

# Sequence Similarity

## Protein Sequence Comparison and Protein Evolution

(What BLAST does/Why BLAST works)

William R. Pearson

[www.people.virginia.edu/~wrp](http://www.people.virginia.edu/~wrp)  
[wrp@virginia.edu](mailto:wrp@virginia.edu)

1

## *Sequence Similarity - Conclusions*

- Homologous sequences share a common ancestor, but most sequences are non-homologous
- Always compare Protein Sequences
- Sequence Homology can be reliably inferred from statistically significant similarity (non-homology cannot from non-similarity)
- Homologous proteins share common structures, but not necessarily common functions
- Sequence statistical significance estimates are accurate (verify this yourself)  $10^{-6} < E() < 10^{-3}$  is statistically significant
- Scoring matrices set evolutionary look back horizons - not every discovery is distant
- PSI-BLAST can be more sensitive, but with lower statistical accuracy

2

*Establishing homology from  
statistically significant similarity*

**Why BLAST works**

- 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

3

**This talk is not about:**

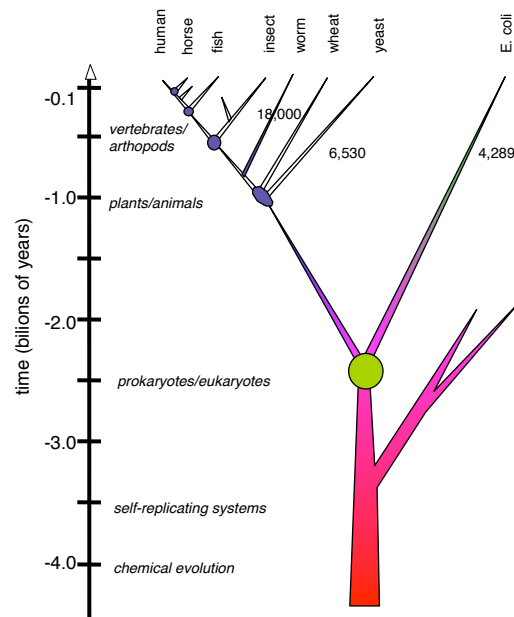
- *Alignment*
  - Alignment quality may be more sensitive to parameter choice
  - Multiple sequences for biologically accurate alignments
- *Inferring Protein Function*
  - Homology (common ancestry) implies common structure (guaranteed), not necessarily common function
  - Homologs have different functions
  - Non-homologs have similar (or identical) functions
- *The best sequences for building trees*
  - Protein sequences are clearly best for establishing homology, but DNA sequences may be better for resolving recent divergence

4

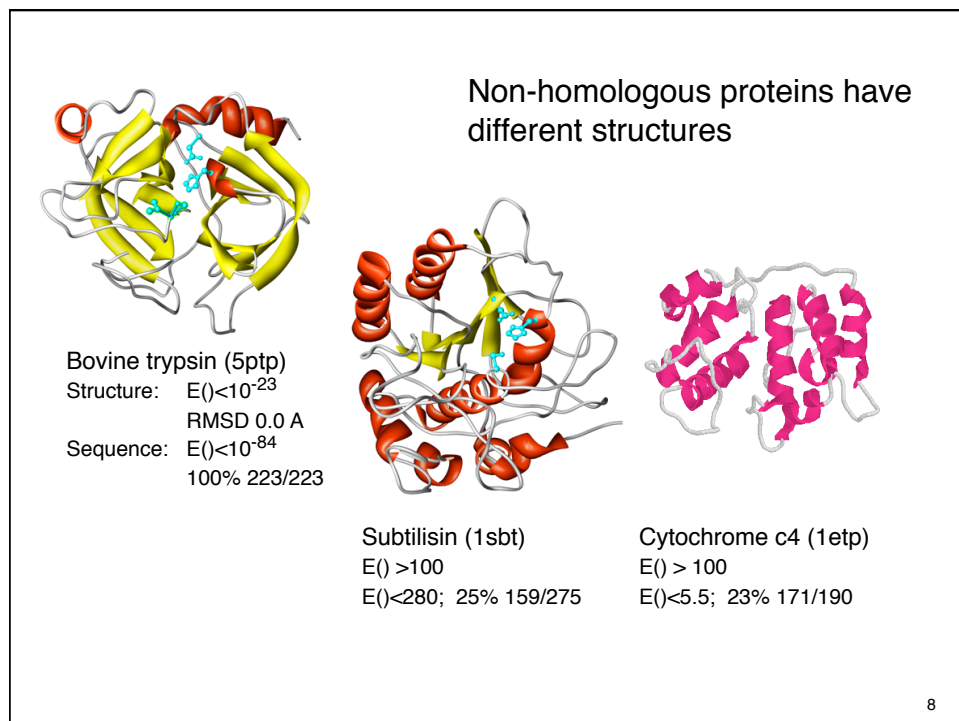
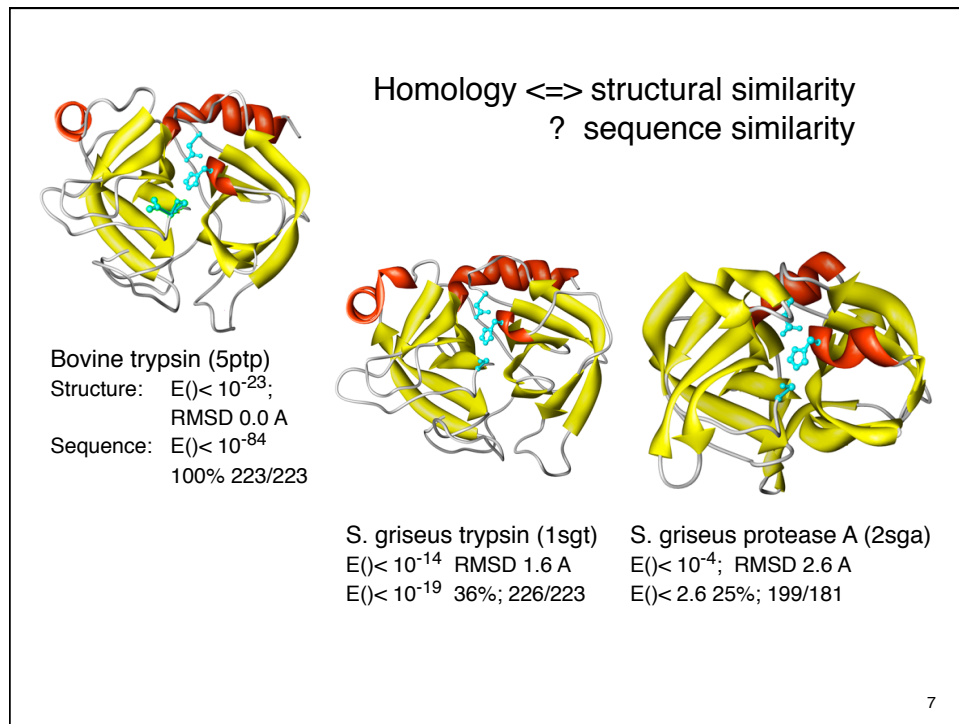
## Protein Evolution and Sequence Similarity

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- Alignment Algorithms/Local sequence alignments
- Similarity scoring matrices
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Improving sensitivity with PSI-BLAST

5



6

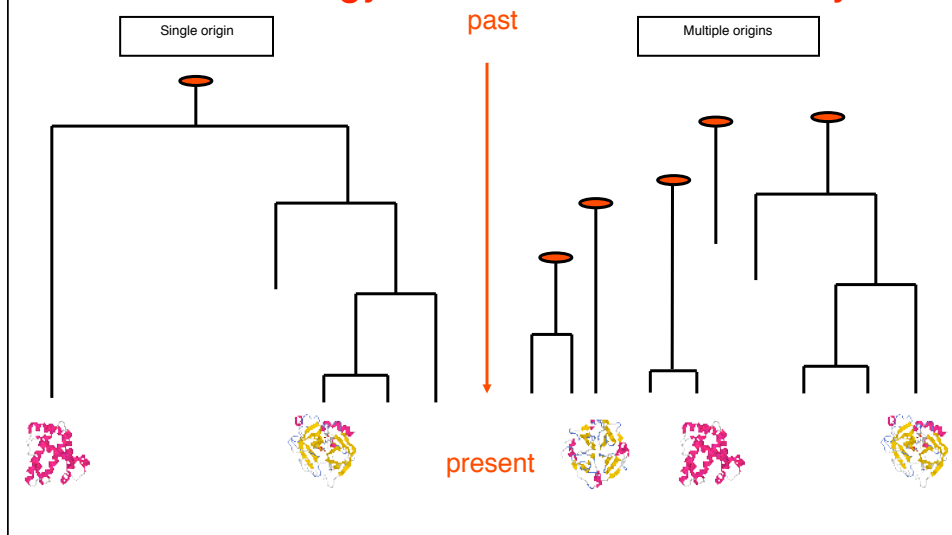


## Homology is confusing I: Homology defined Three(?) Ways

- Proteins/genes/DNA that share a common ancestor
- Specific positions/columns in a multiple sequence alignment that have a 1:1 relationship over evolutionary history
  - sequences are *50% homologous* ???
- Specific (morphological/functional) characters that share a recent divergence (clade)
  - bird/bat/butterfly wings are/are not homologous

9

## Homology is confusing II: Are All Sequences Homologous? **No Homology without excess similarity**



## Homology defined *My Way*

- Sequences are inferred to share a common ancestor based on statistically significant *excess* similarity. Any evidence of *excess* similarity can be used to infer homology
- Lack of evidence *cannot* be used to infer non-homology.
  - Proteins with different structures are non-homologous
- There are always two alternative hypotheses: homology (common ancestry), or convergence – one must weigh the evidence for each hypothesis (convergence is the *null* hypothesis).

