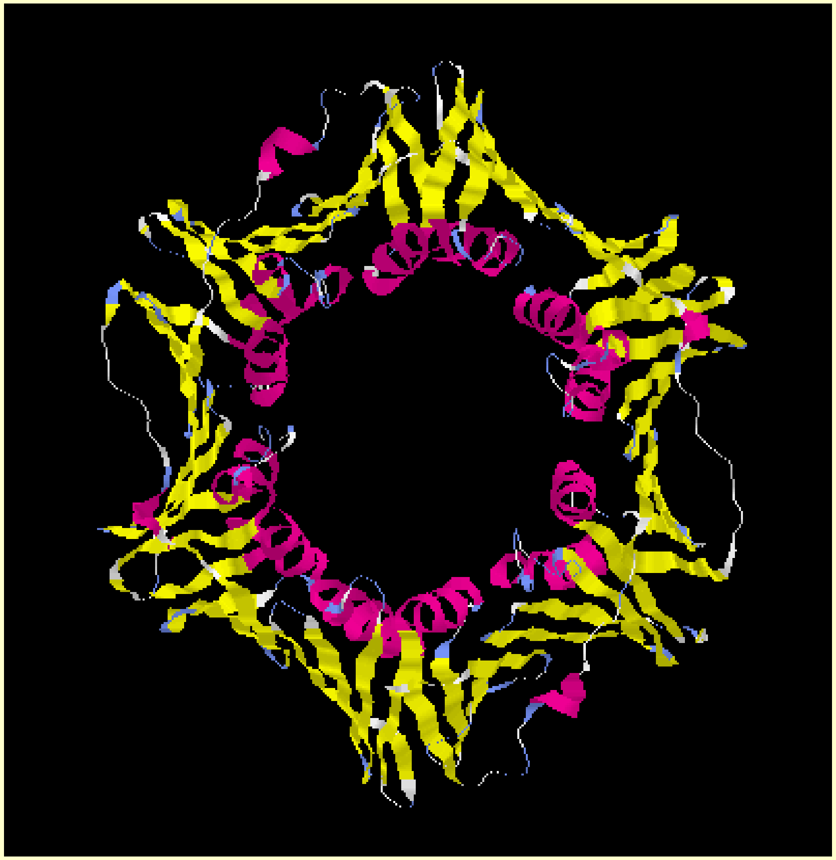
Computational Molecular Biology

Biochem 218 – BioMedical Informatics 231

http://biochem218.stanford.edu/

Rapid Sequence Similarity Search



Doug Brutlag
Professor Emeritus
Biochemistry & Medicine (by courtesy)

Needleman-Wunsch Sequence Alignment

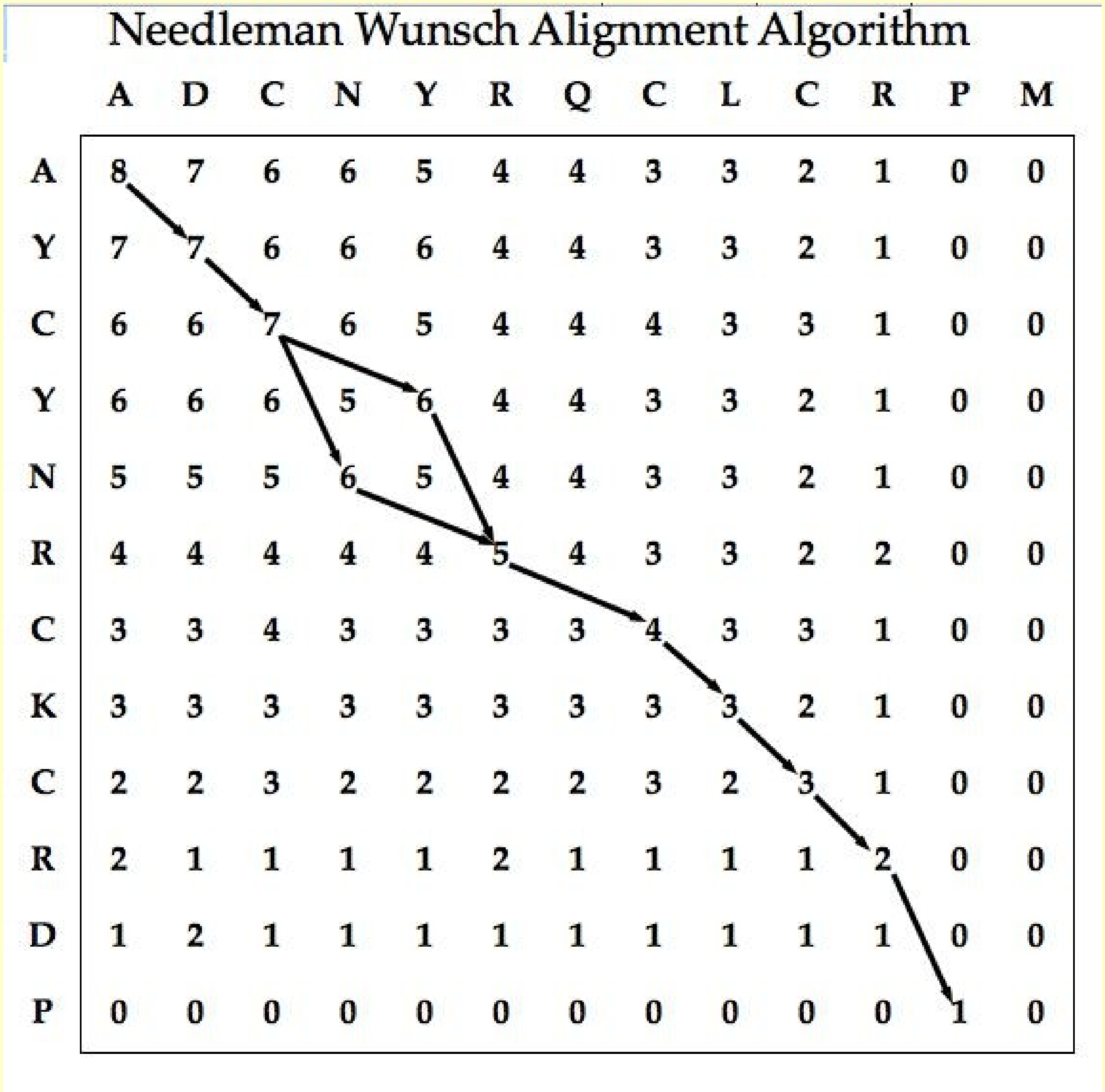
```
  X        220       230       240       250           X
  F--SGGNTHIYMNHVEQCKEILRREPKELCELVISGLPYKFRYLSTKE-QLK-Y
  |  : |::|||:||:|  |  | |||: : :| |     |   ::::: |:: |
GDFIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKAEDFQIEGYN
  X         260       270       280       290          X
```

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

where:

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

Needleman-Wunsch Alignment Algorithm
Trace Back



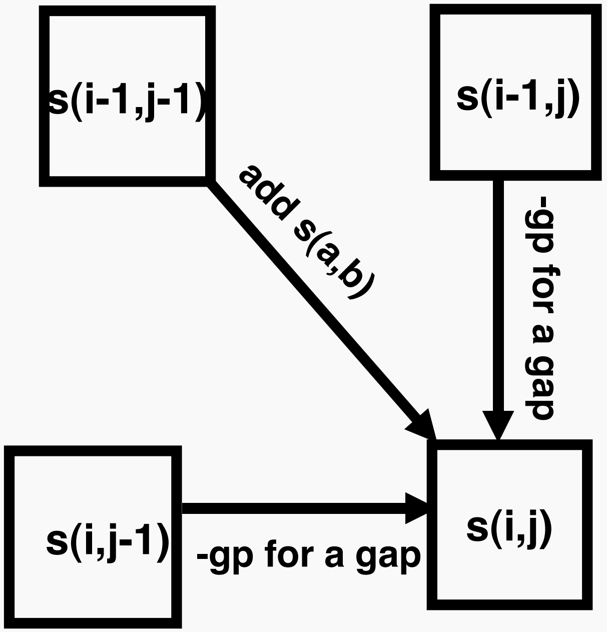
Needleman Wunsch Alignment Algorithm

	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

Doug Brutlag 2010

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

Doug Brutlag 2010

Computer Time and Space Requirements

- Needleman-Wunsch
  - O(N*M) time and O(N*M) space
- Smith-Waterman
  - O(N*M) time and O(N*M) space

Gotoh's Improvement

		Previous Column	Current Column
			$VG(i-2,j)$
Previous Row		$S(i-1,j-2)$	$S(i-1,j)$
Current Row	$HG(i,j-2)$	$S(i,j-1)$	$s(i,j)$

$$S(i,j) = Max \begin{cases} S(i-1,j-1) + s(i,j) \\ S(i-1,j) - GP \\ S(i,j-1) - GP \\ VG(i-2,j) - GEP \\ HG(i,j-2) - GEP \\ 0 \end{cases}$$

s(*i*,*j*)	=	Dayhoff score for amino acids *i* and *j*
S(*i*,*j*)	=	accumulated maximum score at location *i*, *j*
S(*i*-1,*j*-1)	=	accumulated maximum score at location *i*-1, *j*-1
S(*i*,*j*-1)	=	accumulated maximum score at location *i*, *j*-1
S(*i*-1,*j*)	=	accumulated maximum score at location *i*-1, *j*
VG(*i*-2,*j*)	=	accumulated score of gap extending to *i*-1,*j*
HG(*i*,*j*-2)	=	accumulated score of gap extending to *i*, *j*-1
GP	=	Gap Penalty
GEP	=	Gap Extension Penalty

Doug Brutlag 2010

Computer Time and Space Requirements

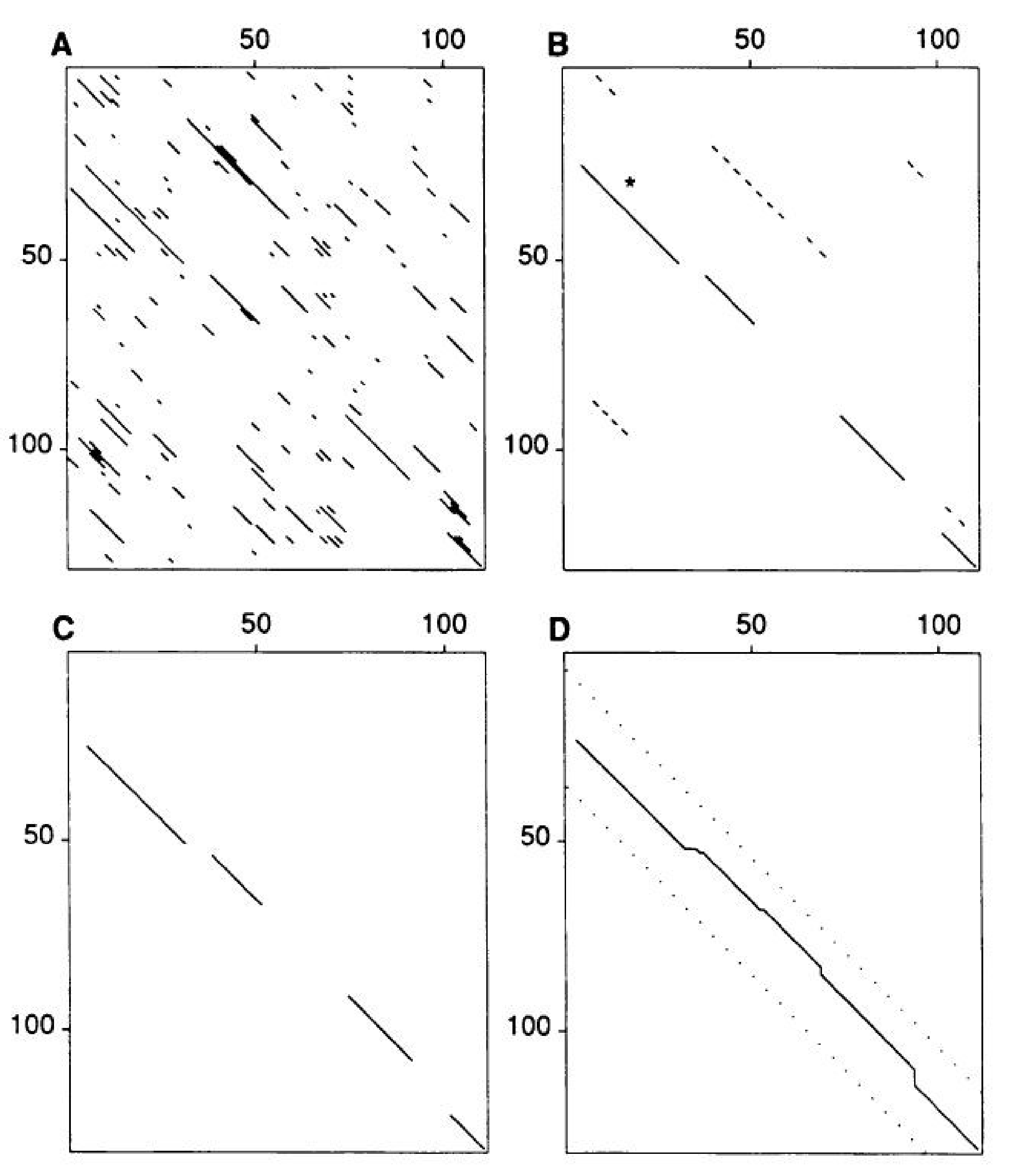
- Needleman-Wunsch
  - O(N*M) time and O(N*M) space
- Smith-Waterman
  - O(N*M) time and O(N*M) space
- Gotoh improvement of Smith-Waterman
  - O(N*M) time and O(N) space
  - Remembers maximum score and its x,y location
  - Must regenerate matrix for alignment
- Myers and Miller (using Hirschberg's method)
  - O(N*M) time and O(N) space
  - Builds optimal alignment

Smith-Waterman Homology Search

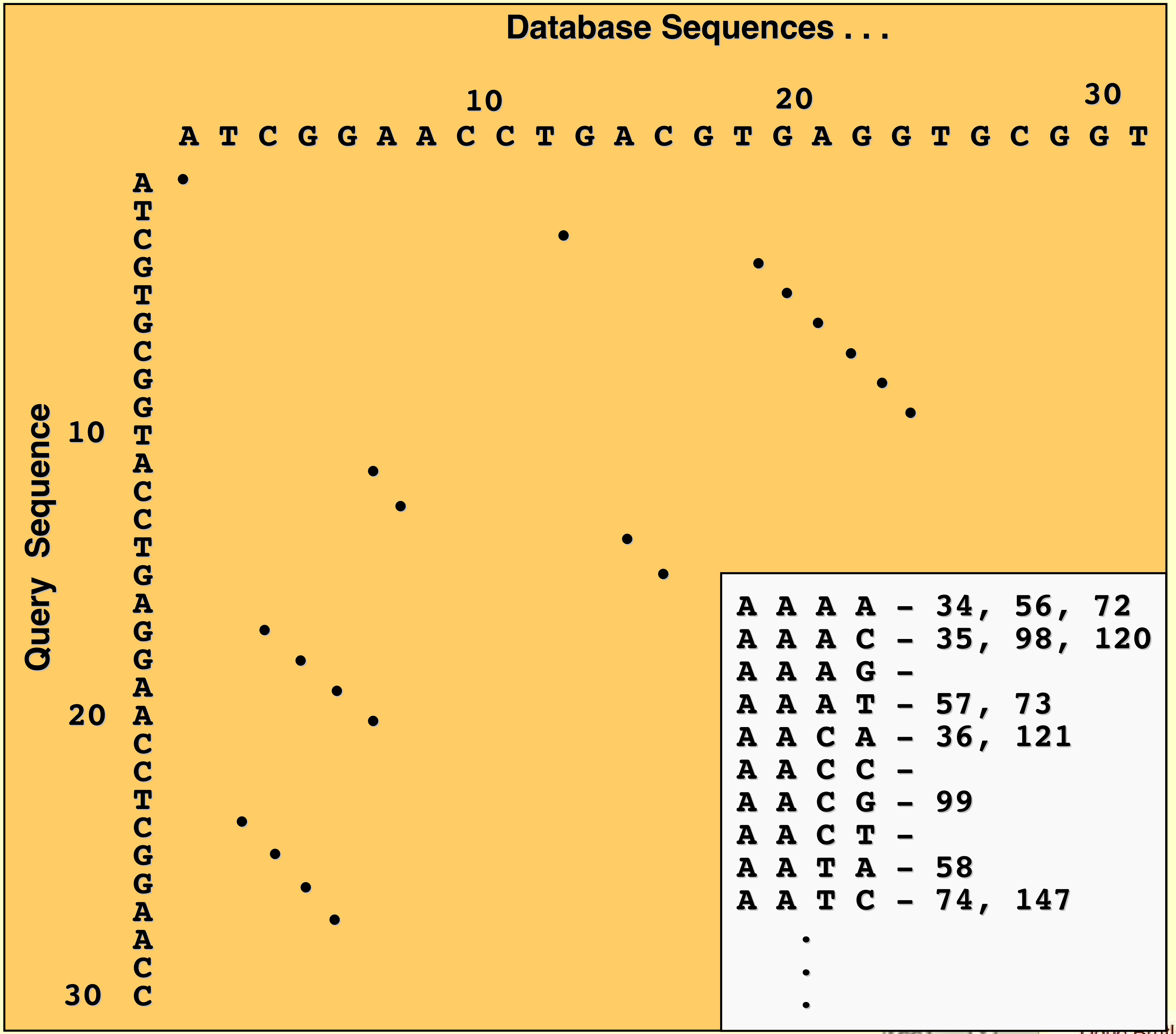
```
Query:          HU-NS1    Maximal Score:  452
PAM Matrix:  200        Gap Penalty:      5       Gap Extension:   0.5

No.   Score   Match Length DB  ID          Description             Pred. No.
  1     452   100.0     90  2  DBHB_ECOLI  DNA-BINDING PROTEIN H   8.74e-86
  2     451    99.8     90  2  DBHB_SALTY  DNA-BINDING PROTEIN H   1.54e-85
  3     336    74.3     90  2  DBHA_ECOLI  DNA-BINDING PROTEIN H   1.64e-57
  4     336    74.3     90  2  DBHA_SALTY  DNA-BINDING PROTEIN H   1.64e-57
  5     328    72.6     90  2  DBH_BACST   DNA-BINDING PROTEIN I   1.35e-55
  6     328    72.6     92  2  DBH_BACSU   DNA-BINDING PROTEIN I   1.35e-55
  7     327    72.3     90  2  DBH_VIBPR   DNA-BINDING PROTEIN H   2.35e-55
  8     302    66.8     90  2  DBH_PSEAE   DNA-BINDING PROTEIN H   2.14e-49
  9     273    60.4     91  2  DBH1_RHILE  DNA-BINDING PROTEIN H   1.47e-42
 10     272    60.2     91  2  DBH_CLOPA   DNA-BINDING PROTEIN H   2.52e-42
 11     263    58.2     90  2  DBH_RHIME   DNA-BINDING PROTEIN H   3.18e-40
 12     261    57.7     91  2  DBH5_RHILE  DNA-BINDING PROTEIN H   9.29e-40
 13     250    55.3     94  2  DBH_ANASP   DNA-BINDING PROTEIN H   3.32e-37
 14     233    51.5     93  2  DBH_CRYPH   DNA-BINDING PROTEIN H   2.70e-33
 15     226    50.0     95  2  DBH_THETH   DNA-BINDING PROTEIN I   1.07e-31
 16     210    46.5     99  3  IHFA_SERMA  INTEGRATION HOST FACT   4.46e-28
 17     206    45.6    100  3  IHFA_RHOCA  INTEGRATION HOST FACT   3.52e-27
 18     205    45.4     99  3  IHFA_SALTY  INTEGRATION HOST FACT   5.90e-27
 19     204    45.1     99  3  IHFA_ECOLI  INTEGRATION HOST FACT   9.87e-27
 20     200    44.2     94  3  IHFB_ECOLI  INTEGRATION HOST FACT   7.71e-26
 21     200    44.2     94  3  IHFB_SERMA  INTEGRATION HOST FACT   7.71e-26
 22     165    36.5     99  5  TF1_BPSP1   TRANSCRIPTION FACTOR    3.42e-18
 23     147    32.5     90  2  DBH_THEAC   DNA-BINDING PROTEIN H   2.12e-14
 24      76    16.8    477  2  GLGA_ECOLI  GLYCOGEN SYNTHASE (EC  3.80e-01
```

Steps in FASTA Method

Lipman & Pearson, Science 1985



FASTA Word Search (Query Hashing)

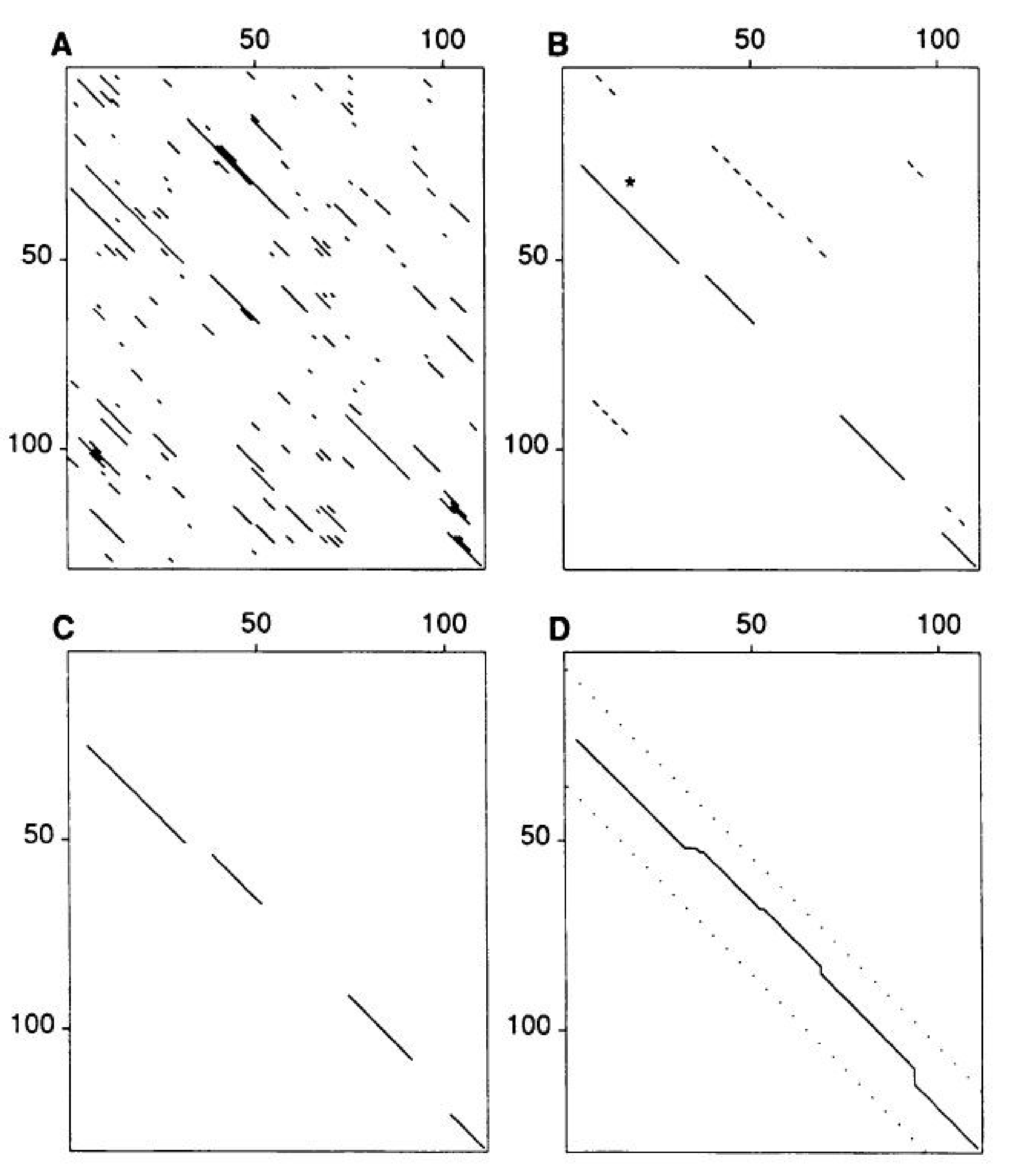

Database Sequences . . .
10
20
30
A T C G G A A C C T G A C G T G A G G T G C G G T
Query Sequence
10
20
30
A T C G T G C G G T A C C T G A G G A A C C T C G G A A C C

A A A A	34, 56, 72
A A A C	35, 98, 120
A A A G	
A A A T	57, 73
A A C A	36, 121
A A C C	
A A C G	99
A A C T	
A A T A	58
A A T C	74, 147

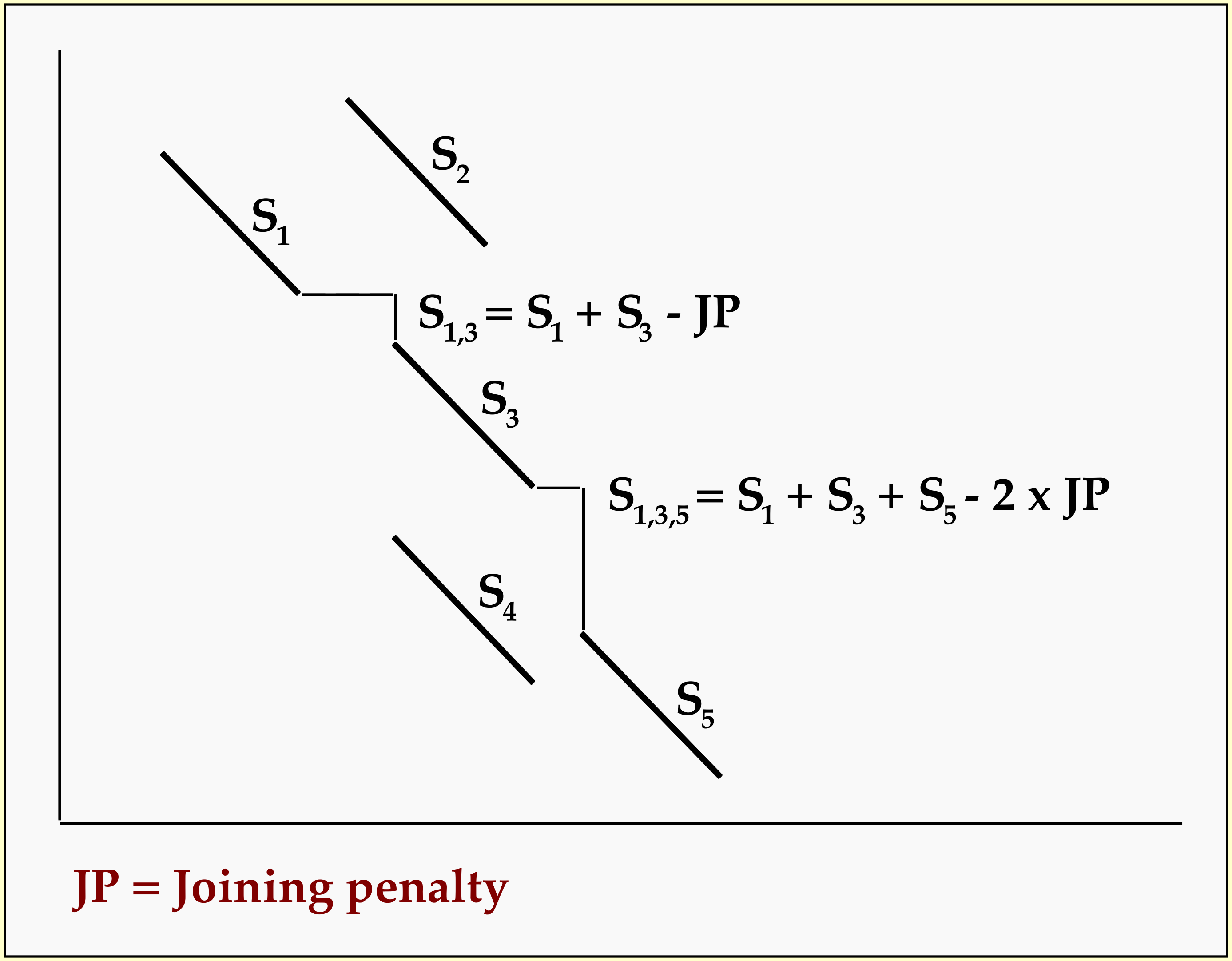
.
.
.

Doug Brutlag 2010

Steps in FASTA Method

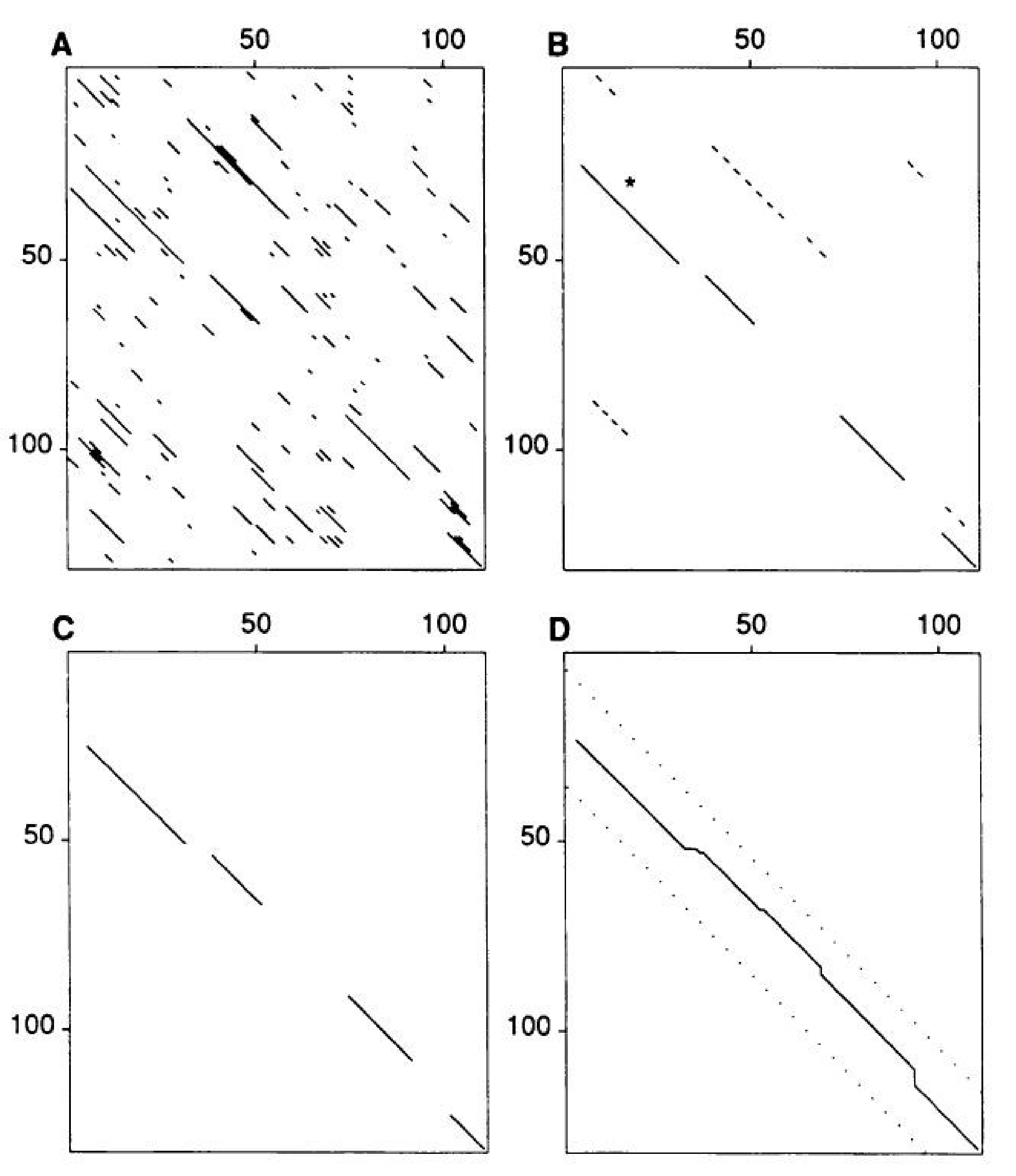


Joining Diagonals of Similarity


S_2
S_1
$S_{1,3} = S_1 + S_3$ - JP
S_3
$S_{1,3,5} = S_1 + S_3 + S_5$ - 2 x JP
S_4
S_5

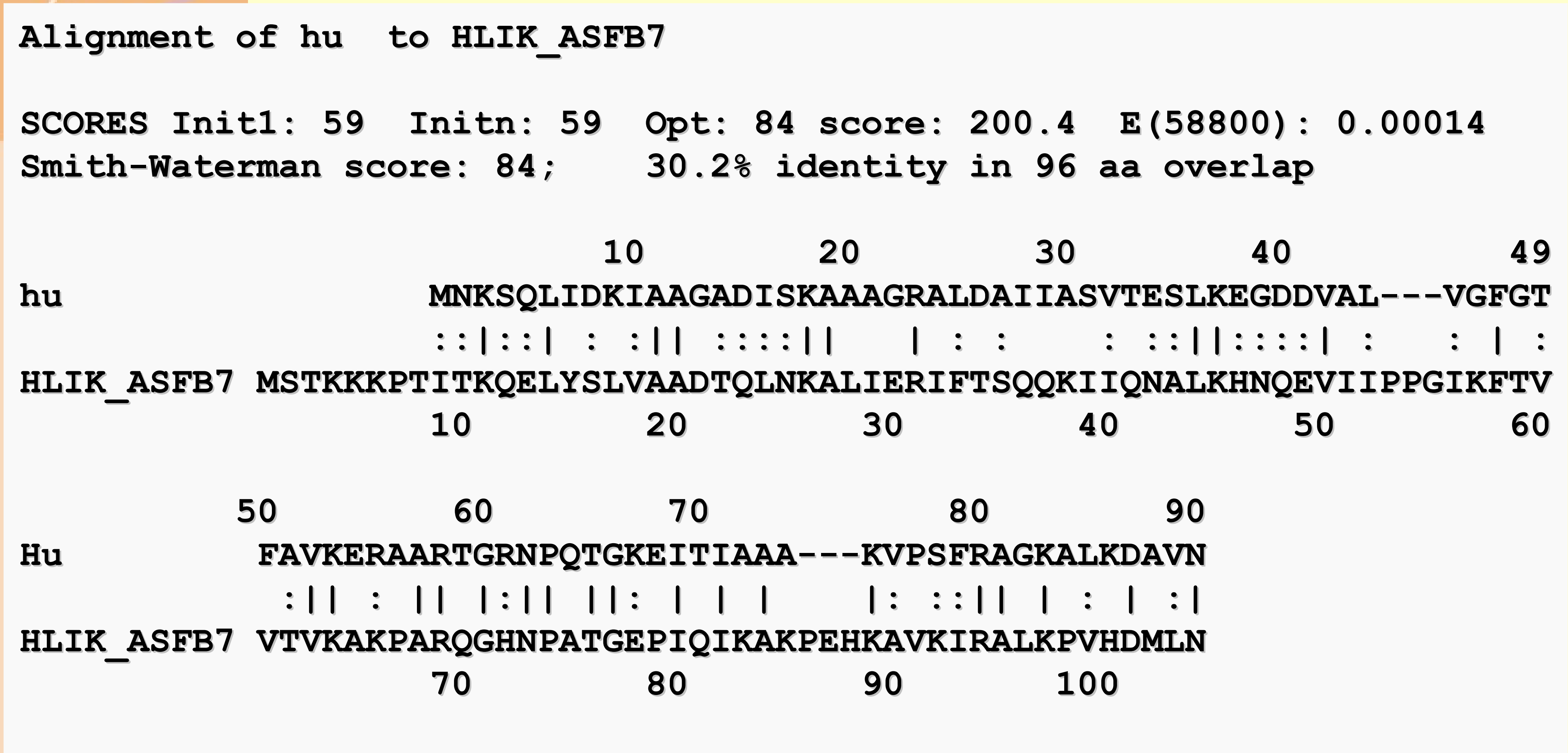
JP = Joining penalty

Steps in FastA Method



FastA Search (cont.)

(HU versus SwissProt)



```
Alignment of hu  to HLIK_ASFB7

SCORES Init1: 59  Initn: 59  Opt: 84 score: 200.4  E(58800): 0.00014
Smith-Waterman score: 84;    30.2% identity in 96 aa overlap

                           10        20        30        40          49
hu                 MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVAL---VGFGT
                   ::|::| : :|| ::::||   | : :    : ::||::::| :   : | :
HLIK_ASFB7 MSTKKKPTITKQELYSLVAADTQLNKALIERIFTSQQKIIQNALKHNQEVIIPPGIKFTV
                   10        20        30        40        50        60

          50        60        70           80        90
Hu         FAVKERAARTGRNPQTGKEITIAAA---KVPSFRAGKALKDAVN
            :|| : || |:|| ||: | | |    |: ::|| | : | :|
HLIK_ASFB7 VTVKAKPARQGHNPATGEPIQIKAKPEHKAVKIRALKPVHDMLN
                   70        80        90       100
```

Original BLAST Algorithm

Altschul et al. J. Mol. Biol. 1990 215, 403-410.

