BLAST
Basic **L**ocal **A**lignment **S**earch **T**ool

Phil McClean
September 2004

An important goal of genomics is to determine if a particular sequence is like another sequence. This is accomplished by comparing the new sequence with sequences that have already been reported and stored in a database. This process is principally one that uses alignment procedures to uncover the "like" sequence in the database. The alignment process will uncover those regions that are identical or closely similar and those regions with little (or any) similarity. Conserved regions might represent motifs that are essential for function. Regions with little similarity could be less essential to function. In a sense, these alignments are used to determine if a database contains a potential homologous sequence to the newly derived sequence. Further, phylogenetic studies are necessary to determine the orthologous/paralogous nature of the two aligned sequences.

Two alignment types are used: ***global*** and ***local***. The global approach compares one whole sequence with other entire sequences. The local method uses a subset of a sequence and attempts to align it to subset of other sequences. The output of a global alignment is a one-to-comparison of two sequences. Local alignments reveal regions that are highly similar, but do not necessarily provide a comparison across the entire two sequences. The global approach is useful when you are comparing a small group of sequences, but becomes become computationally expensive as the number of sequence in the comparison increases. Local alignments use heuristic programming methods that are better suited to successfully searching very large databases, but they do not necessarily give the most optimum solution. Even given this limitation, local alignments are very important to the field of genomics because they can uncover regions of homology that are related by descent between two otherwise diverse sequences.

Here are examples of global and local alignments. The global alignment looks for comparison over the entire range of the two sequences involved.

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

As you can see only a portion of this 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
```

And now the extended alignment:

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

The most common local alignment tool is BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) developed by Altschul et al. (1990. J Mol Biol 215:403). The operative phrase in the phrase is ***local alignment***. The BLAST is a set of algorithms that attempt to find a short fragment of a ***query*** sequence that aligns perfectly with a fragment of a ***subject*** sequence found in a ***database***. That initial alignment must be greater than a ***neighborhood score threshold (T)***. For the original BLAST algorithm, the fragment is then used as a seed to extend the alignment in both directions. The alignment is extended in both directions until the T score for the aligned segment does not continue to increase. Said another way, BLAST looks for short sequences in the query that matches short sequences found in the database.

The first step of the BLAST algorithm is to 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 characters. The words are constructed by using a sliding window of three characters. For example, twelve amino acids near the amino terminal of the *Aradbidopsis thaliana* 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
```

These words are then compared against a sequence in a database. Here is an example of a word match with rabbit muscle phoshoglucomutase (subject line):

```
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. 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 phophoglucomutase:

```
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.

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.

For each alignment reported, an ***Expect (e) Value*** is reported. This value is a function of the S value and the database size. An e value of 1 means that one alignment using a query of this size will by chance produce a S score of this value in a database of this size. As you can imagine an e value of –10 ($=1x10^{-10}$) means that it is much more unlikely that random chance lead to this current alignment compared to an alignment with an e value of 1. The expect value is often considered to be a probability. In other words, the probability of achieving a score of this value using a sequence of this length against a database of this size is equal to the expect value. Therefore, a lower e value means that alignment is significant at a specific probability level. It is important that you note that the expect value is specific to a database of a certain size. This means, that if you perform your BLAST alignment at a later date, you e value might change because the size of the database has changed.

In general, if you see an e value of –30, you can be assured that your sequence is homologous to the sequence to which aligned in this database. Furthermore, e values of –5 are often considered significant enough when annotating a genome.

The example above describes the process of using a protein query to search for alignments in protein database. 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. The two matrices that are used follow:

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

The amino acid scoring matrix is more complex. 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 = **BLO**cks **SU**bstitution **M**atrix) 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 below, you can see that 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.

BLOSUM 62 Amino Acid Matrix

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

So why is it called 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.

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
```

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 matchs 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 values are calculated.

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.

As mentioned above, BLAST is actually a collection of algorithms. So when you do a BLAST search you actually need to specify the type of search that you will perform. The following table outlines each algorithm and the nature of the query and database used.

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

Basic **L**ocal **A**lignment **S**earch **T**ool

Alignments

- used to uncover homologies between sequences
- combined with phylogenetic studies
- can determine orthologous and paralogous relationships

Global Alignments

- compares one whole sequence with other entire sequence
- computationally expensive

Local Alignment

- uses a subset of a sequence and attempts to align it to subset of other sequences
- computationally less expensive

Global Alignment Example

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

only a portion of this two sequences can be aligned

Local Alignment Example

a small seed is uncovered

The initial seed for the alignment:

