CISC 636 Computational Biology & Bioinformatics (Fall 2016)

Sequence pairwise alignment

Score statistics: E-value and p-value

Heuristic algorithms: BLAST and FASTA

Database search: gene finding and annotations

Significance of scores

Goals for sequence alignments:

(1) whether and

(2) how two sequences are related.

It is rare that you have just two particular sequences to compare. More often, you have one query sequence and a large database of sequences.

Database searching: find all sequences in the database that are related to the query sequence.

Solution:

(1) For each sequence in the database, use Smith-Waterman/FASTA/BLAST to align with the query sequence and return the score of the optimal alignment.

(2) Rank the sequences by the score.

Q: how good is a score?

Score statistics [Karlin & Altschul 1990]

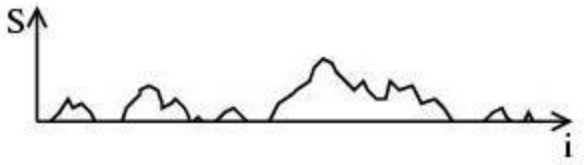
-The score of an ungapped alignment is

$$H_{i,j} = \max\{H_{i-1,j-1} + s(x_i,y_i), 0\}$$

- $\sum_{a,b \in \text{alphabet}} s(a,b)p(a)p(b) < 0 \Rightarrow$ most regions receive zero score.

-The scores of individual sites are independent.

-The landscape of non-zero regions are "islands" in the sea.



-The optimal alignment score S is the global maximum of these island peaks:

$$S = \max\{\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_\kappa\}$$

– The island peak σ_i satisfies a Poisson distribution:

$$\Pr(\sigma_i > x) = (\text{constant})\, e^{-\lambda x}$$

– The parameter λ is the positive root of the equation

$$\sum_{a,b \in \text{alphabet}} e^{\lambda s(a,b)}\, p(a)\, p(b) = 1$$

The probability that the maximum S is smaller than x is

$$P(S < x) = \prod_i [1 - Pr(\sigma_i > x))] \rightarrow \exp[- \kappa e^{-\lambda x}] \text{ when } \kappa \rightarrow \infty.$$

This is a form of **Extreme Value Distribution**.

p-value = probability of at least one sequence scoring with S > x in the given database.

$$P(S > x) = 1- \exp[- \kappa e^{-\lambda x}].$$

E-value = expected number of matches with scores better than S in a database search.

$$E(S) = kmn e^{-\lambda S}.$$

Notes:

- All of the above discussions only applicable to local alignments.
- For gapped local alignments, the same statistics are believed to apply, although not proved.
- The trick is to learn parameters λ and K. These values depend upon the substitution matrix and sequence compositions, and can be estimated from randomly generated data.
- Score statistics for global alignments are not well known.

Q: What is a bit score in the blast search result?

A: The bit score is defined as $S' = (\lambda S - \ln K)/\ln 2$

it is then convenient to calculate the e-value

$$E(S) = mn\, 2^{-S'}$$

Question: What threshold to use?

Answer: No absolute answer, it varies.

E.g., E-value << 1.

$$E(S) = kmn\, e^{-\lambda S}. \Rightarrow S > T + \log(mn)/\lambda.$$

More details are referred to the text and the following.

1. Y.K. Yu & T. Hwa, "Statistical significance of probabilistic sequence alignment and related local hidden Markov models", J. Computational Biology 8(2001)249-282.
2. http://blast.wustl.edu/doc/infotheory.html

```
>gi|7434520|pir||G64632 acetate kinase - Helicobacter pylori (strain 26695)
          Length = 388

 Score = 35.8 bits (81), Expect = 0.10
 Identities = 21/51 (41%), Positives = 29/51 (56%), Gaps = 2/51 (3%)

Query: 1  VLVLNCGSSSLKFAIIDAVNGEEYLSGLAECF--HLPEARIKWKMDGNKQE 49
          +LVLN GSSS+KF + D    +   SGLAE     + + +IK  +  N QE
Sbjct: 3  ILVLNLGSSSIKFKLFDMKENKPLASGLAEKIGEEIGQLKIKSHLHHNDQE 53
```

Heuristic alignment algorithms

- motivation: speed

  sequence DB ~ O(100,000,000) basepair

  query sequence 1000 basepair

  O(nm) time complexity => 10^{11} matrix cells in dynamic programming table

  if 10,000,000 cells/second => 10000 seconds ~ 3 hours.

