

Sequence Homology and Analysis

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



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

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 Homology? How do we recognize it?
- How do we measure sequence similarity by alignment and scoring matrices?
- DNA vs Protein Comparisons
- How do we know when two sequences share statistically significant similarity?
- How can we verify the statistics?
- Alignment Programs

History of Sequence Similarity

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

Search Algorithms

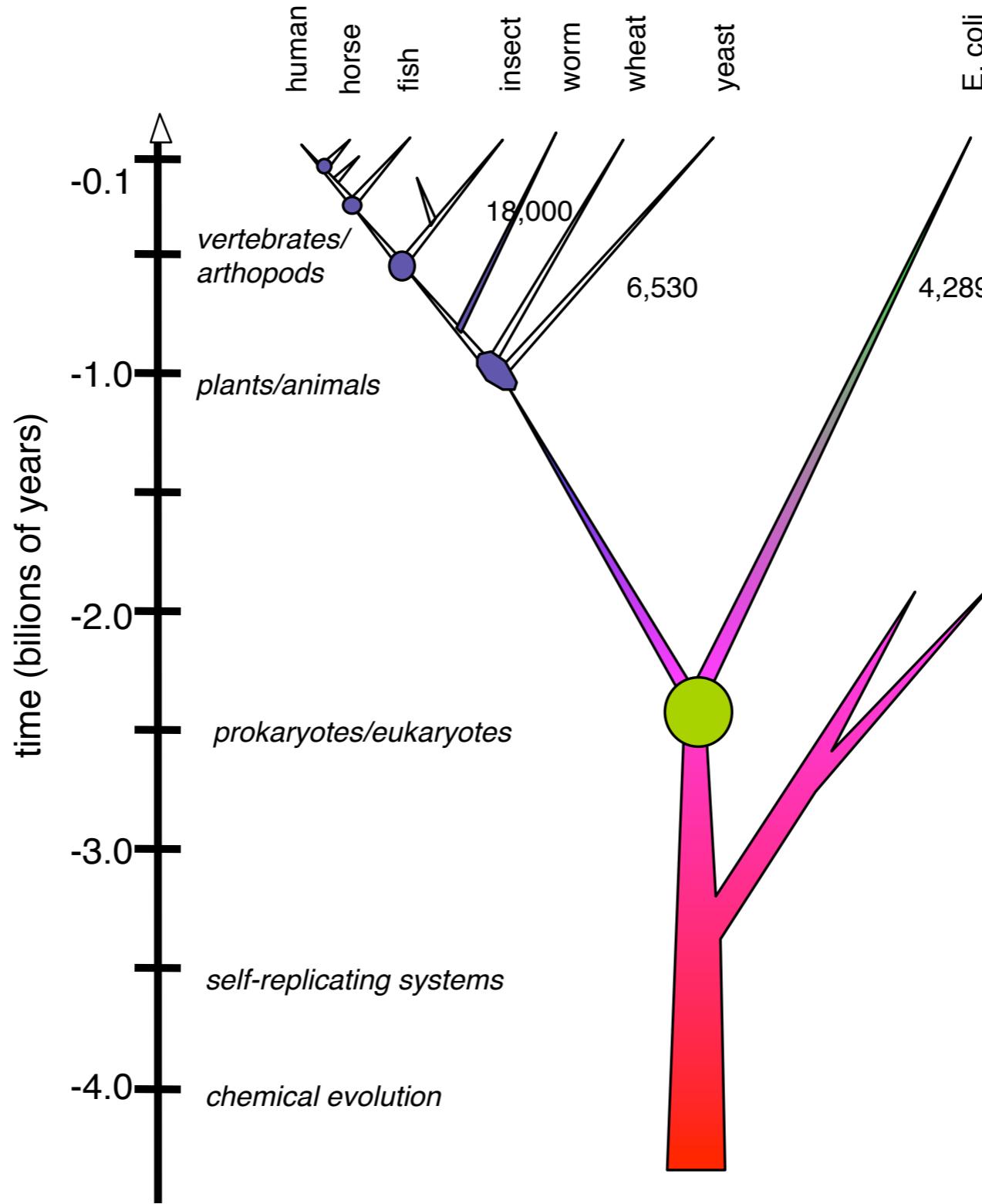
Algorithm	Value Calculated	Scoring Matrix	Gap penalty	Time Requirement	Reference
Needleman-Wunsch	Global similarity	Any	Penalty/Gap	O(n ²)	Needleman and Wunsch, 1970
Sellers	Global distance	Unity	Penalty/Gap	O(n ²)	Sellers, 1974
Smith-Waterman	Local Similarity	$S_{ij} < 0.0$	Affine (q+rk)	O(n ²)	Smith and Waterman, 1981 Gotoh, 1982
SRCHN	Approx. local similarity	diagonal	Penalty/Gap	O(n) – O(n ²)	Wilbur and Lipman, 1983
FASTP/FASTA	Approx. local similarity	$S_{ij} < 0.0$	Limit Size (q+rk)	O(n ²)/K	Lipman and Pearson, 1985, Pearson and Lipman, 1988
BLAST	Maximum Segment Score	$S_{ij} < 0.0$	Multiple Segment	O(n ²)/K	Altschul et al 1990
BLAST2.0	Approx. local similarity	$S_{ij} < 0.0$	(q+rk)	O(n ²)/K	Altschul et al 1997

- What is Homology? How do we recognize it?
- How do we measure sequence similarity by alignment and scoring matrices?
- DNA vs Protein Comparisons
- How do we know when two sequences share statistically significant similarity?
- How can we verify the statistics?
- Alignment Programs

Establishing homology from statistically significant similarity

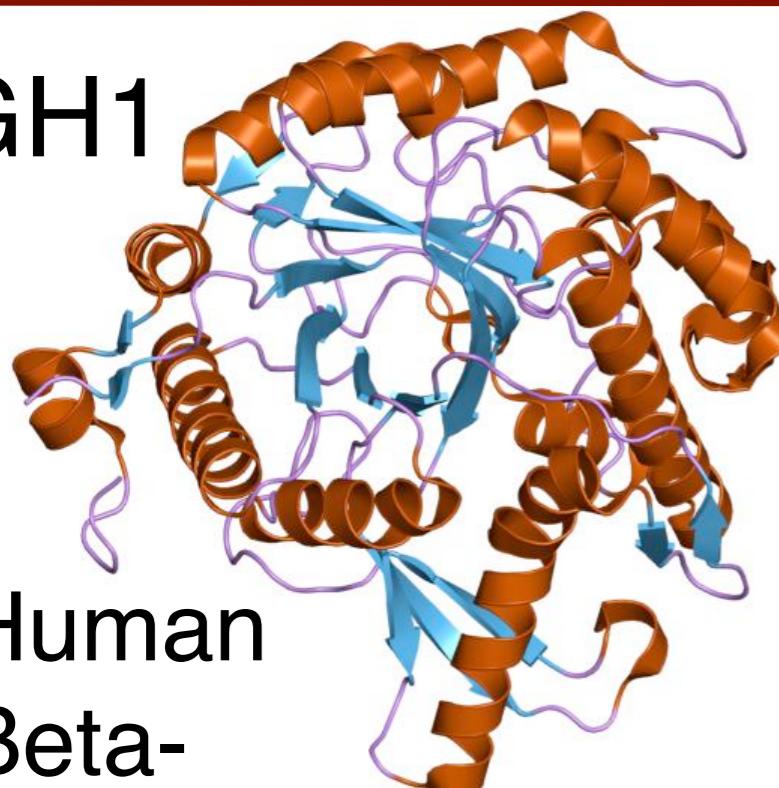
- For most proteins, homologs are easily found over long evolutionary distances (500 My – 2 By) using standard approaches (BLAST, FASTA)
- Difficult for distant relationships or very short domains
- Most default search parameters are optimized for distant relationships and work well

Homologous Sequences Share a Common Ancestor



When do we infer homology?

GH1



Human
Beta-
glucosidase

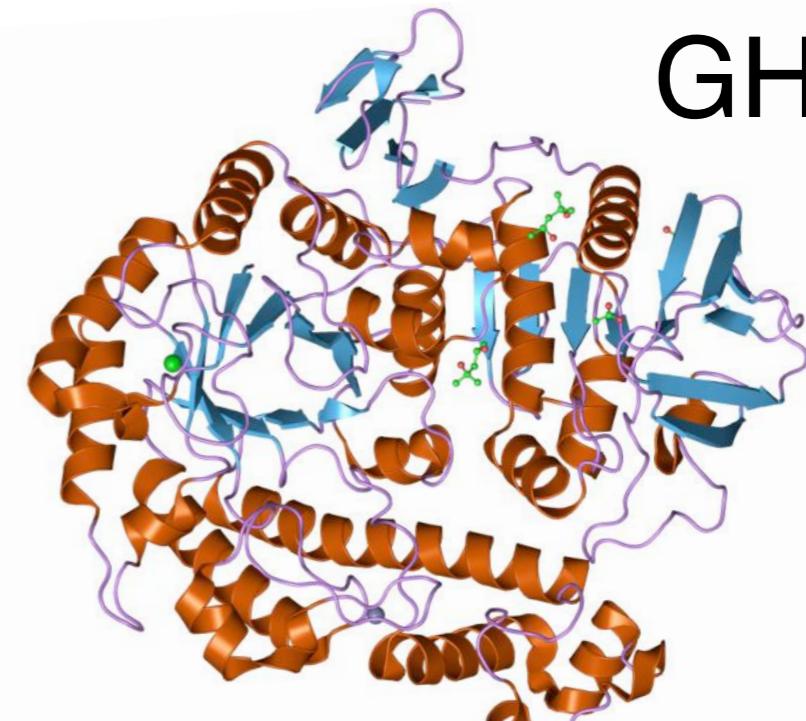
Sequence: Score = 205 bits
Expect = 3e-57
%ID = 30%
Structure: RMSD = 2.63
Score = 1044
P-Value = 0

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

Lactococcus lactis
Beta-galactosidase



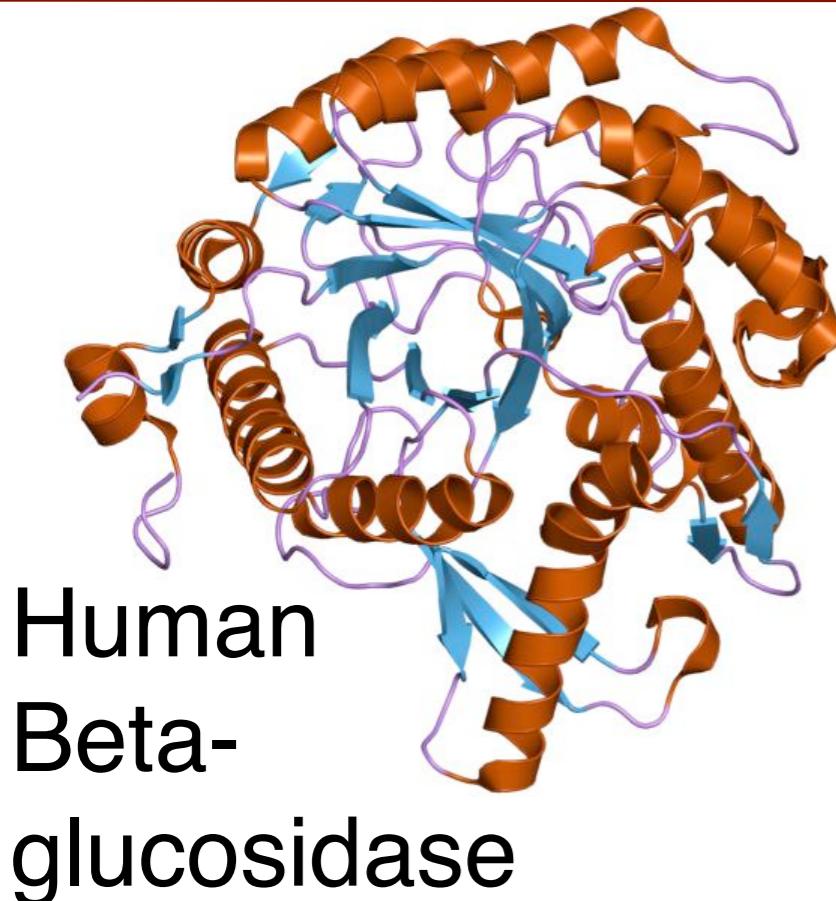
GH42



Human Beta-
galactosidase

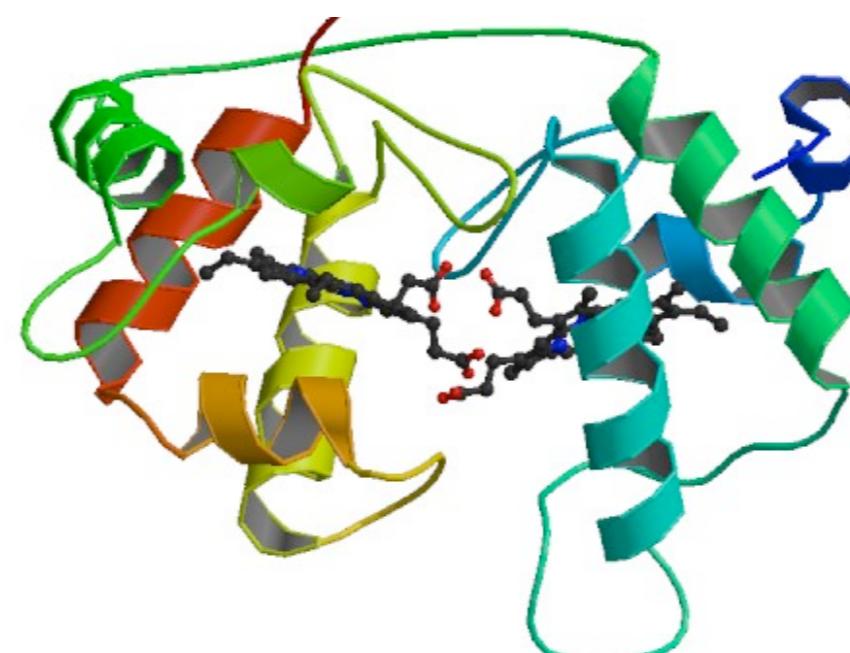
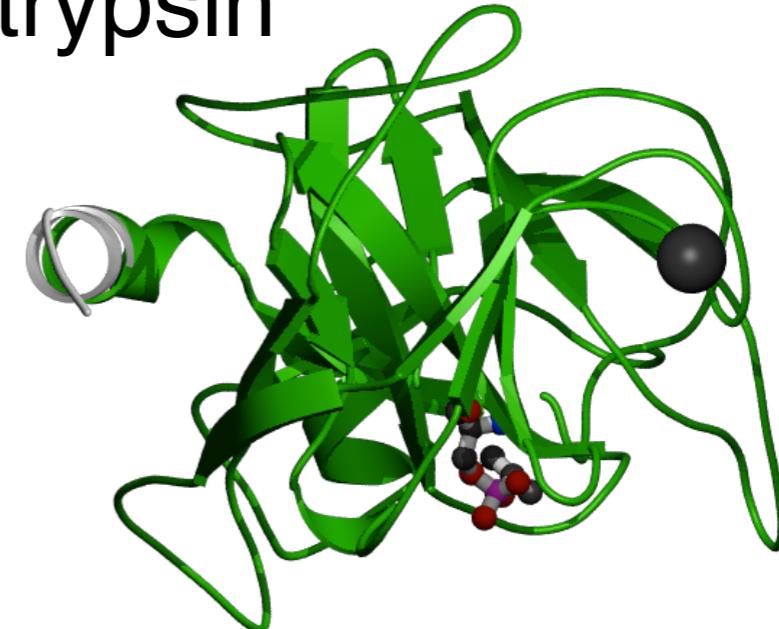
Sequence: Score = 22.3
Expect = 0.051
%ID = 25%
Structure: RMSD = 393.5
Score = 393.5
P-Value = 1.65e-07

When do we infer non-homology?



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%

Bovine trypsin



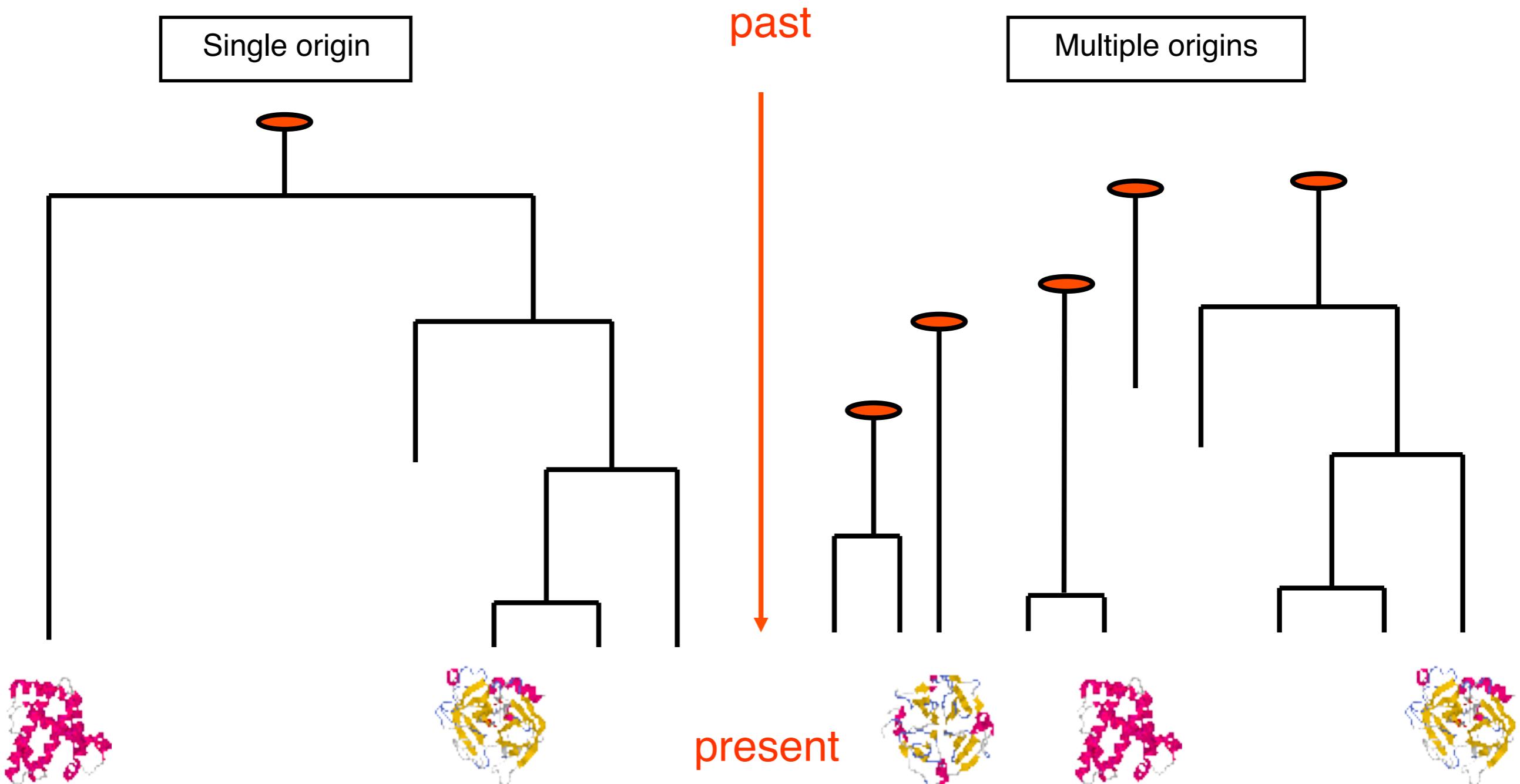
CYTOCHROME
C4

Non-homologous Proteins have
different structures

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 morphological/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)

What BLAST Does

?

