Week 3: Blast Algorithm, theory and practice

Babu Guda, Ph.D.
Department of Genetics, Cell Biology & Anatomy
Bioinformatics and Systems Biology Core
University of Nebraska Medical Center

Database Searching

• Boolean operators
  • AND, OR, NOT
Searching Biological Databases

- Searching is done to find the relatedness between the query and the entries in the database
- For nucleic acids and proteins, the relatedness is defined by “homology”. A ‘Query’ sequence is used to search against each entry, a ‘subject’ in a database
- Two sequences are said to be homologous when they possess sequence identity above a certain threshold
- Thresholds can be defined by length, percentage identity, E-value, Bit-score, r.m.s.d. (for structures), etc., or a combination of one or more of these, depending on the objective of the search
BLAST output

• Pair-wise Alignments

Score = 92.0 bits (63), Expect = 6.3
Identities = 12/34 (35%), Positives = 22/34 (65%), Gaps = 2/34 (6%)

Query: S7 ISLTER1FPAFPKIL--LAVUPOMYNQQV S8
I++P++7 + L + W+7 -LV+-6 +
Sbjct: 34 LAJAPVPKMFGAEVFVIFVIGLGVLYLG 67

Basic Elements in Searching Biological Databases

• Sensitivity versus Specificity/selectivity
• Scoring Scheme, Gap penalties
• Distance/Substitution Matrices (PAM, BLOSUM Series)
• Search Parameters (E-value, Bit score)
• Handling Data Quality Issues (Filtering, Clustering)
• Type of Algorithm (Smith-Waterman, Needleman-Wunsch)
Sensitivity vs. Specificity

- **Sensitivity:**
  - Attempts to report ALL true positives
  - Sensitivity = True Positives / (True Positives + False Negatives)
    \[ \text{Sensitivity} = \frac{TP}{TP + FN} \]
  - (1-sensitivity) gives false negative rate

- **Specificity:**
  - Attempts to report ALL true negatives
  - Specificity = True Negatives / (True Negatives + False Positives)
    \[ \text{Specificity} = \frac{TN}{TN + FP} \]
  - (1-specificity) gives false positive rate

- Known positives are used to test sensitivity and known negatives for specificity
- In database searching, specificity and sensitivity always compete. Depending on the objective of the search, a tradeoff point should be chosen.

Basic Elements in Searching Biological Databases

- Sensitivity versus Specificity/selectivity
- Scoring Scheme, Gap penalties
- Distance/Substitution Matrices (PAM, BLOSSUM Series)
- Search Parameters (E-value, Bit score)
- Handling Data Quality Issues (Filtering, Clustering)
- Type of Algorithm (Smith-Waterman, Needleman-Wunsch)
Scoring Scheme

- Match – Match between identical letters or letters of the same group
- Mismatch – Match between letters of different groups
- Gap – Match between a letter and a gap
- Alignment score is the sum of match, mismatch and gap penalty scores

Say, you are aligning two sequences A and B

<table>
<thead>
<tr>
<th>Sequence A : PQVNTVNRT</th>
<th>Sequence B : PVNRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-Q-V-N-T-V-N-R-T</td>
<td>P-V-N-R-T</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Match 5, Mismatch 2 Gaps 0

<table>
<thead>
<tr>
<th>Scoring Scheme</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match 5, Mismatch 2 Gaps 0</td>
<td>22</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Match 5, Mismatch 2 Gaps -3</td>
<td>7</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Match 5, Mismatch 2 GapI -5, GapE -1</td>
<td>9</td>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>

Gap Penalties

- Proportional gap penalties
  - Penalty $P = na$, where, $n$ is number of gaps and $a$ is gap penalty
- Affine gap penalties
  - Penalty $P = a + mb$, where, $a$ is gap initiation penalty
  - $b$ is gap extension penalty, and $m$ represents the number of extended gaps

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Proportional penalty</th>
<th>Affine penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPAPRAEWVSTVHGSTIEQNEQI</td>
<td>EPAPRAEWVSTVHGSTIEQNEQI</td>
<td></td>
</tr>
<tr>
<td>E--P-A--WV-----ST-E---QI</td>
<td>E--P-A--WV-----ST-E---QI</td>
<td></td>
</tr>
<tr>
<td>Proportional penalty= -65</td>
<td>Proportional penalty= -65</td>
<td></td>
</tr>
<tr>
<td>Affine penalty = (-10)+(-11) = -21</td>
<td>Affine penalty = (-30)+(-7) = -37</td>
<td></td>
</tr>
</tbody>
</table>

- In the proportional scheme, both these alignments score the same, while in the affine penalty scheme, the first one scores much higher

