Programming for Biology
Similarity Searching II –

Practical search strategies

Bill Pearson
wrp@virginia.edu

Protein Evolution and Sequence Similarity

Similarity Searching I

- What is Homology and how do we recognize it?
- How do we measure sequence similarity – alignments and scoring matrices?
- DNA vs protein comparison

Similarity Searching II

- More effective similarity searching
  - Smaller databases
  - Appropriate scoring matrices
  - Using annotation/domain information

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
6. More sensitive methods (PSI-BLAST, HMMER)

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?

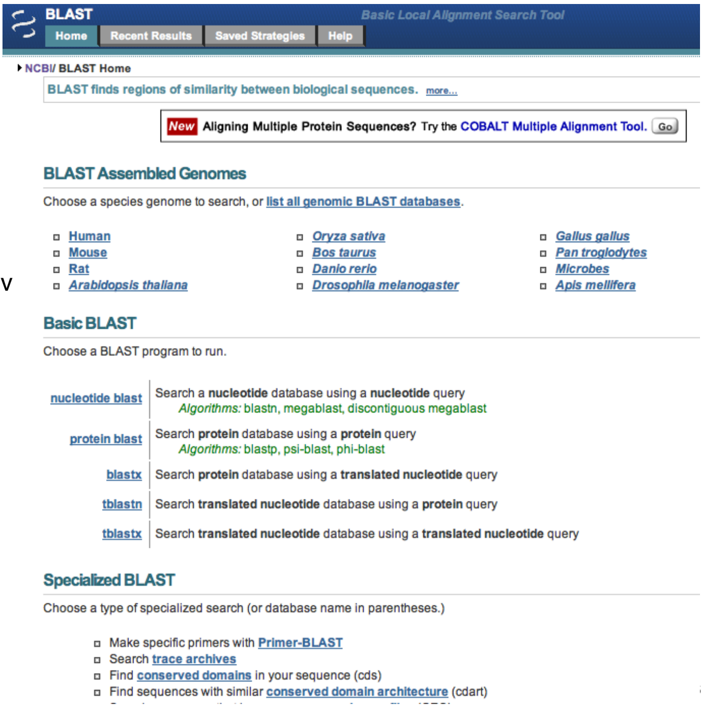
2. What program to run?

- 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)?
  - TBLASTN YYY vs XXX genome
  - TFASTX YYY vs XXX genome
- Does my protein contain repeated domains?
  - LALIGN (UVa http://fasta.bioch.virginia.edu)

5

NCBI BLAST Server

blast.ncbi.nlm.nih.gov



BLAST finds regions of similarity between biological sequences. more...

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool.

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, discontiguous 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

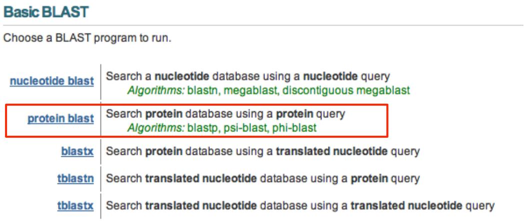
Specialized BLAST

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.

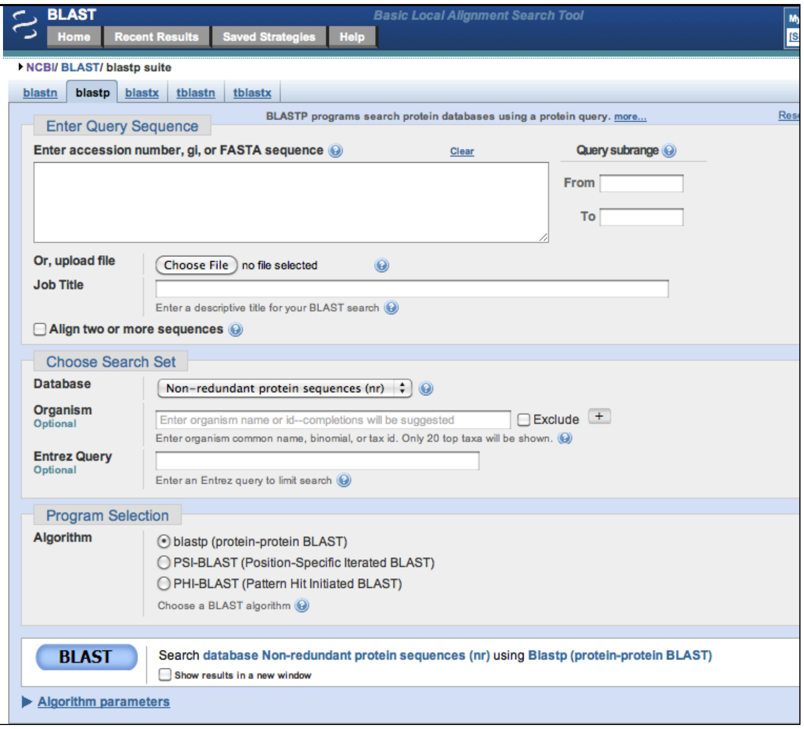
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

What is wrong with this picture?

