

Protein Evolution and Similarity Searching

Establishing Homology

BIOC 8142

February 27, 2013

Bill Pearson wrp@virginia.edu 4-2818 Jordan 6-057

Goals of today's lecture:

- understand why and how homology is inferred; the meaning of "expectation value"
- understand sequence similarity, and why protein comparison is more sensitive than DNA sequence comparison
- why searches of smaller databases are more sensitive
- when to change the scoring matrix

1

What should you do to reinforce the lecture material?

- Pearson, (2000) "Protein Evolution and Sequence Comparison" ISMB2000 tutorial (collab)
- Koonin and Galperin (2003) Sequence – Evolution – Function www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=sef Chapter 4, section 4.2, Principles of sequence similarity searches (collab)
- Doolittle (1981) "Similar amino acid sequences: Chance or common ancestry" Science 214:149-159 (collab)
- Altschul et al. (1994) "Issues in searching molecular sequence databases" Nature Genet. 6:119-129 (collab)
- Altschul, (1991) "Amino-acid scoring matrices from an information theoretic perspective" J. Mol. Biol. 219:555-565

2

Sequence Similarity - Conclusions

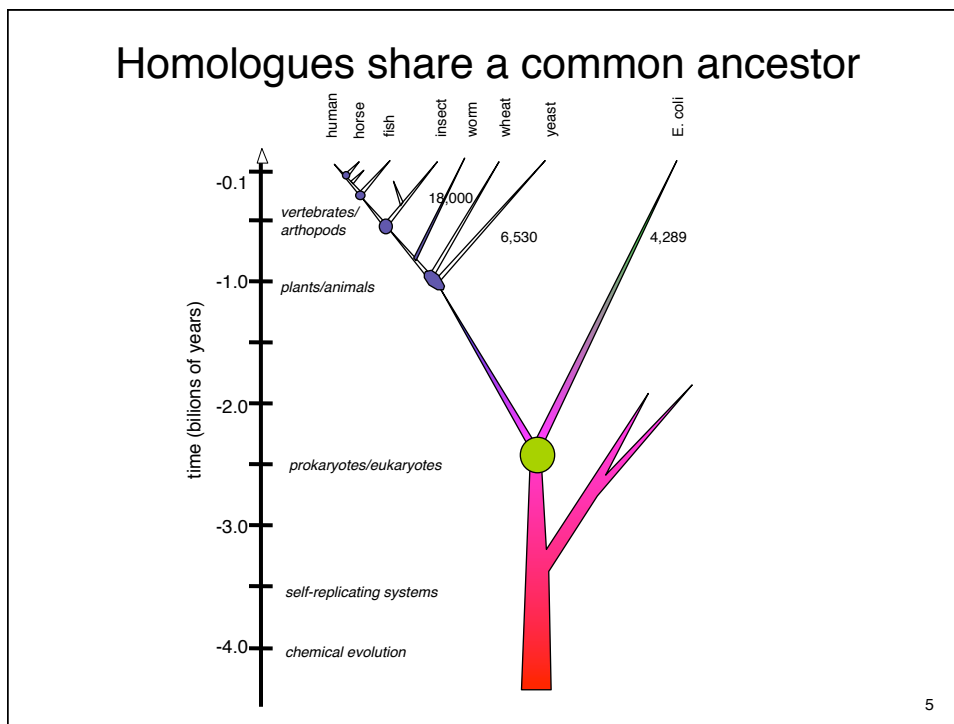
- Homologous sequences share a common ancestor, but most sequences are non-homologous
- Always compare Protein Sequences
- *BLAST works* – Sequence Homology can be reliably inferred from statistically significant similarity (non-homology cannot from non-similarity)
- Homologous proteins share common structures, but not necessarily common functions (other resources required)
- Sequence statistical significance estimates are accurate (verify this yourself) $10^{-6} < E() < 10^{-3}$ is statistically significant

3

Protein Evolution and Sequence Similarity

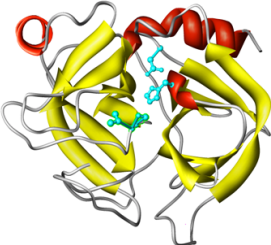
- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison
- Alignment Algorithms/Local sequence alignments
- Similarity scoring matrices
- When are we certain that an alignment is significant - similarity score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

4

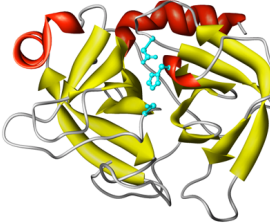


When do we infer homology?

