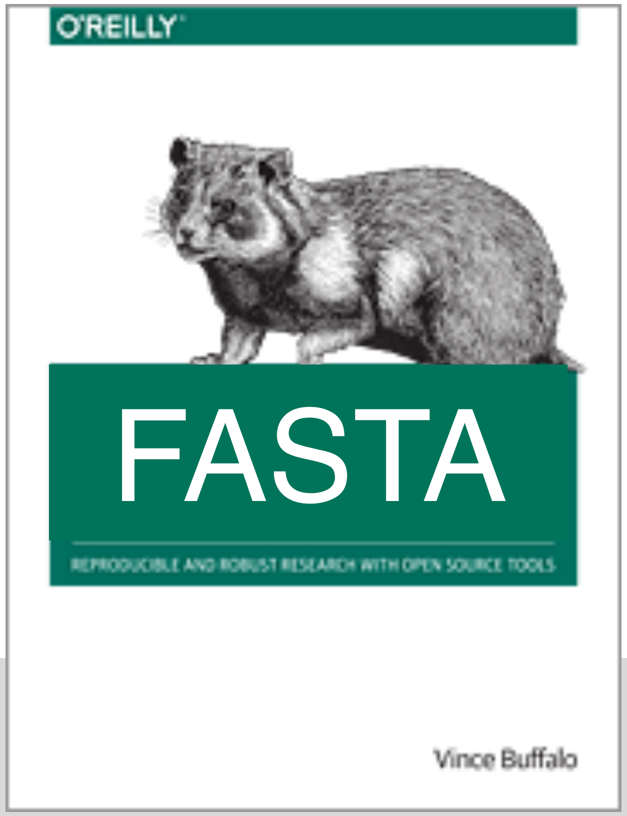
Sequence Homology and Analysis

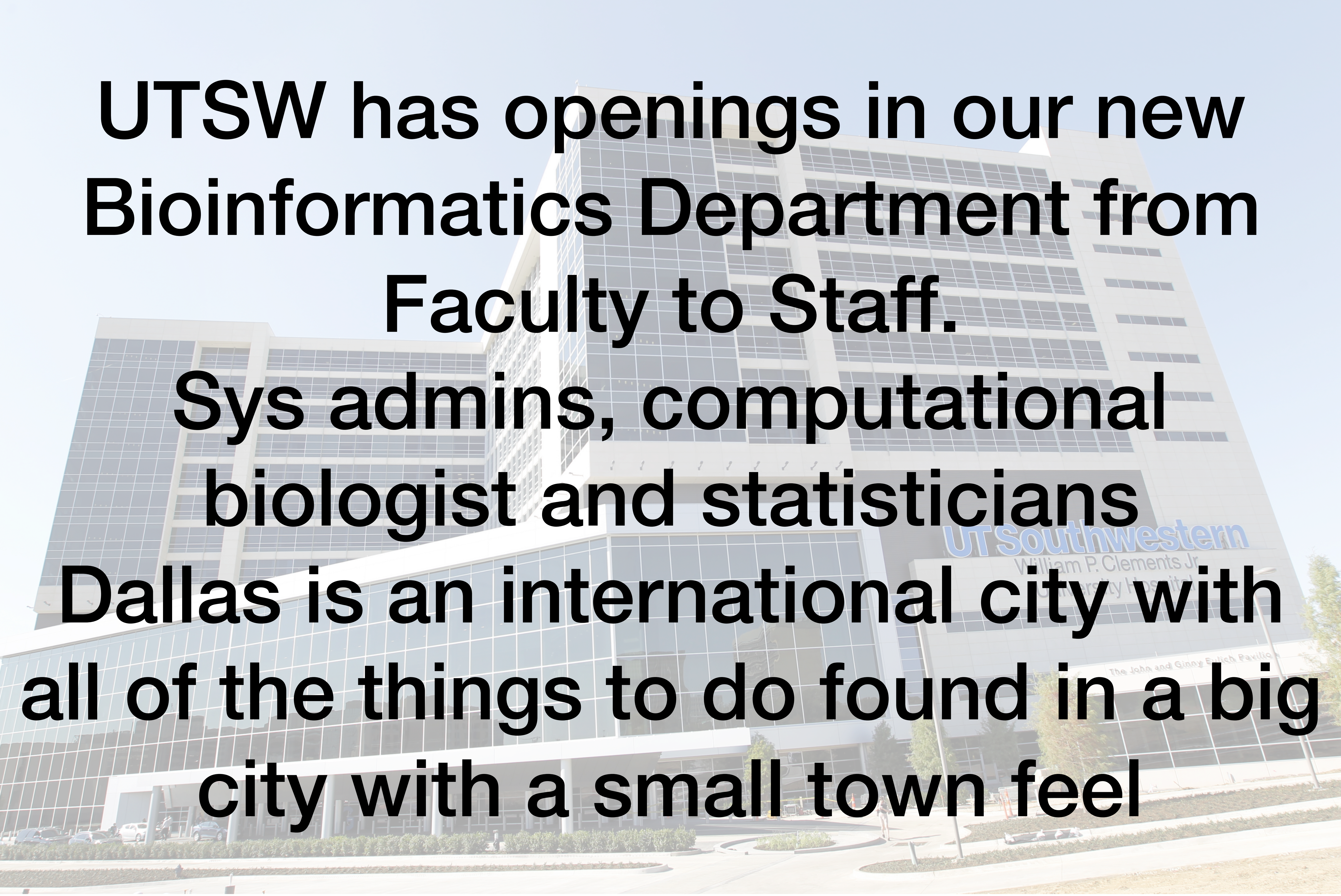
Understanding How FASTA and BLAST work to optimize your sequence similarity searches.

An Essential Guide to the Basic Local Alignment Search Tool
BLAST
O'REILLY
Ian Korf, Mark Yandell & Joseph Bedell



Brandi Cantarel, Ph.D
UTSW, Department of Bioinformatics
Programming for Biology 2018

UTSouthwestern
Medical Center
Lyda Hill Department of Bioinformatics



UTSW has openings in our new Bioinformatics Department from Faculty to Staff.

Sys admins, computational biologist and statisticians

Dallas is an international city with all of the things to do found in a big city with a small town feel

Take Home Messages

1. *Homologous* sequences share a common ancestor, but most sequences are *non-homologous*
2. Compare protein sequence for distant comparison and DNA for close comparisons
3. Sequence Homology can be reliably inferred from statistically significant similarity (non-homology cannot from non-similarity)
4. Homologous proteins share common structures, but not necessarily common functions
5. Sequence statistical significance estimates are accurate (verify this yourself)
6. Smaller databases increase search sensitivity
7. Statistical accuracy can be evaluated by examining the "highest scoring unrelated sequence" or by random shuffles

What is Understanding Homology Important

- Most gene databases use sequence similarity to infer gene function (with a few exceptions)
    - In the absence of high-throughput biochemistry experimentation, homology is used to predict gene function and pathway assignments.
- Many of these predictions are correct however distinguishing orthologs (deviation from speciation) and paralogs (deviation from gene duplication) is difficult
- E-values are more reliable than percent identity

What is Homology?

How do we recognize it?

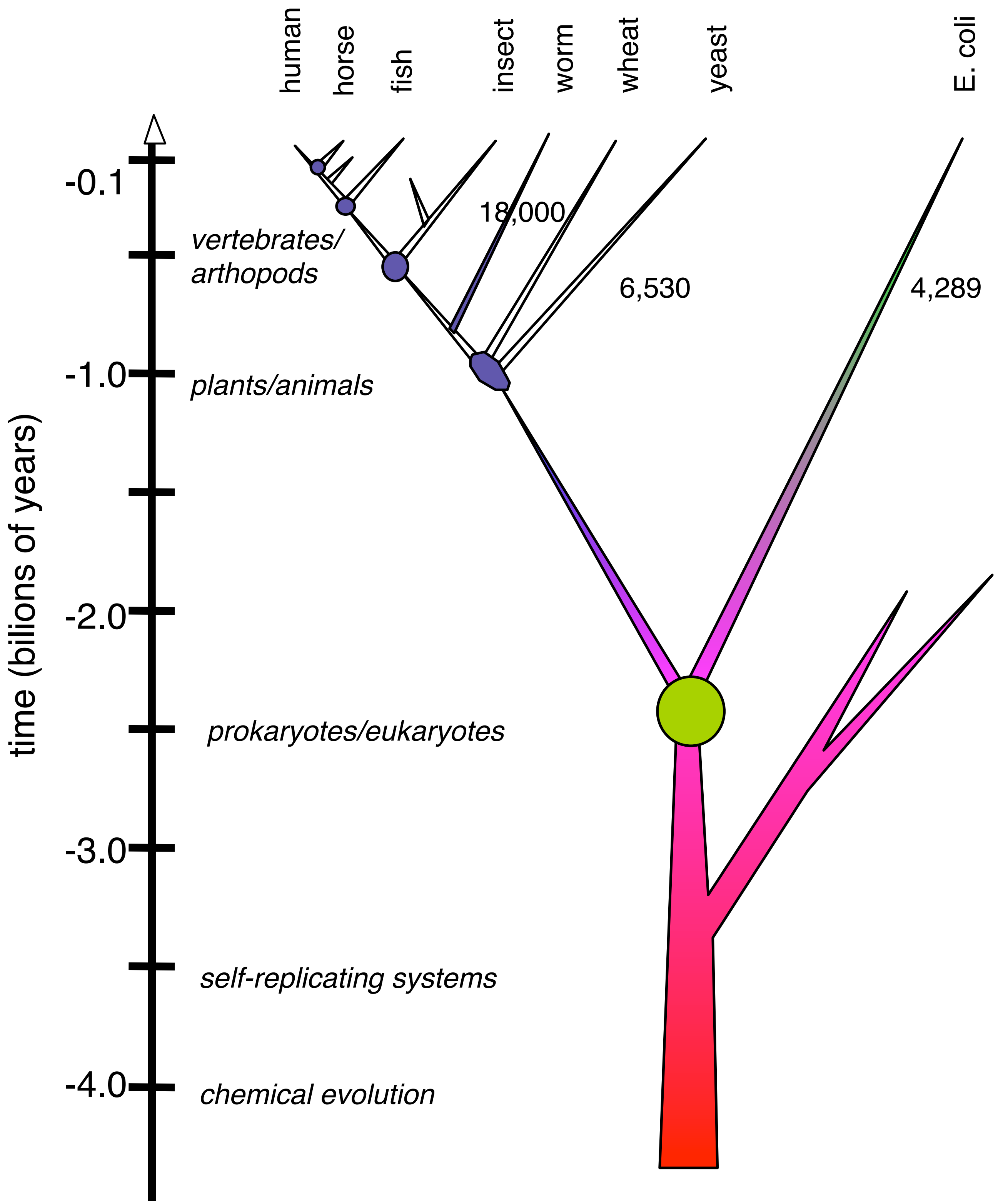
History of Sequence Similarity

1950s	Fredrick Sanger Determines the Sequence of Insulin
1970	
1979	Margaret Dayhoff compiled one of the first protein sequence database, manually aligns sequences and creates the first protein substitution matrices
1981	Temple Smith and Michael Waterman proposed an optimal alignment algorithm for local alignments
1985	William Pearson and David Lipman implement an optimization of the Smith-Waterman Algorithm using short exact matching words (FASTA)
1990	Stephen Altshul, Warren Gish, Webb Miller, Eugene Myers and David Lipman, propose a faster alignment tool to identify high identity matches without GAPS
1992	Randall Smith and Temple Smith implement pattern-induced multiple-sequence alignment (PUMA)
1995	Sean Eddy implements multiple sequence alignments using hidden Markov Models (HMMER)
1996	Warren Gish, branches the development of BLAST with WU-BLAST
1997	Gapped BLAST and PSI-BLAST

Establishing homology from statistically significant similarity

- For most proteins, homologs are easily found over long evolutionary distances (500 My – 2 By) using standard approaches (BLAST, FASTA)
- Difficult for distant relationships or very short domains
- Most default search parameters are optimized for distant relationships and work well

Homologous Sequences Share a Common Ancestor

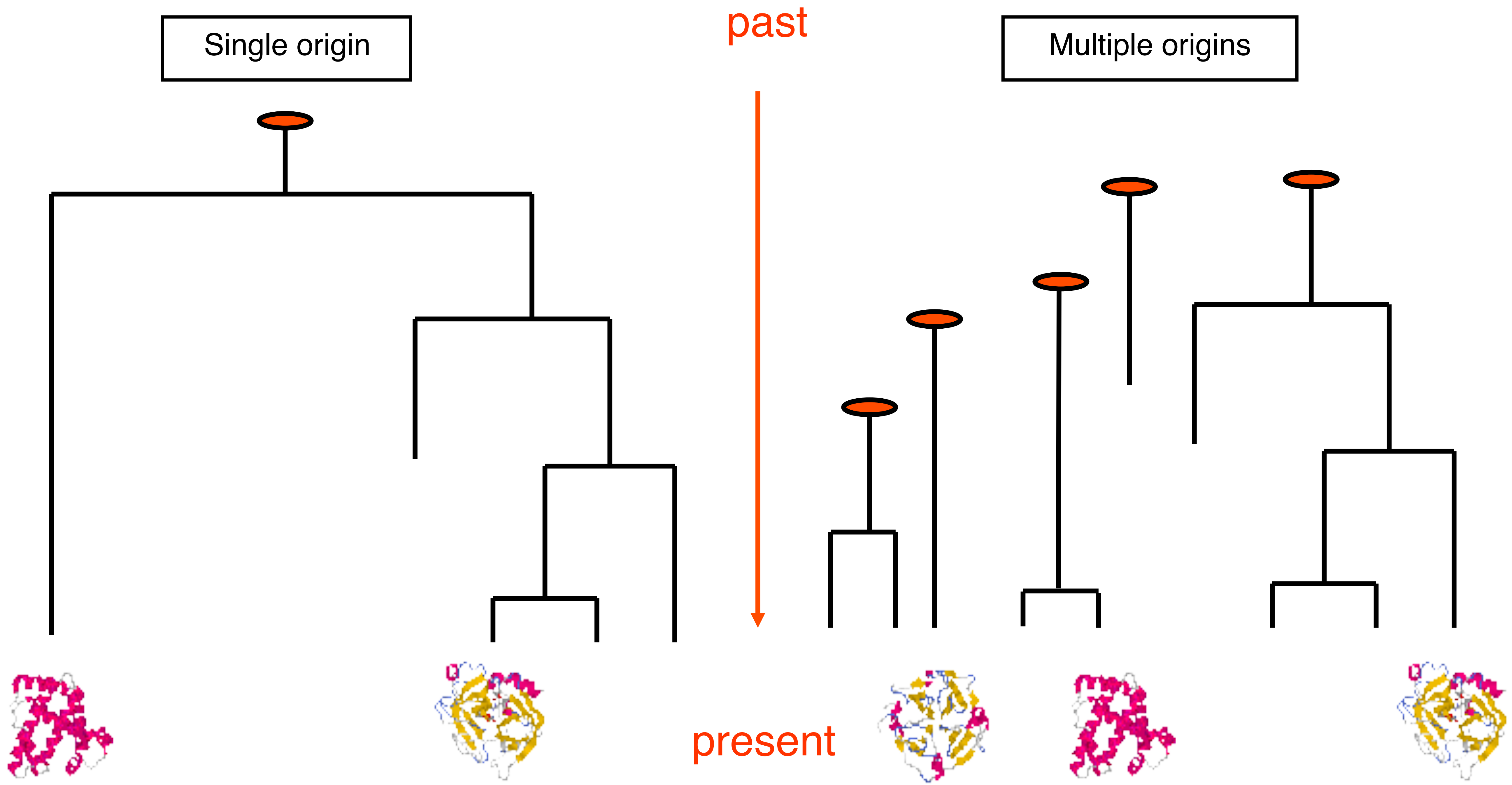


Homology is Confusing:
Ways we have seen it defined

- Protein/Genes/DNA that share a common ancestor
- Specific positions/columns in a multiple sequence alignment that have a 1:1 relationship over evolutionary history
  - Is it possible to be 50% homologous?
- Specific mophological/functional characteristics that share a recent divergence (clade)
  - Are all wings homologous (bat, butterfly, eagle)?

Homology is Confusing:

Are all sequences homologous?

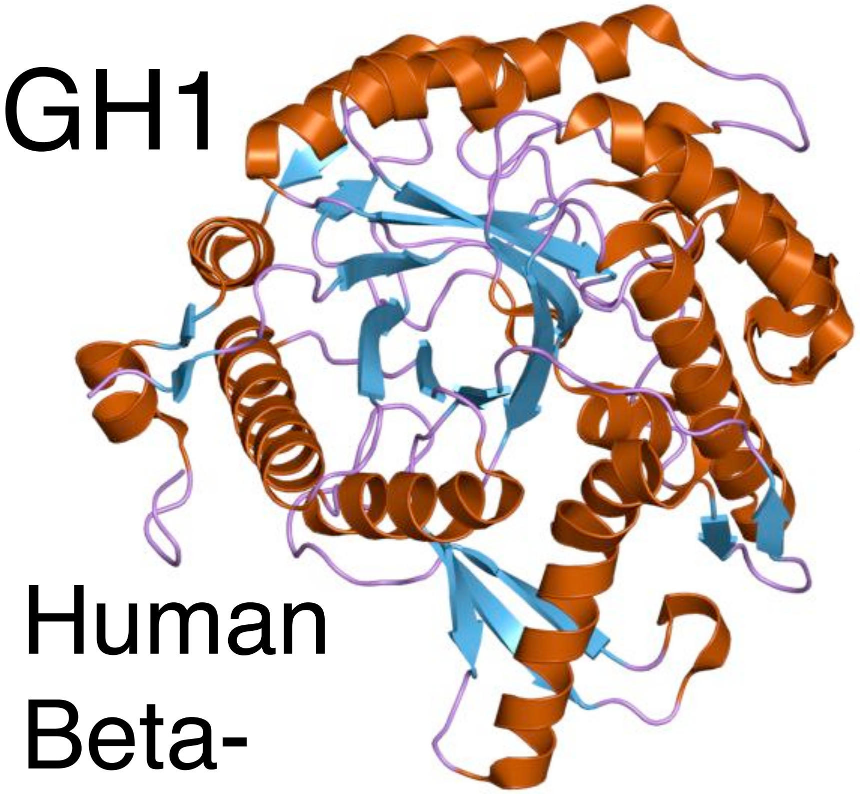


Homology Using Sequence/Structural Comparisons

- Homology is shared ancestry
- Convergence are independent events resulting in the same outcome.
- Sequences are inferred to share a common ancestor based on statistically significant excess similarity
- Any evidence of this excess similarity can be used to infer homology (sequence or structure)
- Lack of evidence cannot be used to infer non-homology
- One must weight the evidence for each hypothesis (Convergence or Homology)

When do we infer homology?

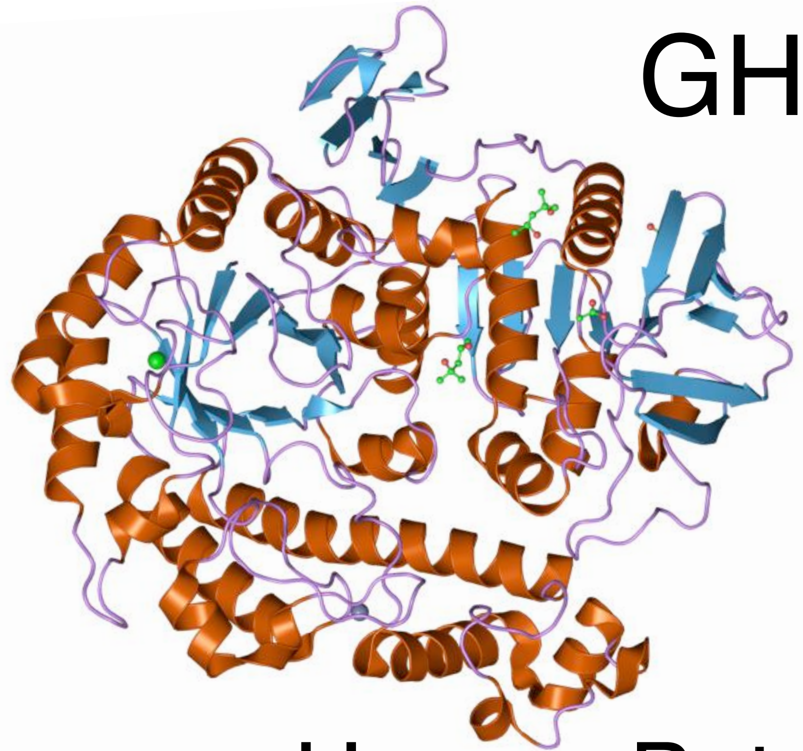
GH1



Human Beta-glucosidase

```
Sequence: Score   = 16.2
     Expect = 1.3
     %ID    = 26%
Structure:RMSD   = 3.80
     Score   = 364.8
     P-Value = 1.97e-02
```

GH42



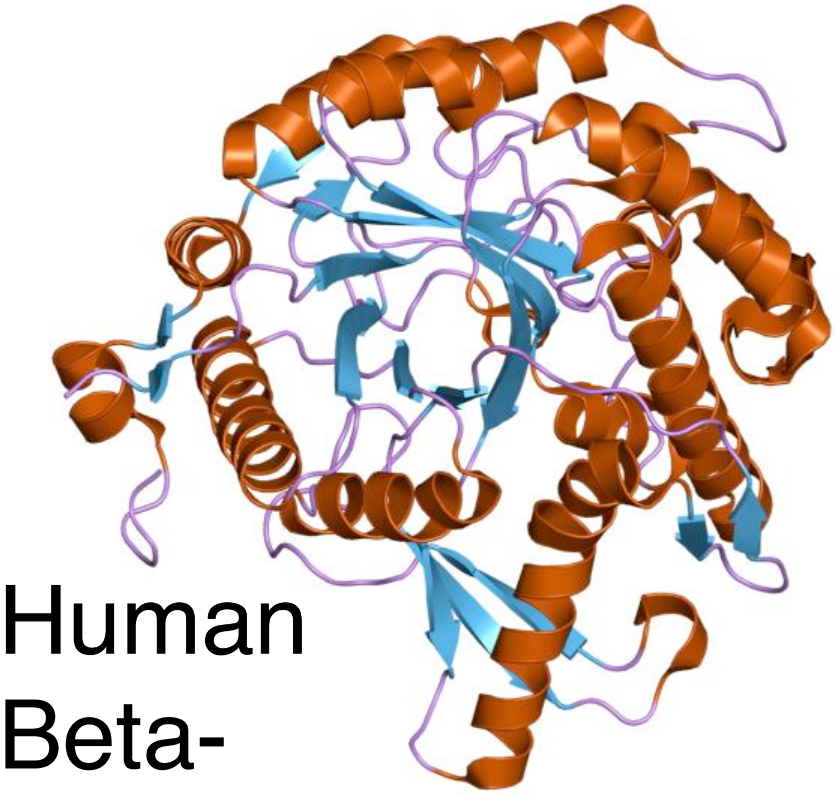
Human Beta-galactosidase

Lactococcus lactis Beta-galactosidase

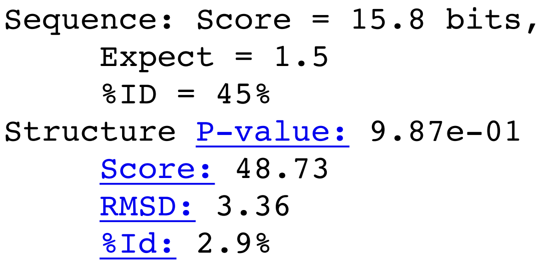
```
Sequence: Score =  205 bits
     Expect = 3e-57
     %ID  = 30%
Structure:RMSD = 2.63
     Score = 1044
     P-Value = 0
```

```
Sequence: Score = 22.3
     Expect = 0.051
     %ID = 25%
Structure: RMSD = 393.5
      Score = 393.5
      P-Value = 1.65e-07
```

When do we infer non-homology?

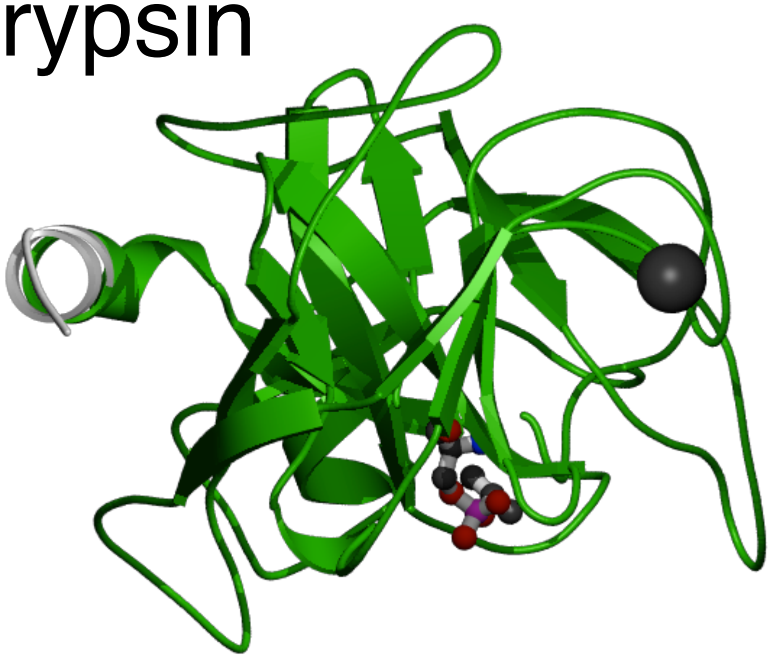


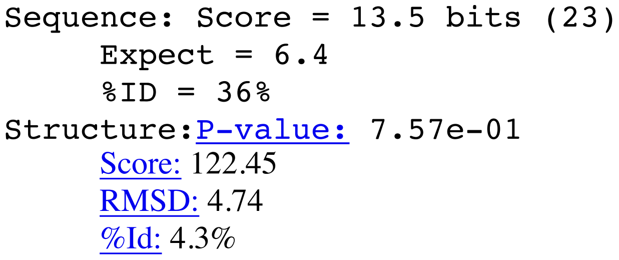
Human Beta-glucosidase



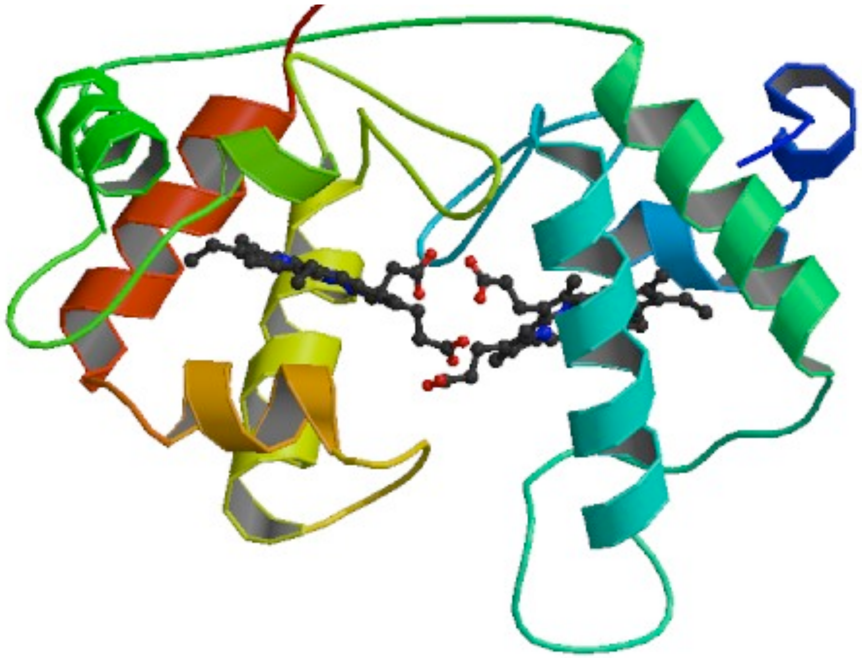
```
Sequence: Score = 15.8 bits,
      Expect = 1.5
      %ID = 45%
Structure P-value: 9.87e-01
      Score: 48.73
      RMSD: 3.36
      %Id: 2.9%
```

Bovine trypsin





```
Sequence: Score = 13.5 bits (23)
      Expect = 6.4
      %ID = 36%
Structure:P-value: 7.57e-01
      Score: 122.45
      RMSD: 4.74
      %Id: 4.3%
```



CYTOCHROME C4

Non-homologous Proteins have different structures

What BLAST Does

?