11

## What BLAST does:

Similarity      ?      Homology  
                          <=>

## Why BLAST works:

Statistical      ?      Biological  
 Significance      <=>      Significance

Divergence      ?      Convergence

12

### Some important dates in history

Origin of the universe	-13.7 <sup>a</sup>
Formation of the solar system	-4.6 ±0.4
First self-replicating system	-3.5 ±0.5
Prokaryotic-eukaryotic divergence	-2.5 ±0.3
Plant-animal divergence	-1.0
Invertebrate-vertebrate divergence	-0.5
Mammalian radiation beginning	-0.1

<sup>a</sup>Billions of years ago

Protein Family	PAMs <sup>a</sup> /100 res. /10 <sup>8</sup> years	Protein Lookback time <sup>b</sup>	
Pseudogenes	400	45 <sup>c</sup>	Primates, Rodents
Fibrinopeptides	90	200	Mammalian Radiation
Lactalbumins	27	670	Vertebrates
Ribonucleases	21	850	Animals
Hemoglobins	12	1.5 <sup>d</sup>	Plants/Animals
Acid Proteases	8	2.3	Prokaryotic/Eukaryotic
Triphosphate isomerase	3	6	Archaea
Glutamate dehydrogenase	1	18	?

<sup>a</sup>PAMs, point accepted mutations. <sup>b</sup>Useful lookback time, 360 PAMs, 15% identity.

<sup>c</sup>Millions of years. <sup>d</sup>Billions of years.

13

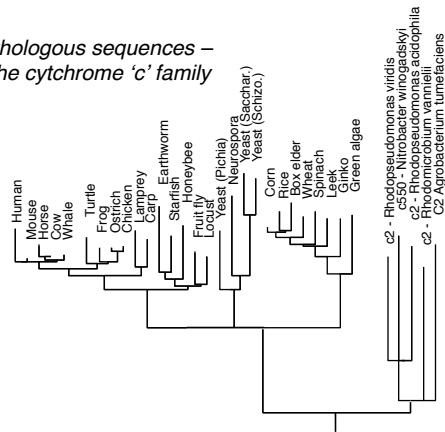
### E. coli proteins vs Human – Ancient Protein Domains

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 hydratase 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
9.2e-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

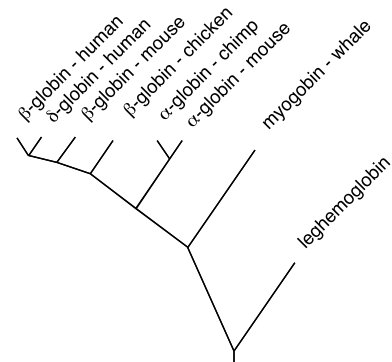
14

## Orthologs and Paralogs – Inferring Function

### Orthologous sequences – the cytochrome 'c' family

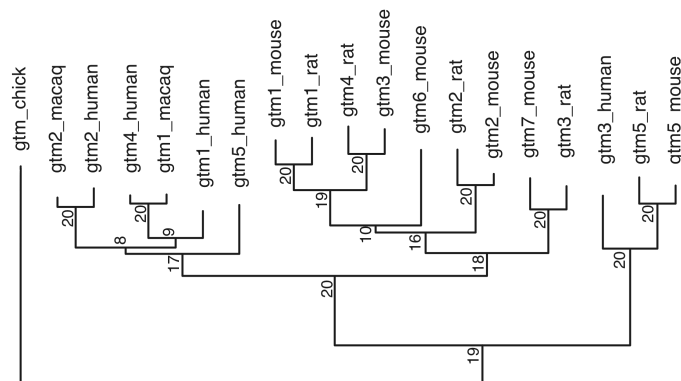


*Paralogous genes – globins*



15

## Orthology can be difficult to infer



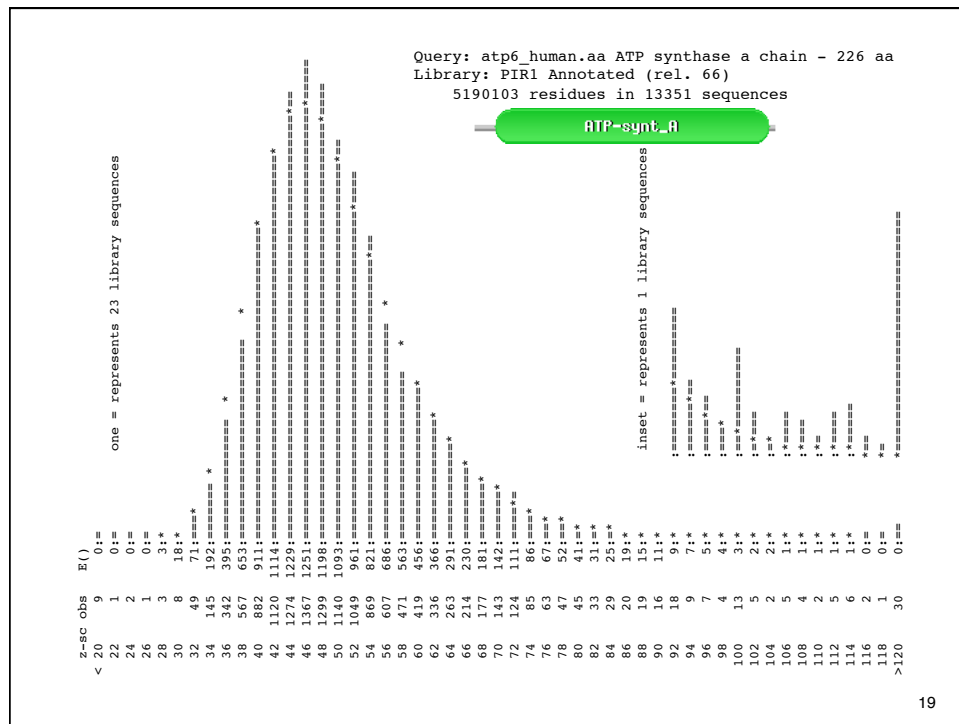


## Orthologs preserve function, but can be difficult to infer

- 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

## Protein Evolution and Sequence Similarity

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- Alignment Algorithms/Local sequence alignments
- Similarity scoring matrices
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Improving sensitivity with PSI-BLAST



## 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

11

## The PAM250 matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
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										
His	-3	-1	-1	0	-1	2	2	1	1	3	6									
Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	-4	-4	-4	-2	-1	-1	-2	7	10	
Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	1

23

## Where do scoring matrices come from?

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

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

## 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

24

```

>>sp|P30391|ATPI_EUGGR Chloroplast ATP synthase a chain precursor (251 aa)
s-w opt: 123 Z-score: 151.3 bits: 35.4 E(): 0.017
Smith-Waterman score: 123; 25.7% identity (57.9% similar) in 214 aa overlap (21-222:50-243)

          10          20          30          40          50          60
human      MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLNNRLITTQQWLKLTQKQMMTM
          .:.: : : : : : : : : : : : : : : : : : : : : : : :
Euglena    VNMFISGIFQIANVEVGQHFYWSILGFQIHGQVLNSWIVILIIGF--LSIYTTKNL--TLVPANKQIFIELVTEFITDI
          10          20          30          40          50          60          70          80

          70          80          90          100         110         120
human      HNTK-GRT---WSLMLVSLIIFIATTNLLG-LLPHSFT--PTTQL---SMNLAMAIPLWAGTVIMGFRSKI-KNALAHF
          .:. :. : : : : : : : : : : : : : : : : : : : : : : :
Euglena     SKTQIGKEYSKWVPYIGTMFLFIFVSNWSGALIPWKIILPNGLGAPTNDINTTAGLAILTSLAYFYAGLNKKGLTYF
          90          100         110         120         130         140         150         160

          130         140         150         160         170         180         190         200
Human       LPQGTPTPLIPMLVIEIETISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVAL
          .:.: . :. : . : : : : : : : : : : : : : : : : : : : : :
Euglena     KKYVQPTPILLPINILEDFT---KPLSLSFRLFGNILADELVVAVLVSL-----VP--LIVPVPLIFLGLF---TSG
          170         180         190         200         210         220

          210         220
human       IQAYVFTLLVSLYLHDNT
          :. :. : . :.
Euglena     IQALIFATLSGSYIGEAMEGHH
          230         240         250

