Similarity Searching II

Algorithms, scoring matrices, statistics

Biol4230 Tues, Jan 31, 2017

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

Goals of today's lecture:

- Quick overview of alignment algorithms
  - local vs global
  - dynamic programming
  - gaps and alignment graphs
  - non-overlapping local alignments
- Where scoring matrices come from
  - scoring matrices as log-odds matrices
  - short alignments, shallow matrices
  - shallow matrices, higher identity alignment
  - matrix "depth" and evolutionary look-back
- Improving search performance - local alignment statistics
  - the extreme value distribution
  - why database size matters
  - evaluating statistical accuracy

To learn more:

- Alignment algorithms:
  - Bioinformatics and Functional Genomics (BFG), Ch. 3 p 76 – 80
- Search sensitivity:
  - Sierk and Pearson (2005) "The limits of protein sequence comparison?" Curr Opin Struct Biol. 15:254-260.
- Statistical accuracy:
  - Sierk and Pearson (2005) Curr Opin Struct Biol. 15:254-260
  - BFG Ch. 3, pp 88 – 90
- Scoring matrices part I
  - BFG Ch. 3, pp. 57 – 76
  - Altschul (1991) J. Mol. Biol. 219:555-565
  - Pearson (2013) Curr Protocols Bioinformatics 3.5.1-3.5.9

Similarity searching II – algorithms, statistics, and scoring matrices

- Global and local alignments
  - Global alignments can be more sensitive for globally similar proteins
  - Local alignments are robust to partial sequences, domain homologies
- Local similarity scores are well described by the extreme value distribution
  - E()-value depends on similarity score AND database size
  - A 50 bit score is almost always significant
  - E()-values are not good measures of evolutionary distance
- Scoring matrices can be designed for long (deep) or short (shallow) evolutionary distances (large/small amounts of change)
  - "shallow" matrices provide more statistical significance for each aligned position, but require higher homologs
  - "deep" matrices can find more distant homologs, but require longer alignments

fasta.bioch.virginia.edu/biol4230 3

Algorithms for sequence alignment

- How do we get from this:

```
>ATP6_HUMAN ATP synthase a chain (ATPase protein 6)
MNENLFASFIAPTILGLPAAVLIIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTKGRTWSL
MLVSLIIFIATTNLLGLLPHSFTPTTQLSMNLAMAIPLWAGTVIMGFRSKIKNALAHFLPQGTPTPLIPM
LVIIETISLLIQPMALAVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTILILLTILEIAVALIQ
AYVFTLLVSLYLHDNT
```

- And this:

```
>sp|P0AB98|ATP6_ECOLI ATP synthase subunit a
MASENMTPQDYIGHHLNNLQLDLRTFSLVDPQNPPATFWTINIDSMFFSVVLGLLFLVLFRSVAKKATSGV
PGKFQTAIELVIGFVNGSVKDMYHGKSKLIAPLALTIFVWVFLMNLMDLLPIDLLPYIAEHVLGLPALRVV
PSADVNVTLSMALGVFILILFYSIKMKGIGGFTKELTLQPFNHWAFIPVNLILEGVSLLSKPVSLGLRLFG
NMYAGELIFILIAGLLPWWSQWILNVPWAIFHILIITLQAFIFMVLTIVYLSMASEEH
```

- To …

fasta.bioch.virginia.edu/biol4230 4

Algorithms for sequence alignment

- To this:

```
>sp|P0AB98|ATP6_ECOLI ATP synthase subunit a; ATP synthase F0 aubunit;
Length=271
 Score = 47.9 bits (178),  Expect = 3e-06
 Identities = 55/199 (27%), Positives = 113/199 (56%), Gaps = 37/199 (18%)

Query  8    SFIAPTILGLPAAVLIIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTKGRTWSLML  72
            S     +LGL   ++++LF   +     +    +  ++ T   + +I + +  +  M++ K +   + +
Sbjct  45   SMFFSVVLGL---LFLVLFRSVAKKATSG-VPGKFQTAIELVIGFVNGSVKDMYHGKSKLIAPLA 105

Query  73   VSLIIFIAT                                    PLWAGTVIMGFRSKI 121
            +++ +++                                        ++    +++ F S
Sbjct  106  LTIFVWVFL                                    GVF---ILILFYSIK 167

Query  122  KNALAHFLP                                    GHLLMHLIGSATLAM 181
               +  F                                      G L+  LI
Sbjct  168  MKGIGGFTK                                    GELIFILIAGLLPWW 232

Query  182  STINLPSTLIIFTILILLTILEIAVALIQAYVFTLLVSLYL 222
            S   L     IF ILI+          +QA++F +L  +YL
Sbjct  233  SQWILNVPWAIFHILIIT---------LQAFIFMVLTIVYL 264
```