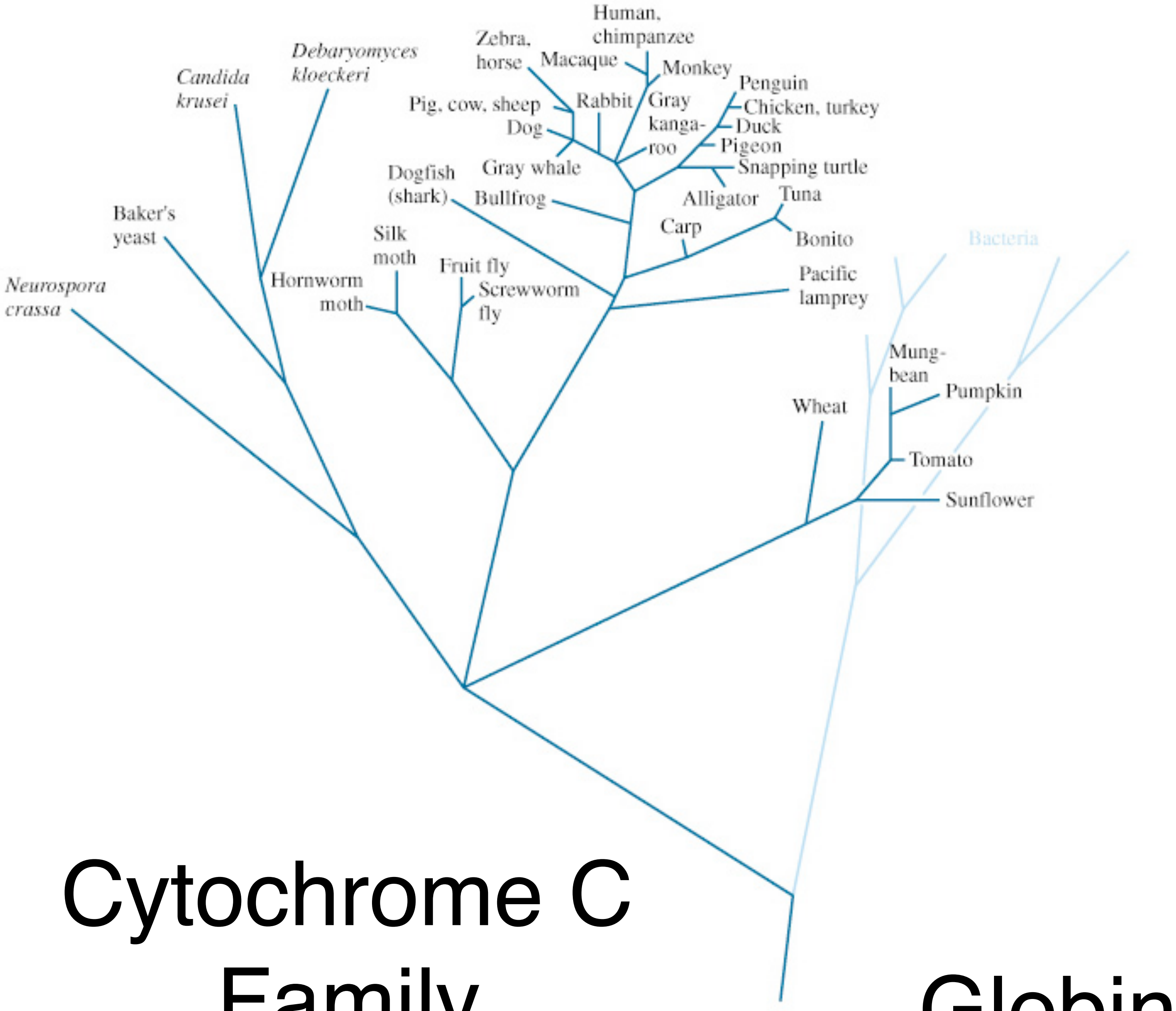
Similarity <=> Homology

?

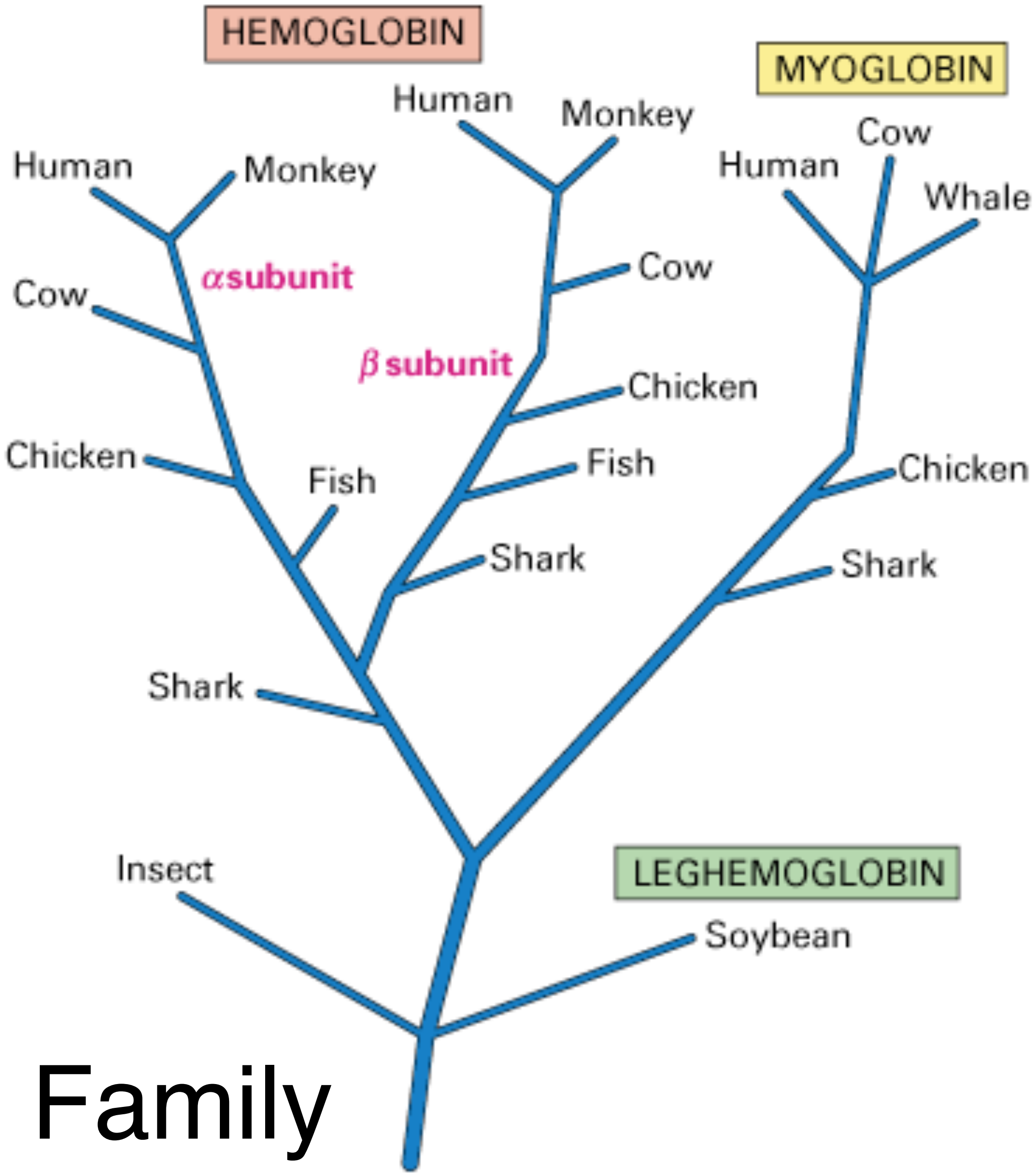
Statistical Significance <=> Biological Significance

Divergence OR Convergence

Orthologs vs Paralogs
Inferring Function

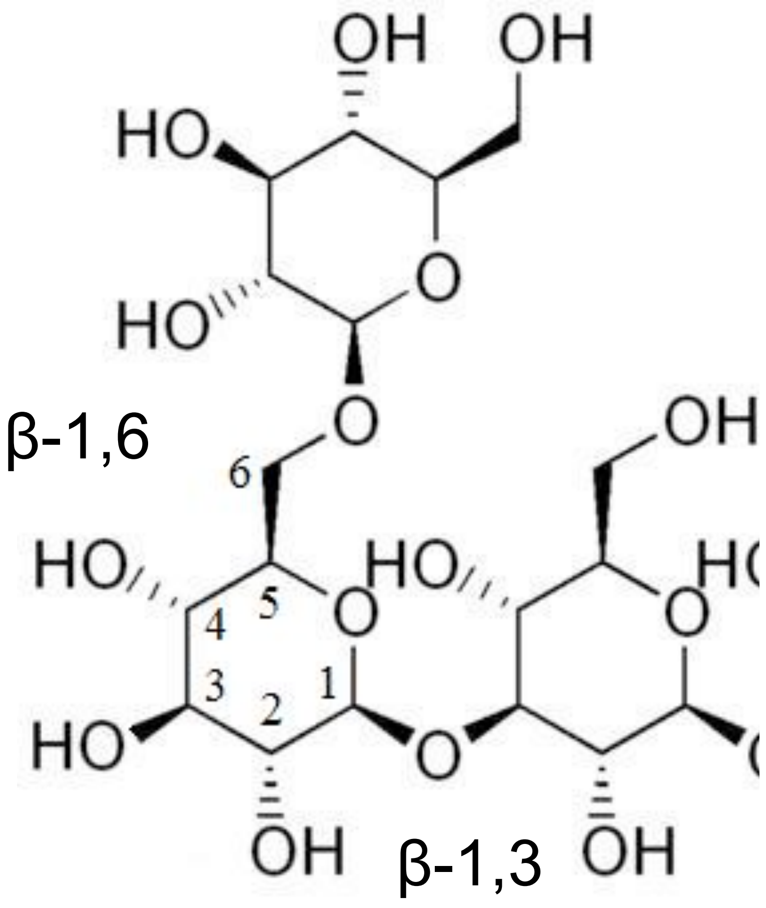


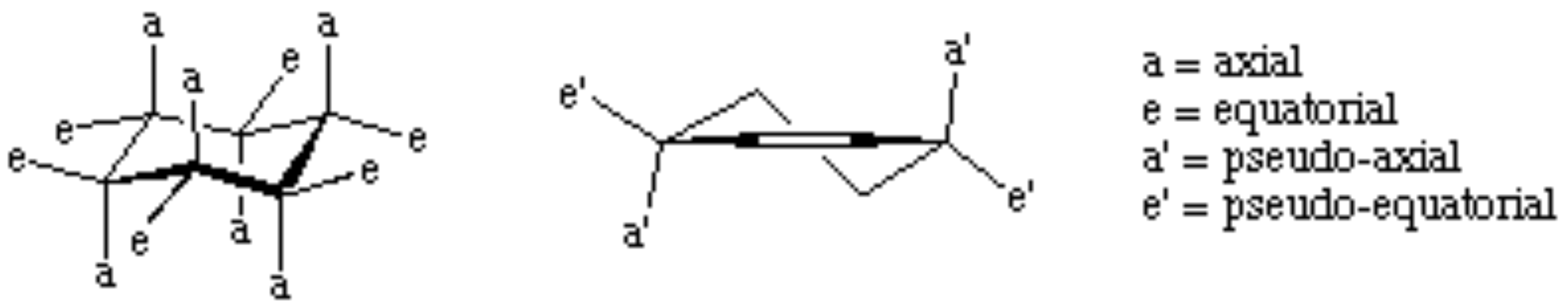
Cytochrome C Family

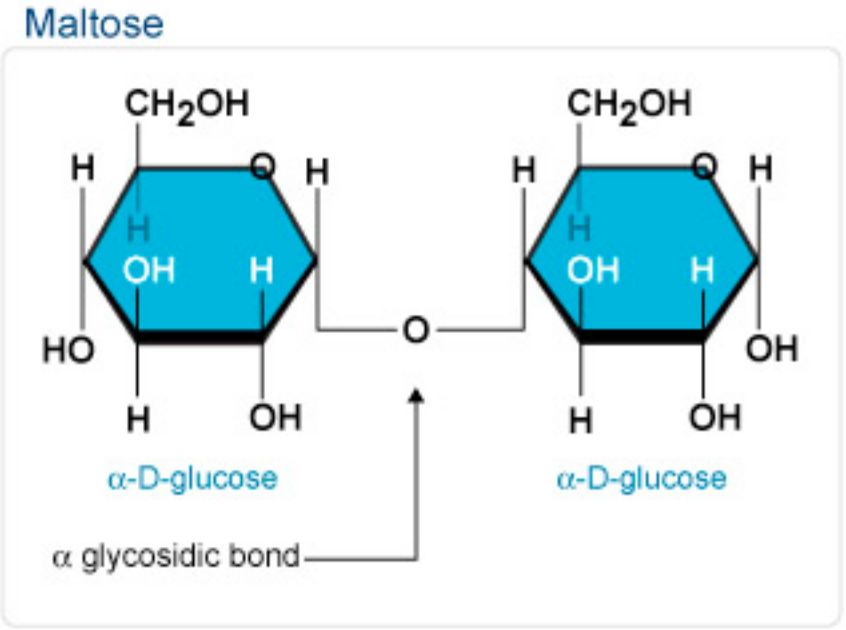


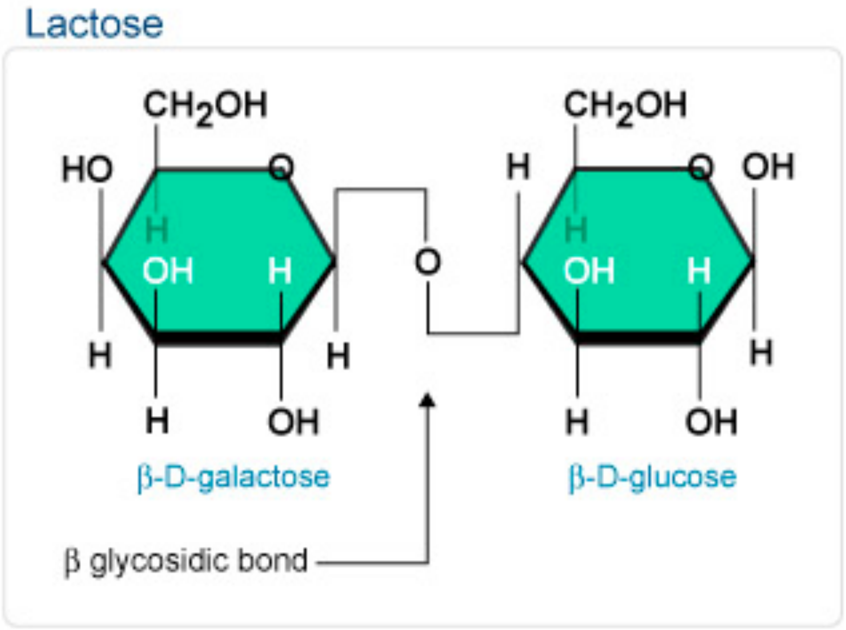
Globin Family

Orthologs vs Paralogs
Inferring Function









Dept. Biol. Penn State ©20

a

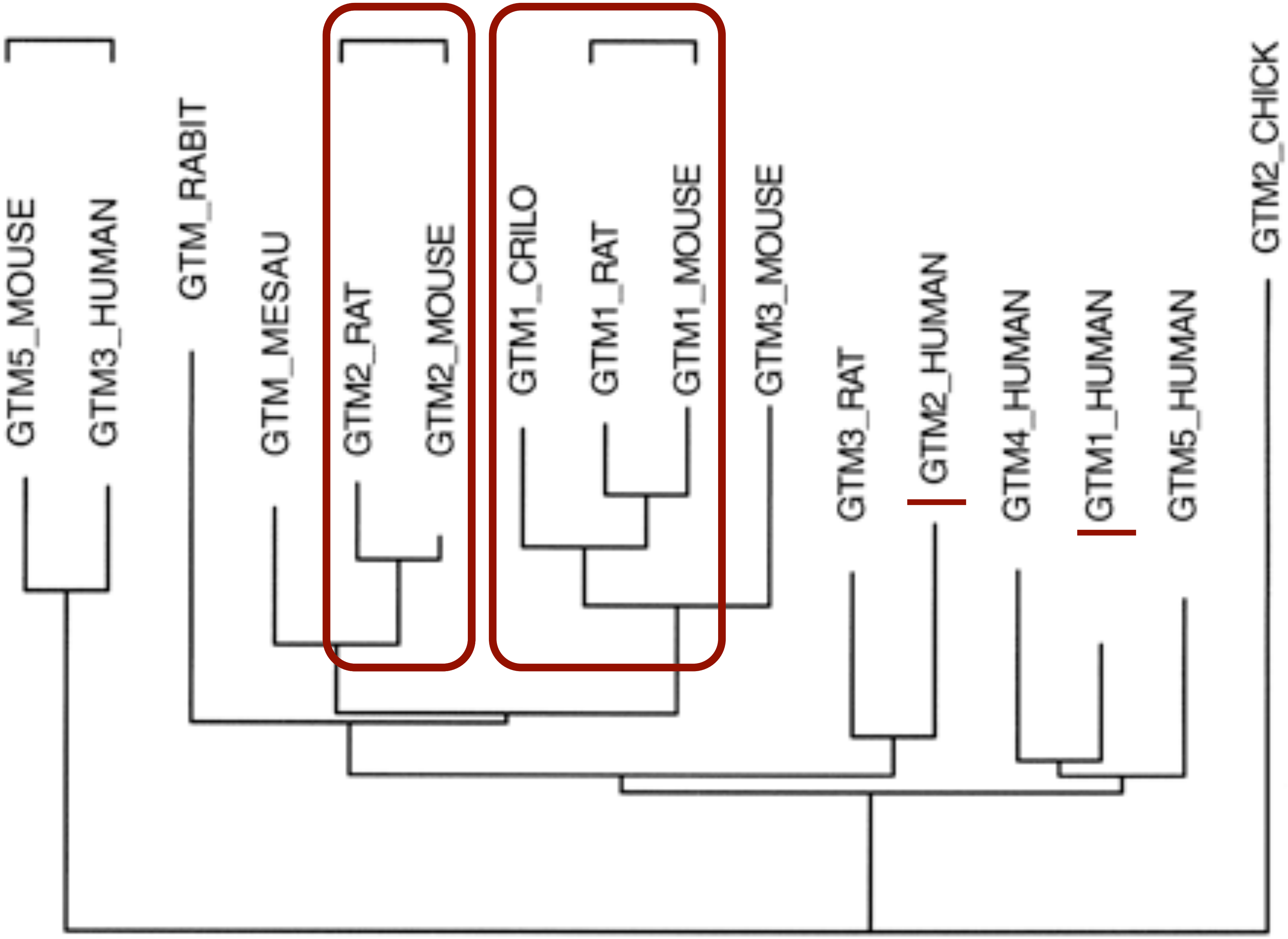
e

Homologs Often Maintain Similar Chemical Functions

Orthology can be difficult to infer

- Over modest distances(human/mouse) post- speciation duplication is common
- Over large distances (human/fly,bacteria), duplication/ loss/replacement may be common
- Homology inferences have false-negatives, but the false-positive rate can be reliably controlled
- Orthology inferences will have both false positives and false negatives
- Paralogous proteins often have similar chemical functions (may act in different pathways)

Orthology can be difficult to infer

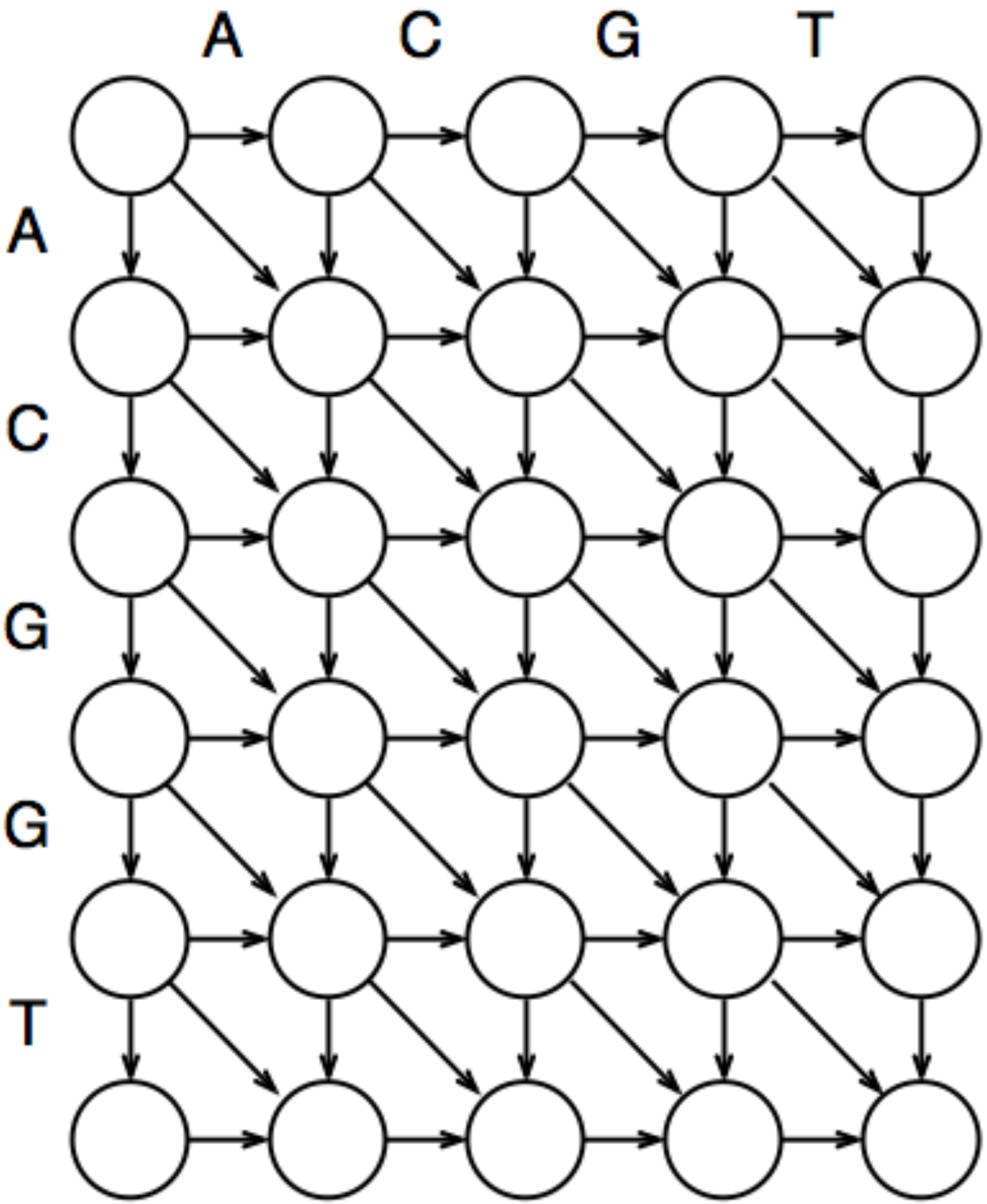


doi: 10.1101/gr.9.4.373 *Genome Res. 1999. 9: 373-382*

How do we measure sequence similarity by alignment and scoring matrices?

