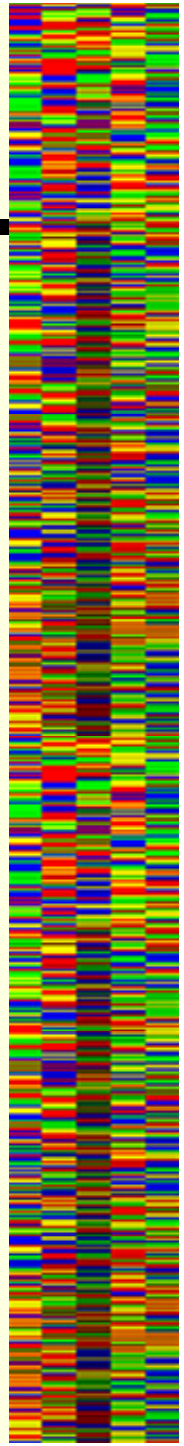
Genomics

3 – 5 September

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

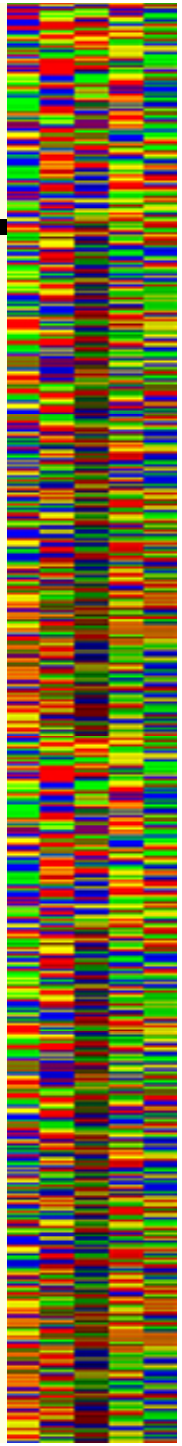
Reading: Mount ch 6 – Sequence database searching for similar sequences

Homework – online test on tree thinking



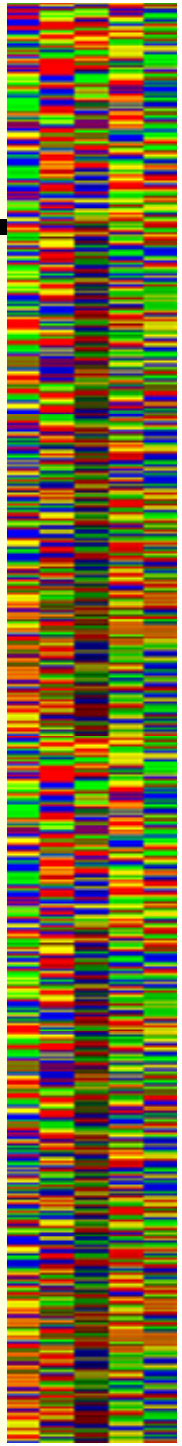
Sequence database searching - FASTA

- ***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

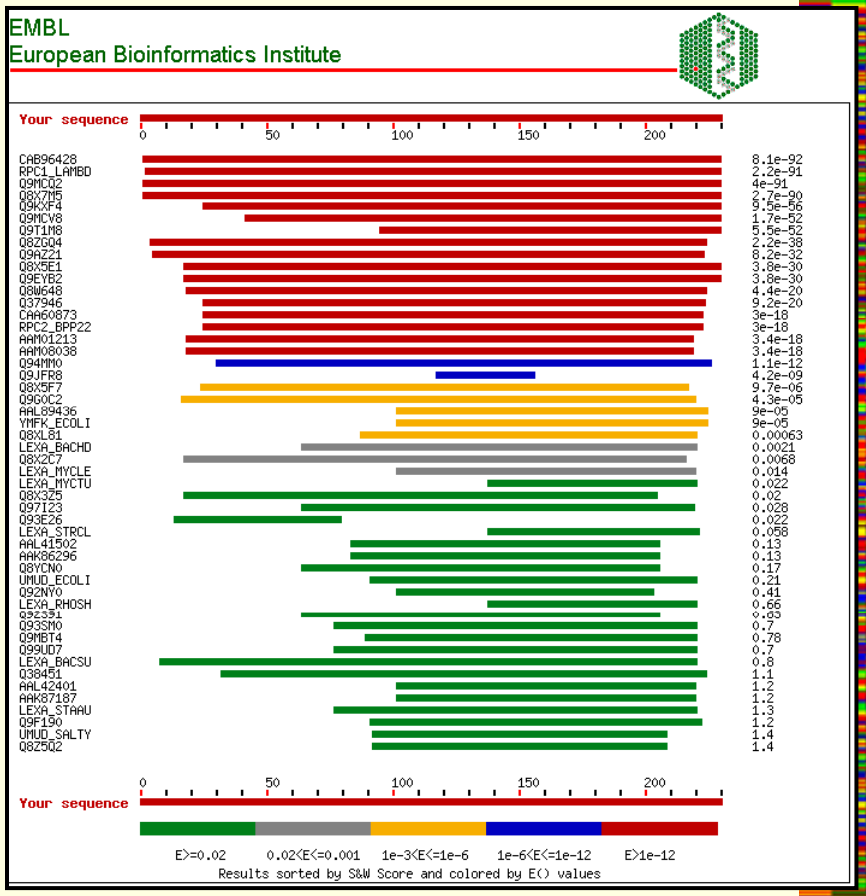
- ***Parameters***
  - ktup (word length)
    - 1 - 2 for proteins
    - 4 - 6 for nucleic acids
  - Scoring Matrix
    - Default is probably BLOSUM 50
  - Gap penalties



Genomics

Sequence database searching - FASTA

- ***Locations of hits on query***



Genomics

FASTA

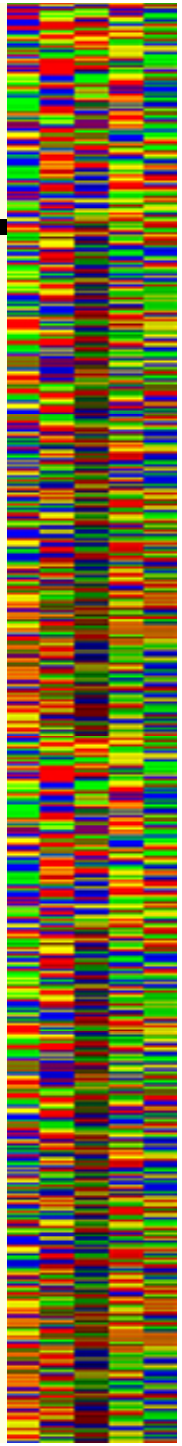
```
 opt      E()
< 20   876     0:==
  22     1     0:=          one = represents 483 library sequences
  24     4     0:=
  26    13     6:*
  28    61    68:*
  30   387   415:*
  32  1600  1603:===*
  34  4494  4348:=========*
  36  9575  8930:==================*=
  38 18356 14757:==============================*========
  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
```



