

## Sensitive similarity searches – PSSMs (PSI-BLAST) and HMMs

- Protein divergence is not uniform over a protein - some parts are more conserved than others
- Position specific scoring matrices can capture the specific patterns of conservation at different sites in a protein
- PSI-BLAST combines searching, multiple alignment, and PSSMs
- Statistical estimates are difficult with PSSMs, use PSI-SEARCH and PSI-PRSS
- Iterative PSSM/HMM searches may be contaminated by Homologous Overextension
- Single models cannot capture diverse families (PFAM Clans)

## Sensitive Similarity Searching – PSSMs, and HMMs

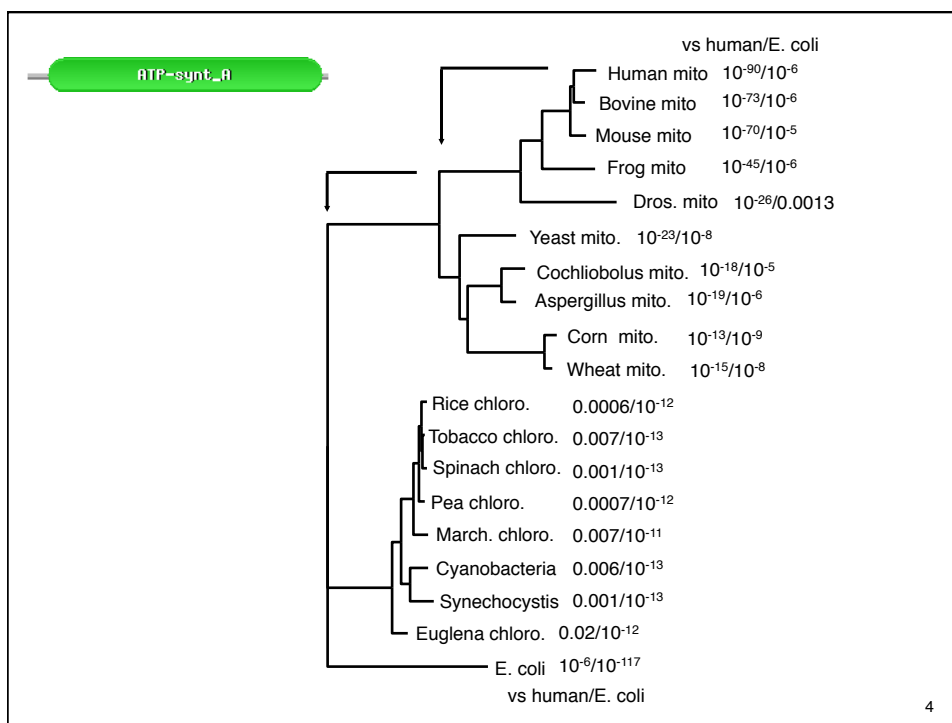
Pevsner, Chapter 5, pp. 145-161

- Gribskov, M. and Veretnik, S. (1996) Identification of sequence pattern with profile analysis. *Methods Enzymol* 266:198-212.
- Henikoff, S. and Henikoff, J. G. (1997) Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.* 6:698-705.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
- A. A. Schaffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul (2001) "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements" *Nucleic Acids Res.* 29:2994-3005
- S. R. Eddy (1998) "Profile Hidden Markov Models" *Bioinformatics* 14:755-763 (\*)
- S. R. Eddy (2008) "A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation" *PLOS Comp. Biol.* 4:e1000069
- M. W. Gonzalez and W. R. Pearson (2010) "Homologous Over-extension: a Challenge for Iterative Similarity Searches" *Nuc. Acids Res.* 38:2177-2189

## Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

Stephen F. Altschul\*, Thomas L. Madden, Alejandro A. Schäffer<sup>1</sup>, Jinghui Zhang, Zheng Zhang<sup>2</sup>, Webb Miller<sup>2</sup> and David J. Lipman

The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities. For protein comparisons, a variety of definitional, algorithmic and statistical refinements described here permits the execution time of the BLAST programs to be decreased substantially while enhancing their sensitivity to weak similarities. A new criterion for triggering the extension of word hits, combined with a new heuristic for generating gapped alignments, yields a gapped BLAST program that runs at approximately three times the speed of the original. In addition, a method is introduced for automatically combining statistically significant alignments produced by BLAST into a position-specific score matrix, and searching the database using this matrix. The resulting Position-Specific Iterated BLAST (PSI-BLAST) program runs at approximately the same speed per iteration as gapped BLAST, but in many cases is much more sensitive to weak but biologically relevant sequence similarities. PSI-BLAST is used to uncover several new and interesting members of the BRCT superfamily.



## ATP synthase - matrices, gaps, algorithms

Matrix:	BLOSUM50	BLOSUM62	BLASTP
Gap open/extend	-10/-2	-11/-1	-11/-1
<b>The best scores are:</b>	<b>bits E(13351)</b>	<b>bits E(13351)</b>	<b>bits E()</b>
ATP6_HUMAN ATP synthase a chai	297.7 1.7e-81	373.6 2.4e-104	296 3e-81
ATP6_BOVIN ATP synthase a chai	252.4 7.2e-68	310.7 2.0e-85	253 2e-68
ATP6_MOUSE ATP synthase a chai	246.4 4.5e-66	302.9 4.4e-83	245 5e-66
ATP6_XENLA ATP synthase a chai	111.9 1.4e-25	125.9 8.7e-30	142 9e-35
ATP6_YEAST ATP synthase a ch	78.7 1.6e-15	90.1 5.7e-19	93 5e-20
ATP6_EMENI ATP synthase a chai	66.3 8.4e-12	76.6 6.8e-15	75 2e-14
ATP6_DROYA ATP synthase a chai	65.6 1.2e-11	75.4 1.4e-14	101 2e-22
ATP6_COCHE ATP synthase a cha	53.6 5.5e-08	60.6 4.6e-10	75 1e-14
ATP6_ECOLI ATP synthase a ch	45.1 2.2e-05	49.1 1.4e-06	42 1e-04
ATP6_TRITI ATP synthase a ch	45.0 3.3e-05	50.7 6.5e-07	83 5e-17
ATP6_TOBAC ATP synthase a chai	40.4 0.00084	47.0 8.6e-06	80 3e-16
ATP6_MAIZE ATP synthase a chai	39.6 0.001	44.9 2.6e-05	
ATP1_PEA Chloroplast ATP syn	35.8 0.013	38.0 0.0028	
ATP1_SPIOL Chloroplast ATP syn	35.5 0.015	38.0 0.0028	
ATP1_ATRBE Chloroplast ATP s	34.0 0.044	36.3 0.0086	
ATP1_MARPO Chloroplast ATP syn	33.2 0.075	34.3 0.036	
*HBA_ODOVI Hemoglobin subunit a		31.9 0.11*	
*AROP_ECOLI Aromatic amino ac	32.1 0.31	31.4 0.5 *	
ATP1_EUGGR Chloroplast ATP syn	31.1 0.32	32.2 0.15	
ATP6_SYNP6 ATP synthase a chai	31.1 0.34	31.8 0.21	
TLCA_RICPR ADP,ATP carrier pro	31.5 0.49	29.7 1.7	
ATP6_SYNY3 ATP synthase a chai	30.6 0.51	31.8 0.22	28 1.9
ATP1_ORYSA Chloroplast ATP	30.1 0.65	32.2 0.15	
*GLUC_MYOSC Glucagon precursor	28.7 0.65	34.4 0.013*	
*VP6_BPPH6 Protein P6	29.1 0.85	28.6 1.3*	
*GLUC_LEPSP Glucagon precursor	27.7 1.	32.7 0.033*	
*ADH1_MOUSE Alcohol dehydrogena	29.8 1.2	34.4 0.013*	

