Programming for Biololgy

Similarity Searching II –

Practical search strategies

Bill Pearson
wrp@virginia.edu

1

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

2

Effective Similarity Searching

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

---

1. Search smaller (comprehensive) databases
   - Less redundancy; higher sensitivity
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

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

4

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)

5

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?

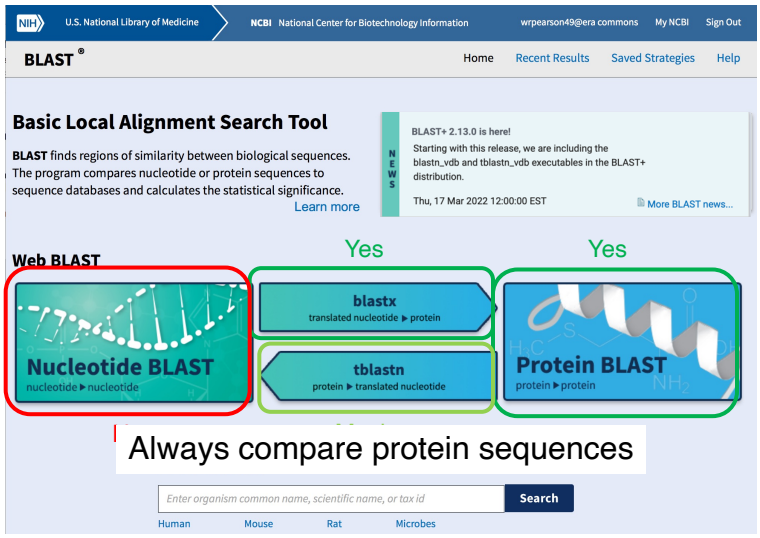
6

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)