Sequence database searching - FASTA

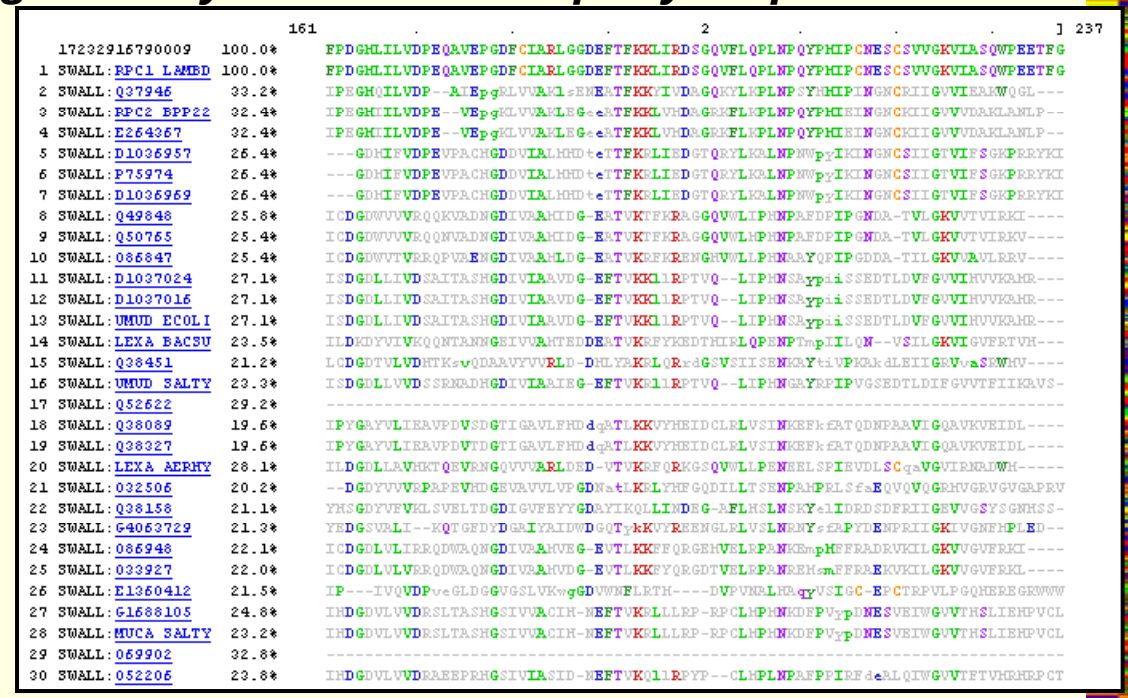
- *Optimized local alignment in region of best region*

```
>>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

Sequence database searching - FASTA

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

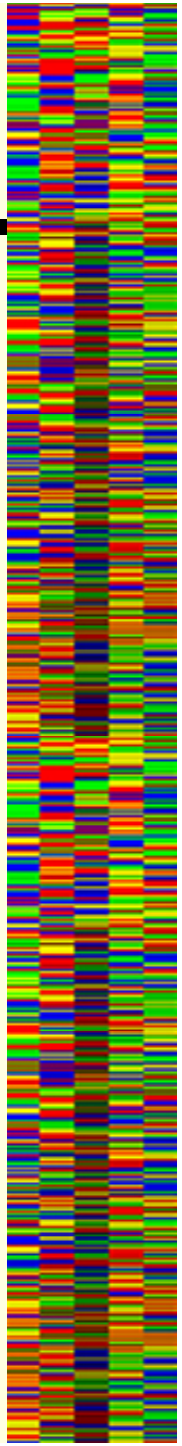


Sequence database searching - FASTA

```
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

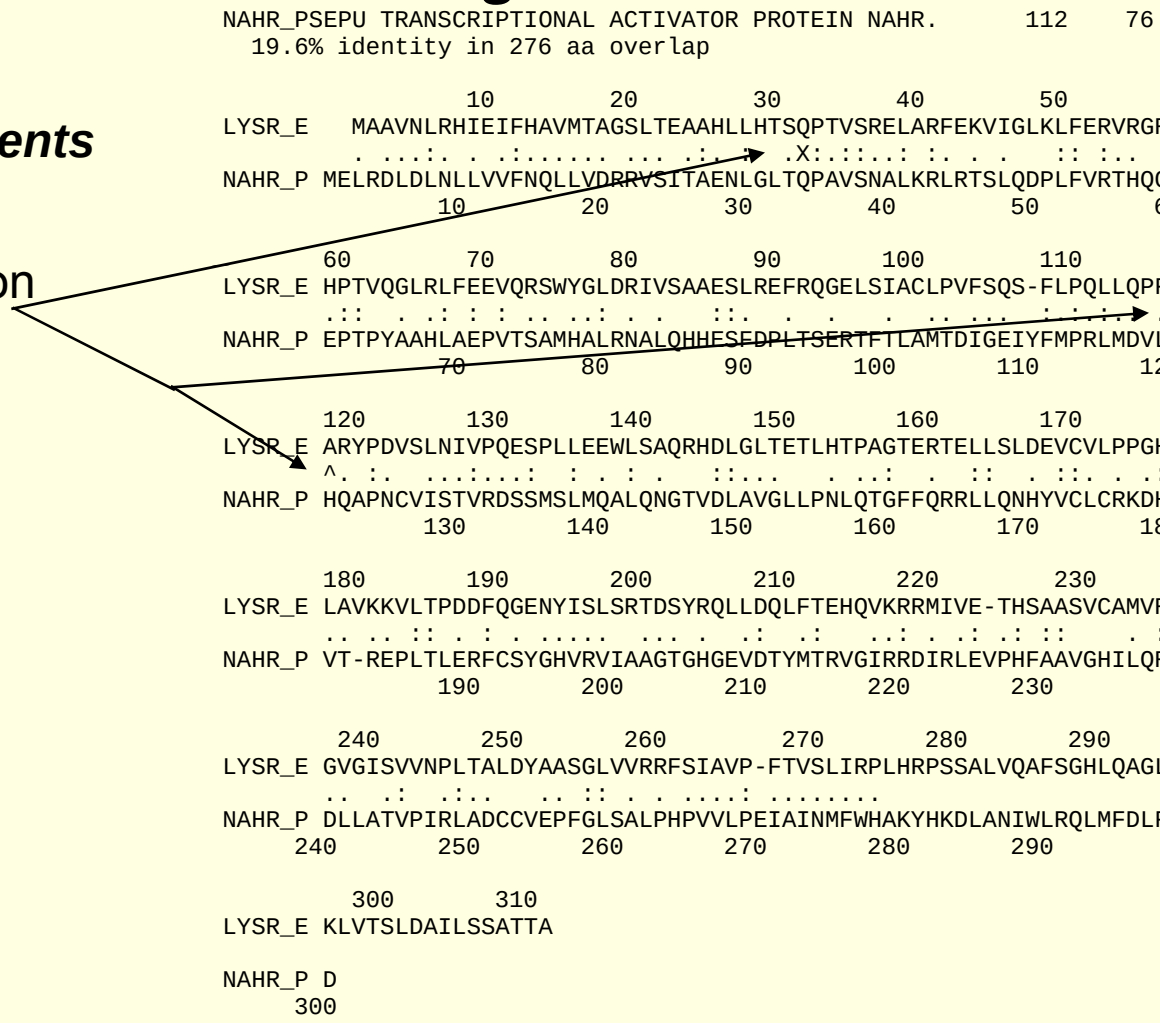
E<1 with Ktup = 2



Sequence database searching - FASTA

- ***Output Alignments***

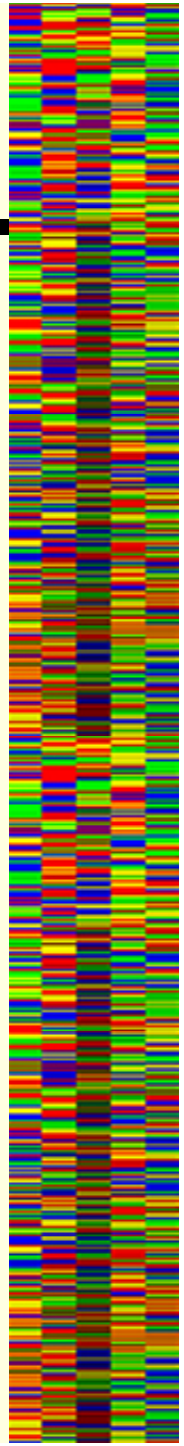
Ends of init1 region



Genomics

Sequence database searching - BLAST

- ***Similar in approach to FASTA, fast scan to remove most unrelated sequences followed by explicit alignments.***
- ***Steps in algorithm***
  - 1. Break down query sequence into overlapping words. For each word, determine a neighborhood of words that, if found in another sequence, would likely to be part of a significant MSP.
  - 2. Scan databases for neighborhood words.
  - 3. If two words are found on the same diagonal within a specified distance, try to extend the word matches into the complete MSP. Significant is easily calculated from Karlin-Altschul equation
  - 4. Perform local dynamic programming alignment around MSP regions



Sequence database searching - BLAST

- ***BLAST uses an explicit equation to predict, the probability of seeing a given score or higher. Assumptions***
  - unrelated sequences are random
  - sequences can be treated as infinite in length
- ***Important parameters***
  - Size of database being searched
  - significance threshold
  - scoring system