Simple Alignments

```
Match: 1
Mismatch: -1
Gap: -2
```

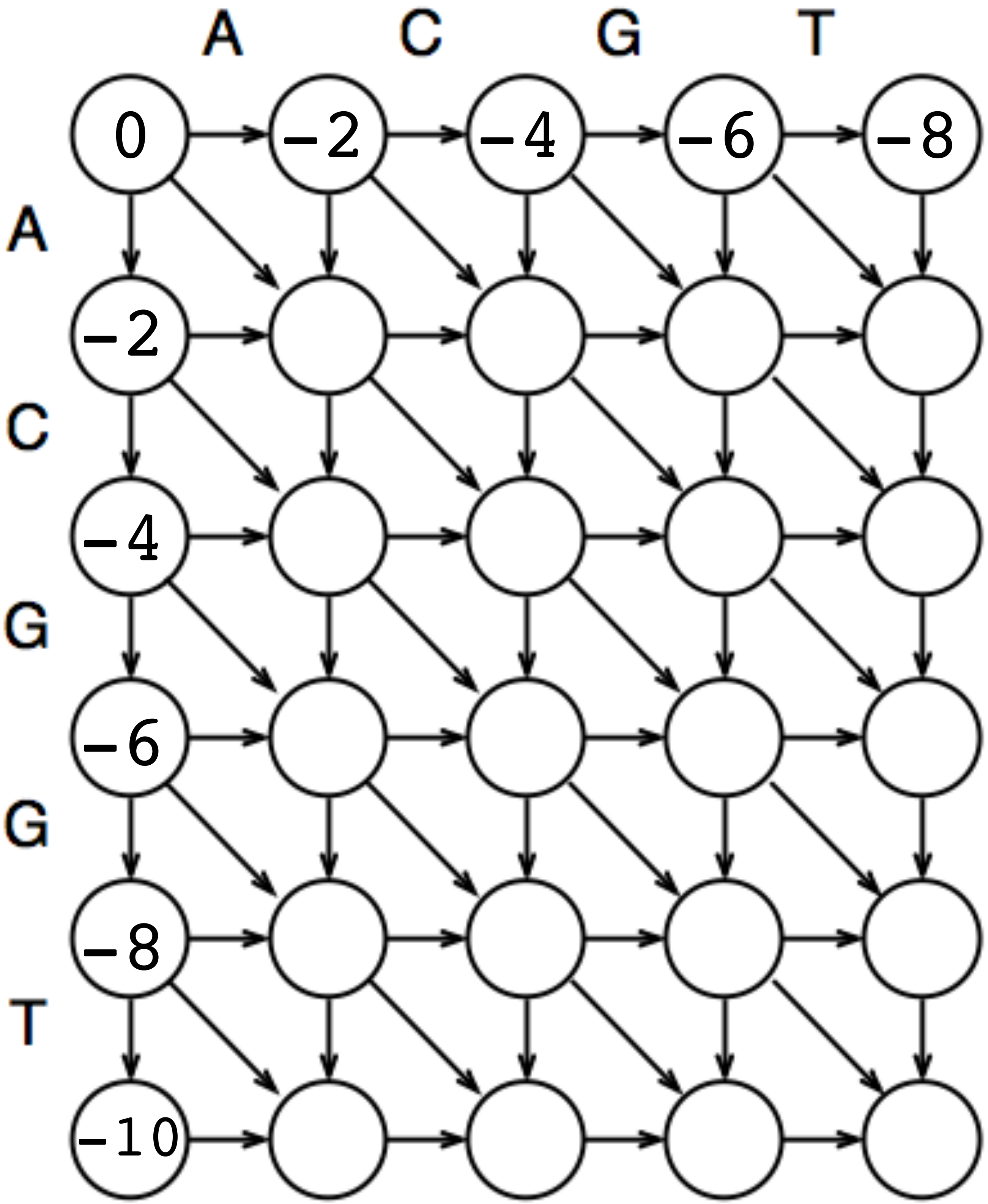


Simple Alignments

```
Match: 1
Mismatch: -1
Gap: -2
```

—————ACGT

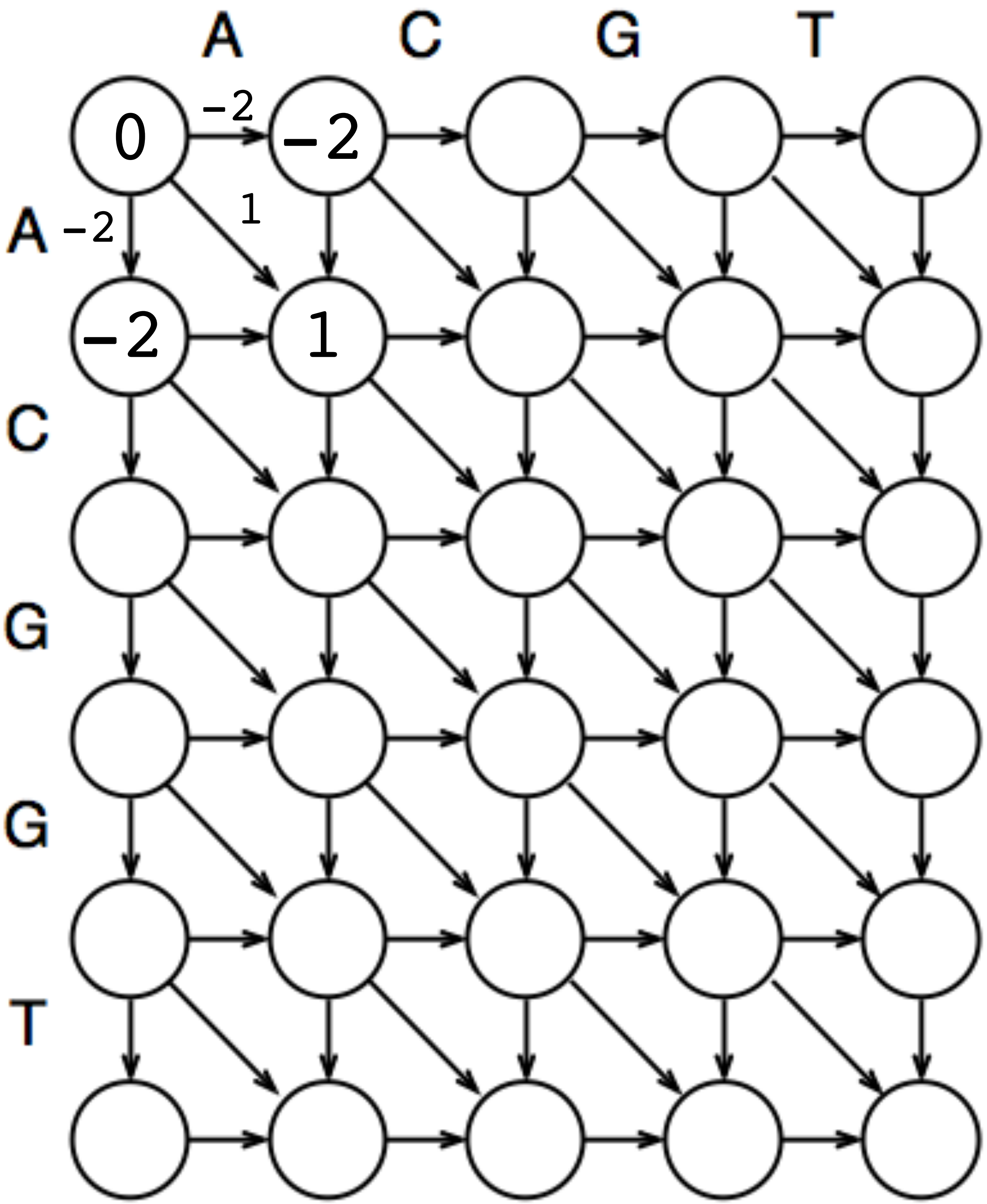
ACGGT————



Simple Alignments

```
Match: 1
Mismatch: -1
Gap: -2
```

A
A

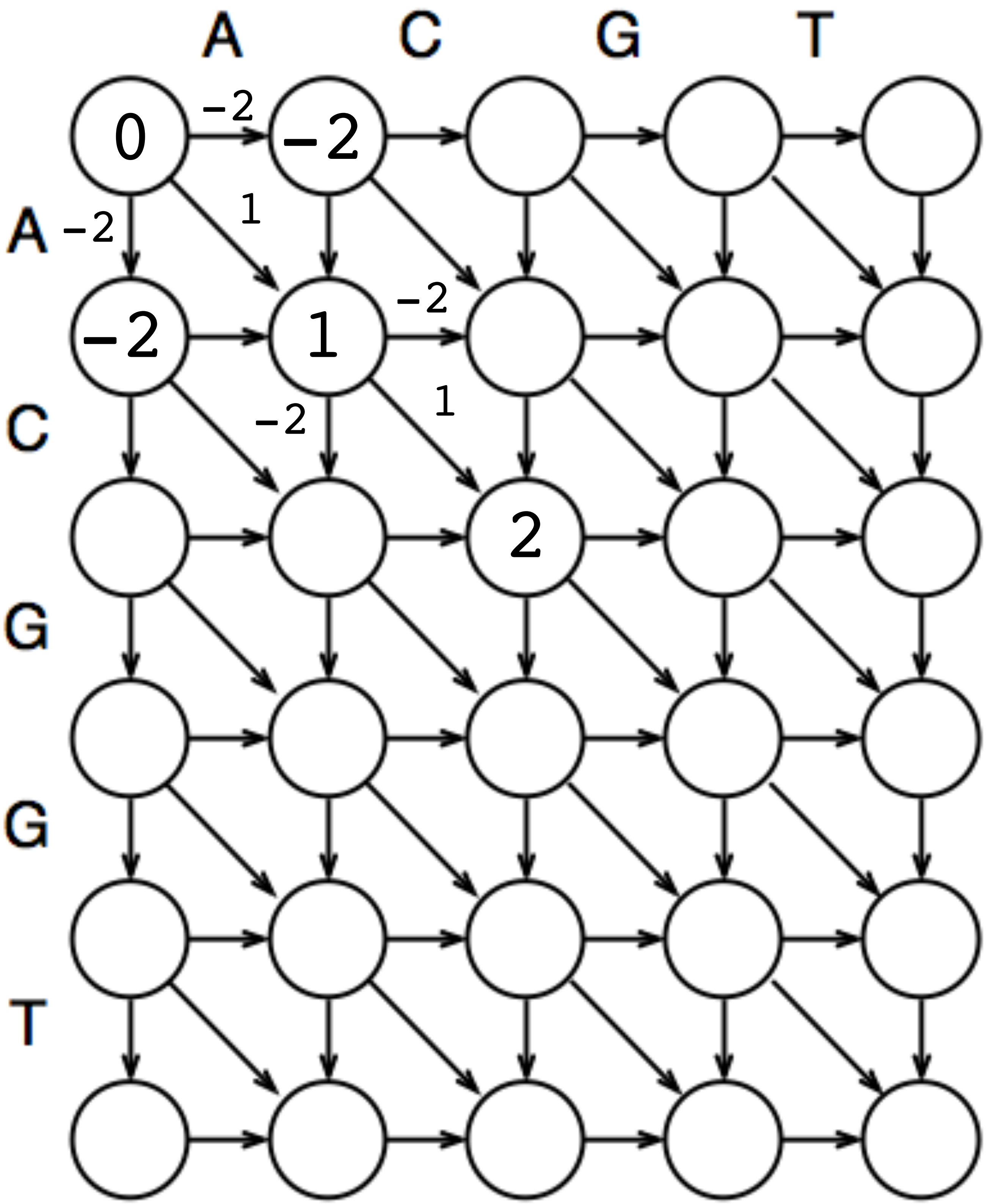


Simple Alignments

Match: 1

Mismatch: -1

Gap: -2

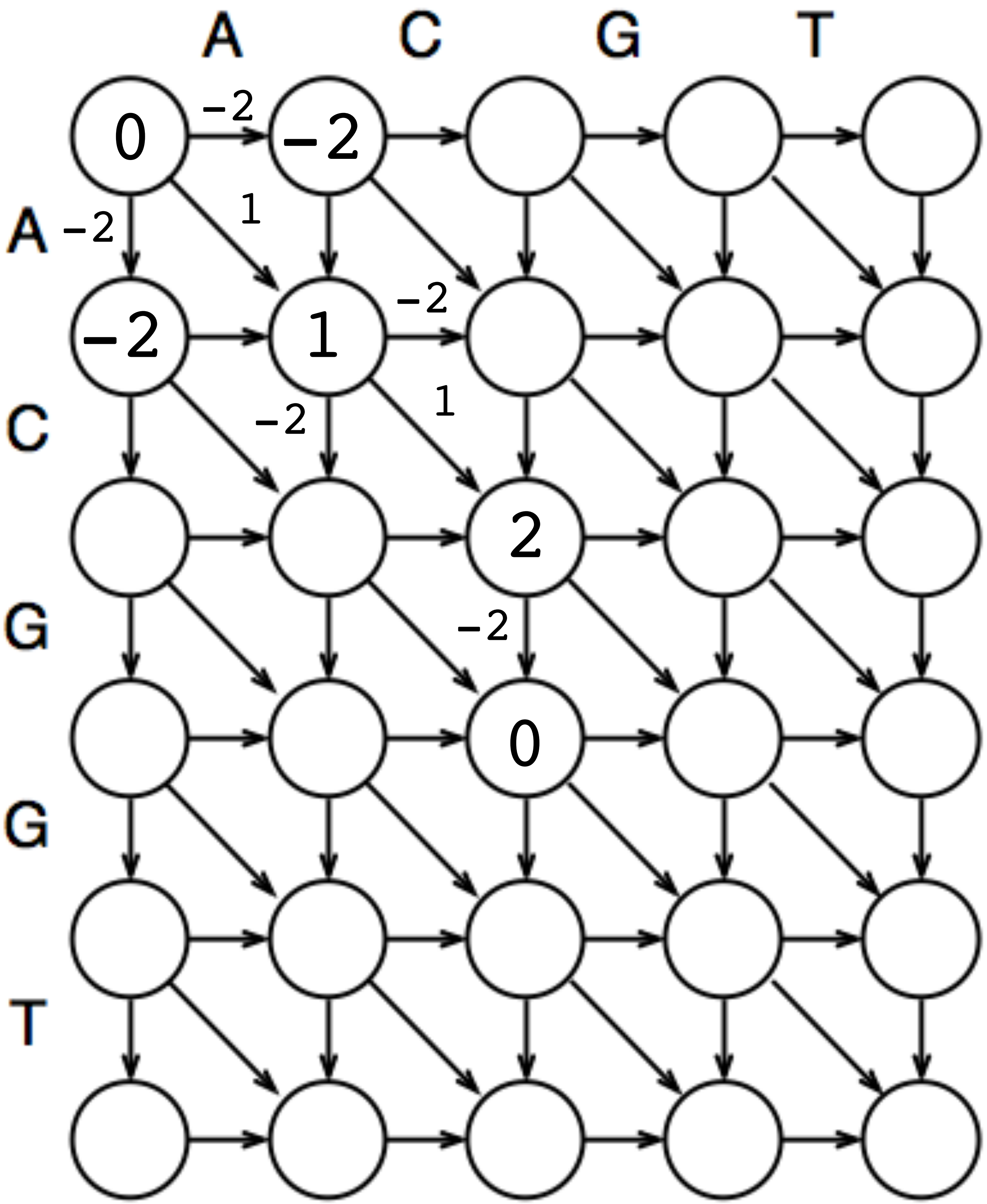


AC

AC

Simple Alignments

```
Match: 1
Mismatch: -1
Gap: -2
```

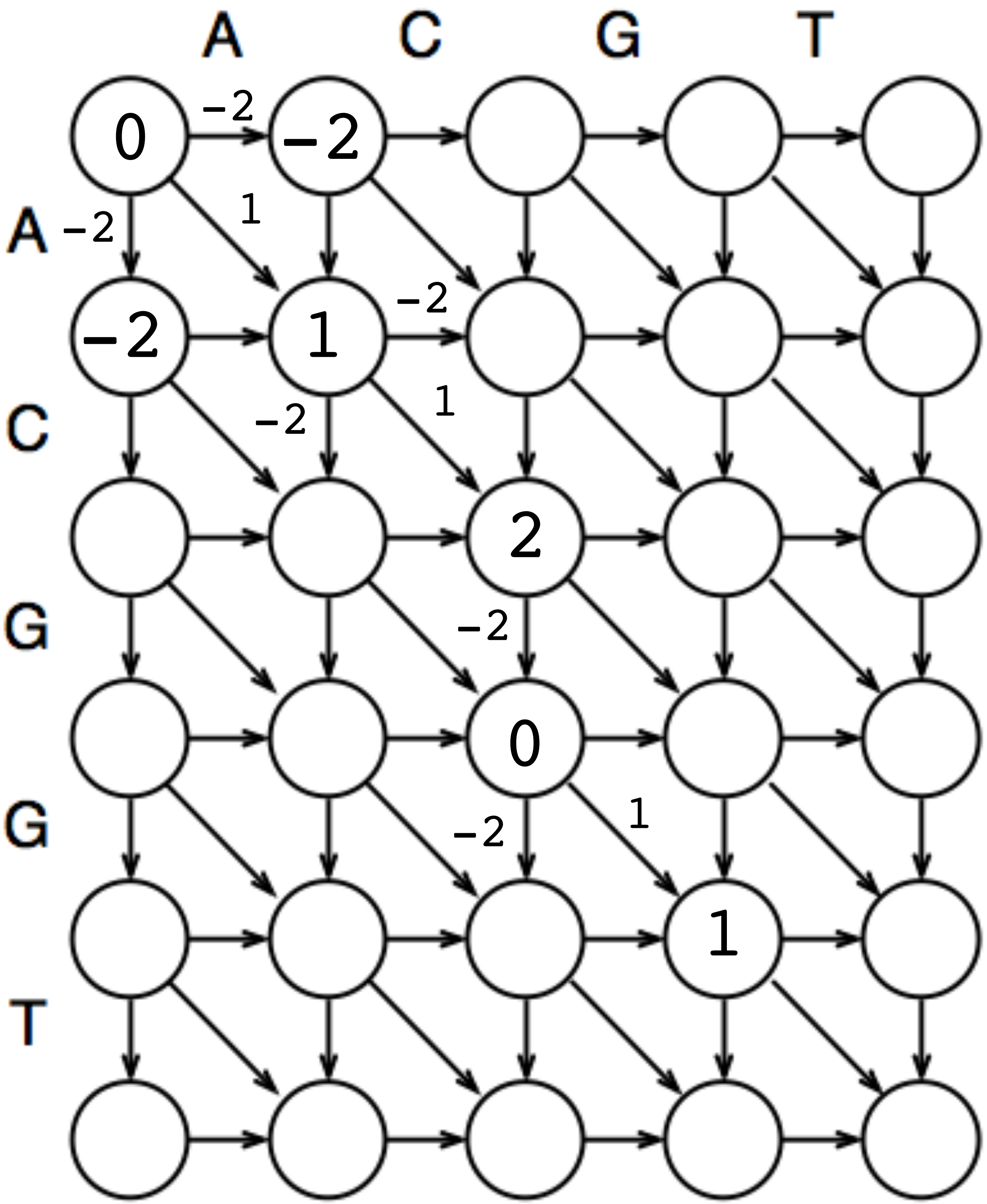


```
AC-
ACG
```

Simple Alignments

```
Match: 1
Mismatch: -1
Gap: -2
```

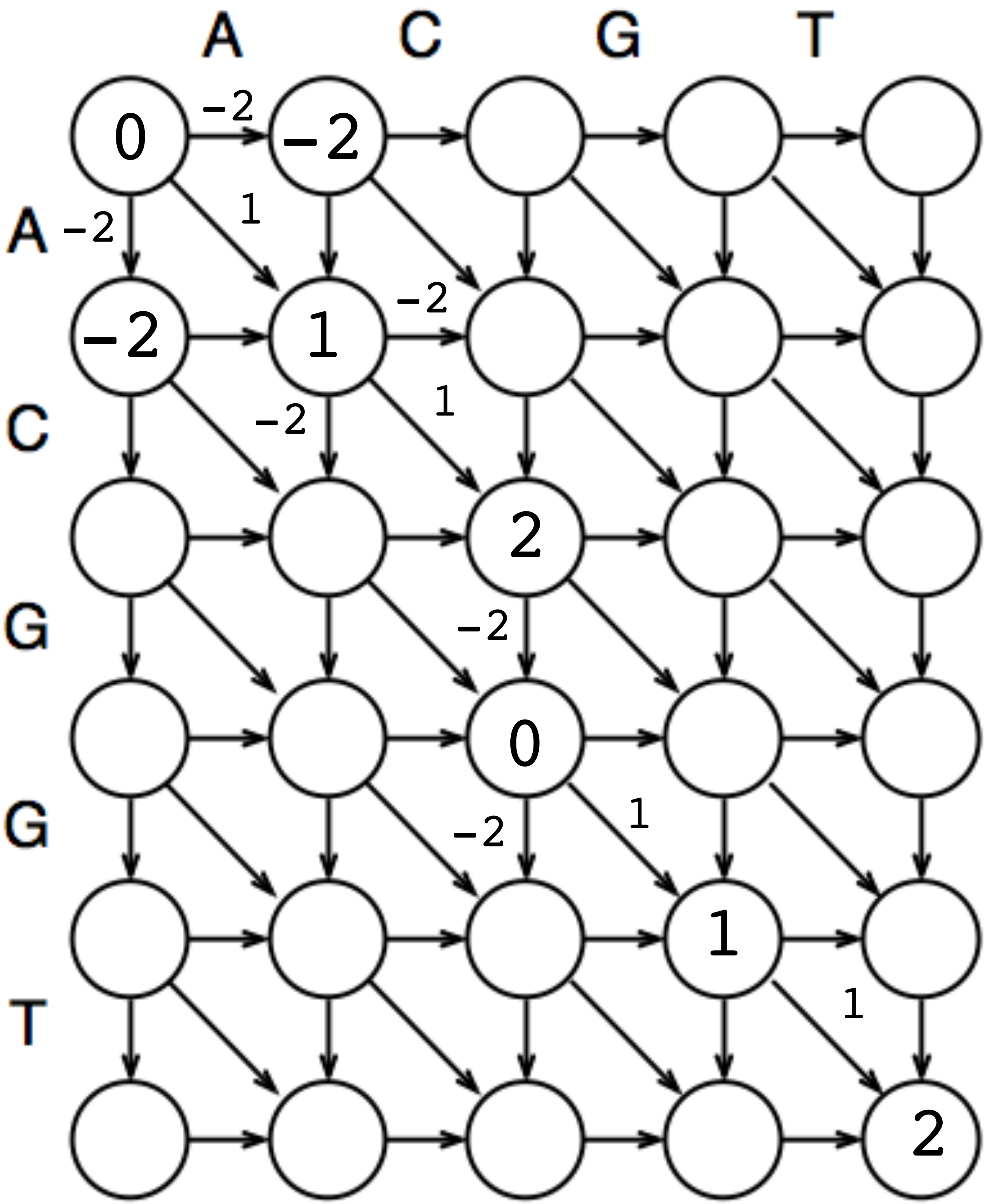
```
AC-G
ACGG
```



Simple Alignments

```
Match: 1
Mismatch: -1
Gap: -2
```

```
AC-GT
ACGGT
```

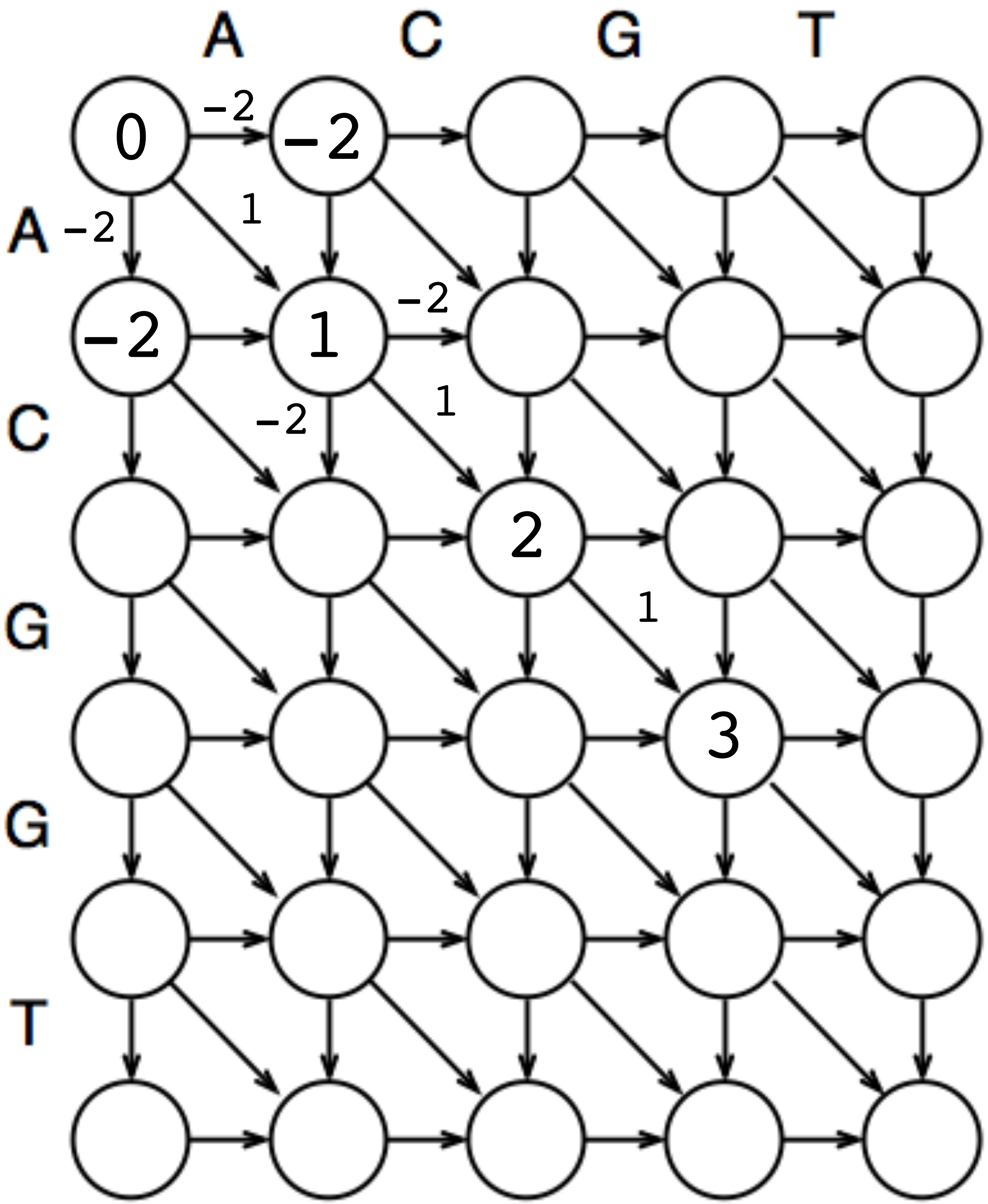


Simple Alignments

```
Match: 1
Mismatch: -1
Gap: -2
```

ACG

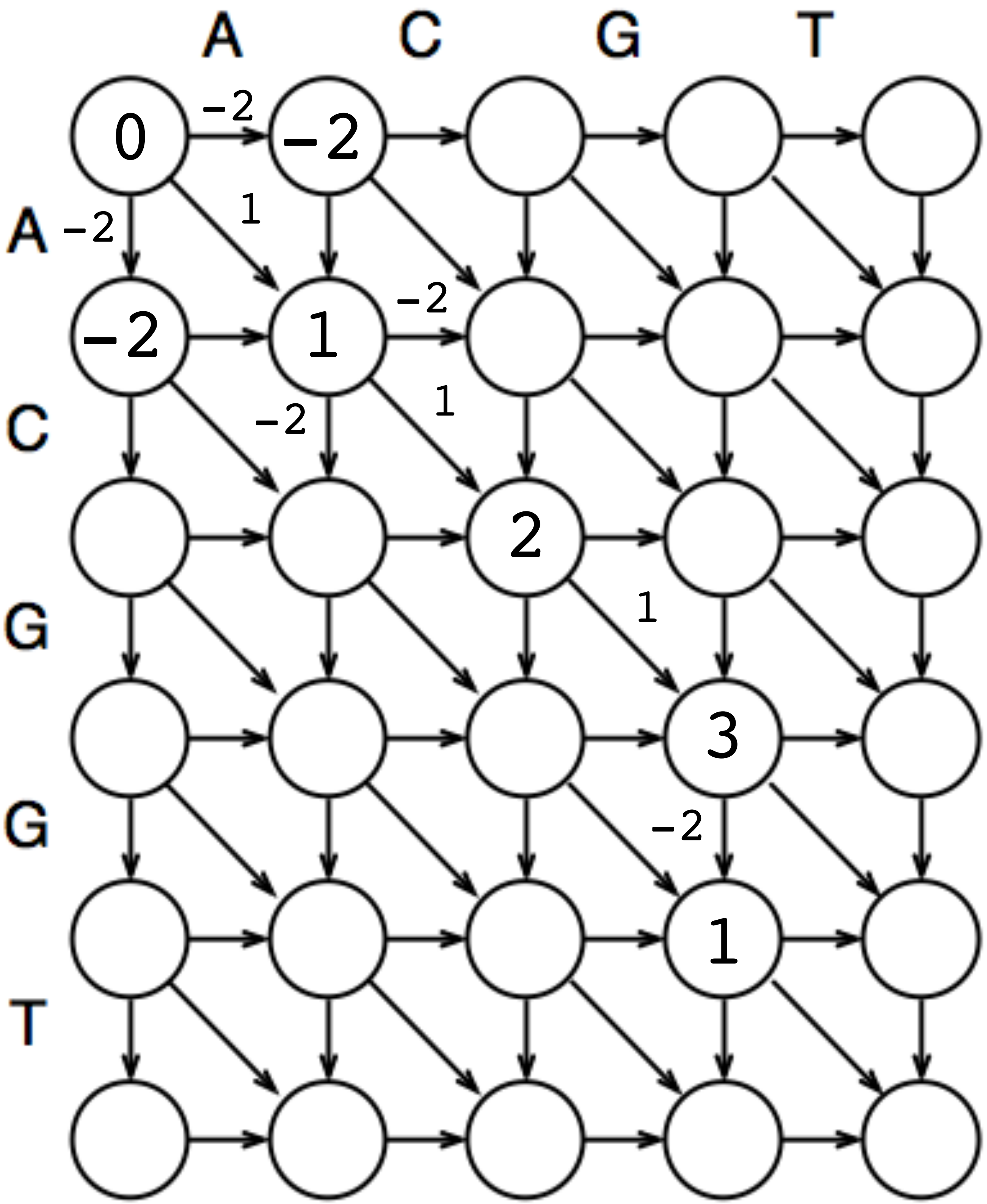
ACG



Simple Alignments

```
Match: 1
Mismatch: -1
Gap: -2
```

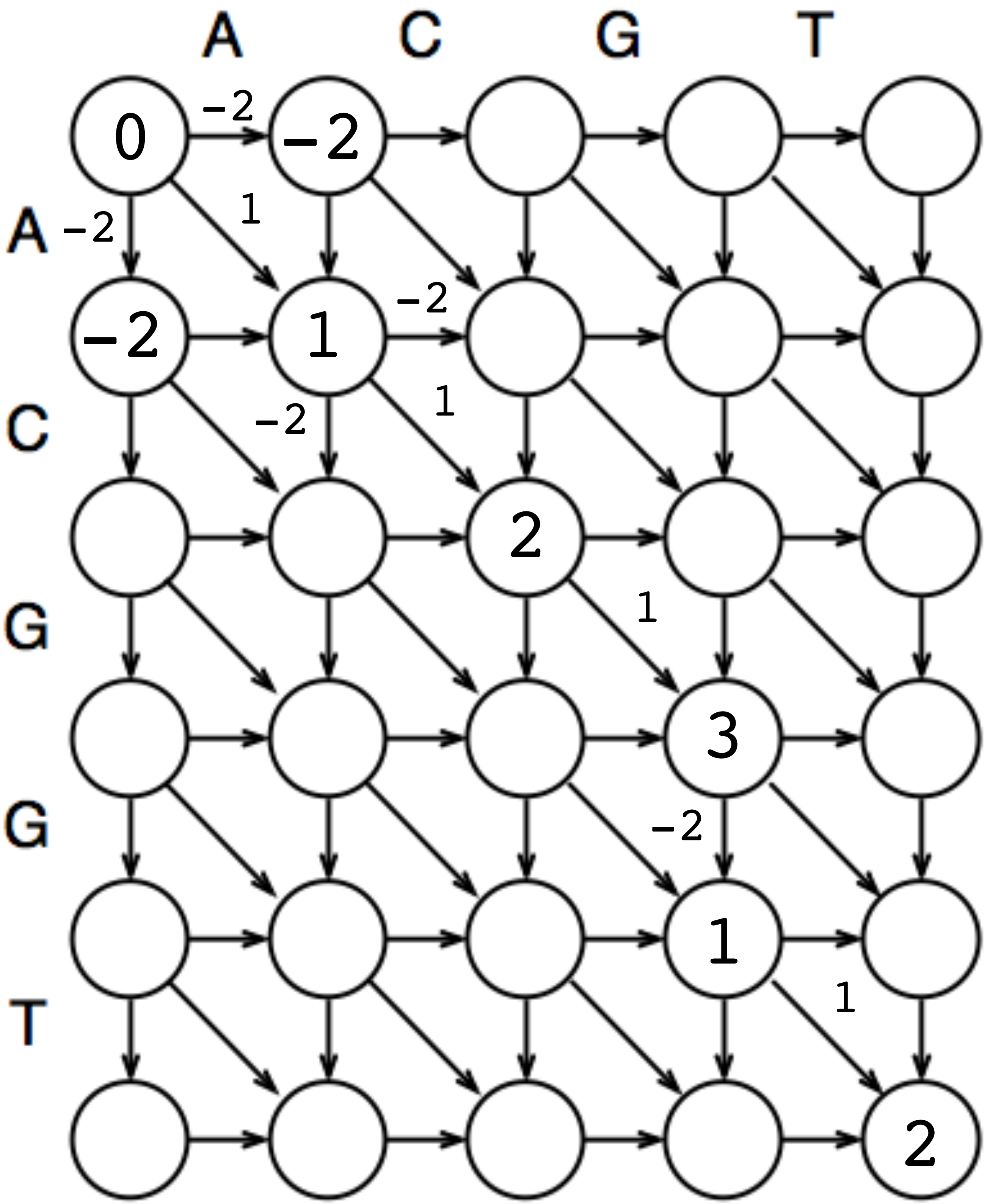
```
ACG-
ACGG
```



Simple Alignments

```
Match: 1
Mismatch: -1
Gap: -2
```

```
ACG-T
ACGGT
```



Global or Local

	P	M	I	L	G	Y	W	N	V	R	G	L
P	:											
P	:											
Y			.			:	.					
T								.	.			
I			:	.					.			.
V		.	.	.					:			.
Y		.				:	.					
F		.	.	.		.	.					.
P	:											
V		.	.	.					:			.
R										:		
G											:	

```
: = Match
. = Similar
- = Gap
```

