	1.01	
V	Val	.02	-.25	-.17	-.21	-.19	-.19	-.18	-.14	-.22	.37	.19	-.24	.18	-.12	-.12	-.10	.03	-.62	-.25	.43
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

Dayhoff's PAM 250 Matrix (1978)



Cys	C	12																			
Ser	S	0	2																		
Thr	T	-2	1	3																	
Pro	P	-3	1	0	6																
Ala	A	-2	1	1	1	2															
Gly	G	-3	1	0	-1	1	5														
Asn	N	-4	1	0	-1	0	0	2													
Asp	D	-5	0	0	-1	0	1	2	4												
Glu	E	-5	0	0	-1	0	0	1	3	4											
Gln	Q	-5	-1	-1	0	0	-1	1	2	2	4										
His	H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
Arg	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
Lys	K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
Met	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
Ile	I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
Leu	L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
Val	V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
Phe	F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Tyr	Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Trp	W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

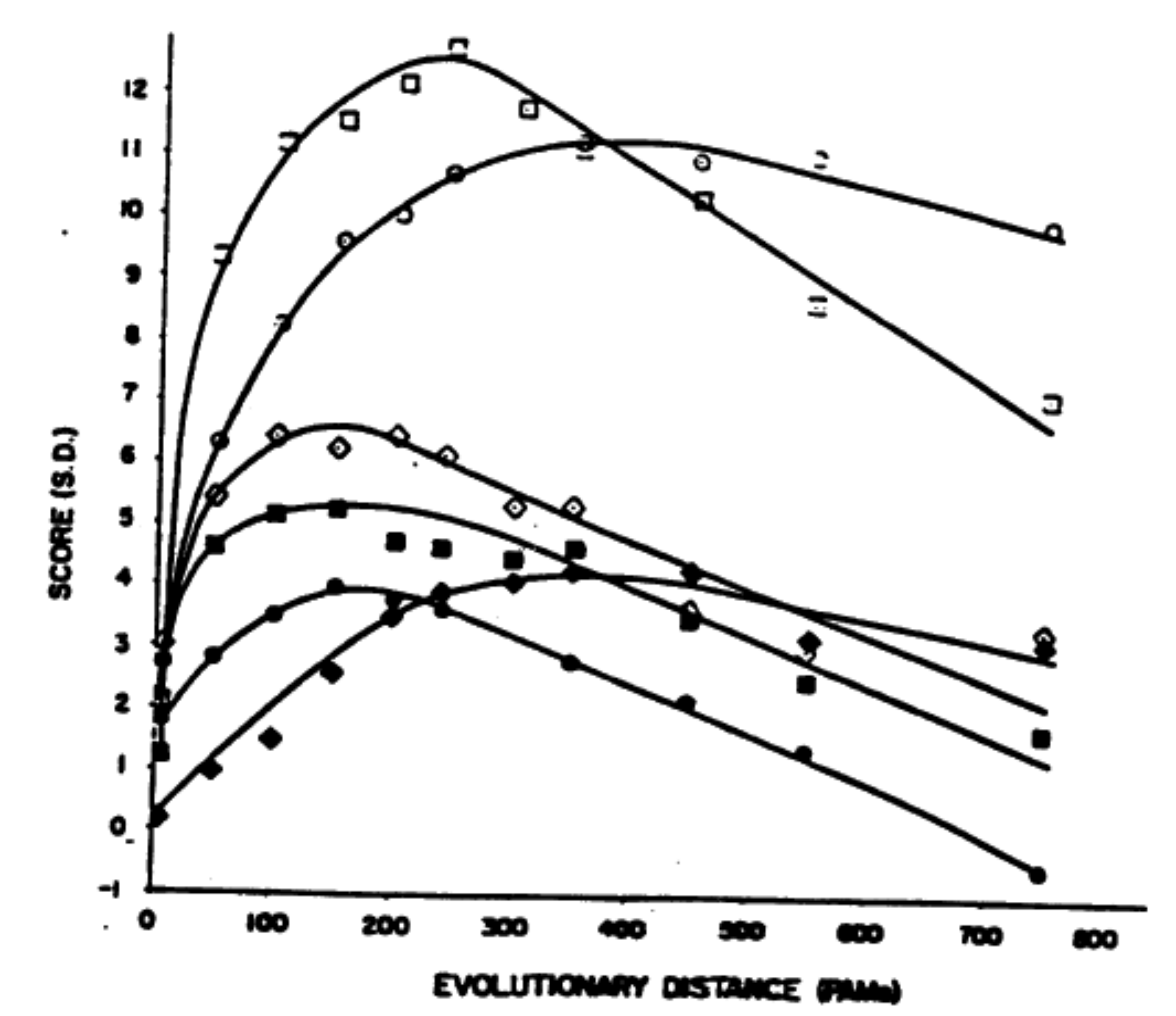
Doug Brutlag 2010

Comparison of Scoring Matrices

Sequences Compared	Unitary Matrix Score (S.D.)	Genetic Code Score (S.D.)	Amino Acid Score (S.D.)	PAM 250 Score (S.D.)
Antibacterial substance A *Streptomyces* vs. Neocarzinostatin *Streptomyces*	3.1	3.2	2.6	2.9
Ferredoxin *Clostridium* vs Ferredoxin *Spirulina*	0.1	1.6	1.8	3.4
a-Hemoglobin Human vs. Myoglobin Human	5.8	6.6	9.9	10.7
a-Hemoglobin Human vs. Globin CTT-III Midge	2.0	2.4	3.2	3.5
Cytochrome C Horse vs. Cytochrome C_6 *Spirulina*	4.5	4.3	7.3	6.1
Cytochrome C Horse vs. Cytochrome C_{553} *Desulfovibrio*	0.2	0.4	0.4	3.9
b2-microglobulin Human vs. IG m chain C4 region Human	3.6	3.3	4.7	4.8

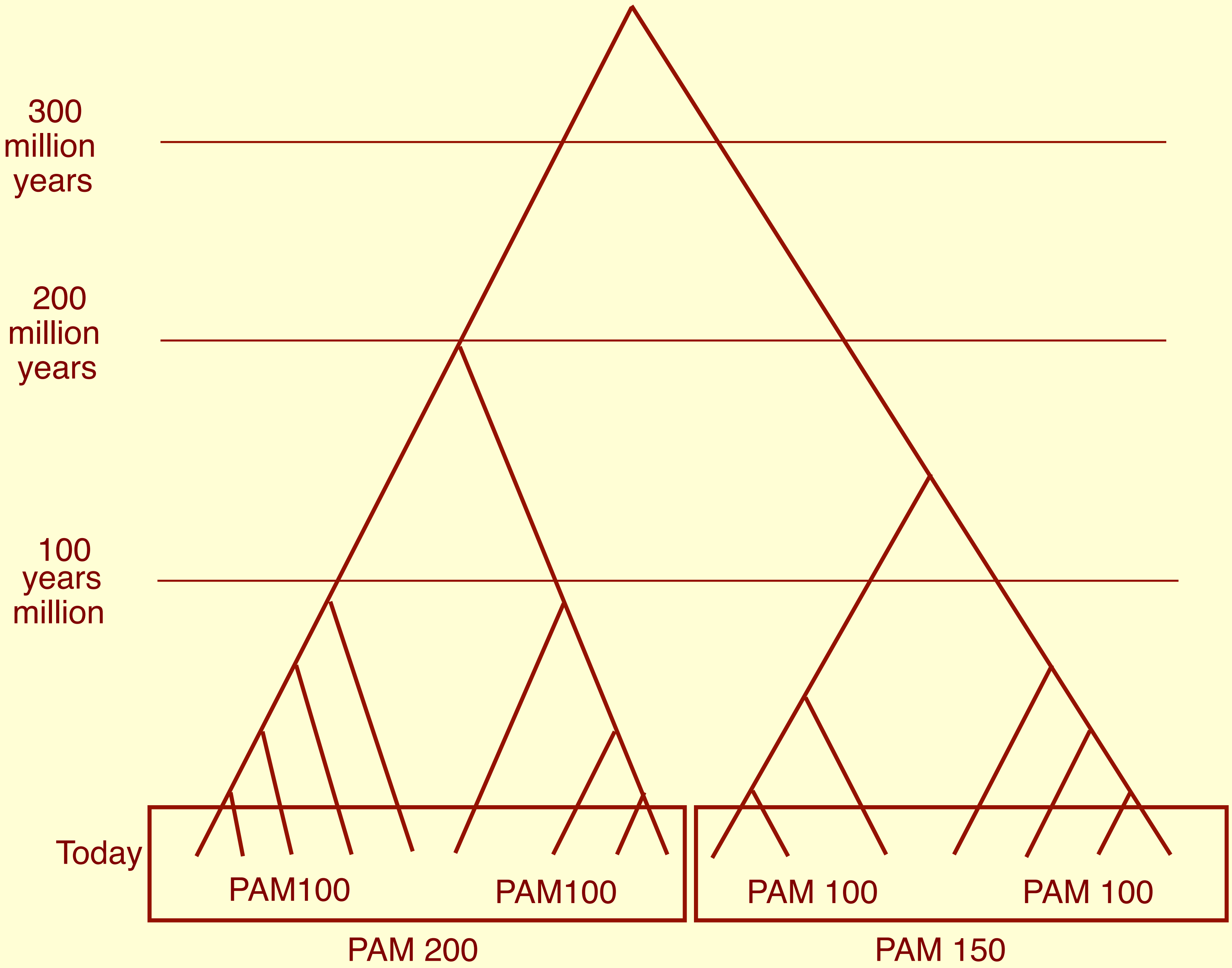
Doug Brutlag 2010

Significance of Alignments vs PAMs



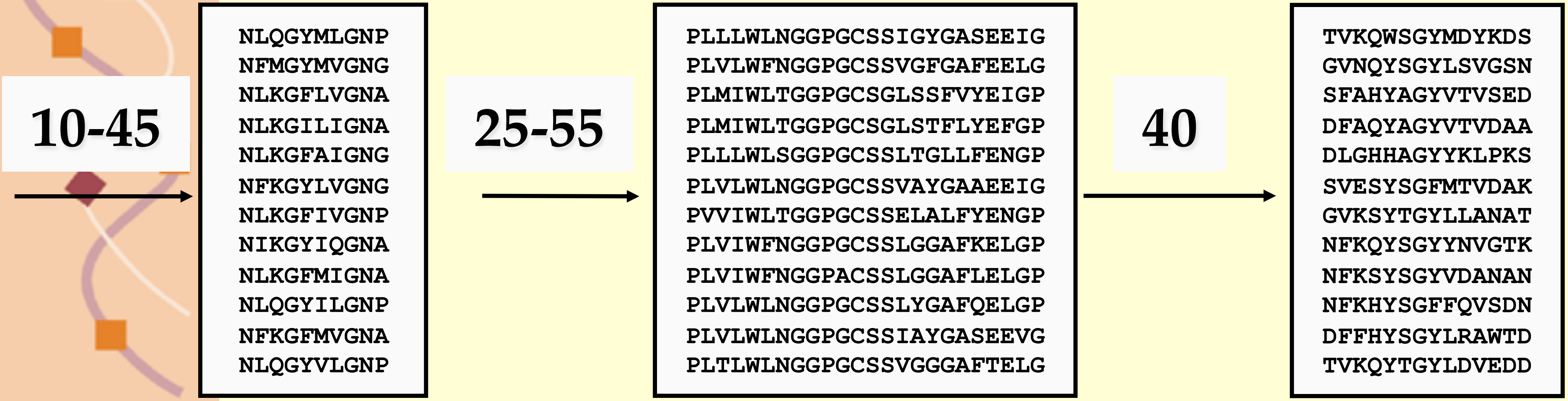
Doug Brutlag 2010

Detecting Evolutionary Relationships



Block Signatures for a Protein Family

http://blocks.fhcrc.org/



10-45

```
NLQGYMLGNP
NFMGYMVGNG
NLKGFLVGNA
NLKGILIGNA
NLKGFAIGNG
NFKGYLVGNG
NLKGFIVGNP
NIKGYIQGNA
NLKGFMIGNA
NLQGYILGNP
NFKGFMVGNA
NLQGYVLGNP
```

25-55

```
PLLLWLNGGPGCSSIGYGASEEIG
PLVLWFNGGPGCSSVGFGAFEELG
PLMIWLTGGPGCSGLSSFVYEIGP
PLMIWLTGGPGCSGLSTFLYEFGP
PLLLWLSGGPGCSSLTGLLFENGP
PLVLWLNGGPGCSSVAYGAAEEIG
PVVIWLTGGPGCSSELALFYENGP
PLVIWFNGGPGCSSLGGAFKELGP
PLVIWFNGGPACSSLGGAFLELGP
PLVLWLNGGPGCSSLYGAFQELGP
PLVLWLNGGPGCSSIAYGASEEVG
PLTLWLNGGPGCSSVGGGAFTELG
```

40

```
TVKQWSGYMDYKDS
GVNQYSGYLSVGSN
SFAHYAGYVTVSED
DFAQYAGYVTVDAA
DLGHHAGYYKLPKS
SVESYSGFMTVDAK
GVKSYTGYLLANAT
NFKQYSGYYNVGTK
NFKSYSGYVDANAN
NFKHYSGFFQVSDN
DFFHYSGYLRAWTD
TVKQYTGYLDVEDD
```