```

25

```

Query: atp6_human.aa ATP synthase a chain - 226 aa
Library: 5190103 residues in 13351 sequences

The best scores are:
  ( len)  s-w bits E(13351) %_id %_sim alen
sp|P00846|ATP6_HUMAN ATP synthase a chain (AT ( 226) 1400 325.8 5.8e-90 1.000 1.000 226
sp|P00847|ATP6_BOVIN ATP synthase a chain (AT ( 226) 1157 270.5 2.5e-73 0.779 0.951 226
sp|P00848|ATP6_MOUSE ATP synthase a chain (AT ( 226) 1118 261.7 1.2e-70 0.757 0.916 226
sp|P00849|ATP6_XENLA ATP synthase a chain (AT ( 226) 745 176.8 4.0e-45 0.533 0.847 229
sp|P00851|ATP6_DROYA ATP synthase a chain (AT ( 224) 473 115.0 1.7e-26 0.378 0.721 222
sp|P00854|ATP6_YEAST ATP synthase a chain pre ( 259) 428 104.7 2.3e-23 0.353 0.694 232
sp|P00852|ATP6_EMENI ATP synthase a chain pre ( 256) 365 90.4 4.8e-19 0.304 0.691 230
sp|P14862|ATP6_COCHE ATP synthase a chain (AT ( 257) 353 87.7 3.2e-18 0.313 0.650 214
sp|P68526|ATP6_TRITI ATP synthase a chain (AT ( 386) 309 77.6 5.1e-15 0.289 0.651 235
sp|P05499|ATP6_TOBAC ATP synthase a chain (AT ( 395) 309 77.6 5.2e-15 0.283 0.635 233
sp|P07925|ATP6_MAIZE ATP synthase a chain (AT ( 291) 283 71.7 2.3e-13 0.311 0.667 180
sp|P0AB98|ATP6_ECOLI ATP synthase a chain (AT ( 271) 178 47.9 3.2e-06 0.233 0.585 236
sp|P0C2Y5|ATPI_ORYSA Chloroplast ATP synth (A ( 247) 144 40.1 0.00062 0.242 0.580 231
sp|P06452|ATPI_PEA Chloroplast ATP synthase a ( 247) 143 39.9 0.00072 0.250 0.586 232
sp|P27178|ATP6_SYNY3 ATP synthase a chain (AT ( 276) 142 39.7 0.00095 0.265 0.571 170
sp|P06451|ATPI_SPIOL Chloroplast ATP synthase ( 247) 138 38.8 0.0016 0.242 0.580 231
sp|P08444|ATP6_SYN6 ATP synthase a chain (AT ( 261) 127 36.3 0.0095 0.263 0.557 167
sp|P69371|ATPI_ATRBE Chloroplast ATP synthase ( 247) 126 36.0 0.01 0.221 0.571 231
sp|P06289|ATPI_MARPO Chloroplast ATP synthase ( 248) 126 36.0 0.011 0.240 0.575 167
sp|P30391|ATPI_EUGGR Chloroplast ATP synthase ( 251) 123 35.4 0.017 0.257 0.579 214

sp|P19568|TLCA_RICPR ADP,ATP carrier protein ( 498) 122 35.0 0.043 0.243 0.579 152

sp|P24966|CYB_TAYTA Cytochrome b ( 379) 113 33.0 0.13 0.234 0.532 158
sp|P03892|NU2M_BOVIN NADH-ubiquinone oxidored ( 347) 107 31.7 0.31 0.261 0.479 211
sp|P68092|CYB_STEAT Cytochrome b ( 379) 104 31.0 0.54 0.277 0.547 137
sp|P03891|NU2M_HUMAN NADH-ubiquinone oxidored ( 347) 103 30.8 0.58 0.201 0.537 149
sp|P00156|CYB_HUMAN Cytochrome b ( 380) 102 30.5 0.74 0.268 0.585 205
sp|P15993|AROP_ECOLI Aromatic amino acid tr ( 457) 103 30.7 0.78 0.234 0.622 111
sp|P24965|CYB_TRANA Cytochrome b ( 379) 101 30.3 0.87 0.234 0.563 158
sp|P29631|CYB_POMTE Cytochrome b ( 308) 99 29.9 0.95 0.274 0.584 113
sp|P24953|CYB_CAPHI Cytochrome b ( 379) 99 29.8 1.2 0.236 0.564 140

```

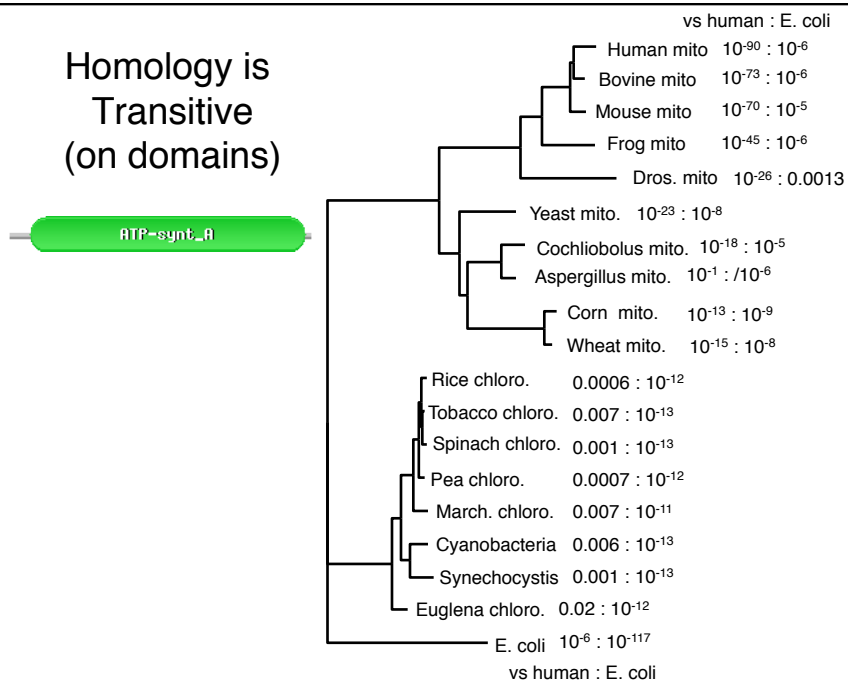
26

Query: atp6\_ecoli.aa ATP synthase a - 271 aa  
Library: 5190103 residues in 13351 sequences

The best scores are:			( len)	s-w	bits	E(13351)	%_id	%_sim	alen
sp P0AB98	ATP6_ECOLI	ATP synthase a chain (AT	( 271)	1774	416.8	3.e-117	1.000	1.000	271
sp P06451	ATPI_SPIOL	Chloroplast ATP synthase	( 247)	274	70.4	5.8e-13	0.270	0.616	211
sp P69371	ATPI_ATRBE	Chloroplast ATP synthase	( 247)	271	69.7	9.3e-13	0.270	0.607	211
sp P08444	ATP6_SYNP6	ATP synthase a chain (AT	( 261)	271	69.7	9.9e-13	0.267	0.600	240
sp P06452	ATPI_PEA	Chloroplast ATP synthase a	( 247)	266	68.5	2.1e-12	0.274	0.614	223
sp P30391	ATPI_EUGGR	Chloroplast ATP synthase	( 251)	265	68.3	2.5e-12	0.298	0.596	225
sp P0C2Y5	ATPI_ORYSA	Chloroplast ATP synthase	( 247)	260	67.2	5.4e-12	0.259	0.603	239
sp P27178	ATP6_SYNY3	ATP synthase a chain (AT	( 276)	260	67.1	6.1e-12	0.264	0.578	258
sp P06289	ATPI_MARPO	Chloroplast ATP synthase	( 248)	250	64.8	2.7e-11	0.261	0.621	211
sp P07925	ATP6_MAIZE	ATP synthase a chain (AT	( 291)	215	56.7	8.7e-09	0.259	0.578	232
sp P68526	ATP6_TRITI	ATP synthase a chain (AT	( 386)	209	55.3	3.1e-08	0.259	0.603	239
sp P00854	ATP6_YEAST	ATP synthase a chain pre	( 259)	204	54.2	4.5e-08	0.235	0.578	277
sp P05499	ATP6_TOBAC	ATP synthase a chain (AT	( 395)	189	50.7	7.8e-07	0.220	0.582	268
sp P00846	ATP6_HUMAN	ATP synthase a chain (AT	( 226)	178	48.2	2.5e-06	0.237	0.589	236
sp P00852	ATP6_EMENI	ATP synthase a chain pre	( 256)	178	48.2	2.8e-06	0.209	0.590	244
sp P00849	ATP6_XENLA	ATP synthase a chain (AT	( 226)	173	47.1	5.5e-06	0.261	0.630	165
sp P00847	ATP6_BOVIN	ATP synthase a chain (AT	( 226)	172	46.8	6.5e-06	0.233	0.581	236
sp P14862	ATP6_COCHE	ATP synthase a chain (AT	( 257)	171	46.6	8.7e-06	0.204	0.608	265
sp P00848	ATP6_MOUSE	ATP synthase a chain (AT	( 226)	166	45.5	1.7e-05	0.259	0.617	193
sp P00851	ATP6_DROYA	ATP synthase a chain (AT	( 224)	139	39.2	0.0013	0.225	0.549	253
sp P24962	CYB_STELO	Cytochrome b	( 379)	125	35.9	0.021	0.223	0.575	193
sp P09716	US17_HCMVA	Hypothetical protein HVL	( 293)	109	32.3	0.21	0.260	0.565	131
sp P68092	CYB_STEAT	Cytochrome b	( 379)	109	32.2	0.27	0.211	0.562	194
sp P24960	CYB_ODOHE	Cytochrome b	( 379)	104	31.1	0.61	0.210	0.555	200
sp P03887	NU1M_BOVIN	NADH-ubiquinone oxidored	( 318)	98	29.7	1.3	0.287	0.545	167
sp P24992	CYB_ANTAM	Cytochrome b	( 379)	99	29.9	1.4	0.192	0.565	193