We need:

(1) Alignment algorithm

(2) Scoring Matrix

(3) Statistical model

fasta.bioch.virginia.edu/biol4230

5

Local, global, and "glocal" alignments

- Global alignments go from include the entire length of both sequences (Needleman-Wunsch, 1970)
  - high global similarity = small sequence distance (100% identity = distance 0)
  - similarity scores can be negative
  - scores are (probably) normally distributed
  - single domain, approx. constant length proteins
  - GGSEARCH calculates "global" alignment scores
- Local alignments find the best match, regardless of the length of the match. (Smith-Waterman, 1981)
  - requires similarity scoring matrix with $E(s_{ij}) < 0.0$
  - all similarity scores are > 0.0
  - scores are extreme value distributed
  - good for partial sequences, homologous domains with sequences
  - BLASTP, FASTA, and SSEARCH generate "local" alignment scores
- "glocal" alignments are "global" in the query (e.g. a domain), but local in the subject
  - a domain within a protein
  - GLSEARCH

fasta.bioch.virginia.edu/biol4230

6

Local, global, and "glocal" alignments

Local – 26.3% id
E() < 0.00024

Globally similar:

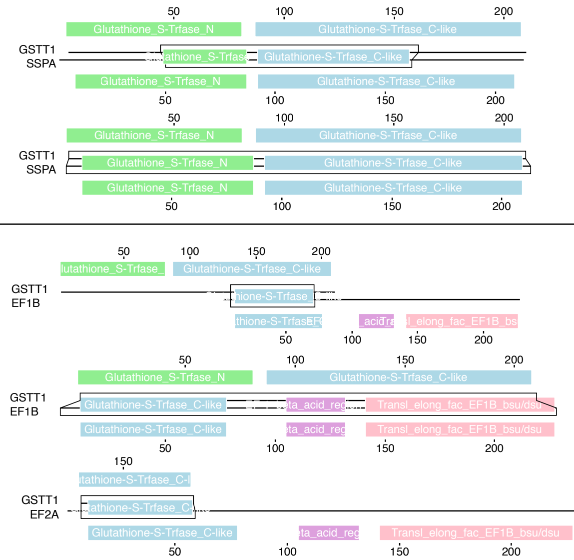
Global – 20.6% id
E() < 10^{-7}

Local – 29.2% id
E() < 9

Locally similar:

Global – 15.3% id
E() < 7000

Glocal – 26.8% id
E() < 0.02



fasta.bioch.virginia.edu/biol4230

7

Dynamic programming for sequence alignment

- Sequence alignments can be *global* – end-to-end, or *local*
- The *Dynamic Programming Algorithm* allows one to examine 2^{2n} alignments (n=100, 10^{77}) in $O(n^2)$ (n=100, $O(n^2)$=10,000) time
- Local alignments can also be used to find duplicated domains in proteins

fasta.bioch.virginia.edu/biol4230

8

Algorithms for Global and Local Similarity Scores

Global:

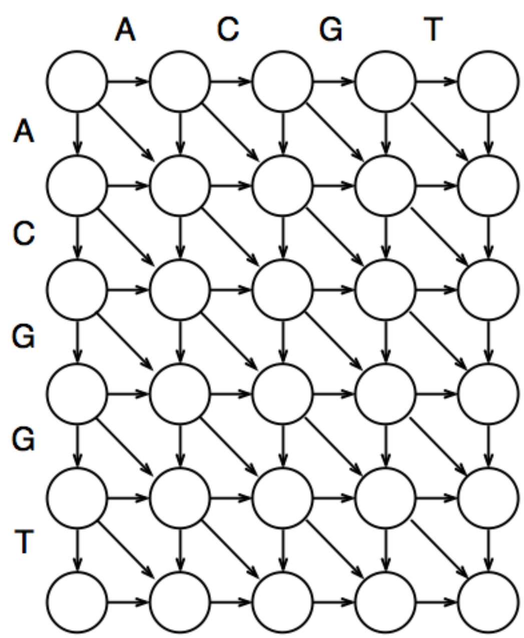
$S(0,0) \leftarrow 0$
for $j \leftarrow 1$ **to** N **do**
  $S(0,j) \leftarrow S(0,j-1) + \sigma\binom{-}{b_j}$
for $i \leftarrow 1$ **to** M **do**
[$S(i,0) \leftarrow S(i-1,0) + \sigma\binom{a_i}{-}$
  for $j \leftarrow 1$ **to** N **do**
    $S(i,j) \leftarrow max[S(i-1,j-1) + \sigma\binom{a_i}{b_j}, S(i-1,j) + \sigma\binom{a_i}{-}, S(i,j-1) + \sigma\binom{-}{b_j}]$
]
write "Global similarity score is" $S(M,N)$

