Genomics

3 – 5 September

September					
	M 1		Labor Day		
4	W 3	MG	Database Searching		Ch. 6
5	F 5	MG	Database Searching	Hw1 due	

Reading: Mount ch 6 – Sequence database searching for similar sequences
focus today: 240-248

Homework – online test on tree thinking

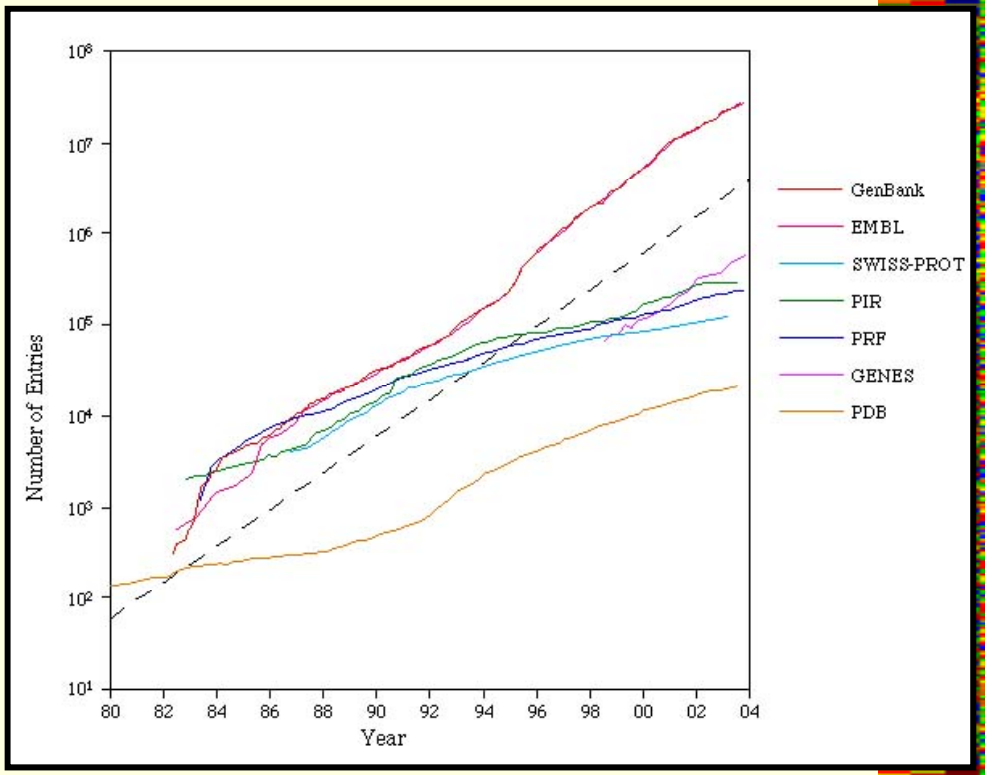
Mount 2nd Edition @ google books

Homework 1
Tree thinking online quiz.

Genomics

Sequence database Searching

- *Big problem is database size*
- *Bigger database means longer search. Alignments are $O(n^{2})$*
- *Bigger database means worse signal to noise ratio*
- *Sequence data doubles every 12-14 months*



Sequence database searching

- ***historically, database searching has been thought of as an approximation of pairwise sequence matching (alignment)***

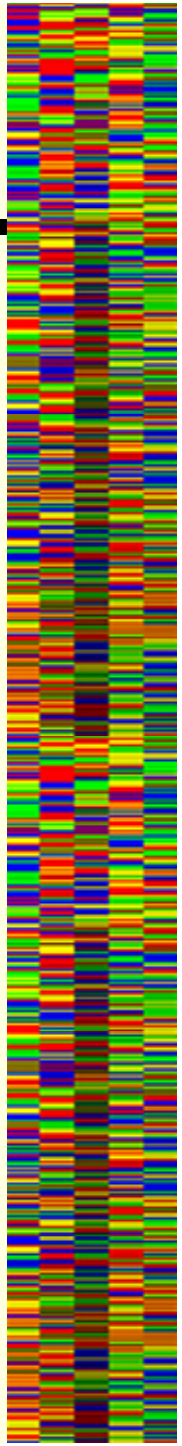
```
  1 VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF.... 46
    .||.::.  | ..||||:|. .:.|.|.| |:| : |.| . |..|
  1 GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLK 50
              .          .          .          .
 47 ..DLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVD 94
      |  .:|.::| || .| .||.. : . :. ...:.:|.: || | ::.
 51 SEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIP 100
              .          .          .          .
 95 PVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR 141
    .  :.::|.|:: .|... |::|.:..::.::| |. ... :.|.|:
101 VKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYK 147
```

- ***however, database searching is different in many ways, and is usually the first operation – it makes no sense to align sequences unless you already have decided they are homologous.***

Genomics

Sequence database searching

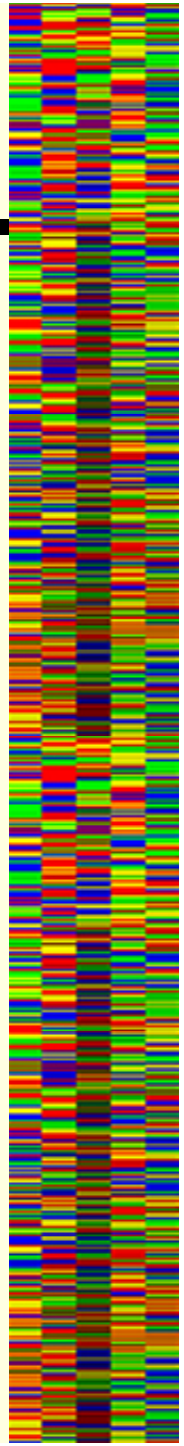
- ***Goal: Find sequences that are homologous***
  - Since we can't see homology,
    Find sequences that are significantly more similar than unrelated sequences should be
- ***Sequence databases are large***
  - 9×10^{10} nucleotides
  - 1.4×10^{8} residues
- ***Checking every possible comparison with our query is***
  - time consuming
  - mostly unnecessary (since most sequences are not a match)
    IF there is a faster way



Genomics

Sequence database searching

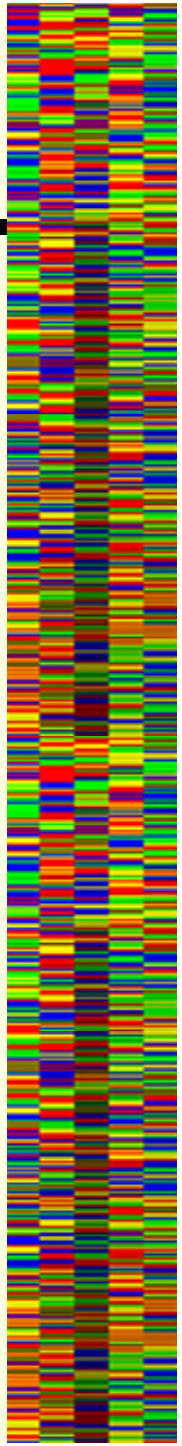
- ***Basic strategy***
- ***Break down database searching into two parts***
  - Skip the uninteresting (=unrelated) sequences
  - Match related sequences to allow information mapping
- ***Only check the database sequences that are most likely to be matches in detail***
- ***Two main programs***
  - FASTA
  - BLAST (most widely used)