5

## Metazoan ATP Synthases

CLUSTAL W (1.81) multiple sequence alignment

```

ATP6_BOVIN  MNENLFTSFITPVILGLPLVTLIVLFPSSLF--PTSNRLVSNRFVTLQQWMLQLVSKQMMSIHNSKGQTWT-LML
ATP6_MOUSE  MNENLFASFITPTMMGFPIVVAIIMFPSSILF--PSSKRLINNRLHSFQHWLVKLIKQMMLIHTPKGRWT-LMI
ATP6_HUMAN  MNENLFASFIAPTILGLPAAVLIILFPPLLI--PTSKYLINNRLITQQWLKLTLSKQMMTMHNTKGRTWS-LML
ATP6_XENLA  MNLSFFDQFMSFVILGILPIAIAMLDPFTLISWPIQSNGFNRLITLQSWFLHNFITIFYLQTSF-GHKWA-LLL
ATP6_DROYA  MMTNLFVDFPSAIFNLSLNWLSLFLGLMI--PSIYWLMPSRYNIFWNSILLTLHKEFKLLGPSGHNGSTFFIF
*  .:* * ..:..:  :  :: *  ..*  ::  .:  :  .*:  .:  ::

ATP6_BOVIN  MSLLIFIGSTNLLGLLPHSFTPTTQLSMNLGMAIPLWAGAVITGFRNKTKASLAHFLPQGTPTPLIPMLVLIETI
ATP6_MOUSE  VSLIMFIGSTNLLGLLPHFTPTTQLSMNLSMAIPLWAGAVITGFRHKLKSSLAHFLPQGTPTISLIPMLIIETI
ATP6_HUMAN  VSLIIFIAITNLLGLLPHSFTPTTQLSMNLAMAIPWAGAVIMGFRSKIKNALAHFLPQGTPTPLIPMLVLIETI
ATP6_XENLA  TSLMLLMSLNLLGLLPYFTPTTQLSLNMGAVLWLVATVIMASKP-TNYALGHLLEPGTPTPLIPVLIETI
ATP6_DROYA  ISLFSLILFNFMGLFPYIFTSTSHLTLTSLALPLWLCFMYLWINHTQHMFAHLVLPQGTPTAILMPFMVCJETI
**::  *:***:  **:::***  ::.  :  :***:***  **::  ***

ATP6_BOVIN  SLFIQPMALAVRLTANITAGHLLIHLIGGATLALMSISTTTALITFTLILLTILEFAVAMIQAYVFTLLVSLYLDNT
ATP6_MOUSE  SLFIQPMALAVRLTANITAGHLLMHLIGGATLVLMMNISPTTATITPIILLLLTILEFAVALIQAYVFTLLVSLYLDNT
ATP6_HUMAN  SLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTLILLTILEIAVALIQAYVFTLLVSLYLDNT
ATP6_XENLA  SLFIRPLALGVRLTANLTAGHLLIQLIATAAFVLLSIMPTVAILTIVLFLLLTLEIAVAMIQAYVFTLLVSLYLDNT
ATP6_DROYA  SNIIRPGLAVRLTANMIAGHLLLTLLGNTGPSMSYLLVTFLLVAQIALLVL--ESAVTMIQSYVFAVLSTLYSSEVN
* :*:  *.*****:  *****:  *.:  :  :  .  :  :  *:*  **::***:***:  *:*  :

```

6

### PSI-BLAST ATP6\_HUMAN - 4 iterations

Results from round:

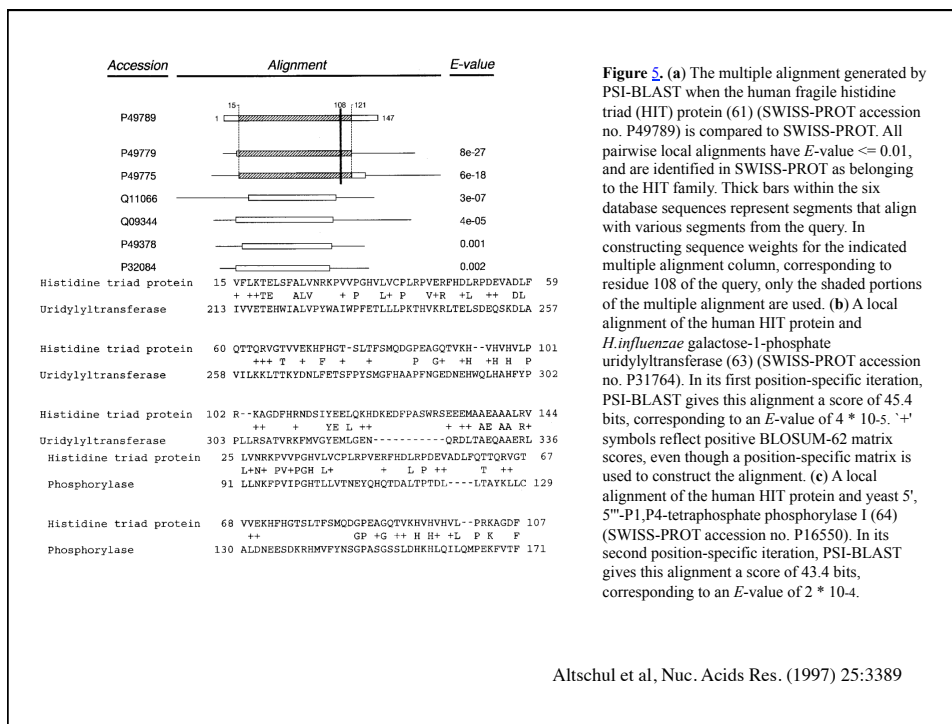
Sequences producing significant alignments:	(1)		(2)		(3)		(4)	
	Score (bits)	E Value	Score (bits)	E Value	Score (bits)	E Value	Score (bits)	E Value
ATP6_HUMAN ATP synthase a chain (ATPase protein 6)	296	3e-81	257	1e-69	241	2e-62	222	5e-59
ATP6_BOVIN ATP synthase a chain (ATPase protein 6)	253	2e-68	257	2e-69	239	8e-65	230	2e-61
ATP6_MOUSE ATP synthase a chain (ATPase protein 6)	245	5e-66	247	3e-66	234	4e-64	225	6e-60
ATP6_XENLA ATP synthase a chain (ATPase protein 6)	142	9e-35	227	1e-60	189	3e-49	177	2e-45
ATP6_DROYA ATP synthase a chain (ATPase protein 6)	101	2e-22	206	3e-54	209	5e-55	196	4e-51
(2)								
ATP6_YEAST ATP synthase a chain precursor (ATPase prot	93	5e-20	97	3e-21	199	4e-52	191	2e-49
ATP6_TRITI ATP synthase a chain (ATPase protein 6)	83	5e-17	96	5e-21	218	1e-57	236	4e-63
(3)								
ATP6_TOBAC ATP synthase a chain (ATPase protein 6)	80	3e-16	90	4e-19	200	2e-52	230	3e-61
ATP6_MAIZE ATP synthase a chain (ATPase protein 6)	76	5e-15	88	1e-18	198	1e-51	219	5e-58
ATP6_COCHE ATP synthase a chain (ATPase protein 6)	75	1e-14	86	9e-18			197	2e-51
ATP6_EMENI ATP synthase a chain precursor (ATPase prot	75	2e-14	84	3e-17	123	5e-29	181	2e-46
(4)								
ATP6_ECOLI ATP synthase a chain (ATPase protein 6)	42	1e-04	40	5e-04	46	8e-06	49	1e-06
ATPI_SPIOL Chloroplast ATP synthase a chain precursor			32	0.12	36	0.006	39	0.001
ATP6_SYNY3 ATP synthase a chain (ATPase protein 6)	28	1.9	32	0.16	44	5e-05	45	1e-05
ATPI_MARPO Chloroplast ATP synthase a chain precursor			31	0.21	44	4e-05	44	3e-05
ATPI_PEA Chloroplast ATP synthase a chain precursor (A			31	0.32	37	0.005		
LAMAZ_MOUSE Laminin subunit alpha-2 precursor (Laminin			31	0.34				
ATPI_ATRBE Chloroplast ATP synthase a chain precursor			31	0.39	41	2e-04		
ATP6_SYNP6 ATP synthase a chain (ATPase protein 6)			28	1.7	41	2e-04		
ATPI_EUGGR Chloroplast ATP synthase a chain precursor					39	0.001		
ATPI_ORYSA Chloroplast ATP synthase a chain precursor			28	1.9	36	0.008		
ATPI_ATRBE Chloroplast ATP synthase a chain precursor					36	0.009	38	0.002
ATP6_ASPAM ATP synthase a chain (ATPase protein 6)							36	0.008
POLG_KUNJM Genome polyprotein [Contains: Capsid protei...	27	5.0						
POL_HTL1C Gag-Pro-Pol polyprotein [Pri160Gag-Pro-Pol] [...	27	5.0						
POLG_DEN2J Genome polyprotein [Contains: Capsid protei...	27	5.2	26	7.0				