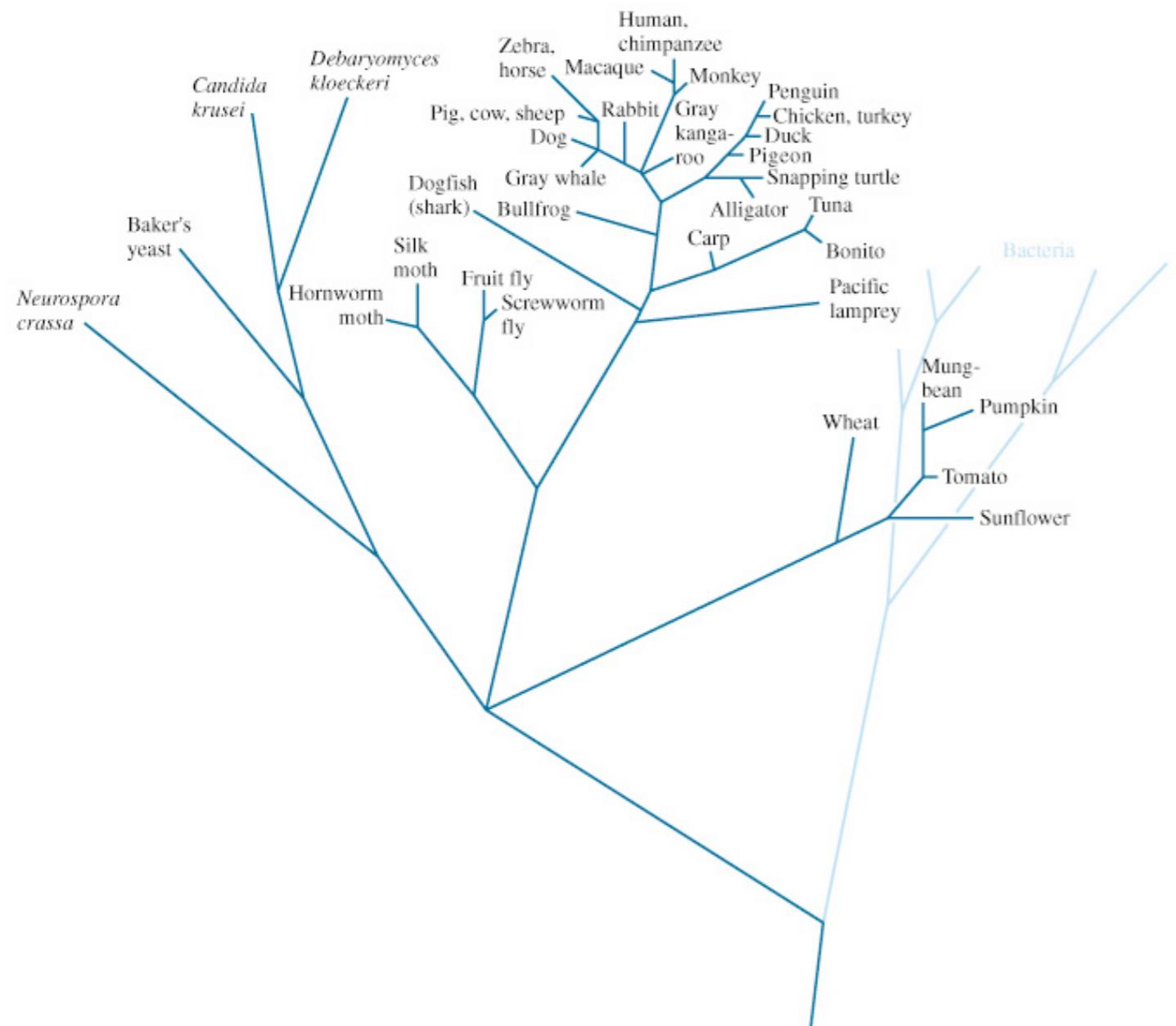
Similarity <=> Homology

?

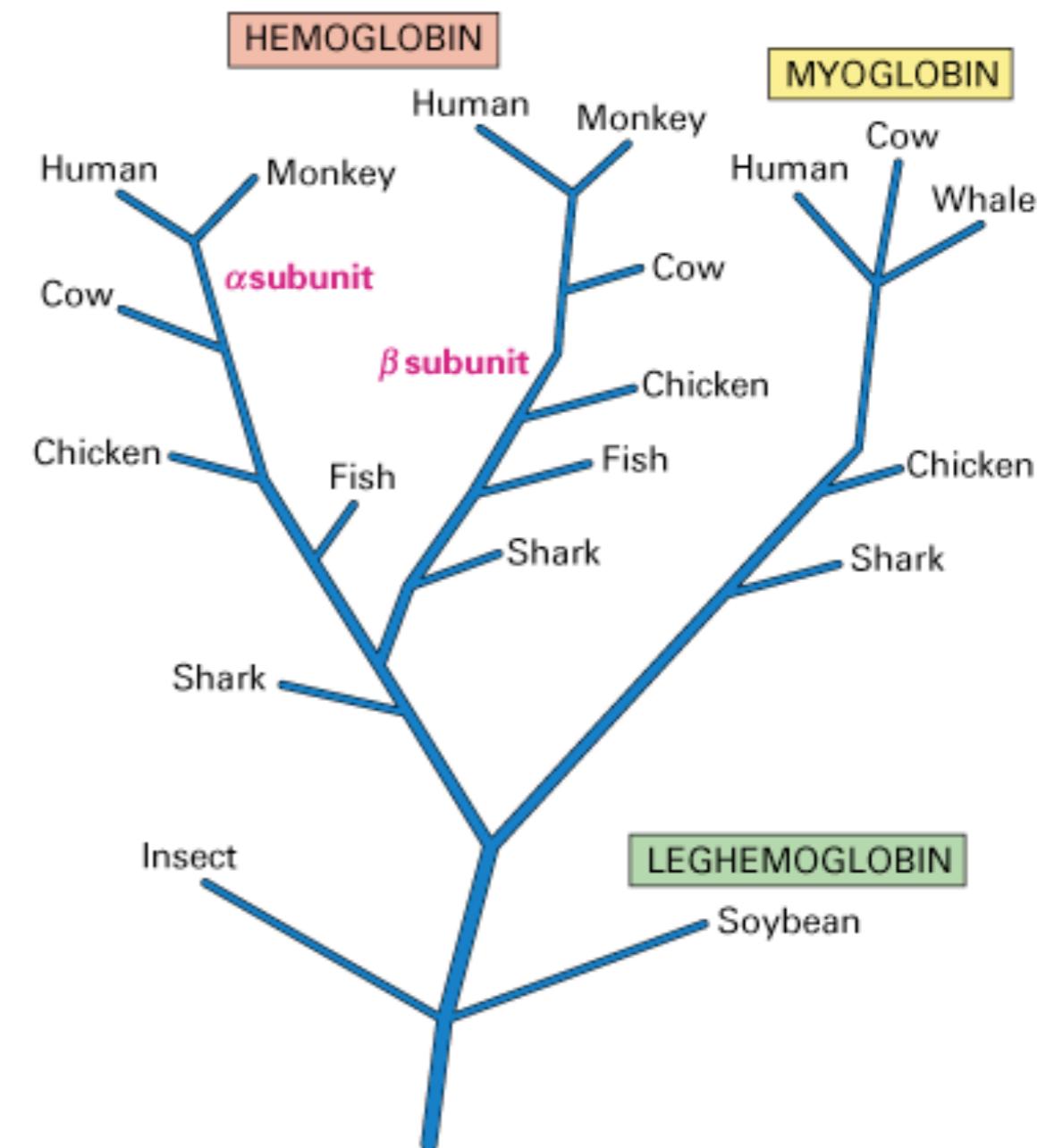
Statistical Significance <=> Biological Significance

Divergence OR Convergence

Orthologs vs Paralogs Inferring Function



Cytochrome C
Family



Globin Family

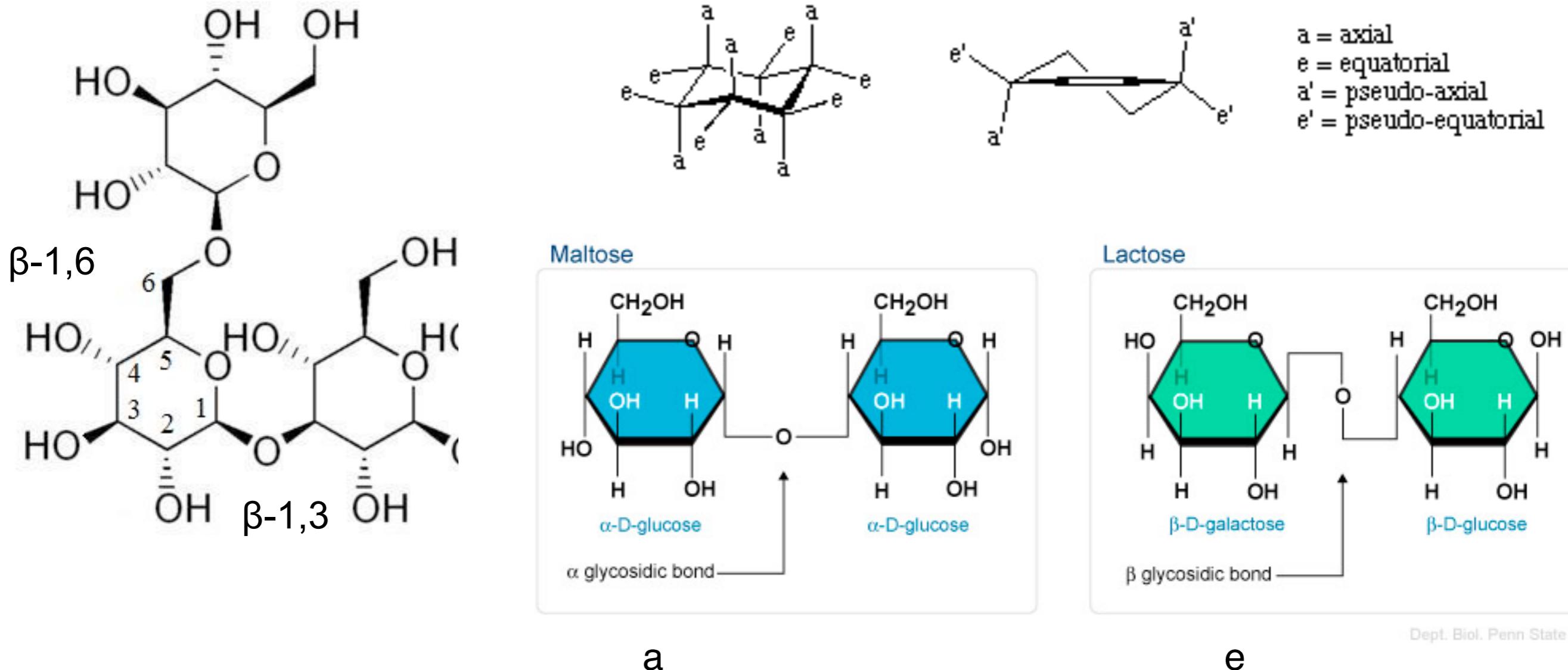
UTSouthwestern

Medical Center

Lyda Hill Department of Bioinformatics

Orthologs vs Paralogs

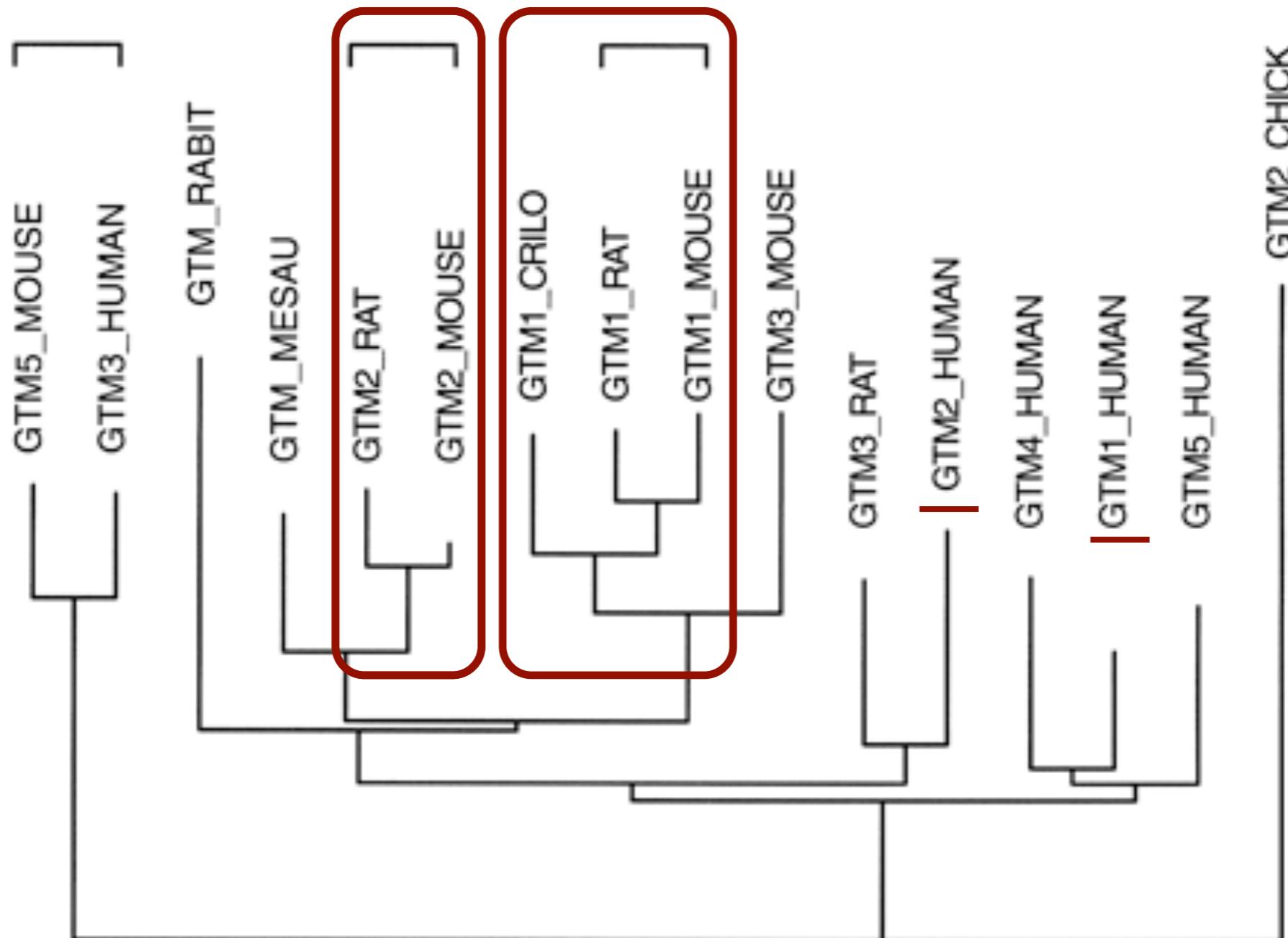
Inferring Function



Dept. Biol. Penn State ©2008

Homologs Often Maintain Similar Chemical Functions

Orthology can be difficult to infer



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

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 functions

Ancient Protein Domain

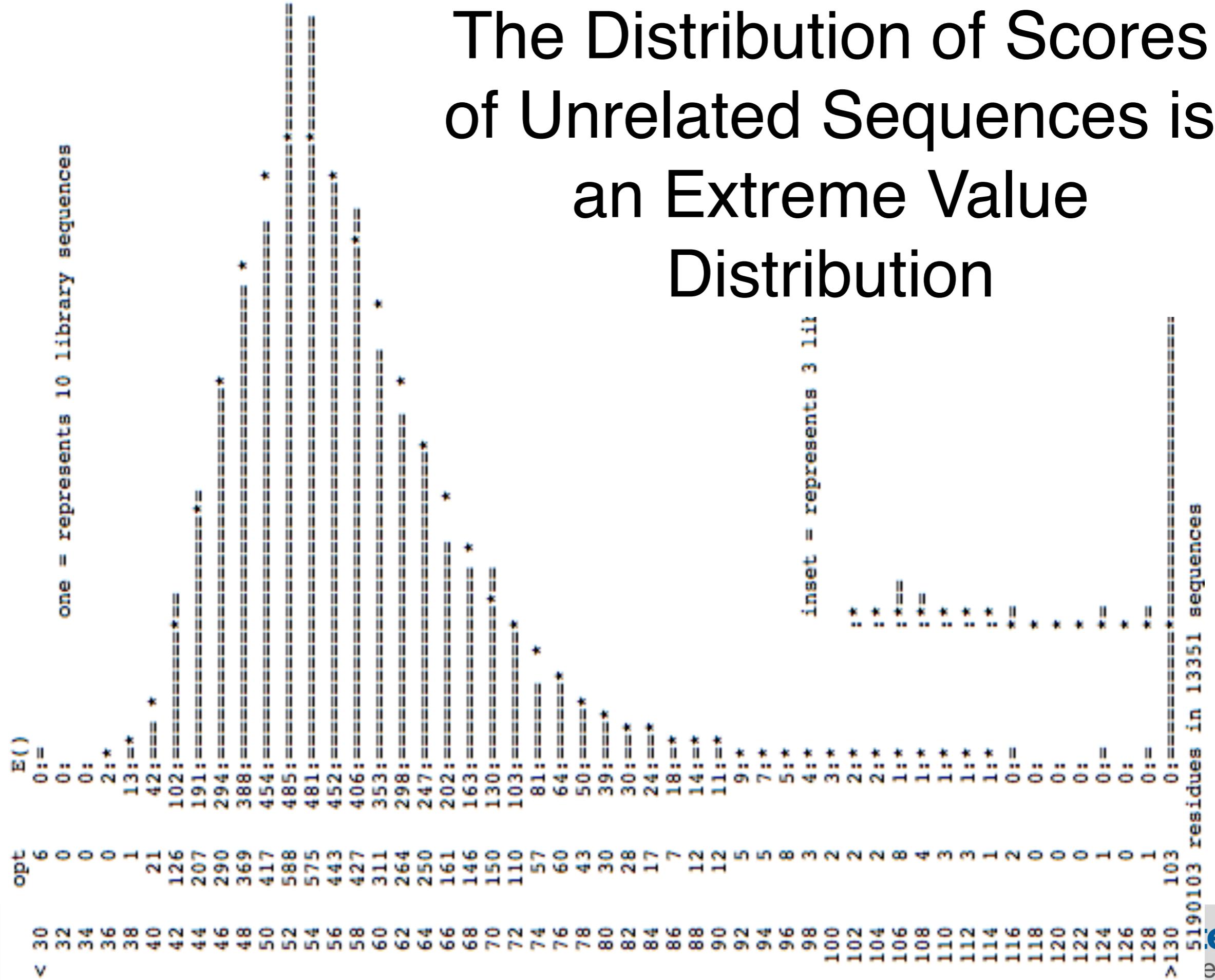
expect	%_id	alen	E coli descr	Human descr	sp_name
2.7e-206	53.8	944	glycine decarboxylase, P	Glycine dehydrogenase [de	GCSP_HUMAN
1.2e-176	59.5	706	methylmalonyl-CoA mutase	Methylmalonyl-CoA mutase,	MUTA_HUMAN
3.8e-176	50.6	803	glycogen phosphorylase [E	Glycogen phosphorylase, l	PHS1_HUMAN
9.9e-173	55.6	1222	B12-dependent homocystein	5-methyltetrahydrofolate-	METH_HUMAN
1.8e-165	41.8	1031	carbamoyl-phosphate synth	Carbamoyl-phosphate synth	CPSM_HUMAN
5.6e-159	65.7	542	glucosephosphate isomeras	Glucose-6-phosphate isome	G6PI_HUMAN
8.1e-143	53.7	855	aconitate hydrase 1 [Esch	Iron-responsive element b	IRE1_HUMAN
2.5e-134	73.0	459	membrane-bound ATP syntha	ATP synthase beta chain,	ATPB_HUMAN
3.3e-121	55.8	550	succinate dehydrogenase,	Succinate dehydrogenase [DHSA_HUMAN
1.5e-113	60.6	401	putative aminotransferase	Cysteine desulfurase, mit	NFS1_HUMAN
4.4e-111	60.9	460	fumarase C= fumarate hydr	Fumarate hydratase, mitoc	FUMH_HUMAN
1.5e-109	56.1	474	succinate-semialdehyde de	Succinate semialdehyde de	SSDH_HUMAN
3.6e-106	44.7	789	maltodextrin phosphorylas	Glycogen phosphorylase, m	PHS2_HUMAN
1.4e-102	53.1	484	NAD+-dependent betaine al	Aldehyde dehydrogenase, E	DHAG_HUMAN
3.8e-98	53.0	449	pyridine nucleotide trans	NAD(P) transhydrogenase,	NNTM_HUMAN
5.8e-96	49.9	489	glycerol kinase [Escheric	Glycerol kinase, testis s	GKP2_HUMAN
2.1e-95	66.8	328	glyceraldehyde-3-phosphat	Glyceraldehyde 3-phosphat	G3P2_HUMAN
5.0e-91	62.5	368	alcohol dehydrogenase cla	Alcohol dehydrogenase cla	ADHX_HUMAN
6.7e-91	56.5	393	protein chain elongation	Elongation factor Tu, mit	EFTU_HUMAN
9.5e-91	56.6	392	protein chain elongation	Elongation factor Tu, mit	EFTU_HUMAN
2.2e-89	59.1	369	methionine adenosyltransf	S-adenosylmethionine synt	METK_HUMAN
6.5e-88	53.3	422	enolase [Escherichia coli	Alpha enolase (2-phospho-	ENOA_HUMAN
9.2e-88	43.3	536	NAD-linked malate dehydro	NADP-dependent malic enzy	MAOX_HUMAN
7.3e-86	55.5	389	2-amino-3-ketobutyrate Co	2-amino-3-ketobutyrate co	KBL_HUMAN
5.2e-83	44.4	543	degrades sigma32, integra	AFG3-like protein 2 (Para	AF32_HUMAN