K, a constant, depends on scoring system and database composition

N, size of search space. query x database length depends on scoring system

Probability, *P*

$$P(S < x) = e^{-KNe^{-\lambda x}}$$

Score, *S*

greater than some value, *x*

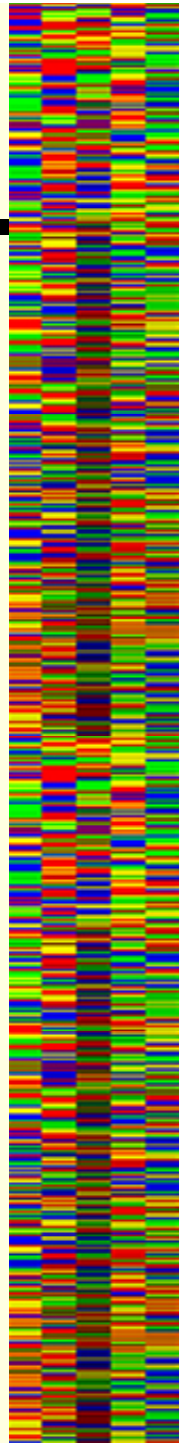
λ, another constant, depends on scoring system

$$\ln[\,P(S < x)] = -KNe^{-\lambda x}$$

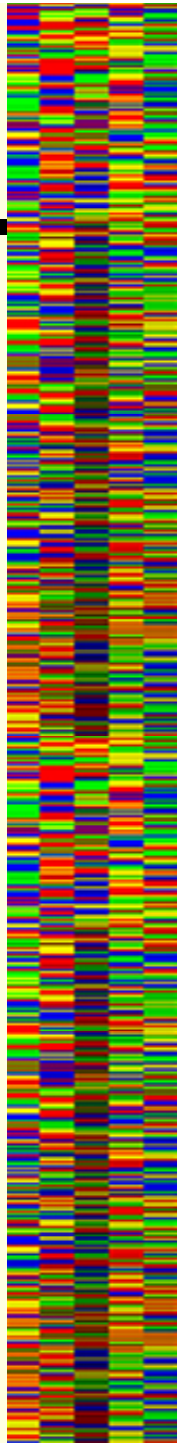
Sequence database searching – BLAST

- ***Step 1 – neighborhood words***

- ***First choose a desired level of significance, and the score, S, needed to achieve that level***
- ***Find short paired regions ("Words" of length w) which align with a similarity score, above a Threshold T***
- ***Choose w and T so that:***
  - maximize SPEED, i.e. minimize computer CPU (Central Processing Unit) time
  - maximize SENSITIVITY, i.e. find ALL "real" (homologs) hits
  - minimize False Negatives (TRUE hits which were NOT reported)
  - maximize SPECIFICITY, i.e. find ONLY "real" (homologs) hits
  - minimize False Positives (FALSE hits which WERE reported)



Genomics



Sequence database searching - BLAST

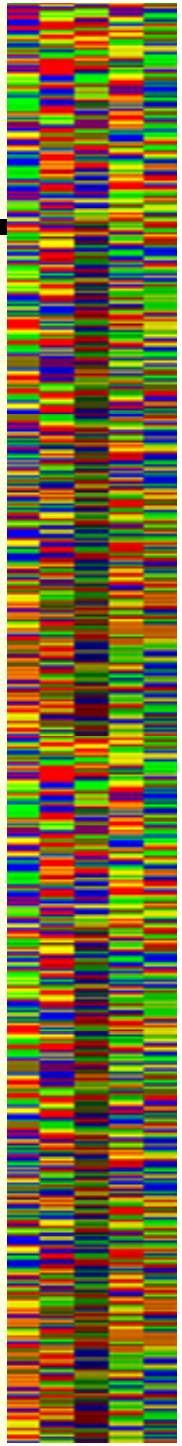
- ***As w (length of a word) increases for a given value of T:***
  - probability of a match decreases and cpu time decreases (Step 2 in BLAST algorithm)
  - get fewer incorrect hits or "false positives", *i.e.* increased SPECIFICITY
  - BUT also miss more correct hits or real homologs, *i.e.* decreased SENSITIVITY
  - ALSO ... neighborhood gets bigger, since Word Scores increase with Word size and hence more will have Scores greater than or equal to T ...
    (Word lists of 296, 3561, 40939 respectively for w = 3, 4, 5 for a 30 amino acid segment)

- ***As T decreases for a given w:***
  - increase the size of neighborhood exponentially:
    - Number of words is proportional to e^{-T}
  - increases the cpu time: cpu time is proportional to Number of Words => cpu time is proportional to e^{-T}
  - get more real hits, *i.e.* increased SENSITIVITY
  - BUT ... also get more incorrect hits, *i.e*. decreased SPECIFICITY
- ***(w,T) = (3,14), (4,16), (5,18) give nearly equivalent SENSITIVITY***

Sequence database searching - BLAST

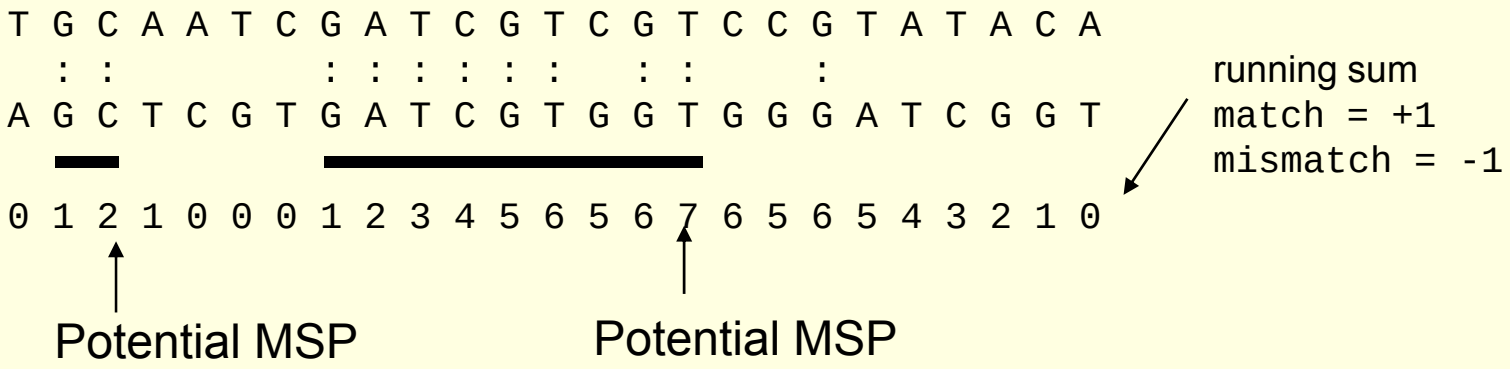
- ***Step 2 – compare neighborhood words to database***

- ***Neighborhood is about 50 times the size of the original protein***
- ***search uses similar idea to lookup table in FASTA, but a more efficient approach using a discrete finite automaton***
- ***BLAST keeps database in memory for even greater speed***
  - multiple queries faster than many individual queries where running on your local computer



Sequence database searching - BLAST

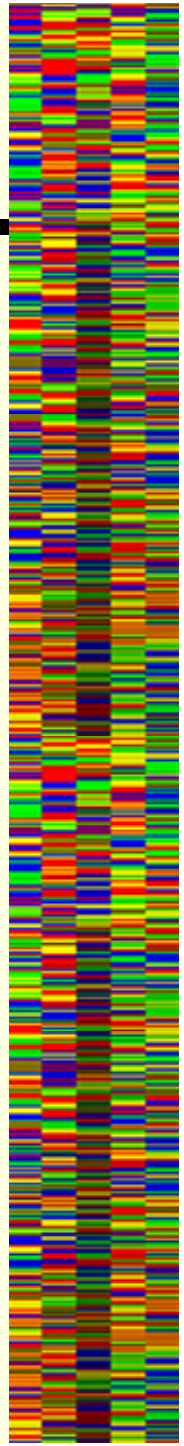
- ***Based on Maximal Segment Pairs (MSP)***
  - Highest scoring pair of identical length segments from two sequences
  - Local alignment without gaps, similar to FASTA local region
  - Expected distribution is known!
- ***Maximal Segment Pair sample calculation***