Local Alignment

```
 PM-ILGYWNVRGL
 :  :. :.  :::
PPYTIV-YFPVRG
```

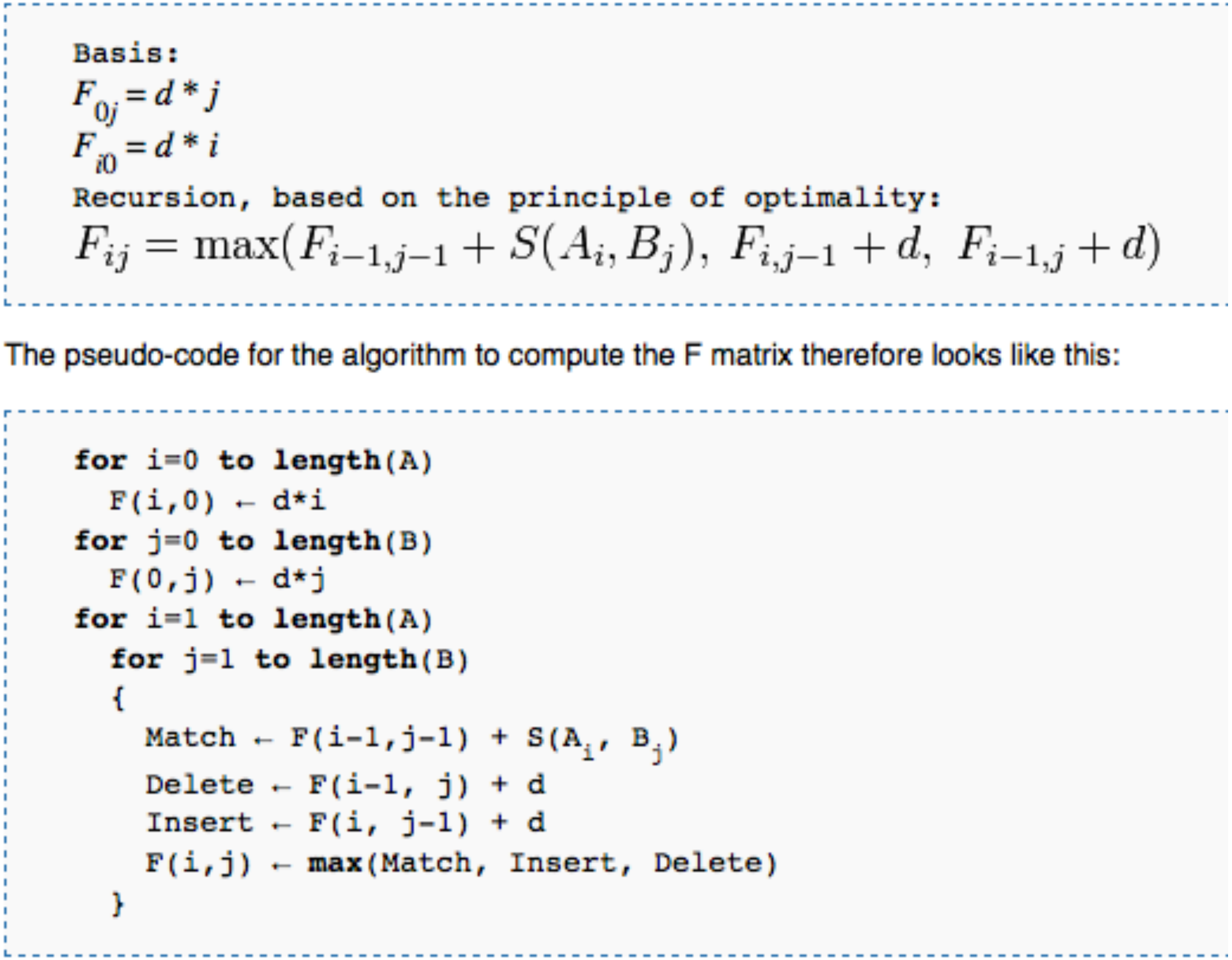
Global Alignment

```
-PMILGYWNVRGL
    . :
PPYTIVYFPVRG-
```

Global Alignments

Global Alignment

```
-PMILGYWNVRGL
 :  . :. :::
PPYTIVYFPVRG-
```



Basis:

$$F_{0j} = d * j$$

$$F_{i0} = d * i$$

Recursion, based on the principle of optimality:

$$F_{ij} = \max(F_{i-1,j-1} + S(A_i, B_j),\ F_{i,j-1} + d,\ F_{i-1,j} + d)$$

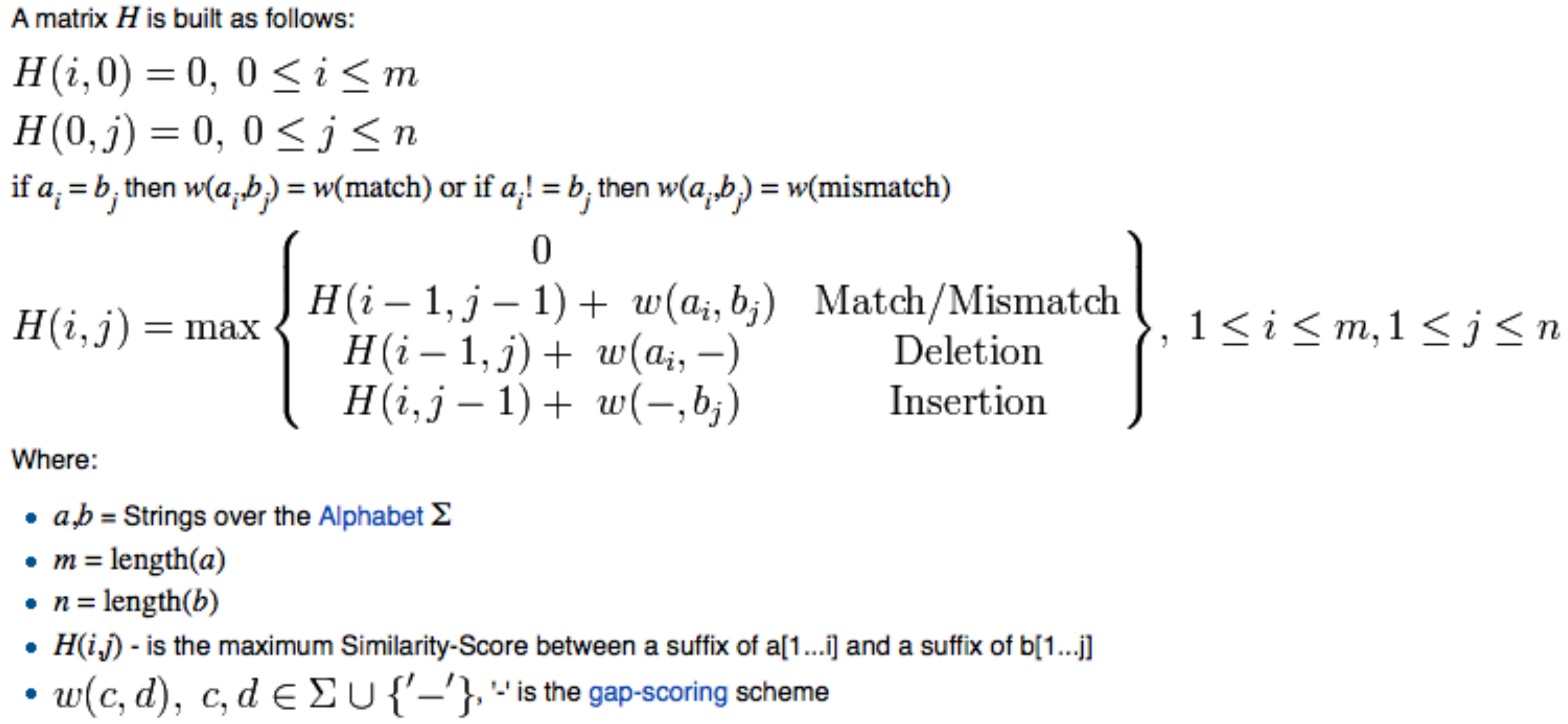
The pseudo-code for the algorithm to compute the F matrix therefore looks like this:

```
for i=0 to length(A)
  F(i,0) ← d*i
for j=0 to length(B)
  F(0,j) ← d*j
for i=1 to length(A)
  for j=1 to length(B)
  {
    Match ← F(i-1,j-1) + S($A_i$, $B_j$)
    Delete ← F(i-1, j) + d
    Insert ← F(i, j-1) + d
    F(i,j) ← max(Match, Insert, Delete)
  }
```

Local Alignments

Local Alignment

```
 AAPMILGYWNVRGLBB
   :  : :: :::
DDPPYTIVYFPVRGCC
```



A matrix H is built as follows:

$$H(i,0) = 0,\ 0 \le i \le m$$

$$H(0,j) = 0,\ 0 \le j \le n$$

if $a_i = b_j$ then $w(a_i,b_j) = w(\text{match})$ or if $a_i != b_j$ then $w(a_i,b_j) = w(\text{mismatch})$

$$H(i,j) = \max \left\{ \begin{matrix} 0 \\ H(i-1,j-1) + \ w(a_i,b_j) & \text{Match/Mismatch} \\ H(i-1,j) + \ w(a_i,-) & \text{Deletion} \\ H(i,j-1) + \ w(-,b_j) & \text{Insertion} \end{matrix} \right\},\ 1 \le i \le m, 1 \le j \le n$$

Where:

- a,b = Strings over the Alphabet Σ
- $m = \text{length}(a)$
- $n = \text{length}(b)$
- $H(i,j)$ - is the maximum Similarity-Score between a suffix of a[1...i] and a suffix of b[1...j]
- $w(c,d),\ c,d \in \Sigma \cup \{'-'\}$, '-' is the gap-scoring scheme

Search Algorithms

Algorithm	Value Calculated	Scoring Matrix	Gap penalty	Time Requirement	Reference
Needleman-Wunsch	Global similarity	Any	Penalty/Gap	O(n2)	Needleman and Wunsch, 1970
Sellers	Global distance	Unity	Penalty/Gap	O(n2)	Sellers, 1974
Smith-Waterman	Local Similarity	Sij < 0.0	Affine (q+rk)	O(n2)	Smith and Waterman, 1981 Gotoh, 1982
SRCHN	Approx. local similarity	diagonal	Penalty/Gap	O(n) – O(n2)	Wilbur and Lipman, 1983
FASTP/FASTA	Approx. local similarity	Sij < 0.0	Limit Size (q+rk)	O(n2)/K	Lipman and Pearson, 1985, Pearson and Lipman, 1988
BLAST	Maximum Segment Score	Sij < 0.0	Multiple Segment	O(n2)/K	Altschul et al 1990
BLAST2.0	Approx. local similarity	Sij < 0.0	(q+rk)	O(n2)/K	Altschul et al 1997

Scoring Matrices For Proteins

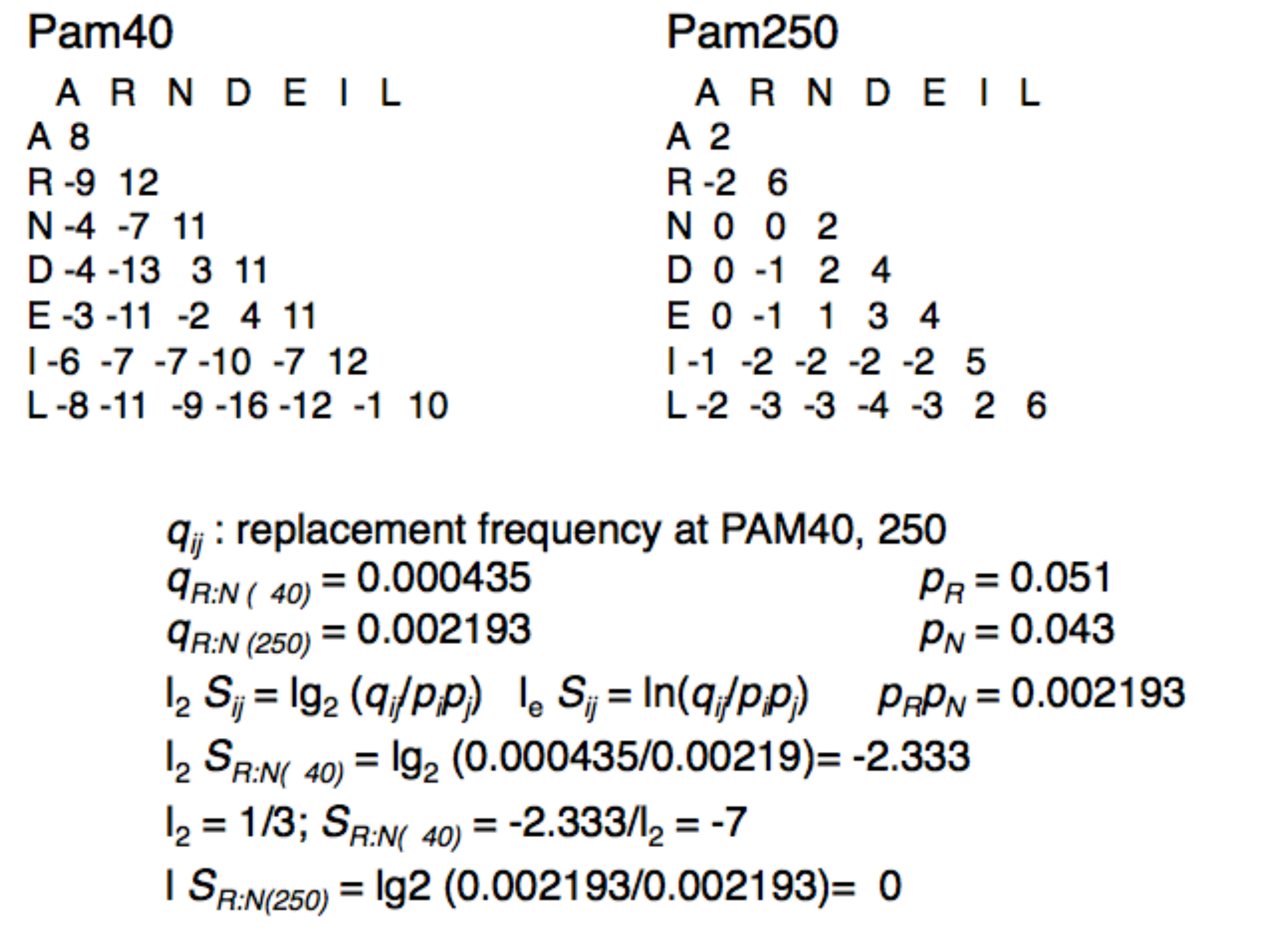
Scoring matrices can set the evolutionary look- back time for a search

- Lower PAM (PAM10/MDM10 ... PAM60) for closer (90% - 50% identity)
- Higher BLOSUM for higher conservation (BLOSUM50 distant, BLOSUM80 conserved)
- Shallow scoring matrices for short domains/short queries (metagenomics)
  - Matrices have "bits/position" (score/position), 40 aa at 0.7 bits/position (BLOSUM62) means 28 bit max score (50 bits significant)
- Deep scoring matrices allow alignments to continue, possibly outside the homologous region

PAM Matrices

- The PAM matrices were introduced by Margret Dayhoff in 1979
- They were based on 1572 observed mutations in 71 families of closely related proteins.
- Each matrix has the twenty standard amino acids in its twenty rows and columns
- The value in a given cell represents the probability of a substitution of one amino acid for another.

Details on Scoring Matrices



Pam40

	A	R	N	D	E	I	L
A	8						
R	-9	12					
N	-4	-7	11				
D	-4	-13	3	11			
E	-3	-11	-2	4	11		
I	-6	-7	-7	-10	-7	12	
L	-8	-11	-9	-16	-12	-1	10

Pam250

	A	R	N	D	E	I	L
A	2						
R	-2	6					
N	0	0	2				
D	0	-1	2	4			
E	0	-1	1	3	4		
I	-1	-2	-2	-2	-2	5	
L	-2	-3	-3	-4	-3	2	6

q_{ij} : replacement frequency at PAM40, 250

$q_{R:N(\ 40)} = 0.000435$ $\qquad$ $p_R = 0.051$

$q_{R:N\ (250)} = 0.002193$ $\qquad$ $p_N = 0.043$

$l_2\ S_{ij} = \lg_2\ (q_{ij}/p_i p_j)$ $\quad$ $l_e\ S_{ij} = \ln(q_{ij}/p_i p_j)$ $\qquad$ $p_R p_N = 0.002193$

$l_2\ S_{R:N(\ 40)} = \lg_2\ (0.000435/0.00219) = -2.333$

$l_2 = 1/3;\ S_{R:N(\ 40)} = -2.333/l_2 = -7$

$l\ S_{R:N(250)} = \lg 2\ (0.002193/0.002193) = \ 0$

PAM Matrices

$$\lambda S = \log\left(\frac{q_{ij}}{p_i p_j}\right)$$

- S is the replacement score of i to j
- λ term is used to scale the matrix so that individual scores can be accurately represented with integers
- qij is Replacement frequency of i to j
- pi is the expected frequency of i

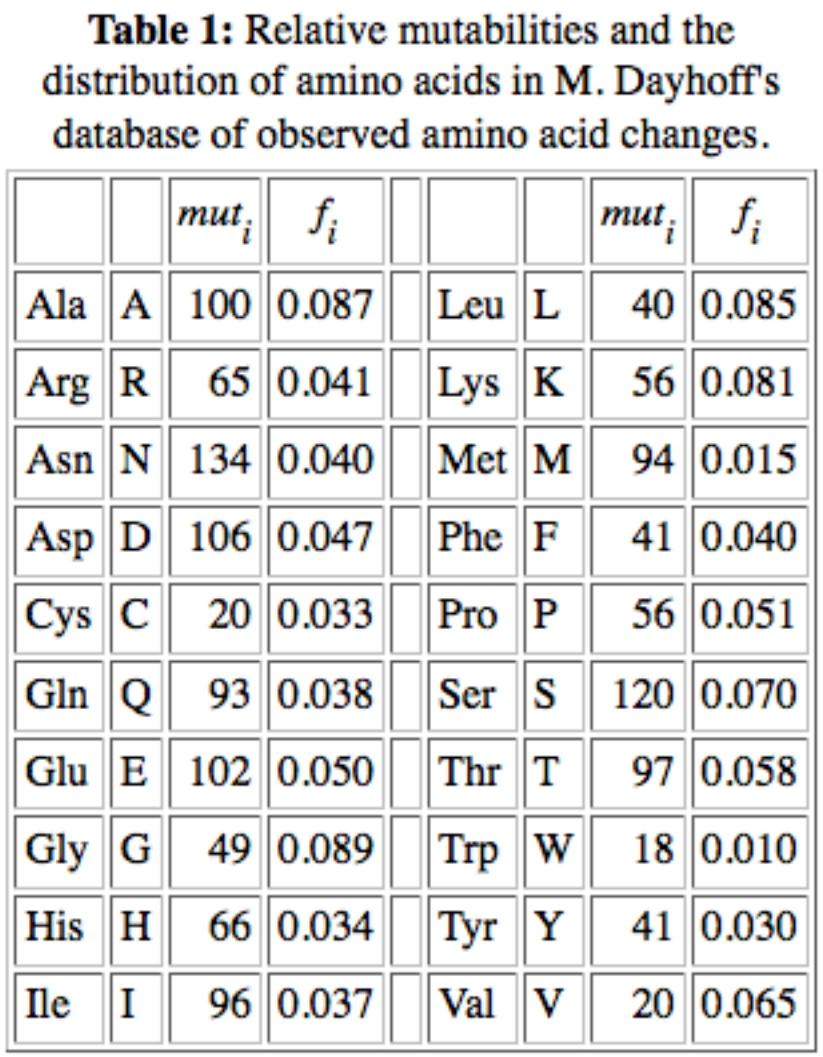


Table 1: Relative mutabilities and the distribution of amino acids in M. Dayhoff's database of observed amino acid changes.

		mut_i	f_i				mut_i	f_i
Ala	A	100	0.087		Leu	L	40	0.085
Arg	R	65	0.041		Lys	K	56	0.081
Asn	N	134	0.040		Met	M	94	0.015
Asp	D	106	0.047		Phe	F	41	0.040
Cys	C	20	0.033		Pro	P	56	0.051
Gln	Q	93	0.038		Ser	S	120	0.070
Glu	E	102	0.050		Thr	T	97	0.058
Gly	G	49	0.089		Trp	W	18	0.010
His	H	66	0.034		Tyr	Y	41	0.030
Ile	I	96	0.037		Val	V	20	0.065

- Scoring matrices can be designed for different evolutionary distances (less=shallow; more=deep)
- Deep matrices allow more substitution

PAM1: Predicts one mutation per 100 aa

PAM40: Predicts 40 mutations per 100 aa

PAM250: Predicts 250 mutations per 100 aa

Details on Scoring Matrices

PAM

- Evolutionary model - extrapolated from PAM1
- PAM20: 20% change (mammals)
- PAM250: 250% change (<20% identity)
- Gap penalties should vary
- shallow matrices (PAM10-40) for short sequences and short distances

BLOSUM

- Empirically determined, no extrapolation (no model)
- BLOSUM45-50 - distant (1/3 bits)
- BLOSUM80 -very highly conserved (not small change), high info/position
- BLOSUM62 - 1/2 bits

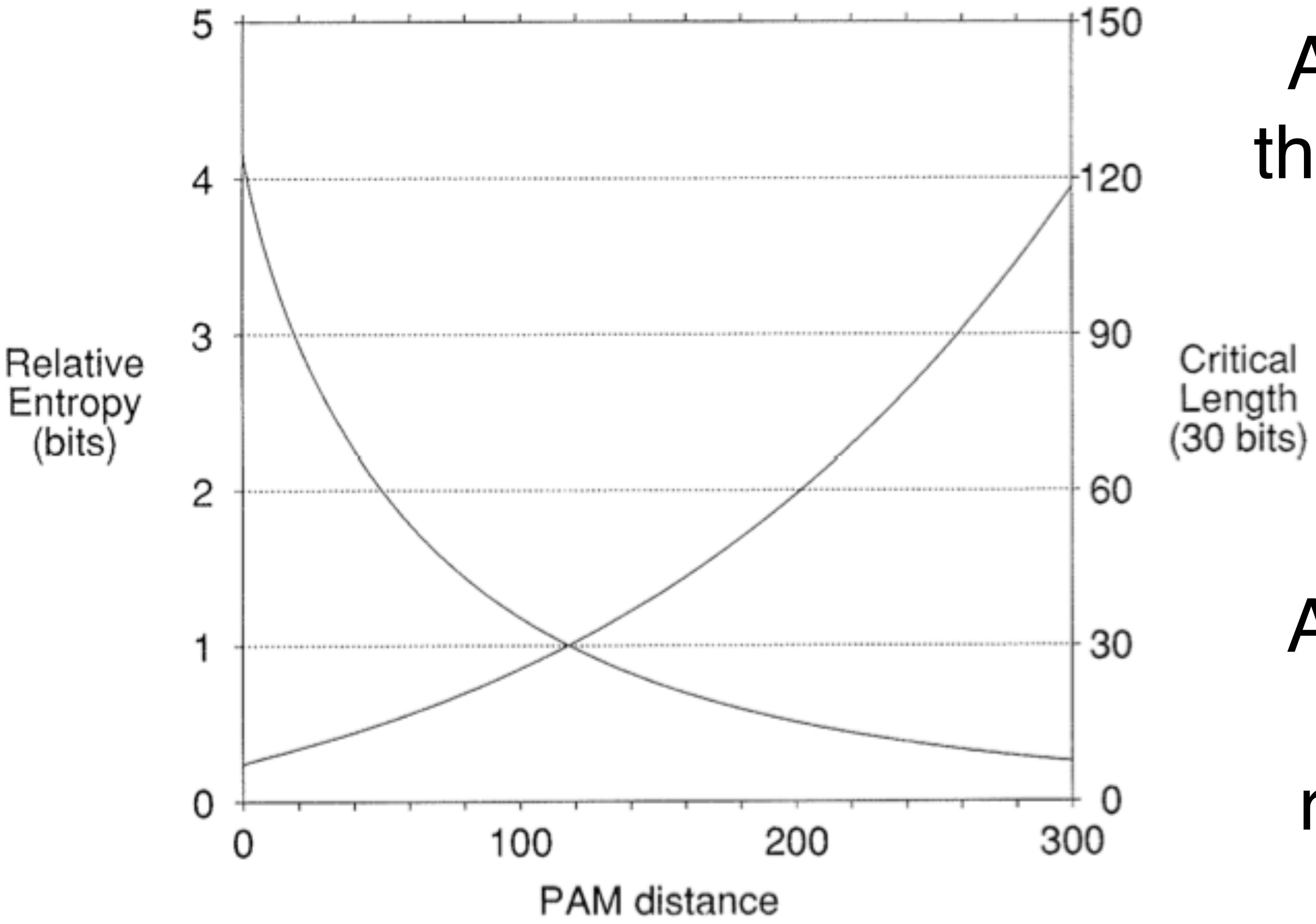
PAM:BLOSUM

PAM100:	BLOSUM90
PAM120:	BLOSUM80
PAM160:	BLOSUM60
PAM200:	BLOSUM52
PAM250:	BLOSUM45

Scoring Matrices

- PAM and BLOSUM matrices greatly improve the sensitivity of protein sequence comparison – low identity with significant similarity
- PAM matrices have an evolutionary model - lower number, less divergence – lower=closer; higher=more distant
- BLOSUM matrices are sampled from conserved regions at different average identity – higher=more conservation
- Short alignments require shallow matrices (closer)
- Shallow matrices set maximum look-back time

Details on Scoring Matrices

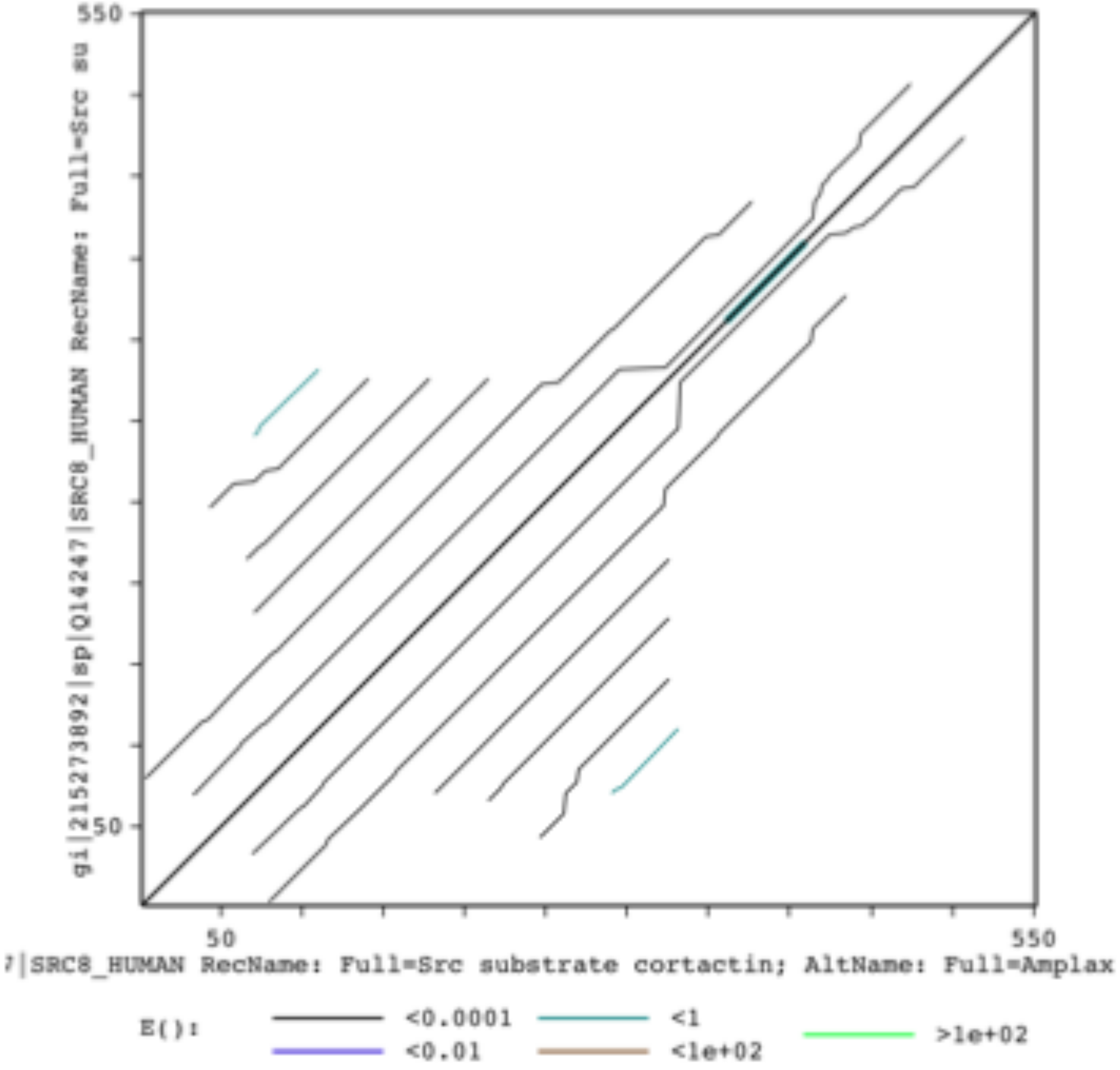


As sequences diverge, there is less information per position

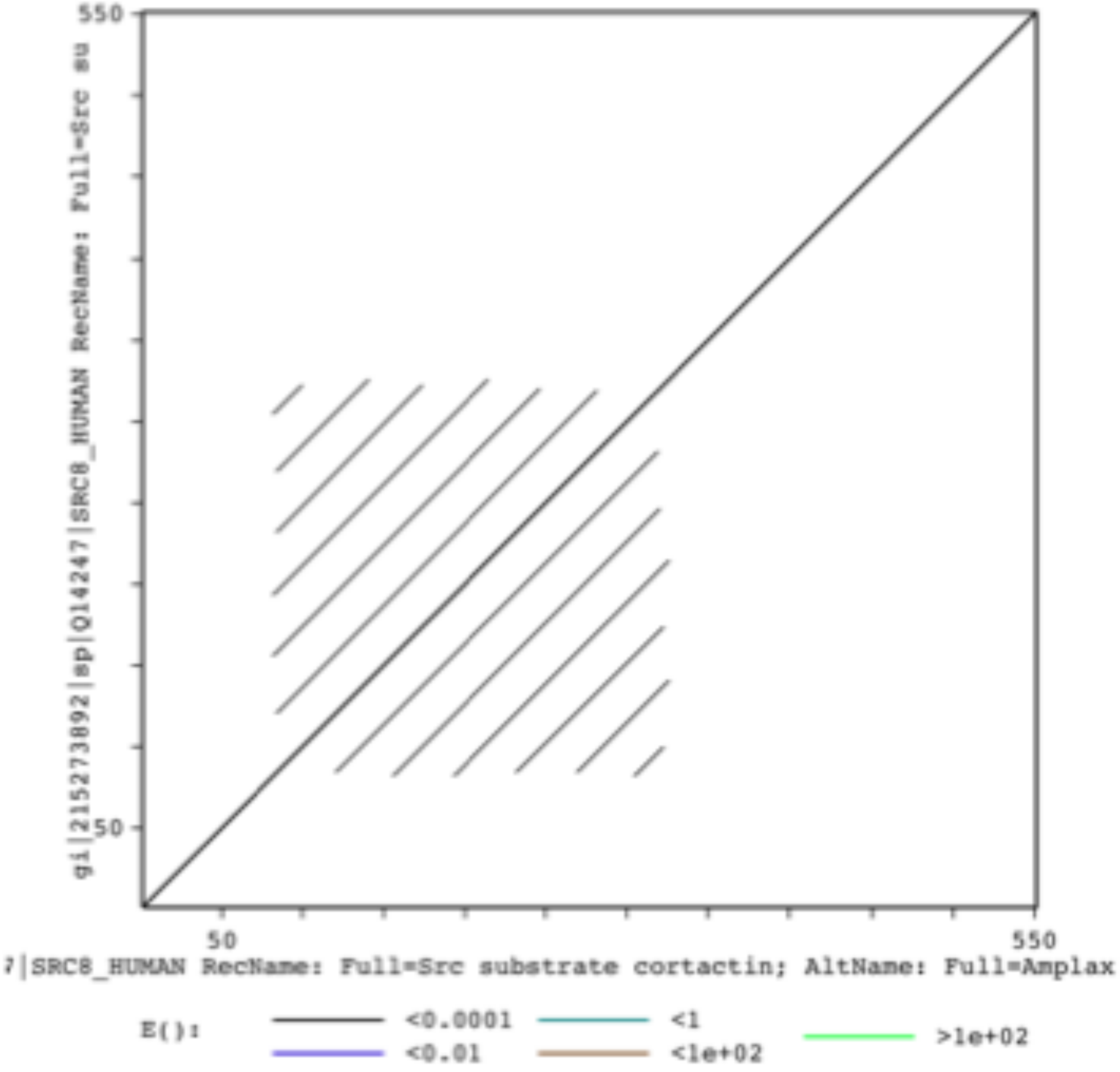
As sequences diverge, longer alignments are required to contain the score threshold

