Programming for Biololgy

Similarity Searching II –

Practical search strategies

Bill Pearson
wrp@virginia.edu

Why is this material important?

- You might be asked to find a homolog
- You might be asked to what your gene/protein does
  - Annotated homologs are missed because databases are large and redundant
  - Short domains and short exons are missed because the “standard” matrix needs long alignments
  - Sometimes, alignments include non-homologous regions

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

---

1. Search smaller (comprehensive) databases
2. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
3. Is every aligned residue homologous?
   - alignment overextension
4. (Tomorrow) All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss

Review – Sequence Similarity - Conclusions

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

Similarity Searching II

1. What question to ask?
2. What program to use?
3. What database to search?
4. When to do something different (changing scoring matrices)
5. Is every aligned domain homologous?
6. (Tomorrow) – 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, …)?

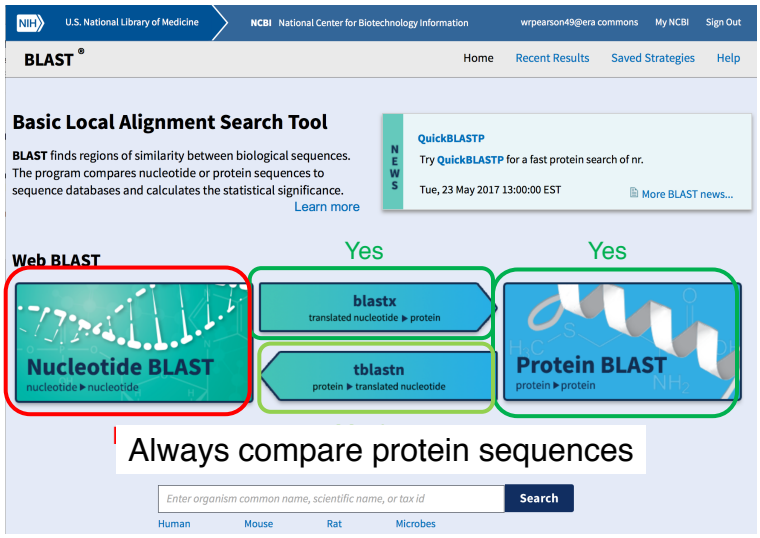
Questions not to ask:

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

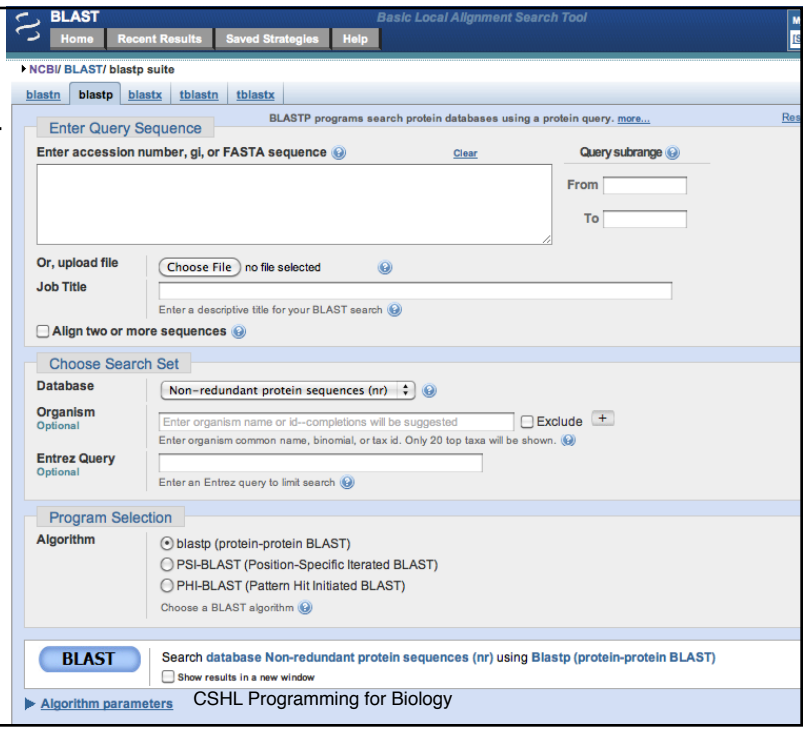
2. What program to run?