BLOSUM Matrices for Sequence Similarity

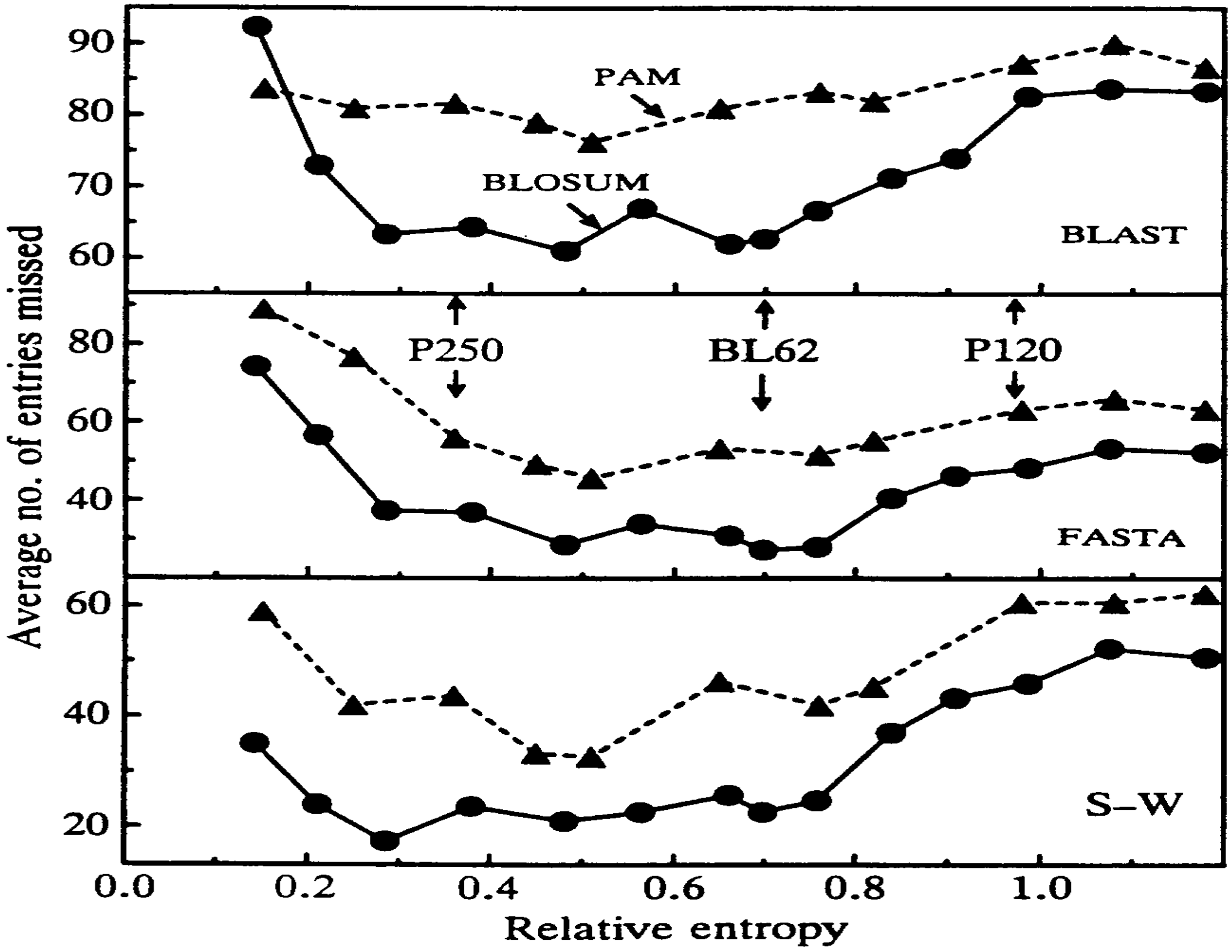
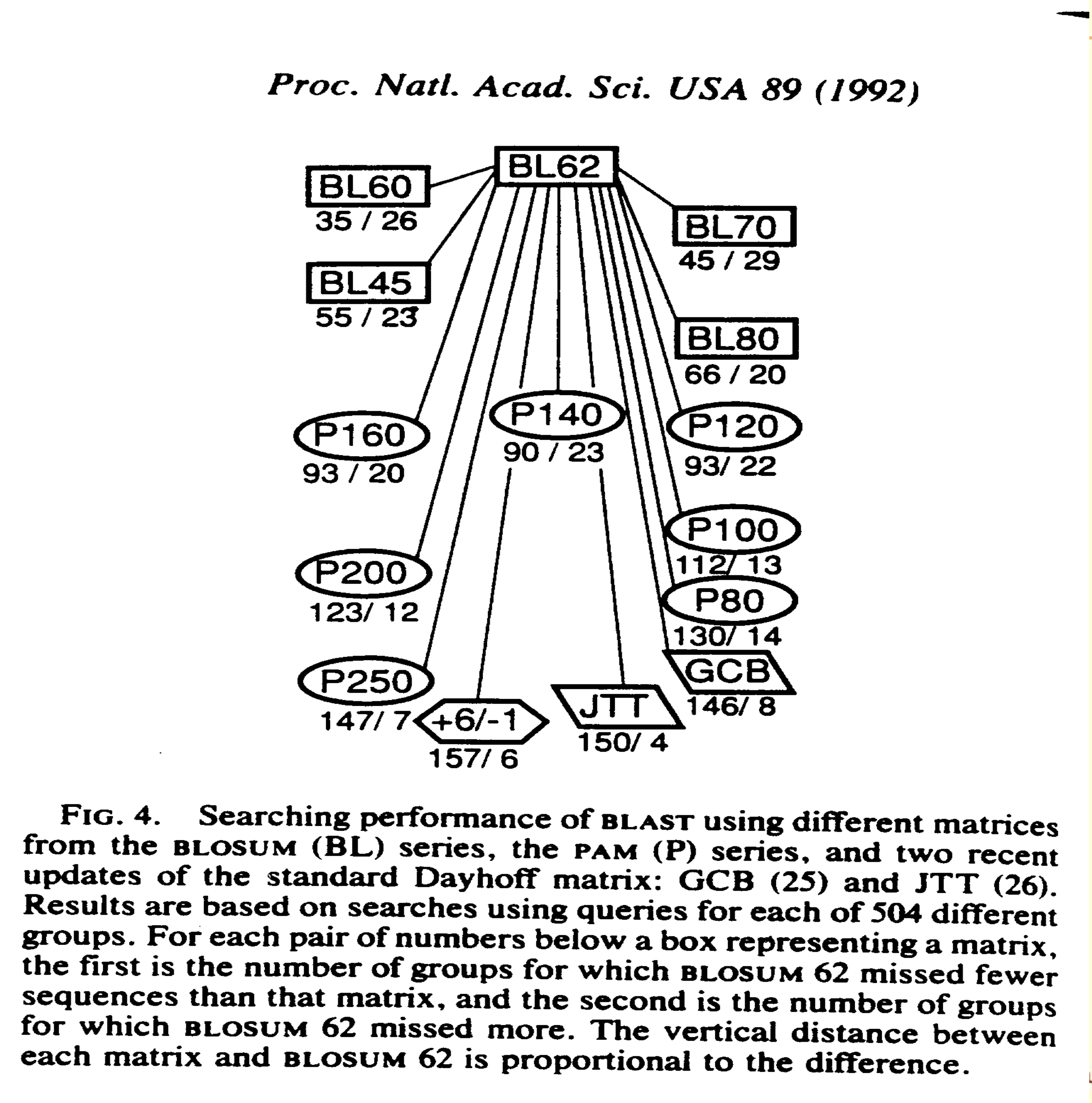


FIG. 3. Searching performance of programs using members of the guanine nucleotide-binding protein-coupled receptor family as queries and matrices from the BLOSUM and PAM series scaled in half-bits (11). Removal of this family from the BLOCKS data base led to a nearly identical matrix with similar performance. Matrices represented (left to right) are BLOSUM (BL) 30, 35, 40, 45, 50, 55, 60, 62, 65, 70, 75, 80, 85, and 90 and PAM (P) 400, 310, 250, 220, 200, 160, 150, 140, 120, 110, and 100. The average numbers of true positive Swiss-Prot entries missed are shown for LSHRRAT, RTARAT, and UL33$HCMVA versus Swiss-Prot 20. Results using BLAST and FASTA or SSEARCH (S–W) are not comparable to each other, since different detection criteria were used for the three programs.

Sequences Missed Using Various Scoring Matrices

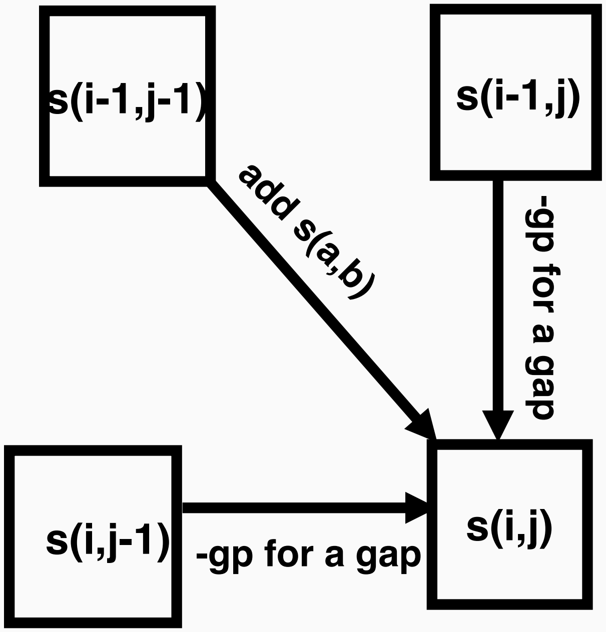


Proc. Natl. Acad. Sci. USA 89 (1992)

FIG. 4. Searching performance of BLAST using different matrices from the BLOSUM (BL) series, the PAM (P) series, and two recent updates of the standard Dayhoff matrix: GCB (25) and JTT (26). Results are based on searches using queries for each of 504 different groups. For each pair of numbers below a box representing a matrix, the first is the number of groups for which BLOSUM 62 missed fewer sequences than that matrix, and the second is the number of groups for which BLOSUM 62 missed more. The vertical distance between each matrix and BLOSUM 62 is proportional to the difference.

Smith-Waterman Algorithm

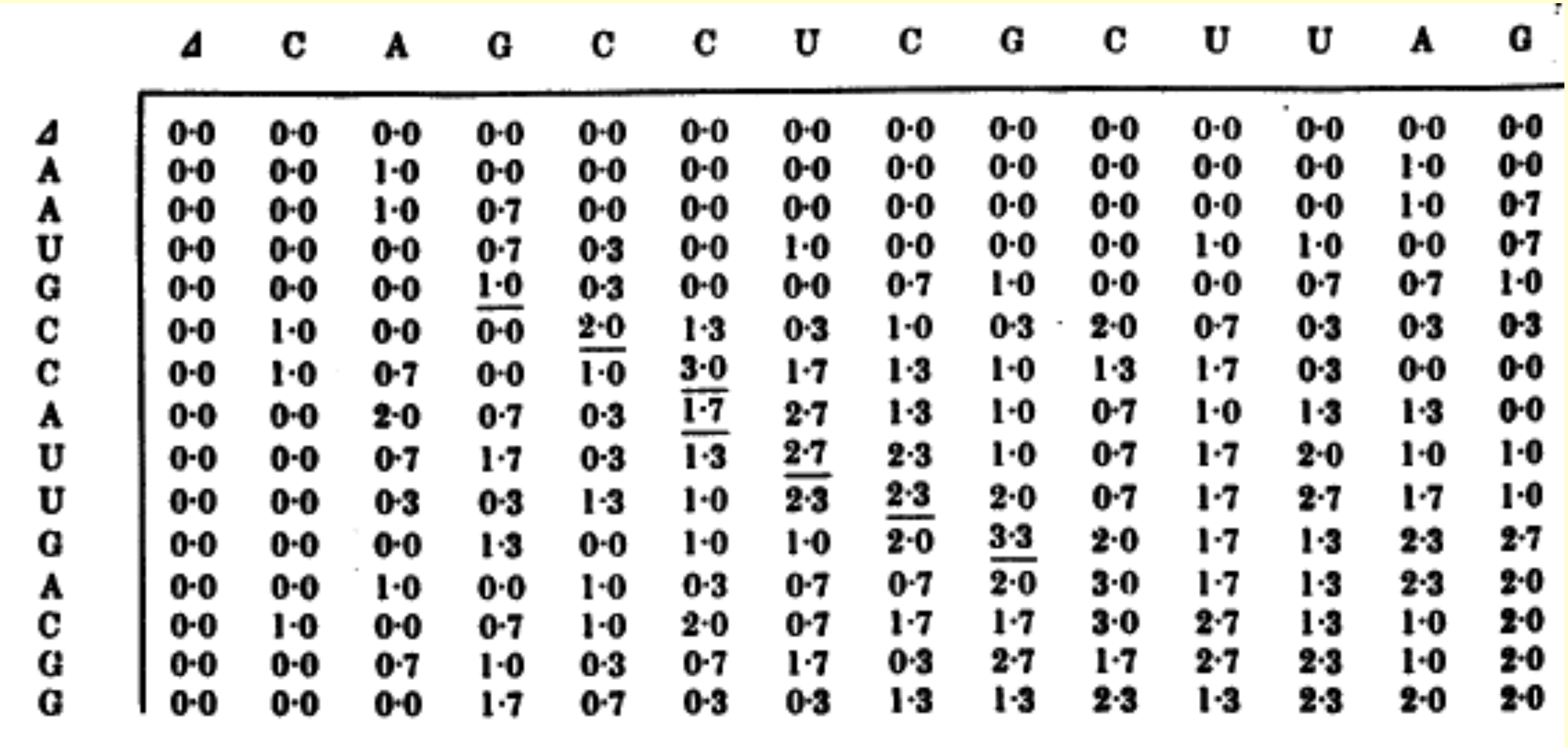
		T	C	A	T	G
	0	0	0	0	0	0
C	0	0	1	0	0	0
A	0	0	0	2	1	0
T	0	1	0	1	3	2
T	0	1	1	0	2	3
G	0	0	1	1	1	3