- What is Homology? How do we recognize it?
- How do we measure sequence similarity by alignment and scoring matrices?
- DNA vs Protein Comparisons
- How do we know when two sequences share statistically significant similarity?
- How can we verify the statistics?
- Alignment Programs

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



Significance Thresholds

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_SYN3	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_SYN6	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	
sp P07864 LDHC_HUMAN	L-lactate dehydrogenase C chain (L	(332)	80	28.4	3.3	0.478	0.783	23	align	
sp Q02655 CYB PODAN	Cytochrome b	(387)	80	28.4	3.9	0.241	0.517	116	align	
sp P00162 CYB_NEUCR	Cytochrome b	(385)	79	28.2	4.6	0.244	0.543	234	align	
sp P06320 TVB7_MOUSE	T-cell receptor beta chain V region	(134)	73	26.6	4.7	0.360	0.540	50	align	
sp P09261 CELF_VZVD	Cell fusion protein precursor	(340)	78	27.9	4.8	0.204	0.512	162	align	
sp P05841 VNS3_BMDNV	Putative nonstructural protein (OR	(167)	74	26.9	4.9	0.464	0.750	28	align	
ref YP_220570.1 NADH	dehydrogenase subunit 3 [Mus musc	(115)	71	26.1	5.8	0.290	0.508	124	align	
sp P02980 TCR2_ECOLI	Tetracycline resistance protein, c	(401)	77	27.7	6.8	0.325	0.650	40	align	
sp P03921 NU5M_MOUSE	NADH-ubiquinone oxidoreductase cha	(607)	79	28.2	7.1	0.321	0.605	81	align	
sp P00160 CYB_XENLA	Cytochrome b	(380)	76	27.4	7.7	0.248	0.545	165	align	
sp P45014 NRFD_HAEIN	Protein nrfD homolog	(321)	75	27.2	7.8	0.221	0.579	145	align	
sp Q00098 KR16_ICHV1	Gene 16 protein kinase	(387)	76	27.4	7.8	0.250	0.484	128	align	

Significance Thresholds

>>sp|P68526|ATP6_TRITI ATP synthase a chain (ATPase protein 6) g (386 aa)
initn: 108 initl: 70 opt: 121 Z-score: 165.5 bits: 38.9 E(13351): 0.0028
Smith-Waterman score: 209; 25.9% identity (60.3% similar) in 239 aa overlap (39-271:150-362)
[Entrez](#) [Lookup](#) [Re-search database](#) [General re-search](#)

	10	20	30	40	50	60	70	
gi 811	MASENMTPQDYIGHHLNNLQLDLRTFSLVDPQNPPATFWTINIDSMFFSVVLGLLFLVLF	RSVAKKATSGVPGKFQTA						
sp P68	AGGGMDNFIQNLPGAYPETPLDQFAIIPIIDLHVGNFYLSFTNEVLYMLLtvvl-vvfl-ffvvtkkgggk	SVPNAWQSL						
	110	120	130	140	150	160	170	180
gi 811	IELVIGFVNNGSVKDMYHG-----KSKLIAPLALTIFVVWVFLMNLMDLLPIDLLPYIAEHVLGLP	ALRVVPSADVNTLSM						
	80	90	100	110	120	130	140	150
sp P68	VELIYDFVNLNVNEQIGGLSGNVKQKFFPRISVT-FTFSLFRNPQGMIPFSFT--VTSHFL-----	ITLAL						
	190	200	210	220	230	240		250
gi 811	ALGVFILILFYSIKMKGIGGFTKELTLQPFNHWAFIPVNLLLEGVSLLSKPVSLGLRFGNMYAGELIF	ILIAGLLPWWS						
	160	170	180	190	200	210	220	230
sp P68	SFSIFIGITIVGFQRHGLHFFS--fllpagvplplapflvlelisYCFRALSLGIRLFANMMAGHS	LVKILSGFA--WT						
	260	270	280	290	300	310	320	
gi 811	QWILNVPWAIFHILLIITLQAFIFMVLTIVYLSMA-SEEH							
	240	250	260	270				
sp P68	MLFLN---NIFYFIGDLGPLFIVLALTGLELGVAISQAHVSTISICIYLNDATNLHQNESFHN							
	330	340	350	360	370	380		

PAM Matrices



- The PAM matrices were introduced by Margret Dayhoff in 1978
- 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.

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
- q_{ij} is Replacement frequency of i to j
- p_i 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

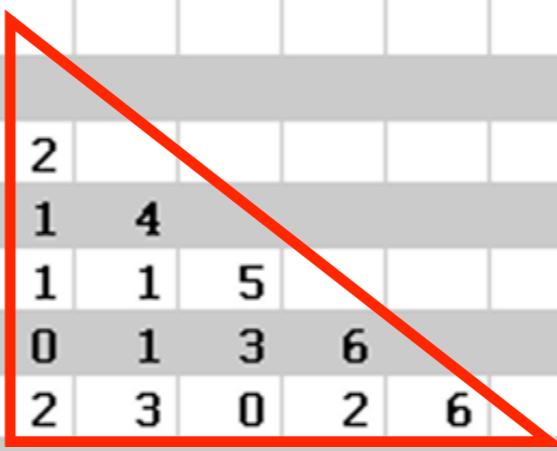
PAM40: Predicts 40 mutations per 100 aa

PAM250: Predicts 250 mutations per 100 aa

PAM Matrices

	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*			
G	5																										G
A	1	2																									A
V	-1	0	4																								V
L	-4	-2	2	6																							L
I	-3	-1	4	2	5																						I
P	0	1	-1	-3	-2	6																					P
S	1	1	-1	-3	-1	1	2																			S	
T	0	1	0	-2	0	0	1	3																		T	
D	1	0	-2	-4	-2	-1	0	0	4																	D	
E	0	0	-2	-3	-2	-1	0	0	3	4																E	
N	0	0	-2	-3	-2	0	1	0	2	1	2															N	
Q	-1	0	-2	-2	-2	0	-1	-1	2	2	1	4														Q	
K	-2	-1	-2	-3	-2	-1	0	0	0	0	1	1	5													K	
R	-3	-2	-2	-3	-2	0	0	-1	-1	-1	0	1	3	6												R	
H	-2	-1	-2	-2	-2	0	-1	-1	1	1	2	3	0	2	6											H	
F	-5	-3	-1	2	1	-5	-3	-3	-6	-5	-3	-5	-5	-4	-2	9										F	
Y	-5	-3	-2	-1	-1	-5	-3	-3	-4	-4	-2	-4	-4	-4	-4	0	7	10								Y	
W	-7	-6	-6	-2	-5	-6	-2	-5	-7	-7	-4	-5	-3	-2	-3	0	0	17								W	
M	-3	-1	2	4	2	-2	-2	-1	-3	-2	-2	-1	0	0	-2	0	-2	-4	6							M	
C	-3	-2	-2	-6	-2	-3	0	-2	-5	-5	-4	-5	-5	-4	-3	-4	0	-8	-5	12						C	
B	0	0	-2	-3	-2	-1	0	0	3	3	2	1	1	-1	1	-4	-3	-5	-2	-4	3				B		
Z	0	0	-2	-3	-2	0	0	-1	3	3	1	3	0	0	2	-5	-4	-6	-2	-5	2	3		Z			
X	-1	0	-1	-1	-1	-1	0	0	-1	-1	0	-1	-1	-1	-2	-2	-4	-1	-3	-1	-1	-1	-1		X		
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	*		
	G	A	V	L	I	P	S	T	D	E	N	Q	K	R	H	F	Y	W	M	C	B	Z	X	*			