- What is your query sequence?
  - protein – BLASTP (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, EBI)

NCBI BLAST Server
blast.ncbi.nlm.nih.gov



NCBI BLAST Server



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)
  - Quest for Orthologs reference proteomes (1,000,000)
- Search a richly annotated protein set (SwissProt, 500,000)
- Always search NR (> 50 million) *LAST*
- Never Search “GenBank” (DNA)

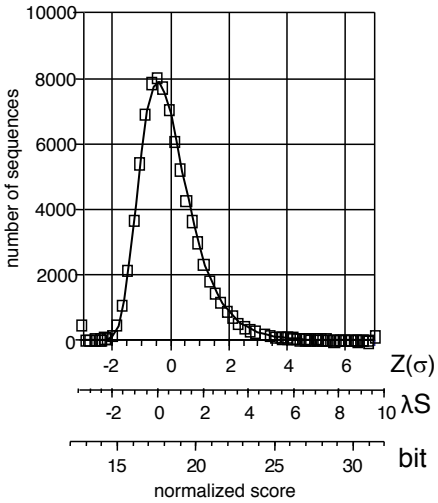
Effective Similarity Searching

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

---

1. Search smaller (comprehensive) databases
2. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
3. Is every aligned residue homologous?
   - alignment overextension
4. (Tomorrow) All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss

Why smaller databases are better – statistics



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

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

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

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

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

Bonferroni correction

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

$P(40 \text{ bits}) = 1.5\times10^{-7}$

$E(40 \mid D=4000) = 6\times10^{-4}$

$E(40 \mid D=60E6) = 9$

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 $P < 0.1$)

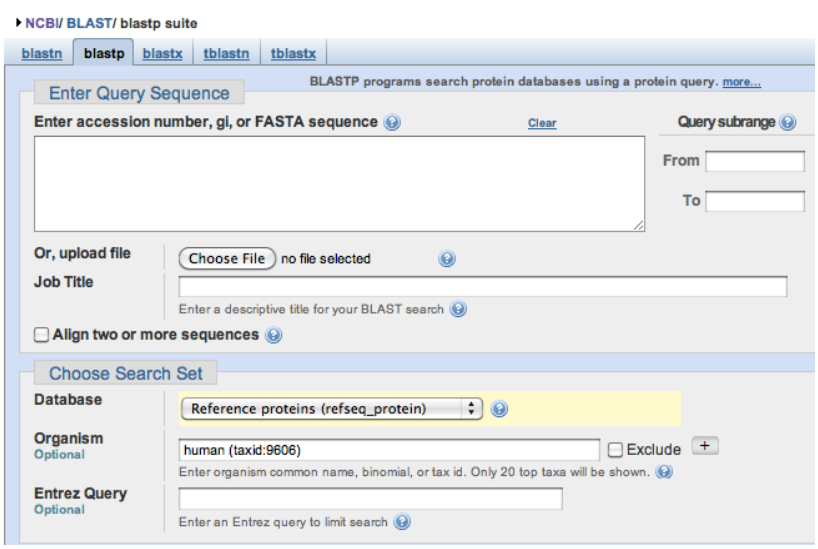
$E(S', S_{bits} \mid D) = PD$

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

dblength = D n

$E(S_{bit}) = m\, \text{dblength}\, 2^{-\text{bits}}$ (BLAST)

NCBI – selecting sequences with Entrez



What is a "bit" score (I)?