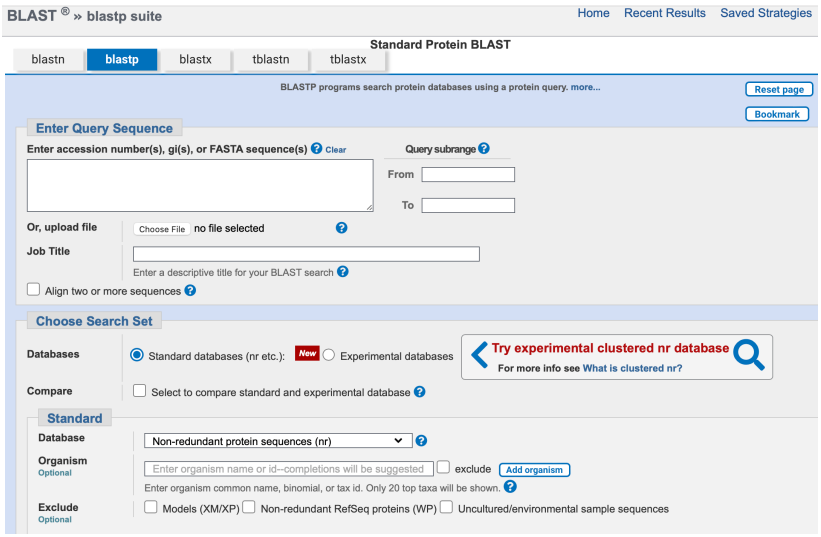
7

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

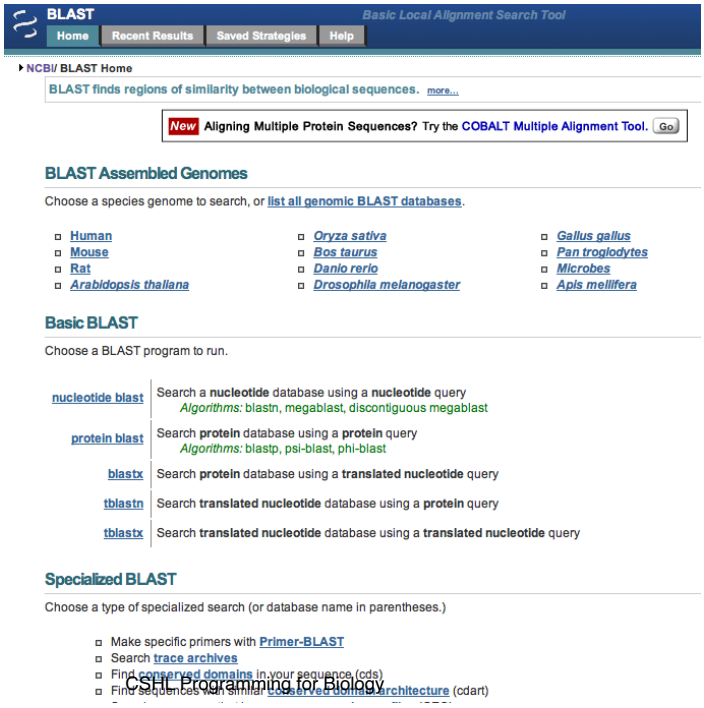


8

NCBI BLAST Server



NCBI BLAST Server



3. What database to search?

- Search the smallest comprehensive database likely to contain your protein
  - vertebrates – human proteins (40,000)
  - NCBI Landmark sequences (human, mouse, no rat)
  - Quest for Orthologs reference proteomes (1,000,000)
- Search a richly annotated protein set (SwissProt: 500,000, NCBI Landmark:)
- Always search NR (> 500 million) *LAST*
- Never Search "GenBank" (DNA)

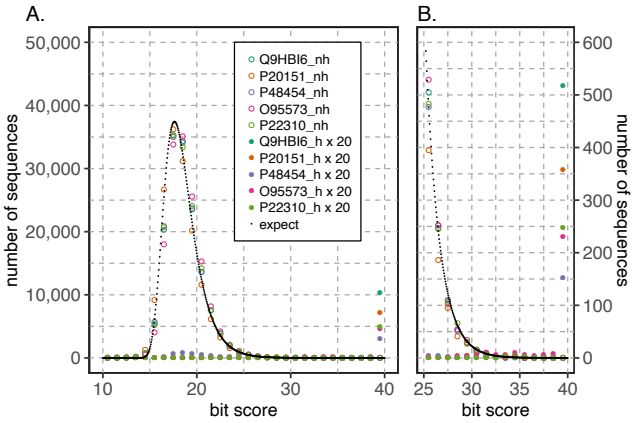
Effective Similarity Searching

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

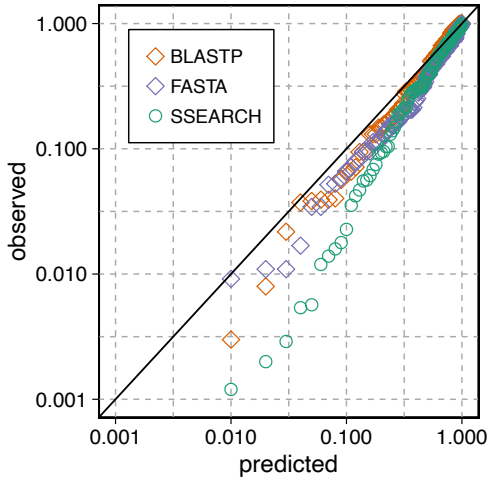
Homology inferences are reliable because similarity statistics are accurate (I) (we know how unrelated sequences behave)



Distributions of similarity scores in searches with 5 human enzymes. Open circles (_nh) show scores for non-homologs. Closed circles show homolog (_h) scores.