7

### Position-Specific Scores ATP Synthase, 4 iterations

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	bits/pos
BL62 Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0.70
46 Q	-2	-1	-2	-2	-4	6	0	1	0	-4	-3	-1	-2	-1	-3	-1	-2	6	4	-3	0.74
%	0	0	0	0	0	54	0	12	0	0	0	0	0	0	0	0	0	13	20	0	
47 Q	-1	-1	3	3	-3	3	3	-2	3	-4	-4	-1	-3	-4	-2	2	-1	-4	-2	-3	0.51
%	0	0	13	20	0	16	19	0	8	0	0	0	0	0	0	24	0	0	0	0	
56 Q	-2	-1	-2	-2	-3	5	2	-4	-1	4	-1	-1	-1	-2	-3	-2	-2	-3	-2	0	0.51
%	0	0	0	0	0	46	13	0	0	41	0	0	0	0	0	0	0	0	0	0	
97 Q	-2	-1	0	-2	-4	4	0	-3	8	-4	-4	-1	-2	-3	-3	-1	-2	-3	0	-4	1.11
%	0	0	0	0	0	35	0	0	65	0	0	0	0	0	0	0	0	0	0	0	
131 Q	3	-1	-1	-1	-2	5	2	-2	-1	-3	-3	0	-2	-4	-2	1	-1	-3	-3	-2	0.52
%	44	0	0	0	0	36	11	0	0	0	0	0	0	0	0	9	0	0	0	0	
152 Q	-2	6	-1	-2	-4	4	0	-3	-1	-4	-3	1	-2	-4	-3	-1	-2	-4	-3	-3	1.00
%	0	77	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
210 Q	-2	0	-1	-1	-4	7	1	-3	0	-4	-3	1	-1	-4	-2	-1	-2	-3	-2	-3	1.13
%	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

8



## PSI-BLAST

NCBI/BLAST/blastp suite

blastn blastp **blasts** tblastn tblastx

BLASTP programs search protein databases

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

From  To

Or, upload file  no file selected

Job Title

Align two or more sequences

Choose Search Set

Database

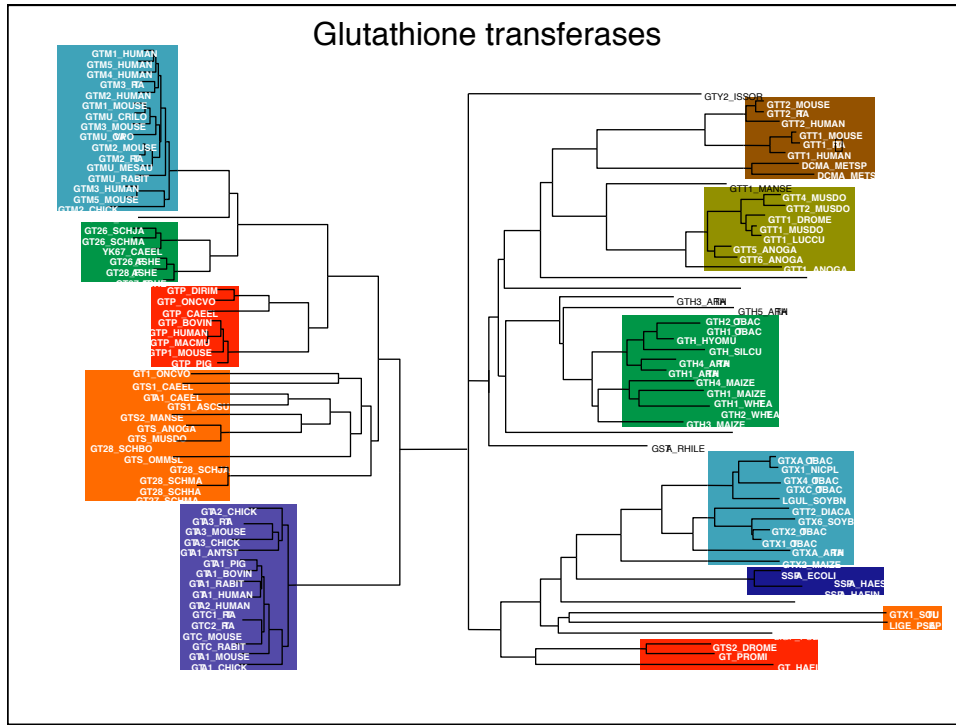
Organism

Exclude  Models (X/M/X/P)  Uncultured/environmental sample sequences

Entrez Query

Program Selection

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



### PSI-BLAST iteration 1 – 171 hits

Accession	Description	Max score	Total score	Query coverage	E value
<a href="#">Q226697.2</a>	RecName: Full=Glutathione S-transferase 3; AltName: Full=	43.5	43.5	89%	0.001
<a href="#">Q555FF3.1</a>	RecName: Full=Putative glutathione S-transferase alpha-1; A	43.1	43.1	71%	0.001
<a href="#">Q91252.2</a>	RecName: Full=Probable glutathione S-transferase 6; AltName	43.1	43.1	73%	0.001
<a href="#">Q91253.1</a>	RecName: Full=Probable glutathione S-transferase 7; AltName	42.7	42.7	92%	0.002
<a href="#">Q16115.1</a>	RecName: Full=Glutathione S-transferase 2; AltName: Full=	42.4	42.4	54%	0.002
<a href="#">Q26624.1</a>	RecName: Full=Glutathione S-transferase class-mu 28 kDa i	41.6	41.6	73%	0.004

Run PSI-Blast iteration 2 with max [500] [Go]

**Sequences with E-value WORSE than threshold**

Accession	Description	Max score	Total score	Query coverage	E value
<a href="#">Q06AXYD.1</a>	RecName: Full=Glutathione S-transferase A6; AltName: Full=	40.4	40.4	95%	0.009
<a href="#">Q27013.1</a>	RecName: Full=S-crystallin 1	39.3	39.3	89%	0.017
<a href="#">Q46434.1</a>	RecName: Full=Glutathione S-transferase 1	38.5	38.5	89%	0.032
<a href="#">Q082451.3</a>	RecName: Full=Probable glutathione S-transferase GSTF2; A	38.1	38.1	65%	0.038
<a href="#">Q04907.4</a>	RecName: Full=Glutathione S-transferase 3; AltName: Full=	37.4	37.4	66%	0.063
<a href="#">Q52828.1</a>	RecName: Full=Protein gstA	37.0	37.0	70%	0.083
<a href="#">Q42769.1</a>	RecName: Full=Glutathione S-transferase PM239X14; AltName	37.0	37.0	28%	0.089
<a href="#">Q30713.3</a>	RecName: Full=Glutathione S-transferase theta-2; AltName:	36.2	36.2	72%	0.16
<a href="#">Q0CG30.1</a>	RecName: Full=Glutathione S-transferase theta-2B; AltName:	35.8	35.8	60%	0.20
<a href="#">Q0CG29.1</a>	RecName: Full=Glutathione S-transferase theta-2; AltName:	35.8	35.8	60%	0.21
<a href="#">Q13155.2</a>	RecName: Full=Aminoacyl tRNA synthase complex-interactin	34.3	34.3	19%	0.52
<a href="#">Q85860.2</a>	RecName: Full=Protein ycf2	33.9	33.9	38%	0.69
<a href="#">Q2N1.00.3</a>	RecName: Full=Glutathione S-transferase theta-1; AltName:	33.5	33.5	55%	1.1
<a href="#">Q9USQ4.1</a>	RecName: Full=NASP-related protein sim3; AltName: Full=C	33.3	33.5	12%	1.1
<a href="#">P13860.1</a>	RecName: Full=Exoglucanase 1; AltName: Full=1,4-beta-cel	33.1	33.1	28%	1.2
<a href="#">P04437.1</a>	RecName: Full=Glutathione S-transferase; AltName: Full=G	32.7	32.7	38%	1.6

high-scoring unrelated

## PSI-BLAST iteration 2 – 313 hits

<input checked="" type="checkbox"/>	<a href="#">Q26624.1</a>	RecName: Full=Glutathione S-transferase class-mu 28 kDa	<a href="#">171</a>	171	92%	4e-42	was 0.004
<input checked="" type="checkbox"/>	<a href="#">P09792.1</a>	RecName: Full=Glutathione S-transferase class-mu 28 kDa	<a href="#">170</a>	170	93%	5e-42	
<input checked="" type="checkbox"/>	<a href="#">P30114.1</a>	RecName: Full=Glutathione S-transferase class-mu 28 kDa	<a href="#">168</a>	168	93%	3e-41	
<b>NEW</b>	<input checked="" type="checkbox"/>	<a href="#">Q18598.3</a>	<a href="#">168</a>	168	92%	3e-41	
<input checked="" type="checkbox"/>	<a href="#">P91252.2</a>	RecName: Full=Probable glutathione S-transferase 6; AltName: Full=G	<a href="#">167</a>	167	93%	4e-41	
<input checked="" type="checkbox"/>	<a href="#">P46428.4</a>	RecName: Full=Glutathione S-transferase; AltName: Full=G	<a href="#">166</a>	166	94%	8e-41	

high-scoring  
unrelated?