1. Scoring matrices (PAM250, BLOSUM62, VTML40) contain "log-odds" scores:
   - $s_{i,j}$ (bits) = $\log_2(q_{i,j}/p_ip_j)$ ($q_{i,j}$ freq. in homologs / p_ip_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)
2. 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.
3. 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: $P(S_{bit} > x) = 1 - \exp(-mn2^{-x}) \sim mn2^{-x}$
   - $400 \times 400 \times 2^{-40} = 160{,}000 / 2^{40}$ ($10^{13.3}$) = 1.5×10^{-7} times
   - Thus, the probability of a 40 bit score in ONE alignment is ~ 10^{-7}

What is a "bit" score (II)?

4. But we did not ONE alignment, we did 4,000, 40,000, 500,000, or 20 million alignments when we searched the database:
   - $E(S_{bit} \mid D)$ = p(40 bits) x 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)
   - $E(40 \mid 500{,}000) = 10^{-7} \times 5 \times 10^5 = 0.05$ (not significant)
   - $E(40 \mid 20 \text{ million}) = 10^{-7} \times 2.0 \times 10^7 = 2.0$ (not significant)

Not significant does not mean not-homologous

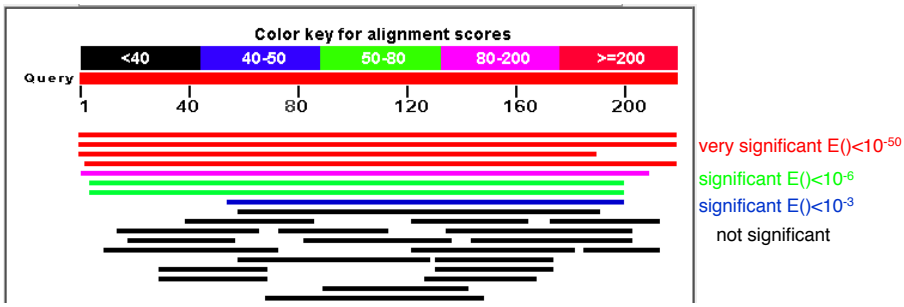
How many “bits” do I need?

E() = p() x database size

E(40 I 4,000) = 10^{-7} x 4,000 = 4 x 10^{-4} (significant)
E(40 I 40,000) = 10^{-7} x 4 x 10^{4} = 4 x 10^{-3} (not significant)
E(40 I 500,000) = 10^{-7} x 5 x 10^{5} = 0.05 (not significant)

To get E() ~ 10^{-3} , how many bits do I need? p = m n 2^{-bits}

bits = –log2(p/(m n)) = –log2(E()/(database_size m n))
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



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

---

1. Search smaller (comprehensive) databases
2. Change the scoring matrix for:
   – short sequences (exons, reads)
   – short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   – high identity (>50% alignments) to reduce over-extension
3. Is every aligned residue homologous?
   – alignment overextension
4. (Tomorrow) 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

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

q_{ij} : homolog frequency wat PAM40, 250

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

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

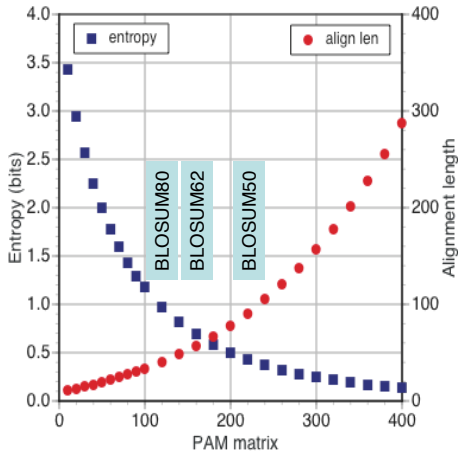
$\lambda_2\ S_{ij} = \lg_2\,(q_{ij}/p_i p_j)$ $\lambda_e\ S_{ij} = \ln(q_{ij}/p_i p_j)$ $p_R p_N = 0.002193$

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

$\lambda_2 = 1/3$; $S_{R:N(40)} = -2.333/l_2 = -7$

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

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)