13

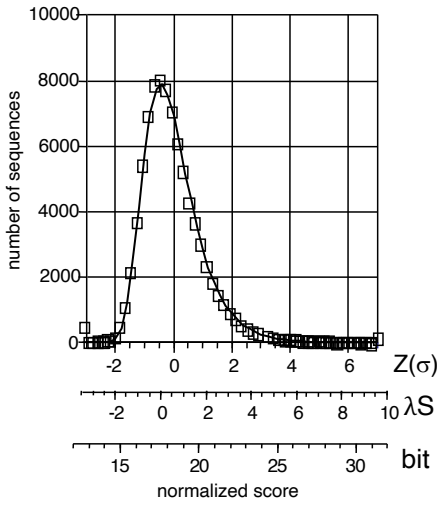
Homology inferences are reliable because similarity statistics are accurate (II) (we know how unrelated sequences behave)



Reported (observed) and expected probabilities of the highest scoring unrelated sequence in searches with 100 human enzymes vs 78 complete proteomes (~1 million sequences).

14

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

$P(B \text{ bits}) = m\ n\ 2^{-B}$
$P(40 \text{ bits}) = 1.5 \times 10^{-7}$
$E(40 \mid D=4000) = 6 \times 10^{-4}$
$E(40 \mid D=500E6) = 75$

15

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 $P < 0.1$)

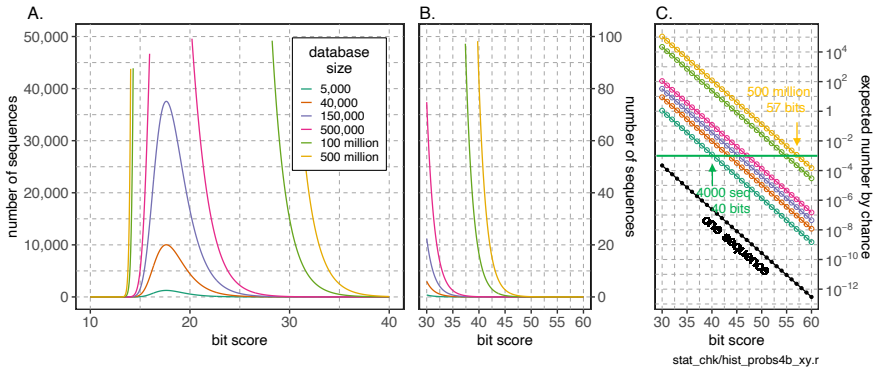
$P(S_{bits} > \text{bits}) = 1 - \exp(-mn2^{-x})$
$P(S_{bits} > \text{bits}) = mn2^{-\text{bits}}$ (for $P < 0.1$)

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

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

16

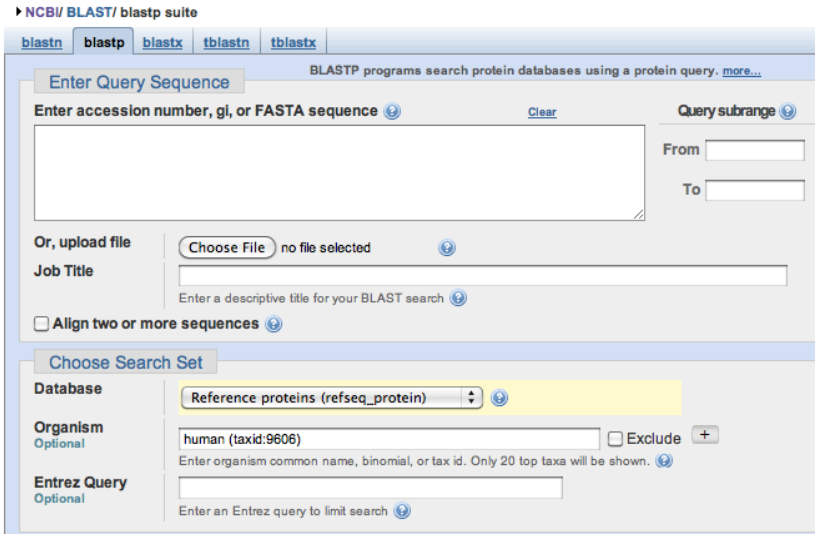
Smaller databases increase sensitivity



(More sophisticated algorithms – PSIBLAST, JACKHMMR – also improve sensitivity, and they work better with large databases