PAM 250



Scoring Alignments

```
>>sp|P07925|ATP6_MAIZE ATP synthase a chain (ATPase protein 6) (291 aa)
initn: 96 initl: 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      MKIVLYYFVNMFISGIFQIANVEVGQHFYWSILGFQIHGQVLINSWIVILIIGFLSIYTTKNL-
          :: : ..... : .. .... .:...: .. : .. : .. : .. : .. :
sp|P07 MERNGEivnnngsiiipggggpvTESPLDQFGIHPILDLNIGK-YYVSFTnls1--smllt1glvlllv-f--vvtkkggg
          10      20      30      40      50      60      70
          70      80      90     100     110     120     130     140
gi|231 TLVPANKQIFIELVTEFITDISKTQIGEKEYS---KWWPYIGTMFLFIFVSNWSGALIPWKIIELPNGELGAPTNDINTT
          :: : ..... : .. .... .: .. : .. : .. : .. : .. : .. : .. : .. :
sp|P07 ksvPNAFQSLVELIYDFVPNLVNEQIGGLSGNVKHKKFFPCISVTFTFSLFRNPQG-MIPFSF-----TVTSHFLIT
          80      90     100     110     120     130
          150     160     170     180     190     200     210
gi|231 AGLAILTSLAYFYAGLNKKGLTYFKKYVQPTPILPIN---ILEDFT---KPLSLSFRLFGNILADELVVAVLVSLVPL
          :... : .. : .. : .. : .. : .. : .. : .. : .. : .. : .. : .. :
sp|P07 LALSFSIFIGITIVGFQRHGLHFFS-f11pagvp1plapflvle1ISHCFRALSSGIRLFANMMAGHSSVKILSGFAWT
          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

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

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

Highest Scoring Unrelated Protein

The best scores are:

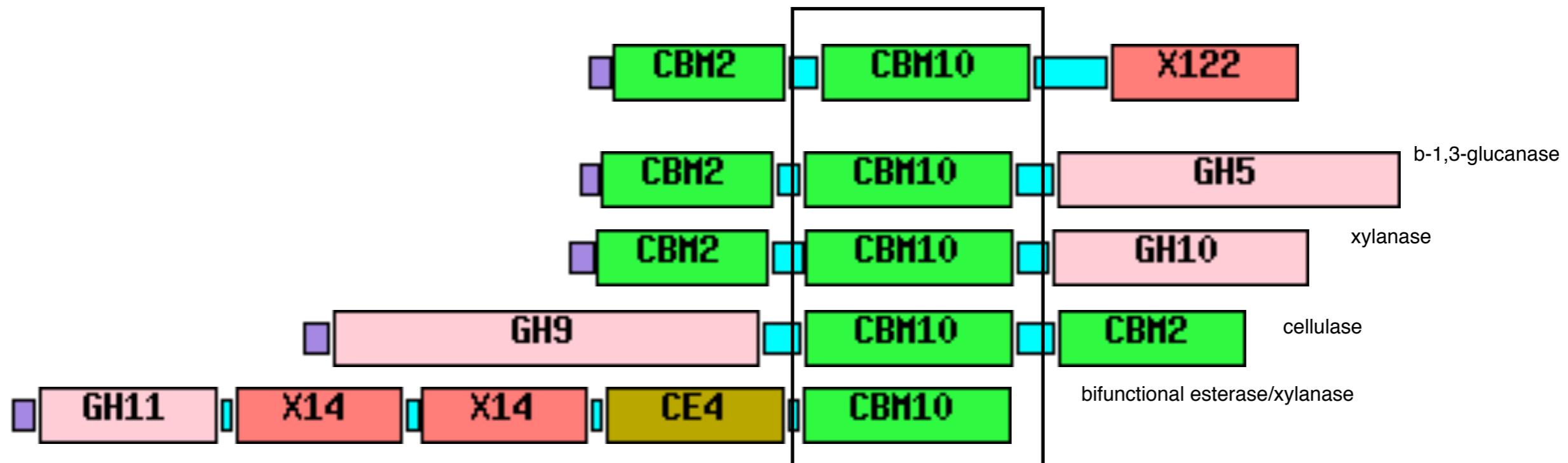
				opt bits	E(13351)	%_id	%_sim	alen	
sp P00846 ATP6_HUMAN	ATP synthase a chain (ATPase prote	(226)	1124	289.8	4.1e-79	1.000	1.000	226	align
sp P00847 ATP6_BOVIN	ATP synthase a chain (ATPase prote	(226)	1075	277.5	2e-75	0.779	0.951	226	align
sp P00848 ATP6_MOUSE	ATP synthase a chain (ATPase prote	(226)	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_SYNYY3	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_SYNYY3	DNA ligase (Polydeoxyribonucleotid	(669)	81	28.6	5.1	0.250	0.570	128	align

Unrelated or Too Distance

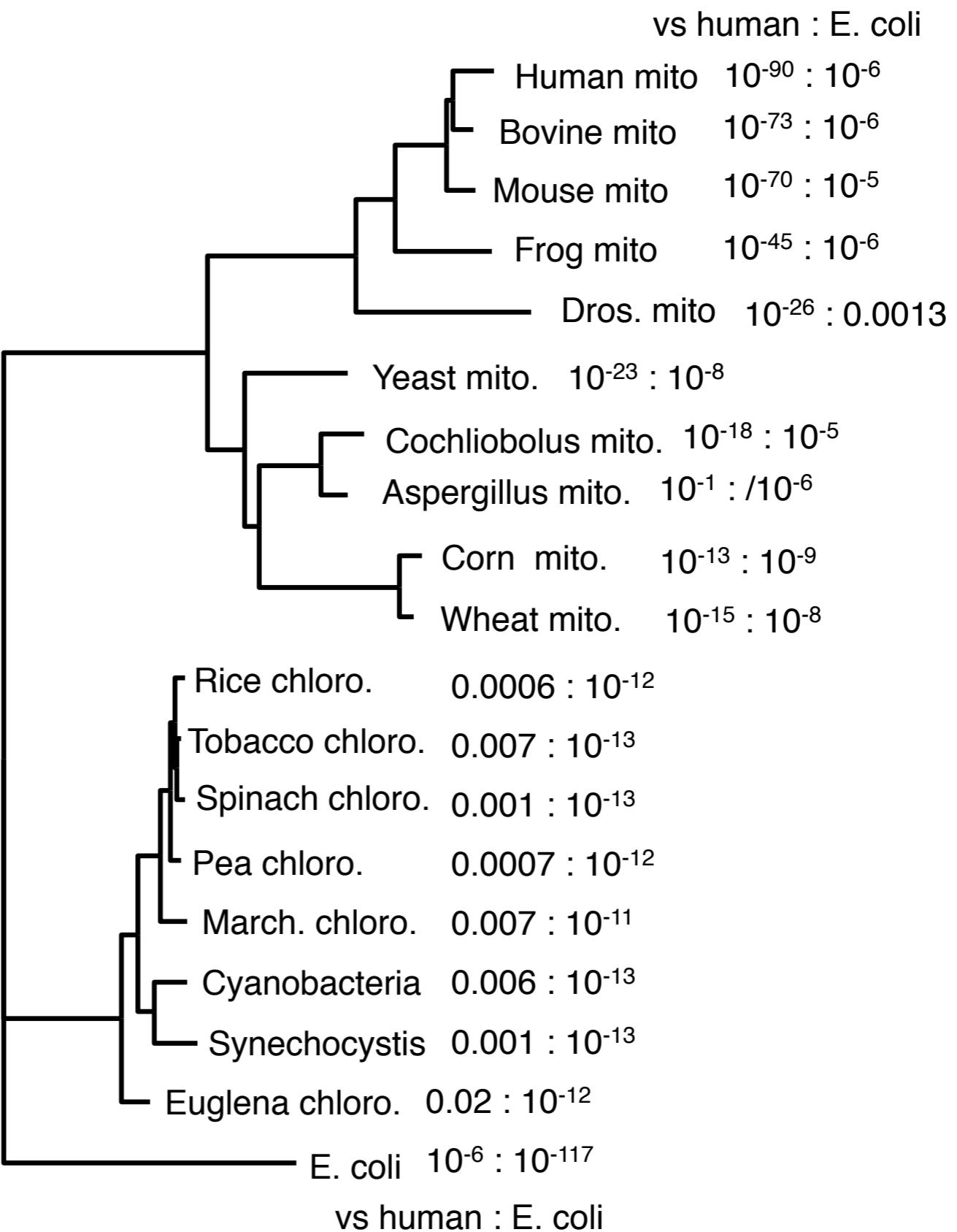
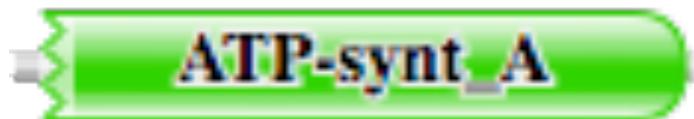
The best scores are:

				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_SYN3	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	

Homology in Domains



Homology is Transitive (in Protein Domains)



Homology in Domains

Xylanase

The best scores are:

			opt bits	E(445410)	%_id	%_sim	alen	
sp P45796.1 XYND_PAEP0	Arabinoxylan arabinofuranohydrol	(635)	1813	412.5	2.6e-113	0.537	0.817	486 align
sp Q45071.2 XYND_BACSU	Arabinoxylan arabinofuranohydrol	(513)	1509	345.0	4.2e-93	0.554	0.812	495 align
sp Q9WXE8.2 XYLO_PRERU	Putative beta-xylosidase; 1,4-be	(518)	563	135.0	7.2e-30	0.384	0.645	276
+-			241	63.5	2.4e-08	0.327	0.633	150 align
sp P77713.1 YAGH_ECOLI	Putative beta-xylosidase; 1,4-be	(536)	334	84.1	1.5e-14	0.305	0.561	321 align
sp P94489.2 XYNB_BACSU	Beta-xylosidase; 1,4-beta-D-xyla	(533)	318	80.6	1.8e-13	0.285	0.555	362 align
sp P07129.2 XYNB_BACPU	Beta-xylosidase; 1,4-beta-D-xyla	(535)	316	80.1	2.4e-13	0.295	0.553	356 align
sp P45982.1 XYLB_BUTFI	Xylosidase/arabinosidase; Includ	(517)	312	79.3	4.3e-13	0.301	0.578	396 align
sp P48791.1 XYNB_PRERU	Beta-xylosidase; 1,4-beta-D-xyla	(319)	228	60.7	1e-07	0.281	0.548	345 align
sp P36917.1 XYNA_THESA	Endo-1,4-beta-xylanase A; Xylan	(1157)	205	55.4	1.5e-05	0.317	0.662	139
+-			198	53.8	4.4e-05	0.261	0.688	138 align
sp P33558.2 XYNA2_CLOSR	Endo-1,4-beta-xylanase A; Xyla	(512)	190	52.2	6.1e-05	0.249	0.558	249 align
sp P38535.1 XYNX_CLOTM	Exoglucanase xynX; 1,4-beta-cell	(1087)	194	52.9	7.6e-05	0.223	0.607	229 align
sp Q8GJ44.2 XYNA1_CLOSR	Endo-1,4-beta-xylanase A; 1,4-b	(651)	190	52.1	7.9e-05	0.322	0.653	118 align
sp P10478.3 XYNZ_CLOTH	Endo-1,4-beta-xylanase Z; Xylan	(837)	187	51.4	0.00017	0.362	0.691	94 align
sp P94522.3 ABNA_BACSU	Arabinan endo-1,5-alpha-L-arabin	(323)	169	47.6	0.00092	0.261	0.540	287 align
sp P48790.1 XYLA_CLOSR	Xylosidase/arabinosidase; Includ	(473)	164	46.4	0.003	0.268	0.523	497 align
sp Q5AZC8.1 ABNB_EMENI	Arabinan endo-1,5-alpha-L-arabin	(400)	153	44.0	0.014	0.290	0.512	252 align

- What is Homology? How do we recognize it?
- How do we measure sequence similarity by alignment and scoring matrices?
- **DNA vs Protein Comparisons**
- How do we know when two sequences share statistically significant similarity?
- How can we verify the statistics?
- Alignment Programs

DNA vs Protein Comparison



A comprehensive non-coding RNA sequence database ver. 3.4

fRNAdb is [Web Service \(SOAP, REST\)](#) [Ready](#).

Total: 510,055 entries



Catalog



Blast



Download



Help

ncRNAs

Population Genetics



Medical Center

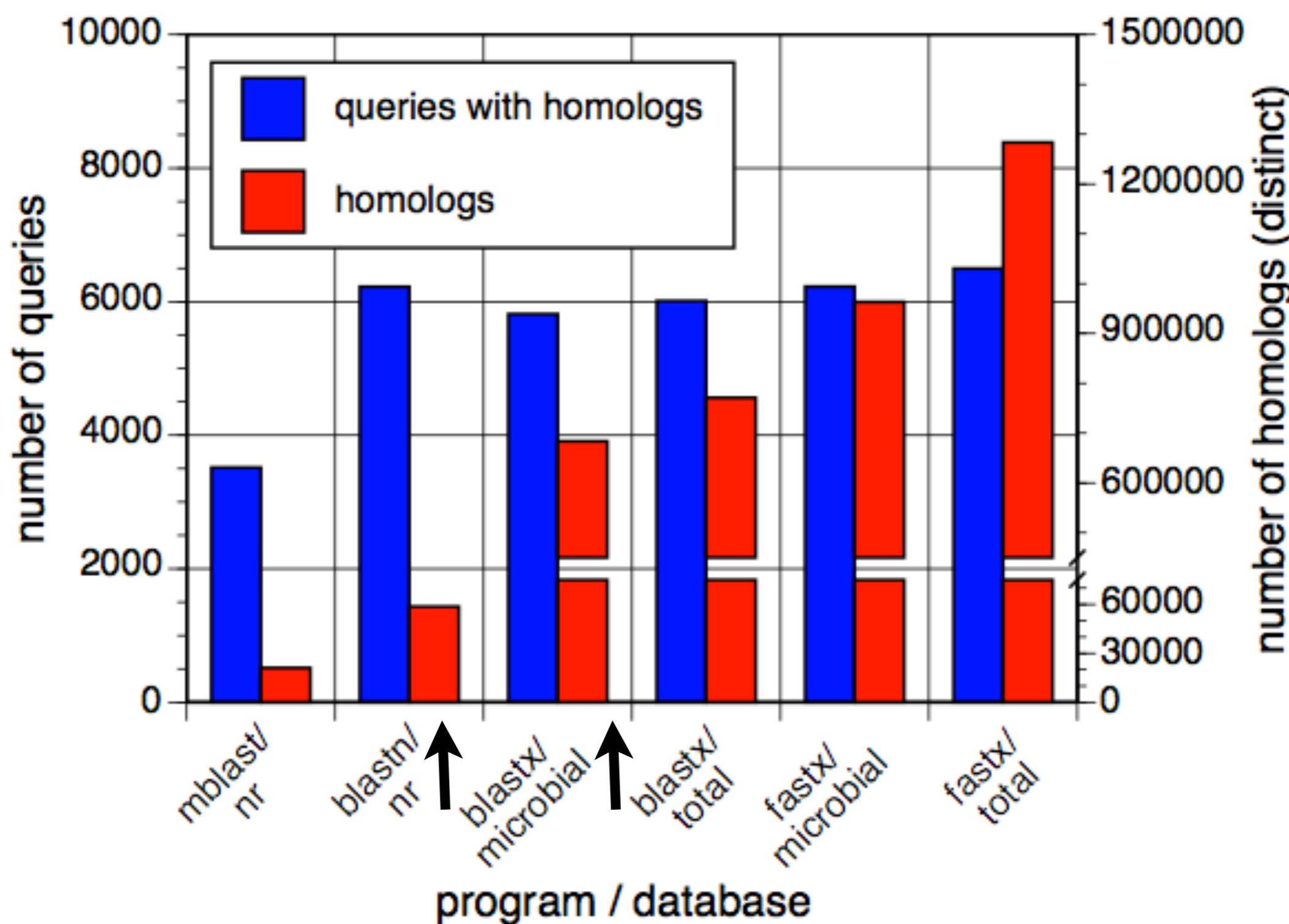
Lyda Hill Department of Bioinformatics

DNA vs Protein Comparison

The best scores are:

		DNA E(188,018)	tfastx3 E(187,524)	prot. 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
MOTGLUSTR	Manduca sexta GST	5.7e-07	2.5e-30	8.0e-25
RLGSTARGN	R.leguminosarum <i>gstA</i>	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 is Homology? How do we recognize it?
- How do we measure sequence similarity by alignment and scoring matrices?
- DNA vs Protein Comparisons
- How do we know when two sequences share statistically significant similarity?
- How can we verify the statistics?
- Alignment Programs

1. What question to ask?
2. What program to use?
3. What database to search?
4. How to avoid mistakes
 1. (what to look out for)
5. When to do something different
 1. (changing scoring matrices)

What question to ask?

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?

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®

Home Rec

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 **GO**

<input type="checkbox"/> Human	<input type="checkbox"/> Rabbit	<input type="checkbox"/> Zebrafish
<input type="checkbox"/> Mouse	<input type="checkbox"/> Chimp	<input type="checkbox"/> Clawed frog
<input type="checkbox"/> Rat	<input type="checkbox"/> Guinea pig	<input type="checkbox"/> Arabidopsis
<input type="checkbox"/> Cow	<input type="checkbox"/> Fruit fly	<input type="checkbox"/> Rice
<input type="checkbox"/> Pig	<input type="checkbox"/> Honey bee	<input type="checkbox"/> Yeast
<input type="checkbox"/> Dog	<input type="checkbox"/> Chicken	<input type="checkbox"/> Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontiguous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> 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

Sequence Alignment Via the Web

BLAST® » blastp suite

Home Recent Results Saved Searches

Standard Protein BLAST

[blastn](#) [blastp](#) [blastx](#) [tblastn](#) [tblastx](#)

[Enter Query Sequence](#)

BLASTP programs search protein databases using a protein query. [more...](#) [Reset page](#)

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 [?](#)

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/ss/>

Sequence Alignment Via the Web

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected

 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/ss/>

or Upload a file: No file chosen

STEP 3 - Set your parameters

PROGRAM

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**)

fasta.bioch.virginia.edu

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**)

10 Minute Break

Take
A BREAK
—(it's—
good for
YOU)

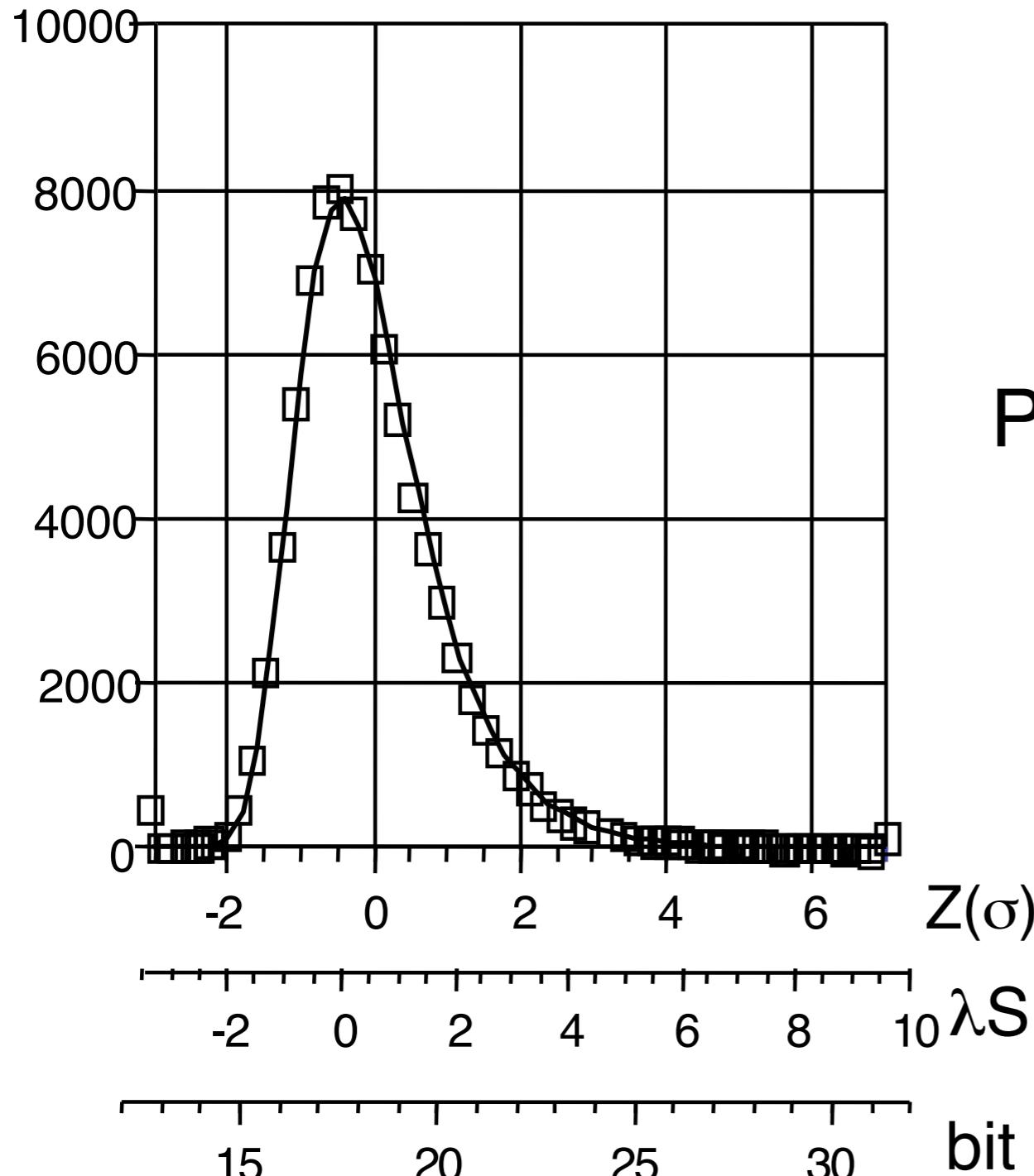
What Database to Search?

- Search the smallest comprehensive database likely to contain your protein
 - 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



$$S' = \lambda S_{\text{raw}} - \ln K m n$$

$$S_{\text{bit}} = (\lambda S_{\text{raw}} - \ln K) / \ln(2)$$

$$P(S' > x) = 1 - \exp(-e^{-x})$$

$$P(S_{\text{bit}} > x) = 1 - \exp(-mn2^{-x})$$

$$E(S' > x | D) = P D$$

$$P(B \text{ bits}) = m n 2^{-B}$$

$$P(40 \text{ bits}) = 1.5 \times 10^{-7}$$

$$E(40 | D=4000) = 6 \times 10^{-4}$$

$$E(40 | D=12E6) = 1.8$$

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	445410165796297		0.0015	10
RefSeq	711441261324908		NS	16

How Can I Choose my DB?

BLAST® » blastp suite

Home Recent Results Saved Searches

Standard Protein BLAST

[blastn](#) [blastp](#) [blastx](#) [tblastn](#) [tblastx](#)

[Enter Query Sequence](#) [Reset page](#)

BLASTP programs search protein databases using a protein query. [more...](#)

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 [?](#)

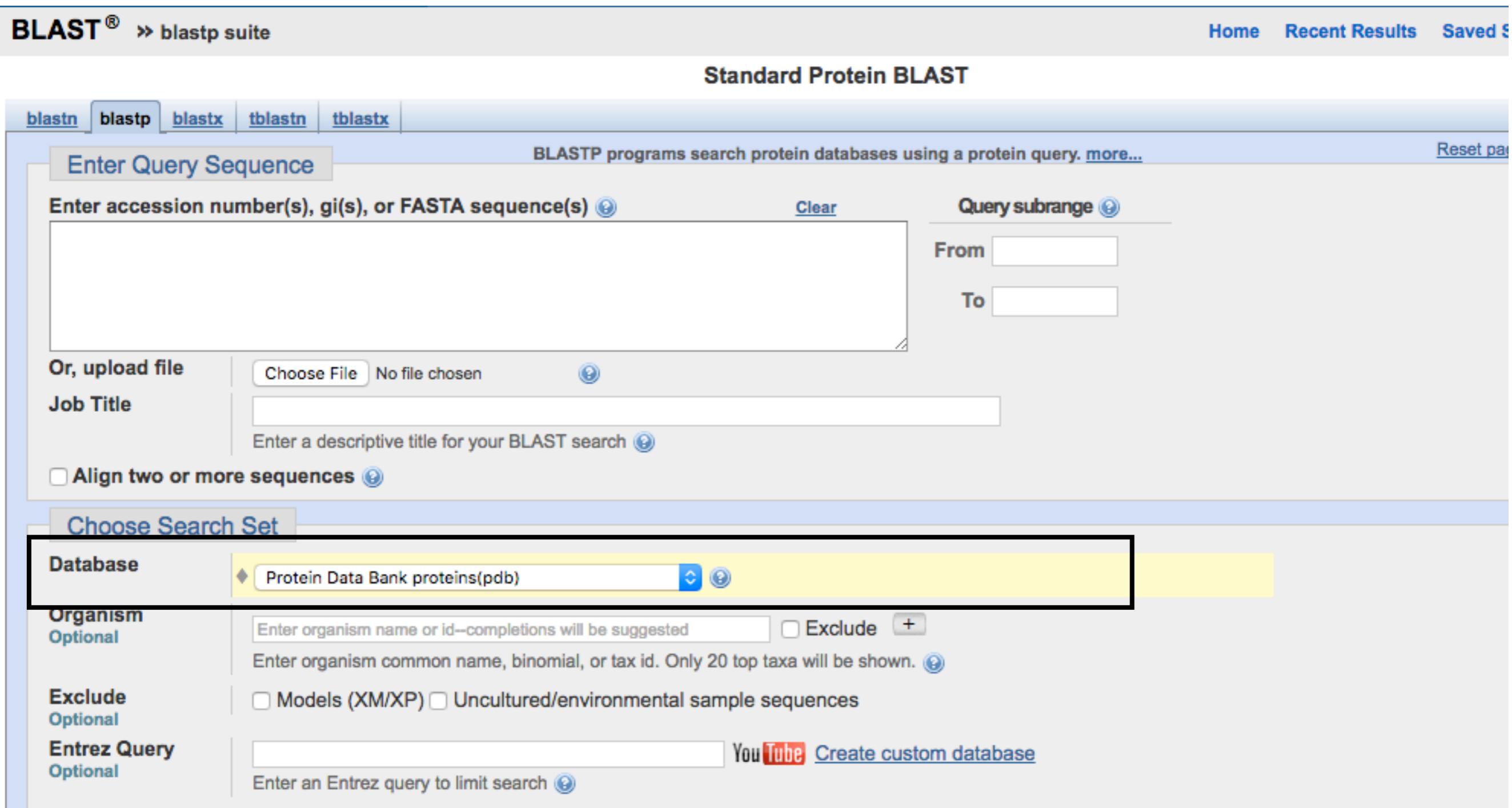
[Choose Search Set](#)

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 [?](#)

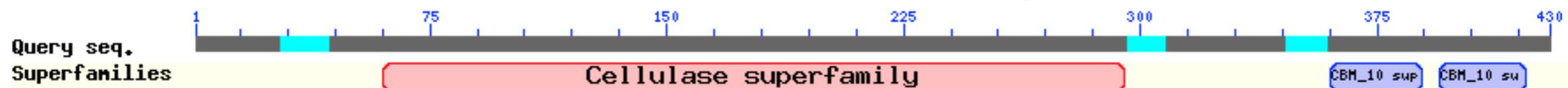


1. What question to ask?
2. What program to use?
3. What database to search?
- 4. How to avoid mistakes**
 1. (what to look out for)
5. When to do something different
 1. (changing scoring matrices)

Imagine you are searching with a protein with multiple domains

Job Title: gb|AAO31759| (430 letters)

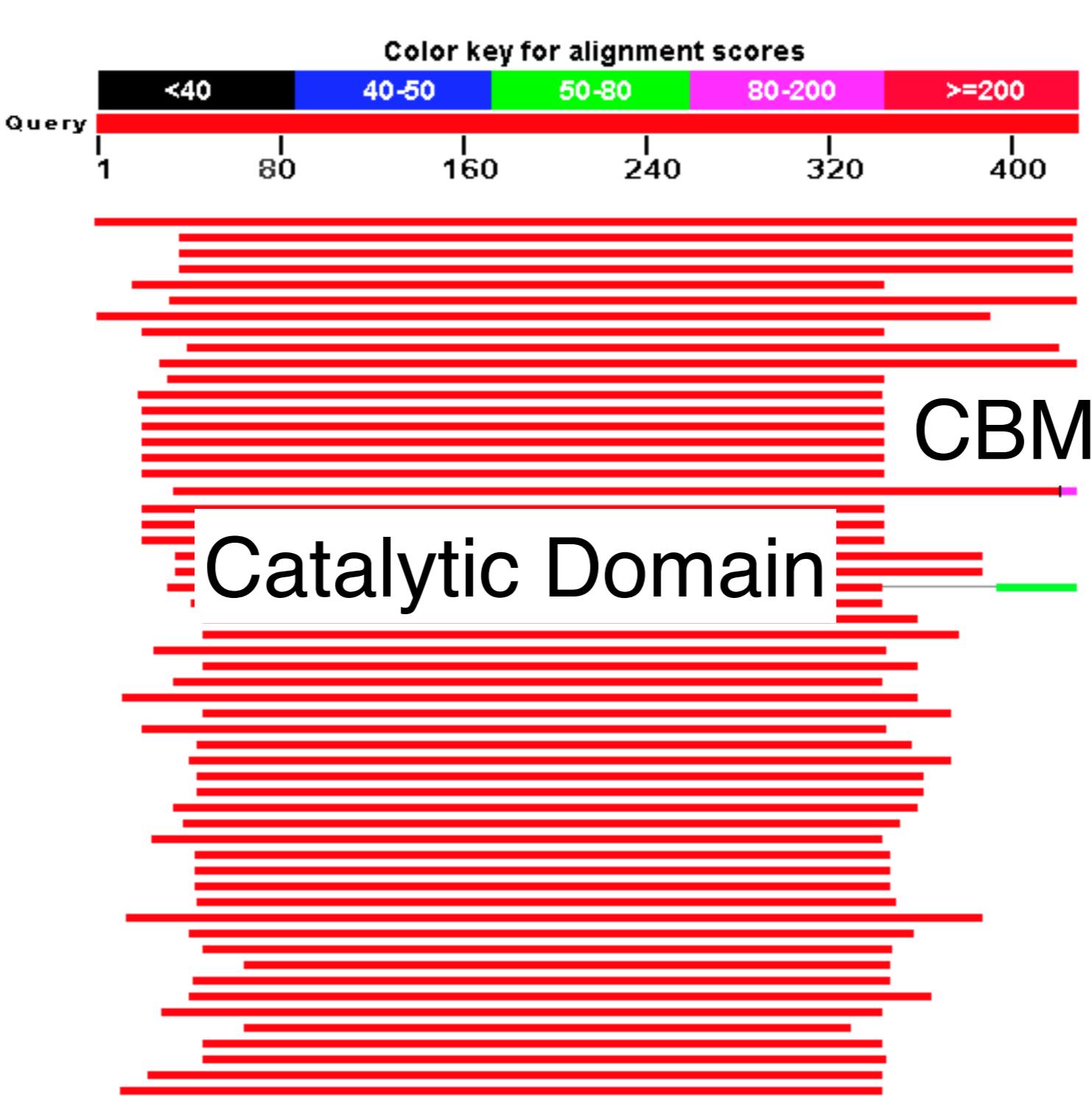
Putative conserved domains have been detected, click on the image below for detailed results.



Request ID	KUZ8VU8K01R
Status	Searching
Submitted at	Thu Feb 16 17:03:07 2012
Current time	Thu Feb 16 17:03:12 2012
Time since submission	00:00:04

This page will be automatically updated in 12 seconds

Not all hits are to the full protein



Look at the Alignment Coverage

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E-value	Max ident	Links
YP_001983792.1	endo- 1,4-beta-mannanase [Cellvibrio japonicus Ueda107]	875	875	100%	0.0	100%	G
ZP_04412299.1	beta-1,4-mannanase [Vibrio cholerae TM 11079-80]	410	410	90%	2e-137	52%	
YP_005049078.1	unnamed protein product [Vibrio furnissii NCTC 11218]	407	407	90%	2e-136	52%	G
ZP_05878245.1	beta-1,4-mannanase [Vibrio furnissii CIP 102972]	407	407	90%	3e-136	52%	
NP_637144.1	mannan endo-1,4-beta-mannosidase [Xanthomonas campestris pv. campestris]	395	395	76%	9e-133	59%	G
YP_525540.1	unnamed protein product [Saccharophagus degradans 2-40]	399	399	92%	7e-131	55%	G
YP_001982936.1	endo- 1,4-beta-mannanase [Cellvibrio japonicus Ueda107]	399	399	90%	1e-130	50%	G
ZP_08181055.1	Cellulase (glycosyl hydrolase family 5) [Xanthomonas vesicatoria ATCC 35937]	387	387	75%	2e-129	58%	
YP_003162168.1	glycoside hydrolase family protein [Jonesia denitrificans DSM 20603]	377	377	88%	1e-124	49%	G
YP_003075599.1	glycoside hydrolase family 5 domain-containing protein [Teredinibacter turneri]	378	511	93%	1e-122	69%	G
ZP_08184376.1	Cellulase (glycosyl hydrolase family 5) [Xanthomonas gardneri ATCC 19865]	369	369	73%	2e-122	59%	
YP_431433.1	endoglucanase [Hahella chejuensis KCTC 2396]	372	372	75%	5e-121	53%	G
ZP_06489984.1	mannan endo-1,4-beta-mannosidase [Xanthomonas campestris pv. musacearum]	364	364	75%	2e-120	57%	
ZP_06486842.1	putative endo-1,4-beta-mannosidase [Xanthomonas campestris pv. vasculorum]	363	363	75%	2e-120	57%	
NP_642123.1	unnamed protein product [Xanthomonas axonopodis pv. citri str. 306]	363	363	75%	2e-120	58%	G
ZP_06704657.1	mannan endo-1,4-beta-mannosidase [Xanthomonas fuscans subsp. aurantifolia]	363	363	75%	5e-120	57%	
ZP_06729989.1	mannan endo-1,4-beta-mannosidase [Xanthomonas fuscans subsp. aurantifolia]	362	362	75%	7e-120	57%	
YP_526130.1	unnamed protein product [Saccharophagus degradans 2-40]	369	457	92%	9e-120	46%	G
YP_004851393.1	mannan endo-1,4-beta-mannosidase [Xanthomonas axonopodis pv. citrumelo]	359	359	75%	1e-118	57%	G
ZP_08186387.1	Cellulase (glycosyl hydrolase family 5) [Xanthomonas perforans 91-118]	358	358	75%	2e-118	57%	

Score

E-value

Coverage

MaxID

BLAST Reports Multiple Highest Scoring Pairs

GENE ID: 8210864 TERTU_2894 | glycoside hydrolase family 5 domain-containing protein [Teredinibacter turnerae T7902] (10 or fewer PubMed links)

Sort alignments for this subject sequence by:
E value Score Percent identity
Query start position Subject start position

Score = 353 bits (906), Expect = 2e-110, Method: Compositional matrix adjust.
Identities = 168/322 (52%), Positives = 227/322 (70%), Gaps = 9/322 (3%)

Query 33	LTALGLMLAAV----SASAGFYVSGKQLREGNGNNFIMRGVNLPHAWFDPRTNQALADIS	88
	L+++ +AAV +A+AGF+V L + N F+MRGVN H W+ RT QAL DI	
Sbjct 70	LSSVAATIAAVCLSTAANAGFHVENGLLDANDKPFVMRCVNHAHTWYEARTQQALIDIE	129
Query 89	ATGANSVRVVLNSNG---RLWSRTPESQVASIISQAKARQLITVLEVHDTTGYGEQT-AAT	144
	+ GAN+VR+VLSNG W R E VA II+Q KA ++I+++EVHD+TGY E+ AA	
Sbjct 130	SVGANAVRIVLNSNGAHGEGWGRDSEQAVAGIIAQMKALEMISIVEVDSTGYPEKAGAAP	189
Query 145	LSEAVDYWIAIRNALIGQEDYVIINIGNEPFGNGQSASTWLNLHRDAINRLRNAGFTHTL	204
	+S AVDYW+ I++ALIG+EDYVIINI NEPFGN SA W++ H++AI RLR AG THTL	
Sbjct 190	MSTAVDYWLDIKDALIGEEDYVIINIANEPFGNTASADDWIDAHKEAITRLRAAGLTHTL	249
Query 205	MVDAANWGQDWENIMRNNASSLFNSDPRRNIFSVHMYEVYPNDTAVNNYMSAF-NSMNL	263
	MVDAANWGQDW+ +MR++A +F DP N++FS+HMY+++ N AV++Y+ F L	
Sbjct 250	MVDAANWGQDWQYVMRDHAQEIFAHDPLANIVFSIHMYQIFNNRQAVDSYLYTFVEDYKL	309
Query 264	PLVVGEFAANHFGSYVDAGSIMARAQQYGFYLGWSWSGNSSNLSALDVVTNFNAGSLTT	323
	PLVVGEF A+H G VD SI+ + Y GYLGSWSGNS + +LD+ N++ L+	
Sbjct 310	PLVVGEFGADHGGEDVDEASILELCELYNLGYLGWSWSGNSSGVESLDITLNYDVNDLSP	369
Query 324	WGNLLINNTNGIRNTSRKATIF 345	
	WG+ LIN+ GIRNT++ A++F	
Sbjct 370	WGDFLINSAYGIRNTAQTAASFV 391	

Score = 51.2 bits (121), Expect = 3e-04, Method: Compositional matrix adjust.
Identities = 20/36 (56%), Positives = 24/36 (67%), Gaps = 1/36 (3%)

Query 396	CNWYGTSY-PICVNTSSGWGWEENNRSCLAASTCAAQ	430
	C WY P+C SGWGWENN+SCI +TCA+Q	
Sbjct 675	CQWYQDPLRPLCTQQDSGWGWEENNQSCIGRTTCASQ	710

Score = 46.6 bits (109), Expect = 0.008, Method: Compositional matrix adjust.
Identities = 17/32 (53%), Positives = 22/32 (69%), Gaps = 0/32 (0%)

Query 396	CNWYGTSYPICVNTSSGWGWEENNRSCLAASTC	427
	CNWYG P+C + GWG EN ++C+ ASTC	
Sbjct 778	CNWYGWIVPVCAFSDQGWGNENGQTCVGASTC	809

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

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

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