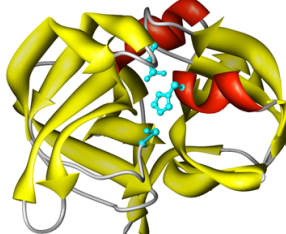
Homology \Leftrightarrow structural similarity
? sequence similarity



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



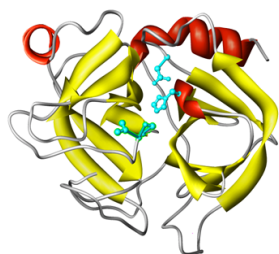
S. griseus trypsin (1sgt)
 $E() < 10^{-14}$ RMSD 1.6 Å
 $E() < 10^{-19}$ 36%; 226/223



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

6

When can we infer non-homology?

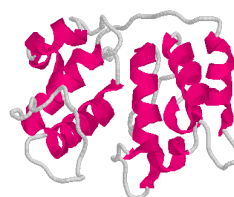


Bovine trypsin (5ptp)
 Structure: $E() < 10^{-23}$
 RMSD 0.0 Å
 Sequence: $E() < 10^{-84}$
 100% 223/223

Non-homologous proteins have different structures

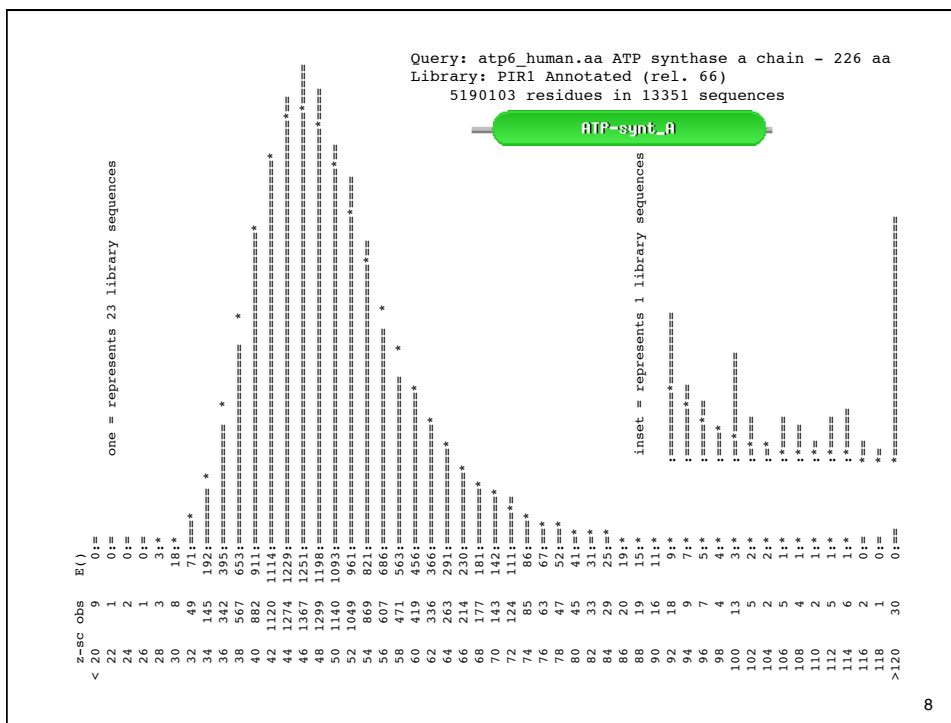


Subtilisin (1sbt)
 $E() > 100$
 $E() < 280$; 25% 159/275



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

7



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

9

```

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

```

10


```

>>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      MNENLFASFIAPTILGLPAAVLIIILFPPLLIPTSKYLINNRLLITQQWLKIKLTSKQMMTM
          .:. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Euglena VNMFISGIFQIANVEVGQHFYWSILGFQIHGQVLINSWIVILLIIGF--LSIYTKNL--TLVPANKQIFIELVTEFITDI
          10          20          30          40          50          60          70          80

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

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

          210         220
human     IQAYVFTLLVSLYLDHNT
          .:. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Euglena  IQALIFATLSGSGYIGEAMEGHH