Genomics

Sequence database searching

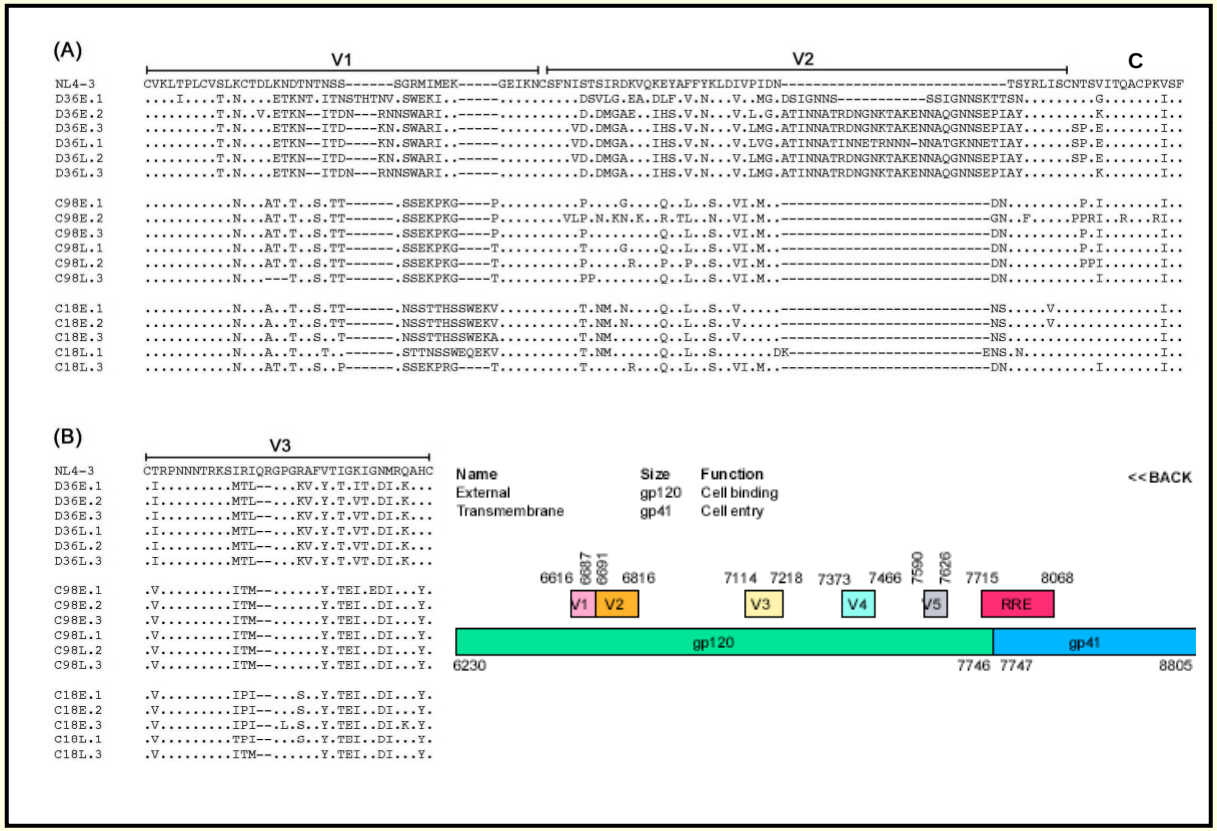
- ***Sequences that are related tend to have their similar bases or residues in clusters***
  - Proteins
    - protein structure and function
  - Genes – many constraints
    - encoded protein
    - exon/intron structure
    - GC content
    - regulatory and other protein binding

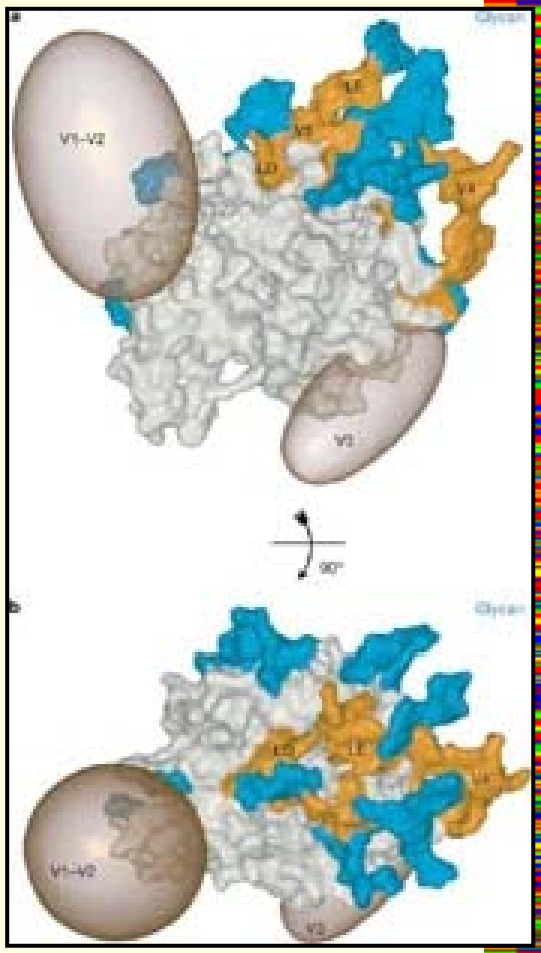


Genomics

HIV Env Gene

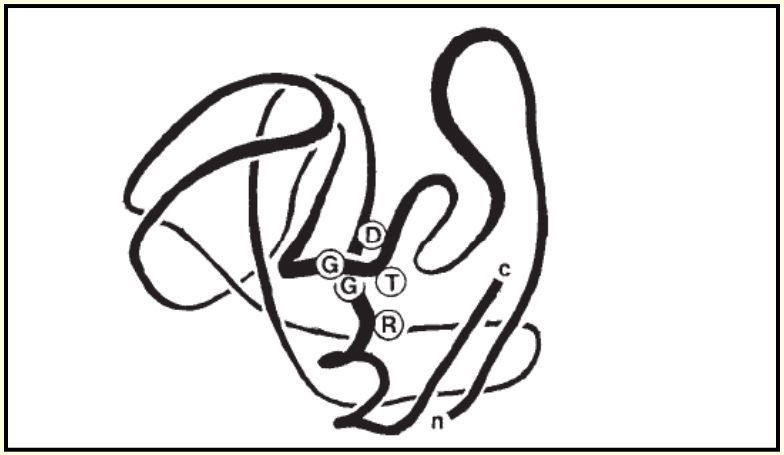
- **Conserved (C) and Variable (V) regions**

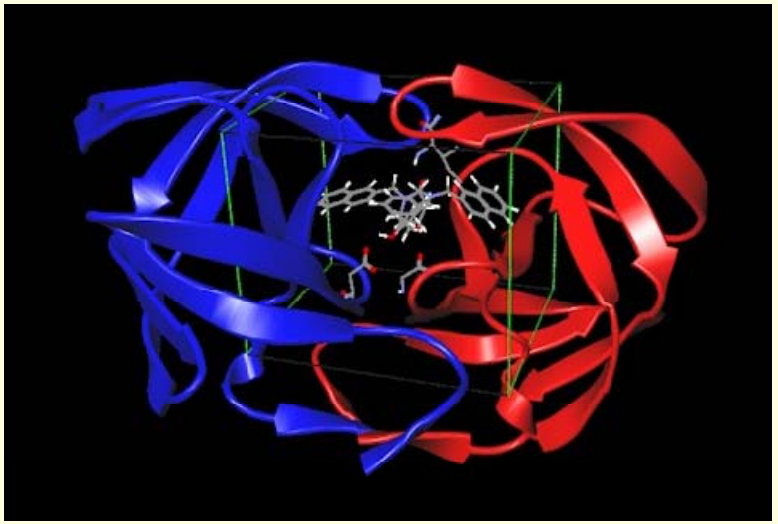




Genomics

HIV Protease





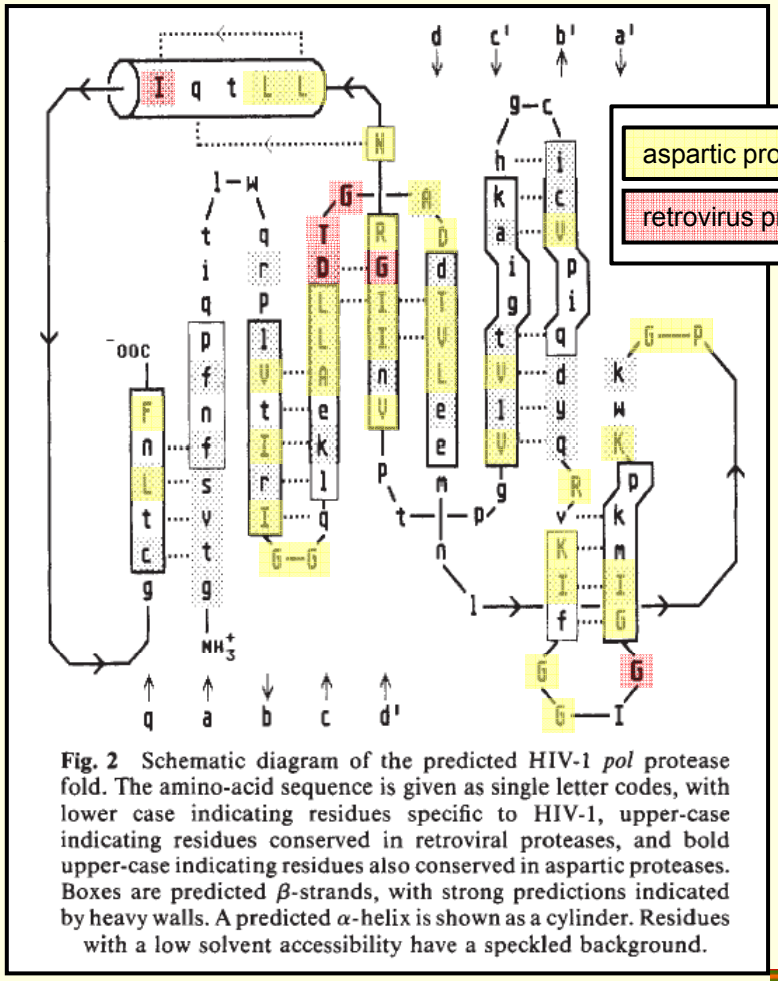


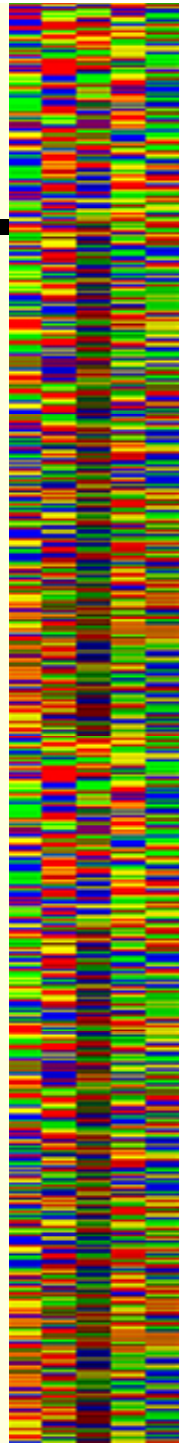
Fig. 2 Schematic diagram of the predicted HIV-1 *pol* protease fold. The amino-acid sequence is given as single letter codes, with lower case indicating residues specific to HIV-1, upper-case indicating residues conserved in retroviral proteases, and bold upper-case indicating residues also conserved in aspartic proteases. Boxes are predicted β-strands, with strong predictions indicated by heavy walls. A predicted α-helix is shown as a cylinder. Residues with a low solvent accessibility have a speckled background.