17

NCBI – selecting sequences with Entrez



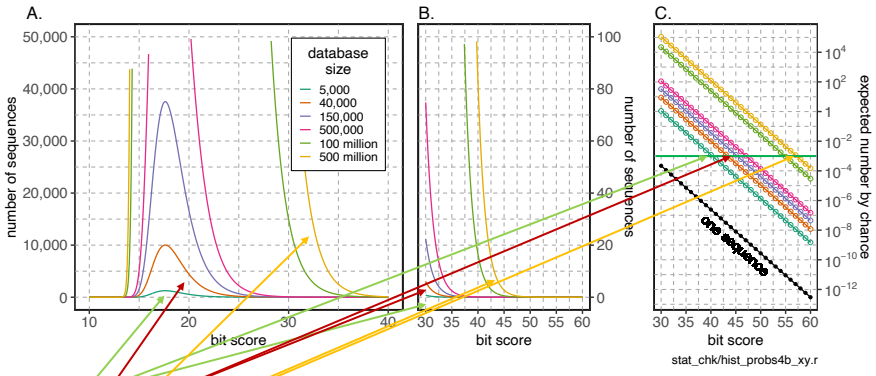
18

3. What database to search?

Database	Size	Bits (0.001)
Landmark	441 thousand	47
SwissProt	480 thousand	47
Refseq_Select	64 million	53
Refseq_Protein	234 million	55
NR (clustered)	242 million	55
NR	510 million	56

19

How many bits do I need?



1. Small (5,000) database, E()<0.001 > 38 bits
2. Genome size (40,000) database, no random scores > 42 bits
3. "Complete" database (500 million), no random scores > 55 bits

(More sophisticated algorithms – PSIBLAST, JACKHMMR – also improve sensitivity, and they work better with large databases

20

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

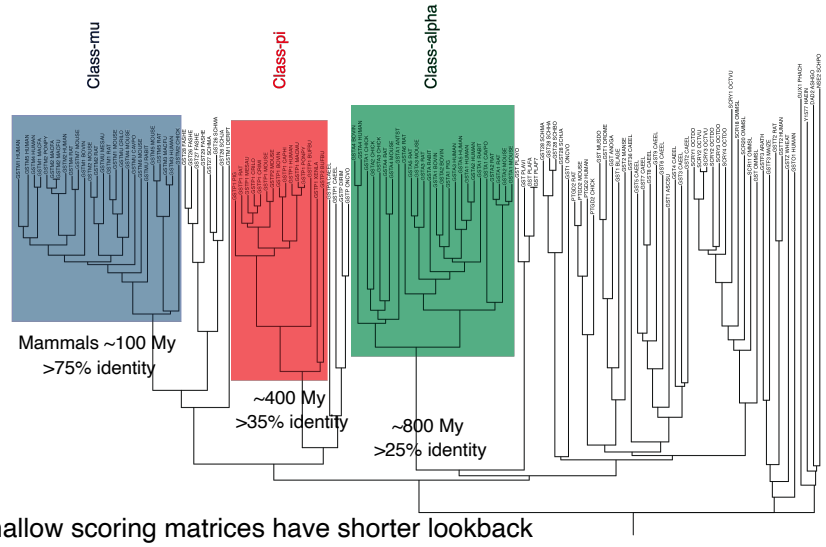
---

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

Scoring matrices – shifting lookback (where do those bits come from?)