Sequence database searching - BLAST

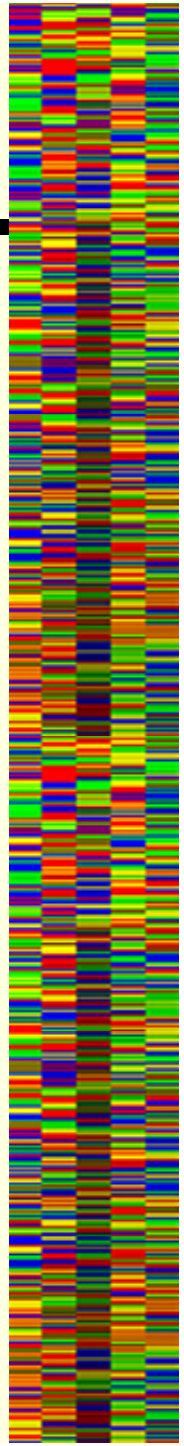
- *Step 3 – Extend word matches*

- *Most expensive step in BLAST algorithm*
- *A diagonal must have at least two matching words to be extended. Extend to end of high scoring segment pair, or HSP. HSPs approximate maximal segment pairs or MSPs. They are only approximate because extension does not continue until running score reaches zero.*
- *Earlier version of BLAST required only one hit. T can be higher, but more extensions and hence slower searches result.*



Sequence database searching - BLAST

- *Step 4 – Local dynamic programming alignment in region of HSPs*
- *Same alignment method, Smith-Waterman algorithm (local dynamic programming), but limited to region that includes the HSP*
- *Significance of alignment with gaps can be evaluated using K and l estimated from alignments of random sequences with same gap penalty and scoring parameters*
- *Ins spite of claims of being "mathematically rigorous" these parameters can only be empirically estimated*

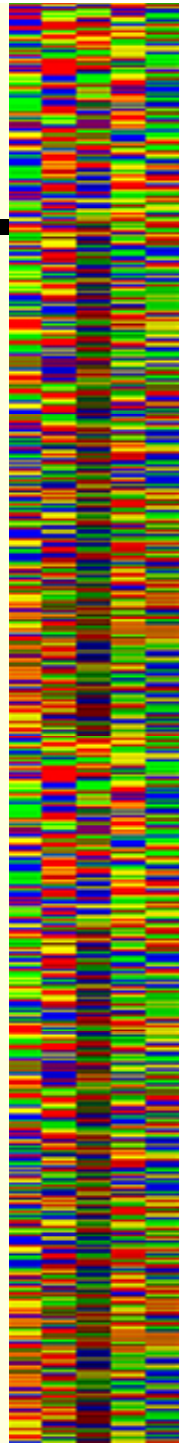


Genomics

Sequence database searching - BLAST

- *Evaluation of significance*

- *Significance is calculated versus theoretic distribution using Karlin-Altschul equation not real sequences.*
- *Assumes sequences are random*
- *Assume database is one long sequence – length effects are not corrected for*
- *Statistics are very inaccurate for short queries (ca. 20 characters). Scores are still ranked correctly, but calculated E-values are far to large. Short queries will typically produce no results – solution is to increase E parameter.*