Pearl & Taylor, Nature 329, 351-354 ,1987

Genomics

Sequence database searching

- ***How do you get rid of non-matching sequences?***
  - Look for sequences that do not have any strong "diagonals" in their match with the query
  - Considerations
    - How to we measure "similarity" or "strong diagonal"?
    - How do we decide how much similarity is needed?
    - How many "diagonals" do unrelated sequences have?
    - How many do related sequences have?

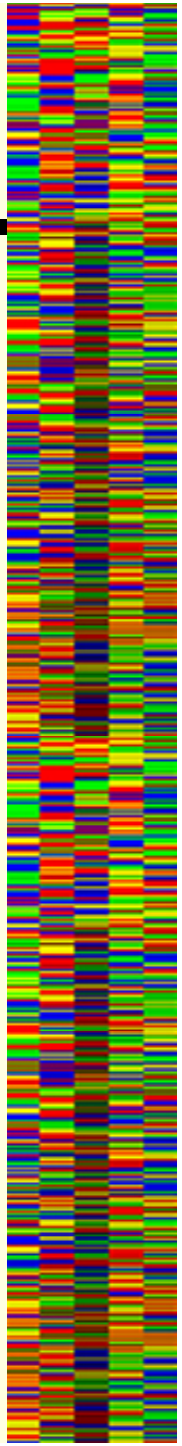


Genomics

Sequence database searching

- ***hypothesis: a pair sequences are unrelated***

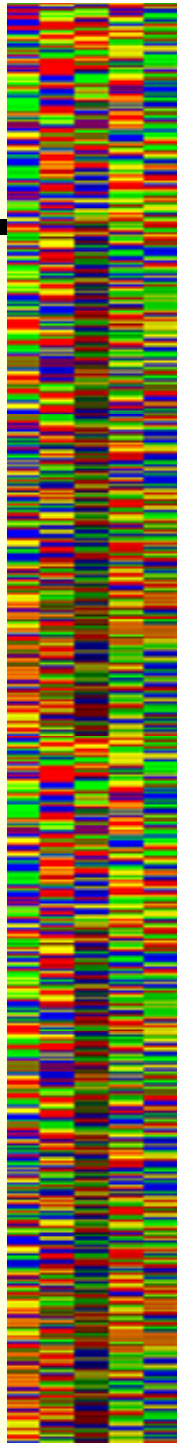
  If ***the pair of sequences have a more significant match than could reasonably be expected from unrelated (not homologous) sequences***
  Then ***they must be related (homologous)***



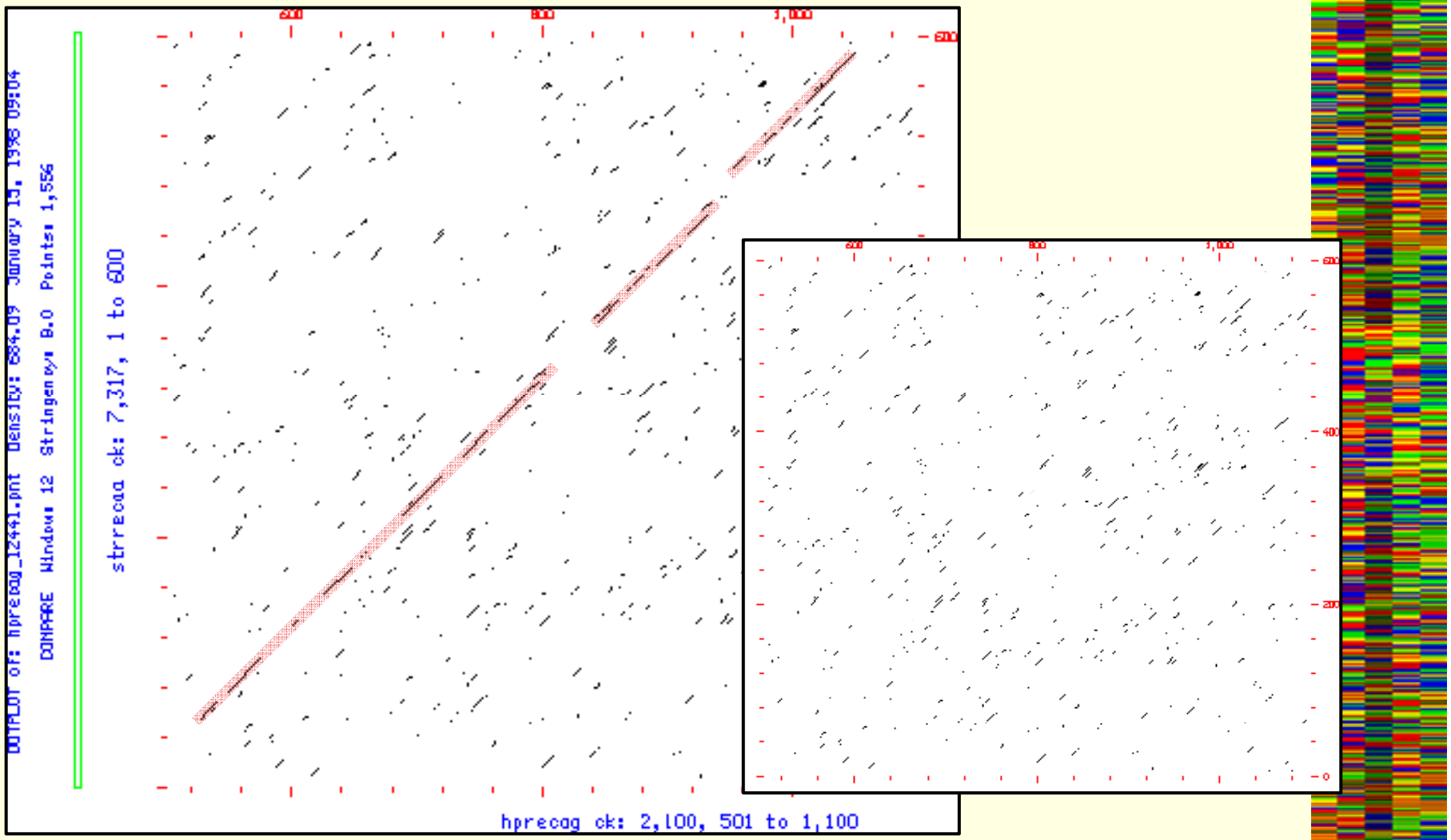
Genomics

Sequence database searching

- ***How good do you have to be?***
  - average size protein ~300 residues in 1.4 x 10^8 residues
    - 300/1.4 x 10^8 residues ~ 1 in 500,000
  - average size gene ~10kbp in 9 x 10^{10} nucleotides
    - 10^4/9 x 10^{10} nucleotides ~ 1 in 9,000,000