The score at s(i,j) is the maximum of:

s(i-1,j-1) + s(a,b)
s(i,j-1) - gap penalty
s(i-1,j) - gap penalty
Zero

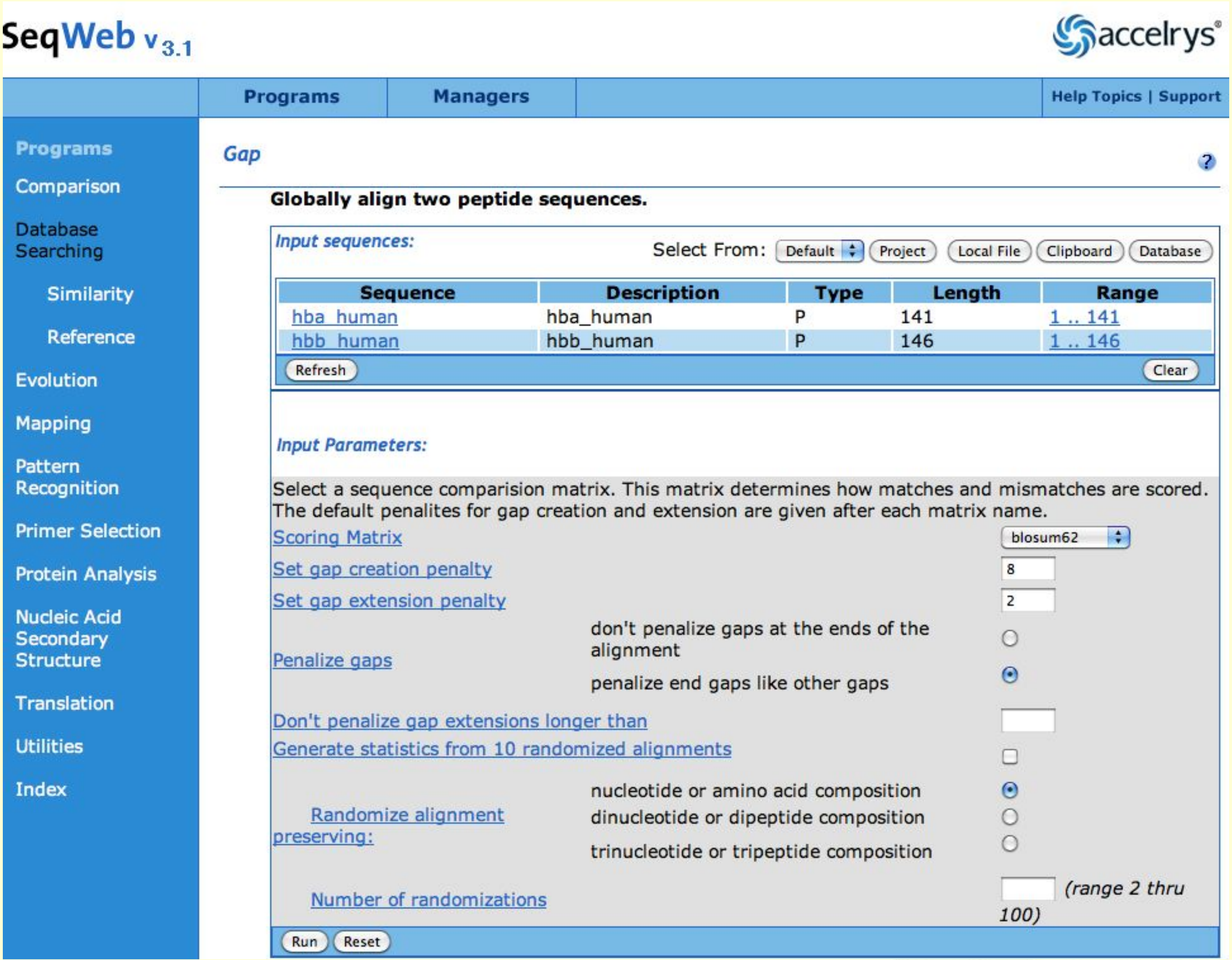
Smith Waterman Score Matrix
(matches=1; mismatches=0; gap=-0.3)



	Δ	C	A	G	C	C	U	C	G	C	U	U	A	G
Δ	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
A	0·0	0·0	1·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	1·0	0·0
A	0·0	0·0	1·0	0·7	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0	1·0	0·7
U	0·0	0·0	0·0	0·7	0·3	0·0	1·0	0·0	0·0	0·0	1·0	1·0	0·0	0·7
G	0·0	0·0	0·0	1·0	0·3	0·0	0·0	0·7	1·0	0·0	0·0	0·7	0·7	1·0
C	0·0	1·0	0·0	0·0	2·0	1·3	0·3	1·0	0·3	2·0	0·7	0·3	0·3	0·3
C	0·0	1·0	0·7	0·0	1·0	3·0	1·7	1·3	1·0	1·3	1·7	0·3	0·0	0·0
A	0·0	0·0	2·0	0·7	0·3	1·7	2·7	1·3	1·0	0·7	1·0	1·3	1·3	0·0
U	0·0	0·0	0·7	1·7	0·3	1·3	2·7	2·3	1·0	0·7	1·7	2·0	1·0	1·0
U	0·0	0·0	0·3	0·3	1·3	1·0	2·3	2·3	2·0	0·7	1·7	2·7	1·7	1·0
G	0·0	0·0	0·0	1·3	0·0	1·0	1·0	2·0	3·3	2·0	1·7	1·3	2·3	2·7
A	0·0	0·0	1·0	0·0	1·0	0·3	0·7	0·7	2·0	3·0	1·7	1·3	2·3	2·0
C	0·0	1·0	0·0	0·7	1·0	2·0	0·7	1·7	1·7	3·0	2·7	1·3	1·0	2·0
G	0·0	0·0	0·7	1·0	0·3	0·7	1·7	0·3	2·7	1·7	2·7	2·3	1·0	2·0
G	0·0	0·0	0·0	1·7	0·7	0·3	0·3	1·3	1·3	2·3	1·3	2·3	2·0	2·0

GAP Align Two Sequences

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=gap-prot



Gap

Globally align two peptide sequences.

Input sequences: Select From: Default

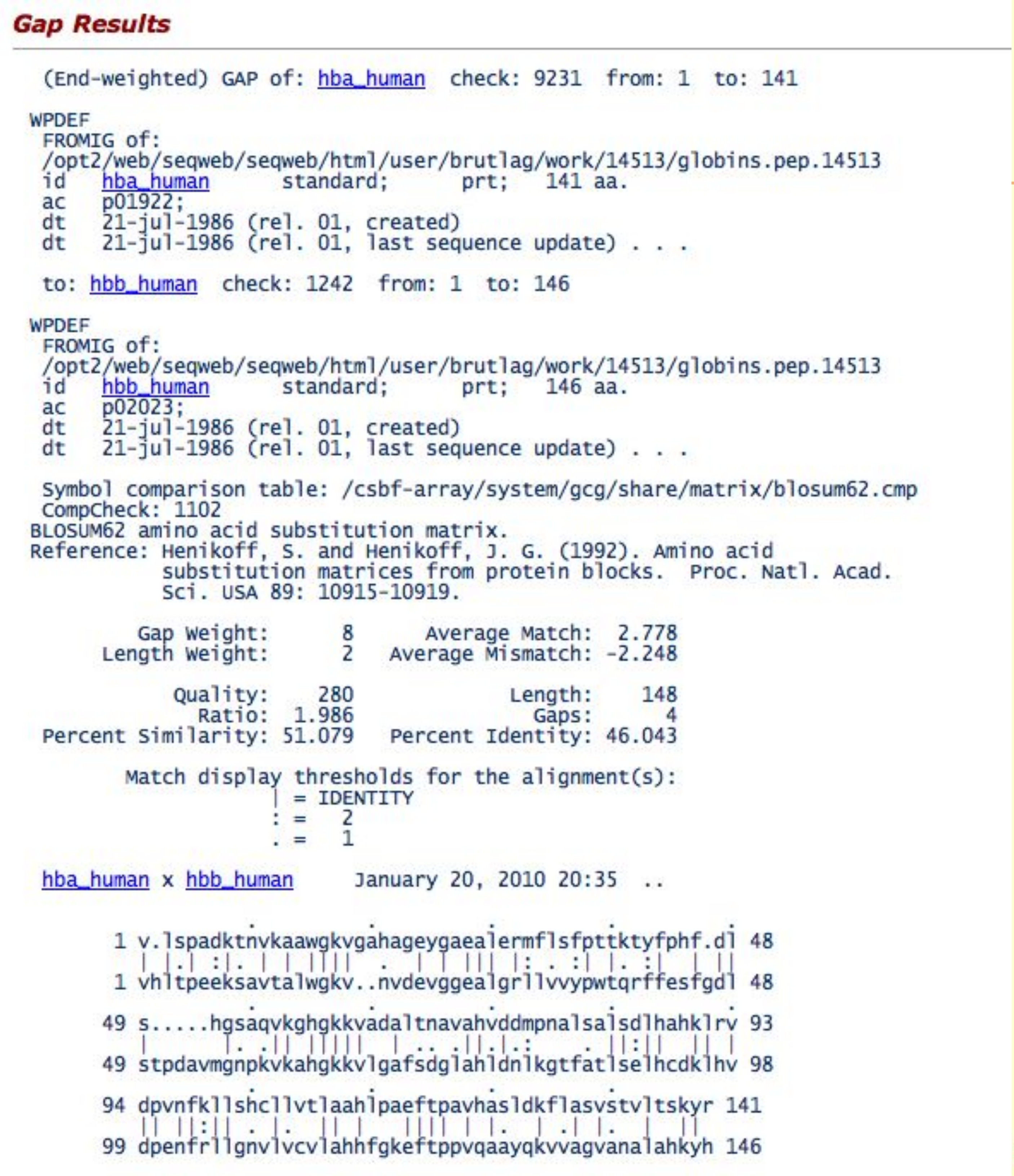
Sequence	Description	Type	Length	Range
hba_human	hba_human	P	141	1 .. 141
hbb_human	hbb_human	P	146	1 .. 146

Input Parameters:

Select a sequence comparision matrix. This matrix determines how matches and mismatches are scored. The default penalites for gap creation and extension are given after each matrix name.