Always compare protein sequences

7

NCBI BLAST Server



NCBI/ BLAST/ blastp suite

blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence — Clear

Query subrange — From — To

Or, upload file — Choose File no file selected

Job Title — Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database — Non-redundant protein sequences (nr)

Organism (Optional) — Enter organism name or id--completions will be suggested — □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional) — Enter an Entrez query to limit search

Program Selection

Algorithm

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

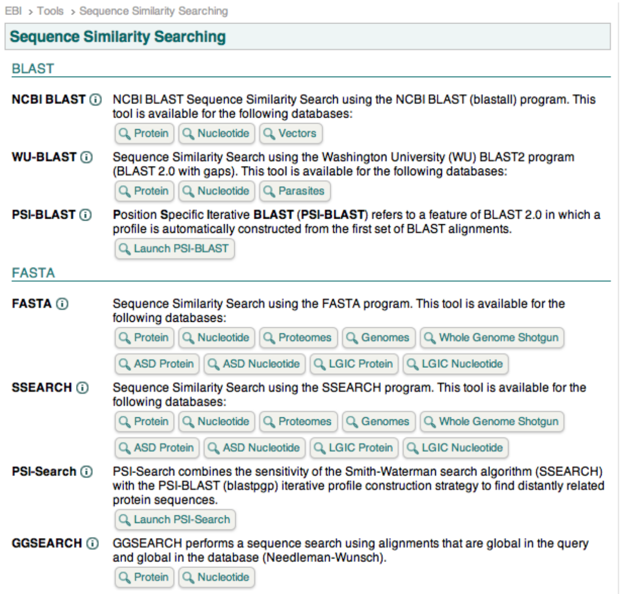
Choose a BLAST algorithm

BLAST — Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

□ Show results in a new window

▶ Algorithm parameters

Searching at the EBI
www.ebi.ac.uk/Tools/sss/



EBI > Tools > Sequence Similarity Searching

Sequence Similarity Searching

BLAST

NCBI BLAST ⓘ NCBI BLAST Sequence Similarity Search using the NCBI BLAST (blastall) program. This tool is available for the following databases:

WU-BLAST ⓘ Sequence Similarity Search using the Washington University (WU) BLAST2 program (BLAST 2.0 with gaps). This tool is available for the following databases:

PSI-BLAST ⓘ Position **S**pecific **I**terative **BLAST** (**PSI-BLAST**) refers to a feature of BLAST 2.0 in which a profile is automatically constructed from the first set of BLAST alignments.

FASTA

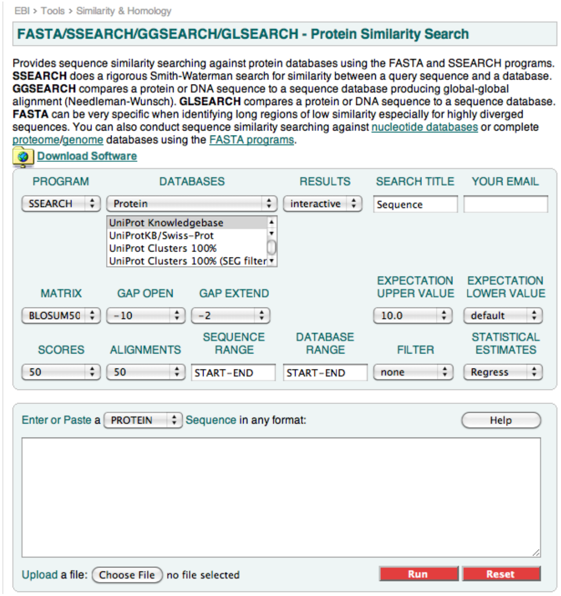
FASTA ⓘ Sequence Similarity Search using the FASTA program. This tool is available for the following databases:

SSEARCH ⓘ Sequence Similarity Search using the SSEARCH program. This tool is available for the following databases:

PSI-Search ⓘ PSI-Search combines the sensitivity of the Smith-Waterman search algorithm (SSEARCH) with the PSI-BLAST (blastpgp) iterative profile construction strategy to find distantly related protein sequences.

GGSEARCH ⓘ GGSEARCH performs a sequence search using alignments that are global in the query and global in the database (Needleman-Wunsch).

Searching at the EBI – ssearch



EBI > Tools > Similarity & Homology

FASTA/SSEARCH/GGSEARCH/GLSEARCH - Protein Similarity Search

Provides sequence similarity searching against protein databases using the FASTA and SSEARCH programs. **SSEARCH** does a rigorous Smith-Waterman search for similarity between a query sequence and a database. **GGSEARCH** compares a protein or DNA sequence to a sequence database producing global-global alignment (Needleman-Wunsch). **GLSEARCH** compares a protein or DNA sequence to a sequence database. **FASTA** can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity searching against nucleotide databases or complete proteome/genome databases using the FASTA programs.

Download Software

PROGRAM	DATABASES	RESULTS	SEARCH TITLE	YOUR EMAIL
SSEARCH	Protein UniProt Knowledgebase UniProtKB/Swiss-Prot UniProt Clusters 100% UniProt Clusters 100% (SEG filter	interactive	Sequence	

MATRIX	GAP OPEN	GAP EXTEND		EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
BLOSUM50	-10	-2		10.0	default
SCORES	**ALIGNMENTS**	**SEQUENCE RANGE**	**DATABASE RANGE**	**FILTER**	**STATISTICAL ESTIMATES**
50	50	START-END	START-END	none	Regress

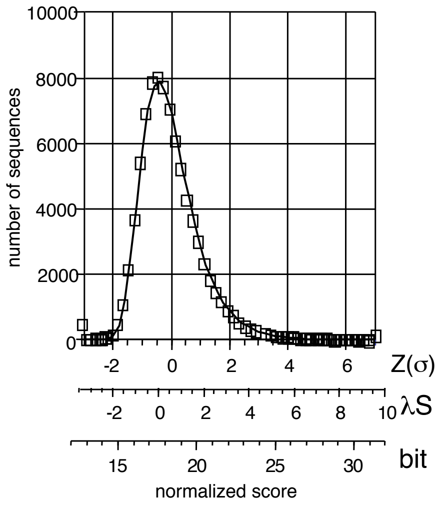
Enter or Paste a PROTEIN Sequence in any format:

Upload a file: Choose File no file selected

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)

Why smaller databases are better – statistics



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

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

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

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

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

What is a “bit” score?