- Basic Local Alignment Search Tool
- Indexes words in database
- Calculates “neighborhood” of each word in query using BLOSUM matrix and probability threshold
- Looks up all words and neighbors from query in database index to find High-scoring Segment Pairs (HSPs)
- 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

Expectation of High-scoring Segment Pairs (HSPs)

Karlin and Altschul PNAS 1990, 87, 2264-2268.

$$\mathrm{Prob}(Score > X) \approx 1 - \exp\{-Ke^{-\lambda X}\}$$

where λ is the root of the equation:

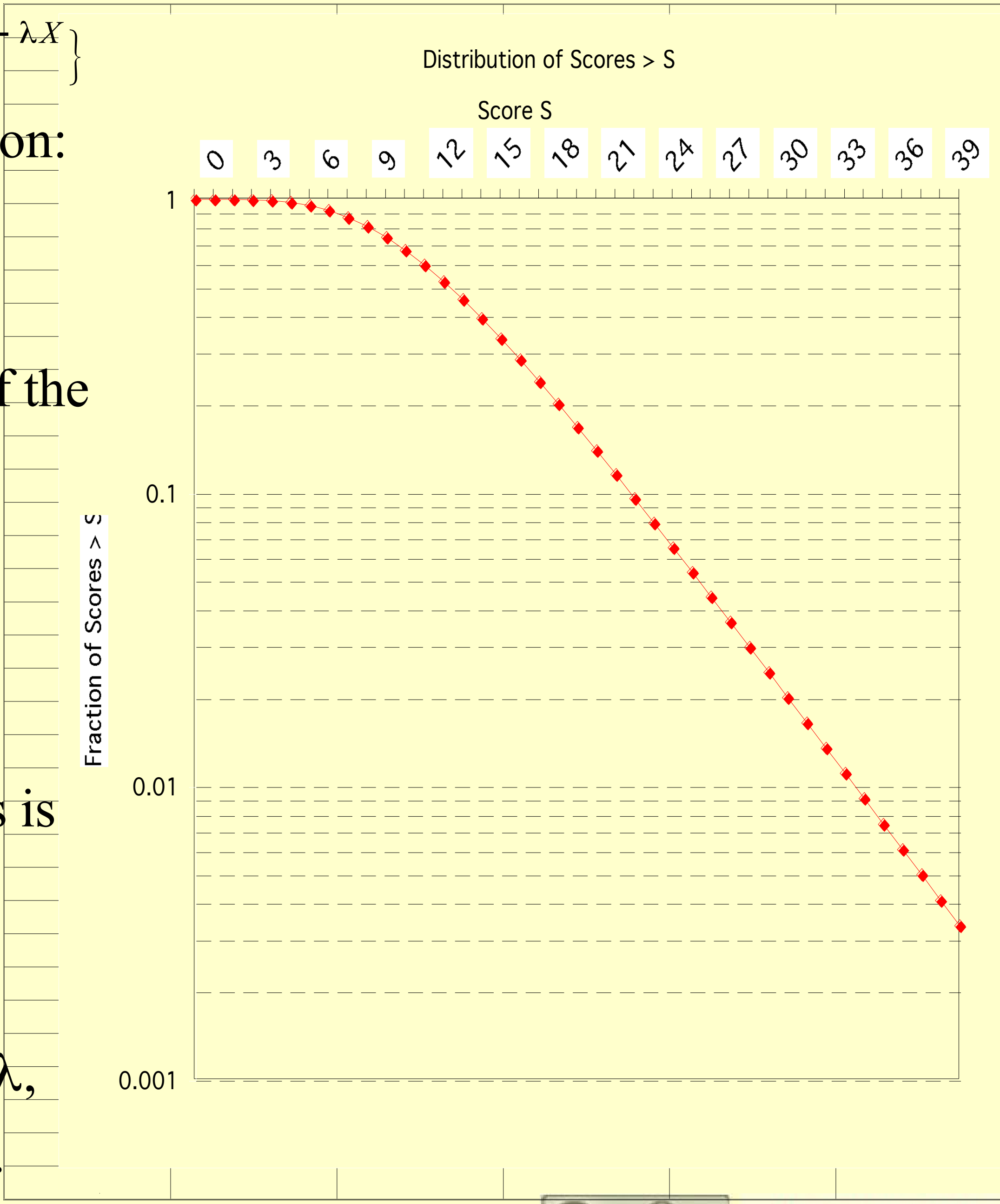
$$\sum_{i=1}^{r}\sum_{j=1}^{r} p_i p_j \exp\{\lambda s_{ij}\} = 1$$

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

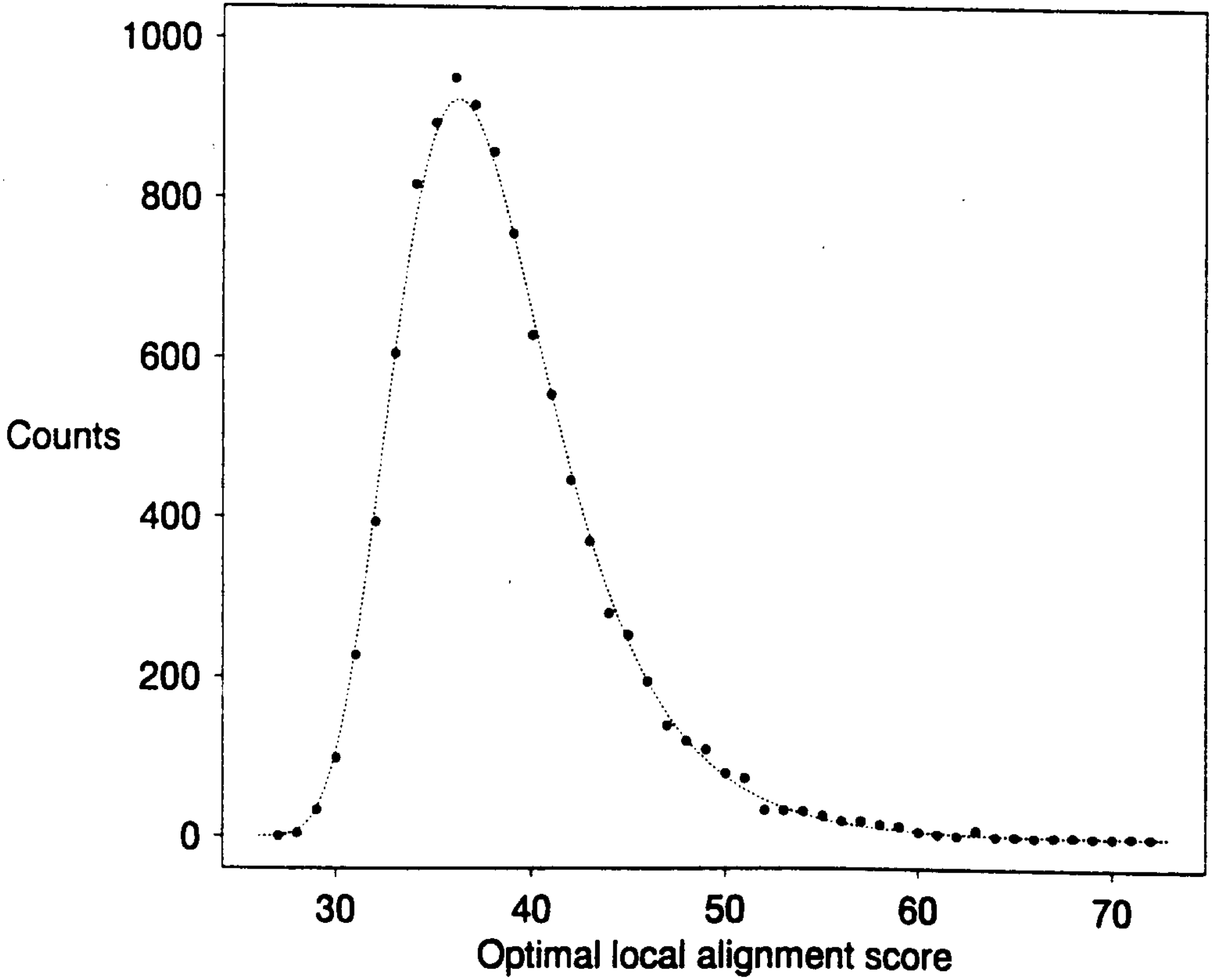
s_{ij} are the similarity scores of two residues i and j.

If the expected value of the scores for random sequences is < 0, 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.



Extreme Value Distribution of Scores

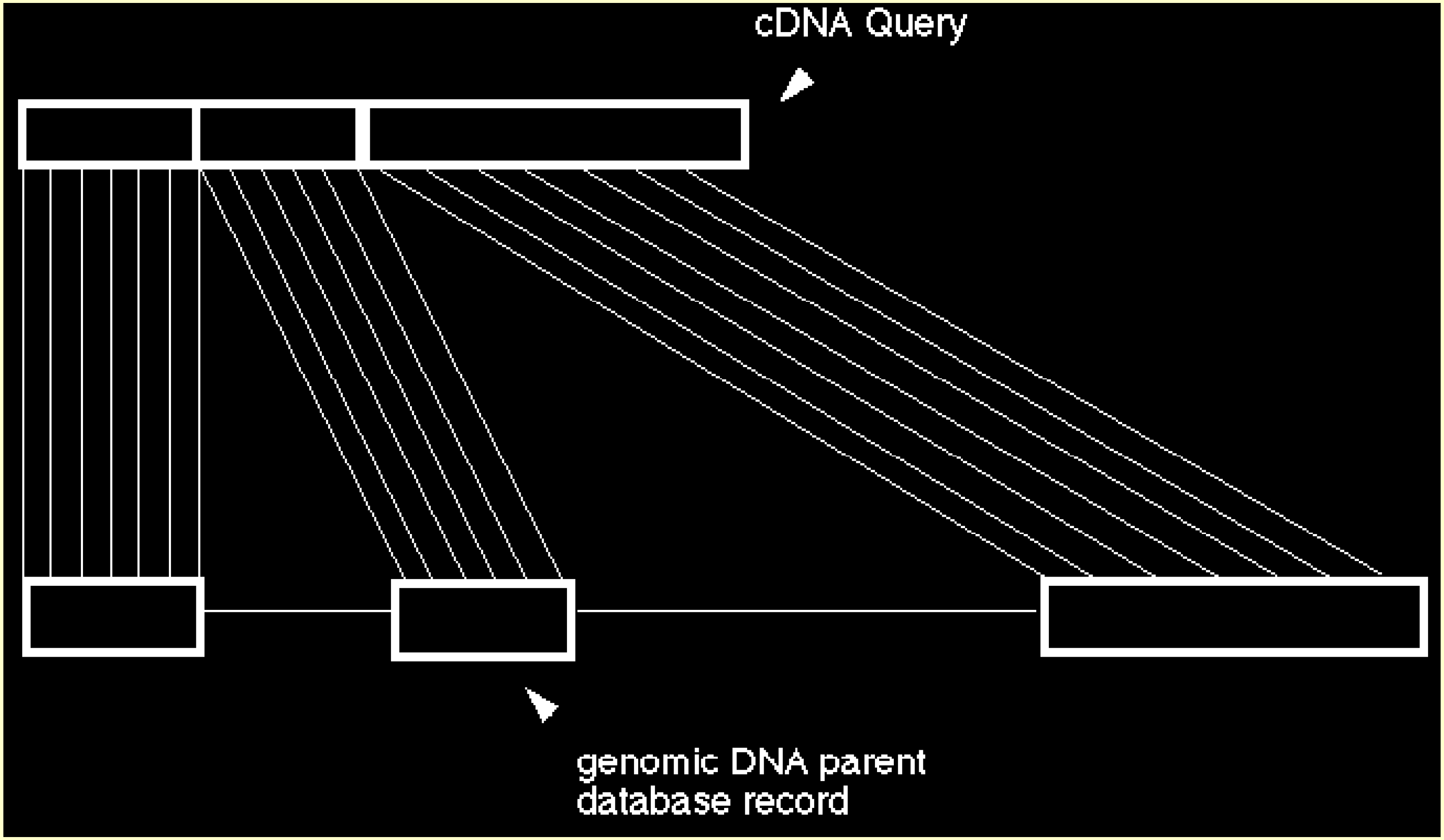


Original BLAST vs Smith & Waterman

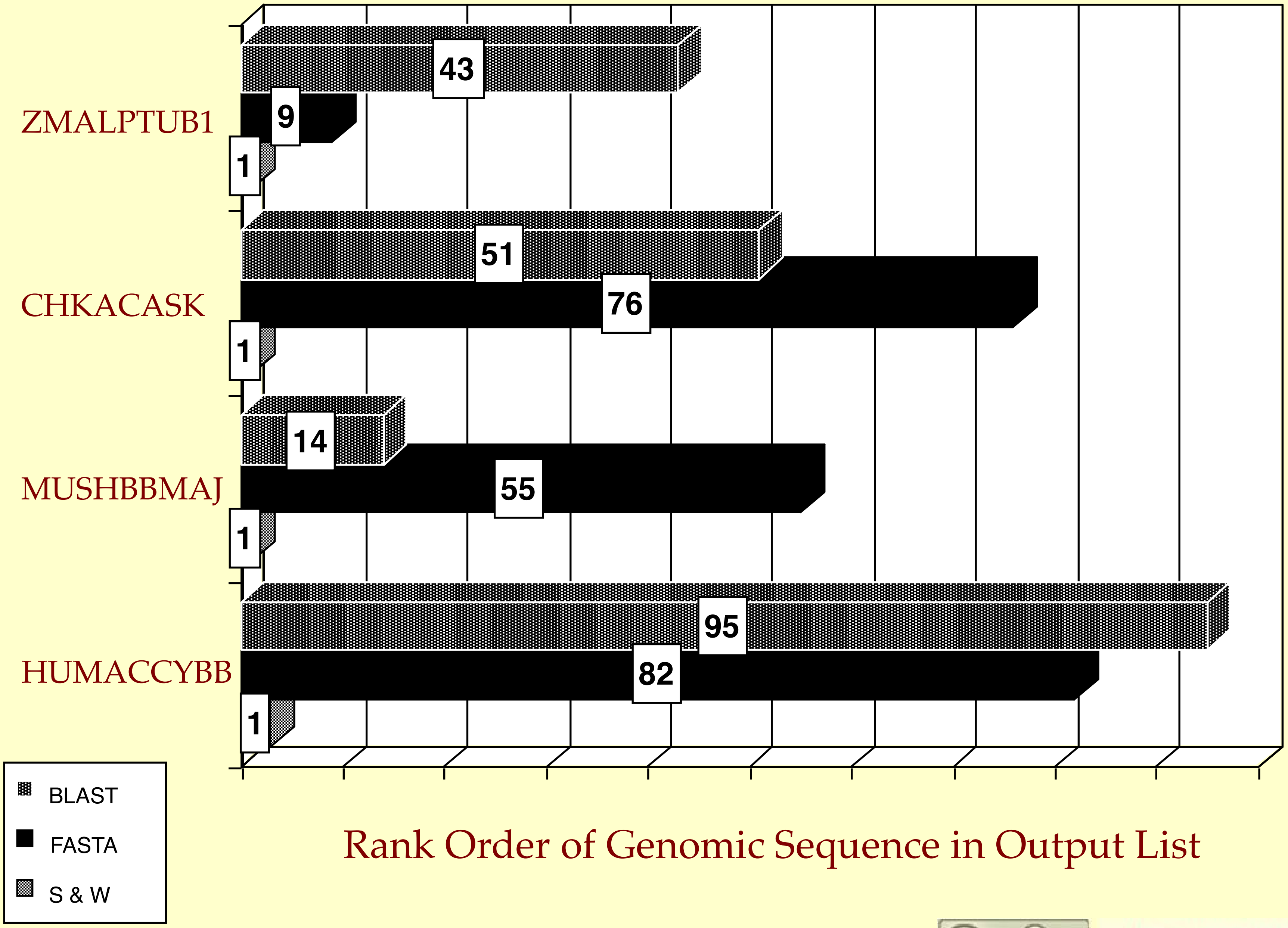
(Metr vs Swiss-Prot: 60 members expected)

		Penalties		Threshold (5% expectation)		
Program	PAM	Gap	Gap Size	Number Right (TP/60)	Number Wrong (FP)	Number Missed (FN/60)
S & W	1	20	5	3	4	57
S & W	50	20	5	27	1	33
S & W	100	20	5	42	1	18
S & W	150	20	5	51	0	9
S & W	200	20	5	53	0	7
S & W	250	20	5	50	0	10
S & W	200	5	5	2	0	58
S & W	200	10	5	53	2	7
S & W	200	20	5	53	0	7
S & W	200	40	5	53	0	7
S & W	200	80	5	51	0	9
BLAST	2	∞	∞	2	0	58
BLAST	50	∞	∞	23	0	37
BLAST	100	∞	∞	32	0	28
BLAST	150	∞	∞	35	0	25
BLAST	200	∞	∞	40	0	20
BLAST	250	∞	∞	35	0	25

cDNA Queries Require Affine Gap Penalties

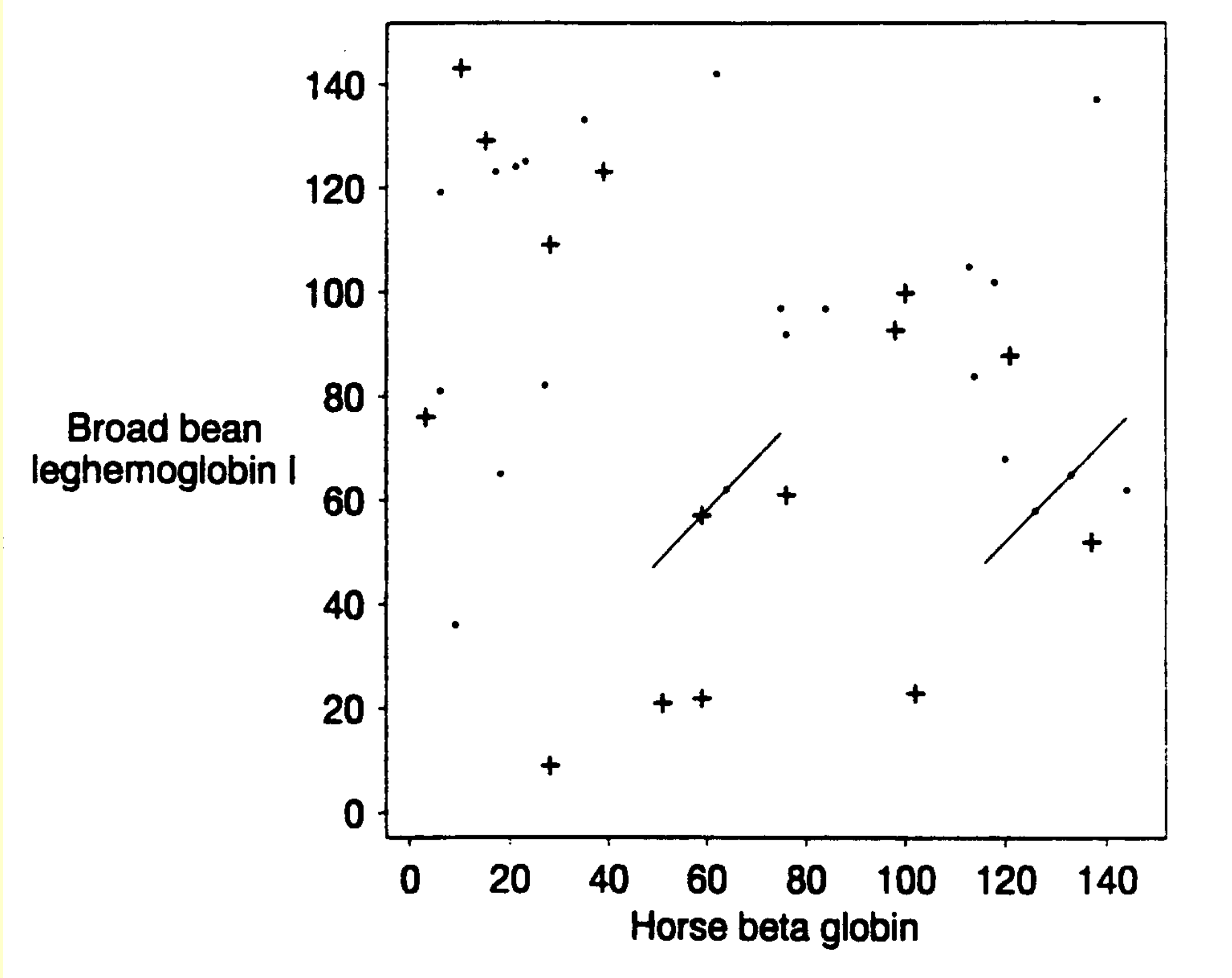


Detecting Genomic Sequences with cDNA Queries

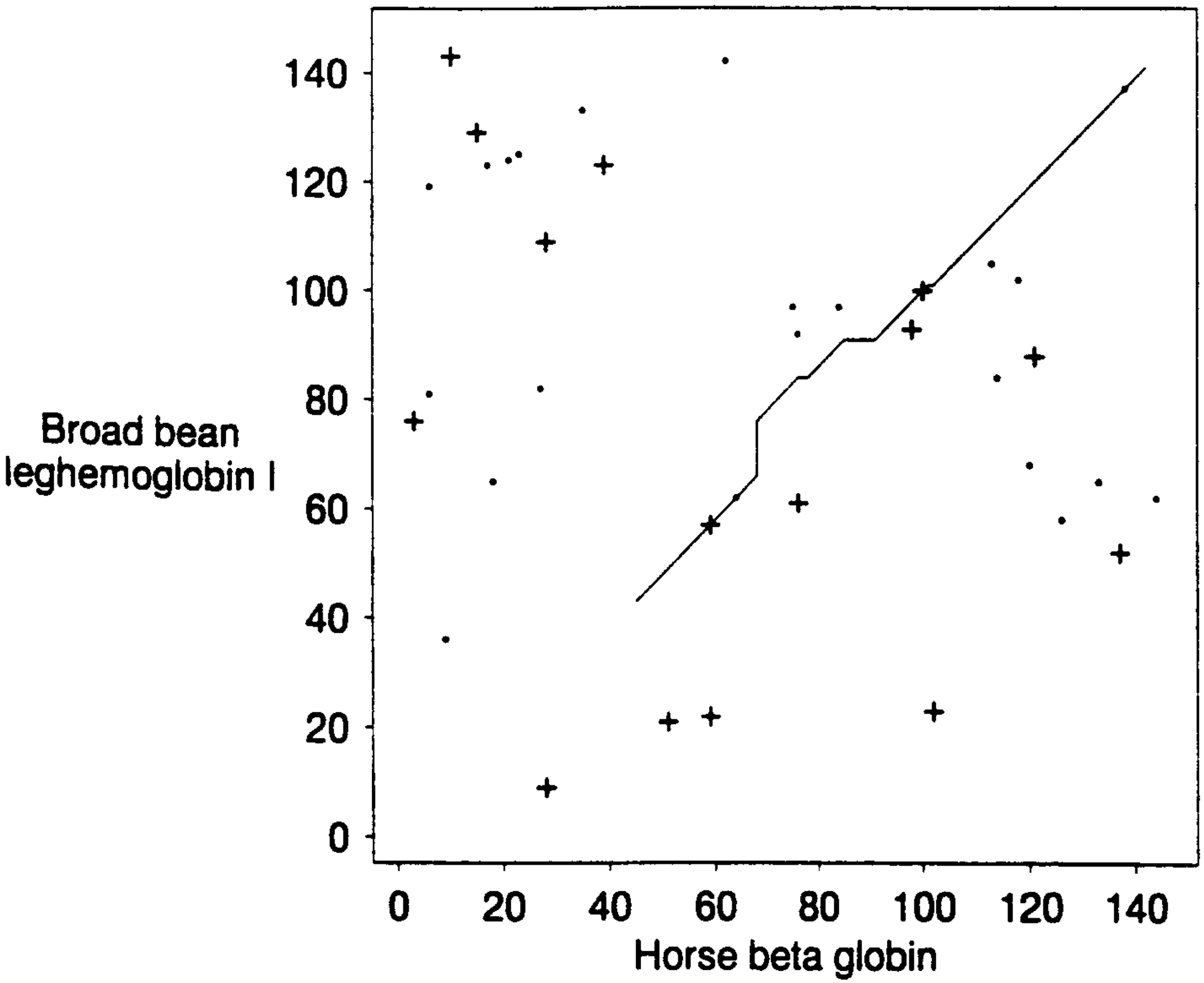


Doug Brutlag 2010

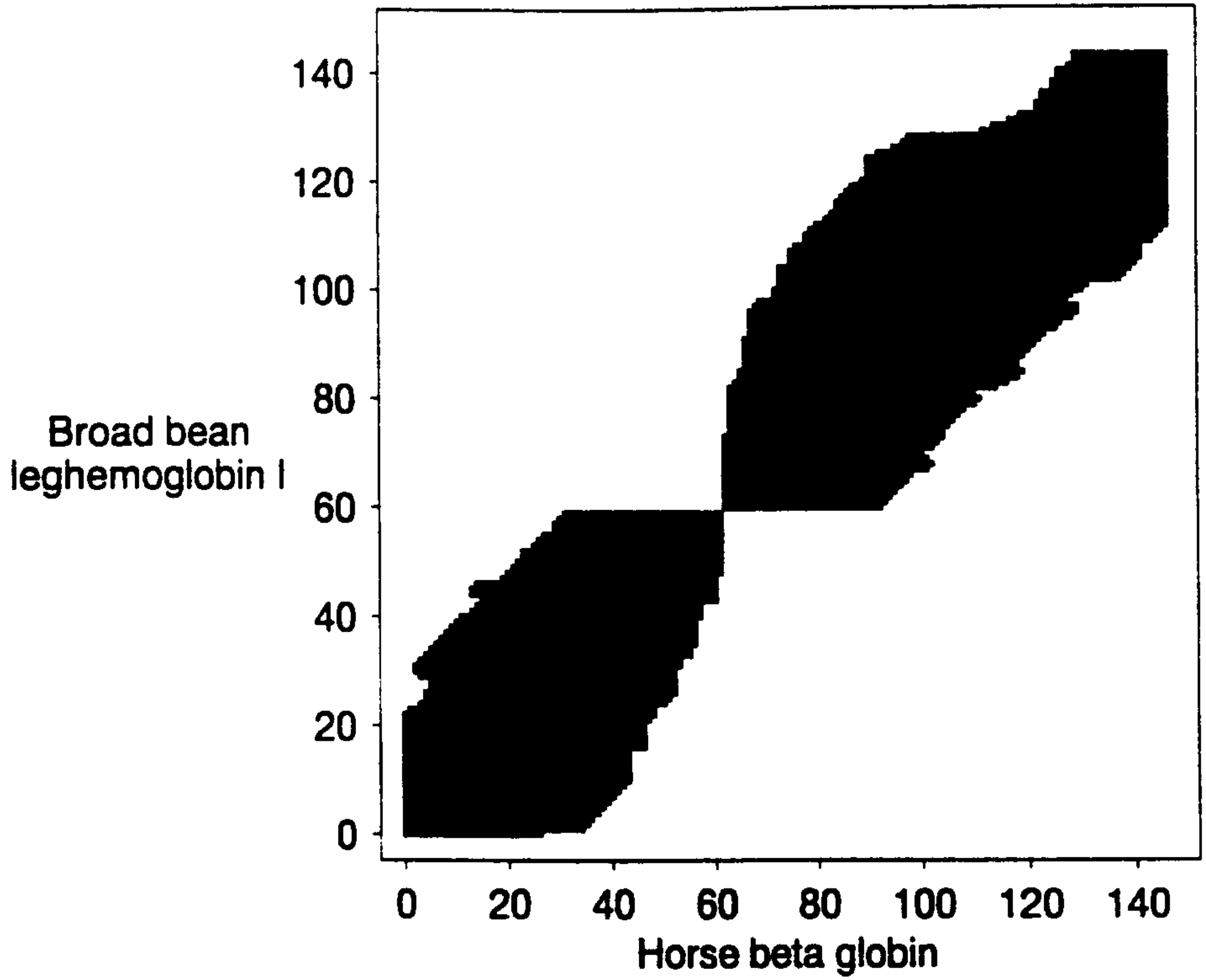
GAPPED BLAST Starts with a Two Hit Approach



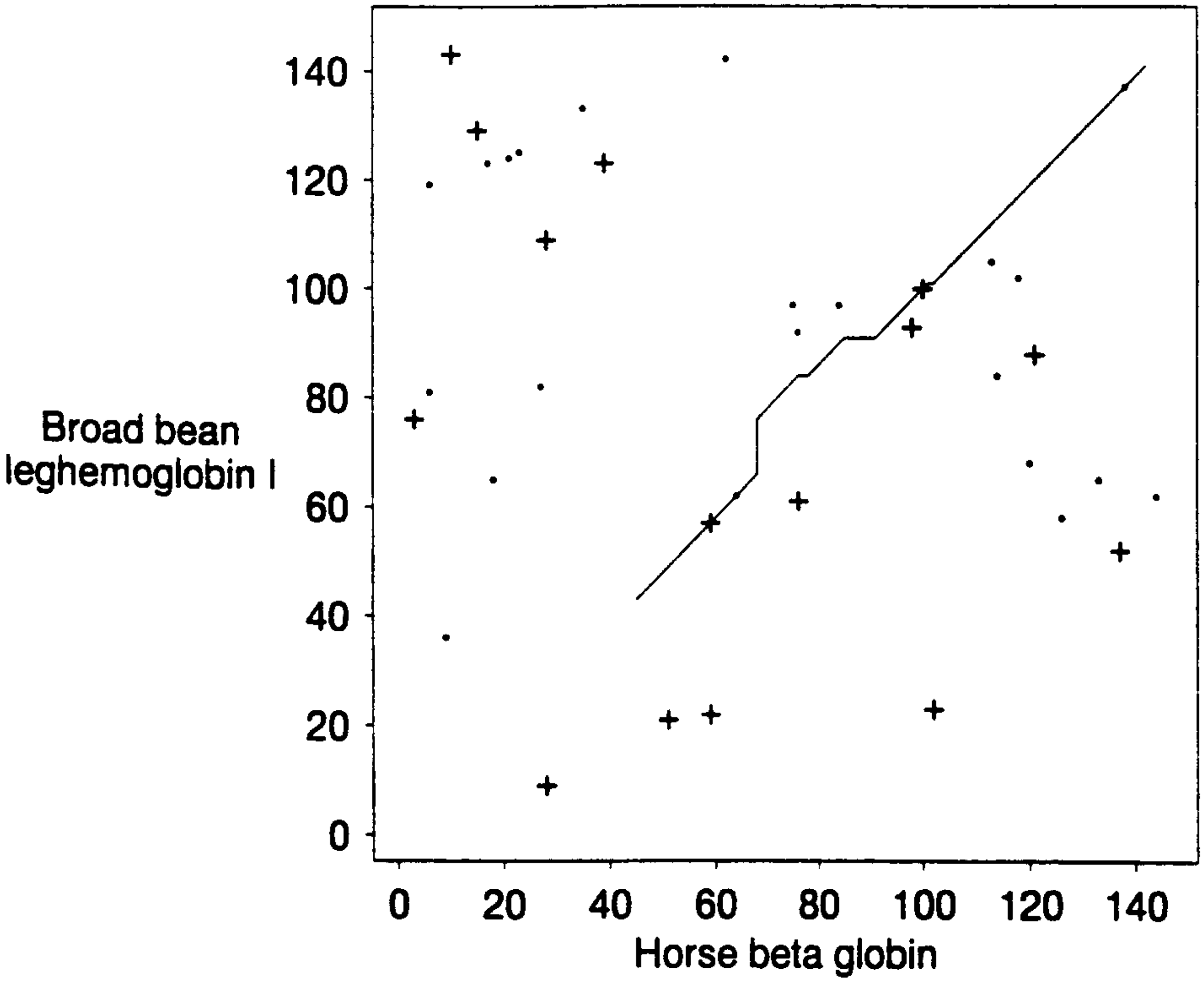
GAPPED BLAST Extension of Two Hit HSP



Region Explored by GAPPED BLAST

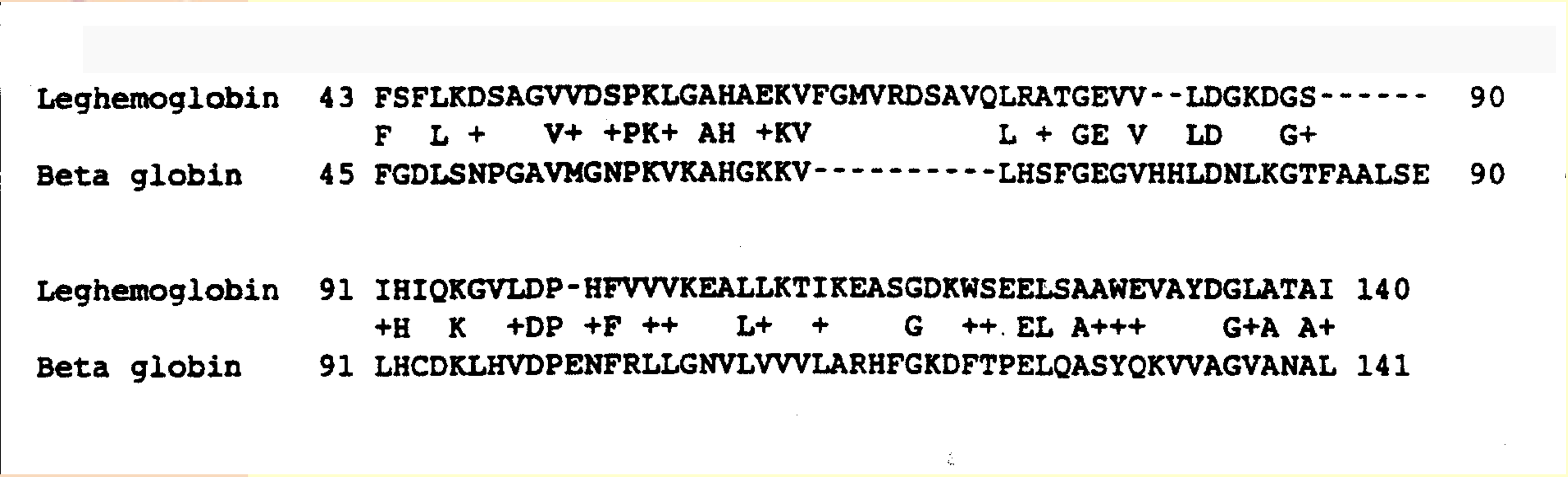


GAPPED BLAST Extension of Two Hit HSP



Doug Brutlag 2010

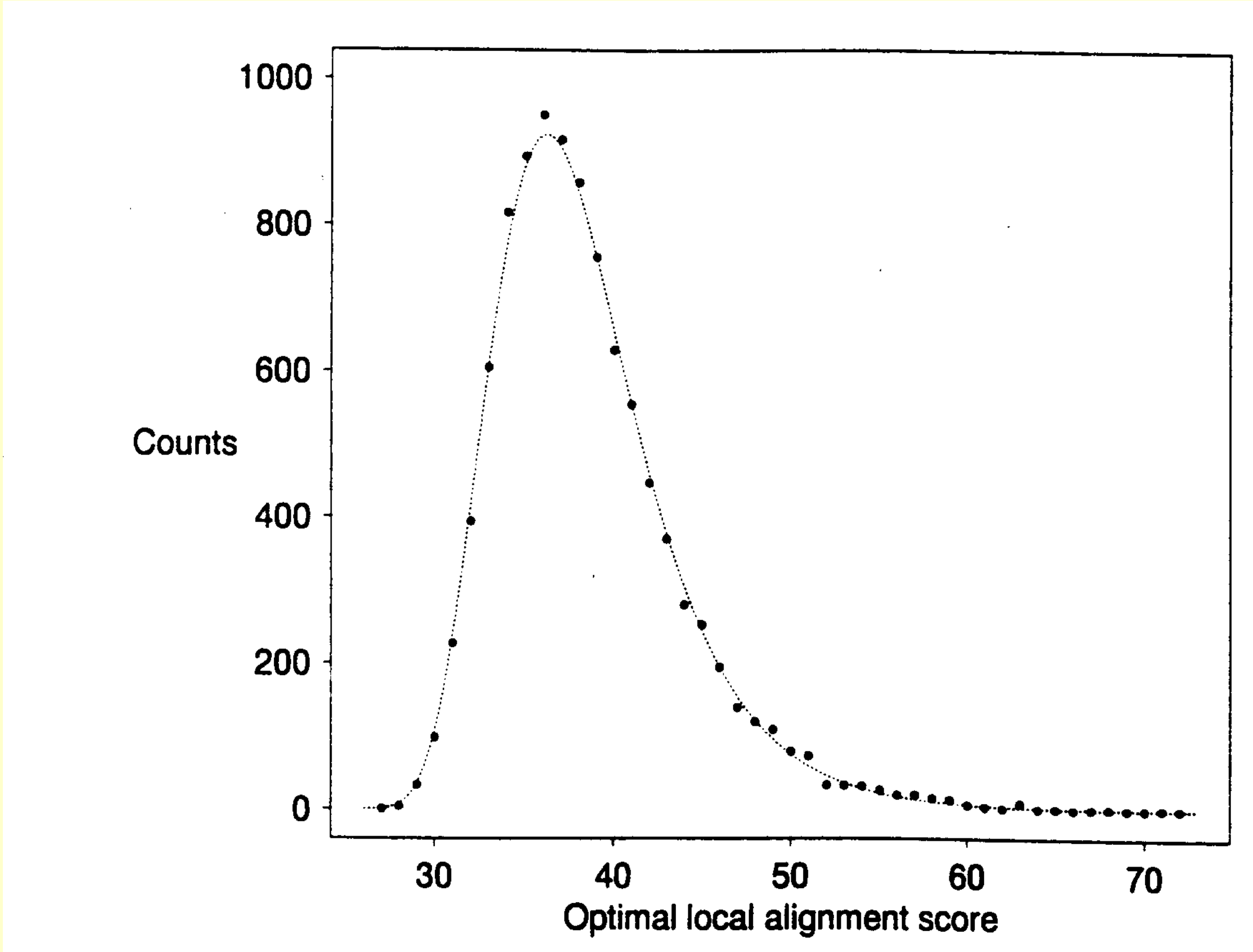
GAPPED BLAST Alignment



```
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
```