Local:

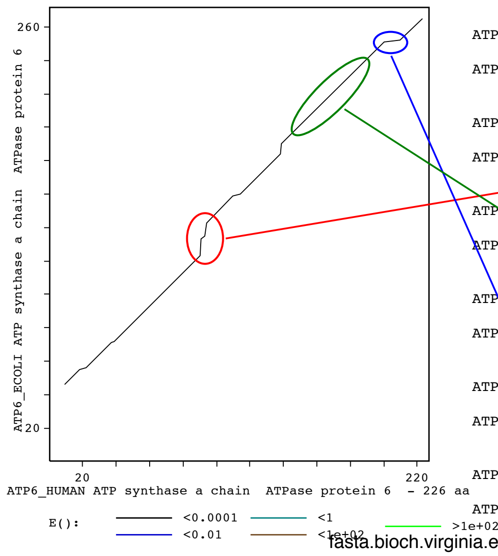
$best \leftarrow 0$
for $j \leftarrow 1$ **to** N **do**
  $S'(0,j) \leftarrow 0$
for $i \leftarrow 1$ **to** M **do**
[$S'(i,0) \leftarrow 0$
  for $j \leftarrow 1$ **to** N **do**
  [$S'(i,j) \leftarrow max[0, S'(i-1,j-1) + \sigma\binom{a_i}{b_j}, S'(i-1,j) + \sigma\binom{a_i}{-}, S'(i,j-1) + \sigma\binom{-}{b_j}]$
    $best \leftarrow max(S'(i,j), best)$
]
write "Local similarity score is" $best$

+1 : match
-1 : mismatch
-2 : gap

align_path2



alignment paths highlight indels

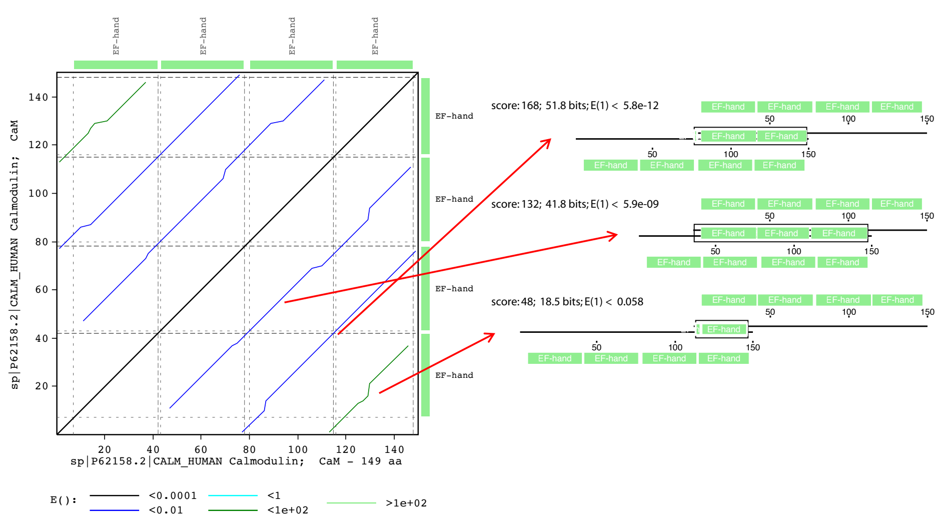



```
>>sp|P0AB98|ATP6_ECOLI ATP synthase (271 aa)
Smith-Waterman score: 178; E(): 2.1e-06
 23.3% identity in 236 aa overlap (8-222:45-264)
                                           10
ATP6_H                         MNENLFASFIAPTILGL
                                      :..  ..:::
ATP6_E HLNNLQLDLRTFSLVDPQNPPATFWTINIDSMFFSVVLGL
             20        30        40        50
                 20        30        40        50
ATP6_H PAAVLIIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQM
          ....::  .   ...  . ... :. . .: ... ..
ATP6_E ---LFLVLFRSVAKKATSG-VPGKFQTAIELVIGFVNGSV
            60        70        80        90
        60        70        80
ATP6_H MTMHNTKGRTWSLMLVSLIIFIATTNLLGLLP--------
        :.. :..  . . .......   ::. :::
ATP6_E KDMYHGKSKLIAPLALTIFVWVFLMNLMDLLPIDLLPYIA
            100       110       120       130
       90               100       110       120
ATP6_H -HSF-------TPTTQLSMNLAMAIPLWAGTVIMGFRSKI
        : .        .:.......:.::. ..   ... : :
ATP6_E EHVLGLPALRVVPSADVNVTLSMALGVF---ILILFYSIK
            140       150          160
           130            140       150
ATP6_H KNALAHFLPQGTPTPL-----IPMLVIIETISLLIQPMAL
        .... :  . :  :.      ::. .:.: .::: .:..:
ATP6_E MKGIGGFTKELTLQPFNHWAFIPVNLILEGVSLLSKPVSL
       170       180       190       200
        160       170       180       190
ATP6_H AVRLTANITAGHLLMHLIGSATLAMSTINLPSTLIIFTIL
       ..:: .:. ::.:.. ::..      :   :     :: ::
ATP6_E GLRLFGNMYAGELIFILIAGLLPWWSQWILNVPWAIFHIL
       210       220       230       240
        200       210       220
ATP6_H ILLTILEIAVALIQAYVFTLLVSLYLHDNT
       :.          .::..: .:. .::
ATP6_E IIT---------LQAFIFMVLTIVYLSMASEEH
       250                 260       270
```