- Scoring matrices (PAM250, BLOSUM62, VTML40) contain “log-odds” scores:
  $s_{i,j}$ (bits) = $\log_2(q_{i,j}/p_i p_j)$ ($q_{i,j}$ freq. in homologs/ $p_i p_j$ freq. by chance)
  $s_{i,j}$ (bits) = 2 -> a residue is 2^2=4-times more likely to occur by homology compared with chance (at one residue)
  $s_{i,j}$ (bits) = -1 -> a residue is 2^{-1} = 1/2 as likely to occur by homology compared with chance (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 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 ~ 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:
  $E(S_{bit} | D)$ = p(40 bits) x database size
  $E(40 | 4{,}000) = 10^{-7} \times 4{,}000 = 4 \times 10^{-4}$ (significant)
  $E(40 | 40{,}000) = 10^{-7} \times 4 \times 10^{4} = 4 \times 10^{-3}$ (not significant)
  $E(40 | 400{,}000) = 10^{-7} \times 4 \times 10^{5} = 4 \times 10^{-2}$ (not significant)
  $E(40 | 16 \text{ million}) = 10^{-7} \times 1.6 \times 10^{7} = 1.6$ (not significant)

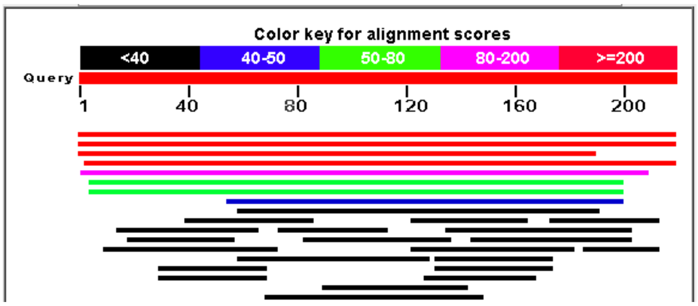
How many “bits” do I need?

E(p | D) = p(40 bits) x 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() ~ 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 ~ $10^{-3}/10^7 = 10^{-10}/160{,}000$ = 50 bits



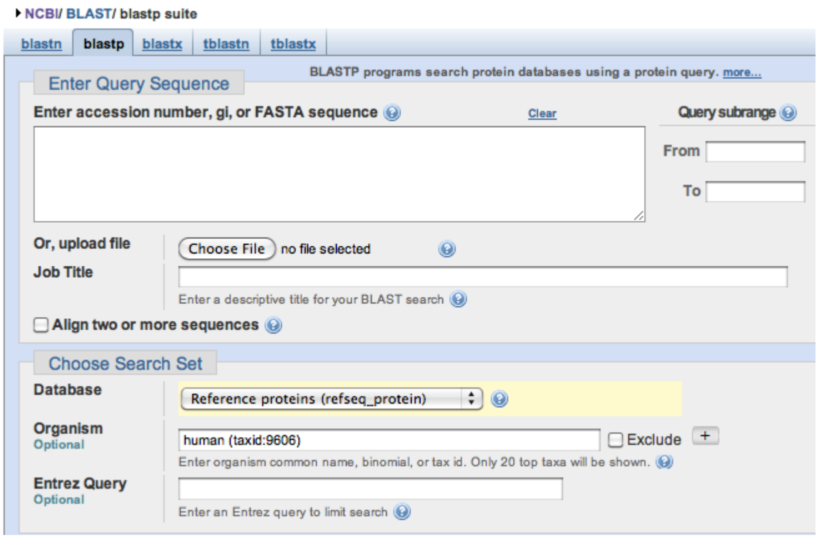
very significant 10^{-50}
significant 10^{-6}
significant 10^{-3}
not significant

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)

15

NCBI – selecting sequences with Entrez



BLASTP programs search protein databases using a protein query.

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Query subrange

From

To

Or, upload file: no file selected

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Database: Reference proteins (refseq_protein)

Organism (Optional): human (taxid:9606) □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional)

Enter an Entrez query to limit search

16

Effective Similarity Searching

1. Always search protein databases (possibly with translated DNA)
2. Use E()-values, not percent identity, to infer homology
   - E() < 0.001 is significant in a single search

---

3. Search smaller (comprehensive) databases
4. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
5. All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss

Scoring matrices

- Scoring matrices can set the evolutionary look-back time for a search
  - Lower PAM (PAM10/VT10 … PAM/VT40) 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 0.45 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

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

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

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

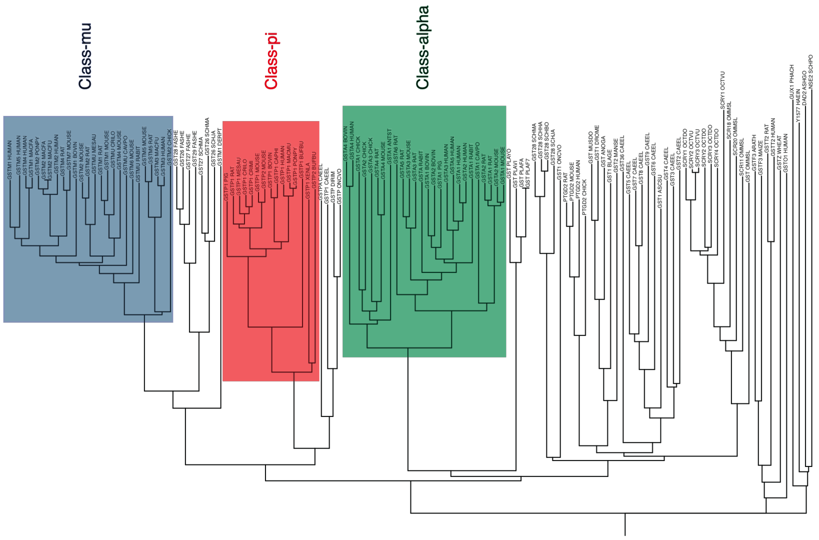
$l_2\ S_{R:N(\ 40)} = lg_2\ (0.000435/0.00219) = -2.333$