<input type="checkbox"/>	<a href="#">Q58248.1</a>	RecName: Full=Uncharacterized protein M0838	<a href="#">33.6</a>	33.6	23%	0.90
<input type="checkbox"/>	<a href="#">P30151.3</a>	RecName: Full=Elongation factor 1-beta; Short=EF-1-beta;	<a href="#">33.6</a>	33.6	30%	0.96
<input type="checkbox"/>	<a href="#">Q6X1Y6.2</a>	RecName: Full=Amiloride-sensitive cation channel 3; AltName: Full=	<a href="#">33.6</a>	33.6	43%	0.98
<input type="checkbox"/>	<a href="#">P95806.1</a>	RecName: Full=2,5-dichlorohydroquinone reductive dechlorin	<a href="#">32.9</a>	32.9	42%	1.4
<input type="checkbox"/>	<a href="#">Q70854.1</a>	RecName: Full=Amiloride-sensitive cation channel 4-A; AltName: Full=	<a href="#">32.9</a>	32.9	43%	1.5
<input type="checkbox"/>	<a href="#">Q5P8W0.1</a>	RecName: Full=3-dehydroquinone dehydratase; Short=3-deh	<a href="#">32.9</a>	32.9	38%	1.5
<input type="checkbox"/>	<a href="#">Q96WL3.1</a>	RecName: Full=Protein URE2	<a href="#">32.9</a>	32.9	74%	1.6
<input type="checkbox"/>	<a href="#">Q9SCX3.1</a>	RecName: Full=Elongation factor 1-beta 2; Short=EF-1-beta	<a href="#">32.5</a>	32.5	27%	1.8
<input type="checkbox"/>	<a href="#">Q94524.1</a>	RecName: Full=Glutathione S-transferase omega-like 2	<a href="#">32.5</a>	32.5	90%	2.1
<input type="checkbox"/>	<a href="#">Q8NEC7.2</a>	RecName: Full=Glutathione S-transferase C-terminal domain	<a href="#">32.5</a>	32.5	27%	2.2
<input type="checkbox"/>	<a href="#">Q9K1N8.1</a>	RecName: Full=Peptide methionine sulfoxide reductase msrA	<a href="#">32.5</a>	32.5	49%	2.3
<input type="checkbox"/>	<a href="#">Q9VUR3.1</a>	RecName: Full=Probable aminoacyl tRNA synthase complex	<a href="#">32.1</a>	32.1	51%	2.4
<input type="checkbox"/>	<a href="#">Q4R6Y8.1</a>	RecName: Full=Glutathione S-transferase C-terminal domain	<a href="#">32.1</a>	32.1	27%	2.4
<input type="checkbox"/>	<a href="#">Q9JWM8.1</a>	RecName: Full=Peptide methionine sulfoxide reductase msrA	<a href="#">32.1</a>	32.1	49%	2.6
<input type="checkbox"/>	<a href="#">ASWH34.1</a>	RecName: Full=DNA-directed RNA polymerase subunit beta'	<a href="#">32.1</a>	32.1	30%	2.8

## Confirming PSI-BLAST hits - reverse SSEARCH

EF1B

The best scores are:

			s-w bits	E(191138)
gi 84028935 sp Q84WM9 EF1B1_ARATH	Elongation facto	( 228)	1216	297.3 3.1e-80
gi 75313298 sp Q9SCX3 EF1B2_ARATH	Elongation factor	( 224)	938	230.8 3.2e-60
gi 232031 sp P29545 EF1B_ORYSA	Elongation factor 1	( 223)	711	176.5 7.3e-44
gi 232033 sp P29546 EF1B_WHEAT	Elongation factor 1	( 216)	699	173.6 5.1e-43
gi 1706587 sp P53787 EF1D_RABIT	Elongation factor	( 280)	380	97.2 6.6e-20
gi 75025468 sp Q9U2H9 EF1B2_CAEL	Probable elongat	( 263)	359	92.2 2e-18
gi 2494267 sp P78590 EF1B_CANAL	Elongation factor	( 213)	336	86.8 7.1e-17
gi 13124243 sp Q9V18 EF1D_DROME	Probable elongati	( 256)	324	83.8 6.5e-16
gi 68845631 sp P32471 EF1B_YEAST	Elongation factor	( 206)	309	80.3 6e-15
gi 461992 sp P34827 EF1B_TRYCR	25 kDa elongation f	( 222)	283	74.1 4.9e-13
gi 3219795 sp O29681 EF1B_ARCFU	Elongation factor	( 88)	125	36.5 0.039
gi 10720007 sp Q9VG97 GSTT3_DROME	Glutathione S-tr	( 199)	122	35.6 0.17
gi 12643923 sp Q9VG98 GSTT2_DROME	Glutathione S-tr	( 215)	109	32.4 1.6
gi 12644404 sp Q10342 YBLE_SCHPO	Hypothetical prot	( 719)	113	33.1 3.5
gi 12643922 sp Q9VG96 GSTT4_DROME	Glutathione S-tr	( 215)	102	30.8 5.2
gi 12643921 sp Q9VG95 GSTT5_DROME	Glutathione S-tr	( 216)	102	30.8 5.2
gi 54037235 sp P67805 GSTT1_DROSI	Glutathione S-tr	( 200)	100	30.3 6.6

LIND

The best scores are:

			s-w bits	E(191138)
gi 32469676 sp P95806 LIND_PSEPA	2,5-dichlorohydro	( 346)	2308	553.1 7.4e-157
gi 38257679 sp Q8TB36 GDAP1_HUMAN	Ganglioside-indu	( 358)	222	58.9 4.3e-08
gi 38257365 sp O88741 GDAP1_MOUSE	Ganglioside-indu	( 358)	200	53.7 1.6e-06
gi 38257738 sp Q96M20 GD1L1_HUMAN	Ganglioside-indu	( 367)	200	53.7 1.7e-06
gi 38257686 sp O8VE33 GD1L1_MOUSE	Ganglioside-induc	( 370)	199	53.5 2e-06
gi 29429205 sp Q03520 PCPC_SPHCR	Tetrachloro-P-hyd	( 248)	191	51.7 4.5e-06
gi 11133278 sp Q9ZV04 GSTZ2_ARATH	Probable glutath	( 223)	118	34.5 0.63
gi 6225842 sp O73888 PTGD2_CHICK	Glutathione-requi	( 199)	116	34.0 0.76
gi 1170090 sp P04907 GSTF3_MAIZE	Glutathione S-tra	( 222)	116	34.0 0.87
gi 38257679 sp Q8TB36 GDAP1_HUMAN	Ganglioside-indu	( 358)	2263	512.0 1.8e-144
gi 38257365 sp O88741 GDAP1_MOUSE	Ganglioside-indu	( 358)	2135	483.5 7.1e-136
gi 38257738 sp Q96M20 GD1L1_HUMAN	Ganglioside-indu	( 367)	1264	289.0 2.5e-77
gi 38257686 sp O8VE33 GD1L1_MOUSE	Ganglioside-induc	( 370)	1246	285.0 4.1e-76
gi 29429205 sp Q03520 PCPC_SPHCR	Tetrachloro-P-hyd	( 248)	237	59.9 1.6e-08
gi 32469676 sp P95806 LIND_PSEPA	2,5-dichlorohydro	( 346)	222	56.4 2.5e-07
gi 417094 sp Q03425 GSTZ2_DIACA	Glutathione S-tran	( 145)	182	47.7 4.3e-05
gi 121736 sp P28342 GSTZ1_DIACA	Glutathione S-tran	( 221)	185	48.3 4.4e-05
gi 11133560 sp Q9KSB2 MAAT_VIBCH	Probable maleylac	( 215)	160	42.7 0.0021