          230         240         250
    
```

```

Query: atp6 human.aa ATP synthase a chain - 226 aa
Library: 5190103 residues in 13351 sequences

The best scores are:
(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

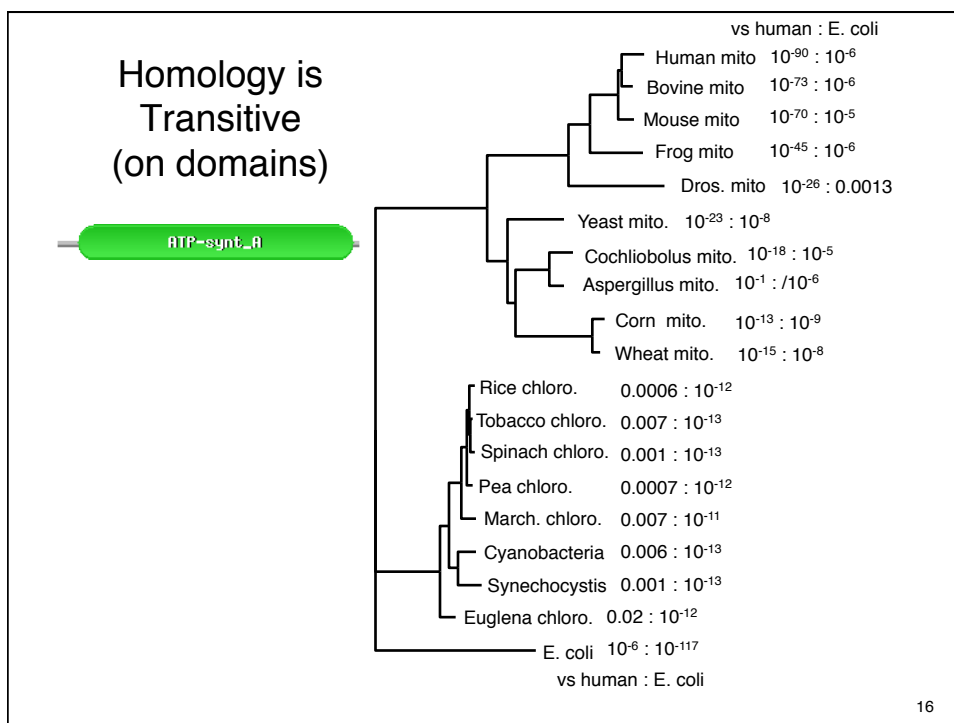
(sp|P24966|CYB_TAYTA Cytochrome b (379) 113 33.0 0.13 0.234 0.532 158)
(sp|P03892|NU2M_BOVIN NADH-ubiquinone oxidored (347) 107 31.7 0.31 0.261 0.479 211)
(sp|P68092|CYB_STEAT Cytochrome b (379) 104 31.0 0.54 0.277 0.547 137)
(sp|P03891|NU2M_HUMAN NADH-ubiquinone oxidored (347) 103 30.8 0.58 0.201 0.537 149)
(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)
    
```

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 NULM_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

15



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	832
PCAF_MOUSE	Histone acetyltransferase PCAF;	(813)	4507 1010	0	0.929	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
B2_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

17

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

18

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

19

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

20

Similarity Searching II

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

21

1. What question to ask?

- Is there an homologous protein (a protein with a similar structure)?
- Does that homologous protein have a similar function?
- Does XXX genome have YYY (kinase, GPCR, ...)?