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

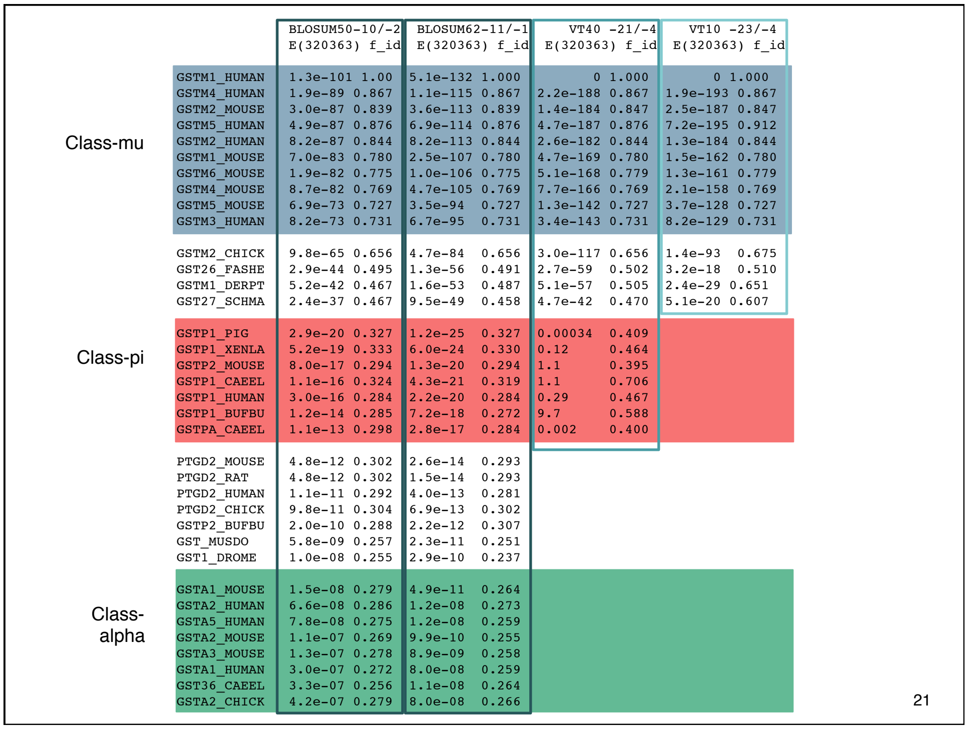
$l\ S_{R:N(250)} = lg2\ (0.002193/0.002193) = \ 0$

19

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



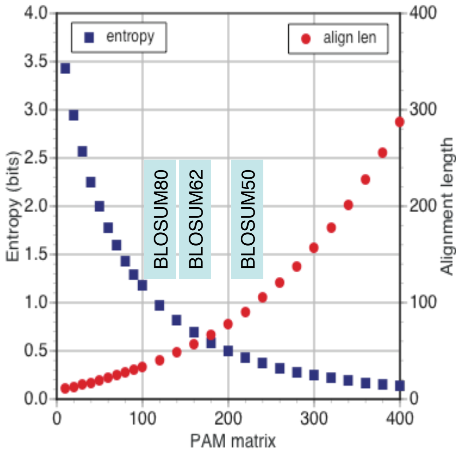
20



		BLOSUM50-10/-2		BLOSUM62-11/-1		VT40 -21/-4		VT10 -23/-4	
		E(320363)	f_id	E(320363)	f_id	E(320363)	f_id	E(320363)	f_id
Class-mu	GSTM1_HUMAN	1.3e-101	1.00	5.1e-132	1.000	0	1.000	0	1.000
	GSTM4_HUMAN	1.9e-89	0.867	1.1e-115	0.867	2.2e-188	0.867	1.9e-193	0.867
	GSTM2_MOUSE	3.0e-87	0.839	3.6e-113	0.839	1.4e-184	0.847	2.5e-187	0.847
	GSTM5_HUMAN	4.9e-87	0.876	6.9e-114	0.876	4.7e-187	0.876	7.2e-195	0.912
	GSTM2_HUMAN	8.2e-87	0.844	8.2e-113	0.844	2.6e-182	0.844	1.3e-184	0.844
	GSTM1_MOUSE	7.0e-83	0.780	2.5e-107	0.780	4.7e-169	0.780	1.5e-162	0.780
	GSTM6_MOUSE	1.9e-82	0.775	1.0e-106	0.775	5.1e-168	0.779	1.3e-161	0.779
	GSTM4_MOUSE	8.7e-82	0.769	4.7e-105	0.769	7.7e-166	0.769	2.1e-158	0.769
	GSTM5_MOUSE	6.9e-73	0.727	3.5e-94	0.727	1.3e-142	0.727	3.7e-128	0.727
	GSTM3_HUMAN	8.2e-73	0.731	6.7e-95	0.731	3.4e-143	0.731	8.2e-129	0.731
	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				