## Profiles – the problems

- Could not compare between two profiles – user had to be involved in interpretation
  - Gap penalties – again the user had to be involved
- 

## HMMs

- Hidden Markov Models have been successfully used for
  - speech recognition
  - passive sonar work
  - other “signal detection” problems

## profile-HMMs

- Anders Krogh in David Haussler’s group.
- Takes the “standard” profiles and uses HMM based “standard” mathematics to solve two problems
  - Profile-HMM scores are comparable (\*)
  - Setting gap costs
- Theoretical framework for what we are doing.
- (\* this is not really true. see later)

## Pairwise Alignment

```

RU1A_HUMAN rrm2   VOAGAAR
PABP_DROME rrm3   EAAEAAV
                   | | |
                   +2 +2 +2
  
```

score matrices:  
20x20, 210  
parameters  
position-  
*independent*

Cys	12								
Ser	0 2								
Thr	-2 1 3								
Pro	-1 1 0 6								
Ala	-2 1 1 1 ②								
Gly	-3 1 0 -1 1 5								
Asn	-4 1 0 -1 0 0 2								
Asp	-5 0 0 -1 0 1 2 4								
Glu	-5 0 0 -1 0 0 1 3 4								
Gln	-5 -1 -1 0 0 -1 1 2 2 4								
C	S	T	P	A	G	N	D	E	Q

## Profile Alignment

```

RU1A_HUMAN rrm1   SSATNAL
RU1A_HUMAN rrm2   VOAGAAR   query
SFR1_HUMAN rrm1   RDAEDAV
SXLFF_DROME rrm1  MDSQRAI

PABP_DROME rrm3   EAAEAAV   target
                   +3 +4
                   0
  
```

profile: 20 scores *per column*  
position-*dependent*

## Where pairwise scores come from –

$$\text{score}(AA) = \log \frac{P(A|A)}{f(A)}$$

“probability of A given an A”  
 the observed probability of seeing an A  
 aligned to an A in real alignments

“frequency of A”  
 the expected frequency of A in any sequence

$$\text{Sc}(AA) = \log_2 \frac{0.64}{0.04} = +4$$

$$\text{Sc}(AE) = \log_2 \frac{0.01}{0.04} = -2$$

## Where profile scores (should) come from

$$\text{score}(A|x) = \log \frac{P(A|\text{position } x)}{f(A)}$$

“probability of A at position x”  
 the observed probability of seeing an A  
 in the consensus column x

$$\text{Sc}(A|6) = \log_2 \frac{1.00}{0.04} = +4.6 \quad \text{Sc}(A|5) = \log_2 \frac{0.04}{0.04} = 0$$

$$\text{Sc}(N|6) = \log_2 \frac{0.00}{0.06} = -\text{inf} \quad \text{Sc}(N|5) = \log_2 \frac{0.06}{0.06} = 0$$

1. what about position-specific gap penalties?
2. how to estimate parameters from small numbers of observations?

## Hidden Markov Models (HMMs)

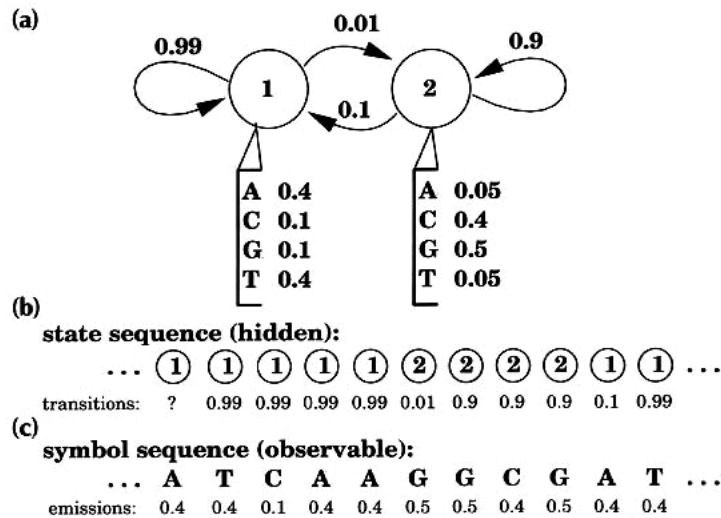
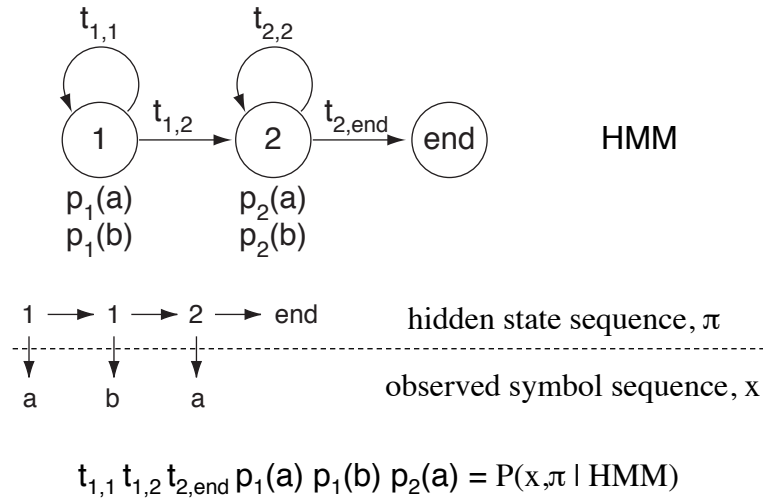
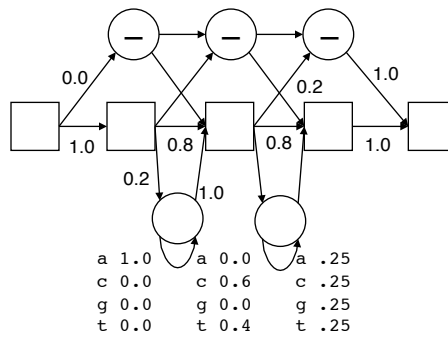


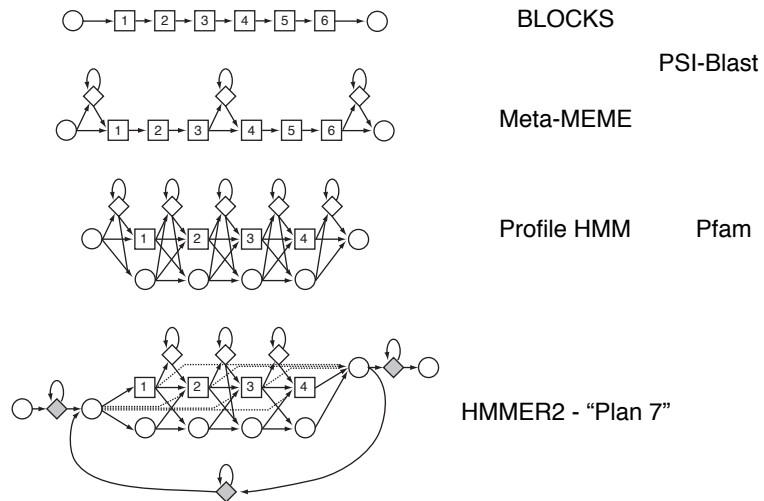
Figure 1 A simple hidden Markov model. A two-state HMM describing DNA sequence with a heterogeneous base composition is shown, following work by Churchill [10]. (a) State 1 (top left) generates AT-rich sequence, and state 2 (top right) generates CG-rich sequence. State transitions and their associated probabilities are indicated by arrows, and symbol emission probabilities for A, C, G and T for each state are indicated below the states. (For clarity, the begin and end states and associated state transitions necessary to model sequences of finite length have been omitted.) (b) This model generates a state sequence as a Markov chain and each state generates a symbol according to its own emission probability distribution (c). The probability of the sequence is the product of the state transitions and the symbol emissions. For a given observed DNA sequence, we are interested in inferring the hidden state sequence that 'generated' it, that is, whether this position is in a CG-rich segment or an AT-rich segment.