fasta.bioch.virginia.edu/biol4230

11

LALIGN – non-overlapping local alignments can identify mobile domains



Scoring matrices

- Scoring matrices are derived from log-odds scores:
  - log(freq. of change in homolog/freq. alignment by chance)
- Scoring matrices can set the evolutionary look-back time for a search
  - Lower PAM (PAM10/VT10 … PAM/VT40) for closer (10% … 50% identity)
  - less evolution, lower frequency of change, higher freq. of 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 0.45 bits/position (BLOSUM62) means 18 bit ave. score (50 bits significant)
- Deep scoring matrices allow alignments to continue, possibly outside the homologous region

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

$$\boxed{\lambda S_{i,j} = \log_b\left(\frac{q_{i,j}}{p_i p_j}\right)}$$

q_{ij} : replacement frequency at PAM40, 250

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

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

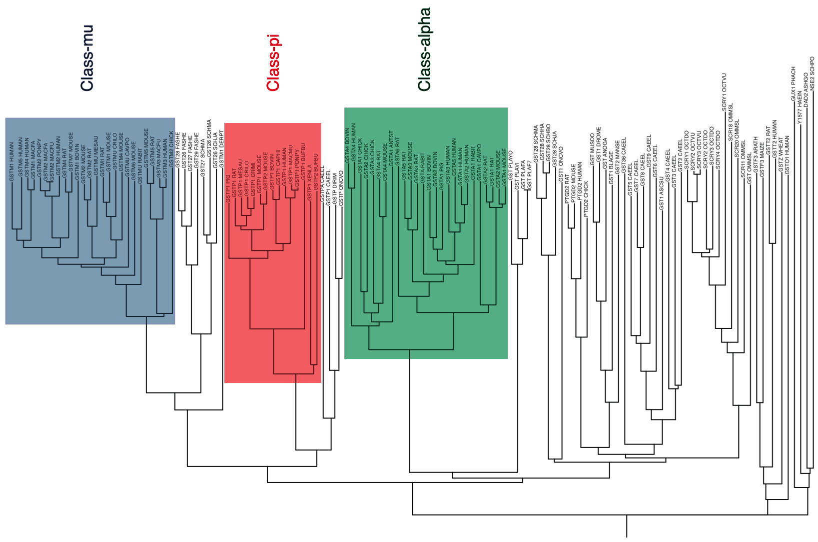
$l_2\ S_{ij} = \lg_2\ (q_{ij}/p_ip_j)$ $l_e\ S_{ij} = \ln(q_{ij}/p_ip_j)$ $p_Rp_N = 0.002193$