Questions not to ask:

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

22

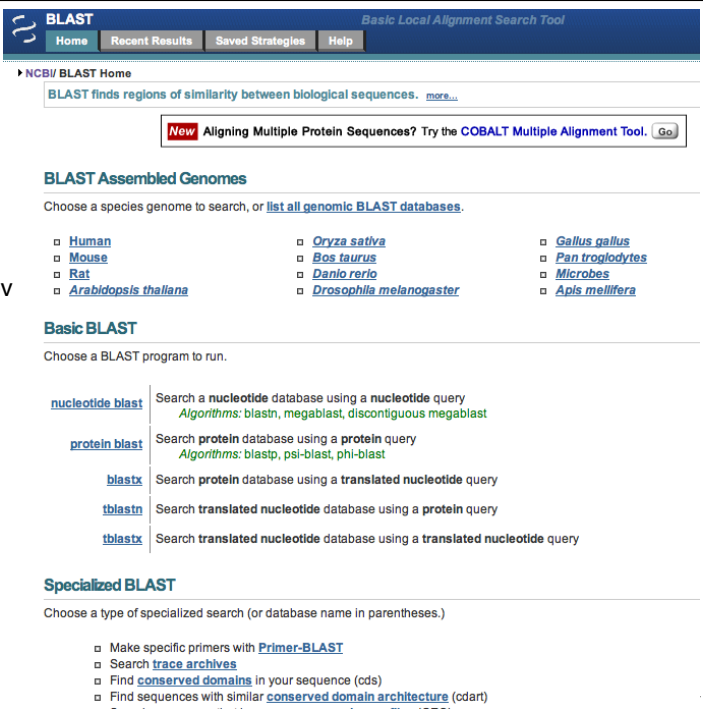
2. What program to run?

- What is your query sequence?
 - protein – BLAST (NCBI), SSEARCH (EBI)
 - protein coding DNA (EST) – BLASTX (NCBI), FASTX (EBI)
 - DNA (structural RNA, repeat family) – BLASTN (NCBI), FASTA (EBI)
- Does XXX genome have YYY (protein)?
 - TBLASTX YYY vs XXX genome
 - TFASTX YYY vs XXX genome
- Does my protein contain repeated domains?
 - LALIGN (UVA <http://fasta.bioch.virginia.edu>)

23

NCBI
BLAST
Server

blast.ncbi.nlm.nih.gov



BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. [Go](#)

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- [Human](#)
- [Mouse](#)
- [Rat](#)
- [Arabidopsis thaliana](#)
- [Oryza sativa](#)
- [Bos taurus](#)
- [Danio rerio](#)
- [Drosophila melanogaster](#)
- [Gallus gallus](#)
- [Pan troglodytes](#)
- [Microbes](#)
- [Apis mellifera](#)

Basic BLAST

Choose a BLAST program to run.

- [nucleotide blast](#) Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast
- [protein blast](#) Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast
- [blastx](#) Search **protein** database using a **translated nucleotide** query
- [tblastn](#) Search **translated nucleotide** database using a **protein** query
- [tblastx](#) Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)

NCBI BLAST Server

blast.ncbi.nlm.nih.gov

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

What is wrong with this picture?

Always compare protein sequences

25

NCBI
BLAST
Server

The screenshot shows the NCBI BLAST Server interface. The 'Basic BLAST' section is active, and the 'protein blast' option is selected. The form includes the following fields and options:

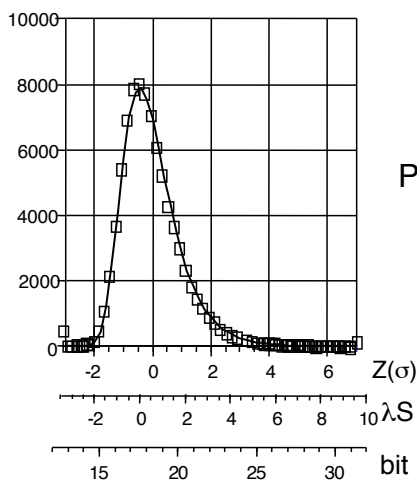
- Enter Query Sequence:** A text input field for the query sequence, with a 'Clear' button and a 'Query subrange' section with 'From' and 'To' input fields.
- Or, upload file:** A 'Choose File' button and a 'no file selected' status.
- Job Title:** A text input field for a descriptive title.
- Align two or more sequences:** A checkbox that is currently unchecked.
- Choose Search Set:**
 - Database:** A dropdown menu set to 'Non-redundant protein sequences (nr)'.
 - Organism:** A text input field with a note: 'Enter organism name or id--completions will be suggested'. There is an 'Exclude' checkbox and a '+' button.
 - Entrez Query:** A text input field with a note: 'Enter an Entrez query to limit search'.
- Program Selection:**
 - Algorithm:** Radio buttons for 'blastp (protein-protein BLAST)', 'PSI-BLAST (Position-Specific Iterated BLAST)', and 'PHI-BLAST (Pattern Hit Initiated BLAST)'. The 'blastp' option is selected.
- BLAST Button:** A blue button labeled 'BLAST' with the text 'Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)'. Below it is a checkbox for 'Show results in a new window'.
- Algorithm parameters:** A link to expand the algorithm parameters.

3. What database to search?

- Search the smallest comprehensive database likely to contain your protein
 - vertebrates – human proteins (40,000)
 - fungi – *S. cerevisiae* (6,000)
 - bacteria – *E. coli*, gram positive, etc. (<100,000)
- Search a richly annotated protein set (SwissProt, 450,000)
- Always search NR (> 12 million) *LAST*
- Never Search “GenBank” (DNA)

27

Why smaller databases are better – statistics



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

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

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

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

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

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

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

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

$$E(40 \mid D=12E6) = 1.8$$

28

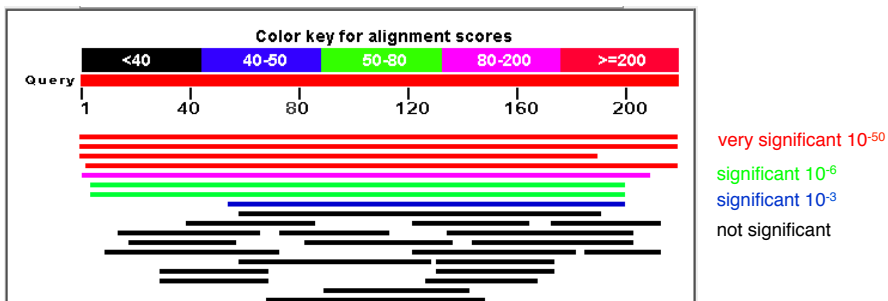
What is a “bit” score?

- Scoring matrices (PAM250, BLOSUM62, VTML40) contain “log-odds” scores:
 - $s_{i,j} \text{ (bits)} = \log_2(q_{i,j}/p_i p_j)$
 - $s_{i,j} \text{ (bits)} = 2 \rightarrow$ a residue is 4-times more likely to occur by homology compared with chance (at one residue)
 - $s_{i,j} \text{ (bits)} = -1 \rightarrow$ a residue is 2-times more likely to occur by chance compared with homology (at one residue)
- An alignment score is the maximum sum of $s_{i,j}$ bit scores across the aligned residues. A 40-bit score is 2^{40} more likely to occur by homology.
- How often should a score occur by chance? In a 400×400 alignment, there are 160,000 places where the alignment could start by chance, so we expect a score of 40 bits would occur:
 - $400 \times 400 \times 2^{-40} = 1.6 \times 10^5 / 2^{40} (10^{13.3}) = 0.8 \times 10^{-8}$ times
 - Thus, the probability of a 40 bit score in ONE alignment is $\sim 10^{-8}$
- But we did not ONE alignment, we did 4,000, 40,000, 400,000, or 16 million):
 - $E(p | D) = p(40 \text{ bits}) \times \text{database size}$
 - $E(40 | 4,000) = 10^{-8} \times 4,000 = 4 \times 10^{-5}$ (significant)
 - $E(40 | 40,000) = 10^{-8} \times 4 \times 10^4 = 4 \times 10^{-4}$ (significant)
 - $E(40 | 400,000) = 10^{-8} \times 4 \times 10^5 = 4 \times 10^{-3}$ (not significant)

29

How many “bits” do I need?