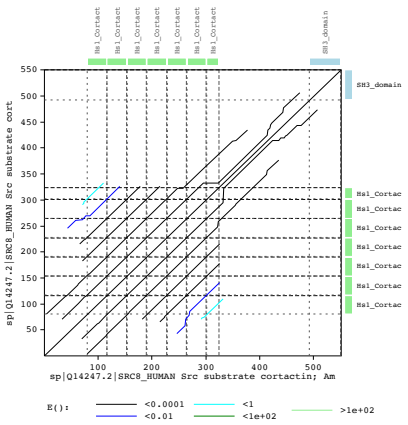
Matrix	target % ident	bits/position	aln len (50 bits)
VT160 -12/-2	23.8	0.26	192
BLOSUM50 -10/-2	25.3	0.23	217
BLOSUM62* -11/-1	28.9	0.45	111
VT120 -11/-1	27.4	1.03	48
VT80 -11/-1	51.9	1.55	32
PAM70* -10/-1	33.8	0.64	78
PAM30* -9/-1	45.5	1.06	47
VT40 -12/-1	72.7	2.76	18
VT20 -15/-2	84.6	3.62	13
VT10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

HMMs can be very "deep"

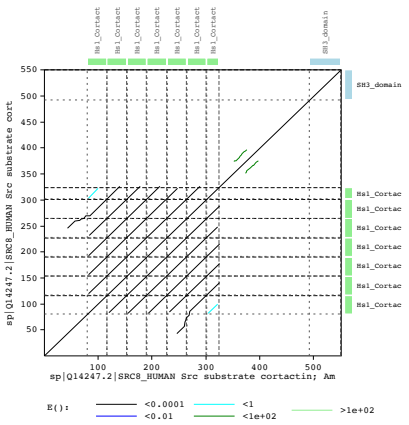
Pearson (2013) Curr. Prot. Bioinformatics 3.5.1

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

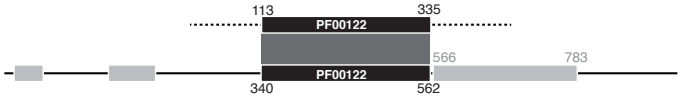
Effective Similarity Searching

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

---

1. Search smaller (comprehensive) databases
2. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
3. Is every aligned residue homologous?
   - alignment overextension
4. (Tomorrow) All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss

Over-extension into random sequence

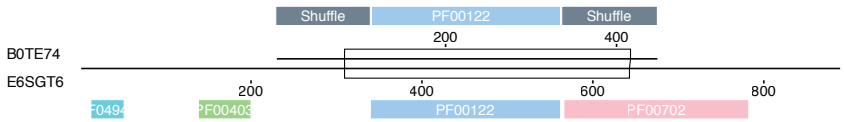



```
> pf26|15978520|E6SGT6|E6SGT6_THEM7 Heavy metal translocating P-type
ATPase EC=3.6.3.4
Length=888

 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
```