```
Query: TMP.q
1>>>gi|28200469|gb|AA031759.1| endo-beta-1,4-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 align
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|Q74706.1|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|O96W08.1|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|P54937.1|GUNA_CLOLO Endoglucanase A; Cellulase A; En ( 517) 166 45.3 0.0024 0.209 0.520 406 align
```

Perform a search with your “suspect”

```
The best scores are:
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_PAEBB 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?

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 protein	(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 subunit	(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

paaggpppqgp	1-8 9-19 20-122	MYPSPVRH IKFTIADTLERIKEENFLQAQYHSIKLEC EKLSNEKTEMQRHYVEMYEMSYGLNVEMHK QTEIAKRLNTLINQLLPFLQADHQQQVLQA VERAKQVTMQELN
liighqqqhgicqqllqqihaqqvpqggppqp mg	123-154 155-292	
rppsrsgssssrstsps	293-308	
akartptpnaaapapgvnpk qmmpqgpppagypgapyqrpa	309-321 322-341 342-362 363-730	LTKDMEKPGTPG DPYQRPPSDPAYGRPPPMPYDPHAHVRTNG IPHPSALTGGKPAYSFHMNGEGLQPVPFP PDALVGVGIPRHARQINTLSHGEVVCAVTI SNPTKYVYTGGKGCVKVWDISQPGNKNPVS QLDCLQRDNYIRSVKLLPDGRTLIVGGEAS NLSIWDLASPPTPRIKAELTSAAPACYALAI SPDSKVCFSCCSDGNIAVWDLHNEILVRQF QGHTDGASCIDISPDSRLWTGGLDNTVRS WDLREGRQLQQHDFSSQIFSLGYCPTGDWL AVGMENSHVVEVLHASKPDKYQLHLHESCVL SLRFAACGKWFVSTGKDNLNAWRTPYGAS IFQSKE TSSVLS CDISTDDKYIVTGSGDKK ATVYEVIY

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

(B) Query sequence: FASTA format

Subset range: _____ Use Subset range

Compare your own sequences:
[Compare sequences](#)

Entrez protein sequence browser
Entrez DNA sequence browser

Or upload query from file: [Browse...](#)

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

(C) Database:

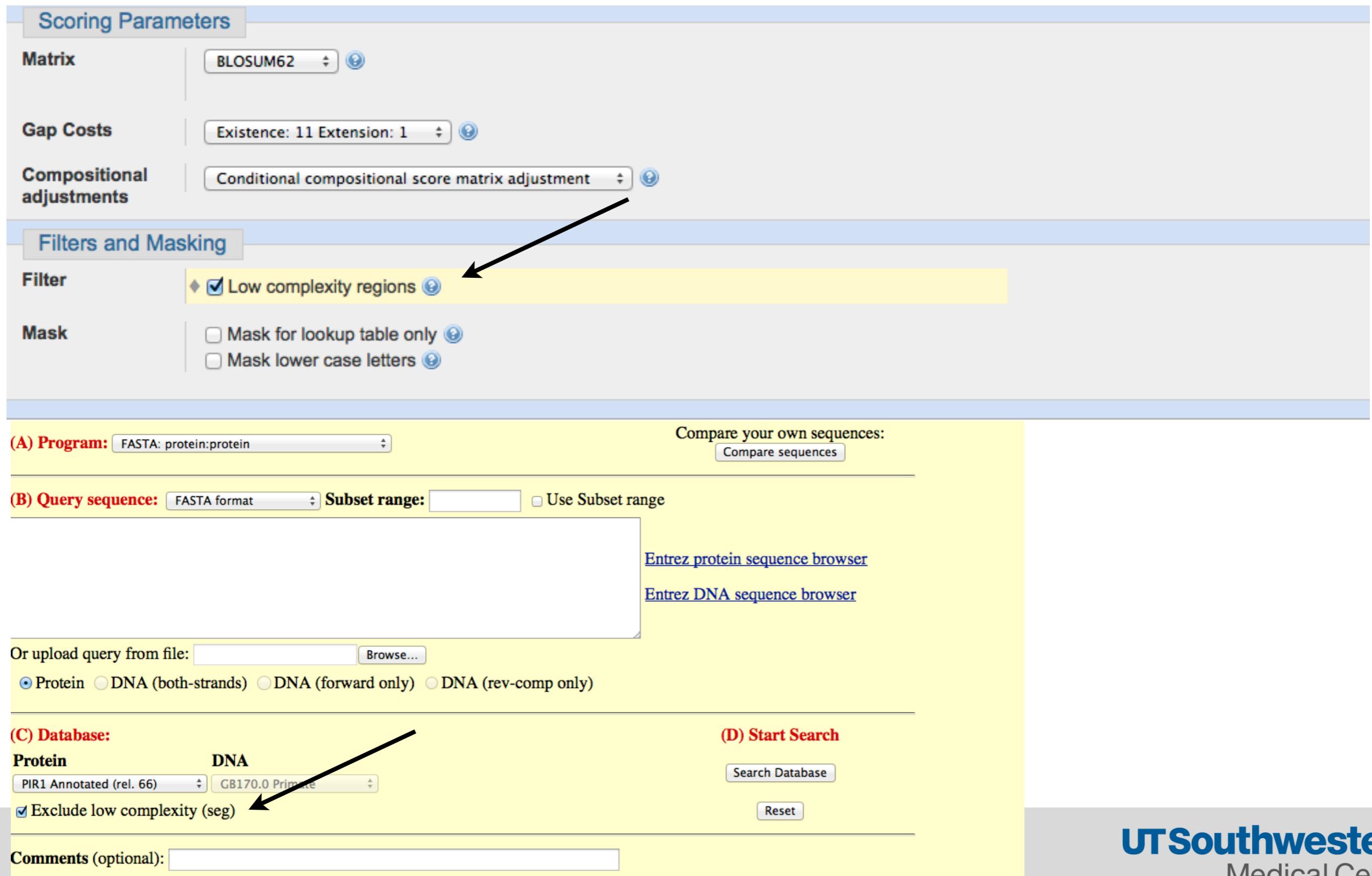
Protein: PIR1 Annotated (rel. 66)
DNA: GB170.0 Primate

Exclude low complexity (seg)

(D) Start Search

[Search Database](#) [Reset](#)

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

Shuffle Sequence

Reset Form

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

Other comparison options:

Scoring matrix: open: ext: Ktup:

Blosum50 (25%) -10 -2 ktup = 2

- What is Homology? How do we recognize it?
- How do we measure sequence similarity by alignment and scoring matrices?
- DNA vs Protein Comparisons
- Determining Homology
- Alignment Algorithms

How to Make an Alignment

PMILGYWNVRGL
:
PPYTIVYFPVRG

PMILGYWNVRGL
:
PPYTIVYFPVRG

PM-ILGYWNVRGL
:
PPYTIV-YFPVRG

PMILGYWNVRGL	
P	:
P	:
Y	.
T	.
I	:
V	..
Y	.
F	...
P	:
V	...
R	:
G	:

Local Alignment

AAPMILGYWNVRGLBB
:
DDPPYTIVYFPVRGCC

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(Ai, Bj)
        Delete ← F(i-1, j) + d
        Insert ← F(i, j-1) + d
        F(i,j) ← max(Match, Insert, Delete)
    }
```

Local Alignments

Local Alignment

AAPMILGYWNVRGLBB
 •••••
DDPPYTIVYFPVRGCCCC

A matrix H is built as follows:

$$H(i, 0) = 0, \quad 0 \leq i \leq m$$

$$H(0, j) = 0, \quad 0 \leq j \leq n$$

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

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

Where:

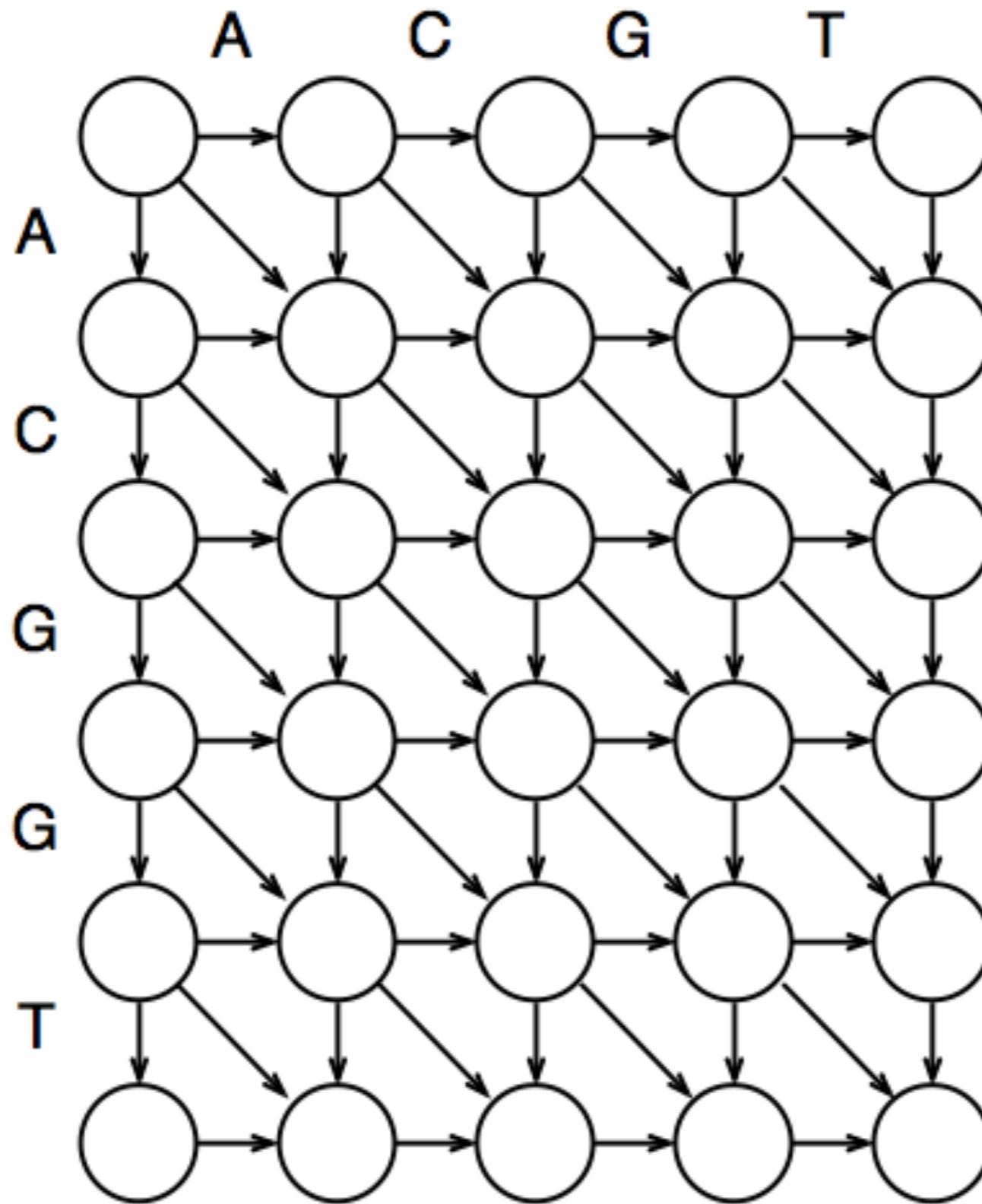
- a, b = Strings over the Alphabet Σ
- $m = \text{length}(a)$
- $n = \text{length}(b)$
- $H(i, j)$ - is the maximum Similarity-Score between a suffix of $a[1..i]$ and a suffix of $b[1..j]$
- $w(c, d), c, d \in \Sigma \cup \{'-\}$, '-' is the gap-scoring scheme

Local Alignments

Match: 1

Mismatch: -1

Gap: -2

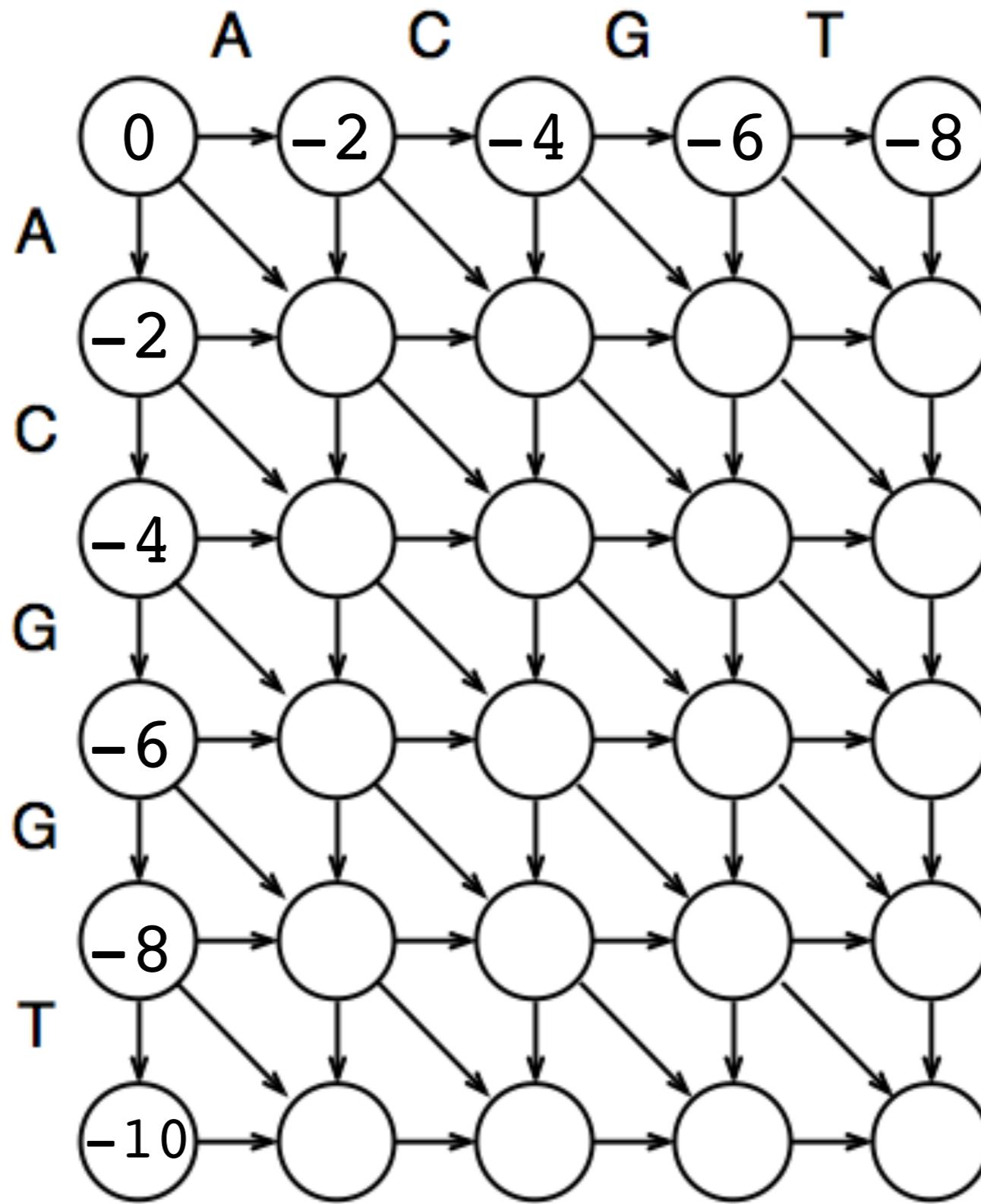


Local Alignments

Match: 1

Mismatch: -1

Gap: -2

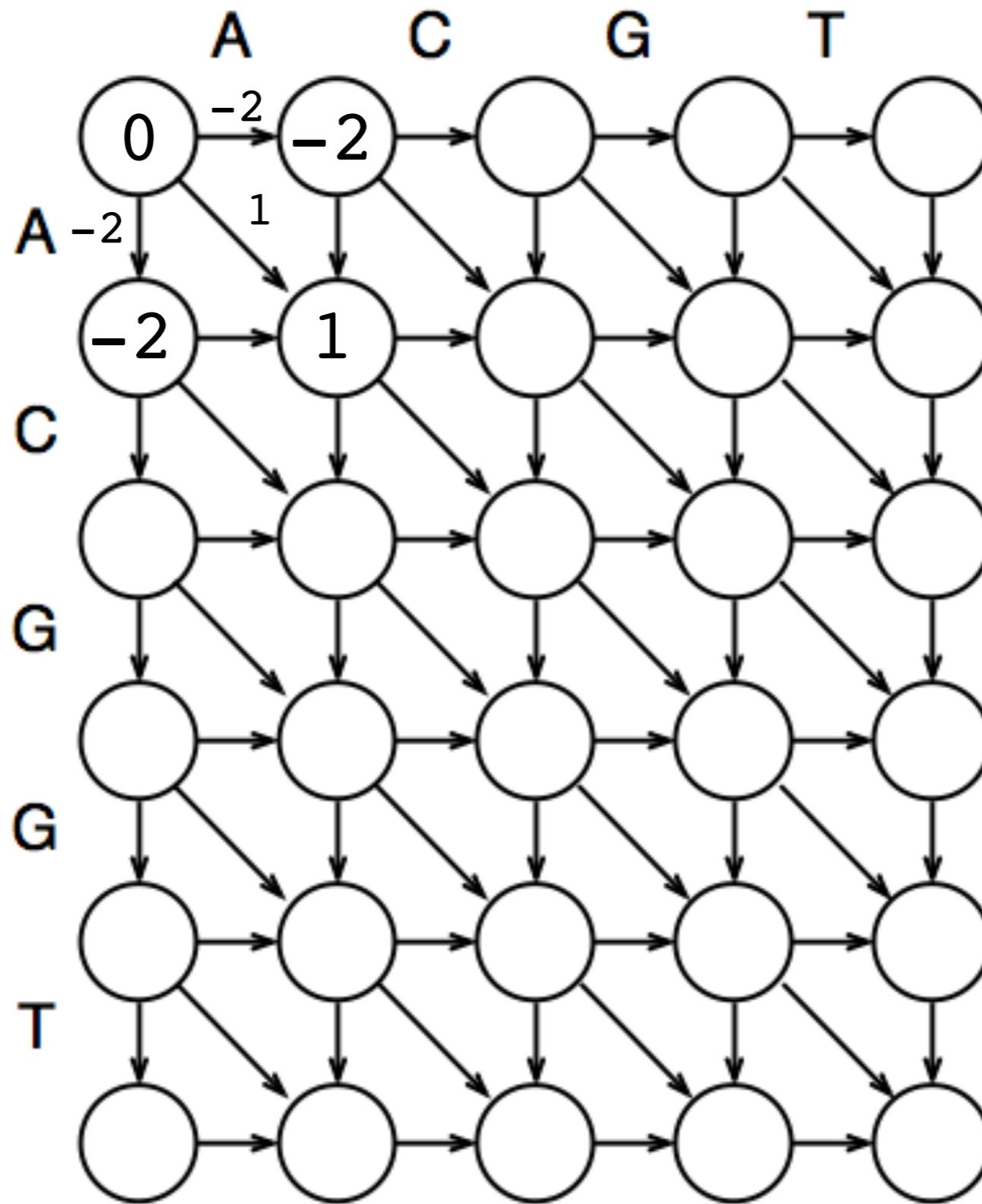


Local Alignments

Match: 1

Mismatch: -1

Gap: -2

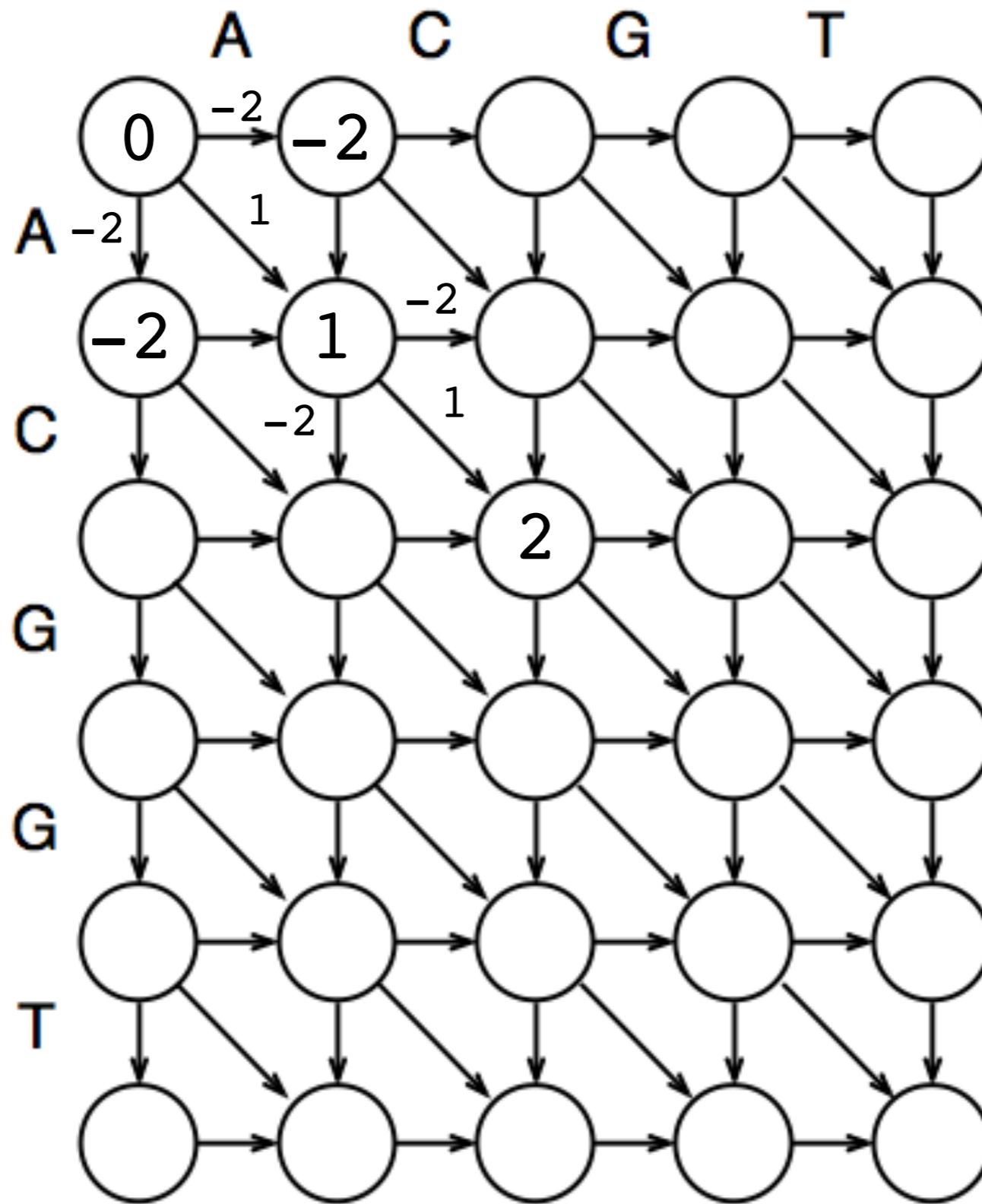


Local Alignments

Match: 1

Mismatch: -1

Gap: -2

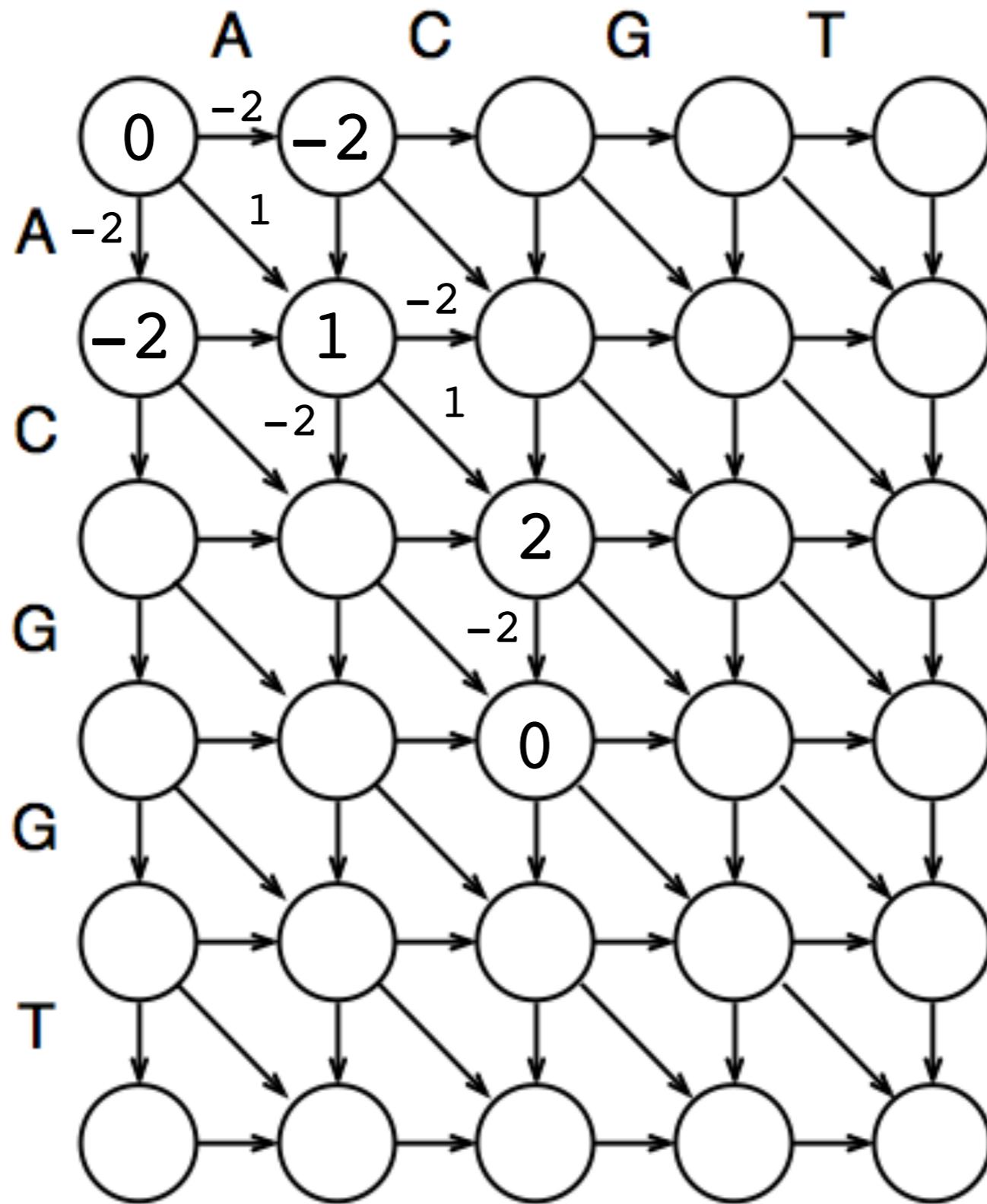


Local Alignments

Match: 1

Mismatch: -1

Gap: -2

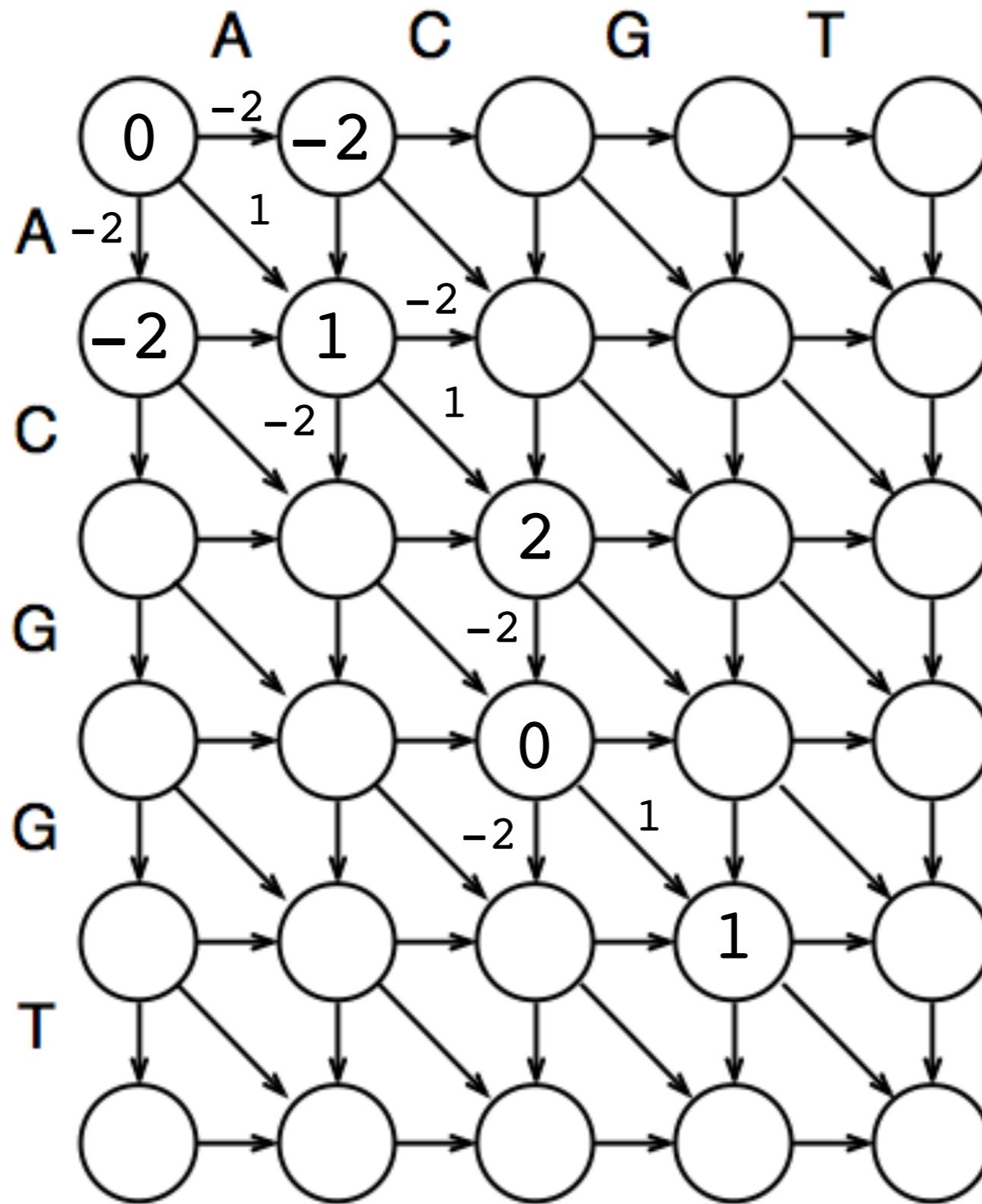


Local Alignments

Match: 1

Mismatch: -1

Gap: -2



Local Alignments

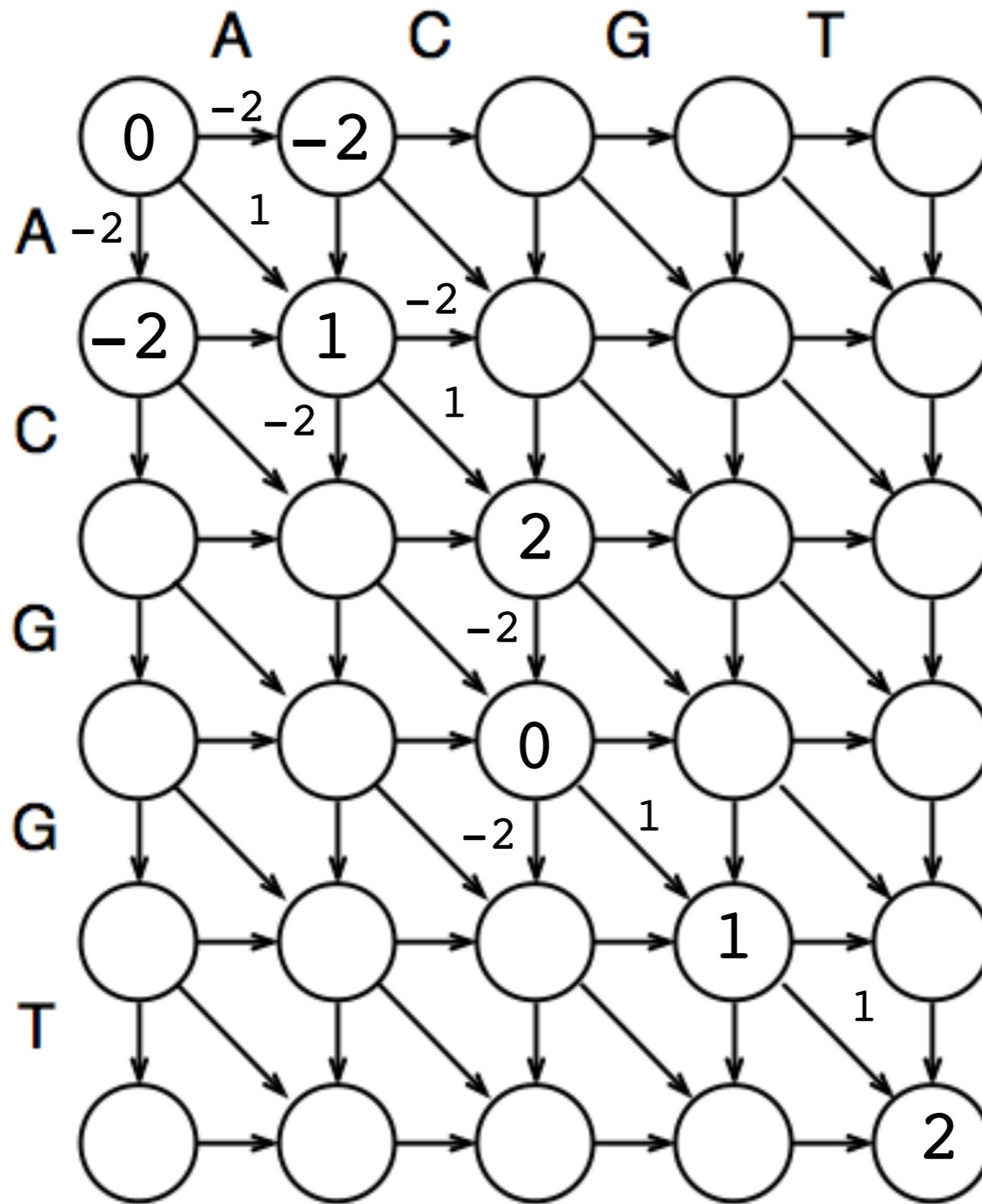
Match: 1

Mismatch: -1

Gap: -2

AC-GT

ACGGT

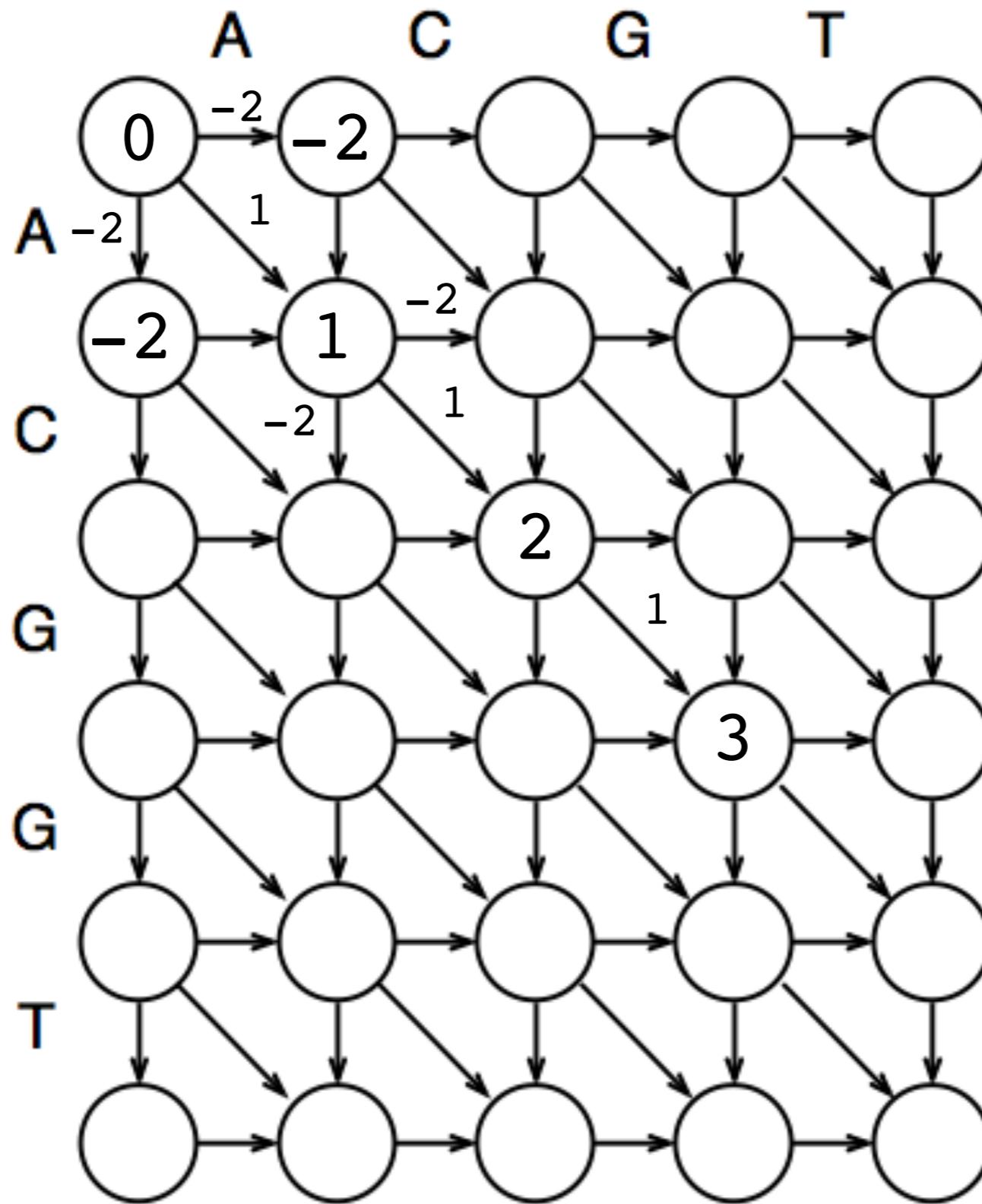


Local Alignments

Match: 1

Mismatch: -1

Gap: -2



Local Alignments

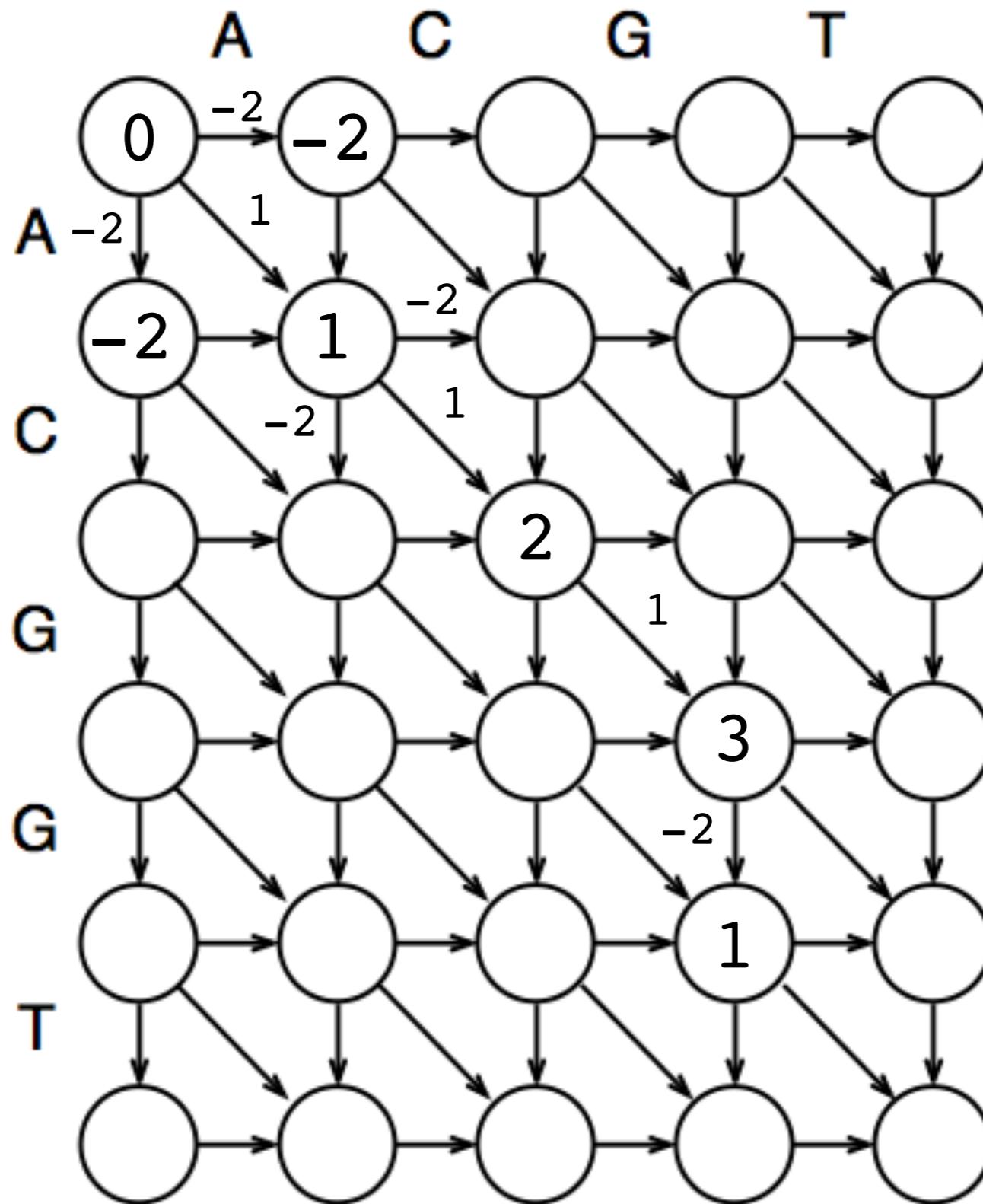
Match: 1

Mismatch: -1