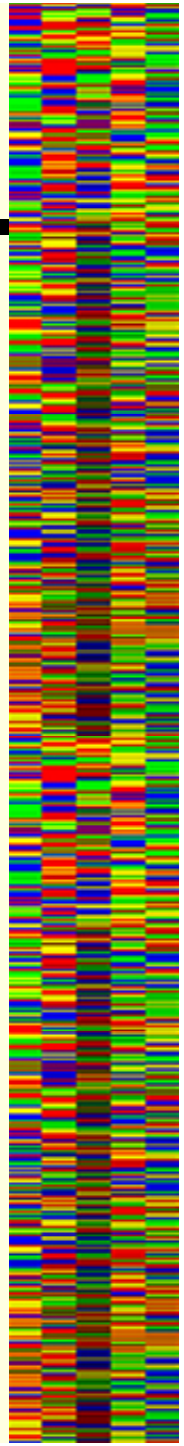


Sequence database searching



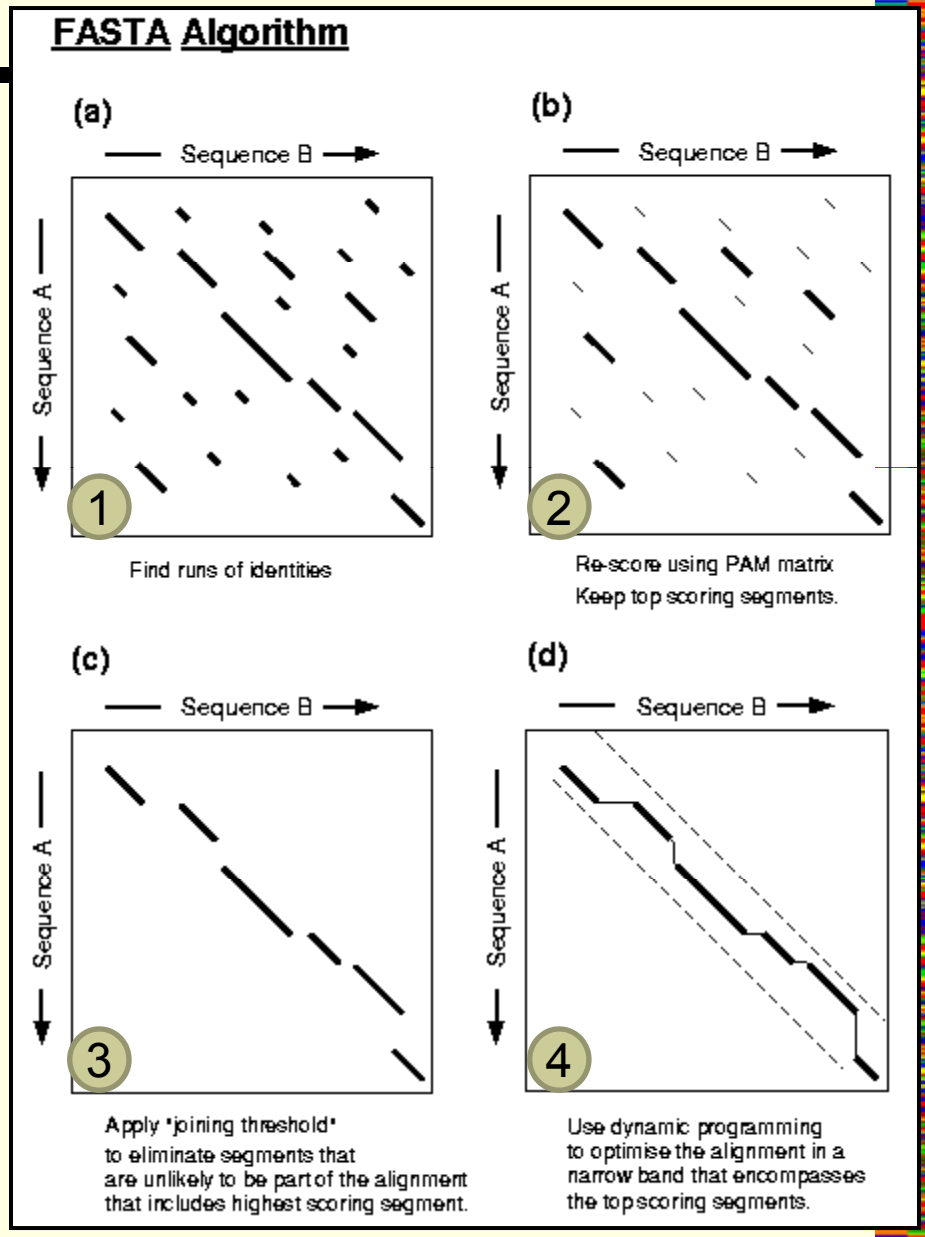
FASTA

- ***Originally developed in the mid-1980s as FASTN and FASTP for nucleic acid and protein, respectively***
- ***Fast approximation of dynamic programming alignment***
- ***Relies on related sequences having "diagonals" with high similarity***
  - Step 1. Find best regions on diagonals
  - Step 2. Rescan 10 best regions with PAM scoring table
  - Step 3. Join initial regions
  - Step 4. Calculate dynamic programming optimal alignment
  - Step 5. Calculate significance of Scores
- ***Heuristic - based on an expert's (Bill Pearson, UVa) experience and intuition***



Sequence database searching - FASTA

- ***Step 1. Find best regions on diagonals***
- ***Step 2. Rescan 10 best regions with scoring table***
- ***Step 3. Join initial regions***
- ***Step 4. Calculate dynamic programming optimal alignment***



Genomics

Sequence database searching - FASTA

- ***Step 1 - Find Initial Regions (Fast part of search)***

- ***Find best regions of diagonals using lookup table***
- ***Lookup table: lists all the words of length ktup and where they occur***

MYSEQVENCEN

CE	9
EN	7,10
EQ	4
MY	1
NC	8
QV	5
SE	3
UE	6
YS	2

HISSEQENCEQ

CE	9
EN	7
EQ	5,10
HI	1
IS	2
NC	8
QE	6
SE	4
SS	3

Sequence database searching - FASTA

- ***Step 1 - Find Initial Regions***

- ***For each matching word (ktup) calculate on which diagonal the match lies - AKA histograming***
  - diagonal = $offset_{database}$ - $offset_{query}$

CE	9	CE	9	0
EN	7,10	EN	7	0, +3
EQ	4	EQ	5,10	-1, -6
MY	1	HI	1	0

- ***Does it already have a region?***
- ***If no, start a region (score=pair score)***
- ***If yes, try to combine them***
  - score > distance to existing region (score = pair scores - distance)

Genomics

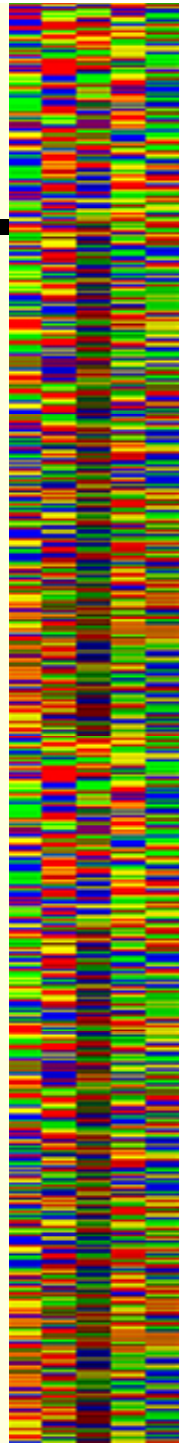
Sequence database searching - FASTA

Positive values are similar residues

A	2																						
R	-2	6																					
N	0	0	2																				
D	0	-1	2	4																			
C	-2	-4	-4	-5	12																		
Q	0	1	1	2	-5	4																	
E	0	-1	1	3	-5	2	4																
G	1	-3	0	1	-3	-1	0	5															
H	-1	2	2	1	-3	3	1	-2	6														
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5													
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6												
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5											
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6										
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9									
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6								
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2							
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3						
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17					
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10				
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4			
B	0	-1	2	3	-4	1	2	0	1	-2	-3	1	-2	-5	-1	0	0	-5	-3	-2	2		
Z	0	0	1	3	-5	3	3	-1	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X

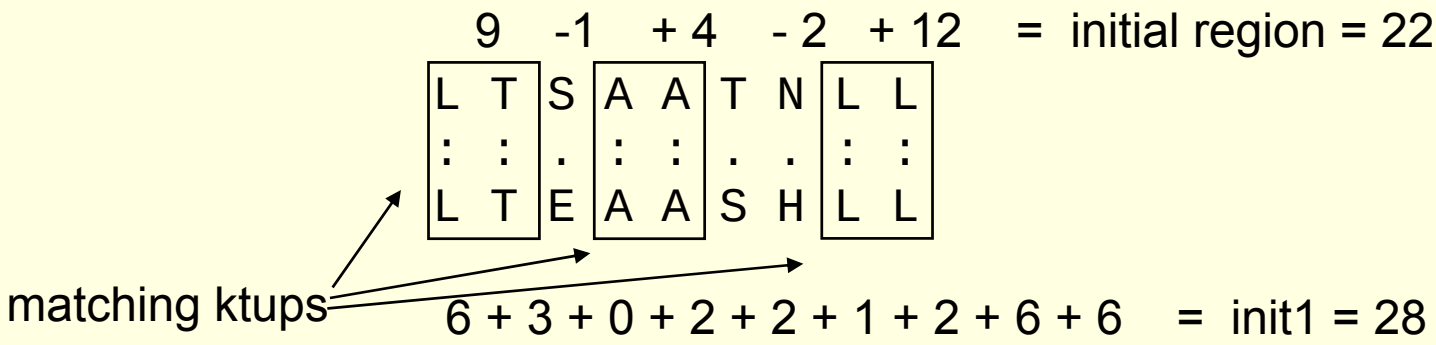
PAM 250 Scoring Table – scores amino acid residues based on chemical and genetic similarity

PROTEIN ONLY!