21

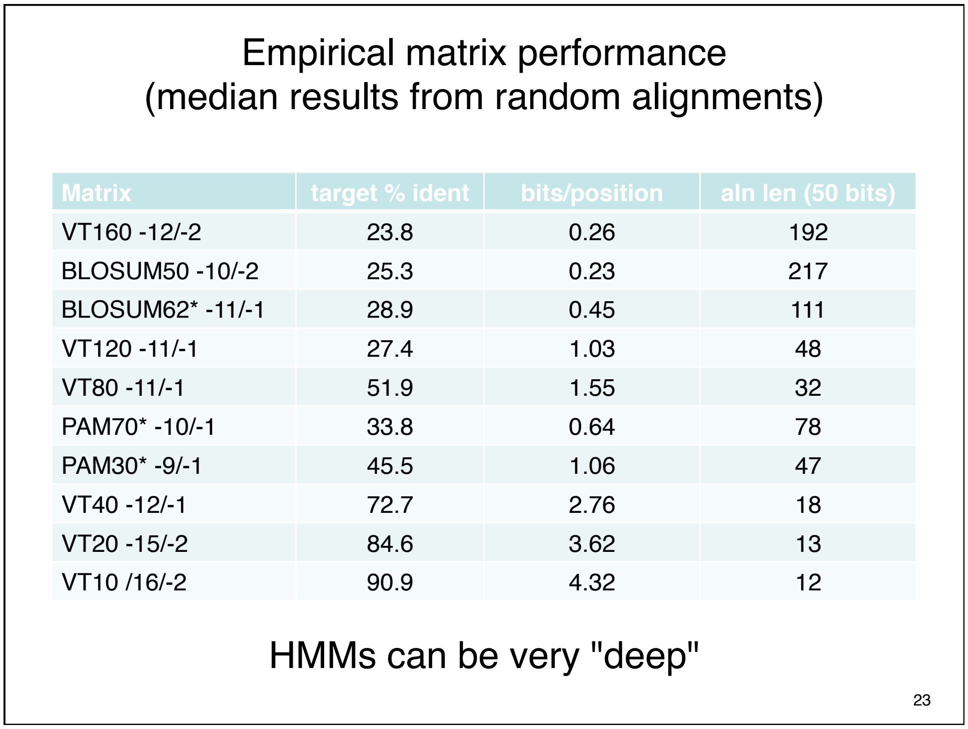
PAM matrices and alignment length



Short domains require “shallow” scoring matrices

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

22

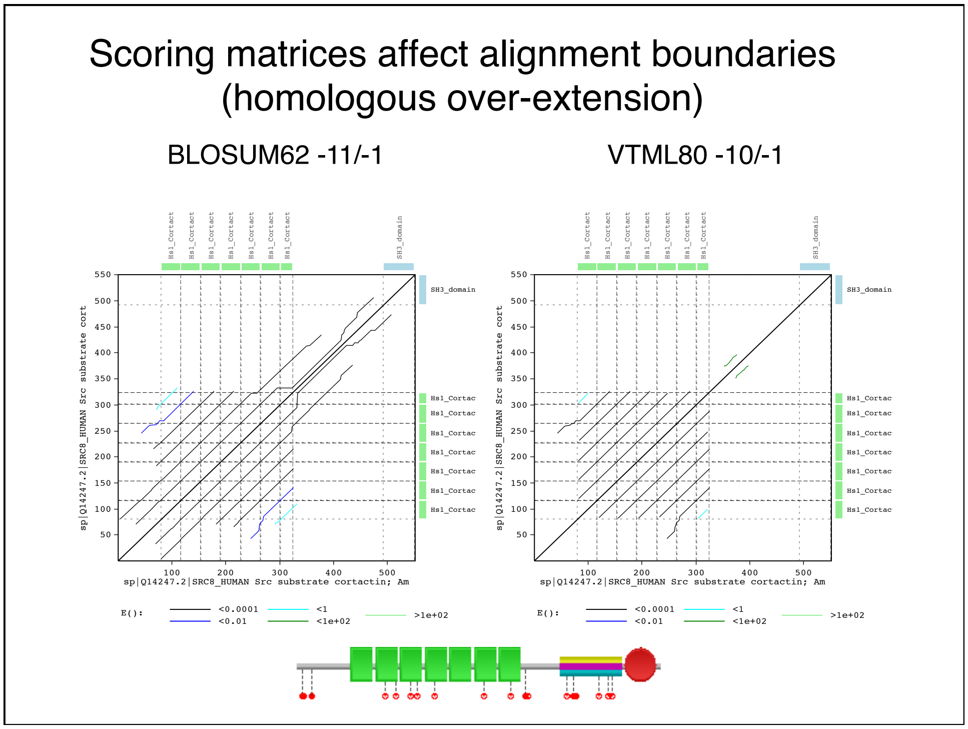


Empirical matrix performance (median results from random alignments)

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"

23



Scoring matrices affect alignment boundaries (homologous over-extension)

BLOSUM62 -11/-1

VTML80 -10/-1

Scoring Matrices - Summary

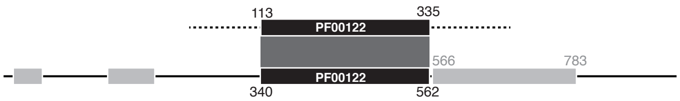
- PAM and BLOSUM matrices greatly improve the sensitivity of protein sequence comparison – low identity with significant similarity
- PAM matrices have an evolutionary model - lower number, less divergence – lower=closer; higher=more distant
- BLOSUM matrices are sampled from conserved regions at different average identity – higher=more conservation
- Shallow matrices set maximum look-back time
- Short alignments (domains, exons, reads) require shallow (higher information content) matrices