Gap: -2

ACG-T

ACGGT



Local Alignments

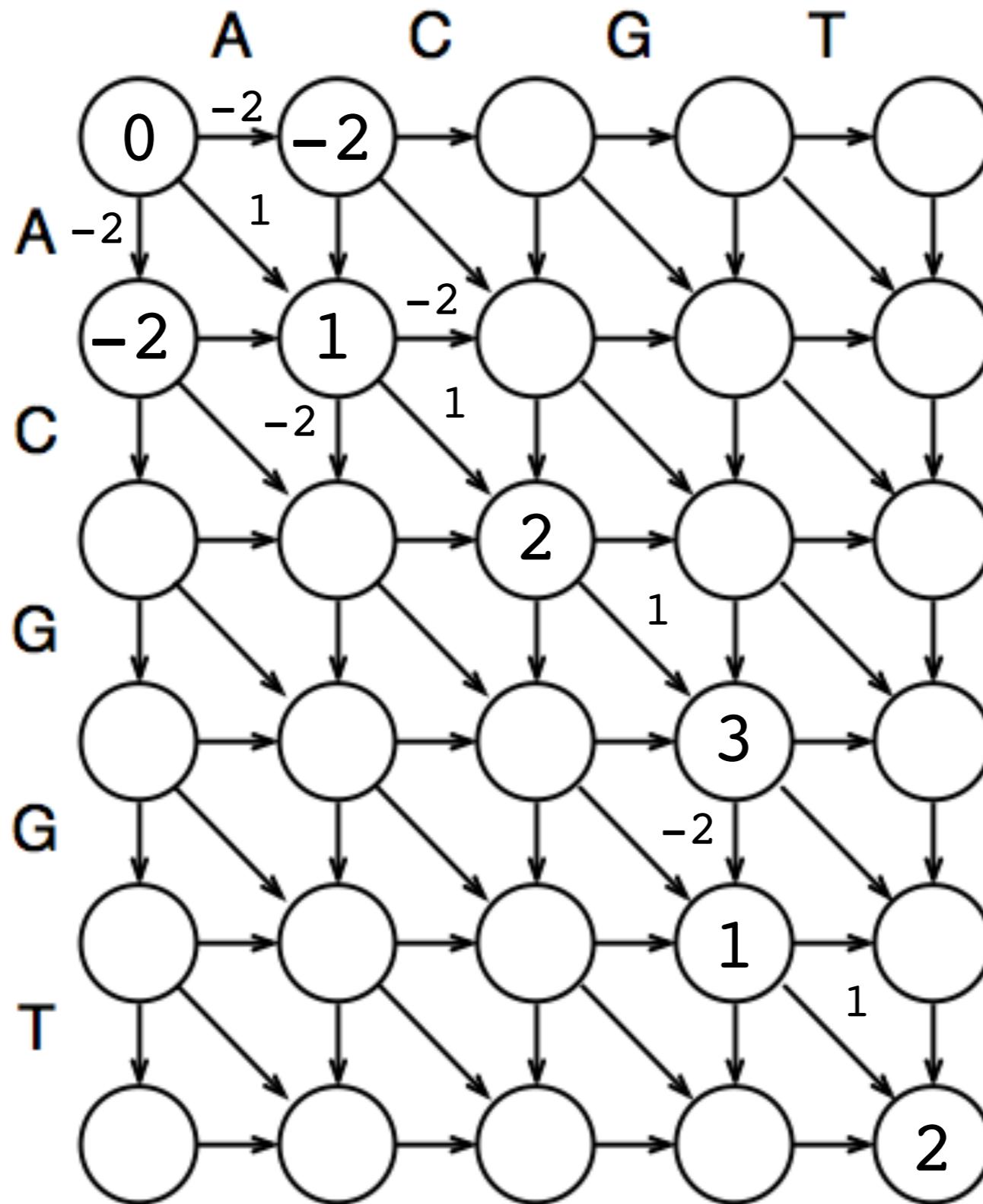
Match: 1

Mismatch: -1

Gap: -2

ACG-T

ACGGT



Scoring Matrices

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

- Lower PAM (PAM10/MDM10 ... PAM60) for closer (10% ...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

Finding Domains Local Alignments

>>>gi|49037474|sp|P62158.2|CALM_HUMAN, 149 aa vs TMP.q2 library

>>sp|P62158.2|CALM_HUMAN Calmodulin; CaM
Waterman-Eggert score: 220; 50.8 bits; E(1) < 1.1e-11
46.1% identity (73.7% similar) in 76 aa overlap (1-76:77-149)
[Entrez Lookup](#) [Re-search database](#) [General re-search](#)

10	20	30	40	50	60	70
gi 490 MADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLQNPTEAELQDMINEVDADGN	TIDFPEFLTMMARK					
sp P62 MKDTDSEEEI---REAFRVFDKGNGYISAAELRHVMTNLGEKLTDDEEVDEMIREA	IDGDGQVN	YEEFVQMMTAK				
80	90	100	110	120	130	140

Calmodulin

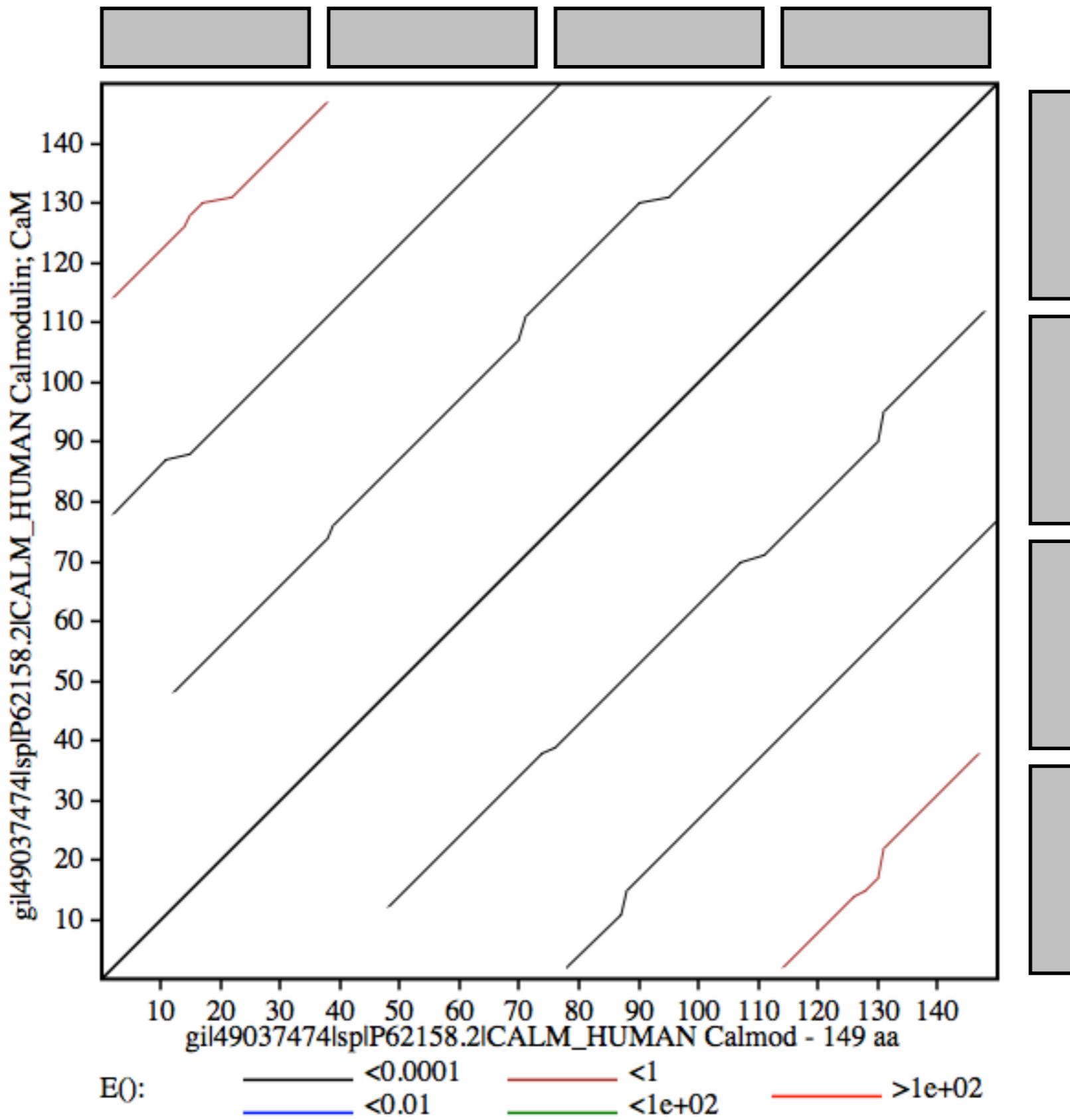
Waterman-Eggert score: 181; 42.6 bits; E(1) < 3.2e-09
34.3% identity (64.8% similar) in 105 aa overlap (11-111:47-147)
[Entrez Lookup](#) [Re-search database](#) [General re-search](#)

20	30	40	50	60	70	80
gi 490 AEFKEAFSLFDKDGDTITTKELGTVM-RSLQNPTEAELQDMINEVDADGN	TIDFPEF---LTMMARKMKD	TSEEEI				
sp P62 AELQDMINEVDADGN	TIDFPEFLTMMARKMKD	TSEEEI	REAFRVFDKGNGYISAAELRHVMTNLGEKLTD	EVDEMI		
50	60	70	80	90	100	110
90	100	110				
gi 490 REAFRVFDKGNGYISAAELRHVMT						
sp P62 REA----DIDGDGQVN	YEEFVQMMT					
130	140					

Waterman-Eggert score: 64; 18.2 bits; E(1) < 0.07
34.2% identity (71.1% similar) in 38 aa overlap (1-37:113-146)
[Entrez Lookup](#) [Re-search database](#) [General re-search](#)

10	20	30