Fall, 2013 GCBA 815
Basic Elements in Searching Biological Databases

- Sensitivity versus Specificity/selectivity
- Scoring Scheme, Gap penalties
- Distance/Substitution Matrices (PAM, BLOSUM Series)
- Search Parameters (E-value, Bit score)
- Handling Data Quality Issues (Filtering, Clustering)
- Type of Algorithm (Smith-Waterman, Needleman-Wunsch)

Terminology in Nucleotide Substitutions

- AGCT: A/G are purines and C/T are pyrimidines
- Transitions vs Transversions
  - Transition: A→G or C→T or vice versa
  - Transversions: A→T or A→C or G→T or G→C or vice versa
- Synonymous vs Non-synonymous substitutions
  - Synonymous: CCC → CCG, both code for ‘Proline’
  - Non-synonymous: UGC → UGG
    Cysteine → Tryptophan
Factors Affecting Amino Acid Replacements

- Each amino acid is coded by a triplet codon
- Each codon can undergo 9 possible single-base substitutions
- So in theory, point mutations in the 61 sense codons can lead to 549 (61 x 9) single-base substitutions
- Of these, 392 are non-synonymous (missense), 134 are synonymous and 23 are nonsense mutations
- Protein substitution matrices are built based on non-synonymous replacements only
- Protein substitution matrices reflect evolutionarily tested and accepted substitutions
• Bias in the frequency of non-synonymous substitutions
  • Structure of the genetic code itself
  • Physical and chemical properties of the amino acid (C, W, M)
  • Variation in the gene copy number or pseudogenes
  • Gene function (Histone has the lowest mutation rate)
  • Species life span (Rodents vs Humans)

• Bias in the frequency of synonymous substitutions
  • Abundance of tRNA species (Rare vs Abundant tRNA species)

Distance/Substitution Matrices

• Unitary matrices/minimum distance matrices
• PAM (Percent Accepted Mutations)
• BLOSUM (BLOcks SUbstitution Matrix)

Terminology
• Global alignment
  • Alignments with the highest score are found at the expense of local similarity
  • Alignments in which the highest scoring subsequences are found anywhere in the alignment
• Local alignment
PAM (Percent Accepted Mutations)

• Developed by Dayhoff and co-workers

• PAM 30, 60, 100, 200, 250

• Built from globally aligned, closely related sequences (85% similarity)

• A database of 1572 changes in 71 groups of closely related proteins

• PAM 1 matrix incorporates amino acid replacements that would be expected if one mutation had occurred per 100 amino acids of sequence i.e., corresponds to roughly one percent divergence in a protein

• Evolutionary tree is built from aligned sequences and relative frequency ($Q_{ij}$) of replacement for each pair of amino acids (i, j) is calculated.

PAM (Percent Accepted Mutations)

• A similarity ratio $R_{ij}$ for each pair is calculated as

$$R_{ij} = \frac{Q_{ij}}{PP_{ij}}$$

where $P_i$ and $P_j$ are observed (natural) frequencies of amino acids $i$ and $j$ in the set of proteins where replacements are counted

• **Substitution score** $S_{ij} = \log_2(R_{ij})$

• $S_{ij} > 0$ means, replacements are favored during evolution

• $S_{ij} < 0$ means, evolutionary selection is against the replacement

<table>
<thead>
<tr>
<th>Real value</th>
<th>Log value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>-1</td>
</tr>
<tr>
<td>0.001</td>
<td>-3</td>
</tr>
</tbody>
</table>

• Average relative frequencies of amino acids in animal proteins

<table>
<thead>
<tr>
<th>A</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>V</th>
<th>W</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.09</td>
<td>1.77</td>
<td>5.22</td>
<td>6.74</td>
<td>3.14</td>
<td>4.71</td>
<td>2.20</td>
<td>4.90</td>
<td>4.99</td>
<td>8.64</td>
<td>2.38</td>
<td>4.23</td>
<td>5.97</td>
<td>4.52</td>
<td>5.92</td>
<td>7.96</td>
<td>5.34</td>
<td>5.95</td>
<td>5.97</td>
<td>2.90</td>
</tr>
</tbody>
</table>

Fall, 2013  GCBA 815
BLOSUM (BLOcks SUBstitution Matrix)

• Developed by Henikoff and Henikoff (1992)
• Blosum 30, 62, 80
• Built from BLOCKS database
• From the most conserved regions of aligned sequences
• 2000 blocks from 500 families
• Blosum 62 is the most popular. Here, 62 means that the sequences used in creating the matrix are at least 62% identical
• Higher Blosum number - Built from closely related sequences
• Lower Blosum number - Built from distant sequences
### Table 2 - The Log odds matrix for BLOSUM 62