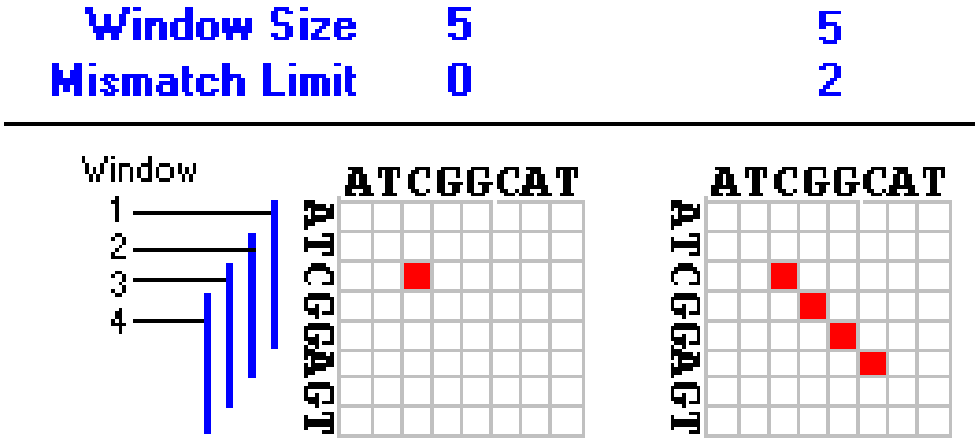
  O(n+m) time => ~ 10 seconds

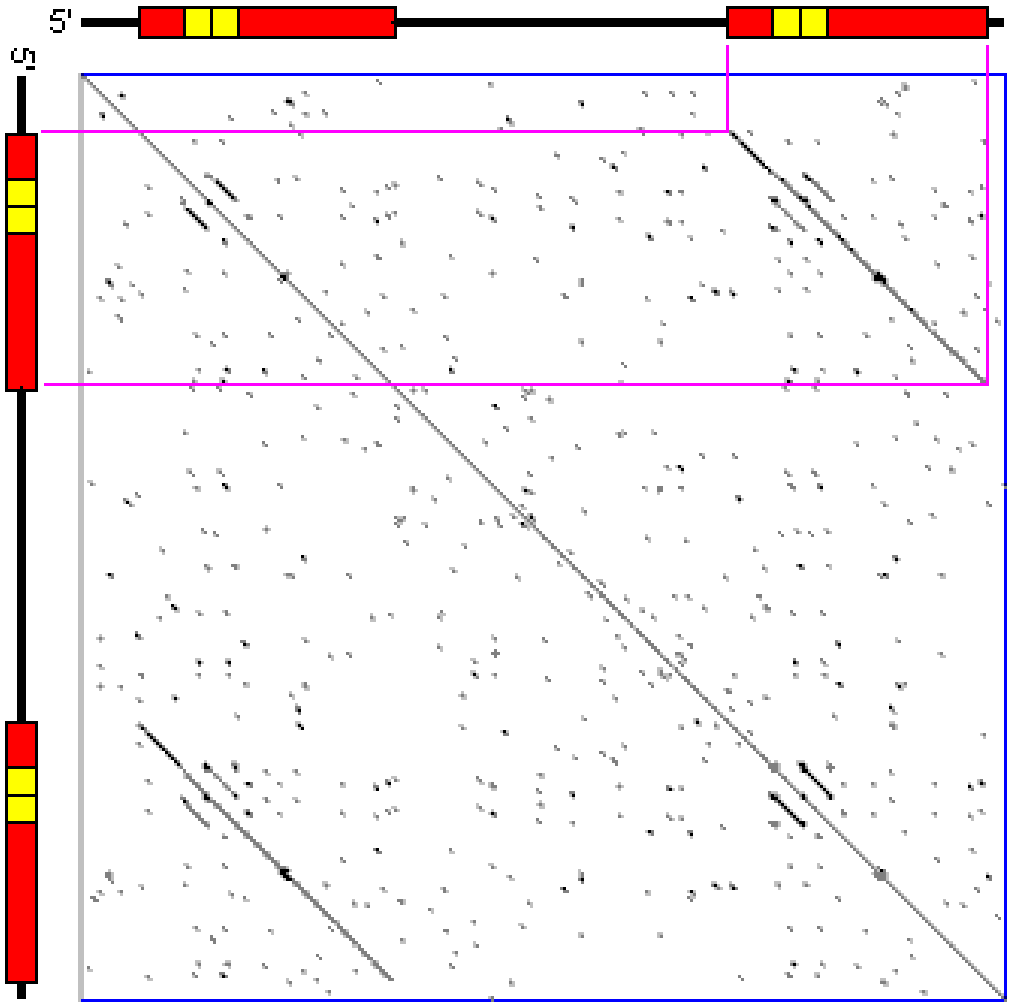
- heuristic versus rigorous

Dot Plots

A matrix comparison of two sequences (or one with itself) is prepared by "sliding" a window of user-defined size along both sequences.

If the two sequences within that window match with a precision set by the mismatch limit, a dot is placed in the middle of the window signifying a match.




5'
5'

FASTA [Pearson & Lipman 1988]

1. k-tup (k=6 for DNA, 2 for proteins)

   rule of thumb: the smaller the word, the slower and more sensitive the search is

2. build a lookup table (also called as hash table or dictionary)

word	query	DB	offset
AAAAAA	xxx	yyyy	zzz
AAAAAT	xxx	yyyy	zzz
AAAAAC	xxx	yyyy	zzz
....	….	…..	….

FASTA cont'd

3. scan through the database: for each every word of size k, look up in the lookup table in step 2. The offsets between the positions of the word in the query and the database entry are calculated and saved.

(implication: same offsets suggest segment of similarity.)

4. join nearby *contiguous* stretches of similarity (diagonal) =>scores init1.

5. join adjacent diagonals into a single long region (by introducing gaps) => scores init*n*.

6. do a dynamic programming algorithm for regions with high init*n* score to determine the *opt* score.

Q: what is the FASTA format?

- Average local alignment score for database sequences in the same length range is determined
- Average score is plotted against log of average length
- The plotted points are fitted to a straight line
- A z value, the number of standard deviations from the fitted line, is calculated for each score
- Extreme value distribution:

$$P(Z > z) = 1 - \exp(-e^{-1.825\, z - 0.5772})$$

- Expected # of sequences in a database of D sequences to have scores higher than z is