27

## Homology is Transitive (on domains)



28

## Homology and Domains – Histone deacetylase PCAF

The best scores are:		s-w bits	E(362341)	%_id	%_sim	alen
PCAF_HUMAN	Histone acetyltransferase PCAF;	( 832)	4876 1092	0	1.000 1.000	832
PCAF_MOUSE	Histone acetyltransferase PCAF;	( 813)	4507 1010	0	0.929 0.974	817
GCNL2_HUMAN	General control of amino acid synthesis protein 5-l	( 837)	3535 793.	0	0.716 0.864	821
GCN5_YEAST	Histone acetyltransferase GCN5	( 439)	1049 240. 5.2e-62	0.469	0.743	354
GCN5_ARATH	Histone acetyltransferase GCN5; AtGCN5	( 568)	956 219. 1.2e-55	0.435	0.733	375
BPTF_HUMAN	Nucleosome-remodeling factor subunit BPTF	(3046)	369 88.3 2.4e-15	0.495	0.773	97
NU301_DROME	Nucleosome-remodeling factor subunit NURF301	(2669)	359 86.2 9.3e-15	0.511	0.787	94
CECR2_HUMAN	Cat eye syndrome critical region protein 2	(1484)	306 74.6 1.6e-11	0.371	0.771	105
BRD4_HUMAN	Bromodomain-containing protein 4; HUNK1 protein	(1362)	288 70.6 2.3e-10	0.379	0.681	116
BRDT_MACFA	Bromodomain testis-specific protein	( 947)	270 66.7 2.3e-09	0.353	0.690	116
FSH_DROME	Homeotic protein female sterile; Fragile-chorion memb	(2038)	276 67.8 2.4e-09	0.341	0.651	129
BRDT_HUMAN	Bromodomain testis-specific protein; RING3-like prot	( 947)	266 65.9 4.3e-09	0.345	0.690	116
Y0777_DICDI	Bromodomain-containing protein DDB_G0280777	(1823)	260 64.3 2.5e-08	0.385	0.725	109
BRDT_MOUSE	Bromodomain testis-specific protein; RING3-like prot	( 956)	247 61.6 8.1e-08	0.328	0.647	116
BAZ2B_HUMAN	Bromodomain adjacent to zinc finger domain protein	(1972)	247 61.3 2e-07	0.343	0.695	105
TAF1_DROME	Transcription initiation factor TFIID subunit 1; Tra	(2129)	230 57.5 3.1e-06	0.349	0.689	106
82_SCHPO	Bromodomain-containing protein C631.02	( 727)	217 55.0 5.9e-06	0.320	0.587	172
BRD9_XENLA	Bromodomain-containing protein 9	( 527)	214 54.5 6.2e-06	0.292	0.579	171
GTE6_ARATH	Transcription factor GTE6; Protein GENERAL TRANSCRIP	( 369)	201 51.7 2.9e-05	0.290	0.601	183
BAZ1B_MOUSE	Bromodomain adjacent to zinc finger domain protein	(1479)	212 53.7 3.1e-05	0.302	0.583	139
K2_SCHPO	Bromodomain-containing protein C1450.02	( 578)	204 52.2 3.3e-05	0.310	0.628	113
TAF1_HUMAN	Transcription initiation factor TFIID subunit 1; Tra	(1872)	212 53.6 4.2e-05	0.339	0.678	115
BAZ1B_HUMAN	Bromodomain adjacent to zinc finger domain protein	(1483)	209 53.0 5e-05	0.397	0.705	78
TIF1A_HUMAN	Transcription intermediary factor 1-alpha; TIF1-al	(1050)	206 52.5 5.1e-05	0.384	0.698	86
BDF2_YEAST	Bromodomain-containing factor 2	( 638)	200 51.3 6.9e-05	0.304	0.607	168

29

## Homology and Domains – Histone deacetylase PCAF

The best scores are: E(362341) alen

PCAF\_HUMAN Histone acetyl ( 832) 0 832



GCN5\_YEAST Histone acetyl ( 439) 5.2e-62 354



BPTF\_HUMAN Nucleosome-rem (3046) 2.4e-15 97



CECR2\_HUMAN Cat eye syndr (1484) 1.6e-11 105



GTE6\_ARATH Transcription ( 369) 2.9e-05 183



30

## Protein Evolution and Sequence Similarity

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- Alignment Algorithms/Local sequence alignments
- Similarity scoring matrices
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Improving sensitivity with PSI-BLAST

31

### *DNA vs protein sequence 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
MOTGLUSTRA	Manduca sexta GST	5.7e-07	2.5e-30	8.0e-25
RLGSTARGN	R.leguminosarum gsta	0.0029	3.2e-13	1.4e-10
HUMGSTT2A	H. sapiens GSTT2	0.32	3.3e-10	2.0e-09
HSGSTT1	H.sapiens GSTT1 mRNA	7.2	8.4e-13	3.6e-10
ECAE000319	E. coli hypothet. prot.	—	4.7e-10	1.1e-09
MYMDCMA	Methyl. dichlorometh. DH	—	1.1e-09	6.9e-07
BCU19883	Burkholderia maleylacetate red.	—	1.2e-09	1.1e-08
NFU43126	Naegleria fowleri GST	—	3.2e-07	0.0056
SP505GST	Sphingomonas paucim	—	1.8e-06	0.0002
EN1838	H. sapiens maleylaceto. iso.	—	2.1e-06	5.9e-06
HSU86529	Human GSTZ1	—	3.0e-06	8.0e-06
SYCCPNC	Synechocystis GST	—	1.2e-05	9.5e-06
HSEF1GMR	H.sapiens EF1g mRNA	—	9.0e-05	0.00065

32



Table 3: DNA and translated DNA similarity searches

Taxonomic Group	blastx	blastn	blastn	
	+3/-3	+1/-3		
Bacteria eubacteria				
. Proteobacteria proteobacteria				
. . Gammaproteobacteria g-proteo.				
. . . Enterobacteriaceae entero.				
. . . . Shigella enterobacteria				
. . . . . Shigella flexneri2a	979	2165	2595	enterobacteria
. . . . Escherichia coli CFT073	976	2130	2508	enterobacteria
. . . . Escherichia coli O157:H7	959	2184	2642	enterobacteria
. . . . Escherichia coli	758	2253	2817	enterobacteria
. . . . Edwardsiella tarda	784	1102	180	enterobacteria
. . Brucella melitensis 16M	496	854	113	a-proteobacter
. . Mesorhizobium loti	60			a-proteobacter
. . Bordetella bronchiseptica RB	330	217		b-proteobacter
. . Geobacter metallireducens ..	53			d-proteobacter
. . Geobacter sulfurreducens PCA	53			d-proteobacter
. Prochlorococcus marinus MIT	517	458		cyanobacteria
. Synechocystis sp. PCC 6803 ...	466	284		cyanobacteria
. Clostridium perfringens str. 13	427			eubacteria
. Streptomyces coelicolor A3(2).	417			high GC Gram+
. Mycobacterium tuberculosis ...	414	311		high GC Gram+
. Listeria innocua .....	414	257		eubacteria
. Listeria monocytogenes .....	414	234		eubacteria
. Enterococcus faecium .....	411			eubacteria
. Streptomyces avermitilis MA4680	409			high GC Gram+
. Lactococcus lactis .....	405	183		eubacteria
. Lactobacillus plantarum WCF51.	390	231		eubacteria
. Bacteroides thetaiotaomicronVPI	387	233		CFB group bact
. Chloroflexus aurantiacus .....	72			GNS bacteria
. Gloeobacter violaceus PCC 7421	48			cyanobacteria
. Streptomyces viridifaciens ...	45			high GC Gram+
. Clostridium tetani E88 .....	45			eubacteria

Bit scores from a blastx and blastn searches presented using the BLAST taxonomy summary option. The DNA sequence (M84025) encoding *E. coli* glutamate decarboxylase used to search the bacterial division of Genbank or Genpept. Species that contain a homolog with a bit score  $\geq 45$  ( $E() < 10^{-3}$  for blastx) are shown. The numbers under the blastx and blastn columns indicate the highest bit-score obtained for that taxonomic group.

33

## Protein Evolution and Sequence Similarity

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- Local sequence alignments
- Similarity scoring matrices
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Improving sensitivity with PSI-BLAST

34

## Local alignments - calmodulin

```

46.1% identity in 76 aa overlap (1-76:77-149); score: 222 E(10000): 2.7e-10
      10      20      30      40      50      60
mchu  MADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTAEALQDMINEVDADG
      : : .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .:
mchu  MKDTSDEEEI---REAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDG
      80      90      100      110      120      130

      70
mchu  NGTIDFPEFLTMMARK
      .: .: .: .: .: .:
mchu  DGQVNYEEFVQMMTAK
      140

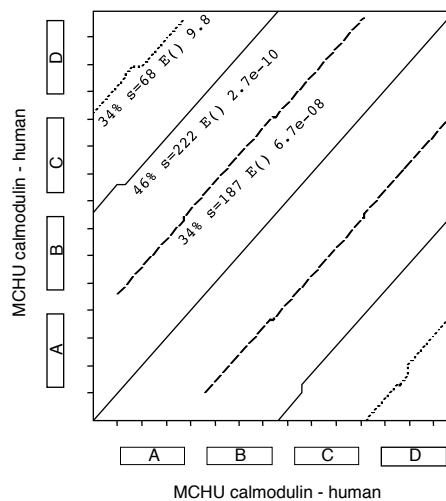
34.3% identity in 105 aa overlap (11-111:47-147); score: 187 E(10000): 6.7e-08
      20      30      40      50      60
mchu  AEFKEAFSLFDKDGDTITTKELGTVM-RSLGQNPTAEALQDMINEVDADGNGTIDFPEF
      : . . . . : : .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .:
mchu  AELQDMINEVDADGNGTIDFPEFLTMMARKMKDTSDEEEIREAFRVFDKDGNGYISAAEL
      50      60      70      80      90      100
      70      80      90      100      110
mchu  ---LTMMARKMKDTSDEEEIREAFRVFDKDGNGYISAAELRHVMT
      .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .:
mchu  RHVMTNLGEKLTDEEVDEMIREA---DIDGDGQVNYEEFVQMMT
      110      120      130      140

34.2% identity in 38 aa overlap (1-37:113-146); score: 68 E(10000): 9.8
      10      20      30
mchu  MADQLTEEQIAEF-KEAFSLFDKDGDTITTKELGTVM
      .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .: .:
mchu  LGEKLTDEEVDEMIREA---DIDGDGQVNYEEFVQMM
      120      130      140

```

35

## Repeated domains with local alignments



36

## Protein Evolution and Sequence Similarity

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- Alignment Algorithms/Local sequence alignments
- **Similarity scoring matrices**
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Improving sensitivity with PSI-BLAST

37

## More about scoring matrices ...

### PAM series:

- 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 series

- 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

38

## Where do scoring matrices come from?

### 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

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

$$p_R = 0.051$$

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

$$p_N = 0.043$$

$$\lambda_2 S_{ij} = \lg_2 (q_{ij}/p_i p_j) \quad \lambda_e S_{ij} = \ln(q_{ij}/p_i p_j) \quad p_R p_N = 0.002193$$