### HMM transitions and emissions are probabilities

a - c g  
 a - t a  
 a - c c  
 a t t t  
 a - c -



### Many approaches are HMM-based (or HMM-like)

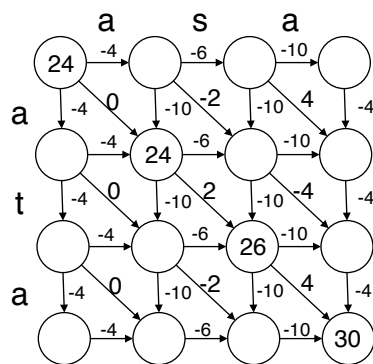


## HMM Algorithms

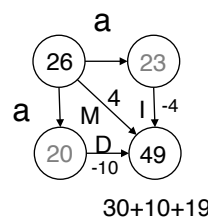
1. The scoring problem:  $P(\text{seq} | \text{model})$   
 "Forward" algorithm (sums over all alignments)
2. The alignment problem:  $\max P(\text{seq}, \text{statepath} | \text{model})$   
 "Viterbi" algorithm
3. The training problem:  
 "Forward-backward" algorithm and  
 Baum-Welch expectation maximization

For profile HMMs, all three algorithms use  $O(MN)$  dynamic programming -- same as "standard" Smith/Waterman and Needleman/Wunsch.

## HMM Alignment



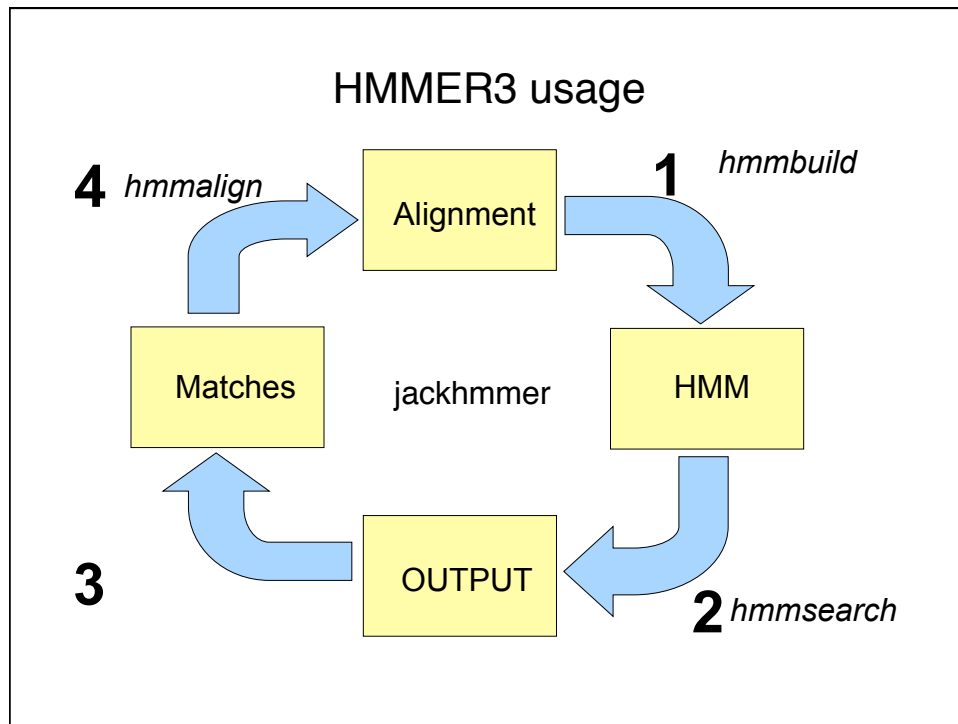
Needleman-Wunsch  
 max log likelihood  
 HMM Viterbi alignment



$$F_j^M(i) = \log \frac{e_{M_j}(x_i)}{q_{x_i}} + \log [a_{M_{j-1}M_j} \exp(F_{j-1}^M(i-1)) + a_{I_{j-1}M_j} \exp(F_{j-1}^I(i-1)) + a_{D_{j-1}M_j} \exp(F_{j-1}^D(i-1))]$$

HMM Forward (score)

$\sum$  probabilities



## 1. Building HMMER models

- Can use HMMER as a black box!
- Usage: *hmmbuild* <hmm> <align>
- Usage: *hmmcalibrate* <hmm> (not necessary for HMMER3)

## 2. HMMER searches

- Search protein database
- Usage: `hmmsearch <hmm> <database>`
  - Option `-A` Use to limit size of output
  - Option `-E` Raise to get more distant matches
- Will take > 20 minutes (hmmsearch v3 as fast as blast)

## 3. HMMER3.0 output

### Header information

```
# hmmsearch :: search profile HMM(s) against a sequence database
# HMMER 3.0b2 (June 2009); http://hmmer.org/
# Copyright (C) 2009 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# -----
# query HMM file:                gstmu.hmm
# target sequence database:      /slib2/blast/pir1.lseg
# -----
```

### 3. HMMER3.0 output

#### Per-sequence scores

```

Query:          gstmu [M=223]
Scores for complete sequences (score includes all domains):
--- full sequence ---   --- best 1 domain ---   -#dom-
  E-value  score  bias  E-value  score  bias  exp  N  Sequence           Description
-----
1.9e-114  380.4  0.2  2.1e-114  380.2  0.1  1.0  1  GSTM2_RAT  Glutathione S-transferase Mu 2 (
7e-82    273.9  1.0  7.7e-82  273.8  0.7  1.0  1  GSTA1_RAT  Glutathione S-transferase alpha-
2.9e-60  203.2  0.0  3.5e-60  202.9  0.0  1.0  1  GSTA4_RAT  Glutathione S-transferase alpha-
1.2e-41  142.3  0.0  1.6e-41  141.9  0.0  1.2  1  GSTP1_RAT  Glutathione S-transferase P (GST
1.1e-38  132.6  0.0  1.5e-38  132.2  0.0  1.2  1  GSTP1_HUMAN  Glutathione S-transferase P (GST
1.3e-13  50.6   0.0  1.7e-13  50.3   0.0  1.2  1  GSTF1_MAIZE  Glutathione S-transferase I (GST
7.4e-09  35.1   0.0  0.00087  18.5   0.0  2.1  2  GSTF3_MAIZE  Glutathione S-transferase III (G
2.5e-05  23.5   0.0  3.6e-05  23.0   0.0  1.2  1  GSTT1_DROME  Glutathione S-transferase 1-1 (G

----- inclusion threshold -----
0.095  11.8  0.0  0.41  9.8  0.0  2.0  1  SSPA_ECO57  Stringent starvation protein Ag
1.7    7.7  0.1  5.7   6.0  0.0  1.8  2  DPOL_EBV   DNA polymerase
2.1    7.4  0.0  4.1   6.5  0.0  1.4  1  SNTD_VIBPA 5'-nucleotidase precursor
2.3    7.3  0.1  4.5   6.4  0.1  1.4  1  Y563_MYCPN Probable GTP-binding protein MG3
3.8    6.6  0.3  17    4.4  0.0  2.2  2  SYEP_HUMAN Bifunctional aminoacyl-tRNA synt
3.8    6.6  0.0  5.6   6.0  0.0  1.2  1  TRPF_YEAST N-(5'-phosphoribosyl)anthranilat
4.2    6.4  0.2  7.6   5.6  0.1  1.3  1  DICA_ECOLI HTH-type transcriptional regulat