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>P</th>
<th>R</th>
<th>K</th>
<th>Q</th>
<th>H</th>
<th>D</th>
<th>E</th>
<th>N</th>
<th>S</th>
<th>T</th>
<th>V</th>
<th>W</th>
<th>Y</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
</tr>
<tr>
<td>C</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>G</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>P</td>
<td>-5</td>
<td>-6</td>
<td>-5</td>
<td>-6</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>R</td>
<td>-5</td>
<td>-5</td>
<td>-5</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>K</td>
<td>-4</td>
<td>-4</td>
<td>-4</td>
<td>-4</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>Q</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>H</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>E</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>N</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>S</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>T</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>V</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>W</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>Y</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>M</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

### Major Differences between PAM and BLOSUM

<table>
<thead>
<tr>
<th>PAM</th>
<th>BLOSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Built from global alignments</td>
<td>Built from local alignments</td>
</tr>
<tr>
<td>Built from small amout of Data</td>
<td>Built from vast amout of Data</td>
</tr>
<tr>
<td>Counting is based on minimum replacement or maximum parsimony</td>
<td>Counting based on groups of related sequences counted as one</td>
</tr>
<tr>
<td>Perform better for finding global alignments and remote homologs</td>
<td>Better for finding local alignments</td>
</tr>
<tr>
<td>Higher PAM series means more divergence</td>
<td>Lower BLOSUM series means more divergence</td>
</tr>
</tbody>
</table>
Basic Elements in Searching Biological Databases

- Sensitivity versus Specificity/selectivity
- Scoring Scheme, Gap penalties
- Distance/Substitution Matrices (PAM, BLOSSUM Series)
- Search Parameters (E-value, Bit score)
- Handling Data Quality Issues (Filtering, Clustering)
- Type of Algorithm (Smith-Waterman, Needleman-Wunsch)

Meaning of E-value in Database Searches

- **E-Value** (Expectation value)
  - The number of equal or higher scores expected at random for a given High Scoring Pair (HSP)
  - E-value of 10 for a match means, in a database of current size, one might expect to see 10 matches with a similar or better score, simply by chance
  - E-value is the most commonly used threshold in database searches. Only those matches with E-values smaller than the set threshold will be reported in the output
  - E-value ranges between 0 to higher, lower the E-value, better the reliability of a match.
Meaning of Bit Score

- **Bit Score**
  - Raw scores have no meaning without the knowledge of the scoring scheme used.
  - Raw scores are normalized to get Bit scores by incorporating information about the scoring scheme used and the search space used (size of database).
  - Bit scores are normalized score and hence it is independent of the size of the database, while E-values are very sensitive to the database size.
  - Generally bit scores of 40 are higher and are considered reliable.

Basic Elements in Searching Biological Databases

- Sensitivity versus Specificity/selectivity
- Scoring Scheme, Gap penalties
- Distance/Substitution Matrices (PAM, BLOSSUM Series)
- Search Parameters (E-value, Bit score)
- Handling Data Quality Issues (Filtering, Clustering)
- Type of Algorithm (Smith-Waterman, Needleman-Wunsch)
Filtering low complexity sequences

• Filters out short repeats and low complexity regions from the query sequences before searching the database
• Filtering helps to obtain statistically significant results and reduce the background noise resulting from matches with repeats and low complexity regions
• The output shows which regions of the query sequence were masked

Filtering low-complex regions

• Low complex regions in protein sequences are those that have repeating residues of the same or a few different residues in a stretch.
• In database searches low complex regions on the query sequence are masked (with X) to avoid random hits and to emphasize on significant hits

Tools used

• SEG
  • NSEG (DNA sequences)
  • PSEG (Protein sequences)
• XNU (For masking repeats <10 in length)
• DUST (DNA sequences)
• CAST (Proteins)
Basic Elements in Searching Biological Databases

- Sensitivity versus Specificity/selectivity
- Scoring Scheme, Gap penalties
- Distance/Substitution Matrices (PAM, BLOSSUM Series)
- Search Parameters (E-value, Bit score)
- Handling Data Quality Issues (Filtering, Clustering)
- Type of Algorithm (Smith-Waterman, Needleman-Wunsch)

Choice of the Searching Algorithm

An ideal algorithm should have

- Good specificity and sensitivity
- Should be fast running
- Should not use too much memory

Greedy algorithms are very sensitive, but very slow. Heuristic algorithms are relatively fast, but loose some sensitivity. It’s always a challenge for a programmer to develop algorithms that fulfill both of these requirements
  • Very greedy algorithm, so very sensitive
  • Implements Dynamic programming
  • Provides global alignment between the two sequences

  • A set of heuristics were applied to the above algorithm to make it less greedy, so it is less sensitive but runs faster
  • Implements Dynamic programming
  • Provide local alignment between two sequences
  • Both BLAST and FASTA use this algorithm with varying heuristics applied in each case