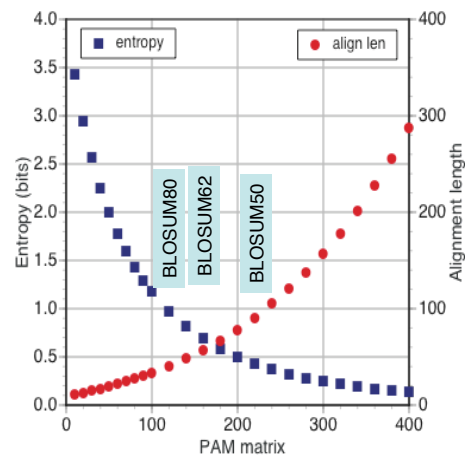
$$\lambda_2 S_{R:N(40)} = \lg_2 (0.000435/0.00219) = -2.333$$

$$\lambda_2 = 1/3; S_{R:N(40)} = -2.333/\lambda_2 = -7$$

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

39

## PAM matrices and alignment length

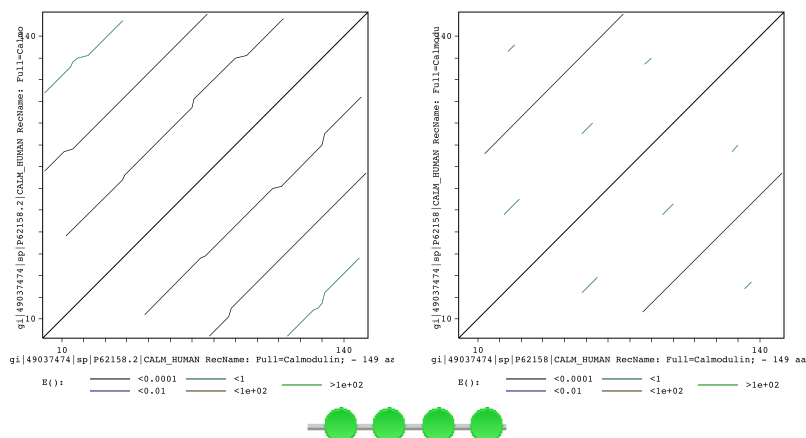


40

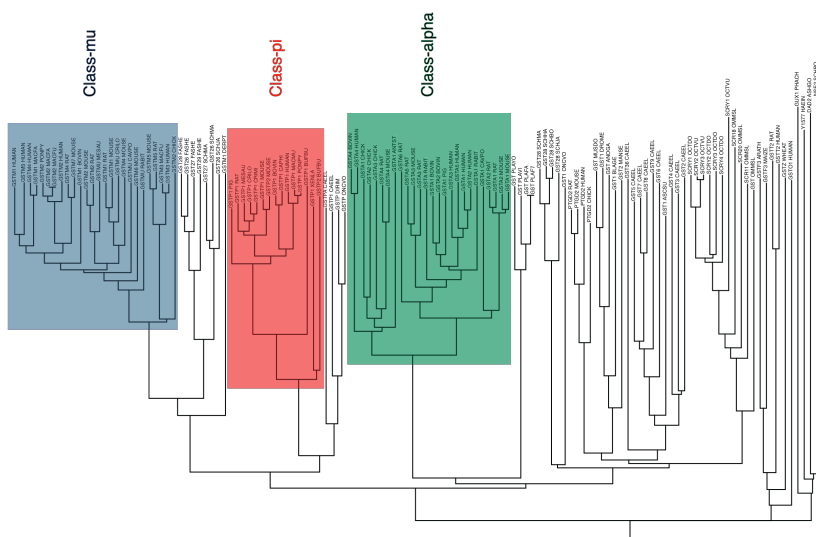
## Scoring matrices set look back time

BLOSUM50 -10/-2

MD20 -26/-4



## Glutathione Transferases (gstm1\_human)



		BLOSUM50-10/-2		BLOSUM62-11/-1		MD40 -21/-4		MD10 -23/-4	
		E(320363)	f_id	E(320363)	f_id	E(320363)	f_id	E(320363)	f_id
Class-mu	GSTM1_HUMAN	1.3e-101	1.00	5.1e-132	1.000	0	1.000	0	1.000
	GSTM4_HUMAN	1.9e-89	0.867	1.1e-115	0.867	2.2e-188	0.867	1.9e-193	0.867
	GSTM2_MOUSE	3.0e-87	0.839	3.6e-113	0.839	1.4e-184	0.847	2.5e-187	0.847
	GSTM5_HUMAN	4.9e-87	0.876	6.9e-114	0.876	4.7e-187	0.876	7.2e-195	0.912
	GSTM2_HUMAN	8.2e-87	0.844	8.2e-113	0.844	2.6e-182	0.844	1.3e-184	0.844
	GSTM1_MOUSE	7.0e-83	0.780	2.5e-107	0.780	4.7e-169	0.780	1.5e-162	0.780
	GSTM6_MOUSE	1.9e-82	0.775	1.0e-106	0.775	5.1e-168	0.779	1.3e-161	0.779
	GSTM4_MOUSE	8.7e-82	0.769	4.7e-105	0.769	7.7e-166	0.769	2.1e-158	0.769
	GSTM5_MOUSE	6.9e-73	0.727	3.5e-94	0.727	1.3e-142	0.727	3.7e-128	0.727
	GSTM3_HUMAN	8.2e-73	0.731	6.7e-95	0.731	3.4e-143	0.731	8.2e-129	0.731
Class-pi	GSTM2_CHICK	9.8e-65	0.656	4.7e-84	0.656	3.0e-117	0.656	1.4e-93	0.675
	GST26_FASHE	2.9e-44	0.495	1.3e-56	0.491	2.7e-59	0.502	3.2e-18	0.510
	GSTM1_DERPT	5.2e-42	0.467	1.6e-53	0.487	5.1e-57	0.505	2.4e-29	0.651
	GST27_SCHMA	2.4e-37	0.467	9.5e-49	0.458	4.7e-42	0.470	5.1e-20	0.607
	GSTP1_PIG	2.9e-20	0.327	1.2e-25	0.327	0.00034	0.409		
	GSTP1_XENLA	5.2e-19	0.333	6.0e-24	0.330	0.12	0.464		
	GSTP2_MOUSE	8.0e-17	0.294	1.3e-20	0.294	1.1	0.395		
	GSTP1_CAEEL	1.1e-16	0.324	4.3e-21	0.319	1.1	0.706		
	GSTP1_HUMAN	3.0e-16	0.284	2.2e-20	0.284	0.29	0.467		
	GSTP1_BUFBU	1.2e-14	0.285	7.2e-18	0.272	9.7	0.588		
Class-alpha	GSTPA_CAEEL	1.1e-13	0.298	2.8e-17	0.284	0.002	0.400		
	PTGD2_MOUSE	4.8e-12	0.302	2.6e-14	0.293				
	PTGD2_RAT	4.8e-12	0.302	1.5e-14	0.293				
	PTGD2_HUMAN	1.1e-11	0.292	4.0e-13	0.281				
	PTGD2_CHICK	9.8e-11	0.304	6.9e-13	0.302				
	GSTP2_BUFBU	2.0e-10	0.288	2.2e-12	0.307				
	GST_MUSDO	5.8e-09	0.257	2.3e-11	0.251				
	GST1_DROME	1.0e-08	0.255	2.9e-10	0.237				
	GSTA1_MOUSE	1.5e-08	0.279	4.9e-11	0.264				
	GSTA2_HUMAN	6.6e-08	0.286	1.2e-08	0.273				

43

## Scoring matrices influence alignment lengths

### A. Search with MJ0050

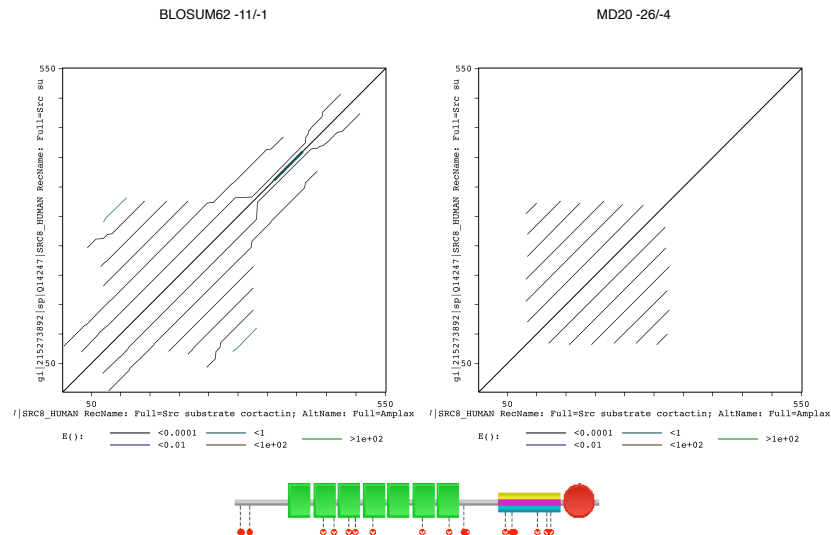
	BLOSUM50 -10/-2				BLOSUM62 -7/-1				BLOSUM62 -11/-1			
	s-w	E()	%_id	alen	s-w	E()	%_id	alen	s-w	E()	%_id	alen
<b>The best scores are:</b>												
NP_416010 glutamate decarb.	250	e-11	24.9	401	216	e-7	25.3	415	137	e-8	22.9	332
NP_417379 glycine decarb.	169	e-05	22.1	420	163	0.001	23.3	430	88	0.004	22.1	331
NP_417025 aminotransferase	122	0.02	23.6	254	119	0.12	24.5	257	76	0.04	23.7	118
NP_414772 aminoacyl-his.	110	0.15	23.4	188	108	0.74	23.2	311	57	6.9	23.4	188
NP_415139 alkyl hydroperoxide	99	1.1	26.9	156	104	1.5	24.5	233	62	2.0	28.9	97