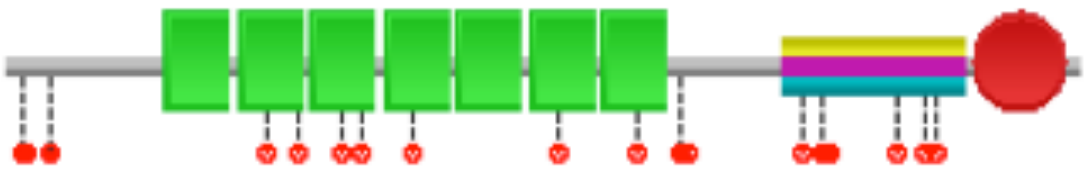
Stringent Score Leads to Short Alignments

BLOSUM62 -11/-1

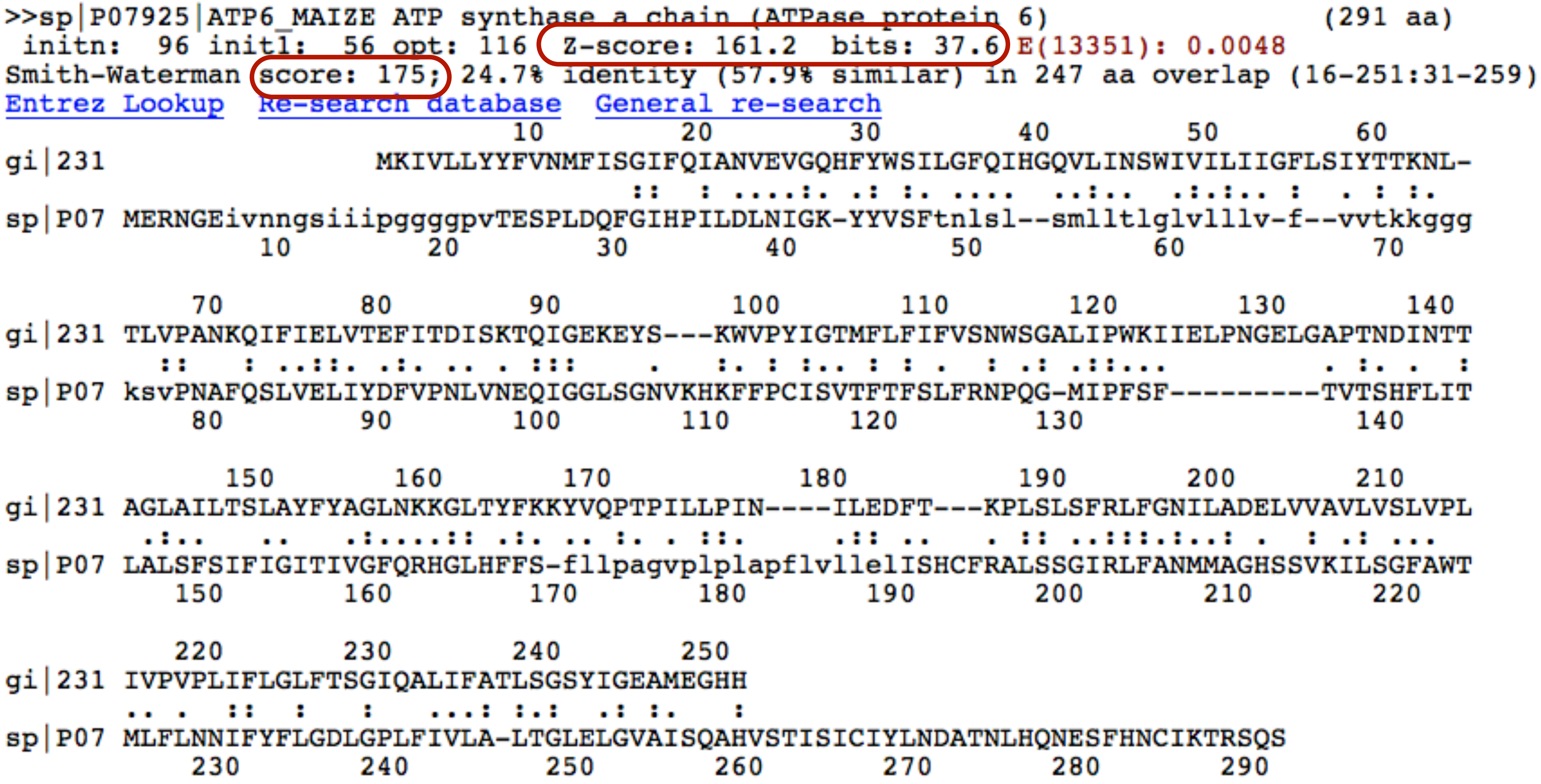


MD20 -26/-4





Scoring Alignments



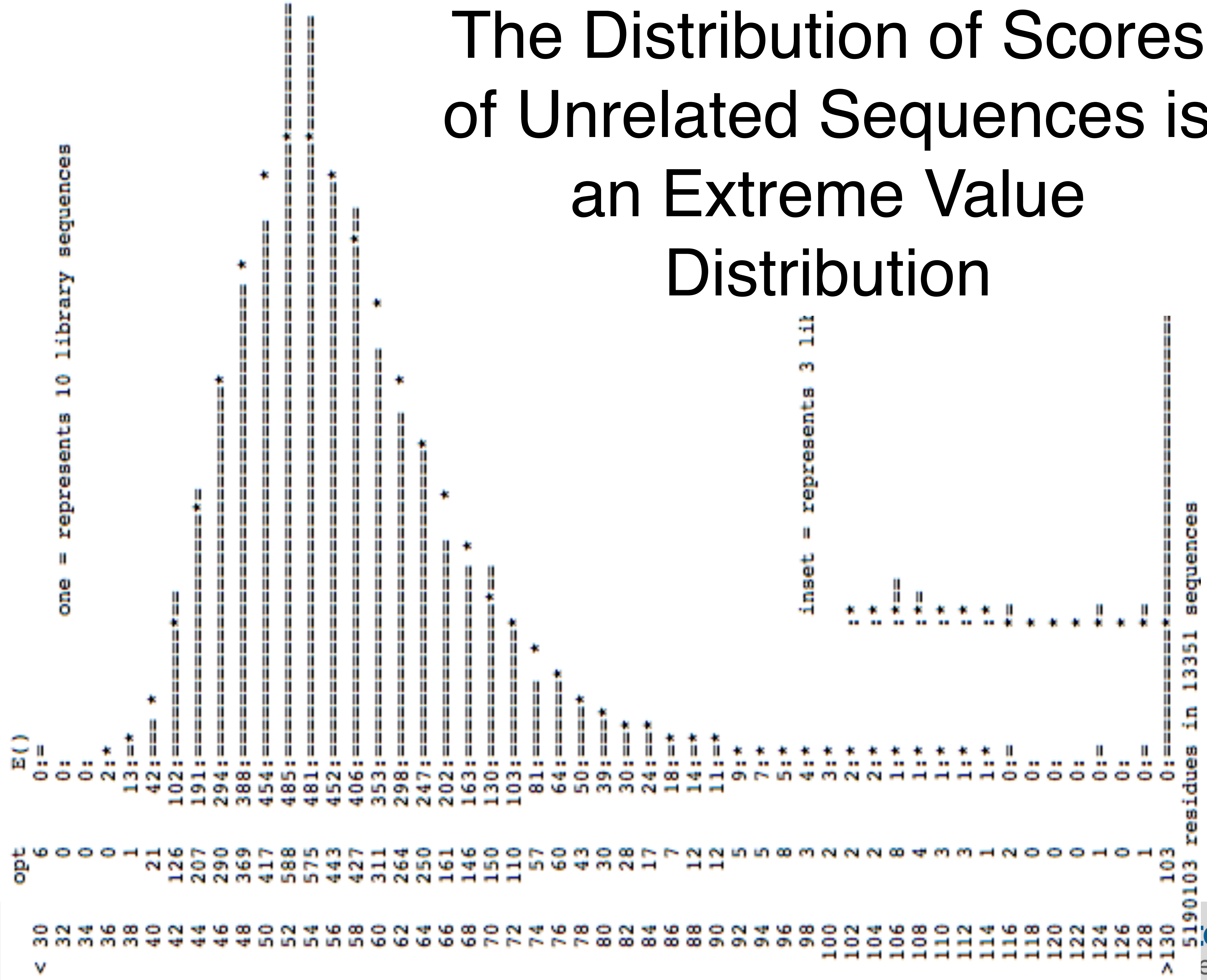
```
>>sp|P07925|ATP6_MAIZE ATP synthase a chain (ATPase protein 6)                (291 aa)
 initn:  96 init1:  56 opt: 116  Z-score: 161.2  bits: 37.6 E(13351): 0.0048
Smith-Waterman score: 175; 24.7% identity (57.9% similar) in 247 aa overlap (16-251:31-259)
Entrez Lookup  Re-search database  General re-search
                              10        20        30        40        50        60
gi|231                MKIVLLYYFVNMFISGIFQIANVEVGQHFYWSILGFQIHGQVLINSWIVILIIGFLSIYTTKNL-
                                     ::  : ....:. .: :. ....  ..:..  .:.:.. :  . : :.
sp|P07 MERNGEivnngsiiipggggpvTESPLDQFGIHPILDLNIGK-YYVSFtnlsl--smlltlglvlllv-f--vvtkkggg
               10        20        30        40         50          60           70

           70        80        90          100       110       120       130       140
gi|231 TLVPANKQIFIELVTEFITDISKTQIGEKEYS---KWVPYIGTMFLFIFVSNWSGALIPWKIIELPNGELGAPTNDINTT
         ::   : ..::. .:. .. . :::    .   :. : :.. : : .  : .: .::...         . :. .  :
sp|P07 ksvPNAFQSLVELIYDFVNPLVNEQIGGLSGNVKHKFFPCISVTFTFSLFRNPQG-MIPFSF---------TVTSHFLIT
           80        90       100       110       120        130                140

             150       160       170           180          190       200       210
gi|231 AGLAILTSLAYFYAGLNKKGLTYFKKYVQPTPILLPIN----ILEDFT---KPLSLSFRLFGNILADELVVAVLVSLVPL
        .:..   ..   .:....:: .:. .. :. . ::.     .:: ..   . :: ..:::.:..: .  : .: ...
sp|P07 LALSFSIFIGITIVGFQRHGLHFFS-fllpagvplplapflvllelISHCFRALSSGIRLFANMMAGHSSVKILSGFAWT
          150       160        170       180       190       200       210       220

          220       230       240       250
gi|231 IVPVPLIFLGLFTSGIQALIFATLSGSYIGEAMEGHH
       ..  .  ::  :   :   ...: :.:  .: :.   :
sp|P07 MLFLNNIFYFLGDLGPLFIVLA-LTGLELGVAISQAHVSTISICIYLNDATNLHQNESFHNCIKTRSQS
           230       240        250       260       270       280       290
```

Alignments are scored using the scoring matrix

Inferring Homology from Statistical Significance

- Real *UNRELATED* sequences have similarity scores that are indistinguishable from *RANDOM* sequences
- If a similarity is NOT *RANDOM,* then it must be NOT *UNRELATED*
- Therefore, NOT *RANDOM* (statistically significant) similarity must reflect *RELATED* sequences

The Distribution of Scores of Unrelated Sequences is an Extreme Value Distribution

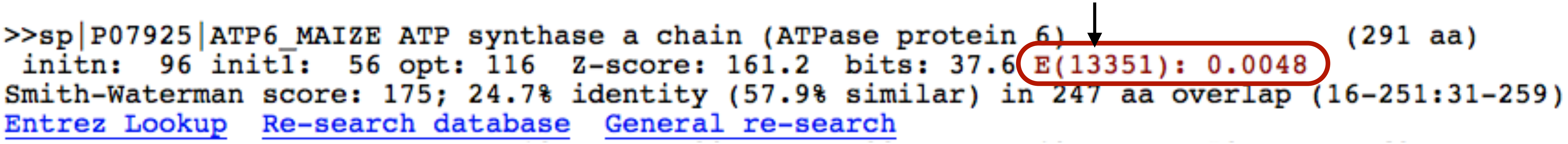


```
      opt      E()
< 30     6     0:=
  32     0     0:            one = represents 10 library sequences
  34     0     0:
  36     0     2:*
  38     1    13:=*
  40    21    42:=== *
  42   126   102:==========*==
  44   207   191:===================*=
  46   290   294:=============================*
  48   369   388:===================================== *
  50   417   454:==========================================   *
  52   588   485:================================================*==========
  54   575   481:================================================*=========
  56   443   452:=============================================*
  58   427   406:========================================*==
  60   311   353:================================   *
  62   264   298:===========================  *
  64   250   247:========================*
  66   161   202:=================   *
  68   146   163:=============== *
  70   150   130:============*==
  72   110   103:==========*
  74    57    81:======  *
  76    60    64:======*
  78    43    50:====*
  80    30    39:===*
  82    28    30:==*
  84    17    24:==*
  86     7    18:=*
  88    12    14:=*
  90    12    11:=*
  92     5     9:*
  94     5     7:*
  96     8     5:*
  98     3     4:*           inset = represents 3 lib
 100     2     3:*
 102     2     2:*          :*
 104     2     2:*          :*
 106     8     1:*          :*==
 108     4     1:*          :*=
 110     3     1:*          :*
 112     3     1:*          :*
 114     1     1:*          :*
 116     2     0:=          *=
 118     0     0:           *
 120     0     0:           *
 122     0     0:           *
 124     1     0:=          *=
 126     0     0:           *
 128     1     0:=          *=
>130   103     0:===========*=======================
 5190103 residues in 13351 sequences
```

What is an Expectation Value

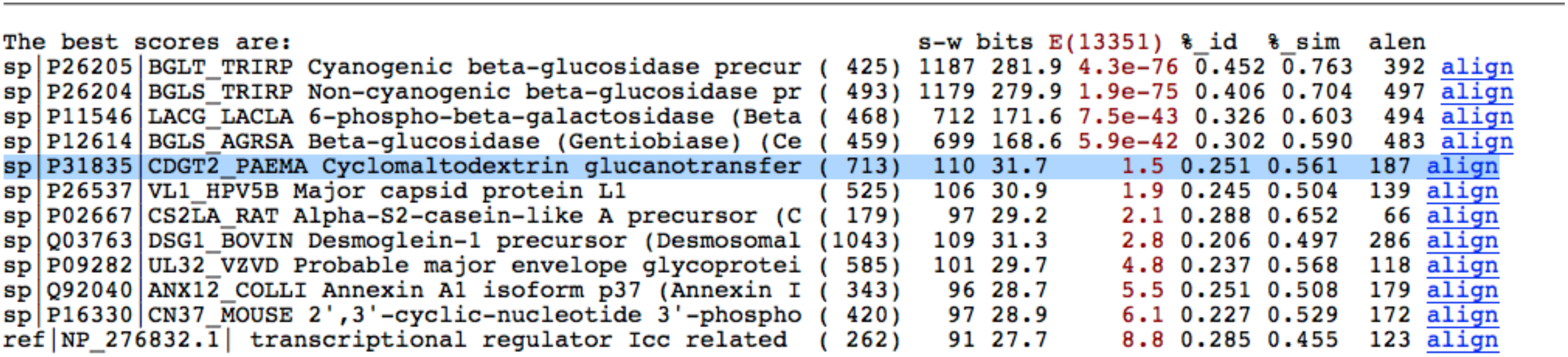
- The Expectation Values is the probability of the score times the number of sequences in your search library
  - The number of times you expect to get that p-value by chance in the search that was performed.

Library Size



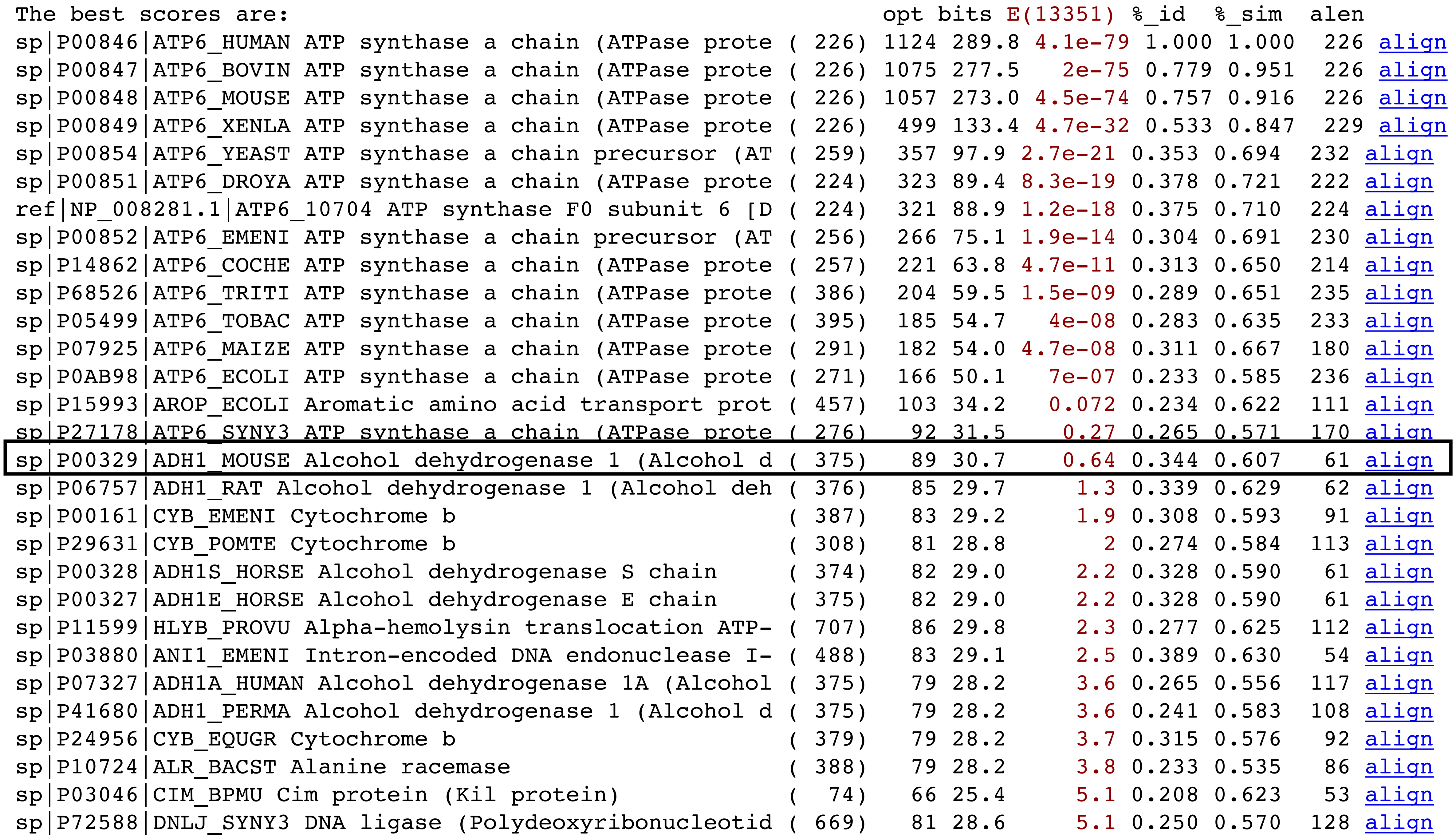
```
>>sp|P07925|ATP6_MAIZE ATP synthase a chain (ATPase protein 6)              (291 aa)
 initn:  96 init1:  56 opt: 116  Z-score: 161.2  bits: 37.6 E(13351): 0.0048
Smith-Waterman score: 175; 24.7% identity (57.9% similar) in 247 aa overlap (16-251:31-259)
Entrez Lookup  Re-search database  General re-search
```

Highest Scoring Unrelated Sequenced E() ~ 1



```
The best scores are:                                         s-w bits E(13351) %_id  %_sim  alen
sp|P26205|BGLT_TRIRP Cyanogenic beta-glucosidase precur ( 425) 1187 281.9 4.3e-76 0.452 0.763  392 align
sp|P26204|BGLS_TRIRP Non-cyanogenic beta-glucosidase pr ( 493) 1179 279.9 1.9e-75 0.406 0.704  497 align
sp|P11546|LACG_LACLA 6-phospho-beta-galactosidase (Beta ( 468)  712 171.6 7.5e-43 0.326 0.603  494 align
sp|P12614|BGLS_AGRSA Beta-glucosidase (Gentiobiase) (Ce ( 459)  699 168.6 5.9e-42 0.302 0.590  483 align
sp|P31835|CDGT2_PAEMA Cyclomaltodextrin glucanotransfer ( 713)  110 31.7       1.5 0.251 0.561  187 align
sp|P26537|VL1_HPV5B Major capsid protein L1             ( 525)  106 30.9       1.9 0.245 0.504  139 align
sp|P02667|CS2LA_RAT Alpha-S2-casein-like A precursor (C ( 179)   97 29.2       2.1 0.288 0.652   66 align
sp|Q03763|DSG1_BOVIN Desmoglein-1 precursor (Desmosomal (1043)  109 31.3       2.8 0.206 0.497  286 align
sp|P09282|UL32_VZVD Probable major envelope glycoprotei ( 585)  101 29.7       4.8 0.237 0.568  118 align
sp|Q92040|ANX12_COLLI Annexin A1 isoform p37 (Annexin I ( 343)   96 28.7       5.5 0.251 0.508  179 align
sp|P16330|CN37_MOUSE 2',3'-cyclic-nucleotide 3'-phospho ( 420)   97 28.9       6.1 0.227 0.529  172 align
ref|NP_276832.1| transcriptional regulator Icc related  ( 262)   91 27.7       8.8 0.285 0.455  123 align
```

Highest Scoring Unrelated Protein



```
The best scores are:                                           opt bits E(13351) %_id  %_sim  alen
sp|P00846|ATP6_HUMAN ATP synthase a chain (ATPase prote ( 226) 1124 289.8 4.1e-79 1.000 1.000  226 align
sp|P00847|ATP6_BOVIN ATP synthase a chain (ATPase prote ( 226) 1075 277.5   2e-75 0.779 0.951  226 align
sp|P00848|ATP6_MOUSE ATP synthase a chain (ATPase prote ( 226) 1057 273.0 4.5e-74 0.757 0.916  226 align
sp|P00849|ATP6_XENLA ATP synthase a chain (ATPase prote ( 226)  499 133.4 4.7e-32 0.533 0.847  229 align
sp|P00854|ATP6_YEAST ATP synthase a chain precursor (AT ( 259)  357 97.9 2.7e-21 0.353 0.694  232 align
sp|P00851|ATP6_DROYA ATP synthase a chain (ATPase prote ( 224)  323 89.4 8.3e-19 0.378 0.721  222 align
ref|NP_008281.1|ATP6_10704 ATP synthase F0 subunit 6 [D ( 224)  321 88.9 1.2e-18 0.375 0.710  224 align
sp|P00852|ATP6_EMENI ATP synthase a chain precursor (AT ( 256)  266 75.1 1.9e-14 0.304 0.691  230 align
sp|P14862|ATP6_COCHE ATP synthase a chain (ATPase prote ( 257)  221 63.8 4.7e-11 0.313 0.650  214 align
sp|P68526|ATP6_TRITI ATP synthase a chain (ATPase prote ( 386)  204 59.5 1.5e-09 0.289 0.651  235 align
sp|P05499|ATP6_TOBAC ATP synthase a chain (ATPase prote ( 395)  185 54.7   4e-08 0.283 0.635  233 align
sp|P07925|ATP6_MAIZE ATP synthase a chain (ATPase prote ( 291)  182 54.0 4.7e-08 0.311 0.667  180 align
sp|P0AB98|ATP6_ECOLI ATP synthase a chain (ATPase prote ( 271)  166 50.1   7e-07 0.233 0.585  236 align
sp|P15993|AROP_ECOLI Aromatic amino acid transport prot ( 457)  103 34.2   0.072 0.234 0.622  111 align
sp|P27178|ATP6_SYNY3 ATP synthase a chain (ATPase prote ( 276)   92 31.5    0.27 0.265 0.571  170 align
sp|P00329|ADH1_MOUSE Alcohol dehydrogenase 1 (Alcohol d ( 375)   89 30.7    0.64 0.344 0.607   61 align
sp|P06757|ADH1_RAT Alcohol dehydrogenase 1 (Alcohol deh ( 376)   85 29.7     1.3 0.339 0.629   62 align
sp|P00161|CYB_EMENI Cytochrome b                        ( 387)   83 29.2     1.9 0.308 0.593   91 align
sp|P29631|CYB_POMTE Cytochrome b                        ( 308)   81 28.8       2 0.274 0.584  113 align
sp|P00328|ADH1S_HORSE Alcohol dehydrogenase S chain     ( 374)   82 29.0     2.2 0.328 0.590   61 align
sp|P00327|ADH1E_HORSE Alcohol dehydrogenase E chain     ( 375)   82 29.0     2.2 0.328 0.590   61 align
sp|P11599|HLYB_PROVU Alpha-hemolysin translocation ATP- ( 707)   86 29.8     2.3 0.277 0.625  112 align
sp|P03880|ANI1_EMENI Intron-encoded DNA endonuclease I- ( 488)   83 29.1     2.5 0.389 0.630   54 align
sp|P07327|ADH1A_HUMAN Alcohol dehydrogenase 1A (Alcohol ( 375)   79 28.2     3.6 0.265 0.556  117 align
sp|P41680|ADH1_PERMA Alcohol dehydrogenase 1 (Alcohol d ( 375)   79 28.2     3.6 0.241 0.583  108 align
sp|P24956|CYB_EQUGR Cytochrome b                        ( 379)   79 28.2     3.7 0.315 0.576   92 align
sp|P10724|ALR_BACST Alanine racemase                    ( 388)   79 28.2     3.8 0.233 0.535   86 align
sp|P03046|CIM_BPMU Cim protein (Kil protein)            (  74)   66 25.4     5.1 0.208 0.623   53 align
sp|P72588|DNLJ_SYNY3 DNA ligase (Polydeoxyribonucleotid ( 669)   81 28.6     5.1 0.250 0.570  128 align
```

Unrelated or Too Distance

```
The best scores are:                                          opt bits E(13351) %_id  %_sim  alen
sp|P0AB98|ATP6_ECOLI ATP synthase a chain (ATPase prote ( 271) 1650 428.4 1.1e-120 1.000 1.000  271 align
sp|P06451|ATPI_SPIOL Chloroplast ATP synthase a chain p ( 247)  161 49.1 1.5e-06 0.270 0.616  211 align
sp|P06289|ATPI_MARPO Chloroplast ATP synthase a chain p ( 248)  161 49.1 1.5e-06 0.261 0.621  211 align
sp|P06452|ATPI_PEA Chloroplast ATP synthase a chain pre ( 247)  158 48.3 2.6e-06 0.274 0.614  223 align
sp|P69371|ATPI_ATRBE Chloroplast ATP synthase a chain p ( 247)  156 47.8 3.7e-06 0.270 0.607  211 align
sp|P00848|ATP6_MOUSE ATP synthase a chain (ATPase prote ( 226)  149 46.0 1.2e-05 0.259 0.617  193 align
sp|P00846|ATP6_HUMAN ATP synthase a chain (ATPase prote ( 226)  148 45.7 1.4e-05 0.237 0.589  236 align
sp|P30391|ATPI_EUGGR Chloroplast ATP synthase a chain p ( 251)  139 43.4 7.6e-05 0.298 0.596  225 align
sp|P00847|ATP6_BOVIN ATP synthase a chain (ATPase prote ( 226)  138 43.2 8.1e-05 0.233 0.581  236 align
sp|P0C2Y5|ATPI_ORYSA Chloroplast ATP synthase a chain p ( 247)  132 41.7 0.00026 0.259 0.603  239 align
sp|P68526|ATP6_TRITI ATP synthase a chain (ATPase prote ( 386)  121 38.9  0.0028 0.259 0.603  239 align
sp|P27178|ATP6_SYNY3 ATP synthase a chain (ATPase prote ( 276)  116 37.6  0.0048 0.264 0.578  258 align
sp|P00854|ATP6_YEAST ATP synthase a chain precursor (AT ( 259)  113 36.8  0.0077 0.235 0.578  277 align
sp|P08444|ATP6_SYNP6 ATP synthase a chain (ATPase prote ( 261)  113 36.8  0.0077 0.267 0.600  240 align
sp|P00852|ATP6_EMENI ATP synthase a chain precursor (AT ( 256)  111 36.3   0.011 0.209 0.590  244 align
sp|P07925|ATP6_MAIZE ATP synthase a chain (ATPase prote ( 291)  109 35.8   0.017 0.259 0.578  232 align
sp|P00851|ATP6_DROYA ATP synthase a chain (ATPase prote ( 224)   98 33.0   0.094 0.225 0.549  253 align
sp|P14862|ATP6_COCHE ATP synthase a chain (ATPase prote ( 257)   91 31.2    0.37 0.204 0.608  265 align
ref|NP_008281.1|ATP6_10704 ATP synthase F0 subunit 6 [D ( 224)   90 31.0    0.39 0.230 0.576  165 align
sp|P09716|US17_HCMVA Hypothetical protein HVLF1         ( 293)   91 31.2    0.42 0.260 0.565  131 align

sp|P12446|MAT_INCJJ Polyprotein p42 [Contains: Protein  ( 374)   85 29.7     1.5 0.247 0.559   93 align

sp|P00849|ATP6_XENLA ATP synthase a chain (ATPase prote ( 226)   79 28.2     2.7 0.261 0.630  165 align

sp|P06974|FLIM_ECOLI Flagellar motor switch protein fli ( 334)   81 28.7     2.8 0.308 0.673   52 align

sp|P05499|ATP6_TOBAC ATP synthase a chain (ATPase prote ( 395)   81 28.7     3.3 0.220 0.582  268 align
```

Proteins Sequences Are Better for Comparing Divergent or Not Well Conserved Genes.