Scoring Matrix		blosum62
Set gap creation penalty		8
Set gap extension penalty		2
Penalize gaps	don't penalize gaps at the ends of the alignment	○
	penalize end gaps like other gaps	●
Don't penalize gap extensions longer than		
Generate statistics from 10 randomized alignments		□
Randomize alignment preserving:	nucleotide or amino acid composition	●
	dinucleotide or dipeptide composition	○
	trinucleotide or tripeptide composition	○
Number of randomizations		*(range 2 thru 100)*

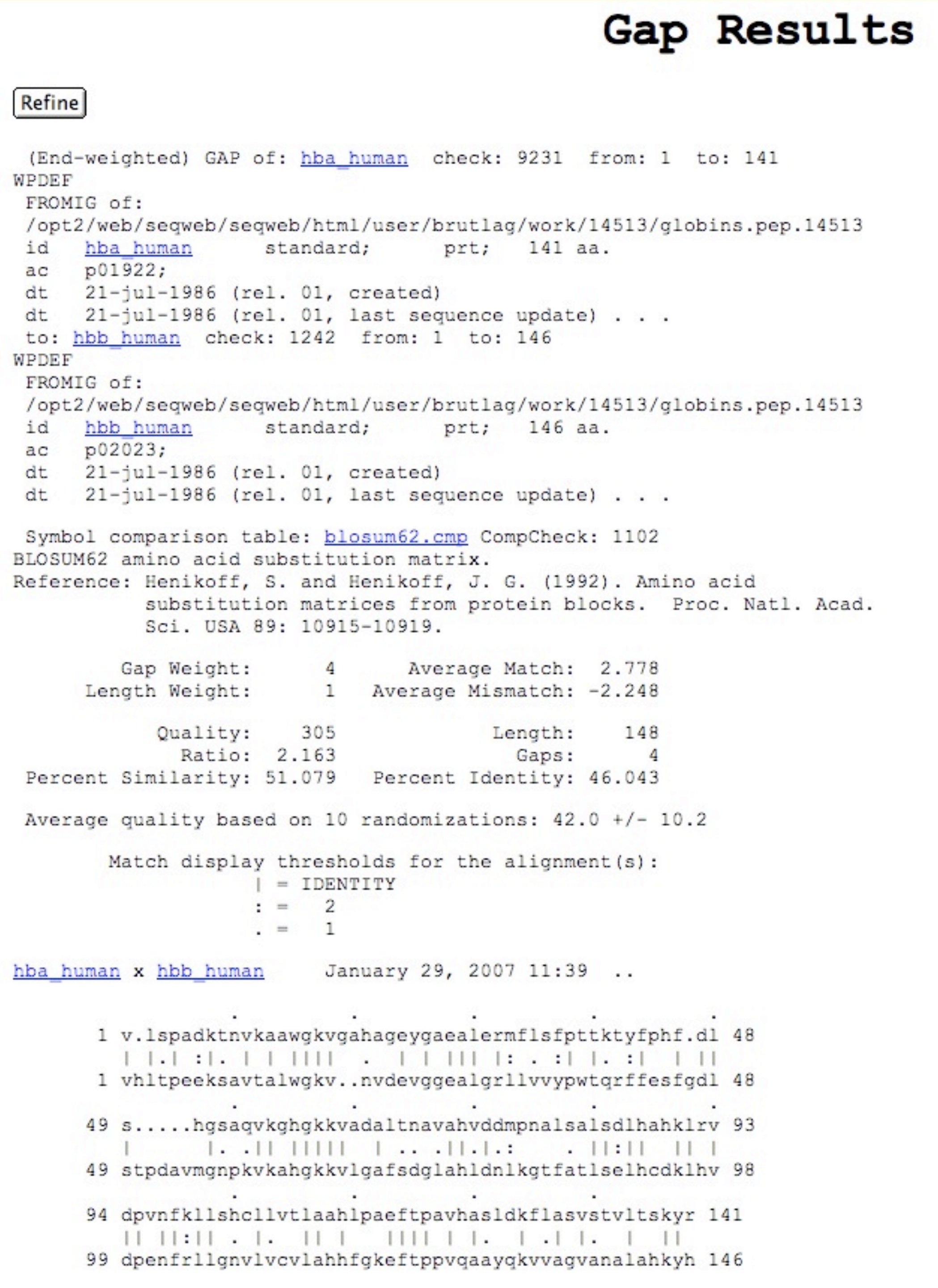
Doug Brutlag 2010



Gap Results

```
 (End-weighted) GAP of: hba_human  check: 9231  from: 1  to: 141

WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hba_human      standard;      prt;   141 aa.
 ac    p01922;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .

 to: hbb_human  check: 1242  from: 1  to: 146

WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hbb_human      standard;      prt;   146 aa.
 ac    p02023;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .

 Symbol comparison table: /csbf-array/system/gcg/share/matrix/blosum62.cmp
 CompCheck: 1102
BLOSUM62 amino acid substitution matrix.
Reference: Henikoff, S. and Henikoff, J. G. (1992). Amino acid
           substitution matrices from protein blocks.  Proc. Natl. Acad.
           Sci. USA 89: 10915-10919.

        Gap Weight:      8      Average Match:  2.778
     Length Weight:      2   Average Mismatch: -2.248

           Quality:    280             Length:    148
             Ratio:  1.986               Gaps:      4
 Percent Similarity: 51.079   Percent Identity: 46.043

        Match display thresholds for the alignment(s):
                     | = IDENTITY
                     : =    2
                     . =    1

 hba_human x hbb_human     January 20, 2010 20:35  ..

                   .         .         .         .         .
        1 v.lspadktnvkaawgkvgahageygaealermflsfpttktyfphf.dl 48
          | |.| :|. | | ||||    .| | ||| |: . :| |. :|  | ||
        1 vhltpeeksavtalwgkv..nvdevggealgrllvvypwtqrffesfgdl 48
                   .         .         .         .         .
       49 s.....hgsaqvkghgkkvadaltnavahvddmpnalsalsdlhahklrv 93
          |      |. .|| |||||  | .. .||.|.:   .  ||:||  || |
       49 stpdavmgnpkvkahgkkvlgafsdglahldnlkgtfatlselhcdklhv 98
                   .         .         .         .
       94 dpvnfkllshcllvtlaahlpaeftpavhasldkflasvstvltskyr 141
          || ||:|| . |.  || |   |||| | |.  | .| |.  |  ||
       99 dpenfrllgnvlvcvlahhfgkeftppvqaayqkvvagvanalahkyh 146
```

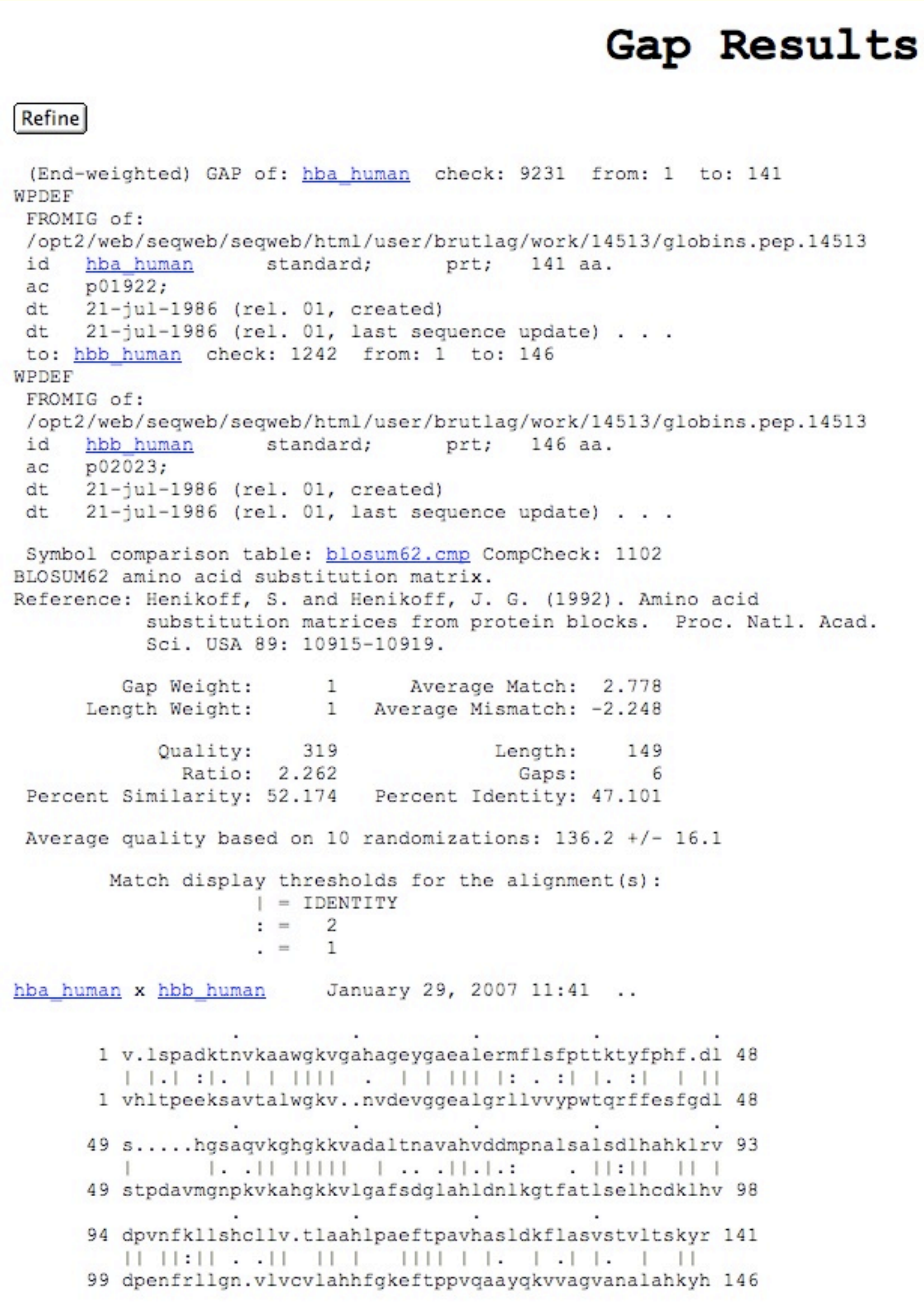
GAP Results (Gap Weight 4)



```
 (End-weighted) GAP of: hba_human  check: 9231  from: 1  to: 141
WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hba_human      standard;      prt;    141 aa.
 ac    p01922;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .
 to: hbb_human  check: 1242  from: 1  to: 146
WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hbb_human      standard;      prt;    146 aa.
 ac    p02023;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .

 Symbol comparison table: blosum62.cmp CompCheck: 1102
BLOSUM62 amino acid substitution matrix.
Reference: Henikoff, S. and Henikoff, J. G. (1992). Amino acid
           substitution matrices from protein blocks.  Proc. Natl. Acad.
           Sci. USA 89: 10915-10919.

        Gap Weight:      4      Average Match:  2.778
     Length Weight:      1   Average Mismatch: -2.248

           Quality:    305             Length:    148
             Ratio:  2.163               Gaps:      4
 Percent Similarity: 51.079   Percent Identity: 46.043

 Average quality based on 10 randomizations: 42.0 +/- 10.2

        Match display thresholds for the alignment(s):
                    | = IDENTITY
                    : =    2
                    . =    1

hba_human x hbb_human     January 29, 2007 11:39  ..

                  .         .         .         .         .
       1 v.lspadktnvkaawgkvgahageygaealermflsfpttktyfphf.dl 48
         | |.| :|. | | ||||  .  | | ||| |: . :| |. :|  | ||
       1 vhltpeeksavtalwgkv..nvdevggealgrllvvypwtqrffesfgdl 48
                  .         .         .         .         .
      49 s.....hgsaqvkghgkkvadaltnavahvddmpnalsalsdlhahklrv 93
         |      |. .|| |||||  | .. .||.|.:    . ||:||  || |
      49 stpdavmgnpkvkahgkkvlgafsdglahldnlkgtfatlselhcdklhv 98
                  .         .         .         .
      94 dpvnfkllshcllvtlaahlpaeftpavhasldkflasvstvltskyr 141
         || ||:|| . |.  || |   |||| | |.  | .| |.  |  ||
      99 dpenfrllgnvlvcvlahhfgkeftppvqaayqkvvagvanalahkyh 146
```

Doug Brutlag 2010

GAP Results (Gap Weight 1)