gi 490 MADQLTEEQIAEF-KEA	FSLFDKDGDTITTKELGTVM	
sp P62 LGEKLTD	EEVDEMIREA	---
120	130	140

Finding Domains Local Alignments



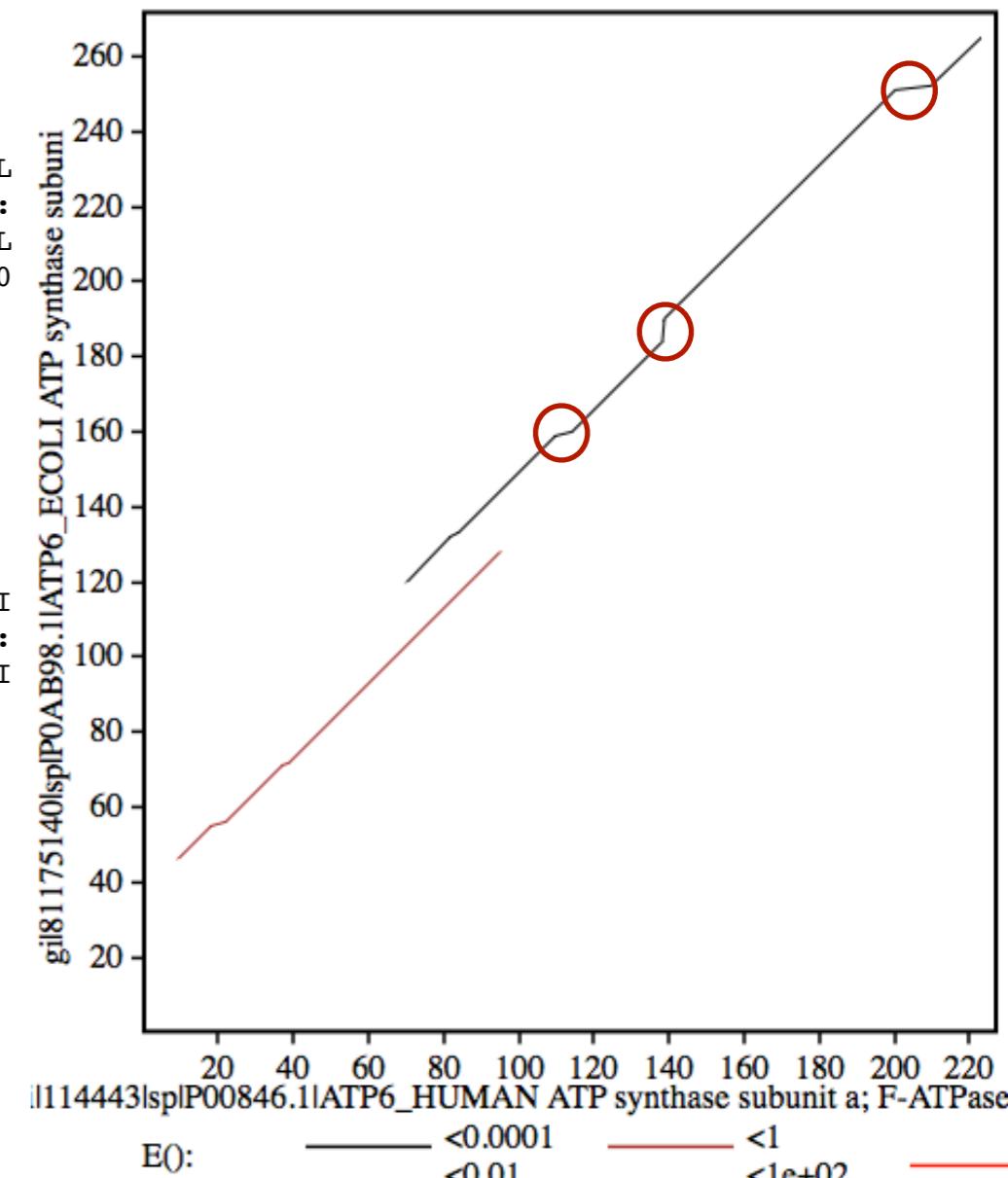
Local Alignments

gi|114 SFIAPTILGLPAAVLIIILFPPLLIPTSKYLIINRLITTQQWLKLTSKQMMTMHNTKGRTWSLMLVSLIIFIATTNLLGL
 sp|P0A SMFFSVVLGGL---LFLVLFRSVAKKATSG-VPGKFQTAIELVIGFVNNGVKDMYHGKSCLIAPLALTIFVWWFLMNLMDL

gi|114 LPHSFTP
 sp|P0A LPIDLLP

gi|114 SLMLVSLIIFIATTNLLGLLPHSFTPTTQLSMNLAMAIPWAGTVIMGFRSKIKNALAHFLPQGTPTPL---IPMLVI
 sp|P0A DLLPIDLLPYIAE-HVLGLPALRVVPSADVNTLSMALCVF---ILILFYSIKMKGIGGFTKELTLQPENHWAFIPVNLI

gi|114 IETISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYL
 sp|P0A LEGVSLLSKPVSLGLRLFGNMYAGELIFILIAGLLPWWSQWILNVPWAIFHILIIT---LQAFIFMVLTIVYL



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

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

$$q_{R:N(40)} = 0.000435$$

$$p_R = 0.051$$

$$q_{R:N(250)} = 0.002193$$

$$p_N = 0.043$$

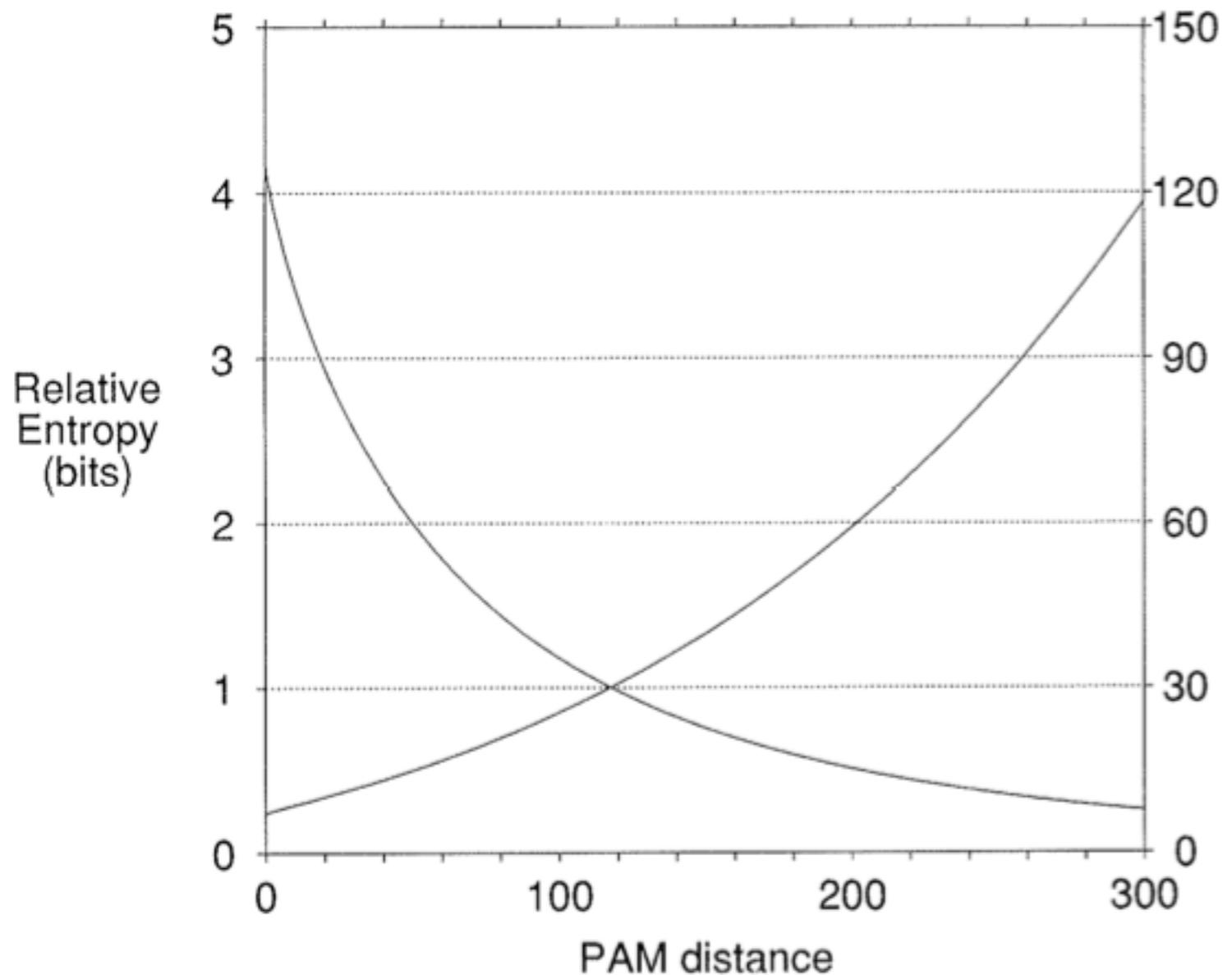
$$\text{I}_2 S_{ij} = \lg_2 (q_{ij}/p_i p_j) \quad \text{I}_e S_{ij} = \ln(q_{ij}/p_i p_j) \quad p_R p_N = 0.002193$$

$$\text{I}_2 S_{R:N(40)} = \lg_2 (0.000435/0.00219) = -2.333$$

$$\text{I}_2 = 1/3; S_{R:N(40)} = -2.333/\text{I}_2 = -7$$

$$\text{I}_2 S_{R:N(250)} = \lg_2 (0.002193/0.002193) = 0$$

Details on Scoring Matrices

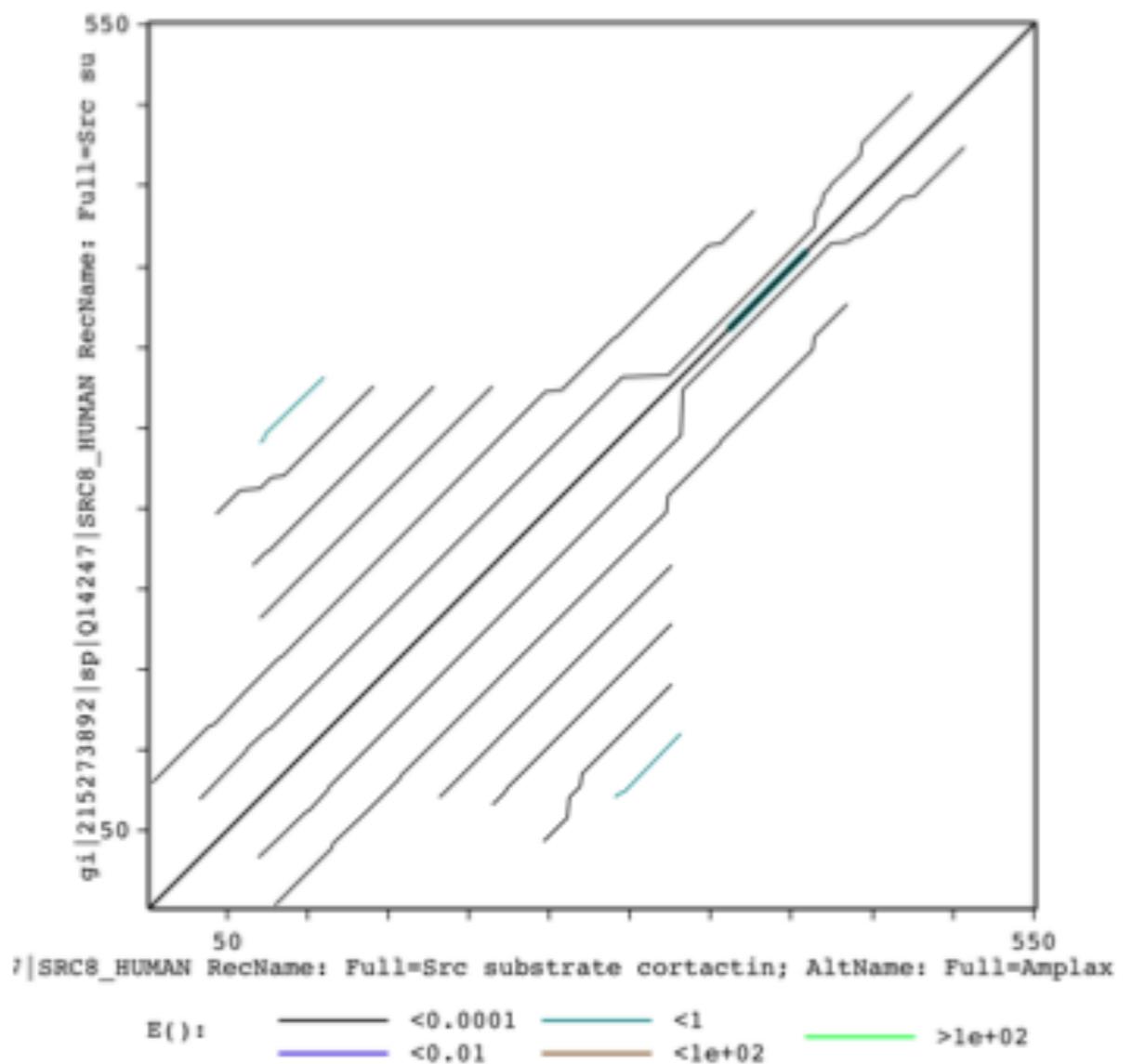


As sequences diverge,
there is less information
per position

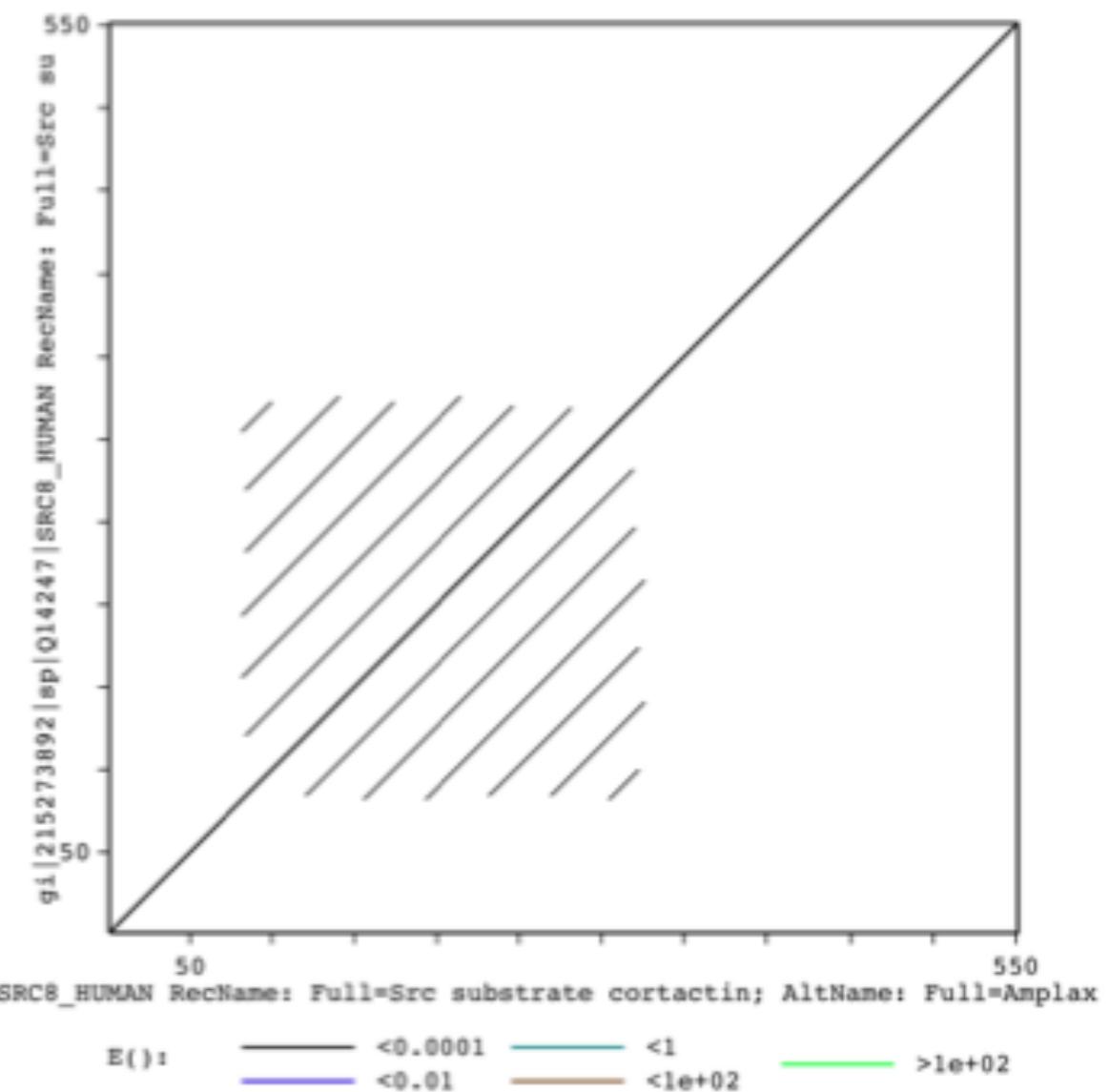
As sequences diverge,
longer alignments are
required to contain the
score threshold

Stringent Score Leads to Short Alignments

BLOSUM62 -11/-1



MD20 -26/-4



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