Effective Similarity Searching Using Annotations

- Modern sequence similarity searching is highly efficient, sensitive, and reliable – homologs are homologs
  - similarity statistics are accurate
  - databases are large
  - most queries will find a significant match

- Improving similarity searches
  - smaller databases
  - appropriate scoring matrices for short reads/assemblies
  - appropriate alignment boundaries
- Extracting more information from annotations
  - homologous over extension
  - scoring sub-alignments to identify homologous domains
- All methods (pairwise, HMM, PSSM) miss homologs
  - all methods find genuine homologs the other methods miss

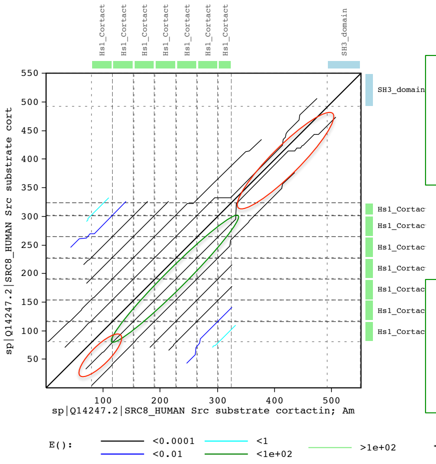
Overextension into random sequence



```
> pf26|15978520|E6SGT6|E6SGT6_THEM7 Heavy metal translocating P-type
ATPase EC=3.6.3.4
Length=888

 Score =  299 bits (766),  Expect = 1e-90, Method: Compositional matrix adjust.
 Identities = 170/341 (50%), Positives = 224/341 (66%), Gaps = 19/341 (6%)
                                          113
Query  84   FLFVNVFAALFNYWPTEGKILMFGKLEKVLITLILLGKTLEAVAKGRTSEAIKKLMGLKA  143
            +L+  V  A    +P+      +F  +  V++ L+ LG  LE  A+GRTSEAIKKL+GL+A
Sbjct  312  WLYSTVAVAFPQIFPSMALAEVFYDVTAVVVALVNLGLALELRARGRTSEAIKKLIGLQA  371
                                          340
Query  144  KRARVIRGGRELDIPVEAVLAGDLVVVRPGEKIPVDGVVEEGASAVDESMLTGESLPVDK  203
            + ARV+R G E+DIPVE VL GD+VVVRPGEKIPVDGVV EG S+VDESM+TGES+PV+
Sbjct  372  RTARVVRDGTEVDIPVEEVLVGDIVVVRPGEKIPVDGVVIEGTSSVDESMITGESIPVEM  431

Query  204  QPGDTVIGATLNKQGSFKFRATKVGRDTALAQIISVVEEAQGSKAPIQRLADTISGYFVP  263
            +PGD VIGAT+N+ GSF+FRATKVG+DTAL+QII +V++AQGSKAPIQR+ D +S YFVP
Sbjct  432  KPGDEVIGATINQTGSFRFRATKVGKDTALSQIIRLVQDAQGSKAPIQRIVDRVSHYFVP  491

Query  264  VVVSLAVITFFVWYFAVAPENFTRALLNFTAVLVIACPCALGLATPTSIMVGTGKGAEKG  323
             V+ LA++   VWY        +  AL+ F   L+IACPCALGLATPTS+ VG GKGAE+G
Sbjct  492  AVLILAIVAAVVWYVFGPEPAYIYALIVFVTTLIIACPCALGLATPTSLTVGIGKGAEQG  551
                     335
Query  324  ILFKGGEHLENAG---------GGAHTEGAENKAELLKTRATGISILVTLGLTAKGRDRS  374
            IL + G+ L+ A            G  T+G     +++   ATG    + L LTA
Sbjct  552  ILIRSGDALQMASRLDVIVLDKTGTITKGKPELTDVVA--ATGFDEDLILRLTA------  603
                     562 566
Query  375  TVAFQKNTGFKLKIPIGQAQLQREVAASESIVISAYPIVGV  415
              A ++ +   L   I +  L R +A  E+   +A P  GV
Sbjct  604  --AIERKSEHPLATAIVEGALARGLALPEADGFAAIPGHGV  642
```

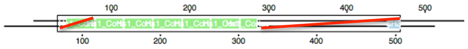
Scoring matrices affect alignment boundaries (homologous over-extension)

BLOSUM62 -11/-1

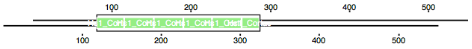


BLOSUM62 -11/-1

```
 32- 42: 69- 79 : Id=0.455; Q= 0.0 : NODOM :0
 43- 79: 80-116 : Id=0.158; Q= 0.0 : Hs1_Cortactin
 80-116:117-153 : Id=0.622; Q=37.4 : Hs1_Cortactin
117-153:154-190 : Id=0.757; Q=50.2 : Hs1_Cortactin
154-190:191-227 : Id=0.811; Q=61.0 : Hs1_Cortactin
191-227:228-264 : Id=0.568; Q=35.3 : Hs1_Cortactin
228-264:265-301 : Id=0.649; Q=41.5 : Hs1_Cortactin
265-287:302-324 : Id=0.565; Q= 8.9 : Hs1_Cortactin
288-458:325-491 : Id=0.165; Q= 0.0 : NODOM
459-473:492-506 : Id=0.200; Q= 0.0 : SH3
```



