COMPUTATIONAL MOLECULAR BIOLOGY
HOMEWORK 4
SHUBHRA GUPTA
ASU ID 993755974
Comparison between FASTA and BLAST

In this paper I am going to compare FASTA with BLAST. These two algorithms address the problem of sequence database search. Motivation is to obtain new DNA or protein sequences by similarity search (If new sequence is not similar to existing sequences).

What is FASTA?

FASTA is a program for database searching by homology. FASTA finds exact local matches and then extends them to get a global alignment. FASTA may be better for less similar sequences.

What is BLAST?

BLAST is a fast, heuristic search tool for sequence databases. BLAST searches involve finding ungapped, locally optimal sequence alignments. BLAST compare an amino acid query sequence against a protein sequence database or a nucleotide query sequence against a nucleotide sequence database, as well as other combinations of protein and nucleic acid comparisons.
Both FASTA and BLAST programs universally used to approximate local alignment and local similarity.

Local alignment: Smith-Waterman algorithm

```
----TGKG----
        | | |
----AGKG----
```

Term approximation has the normal colloquial meaning, rather than mathematical meaning of a bounded-error approximation algorithm.

Global alignment: Needleman-Wunsch algorithm

```
LGPSSKQTGKGS-SRIWDN
|         |       | | |       |       |
LN-ITKSAGKGAIMRLGDA
```

Main difference between FASTA and BLAST:

In BLAST substrings of the query sequence and the database sequence, the score of the pair is the highest, but there is no gap alignment allowed between them. Although penalized in FASTA, gaps are allowed.

FASTA

FASTA is a short of “fast-all” or “FastA”. FASTA is better for nucleotides search than for proteins
Wilbur and Lipman in 1983, Lipman and Pearson in 1985, Pearson and Lipman in 1988 developed FASTA program. FASTA is continued to improve by Pearson.
FASTA is a heuristic program for rapid alignment of pairs of protein and DNA sequences. FASTA compare an input DNA or protein sequence to all of the sequences in a target sequence database and then report the best-matched sequences and local alignment of these matched sequences with the input sequence. Rather than comparing individual residues in the two sequences, FASTA searches for matching sequence patterns or words, called k-tuples. E.g.

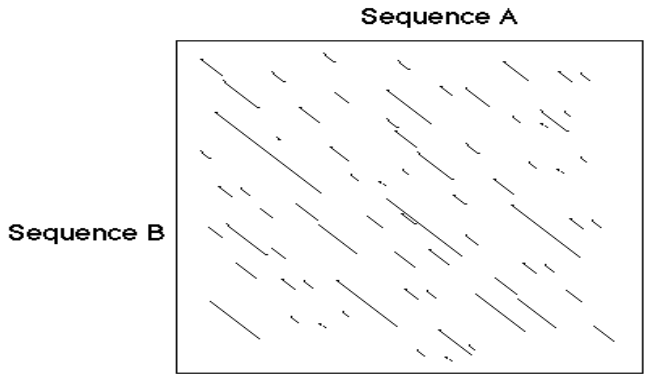
Protein 1 n c s p t a
| | |
Protein 2 a c s p r k
Two sequences that share a pattern c-s-p (common word)

FASTA provides a rapid way to find short stretches of similar sequence between a new sequences and any sequence in a database. Each sequence is broken down into short words a few sequence characters long and these words are organized into a table indicating where they are in the sequence. If one or more words are present in both sequences, and especially if several words can be joined, the sequences must be similar in those regions. This method is used from the web site http://fasta.bioch.virginia.edu/fasta/

Suppose sequence A - - w I v - -
sequence B - - w I v - -

In the above e.g. FASTA looks for short regions in these two amino acid sequences that match and then tries to extend the alignment to the right and left.

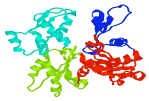
FASTA searches for all possible words of the same length. FASTA theoretically provides a more sensitive search of DNA sequence databases.
Sensitivity:
FASTA > BLAST
Speed:
BLAST > FASTA



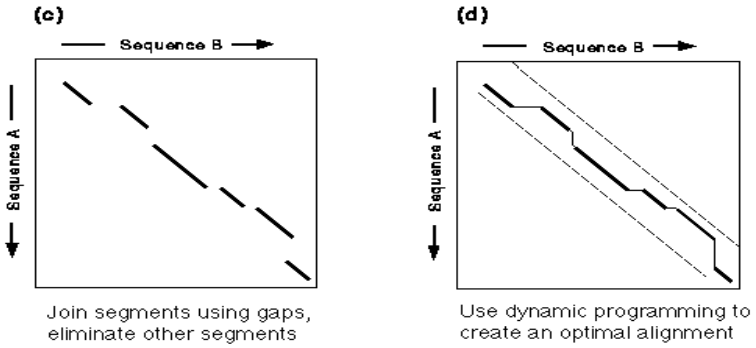
FASTA algorithm: Basic steps
Step1:
Set a word size, usually 6 for DNA and 2 for protein.
Hashing: FASTA locates regions of the query sequence and matching regions in the database sequences that have high densities of exact word matches (without gaps). The length of the matched word is called the k-tuple parameter.

Step 2:
Scoring: The ten highest scoring regions are rescored using the BLOSUM50 scoring matrix. The score for such a pair of regions is saved as the $init_1$ score.

Step 3:
Introduction of Gaps: FASTA determines if any of the initial regions from different diagonals may be joined together to form an approximate alignment with gaps. Only non-overlapping regions may be joined. The score for the joined regions is the sum of the scores of the initial regions minus a joining penalty for each gap. The score of the highest scoring region, at the end of this step, is saved as the init n.



FASTA (4)



Step 4:
Alignment: After computing the initial scores, FASTA determines the best segment of similarity between the query sequence and the search set sequence, using a variation of the Smith-Waterman algorithm. The score for this alignment is the opt score.

Step 5:
Random Sequence Simulation: In order to evaluate the significance of such alignment FASTA empirically estimates the score distribution from the alignment of many random pairs of sequences. More precisely, the characters of the query sequences are reshuffled (to maintain bias due to length and character composition) and searched against a random subset of the database. This empirical distribution is extrapolated, assuming it is an extreme value distribution, and each alignment to the real query is assigned a Z-score and an E-score.

Modifications:
In step4, use a band around init1.

In terms of Statistics:
FASTA calculates significance “on the fly”. This can be problematic if the dataset is small.

To identify an unknown protein sequence use either of these: FastA3, Ssearch3 or tFastX3. FASTA3 has improved methods of aligning sequences and of calculating the statistical significance of alignment.

http://www.ebi.ac.uk/fasta33/index.html

BLAST

BLAST is for "basic local alignment search tool". BLAST is better for proteins search than for nucleotides.
S. Altchul et al developed BLAST program in 1990.
The BLAST algorithm was developed as a new way to perform a sequence similarity search by an algorithm that is faster and sensitive than FASTA. This method is widely used sequence analysis facility in the world and provides similarity searching to all currently available sequences from the web site of the National Center for Biotechnology information (NCBI).
http://www.ncbi.nlm.nih.gov/BLAST/

Suppose sequence A - - w I v - -
sequence B - - w I v - -

In the above e.g. BLAST works similarly as FASTA, but only examines matched patterns of length 3 of the more significant amino acid substitutions.

BLAST search for common words or k-tuples in the query sequences and each database sequence.

L P P Q G L L query sequence
M P P E G L L database sequence
2 7 <7 2 6> 4 4
←------ -----→
extension to left extension to right
BLOSUM62.scores. word score = 15 for PQ G

BLAST algorithm: Basic steps
Step1:
Set a word size, usually 11 for DNA and 3 for protein. Given query sequence, compile the list of possible words, which form with words in high scoring word pairs (Filter out low complexity regions)

Step 2:
Scan database for exact matching with the list of words complied in step 1.
e.g. qlnfsagw -> (ql, ln, nf, fs, sa, ag, gw)
Extend the list (using some threshold T)

Step 3:
Scan through the string and whenever a word in the list is found try to extend it in both directions (no gaps) to get to a score beyond a threshold S. While extending use a parameter L that defines how long an extension will be tried to raise the score over S.

Modification of step 3:
-Original BLAST: Extension is continued as long as the score continued to increase.
-Another version
-BLAST2 (gapped BLAST): - Lower value of T is used.
- After extension try to combine (allowing gaps)
- Find maximal scoring segment.
This program uses the BLASTP or BLASTN algorithms for aligning two sequences.

In terms of Statistics:
BLAST calculates probabilities and this can fail if some assumptions are invalid for that search.

There are versions of BLAST for searching nucleic acid and protein databases, which can be used to translate DNA sequences prior to comparing them to protein sequence databases in 1997.

Recent improvement in BLAST is GAPPED-BLAST (three times faster than the original BLAST) and PSI-BLAST (position-specific-iterated BLAST). The GAPPED-BLAST algorithm allows gaps to be introduced into the alignments. That means that similar regions are not broken into several segments (as in the older versions). This method reflects biological relationships much better than ordinary BLAST.

Other Blast program:

Program	Search sequence	type of alignment
BLASTN	Nucleotide	gapped
BLASTP	Protein	gapped
BLASTX	Protein, Nucleotide	each framed gapped
TBLASTN	translated nucleic acid	each framed gapped

Tips for database searches:
1. Use the latest version of database.
2. First Run BLAST, then depending on the results, run a finer tools (FASTA or others)

References:
Pevzner, P. A., Computational Molecular Biology-An Algorithmic Approach, the MIT Press Cambridge, 2000.
Gusfield, D., Algorithms on Strings, Trees and Sequences, Cambridge University Press, 1997.
Mount, D. W., Bioinformatics- Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, 2000.

Other References:
http://www.roselab.jhu.edu/~przytyck/Lect03b-2002.ppt
http://ludwig.chem.wesleyan.edu/~model/chem389/Database.htm
http://www.public.asu.edu/~cbaral/ - Class notes/slides- FASTA and BLAST links
www.bioinformaticsonline.org