Sequence database searching - FASTA

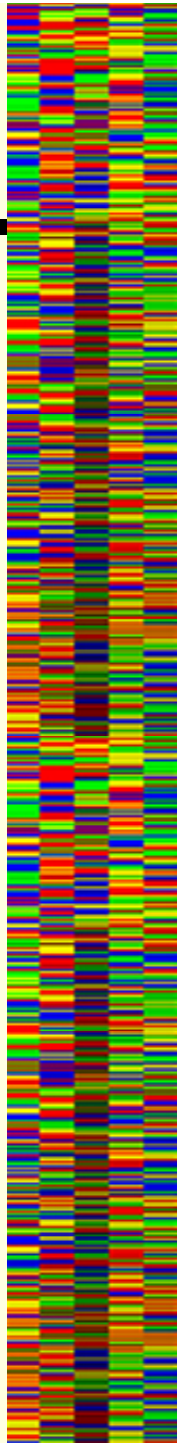
- *Step 2 - rescore 10 best initial diagonals*
- *Best single diagonal is init1 score*



Sequence database searching - FASTA

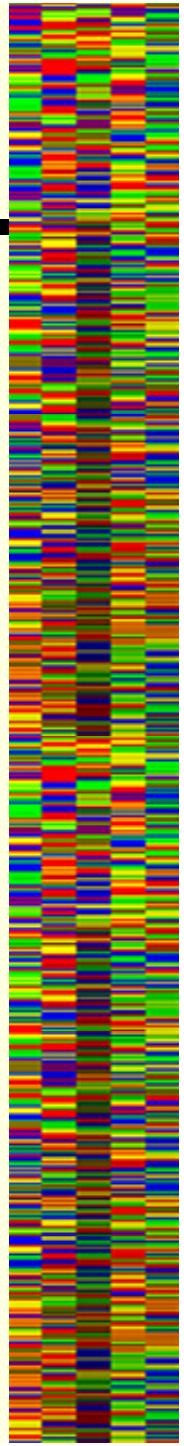
- ***Step 3 - Join Initial Diagonals if possible***

- ***Only consider sequences with init1 > cutoff***
- ***If distance between diagonals < gappen***
  - cutoff and gappen depend on length and ktup
  - cutoff is about 1 std. deviation above predicted average score
    - 28 for ktup=2 length=200
- ***Best joined score is initn***



Sequence database searching - FASTA

- ***Step 4 - Complete sequence matching (dynamic programming alignment)***
- ***Local dynamic programming alignment of sequences with init1 > cutoff***
  - alignment in band around init1
- ***This is the opt score***

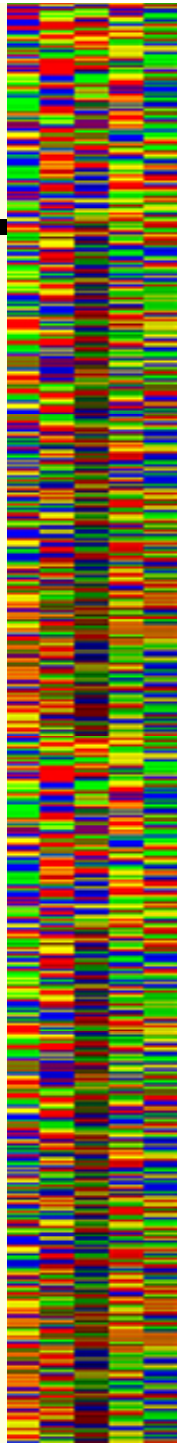


Sequence database searching - FASTA

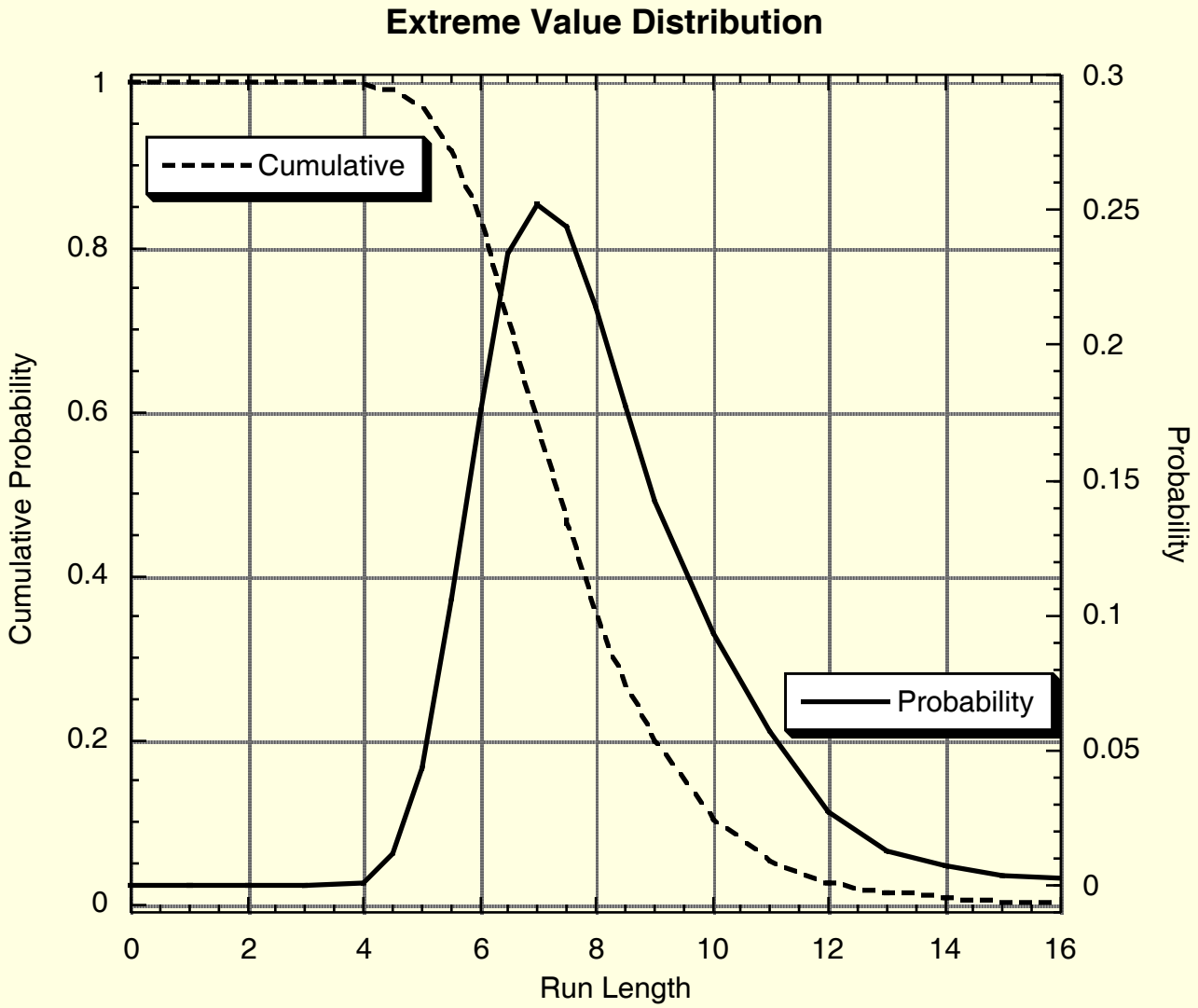
- ***Step 5 - Significance of scores – use internal unrelated sequence to estimate unrelated distribution***

- ***Calculate averages scores in pools of length***
- ***Fit a linear regression line to log(length) vs. score***
- ***Calculate Z scores for dispersion around line***
- ***Remove large Z scores and repeat several times***
  - Removes homologous sequences
- ***Convert to extreme value distribution P-value using***

