Sequence Similarity

Protein Sequence Comparison and Protein Evolution

(What BLAST does/Why BLAST works)

William R. Pearson
www.people.virginia.edu/~wrp
wrp@virginia.edu

1

20 Years of Biological Sequence Comparison

Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 726–730, February 1983
Biochemistry

***PNAS* (1983) 80:726**

Rapid similarity searches of nucleic acid and protein data banks

(global homology/optimal alignment)

W. J. WILBUR AND DAVID J. LIPMAN

Mathematical Research Branch, National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases, National Institutes of Health, Building 31 Room 4B-54, Bethesda, Maryland 20205

RESEARCH ARTICLE

***Science* (1985) 227:1435**

Rapid and Sensitive Protein Similarity Searches

David J. Lipman and William R. Pearson

J. Mol. Biol. (1990) 215, 403–410

***J. Mol. Biol.* (1990) 215:403**

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

2

Sequence Similarity - Conclusions

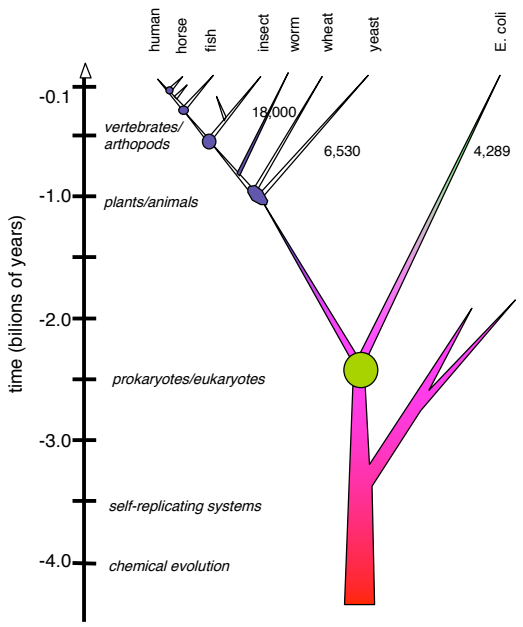
- Always compare Protein Sequences
- Sequence Homology can be reliably inferred from statistically significant similarity (non-homology cannot from non-similarity)
- Homologous proteins share common structures, but not necessarily common functions
- Sequence statistical significance estimates are accurate (verify this yourself)$10^{-6} < E() < 10^{-3}$ is statistically significant
- Structure comparison is more sensitive than sequence comparison, but less reliable for establishing homology

3

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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Sequence comparison vs structure comparison, reliability and sensitivity

4



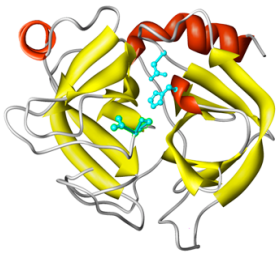
5

E. coli proteins vs Human – Ancient Protein Domains

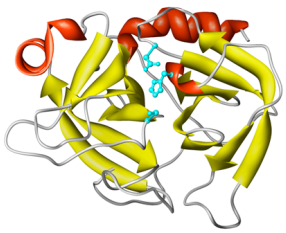
expect	%_id	alen	E coli descr	Human descr	sp_name
2.7e-206	53.8	944	glycine decarboxylase, P	Glycine dehydrogenase [de	GCSP_HUMAN
1.2e-176	59.5	706	methylmalonyl-CoA mutase	Methylmalonyl-CoA mutase,	MUTA_HUMAN
3.8e-176	50.6	803	glycogen phosphorylase [E	Glycogen phosphorylase, l	PHS1_HUMAN
9.9e-173	55.6	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 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

6

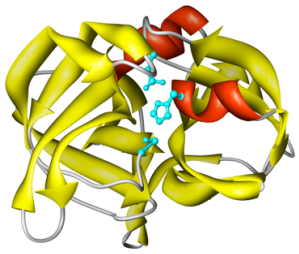
Homology => structural similarity
? sequence similarity



Bovine trypsin (5ptp)
Structure: E()< 10^{-23}; RMSD 0.0 A
Sequence: E()< 10^{-84}
100% 223/223



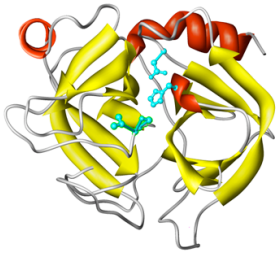
S. griseus trypsin (1sgt)
E()< 10^{-14} RMSD 1.6 A
E()< 10^{-19} 36%; 226/223



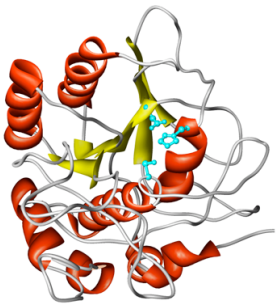
S. griseus protease A (2sga)
E()< 10^{-4}; RMSD 2.6 A
E()< 2.6 25%; 199/181

7

Non-homologous proteins have different structures



Bovine trypsin (5ptp)
Structure: E()<10^{-23}
RMSD 0.0 A
Sequence: E()<10^{-84}
100% 223/223



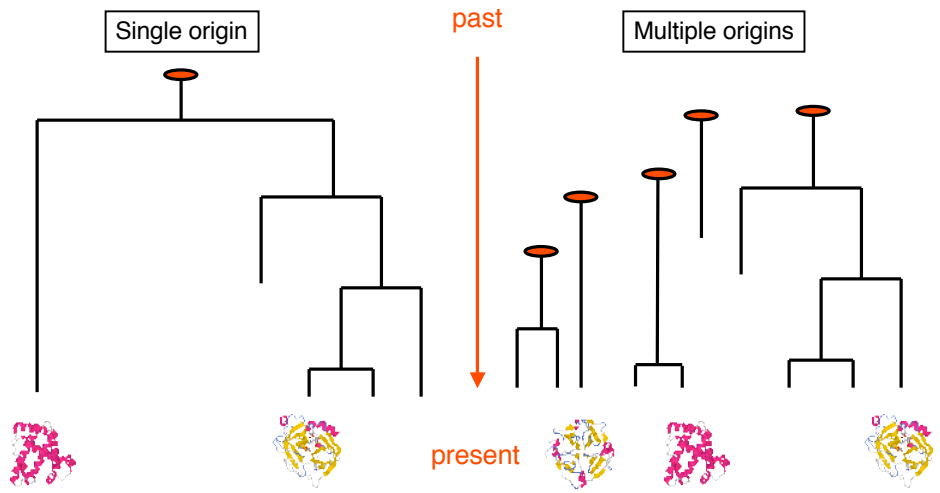
Subtilisin (1sbt)
E() >100
E()<280; 25% 159/275

Cytochrome c4 (1etp)
E() > 100
E()<5.5; 23% 171/190

8

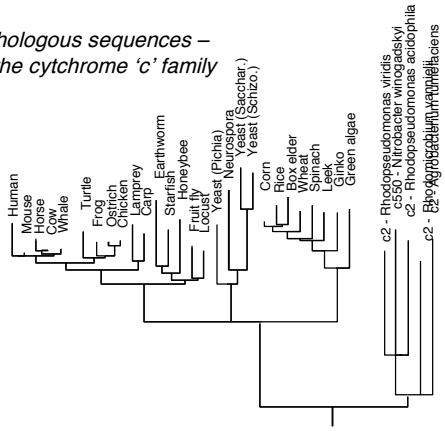
Are All Sequences Homologous?

No Homology without excess similarity

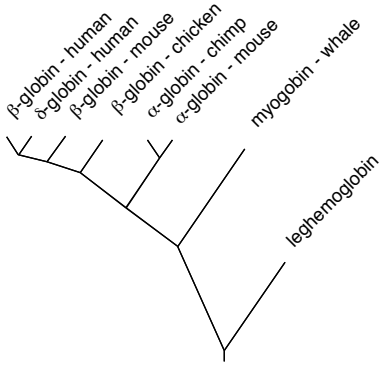


Orthologs and Paralogs – Inferring Function

Orthologous sequences – the cytchrome 'c' family



Paralogous genes – globins

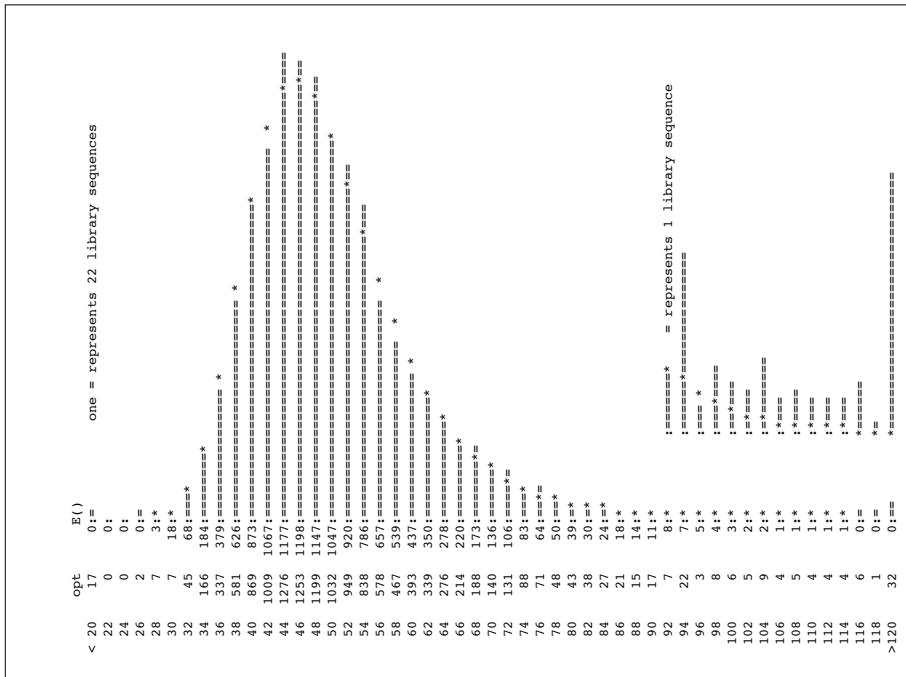


10

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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Sequence comparison vs structure comparison, reliability and sensitivity

11



12

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

13

The best scores are:	s-w bits	E(14548)	% id	alen
PWHU6 H+-trans. ATP syn. - human mito.	400 326.7	3.3e-90	1.000	226
PWBO6 H+-trans. ATP syn. - cow mito.	157 271.3	1.6e-73	0.779	226
PWMS6 H+-trans. ATP syn. - mouse mito.	118 262.4	7.6e-71	0.757	226
PWXL6 H+-trans. ATP syn. - frog mito.	745 177.3	3.1e-45	0.533	229
PWFF6 H+-trans. ATP syn. - D. melanog.	471 114.8	2.0e-26	0.378	222
PWBY3 H+-trans. ATP syn. - yeast mito.	438 107.3	4.4e-24	0.362	232
PWAS6N H+-trans. ATP syn. - E. nidulans	365 90.6	4.4e-19	0.304	230
PWKQ6 H+-trans. ATP syn. - H. maydis	353 87.9	3.0e-18	0.313	214
PWWT6 H+-trans. ATP syn. - wheat mito.	309 77.8	4.9e-15	0.292	233
PWNT6M H+-trans. ATP syn. - tobacco	309 77.8	5.0e-15	0.283	233
PWZM6M H+-trans. ATP syn. - corn mito.	283 71.9	2.2e-13	0.311	180
LWEC6 H+-trans. ATP syn. - E. coli	178 48.0	3.3e-06	0.237	236
LWRZ6 H+-trans. ATP syn. - rice chloro.	144 40.2	0.00063	0.242	231
PWPMA6 H+-trans. ATP syn. - pea chloro.	143 40.0	0.00074	0.250	232
PWYBAA H+-trans. ATP syn. - Cyano. syn.	142 39.7	0.00099	0.265	170
PWSPA6 H+-trans. ATP syn. - spinach	138 38.9	0.0016	0.238	231
PWYCA6 H+-trans. ATP syn. - Synecho.	127 36.3	0.0099	0.263	167
LWNT6 H+-trans. ATP syn. - tobacco	126 36.1	0.011	0.221	231
LWLV6 H+-trans. ATP syn. - liverwort	126 36.1	0.011	0.244	168
PWEGAC H+-trans. ATP syn. - euglena	123 35.4	0.018	0.257	214
JQ0026 ATP/ADP translocase tlc1 - Ricket	122 35.1	0.045	0.247	154
S17420 ubiquinol--cytochrome-c reductase	113 33.1	0.14	0.228	158
QXBO2M NADH dehydrogenase (ubiquinone)	107 31.7	0.32	0.261	211
S17415 ubiquinol--cytochrome-c reductase	105 31.3	0.49	0.277	137
S17417 ubiquinol--cytochrome-c reductase	104 31.0	0.57	0.277	137
DNHUN2 NADH dehydrogenase (ubiquinone)	103 30.8	0.61	0.201	149
CBHU ubiquinol--cytochrome-c reductase	102 30.6	0.79	0.268	205
QRECAA aromatic amino acid trans. prot.	103 30.8	0.82	0.234	111
S17419 ubiquinol--cytochrome-c reductase	101 30.3	0.92	0.234	158

14

```
>>LWEC6 H+-transporting ATP synthase (EC 3.6.1.34) protein 6 - Escherichia coli        (272 aa)
 s-w opt: 178  Z-score: 218.7  bits: 48.0 E(): 3.3e-06
Smith-Waterman score: 178;  23.729% identity (28.141% ungapped) in 236 aa overlap (8-222:45-264)

                    10        20        30        40        50        60        70        80
PWHU6        MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTKGRTWSLMLVSLIIFIA
                   :..  ..:::   ....::  .   ...  . ... :. . .: ... ..  :.. :.. . . .......
LWEC6  QNPPATFWTINIDSMFFSVVLGL---LFLVLFRSVAKKATSG-VPGKFQTAIELVIGFVNGSVKDMYHGKSKLIAPLALTIFVWVF
              40        50          60        70         80        90      100       110

                     90              100       110       120       130            140
PWHU6  TTNLLGLLP---------HSF-------TPTTQLSMNLAMAIPLWAGTVIMGFRSKIKNALAHFLPQGTPTPL-----IPMLVIIE
        ::. :::          : .       .:.......:.::. ..   ... : :  .... :  . :  :.     ::. .:.:
LWEC6  LMNLMDLLPIDLLPYIAEHVLGLPALRVVPSADVNVTLSMALGVF---ILILFYSIKMKGIGGFTKELTLQPFNHWAFIPVNLILE
         120       130       140       150          160       170       180         190

        150       160       170       180       190       210       220
PWHU6  TISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYLHDNT
       .::: .:..:..:: .:. ::.:.. ::..     :   :      :: :::.          .::..: .:. .::
LWEC6  GVSLLSKPVSLGLRLFGNMYAGELIFILIAGLLPWWSQWILNVPWAIFHILIIT---------LQAFIFMVLTIVYLSMASEEH
        200       210       220       230       240       250               260       270
```

15

The PAM250 matrix

```
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  0 -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 17
      C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  Y  W
```

16

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

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

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

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

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

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

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

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

17

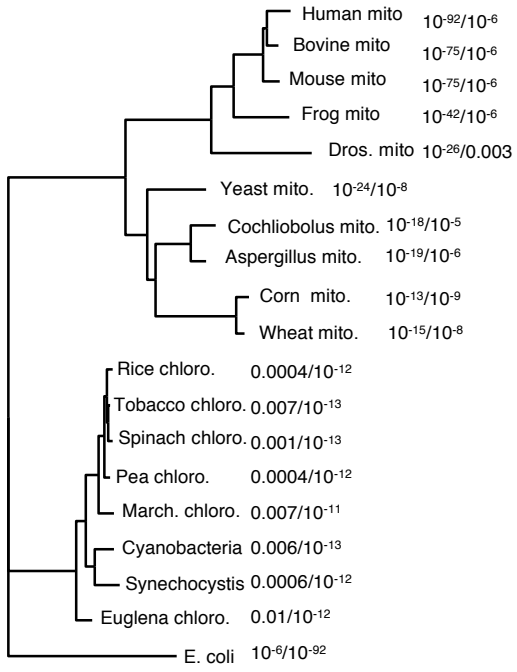
```
>PWEGAC H+-transporting ATP synthase (EC 3.6.1.34) chain a - Euglena gracilis chloroplast (252 aa)
 s-w opt: 123  Z-score: 151.6  bits: 35.4 E(): 0.018
Smith-Waterman score: 123;  25.701% identity (30.220% ungapped) in 214 aa overlap (21-222:50-243)

                      10        20        30        40        50        60            70
PWHU6           MNENLFASFIAPTILGLPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTK-GRT----WSLM
                                .::: :  : : :.: :  . . ...: .:.:... .  . .:. :.     :  .
PWEGAC IANVEVGQHFYWSILGFQIHGQVLINSWIVILIIGF--LSIYTTKNL--TLVPANKQIFIELVTEFITDISKTQIGEKEYSKWVPY
      20        30        40        50          60          70        80        90      100

           80         90             100       110       120       130      140       150
PWHU6  LVSLIIFIATTNLLG-LLPHSFT--PTTQL---SMNLAMAIPLWAGTVIMGFRSKI-KNALAHFLPQGTPTPLIPMLVIIETISLL
       . ...:: ..:  : :.: ..   :. .:   .  ..  .  :   : .  : . . :..:..:     :::.. . :.: ..
PWEGAC IGTMFLFIFVSNWSGALIPWKIIELPNGELGAPTNDINTTAGLAILTSLAYFYAGLNKKGLTYFKKYVQPTPILLPINILEDFT--
          110       120       130       140       150       160       170       180

            160       170       180       190       200       210       220
PWHU6  IQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYLHDNT
        .:..:. :: .:: : .:.. .. :         .:  ::. . ::.: ..   .. ::: .:. : . :.
PWEGAC -KPLSLSFRLFGNILADELVVAVLVSL--------VP--LIVPVPLIFLGLF---TSGIQALIFATLSGSYIGEAMEGHH
       190        200       210               220          230       240       250
```

18



19

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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Sequence comparison vs structure comparison, reliability and sensitivity

20

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

21

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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Sequence comparison vs structure comparison, reliability and sensitivity

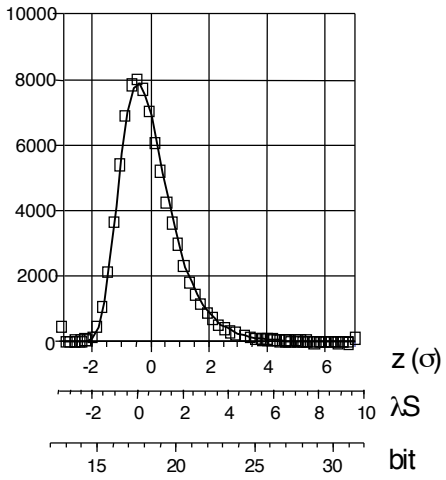
22

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

23

Extreme value distribution



$S' = \lambda S - \ln K\ m\ n$

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

$E(S'>x \mid D) = P\ D$

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

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

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

$E(40 \mid D=1.4E6) = 0.2$

24

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-req.    ( 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 OL1  ( 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 l  ( 593)  103  30.5  *9.9*   0.227  185
```

25

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 chai  ( 341)  237  46.6  3.5e-05
RGBOB1 GTP-binding regulatory protein beta-1 chai  ( 341)  237  46.6  3.5e-05
RGHUB3 GTP-binding regulatory protein beta-3 chai  ( 341)  233  46.0  5.2e-05
RGMSB4 GTP-binding regulatory protein beta-4 chai  ( 341)  232  45.8  5.7e-05
PIHUPF salivary proline-rich glycoprotein precurs  ( 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 (fragme  ( 403)  195  40.5 *0.0027*
WMBEW6 capsid protein - human herpesvirus 1 (stra  ( 636)  192  40.2 *0.0051*
W4WLB5 E4 protein - human papillomavirus type 5b   ( 246)  170  36.6 *0.024*
OZZQMY circumsporozoite protein precursor - Plasm  ( 368)  172  37.1 *0.026*
FOMVME gag polyprotein - murine leukemia virus (s  ( 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 chai  ( 341)  233  56.5 3.6e-08
RGMSB4 GTP-binding regulatory protein beta-4 chai  ( 341)  232  56.3 4.1e-08
RGHUB2 GTP-binding regulatory protein beta-2 chai  ( 341)  228  55.5 7.2e-08
RGBOB1 GTP-binding regulatory protein beta-1 chai  ( 341)  225  54.9 1.1e-07
RGFFB  GTP-binding regulatory protein beta chain   ( 347)  223  54.5 1.5e-07
BVBYMS MSI1 protein - yeast (Saccharomyces cerevi  ( 423)  135  37.0 *0.033*
ERHUAH coatomer complex alpha chain homolog - hum  (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
```

26

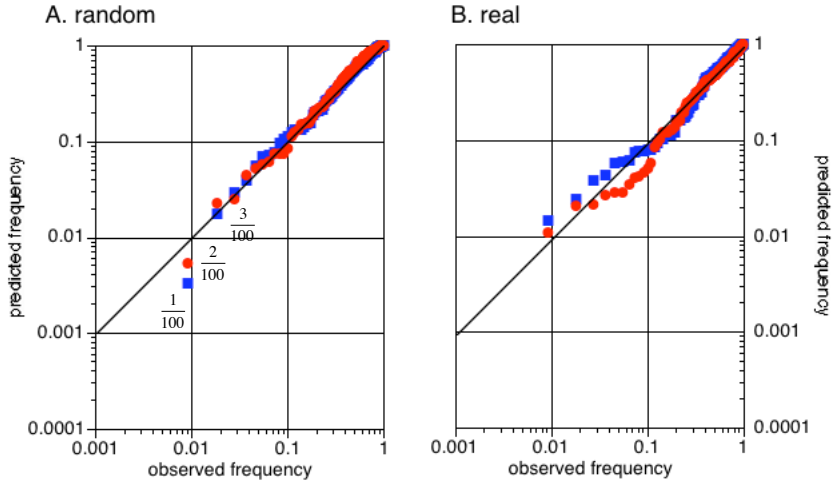
`pseg` removes low-complexity regions

```
>gi|17380405|sp|P16371|GROU_DROME Groucho protein (Enhancer of split M9/10)

                              1-8     MYPSPVRH
                  paaggpppqgp 9-19
                             20-131   IKFTIADTLERIKEEFNFLQAQYHSIKLEC
                                      EKLSNEKTEMQRHYVMYYEMSYGLNVEMHK
                                      QTEIAKRLNTLINQLLPFLQADHQQQVLQA
                                      VERAKQVTMQELNLIIGQQIHA
                 qqvpggppqpmg 132-143
                             144-281  ALNPFGALGATMGLPHGPQGLLNKPPEHHR
                                      PDIKPTGLEGPAAAEERLRNSVSPADREKY
                                      RTRSPLDIENDSKRRKDEKLQEDEGEKSDQ
                                      DLVVDVANEMESHSPRPNGEHVSMEVRDRE
                                      SLNGERLEKPSSSGIKQE
          rppsrsgssssrstps 282-297
                             298-310  LKTKDMEKPGTPG
       akartptpnaaapapgvnpk 311-330
     qmmpqgpppagypgapyqrpa 331-351
                             352-719  DPYQRPPSDPAYGRPPPMPYDPHAHVRTNG
                                      IPHPSALTGGKPAYSFHMNGEGSLQPVPFP
                                      PDALVGVGIPRHARQINTLSHGEVVCAVTI
                                      SNPTKYVYTGGKGCVKVWDISQPGNKNPVS
                                      QLDCLQRDNYIRSVKLLPDGRTLIVGGEAS
                                      NLSIWDLASPTPRIKAELTSAAPACYALAI
                                      SPDSKVCFSCCSDGNIAVWDLHNEILVRQF
                                      QGHTDGASCIDISPDGSRLWTGGLDNTVRS
                                      WDLREGRQLQQHDFSSQIFSLGYCPTGDWL
                                      AVGMENSHVEVLHASKPDKYQLHLHESCVL
                                      SLRFAACGKWFVSTGKDNLLNAWRTPYGAS
                                      IFQSKETSSVLSCDISTDDKYIVTGSGDKK
                                      ATVYEVIY
```

27

Protein Sequence Comparison Statistics are Accurate



28

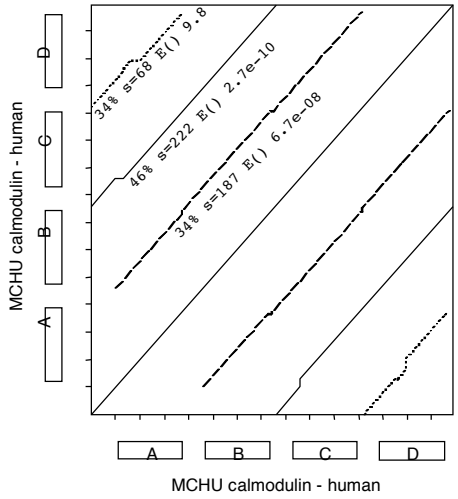
Local alignments - calmodulin

```
 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    MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADG
        : :  .::.:   .::: .::::::.: :.. ::  :: .::.. :. :...:: :.: ::
mchu    MKDTDSEEEI---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    AEFKEAFSLFDKDGDGTITTKELGTVM-RSLGQNPTEAELQDMINEVDADGNGTIDFPEF
        ::... ..  : ::.:::   :. :.: :.. .. .: :... .   : :::: :.  :.
mchu    AELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAEL
          50        60        70        80        90       100
          70        80        90       100       110
mchu    ---LTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMT
           .: ...:. : . .: ::::    : ::.: ..  :. ..::
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-KEAFSLFDKDGDGTITTKELGTVM
        ....::.:.. :. .::    : :::: .. .:.  .:
mchu    LGEKLTDEEVDEMIREA----DIDGDGQVNYEEFVQMM
             120           130       140
```


Repeated domains with local alignments



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
- When are we certain that an alignment is significant - similarity score statistics?
- When to trust similarity statistics?
- Sequence comparison vs structure comparison, reliability and sensitivity

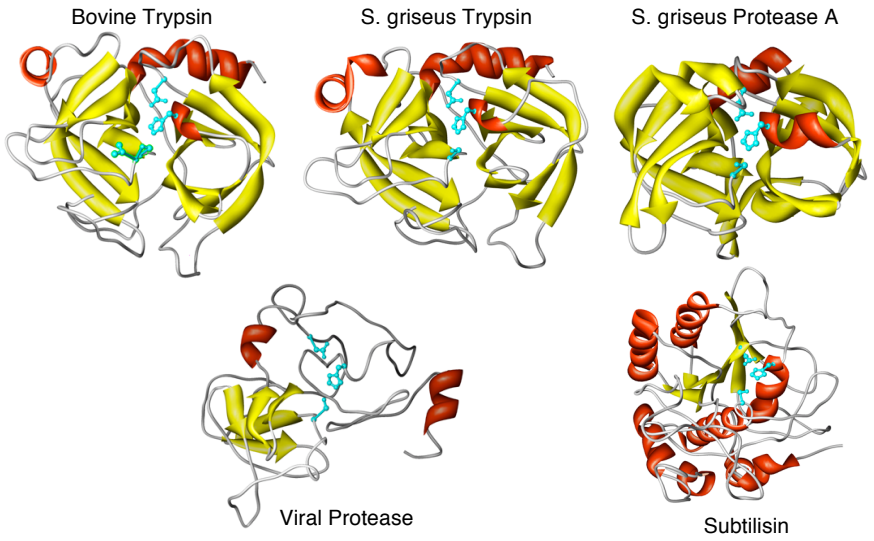
31

Homology from Similarity – Sequence or Structure?

- Structure comparison is the “gold standard” for establishing homology – structures change more slowly than sequence
- Structure comparison problems –
  - Structures are not unique (differ by > 1.5 Å for identical sequences)
  - No optimal alignment algorithm
  - Poor understanding of statistics - no “random” structures
- Statistical significance of structural similarity rarely quantified - homology vs analogy (convergence).

32

Homologs, Topologs, and Convergence

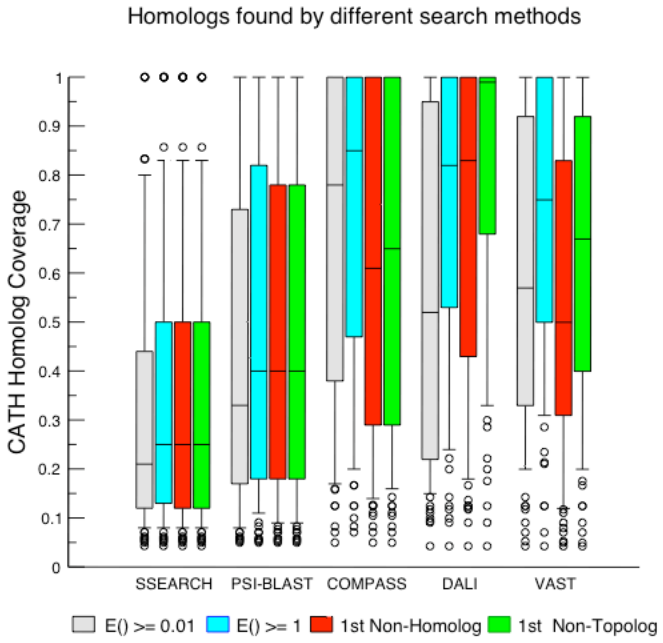


33

Homlogy, Similarity, and Convergence – Serine Proteases

		CATH Homology			Topology	Convergent
		Bovine Trypsin	S. griseus Trypsin	S. griseus Protease A	Viral Protease	Subtilisin
		5PTP vs. :	1SGT	2SGA	1BEF	1SBT
Structure/ Structure →	**Dali**	Z	32.7	13.7	8.8	<2
		E(2775)	10^{-14}	10^{-4}	0.02	>100
		N_{align} (%id)	209 (34)	147 (19)	131 (10)	N/A
		RMSD (Å)	1.4	2.8	2.9	N/A
Structure/ Structure →	**VAST**	E(2775)	10^{-21}	0.017 [a]	1.94	N/A
		N_{align} (%id)	208 (34)	130 (22)	122 (14)	N/A
		RMSD (Å)	1.5	2.3	2.8	N/A
Profile/ Profile ←	**COMPASS**	E(10000)	10^{-114}	10^{-13}	0.056	13
Profile/ Sequence →	**PSI-BLAST**	E(2775)	10^{-48}	2.5	>10	>10
		N_{align}	231	40	N/A	N/A
Sequence/ Sequence ←	**SSEARCH**	E(10000)	10^{-19}	2.6	>10	>10
		N_{align} (%id)	223 (36)	181 (25)	68 (33)	159 (25)

34



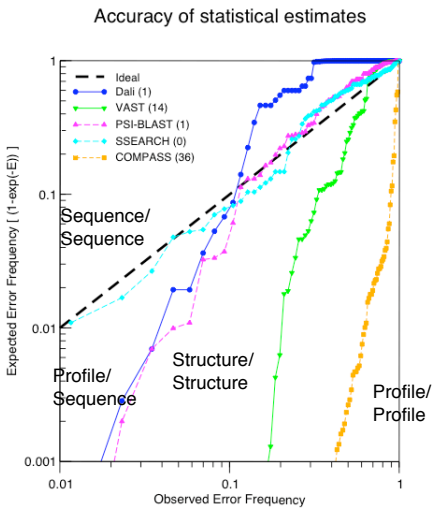
35

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

1. *Should Unrelated Structures have E() ≥ 1 ?*
2. *Are there "chance" Structural Similarities?*

36



- 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*

37

Structure Comparison Statistics

- Most structure comparison methods report very significant structural similarity for non-homologous proteins (*unrelated ≠ random*)
- These significance estimates are used to infer *ancient domain homologies*, which are preferred to *multiple independent origins*
- Dali produces relatively accurate estimates, and is one of the most sensitive search methods – thus, unrelated structures may be random
- If structural similarity can be random, there may be many more possible structures than existing ones

38

Sequence Similarity - Conclusions

- 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
- Protein sequence statistical significance estimates are accurate (verify)$10^{-6} < E() < 10^{-3}$ is statistically significant
- Searching smaller libraries improves sensitivity
- Structure comparison is more sensitive than sequence comparison, but less reliable for establishing homology.

39

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. Why is statistical significance important when inferring homology?
3. 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)?
4. 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 eubacterial 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?
5. Why is structure comparison considered more sensitive than sequence comparison? Why is it less “selective”?

40