$l_2\ S_{R:N(40)} = \lg_2\ (0.000435/0.00219) = -2.333$

$l_2 = 1/3$; $S_{R:N(40)} = -2.333/l_2 = -7$

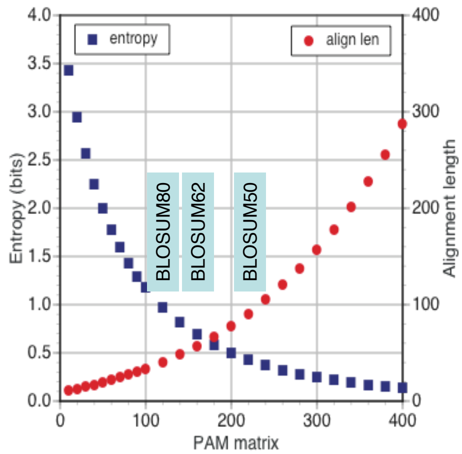
$l\ S_{R:N(250)} = \lg2\ (0.002193/0.002193) = 0$

Scoring matrices set look back time: Glutathione Transferases (gstm1_human)



		BLOSUM50-10/-2 E(320363)	f_id	BLOSUM62-11/-1 E(320363)	f_id	VT40 -21/-4 E(320363)	f_id	VT10 -23/-4 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
	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
Class-pi	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		
	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				
Class-alpha	GSTA1_MOUSE	1.5e-08	0.279	4.9e-11	0.264				
	GSTA2_HUMAN	6.6e-08	0.286	1.2e-08	0.273				
	GSTA5_HUMAN	7.8e-08	0.275	1.2e-08	0.259				
	GSTA2_MOUSE	1.1e-07	0.269	9.9e-10	0.255				
	GSTA3_MOUSE	1.3e-07	0.278	8.9e-09	0.258				
	GSTA1_HUMAN	3.0e-07	0.272	8.0e-08	0.259				
	GST36_CAEEL	3.3e-07	0.256	1.1e-08	0.264				
	GSTA2_CHICK	4.2e-07	0.279	8.0e-08	0.266				

PAM matrices and alignment length



Short domains require "shallow" scoring matrices

Altschul (1991) "Amino acid substitution matrices from an information theoretic perspective" J. Mol. Biol. 219:555-565

Empirical matrix performance (median results from random alignments)

Matrix	target % ident	bits/position	aln len (50 bits)
VT160 -12/-2	23.8	0.26	192
BLOSUM50 -10/-2	25.3	0.23	217
BLOSUM62* -11/-1	28.9	0.45	111
VT120 -11/-1	27.4	1.03	48
VT80 -11/-1	51.9	1.55	32
PAM70* -10/-1	33.8	0.64	78
PAM30* -9/-1	45.5	1.06	47
VT40 -12/-1	72.7	2.76	18
VT20 -15/-2	84.6	3.62	13
VT10 -16/-2	90.9	4.32	12

HMMs can be very "deep"

Pearson (2013) Curr Protoc. Bioinfo 3.5.1-3.5.9

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
- Shallow matrices set maximum look-back time
- Short alignments (domains, exons, reads) require shallow (higher information content) matrices

Improving Similarity Searching (Similarity Statistics)

- What gets missed? / What shouldn't be found
  - comparing sequence and structural similarity
  - what is a "non-homolog"?
- Homology from "significance" – local alignment statistics
  - E()-values and bit-scores
- Use protein databases
  - smaller
  - more sensitive
  - better statistics

How well does BLAST work?

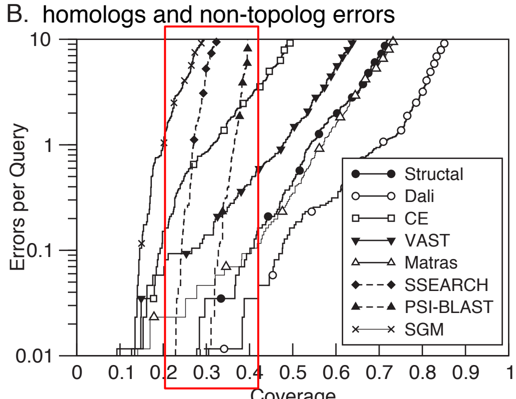
Gold standard – homologous proteins ALWAYS share statistically significant structural similarity

- databases of structures: SCOP (structural classification of proteins)
- CATH (Class, Architecture, Topology, Homology)
  - All "Homologs" are "homologous"
  - Some "Topologs" might be homologous
  - Architecture without similar topology, non-homologous

How well are homologs identified?