```

### 3. HMMER3.0 output

#### Alignments and Domains

```

>> gi|121695|sp|P12653|GSTF1_MAIZE  Glutathione S-transferase I (GST-I) (GST-29) (GST class-phi)
#   score  bias  c-Evalue  i-Evalue  hmmfrom  hmm to  alifrom  ali to  envfrom  env to  acc
---
1 !  50.3  0.0  1.9e-16  1.7e-13  57      203 ..  50      213 ..  15      214 .] 0.76

Alignments for each domain:
== domain 1  score: 50.3 bits; conditional E-value: 1.9e-16

gstmu 57 LdfpnlPylidgkklivgsnaiLryiarkyn...lcGedekirvdlengimdl.....riqLlklCydee 121
      f ++P l dg + s+ai +y ark l + +e ++vd+ ++ +++ +++++
GSTF1_MAIZE 50 NPPGQVPALQDGDLYLFESRAICKYAARKNKpelLREGNLEEAMVDVWIEVEANQytaaInpilfQVLISPLGGTT 127
4699*****9754434555789999*****9986666543344443344444555567 PP

gstmu 122 keklkakyleelkeklklfsvkLgkkdyLvGnkltfvdflllydv.LdrnrildpslldafPnLkalisrfe...aLek 195
      +k + le+l+ l+ ++ L k +yl+G+ l+ +d+ + v L +s+lda+P +ka+ s + +++++
GSTF1_MAIZE 128 DQKVV DENLEKlKkVLEVYEARLTKCKYLAGDFLSLADLNHVSVtLCLFATPYASVLDAYPHVKAWWSGLMerpSVQK 205
7888999*****988875778888899*****9998761224678 PP

gstmu 196 ikaylkss 203
      ++a +k+s
GSTF1_MAIZE 206 VAALMKPS 213
88888876 PP

```

### 3. HMMER3.0 output

#### Alignments and Domains

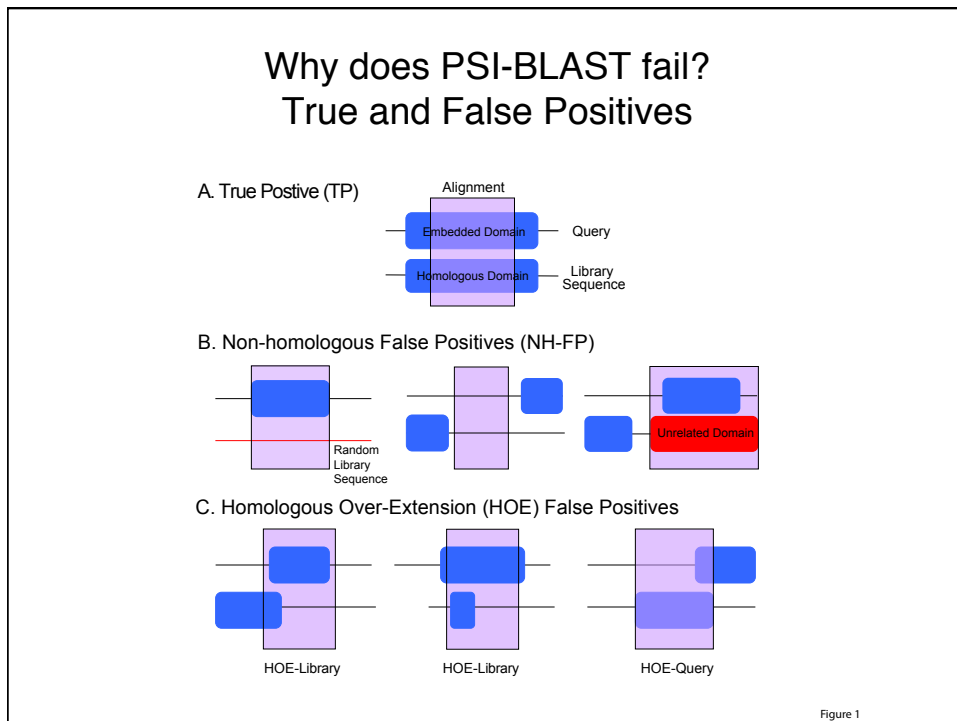
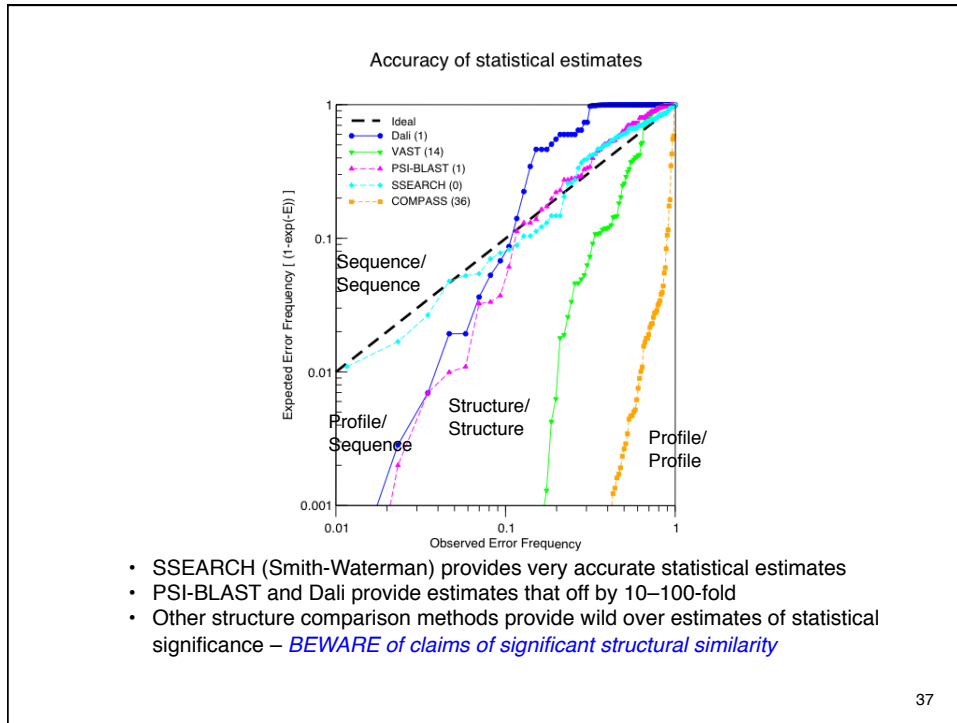
```
>> gi|1170090|sp|P04907|GSTF3_MAIZE Glutathione S-transferase III (GST-III) (GST class-phi)
#   score bias  c-Evalue  i-Evalue hmfrom  hm  to   alifrom  ali  to   envfrom  env  to   acc
-----
 1 !  18.5   0.0   9.8e-07   0.00087   59   91 ..   52   84 ..   26   91 .. 0.91
 2 !  14.8   0.0   1.3e-05   0.012    127  194 ..  134  203 ..  116  219 .. 0.83

Alignments for each domain:
== domain 1   score: 18.5 bits; conditional E-value: 9.8e-07
   gstmu 59 fpnlPylidgkkkivgsnaiLryiarkynlcGe 91
   f ++P l+dg + + s+ai ryia+ky  G
GSTF3_MAIZE 52 FGQIPALVDCDEVLFEFRAINRYIASKYASEGT 84
   99*****87776 PP

== domain 2   score: 14.8 bits; conditional E-value: 1.3e-05
   gstmu 127 akyleelkeklkfskvLgkkdylvGnkltfvdf..llydvLdrnrildpslldafPnLkalisrfeaLe 194
   k e+l + l ++ L +++yl+G+ +t +d  l  L  r  p ++ a P +ka+  + a +
GSTF3_MAIZE 134 EKHAEQLAKVLDVYEAHLARNKYLAGEFTLADanhALLPALTSARPPRPGCVAARPHVKAWWEAIAARP 203
   577899*****99722467789999999999*****9999877766655 PP
```

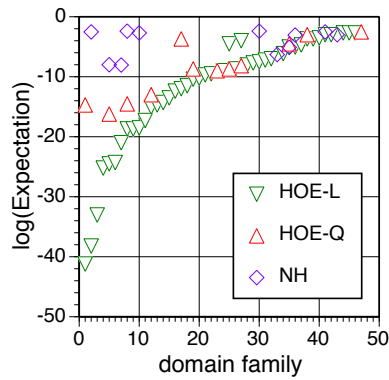
### jackhmmer (3.0) puts it all together

```
# jackhmmer :: iteratively search a protein sequence against a protein database
# HMMER 3.0b2 (June 2009); http://hmmer.org/
# Copyright (C) 2009 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# -----
# query sequence file:           /Users/wrp/Devel/fa35h_svn/seq/gtml_human.aa
# target sequence database:      /slib2/blast/swissprot.lseg
# -----
Query:      gtml_human [L=218]
Description: GLUTATHIONE S-TRANSFERASE MU 1 (EC 2.5.1.18) (GSTM1-1) (HB SUBUNI)
=====
@@ New targets included: 113
@@ Round: 2
@@ Included in MSA: 115 subsequences (query + 114 subseqs from 113 targets)
=====
@@ New targets included: 155
@@ Round: 3
@@ Included in MSA: 279 subsequences (query + 278 subseqs from 268 targets)
=====
@@ New targets included: 100
@@ Round: 4
@@ Included in MSA: 383 subsequences (query + 382 subseqs from 368 targets)
=====
@@ New targets included: 17
@@ Round: 5
@@ Included in MSA: 387 subsequences (query + 386 subseqs from 384 targets)
=====
```