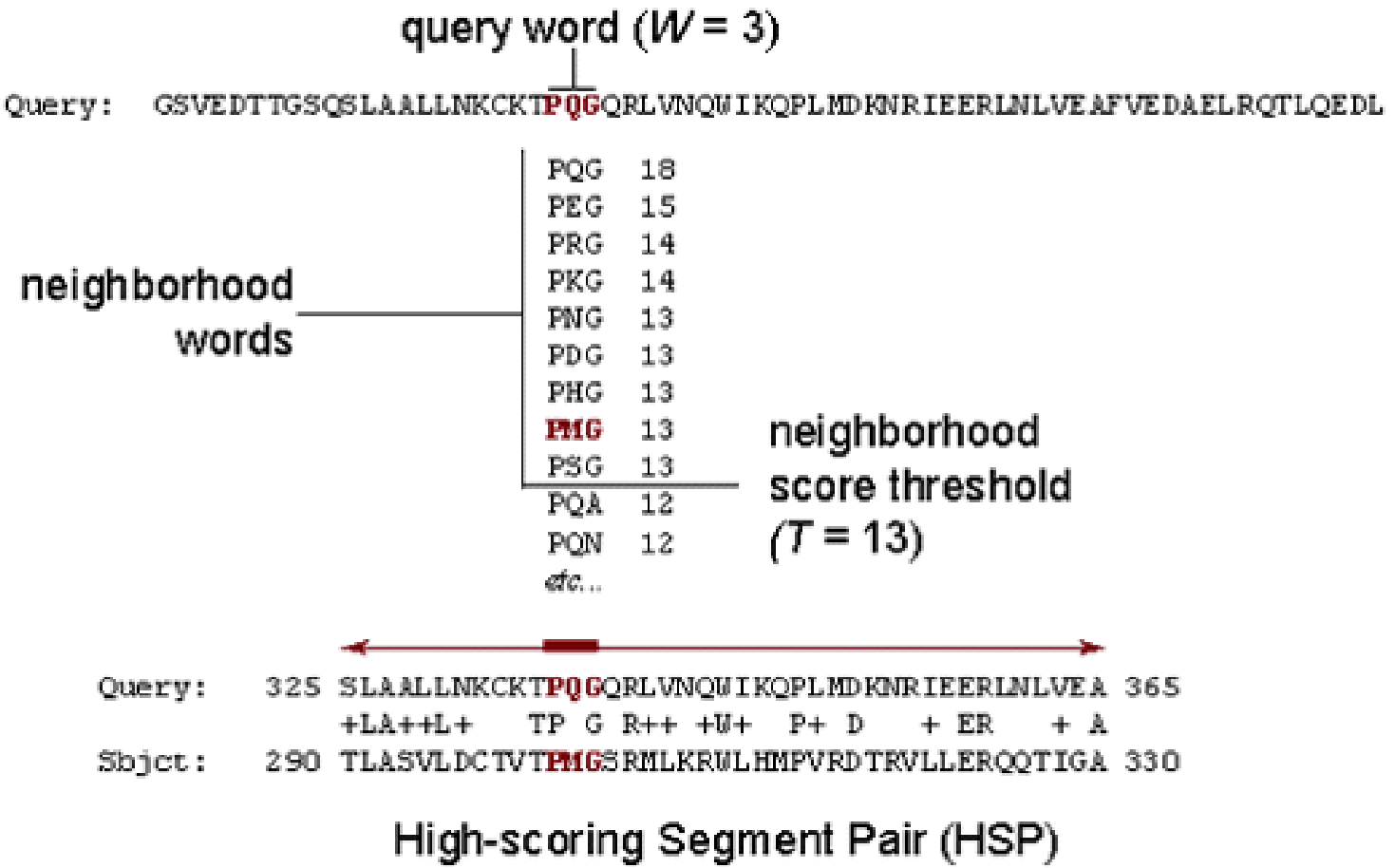
$$E(Z > z) = D \times P(Z > z)$$

```
z-score  obs     exp
          (=)     (*)

< 20    863      0:=====
  22      0      0:
  24      0      0:
  26      2      2:*
  28     14     25:*
  30     81    149:*
  32    306    577:== *
  34   1045   1564:======   *
  36   2925   3213:==================*
  38   5368   5310:==============================*
  40   7971   7407:=========================================*===
  42   9957   9054:=================================================*=====
  44  10706   9987:=====================================================*====
  46  10069  10172:======================================================*
  48   9611   9739:=====================================================*
  50   8595   8887:================================================*
  52   7636   7813:============================================*
  54   6559   6674:=======================================*
  56   5262   5574:=============================== *
  58   4590   4576:==========================*
  60   3638   3707:====================*
  62   2916   2972:================*
  64   2320   2364:=============*
  66   1907   1868:==========*
  68   1368   1469:========*
  70   1122   1152:======*
  72    837    900:=====*
  74    631    702:===*
  76    483    546:===*
  78    349    424:==*
  80    299    330:=*
  82    213    252:=*
  84    132    200:=*
  86    112    155:*
  88     87    120:*
  90     74     93:*
  92     47     72:*          :=== *
  94     27     55:*          :== *
  96     29     43:*          :==*
  98     25     33:*          :=*
 100     24     26:*          :=*
 102      9     20:*          :=*
 104     11     15:*          :*
 106      6     12:*          :*
 108      7      9:*          :*
 110      4      7:*          :*
 112      3      6:*          :*
 114      2      4:*          :*
 116      3      3:*          :*
 118      1      3:*          :*
>120    829      2:*====      :*=======================================

Joining threshold: 36, opt. threshold: 24, opt. width:  16, reg.-scaled

The best scores are:                      init1 initn   opt    z-sc E(108291)..

PIR1:HGCZG
! hemoglobin gamma-G chain - chimpanzee    971   971   971  1145.0  6.2e-57
```

