**Course: Advanced Bioinformatics**

**Module title: FASTA**

**Module no. : 33**

This document provides information about FASTA. FASTA was the first rapid search method developed for database searching. FASTA uses a heuristic algorithm to speed up the process of locating similar regions. Unlike dynamic programming, FASTA is not guaranteed to lead to the optimal solution. However, the search time is roughly 50 times faster than DP solutions.

#### FASTA Algorithm

In the initial stage of searching for regions of similarity, FASTA uses a hashing approach. For each of the sequences being compared, a table is constructed showing the positions of each word of length k, or k-tuple. The relative positions of each word in the two sequences are calculated by subtracting the position of the first sequence from the position of the second. Words having the same offset are in phase and reveal a longer region of alignment between the two sequences.

Step 2: the ten regions with the highest density of identities are identified. The ends of each region is trimmed to include only residues contributing to the highest score. Each resulting region is now a partial alignment without gaps. Each is given a score (init1 score)

Step 3: If there are several initial regions with scores greater than a cutoff value, check to see if the trimmed initial regions can be joined to form an approximate alignment with gaps. A similarity score is calculated as the sum of the init1 scores for each of the initial regions minus a penalty for each gap. (initn score)

Step 4: Construct a needleman-wunch optimal alignment of the query sequence and the library sequence, considering only those residues that lie in a band 32 residues wide, centered on the best initial region found in step 2 (opt score)

After locating the k-tuples and grouping the ones with the same offset together, an optimization step is invoked to piece together k-tuple alignments allowing gaps.

Using this approach, the search time increases linearly with the size of the query and target sequences. Compared to the polynomial increase with dynamic programming, FASTA presents a much faster alternative, particularly as the sequence size increases.

For DNA and RNA sequences, the typical size of the k-tuple in the FASTA algorithm is 4-6, while in protein sequences it is 1 or 2. The larger the k-tuple, the faster FASTA will run, but the less thorough it will be in determining regions of similarity.

**Significance of fasta scores**

In order to determine the significance of an alignment for a target database and a query sequence, FASTA calculates the u and lambda parameters for the extreme value distribution, which will vary with the length and the composition of the sequences being compared. The steps to calculate z-scores for each possible score is calculated as follows:

1. The average score for database sequences in the same length range is determined.
2. The average score is plotted against the logarithm of average sequence length in each length range.
3. The points are then fitted to a straight line by linear regression.
4. A z score, the number of standard deviations from the fitted line, is calculated for each score.
5. High-scoring, presumably related sequences, and also very low scoring alignments that do not fit the straight line are removed from consideration.
6. Steps 1-5 are repeated one or more times.
7. The known statistical distribution of alignment scores is used to calculated the probability that a Z score between unrelated or random sequences of the same lengths as the query and database sequence could be greater than z, which follow an extreme value distribution such that: (Pearson, 2000 ISMB)



The expectation of observing a Z-score greater than z in a database of D sequences is:



1. Z scores are then normalized to z’ = 50 + 10z so that an alignment score with a standard deviation of 5 now has a normalized score of 100.
2. The significance of the alignment score between a sequence and a database can be further analyzed by aligning a sequence with a shuffled library.

# HISTOGRAM OF FASTA DATA

One of the items reported in the FASTA output is a histogram showing a graphical representation of the distribution of the normalized scores when matched with the query sequence. These scores are expected to fall approximately into a normal distribution, and any significant matches will fall outside the normal curve.

The first column listed in the fasta score distribution is the z’ score, which is a z score normalized to a mean of 50 and a standard deviation of 10. The second column lists the number of optimized scores found in that range. The third column lists the number of expected sequences to lie within a range, given an extreme value distribution and the calculated values of u and lambda.

The “=” signs give an approximate curve for the actual distribution, while the “\*” indicates the expected score distribution.

The z’-scores greater than 120 are considered to be high-scoring alignments.

opt E()

< 20 188 0:==

22 0 0: one = represents 109 library sequences

24 0 0:

26 2 1:\*

28 7 15:\*

30 28 91:\*

32 200 353:== \*

34 841 958:========\*

36 2217 1968:==================\*==

38 3746 3253:=============================\*=====

40 5360 4538:=========================================\*========

42 6055 5547:==================================================\*=====

44 6496 6119:========================================================\*===

46 5820 6232:====================================================== \*

48 5469 5966:=================================================== \*

50 4820 5444:============================================= \*

52 4202 4787:======================================= \*

54 3815 4089:=================================== \*

56 3271 3415:===============================\*

58 2755 2804:=========================\*

60 2268 2271:====================\*

62 1813 1821:================\*

64 1500 1448:=============\*

66 1233 1145:==========\*=

68 951 900:========\*

70 746 706:======\*

72 699 551:=====\*=

74 460 430:===\*=

76 337 335:===\*

78 287 260:==\*

80 244 202:=\*=

82 185 154:=\*

84 115 122:=\*

86 114 95:\*=

88 75 73:\* inset = represents 1 library sequences

90 70 57:\*

92 48 44:\* :=======================================\*

94 26 34:\* :========================== \*

96 33 26:\* :=========================\*=======

98 14 20:\* :============== \*

100 10 16:\* :========== \*

102 7 12:\* :======= \*

104 6 9:\* :====== \*

106 5 7:\* :===== \*

108 2 6:\* :== \*

110 2 4:\* :== \*

112 1 3:\* := \*

114 0 3:\* : \*

116 0 2:\* : \*

118 0 2:\* : \*

>120 27 1:\* :\*==========================

After the histogram is calculation of the Kolmogorov-Smirnov statistic, which yields some information into the deviation between the observed and expected distributions. If the deviation is significant enough, then the alignment should be performed again with different gap penalties.

After the statistics is a list of the best scoring hits. Note that FASTA presents at most one highest scoring hit per sequence, whereas other alignment programs may present many. Listed in the hits section are the description of the sequence, the z’ score, the initn, initl, and opt scores (note the initn score is the extended hit score; the init1 score is the initial hit score; the opt score is the score calculated by stringing together regions with gaps – see Figure 7.2 of Mount for a more in-depth explanation) and the E score (calculated as an estimate of the likelihood of a match occurring by chance).

The best scores are: initn init1 opt z-sc E(66345)

MERR\_PSEAE mercuric resistance operon regu ( 144) 928 928 928 1129.8 0

MERR\_SHIFL mercuric resistance operon regu ( 144) 871 871 871 1061.3 0

MERR\_SERMA mercuric resistance operon regu ( 144) 810 810 810 988.1 0

MERR\_STAAU mercuric resistance operon regu ( 135) 292 172 298 373.6 3.5e-14

MERR\_BACSR (strain rc607). mercuric resist ( 132) 241 198 289 363.0 1.4e-13

YHDM\_ECOLI hypothetical transcriptional re ( 141) 175 175 276 347.0 1.1e-12

After the list of the highest scoring hits are the smith-waterman alignments between the query and the highest scoring hits. A ‘:’ marks conservation; ‘.’ denotes a conservative substitution:

>>MERR\_STAAU mercuric resistance operon regulatory protei (135 aa)

initn: 292 init1: 172 opt: 298 Z-score: 373.6 expect() 3.5e-14

Smith-Waterman score: 298; 36.923% identity in 130 aa overlap

10 20 30 40 50 60

MerR MENNLENLTIGVFAKAAGVNVETIRFYQRKGLLLEPDKPYGSIRRYGEADVTRVRFVKSA

. :. .::: :: ::.:.:.::::. : . .. : :.: . ::::.:

MERR\_S MGMKISELAKACDVNKETVRYYERKGLIAGPPRNESGYRIYSEETADRVRFIKRM

10 20 30 40 50

70 80 90 100 110

MerR QRLGFSLDEIAELLRL--EDGTHCEEASSLAEHKLKDVREKMADLARMEAVLSELVCACH

..: ::: :: :. . .:: .:.. ... .: :....:. : :.. .: :: :

MERR\_S KELDFSLKEIHLLFGVVDQDGERCKDMYAFTVQKTKEIERKVQGLLRIQRLLEELKEKCP

60 70 80 90 100 110

120 130 140

MerR ARRGNVSCPLIASLQGGASLAGSAMP

... .::.: .:.::

MERR\_S DEKAMYTCPIIETLMGGPDK

120 130

***FASTA Programs***

FASTA – compares a query protein sequence to a protein sequence library or a DNA sequence to a DNA sequence library.

TFASTA – compares a query protein sequence to a DNA sequence library, after the DNA sequence library has been translated in all six reading frames.

FASTF – compares a set of ordered peptide fragments, obtained from analysis of a protein by cleavage and sequencing of protein bands resolved by electrophoresis, against a protein database

TFASTF – compares a set of ordered peptide fragments, obtained from analysis of a protein by cleavage and sequencing of protein bands resolved by electrophoresis, against a DNA database

FASTS – compares a set of ordered peptide fragments, obtained from mass-spectometry analysis of a protein, against a protein database.

TFASTS – compares a set of ordered peptide fragments, obtained from mass-spectometry analysis of a protein, against a DNA database.

Example

>mgstm1

MGCEN,MIDYP,MLLAY,MLLGY

FASTX, FASTY – compares a query DNA sequence to a protein sequence database, translating the DNA sequence in all six reading frames and allowing frameshifts.

TFASTX, TFASTY – Compares a protein sequence to a DNA sequence or DNA sequence library, such that the DNA sequence is translated in all six reading frames, and the protein query sequence is compared to each of the six derived protein sequences. The DNA sequence is translated from one end to the other; termination codons are translated into unknown amino acids.

LALIGN, LFASTA – Same as the FASTA program, except that multiple aligning regions may be reported for each sequence.

PLALIGN – dot plot algorithm available through the fasta suite

FAST-pat, FAST-swap: compares a sequence to a pattern database

FAST-swap