Doug Brutlag 2010

Extreme Value Distribution of Scores

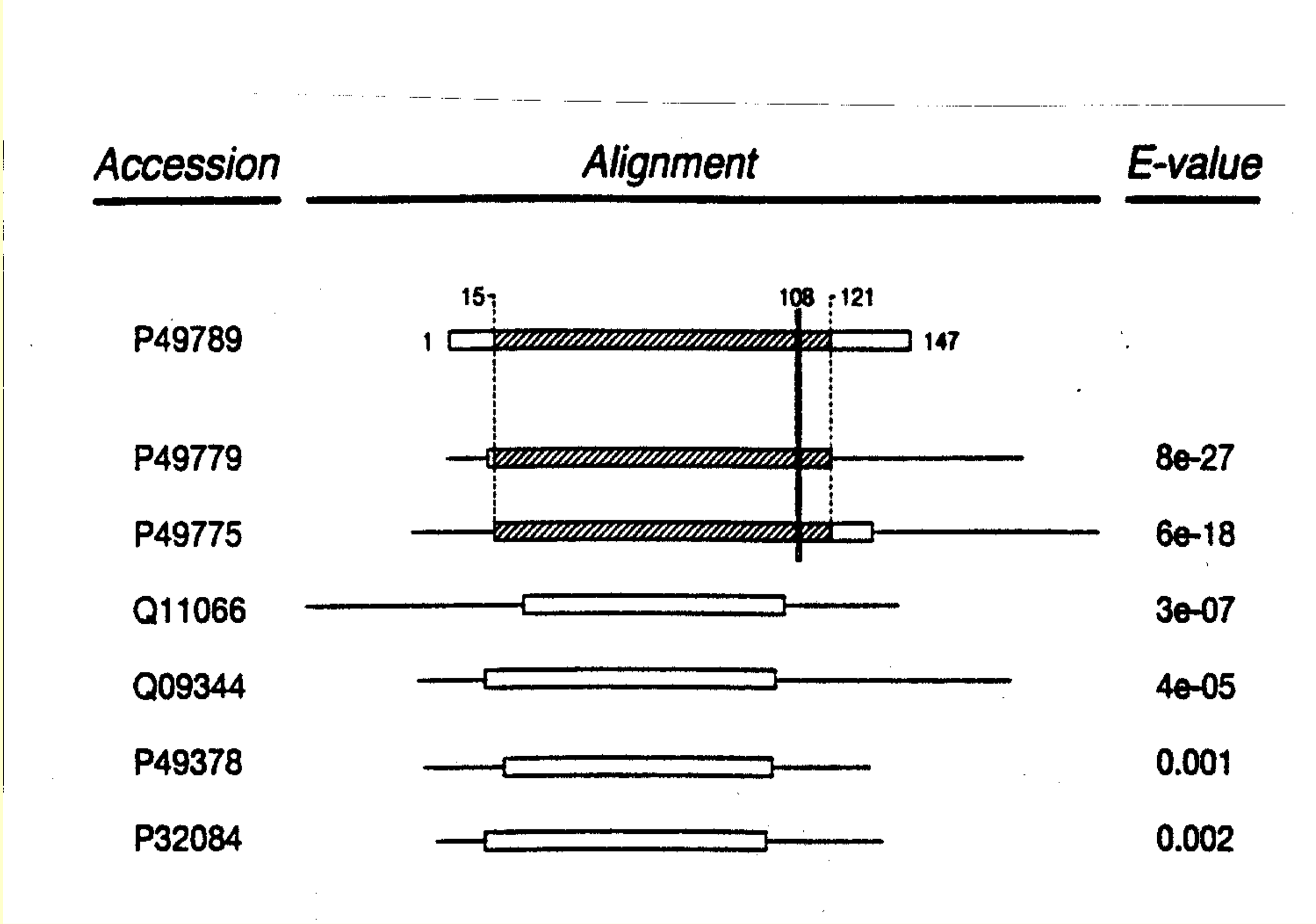


Gapped BLAST Advanced Settings

http://www.ncbi.nlm.nih.gov/BLAST/

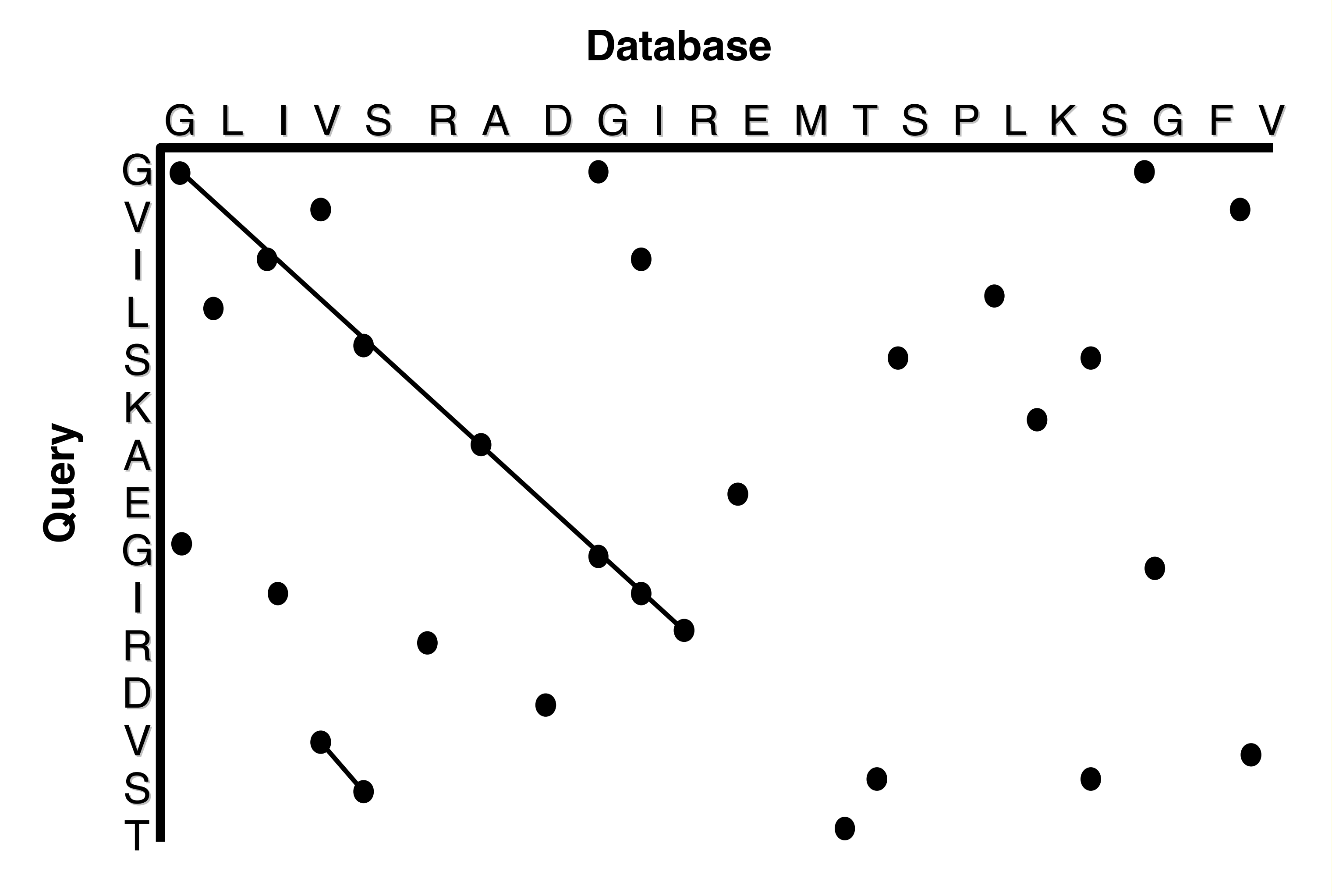
- -G Cost to open gap [Integer]
  - default = 5 for nucleotides 11 proteins
- -E Cost to extend gap [Integer]
  - default = 2 nucleotides 1 proteins
- -q Penalty for nucleotide mismatch [Integer]
  - default = -3
- -r reward for nucleotide match [Integer]
  - default = 1
- -e expect value [Real]
  - default = 10
- -W wordsize [Integer]
  - default = 11 nucleotides 3 proteins

PSI-BLAST Alignment



Accession	Alignment	E-value
P49789		
P49779		8e-27
P49775		6e-18
Q11066		3e-07
Q09344		4e-05
P49378		0.001
P32084		0.002

Dynamic Programming

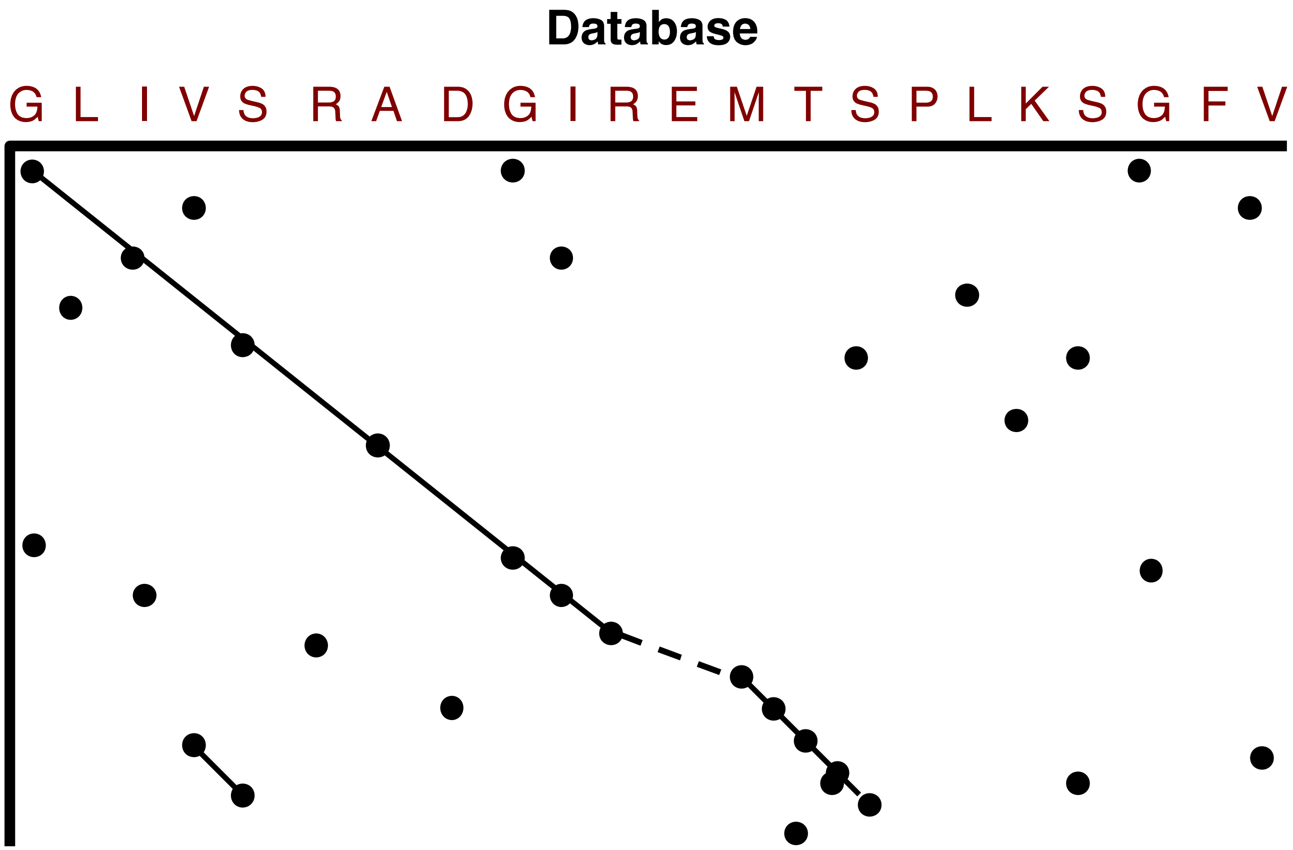


Doug Brutlag 2010

Generalized Dynamic Programming

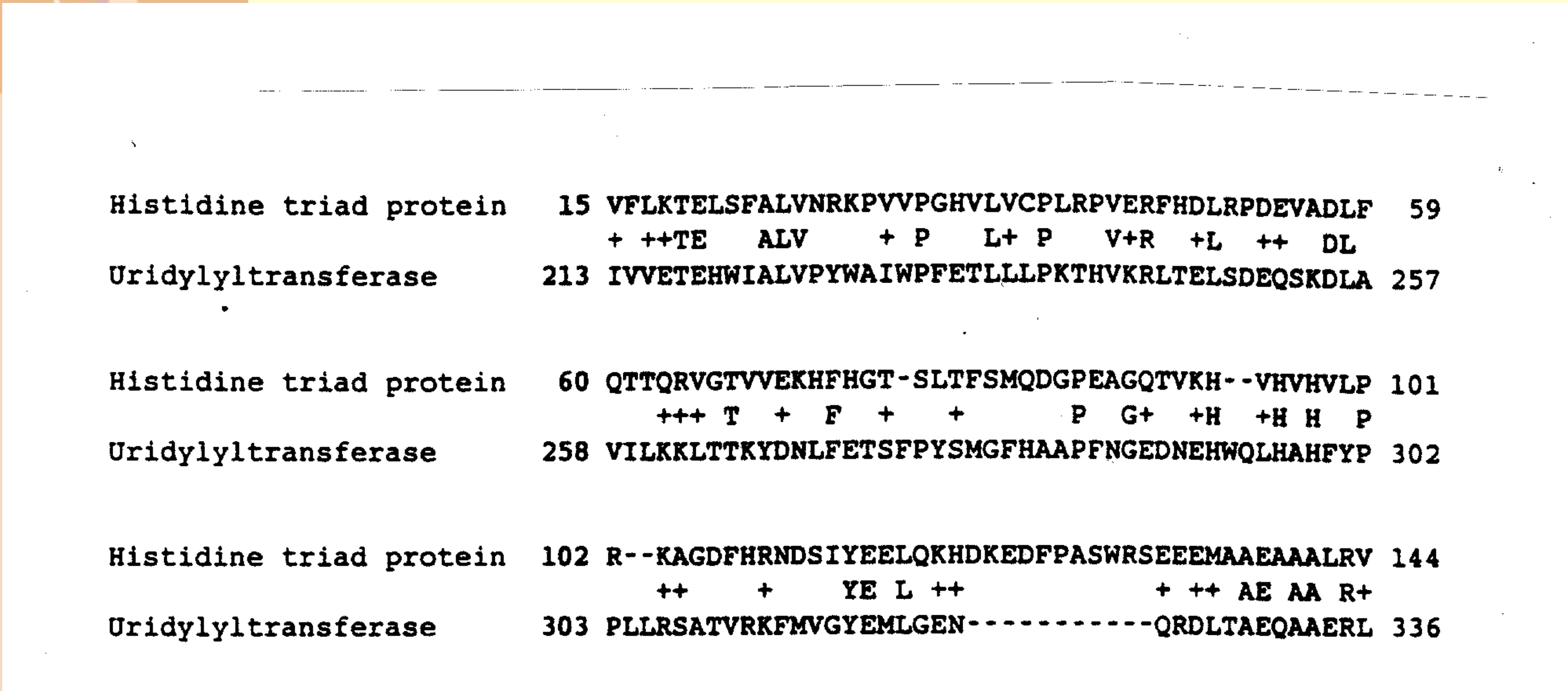
Query

```
  A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
(3  4  5  1-5  2  7  1  9-3  5  0-6  1  2  5  6-7  3  4)
(3  0  5  9-5-3  2  2-3-3  2  0-2  1  1-5  6-7  3  4)
(1  3  5  2-5-3  2  2-3  2  2  0-2  1  1-5  6-7  3  4)
(6  4-3  0  2-1-3-1  4  3-5  1-3  3  4-5  2-3  2-1)
(1  3  5  2-5-3  2  2-3  2  2  0-2  1  1-5  6-7  3  4)
(3  0  5  9-5-3  2  2-3-3  2  0-2  1  1-5  6-7  3  4)
(2-3  4-2  5  2-3  1-1  0  2  5-4  2-3  4  5-1  0  4)
(6  4-3  0  2-1-3-1  4  3-5  1-3  3  4-5  2-3  2-1)
(2-3  4-2  5  2-3  1-1  0  2  5-4  2-3  4  5-1  0  4)
(1  3  5  2-5-3  2  2-3  2  2  0-2  1  1-5  6-7  3  4)
(3  4  5  1-5  2  7  1  9-3  5  0-6  1  2  5  6-7  3  4)
(1  3  5  2-5-3  2  2-3  2  2  0-2  1  1-5  6-7  3  4)
(6  4-3  0  2-1-3-1  4  3-5  1-3  3  4-5  2-3  2-1)
(2-3  4-2  5  2-3  1-1  0  2  5-4  2-3  4  5-1  0  4)
(2-3  4-2  5  2-3  1-1  0  2  5-4  2-3  4  5-1  0  4)
(6  4-3  0  2-1-3-1  4  3-5  1-3  3  4-5  2-3  2-1)
(3  0  5  9-5-3  2  2-3-3  2  0-2  1  1-5  6-7  3  4)
(1  3  5  2-5-3  2  2-3  2  2  0-2  1  1-5  6-7  3  4)
```



Doug Brutlag 2010

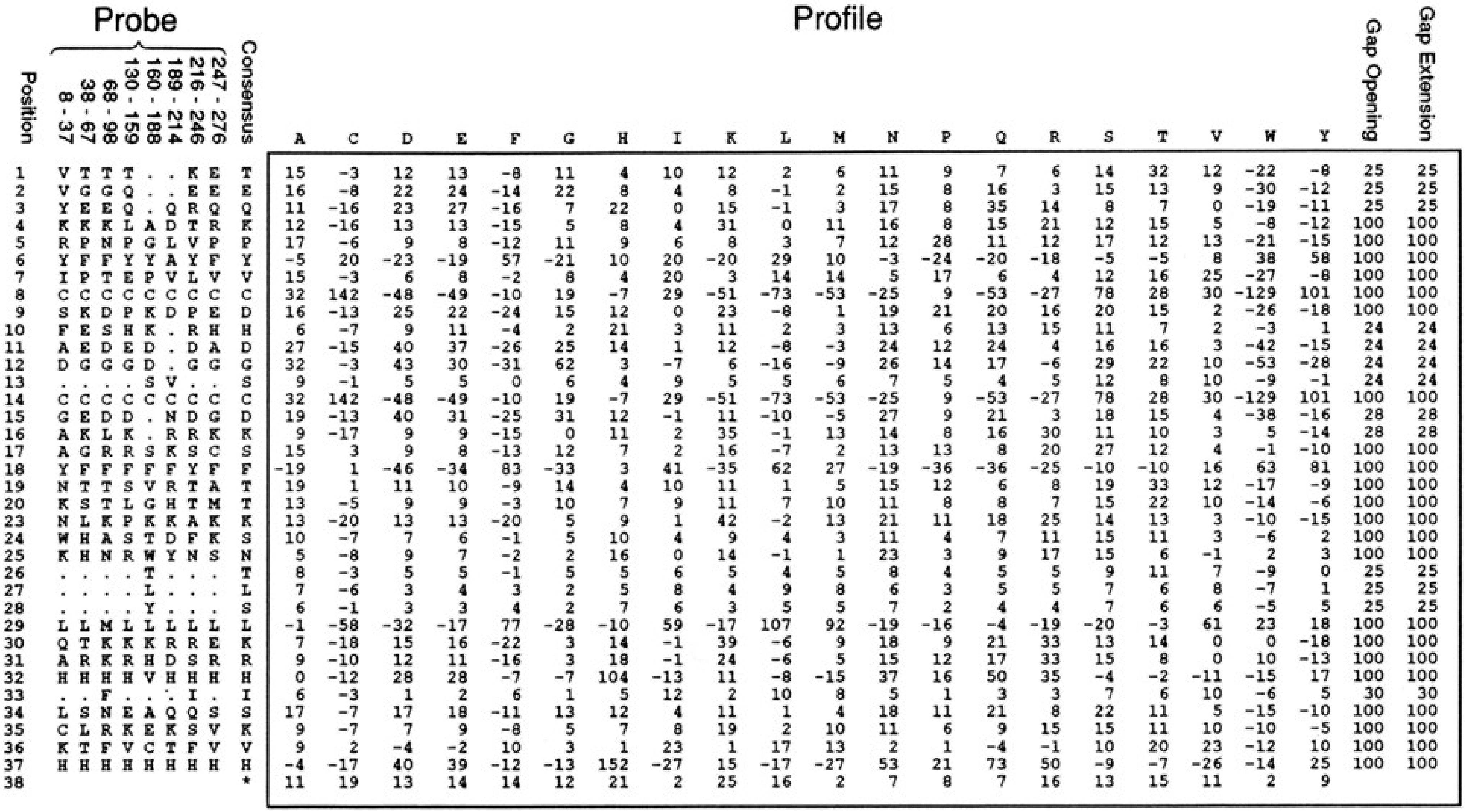
PSI-BLAST Alignment I



```
Histidine triad protein    15 VFLKTELSFALVNRKPVVPGHVLVCPLRPVERFHDLRPDEVADLF  59
                              + ++TE   ALV    + P   L+ P   V+R  +L  ++  DL
Uridylyltransferase       213 IVVETEHWIALVPYWAIWPFETLLLPKTHVKRLTELSDEQSKDLA 257

Histidine triad protein    60 QTTQRVGTVVEKHFHGT-SLTFSMQDGPEAGQTVKH--VHVHVLP 101
                                 +++ T  +  F  +   +      P  G+  +H  +H H  P
Uridylyltransferase       258 VILKKLTTKYDNLFETSFPYSMGFHAAPFNGEDNEHWQLHAHFYP 302

Histidine triad protein   102 R--KAGDFHRNDSIYEELQKHDKEDFPASWRSEEEMAAEAAALRV 144
                                 ++    +    YE L ++           + ++ AE AA R+
Uridylyltransferase       303 PLLRSATVRKFMVGYEMLGEN-----------QRDLTAEQAAERL 336
```

Doug Brutlag 2010

Sequence Profile

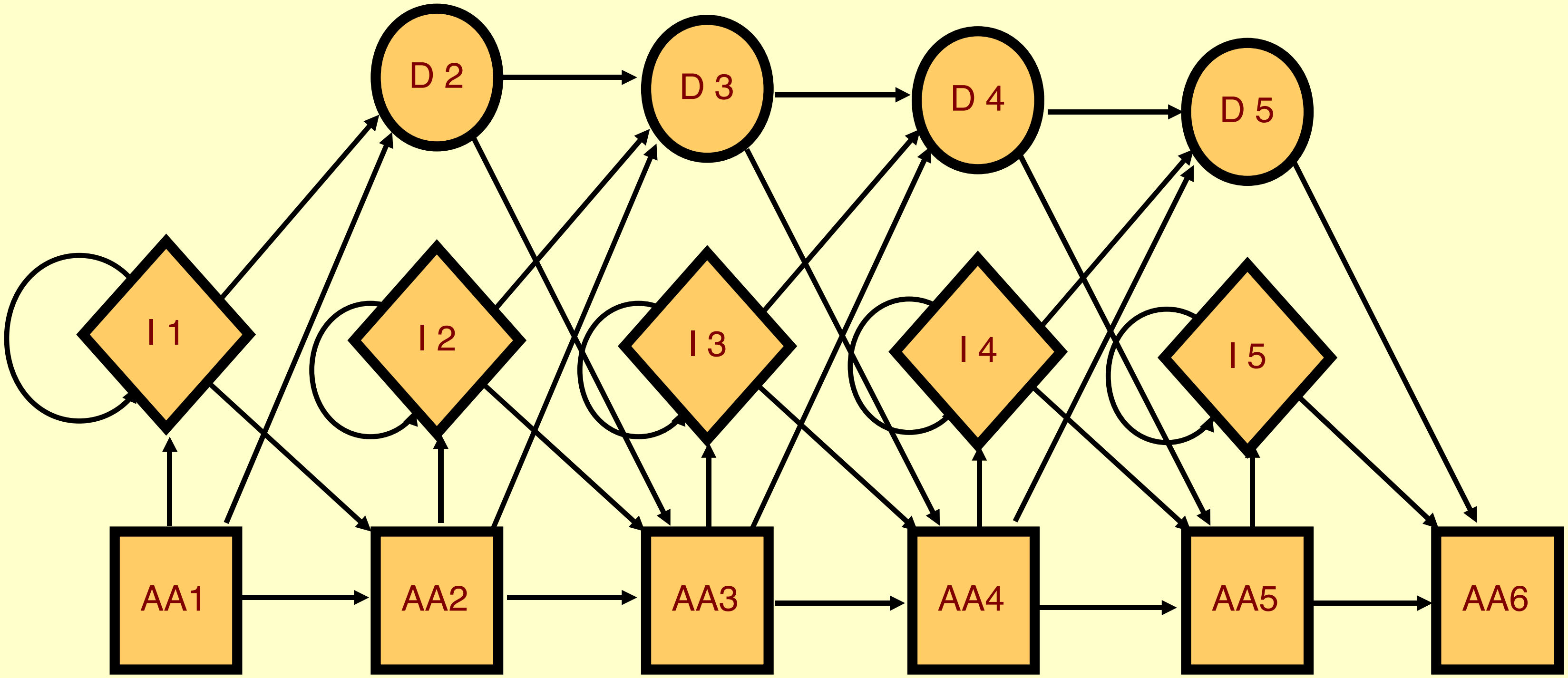


	Probe									Profile																					
Position	8 - 37	38 - 67	68 - 98	130 - 159	160 - 188	189 - 214	216 - 246	247 - 276	Consensus	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Gap Opening	Gap Extension
1	V	T	T	T	.	.	K	E	T	15	-3	12	13	-8	11	4	10	12	2	6	11	9	7	6	14	32	12	-22	-8	25	25
2	V	G	G	Q	.	.	E	E	E	16	-8	22	24	-14	22	8	4	8	-1	2	15	8	16	3	15	13	9	-30	-12	25	25
3	Y	E	E	Q	.	Q	R	Q	Q	11	-16	23	27	-16	7	22	0	15	-1	3	17	8	35	14	8	7	0	-19	-11	25	25
4	K	K	K	L	A	D	T	R	K	12	-16	13	13	-15	5	8	4	31	0	11	16	8	15	21	12	15	5	-8	-12	100	100
5	R	P	N	P	G	L	V	P	P	17	-6	9	8	-12	11	9	6	8	3	7	12	28	11	12	17	12	13	-21	-15	100	100
6	Y	F	F	Y	Y	A	Y	F	Y	-5	20	-23	-19	57	-21	10	20	-20	29	10	-3	-24	-20	-18	-5	-5	8	38	58	100	100
7	I	P	T	E	P	V	L	V	V	15	-3	6	8	-2	8	4	20	3	14	14	5	17	6	4	12	16	25	-27	-8	100	100
8	C	C	C	C	C	C	C	C	C	32	142	-48	-49	-10	19	-7	29	-51	-73	-53	-25	9	-53	-27	78	28	30	-129	101	100	100
9	S	K	D	P	K	D	P	E	D	16	-13	25	22	-24	15	12	0	23	-8	1	19	21	20	16	20	15	2	-26	-18	100	100
10	F	E	S	H	K	.	R	H	H	6	-7	9	11	-4	2	21	3	11	2	3	13	6	13	15	11	7	2	-3	1	24	24
11	A	E	D	E	D	.	D	A	D	27	-15	40	37	-26	25	14	1	12	-8	-3	24	12	24	4	16	16	3	-42	-15	24	24
12	D	G	G	G	D	.	G	G	G	32	-3	43	30	-31	62	3	-7	6	-16	-9	26	14	17	-6	29	22	10	-53	-28	24	24
13	.	.	.	.	S	V	.	.	S	9	-1	5	5	0	6	4	9	5	5	6	7	5	4	5	12	8	10	-9	-1	24	24
14	C	C	C	C	C	C	C	C	C	32	142	-48	-49	-10	19	-7	29	-51	-73	-53	-25	9	-53	-27	78	28	30	-129	101	100	100
15	G	E	D	D	.	N	D	G	D	19	-13	40	31	-25	31	12	-1	11	-10	-5	27	9	21	3	18	15	4	-38	-16	28	28
16	A	K	L	K	.	R	R	K	K	9	-17	9	9	-15	0	11	2	35	-1	13	14	8	16	30	11	10	3	5	-14	28	28
17	A	G	R	R	S	K	S	C	S	15	3	9	8	-13	12	7	2	16	-7	2	13	13	8	20	27	12	4	-1	-10	100	100
18	Y	F	F	F	F	F	Y	F	F	-19	1	-46	-34	83	-33	3	41	-35	62	27	-19	-36	-36	-25	-10	-10	16	63	81	100	100
19	N	T	T	S	V	R	T	A	T	19	1	11	10	-9	14	4	10	11	1	5	15	12	6	8	19	33	12	-17	-9	100	100
20	K	S	T	L	G	H	T	M	T	13	-5	9	9	-3	10	7	9	11	7	10	11	8	8	7	15	22	10	-14	-6	100	100
23	N	L	K	P	K	K	A	K	K	13	-20	13	13	-20	5	9	1	42	-2	13	21	11	18	25	14	13	3	-10	-15	100	100
24	W	H	A	S	T	D	F	K	S	10	-7	7	6	-1	5	10	4	9	4	3	11	4	7	11	15	11	3	-6	2	100	100
25	K	H	N	R	W	Y	N	S	N	5	-8	9	7	-2	2	16	0	14	-1	1	23	3	9	17	15	6	-1	2	3	100	100
26	.	.	.	.	T	.	.	.	T	8	-3	5	5	-1	5	5	6	5	4	5	8	4	5	5	9	11	7	-9	0	25	25
27	.	.	.	.	L	.	.	.	L	7	-6	3	4	3	2	5	8	4	9	8	6	3	5	5	7	6	8	-7	1	25	25
28	.	.	.	.	Y	.	.	.	S	6	-1	3	3	4	2	7	6	3	5	5	7	2	4	4	7	6	6	-5	5	25	25
29	L	L	M	L	L	L	L	L	L	-1	-58	-32	-17	77	-28	-10	59	-17	107	92	-19	-16	-4	-19	-20	-3	61	23	18	100	100
30	Q	T	K	K	K	R	R	E	K	7	-18	15	16	-22	3	14	-1	39	-6	9	18	9	21	33	13	14	0	0	-18	100	100
31	A	R	K	R	H	D	S	R	R	9	-10	12	11	-16	3	18	-1	24	-6	5	15	12	17	33	15	8	0	10	-13	100	100
32	H	H	H	H	V	H	H	H	H	0	-12	28	28	-7	-7	104	-13	11	-8	-15	37	16	50	35	-4	-2	-11	-15	17	100	100
33	.	.	F	.	.	.	I	.	I	6	-3	1	2	6	1	5	12	2	10	8	5	1	3	3	7	6	10	-6	5	30	30
34	L	S	N	E	A	Q	Q	S	S	17	-7	17	18	-11	13	12	4	11	1	4	18	11	21	8	22	11	5	-15	-10	100	100
35	C	L	R	K	E	K	S	V	K	9	-7	7	9	-8	5	7	8	19	2	10	11	6	9	15	15	11	10	-10	-5	100	100
36	K	T	F	V	C	T	F	V	V	9	2	-4	-2	10	3	1	23	1	17	13	2	1	-4	-1	10	20	23	-12	10	100	100
37	H	H	H	H	H	H	H	H	H	-4	-17	40	39	-12	-13	152	-27	15	-17	-27	53	21	73	50	-9	-7	-26	-14	25	100	100
38									*	11	19	13	14	14	12	21	2	25	16	2	7	8	7	16	13	15	11	2	9		

Doug Brutlag 2010

Hidden Markov Models (after Haussler)

http://www.cse.ucsc.edu/research/compbio/HMM-apps/HMM-applica

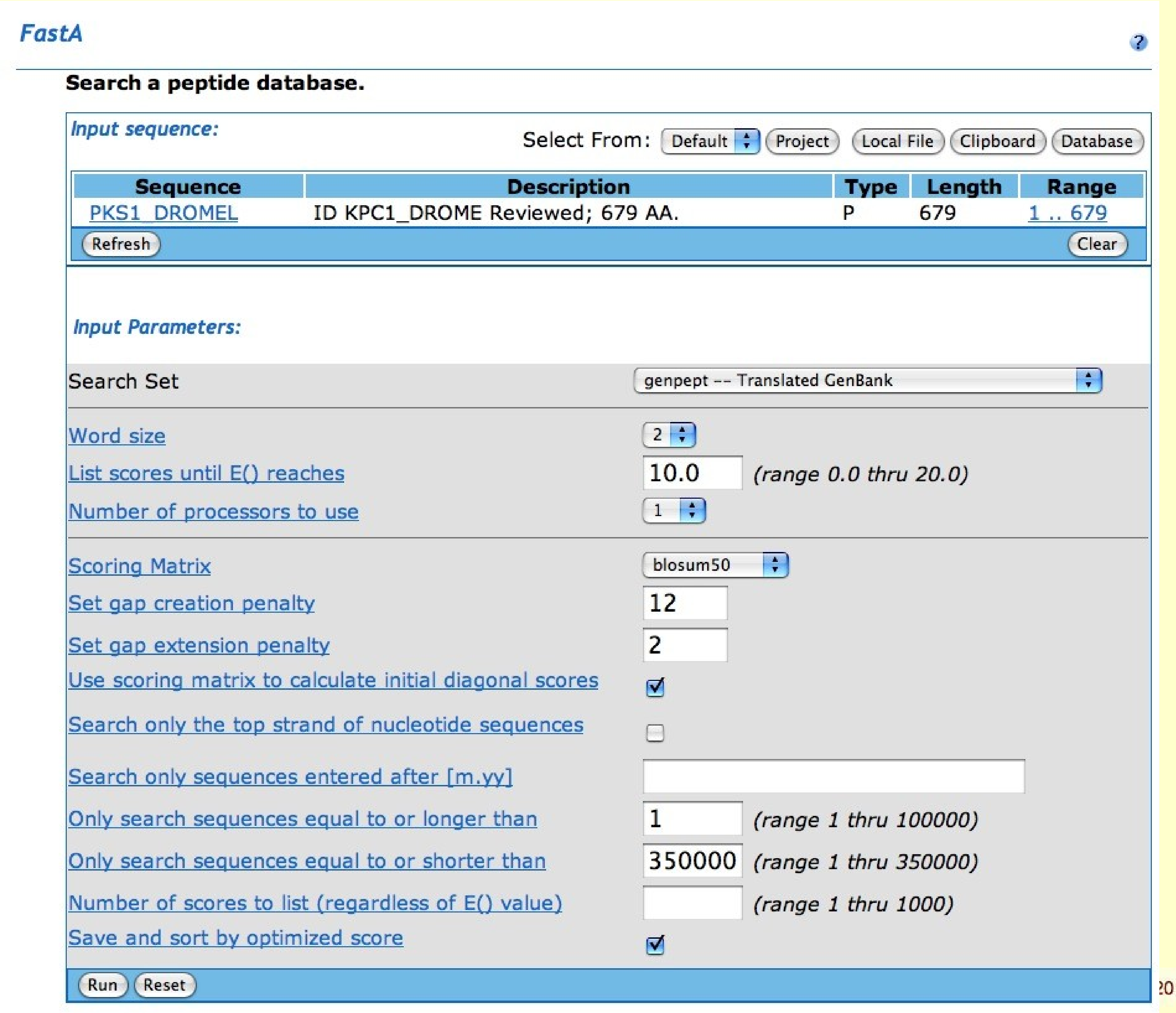


SeqWeb v3.1

accelrys®

FastA Protein

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



FastA

Search a peptide database.

Input sequence:

Select From: Default Project Local File Clipboard Database

Sequence	Description	Type	Length	Range
PKS1_DROMEL	ID KPC1_DROME Reviewed; 679 AA.	P	679	1 .. 679

Refresh Clear

Input Parameters:

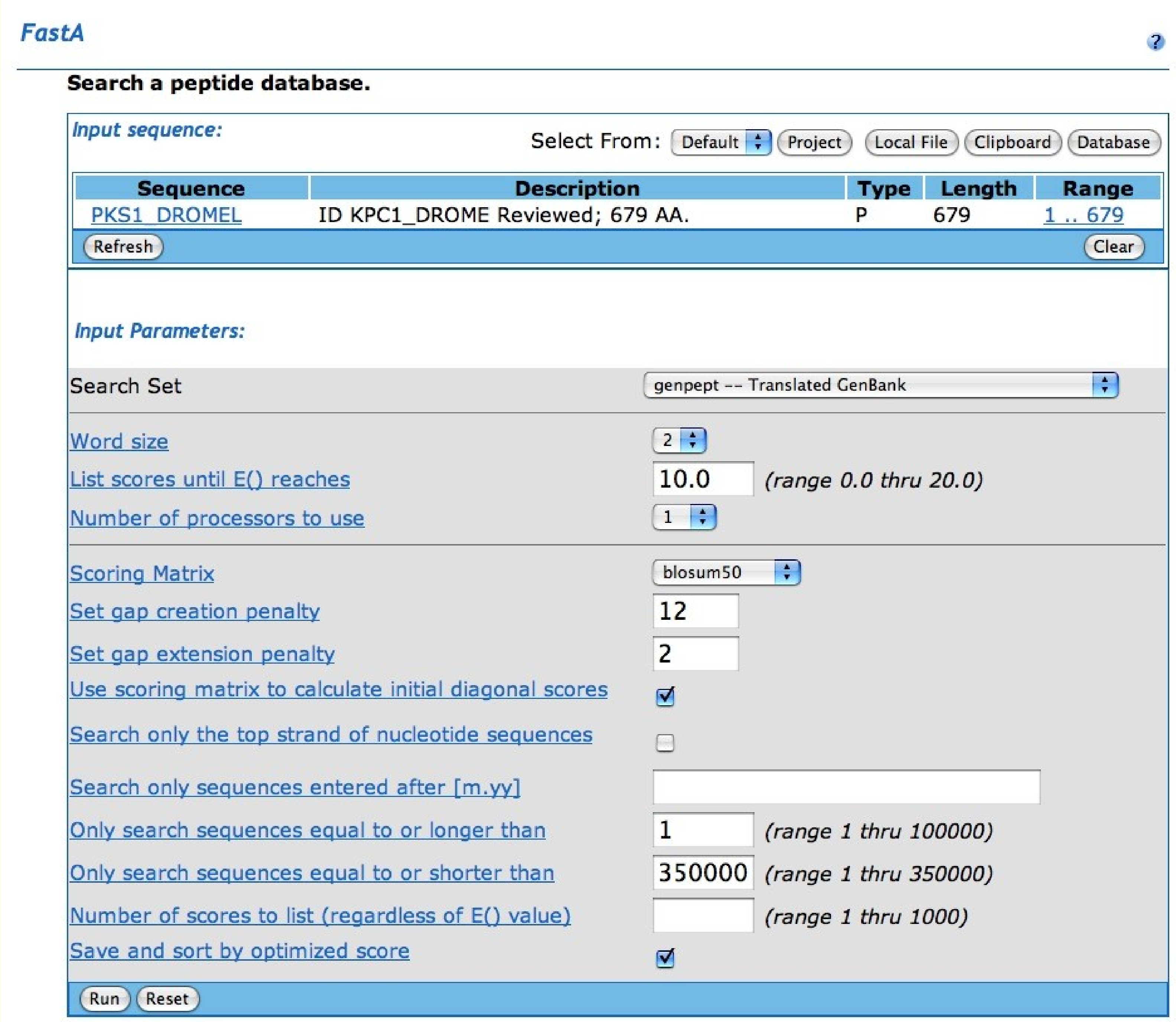
Search Set	genpept -- Translated GenBank	
Word size	2	
List scores until E() reaches	10.0	*(range 0.0 thru 20.0)*
Number of processors to use	1	
Scoring Matrix	blosum50	
Set gap creation penalty	12	
Set gap extension penalty	2	
Use scoring matrix to calculate initial diagonal scores	☑	
Search only the top strand of nucleotide sequences	☐	
Search only sequences entered after [m.yy]		
Only search sequences equal to or longer than	1	*(range 1 thru 100000)*
Only search sequences equal to or shorter than	350000	*(range 1 thru 350000)*
Number of scores to list (regardless of E() value)		*(range 1 thru 1000)*
Save and sort by optimized score	☑	

Run Reset

2010

SeqWeb FASTA

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



FastA

Search a peptide database.

Input sequence:

Select From: Default | Project | Local File | Clipboard | Database

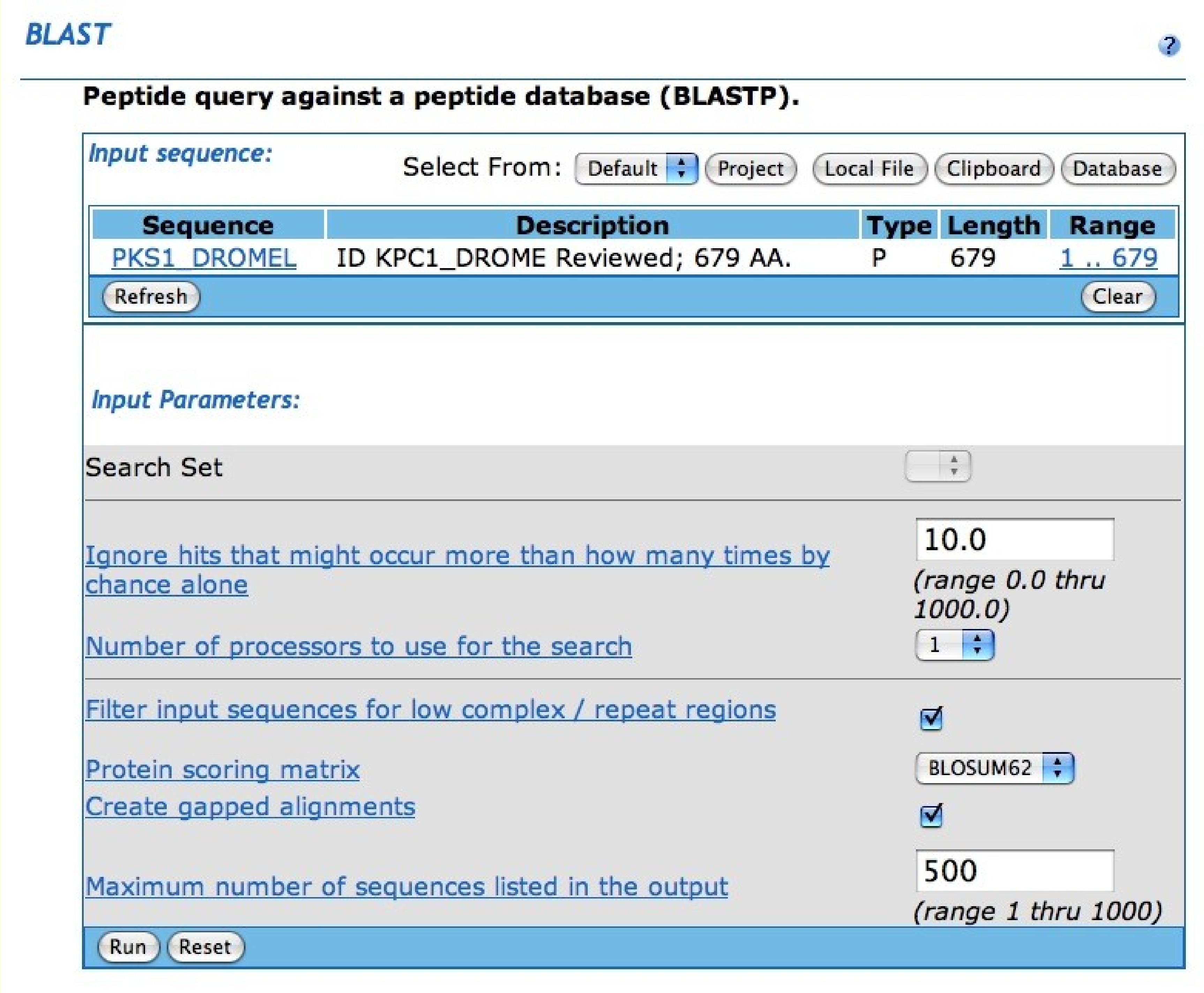
Sequence	Description	Type	Length	Range
PKS1_DROMEL	ID KPC1_DROME Reviewed; 679 AA.	P	679	1 .. 679

Input Parameters:

Search Set	genpept -- Translated GenBank	
Word size	2	
List scores until E() reaches	10.0	*(range 0.0 thru 20.0)*
Number of processors to use	1	
Scoring Matrix	blosum50	
Set gap creation penalty	12	
Set gap extension penalty	2	
Use scoring matrix to calculate initial diagonal scores	☑	
Search only the top strand of nucleotide sequences	□	
Search only sequences entered after [m.yy]		
Only search sequences equal to or longer than	1	*(range 1 thru 100000)*
Only search sequences equal to or shorter than	350000	*(range 1 thru 350000)*
Number of scores to list (regardless of E() value)		*(range 1 thru 1000)*
Save and sort by optimized score	☑	

SeqWeb BLAST Protein

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



BLAST

Peptide query against a peptide database (BLASTP).

Input sequence:

Select From:

Sequence	Description	Type	Length	Range
PKS1_DROMEL	ID KPC1_DROME Reviewed; 679 AA.	P	679	1 .. 679

Input Parameters:

Search Set

Ignore hits that might occur more than how many times by chance alone: 10.0 *(range 0.0 thru 1000.0)*

Number of processors to use for the search: 1

Filter input sequences for low complex / repeat regions ☑

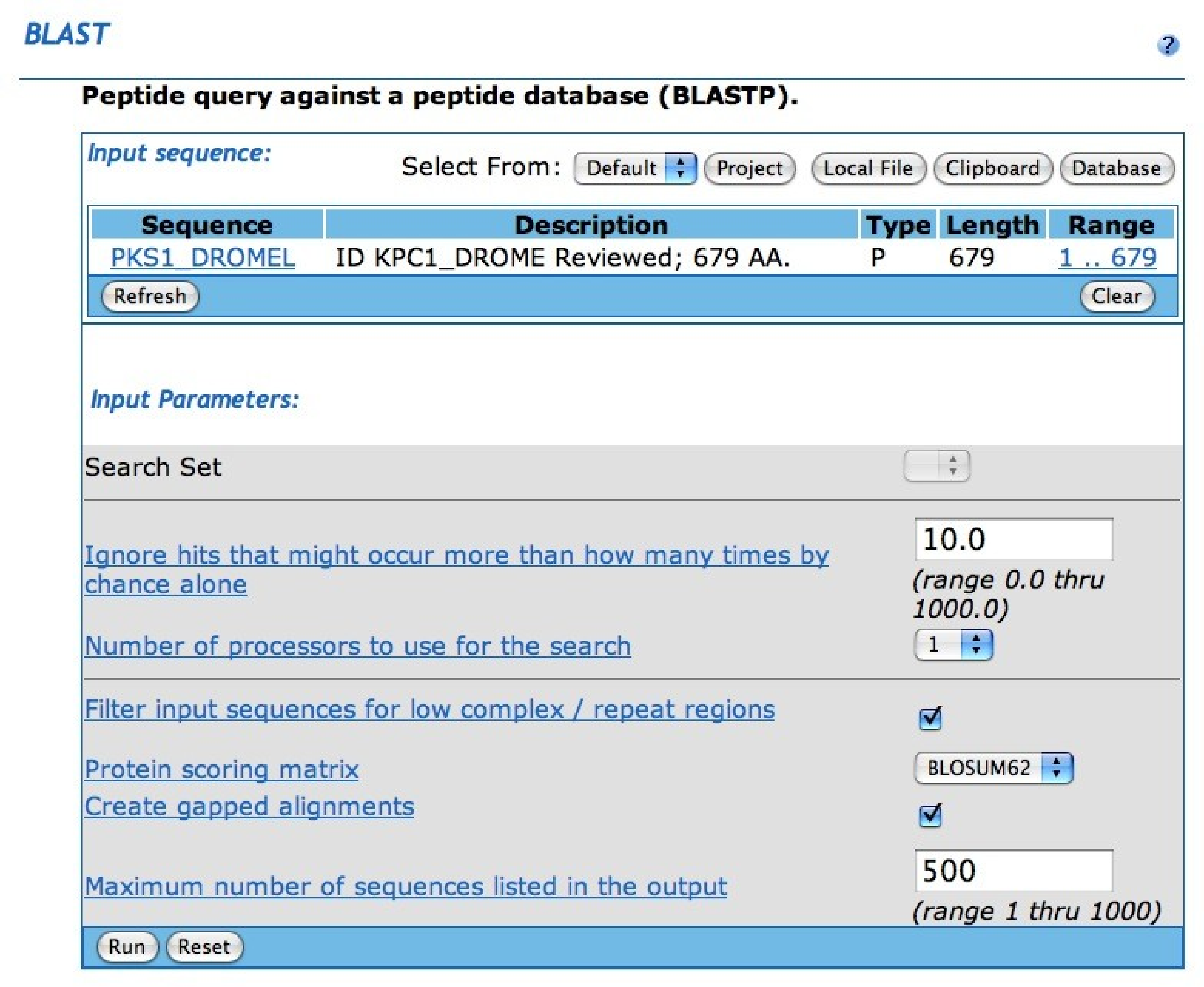
Protein scoring matrix: BLOSUM62

Create gapped alignments ☑

Maximum number of sequences listed in the output: 500 *(range 1 thru 1000)*

SeqWeb BLASTP

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



BLAST

Peptide query against a peptide database (BLASTP).

Input sequence:

Sequence	Description	Type	Length	Range
PKS1_DROMEL	ID KPC1_DROME Reviewed; 679 AA.	P	679	1 .. 679

Input Parameters:

Search Set

Ignore hits that might occur more than how many times by chance alone: 10.0 *(range 0.0 thru 1000.0)*

Number of processors to use for the search: 1

Filter input sequences for low complex / repeat regions ☑

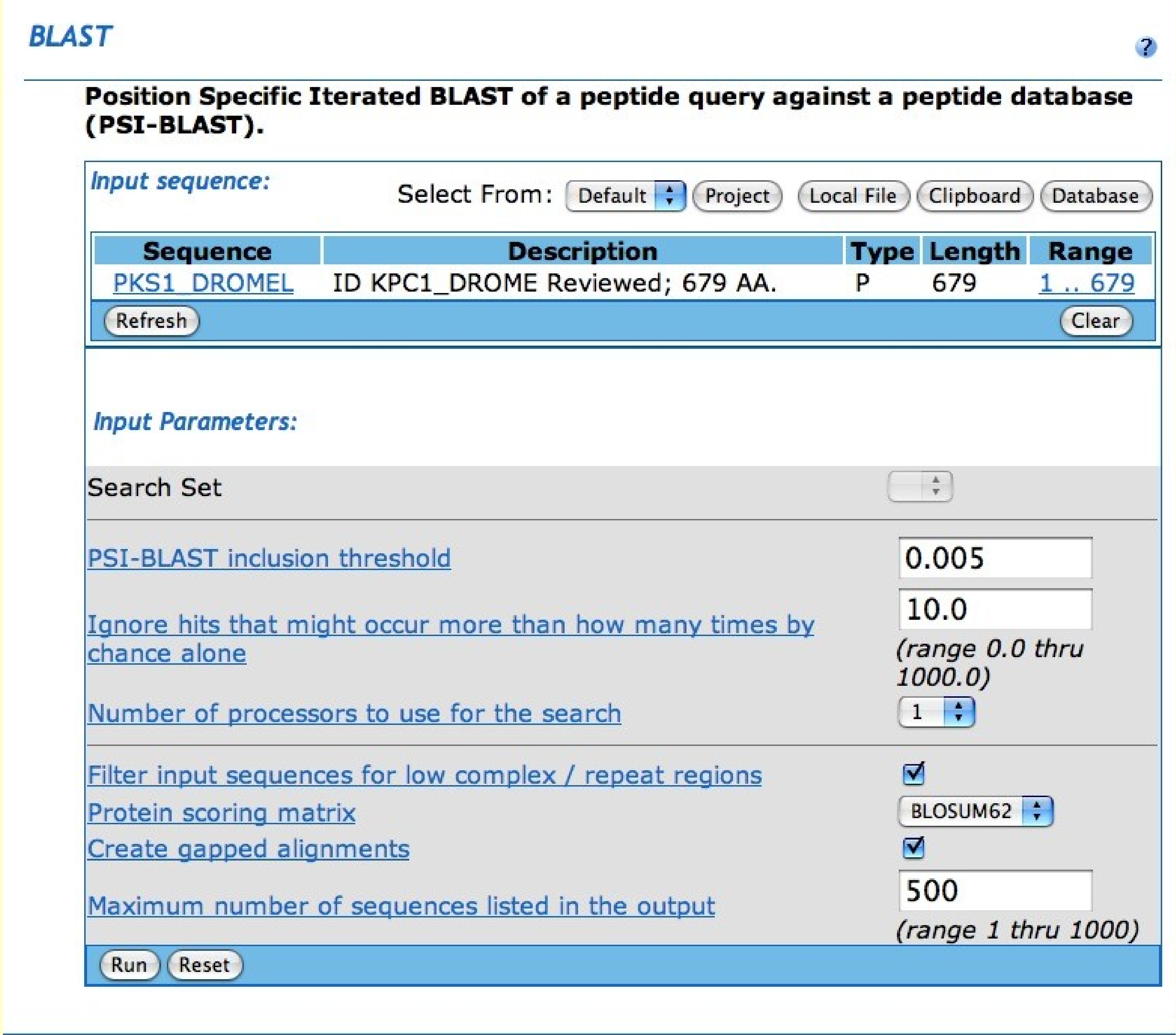
Protein scoring matrix: BLOSUM62

Create gapped alignments ☑

Maximum number of sequences listed in the output: 500 *(range 1 thru 1000)*

SeqWeb PSI-BLAST Protein

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



BLAST

Position Specific Iterated BLAST of a peptide query against a peptide database (PSI-BLAST).

Input sequence:

Select From: Default | Project | Local File | Clipboard | Database

Sequence	Description	Type	Length	Range
PKS1_DROMEL	ID KPC1_DROME Reviewed; 679 AA.	P	679	1 .. 679

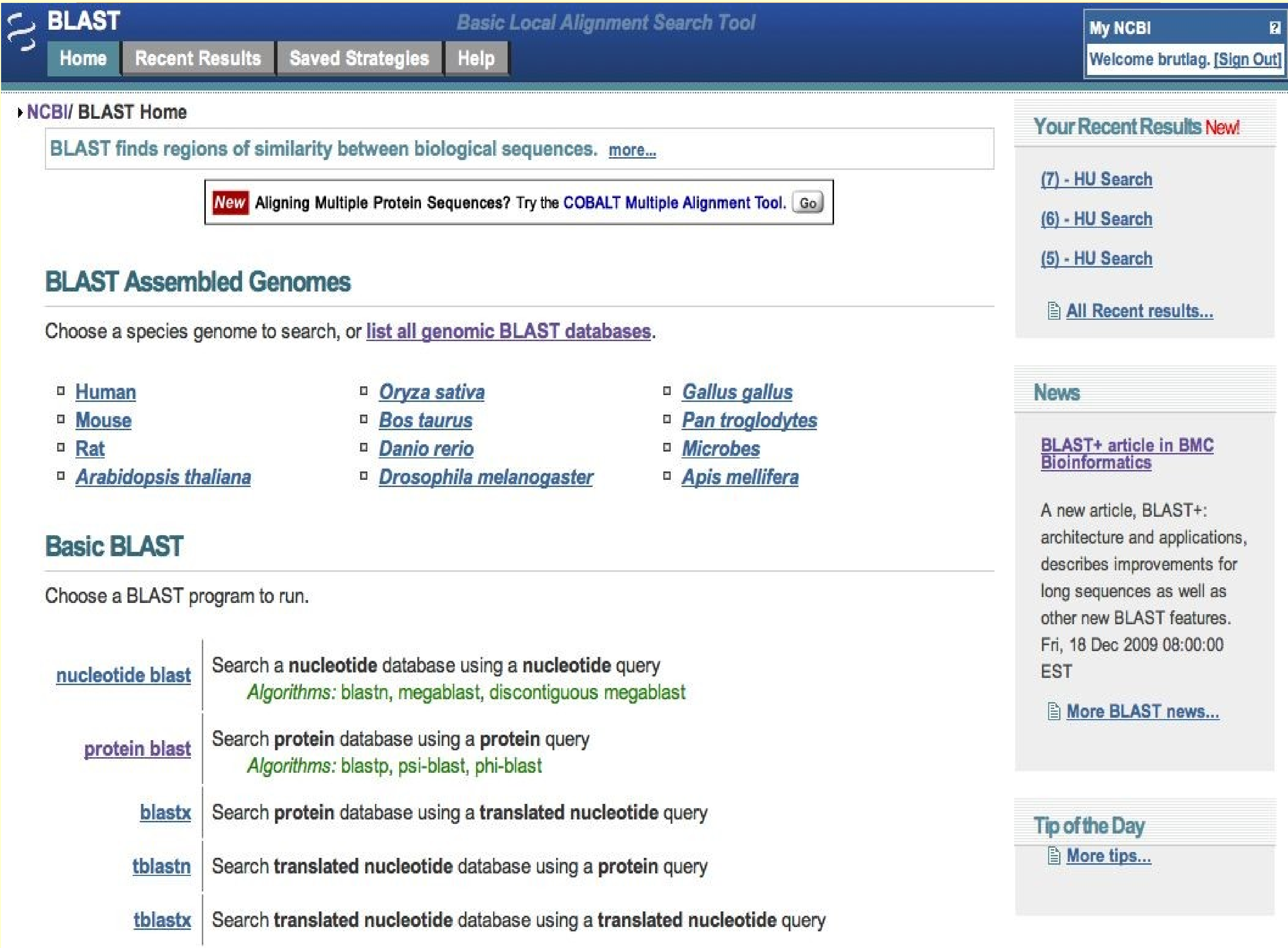
Refresh | Clear

Input Parameters:

Search Set	
PSI-BLAST inclusion threshold	0.005
Ignore hits that might occur more than how many times by chance alone	10.0 *(range 0.0 thru 1000.0)*
Number of processors to use for the search	1
Filter input sequences for low complex / repeat regions	☑
Protein scoring matrix	BLOSUM62
Create gapped alignments	☑
Maximum number of sequences listed in the output	500 *(range 1 thru 1000)*

NCBI BLAST Home Page

http://blast.ncbi.nlm.nih.gov/



BLAST finds regions of similarity between biological sequences. more...

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool.

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Your Recent Results New!

(7) - HU Search

(6) - HU Search

(5) - HU Search

All Recent results...

News

BLAST+ article in BMC Bioinformatics

A new article, BLAST+: architecture and applications, describes improvements for long sequences as well as other new BLAST features.
Fri, 18 Dec 2009 08:00:00 EST

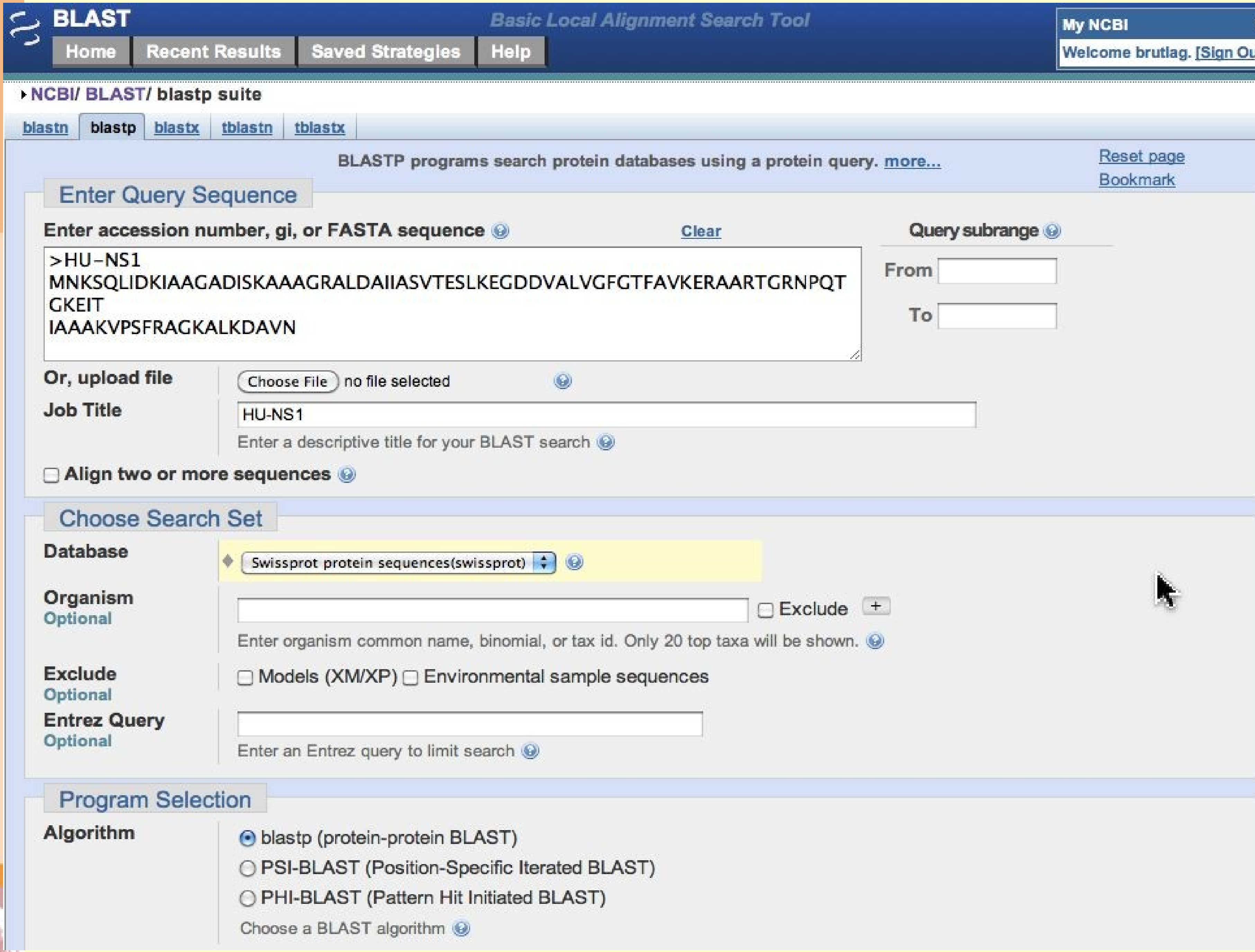
More BLAST news...

Tip of the Day

More tips...

NCBI BLAST Input

http://blast.ncbi.nlm.nih.gov/



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>HU-NS1
MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTGRNPQT
GKEIT
IAAAKVPSFRAGKALKDAVN
```

Query subrange

From

To

Or, upload file: no file selected

Job Title: HU-NS1

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Swissprot protein sequences(swissprot)

Organism Optional: □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

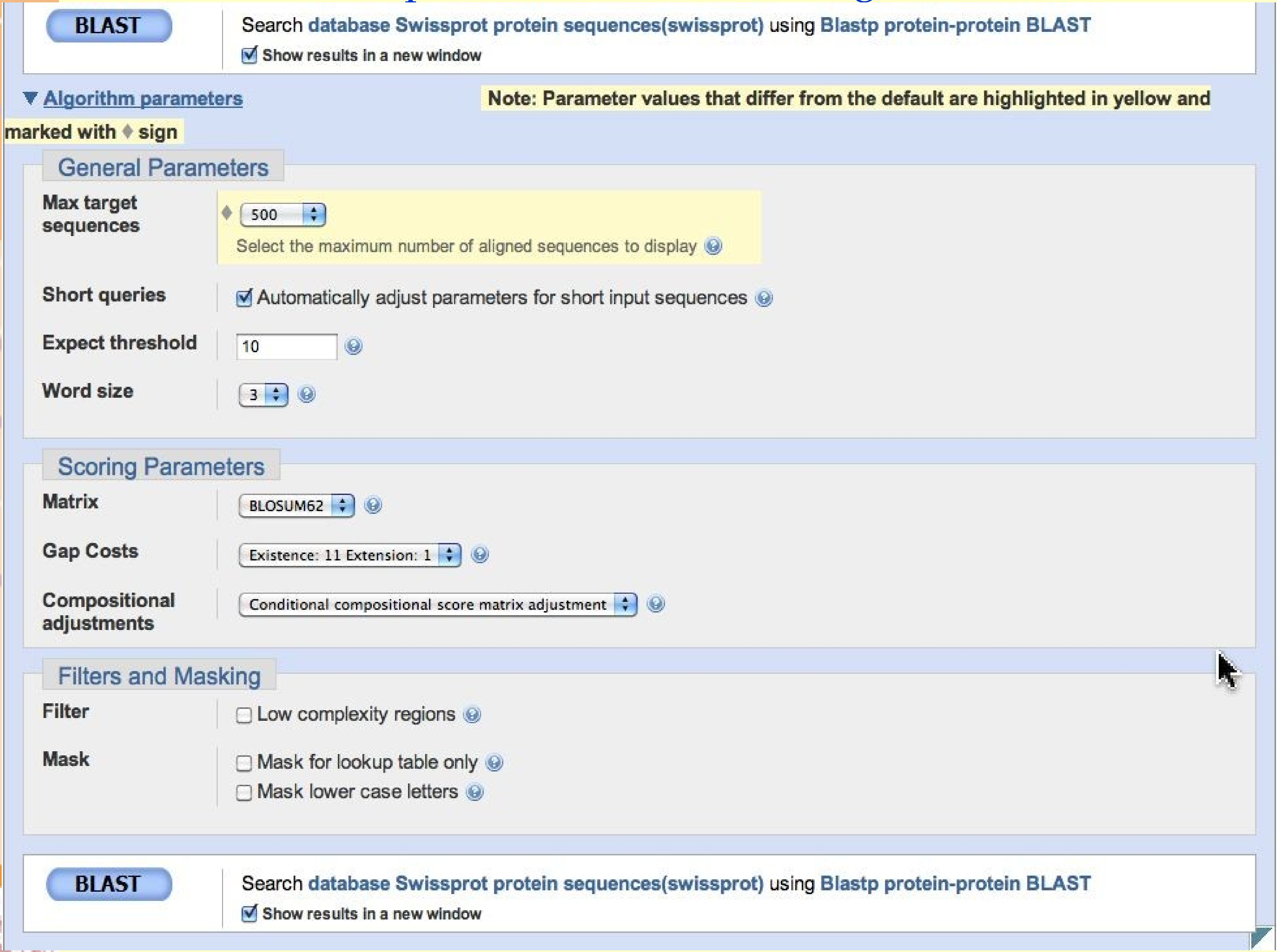
- ● blastp (protein-protein BLAST)
- ○ PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

Doug Brutlag 2010

NCBI BLAST Parameters

http://blast.ncbi.nlm.nih.gov/



Search database Swissprot protein sequences(swissprot) using Blastp protein-protein BLAST

☑ Show results in a new window

▼ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

General Parameters

Max target sequences: ♦ 500

Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: □ Low complexity regions

Mask: □ Mask for lookup table only

□ Mask lower case letters

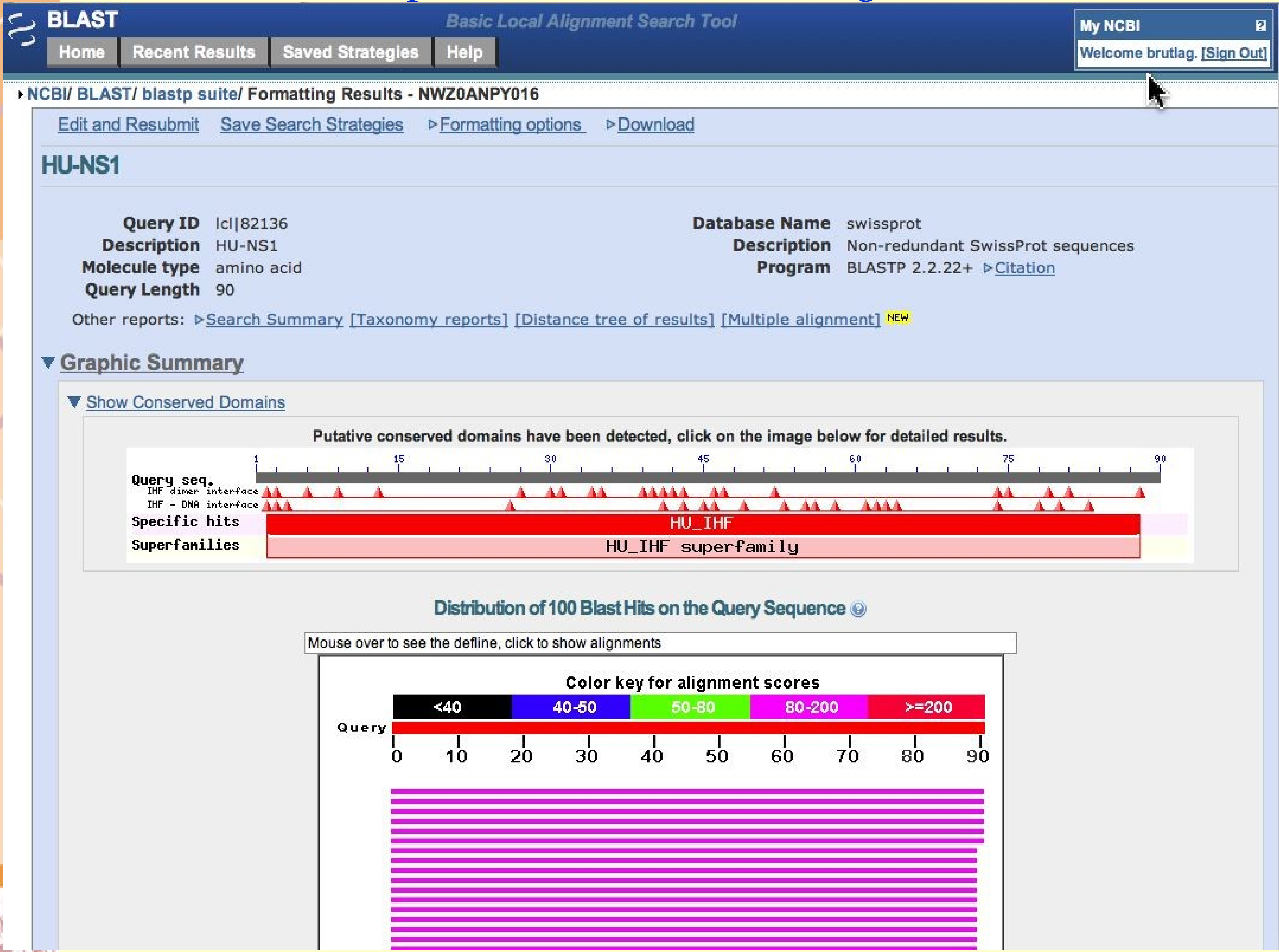
Search database Swissprot protein sequences(swissprot) using Blastp protein-protein BLAST

☑ Show results in a new window

Doug Brutlag 2010

NCBI BLAST Conserved Domains

http://blast.ncbi.nlm.nih.gov/


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
Welcome brutlag. [Sign Out]

NCBI/ BLAST/ blastp suite/ Formatting Results - NWZ0ANPY016

HU-NS1

Query ID lcl|82136
Description HU-NS1
Molecule type amino acid
Query Length 90

Database Name swissprot
Description Non-redundant SwissProt sequences
Program BLASTP 2.2.22+ Citation

Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] NEW

Graphic Summary

Show Conserved Domains

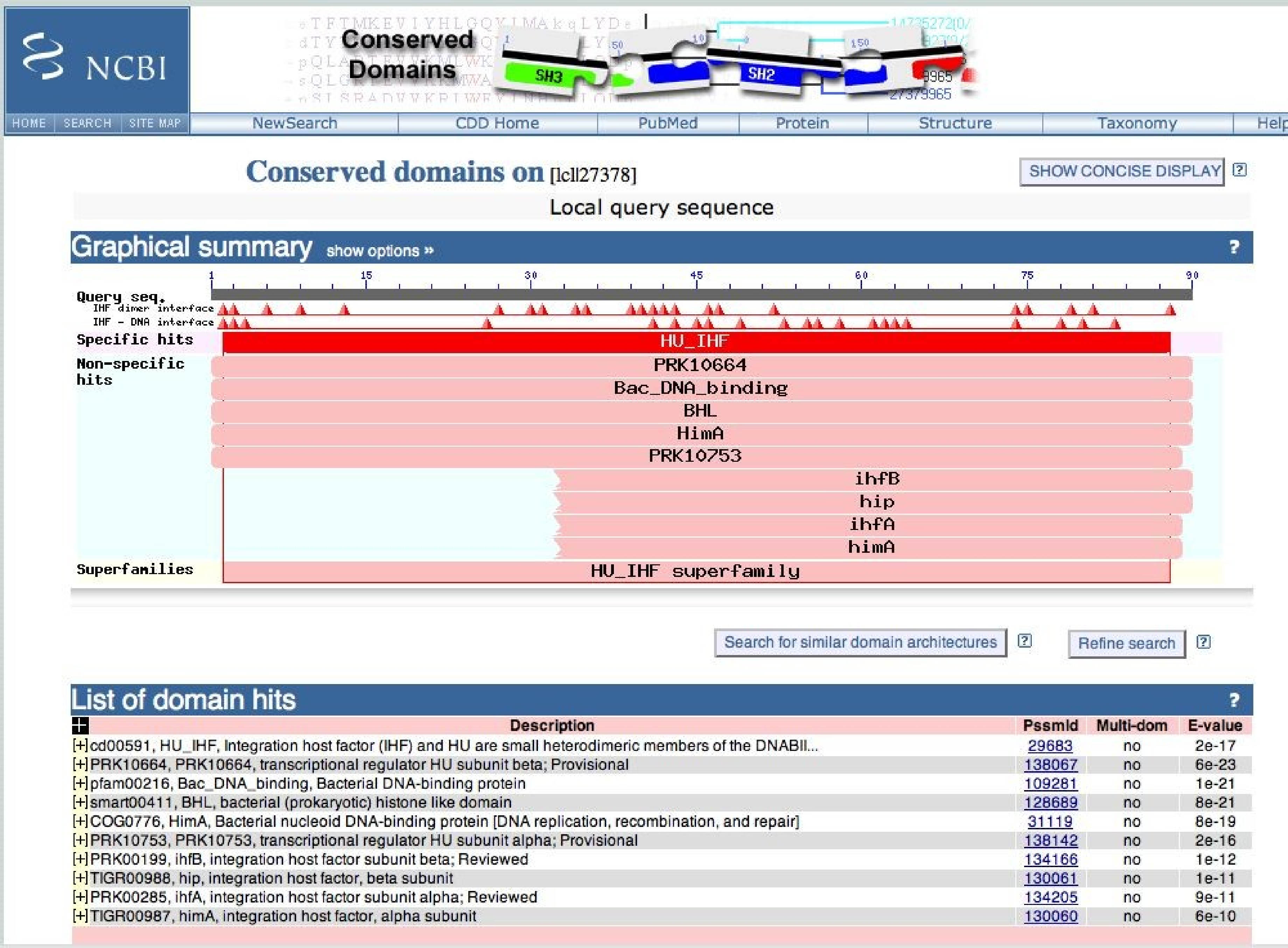
Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

NCBI BLAST Conserved Domains

http://blast.ncbi.nlm.nih.gov/



Conserved domains on [lcl27378]

Local query sequence

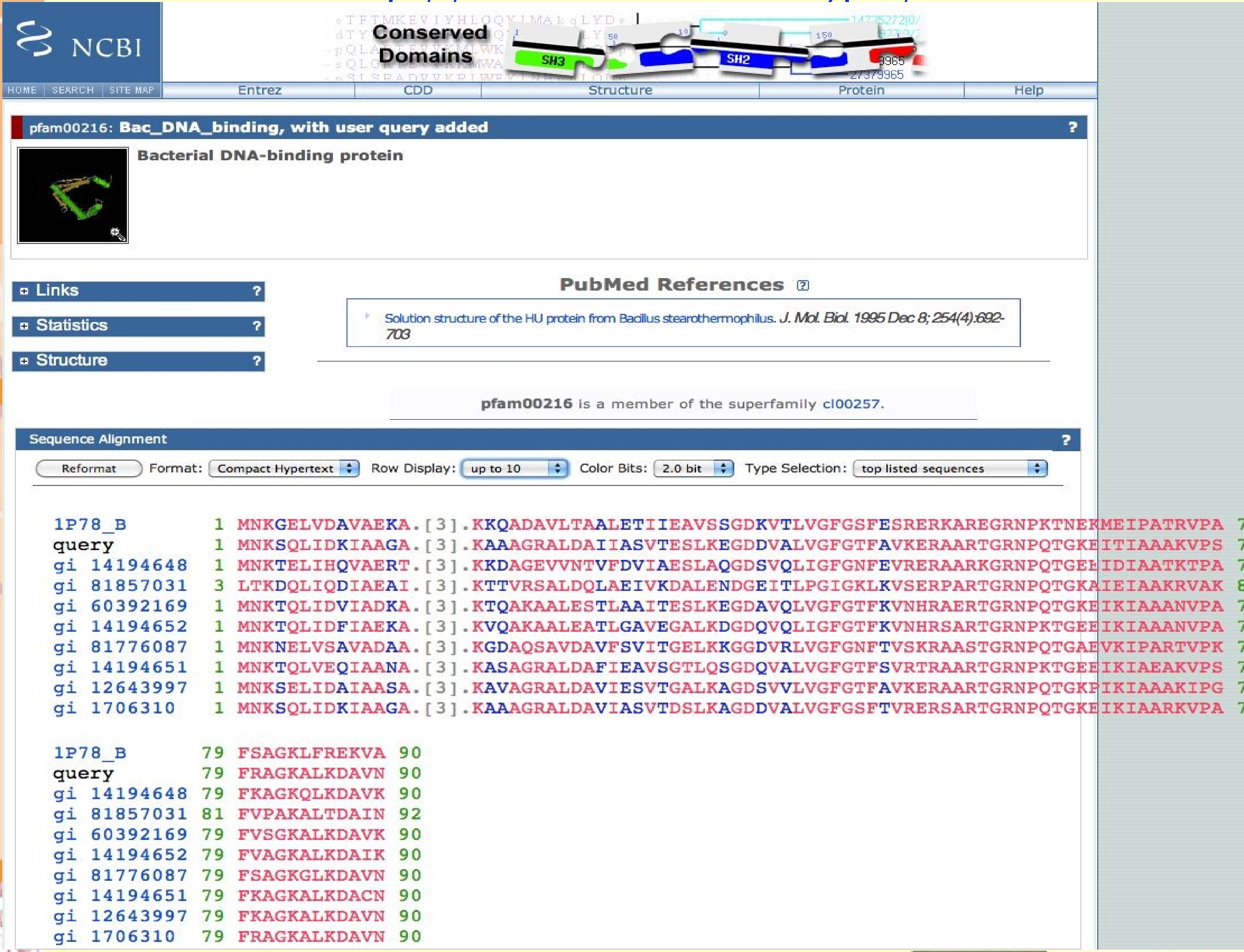
Graphical summary

List of domain hits

Description	PssmId	Multi-dom	E-value
[+]cd00591, HU_IHF, Integration host factor (IHF) and HU are small heterodimeric members of the DNABII...	29683	no	2e-17
[+]PRK10664, PRK10664, transcriptional regulator HU subunit beta; Provisional	138067	no	6e-23
[+]pfam00216, Bac_DNA_binding, Bacterial DNA-binding protein	109281	no	1e-21
[+]smart00411, BHL, bacterial (prokaryotic) histone like domain	128689	no	8e-21
[+]COG0776, HimA, Bacterial nucleoid DNA-binding protein [DNA replication, recombination, and repair]	31119	no	8e-19
[+]PRK10753, PRK10753, transcriptional regulator HU subunit alpha; Provisional	138142	no	2e-16
[+]PRK00199, ihfB, integration host factor subunit beta; Reviewed	134166	no	1e-12
[+]TIGR00988, hip, integration host factor, beta subunit	130061	no	1e-11
[+]PRK00285, ihfA, integration host factor subunit alpha; Reviewed	134205	no	9e-11
[+]TIGR00987, himA, integration host factor, alpha subunit	130060	no	6e-10

Bacterial DNA-Binding Protein

http://blast.ncbi.nlm.nih.gov/



pfam00216: Bac_DNA_binding, with user query added

Bacterial DNA-binding protein

PubMed References

Solution structure of the HU protein from Bacillus stearothermophilus. *J. Mol. Biol. 1995 Dec 8; 254(4):692-703*

pfam00216 is a member of the superfamily cl00257.

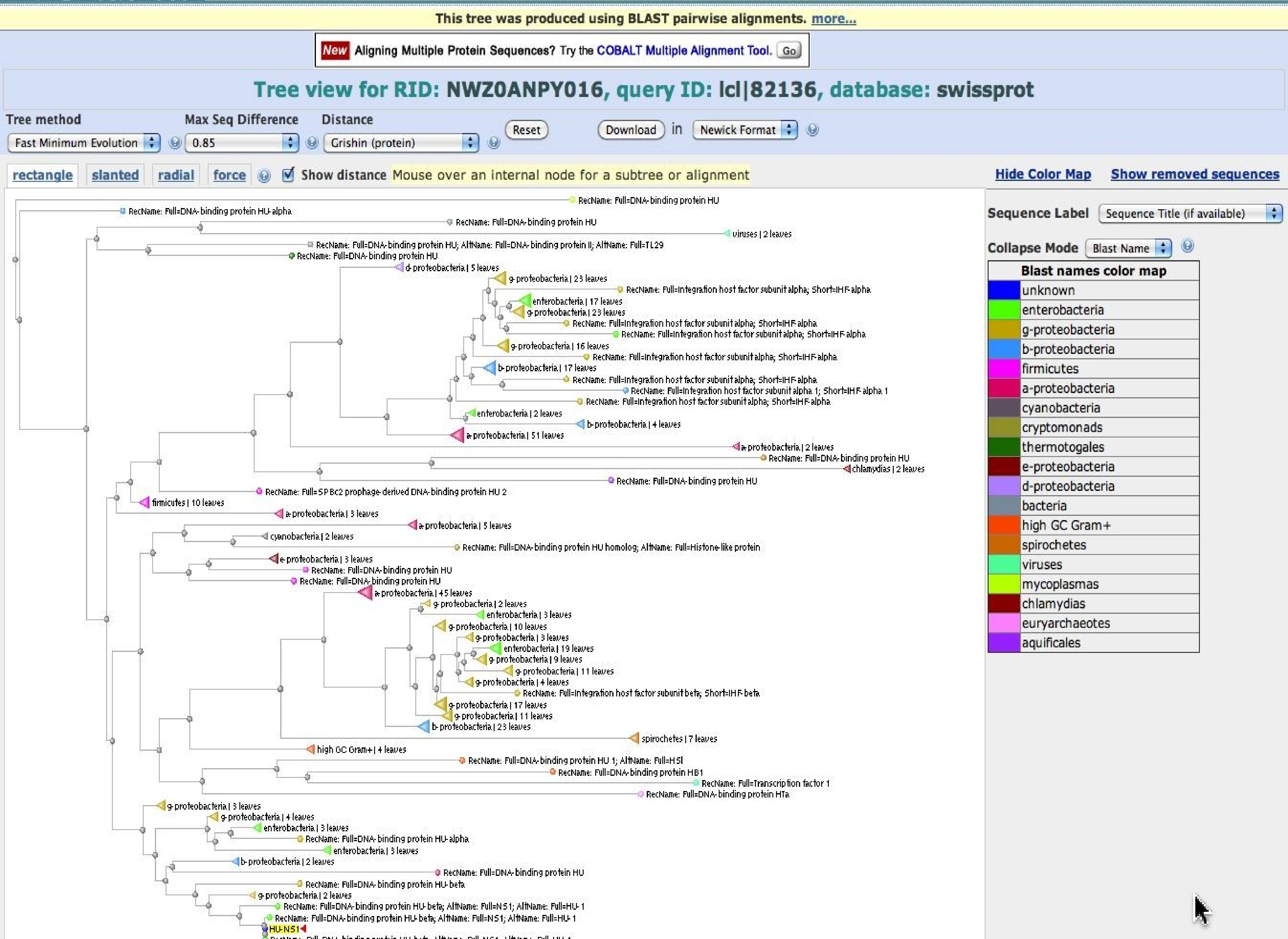
Sequence Alignment

Format: Compact Hypertext Row Display: up to 10 Color Bits: 2.0 bit Type Selection: top listed sequences

```
1P78_B        1 MNKGELVDAVAEKA.[3].KKQADAVLTAALETIIEAVSSGDKVTLVGFGSFESRERKAREGRNPKTNEKMEIPATRVPA 7
query         1 MNKSQLIDKIAAGA.[3].KAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTGRNPQTGKEITIAAAKVPS 7
gi 14194648   1 MNKTELIHQVAERT.[3].KKDAGEVVNTVFDVIAESLAQGDSVQLIGFGNFEVRERAARKGRNPQTGEHIDIAATKTPA 7
gi 81857031   3 LTKDQLIQDIAEAI.[3].KTTVRSALDQLAEIVKDALENDGEITLPGIGKLKVSERPARTGRNPQTGKAIEIAAKRVAK 8
gi 60392169   1 MNKTQLIDVIADKA.[3].KTQAKAALESTLAAITESLKEGDAVQLVGFGTFKVNHRAERTGRNPQTGKEIKIAAANVPA 7
gi 14194652   1 MNKTQLIDFIAEKA.[3].KVQAKAALEATLGAVEGALKDGDQVQLIGFGTFKVNHRSARTGRNPKTGEEIKIAAANVPA 7
gi 81776087   1 MNKNELVSAVADAA.[3].KGDAQSAVDAVFSVITGELKKGGDVRLVGFGNFTVSKRAASTGRNPQTGAEVKIPARTVPK 7
gi 14194651   1 MNKTQLVEQIAANA.[3].KASAGRALDAFIEAVSGTLQSGDQVALVGFGTFSVRTRAARTGRNPKTGEEIKIAEAKVPS 7
gi 12643997   1 MNKSELIDAIAASA.[3].KAVAGRALDAVIESVTGALKAGDSVVLVGFGTFAVKERAARTGRNPQTGKPIKIAAAKIPG 7
gi 1706310    1 MNKSQLIDKIAAGA.[3].KAAAGRALDAVIASVTDSLKAGDDVALVGFGSFTVRERSARTGRNPQTGKEIKIAARKVPA 7

1P78_B       79 FSAGKLFREKVA 90
query        79 FRAGKALKDAVN 90
gi 14194648  79 FKAGKQLKDAVK 90
gi 81857031  81 FVPAKALTDAIN 92
gi 60392169  79 FVSGKALKDAVK 90
gi 14194652  79 FVAGKALKDAIK 90
gi 81776087  79 FSAGKGLKDAVN 90
gi 14194651  79 FKAGKALKDACN 90
gi 12643997  79 FKAGKALKDAVN 90
gi 1706310   79 FRAGKALKDAVN 90
```

Doug Brutlag 2010

NCBI Blast Distance Tree



This tree was produced using BLAST pairwise alignments. more...

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. Go

Tree view for RID: NWZ0ANPY016, query ID: lcl|82136, database: swissprot

Tree method: Fast Minimum Evolution

Max Seq Difference: 0.85

Distance: Grishin (protein)

Reset

Download in Newick Format

rectangle slanted radial force ☑ Show distance Mouse over an internal node for a subtree or alignment

Hide Color Map Show removed sequences

Sequence Label: Sequence Title (if available)

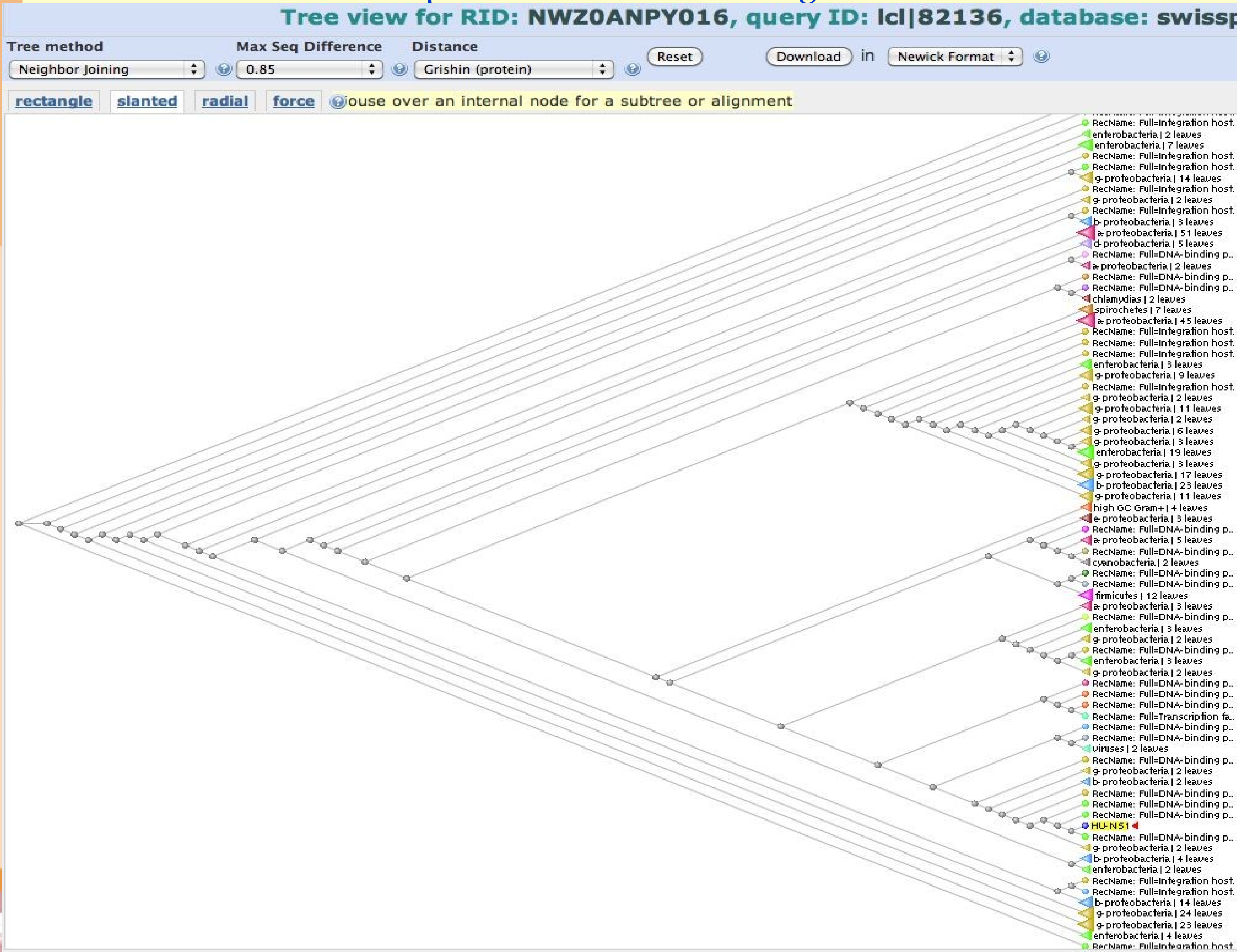
Collapse Mode: Blast Name

Blast names color map	
🟦	unknown
🟩	enterobacteria
🟨	g-proteobacteria
🟦	b-proteobacteria
🟪	firmicutes
🟥	a-proteobacteria
⬛	cyanobacteria
🟫	cryptomonads
🟩	thermotogales
🟥	e-proteobacteria
🟪	d-proteobacteria
⬜	bacteria
🟧	high GC Gram+
🟧	spirochetes
🟩	viruses
🟩	mycoplasmas
🟥	chlamydias
🟪	euryarchaeotes
🟪	aquificales

RecName: Full=DNA-binding protein HU
RecName: Full=DNA-binding protein HU-alpha
RecName: Full=DNA-binding protein HU
viruses | 2 leaves
RecName: Full=DNA-binding protein HU; AltName: Full=DNA-binding protein II; AltName: Full=TL29
RecName: Full=DNA-binding protein HU
d-proteobacteria | 5 leaves
g-proteobacteria | 23 leaves
RecName: Full=Integration host factor subunit alpha; Short=IHF-alpha
enterobacteria | 17 leaves
g-proteobacteria | 23 leaves
RecName: Full=Integration host factor subunit alpha; Short=IHF-alpha
RecName: Full=Integration host factor subunit alpha; Short=IHF-alpha
g-proteobacteria | 16 leaves
RecName: Full=Integration host factor subunit alpha; Short=IHF-alpha
b-proteobacteria | 17 leaves
RecName: Full=Integration host factor subunit alpha; Short=IHF-alpha
RecName: Full=Integration host factor subunit alpha 1; Short=IHF-alpha 1
RecName: Full=Integration host factor subunit alpha; Short=IHF-alpha
enterobacteria | 2 leaves
b-proteobacteria | 4 leaves
a-proteobacteria | 51 leaves
a-proteobacteria | 2 leaves
RecName: Full=DNA-binding protein HU
chlamydias | 2 leaves
RecName: Full=DNA-binding protein HU
RecName: Full=SPBc2 prophage-derived DNA-binding protein HU 2
firmicutes | 10 leaves
a-proteobacteria | 3 leaves
a-proteobacteria | 5 leaves
cyanobacteria | 2 leaves
RecName: Full=DNA-binding protein HU homolog; AltName: Full=Histone-like protein
e-proteobacteria | 3 leaves
RecName: Full=DNA-binding protein HU
RecName: Full=DNA-binding protein HU
a-proteobacteria | 45 leaves
g-proteobacteria | 2 leaves
enterobacteria | 3 leaves
g-proteobacteria | 10 leaves
g-proteobacteria | 3 leaves
enterobacteria | 19 leaves
g-proteobacteria | 9 leaves
g-proteobacteria | 11 leaves
g-proteobacteria | 4 leaves
RecName: Full=Integration host factor subunit beta; Short=IHF-beta
g-proteobacteria | 17 leaves
g-proteobacteria | 11 leaves
b-proteobacteria | 23 leaves
spirochetes | 7 leaves
high GC Gram+ | 4 leaves
RecName: Full=DNA-binding protein HU 1; AltName: Full=HSl
RecName: Full=DNA-binding protein HB1
RecName: Full=Transcription factor 1
RecName: Full=DNA-binding protein HTa
g-proteobacteria | 3 leaves
g-proteobacteria | 4 leaves
enterobacteria | 3 leaves
RecName: Full=DNA-binding protein HU-alpha
enterobacteria | 3 leaves
b-proteobacteria | 2 leaves
RecName: Full=DNA-binding protein HU
RecName: Full=DNA-binding protein HU-beta
g-proteobacteria | 2 leaves
RecName: Full=DNA-binding protein HU-beta; AltName: Full=NS1; AltName: Full=HU-1
RecName: Full=DNA-binding protein HU-beta; AltName: Full=NS1; AltName: Full=HU-1
HU-NS1

NCBI Blast Distance Tree

http://blast.ncbi.nlm.nih.gov/



Tree view for RID: NWZ0ANPY016, query ID: lcl|82136, database: swissp

Tree method: Neighbor Joining

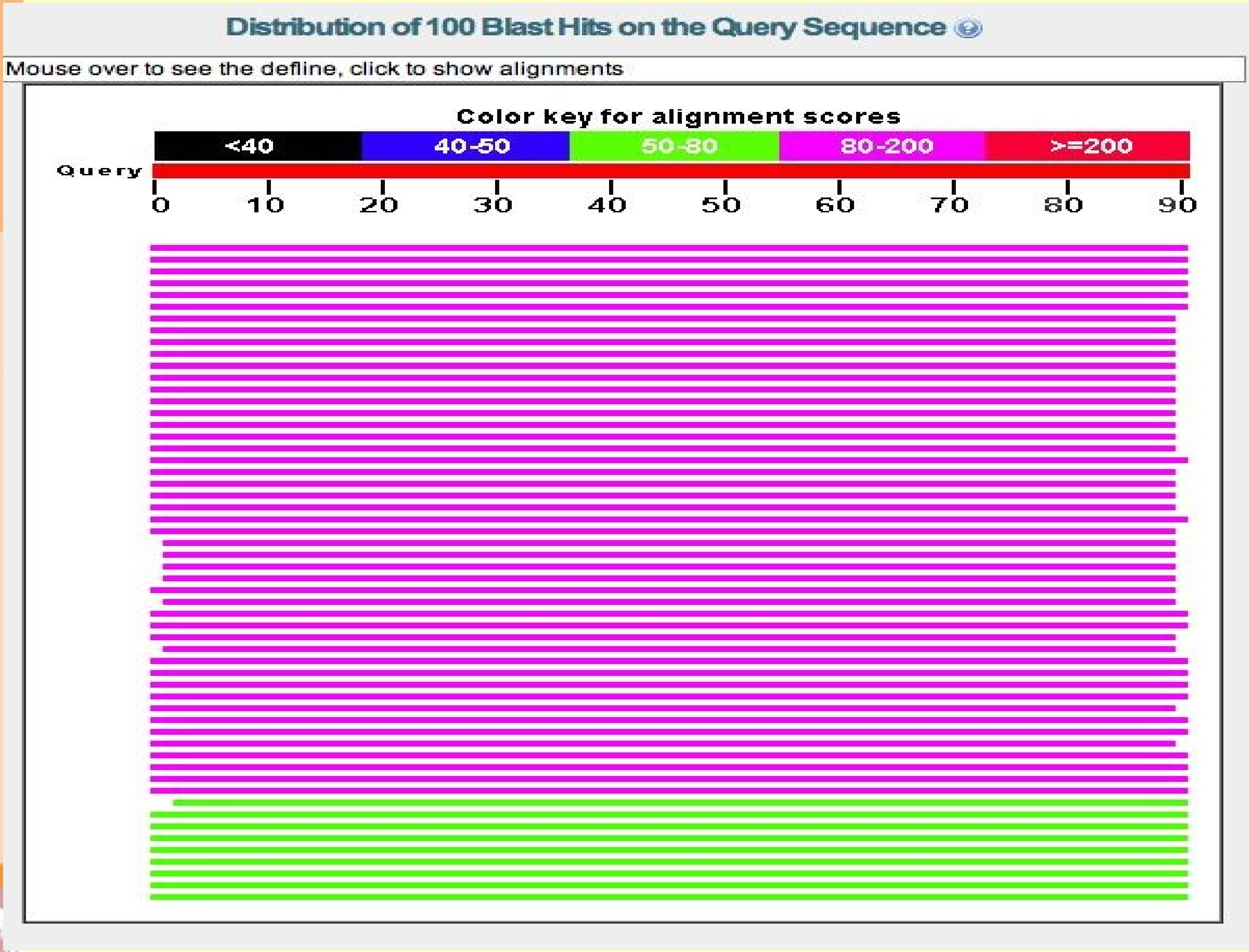
Max Seq Difference: 0.85

Distance: Grishin (protein)

Reset
Download
in
Newick Format
rectangle
slanted
radial
force
Mouse over an internal node for a subtree or alignment
RecName: Full=Integration host.
enterobacteria | 2 leaves
enterobacteria | 7 leaves
RecName: Full=Integration host.
RecName: Full=Integration host.
g-proteobacteria | 14 leaves
RecName: Full=Integration host.
g-proteobacteria | 2 leaves
RecName: Full=Integration host.
b-proteobacteria | 3 leaves
a-proteobacteria | 51 leaves
d-proteobacteria | 5 leaves
RecName: Full=DNA-binding p..
a-proteobacteria | 2 leaves
RecName: Full=DNA-binding p..
RecName: Full=DNA-binding p..
chlamydias | 2 leaves
spirochetes | 7 leaves
a-proteobacteria | 45 leaves
RecName: Full=Integration host.
RecName: Full=Integration host.
RecName: Full=Integration host.
enterobacteria | 3 leaves
g-proteobacteria | 9 leaves
RecName: Full=Integration host.
g-proteobacteria | 2 leaves
g-proteobacteria | 11 leaves
g-proteobacteria | 2 leaves
g-proteobacteria | 6 leaves
g-proteobacteria | 3 leaves
enterobacteria | 19 leaves
g-proteobacteria | 3 leaves
g-proteobacteria | 17 leaves
b-proteobacteria | 23 leaves
g-proteobacteria | 11 leaves
high GC Gram+ | 4 leaves
e-proteobacteria | 3 leaves
RecName: Full=DNA-binding p..
a-proteobacteria | 5 leaves
RecName: Full=DNA-binding p..
cyanobacteria | 2 leaves
RecName: Full=DNA-binding p..
RecName: Full=DNA-binding p..
firmicutes | 12 leaves
a-proteobacteria | 3 leaves
RecName: Full=DNA-binding p..
enterobacteria | 3 leaves
g-proteobacteria | 2 leaves
RecName: Full=DNA-binding p..
enterobacteria | 3 leaves
g-proteobacteria | 2 leaves
RecName: Full=DNA-binding p..
RecName: Full=DNA-binding p..
RecName: Full=DNA-binding p..
RecName: Full=Transcription fa..
RecName: Full=DNA-binding p..
RecName: Full=DNA-binding p..
viruses | 2 leaves
RecName: Full=DNA-binding p..
g-proteobacteria | 2 leaves
b-proteobacteria | 2 leaves
RecName: Full=DNA-binding p..
RecName: Full=DNA-binding p..
RecName: Full=DNA-binding p..
HU-NS1
RecName: Full=DNA-binding p..
g-proteobacteria | 2 leaves
b-proteobacteria | 4 leaves
enterobacteria | 2 leaves
RecName: Full=Integration host.
RecName: Full=Integration host.
b-proteobacteria | 14 leaves
g-proteobacteria | 24 leaves
g-proteobacteria | 23 leaves
enterobacteria | 4 leaves
RecName: Full=Integration host.

BLAST High Scores

http://blast.ncbi.nlm.nih.gov/



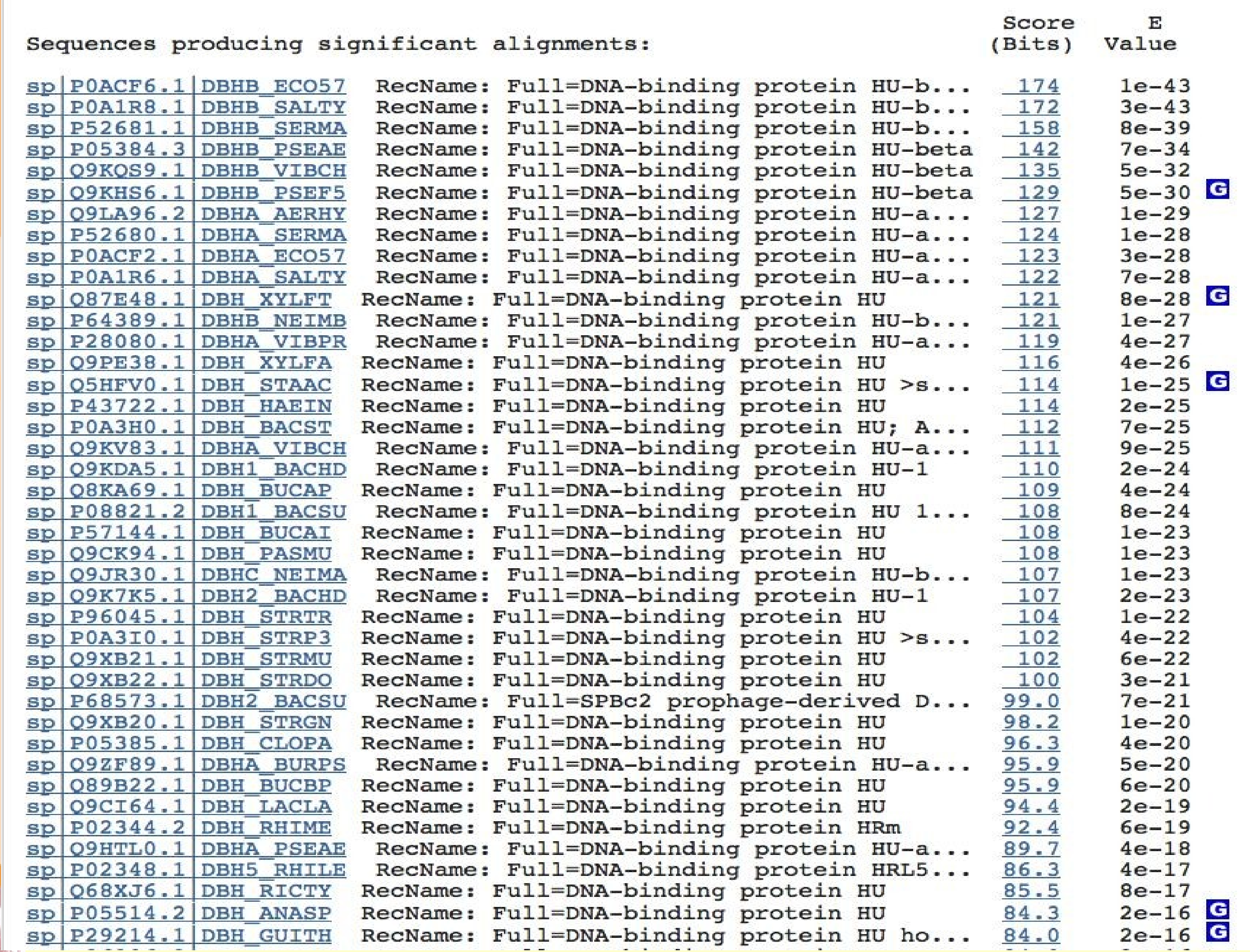
Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

Doug Brutlag 2010

BLAST High Scores

http://blast.ncbi.nlm.nih.gov/

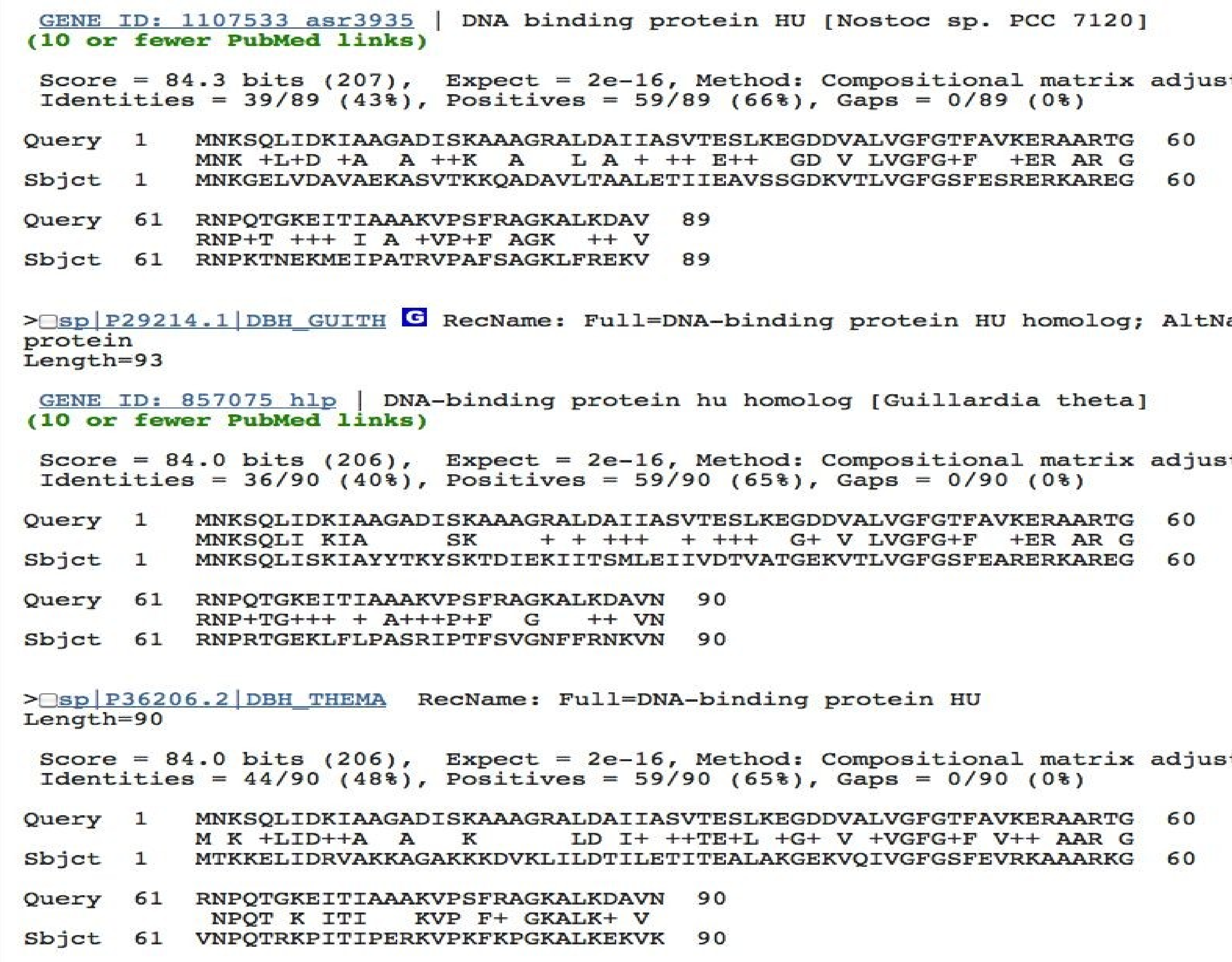


```
                                                                           Score     E
Sequences producing significant alignments:                                (Bits)  Value

sp|P0ACF6.1|DBHB_ECO57  RecName: Full=DNA-binding protein HU-b...           174     1e-43
sp|P0A1R8.1|DBHB_SALTY  RecName: Full=DNA-binding protein HU-b...           172     3e-43
sp|P52681.1|DBHB_SERMA  RecName: Full=DNA-binding protein HU-b...           158     8e-39
sp|P05384.3|DBHB_PSEAE  RecName: Full=DNA-binding protein HU-beta           142     7e-34
sp|Q9KQS9.1|DBHB_VIBCH  RecName: Full=DNA-binding protein HU-beta           135     5e-32
sp|Q9KHS6.1|DBHB_PSEF5  RecName: Full=DNA-binding protein HU-beta           129     5e-30 G
sp|Q9LA96.2|DBHA_AERHY  RecName: Full=DNA-binding protein HU-a...           127     1e-29
sp|P52680.1|DBHA_SERMA  RecName: Full=DNA-binding protein HU-a...           124     1e-28
sp|P0ACF2.1|DBHA_ECO57  RecName: Full=DNA-binding protein HU-a...           123     3e-28
sp|P0A1R6.1|DBHA_SALTY  RecName: Full=DNA-binding protein HU-a...           122     7e-28
sp|Q87E48.1|DBH_XYLFT  RecName: Full=DNA-binding protein HU                 121     8e-28 G
sp|P64389.1|DBHB_NEIMB  RecName: Full=DNA-binding protein HU-b...           121     1e-27
sp|P28080.1|DBHA_VIBPR  RecName: Full=DNA-binding protein HU-a...           119     4e-27
sp|Q9PE38.1|DBH_XYLFA  RecName: Full=DNA-binding protein HU                 116     4e-26
sp|Q5HFV0.1|DBH_STAAC  RecName: Full=DNA-binding protein HU >s...           114     1e-25 G
sp|P43722.1|DBH_HAEIN  RecName: Full=DNA-binding protein HU                 114     2e-25
sp|P0A3H0.1|DBH_BACST  RecName: Full=DNA-binding protein HU; A...           112     7e-25
sp|Q9KV83.1|DBHA_VIBCH  RecName: Full=DNA-binding protein HU-a...           111     9e-25
sp|Q9KDA5.1|DBH1_BACHD  RecName: Full=DNA-binding protein HU-1              110     2e-24
sp|Q8KA69.1|DBH_BUCAP  RecName: Full=DNA-binding protein HU                 109     4e-24
sp|P08821.2|DBH1_BACSU  RecName: Full=DNA-binding protein HU 1...           108     8e-24
sp|P57144.1|DBH_BUCAI  RecName: Full=DNA-binding protein HU                 108     1e-23
sp|Q9CK94.1|DBH_PASMU  RecName: Full=DNA-binding protein HU                 108     1e-23
sp|Q9JR30.1|DBHC_NEIMA  RecName: Full=DNA-binding protein HU-b...           107     1e-23
sp|Q9K7K5.1|DBH2_BACHD  RecName: Full=DNA-binding protein HU-1              107     2e-23
sp|P96045.1|DBH_STRTR  RecName: Full=DNA-binding protein HU                 104     1e-22
sp|P0A3I0.1|DBH_STRP3  RecName: Full=DNA-binding protein HU >s...           102     4e-22
sp|Q9XB21.1|DBH_STRMU  RecName: Full=DNA-binding protein HU                 102     6e-22
sp|Q9XB22.1|DBH_STRDO  RecName: Full=DNA-binding protein HU                 100     3e-21
sp|P68573.1|DBH2_BACSU  RecName: Full=SPBc2 prophage-derived D...           99.0    7e-21
sp|Q9XB20.1|DBH_STRGN  RecName: Full=DNA-binding protein HU                 98.2    1e-20
sp|P05385.1|DBH_CLOPA  RecName: Full=DNA-binding protein HU                 96.3    4e-20
sp|Q9ZF89.1|DBHA_BURPS  RecName: Full=DNA-binding protein HU-a...           95.9    5e-20
sp|Q89B22.1|DBH_BUCBP  RecName: Full=DNA-binding protein HU                 95.9    6e-20
sp|Q9CI64.1|DBH_LACLA  RecName: Full=DNA-binding protein HU                 94.4    2e-19
sp|P02344.2|DBH_RHIME  RecName: Full=DNA-binding protein HRm                92.4    6e-19
sp|Q9HTL0.1|DBHA_PSEAE  RecName: Full=DNA-binding protein HU-a...           89.7    4e-18
sp|P02348.1|DBH5_RHILE  RecName: Full=DNA-binding protein HRL5...           86.3    4e-17
sp|Q68XJ6.1|DBH_RICTY  RecName: Full=DNA-binding protein HU                 85.5    8e-17
sp|P05514.2|DBH_ANASP  RecName: Full=DNA-binding protein HU                 84.3    2e-16 G
sp|P29214.1|DBH_GUITH  RecName: Full=DNA-binding protein HU ho...           84.0    2e-16 G
```

Doug Brutlag 2010

BLAST Low Scores

http://blast.ncbi.nlm.nih.gov/

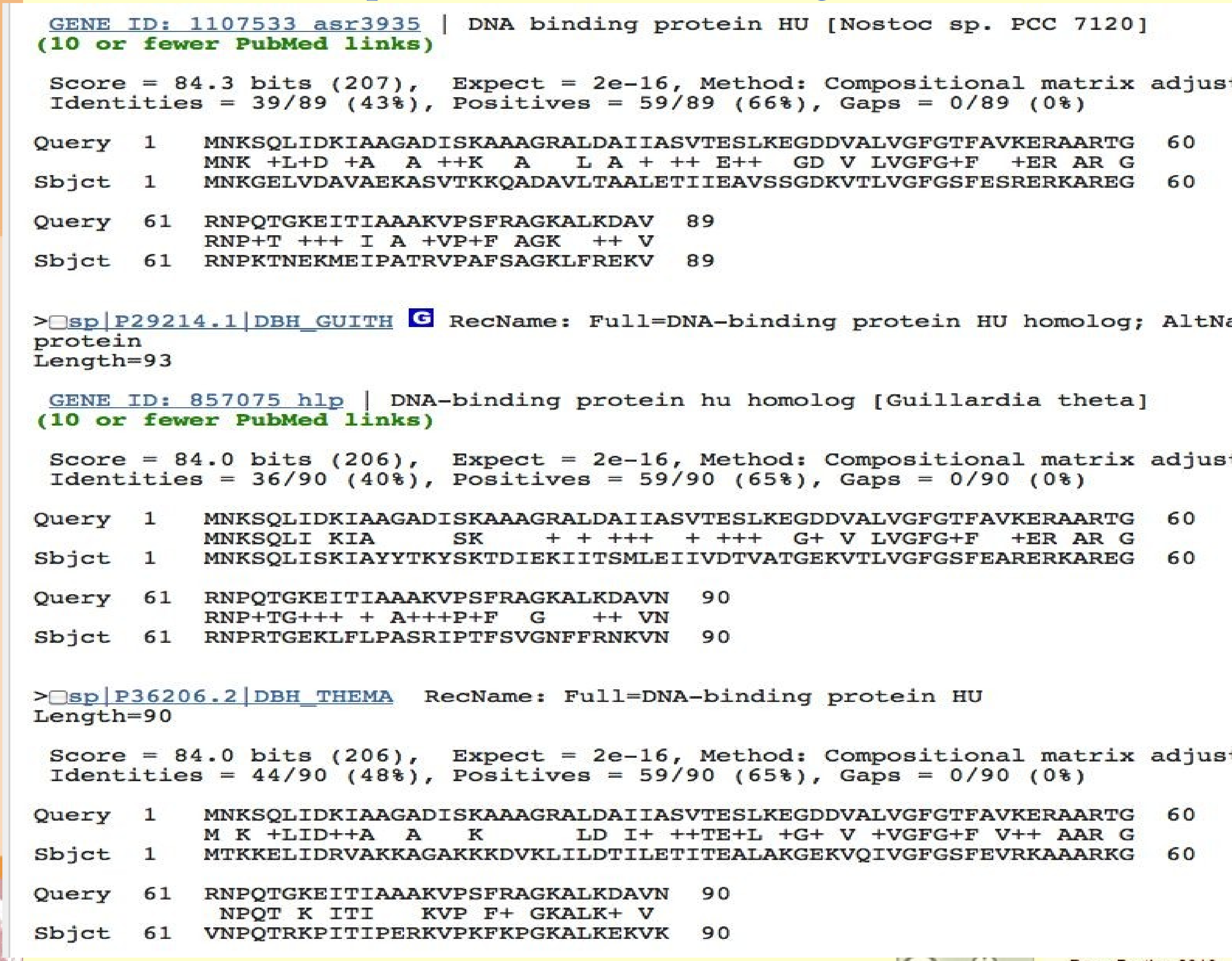


```
 GENE ID: 1107533 asr3935 | DNA binding protein HU [Nostoc sp. PCC 7120]
(10 or fewer PubMed links)

 Score = 84.3 bits (207),  Expect = 2e-16, Method: Compositional matrix adjus
 Identities = 39/89 (43%), Positives = 59/89 (66%), Gaps = 0/89 (0%)

Query  1    MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTG  60
            MNK +L+D +A  A ++K  A    L A + ++ E++  GD V LVGFG+F  +ER AR G
Sbjct  1    MNKGELVDAVAEKASVTKKQADAVLTAALETIIEAVSSGDKVTLVGFGSFESRERKAREG  60

Query  61   RNPQTGKEITIAAAKVPSFRAGKALKDAV  89
            RNP+T +++ I A +VP+F AGK  ++ V
Sbjct  61   RNPKTNEKMEIPATRVPAFSAGKLFREKV  89


>□sp|P29214.1|DBH_GUITH G RecName: Full=DNA-binding protein HU homolog; AltNa
protein
Length=93

 GENE ID: 857075 hlp | DNA-binding protein hu homolog [Guillardia theta]
(10 or fewer PubMed links)

 Score = 84.0 bits (206),  Expect = 2e-16, Method: Compositional matrix adjus
 Identities = 36/90 (40%), Positives = 59/90 (65%), Gaps = 0/90 (0%)

Query  1    MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTG  60
            MNKSQLI KIA     SK    + + +++  + +++  G+ V LVGFG+F  +ER AR G
Sbjct  1    MNKSQLISKIAYYTKYSKTDIEKIITSMLEIIVDTVATGEKVTLVGFGSFEARERKAREG  60

Query  61   RNPQTGKEITIAAAKVPSFRAGKALKDAVN  90
            RNP+TG+++ + A+++P+F  G   ++ VN
Sbjct  61   RNPRTGEKLFLPASRIPTFSVGNFFRNKVN  90


>□sp|P36206.2|DBH_THEMA  RecName: Full=DNA-binding protein HU
Length=90

 Score = 84.0 bits (206),  Expect = 2e-16, Method: Compositional matrix adjus
 Identities = 44/90 (48%), Positives = 59/90 (65%), Gaps = 0/90 (0%)

Query  1    MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTG  60
            M K +LID++A  A   K      LD I+ ++TE+L +G+ V +VGFG+F V++ AAR G
Sbjct  1    MTKKELIDRVAKKAGAKKKDVKLILDTILETITEALAKGEKVQIVGFGSFEVRKAAARKG  60

Query  61   RNPQTGKEITIAAAKVPSFRAGKALKDAVN  90
             NPQT K ITI   KVP F+ GKALK+ V
Sbjct  61   VNPQTRKPITIPERKVPKFKPGKALKEKVK  90
```

Doug Brutlag 2010

BLAST Typical Alignments

http://blast.ncbi.nlm.nih.gov/

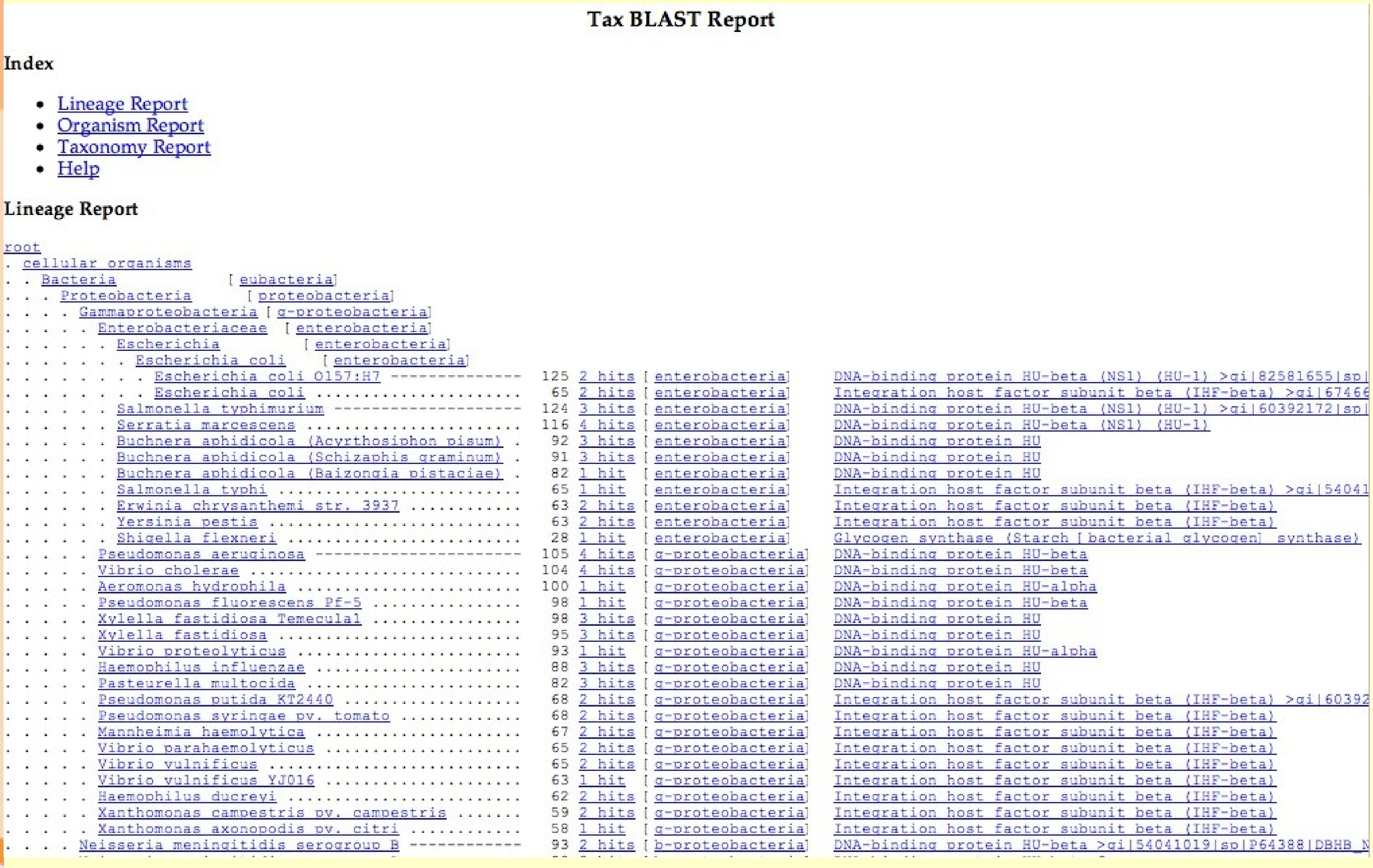


```
 GENE ID: 1107533 asr3935 | DNA binding protein HU [Nostoc sp. PCC 7120]
(10 or fewer PubMed links)

 Score = 84.3 bits (207),  Expect = 2e-16, Method: Compositional matrix adjus
 Identities = 39/89 (43%), Positives = 59/89 (66%), Gaps = 0/89 (0%)

Query  1    MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTG  60
            MNK +L+D +A  A ++K  A   L A + ++ E++  GD V LVGFG+F  +ER AR G
Sbjct  1    MNKGELVDAVAEKASVTKKQADAVLTAALETIIEAVSSGDKVTLVGFGSFESRERKAREG  60

Query  61   RNPQTGKEITIAAAKVPSFRAGKALKDAV  89
            RNP+T +++ I A +VP+F AGK  ++ V
Sbjct  61   RNPKTNEKMEIPATRVPAFSAGKLFREKV  89


>☐sp|P29214.1|DBH_GUITH  G  RecName: Full=DNA-binding protein HU homolog; AltNa
protein
Length=93

 GENE ID: 857075 hlp | DNA-binding protein hu homolog [Guillardia theta]
(10 or fewer PubMed links)

 Score = 84.0 bits (206),  Expect = 2e-16, Method: Compositional matrix adjus
 Identities = 36/90 (40%), Positives = 59/90 (65%), Gaps = 0/90 (0%)

Query  1    MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTG  60
            MNKSQLI KIA     SK    + + +++  + +++  G+ V LVGFG+F  +ER AR G
Sbjct  1    MNKSQLISKIAYYTKYSKTDIEKIITSMLEIIVDTVATGEKVTLVGFGSFEARERKAREG  60

Query  61   RNPQTGKEITIAAAKVPSFRAGKALKDAVN  90
            RNP+TG+++ + A+++P+F  G    ++ VN
Sbjct  61   RNPRTGEKLFLPASRIPTFSVGNFFRNKVN  90


>☐sp|P36206.2|DBH_THEMA  RecName: Full=DNA-binding protein HU
Length=90

 Score = 84.0 bits (206),  Expect = 2e-16, Method: Compositional matrix adjus
 Identities = 44/90 (48%), Positives = 59/90 (65%), Gaps = 0/90 (0%)

Query  1    MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTG  60
            M K +LID++A  A   K      LD I+ ++TE+L +G+ V +VGFG+F V++ AAR G
Sbjct  1    MTKKELIDRVAKKAGAKKKDVKLILDTILETITEALAKGEKVQIVGFGSFEVRKAAARKG  60

Query  61   RNPQTGKEITIAAAKVPSFRAGKALKDAVN  90
             NPQT K ITI   KVP F+ GKALK+ V
Sbjct  61   VNPQTRKPITIPERKVPKFKPGKALKEKVK  90
```

Doug Brutlag 2010

NCBI Blast Taxonomy Report

http://www.ncbi.nlm.nih.gov/BLAST/



Tax BLAST Report

Index

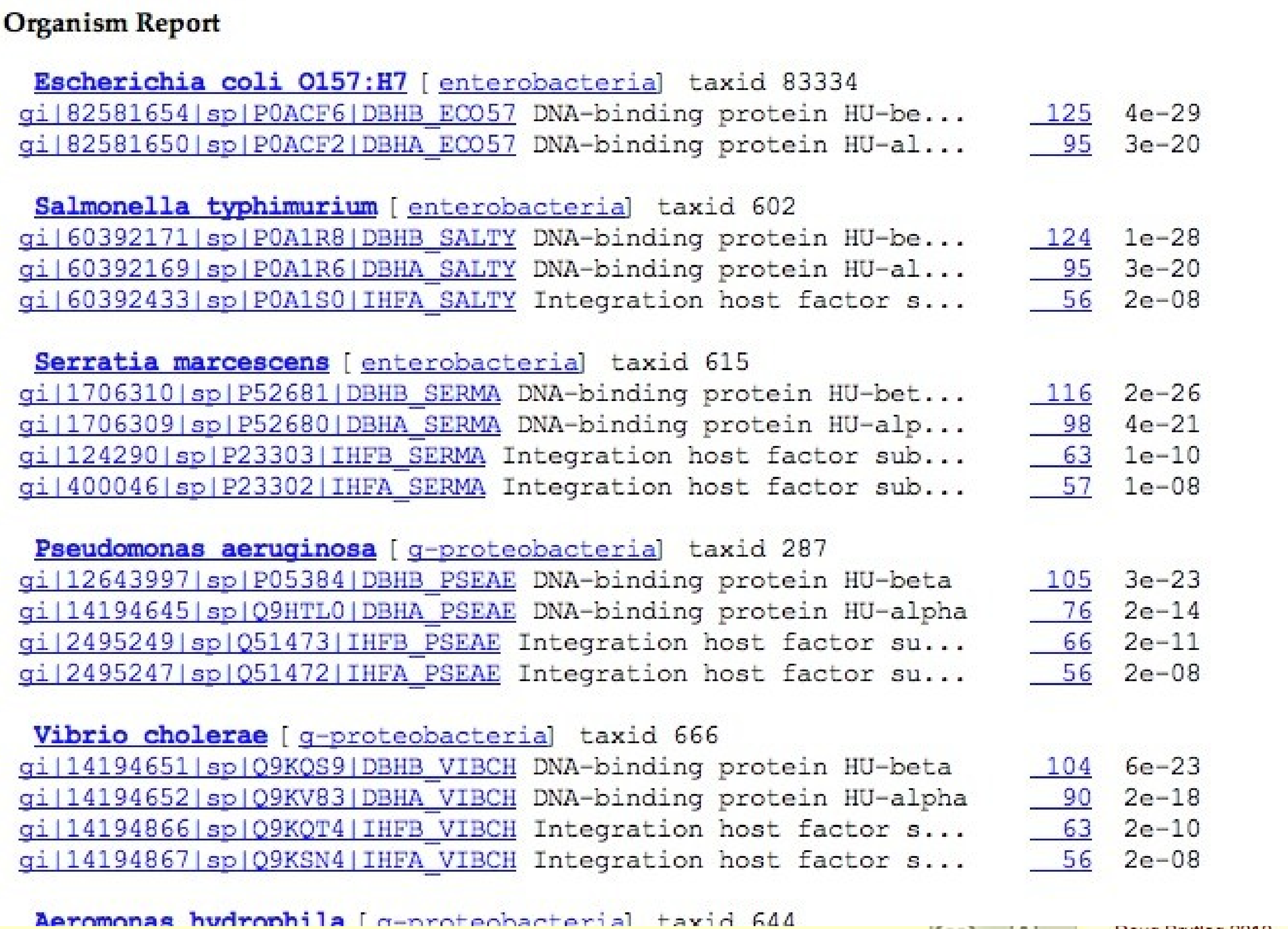
- Lineage Report
- Organism Report
- Taxonomy Report
- Help

Lineage Report

```
root
. cellular organisms
. . Bacteria                  [ eubacteria]
. . . Proteobacteria           [ proteobacteria]
. . . . Gammaproteobacteria [ g-proteobacteria]
. . . . . Enterobacteriaceae  [ enterobacteria]
. . . . . . Escherichia           [ enterobacteria]
. . . . . . . Escherichia coli    [ enterobacteria]
. . . . . . . . Escherichia coli O157:H7 -------------- 125 2 hits [ enterobacteria]    DNA-binding protein HU-beta (NS1) (HU-1) >gi|82581655|sp|
. . . . . . . . Escherichia coli ......................  65 2 hits [ enterobacteria]    Integration host factor subunit beta (IHF-beta) >gi|67466
. . . . . . Salmonella typhimurium -------------------- 124 3 hits [ enterobacteria]    DNA-binding protein HU-beta (NS1) (HU-1) >gi|60392172|sp|
. . . . . . Serratia marcescens ....................... 116 4 hits [ enterobacteria]    DNA-binding protein HU-beta (NS1) (HU-1)
. . . . . . Buchnera aphidicola (Acyrthosiphon pisum) .  92 3 hits [ enterobacteria]    DNA-binding protein HU
. . . . . . Buchnera aphidicola (Schizaphis graminum) .  91 3 hits [ enterobacteria]    DNA-binding protein HU
. . . . . . Buchnera aphidicola (Baizongia pistaciae) .  82 1 hit  [ enterobacteria]    DNA-binding protein HU
. . . . . . Salmonella typhi ..........................  65 1 hit  [ enterobacteria]    Integration host factor subunit beta (IHF-beta) >gi|54041
. . . . . . Erwinia chrysanthemi str. 3937 ............  63 2 hits [ enterobacteria]    Integration host factor subunit beta (IHF-beta)
. . . . . . Yersinia pestis ...........................  63 2 hits [ enterobacteria]    Integration host factor subunit beta (IHF-beta)
. . . . . . Shigella flexneri .........................  28 1 hit  [ enterobacteria]    Glycogen synthase (Starch [bacterial glycogen] synthase)
. . . . . Pseudomonas aeruginosa ---------------------- 105 4 hits [ g-proteobacteria]  DNA-binding protein HU-beta
. . . . . Vibrio cholerae ............................. 104 4 hits [ g-proteobacteria]  DNA-binding protein HU-beta
. . . . . Aeromonas hydrophila ........................ 100 1 hit  [ g-proteobacteria]  DNA-binding protein HU-alpha
. . . . . Pseudomonas fluorescens Pf-5 ................  98 1 hit  [ g-proteobacteria]  DNA-binding protein HU-beta
. . . . . Xylella fastidiosa Temecula1 ................  98 3 hits [ g-proteobacteria]  DNA-binding protein HU
. . . . . Xylella fastidiosa ..........................  95 3 hits [ g-proteobacteria]  DNA-binding protein HU
. . . . . Vibrio proteolyticus ........................  93 1 hit  [ g-proteobacteria]  DNA-binding protein HU-alpha
. . . . . Haemophilus influenzae ......................  88 3 hits [ g-proteobacteria]  DNA-binding protein HU
. . . . . Pasteurella multocida .......................  82 3 hits [ g-proteobacteria]  DNA-binding protein HU
. . . . . Pseudomonas putida KT2440 ...................  68 2 hits [ g-proteobacteria]  Integration host factor subunit beta (IHF-beta) >gi|60392
. . . . . Pseudomonas syringae pv. tomato .............  68 2 hits [ g-proteobacteria]  Integration host factor subunit beta (IHF-beta)
. . . . . Mannheimia haemolytica ......................  67 2 hits [ g-proteobacteria]  Integration host factor subunit beta (IHF-beta)
. . . . . Vibrio parahaemolyticus .....................  65 2 hits [ g-proteobacteria]  Integration host factor subunit beta (IHF-beta)
. . . . . Vibrio vulnificus ...........................  65 2 hits [ g-proteobacteria]  Integration host factor subunit beta (IHF-beta)
. . . . . Vibrio vulnificus YJ016 .....................  63 1 hit  [ g-proteobacteria]  Integration host factor subunit beta (IHF-beta)
. . . . . Haemophilus ducreyi .........................  62 2 hits [ g-proteobacteria]  Integration host factor subunit beta (IHF-beta)
. . . . . Xanthomonas campestris pv. campestris .......  59 2 hits [ g-proteobacteria]  Integration host factor subunit beta (IHF-beta)
. . . . . Xanthomonas axonopodis pv. citri ............  58 1 hit  [ g-proteobacteria]  Integration host factor subunit beta (IHF-beta)
. . . . Neisseria meningitidis serogroup B ------------  93 2 hits [ b-proteobacteria]  DNA-binding protein HU-beta >gi|54041019|sp|P64388|DBHB_N
```

NCBI Blast Organism Report

http://www.ncbi.nlm.nih.gov/BLAST/

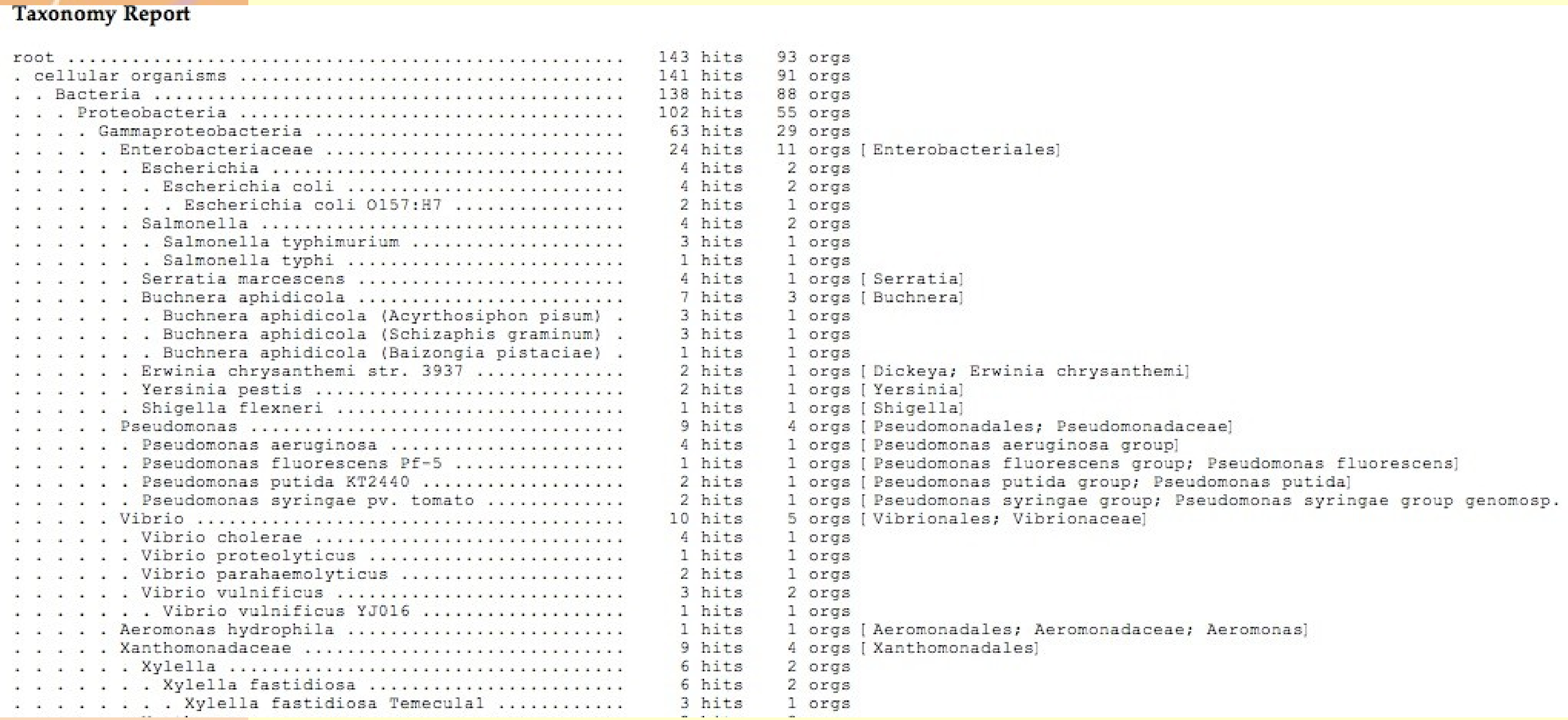


Organism Report

```
 Escherichia coli O157:H7 [ enterobacteria]  taxid 83334
gi|82581654|sp|P0ACF6|DBHB_ECO57 DNA-binding protein HU-be...     125   4e-29
gi|82581650|sp|P0ACF2|DBHA_ECO57 DNA-binding protein HU-al...      95   3e-20

 Salmonella typhimurium [ enterobacteria]  taxid 602
gi|60392171|sp|P0A1R8|DBHB_SALTY DNA-binding protein HU-be...     124   1e-28
gi|60392169|sp|P0A1R6|DBHA_SALTY DNA-binding protein HU-al...      95   3e-20
gi|60392433|sp|P0A1S0|IHFA_SALTY Integration host factor s...      56   2e-08

 Serratia marcescens [ enterobacteria]  taxid 615
gi|1706310|sp|P52681|DBHB_SERMA DNA-binding protein HU-bet...     116   2e-26
gi|1706309|sp|P52680|DBHA_SERMA DNA-binding protein HU-alp...      98   4e-21
gi|124290|sp|P23303|IHFB_SERMA Integration host factor sub...      63   1e-10
gi|400046|sp|P23302|IHFA_SERMA Integration host factor sub...      57   1e-08

 Pseudomonas aeruginosa [ g-proteobacteria]  taxid 287
gi|12643997|sp|P05384|DBHB_PSEAE DNA-binding protein HU-beta      105   3e-23
gi|14194645|sp|Q9HTL0|DBHA_PSEAE DNA-binding protein HU-alpha      76   2e-14
gi|2495249|sp|Q51473|IHFB_PSEAE Integration host factor su...      66   2e-11
gi|2495247|sp|Q51472|IHFA_PSEAE Integration host factor su...      56   2e-08

 Vibrio cholerae [ g-proteobacteria]  taxid 666
gi|14194651|sp|Q9KQS9|DBHB_VIBCH DNA-binding protein HU-beta      104   6e-23
gi|14194652|sp|Q9KV83|DBHA_VIBCH DNA-binding protein HU-alpha      90   2e-18
gi|14194866|sp|Q9KQT4|IHFB_VIBCH Integration host factor s...      63   2e-10
gi|14194867|sp|Q9KSN4|IHFA_VIBCH Integration host factor s...      56   2e-08

 Aeromonas hydrophila [ g-proteobacteria]  taxid 644
```

NCBI Blast Taxomomy Report

http://www.ncbi.nlm.nih.gov/BLAST/

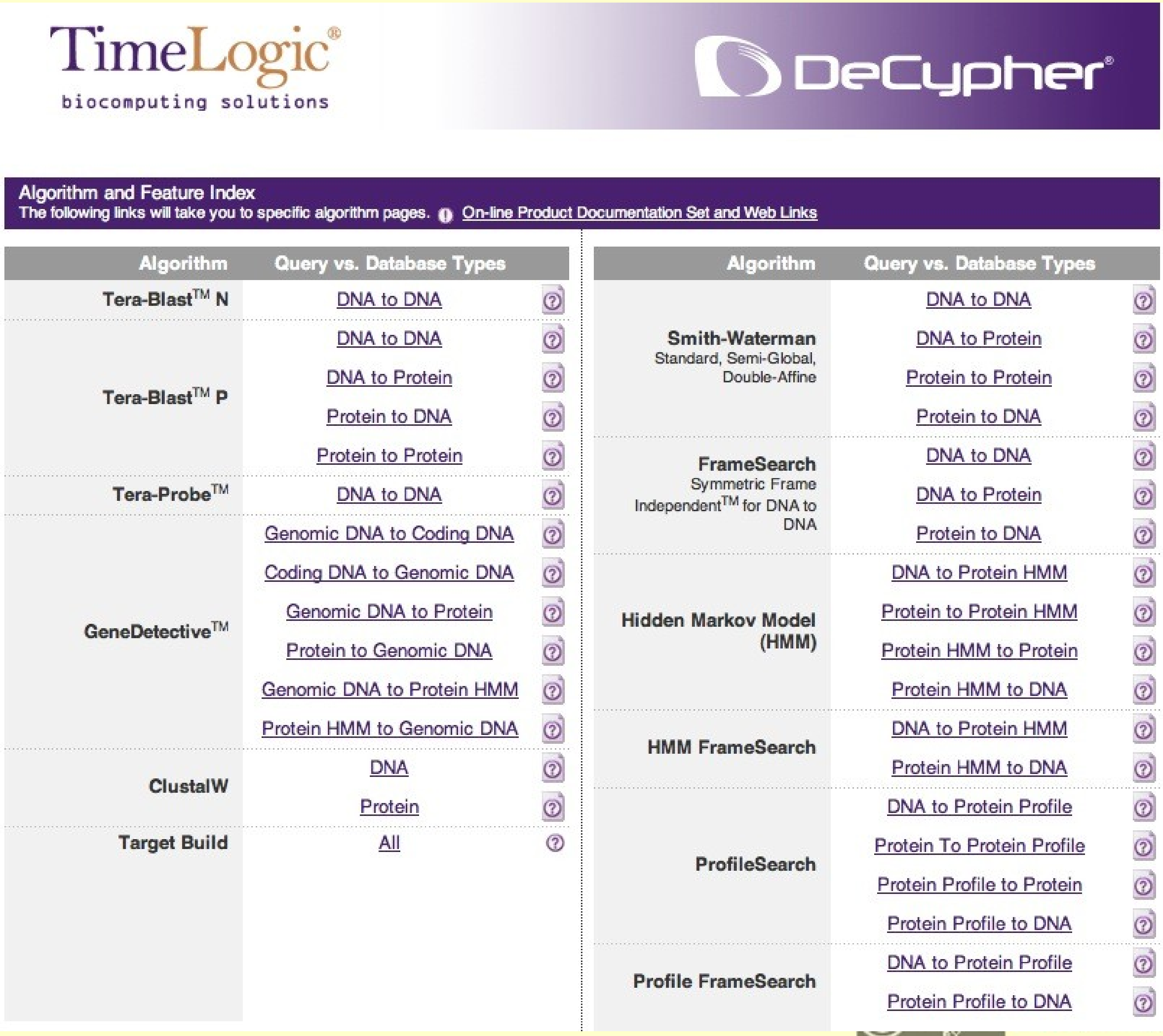


Taxonomy Report

```
root ....................................................  143 hits   93 orgs
. cellular organisms ....................................  141 hits   91 orgs
. . Bacteria ............................................  138 hits   88 orgs
. . . Proteobacteria ....................................  102 hits   55 orgs
. . . . Gammaproteobacteria .............................   63 hits   29 orgs
. . . . . Enterobacteriaceae ............................   24 hits   11 orgs [ Enterobacteriales]
. . . . . . Escherichia .................................    4 hits    2 orgs
. . . . . . . Escherichia coli ..........................    4 hits    2 orgs
. . . . . . . . Escherichia coli O157:H7 ................    2 hits    1 orgs
. . . . . . Salmonella ..................................    4 hits    2 orgs
. . . . . . . Salmonella typhimurium ....................    3 hits    1 orgs
. . . . . . . Salmonella typhi ..........................    1 hits    1 orgs
. . . . . . Serratia marcescens .........................    4 hits    1 orgs [ Serratia]
. . . . . . Buchnera aphidicola .........................    7 hits    3 orgs [ Buchnera]
. . . . . . . Buchnera aphidicola (Acyrthosiphon pisum) .    3 hits    1 orgs
. . . . . . . Buchnera aphidicola (Schizaphis graminum) .    3 hits    1 orgs
. . . . . . . Buchnera aphidicola (Baizongia pistaciae) .    1 hits    1 orgs
. . . . . . Erwinia chrysanthemi str. 3937 ..............    2 hits    1 orgs [ Dickeya; Erwinia chrysanthemi]
. . . . . . Yersinia pestis .............................    2 hits    1 orgs [ Yersinia]
. . . . . . Shigella flexneri ...........................    1 hits    1 orgs [ Shigella]
. . . . . Pseudomonas ...................................    9 hits    4 orgs [ Pseudomonadales; Pseudomonadaceae]
. . . . . . Pseudomonas aeruginosa ......................    4 hits    1 orgs [ Pseudomonas aeruginosa group]
. . . . . . Pseudomonas fluorescens Pf-5 ................    1 hits    1 orgs [ Pseudomonas fluorescens group; Pseudomonas fluorescens]
. . . . . . Pseudomonas putida KT2440 ...................    2 hits    1 orgs [ Pseudomonas putida group; Pseudomonas putida]
. . . . . . Pseudomonas syringae pv. tomato .............    2 hits    1 orgs [ Pseudomonas syringae group; Pseudomonas syringae group genomosp.
. . . . . Vibrio ........................................   10 hits    5 orgs [ Vibrionales; Vibrionaceae]
. . . . . . Vibrio cholerae .............................    4 hits    1 orgs
. . . . . . Vibrio proteolyticus ........................    1 hits    1 orgs
. . . . . . Vibrio parahaemolyticus .....................    2 hits    1 orgs
. . . . . . Vibrio vulnificus ...........................    3 hits    2 orgs
. . . . . . . Vibrio vulnificus YJ016 ...................    1 hits    1 orgs
. . . . . Aeromonas hydrophila ..........................    1 hits    1 orgs [ Aeromonadales; Aeromonadaceae; Aeromonas]
. . . . . Xanthomonadaceae ..............................    9 hits    4 orgs [ Xanthomonadales]
. . . . . . Xylella .....................................    6 hits    2 orgs
. . . . . . . Xylella fastidiosa ........................    6 hits    2 orgs
. . . . . . . . Xylella fastidiosa Temecula1 ............    3 hits    1 orgs
```

Decypher Search Engine

http://decypher.stanford.edu/



Algorithm and Feature Index

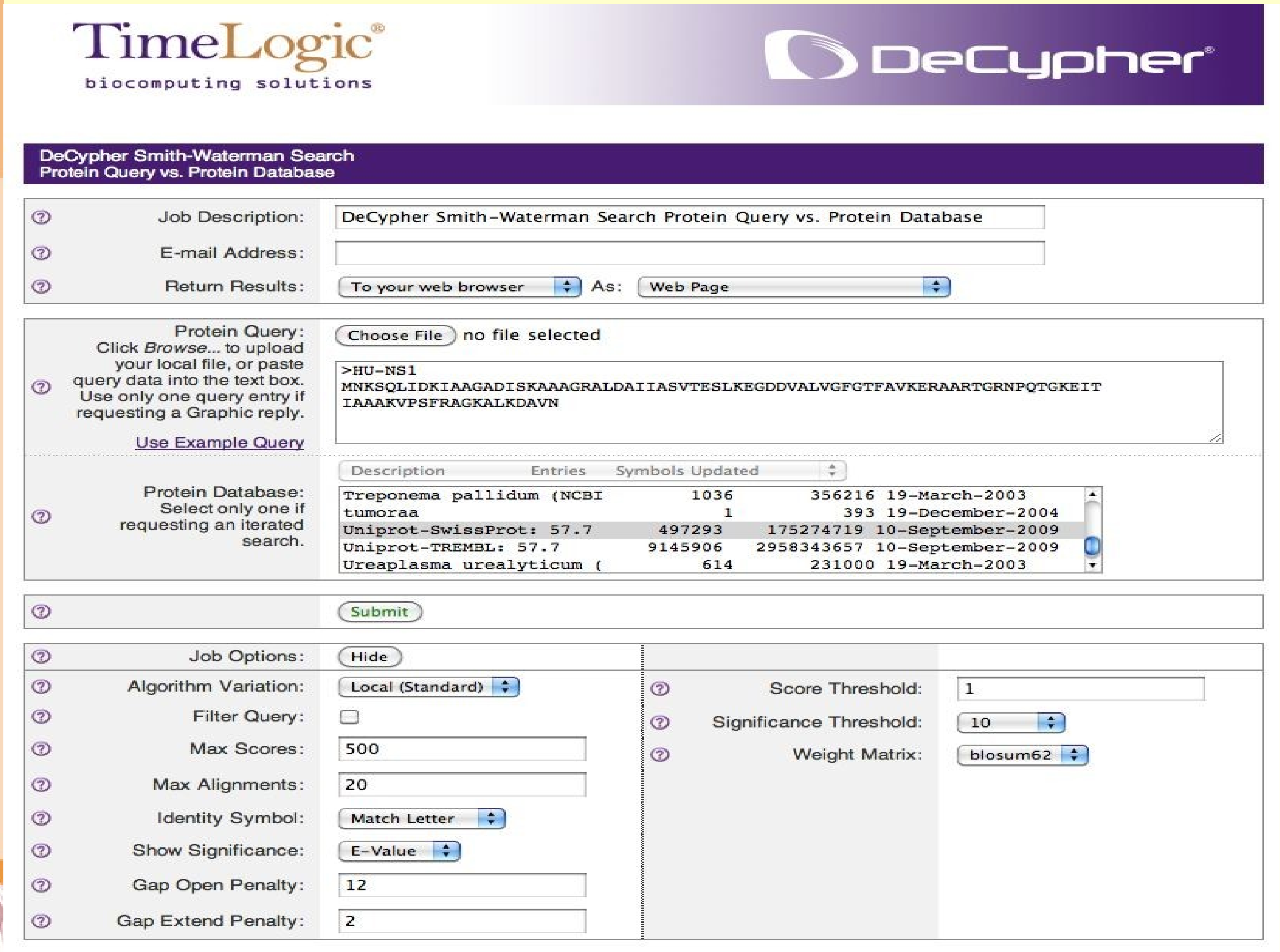
The following links will take you to specific algorithm pages. On-line Product Documentation Set and Web Links

Algorithm	Query vs. Database Types
Tera-Blast^{TM} N	DNA to DNA
Tera-Blast^{TM} P	DNA to DNA
	DNA to Protein
	Protein to DNA
	Protein to Protein
Tera-Probe^{TM}	DNA to DNA
GeneDetective^{TM}	Genomic DNA to Coding DNA
	Coding DNA to Genomic DNA
	Genomic DNA to Protein
	Protein to Genomic DNA
	Genomic DNA to Protein HMM
	Protein HMM to Genomic DNA
ClustalW	DNA
	Protein
Target Build	All

Algorithm	Query vs. Database Types
Smith-Waterman Standard, Semi-Global, Double-Affine	DNA to DNA
	DNA to Protein
	Protein to Protein
	Protein to DNA
FrameSearch Symmetric Frame Independent^{TM} for DNA to DNA	DNA to DNA
	DNA to Protein
	Protein to DNA
Hidden Markov Model (HMM)	DNA to Protein HMM
	Protein to Protein HMM
	Protein HMM to Protein
	Protein HMM to DNA
HMM FrameSearch	DNA to Protein HMM
	Protein HMM to DNA
ProfileSearch	DNA to Protein Profile
	Protein To Protein Profile
	Protein Profile to Protein
	Protein Profile to DNA
Profile FrameSearch	DNA to Protein Profile
	Protein Profile to DNA

Decypher Search Engine Input

http://decypher.stanford.edu/



DeCypher Smith-Waterman Search
Protein Query vs. Protein Database

Job Description:	DeCypher Smith-Waterman Search Protein Query vs. Protein Database
E-mail Address:	
Return Results:	To your web browser As: Web Page

Protein Query: Click *Browse...* to upload your local file, or paste query data into the text box. Use only one query entry if requesting a Graphic reply.

Use Example Query

Choose File no file selected

```
>HU-NS1
MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTGRNPQTGKEIT
IAAAKVPSFRAGKALKDAVN
```

Protein Database: Select only one if requesting an iterated search.

Description	Entries	Symbols	Updated
Treponema pallidum (NCBI	1036	356216	19-March-2003
tumoraa	1	393	19-December-2004
Uniprot-SwissProt: 57.7	497293	175274719	10-September-2009
Uniprot-TREMBL: 57.7	9145906	2958343657	10-September-2009
Ureaplasma urealyticum (	614	231000	19-March-2003

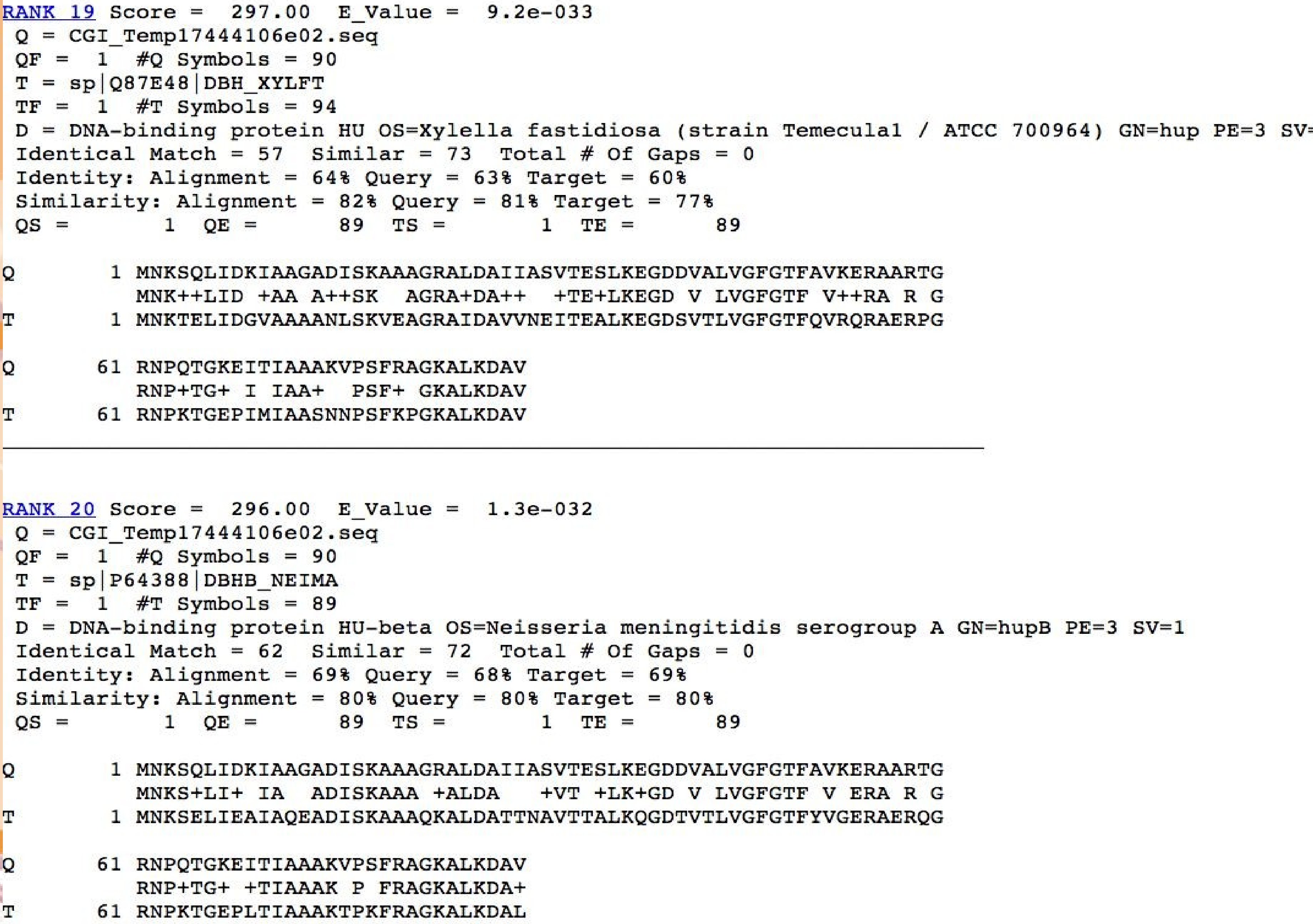
Submit

Job Options:	Hide
Algorithm Variation:	Local (Standard)
Filter Query:	□
Max Scores:	500
Max Alignments:	20
Identity Symbol:	Match Letter
Show Significance:	E-Value
Gap Open Penalty:	12
Gap Extend Penalty:	2
Score Threshold:	1
Significance Threshold:	10
Weight Matrix:	blosum62

Doug Brutlag 2010

Decypher Search Engine Results

http://decypher.stanford.edu/

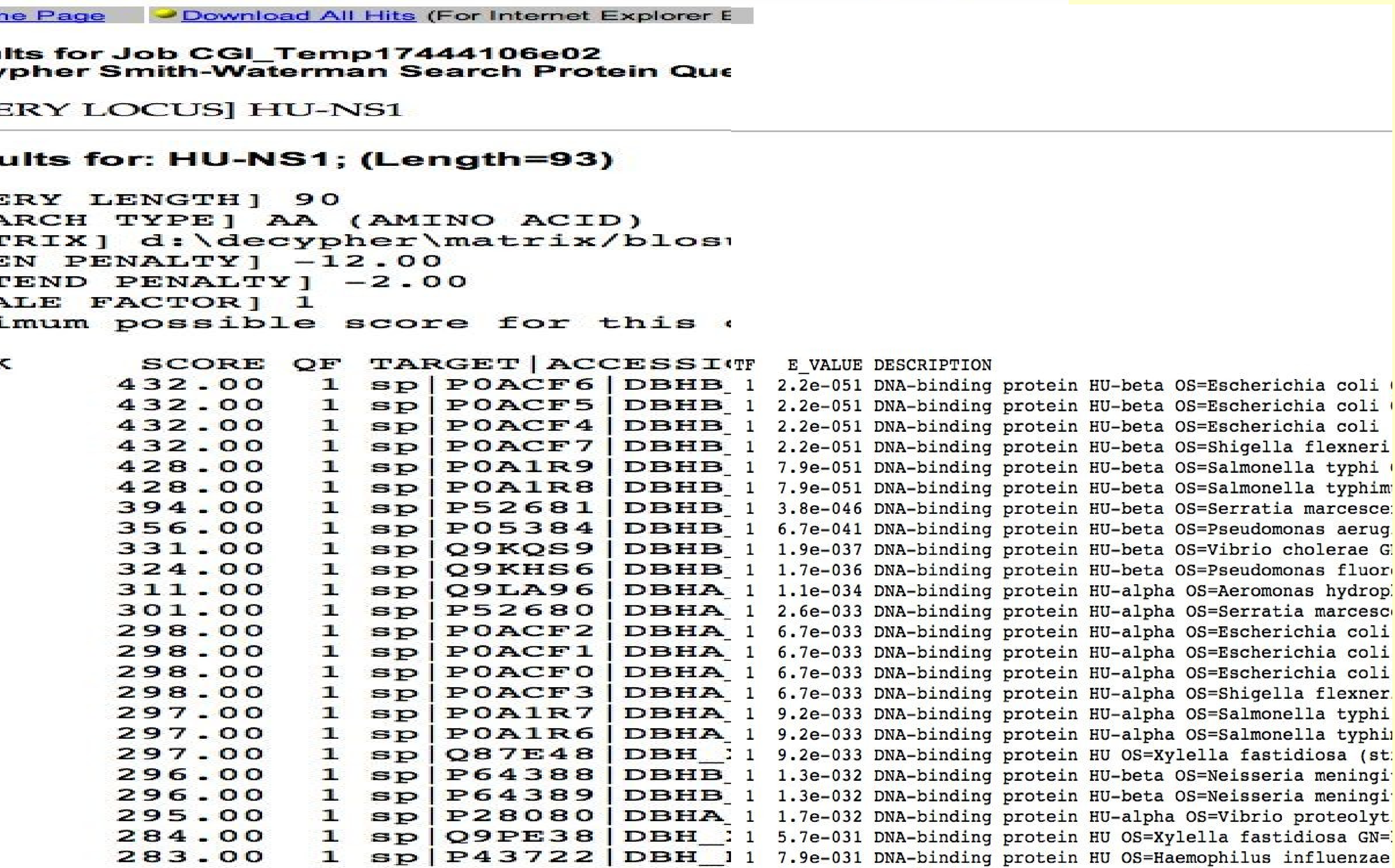


```
RANK 19 Score =  297.00  E_Value =  9.2e-033
 Q = CGI_Temp17444106e02.seq
 QF =  1  #Q Symbols = 90
 T = sp|Q87E48|DBH_XYLFT
 TF =  1  #T Symbols = 94
 D = DNA-binding protein HU OS=Xylella fastidiosa (strain Temecula1 / ATCC 700964) GN=hup PE=3 SV=
 Identical Match = 57  Similar = 73  Total # Of Gaps = 0
 Identity: Alignment = 64% Query = 63% Target = 60%
 Similarity: Alignment = 82% Query = 81% Target = 77%
 QS =        1  QE =        89  TS =        1  TE =        89

Q       1 MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTG
          MNK++LID +AA A++SK  AGRA+DA++  +TE+LKEGD V LVGFGTF V++RA R G
T       1 MNKTELIDGVAAAANLSKVEAGRAIDAVVNEITEALKEGDSVTLVGFGTFQVRQRAERPG

Q      61 RNPQTGKEITIAAAKVPSFRAGKALKDAV
          RNP+TG+ I IAA+  PSF+ GKALKDAV
T      61 RNPKTGEPIMIAASNNPSFKPGKALKDAV
```

```
RANK 20 Score =  296.00  E_Value =  1.3e-032
 Q = CGI_Temp17444106e02.seq
 QF =  1  #Q Symbols = 90
 T = sp|P64388|DBHB_NEIMA
 TF =  1  #T Symbols = 89
 D = DNA-binding protein HU-beta OS=Neisseria meningitidis serogroup A GN=hupB PE=3 SV=1
 Identical Match = 62  Similar = 72  Total # Of Gaps = 0
 Identity: Alignment = 69% Query = 68% Target = 69%
 Similarity: Alignment = 80% Query = 80% Target = 80%
 QS =        1  QE =        89  TS =        1  TE =        89

Q       1 MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTG
          MNKS+LI+ IA  ADISKAAA +ALDA    +VT +LK+GD V LVGFGTF V ERA R G
T       1 MNKSELIEAIAQEADISKAAAQKALDATTNAVTTALKQGDTVTLVGFGTFYVGERAERQG

Q      61 RNPQTGKEITIAAAKVPSFRAGKALKDAV
          RNP+TG+ +TIAAAK P FRAGKALKDA+
T      61 RNPKTGEPLTIAAAKTPKFRAGKALKDAL
```

Decypher Search Engine Results

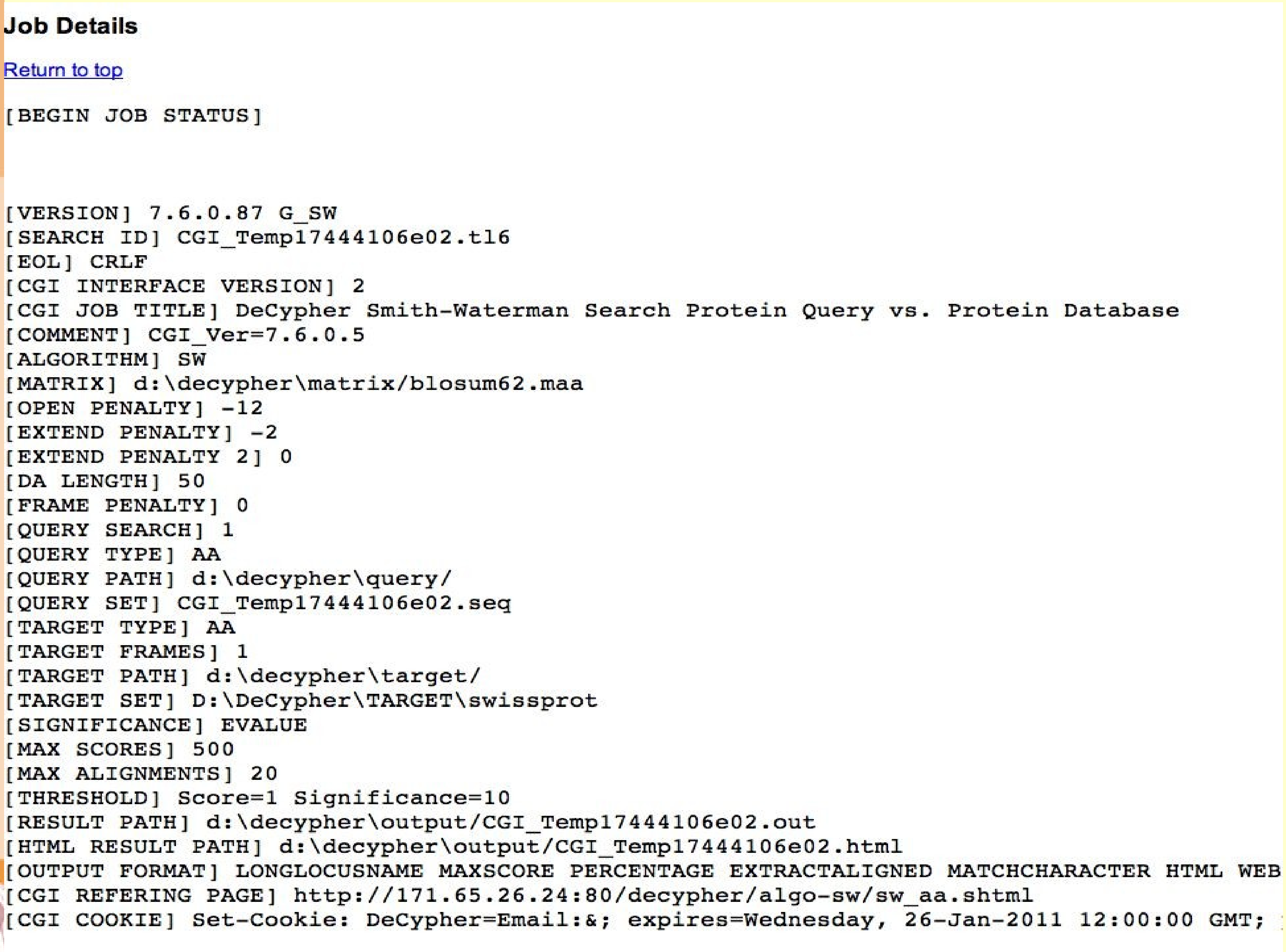
http://decypher.stanford.edu/



```
ne Page    Download All Hits (For Internet Explorer E

lts for Job CGI_Temp17444106e02
ypher Smith-Waterman Search Protein Que

ERY LOCUS] HU-NS1

ults for: HU-NS1; (Length=93)

ERY LENGTH] 90
ARCH TYPE] AA (AMINO ACID)
TRIX] d:\decypher\matrix/blos
EN PENALTY] -12.00
TEND PENALTY] -2.00
ALE FACTOR] 1
imum possible score for this

K       SCORE  QF  TARGET|ACCESSI TF  E_VALUE DESCRIPTION
       432.00   1  sp|P0ACF6|DBHB_ 1  2.2e-051 DNA-binding protein HU-beta OS=Escherichia coli
       432.00   1  sp|P0ACF5|DBHB_ 1  2.2e-051 DNA-binding protein HU-beta OS=Escherichia coli
       432.00   1  sp|P0ACF4|DBHB_ 1  2.2e-051 DNA-binding protein HU-beta OS=Escherichia coli
       432.00   1  sp|P0ACF7|DBHB_ 1  2.2e-051 DNA-binding protein HU-beta OS=Shigella flexneri
       428.00   1  sp|P0A1R9|DBHB_ 1  7.9e-051 DNA-binding protein HU-beta OS=Salmonella typhi
       428.00   1  sp|P0A1R8|DBHB_ 1  7.9e-051 DNA-binding protein HU-beta OS=Salmonella typhim
       394.00   1  sp|P52681|DBHB_ 1  3.8e-046 DNA-binding protein HU-beta OS=Serratia marcesce
       356.00   1  sp|P05384|DBHB_ 1  6.7e-041 DNA-binding protein HU-beta OS=Pseudomonas aerug
       331.00   1  sp|Q9KQS9|DBHB_ 1  1.9e-037 DNA-binding protein HU-beta OS=Vibrio cholerae G
       324.00   1  sp|Q9KHS6|DBHB_ 1  1.7e-036 DNA-binding protein HU-beta OS=Pseudomonas fluor
       311.00   1  sp|Q9LA96|DBHA_ 1  1.1e-034 DNA-binding protein HU-alpha OS=Aeromonas hydrop
       301.00   1  sp|P52680|DBHA_ 1  2.6e-033 DNA-binding protein HU-alpha OS=Serratia marcesc
       298.00   1  sp|P0ACF2|DBHA_ 1  6.7e-033 DNA-binding protein HU-alpha OS=Escherichia coli
       298.00   1  sp|P0ACF1|DBHA_ 1  6.7e-033 DNA-binding protein HU-alpha OS=Escherichia coli
       298.00   1  sp|P0ACF0|DBHA_ 1  6.7e-033 DNA-binding protein HU-alpha OS=Escherichia coli
       298.00   1  sp|P0ACF3|DBHA_ 1  6.7e-033 DNA-binding protein HU-alpha OS=Shigella flexner
       297.00   1  sp|P0A1R7|DBHA_ 1  9.2e-033 DNA-binding protein HU-alpha OS=Salmonella typhi
       297.00   1  sp|P0A1R6|DBHA_ 1  9.2e-033 DNA-binding protein HU-alpha OS=Salmonella typhi
       297.00   1  sp|Q87E48|DBH__ 1  9.2e-033 DNA-binding protein HU OS=Xylella fastidiosa (st
       296.00   1  sp|P64388|DBHB_ 1  1.3e-032 DNA-binding protein HU-beta OS=Neisseria meningi
       296.00   1  sp|P64389|DBHB_ 1  1.3e-032 DNA-binding protein HU-beta OS=Neisseria meningi
       295.00   1  sp|P28080|DBHA_ 1  1.7e-032 DNA-binding protein HU-alpha OS=Vibrio proteolyt
       284.00   1  sp|Q9PE38|DBH__ 1  5.7e-031 DNA-binding protein HU OS=Xylella fastidiosa GN=
       283.00   1  sp|P43722|DBH__ 1  7.9e-031 DNA-binding protein HU OS=Haemophilus influenzae
```

Decypher Search Engine Summary

http://decypher.stanford.edu/



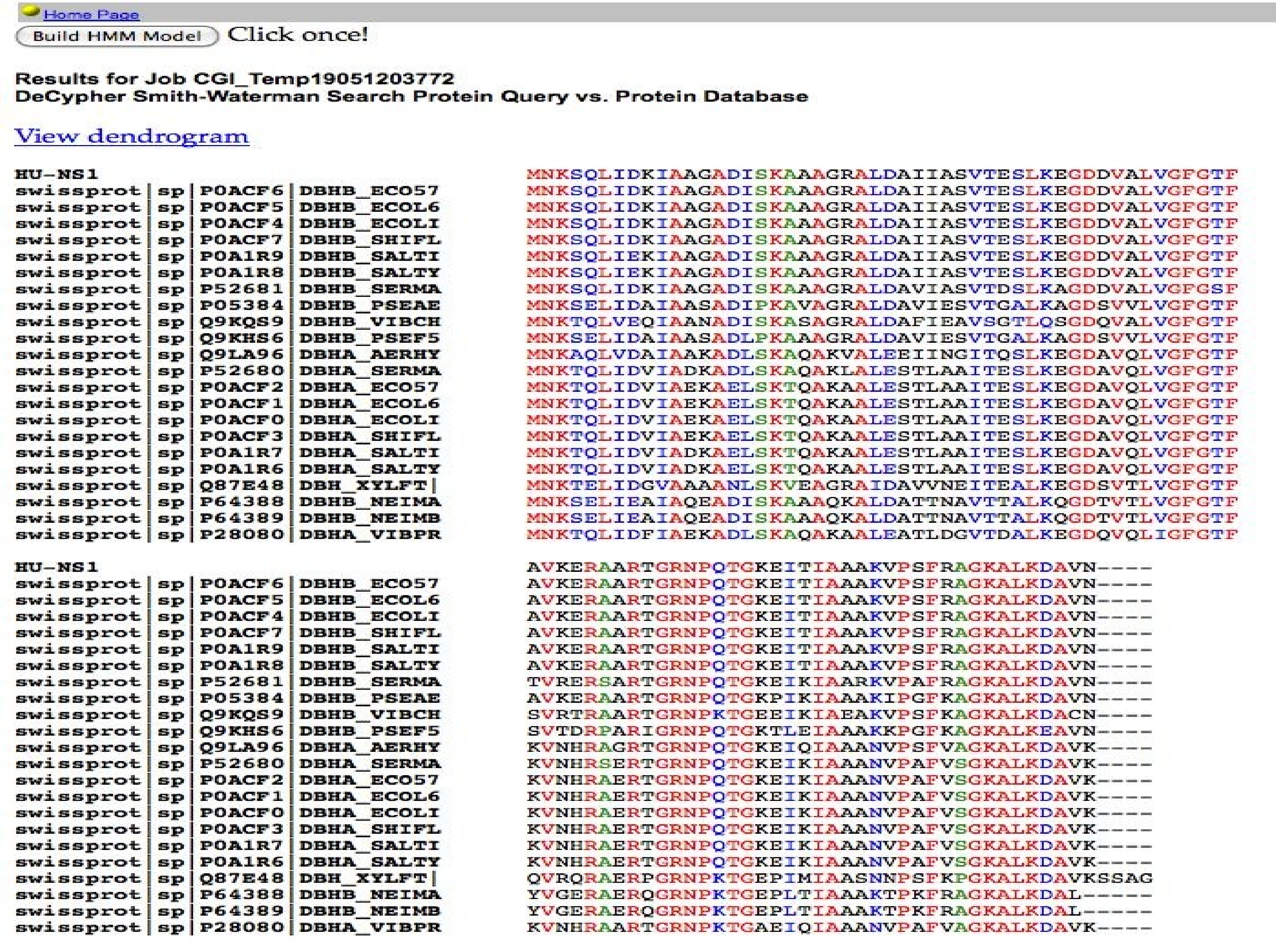
Job Details

Return to top

```
[BEGIN JOB STATUS]



[VERSION] 7.6.0.87 G_SW
[SEARCH ID] CGI_Temp17444106e02.tl6
[EOL] CRLF
[CGI INTERFACE VERSION] 2
[CGI JOB TITLE] DeCypher Smith-Waterman Search Protein Query vs. Protein Database
[COMMENT] CGI_Ver=7.6.0.5
[ALGORITHM] SW
[MATRIX] d:\decypher\matrix/blosum62.maa
[OPEN PENALTY] -12
[EXTEND PENALTY] -2
[EXTEND PENALTY 2] 0
[DA LENGTH] 50
[FRAME PENALTY] 0
[QUERY SEARCH] 1
[QUERY TYPE] AA
[QUERY PATH] d:\decypher\query/
[QUERY SET] CGI_Temp17444106e02.seq
[TARGET TYPE] AA
[TARGET FRAMES] 1
[TARGET PATH] d:\decypher\target/
[TARGET SET] D:\DeCypher\TARGET\swissprot
[SIGNIFICANCE] EVALUE
[MAX SCORES] 500
[MAX ALIGNMENTS] 20
[THRESHOLD] Score=1 Significance=10
[RESULT PATH] d:\decypher\output/CGI_Temp17444106e02.out
[HTML RESULT PATH] d:\decypher\output/CGI_Temp17444106e02.html
[OUTPUT FORMAT] LONGLOCUSNAME MAXSCORE PERCENTAGE EXTRACTALIGNED MATCHCHARACTER HTML WEB
[CGI REFERING PAGE] http://171.65.26.24:80/decypher/algo-sw/sw_aa.shtml
[CGI COOKIE] Set-Cookie: DeCypher=Email:&; expires=Wednesday, 26-Jan-2011 12:00:00 GMT;
```

Decypher Search Sequence Alignments

http://decypher.stanford.edu/



Build HMM Model Click once!

Results for Job CGI_Temp19051203772
DeCypher Smith-Waterman Search Protein Query vs. Protein Database

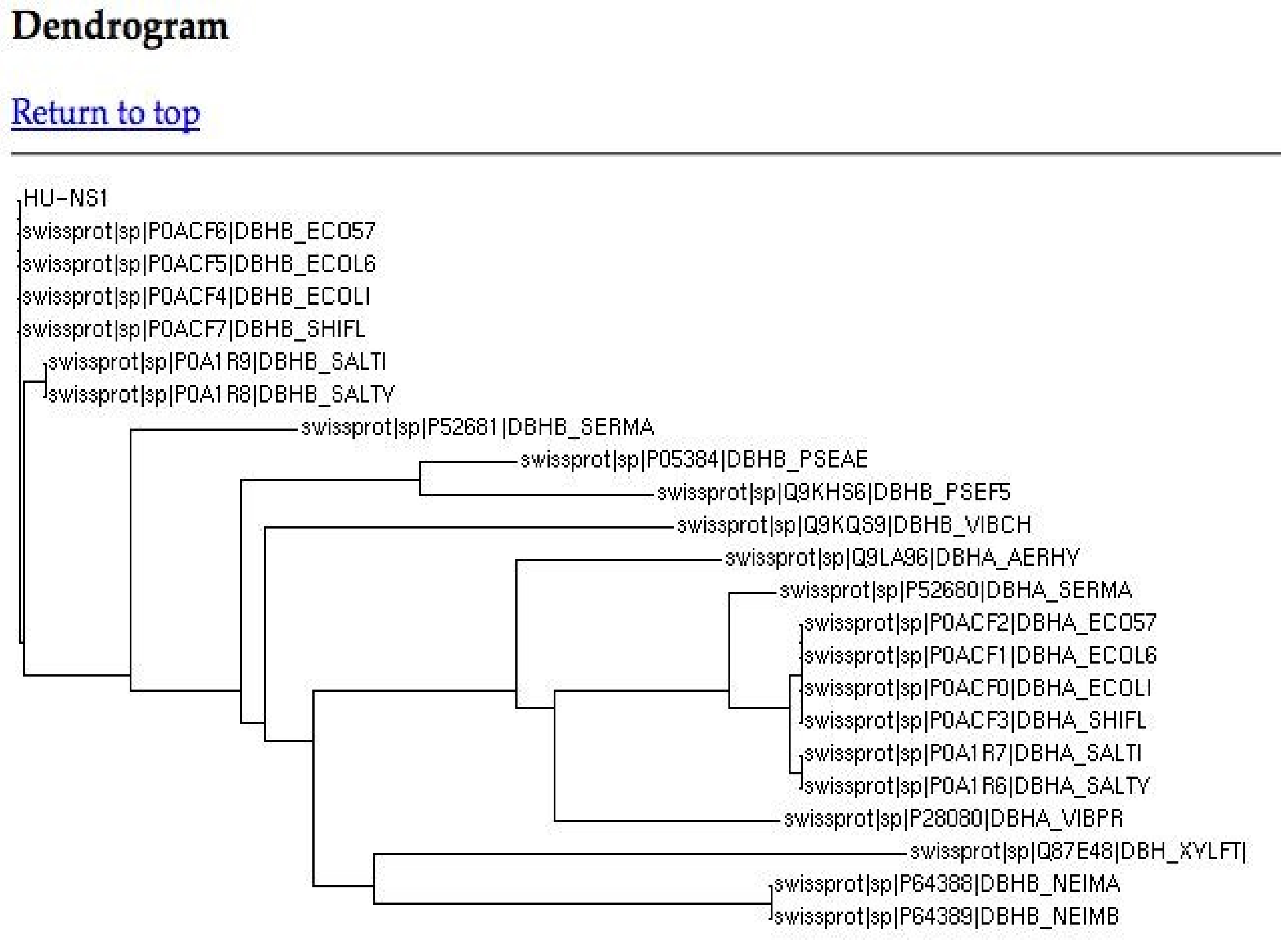
View dendrogram

```
HU-NS1                                  MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTF
swissprot|sp|P0ACF6|DBHB_ECO57          MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTF
swissprot|sp|P0ACF5|DBHB_ECOL6          MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTF
swissprot|sp|P0ACF4|DBHB_ECOLI          MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTF
swissprot|sp|P0ACF7|DBHB_SHIFL          MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTF
swissprot|sp|P0A1R9|DBHB_SALTI          MNKSQLIEKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTF
swissprot|sp|P0A1R8|DBHB_SALTY          MNKSQLIEKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTF
swissprot|sp|P52681|DBHB_SERMA          MNKSQLIDKIAAGADISKAAAGRALDAVIASVTDSLKAGDDVALVGFGSF
swissprot|sp|P05384|DBHB_PSEAE          MNKSELIDAIAASADIPKAVAGRALDAVIESVTGALKAGDSVVLVGFGTF
swissprot|sp|Q9KQS9|DBHB_VIBCH          MNKTQLVEQIAANADISKASAGRALDAFIEAVSGTLQSGDQVALVGFGTF
swissprot|sp|Q9KHS6|DBHB_PSEF5          MNKSELIDAIAASADLPKAAAGRALDAVIESVTGALKAGDSVVLVGFGTF
swissprot|sp|Q9LA96|DBHA_AERHY          MNKAQLVDAIAAKADLSKAQAKVALEEIINGITQSLKEGDAVQLVGFGTF
swissprot|sp|P52680|DBHA_SERMA          MNKTQLIDVIADKADLSKAQAKLALESTLAAITESLKEGDAVQLVGFGTF
swissprot|sp|P0ACF2|DBHA_ECO57          MNKTQLIDVIAEKAELSKTQAKAALESTLAAITESLKEGDAVQLVGFGTF
swissprot|sp|P0ACF1|DBHA_ECOL6          MNKTQLIDVIAEKAELSKTQAKAALESTLAAITESLKEGDAVQLVGFGTF
swissprot|sp|P0ACF0|DBHA_ECOLI          MNKTQLIDVIAEKAELSKTQAKAALESTLAAITESLKEGDAVQLVGFGTF
swissprot|sp|P0ACF3|DBHA_SHIFL          MNKTQLIDVIAEKAELSKTQAKAALESTLAAITESLKEGDAVQLVGFGTF
swissprot|sp|P0A1R7|DBHA_SALTI          MNKTQLIDVIADKAELSKTQAKAALESTLAAITESLKEGDAVQLVGFGTF
swissprot|sp|P0A1R6|DBHA_SALTY          MNKTQLIDVIADKAELSKTQAKAALESTLAAITESLKEGDAVQLVGFGTF
swissprot|sp|Q87E48|DBH_XYLFT|          MNKTELIDGVAAAANLSKVEAGRAIDAVVNEITEALKEGDSVTLVGFGTF
swissprot|sp|P64388|DBHB_NEIMA          MNKSELIEAIAQEADISKAAAQKALDATTNAVTTALKQGDTVTLVGFGTF
swissprot|sp|P64389|DBHB_NEIMB          MNKSELIEAIAQEADISKAAAQKALDATTNAVTTALKQGDTVTLVGFGTF
swissprot|sp|P28080|DBHA_VIBPR          MNKTQLIDFIAEKADLSKAQAKAALEATLDGVTDALKEGDQVQLIGFGTF

