Similarity searching / sequence alignment summary

Biol4230 Thurs, February 22, 2016

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

What have we covered?

- Homology – excess similiarity
  - but no excess similarity ≠ non-homology
  - what is an Expectation E() value?
  - DNA vs protein searches?
- Alignment scores
  - use scoring matrix not identity (for proteins)
  - why is protein comparison more sensitive?
- BLAST lab I:
  - non-significant ≠ not-homologous
  - domains show homology when pairwise score does not (why?)
  - are parts of domains missing when only part aligns?

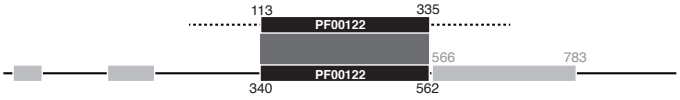
Similarity searching summary (2)

- Quick overview of alignment algorithms
  - local vs global
  - dynamic programming
  - non-overlapping local alignments
- Improving search performance - local alignment statistics
  - the extreme value distribution
  - why database size matters
  - evaluating statistical accuracy – what is the "control?"
- What are E()-values good for? Not good for?
- Where scoring matrices come from
  - scoring matrices as log-odds matrices
  - shallow matrices: short higher identity alignments / deep matrices: long alignments, lower identity alignments – WHY??
    - shallow matrices, higher identity alignment (less over-extension)
- Blast lab II –
  - local alignments of duplicated domains?
  - alignment over-extension

Domains

- domain definitions –
  - domains are "atomic" – mobile structural units
  - why do only parts of domains align?
- InterPro, a "meta"-database of domain databases, and Pfam
  - when do the domain databases agree? where do they disagree?
- Where do pairwise scoring matrices come from? –
  - log(odds) [f-homology/f-chance]
  - which part changes for different amounts of divergence?
- What are position specific scoring matrices (PSSMs)
  - [f-position/f-chance] PSI-BLAST
  - what are the starting values? which part changes?
- What mistakes do Iterative methods (PSI-BLAST) make?
  - alignment over-extension (which can lead to …)
  - multiple alignment (PSSM) contamination

Over-extension into random sequence



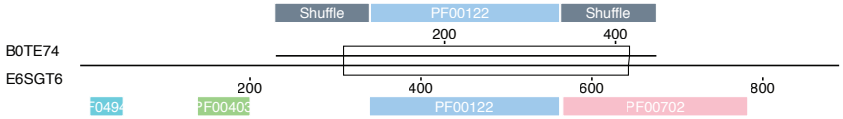

```
> pf26|15978520|E6SGT6|E6SGT6_THEM7 Heavy metal translocating P-type
ATPase EC=3.6.3.4
Length=888

 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)
```

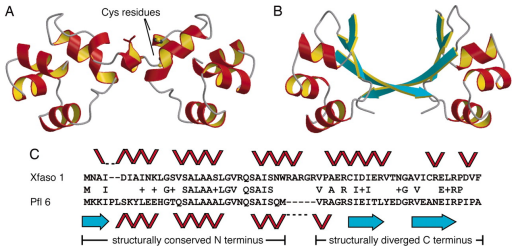
$Q=-10\log_{10}(p)$




```
                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
```


Deep scoring matrices cause overextension

Roessler C G et al. PNAS (2008)105:2343



BLOSUM80 40% id

```
Region: 12-37:14-34 : score=50; bits=26.0; LPr=4.1 :  DOMAIN_N: alpha
Region: 42-60:39-57 : score=26; bits=14.5; LPr=0.7 :  DOMAIN_N: beta
Smith-Waterman score: 82; E(1) < 2e-9; 49.0% identity (61.2% similar) in 49 aa overlap
                  10        20        30        40        50        60        70
3BD1     MNAIDIAINKLGSVSALAASLGVRQSAISNWRARGRVPAERCIDIERVTNGAVICRELRPDVFGASPAGHRPEASNAAA
                    :. :::::.::: :::::.      : : : :.:    .: :   :.::
2PIJ   XKKIPLSKYLEEHGTQSALAAALGVNQSAISQ-----MVRAGRSIEITLYEDGRVEANEIRPIPARPKRTAA
             [  10        20        30        ]    [0        50        60    ]
```