### B. Search with MJ1633

	BLOSUM50 -10/-2				BLOSUM62 -7/-1				BLOSUM62 -11/-1			
	s-w	E()	%_id	alen	s-w	E()	%_id	alen	s-w	E()	%_id	alen
<b>The best scores are:</b>												
NP_417809 KefB	196	e-06	28.2	177	162	0.02	27.3	176	143	e-8	34.4	96
NP_414589 K+ antiporter	175	e-04	25.4	142	141	0.2	24.7	166	131	e-7	25.4	142
NP_415011 transport protein	133	0.03	23.2	142	113	4.4	23.2	142	89	0.005	23.2	142
NP_417748 TrkA	128	0.04	23.7	135	114	2.9	22.2	176	99	e-3	21.8	133
NP_416807 NAD(P) binding	103	0.98	26.1	92					70	0.29	26.1	92

44

## Scoring matrices influence alignment lengths



## Similarity Scoring Matrices - Summary

- Similarity scoring matrices are “log-odds” matrices, reporting the “odds” that an alignment reflects homology rather than chance
- One can predict evolutionary changes using a simple random model, which can generate mutation frequencies at any evolutionary distance
- The optimal scoring matrix has an evolutionary distance that matches that of the alignment. Matrices that are shallower than the true distance produce short alignments, while matrices that are deeper produce long alignments.
- Shallower scoring matrices have more information content, or “bits/residue”, and thus can be used to find shorter domains
- Scoring matrices set evolutionary look back times

### *Scoring Matrices - Summary*

- 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
- Shallow matrices set maximum look-back time

47

### Protein Evolution and Sequence Similarity

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- Alignment Algorithms/Local sequence alignments
- Similarity scoring matrices
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Improving sensitivity with PSI-BLAST

48

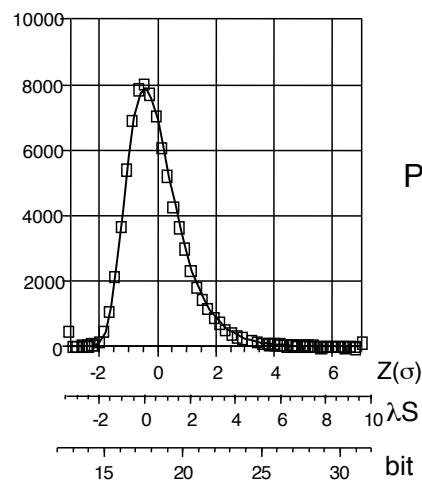


## Inferring Homology from Statistical Significance

- 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

49

### Extreme value distribution



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

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

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

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

$$E(S' > x \text{ ID}) = P D$$

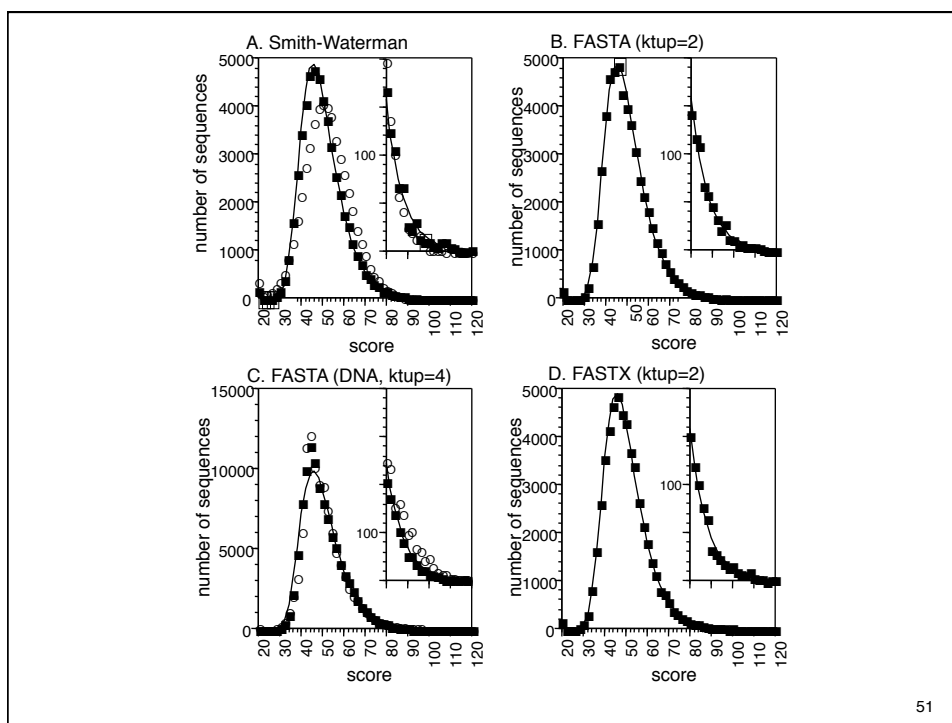
$$P(B \text{ bits}) = m n 2^{-B}$$

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

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

$$E(40 \mid D=4E6) = 0.6$$

50



## Smith-Waterman (ssearch)

The best scores are:

			s-w	bits	E(115640)	%_id	alen
GTM1_MOUSE	Glutathione S-trans	( 218)	1497	363.5	2e-100	1.000	218
GTM2_CHICK	Glutathione S-trans	( 220)	958	234.9	1.1e-61	0.619	218
GTP_HUMAN	Glutathione S-trans	( 210)	356	91.2	1.8e-18	0.308	211
PGD2_MOUSE	Glutathione-reg.	( 199)	262	68.8	9.7e-12	0.319	204
GTA1_MOUSE	Glutathione S-trans	( 223)	229	60.9	2.6e-09	0.284	225
SC1_OCTDO	S-crystallin 1 OLI	( 215)	228	60.7	3.0e-09	0.269	219
GTS_MUSDO	Glutathione S-trans	( 241)	228	60.6	3.4e-09	0.264	201
GTS1_CAEEL	Prob. Glut. S-trans	( 210)	220	58.8	1.1e-08	0.284	225
GTS_OMMSL	Glutathione S-trans	( 203)	196	53.0	5.5e-07	0.258	209
GTH3_ARATH	Glutathione S-trans	( 215)	142	40.1	0.0045	0.310	126
GTT2_HUMAN	Glutathione S-trans	( 244)	132	37.7	0.027	0.257	167
GT24_DROME	Glutathione S-trans	( 216)	131	37.5	0.028	0.255	153
YFCG_ECOLI	Hypothetical GST	( 215)	112	33.0	0.64	0.235	187
YJY1_YEAST	hypothetical 30.5	( 261)	110	32.4	*1.1*	0.248	149
DCMA_METS1	dichloromethane DM	( 267)	103	30.8	3.7	0.214	210
YA42_HAEIN	Hypothetical prot.	( 617)	108	31.7	*4.6*	0.283	120
GTO1_RAT	Glutathione trans	( 241)	100	30.1	5.4	0.234	158
DP41_BACHD	DNA polymerase I	( 413)	104	30.8	*5.4*	0.234	184
GTH1_WHEAT	Glutathione S-trans	( 229)	98	29.6	7.0	0.246	171
LGUL_SOYBN	Lactoylglutathione	( 219)	97	29.4	7.8	0.200	190
VP2_AHSV3	outer capsid prot	(1057)	108	31.5	*8.9*	0.205	200
GTH5_ARATH	Glutathione S-trans	( 218)	96	29.2	9.2	0.258	66
DCMA_METSP	dichloromethane DM	( 288)	98	29.5	9.3	0.195	200
GTXA_ARATH	Glutathione S-trans	( 224)	96	29.1	9.5	0.248	125
SLT_HAEIN	Putative soluble 1	( 593)	103	30.5	*9.9*	0.227	185

52

## FASTA search – low complexity regions

Search with complete grou\_drome:

The best scores are:					opt	bits	E(14548)
RGHUB1	GTP-binding regulatory protein beta-1 chain	( 341)	237	46.6	3.5e-05		
RGBOB1	GTP-binding regulatory protein beta-1 chain	( 341)	237	46.6	3.5e-05		
RGHUB3	GTP-binding regulatory protein beta-3 chain	( 341)	233	46.0	5.2e-05		
RGMSB4	GTP-binding regulatory protein beta-4 chain	( 341)	232	45.8	5.7e-05		
PIHUPF	salivary proline-rich glycoprotein precursor	( 252)	224	44.5	*0.00010*		
RGFFB	GTP-binding regulatory protein beta chain	( 347)	223	44.5	0.00014		
PIRT3	acidic proline-rich protein precursor - rat	( 207)	199	40.8	*0.0011*		
PIHUB6	salivary proline-rich protein precursor PR	( 393)	203	41.6	*0.0012*		
CGBO2S	collagen alpha 2(I) chain - bovine (fragment)	( 403)	195	40.5	*0.0027*		
WMBEW6	capsid protein - human herpesvirus 1 (strain)	( 636)	192	40.2	*0.0051*		
W4WLB5	E4 protein - human papillomavirus type 5b	( 246)	170	36.6	*0.024*		
OZZQMY	circumsporozoite protein precursor - Plasmodium	( 368)	172	37.1	*0.026*		
FOMVME	gag polyprotein - murine leukemia virus (strain)	( 537)	161	35.6	*0.10*		

Search with seg-ed grou\_drome: (low complexity regions removed)

The best scores are:					opt	bits	E(14548)
RGHUB3	GTP-binding regulatory protein beta-3 chain	( 341)	233	56.5	3.6e-08		
RGMSB4	GTP-binding regulatory protein beta-4 chain	( 341)	232	56.3	4.1e-08		
RGHUB2	GTP-binding regulatory protein beta-2 chain	( 341)	228	55.5	7.2e-08		
RGBOB1	GTP-binding regulatory protein beta-1 chain	( 341)	225	54.9	1.1e-07		
RGFFB	GTP-binding regulatory protein beta chain	( 347)	223	54.5	1.5e-07		
BVBYS	MSI1 protein - yeast (Saccharomyces cerevisiae)	( 423)	135	37.0	*0.033*		
ERHUAH	coatamer complex alpha chain homolog - human	(1225)	134	37.1	*0.088*		
A28468	chromogranin A precursor - human	( 458)	122	34.4	*0.21*		
RGOOBE	GTP-binding regulatory protein beta chain	( 342)	120	33.9	0.22	53	