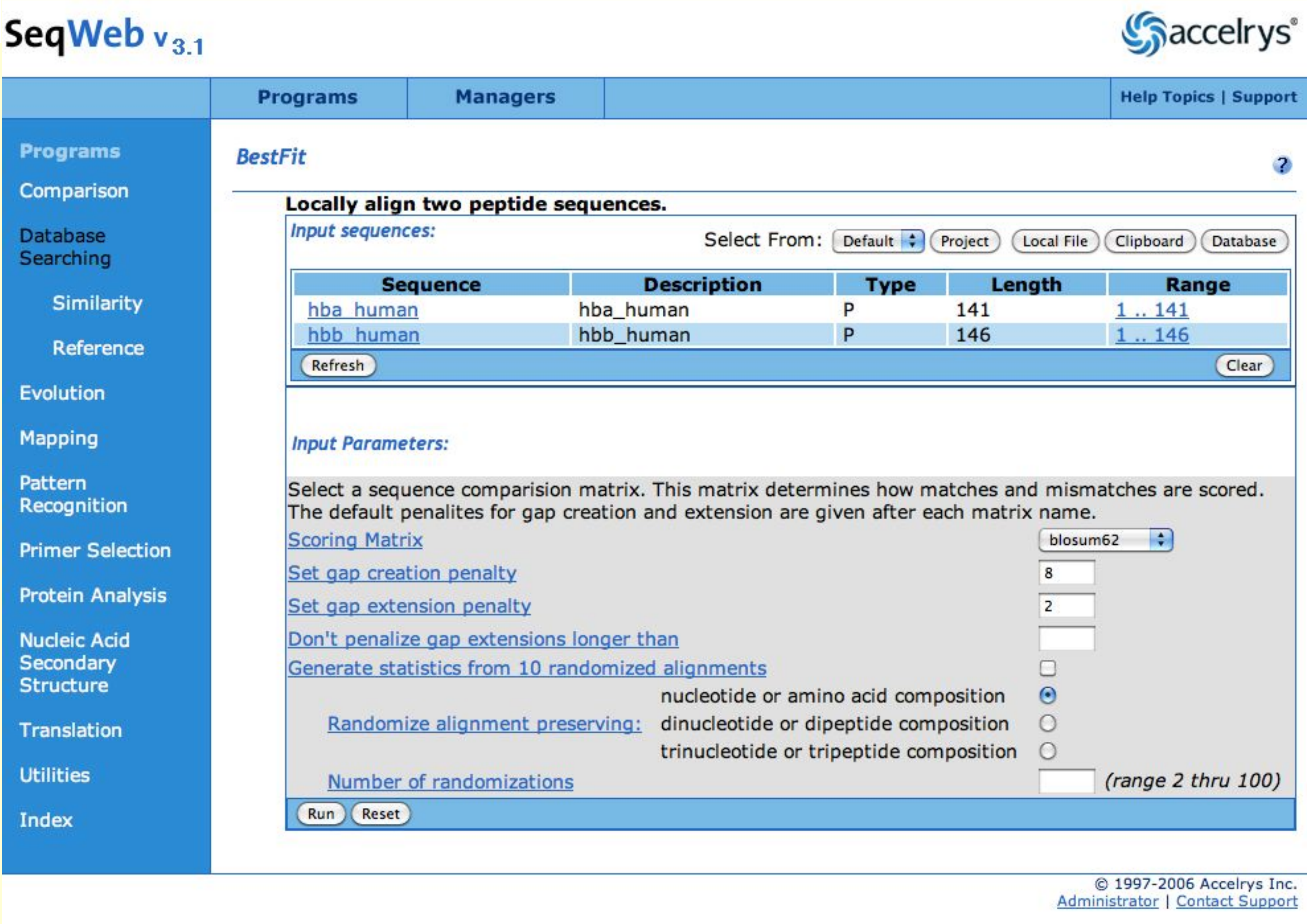


Gap Results

```
 (End-weighted) GAP of: hba_human  check: 9231  from: 1  to: 141
WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hba_human      standard;      prt;   141 aa.
 ac    p01922;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .
 to: hbb_human  check: 1242  from: 1  to: 146
WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hbb_human      standard;      prt;   146 aa.
 ac    p02023;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .

 Symbol comparison table: blosum62.cmp CompCheck: 1102
BLOSUM62 amino acid substitution matrix.
Reference: Henikoff, S. and Henikoff, J. G. (1992). Amino acid
           substitution matrices from protein blocks.  Proc. Natl. Acad.
           Sci. USA 89: 10915-10919.

         Gap Weight:      1      Average Match:  2.778
      Length Weight:      1   Average Mismatch: -2.248

            Quality:    319            Length:    149
              Ratio:  2.262              Gaps:      6
 Percent Similarity: 52.174   Percent Identity: 47.101

 Average quality based on 10 randomizations: 136.2 +/- 16.1

        Match display thresholds for the alignment(s):
                    | = IDENTITY
                    : =    2
                    . =    1

hba_human x hbb_human     January 29, 2007 11:41  ..

                  .         .         .         .         .
        1 v.lspadktnvkaawgkvgahageygaealermflsfpttktyfphf.dl 48
          | |.| :|. | | ||||  .  | | ||| |: . :| |. :|  | ||
        1 vhltpeeksavtalwgkv..nvdevggealgrllvvypwtqrffesfgdl 48
                  .         .         .         .         .
       49 s.....hgsaqvkghgkkvadaltnavahvddmpnalsalsdlhahklrv 93
          |      |. .|| |||||  | .. .||.|.:    . ||:||  || |
       49 stpdavmgnpkvkahgkkvlgafsdglahldnlkgtfatlselhcdklhv 98
                  .         .         .         .
       94 dpvnfkllshcllv.tlaahlpaeftpavhasldkflasvstvltskyr 141
          || ||:|| . .||  || |   |||| | |.  | .| |.  |  ||
       99 dpenfrllgn.vlvcvlahhfgkeftppvqaayqkvvagvanalahkyh 146
```

BestFit Parameters

http://seqweb.stanford.edu:81/gcg-bin/analysis.cgi?program=bestfit-prot



BestFit

Locally align two peptide sequences.

Input sequences: Select From:

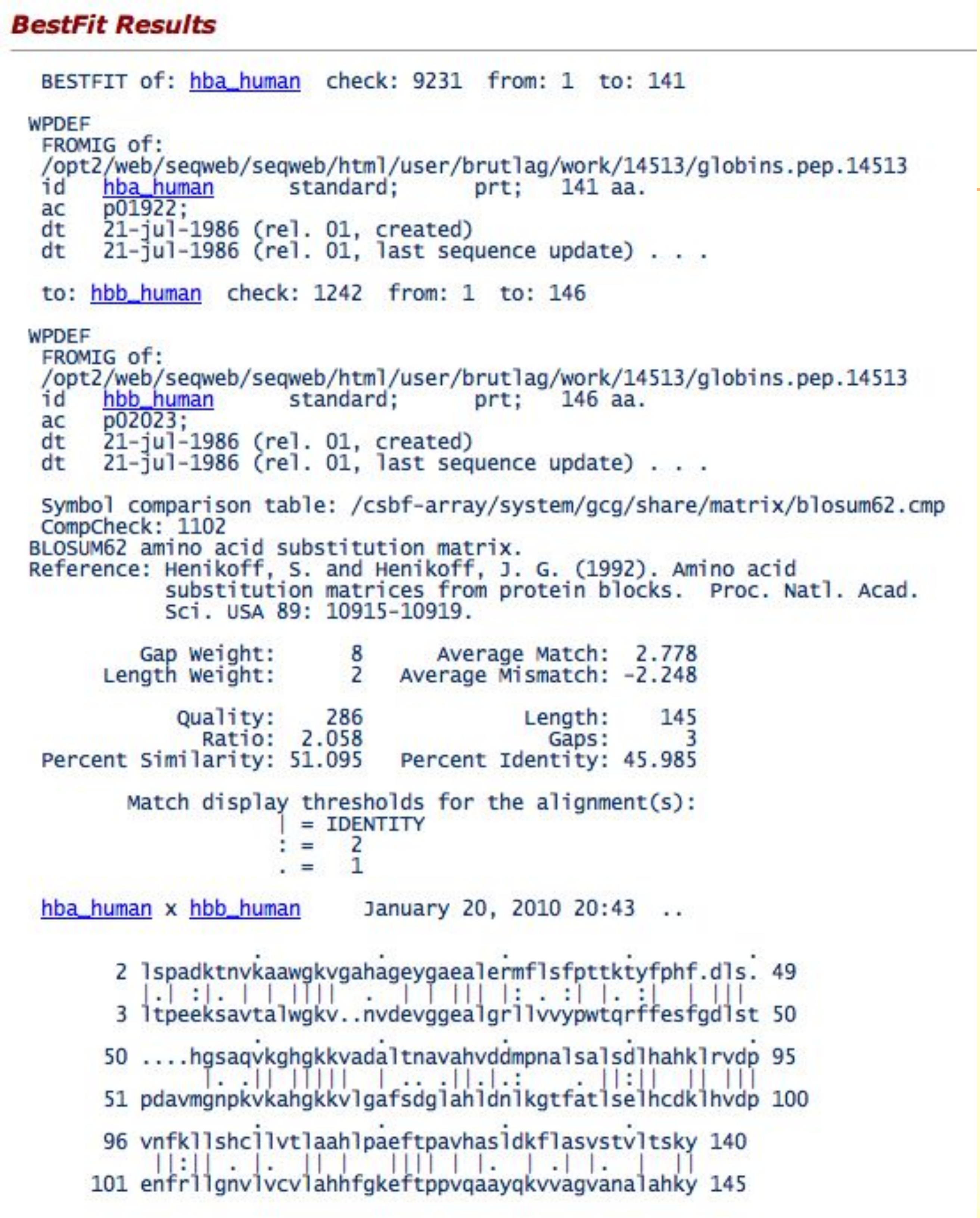
Sequence	Description	Type	Length	Range
hba_human	hba_human	P	141	1 .. 141
hbb_human	hbb_human	P	146	1 .. 146

Input Parameters:

Select a sequence comparision matrix. This matrix determines how matches and mismatches are scored. The default penalites for gap creation and extension are given after each matrix name.

Scoring Matrix	blosum62
Set gap creation penalty	8
Set gap extension penalty	2
Don't penalize gap extensions longer than	
Generate statistics from 10 randomized alignments	□
Randomize alignment preserving: nucleotide or amino acid composition	●
dinucleotide or dipeptide composition	○
trinucleotide or tripeptide composition	○
Number of randomizations	(range 2 thru 100)

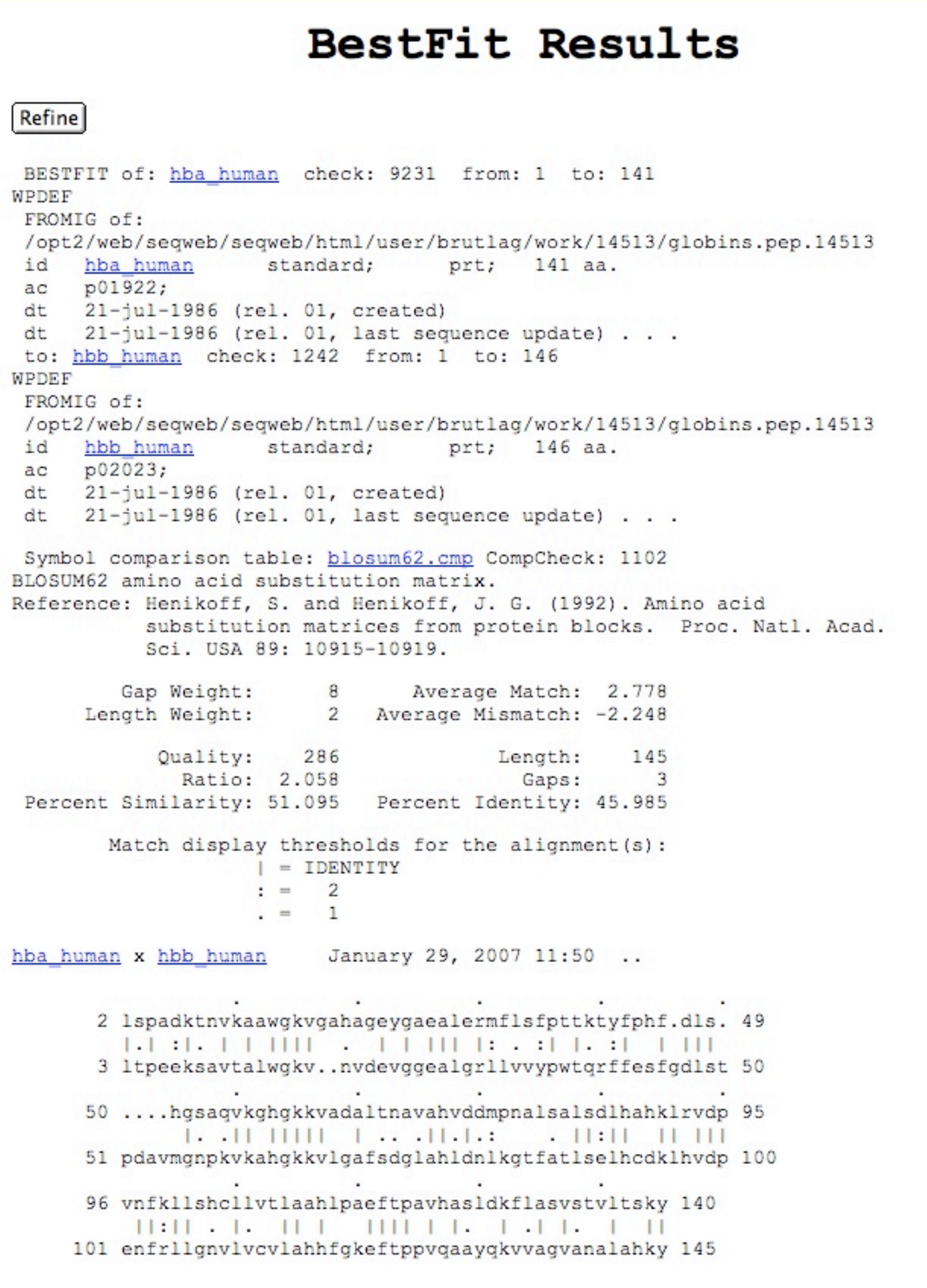
© 1997-2006 Accelrys Inc.
Administrator | Contact Support