$$P(Z>z) = 1 - \exp(e^{-1.2825z-0.5772})$$



Sequence database searching - FASTA



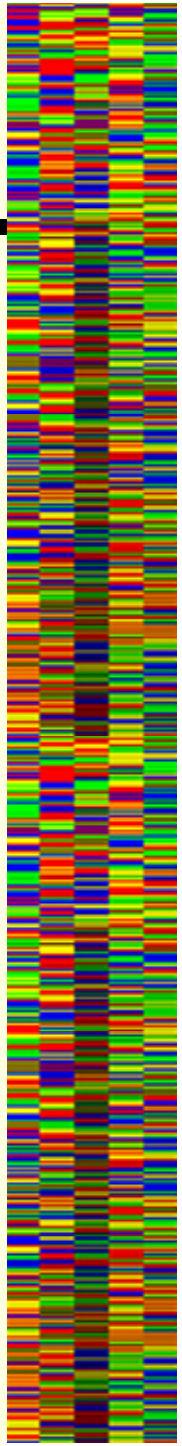
Genomics

FASTA

```
 opt      E()
< 20   876     0:==
  22     1     0:=          one = represents 483 library sequences
  24     4     0:=
  26    13     6:*
  28    61    68:*
  30   387   415:*
  32  1600  1603:===*
  34  4494  4348:=========*
  36  9575  8930:==================*=
  38 18356 14757:==============================*========
  40 23173 20585:==========================================*=====
  42 28394 25163:====================================================*======
  44 28936 27757:=========================================================*==
  46 27839 28271:==========================================================*
  48 25418 27066:=====================================================   *
  50 22278 24698:===============================================    *
  52 19638 21714:=========================================   *
  54 16240 18547:==================================    *
  56 16567 15493:================================*==
  58 12196 12719:==========================*
  60  9400 10303:==================== *
  62  7669  8260:================ *
  64  6124  6569:=============*
  66  4978  5192:==========*
  68  3972  4084:========*
  70  3235  3201:======*
  72  2232  2501:=====*
  74  1799  1950:====*
  76  1351  1518:===*
  78   951  1180:==*
  80   812   916:=*
  82   655   701:=*
  84   476   555:=*
  86   372   429:*
  88   301   332:*          inset = represents 3 library sequences
  90   213   257:*
  92   146   199:*         :=======================================*
  94   125   154:*         :=======================================*
  96    64   119:*         :======================                 *
  98    65    92:*         :======================        *
 100    51    71:*         :=================      *
 102    33    55:*         :===========       *
 104    14    43:*         :=====       *
 106    16    33:*         :======    *
 108    21    26:*         :======= *
 110     6    20:*         :==    *
 112     5    15:*         :==  *
 114     6    12:*         :== *
 116     8     9:*         :==*
 118     4     7:*         :==*
>120    46     6:*         :=*=============
94838015 residues in 301196 sequences
 statistics extrapolated from 50000 to 300969 sequences
  Expectation_n fit: rho(ln(x))= 5.8358+/-0.000529; mu= 2.8541+/- 0.030;
 mean_var=75.4981+/-14.859, 0's: 149 Z-trim: 36  B-trim: 0 in 0/64
 Kolmogorov-Smirnov  statistic: 0.0378 (N=29) at  44
```

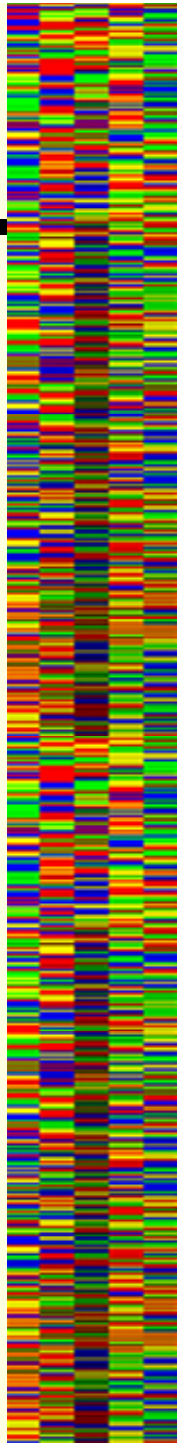
Sequence database searching - FASTA

- ***Evaluation of significance compares to the distribution of real unrelated proteins***
  - most sequences are unrelated so they can be used to judge how significant the scores are
  - Scores are normalized for length, then fit to extreme value distribution, i.e. they are corrected for length of database sequence
  - Unrelated sequence model has all properties of true sequences
- ***Look at the score histogram***
  - Look at where clearly unrelated sequences score
  - Look at where clearly related sequences score
- ***What might fool you?***
  - unusual compositions
  - transmembrane sequences
  - repeated sequences



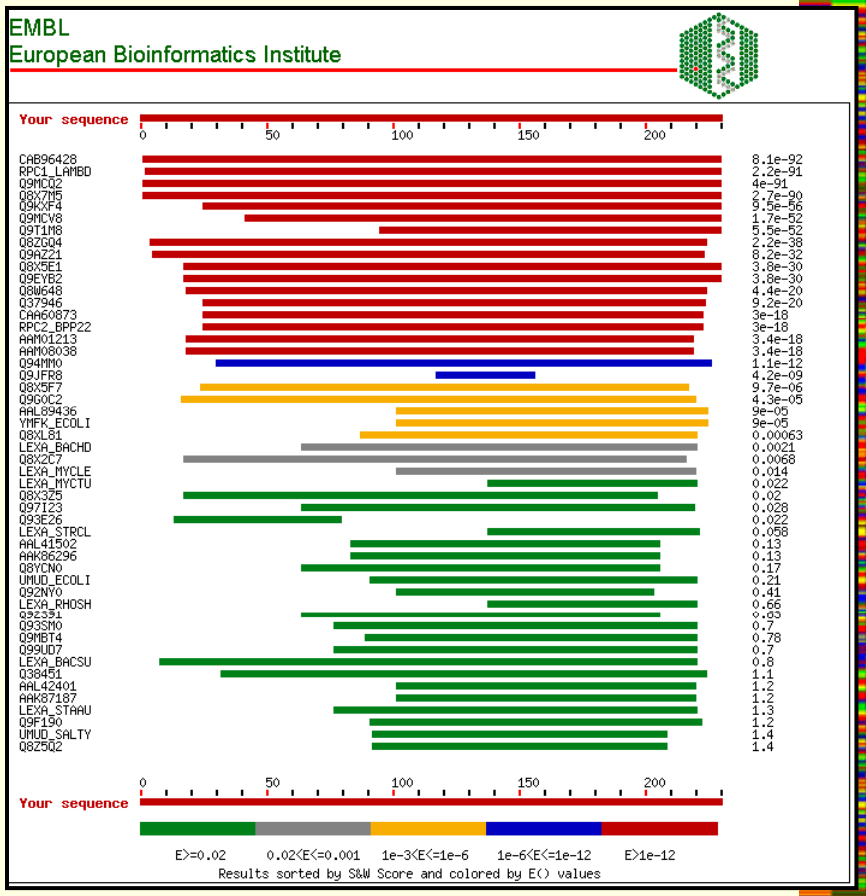
Sequence database searching - FASTA

- ***Parameters***
  - ktup (word length)
    - 1 - 2 for proteins
    - 4 - 6 for nucleic acids
  - Scoring Matrix
    - Default is probably BLOSUM 50
  - Gap penalties



Sequence database searching - FASTA

- *Locations of hits on query*



Genomics

FASTA

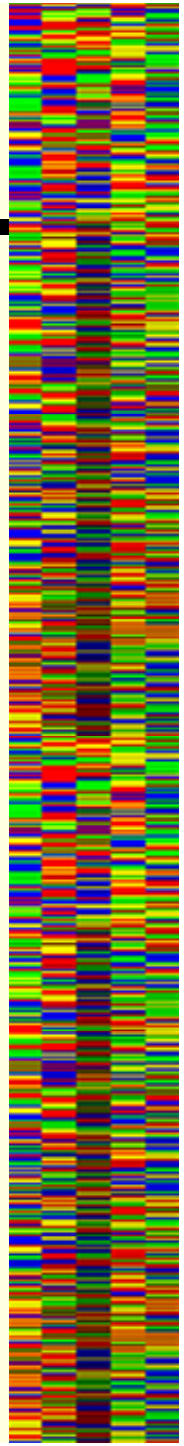
```
 opt      E()
< 20   876     0:==
  22     1     0:=          one = represents 483 library sequences
  24     4     0:=
  26    13     6:*
  28    61    68:*
  30   387   415:*
  32  1600  1603:===*
  34  4494  4348:=========*
  36  9575  8930:==================*=
  38 18356 14757:==============================*========
  40 23173 20585:==========================================*=====
  42 28394 25163:====================================================*======
  44 28936 27757:=========================================================*==
  46 27839 28271:==========================================================*
  48 25418 27066:=====================================================    *
  50 22278 24698:===============================================    *
  52 19638 21714:=========================================    *
  54 16240 18547:==================================    *
  56 16567 15493:================================*==
  58 12196 12719:==========================*
  60  9400 10303:==================== *
  62  7669  8260:================ *
  64  6124  6569:=============*
  66  4978  5192:==========*
  68  3972  4084:========*
  70  3235  3201:======*
  72  2232  2501:=====*
  74  1799  1950:====*
  76  1351  1518:===*
  78   951  1180:==*
  80   812   916:=*
  82   655   701:=*
  84   476   555:=*
  86   372   429:*
  88   301   332:*          inset = represents 3 library sequences
  90   213   257:*
  92   146   199:*         :=======================================*
  94   125   154:*         :=======================================*
  96    64   119:*         :======================                 *
  98    65    92:*         :======================        *
 100    51    71:*         :=================      *
 102    33    55:*         :===========      *
 104    14    43:*         :=====         *
 106    16    33:*         :======    *
 108    21    26:*         :======= *
 110     6    20:*         :==    *
 112     5    15:*         :==  *
 114     6    12:*         :== *
 116     8     9:*         :==*
 118     4     7:*         :==*
>120    46     6:*         :=*=============
94838015 residues in 301196 sequences
 statistics extrapolated from 50000 to 300969 sequences
  Expectation_n fit: rho(ln(x))= 5.8358+/-0.000529; mu= 2.8541+/- 0.030;
 mean_var=75.4981+/-14.859, 0's: 149 Z-trim: 36  B-trim: 0 in 0/64
 Kolmogorov-Smirnov  statistic: 0.0378 (N=29) at  44
```