## pseg removes low-complexity regions

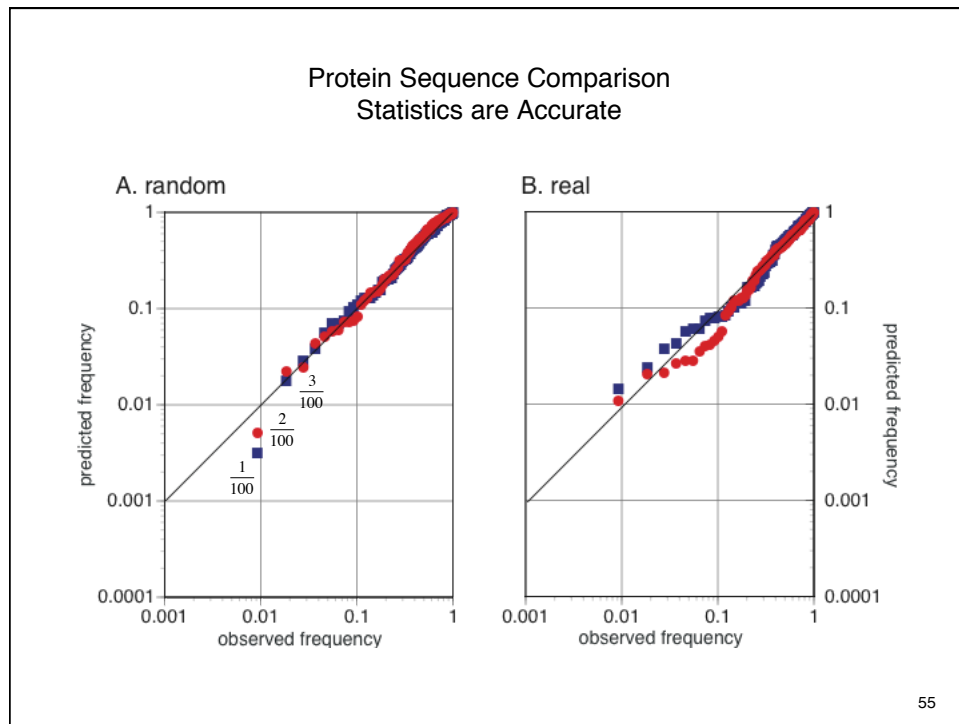
>gi|17380405|sp|P16371|GROU\_DROME Groucho protein (Enhancer of split M9/10)

```

      1-8      MYSPVRH
      9-19
paaggpppggp      20-131      IKFTIADTLERIKEEFNFLQAQYHSIKLEC
      EKLSNEKTEMQRHYVVMYEMSYGLNVMHK
      QTEIAKRLNTLINQLLPFLQADHQQQVLQA
      VERAKQVTMQELNLIIGQIHA
      qqvpggppqpmg      132-143
      144-281      ALNPFGLGATMGLPHGPGQLLNKPPPEHHR
      PDIKPTGLEGPAAAEERLNSVSPADREKY
      RTRSPDLIENDSKRRKDEKLQDEGEKSDQ
      DLVVDVANEMESHSPRPNGEHVSMVRDRE
      SLNGERLEKPSSSGIKQE
      rppsrsgsssrstps      282-297
      298-310      LTKDMEKPGTPG
      akartptpnaaapagvnpk      311-330
      qmmpqgpppagypgapyqrpa      331-351
      352-719      DPYQRPPSDPAYGRPPMPYDPHAHVRTNG
      IPHPSALTGGKPAYSFHMNGESLQPVFPF
      PDALVGVGIPRHARQINTLSHGEVVCVITI
      SNPTKYVYTGKGCVKVWDISQPGNKNPVS
      QLDCLQRDNYIRSVKLLPDGRTLIVGGEAS
      NLSIWDLASPTPRIKAELTSAAPACYALAI
      SPDSKVCFSCCSDGNIAVWDLHNEILVRQF
      QGHTDGASCIDISPDGSLWTGGLDNTVRS
      WDLREGRLQQHDFSSQIFSLGYCPTGDWL
      AVGMENSHVEVLHASKPKDYQLHLHESCVL
      SLRFAACGKWFVSTGKDNLLNAWRTPYGAS
      IFQSKETSSVLSCDISTDDKYIVTSGSGDKK
      ATVYEVII

```

54



### Statistical estimates from random shuffles

- BLAST estimates statistical significance from simulations of “normal” (average composition) proteins
- FASTA estimates statistical significance from the distribution of similarity scores obtained during the database search (selects 60,000 unrelated sequence scores from the database of *real* proteins)
- What if the sequences are different from most proteins, but similar to each other, e.g. membrane proteins?
- PRSS estimates statistical significance by producing hundreds of shuffled (random) sequences with the same length and composition, and then estimates  $\lambda$  and K from comparisons against those proteins

### prss - uniform and window shuffle

```

>lweec6 H+-transporting ATP synthase (EC 3.6.1.34) protein 6 - Escherichia coli
MASENMTPOD YIGHHLNNOQ LDLRTFSLVD PQNPATFWT INIDSMFFSV VLGLFLVLVF
RSVAKKATSG VPGKFQTAIE LVIGFVNGSV KDMYHGKSKL IAPLALTIFV WVFLLMNLMDL
LPIDLLPYTA EHVGLPALR VVPSADVNTV LSMALGVFIL ILFYSIKMKG IGGFTKELTL
QPFHMAFIP VNFLEGVSL LSKPVSLGLR LFGNMYAGEL IFILIAGLLP WWSQWILNVP
WAIFHILIT LQAFTEMLT IVYLSMASEE H

>lweec6_0 shuffled
GMPISVLLFK PPEVLLVFL SVMTNFPW GGFIMKGFKI VSFVGWVRV AVAGHLALYK
ITRDVNIKS AVFGSALLHP LLLQLSELNI VFNLLNIKI RTAYVHGML LSHIPLEPAS
GEGVFSDDL IITWNSASVL SGLDMFANIA LLGNPLMTN IVIILQRKFI ATTKFSLADI
HLHKQYSWDG MMSHTLIIFS ALELVWQNGD IFIPLNEYIL PFTLYVPNW ITQALVVALV
ELPGQQIDAE PLFLLPFPF EKTWYGDIMF L

PRSS34 - 1000 shuffles; uniform shuffle
unshuffled s-w score: 178; bits(s=178|n_l=271): 34.8 p(178) < 2.005e-06
For 10000 sequences, a score >= 178 is expected 0.02005 times

>lweec6_0 shuffled window: 10
EDSMANTMPO HONILGYHLN DLRTSDFVLL FTQAPWPTPN SMNIDIVFSF VLLVLLFFGL
SRGAVKATKS EQVTGIKFP VVSGVILGFN HDKGMSLYKK VLPFIIFLAAT DWLMNFVLLM
IIDLYLLAPP ERVGHPLAL APNVVVSVD MLFLIGSALV IFSLMKGIKY TTIFGLEKGL
QAWNFFPHIP NLSVEVGLLI GLPVRSSLKL MFEELAGNGY PFGILILILA SLINWVPWQW
IAIWTIFHL VQMTFFLAIL VSESELMIYA H

PRSS34 - 1000 shuffles; window shuffle, window size: 20
unshuffled s-w score: 178; bits(s=178|n_l=271): 34.5 p(178) < 2.601e-06
For 10000 sequences, a score >= 178 is expected 0.02602 times

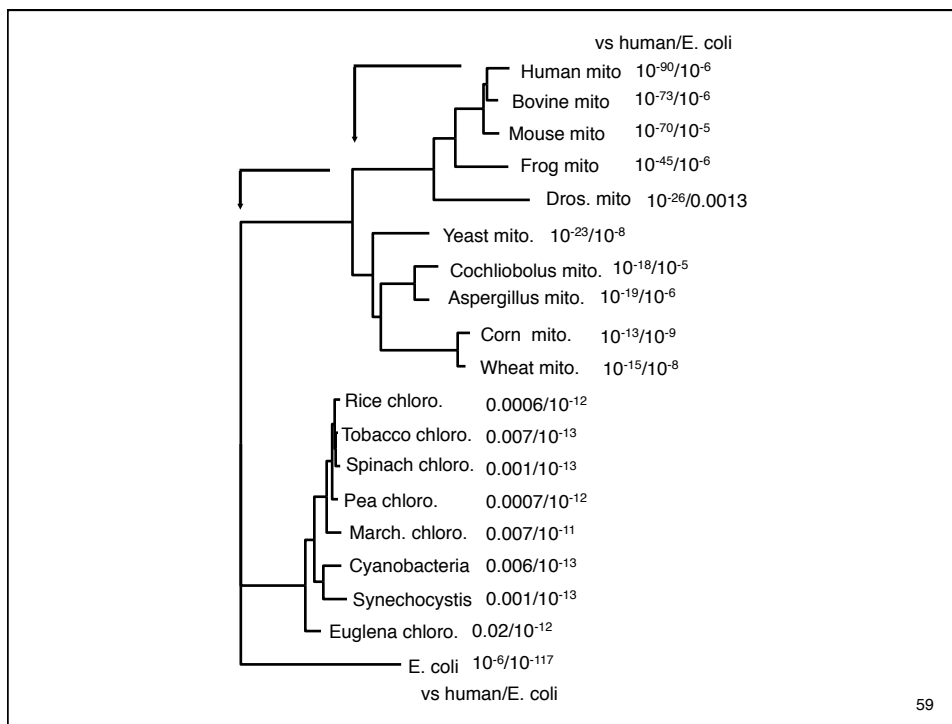
```

57

## Protein Evolution and Sequence Similarity

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- Alignment Algorithms/Local sequence alignments
- Similarity scoring matrices
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Improving sensitivity with PSI-BLAST

58



## 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	
ATPI_PEA Chloroplast ATP syn	35.8 0.013	38.0 0.0028	
ATPI_SPIOL Chloroplast ATP syn	35.5 0.015	38.0 0.0028	
ATPI_ATRBE Chloroplast ATP s	34.0 0.044	36.3 0.0086	
ATPI_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 *	
ATPI_EUGGR Chloroplast ATP syn	31.1 0.32	32.2 0.15	
ATP6_SYN6 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
ATPI_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_LEFSP Glucagon precursor	27.7 1.	32.7 0.033*	
*ADH1_MOUSE Alcohol dehydrogena	29.8 1.2	34.4 0.013*	

## Metazoan ATP Synthases

CLUSTAL W (1.81) multiple sequence alignment

```

ATP6_BOVIN  MNENLFTSFITPVILGLPLVTLIVLFPSSLF--PTSNRLVSNRFVTLQQWMLQLVSKQMMSIHNSKGQWTW-LML
ATP6_MOUSE  MNENLFASFITPTMMGFPIVVAIIMPSSLF--PSSKRLINNRLHSFQHWLVKLIKQMMLIHTPKGRWTW-LMI
ATP6_HUMAN  MNENLFASFIAPTILGLPAAVLIILFPPLLI--PTSKYLINNRLITTQQWLIKLTQSKQMMTMHNTKGRTW-S-LML
ATP6_XENLA  MNLSFFDQFMSPVILGIPLIAIAMDPTTLISWPIQSNGFNNRLITLQSWFLHNFTTIFYQLTSP-GHKWA-LLL
ATP6_DROYA  MMTNLFVSFDPFAIFNLSLNWLSTFLGLLMI--PSIYWLMPSTRYNIFWNSILLTLHKEFKTLLGPSGHNGSTFIF
          *  .:*  *  ...:..:  :  :  :  *  .  .*  :  :  :  .  *:  :  :  :

ATP6_BOVIN  MSLILFIGSTNLLGLLPHSFTPTTQLSMNLGMAIPLWAGAVITGFRNKTKASLAHFLPQGTPTPLIPMLVIIETI
ATP6_MOUSE  VSLIMFIGSTNLLGLLPHFTFTPTTQLSMNLGMAIPLWAGAVITGFRHKLKSSLAHFLPQGTPTPLIPMLVIIETI
ATP6_HUMAN  VSLIIFIATNLLGLLPHSFTPTTQLSMNLGMAIPLWAGAVITGFRSKIKNALAHFLPQGTPTPLIPMLVIIETI
ATP6_XENLA  TSLMLLLSLNLLGLLPYTFTPTTQLSLNMGAVPLWLATVIMASKP-TNYALGHLLEPGTPTPLIPVLIIEIETI
ATP6_DROYA  ISLSLILFNNFMGLPPYIFTSTSHLTTLTSLALPLWLCFMYLWGINHTQHMFAHLPVQGTPTAILMPFMVCIETI
          **  :  :  *:***:  **:***:***:***:***  :  :  :  :  :  :  :  :  :  :  :  :

ATP6_BOVIN  SLFIQPMALAVRLTANITAGHLLIHLIGGATLALMSISTTTALITFTILILLTILEFAVAMIQAYVFTLLVSLYLDNT
ATP6_MOUSE  SLFIQPMALAVRLTANITAGHLLMHLIGGATLVMNISPTTATITFIILLTILEFAVALIQAYVFTLLVSLYLDNT
ATP6_HUMAN  SLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYLDNT
ATP6_XENLA  SLFIRPLALGVRLTANLTAGHLLIQLIATAAFVLLSIMPTVAILTIVLFLTLLEIAVAMIQAYVFTLLVSLYLDNT
ATP6_DROYA  SNIRPGLTAVRLTANMIAHLLLTLLGNTGPSMSYLLVTLFLVAQIALLLV--ESAVTMIQSYVFAVLSTLYSSEVN
          *  :*:  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :  :

```

61

## PSI-BLAST ATP6\_HUMAN - 4 iterations

Results from round:		(1)		(2)		(3)		(4)	
Sequences producing significant alignments:		Score	E	Score	E	Score	E	Score	E
		(bits)	Value	(bits)	Value	(bits)	Value	(bits)	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				
LAMA2_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_SYN6	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_HTLIC	Gag-Pro-Pol polyprotein (Pr160Gag-Pro-Pol) [...	27	5.0						
POLG_DEN2J	Genome polyprotein [Contains: Capsid protei...	27	5.2	26	7.0				

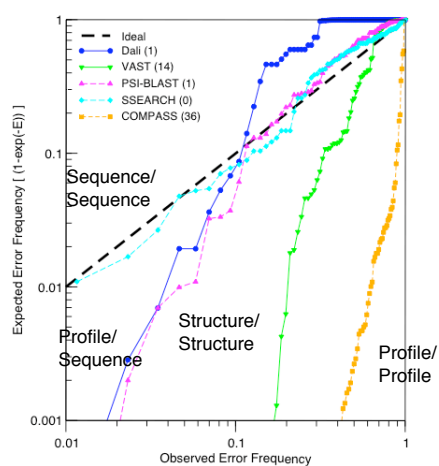
62

### 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	

63

### Accuracy of statistical estimates



- SSEARCH (Smith-Waterman) provides very accurate statistical estimates
- PSI-BLAST and Dali provide estimates that off by 10–100-fold
- Other structure comparison methods provide wild over estimates of statistical significance – *BEWARE of claims of significant structural similarity*

64



## Sensitive searches with PSI-BLAST

- PSI-BLAST improves sensitivity by building a Position Specific Scoring Matrix (PSSM)
  - models ancestral sequence (consensus distribution)
  - similar to PFAM HMM (but less sophisticated weights, gaps)
- Sensitivity improves with additional iterations
  - model moves to base of tree
- Statistical estimates are difficult
  - once a sequence is in, it is “significant” - validation must be done before a sequence is included
- Very diverse families may not produce a well defined PSSM
  - similar problems with HMMs have lead to “clans”

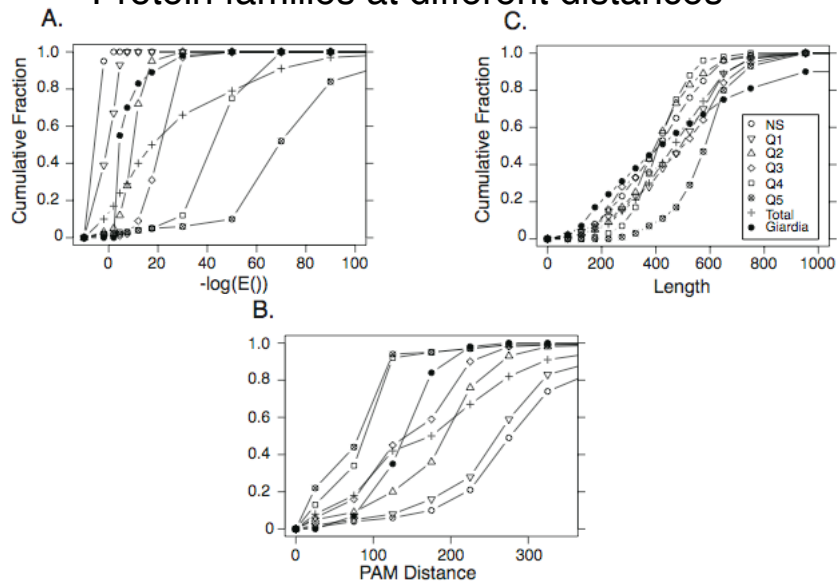
65

## *Sequence Similarity - Conclusions*

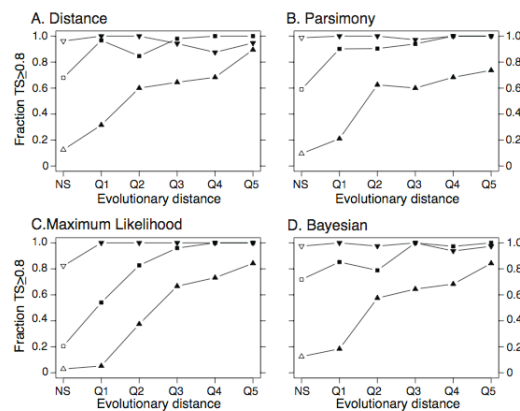
- Homologous sequences share a common ancestor, but most sequences are non-homologous
- Always compare Protein Sequences
- Sequence Homology can be reliably inferred from statistically significant similarity (non-homology cannot from non-similarity)
- Homologous proteins share common structures, but not necessarily common functions
- Sequence statistical significance estimates are accurate (verify this yourself)  $10^{-6} < E() < 10^{-3}$  is statistically significant
- Scoring matrices set evolutionary look back horizons - not every discovery is distant
- PSI-BLAST can be more sensitive, but with lower statistical accuracy

66

## Homology and Accuracy I Protein families at different distances



## Homology and Accuracy II Tree accuracy with different methods



### Discussion (exam) questions

1. What is the difference between similarity and homology? When does high identity not imply homology? What conclusions can be drawn from homology?
2. What is the range of an expectation value ( $E$ )-value)? If you compare a sequence to 50,000 random(unrelated) sequences, what should the expectation value for the highest of the 50,000 similarity scores be (on average)?
3. In a sequence similarity database search, you identify a statistically significant similarity ( $E < 0.005$ ), but the alignment is relatively short (50 aa). How might you determine whether the alignment reflects a genuine homology, or a random sequence match?
4. What scoring matrix should be used to identify protein orthologs that have diverged over the past 100 My (e.g. human/mouse)?
5. When the *M. janaschii* genome was first sequenced, Venter and his colleagues stated that almost 60% of the open reading frames (proteins or genes) were novel to this organism. (For bacteria like *E. coli* or *H. influenzae*, a similar number would be 20 - 40%.) On what would they base such a statement? Is it likely to be correct?

69