- 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

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



Shallow scoring matrices have shorter lookback times because they penalize more for non-identity

23

Scoring matrices and alignment length

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

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

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

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

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

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

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

24

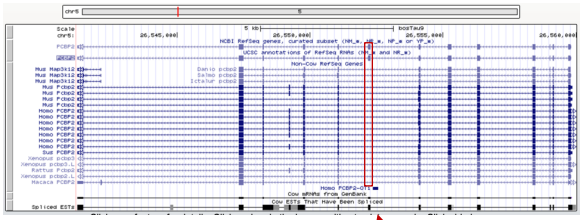
Empirical matrix performance (median results from random alignments)

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, alignment length, and exon detection – bovine PCBP2



```
human  MDTGVIEGGLNVTLTIRLLMHGKEVGSIIGKKGESVKKMREESGARINISEGNCPERIITLAGPTNAIFKAFAMIIDKLE
bovin  MDTGVIEGGLNVTLTIRLLMHGKEVGSIIGKKGESVKKMREESGARINISEGNCPERIITLAGPTNAIFKAFAMIIDKLE
human  EDISSSMTNSTAASRPPVTLRLVVPASQCGSLIGKGGCKIKEIRESTGAQVQVAGDMLPNSTERAITIAGIPQSIIECVK
bovin  EDISSSMTNSTAASRPPVTLRLVVPASQCGSLIGKGGCKIKEIRESTGAQVQVAGDMLPNSTERAITIAGIPQSIIECVK
human  QICVVMLETLSQSPPKGVTIPYRPKPSSSPVIFAGGQDRYSTGSDSASFPHTTPSMCLNPDLEGPPLEAYTIQGQYAIPQ
bovin  QICVVMLE----SPPKGVTIPYRPKPSSSPVIFAGGQDRYSTGSDSASFPHTTPSMCLNPDLEGPPLEAYTIQGQYAIPQ
human  PDLTKLHQLAMQQSHFPMTHGNTGFSGIESSSPEVKGTWGLDASAQTTSHELTIPNDLIGCIIGRQGAKINEIRQMSGAQ
bovin  PDLTKLHQLAMQQSHFPMTHGNTGFSA----------GLDASAQTTSHELTIPNDLIGCIIGRQGAKINEIRQMSGAQ
human  IKIANPVEGSTDRQVTITGSAASISLAQYLINVRLSSETGGMGSS
bovin  IKIANPVEGSTDRQVTITGSAASISLAQYLINVRLSSETGGMGSS
```

Is this exon really missing from cow?

Scoring matrices, alignment length, and exon detection – bovine PCBP2

name	start	end	len	VT10	bits	BP62	Bits	PAM30	Bits
ex_1	1	23	23	=	58	+5	45	+5	57
ex_2	24	31	8	=	30	–	<25	=	19
ex_3	32	42	11	=	36	+1	27	=	26
ex_4	43	81	39	=	88	=	61	+5	95
ex_5	82	125	44	=	96	=	66	=	104
ex_6	126	168	43	=	96	+5	65	+7	106
ex_7	169	197	29	=	69	+2	50	+2	70
ex_8	198	228	31	=	76	+43	53	=	80
ex_9	229	242	14	+4	40	+4	32	+2	37
ex_10	243	266	24	+5	60	+87	45	+5	68
ex_11	267	280	14	=	42	+38	32	=	43
ex_12	281	297	17	=	49	+2	34	–	
ex_13	298	354	57	=	120	=	78	+9	135
ex_14	355	365	11	=	37	=	32	–	

VT10

```
qRegion: bits=71.5; Id=1.000; exon_8-8

       ][00       210       220       ]
sp|Q15 QDRYSTGSDSASFPHTTPSMCLNPDLEGPPLE
       ::::::::::::::::::::::::::::::::
chr5:2 QDRYSTGSDSASFPHTTPSMCLNPDLEGPPLE
```