Genomics

FASTA

- *Top Scores*

```
FASTA (3.15 August, 1998) function (optimized,
/ebi/services/idata/appbin/matrix/aa/blo matrix) ktup: 2
 join: 36, opt: 24, gap-pen: -12/ -2, width:  16 reg.-scaled
 Scan time: 86.867
The best scores are:                                      initn init1 opt z-sc E(300969)
SWALL:RPC1_LAMBD P03034 REPRESSOR PROTEI  ( 236) 1558 1558 1558 1803.1 3.8e-93
SWALL:Q37946 Q37946 REPRESSOR PROTEIN C2  ( 220)  321  156  425 499.6 1.5e-20
SWALL:RPC2_BPP22 P03035 REPRESSOR PROTEI  ( 216)  305  146  401 472.1 5.2e-19
SWALL:E264367 E264367 BACTERIOPHAGE ES18  ( 216)  305  146  401 472.1 5.2e-19
SWALL:D1036957 D1036957 REPRESSOR PROTEI  ( 224)  158   74  188 226.7 2.4e-05
SWALL:P75974 P75974 FROM BASES 1195814 T  ( 224)  158   74  188 226.7 2.4e-05
SWALL:D1036969 D1036969 REPRESSOR PROTEI  ( 224)  158   74  188 226.7 2.4e-05
SWALL:Q49848 Q49848 LEXA. 11/98           ( 235)   84   61  153 186.1 0.0044
SWALL:Q50765 Q50765 LEXA GENE. 11/98      ( 217)   64   64  150 183.2 0.0064
SWALL:O86847 O86847 LEXA PROTEIN. 11/98   ( 264)   34   34  144 175.0 0.018
SWALL:D1037024 D1037024 UMUD PROTEIN. .   ( 139)   69   69  132 165.5 0.062
SWALL:D1037016 D1037016 UMUD PROTEIN. .   ( 139)   69   69  132 165.5 0.062
SWALL:UMUD_ECOLI P04153 UMUD PROTEIN (EC  ( 139)   69   69  132 165.5 0.062
SWALL:LEXA_BACSU P31080 SOS REGULATORY P  ( 205)   94   65  125 154.8 0.24
SWALL:Q38451 Q38451 PUTATIVE REPRESSOR.   ( 240)   33   33  124 152.6 0.33
SWALL:UMUD_SALTY P22493 UMUD PROTEIN (EC  ( 139)   56   56  119 150.5 0.42
SWALL:Q52622 Q52622 REGULATORY TRANSCRIP  (  84)   94   94  116 150.5 0.43
SWALL:Q38089 Q38089 REPRESSOR PROTEIN. 1  ( 278)   46   46  122 149.3  0.5
SWALL:Q38327 Q38327 REPRESSOR PROTEIN. 1  ( 297)   46   46  122 148.9 0.52
SWALL:LEXA_AERHY Q44069 LEXA REPRESSOR (  ( 207)  113   55  119 147.9  0.6
SWALL:O32506 O32506 LEXA PROTEIN. 11/98   ( 210)   91   61  119 147.8 0.61
SWALL:Q38158 Q38158 REPRESSOR PROTEIN. 1  ( 256)   33   33  117 144.1 0.97
SWALL:G4063729 G4063729 UMUD MUCA HOMOLO  ( 224)   27   27  115 142.7  1.2
SWALL:O86948 O86948 LEXA REPRESSOR (EC 3  ( 197)   56   56  114 142.4  1.2
SWALL:O33927 O33927 LEXA. 11/98           ( 197)   52   52  110 137.8  2.2
SWALL:E1360412 E1360412 APS REDUCTASE PR  ( 454)   52   52  114 136.8  2.5
SWALL:G1688105 G1688105 MUCAB PROTEINS.   ( 145)   29   29  107 136.4  2.6
SWALL:MUCA_SALTY P07376 MUCA PROTEIN (EC  ( 146)   29   29  105 134.1  3.5
SWALL:O69902 O69902 PUTATIVE TRANSCRIPTI  (  63)   77   77  100 134.0  3.6
SWALL:O52206 O52206 MUCA. 11/98           ( 144)   61   38  104 133.0    4
SWALL:Q38607 Q38607 REPRESSOR PROTEIN. 1  ( 286)  103   71  108 133.0    4
```

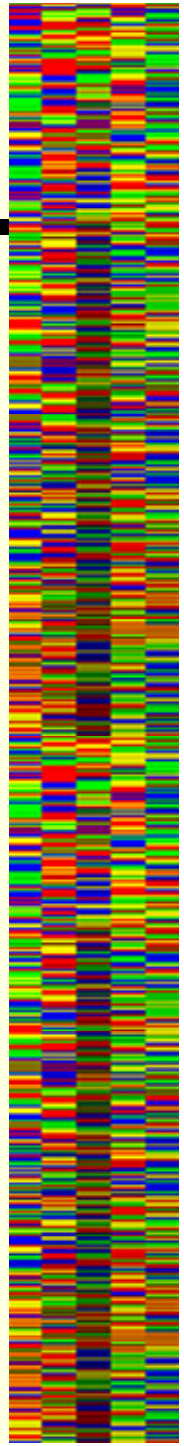


Genomics

FASTA

- ***Optimized local alignment in region of best region***

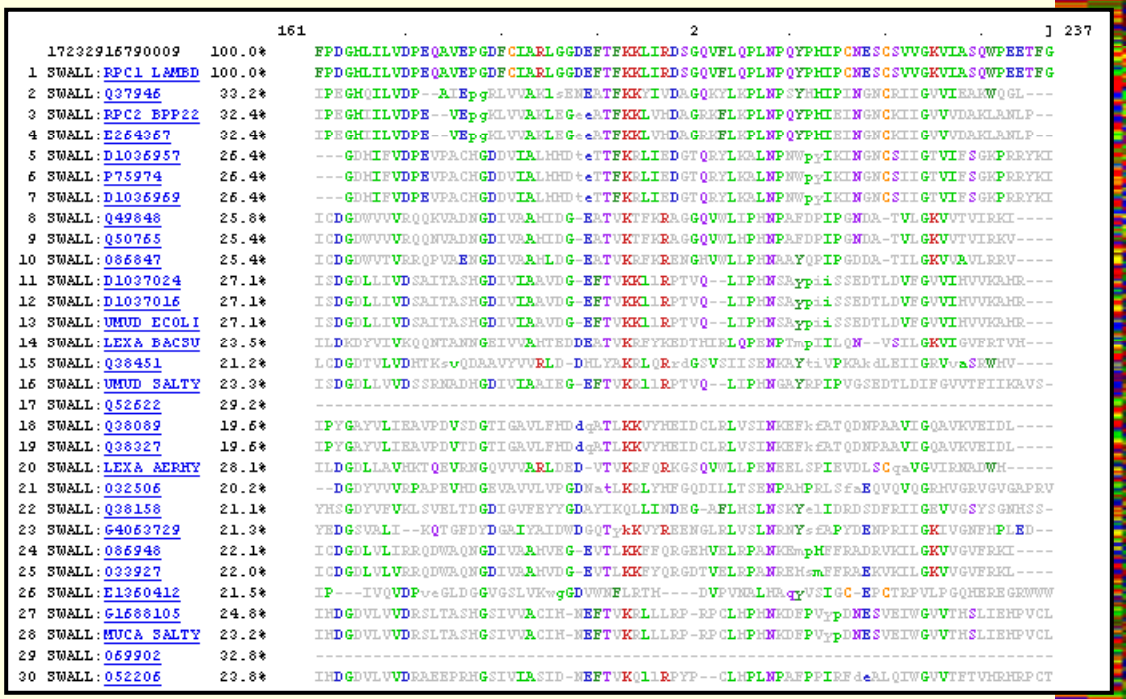
```
>>SWALL:Q49848 Q49848 LEXA. 11/98                          (235 aa)
 initn:  84 init1:  61 opt: 153 Z-score: 186.1 expect() 0.0044
Smith-Waterman score: 168;  31.200% identity in 125 aa overlap

           80        90       100       110       120       130
gi|133 EEFSPSIAREIYEMYEAVSMQPSLRSEYEYPVFSHVQAGMFSPELRTFTKGDAERWVSTT
                                        :..... ::  ::    :..: .:
SWALL: VRGVEETQAAGPAVLTEVAGSDVLPEPTFVPILGRIAAG--SP---IFAEGTVEDIFPLP
         90       100       110       120             130       140

            140       150       160       170       180       190
gi|133 KK--ASDSAFWLEVEGNSMTAPTGSKPSFPDGMLILVDPEQAVEPGDFCIARLGGDEFTF
       ..  .  . : :.: :.::.     . .. ::  ..:  ..... ::.  : . : : :
SWALL: RELVGEGTLFLLKVTGDSMV-----EAAICDGDWVVVRQQKVADNGDIVAAMIDG-EATV
             150       160            170       180       190

            200       210       220       230
gi|133 KKLIRDSGQVFLQPLNPQYPMIPCNESCSVVGKVIASQWPEETFG
       : . : .::::: : :: .  :: :.. .:::::..
SWALL: KTFKRAGGQVWLIPHNPAFDPIPGNDA-TVLGKVVTVIRKI
         200       210       220       230
```



