CS123A Bioinformatics Module 2 – Week 5 – Presentation 2

Leonard Wesley

Computer Science Dept

San Jose State Univ

Agenda

- BLAST
  - Local alignment algorithm

BLAST: Basic Local Alignment Search Tool

Why BLAST?

- Needleman–Wunsch (1970) global alignment algorithm is not used for database searches because we are usually more interested in identifying locally matching regions such as protein domains.

- The Smith–Waterman (1981) local alignment algorithm finds optimal pairwise alignments, but we cannot use it for database searches generally because it is too computationally intensive.

- BLAST offers a local alignment strategy having both speed and sensitivity, as described in this chapter. It also offers convenient accessibility on the World Wide Web or as a command-line tool.

BLAST Can Be Used For …

- *Determining what orthologs and paralogs are known for a particular protein or nucleic acid sequence.*
- *Determining what proteins or genes are present in a particular organism.*
- *Determining the identity of a DNA or protein sequence.*
- *Discovering new genes.*
- *Determining what variants have been described for a particular gene or protein.*
- *… MORE*

BLAST Starts With A Seed Word

- Can be performed for DNA, Proteins, …etc.
- Global alignment looks for comparison over the entire range of the two sequences involved.

```
GCATTACTAATATATTAGTAAATCAGAGTAGTA
        |||||||||  ||
AAGCGAATAATATATTTATACTCAGATTATTGCGCG
```

We see only a portion of these two sequences can be aligned.

- By contrast, when a local alignment is performed, a small seed is uncovered that can be used to quickly extend the alignment. The initial seed for the alignment:

```
          TAT
          |||
AAGCGAATAATATATTTATACTCAGATTATTGCGCG
```

BLAST *(cont.)*

- now the extended alignment:

```
          TATATATTAGTA
          ||||||||| ||
AAGCGAATAATATATTTATACTCAGATTATTGCGCG
```

BLAST First steps …

- Break the query into short words of a specific length.
- A word is a series of characters from the query sequences. The default length of the search is three (3) characters.
- The words are constructed by using a sliding window of three characters. For example, twelve amino acids near the amino terminal of the *A.thaliana (*a small flowering plant native to Eurasia and Africa*)* protein phosphoglucomutase sequence are:

  NYLENFVQATFN

- This sequence is broken down into three character words by selecting the first amino acid characters, moving over one character, selecting the next three amino acid characters, and so on to create the following seven words:

  NYL YLE LEN ENF NFV FVQ VQA QAT ATF TFN

BLAST *(cont.)*

- These words are then compared against a sequence in a database. Here is an example of a word match with rabbit muscle phosphoglucomutase :

```
Query              ENF
Subject    SSTNYAENTIQSIISTVEPAQR
```

- This search is performed for all words. For the original BLAST search, those words whose T value was greater than 18 were used as seeds to extend the alignment.

- The T value is derived by using a scoring matrix. The BLOSUM 62 matrix is the default for protein searches and will be discussed later.

BLAST *(cont.)*

- The alignment is extended in both directions until the alignment score decreases in value.
- As an example, consider the following alignment between the A. thaliana and rabbit muscle phosphoglucomutase :

```
Query      NLYENFVQATFNALTAEKV
           NY ENF+Q+  + +   +
Subject    NYAENTIQSIISTVEPAQR
```

- The centerline provides the following information. A letter designates an identity (or high similarity) between the two sequences. A “+” means the two sequences are similar but not highly similar. If no symbol is given between the two sequences, then a non-similar substitution has occurred.

BLAST *(cont.)*

- Those alignments whose T score does not decrease are then compared with scores obtained by random searches.
- Those alignments whose score is above the cutoff are called a High Scoring Segment Pair (HSP).
- Once this alignment process is completed for a query and each subject sequence in the database, a report is generated.
- This report provides a list of those alignments (default size of 50) with a value greater than the S cutoff value.

BLAST *(cont.)*

- Alignments are also possible between a nucleotide query and a nucleotide database.

- The entire BLAST process described above is the same for nucleotide searches except the default word size is eleven and a different scoring matrix is applied.

- Scoring matrices are used to obtain the S value.

- For nucleotides, these are simple; each identical match is given the same score, and all mismatches are given a penalty (negative) score.

BLAST Nucleotide Scoring Matrices

BLAST Nucleotide Matrix ("Ungapped Alignment")

	A	T	C	G
A	5			
T	-4	5		
C	-4	-4	5	
G	-4	-4	-4	5

BLAST Nucleotide Matrix ("Gapped Alignment")

	A	T	C	G
A	1			
T	-3	1		
C	-3	-3	1	
G	-3	-3	-3	1

BLAST BLOSUM62 Amino Acid Scoring Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

B = Asparagine (N) or Aspartic acid (D)

Z = Glutamine (Q) or Glutamic acid (E)

X = Any amino acid

* = gap/terminator

History Of BLOSUM Matrix

- Henikoff and Henikoff (1992. PNAS 89:10915-10919) studied 2000 aligned blocks of 500 groups of related proteins.
- They determined the different types of amino acid substitutions that occurred in these proteins. From this study, they developed the BLOSUM 62 matrix. (BLOSUM = BLOcks SUbstitution Matrix)
- This matrix gives a score (positive value) or penalty (negative value) for each amino acid identity or substitution between two aligned sequences.
- From the table, not all identities or substitutions are of equal value. This is because the comparison the authors did between the many proteins gave an indication of the likelihood that a specific substitution might occur.

Why BLOSUM 62?

- If you were to score an alignment between two amino acid sequences that were 62% identical, their BLOSUM 62 score would be 1.

- Similar matrices are also available if you require a higher or lower percent identity. These are BLOSUM 45 and BLOSUM 80.

- The BLOSUM 45 matrix should be used if you are looking for distantly related sequences, whereas the BLOSUM 80 matrix is appropriate for searches involving highly conserved sequences. For protein alignments, the BLAST algorithm uses BLOSUM 62 as the default matrix.

Deriving BLAST Scores

- Using the BLOSUM62 matrix, we can then derive a score for the following alignment.

```
Query      NLYENFVQATF
           NY ENF+Q+
Subject    NYAENTIQSII
```

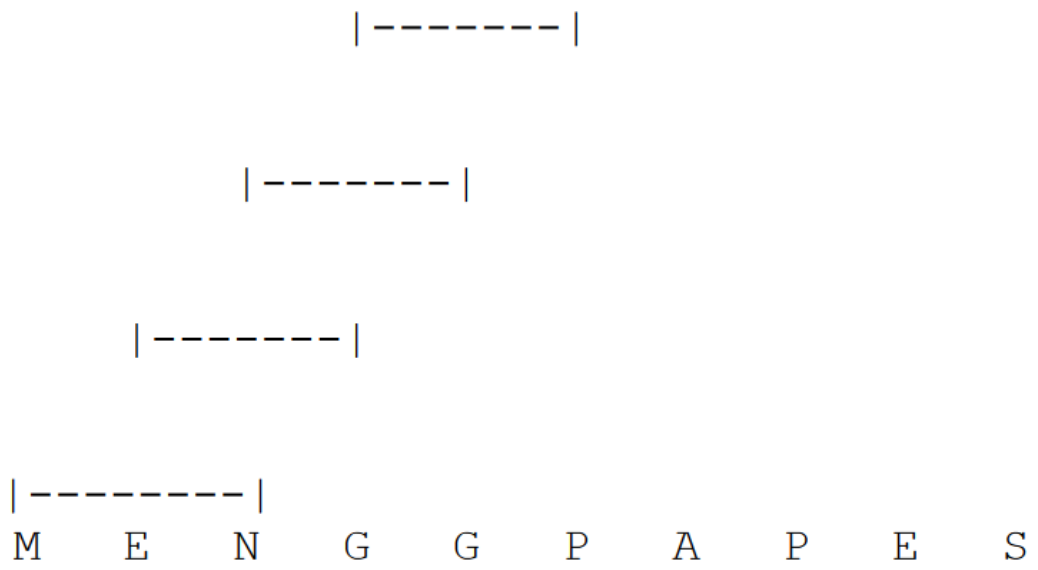
- Going from left to right the score is summed as follows:

```
Query      N  L  Y  E  N  F  V  Q  A  T  F
Subject    N  Y  A  E  N  T  I  Q  S  I  I
Score      5 -1 -2  5  6 -2  3  5  1 -1  0

Score = 19
```

In-Class BLAST Alignment Exercise

- Consider the following AA sequence that you wish to align.



Calculate T Scores

- Calculate T Scores against the following sequence

- I P A G G P A P E S

T Score ANSWER

Extend Seed

- Start with seed with T score = 19.
- Extend one letter in each direction. Then calculate alignment score.
- Stop extending if the updated score drops below previous score.

Extension ANSWER

1. Original alignment: T Score =19

```
                G   G   P
1  -1  -2   6   6   7   4   7   5   4   BLOSUM 62 Score
I   P   A   G   G   P   A   P   E   S
```

2. Extend one amino acid in each direction: T Score = 21

```
            N   G   G   P   A
1  -1  -2   6   6   7   4   7   5   4   BLOSUM 62 Score
I   P   A   G   G   P   A   P   E   S
```

3. Stop when next extension drops off below value X compared to previous score

Continue Calculating Score for Remaining Part Of Seq

M E **N G G P A** P E S

I P A G G P A P E S

BLAST2

- BLAST2 (1997. Nucleic Acids Research 25:3389-3402) takes a different (and three-times faster) approach than the original BLAST algorithm.

- As with the original BLAST it looks for matches to the three character words, but the T value is lower.

- It then identifies two words that lie next to each other and uses those neighboring words as the seed to extend the alignment.

- As with the original BLAST procedure, S scores are obtained, and expect (E) values are calculated.

BLAST 2 Advantages

- Another feature introduced with BLAST2 was the ability to add gaps to the alignment.

- Because gaps are evidence of evolutionary differences between sequences (assuming they are not sequencing errors), gap penalties are used to reduce the score value.

- The default for protein searches is a reduction of 11 for the introduction of a gap, and a reduction of 1 for each gap added at that same gap location.

- Gaps are useful because you can actually increase the score of a local alignment, even when gap penalties are included in the score.

BLAST Is A Collection Of Algorithms

Search	Query	Database
blastn	nucleotide	nucleotide
blastx	translated nucleotide in all six frames	protein
tblastx	translated nucleotide in all six frames	translated nucleotide in all six frames
blastp	protein	protein