BestFit Results

```
 BESTFIT of: hba_human  check: 9231  from: 1  to: 141

WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id   hba_human      standard;      prt;   141 aa.
 ac   p01922;
 dt   21-jul-1986 (rel. 01, created)
 dt   21-jul-1986 (rel. 01, last sequence update) . . .

 to: hbb_human  check: 1242  from: 1  to: 146

WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id   hbb_human      standard;      prt;   146 aa.
 ac   p02023;
 dt   21-jul-1986 (rel. 01, created)
 dt   21-jul-1986 (rel. 01, last sequence update) . . .

 Symbol comparison table: /csbf-array/system/gcg/share/matrix/blosum62.cmp
 CompCheck: 1102
BLOSUM62 amino acid substitution matrix.
Reference: Henikoff, S. and Henikoff, J. G. (1992). Amino acid
           substitution matrices from protein blocks.  Proc. Natl. Acad.
           Sci. USA 89: 10915-10919.

         Gap Weight:      8      Average Match:  2.778
      Length Weight:      2   Average Mismatch: -2.248

            Quality:    286             Length:    145
              Ratio:  2.058               Gaps:      3
 Percent Similarity: 51.095   Percent Identity: 45.985

        Match display thresholds for the alignment(s):
                    | = IDENTITY
                    : =   2
                    . =   1

 hba_human x hbb_human     January 20, 2010 20:43  ..


                  .         .         .         .         .
       2 lspadktnvkaawgkvgahageygaealermflsfpttktyfphf.dls. 49
         |.| :|. | | ||||  .  | | ||| |: . :| |. :|  | |||
       3 ltpeeksavtalwgkv..nvdevggealgrllvvypwtqrffesfgdlst 50

                  .         .         .         .         .
      50 ....hgsaqvkghgkkvadaltnavahvddmpnalsalsdlhahklrvdp 95
              |. .|| |||||  | .. .||.|.:    . ||:||  || |||
      51 pdavmgnpkvkahgkkvlgafsdglahldnlkgtfatlselhcdklhvdp 100

                  .         .         .         .
      96 vnfkllshcllvtlaahlpaeftpavhasldkflasvstvltsky 140
          ||:|| . |.  || |   |||| | |.  | .| |.  |  ||
     101 enfrllgnvlvcvlahhfgkeftppvqaayqkvvagvanalahky 145
```

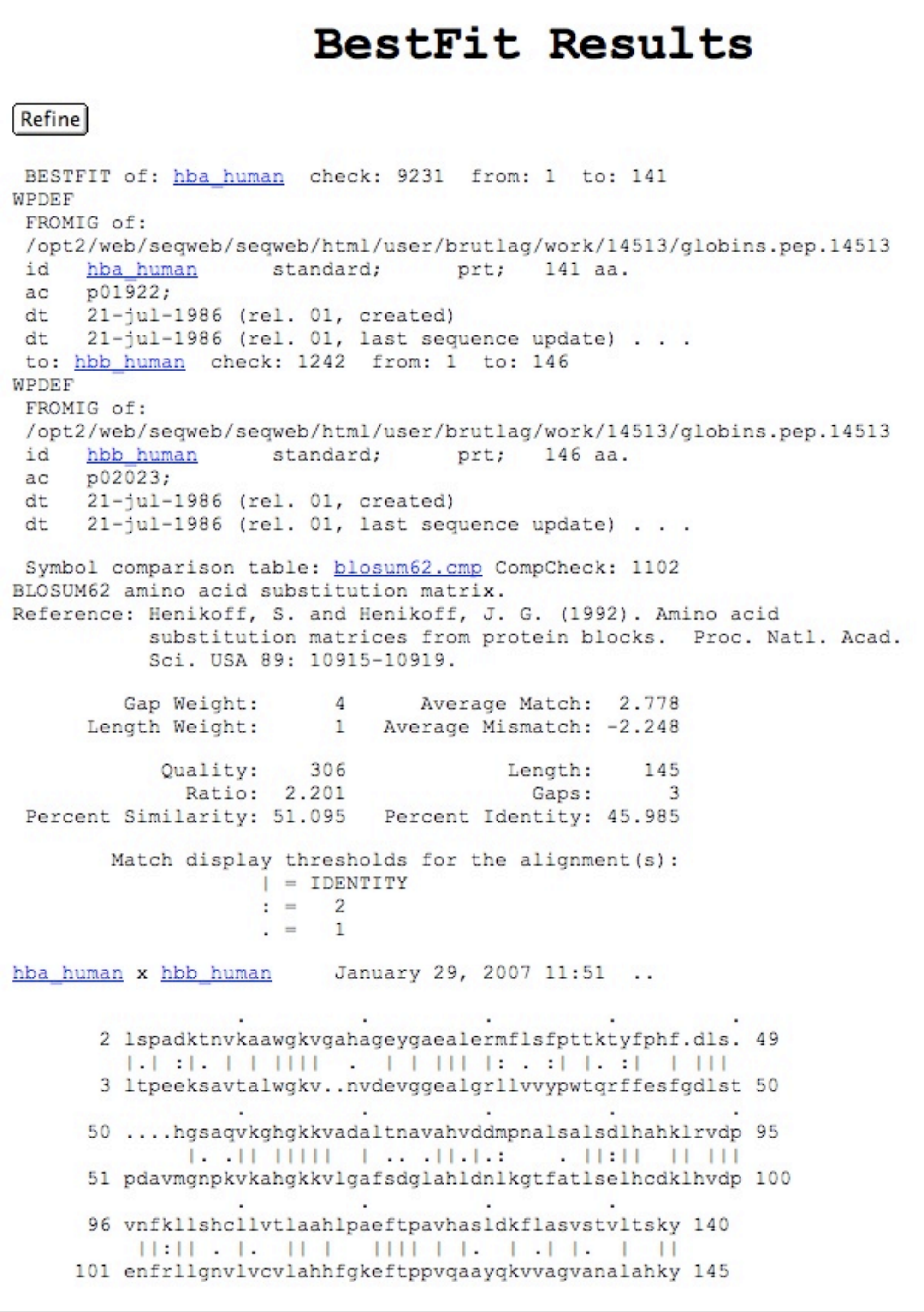
BestFit Results (Gap Weight 8)



BestFit Results

```
 BESTFIT of: hba_human  check: 9231  from: 1  to: 141
WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id   hba_human      standard;      prt;   141 aa.
 ac   p01922;
 dt   21-jul-1986 (rel. 01, created)
 dt   21-jul-1986 (rel. 01, last sequence update) . . .
 to: hbb_human  check: 1242  from: 1  to: 146
WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id   hbb_human      standard;      prt;   146 aa.
 ac   p02023;
 dt   21-jul-1986 (rel. 01, created)
 dt   21-jul-1986 (rel. 01, last sequence update) . . .

 Symbol comparison table: blosum62.cmp CompCheck: 1102
BLOSUM62 amino acid substitution matrix.
Reference: Henikoff, S. and Henikoff, J. G. (1992). Amino acid
           substitution matrices from protein blocks.  Proc. Natl. Acad.
           Sci. USA 89: 10915-10919.

         Gap Weight:      8      Average Match:  2.778
      Length Weight:      2   Average Mismatch: -2.248

            Quality:    286             Length:    145
              Ratio:  2.058               Gaps:      3
 Percent Similarity: 51.095   Percent Identity: 45.985

        Match display thresholds for the alignment(s):
                    | = IDENTITY
                    : =     2
                    . =     1

hba_human x hbb_human     January 29, 2007 11:50  ..

                   .         .         .         .         .
        2 lspadktnvkaawgkvgahageygaealermflsfpttktyfphf.dls. 49
          |.| :|. | | ||||  .  | | ||| |: . :| |. :|  | |||
        3 ltpeeksavtalwgkv..nvdevggealgrllvvypwtqrffesfgdlst 50
                   .         .         .         .         .
       50 ....hgsaqvkghgkkvadaltnavahvddmpnalsalsdlhahklrvdp 95
               |. .|| |||||  | .. .||.|.:    . ||:||  || |||
       51 pdavmgnpkvkahgkkvlgafsdglahldnlkgtfatlselhcdklhvdp 100
                   .         .         .         .
       96 vnfkllshcllvtlaahlpaeftpavhasldkflasvstvltsky 140
           ||:|| . |.  || |   |||| | |.  | .| |.  |  ||
      101 enfrllgnvlvcvlahhfgkeftppvqaayqkvvagvanalahky 145
```

BestFit Results (Gap Weight 4)



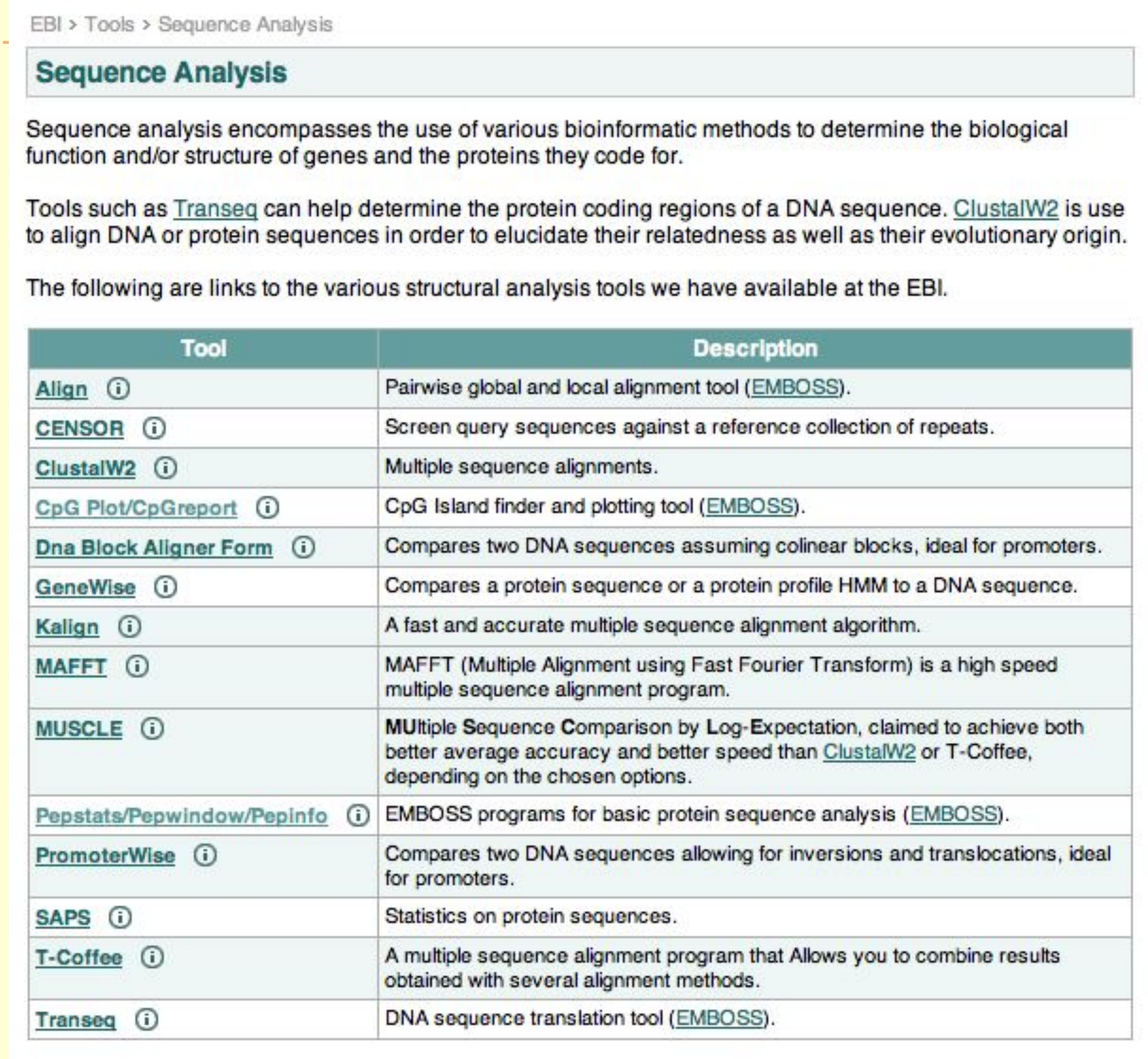
BestFit Results

```
 BESTFIT of: hba_human  check: 9231  from: 1  to: 141
WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hba_human      standard;      prt;   141 aa.
 ac    p01922;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .
 to: hbb_human  check: 1242  from: 1  to: 146
WPDEF
 FROMIG of:
 /opt2/web/seqweb/seqweb/html/user/brutlag/work/14513/globins.pep.14513
 id    hbb_human      standard;      prt;   146 aa.
 ac    p02023;
 dt    21-jul-1986 (rel. 01, created)
 dt    21-jul-1986 (rel. 01, last sequence update) . . .

 Symbol comparison table: blosum62.cmp CompCheck: 1102
BLOSUM62 amino acid substitution matrix.
Reference: Henikoff, S. and Henikoff, J. G. (1992). Amino acid
           substitution matrices from protein blocks.  Proc. Natl. Acad.
           Sci. USA 89: 10915-10919.

         Gap Weight:      4      Average Match:  2.778
      Length Weight:      1   Average Mismatch: -2.248

            Quality:    306             Length:    145
              Ratio:  2.201               Gaps:      3
 Percent Similarity: 51.095   Percent Identity: 45.985

        Match display thresholds for the alignment(s):
                    | = IDENTITY
                    : =    2
                    . =    1

hba_human x hbb_human     January 29, 2007 11:51  ..

                        .         .         .         .         .
          2 lspadktnvkaawgkvgahageygaealermflsfpttktyfphf.dls. 49
            |.| :|. | | ||||  .  | | ||| |: . :| |. :|  | |||
          3 ltpeeksavtalwgkv..nvdevggealgrllvvypwtqrffesfgdlst 50
                        .         .         .         .         .
         50 ....hgsaqvkghgkkvadaltnavahvddmpnalsalsdlhahklrvdp 95
                 |. .|| |||||  | .. .||.|.:    . ||:||  || |||
         51 pdavmgnpkvkahgkkvlgafsdglahldnlkgtfatlselhcdklhvdp 100
                        .         .         .         .
         96 vnfkllshcllvtlaahlpaeftpavhasldkflasvstvltsky 140
             ||:|| . |.  || |   |||| | |.  | .| |.  |  ||
        101 enfrllgnvlvcvlahhfgkeftppvqaayqkvvagvanalahky 145
```