Genomics

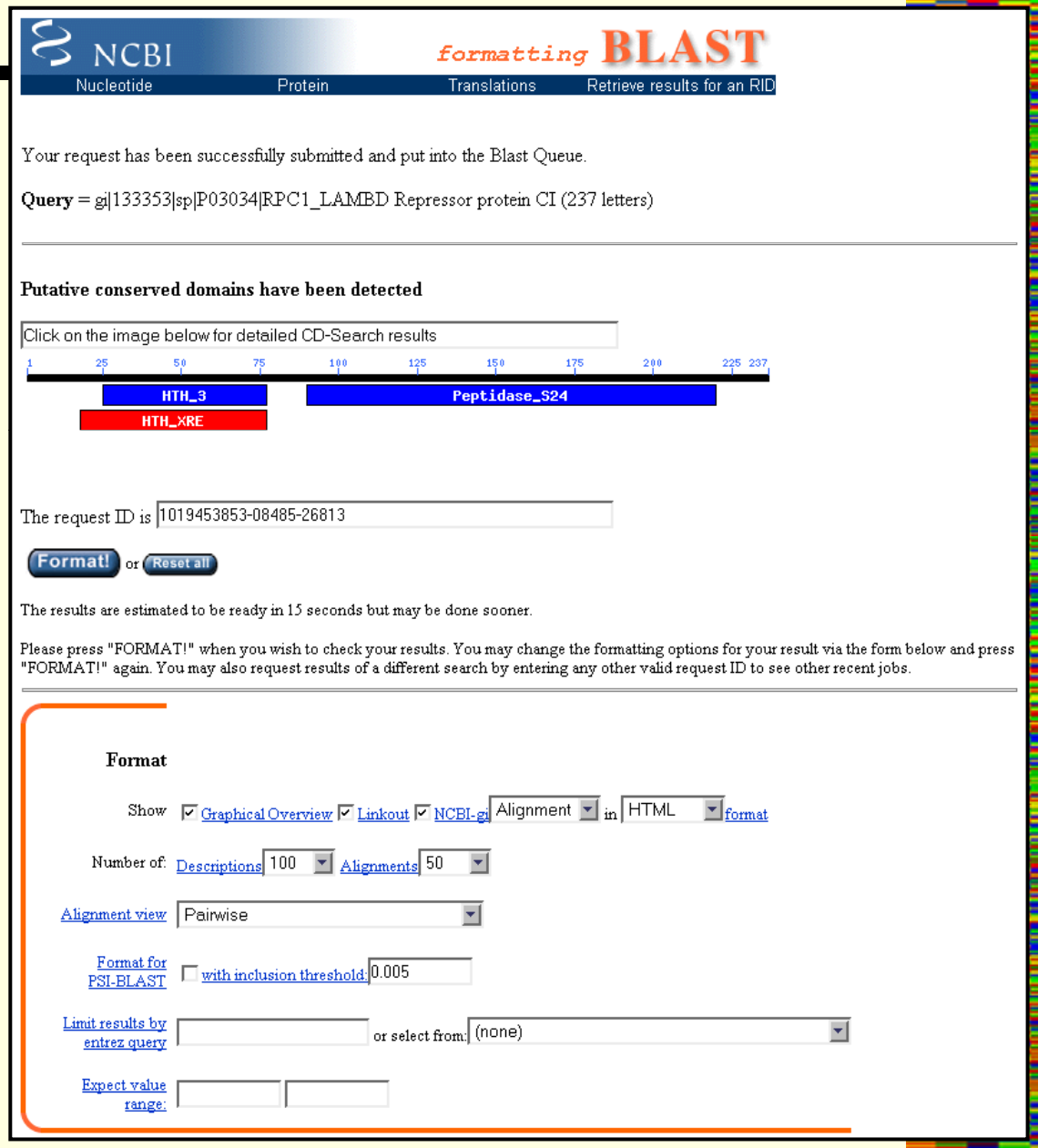
Sequence database searching - BLAST

- www.ncbi.nlm.nih.gov/blast/

Genomics

Sequence database searching - BLAST

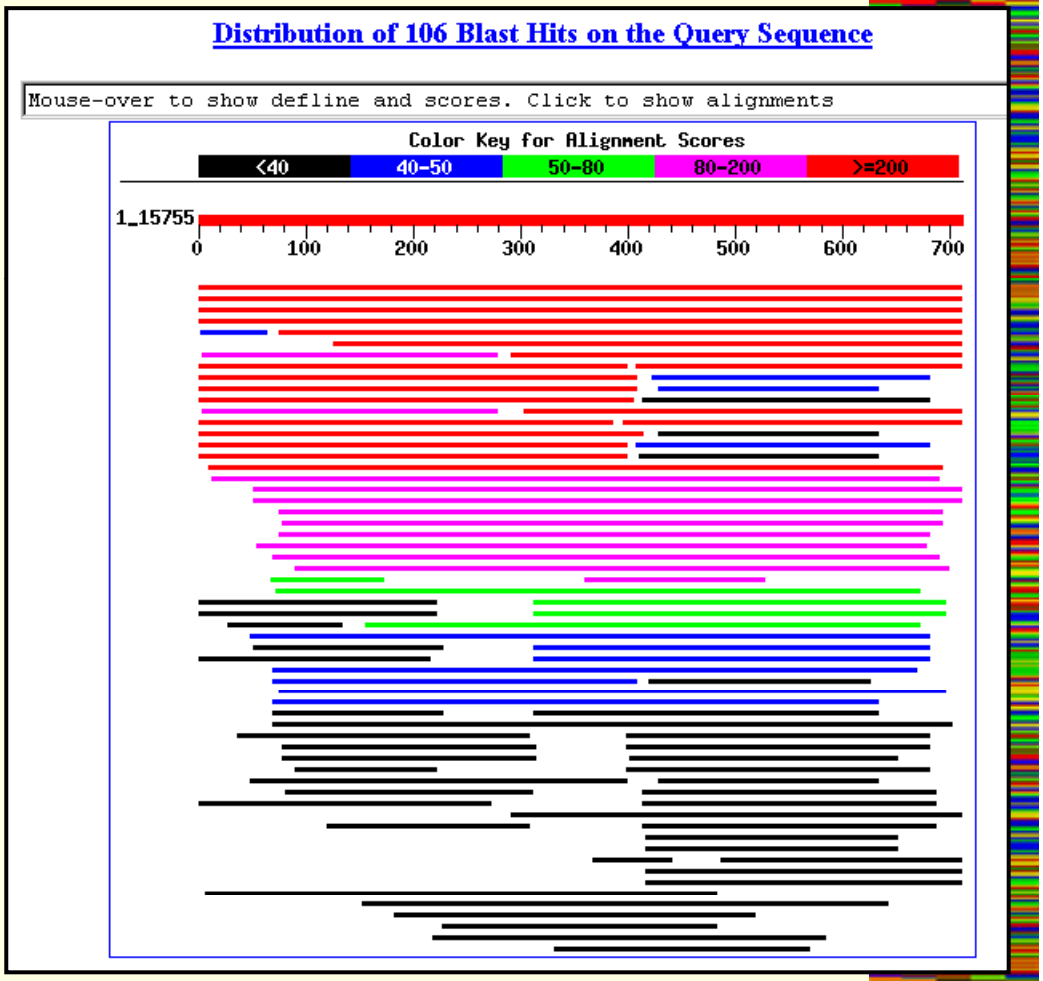
- *NCBI BLAST interface, intermediate result after submitting query*



Genomics

Sequence database searching - BLAST

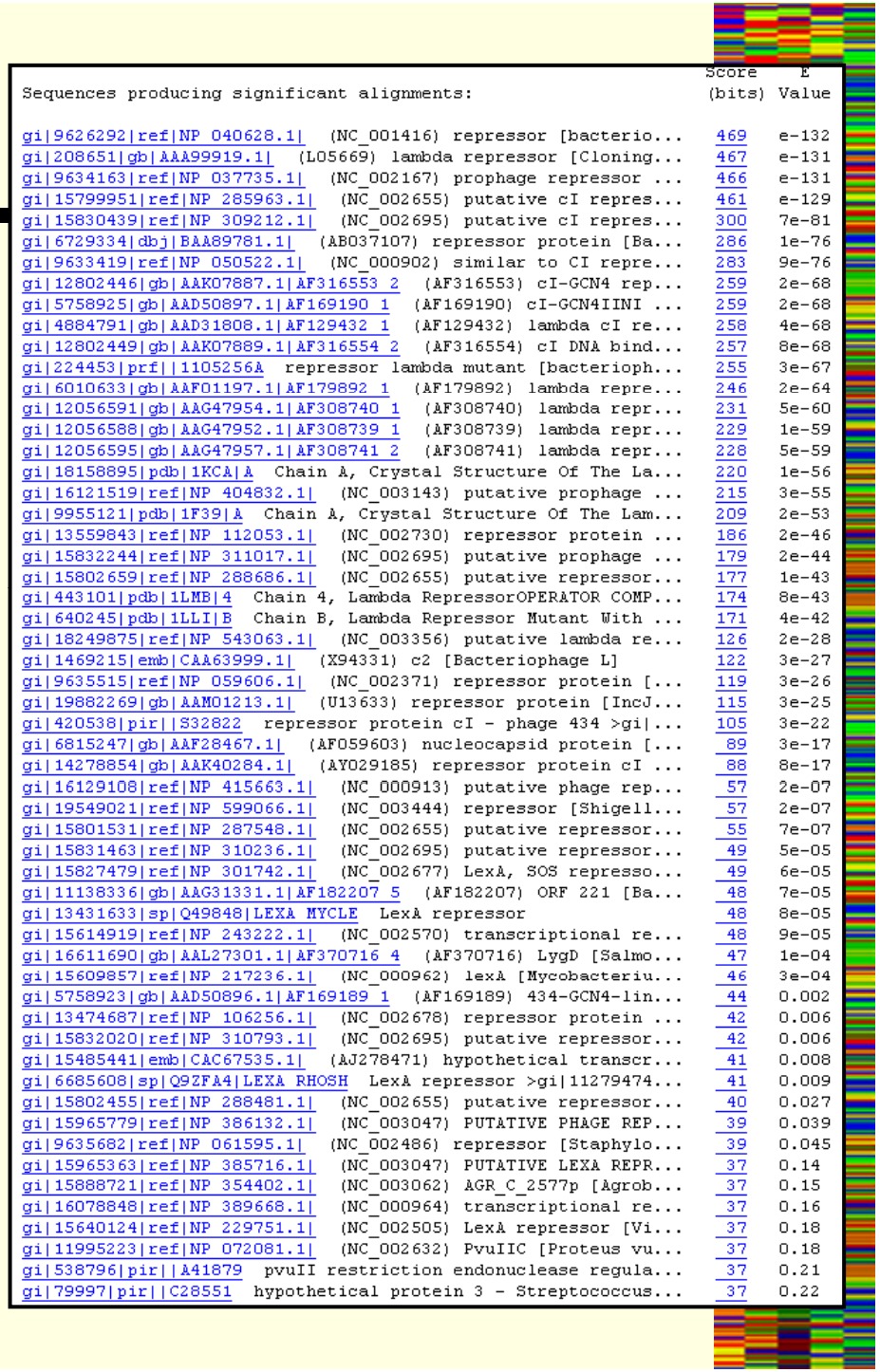
- *locations of matches.*
- *multiple segments on one line are from different sequences unless connected by shaded bar.*