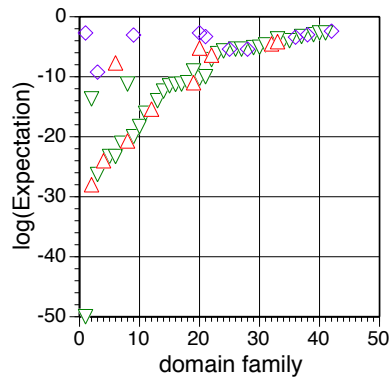


## Why does PSI-BLAST fail? E()-value of highest scoring non-homolog

A. Hard families



B. Sampled families



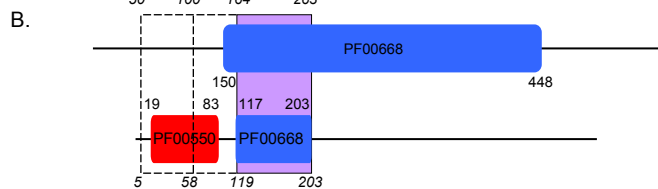
## Why does PSI-BLAST Fail? Homologous Over Extension

A. >Q71EG8\_BACSU|1368305 Length = 203  
Score = 137 bits (346), Expect = 3e-31, Method: Composition-based stats.  
Identities = 25/214 (11%), Positives = 59/214 (27%), Gaps = 17/214 (7%)

```

Query: 52 |NHAPNKGSGFNNTLYKTVALLLALFPVTATRISEFSYQHAYLDNLTTEHAWPY--RQIFLR 110
          + N T K AL + I + ++ ++ ++
Sbjct: 5 |ADRQAYTAPRNVTEMKLCALWEEVLKNGPVGIRDHFFERGGHSLKATALVSRIAKEFGVQ 64
          5th, 4th Iter 3rd Iter
Query: 111 |QWHSGERPYHKICDTSQLYSNGALGMVTAYSACLARVYVNPVT-QAQYYWHEFSLHSEKV 169
          + + + + + + + P A+ +L V
Sbjct: 65 |VPLQDIFARPTVEELASVIQDLEESPYESIQAQKGTGHLPGVFGAETDVCAPTALEDGGV 124
          2nd Iter
Query: 170 |VSTVAHYLDIQGNVDQEAALCKAITMVISETDVLVSRFKREERLEFLQOPNQAAATPQLKF 229
          + L+ G +D+ L + ++ + L F+ + P+Q+ + + QL
Sbjct: 125 |GYNMPAVLELTGPLDRGRLEETFRQLVERHESLRTSFETGPD-GEFVQRIHDSVPFQL-- 181

Query: 230 |IDLQTPDPFNTALQLMRADVESPLNLLTQLLSA 263
          D +A + P L L
Sbjct: 182 |-----DEAESADAFV-----RPFCEEGLPLR 203
    
```



## Preventing extension improves specificity

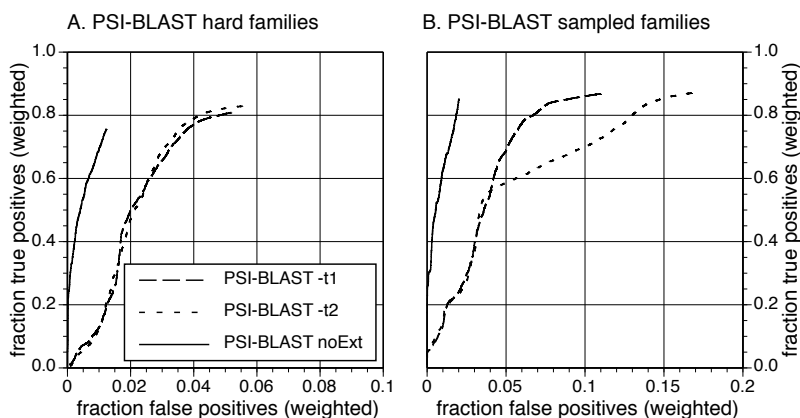


Figure 7

## Why does PSI-BLAST fail?

- Homologous over-extension causes the most initial errors, and the most errors at iteration 5
- Homologous over-extension is an *alignment failure*, not a statistics failure, and applies to any iterative method
- Preventing subsequent extension dramatically improves specificity
- More appropriate alignment matrices can should work even better
- By preventing PSSM corruption through over-extension, look-back time can be improved

Gonzalez and Pearson (2010) Nuc. Acids Res.



## Missing domains, too far from the model

There are 6566 sequences with the following architecture: GST\_N, GST\_C

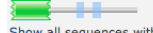
[DCMA\\_METS1](#) [Methylophilus sp. (strain DM11)] Dichloromethane dehalogenase EC=4.5.1.3 (267 residues)



[Show all sequences with this architecture.](#)

There are 1741 sequences with the following architecture: GST\_N

[GSTX2\\_MAIZE](#) [Zea mays (Maize)] Probable glutathione S-transferase B22 EC=2.5.1.18 (236 residues)



[Show all sequences with this architecture.](#)

The best scores are:

sp	id	len	s-w	bits	E(437847)	%_id	%_sim	alen
sp P50472.1	GSTX2_MAIZE (236)	1371	338.8	2.4e-92	1.000	1.000	236	
sp P32110.1	GSTX6_SOYBN (225)	437	112.3	3.6e-24	0.367	0.664	226	
sp Q06398.2	GSTU6_ORYSJ (236)	414	106.7	1.8e-22	0.449	0.705	227	
sp P49332.1	GSTXC_TOBAC (221)	363	94.4	8.9e-19	0.356	0.616	219	
sp P46417.1	GSTX3_SOYBN (219)	324	84.9	6.2e-16	0.330	0.606	218	
sp P32111.1	GSTX1_SOLTU (217)	322	84.4	8.6e-16	0.308	0.608	227	
sp P46421.1	GSTXA_ARATH (224)	318	83.4	1.8e-15	0.311	0.605	228	
sp Q03662.1	GSTX1_TOBAC (223)	304	80.1	1.8e-14	0.318	0.583	223	
sp P50471.1	GSTX1_NICPL (219)	284	75.2	5.2e-13	0.338	0.572	222	
sp P78417.2	GSTO1_HUMAN (241)	222	60.1	1.9e-08	0.273	0.567	231	
sp A2XMN2.1	GSTU1_ORYS (231)	201	55.1	6.3e-07	0.351	0.596	225	
sp Q9N1F5.2	GSTO1_PIG (241)	194	53.4	2.2e-06	0.278	0.565	209	
sp Q8K2Q2.1	GSTO2_MOUSE (248)	194	53.4	2.2e-06	0.279	0.528	233	
sp P30347.1	LIGF_PSEPA (257)	192	52.9	3.2e-06	0.351	0.605	114*	
sp O09131.2	GSTO1_MOUSE (240)	191	52.6	3.6e-06	0.261	0.526	234	
sp P46420.2	GSTF4_MAIZE (223)	187	51.7	6.4e-06	0.309	0.541	220	
sp P57109.1	MAAI_PSEAE (212)	181	50.2	1.7e-05	0.301	0.551	216	
sp Q6AXV9.1	GSTO2_RAT (248)	181	50.2	2e-05	0.271	0.543	210	
sp P46423.1	GSTF_HYOMU (212)	171	47.8	8.9e-05	0.258	0.524	225	

45

## Profiles, PSSMs, and PSI-BLAST - Summary

- Protein divergence is not uniform over a protein - some parts are more conserved than others
- Position specific scoring matrices can capture the specific patterns of conservation at different sites in a protein
- PSI-BLAST combines searching, multiple alignment, and PSSMs
- Statistical estimates are difficult with PSSMs, use PSI-SEARCH and PSI-PRSS
- Iterative PSSM/HMM searches may be contaminated by Homologous Overextension
- Single models cannot capture diverse families (PFAM Clans)