EMBL-EBI

EMBL-EBI Sequence Analysis Tools

http://www.ebi.ac.uk/Tools/sequence.html



EBI > Tools > Sequence Analysis

Sequence Analysis

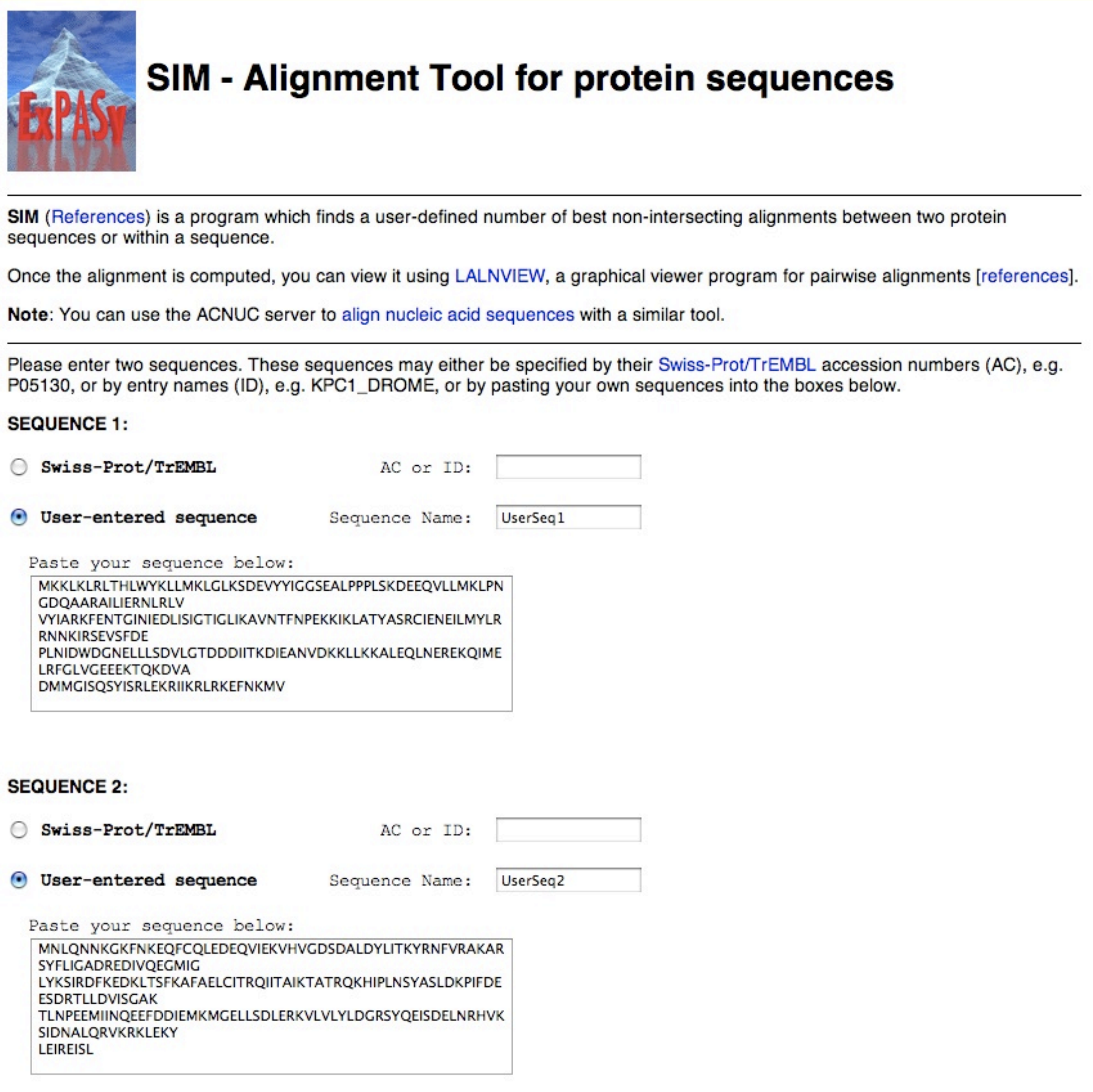
Sequence analysis encompasses the use of various bioinformatic methods to determine the biological function and/or structure of genes and the proteins they code for.

Tools such as Transeq can help determine the protein coding regions of a DNA sequence. ClustalW2 is use to align DNA or protein sequences in order to elucidate their relatedness as well as their evolutionary origin.

The following are links to the various structural analysis tools we have available at the EBI.

Tool	Description
Align ⓘ	Pairwise global and local alignment tool (EMBOSS).
CENSOR ⓘ	Screen query sequences against a reference collection of repeats.
ClustalW2 ⓘ	Multiple sequence alignments.
CpG Plot/CpGreport ⓘ	CpG Island finder and plotting tool (EMBOSS).
Dna Block Aligner Form ⓘ	Compares two DNA sequences assuming colinear blocks, ideal for promoters.
GeneWise ⓘ	Compares a protein sequence or a protein profile HMM to a DNA sequence.
Kalign ⓘ	A fast and accurate multiple sequence alignment algorithm.
MAFFT ⓘ	MAFFT (Multiple Alignment using Fast Fourier Transform) is a high speed multiple sequence alignment program.
MUSCLE ⓘ	**MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation, claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.
Pepstats/Pepwindow/Pepinfo ⓘ	EMBOSS programs for basic protein sequence analysis (EMBOSS).
PromoterWise ⓘ	Compares two DNA sequences allowing for inversions and translocations, ideal for promoters.
SAPS ⓘ	Statistics on protein sequences.
T-Coffee ⓘ	A multiple sequence alignment program that Allows you to combine results obtained with several alignment methods.
Transeq ⓘ	DNA sequence translation tool (EMBOSS).

SIM Alignment Tool



SIM - Alignment Tool for protein sequences

SIM (References) is a program which finds a user-defined number of best non-intersecting alignments between two protein sequences or within a sequence.

Once the alignment is computed, you can view it using LALNVIEW, a graphical viewer program for pairwise alignments [references].

Note: You can use the ACNUC server to align nucleic acid sequences with a similar tool.

Please enter two sequences. These sequences may either be specified by their Swiss-Prot/TrEMBL accession numbers (AC), e.g. P05130, or by entry names (ID), e.g. KPC1_DROME, or by pasting your own sequences into the boxes below.

SEQUENCE 1:

○ **Swiss-Prot/TrEMBL** AC or ID:

● **User-entered sequence** Sequence Name: UserSeq1

Paste your sequence below:

```
MKKLKLRLTHLWYKLLMKLGLKSDEVYYIGGSEALPPPLSKDEEQVLLMKLPN
GDQAARAILIERNLRLV
VYIARKFENTGINIEDLISIGTIGLIKAVNTFNPEKKIKLATYASRCIENEILMYLR
RNNKIRSEVSFDE
PLNIDWDGNELLLSDVLGTDDDIITKDIEANVDKKLLKKALEQLNEREKQIME
LRFGLVGEEEKTQKDVA
DMMGISQSYISRLEKRIIKRLRKEFNKMV
```

SEQUENCE 2:

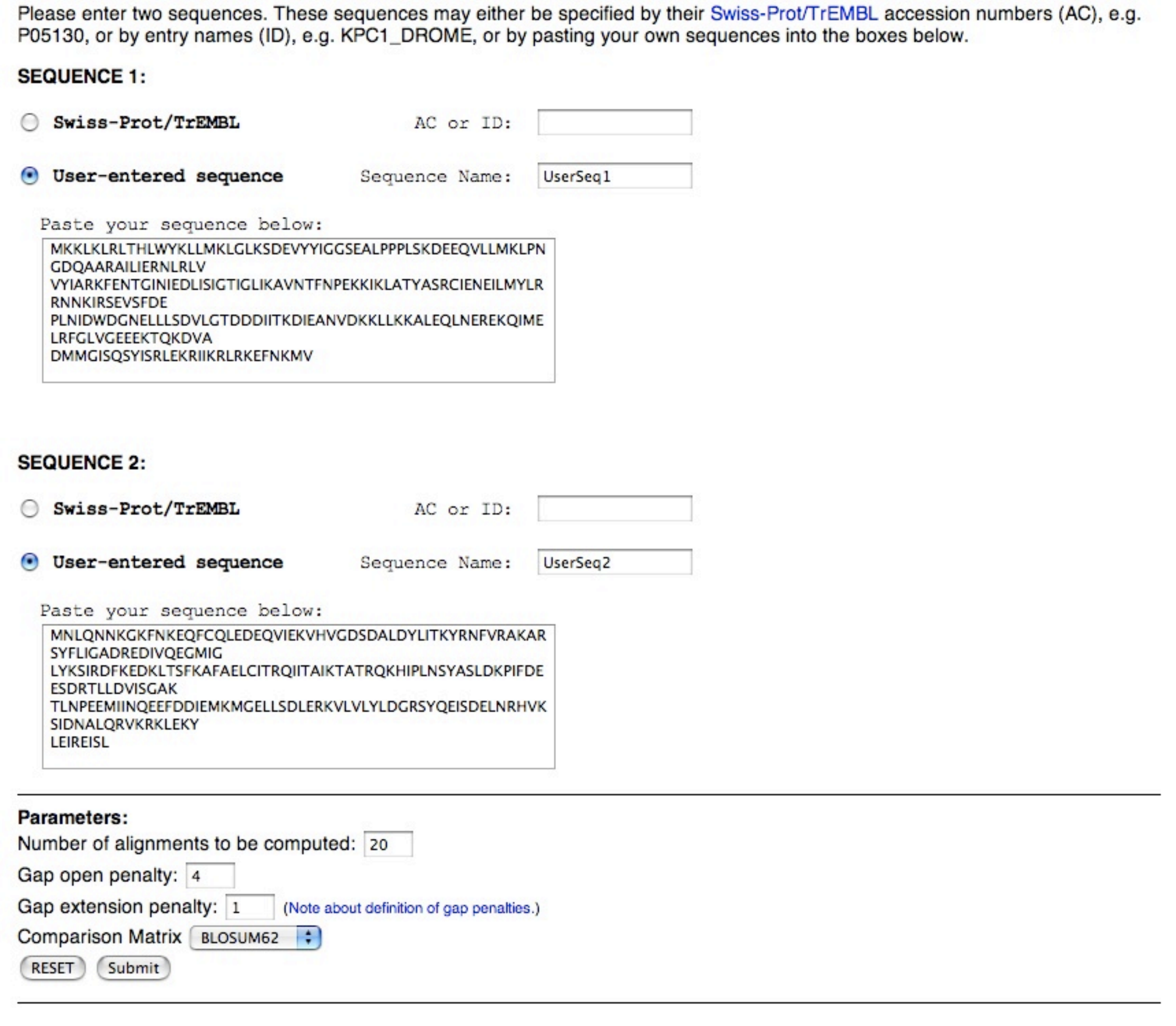
○ **Swiss-Prot/TrEMBL** AC or ID:

● **User-entered sequence** Sequence Name: UserSeq2

Paste your sequence below:

```
MNLQNNKGKFNKEQFCQLEDEQVIEKVHVGDSDALDYLITKYRNFVRAKAR
SYFLIGADREDIVQEGMIG
LYKSIRDFKEDKLTSFKAFAELCITRQIITAIKTATRQKHIPLNSYASLDKPIFDE
ESDRTLLDVISGAK
TLNPEEMIINQEEFDDIEMKMGELLSDLERKVLVLYLDGRSYQEISDELNRHVK
SIDNALQRVKRKLEKY
LEIREISL
```

SIM Input Parameters



Please enter two sequences. These sequences may either be specified by their Swiss-Prot/TrEMBL accession numbers (AC), e.g. P05130, or by entry names (ID), e.g. KPC1_DROME, or by pasting your own sequences into the boxes below.