BP62

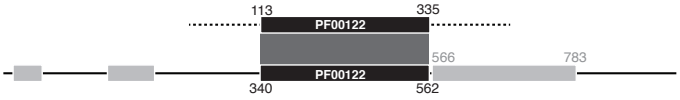
```
qRegion: 181-197 : bits=1.6;  Id=0.500:  exon_7-7
qRegion: 198-228 : bits=52.7; Id=1.000:  exon_8-8
qRegion: 229-242 : bits=0.0;  Id=0.333;  exon_9-9
qRegion: 243-255 : bits=5.4;  Id=0.286;  exon_10-10

                190      ][00       210       220       ][0       240 ][    250
sp|Q15 PYRPKPSSSPVIFAGGQDRYSTGSDSASFPHTTPSMCLNPDLEGPPLEAYTIQGQYAIPQPDLTKLHQLAMQQSH
       :.: : :  :      ::::::::::::::::::::::::::::::::   ..:.  . .: ::.  .:. . .:
chr5:2 PWRLKSSIYP------QDRYSTGSDSASFPHTTPSMCLNPDLEGPPLE---VRGD--VQSPRLTQSFRLSRDCQH
```

Shallow matrix (MD10) finds exon and exon boundaries

Deep (sensitive) matrix (BP62) finds exon, but overextends exon boundaries

The exon is present in cow, but not detected because it is short

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

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

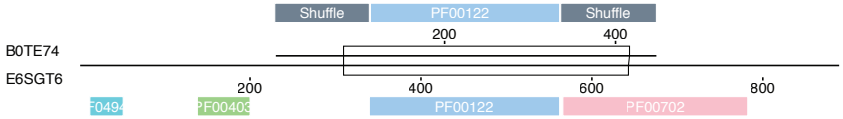
1. Always search protein databases (possibly with DNA – blastx, fastx)
2. Use E()-values, not percent identity, to infer homology
   - E() < 0.001 is significant in a single search

---

1. Search smaller (comprehensive) databases
   - Less redundancy; better sensitivity
2. Change the scoring matrix for:
   - short evolutionary distances (mammals, vertebrates, a-proteobacteria)
   - short sequences (exons, reads)
   - high identity (>50% alignments) to reduce over-extension

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)
```

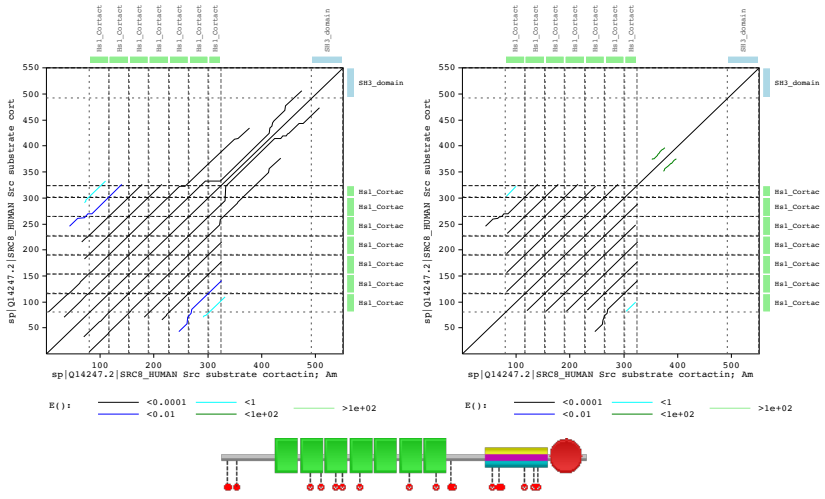



```
                50        60        70        80        90       100       110 ][    120
B0TE74 LPIVPGTMALGVQSDGKDETLLVALEVPVDRERVGPAIVDAFRFLFVNVFAALFNYWPTEGKILMFGKLEKVLITLILLG
                                                  :  .:. .: .:. . .:. .   .:  .  :...:. ::
E6SGT6 ILGLLTLPVMLWSGSHFFNGMWQGLKHRQANMHTLISIGIAAAWLYSTVAVAFPQIFPSMALAEVFYDvtavvvalvnlg
       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

VTML80 -10/-1



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

33

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

34

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	29.7	67	2.1
BLOSUM50	34.0	121	1.2
BLOSUM62* -11/-1	31.2	90	0.37
VT80	66.7	50	1.8
VT40	72.7	11	1.3