- $E(p | D) = p(40 \text{ bits}) \times \text{database size}$
- $E(40 | 4,000) = 10^{-8} \times 4,000 = 4 \times 10^{-5}$ (significant)
- $E(40 | 40,000) = 10^{-8} \times 4 \times 10^4 = 4 \times 10^{-4}$ (significant)
- $E(40 | 400,000) = 10^{-8} \times 4 \times 10^5 = 4 \times 10^{-3}$ (not significant)
- To get $E() \sim 10^{-3}$:
 - genome (10,000) $p \sim 10^{-3}/10^4 = 10^{-7}/160,000 = 40 \text{ bits}$
 - SwissProt (500,000) $p \sim 10^{-3}/10^6 = 10^{-9}/160,000 = 47 \text{ bits}$
 - Uniprot/NR (10^7) $p \sim 10^{-3}/10^7 = 10^{-10}/160,000 = 50 \text{ bits}$



30

Similarity Searching II

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

31

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

32

Smith-Waterman (ssearch)

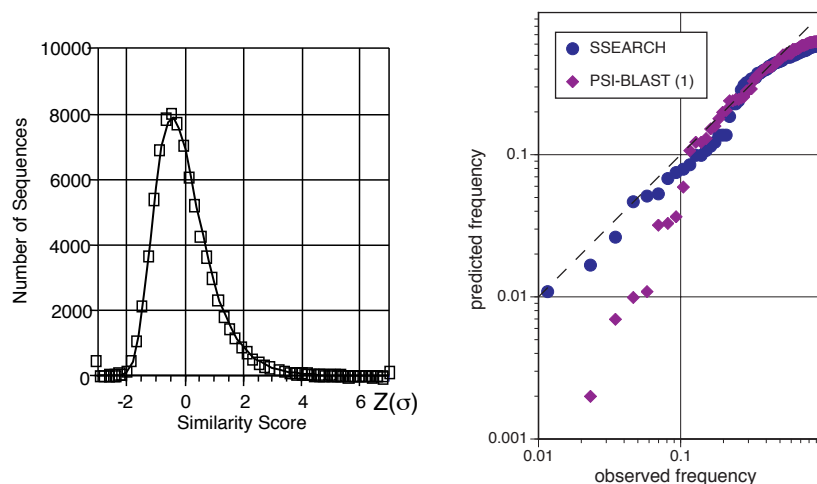
The best scores are:

			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 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
GTT2_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

Highest scoring unrelated sequence E() ~ 1.0

33

Homology inferences are reliable because Pairwise similarity statistics are accurate



Pairwise similarity statistics are accurate because
protein sequences are random

Validating homologs/statistics

- In general, BLASTP statistical estimates are accurate
- The most common errors occur because of low-complexity regions, or biased amino-acid composition
- To confirm statistical accuracy, find the highest scoring non homolog
 - No need to test every hit, test hits that are surprising
 - Confirm homology/non-homology by searching against a different comprehensive database, e.g. SwissProt, or refseq.
 - Non-homologs will find many significant members of other families, but not the family you are testing for
- Statistical estimates can be confirmed with shuffles (see ISMB2000 tutorial, fasta.bioch.virginia.edu/fasta_www2 shuffle link)

35

Scoring matrices

- Scoring matrices can set the evolutionary look-back time for a search
 - Lower PAM (PAM10/MDM10 ... PAM60) for closer (10% ... 50% identity)
 - Higher BLOSUM for higher conservation (BLOSUM50 distant, BLOSUM80 conserved)
- Shallow scoring matrices for short domains/short queries (metagenomics)
 - Matrices have “bits/position” (score/position), 40 aa at 0.7 bits/position (BLOSUM62) means 28 bit max score (50 bits significant)
- Deep scoring matrices allow alignments to continue, possibly outside the homologous region

36

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

37

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

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

$$p_R = 0.051$$

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

$$p_N = 0.043$$

$$I_2 S_{ij} = \lg_2 (q_{ij}/p_i p_j) \quad I_e S_{ij} = \ln(q_{ij}/p_i p_j) \quad p_R p_N = 0.002193$$