Basic **L**ocal **A**lignment **S**earch **T**oolkit [Altschul et al, 1990]

1. A list of neighborhood words of fixed length (3 for protein and 11 for DNA) that match the query with score > a threshold.
2. Scan the database sequences and look for words in the list; once find a spot, try a "hit extension" process to extend the possible match as an ungapped alignment in both directions, stopping at the maximum scoring extension.

The BLAST Search Algorithm



Variants of BLAST search

- BLASTP: protein vs. protein
- BLASTN: nucleotide vs. nucleotide
- BLASTX: nucleotide (translated to protein) vs. protein
- TBLASTN: protein vs. nucleotide (translated to protein)
- TBLASTX: nucleotide (translated to protein) vs. nucleotide (translated to protein)

Note: Since proteins are strings of 20 alphabets the odds of having false positive matches is significantly lower than that of DNA sequences, which are strings of 4 alphabets.

Open Reading Frame (ORF)

Universal Genetic Code

5'	2nd Position				3'
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Val	Ala	Asp	Gly	C

ACCUUAGCGUA

Thr Leu Ala

Reading frame 1

ACCUUAGCGUA

Pro Stop Arg

Reading frame 2

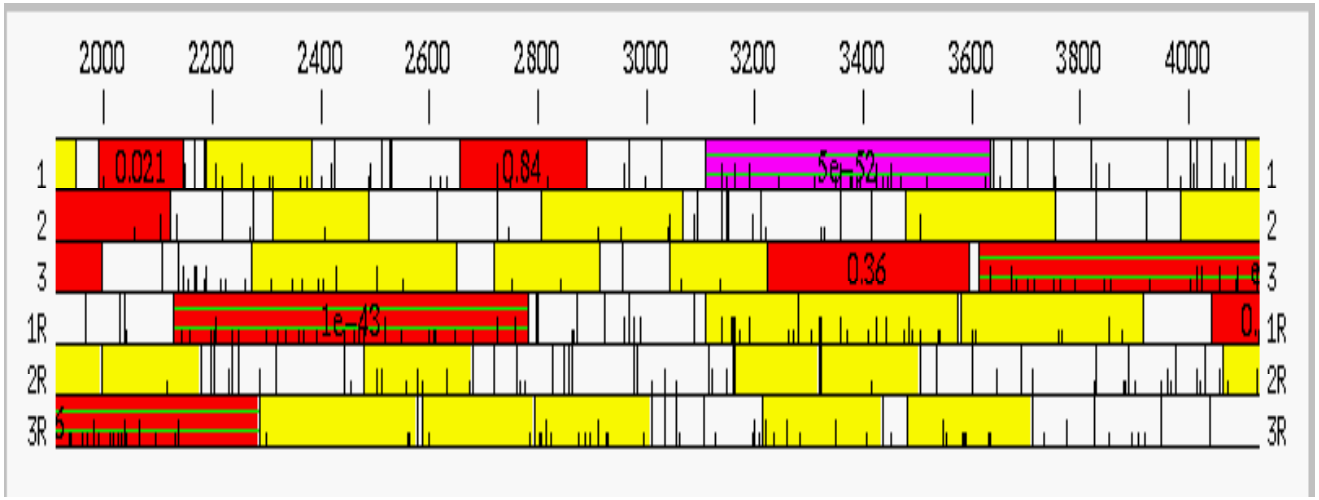
ACCUUAGCGUA

Leu Ser Val

Reading frame 3

Three other reading frames on the reverse complementary strand

Using BLAST to identify genes



Comparative method:

- identify possible ORFs (e.g. with length > 50 bp)
- search against Genbank for homologs and use a threshold of E-value (e.g., e^{-5}) to call putative genes.