VTML40 70% id

```
Region: 12-29:14-31 : score=82; bits=42.9; LPr=9.2 :  DOMAIN_N: alpha
Smith-Waterman score: 82; E(1) < 6.5e-10 77.8% identity (88.9% similar) in 18 aa overlap
                  10        20        30        40        50        60        70
3BD1     MNAIDIAINKLGSVSALAASLGVRQSAISNWRARGRVPAERCIDIERVTNGAVICRELRPDVFGASPAGHRPEASNAAA
                    :. :::::.::: :::::
2PIJ   XKKIPLSKYLEEHGTQSALAAALGVNQSAISQMVRAGRSIEITLYEDGRVEANEIRPIPARPKRTAA
             [  10        20        30  ]     40        50        60
```

Empirical matrix performance
(median results from random alignments)

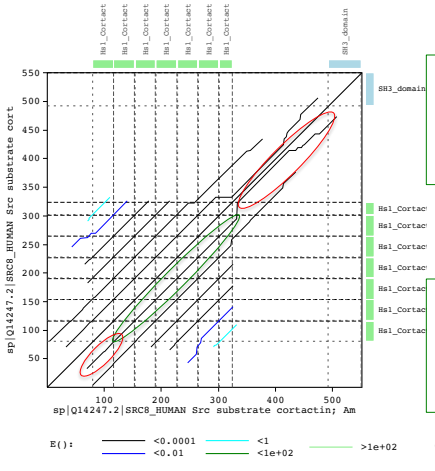
Matrix	target % ident	bits/position	aln len (50 bits)
VT160 -12/-2	23.8	0.26	192
BLOSUM50 -10/-2	25.3	0.23	217
BLOSUM62* -11/-1	28.9	0.45	111
VT120 -11/-1	27.4	1.03	48
VT80 -11/-1	51.9	1.55	32
PAM70* -10/-1	33.8	0.64	78
PAM30* -9/-1	45.5	1.06	47
VT40 -12/-1	72.7	2.76	18
VT20 -15/-2	84.6	3.62	13
VT10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

HMMs can be very "deep"
What is a "deep" matrix? a "shallow" matrix?

Pearson (2013) Curr. Protoc. Bioinformatics 3.5.1

Scoring matrices affect alignment boundaries
(homologous over-extension)

BLOSUM62 -11/-1



BLOSUM62 -11/-1

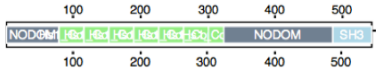
```
32- 42: 69- 79 : Id=0.455; Q= 0.0 : NODOM :0
43- 79: 80-116 : Id=0.158; Q= 0.0 : Hs1_Cortactin
 80-116:117-153 : Id=0.622; Q=37.4 : Hs1_Cortactin
117-153:154-190 : Id=0.757; Q=50.2 : Hs1_Cortactin
154-190:191-227 : Id=0.811; Q=61.0 : Hs1_Cortactin
191-227:228-264 : Id=0.568; Q=35.3 : Hs1_Cortactin
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
```

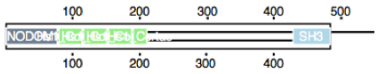
VTML80 -10/-1

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
```



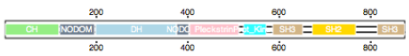

```
>>sp|SRC8_HUMAN Src substrate cortactin (550 aa)
>>sp|HCLS1_MOUSE Hematopoiet ln cell-sp (486 aa)
44.1% id (1-548:1-485)  E(454402): 4.1e-61