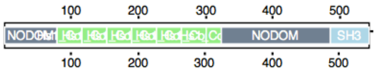
```
 82-116:119-153 : Id=0.657; Q=102.2 : Hs1_Cortactin
117-153:154-190 : Id=0.757; Q=138.0 : Hs1_Cortactin
154-190:191-227 : Id=0.811; Q=164.6 : Hs1_Cortactin
191-227:228-264 : Id=0.568; Q= 91.9 : Hs1_Cortactin
228-264:265-301 : Id=0.649; Q=112.4 : Hs1_Cortactin
265-287:302-324 : Id=0.565; Q= 36.7 : Hs1_Cortactin
```



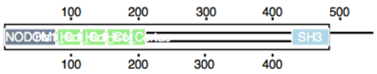
VTML80 -10/-1

Scoring domains highlights over extension

```
>>sp|SRC8_HUMAN Src substrate cortactin; (550 aa)
>>sp|SRC8_CHICK Src substrate p85;  Cort (563 aa)
84.7% id (1-550:11-563) E(454402): 1.2e-159

  1- 79: 11- 88 Id=0.873; Q=281.4 : NODOM
 80-116: 89-125 Id=1.000; Q=133.2 : Hs1_Cortactin
117-153:126-162 Id=0.946; Q=121.0 : Hs1_Cortactin
154-190:163-199 Id=0.973; Q=127.1 : Hs1_Cortactin
191-227:200-236 Id=0.973; Q=128.3 : Hs1_Cortactin
228-264:237-273 Id=0.973; Q=137.5 : Hs1_Cortactin
265-301:274-310 Id=0.892; Q=117.3 : Hs1_Cortactin
302-324:311-333 Id=0.957; Q= 69.6 : Hs1_Cortactin
325-491:334-504 Id=0.632; Q=386.6 : NODOM
492-550:505-563 Id=0.966; Q=226.3 : SH3
```



```
>>sp|SRC8_HUMAN Src substrate cortactin (550 aa)
>>sp|HCLS1_MOUSE Hematopoiet ln cell-sp (486 aa)
44.1% id (1-548:1-485)  E(454402): 4.1e-61

  1- 79:   1- 78 Id=0.671; Q=213.0 : NODOM
 80-116: 79-115 Id=0.757; Q= 97.9 : Hs1_Cortactin
117-153:116-152 Id=0.703; Q= 94.8 : Hs1_Cortactin
154-190:153-189 Id=0.703; Q= 97.3 : Hs1_Cortactin
191-213:190-212 Id=0.826; Q= 60.5 : Hs1_Cortactin



214-491:213-428 Id=0.179; Q=  0.0 : NODOM :0
492-548:429-485 Id=0.719; Q=173.2 : SH3
```

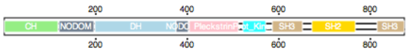


$Q = -10 \log(p)$

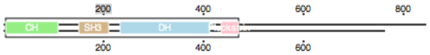
$Q>30.0$ -> $p < 0.001$

Over extension or distant homologs?

```
>>sp|VAV_HUMAN Proto-oncogene vav  (845 aa)
>>sp|VAV2_HUMAN Guanine nt EF VAV  (878 aa)

49.3% id (1-840:1-875) E(454402): 4.1e-210

  1-119:   1-119 :Id=0.689; Q=432.7 : CH
120-193:120-197 :Id=0.444; Q=117.5 : NODOM
194-373:198-376 :Id=0.494; Q=466.0 : DH
374-401:377-404 :Id=0.607; Q= 48.7 : NODOM
402-504:405-512 :Id=0.509; Q=275.7 : Pleckstrin
505-514:513-522 :Id=0.600; Q=  0.0 : NODOM
515-564:523-572 :Id=0.640; Q=175.6 : PE/DAG-bd
579-591:573-585 :Id=0.154; Q=  0.0 : NODOM
592-659:586-652 :Id=0.420; Q=101.4 : SH3
659-670:653-672 :Id=0.158; Q=  0.0 : NODOM
671-765:673-767 :Id=0.516; Q=241.2 : SH2
766-784:768-815 :Id=0.125; Q=  0.0 : NODOM
784-840:816-875 :Id=0.593; Q=162.7 : SH3
```



```
>>sp|VAV_HUMAN Proto-oncogene vav  (845 aa)
>>sp|Q5ZLR6.1|ARHG6_CHICK RhoGEF   (764 aa)

24.9% id (6-433:6-472) E(454402): 1.1e-12

  6-119:   6-110 :Id=0.325; Q=97.8 : CH
120-155:111-151 :Id=0.195; Q= 0.0 : NODOM
155-180:152-211 :Id=0.169; Q= 0.0 : SH3
181-195:212-232 :Id=0.190; Q= 0.0 : NODOM
196-373:233-413 :Id=0.265; Q=74.1 : DH
374-395:414-434 :Id=0.174; Q= 0.0 : NODOM
396-433:435-472 :Id=0.211; Q= 0.0 : Pleckstrin
```