Mills and Pearson (2013) Bioinformatics 29:3007

Sub-alignment scoring detects over-extension

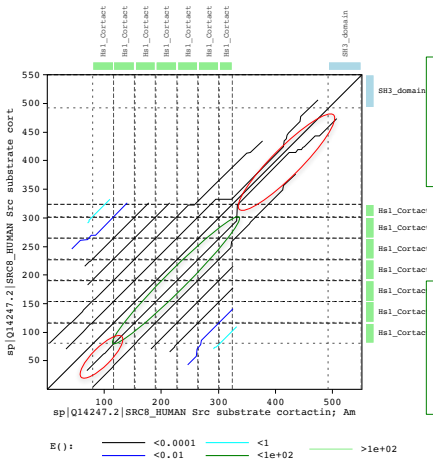
```
>>sp|E6SGT6|E6SGT6_THEM7 Heavy metal translocating P-type ATPase EC=3.6.3.4   (888 aa)
 qRegion: 81-112:309-340 : score=15; bits=12.3; Id=0.219; Q=0.0 :  Shuffle
 qRegion: 113-335:341-563 : score=736; bits=232.8; Id=0.641; Q=644.7 :  PF00122
 qRegion: 336-415:564-642 : score=14; bits=12.0; Id=0.236; Q=0.0 :  Shuffle
  Region: 81-111:309-339 : score=11; bits=11.1; Id=0.194; Q=0.0 :  NODOM :0
  Region: 112-334:340-562 : score=736; bits=232.8; Id=0.641; Q=644.7 :  PF00122  Pfam
  Region: 338-415:566-642 : score=16; bits=12.6; Id=0.244; Q=0.0 :  PF00702  Pfam
 s-w opt: 632  Z-score: 1048.6  bits: 204.2 E(274545): 3.7e-51
Smith-Waterman score: 765; 49.7% identity (73.3% similar) in 344 aa overlap (81-415:309-642)
```




```
                50        60        70        80        90       100       110 ][     120
B0TE74 LPIVPGTMALGVQSDGKDETLLVALEVPVDRERVGPAIVDAFRFLFVNVFAALFNYWPTEGKILMFGKLEKVLITLILLG
                                                   :  .:. .: .:. . .:. .   .:  .  :...:. ::
E6SGT6 ILGLLTLPVMLWSGSHFFNGMWQGLKHRQANMHTLISIGIAAAWLYSTVAVAFPQIFPSMALAEVFYDvtavvvalvnlg
       270       280       290       300       310       320       330       3][
                                                ...
                290       300       310       320       330     ][        340       350
B0TE74 APENFTRALLNFTAVLVIACPCALGLATPTSIMVGTGKGAEKGILFKGGEHLENAGG---------GAHTEGAENKAELL
          .  ::. :...:.:::::::::::::. :: :::::.:::...:. :.  :.          :. :.:  . ....
E6SGT6 PEPAYIYALIVFVTTLIIACPCALGLATPTSLTVGIGKGAEQGILIRSGDALQMASRLDVIVLDKTGTITKGKPELTDVV
       510       520       530       540       550       560 ]   [ 570       580
              360       370       380       390       400       410       420       430
B0TE74 KTRATGISILVTLGLTAKGRDRSTVAFQKNTGFKLKIPIGQAQLQREVAASESIVISAYPIVGVVVDSLVTTAFLAVEEI
          :::..  . : :::        :.....   :   : .. : : .:  :.  ..: :  ::
E6SGT6 --AATGFDEDLILRLTA--------AIERKSEHPLATAIVEGALARGLALPEADGFAAIPGHGVEAQVEGHHVLVGNERL
            590       600                 610       620       630       640       650
```

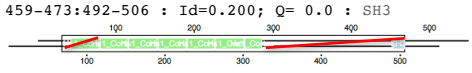

Scoring matrices affect alignment boundaries (homologous over-extension)

BLOSUM62 -11/-1

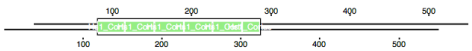


BLOSUM62 -11/-1

```
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
```

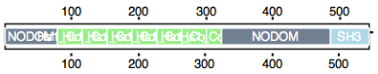


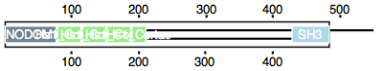
VTML80 -10/-1

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

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

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

---

1. Search smaller (comprehensive) databases
2. Change the scoring matrix for:
   - short sequences (exons, reads)
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - high identity (>50% alignments) to reduce over-extension
3. Is every aligned residue homologous?
   - alignment overextension
4. (Tomorrow) All methods (pairwise, HMM, PSSM) miss homologs, and find homologs the other methods miss

workshop II – parsing blast results

Goto:

fasta.bioch.virginia.edu/mol_evol/pfb_python_matrices.html

Your goal is to reproduce a version of this table:

Matrix	target % ident	align_len	evalue
VT160 -12/-2	23.8		
BLOSUM50 -10/-2	25.3		
BLOSUM62* -11/-1	28.9		
VT120 -11/-1	27.4		
VT80 -11/-1	51.9		