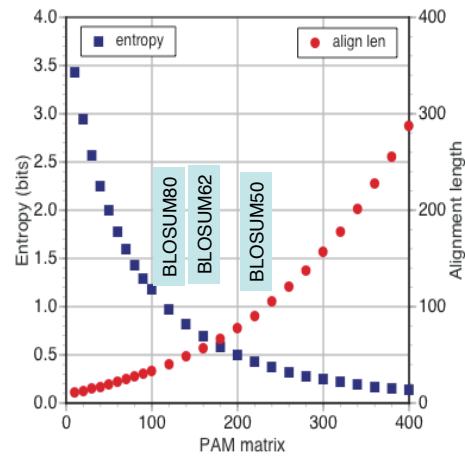
$$I_2 S_{R:N(40)} = \lg_2 (0.000435/0.002193) = -2.333$$

$$I_2 = 1/3; S_{R:N(40)} = -2.333/I_2 = -7$$

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

38

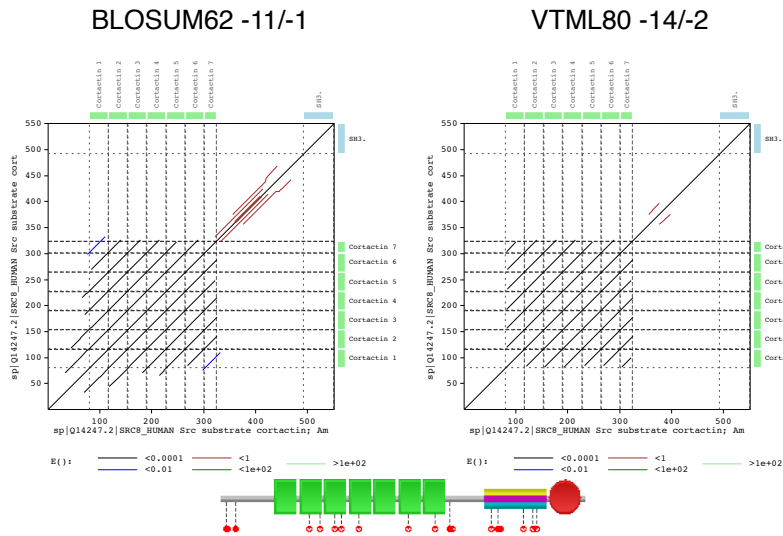
PAM matrices and alignment length



Short domains require “shallow” scoring matrices

39

Scoring matrices affect alignment boundaries



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

41

Sequence Similarity II - Conclusions

- Always compare Protein Sequences
 - use BLASTP or SSEARCH for protein-protein
 - blastx or fastx for DNA protein
- Search small (comprehensive) databases – never search NR or GenBank
- Scoring matrices set evolutionary look back horizons - not every discovery is distant
- Shallow scoring matrices for short domains
- Accurate statistics => highest unrelated E() ~ 1.0

42

Homework

The focus of these exercises is to explore the accuracy of BLAST (and SSEARCH) evaluates.

1. Write a program to generate random protein sequences of length 300. Produce the random sequences using two strategies using functions in the `~wrp/new_data/random_aa.pm` perl module (which will not be available until Friday). (a) `random_aa($length)` returns a random sequence; (b) `shuffle_aa($sequence)` takes an existing sequence and shuffles it.
2. Produce 20 random sequences from method 1a and 20 using method 1b. Use these random sequence to search Swissprot with `blastp`. For each search, save the evaluate of the highest scoring sequence.
3. Plot (using whatever method you like) the 20 evaluates from each method, sorted low to high vs 1/20, 2/20, ..., 20/20 on the x-axis.
4. Also do the searching/plotting analysis using `/seqprg/bin/ssearch36` (use the `-m 8` option to produce blast tabular output; use `ssearch36` to see the command line options; libraries for `ssearch` are in `/data/slib/fa_dbs/`, e.g. `swissprot.lseg`).