  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
```



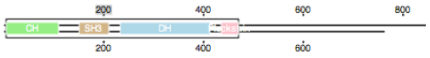
Q = -10 log(p)
Q>30.0 -> p < 0.001

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
```



Alignment statistics II / Algorithms II

- Foundation of homology from excess similarity
  - Unrelated sequence similarity scores are indistinguishable from "random" scores
  - Not-random ➔ not unrelated
- what is the probability of an alignment score?
  - given two sequences
  - after a database search
  - after N (100-10,000) database searches
- Hidden Markov Models
  - transition state models
  - profile HMMs

Multiple sequence alignment

- *No multiple alignments without* ***HOMOLOGY***
- Multiple sequence alignments can resolve ambiguous gaps – largely used to specify gap positions
- Optimal methods are $O(n^k)$ – impractical for > 5 sequences
- Most programs build successive pair-wise alignments (progressive alignment) – Clustal-W (Clustal-Omega), T-coffee, MUSCLE
- Simple progressive alignment methods fix gaps early, after which they cannot be moved
- Iterative approaches required to adjust gaps
- Tree-based alignments bring a more phylogenetic perspective
- What is the "correct" answer?

Multiple sequence alignment

- Why multiple sequence alignment (MSA)?
  - identify conserved (functional?) positions among related sequences
  - input to evolutionary tree methods
- MSA computational complexity
  - Models for MSA: tree-based, Sum-of-pairs, star
  - "optimal" $O(N^k)$ (k sequences of length N)
  - progressive: $O(k^2N^2)$
  - progressive/iterative: $O(k^2N^2)$
- Evaluating MSA accuracy
  - BALIBASE
  - are structural alignments correct?

First exam sample questions– 2 hours, collab Due Monday, Feb. 26 at 5:00 PM

1. Statistical estimates based on sequence shuffling on the fasta.bioch web site typically shows the expectation value as E(10,000).
   a. What does E(10,000) mean?
   b. Since only two sequences are being compared, why does it make sense to present E(10,000)? What E() context would be more appropriate?
2. In the similarity searching exercise, you were asked to find the highest scoring non-homolog in the search.
   a. If the statistical estimates are accurate, what should the Expect (E()-value) be for the highest scoring unrelated sequence (approx.)?
   b. are all sequences with scores worse than the highest scoring non-homolog non-homologous?
3. Expectation values -
   a. What is the range of Expect values (smallest and largest) in a database search of the human proteome, with 44,000 proteins?
   b. Expect values are corrected by the size of the database for a single query; E()<0.001 means that a score this good would occur less than once in 1000 searches by chance. What Expect threshold should you choose if you wanted a 1% (0.01) chance of getting a similarity score by chance after a large scale genome analysis that required 10,000 searches?
   c. What kinds of errors might occur because you adjusted the Expect threshold to the value you chose in part (b)?

First exam sample questions– 2 hours, collab Due Monday, Feb. 26 at 5:00 PM

4. A Pfam annotation suggests that a domain with model length 200 aligns in two places to a 150 residue protein. One location has (seq_start,seq_end) = (1,60), with (hmm_start,hmm_end) = (11,70), while the other location has (seq_start, seq_end)=(61,150) and (hmm_start, hmm_end) = (111,200).
   a) Do these mappings of domain regions make biological sense? Why or why not?
   b) Give an explanation for the annotation that makes biological sense.
   c) Give an explanation for the annotation that suggests some kind of artifact.
5. What is the expectation (E()) for a pairwise alignment with a score of 45 bits between two average length proteins (400 aa) in a search of the human proteome (44,000 proteins)
   a) If the 45 bit score were produced by a 200 residue alignment, what is the expected percent identity (approximately) and what scoring matrix should be used?
   b) If the score were produced by a 50 residue alignment, what would be the best scoring matrix and expected percent identity?
6. Why would raising the gap penalty improve the E()-value for very closely related sequences, but reduce the significance (increase the E()-value) for distantly related sequences?