Question you can ask using sequence similarity

Is there an homologous protein?

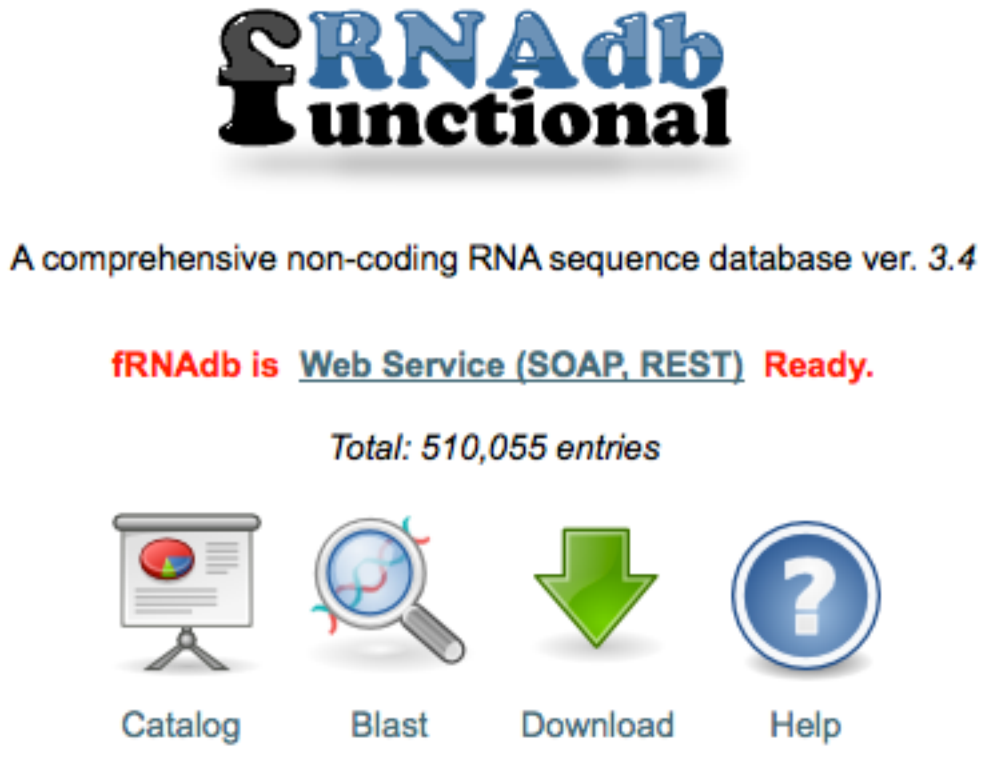
- Does that homologous protein have a similar domain?
- Does XXX genome have YYY (kinase, GPCR, ...)?

Questions not to ask:

- Does this DNA sequence have a similar regulatory element (too short – never significant)?
- Does (non-significant) protein have the same function/ modification/antigenic site?

DNA or Protein

- DNA is better when comparing genomes with few variants (populations within a species) or highly conserved genes or RNA genes.
- Otherwise use Protein Sequences



A comprehensive non-coding RNA sequence database ver. 3.4

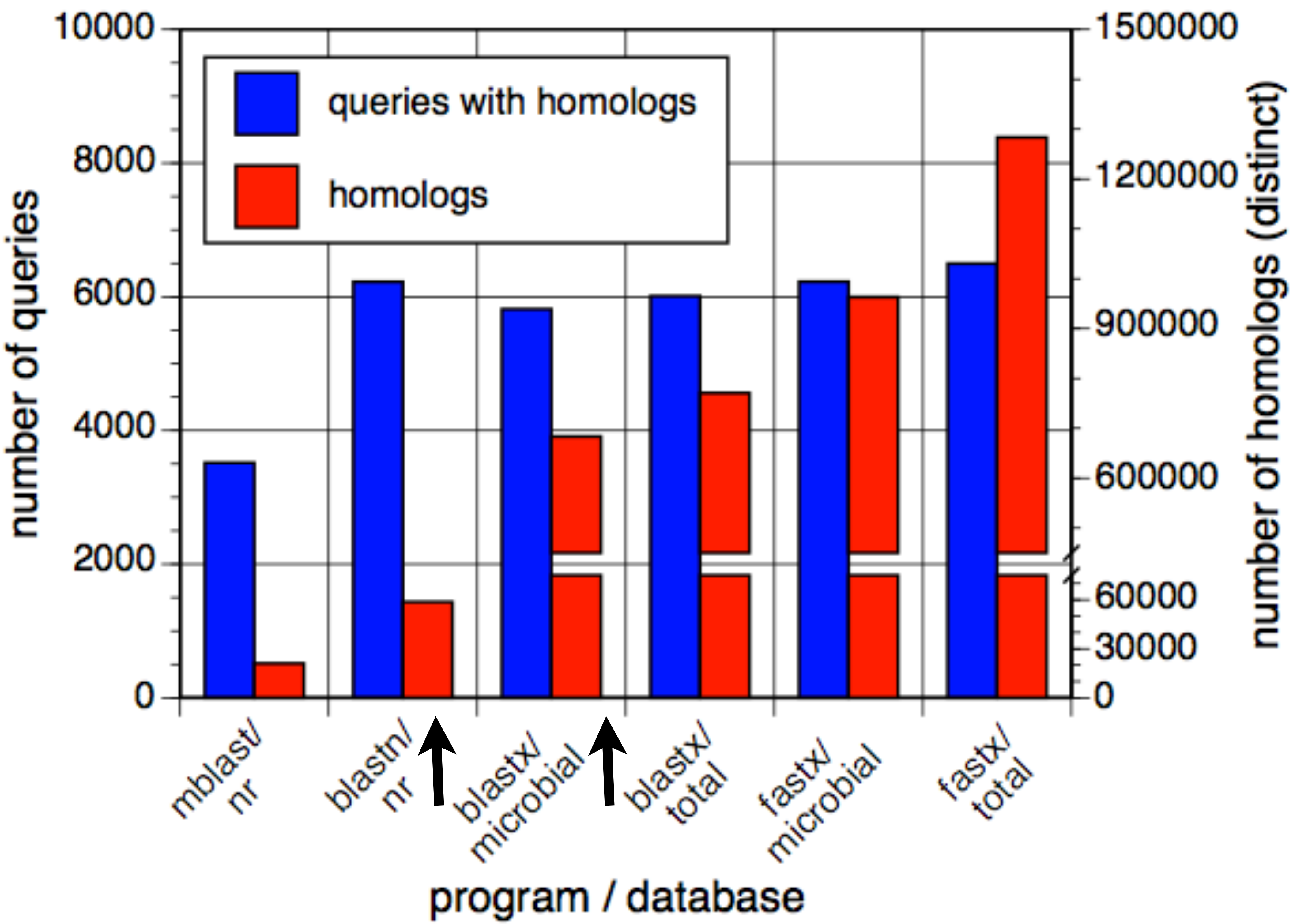
fRNAdb is Web Service (SOAP, REST) Ready.

Total: 510,055 entries

Protein has a longer look back

The best scores are:		DNA	tfastx3	prot.
		E(188,018)	E(187,524)	E(331,956)
DMGST	D.melanogaster GST1-1	1.3e-164	4.1e-109	1.0e-109
MDGST1	M.domestica GST-1 gene	2e-77	3.0e-95	1.9e-76
LUCGLTR	Lucilia cuprina GST	1.5e-72	5.2e-91	3.3e-73
MDGST2A	M.domesticus GST-2 mRNA	9.3e-53	1.4e-77	1.6e-62
MDNF1	M.domestica nf1 gene. 10	4.6e-51	2.8e-77	2.2e-62
MDNF6	M.domestica nf6 gene. 10	2.8e-51	4.2e-77	3.1e-62
MDNF7	M.domestica nf7 gene. 10	6.1e-47	9.2e-77	6.7e-62
AGGST15	A.gambiae GST mRNA	3.1e-58	4.2e-76	4.3e-61
CVU87958	Culicoides GST	1.8e-41	4.0e-73	3.6e-58
AGG3GST11	A.gambiae GST1-1 mRNA	1.5e-46	2.8e-55	1.1e-43
BMO6502	Bombyx mori GST mRNA	1.1e-23	8.8e-50	5.7e-40
AGSUGST12	A.gambiae GST1-1 gene	2.3e-16	4.5e-46	5.1e-37
MOTGLUSTRA	Manduca sexta GST	5.7e-07	2.5e-30	8.0e-25
RLGSTARGN	R.legominosarum *gstA*	0.0029	3.2e-13	1.4e-10
HUMGSTT2A	H. sapiens GSTT2	0.32	3.3e-10	2.0e-09
HSGSTT1	H.sapiens GSTT1 mRNA	7.2	8.4e-13	3.6e-10
ECAE000319	E. coli hypothet. prot.	—	4.7e-10	1.1e-09
MYMDCMA	Methyl. dichlorometh. DH	—	1.1e-09	6.9e-07
BCU19883	Burkholderia maleylacetate red.	—	1.2e-09	1.1e-08
NFU43126	Naegleria fowleri GST	—	3.2e-07	0.0056
SP505GST	Sphingomonas paucim	—	1.8e-06	0.0002
EN1838	H. sapiens maleylaceto. iso.	—	2.1e-06	5.9e-06
HSU86529	Human GSTZ1	—	3.0e-06	8.0e-06
SYCCPNC	Synechocystis GST	—	1.2e-05	9.5e-06
HSEF1GMR	H.sapiens EF1g mRNA	—	9.0e-05	0.00065

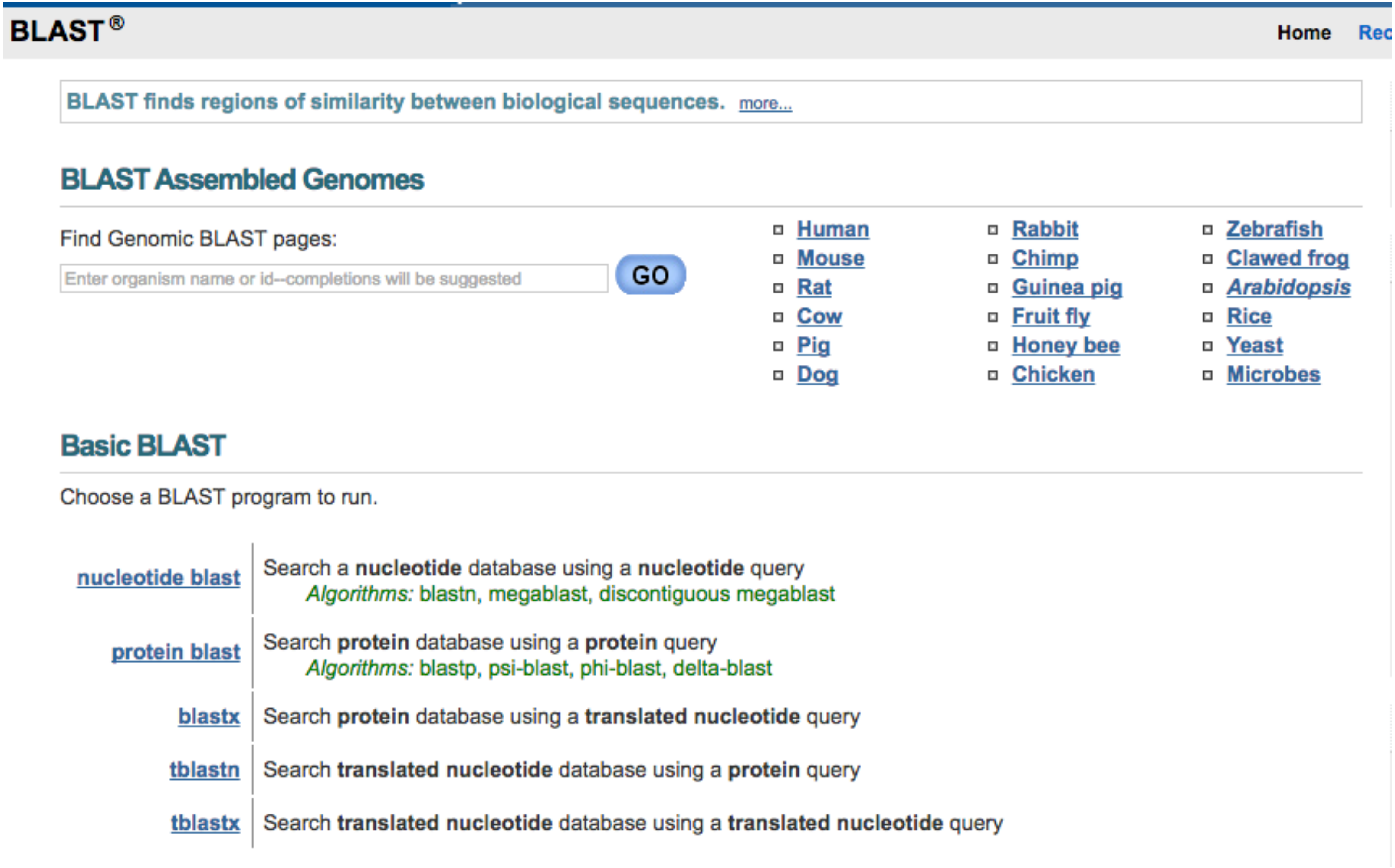
BlastX vs BlastN



What program do I use?

- What is your query sequence?
  - protein: BLAST (NCBI), SSEARCH (EBI)
  - DNA vs Protein: BLASTX (NCBI), FASTX (EBI)
- DNA (structural RNA, repeat family)
  - BLASTN (NCBI), FASTA (EBI)
- Does XXX genome have YYY (protein)?
  - TBLASTN YYY vs XXX genome
  - TFASTX YYY vs XXX genome
- Is Sequence X homologous to Y?
  - BL2SEQ (NCBI), LALIGN, PRSS
- Does my protein contain repeated domains?
  - LALIGN

Sequence Alignment Via the Web



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

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested

- □ Human
- □ Mouse
- □ Rat
- □ Cow
- □ Pig
- □ Dog
- □ Rabbit
- □ Chimp
- □ Guinea pig
- □ Fruit fly
- □ Honey bee
- □ Chicken
- □ Zebrafish
- □ Clawed frog
- □ *Arabidopsis*
- □ Rice
- □ Yeast
- □ Microbes

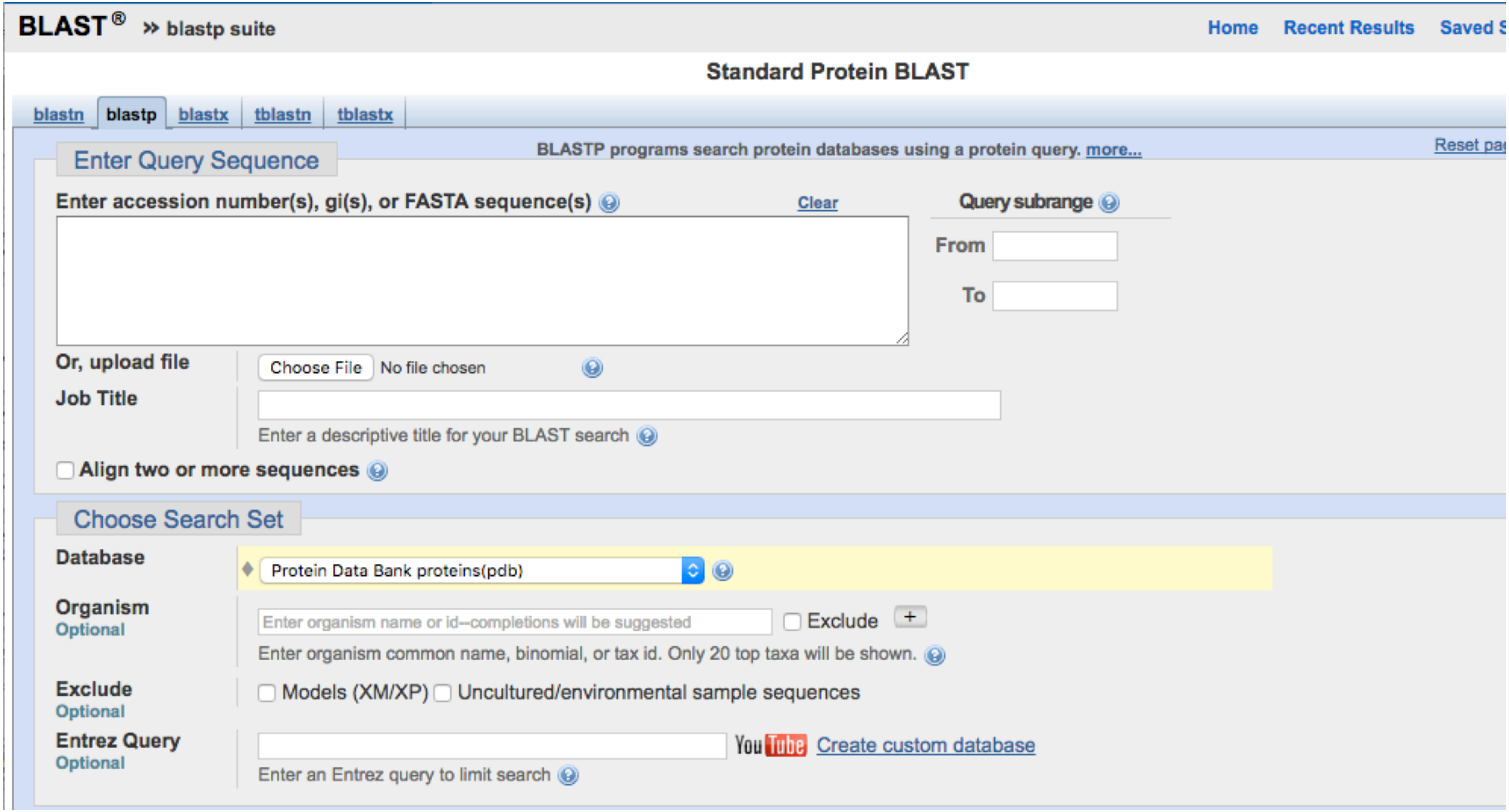
Basic BLAST

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, delta-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

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

Sequence Alignment Via the Web



BLAST® » blastp suite

Home Recent Results Saved S

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

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

Reset pa

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Protein Data Bank proteins(pdb)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

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

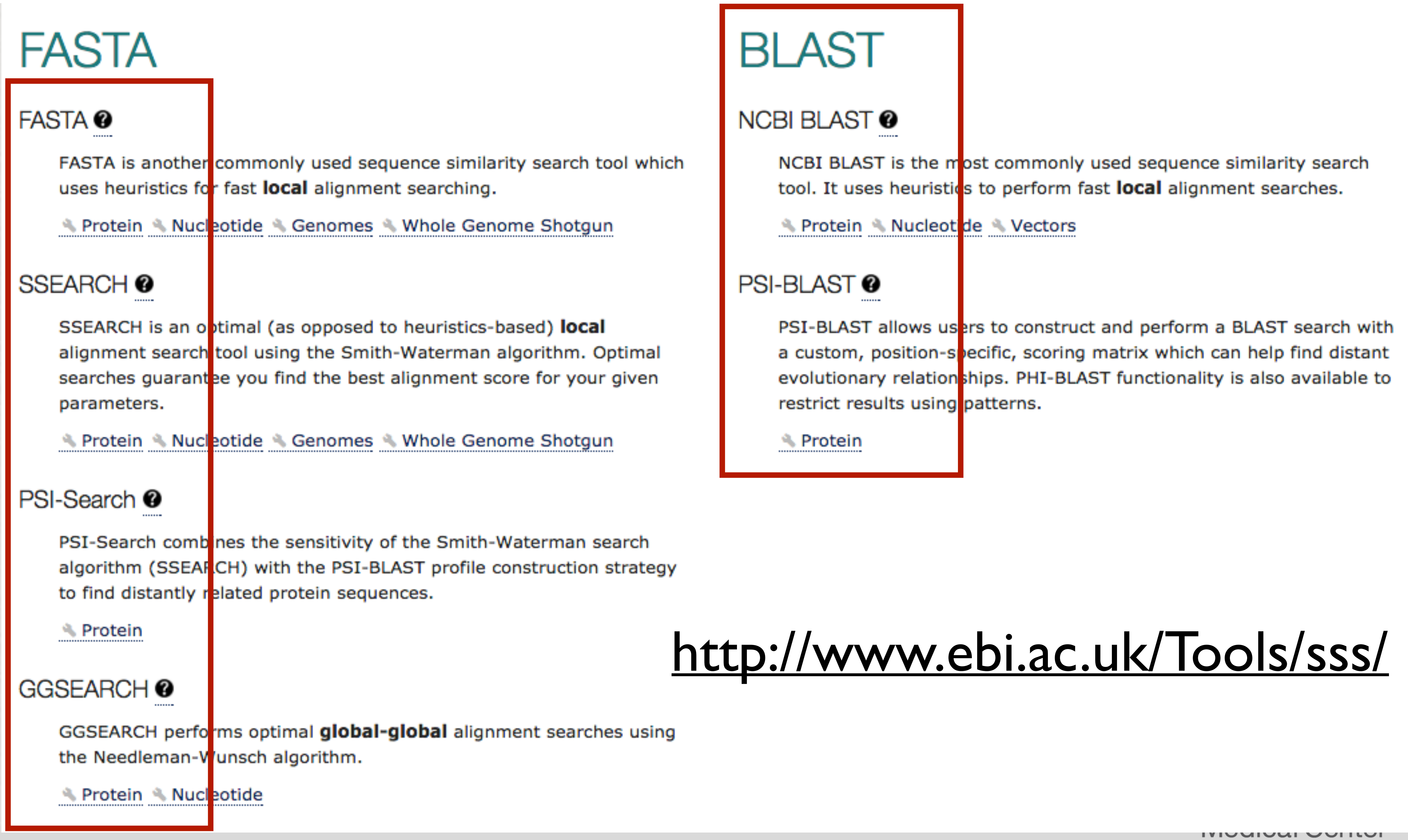
Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query Optional: YouTube Create custom database

Enter an Entrez query to limit search

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

Sequence Alignment Via the Web



FASTA

FASTA

FASTA is another commonly used sequence similarity search tool which uses heuristics for fast **local** alignment searching.

Protein · Nucleotide · Genomes · Whole Genome Shotgun

SSEARCH

SSEARCH is an optimal (as opposed to heuristics-based) **local** alignment search tool using the Smith-Waterman algorithm. Optimal searches guarantee you find the best alignment score for your given parameters.

Protein · Nucleotide · Genomes · Whole Genome Shotgun

PSI-Search

PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH) with the PSI-BLAST profile construction strategy to find distantly related protein sequences.

Protein

GGSEARCH

GGSEARCH performs optimal **global-global** alignment searches using the Needleman-Wunsch algorithm.

Protein · Nucleotide

BLAST

NCBI BLAST

NCBI BLAST is the most commonly used sequence similarity search tool. It uses heuristics to perform fast **local** alignment searches.

Protein · Nucleotide · Vectors

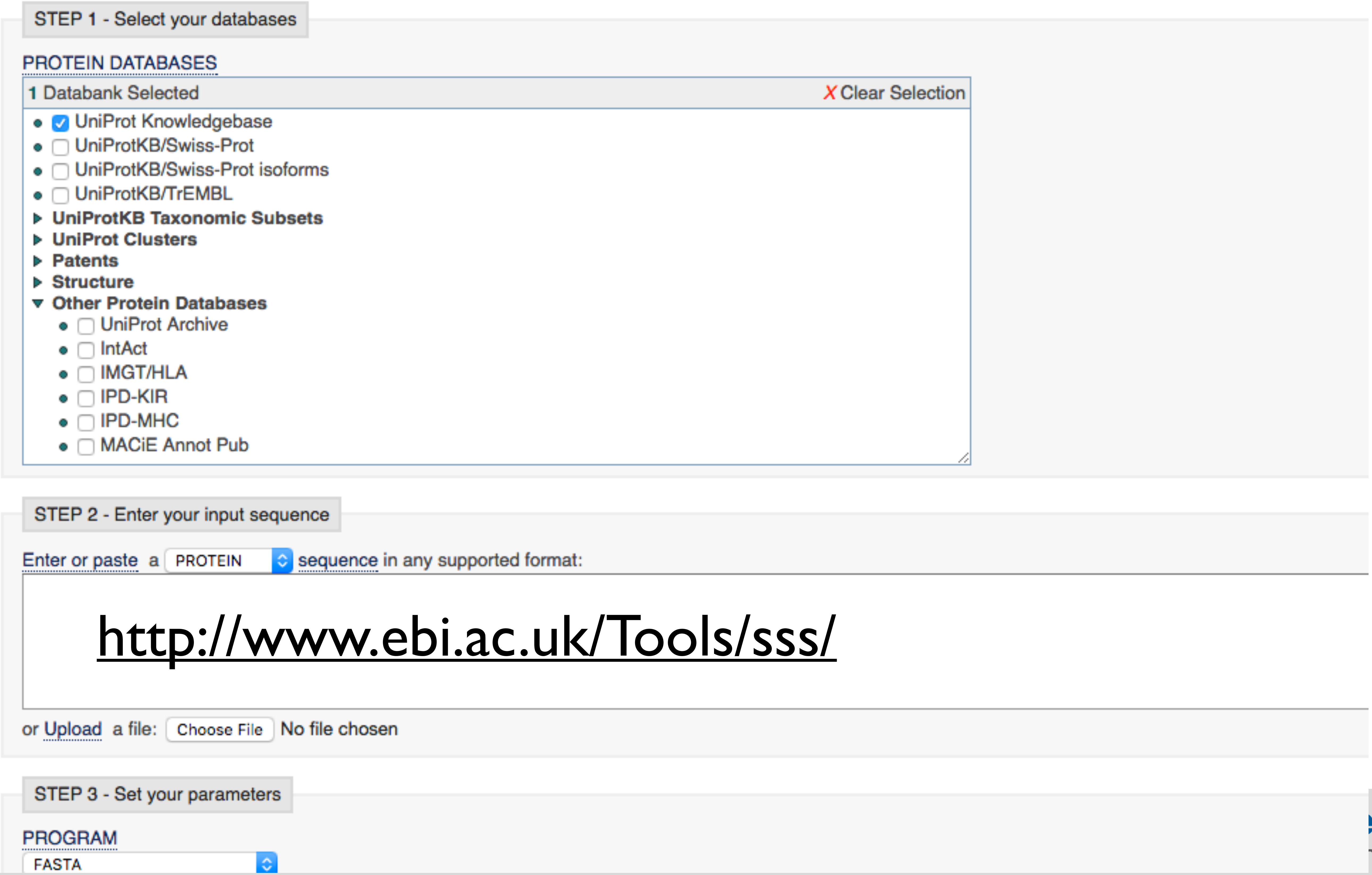
PSI-BLAST

PSI-BLAST allows users to construct and perform a BLAST search with a custom, position-specific, scoring matrix which can help find distant evolutionary relationships. PHI-BLAST functionality is also available to restrict results using patterns.

Protein

http://www.ebi.ac.uk/Tools/sss/

Sequence Alignment Via the Web



STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected X Clear Selection

- ☑ UniProt Knowledgebase
- □ UniProtKB/Swiss-Prot
- □ UniProtKB/Swiss-Prot isoforms
- □ UniProtKB/TrEMBL
- ▶ **UniProtKB Taxonomic Subsets**
- ▶ **UniProt Clusters**
- ▶ **Patents**
- ▶ **Structure**
- ▼ **Other Protein Databases**
  - □ UniProt Archive
  - □ IntAct
  - □ IMGT/HLA
  - □ IPD-KIR
  - □ IPD-MHC
  - □ MACiE Annot Pub

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

http://www.ebi.ac.uk/Tools/sss/

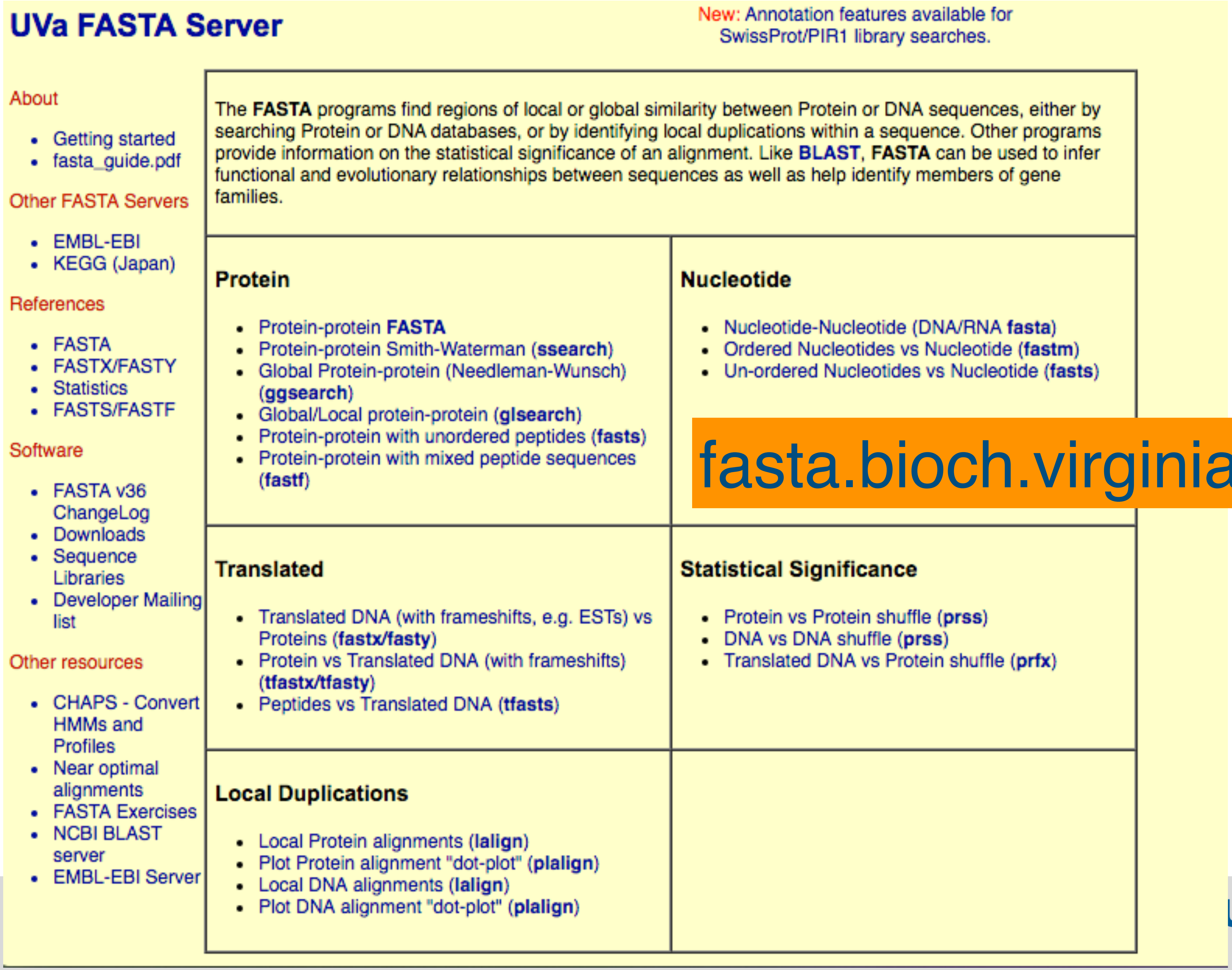
or Upload a file: Choose File No file chosen

STEP 3 - Set your parameters

PROGRAM

FASTA

Sequence Alignment Via the Web



UVa FASTA Server

New: Annotation features available for SwissProt/PIR1 library searches.

About

- Getting started
- fasta_guide.pdf

Other FASTA Servers

- EMBL-EBI
- KEGG (Japan)

References

- FASTA
- FASTX/FASTY
- Statistics
- FASTS/FASTF

Software

- FASTA v36 ChangeLog
- Downloads
- Sequence Libraries
- Developer Mailing list

Other resources

- CHAPS - Convert HMMs and Profiles
- Near optimal alignments
- FASTA Exercises
- NCBI BLAST server
- EMBL-EBI Server

The **FASTA** programs find regions of local or global similarity between Protein or DNA sequences, either by searching Protein or DNA databases, or by identifying local duplications within a sequence. Other programs provide information on the statistical significance of an alignment. Like **BLAST**, **FASTA** can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Protein

- Protein-protein **FASTA**
- Protein-protein Smith-Waterman (**ssearch**)
- Global Protein-protein (Needleman-Wunsch) (**ggsearch**)
- Global/Local protein-protein (**glsearch**)
- Protein-protein with unordered peptides (**fasts**)
- Protein-protein with mixed peptide sequences (**fastf**)

Nucleotide

- Nucleotide-Nucleotide (DNA/RNA **fasta**)
- Ordered Nucleotides vs Nucleotide (**fastm**)
- Un-ordered Nucleotides vs Nucleotide (**fasts**)

Translated

- Translated DNA (with frameshifts, e.g. ESTs) vs Proteins (**fastx/fasty**)
- Protein vs Translated DNA (with frameshifts) (**tfastx/tfasty**)
- Peptides vs Translated DNA (**tfasts**)

Statistical Significance

- Protein vs Protein shuffle (**prss**)
- DNA vs DNA shuffle (**prss**)
- Translated DNA vs Protein shuffle (**prfx**)

Local Duplications

- Local Protein alignments (**lalign**)
- Plot Protein alignment "dot-plot" (**plalign**)
- Local DNA alignments (**lalign**)
- Plot DNA alignment "dot-plot" (**plalign**)

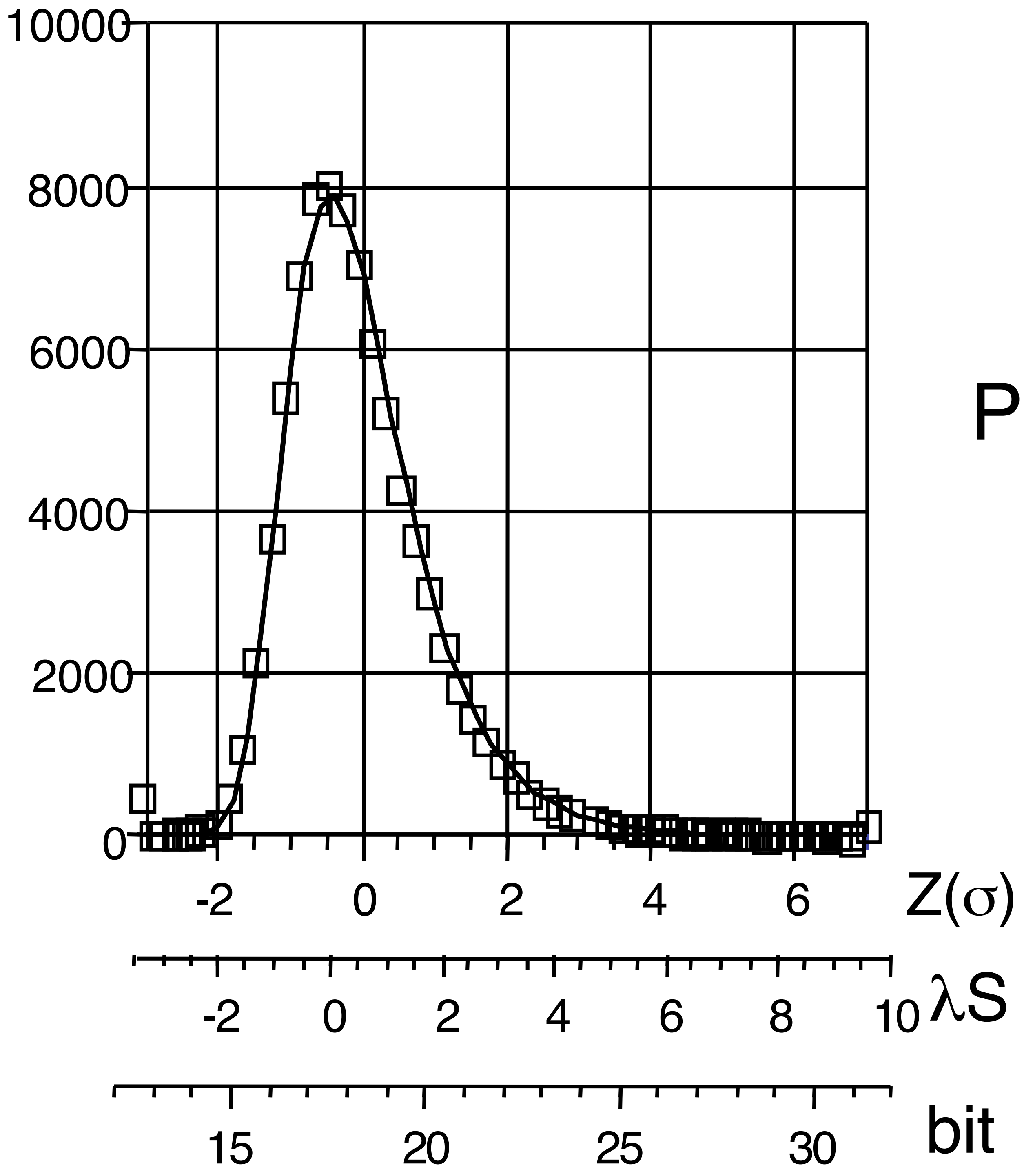
fasta.bioch.virginia.edu

What Database to Search?

- Search the smallest comprehensive database likely to contain your protein of interest
    - vertebrates – human proteins (40,000)
    - fungi – S. cerevisiae (6,000)
    - bacteria – E. coli, gram positive, etc. (<100,000)
- Search a richly annotated protein set (SwissProt, 450,000)
- Always search NR (> 12 million) *LAST*

Never Search “GenBank” (DNA)

DB Size Matters
Smaller is Better



$$S' = \lambda S_{raw} - \ln K\,m\,n$$
$$S_{bit} = (\lambda S_{raw} - \ln K)/\ln(2)$$
$$P(S'>x) = 1 - \exp(-e^{-x})$$
$$P(S_{bit} > x) = 1 - \exp(-mn2^{-x})$$
$$E(S'>x \mid D) = P\,D$$

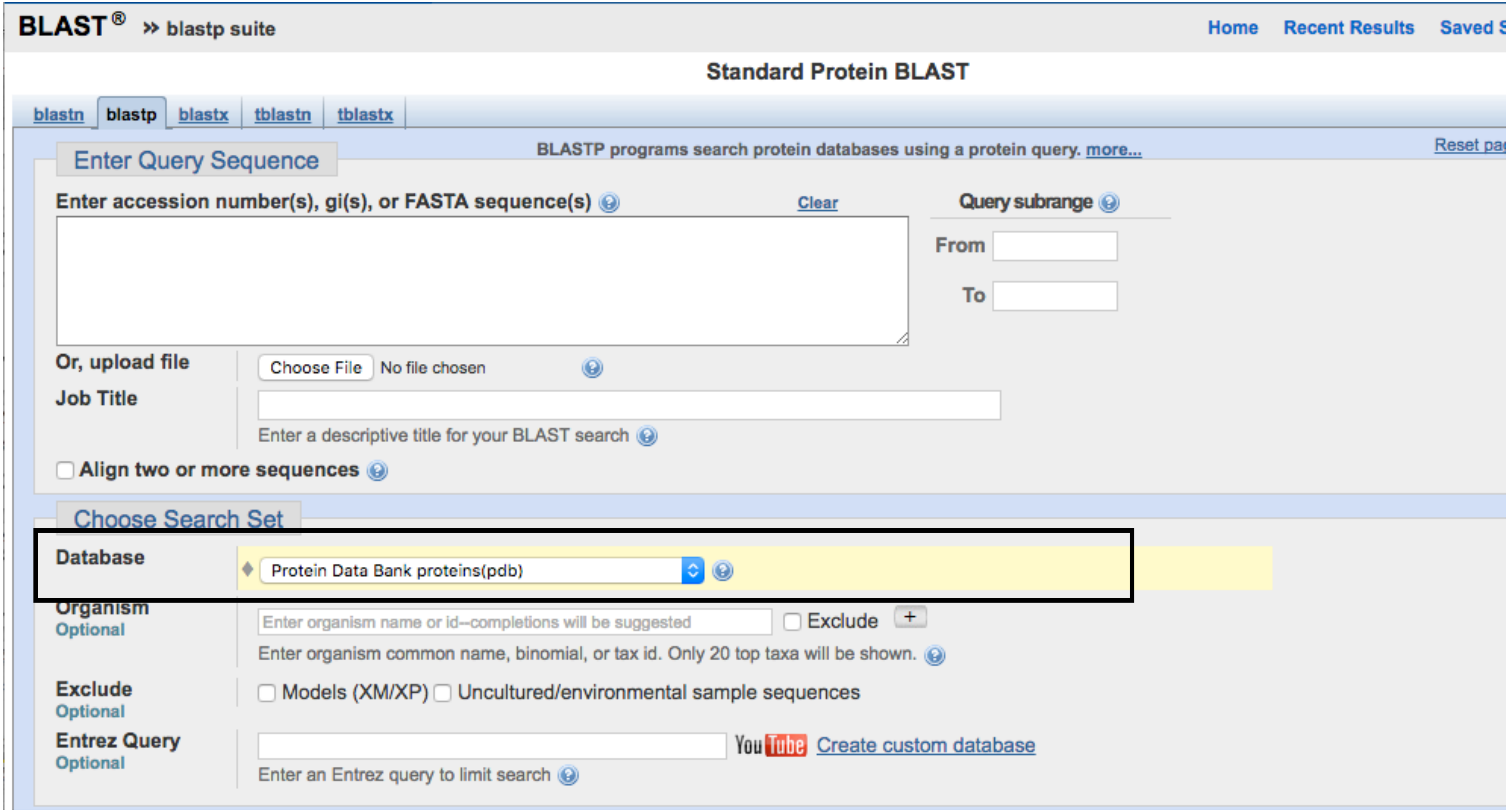
$$P(B\ \text{bits}) = m\,n\,2^{-B}$$
$$P(40\ \text{bits}) = 1.5\text{x}10^{-7}$$
$$E(40 \mid D=4000) = 6\text{x}10^{-4}$$
$$E(40 \mid D=12E6) = 1.8$$

DB Size Matters
Smaller is Better

```
gi|114443|sp|P00846.1|ATP6_HUMAN ATP synthase subunit a; F-ATPase - 226 aa
                                              vs
gi|16131606|ref|NP_418194.1| F0 sector of membrane-bound ATP synthase, subunit a [Escherichia
coli str. K-12 subst - 271 aa

                initn: 159 init1: 104 opt: 148  Z-score: 212.5  bits: 46.8
               Smith-Waterman score: 178; 23.7% identity (58.9% similar) in 236 aa overlap (45-264:8-222)
```

Database	Entries	Length	E()	Time (s)
E.Coli	4237	1350094	3.8E-07	<0.5
Human Ref	38000	17401176	1.9E-05	1
SwissProt	445410	165796297	0.0015	10
RefSeq	711441	261324908	NS	16

How Can I Choose my DB?



BLAST® » blastp suite

Home Recent Results Saved S

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

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

Reset pa

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Protein Data Bank proteins(pdb)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

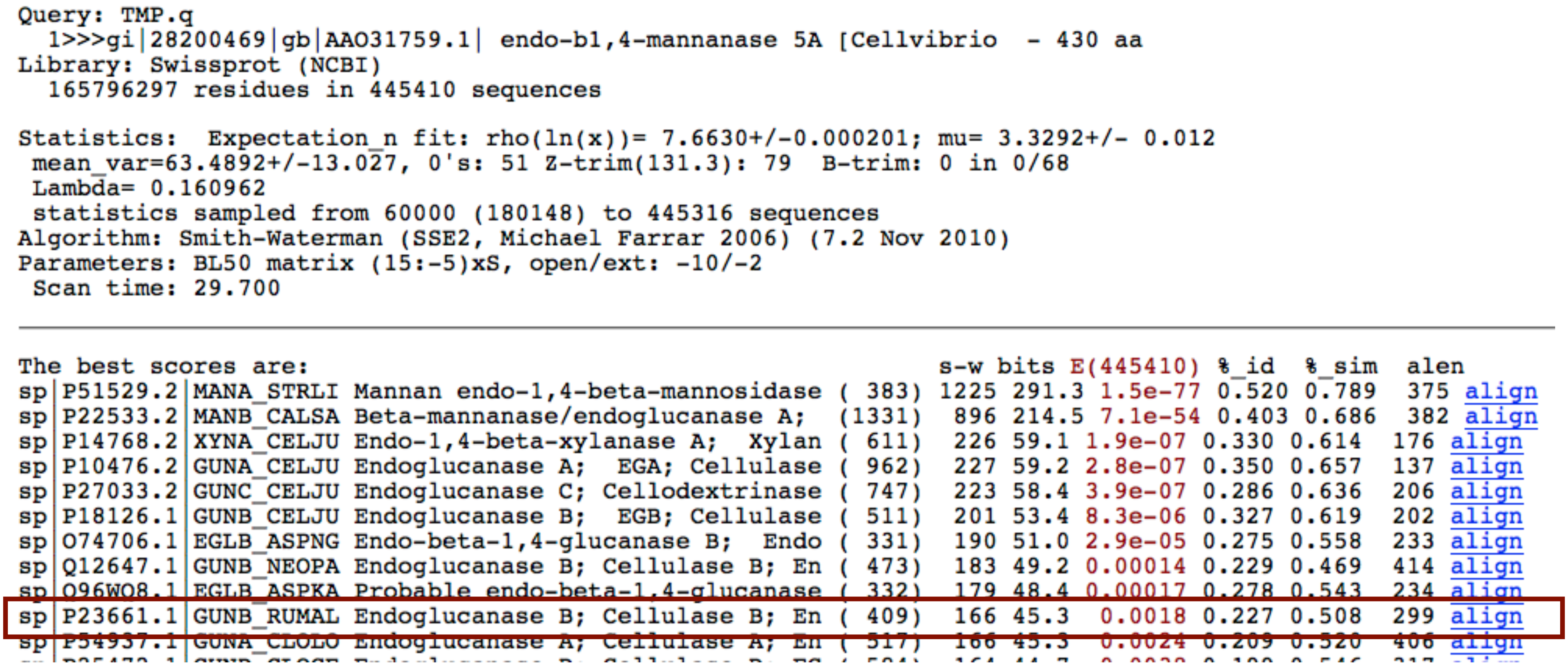
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

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

Entrez Query Optional: You Tube Create custom database

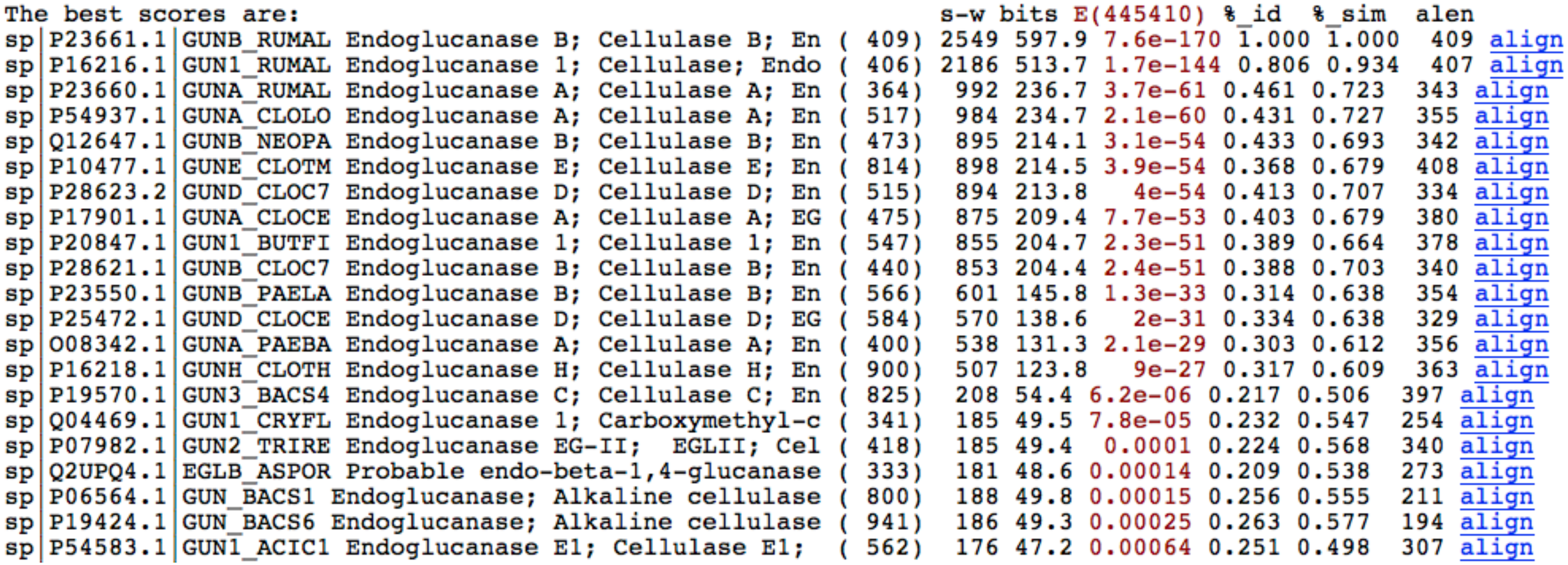
Enter an Entrez query to limit search

How can you tell what is the Highest Scoring Unrelated Hit?



```
Query: TMP.q
  1>>>gi|28200469|gb|AAO31759.1| endo-b1,4-mannanase 5A [Cellvibrio  - 430 aa
Library: Swissprot (NCBI)
  165796297 residues in 445410 sequences

Statistics:  Expectation_n fit: rho(ln(x))= 7.6630+/-0.000201; mu= 3.3292+/- 0.012
 mean_var=63.4892+/-13.027, 0's: 51 Z-trim(131.3): 79  B-trim: 0 in 0/68
 Lambda= 0.160962
 statistics sampled from 60000 (180148) to 445316 sequences
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (7.2 Nov 2010)
Parameters: BL50 matrix (15:-5)xS, open/ext: -10/-2
 Scan time: 29.700

The best scores are:                                      s-w bits E(445410) %_id  %_sim  alen
sp|P51529.2|MANA_STRLI Mannan endo-1,4-beta-mannosidase ( 383) 1225 291.3 1.5e-77 0.520 0.789  375 align
sp|P22533.2|MANB_CALSA Beta-mannanase/endoglucanase A;  (1331)  896 214.5 7.1e-54 0.403 0.686  382 align
sp|P14768.2|XYNA_CELJU Endo-1,4-beta-xylanase A;  Xylan ( 611)  226 59.1 1.9e-07 0.330 0.614  176 align
sp|P10476.2|GUNA_CELJU Endoglucanase A;  EGA; Cellulase ( 962)  227 59.2 2.8e-07 0.350 0.657  137 align
sp|P27033.2|GUNC_CELJU Endoglucanase C; Cellodextrinase ( 747)  223 58.4 3.9e-07 0.286 0.636  206 align
sp|P18126.1|GUNB_CELJU Endoglucanase B;  EGB; Cellulase ( 511)  201 53.4 8.3e-06 0.327 0.619  202 align
sp|O74706.1|EGLB_ASPNG Endo-beta-1,4-glucanase B;  Endo ( 331)  190 51.0 2.9e-05 0.275 0.558  233 align
sp|Q12647.1|GUNB_NEOPA Endoglucanase B; Cellulase B; En ( 473)  183 49.2 0.00014 0.229 0.469  414 align
sp|Q96WQ8.1|EGLB_ASPKA Probable endo-beta-1,4-glucanase ( 332)  179 48.4 0.00017 0.278 0.543  234 align
sp|P23661.1|GUNB_RUMAL Endoglucanase B; Cellulase B; En ( 409)  166 45.3  0.0018 0.227 0.508  299 align
sp|P54937.1|GUNA_CLOLO Endoglucanase A; Cellulase A; En ( 517)  166 45.3  0.0024 0.209 0.520  406 align
sp|P25472.1|GUNB_CLOCE Endoglucanase B; Cellulase B; EG ( 584)  164 44.7  0.0028 0.190 0.546  317 align
```

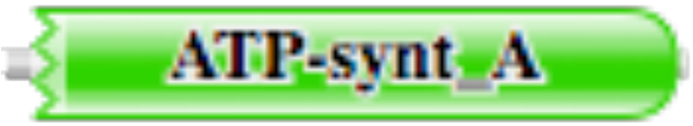
Perform a search with your "suspect"



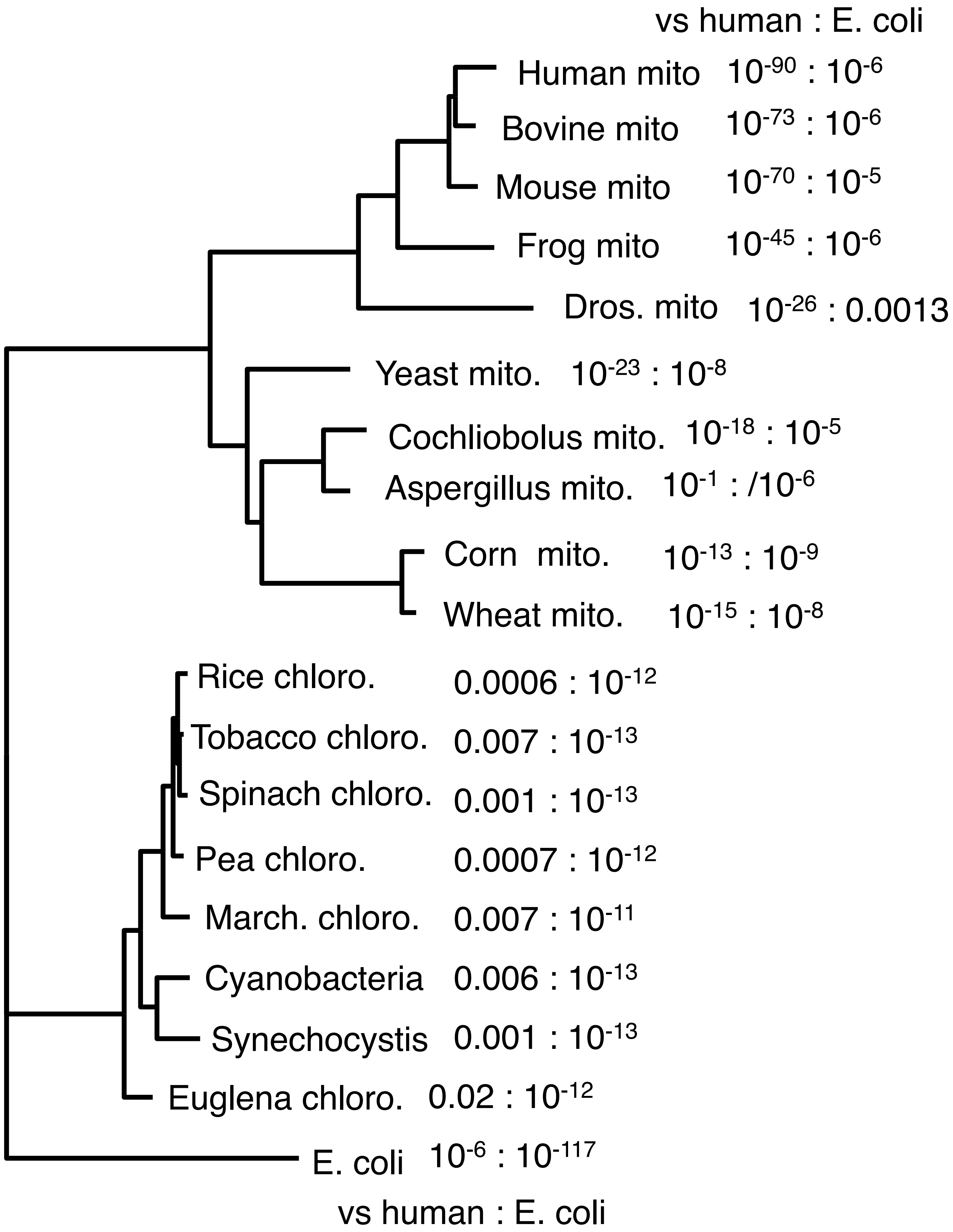
```
The best scores are:                                        s-w bits E(445410) %_id  %_sim  alen
sp|P23661.1|GUNB_RUMAL Endoglucanase B; Cellulase B; En ( 409) 2549 597.9 7.6e-170 1.000 1.000  409 align
sp|P16216.1|GUN1_RUMAL Endoglucanase 1; Cellulase; Endo ( 406) 2186 513.7 1.7e-144 0.806 0.934  407 align
sp|P23660.1|GUNA_RUMAL Endoglucanase A; Cellulase A; En ( 364)  992 236.7 3.7e-61 0.461 0.723  343 align
sp|P54937.1|GUNA_CLOLO Endoglucanase A; Cellulase A; En ( 517)  984 234.7 2.1e-60 0.431 0.727  355 align
sp|Q12647.1|GUNB_NEOPA Endoglucanase B; Cellulase B; En ( 473)  895 214.1 3.1e-54 0.433 0.693  342 align
sp|P10477.1|GUNE_CLOTM Endoglucanase E; Cellulase E; En ( 814)  898 214.5 3.9e-54 0.368 0.679  408 align
sp|P28623.2|GUND_CLOC7 Endoglucanase D; Cellulase D; En ( 515)  894 213.8    4e-54 0.413 0.707  334 align
sp|P17901.1|GUNA_CLOCE Endoglucanase A; Cellulase A; EG ( 475)  875 209.4 7.7e-53 0.403 0.679  380 align
sp|P20847.1|GUN1_BUTFI Endoglucanase 1; Cellulase 1; En ( 547)  855 204.7 2.3e-51 0.389 0.664  378 align
sp|P28621.1|GUNB_CLOC7 Endoglucanase B; Cellulase B; En ( 440)  853 204.4 2.4e-51 0.388 0.703  340 align
sp|P23550.1|GUNB_PAELA Endoglucanase B; Cellulase B; En ( 566)  601 145.8 1.3e-33 0.314 0.638  354 align
sp|P25472.1|GUND_CLOCE Endoglucanase D; Cellulase D; EG ( 584)  570 138.6    2e-31 0.334 0.638  329 align
sp|O08342.1|GUNA_PAEBA Endoglucanase A; Cellulase A; En ( 400)  538 131.3 2.1e-29 0.303 0.612  356 align
sp|P16218.1|GUNH_CLOTH Endoglucanase H; Cellulase H; En ( 900)  507 123.8    9e-27 0.317 0.609  363 align
sp|P19570.1|GUN3_BACS4 Endoglucanase C; Cellulase C; En ( 825)  208 54.4 6.2e-06 0.217 0.506  397 align
sp|Q04469.1|GUN1_CRYFL Endoglucanase 1; Carboxymethyl-c ( 341)  185 49.5 7.8e-05 0.232 0.547  254 align
sp|P07982.1|GUN2_TRIRE Endoglucanase EG-II;  EGLII; Cel ( 418)  185 49.4  0.0001 0.224 0.568  340 align
sp|Q2UPQ4.1|EGLB_ASPOR Probable endo-beta-1,4-glucanase ( 333)  181 48.6 0.00014 0.209 0.538  273 align
sp|P06564.1|GUN_BACS1 Endoglucanase; Alkaline cellulase ( 800)  188 49.8 0.00015 0.256 0.555  211 align
sp|P19424.1|GUN_BACS6 Endoglucanase; Alkaline cellulase ( 941)  186 49.3 0.00025 0.263 0.577  194 align
sp|P54583.1|GUN1_ACIC1 Endoglucanase E1; Cellulase E1;  ( 562)  176 47.2 0.00064 0.251 0.498  307 align
```

Is a hit from your original search in the re-search?