```
           TAT
           |||
AAGCGAATAATATATTTATACTCAGATTATTGCGCG
```

And now the extended alignment:

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

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool

Altschul et al. (1990. J Mol Biol 215:403)

- Set of alignment algorithms
- Use the same search protocol to:
    - Find a short fragment of a query sequence
    - That aligns with a fragment of a subject sequence found in a database

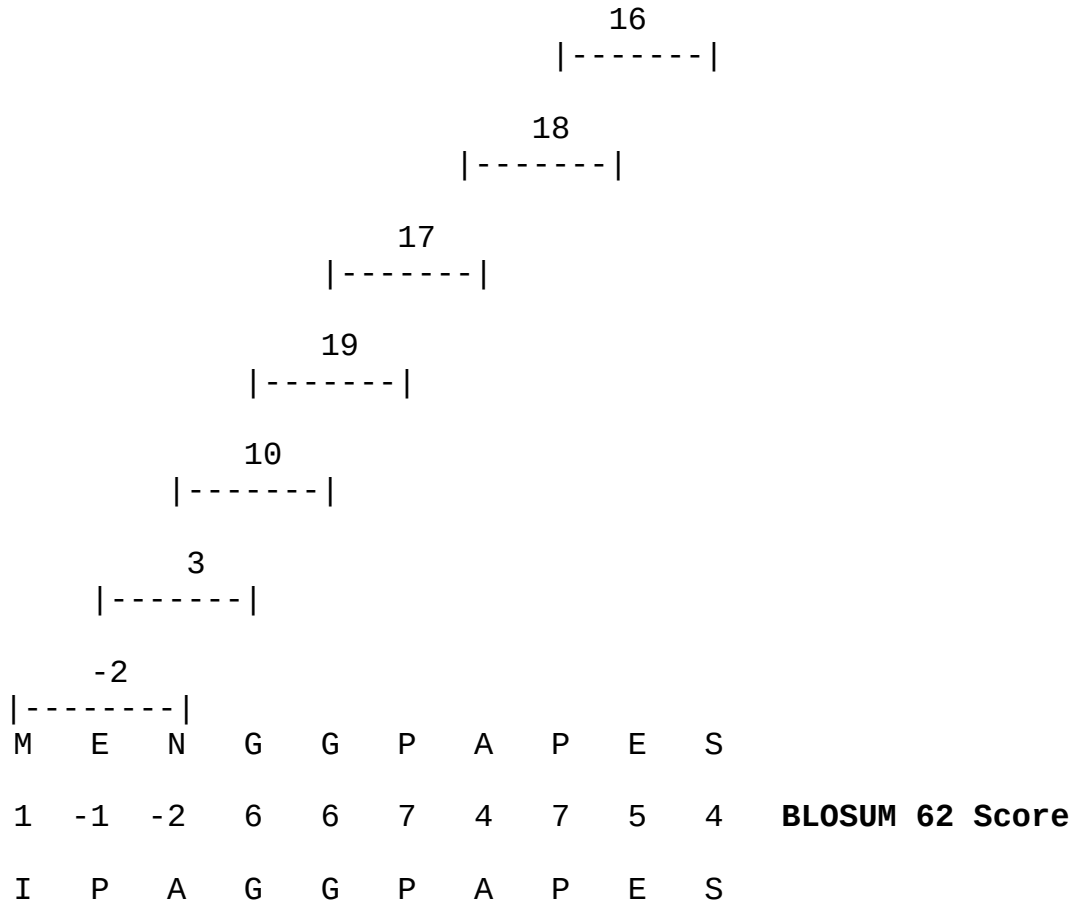
General Concept for Original BLAST Program

- Sequence (query) is broken into words of ***length W***
- Align all words with sequences in the database
- Calculate ***score T*** for each word that aligns with a sequence in the database using a substitution matrix
- Discard words whose T value is below a ***neighborhood score threshold***
- Extend words in both directions until score falls by ***dropoff value X*** when compared to previous best score

BLAST Words

- Three characters in length
- Complied by using a sliding window

Align all words and calculate T score



Build Alignment

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

Points to Remember

- The T score is converted into a bits score by a complicated formula
- The X value is based on the bit score

Search Matrices

BLOSUM 62 matrix. (BLOSUM = BLOcks SUbstitution Matrix)

- Henikoff and Henikoff (1992. PNAS 89:10915-10919)
- Studied 2000 aligned blocks of 500 groups of related proteins
  - Determined the different types of amino acid substitutions that occurred in these proteins
  - Developed the matrix based on the study
  - Positive value
    - Identities or high similarities
  - Negative value
    - Penalty
    - Non similar substitutions

BLOSUM 62 Amino Acid Matrix

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

So why is it called BLOSUM 62?

- Alignment between two amino acid sequences that were 62% identical
- BLOSUM 62 score would be 1

- BLOSUM 45
  - 45% identical
  - distantly related sequences

- BLOSUM 80
  - 80% identity
  - highly conserved sequences

BLAST Statistics

Score (bits)

- A statistical conversion of the score derived by summing using the substitution matrix

Expect (e) Value

- Function of the S value and the database size
- An e value of 1
- One alignment using a query of this size will by chance produce a S score of this value in a database of this size

E value of –10 (=1x10^{-10})

- Unlikely that random chance lead to this current alignment compared to an alignment with an e value of 1
- Often considered to be a probability

NOTE:

- Expect value is specific to a database of a certain size
  - Later searches may give different value
- Why?
  - Database size has changed.

Rules of thumb:

- **E value of –30 or less**
  - Sequences are homologous
- **E values of –5**
    - Often considered significant enough when annotating a genome

BLAST2

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

- Same word search
- Lower T value
- **Neighboring words discovered**
    - Must be at a distance less than A (default 40)
- Alignment extended from the neighboring words

Gap penalties

- New in BLAST2
- **Allow for better alignments**
- Default for amino acid search
- Introducing a gap
    - -11
- Extending that gap
    - -1

BLAST 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

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