Genomics

Sequence database searching - BLAST

- *List of top scores*



Sequences producing significant alignments:	Score (bits)	E Value
gi|9626292|ref|NP_040628.1| (NC_001416) repressor [bacterio...	469	e-132
gi|208651|gb|AAA99919.1| (L05669) lambda repressor [Cloning...	467	e-131
gi|9634163|ref|NP_037735.1| (NC_002167) prophage repressor ...	466	e-131
gi|15799951|ref|NP_285963.1| (NC_002655) putative cI repres...	461	e-129
gi|15830439|ref|NP_309212.1| (NC_002695) putative cI repres...	300	7e-81
gi|6729334|dbj|BAA89781.1| (AB037107) repressor protein [Ba...	286	1e-76
gi|9633419|ref|NP_050522.1| (NC_000902) similar to CI repre...	283	9e-76
gi|12802446|gb|AAK07887.1|AF316553_2 (AF316553) cI-GCN4 rep...	259	2e-68
gi|5758925|gb|AAD50897.1|AF169190_1 (AF169190) cI-GCN4IINI ...	259	2e-68
gi|4884791|gb|AAD31808.1|AF129432_1 (AF129432) lambda cI re...	258	4e-68
gi|12802449|gb|AAK07889.1|AF316554_2 (AF316554) cI DNA bind...	257	8e-68
gi|224453|prf||1105256A repressor lambda mutant [bacterioph...	255	3e-67
gi|6010633|gb|AAF01197.1|AF179892_1 (AF179892) lambda repre...	246	2e-64
gi|12056591|gb|AAG47954.1|AF308740_1 (AF308740) lambda repr...	231	5e-60
gi|12056588|gb|AAG47952.1|AF308739_1 (AF308739) lambda repr...	229	1e-59
gi|12056595|gb|AAG47957.1|AF308741_2 (AF308741) lambda repr...	228	5e-59
gi|18158895|pdb|1KCA|A Chain A, Crystal Structure Of The La...	220	1e-56
gi|16121519|ref|NP_404832.1| (NC_003143) putative prophage ...	215	3e-55
gi|9955121|pdb|1F39|A Chain A, Crystal Structure Of The Lam...	209	2e-53
gi|13559843|ref|NP_112053.1| (NC_002730) repressor protein ...	186	2e-46
gi|15832244|ref|NP_311017.1| (NC_002695) putative prophage ...	179	2e-44
gi|15802659|ref|NP_288686.1| (NC_002655) putative repressor...	177	1e-43
gi|443101|pdb|1LMB|4 Chain 4, Lambda RepressorOPERATOR COMP...	174	8e-43
gi|640245|pdb|1LLI|B Chain B, Lambda Repressor Mutant With ...	171	4e-42
gi|18249875|ref|NP_543063.1| (NC_003356) putative lambda re...	126	2e-28
gi|1469215|emb|CAA63999.1| (X94331) c2 [Bacteriophage L]	122	3e-27
gi|9635515|ref|NP_059606.1| (NC_002371) repressor protein [...	119	3e-26
gi|19882269|gb|AAM01213.1| (U13633) repressor protein [IncJ...	115	3e-25
gi|420538|pir||S32822 repressor protein cI - phage 434 >gi|...	105	3e-22
gi|6815247|gb|AAF28467.1| (AF059603) nucleocapsid protein [...	89	3e-17
gi|14278854|gb|AAK40284.1| (AY029185) repressor protein cI ...	88	8e-17
gi|16129108|ref|NP_415663.1| (NC_000913) putative phage rep...	57	2e-07
gi|19549021|ref|NP_599066.1| (NC_003444) repressor [Shigell...	57	2e-07
gi|15801531|ref|NP_287548.1| (NC_002655) putative repressor...	55	7e-07
gi|15831463|ref|NP_310236.1| (NC_002695) putative repressor...	49	5e-05
gi|15827479|ref|NP_301742.1| (NC_002677) LexA, SOS represso...	49	6e-05
gi|11138336|gb|AAG31331.1|AF182207_5 (AF182207) ORF 221 [Ba...	48	7e-05
gi|13431633|sp|Q49848|LEXA_MYCLE LexA repressor	48	8e-05
gi|15614919|ref|NP_243222.1| (NC_002570) transcriptional re...	48	9e-05
gi|16611690|gb|AAL27301.1|AF370716_4 (AF370716) LygD [Salmo...	47	1e-04
gi|15609857|ref|NP_217236.1| (NC_000962) lexA [Mycobacteriu...	46	3e-04
gi|5758923|gb|AAD50896.1|AF169189_1 (AF169189) 434-GCN4-lin...	44	0.002
gi|13474687|ref|NP_106256.1| (NC_002678) repressor protein ...	42	0.006
gi|15832020|ref|NP_310793.1| (NC_002695) putative repressor...	42	0.006
gi|15485441|emb|CAC67535.1| (AJ278471) hypothetical transcr...	41	0.008
gi|6685608|sp|Q9ZFA4|LEXA_RHOSH LexA repressor >gi|11279474...	41	0.009
gi|15802455|ref|NP_288481.1| (NC_002655) putative repressor...	40	0.027
gi|15965779|ref|NP_386132.1| (NC_003047) PUTATIVE PHAGE REP...	39	0.039
gi|9635682|ref|NP_061595.1| (NC_002486) repressor [Staphylo...	39	0.045
gi|15965363|ref|NP_385716.1| (NC_003047) PUTATIVE LEXA REPR...	37	0.14
gi|15888721|ref|NP_354402.1| (NC_003062) AGR_C_2577p [Agrob...	37	0.15
gi|16078848|ref|NP_389668.1| (NC_000964) transcriptional re...	37	0.16
gi|15640124|ref|NP_229751.1| (NC_002505) LexA repressor [Vi...	37	0.18
gi|11995223|ref|NP_072081.1| (NC_002632) PvuIIC [Proteus vu...	37	0.18
gi|538796|pir||A41879 pvuII restriction endonuclease regula...	37	0.21
gi|79997|pir||C28551 hypothetical protein 3 - Streptococcus...	37	0.22

Genomics

Sequence database searching - BLAST

- *Alignments*
- *Middle line shows matches*

```
>gi|15830439|ref|NP_309212.1|  (NC_002695) putative cI repressor protein [Escherichia coli
          O157:H7]
 gi|7649844|dbj|BAA94122.1|  (AP000422) CI protein [Escherichia coli O157:H7]
 gi|13360645|dbj|BAB34608.1|  (AP002554) putative cI repressor protein [Escherichia coli O157:H7]
          Length = 217

 Score =  300 bits (769), Expect = 7e-81
 Identities = 150/212 (70%), Positives = 168/212 (78%), Gaps = 4/212 (1%)

Query: 26  KKNELGLSQESVADKMGMGQSGVGALFNGINALNAYNAALLAKILKVSVEEFSPSIAREI 85
           +  ELG++QE +A+++GM Q G+G    G    +  +  ++ K L +    F+
Sbjct: 10  RMKELGITQEKLAEELGMTQGGIGHWLRGSRHPSLSDIGVVFKYLGIDNISFNHDGTFSP 69

Query: 86  YEMYEAVSMQPSLRSEYEYPVFSHVQAGMFSPELRTFTKGDAERWVSTTKKASDSAFWLE 145
              Y +      ++ +YEYPVFSHVQAGMFSPELRTFTKGDAER VSTTKKASDSAFWLE
Sbjct: 70  VGEYSSA----PVKKQYEYPVFSHVQAGMFSPELRTFTKGDAERLVSTTKKASDSAFWLE 125

Query: 146 VEGNSMTAPTGSKPSFPDGMLILVDPEQAVEPGDFCIARLGGDEFTFKKLIRDSGQVFLQ 205
           VEGNSMTAPTGSKPSFPDGMLILVDPEQAVEPGDFCIARLGGDEFTFKKLIRDSGQVFLQ
Sbjct: 126 VEGNSMTAPTGSKPSFPDGMLILVDPEQAVEPGDFCIARLGGDEFTFKKLIRDSGQVFLQ 185

Query: 206 PLNPQYPMIPCNESCSVVGKVIASQWPEETFG 237
           PLNPQYPMIPCNESCSVVGKVIA QWPEETFG
Sbjct: 186 PLNPQYPMIPCNESCSVVGKVIAGQWPEETFG 217




>gi|6729334|dbj|BAA89781.1|  (AB037107) repressor protein [Bacteriophage VT2-Sa]
          Length = 191

 Score =  286 bits (732), Expect = 1e-76
 Identities = 144/195 (73%), Positives = 155/195 (78%), Gaps = 4/195 (2%)

Query: 43  MGQSGVGALFNGINALNAYNAALLAKILKVSVEEFSPSIAREIYEMYEAVSMQPSLRSEY 102
           M Q G+G    G    +  +  ++ K L +    F+          Y +      ++ +Y
Sbjct: 1   MTQGGIGHWLRGSRHPSLSDIGVVFKYLGIDNISFNHDGTFSPVGEYSSA----PVKKQY 56

Query: 103 EYPVFSHVQAGMFSPELRTFTKGDAERWVSTTKKASDSAFWLEVEGNSMTAPTGSKPSFP 162
           EYPVFSHVQAGMFSPELRTFTKGDAER VSTTKKASDSAFWLEVEGNSMTAPTGSKPSFP
Sbjct: 57  EYPVFSHVQAGMFSPELRTFTKGDAERLVSTTKKASDSAFWLEVEGNSMTAPTGSKPSFP 116

Query: 163 DGMLILVDPEQAVEPGDFCIARLGGDEFTFKKLIRDSGQVFLQPLNPQYPMIPCNESCSV 222
           DGMLILVDPEQAVEPGDFCIARLGGDEFTFKKLIRDSGQVFLQPLNPQYPMIPCNESCSV
Sbjct: 117 DGMLILVDPEQAVEPGDFCIARLGGDEFTFKKLIRDSGQVFLQPLNPQYPMIPCNESCSV 176

Query: 223 VGKVIASQWPEETFG 237
           VGKVIASQWPEETFG
Sbjct: 177 VGKVIASQWPEETFG 191
```