Homology through Transitivity



How do you pick the right sequence homologous to both?



vs human : E. coli

Human mito 10^{-90} : 10^{-6}

Bovine mito 10^{-73} : 10^{-6}

Mouse mito 10^{-70} : 10^{-5}

Frog mito 10^{-45} : 10^{-6}

Dros. mito 10^{-26} : 0.0013

Yeast mito. 10^{-23} : 10^{-8}

Cochliobolus mito. 10^{-18} : 10^{-5}

Aspergillus mito. 10^{-1} : /10^{-6}

Corn mito. 10^{-13} : 10^{-9}

Wheat mito. 10^{-15} : 10^{-8}

Rice chloro. 0.0006 : 10^{-12}

Tobacco chloro. 0.007 : 10^{-13}

Spinach chloro. 0.001 : 10^{-13}

Pea chloro. 0.0007 : 10^{-12}

March. chloro. 0.007 : 10^{-11}

Cyanobacteria 0.006 : 10^{-13}

Synechocystis 0.001 : 10^{-13}

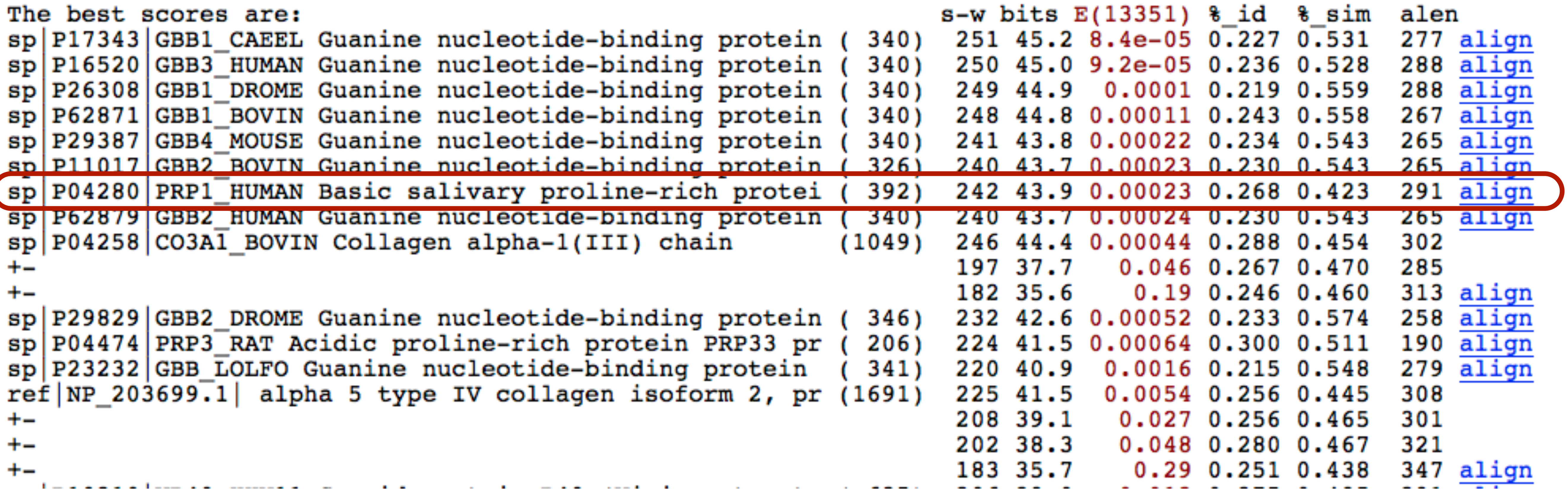
Euglena chloro. 0.02 : 10^{-12}

E. coli 10^{-6} : 10^{-117}

vs human : E. coli

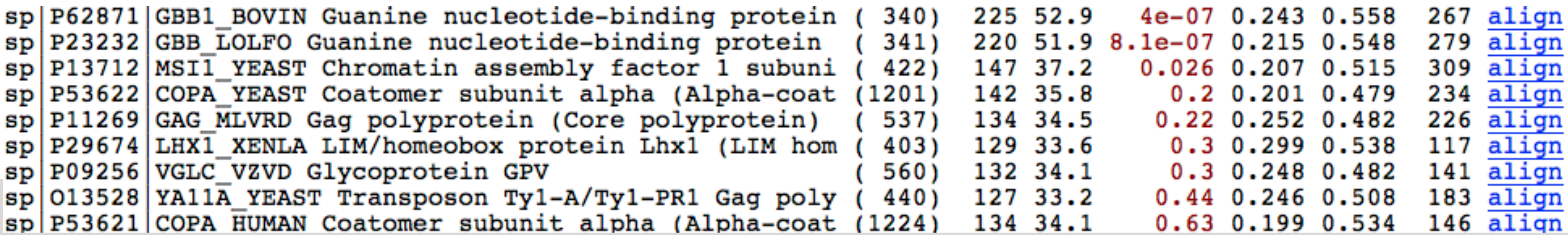
Unrelated ≠ Random

low complexity sequence



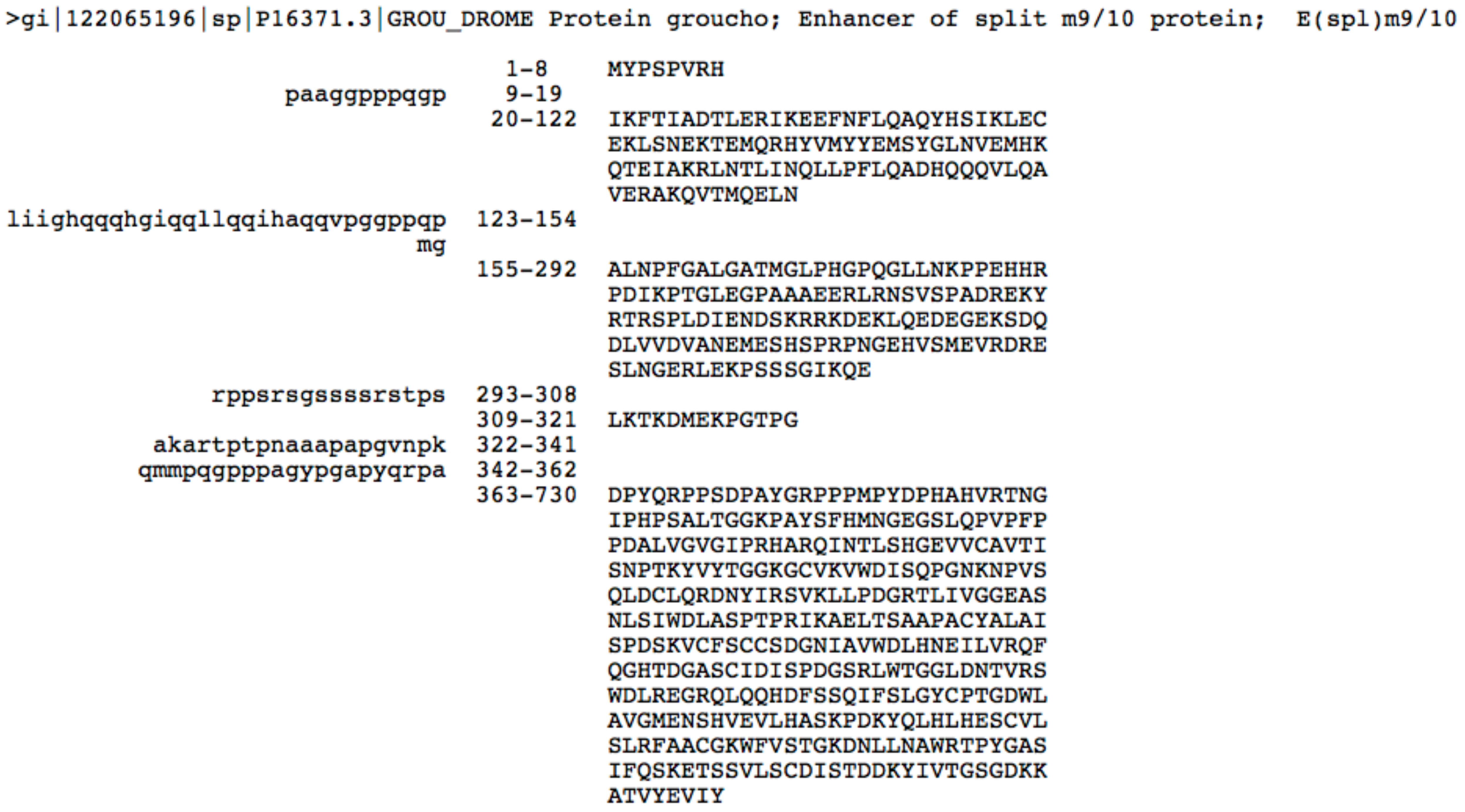
```
The best scores are:                                         s-w bits E(13351) %_id  %_sim  alen
sp|P17343|GBB1_CAEEL Guanine nucleotide-binding protein ( 340)  251 45.2 8.4e-05 0.227 0.531  277 align
sp|P16520|GBB3_HUMAN Guanine nucleotide-binding protein ( 340)  250 45.0 9.2e-05 0.236 0.528  288 align
sp|P26308|GBB1_DROME Guanine nucleotide-binding protein ( 340)  249 44.9  0.0001 0.219 0.559  288 align
sp|P62871|GBB1_BOVIN Guanine nucleotide-binding protein ( 340)  248 44.8 0.00011 0.243 0.558  267 align
sp|P29387|GBB4_MOUSE Guanine nucleotide-binding protein ( 340)  241 43.8 0.00022 0.234 0.543  265 align
sp|P11017|GBB2_BOVIN Guanine nucleotide-binding protein ( 326)  240 43.7 0.00023 0.230 0.543  265 align
sp|P04280|PRP1_HUMAN Basic salivary proline-rich protei ( 392)  242 43.9 0.00023 0.268 0.423  291 align
sp|P62879|GBB2_HUMAN Guanine nucleotide-binding protein ( 340)  240 43.7 0.00024 0.230 0.543  265 align
sp|P04258|CO3A1_BOVIN Collagen alpha-1(III) chain       (1049)  246 44.4 0.00044 0.288 0.454  302
+-                                                              197 37.7   0.046 0.267 0.470  285
+-                                                              182 35.6    0.19 0.246 0.460  313 align
sp|P29829|GBB2_DROME Guanine nucleotide-binding protein ( 346)  232 42.6 0.00052 0.233 0.574  258 align
sp|P04474|PRP3_RAT Acidic proline-rich protein PRP33 pr ( 206)  224 41.5 0.00064 0.300 0.511  190 align
sp|P23232|GBB_LOLFO Guanine nucleotide-binding protein  ( 341)  220 40.9  0.0016 0.215 0.548  279 align
ref|NP_203699.1| alpha 5 type IV collagen isoform 2, pr (1691)  225 41.5  0.0054 0.256 0.445  308
+-                                                              208 39.1   0.027 0.256 0.465  301
+-                                                              202 38.3   0.048 0.280 0.467  321
+-                                                              183 35.7    0.29 0.251 0.438  347 align
```

Filter Low Complexity (SEG)



```
sp|P62871|GBB1_BOVIN Guanine nucleotide-binding protein ( 340)  225 52.9   4e-07 0.243 0.558  267 align
sp|P23232|GBB_LOLFO Guanine nucleotide-binding protein  ( 341)  220 51.9 8.1e-07 0.215 0.548  279 align
sp|P13712|MSI1_YEAST Chromatin assembly factor 1 subuni ( 422)  147 37.2   0.026 0.207 0.515  309 align
sp|P53622|COPA_YEAST Coatomer subunit alpha (Alpha-coat (1201)  142 35.8     0.2 0.201 0.479  234 align
sp|P11269|GAG_MLVRD Gag polyprotein (Core polyprotein)  ( 537)  134 34.5    0.22 0.252 0.482  226 align
sp|P29674|LHX1_XENLA LIM/homeobox protein Lhx1 (LIM hom ( 403)  129 33.6     0.3 0.299 0.538  117 align
sp|P09256|VGLC_VZVD Glycoprotein GPV                    ( 560)  132 34.1     0.3 0.248 0.482  141 align
sp|O13528|YA11A_YEAST Transposon Ty1-A/Ty1-PR1 Gag poly ( 440)  127 33.2    0.44 0.246 0.508  183 align
sp|P53621|COPA_HUMAN Coatomer subunit alpha (Alpha-coat (1224)  134 34.1    0.63 0.199 0.534  146 align
```

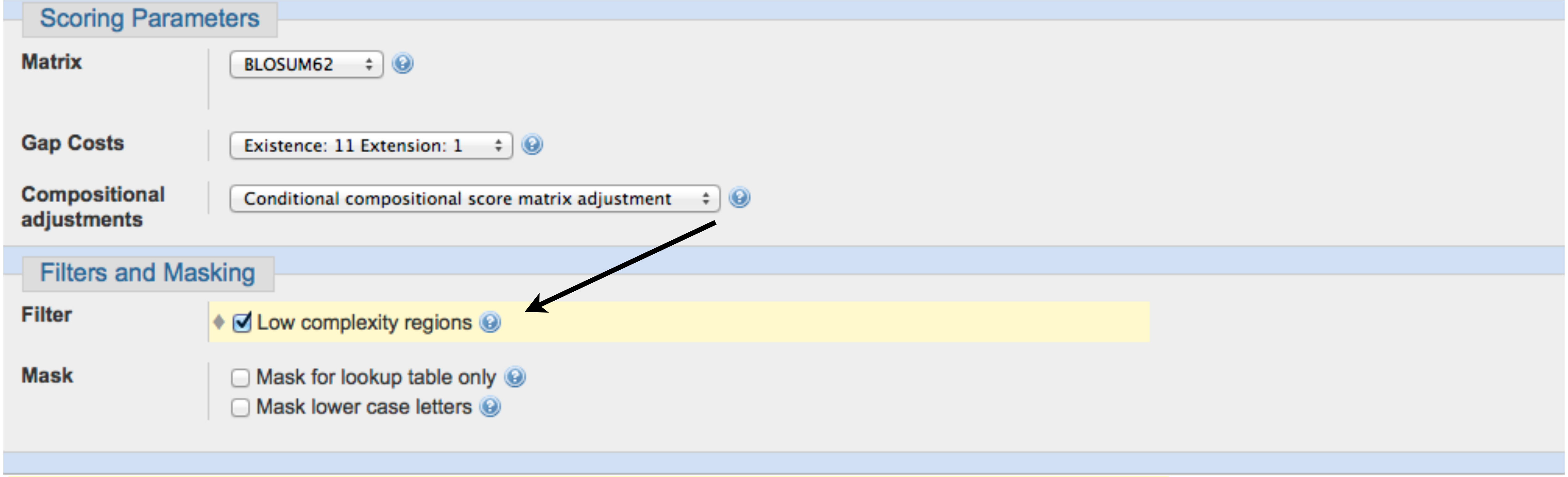
SEG Remove Low Complexity



```
>gi|122065196|sp|P16371.3|GROU_DROME Protein groucho; Enhancer of split m9/10 protein;  E(spl)m9/10

                                  1-8    MYPSPVRH
                   paaggpppqgp    9-19
                                 20-122  IKFTIADTLERIKEEFNFLQAQYHSIKLEC
                                         EKLSNEKTEMQRHYVMYYEMSYGLNVEMHK
                                         QTEIAKRLNTLINQLLPFLQADHQQQVLQA
                                         VERAKQVTMQELN
liighqqhgiqqllqqihaqqvpggppqp  123-154
                           mg
                                155-292  ALNPFGALGATMGLPHGPQGLLNKPPEHHR
                                         PDIKPTGLEGPAAAEERLRNSVSPADREKY
                                         RTRSPLDIENDSKRRKDEKLQEDEGEKSDQ
                                         DLVVDVANEMESHSPRPNGEHVSMEVRDRE
                                         SLNGERLEKPSSSGIKQE
              rppsrsgssssrstps  293-308
                                309-321  LKTKDMEKPGTPG
          akartptpnaaapapgvnpk  322-341
         qmmpqgpppagypgapyqrpa  342-362
                                363-730  DPYQRPPSDPAYGRPPPMPYDPHAHVRTNG
                                         IPHPSALTGGKPAYSFHMNGEGSLQPVPFP
                                         PDALVGVGIPRHARQINTLSHGEVVCAVTI
                                         SNPTKYVYTGGKGCVKVWDISQPGNKNPVS
                                         QLDCLQRDNYIRSVKLLPDGRTLIVGGEAS
                                         NLSIWDLASPTPRIKAELTSAAPACYALAI
                                         SPDSKVCFSCCSDGNIAVWDLHNEILVRQF
                                         QGHTDGASCIDISPDGSRLWTGGLDNTVRS
                                         WDLREGRQLQQHDFSSQIFSLGYCPTGDWL
                                         AVGMENSHVEVLHASKPDKYQLHLHESCVL
                                         SLRFAACGKWFVSTGKDNLLNAWRTPYGAS
                                         IFQSKETSSVLSCDISTDDKYIVTGSGDKK
                                         ATVYEVIY
```

http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=misc1&pgm=seg

SEG Remove Low Complexity



Scoring Parameters

Matrix: BLOSUM62

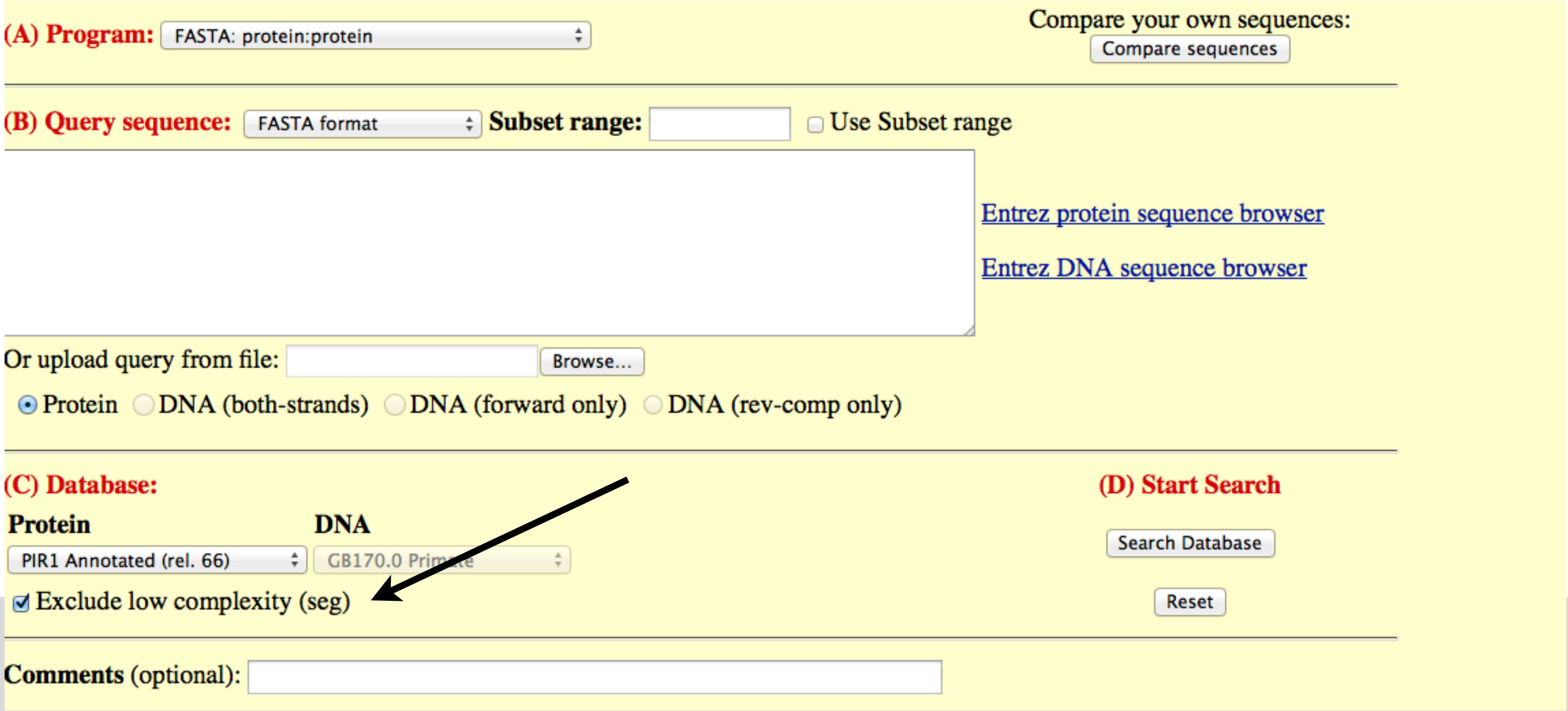
Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: ☑ Low complexity regions

Mask: □ Mask for lookup table only
□ Mask lower case letters



(A) Program: FASTA: protein:protein

Compare your own sequences:

(B) Query sequence: FASTA format **Subset range:** □ Use Subset range

Or upload query from file:

● Protein ○ DNA (both-strands) ○ DNA (forward only) ○ DNA (rev-comp only)

(C) Database:

Protein PIR1 Annotated (rel. 66)

DNA GB170.0 Primate

☑ Exclude low complexity (seg)

(D) Start Search

Comments (optional):

Validating Stats

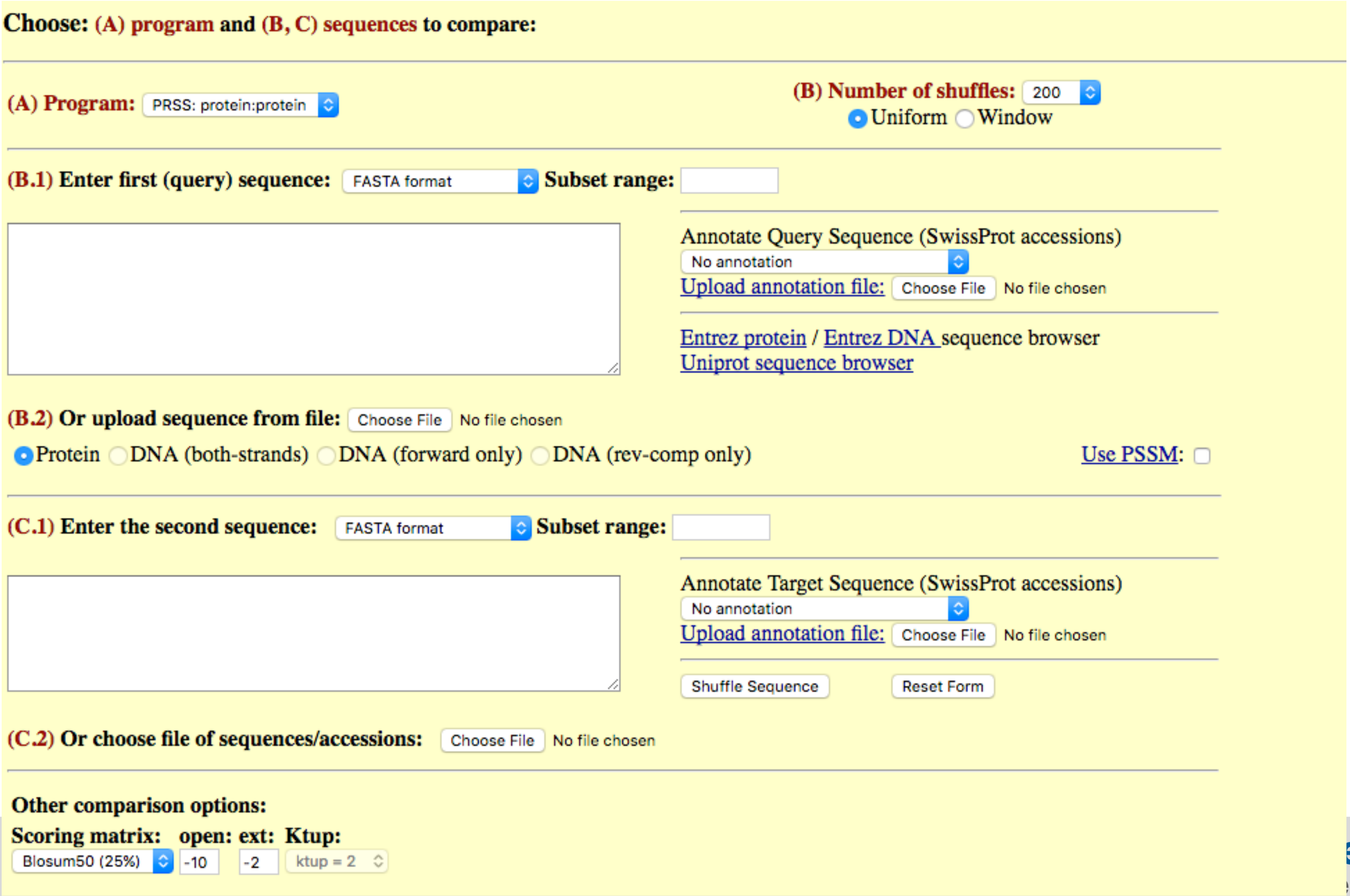
- In general,BLASTP statistical estimates are accurate

  The most common errors occur because of low- complexity regions, or biased amino-acid composition

  To confirm statistical accuracy, find the highest scoring non homolog

  - No need to test every hit, test hits that are surprising
  - Confirm homology/non-homology by searching against a different comprehensive database, e.g. SwissProt, or refseq.
  - Non-homologs will find many significant members of other families, but not the family you are testing for
  - Statistical estimates can be confirmed with shuffles

Validating Stats



Choose: (A) program and (B, C) sequences to compare:

(A) Program: PRSS: protein:protein

(B) Number of shuffles: 200
● Uniform ○ Window

(B.1) Enter first (query) sequence: FASTA format **Subset range:**

Annotate Query Sequence (SwissProt accessions)
No annotation
Upload annotation file: Choose File No file chosen

Entrez protein / Entrez DNA sequence browser
Uniprot sequence browser

(B.2) Or upload sequence from file: Choose File No file chosen

● Protein ○ DNA (both-strands) ○ DNA (forward only) ○ DNA (rev-comp only) Use PSSM: □

(C.1) Enter the second sequence: FASTA format **Subset range:**

Annotate Target Sequence (SwissProt accessions)
No annotation
Upload annotation file: Choose File No file chosen

(C.2) Or choose file of sequences/accessions: Choose File No file chosen

Other comparison options:

Scoring matrix: Blosum50 (25%) **open:** -10 **ext:** -2 **Ktup:** ktup = 2

Alignment Summary

- Compare Protein Sequences for long distances, DNA for close relationships.
- Sequence statistical significance estimates are accurate (verify this yourself)10-6 < E() < 10-3 is statistically significant
- Local sequence alignments find the best region (so that extending the region reduces the score). Global alignments go from end-to-end.
- The Smith-Waterman algorithm produces local alignments with affine gaps in time O(nm) and space O (n).
- BLAST and FASTA try to approximate Smith- Waterman scores for homologous sequences
- Smaller databases increase search sensitivity
- Statistical accuracy can be evaluated by examining the “highest scoring unrelated sequence” or by random shuffles

Workshop
Time

https://bcantarel.github.io/cshl_homology_workshop1