Genomics

FASTA

- *Mview*
- *Example of a leader-follower multiple sequence alignment*
- *All sequences are aligned only to the FASTA query sequence*



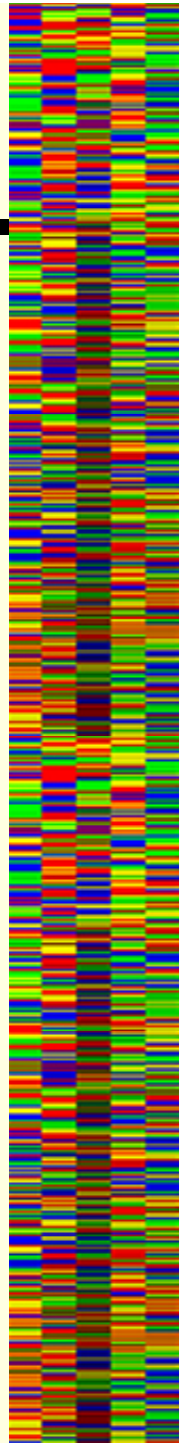
Genomics

FASTA

```
FASTA (3.15 August, 1998) function (optimized, /ebi/services/idata/appbin/matrix/aa/blo matrix) ktup: 1
 join: 42, opt: 30, gap-pen: -12/ -2, width:  32 reg.-scaled
 Scan time: 230.300
The best scores are:                                      initn init1 opt z-sc E(300954)
SWALL:RPC1_LAMBD P03034 REPRESSOR PROTEI  ( 236) 1558 1558 1558 1889.0 6.2e-98
SWALL:Q37946 Q37946 REPRESSOR PROTEIN C2  ( 220)  422  176  425 516.3 1.8e-21
SWALL:E264367 E264367 BACTERIOPHAGE ES18  ( 216)  314  174  401 487.4 7.3e-20
SWALL:RPC2_BPP22 P03035 REPRESSOR PROTEI  ( 216)  314  174  401 487.4 7.3e-20
SWALL:P75974 P75974 FROM BASES 1195814 T  ( 224)  135   92  188 229.0 1.8e-05
SWALL:D1036969 D1036969 REPRESSOR PROTEI  ( 224)  135   92  188 229.0 1.8e-05
SWALL:D1036957 D1036957 REPRESSOR PROTEI  ( 224)  135   92  188 229.0 1.8e-05
SWALL:Q49848 Q49848 LEXA. 11/98           ( 235)   90   90  168 204.4 0.00043
SWALL:Q50765 Q50765 LEXA GENE. 11/98      ( 217)   93   93  150 183.1 0.0065
SWALL:O69979 O69979 SOS REGULATORY PROTE  ( 234)   79   79  148 180.2 0.0095
SWALL:O86847 O86847 LEXA PROTEIN. 11/98   ( 264)   80   80  144 174.4 0.02
SWALL:RPC1_BPD3 Q37906 REPRESSOR PROTEIN  ( 223)   65   65  138 168.4 0.043
SWALL:LEXA_BACSU P31080 SOS REGULATORY P  ( 205)  114   86  135 165.4 0.063
SWALL:UMUD_ECOLI P04153 UMUD PROTEIN (EC  ( 139)   69   69  132 164.6 0.07
SWALL:D1037024 D1037024 UMUD PROTEIN. .   ( 139)   69   69  132 164.6 0.07
SWALL:D1037016 D1037016 UMUD PROTEIN. .   ( 139)   69   69  132 164.6 0.07
SWALL:LEXA_SALTY P29831 LEXA REPRESSOR (  ( 202)   95   58  130 159.4 0.14
SWALL:LEXA_ECOLI P03033 LEXA REPRESSOR (  ( 202)   95   58  128 157.0 0.19
SWALL:LEXA_AERHY Q44069 LEXA REPRESSOR (  ( 207)  108   63  128 156.8 0.19
SWALL:Q54446 Q54446 HYPOTHETICAL 26.5 KD  ( 228)   82   82  127 154.9 0.24
SWALL:SAMA_SALTY P23831 SAMA PROTEIN (EC  ( 140)   87   52  122 152.4 0.33
SWALL:Q38451 Q38451 PUTATIVE REPRESSOR.   ( 240)   56   56  124 150.9 0.41
SWALL:Q52622 Q52622 REGULATORY TRANSCRIP  (  84)  106  106  116 148.9 0.52
SWALL:UMUD_SALTY P22493 UMUD PROTEIN (EC  ( 139)   66   66  119 148.9 0.53
SWALL:Q38089 Q38089 REPRESSOR PROTEIN. 1  ( 278)   46   46  122 147.4 0.64
SWALL:Q38327 Q38327 REPRESSOR PROTEIN. 1  ( 297)   46   46  122 146.9 0.68
SWALL:O32506 O32506 LEXA PROTEIN. 11/98   ( 210)   92   62  119 145.8 0.78
SWALL:O33927 O33927 LEXA. 11/98           ( 197)   52   52  117 143.9    1
SWALL:IMPA_SALTY P18641 IMPA PROTEIN (EC  ( 145)   60   60  115 143.7    1
SWALL:G4138833 G4138833 IMPA. 1/99        ( 145)   60   60  115 143.7    1
SWALL:E1360412 E1360412 APS REDUCTASE PR  ( 454)   77   77  121 142.5  1.2
SWALL:Q38158 Q38158 REPRESSOR PROTEIN. 1  ( 256)   73   61  117 141.9  1.3
SWALL:G4063729 G4063729 UMUD MUCA HOMOLO  ( 224)   50   50  115 140.5  1.5
SWALL:O86948 O86948 LEXA REPRESSOR (EC 3  ( 197)   56   56  114 140.2  1.6
SWALL:O64370 O64370 REPRESSOR. 8/98       ( 224)   60   60  113 138.1  2.1
SWALL:LEXA_PSEAE P37452 LEXA REPRESSOR (  ( 204)   66   66  110 135.1  3.1
SWALL:E1358521 E1358521 LEXA REPRESSOR (  ( 204)   66   66  110 135.1  3.1
SWALL:G1688105 G1688105 MUCAB PROTEINS.   ( 145)   45   45  107 134.0  3.5
```

Ktup=1

E<1 with Ktup = 2



Genomics

FASTA

- ***Output Alignments***

Ends of init1 region

```
NAHR_PSEPU TRANSCRIPTIONAL ACTIVATOR PROTEIN NAHR.      112    76   185
  19.6% identity in 276 aa overlap

                10        20        30        40        50
LYSR_E   MAAVNLRHIEIFHAVMTAGSLTEAAHLLHTSQPTVSRELARFEKVIGLKLFERVRGRL
          . ...:. . .:..... ... .:.  .X:.::..: :. . .   :: :..  .
NAHR_P MELRDLDLNLLVVFNQLLVDRRVSITAENLGLTQPAVSNALKRLRTSLQDPLFVRTHQGM
              10        20        30        40        50        60

        60        70        80        90       100       110
LYSR_E HPTVQGLRLFEEVQRSWYGLDRIVSAAESLREFRQGELSIACLPVFSQS-FLPQLLQPFL
       .::  . .: : : .. ..: . .   ::.  .  .   .  .. ...  : :.:. .v
NAHR_P EPTPYAAHLAEPVTSAMHALRNALQHHESFDPLTSERTFTLAMTDIGEIYFMPRLMDVLA
              70        80        90       100       110       120

       120       130       140       150       160       170
LYSR_E ARYPDVSLNIVPQESPLLEEWLSAQRHDLGLTETLHTPAGTERTELLSLDEVCVLPPGHP
       ^. :.  ...:...:  : . : .   ::...   . ..:  .  ::  . ::. . .::
NAHR_P HQAPNCVISTVRDSSMSLMQALQNGTVDLAVGLLPNLQTGFFQRRLLQNHYVCLCRKDHP
             130       140       150       160       170       180

       180       190       200       210       220        230
LYSR_E LAVKKVLTPDDFQGENYISLSRTDSYRQLLDQLFTEHQVKRRMIVE-THSAASVCAMVRA
       .. .. :: . : . .....  ... .  .:  .:   ..: . .: .: ::    . :.
NAHR_P VT-REPLTLERFCSYGHVRVIAAGTGHGEVDTYMTRVGIRRDIRLEVPHFAAVGHILQRT
              190       200       210       220       230

        240       250       260       270       280       290
LYSR_E GVGISVVNPLTALDYAASGLVVRRFSIAVP-FTVSLIRPLHRPSSALVQAFSGHLQAGLP
       ..  .:  .:..   .. :: . . ....: ........
NAHR_P DLLATVPIRLADCCVEPFGLSALPHPVVLPEIAINMFWHAKYHKDLANIWLRQLMFDLFT
     240       250       260       270       280       290

         300       310
LYSR_E KLVTSLDAILSSATTA

NAHR_P D
     300
```