- Structure comparison:
  - DALI, VAST, MATRAS, CE, STRUCTAL, SGM
- Pairwise sequence comparison:
  - SSEARCH
- Model-based sequence comparison:
  - PSI-BLAST

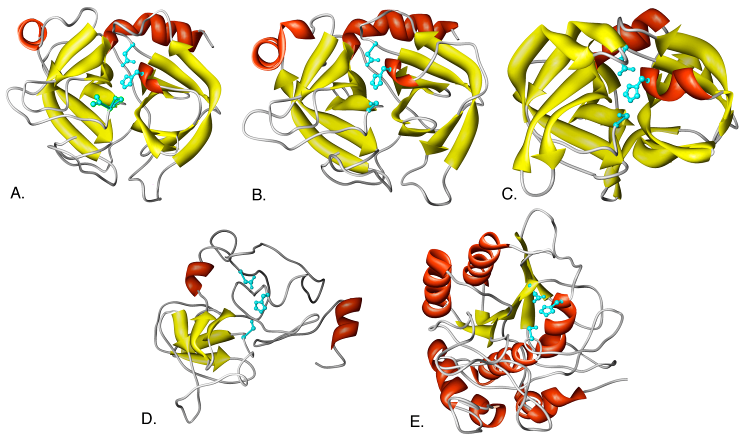
B. homologs and non-topolog errors



Sierk and Pearson (2004) Prot. Sci. 13:773

What is a non-homolog?

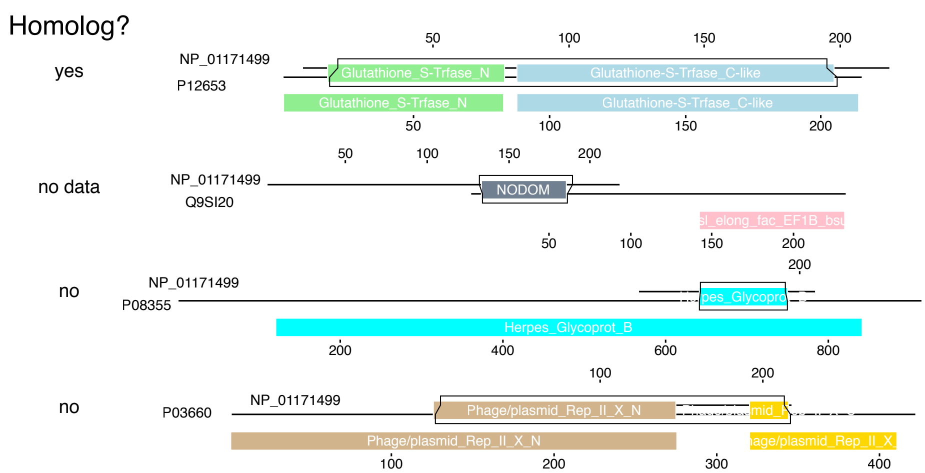
Five serine proteases: three trysin like (A, B, C, homologs), subtilisin (E, non-homolog), and ? (D)



fasta.bioch.virginia.edu/biol4230

23

Non-homologs have different domains



domain annotations use methods that are more sensitive than pairwise sequence alignment

fasta.bioch.virginia.edu/biol4230

24

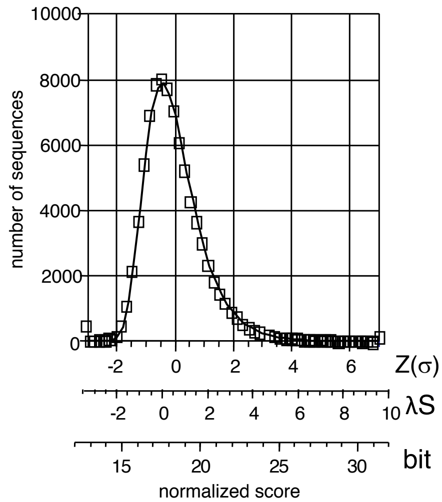
Improving sensitivity by improving statistical significance

- Local similarity scores follow the "extreme value distribution"
  - unrelated ➔ random, thus:
  - not random ➔ homologous
  - random == extreme value distribution
- improve sensitivity with smaller databases
- can we trust the statistics?