HU-NS1                                  AVKERAARTGRNPQTGKEITIAAAKVPSFRAGKALKDAVN----
swissprot|sp|P0ACF6|DBHB_ECO57          AVKERAARTGRNPQTGKEITIAAAKVPSFRAGKALKDAVN----
swissprot|sp|P0ACF5|DBHB_ECOL6          AVKERAARTGRNPQTGKEITIAAAKVPSFRAGKALKDAVN----
swissprot|sp|P0ACF4|DBHB_ECOLI          AVKERAARTGRNPQTGKEITIAAAKVPSFRAGKALKDAVN----
swissprot|sp|P0ACF7|DBHB_SHIFL          AVKERAARTGRNPQTGKEITIAAAKVPSFRAGKALKDAVN----
swissprot|sp|P0A1R9|DBHB_SALTI          AVKERAARTGRNPQTGKEITIAAAKVPSFRAGKALKDAVN----
swissprot|sp|P0A1R8|DBHB_SALTY          AVKERAARTGRNPQTGKEITIAAAKVPSFRAGKALKDAVN----
swissprot|sp|P52681|DBHB_SERMA          TVRERSARTGRNPQTGKEIKIAARKVPAFRAGKALKDAVN----
swissprot|sp|P05384|DBHB_PSEAE          AVKERAARTGRNPQTGKPIKIAAAKIPGFKAGKALKDAVN----
swissprot|sp|Q9KQS9|DBHB_VIBCH          SVRTRAARTGRNPKTGEEIKIAEAKVPSFKAGKALKDACN----
swissprot|sp|Q9KHS6|DBHB_PSEF5          SVTDRPARIGRNPQTGKTLEIAAAKKPGFKAGKALKEAVN----
swissprot|sp|Q9LA96|DBHA_AERHY          KVNHRAGRTGRNPQTGKEIQIAAANVPSFVAGKALKDAVK----
swissprot|sp|P52680|DBHA_SERMA          KVNHRSERTGRNPQTGKEIKIAAANVPAFVSGKALKDAVK----
swissprot|sp|P0ACF2|DBHA_ECO57          KVNHRAERTGRNPQTGKEIKIAAANVPAFVSGKALKDAVK----
swissprot|sp|P0ACF1|DBHA_ECOL6          KVNHRAERTGRNPQTGKEIKIAAANVPAFVSGKALKDAVK----
swissprot|sp|P0ACF0|DBHA_ECOLI          KVNHRAERTGRNPQTGKEIKIAAANVPAFVSGKALKDAVK----
swissprot|sp|P0ACF3|DBHA_SHIFL          KVNHRAERTGRNPQTGKEIKIAAANVPAFVSGKALKDAVK----
swissprot|sp|P0A1R7|DBHA_SALTI          KVNHRAERTGRNPQTGKEIKIAAANVPAFVSGKALKDAVK----
swissprot|sp|P0A1R6|DBHA_SALTY          KVNHRAERTGRNPQTGKEIKIAAANVPAFVSGKALKDAVK----
swissprot|sp|Q87E48|DBH_XYLFT|          QVRQRAERPGRNPKTGEPIMIAASNNPSFKPGKALKDAVKSSAG
swissprot|sp|P64388|DBHB_NEIMA          YVGERAERQGRNPKTGEPLTIAAAKTPKFRAGKALKDAL-----
swissprot|sp|P64389|DBHB_NEIMB          YVGERAERQGRNPKTGEPLTIAAAKTPKFRAGKALKDAL-----
swissprot|sp|P28080|DBHA_VIBPR          KVNHRAARTGRNPKTGAEIQIAAANVPAFVAGKALKDAVK----
```

TimeLogic®
biocomputing solutions

Decypher Search Sequence Alignments

http://decypher.stanford.edu/



Dendrogram

Return to top

HU-NS1
swissprot|sp|P0ACF6|DBHB_ECO57
swissprot|sp|P0ACF5|DBHB_ECOL6
swissprot|sp|P0ACF4|DBHB_ECOLI
swissprot|sp|P0ACF7|DBHB_SHIFL
swissprot|sp|P0A1R9|DBHB_SALTI
swissprot|sp|P0A1R8|DBHB_SALTY
swissprot|sp|P52681|DBHB_SERMA
swissprot|sp|P05384|DBHB_PSEAE
swissprot|sp|Q9KHS6|DBHB_PSEF5
swissprot|sp|Q9KQS9|DBHB_VIBCH
swissprot|sp|Q9LA96|DBHA_AERHY
swissprot|sp|P52680|DBHA_SERMA
swissprot|sp|P0ACF2|DBHA_ECO57
swissprot|sp|P0ACF1|DBHA_ECOL6
swissprot|sp|P0ACF0|DBHA_ECOLI
swissprot|sp|P0ACF3|DBHA_SHIFL
swissprot|sp|P0A1R7|DBHA_SALTI
swissprot|sp|P0A1R6|DBHA_SALTY
swissprot|sp|P28080|DBHA_VIBPR
swissprot|sp|Q87E48|DBH_XYLFT|
swissprot|sp|P64388|DBHB_NEIMA
swissprot|sp|P64389|DBHB_NEIMB

General DNA Similarity Search Principles

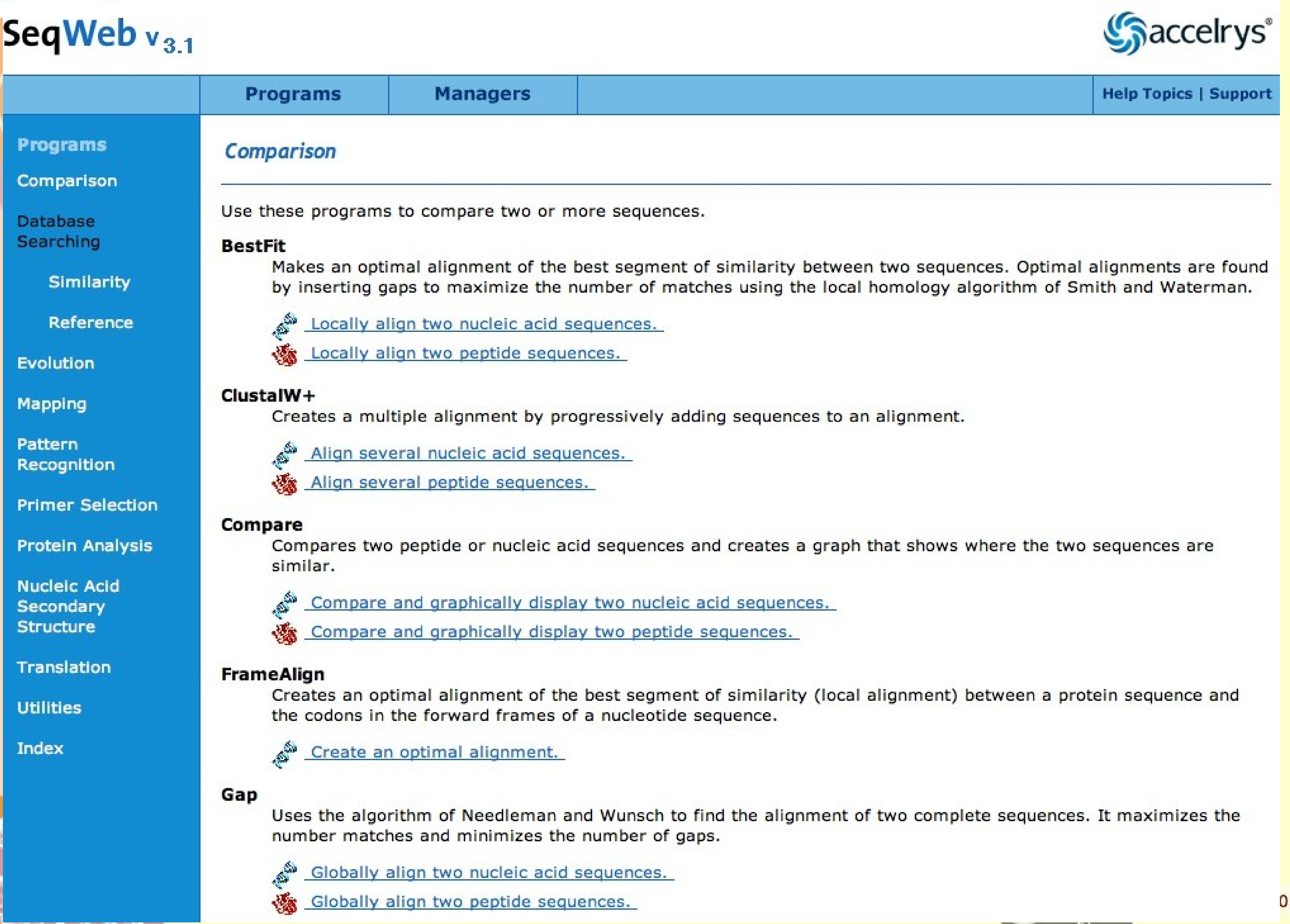
- Search both Strands
- Translate ORFs
- Use most sensitive search possible
  - UnGapped BLAST for infinite gap penalty (PCR & CHIP oligos)
  - Gapped BLAST for most searches
  - Smith Waterman or megaBLAST or discontinuous MegaBLAST for cDNA/genome comparisons
  - cDNA =>Zero gap-length penalty
  - Consider using transition matrices
  - Ensure that expected value of score is negative
- Examine results with exp. between 0.05 and 10
- Reevaluate results of borderline significance using limited query

General Protein Similarity Search Principles

- Chose between local or global search algorithm
- Use most sensitive search algorithm available
  - Original BLAST for no gaps
  - Smith-Waterman for most flexibility
  - Gapped BLAST for well delimited regions
  - PSI-BLAST for families
  - Initially BLOSUM62 and default gap penalties
  - If no significant results, use BLOSUM30 and lower gap penalties
  - Ensure expected score is negative
- Examine results between exp. 0.05 and 10 for biological significance
- Beware of long hits or those with unusual amino acid composition
- Reevaluate results of borderline significance using limited query

SeqWeb Comparison Programs

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



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.

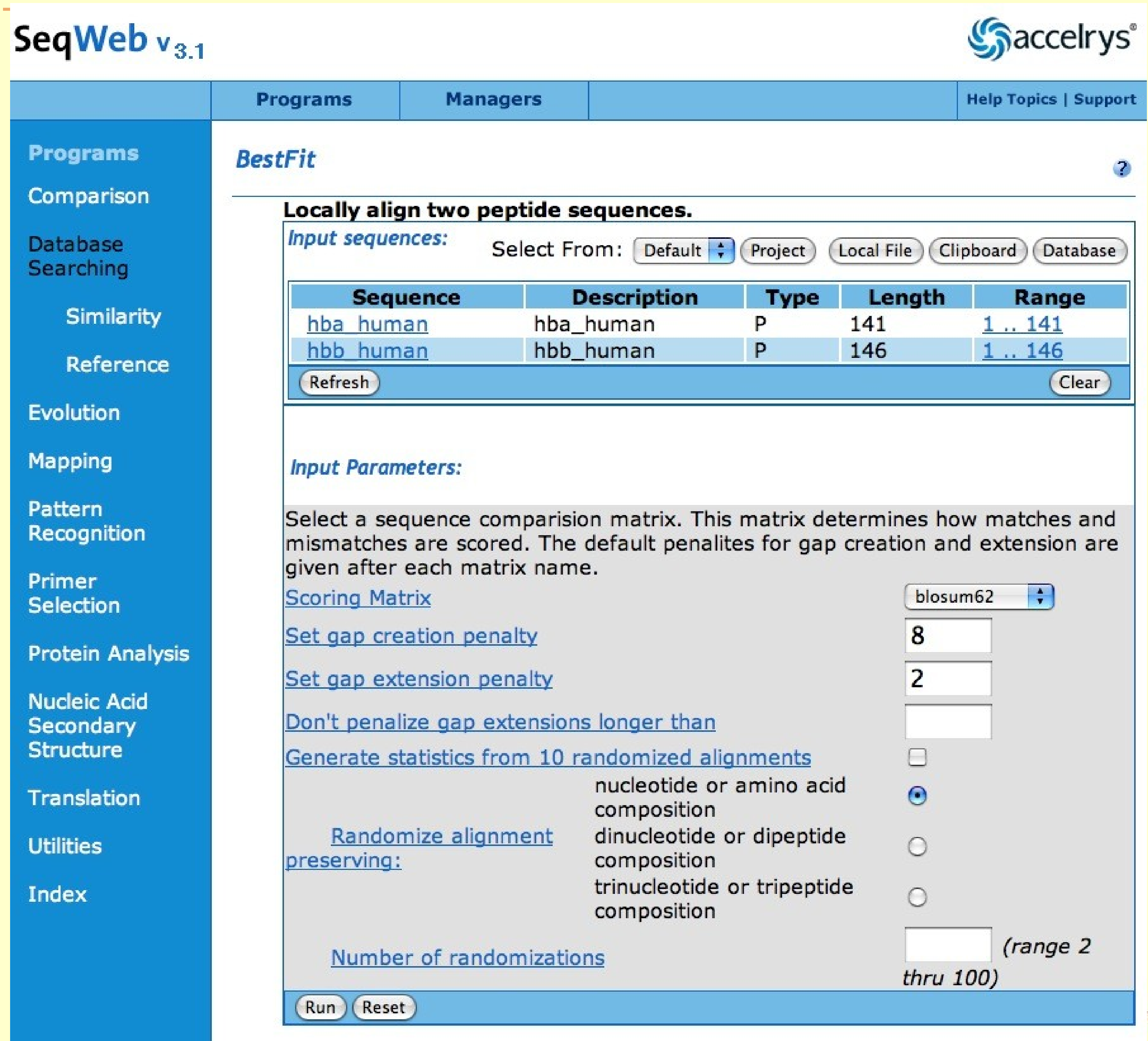
Globally align two peptide sequences.

SeqWeb v3.1

accelrys

SeqWeb BestFit Protein Program

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



BestFit

Locally align two peptide sequences.

Input sequences: Select From:

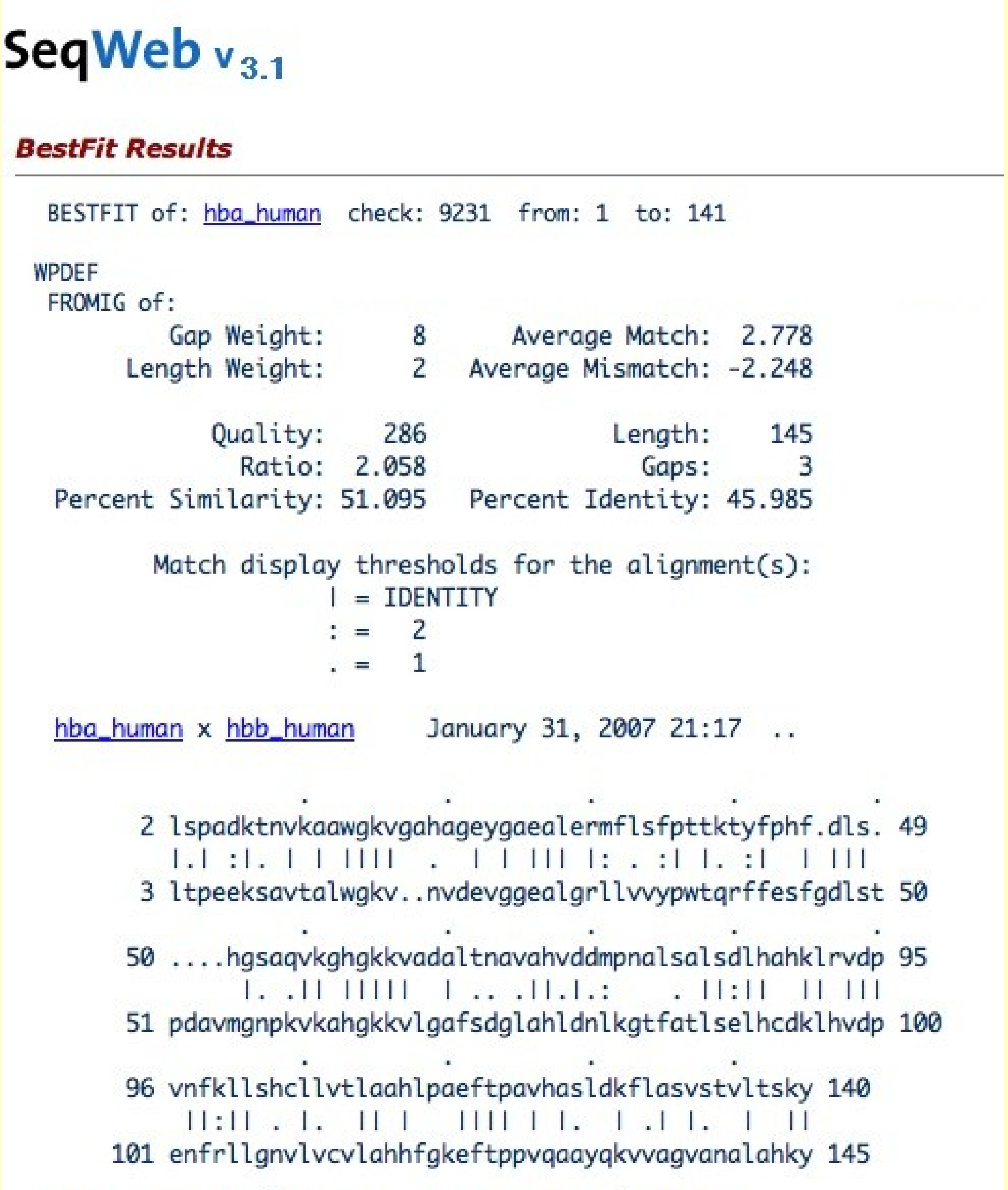
Sequence	Description	Type	Length	Range
hba_human	hba_human	P	141	1 .. 141
hbb_human	hbb_human	P	146	1 .. 146

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)*

BestFit Alignments (Gap =8)

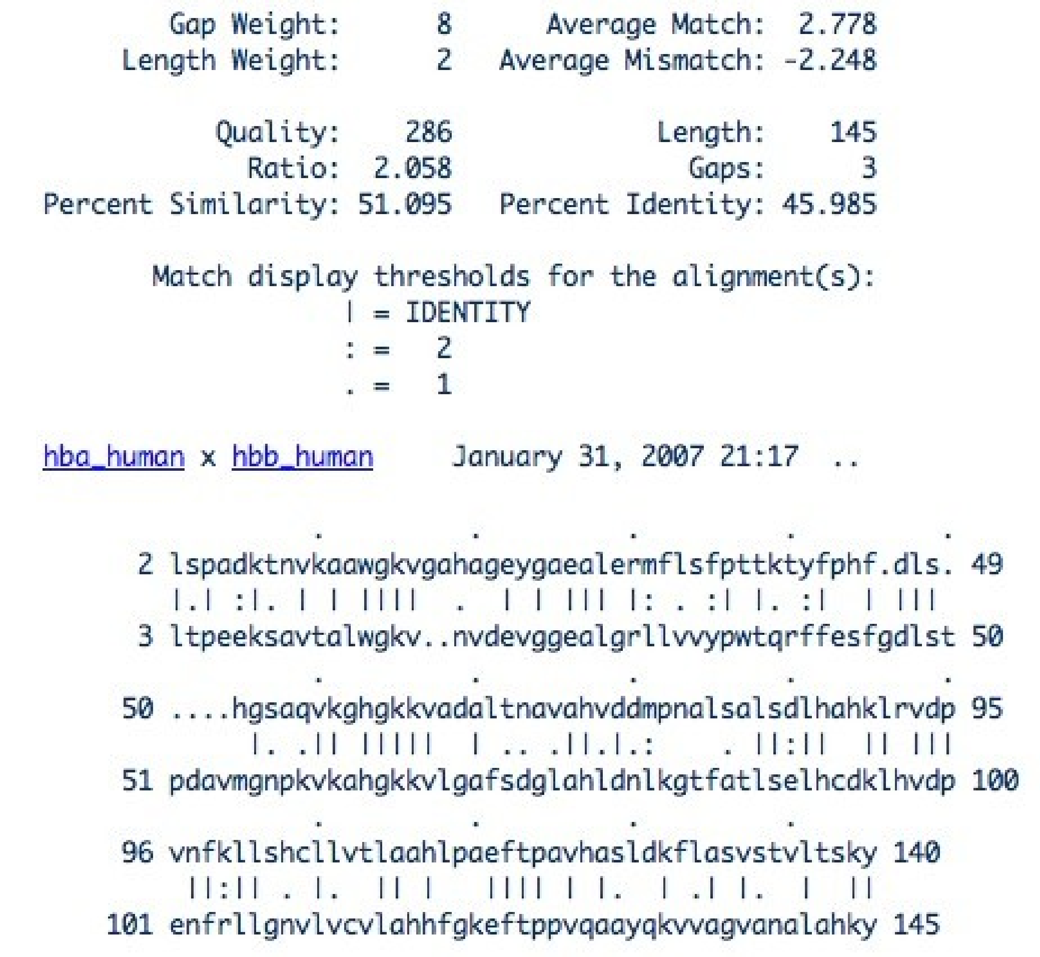


SeqWeb v3.1

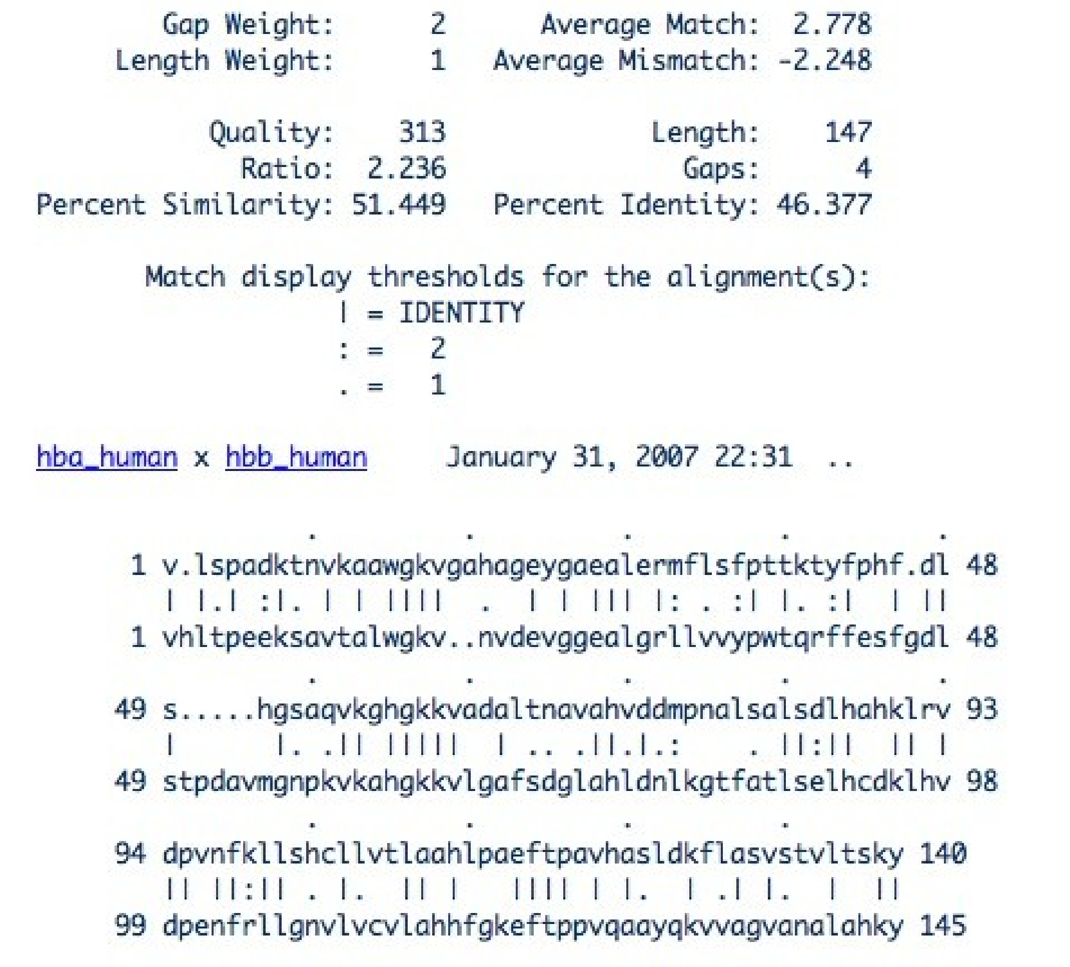
BestFit Results

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

WPDEF
 FROMIG of:
         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 31, 2007 21:17  ..

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

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

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

BestFit Alignments (Gap =8&2)



```
         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 31, 2007 21:17  ..


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



```
         Gap Weight:      2      Average Match:  2.778
      Length Weight:      1   Average Mismatch: -2.248

            Quality:    313             Length:    147
              Ratio:  2.236               Gaps:      4
 Percent Similarity: 51.449   Percent Identity: 46.377

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

hba_human x hbb_human      January 31, 2007 22:31  ..


                .         .         .         .         .
 1 v.lspadktnvkaawgkvgahageygaealermflsfpttktyfphf.dl 48
   | |.| :|. | | ||||  .  | | ||| |: . :| |. :|  | ||
 1 vhltpeeksavtalwgkv..nvdevggealgrllvvypwtqrffesfgdl 48
                .         .         .         .         .
49 s.....hgsaqvkghgkkvadaltnavahvddmpnalsalsdlhahklrv 93
   |      |. .|| |||||  | .. .||.|.:    . ||:||  || |
49 stpdavmgnpkvkahgkkvlgafsdglahldnlkgtfatlselhcdklhv 98
                .         .         .         .
94 dpvnfkllshcllvtlaahlpaeftpavhasldkflasvstvltsky 140
   || ||:|| . |.  || |   |||| | |.  | .| |.  |  ||
99 dpenfrllgnvlvcvlahhfgkeftppvqaayqkvvagvanalahky 145
```

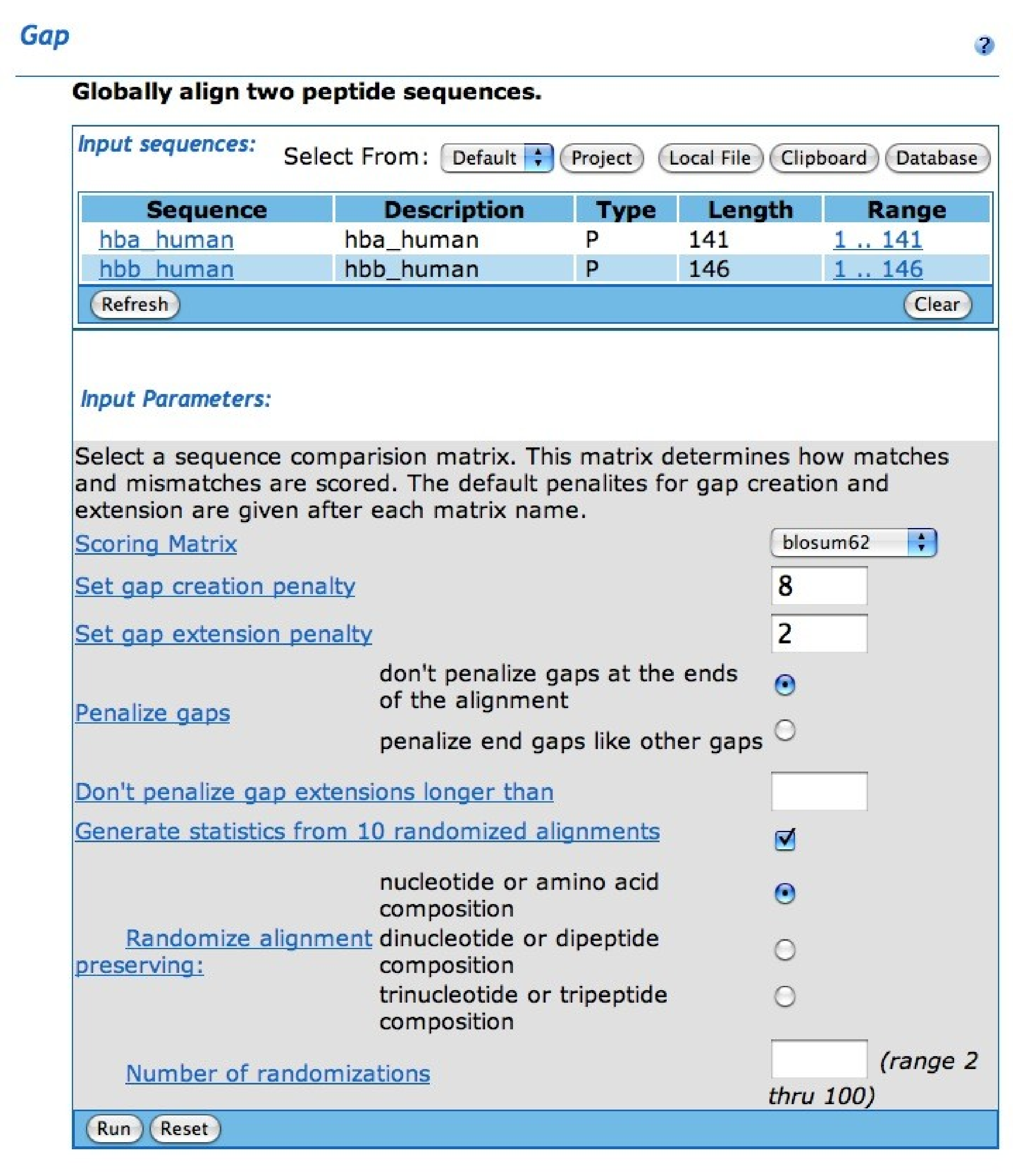
Doug Brutlag 2010

SeqWeb v3.1

accelrys®

SeqWeb Gap Protein Alignments

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

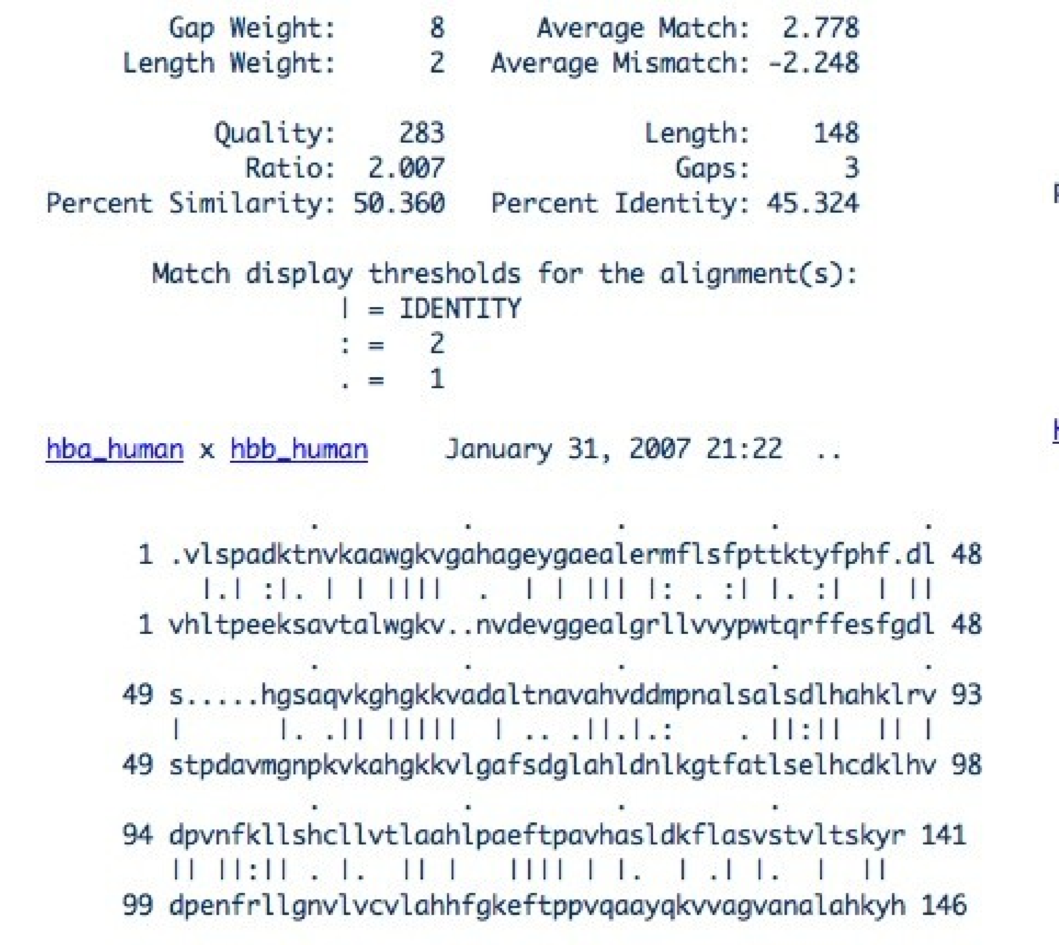
Input Parameters:

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.

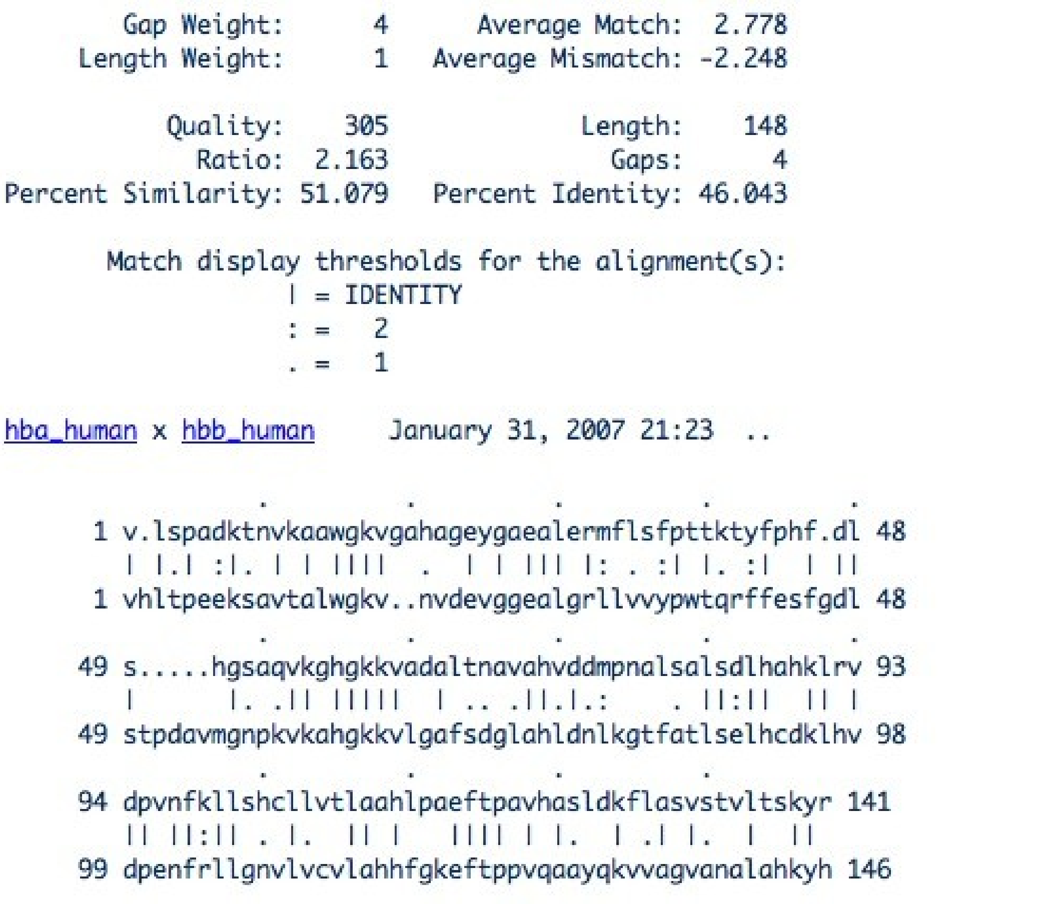
Scoring Matrix		blosum62
Set gap creation penalty		8
Set gap extension penalty		2
Penalize gaps	don't penalize gaps at the ends of the alignment	●
	penalize end gaps like other gaps	○
Don't penalize gap extensions longer than		
Generate statistics from 10 randomized alignments		☑
Randomize alignment 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 (Gap 8&4)



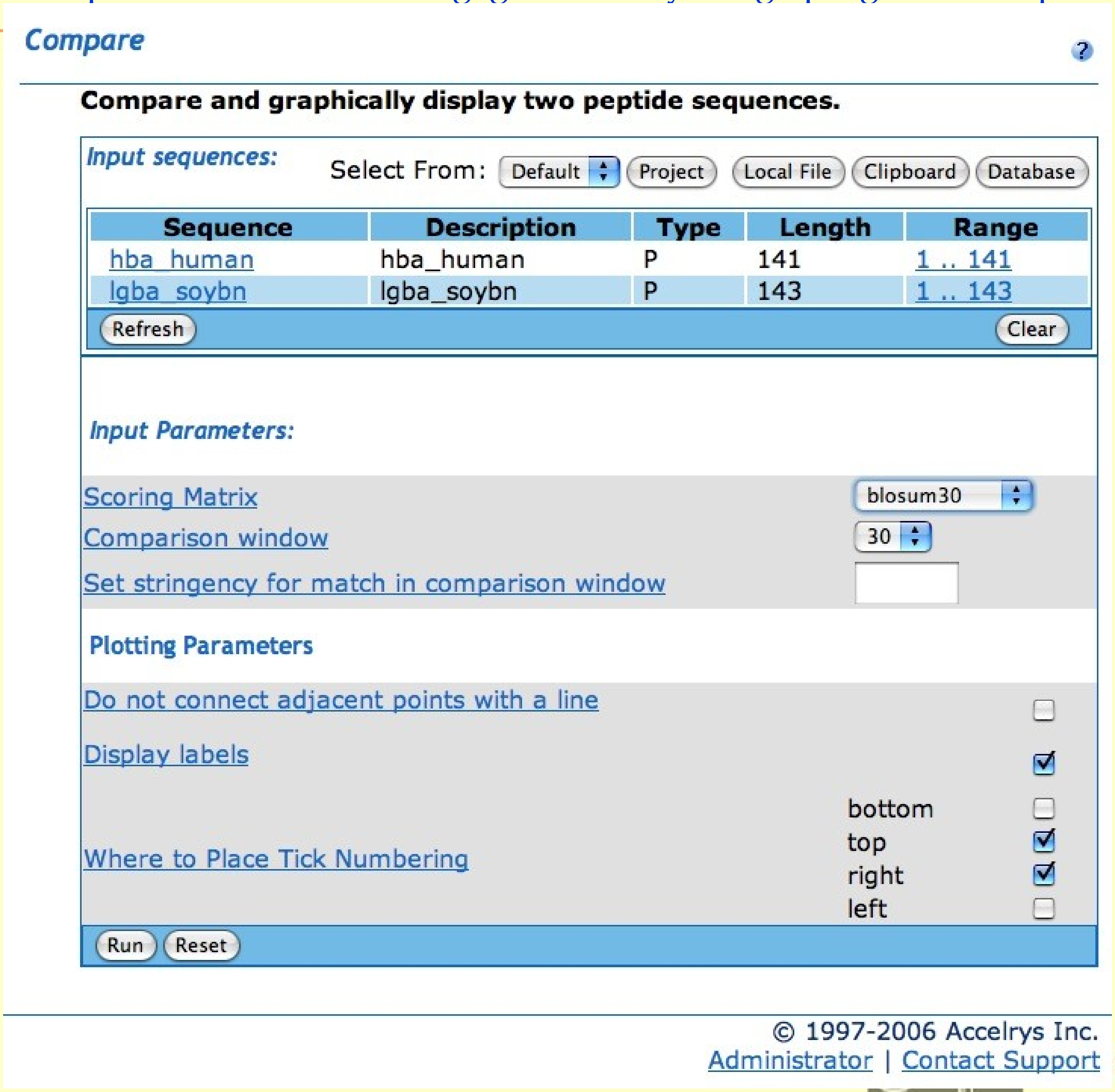
```
        Gap Weight:      8      Average Match:   2.778
     Length Weight:      2   Average Mismatch:  -2.248

           Quality:    283             Length:     148
             Ratio:  2.007               Gaps:       3
Percent Similarity: 50.360   Percent Identity:  45.324

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

hba_human x hbb_human      January 31, 2007 21:22  ..

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



```
        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

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

hba_human x hbb_human      January 31, 2007 21:23  ..

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

SeqWeb Compare Proteins

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



Compare

Compare and graphically display two peptide sequences.

Input sequences: Select From:

Sequence	Description	Type	Length	Range
hba_human	hba_human	P	141	1 .. 141
lgba_soybn	lgba_soybn	P	143	1 .. 143

Input Parameters:

Scoring Matrix: blosum30

Comparison window: 30

Set stringency for match in comparison window:

Plotting Parameters

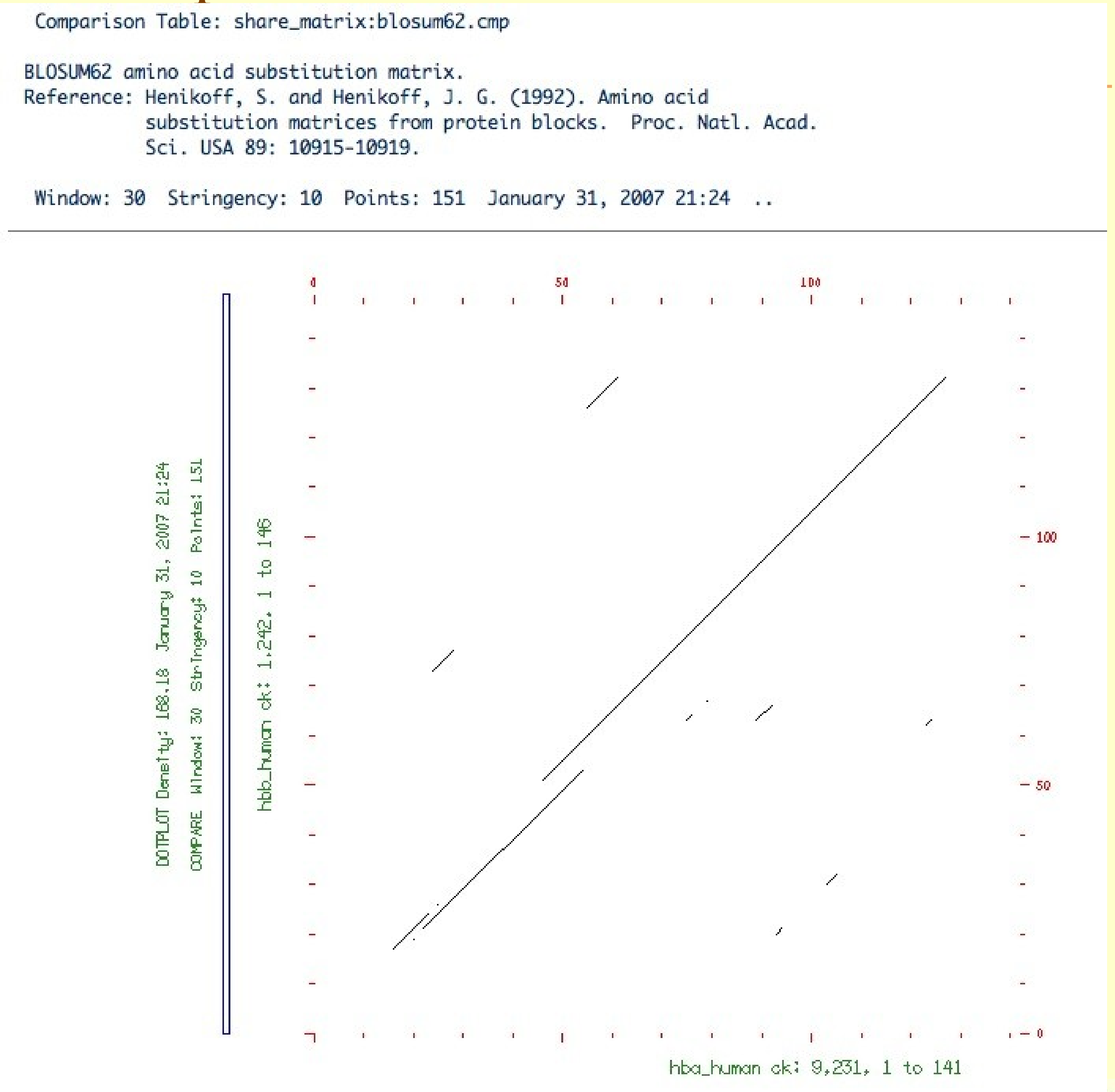
Do not connect adjacent points with a line □

Display labels ☑

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

Administrator | Contact Support

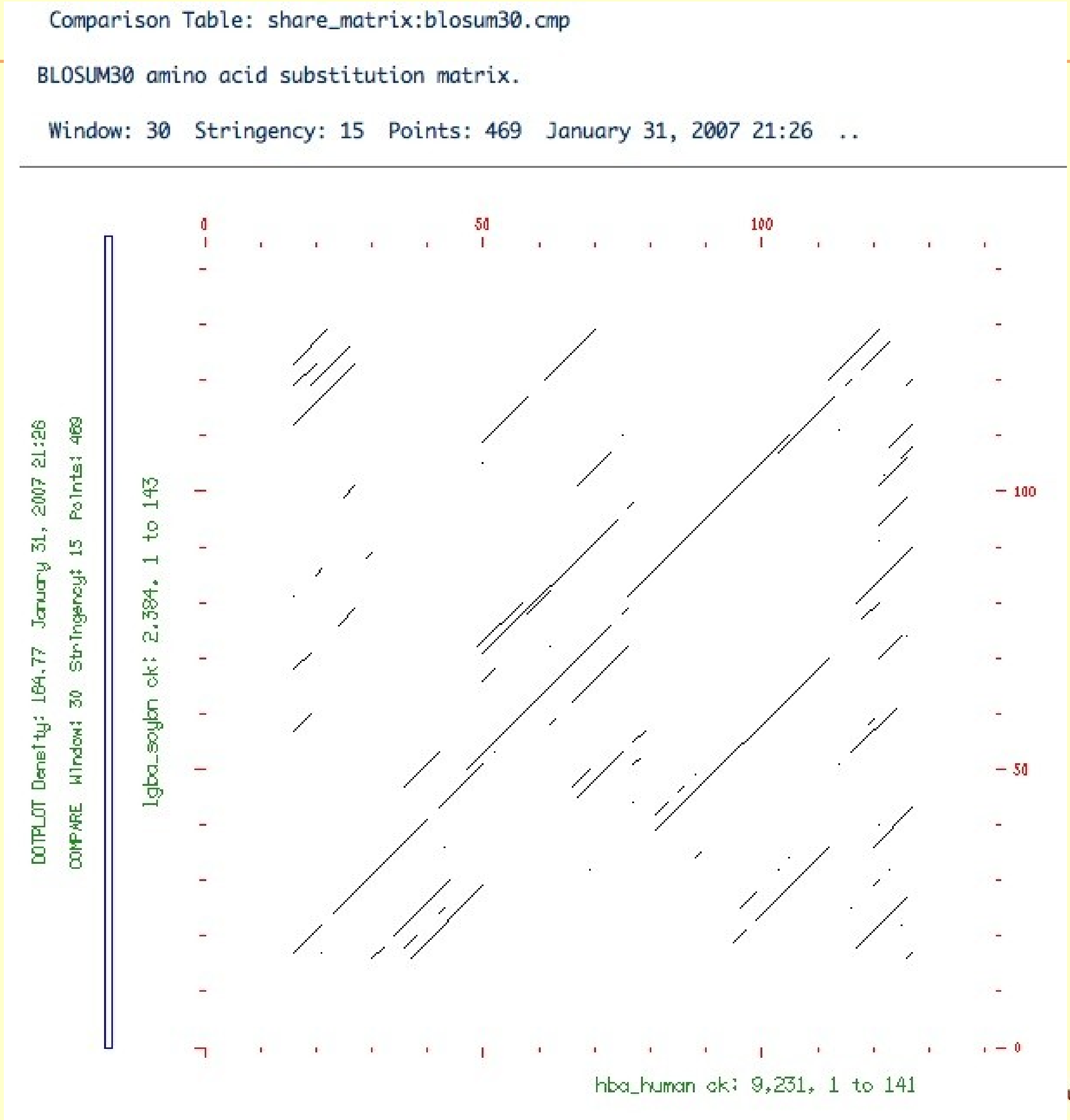
Compare HHA and HHB Human



```
 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 31, 2007 21:24  ..
```

DOTPLOT Density: 168.18 January 31, 2007 21:24
COMPARE Window: 30 Stringency: 10 Points: 151
hbb_human ck: 1,242, 1 to 146
0
50
100
100
50
0
hba_human ck: 9,231, 1 to 141

Compare HHA to Soybean HB



Comparison Table: share_matrix:blosum30.cmp

BLOSUM30 amino acid substitution matrix.

Window: 30 Stringency: 15 Points: 469 January 31, 2007 21:26 ..

0
50
100
DOTPLOT Density: 164.77 January 31, 2007 21:26
COMPARE Window: 30 Stringency: 15 Points: 469
lgba_soybn ck: 2,384, 1 to 143
100
50
0
hba_human ck: 9,231, 1 to 141

ug Brutlag 2010