Sequence database searching - BLAST

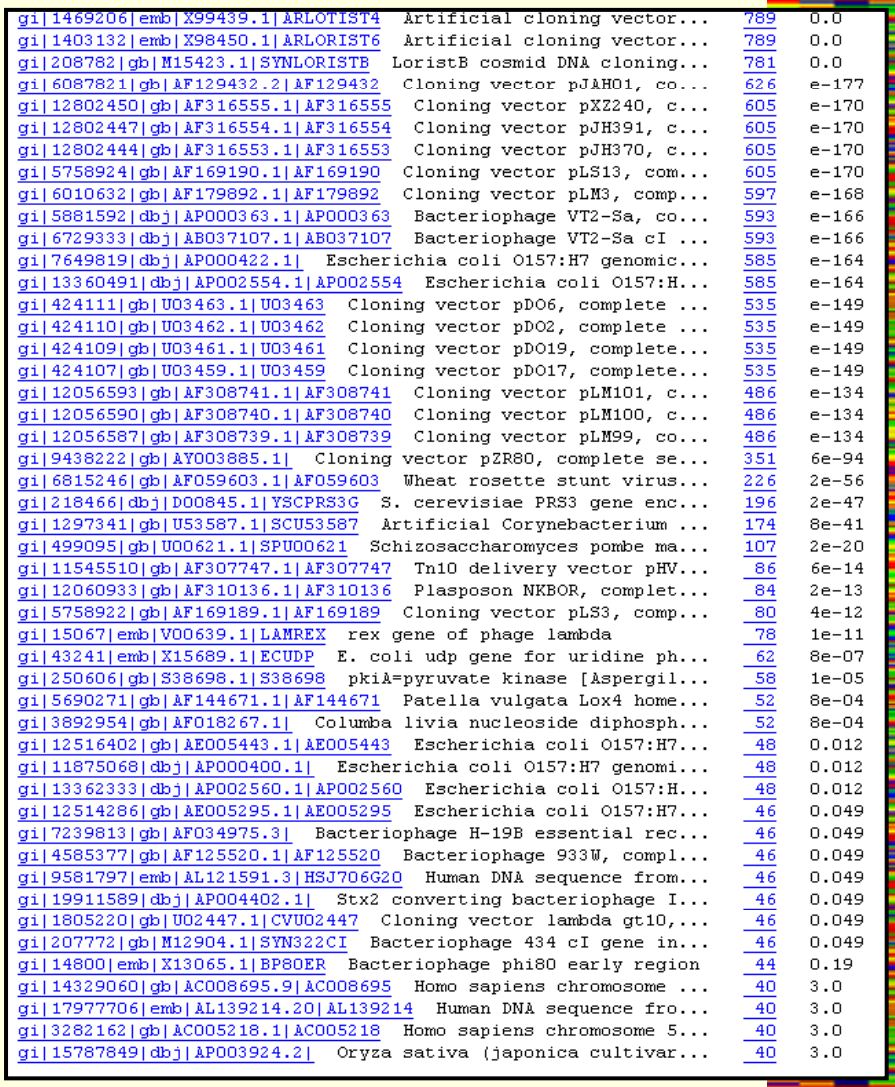
- ***Detailed information on parameters from end of results***

```
  Database: All non-redundant GenBank CDS
  translations+PDB+SwissProt+PIR+PRF
    Posted date:  Apr 17, 2002 10:49 AM
  Number of letters in database: 289,186,171
  Number of sequences in database:  922,227

Lambda     K      H
   0.314    0.132    0.379

Gapped
Lambda     K      H
   0.267   0.0410    0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 138,050,983
Number of Sequences: 922227
Number of extensions: 5863505
Number of successful extensions: 10977
Number of sequences better than 10.0: 111
Number of HSP's better than 10.0 without gapping: 59
Number of HSP's successfully gapped in prelim test: 52
Number of HSP's that attempted gapping in prelim test: 10885
Number of HSP's gapped (non-prelim): 111
length of query: 237
length of database: 289,186,171
effective HSP length: 117
effective length of query: 120
effective length of database: 181,285,612
effective search space: 21754273440
effective search space used: 21754273440
T: 11
A: 40
X1: 16 ( 7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 42 (22.0 bits)
S2: 69 (31.2 bits)
```

Genomics

Sequence database searching - BLAST

- *Nucleotide comparison*
- *Note that there are fewer matches than in the protein/protein search*



gi|1469206|emb|X99439.1|ARLOTIST4	Artificial cloning vector...	789	0.0
gi|1403132|emb|X98450.1|ARLORIST6	Artificial cloning vector...	789	0.0
gi|208782|gb|M15423.1|SYNLORISTB	LoristB cosmid DNA cloning...	781	0.0
gi|6087821|gb|AF129432.2|AF129432	Cloning vector pJAH01, co...	626	e-177
gi|12802450|gb|AF316555.1|AF316555	Cloning vector pXZ240, c...	605	e-170
gi|12802447|gb|AF316554.1|AF316554	Cloning vector pJH391, c...	605	e-170
gi|12802444|gb|AF316553.1|AF316553	Cloning vector pJH370, c...	605	e-170
gi|5758924|gb|AF169190.1|AF169190	Cloning vector pLS13, com...	605	e-170
gi|6010632|gb|AF179892.1|AF179892	Cloning vector pLM3, comp...	597	e-168
gi|5881592|dbj|AP000363.1|AP000363	Bacteriophage VT2-Sa, co...	593	e-166
gi|6729333|dbj|AB037107.1|AB037107	Bacteriophage VT2-Sa cI ...	593	e-166
gi|7649819|dbj|AP000422.1|	Escherichia coli O157:H7 genomic...	585	e-164
gi|13360491|dbj|AP002554.1|AP002554	Escherichia coli O157:H...	585	e-164
gi|424111|gb|U03463.1|U03463	Cloning vector pDO6, complete ...	535	e-149
gi|424110|gb|U03462.1|U03462	Cloning vector pDO2, complete ...	535	e-149
gi|424109|gb|U03461.1|U03461	Cloning vector pDO19, complete...	535	e-149
gi|424107|gb|U03459.1|U03459	Cloning vector pDO17, complete...	535	e-149
gi|12056593|gb|AF308741.1|AF308741	Cloning vector pLM101, c...	486	e-134
gi|12056590|gb|AF308740.1|AF308740	Cloning vector pLM100, c...	486	e-134
gi|12056587|gb|AF308739.1|AF308739	Cloning vector pLM99, co...	486	e-134
gi|9438222|gb|AY003885.1|	Cloning vector pZR80, complete se...	351	6e-94
gi|6815246|gb|AF059603.1|AF059603	Wheat rosette stunt virus...	226	2e-56
gi|218466|dbj|D00845.1|YSCPRS3G	S. cerevisiae PRS3 gene enc...	196	2e-47
gi|1297341|gb|U53587.1|SCU53587	Artificial Corynebacterium ...	174	8e-41
gi|499095|gb|U00621.1|SPU00621	Schizosaccharomyces pombe ma...	107	2e-20
gi|11545510|gb|AF307747.1|AF307747	Tn10 delivery vector pHV...	86	6e-14
gi|12060933|gb|AF310136.1|AF310136	Plasposon NKBOR, complet...	84	2e-13
gi|5758922|gb|AF169189.1|AF169189	Cloning vector pLS3, comp...	80	4e-12
gi|15067|emb|V00639.1|LAMREX	rex gene of phage lambda	78	1e-11
gi|43241|emb|X15689.1|ECUDP	E. coli udp gene for uridine ph...	62	8e-07
gi|250606|gb|S38698.1|S38698	pkiA=pyruvate kinase [Aspergil...	58	1e-05
gi|5690271|gb|AF144671.1|AF144671	Patella vulgata Lox4 home...	52	8e-04
gi|3892954|gb|AF018267.1|	Columba livia nucleoside diphosph...	52	8e-04
gi|12516402|gb|AE005443.1|AE005443	Escherichia coli O157:H7...	48	0.012
gi|11875068|dbj|AP000400.1|	Escherichia coli O157:H7 genomi...	48	0.012
gi|13362333|dbj|AP002560.1|AP002560	Escherichia coli O157:H...	48	0.012
gi|12514286|gb|AE005295.1|AE005295	Escherichia coli O157:H7...	46	0.049
gi|7239813|gb|AF034975.3|	Bacteriophage H-19B essential rec...	46	0.049
gi|4585377|gb|AF125520.1|AF125520	Bacteriophage 933W, compl...	46	0.049
gi|9581797|emb|AL121591.3|HSJ706G20	Human DNA sequence from...	46	0.049
gi|19911589|dbj|AP004402.1|	Stx2 converting bacteriophage I...	46	0.049
gi|1805220|gb|U02447.1|CVU02447	Cloning vector lambda gt10,...	46	0.049
gi|207772|gb|M12904.1|SYN322CI	Bacteriophage 434 cI gene in...	46	0.049
gi|14800|emb|X13065.1|BP80ER	Bacteriophage phi80 early region	44	0.19
gi|14329060|gb|AC008695.9|AC008695	Homo sapiens chromosome ...	40	3.0
gi|17977706|emb|AL139214.20|AL139214	Human DNA sequence fro...	40	3.0
gi|3282162|gb|AC005218.1|AC005218	Homo sapiens chromosome 5...	40	3.0
gi|15787849|dbj|AP003924.2|	Oryza sativa (japonica cultivar...	40	3.0