Smaller databases for more sensitive searches which 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 (> 80 million) *LAST*
- Never Search “GenBank” (DNA)

Why smaller databases are better (more sensitive) – statistics



$S' = \lambda S_{raw} - \ln K\, m\, n$
$S_{bit} = (\lambda S_{raw} - \ln K)/\ln(2)$
$P(S'>x) = 1 - \exp(-e^{-x})$
$P(S_{bit} > x) = 1 - \exp(-mn2^{-x})$
$E(S'>x \mid D) = P\, D$

$P(B \text{ bits}) = m\, n\, 2^{-B}$
$P(40 \text{ bits}) = 1.5\times10^{-7}$
$E(40 \mid D=4000) = 6\times10^{-4}$
$E(40 \mid D=80E6) = 12$

fasta.bioch.virginia.edu/biol4230 27

Local similarity statistics

$S' = \lambda S_{raw} - \ln K\, m\, n$ m: query length, n: subj length
$S_{bit} = (\lambda S_{raw} - \ln K)/\ln(2)$
$P(S'>x) = 1 - \exp(-e^{-x})$
$P(S'>x) = e^{-x}$ (for $P < 0.1$)

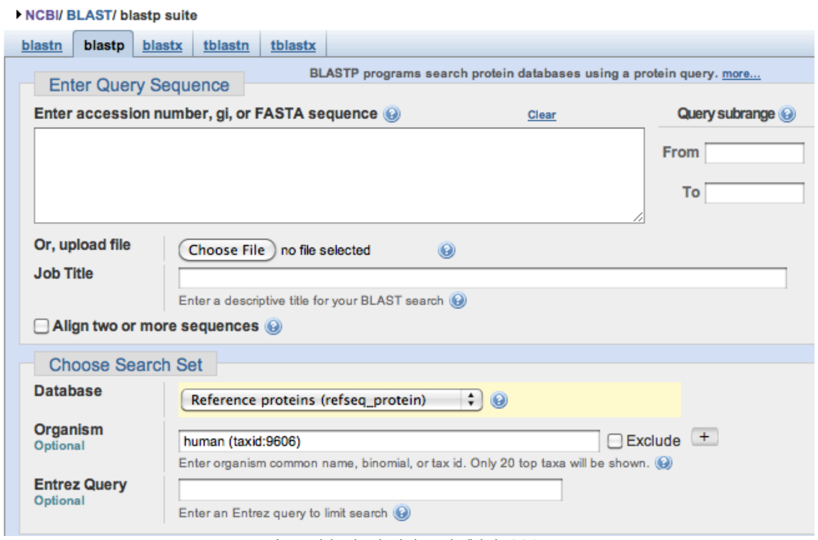
$P(S_{bits} > \text{bits}) = 1 - \exp(-mn2^{-x})$
$P(S_{bits} > \text{bits}) = mn2^{-\text{bits}}$ (for $P < 0.1$)

$E(S', S_{bits} \mid D) = P\, D$
$E(S_{bits} \mid D) = D\, mn2^{-\text{bits}}$ Bonferroni correction

dblength = $\sum n$ or (Dn)
$E(S_{bit}) = m\, \text{dblength}\, 2^{-\text{bits}}$ (BLAST formula)

BIMS6000 - Searching II 28

NCBI – selecting sequences with Entrez



Bits and significance

- 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 than by chance.
- 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: $P(S_{bit} > x) = 1 - \exp(-mn2^{-x}) \sim mn2^{-x}$

  $400 \times 400 \times 2^{-40} = 1.6 \times 10^5 / 2^{40} (10^{13.3}) = 1.5 \times 10^{-7}$ times

  Thus, the probability of a 40 bit score in ONE alignment is $\sim 10^{-7}$
- But we did not ONE alignment, we did 4,000, 40,000, 400,000, or 16 million alignments when we searched the database:

  $E(S_{bit} \mid D) = p(40\text{ bits}) \times \text{database size}$

  $E(40 \mid 4{,}000) = 10^{-7} \times 4{,}000 = 4 \times 10^{-4}$ (significant)

  $E(40 \mid 40{,}000) = 10^{-7} \times 4 \times 10^4 = 4 \times 10^{-3}$ (not significant)

  $E(40 \mid 400{,}000) = 10^{-7} \times 4 \times 10^5 = 4 \times 10^{-2}$ (not significant)

  $E(40 \mid 16\text{ million}) = 10^{-7} \times 1.6 \times 10^7 = 1.6$ (not significant)

How many “bits” do I need?

E(p I D) = p(40 bits) x database size

E(40 I 4,000) = 10^{-8} x 4,000 = 4 x 10^{-5} (significant)

E(40 I 40,000) = 10^{-8} x 4 x 10^{4} = 4 x 10^{-4} (significant)

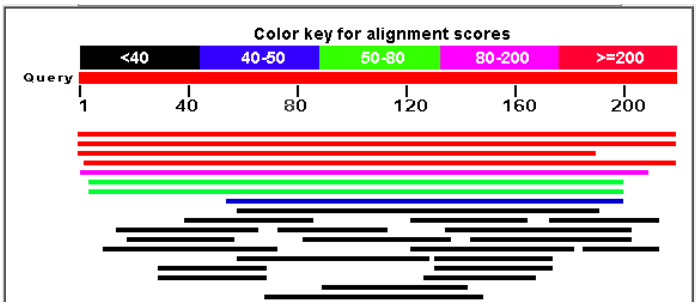
E(40 I 400,000) = 10^{-8} x 4 x 10^{5} = 4 x 10^{-3} (not significant)

To get E() ~ 10^{-3} :

genome (10,000) p ~ $10^{-3}/10^{4}$ = 10^{-7}/160,000 = 40 bits

SwissProt (500,000) p ~ $10^{-3}/10^{6}$ = 10^{-9}/160,000 = 47 bits

Uniprot/NR (10^{7}) p ~ $10^{-3}/10^{7}$ = 10^{-10}/160,000 = $\boxed{\text{50 bits}}$



very significant 10^{-50}

significant 10^{-6}

significant 10^{-3}

not significant

fasta.bioch.virginia.edu/biol4230 31

Should you trust the E()-value??
(what is the *control* for this *experiment*)

- The inference of homology from statistically significant similarity depends on the observation that unrelated sequences look like random sequences
  - Is this ALWAYS true?
  - How can we recognize when it is not true?
- If unrelated==random, then the E()-value of the highest scoring unrelated sequence should be E() ~ 1.0
- Statistical estimates can also be confirmed by searches against shuffled sequences

fasta.bioch.virginia.edu/biol4230 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 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
```

33

Breaking the statistics: 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
```

34

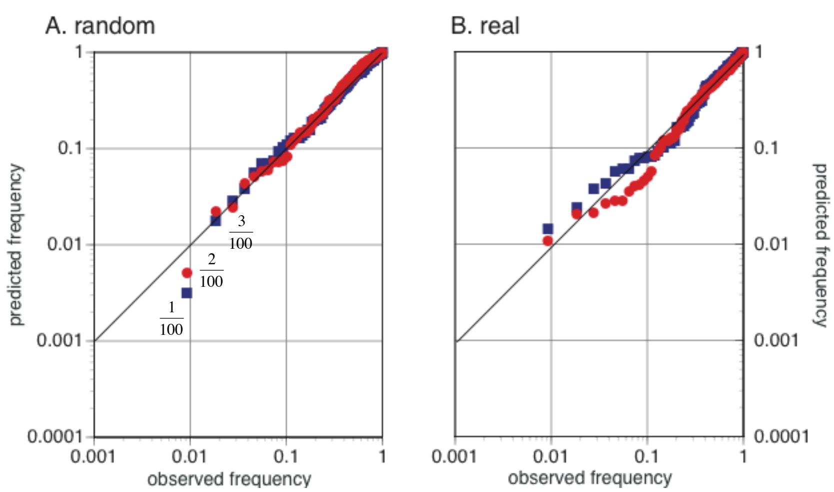
pseg removes low-complexity regions

>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

35

Protein Sequence Comparison Statistics are Accurate



36

E()-values when??

- E()-values (BLAST expect) provide accurate statistical estimates of similarity by chance
  - non-random -> not unrelated (homologous)
  - E()-values are accurate (0.001 happens 1/1000 by chance)
  - E()-values factor in (and depend on) sequence lengths and database size
- E()-values are NOT a good proxy for evolutionary distance
  - doubling the length/score SQUARES the E()-value
  - percent identity (corrected) reflects distance (given homology)

Similarity searching II – algorithms, statistics, and scoring matrices

- Global and local alignments
  - Global alignments can be more sensitive for globally similar proteins
  - Local alignments are robust to partial sequences, domain homologies
- Scoring matrices can be designed for long (deep) or short (shallow) evolutionary distances (large/small amounts of change)
  - "shallow" matrices provide more statistical significance for each aligned position, but require higher homologs
  - "deep" matrices can find more distant homologs, but require longer alignments
- Local similarity scores are well described by the extreme value distribution
  - E()-value depends on similarity score AND database size
  - A 50 bit score is almost always significant
  - E()-values are not good measures of evolutionary distance