SEQUENCE 1:

○ **Swiss-Prot/TrEMBL** AC or ID:

● **User-entered sequence** Sequence Name: UserSeq1

Paste your sequence below:

```
MKKLKLRLTHLWYKLLMKLGLKSDEVYYIGGSEALPPPLSKDEEQVLLMKLPN
GDQAARAILIERNLRLV
VYIARKFENTGINIEDLISIGTIGLIKAVNTFNPEKKIKLATYASRCIENEILMYLR
RNNKIRSEVSFDE
PLNIDWDGNELLLSDVLGTDDDIITKDIEANVDKKLLKKALEQLNEREKQIME
LRFGLVGEEEKTQKDVA
DMMGISQSYISRLEKRIIKRLRKEFNKMV
```

SEQUENCE 2:

○ **Swiss-Prot/TrEMBL** AC or ID:

● **User-entered sequence** Sequence Name: UserSeq2

Paste your sequence below:

```
MNLQNNKGKFNKEQFCQLEDEQVIEKVHVGDSDALDYLITKYRNFVRAKAR
SYFLIGADREDIVQEGMIG
LYKSIRDFKEDKLTSFKAFAELCITRQIITAIKTATRQKHIPLNSYASLDKPIFDE
ESDRTLLDVISGAK
TLNPEEMIINQEEFDDIEMKMGELLSDLERKVLVLYLDGRSYQEISDELNRHVK
SIDNALQRVKRKLEKY
LEIREISL
```

Parameters:

Number of alignments to be computed: 20

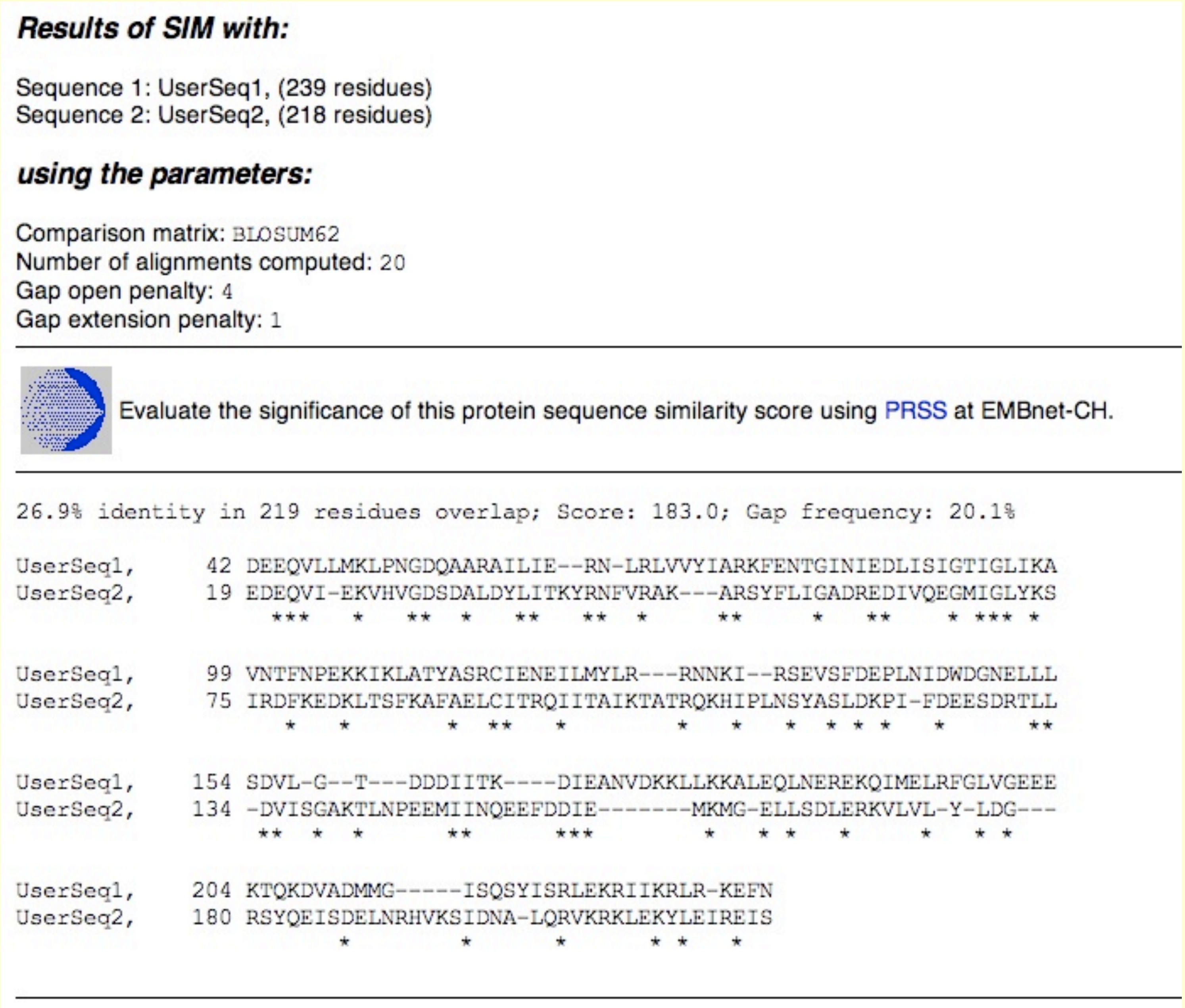
Gap open penalty: 4

Gap extension penalty: 1 (Note about definition of gap penalties.)

Comparison Matrix: BLOSUM62

Doug Brutlag 2010

SIM Results (1)



Results of SIM with:

Sequence 1: UserSeq1, (239 residues)
Sequence 2: UserSeq2, (218 residues)

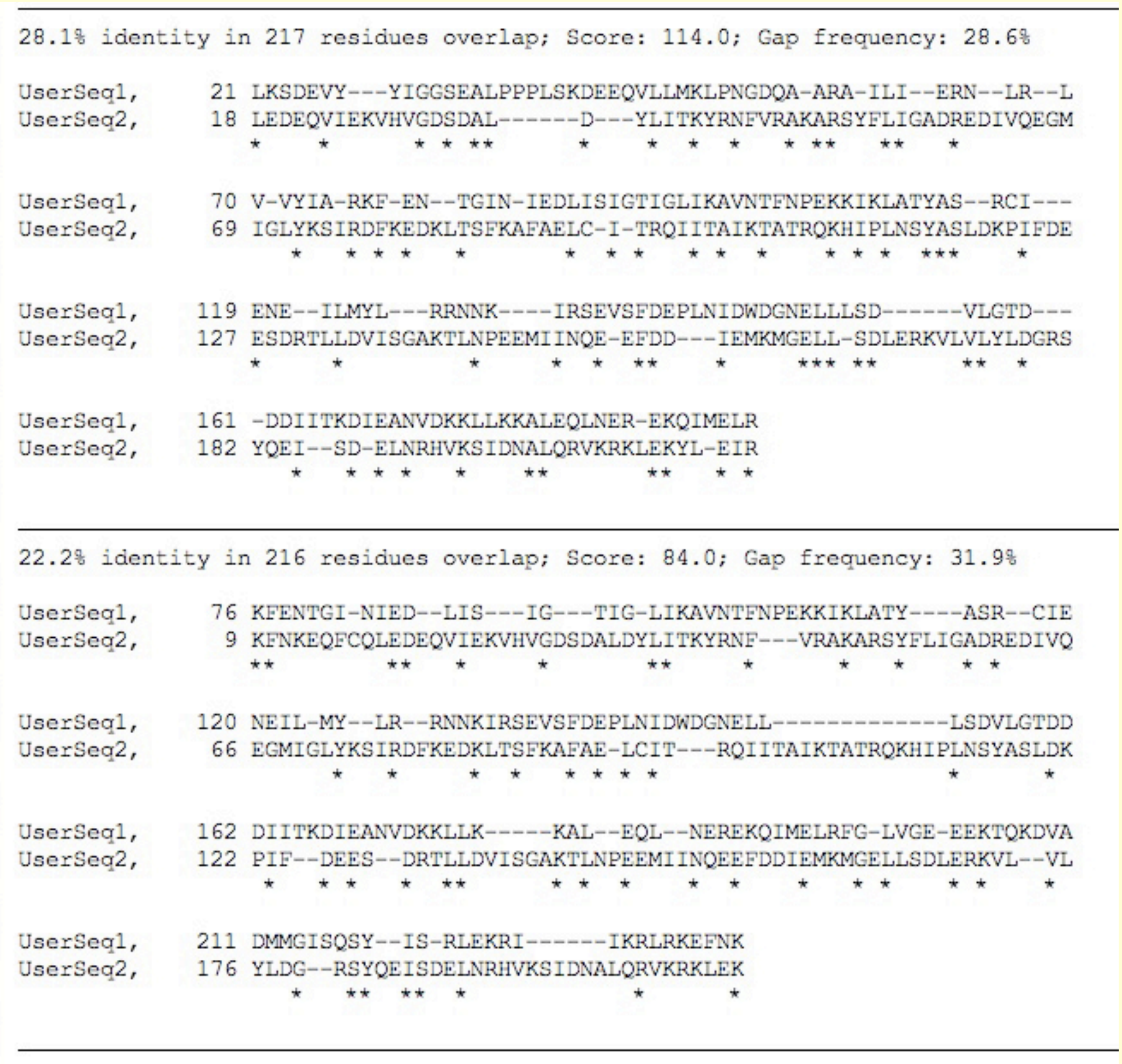
using the parameters:

Comparison matrix: BLOSUM62
Number of alignments computed: 20
Gap open penalty: 4
Gap extension penalty: 1

Evaluate the significance of this protein sequence similarity score using PRSS at EMBnet-CH.

```
26.9% identity in 219 residues overlap; Score: 183.0; Gap frequency: 20.1%

UserSeq1,      42 DEEQVLLMKLPNGDQAARAILIE--RN-LRLVVYIARKFENTGINIEDLISIGTIGLIKA
UserSeq2,      19 EDEQVI-EKVHVGDSDALDYLITKYRNFVRAK---ARSYFLIGADREDIVQEGMIGLYKS
                    ***   *   **  *   **   **  *     **     *   **    * *** *

UserSeq1,      99 VNTFNPEKKIKLATYASRCIENEILMYLR---RNNKI--RSEVSFDEPLNIDWDGNELLL
UserSeq2,      75 IRDFKEDKLTSFKAFAELCITRQIITAIKTATRQKHIPLNSYASLDKPI-FDEESDRTLL
                     *   *       *  **   *        *   *   *  * * *   *      **

UserSeq1,     154 SDVL-G--T---DDDIITK----DIEANVDKKLLKKALEQLNEREKQIMELRFGLVGEEE
UserSeq2,     134 -DVISGAKTLNPEEMIINQEEFDDIE-------MKMG-ELLSDLERKVLVL-Y-LDG---
                   **  *  *      **      ***        *   * *   *     *   * *

UserSeq1,     204 KTQKDVADMMG-----ISQSYISRLEKRIIKRLR-KEFN
UserSeq2,     180 RSYQEISDELNRHVKSIDNA-LQRVKRKLEKYLEIREIS
                         *        *      *      * *   *
```

SIM Results (2)



```
28.1% identity in 217 residues overlap; Score: 114.0; Gap frequency: 28.6%

UserSeq1,     21 LKSDEVY---YIGGSEALPPPLSKDEEQVLLMKLPNGDQA-ARA-ILI--ERN--LR--L
UserSeq2,     18 LEDEQVIEKVHVGDSDAL------D---YLITKYRNFVRAKARSYFLIGADREDIVQEGM
                 *    *      * * **      *    *  *  *   * **   **   *

UserSeq1,     70 V-VYIA-RKF-EN--TGIN-IEDLISIGTIGLIKAVNTFNPEKKIKLATYAS--RCI---
UserSeq2,     69 IGLYKSIRDFKEDKLTSFKAFAELC-I-TRQIITAIKTATRQKHIPLNSYASLDKPIFDE
                    *   * * *   *       *  * *   * *  *    * * *  ***    *

UserSeq1,    119 ENE--ILMYL---RRNNK----IRSEVSFDEPLNIDWDGNELLLSD------VLGTD---
UserSeq2,    127 ESDRTLLDVISGAKTLNPEEMIINQE-EFDD---IEMKMGELL-SDLERKVLVLYLDGRS
                 *     *         *     *  *  **    *     *** **      **  *

UserSeq1,    161 -DDIITKDIEANVDKKLLKKALEQLNER-EKQIMELR
UserSeq2,    182 YQEI--SD-ELNRHVKSIDNALQRVKRKLEKYL-EIR
                    *   * * *   *    **       **   * *

22.2% identity in 216 residues overlap; Score: 84.0; Gap frequency: 31.9%

UserSeq1,     76 KFENTGI-NIED--LIS---IG---TIG-LIKAVNTFNPEKKIKLATY----ASR--CIE
UserSeq2,      9 KFNKEQFCQLEDEQVIEKVHVGDSDALDYLITKYRNF---VRAKARSYFLIGADREDIVQ
                 **        **   *     *       **     *      *   *    * *

UserSeq1,    120 NEIL-MY--LR--RNNKIRSEVSFDEPLNIDWDGNELL-------------LSDVLGTDD
UserSeq2,     66 EGMIGLYKSIRDFKEDKLTSFKAFAE-LCIT---RQIITAIKTATRQKHIPLNSYASLDK
                       *   *     *  *   * * * *                     *      *

UserSeq1,    162 DIITKDIEANVDKKLLK-----KAL--EQL--NEREKQIMELRFG-LVGE-EEKTQKDVA
UserSeq2,    122 PIF--DEES--DRTLLDVISGAKTLNPEEMIINQEEFDDIEMKMGELLSDLERKVL--VL
                  *   * *   *  **      * *  *    *  *    *   * *    * *    *

UserSeq1,    211 DMMGISQSY--IS-RLEKRI------IKRLRKEFNK
UserSeq2,    176 YLDG--RSYQEISDELNRHVKSIDNALQRVKRKLEK
                    *   **  **  *            *      *
```

Doug Brutlag 2010