Genomics

Sequence database searching - BLAST

blastp

- compares an amino acid query sequence against a protein sequence database

blastn

- compares a nucleotide query sequence against a nucleotide sequence database

blastx

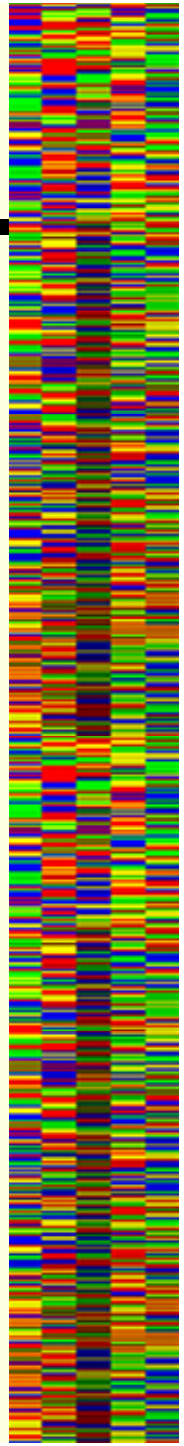
- compares a nucleotide query sequence translated in all reading frames against a protein sequence database, e.g., unknown genomic sequence

tblastn

- compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames, e.g., trying to find exons in genomic DNA

tblastx

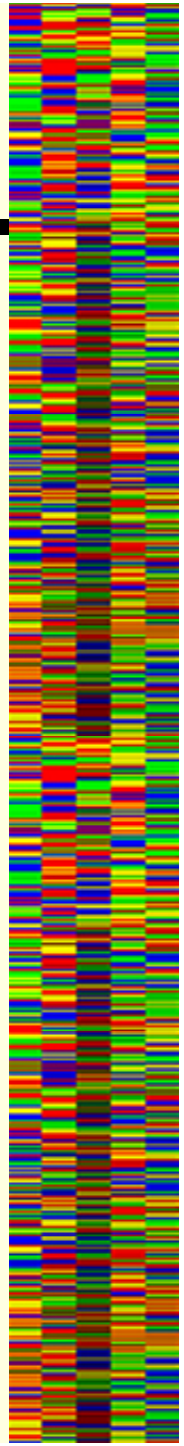
- compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database, e.g. two unknown genes where you don't know the exons



Genomics

Sequence database search - Filtering

- ***Purpose – remove parts of query that are likely to obscure correct results***
  - repeats
  - low-entropy sequences – regions of low compositional complexity – regions that only have one or two sequence characters
- ***BLASTN uses dust program***
  - low complexity sequences (similar to seg)
  - SINES,LINES, known repeated sequences, vector sequences?
- ***BLASTP uses seg program***
  - Altschul, S. F., M. S. Boguski, W. Gish, J. C. Wootton (1994). Issues in searching molecular sequence databases. Nat Genet 6: 119-129.
  - Wootton, J. C. and S. Federhen (1993). Statistics of local complexity in amino acid sequences and sequence databases. Computers in Chemistry 17:149-163.
  - Wootton, J. C. and S. Federhen (1996). Analysis of compositionally biased regions in sequence databases. Methods in Enzymology 266: 554-571.
- ***Filtered regions are marked with Xs in output, and are not included in search or alignment so a perfect match may not have 100% identity***
  - filtered regions don't count as matches!



Sequence database search - Filtering

- Calculate compositional complexity, K

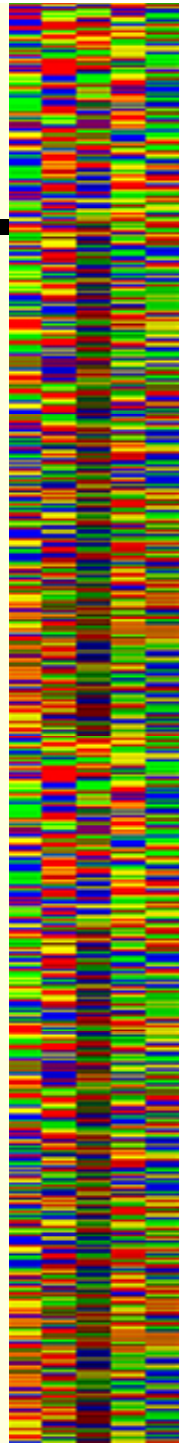
$$K = 1 / L \log_N(L!/\Pi n_i!)$$

where L is the window length and N is the alphabet size (4 for DNA, 20 for protein) and n_i is the count of each kind of character

- For GAGAGAGA, $n_G=4$ $n_A=4$ $n_T=n_C=0$

$K = 1/ 8 \log_4(8!/4!4!0!0!) = 1/8 \log_4(40320 / 576) = 1 / 8 \log_4 (70) = 1 / 8 x3.06$
$K = 0.041$

- Note that only composition, not order is important
- More examples in book



Genomics

Sequence database search - Filtering

- ***Seg filtering of human MTG8***
- ***putative transcription factor in** acute myeloid leukemia*

Human MGT8a protein

Low-complexity segments		High-complexity segments
	1-24	MPDRTEKHSTMPDSPVDVKTQSRL
tpptmpppptt	25-35	
	36-60	QGAPRTSSFTPTTLTNGTSHSPTAL
ngapsppngfsngpsssssssslanqqlpp	61-89	
	90-258	ACGARQLSKLKRFLTTLQQFGNDISPEIGE RVRTLVLGLVNSTLTIEEFHSKLQEATNFP LRPFVIPFLKANLPLLQRELLHCARLAKQN PAQYLAQHEQLLLDASTTSPVDSSELLLDV NENGKRRTPDRTKENGFDREPLHSEHPSKR PCTISPGQRYSPNNGLSYQ
pnglphptpppp	259-270	
	271-377	QHYRLDDMAIAHHYRDSYRHPSHRDLRDRN RPMGLHGTRQEEMIDHRLTDREWAEEWKHL DHLLNCIMDMVEKTRRSLTVLRRCQEADRE ELNYWIRRYSDAEDLKK
gggsssshs	378-386	
	387-554	RQQSPVNPDPVALDAHREFLHRPASGYVPE EIWKKAEEAVNEVKRQAMTELQKAVSEAER KAHDMITTERAKMERTVAEAKRQAAEDALA VINQQEDSSESCWNCGRKASETCSGCNTAR YCGSFCQHKDWEKHHHICGQTLQAQQQGDT PAVSSSVTPNSGAGSPMD
tppaatprsttpgtpstiettp	555-576	
	577-577	R