Homology, non-homology, and over-extension

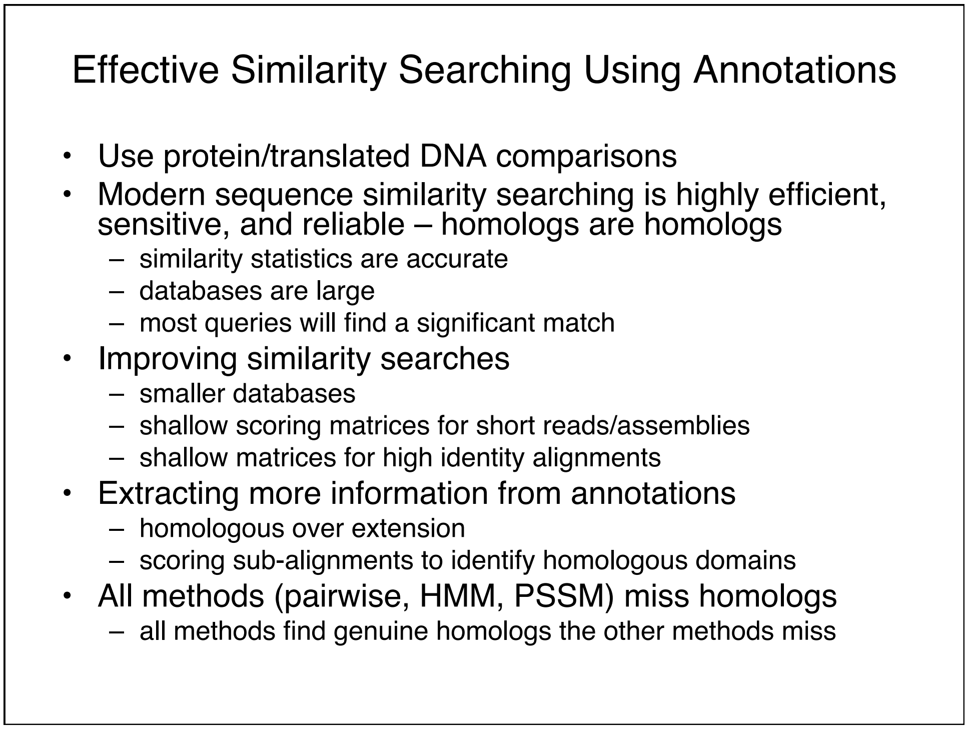
- Sequences that share statistically significant sequence similarity are homologous (simplest explanation)
- But not all regions of the alignment contribute uniformly to the score
  - lower identity/Q-value because of non-homology (over-extension) ?
  - lower identity/Q-value because more distant relationship (domains have different ages) ?
- Test by searching with isolated region
  - can the *distant domain (?)* find closer (significant) homologs?
- Similar (homology) or distinct (non-homology) structure is the gold standard
- Multiple sequence alignment can obscure over-extension
  - if the alignment is over-extended, part of the alignment is NOT homologous

31

Effective Similarity Searching

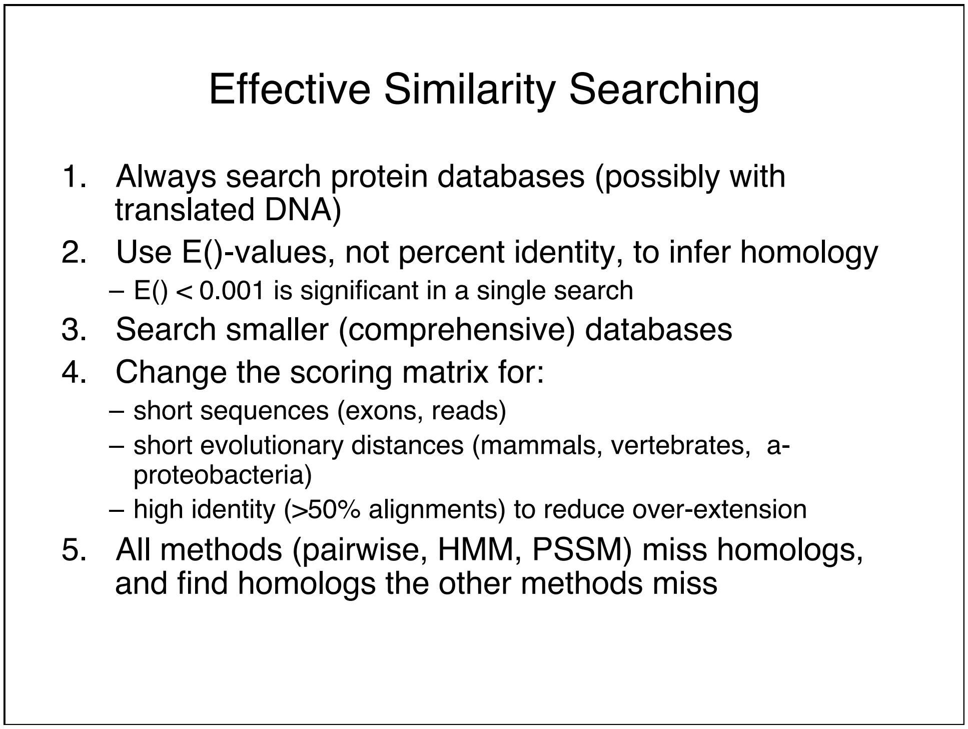
1. Always search protein databases (possibly with translated DNA)
2. Use E()-values, not percent identity, to infer homology
   - E() < 0.001 is significant in a single search

3. Search smaller (comprehensive) databases
4. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
5. All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss



Effective Similarity Searching Using Annotations

- Use protein/translated DNA comparisons
- Modern sequence similarity searching is highly efficient, sensitive, and reliable – homologs are homologs
  - similarity statistics are accurate
  - databases are large
  - most queries will find a significant match
- Improving similarity searches
  - smaller databases
  - shallow scoring matrices for short reads/assemblies
  - shallow matrices for high identity alignments
- Extracting more information from annotations
  - homologous over extension
  - scoring sub-alignments to identify homologous domains
- All methods (pairwise, HMM, PSSM) miss homologs
  - all methods find genuine homologs the other methods miss



Effective Similarity Searching

1. Always search protein databases (possibly with translated DNA)
2. Use E()-values, not percent identity, to infer homology
   - E() < 0.001 is significant in a single search
3. Search smaller (comprehensive) databases
4. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
5. All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss