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

Local Alignment

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

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

Search Matrices

BLOSUM 62 matrix. (BLOSUM = BLOcks SUbstitution Matrix)

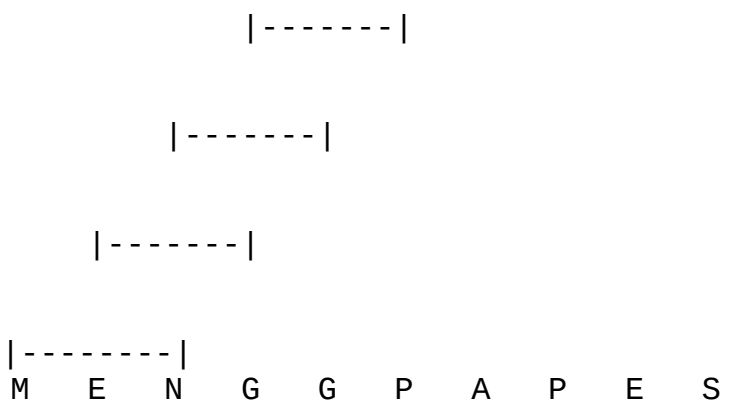
- Henikoff and Henikoff (1992. PNAS 89:10915-10919)
- Studied 2000 aligned blocks of 500 groups of related proteins
  - Determined the different types of amino acid substitutions that occurred in these proteins
  - Developed the matrix based on the study
  - Positive value
    - Identities or high similarities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

BLOSUM 62 Amino Acid Matrix

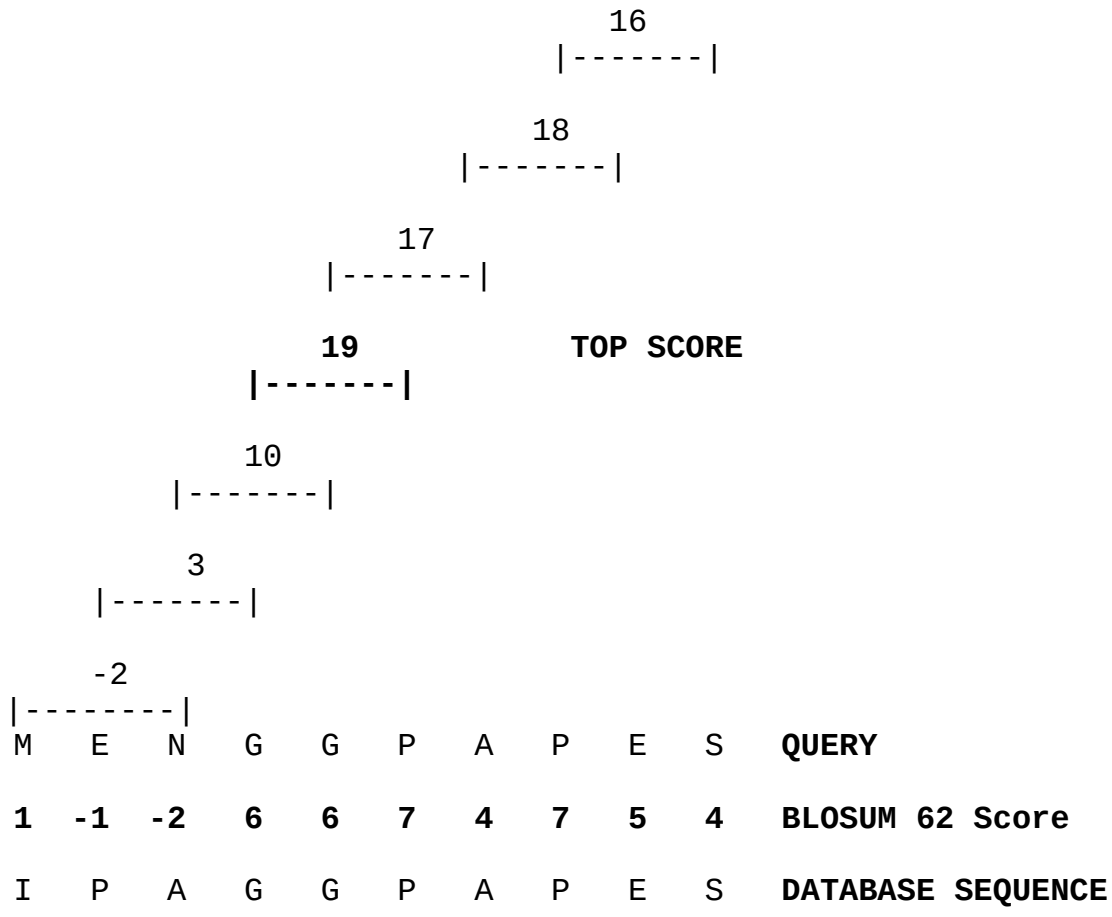
	A	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C		9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D			6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-2	-3	-4	-3
E				5	-3	-2	0	-3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

BLAST Words

- Three characters in length for proteins
- 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

BLAST Statistics

Score (bits)

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

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

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

Homology, Orthology, Parology, Identity, Similarity

How do we define the relationship between two nucleotide or protein sequences?

Homology

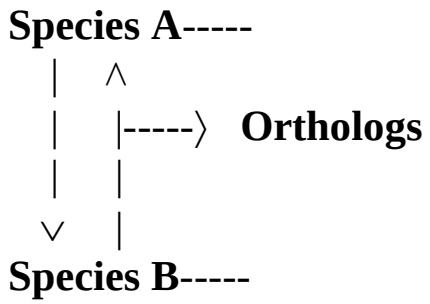
Two sequences are said to show homology if they are identical by descent in a taxonomic lineage or the result of within species duplication

Homologs

Two sequences that exhibit homology

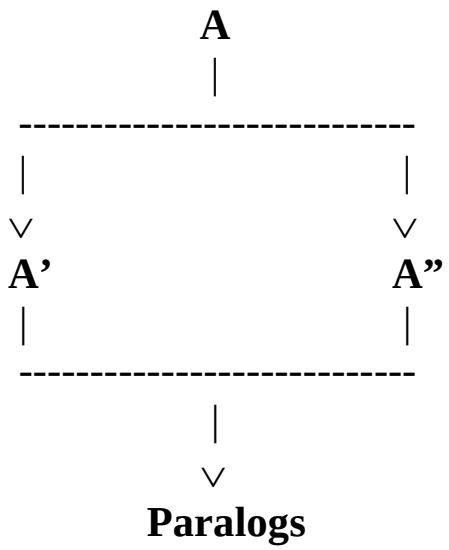
Orthologs

Two sequences related by descent in a species lineage



Paralogs

Two sequences related by a duplication event within a species



Describing the relationship between orthologs and paralogs

Important note

Two sequences are homologous or not homologous; ***there is no percentage of homology***

Identity

The % of exact nucleotide or amino acid matches between to sequences

Similarity

The % of identical or similar amino acids between to protein sequences

Examples of similar amino acids

Valine/Leucine
Valine/Isoleucine
Threonine/Serine