FASTA (FAST Algorithm)
  • The fist step is application of heuristics and the second step is using dynamic programming
    • First, the query sequence and the database sequence are cut into defined length words and a word matching is performed in all-to-all combinations
    • Word size is 2 for proteins and 6 for nucleic acids
    • If the initial score is above a threshold, the second score is computed by joining fragments and using gaps of less than some maximum length
    • If this second score is above some threshold, Smith-Waterman alignment is performed within the regions of high identities (known as high-scoring pairs)
Protein and nucleotide substitution matrices

BLAST (Basic Local Alignment Search Tool)

- The first step is application of heuristics and the second step is using dynamic programming
  - First, the query sequence and the database sequence are cut into defined length words and a word matching is performed in all combinations
  - Words that score above a threshold are used to extend the word list

Expanded list - 47 words

<table>
<thead>
<tr>
<th>Word</th>
<th>Expanded List</th>
</tr>
</thead>
<tbody>
<tr>
<td>ql</td>
<td>ql, qm, hl</td>
</tr>
<tr>
<td>ln</td>
<td>ln</td>
</tr>
<tr>
<td>nf</td>
<td>nf, af, ny, df, qf, ef, gf, hf, kf, sf, tf</td>
</tr>
<tr>
<td>fs</td>
<td>fs, fa, fn, fd, fg, fp, ft, ys</td>
</tr>
<tr>
<td>gw</td>
<td>gw, aw, rw, nw, dw, qw, ew, hw, iw, kw, mw, pw, sw, tw, vw</td>
</tr>
</tbody>
</table>

Creating a Word List for BLAST, Word Size = 2

Adiposona borealis L. - Nigrotylocoelia
BLAST continued ...

- BLAST is a local alignment algorithm
- Several High Scoring Segments are found, with the maximum scoring segment used to define a band in the path graph
- Smith-Waterman algorithm is performed on several possible segments to obtain optimal alignment
- The word size for Protein is 3 and for Nucleic acid is 11
Comparison of BLAST and FASTA

- BLAST uses an expanded list to compensate for the loss of sensitivity from increased word size.
- BLAST is more sensitive than FASTA for protein searches while FASTA is more sensitive than BLAST for nucleic acid searches.
- Both BLAST and FASTA run faster than the original Needleman-Wauch algorithm at the cost of loss of sensitivity.
- Both algorithms fail to find optimal alignments that fall outside of the defined band width.

Essential Elements of an Alignment Algorithm

- Defining the problem (Global, semi-global, local alignment).
- Scoring scheme (Gap penalties).
- Distance Matrix (PAM, BLOSUM series).
- Scoring/Target function (How scores are calculated).
- Good programming language to test the algorithm.
Types of Alignments

- **Global** - When two sequences are of approximately equal length. Here, the goal is to obtain maximum score by completely aligning them.

- **Semi-global** - When one sequence matches towards one end of the other.
  - Ex. Searches for 5’ or 3’ regulatory sequences.

- **Local** - When one sequence is a sub-string of the other or the goal is to get maximum local score.
  - Protein motif searches in a database.

Local vs Global alignments

![Local vs Global alignments diagram](image-url)
Pair-wise sequence comparison

Different Alignment Paths

Sequence T

S

T

S

T

S
Whole genome sequence comparison

Scoring System for Alignments

**Scoring Weights**
- Matches +10 - These are arbitrary values, but the real values come from distance matrices (PAM, BLOSUM etc)
- Mismatches +4

**Gap Penalties**
- Gap Initiation -10 - Arbitrarily chosen but, optimized
- Gap Extension -2 for a particular Distance matrix

**Rules of Thumb for Affine Gap Penalties**
- Gap Initiation Penalty should be 2 to 3 times the largest negative score in a distance matrix table
- Gap Extension Penalty should be 0.3 to 0.1 times the Initiation Penalty
Similarity Score

- Similarity score is the sum of pair-wise scores for matches/mismatches and gap penalties

Scoring Weights

- Matches: +10
- Mismatches: +4
- Gap Initiation: -10
- Gap Extension: -2

Global Alignment
EPSGFPAWVSTVHGQEQI
E----PAWVST-----QI
Score: (9 x 10) + (0 x 4) - (2 x -10 + 7 x -2)
Score = 90 - 34 = 56

Scoring/Target function

- The scoring function calculates the similarity score. The goal of the algorithm is to maximize similarity score

Global Alignment
EPSGFPAWVSTVHGQEQI
E----PAWVST-----QI
Score: (9 x 10) + (2 x -10 + 7 x -2)
Score = 90 + (-34) = 56

Scoring Weights

- Matches: +10
- Mismatches: +4
- Gap Initiation: -10
- Gap Extension: -2

Local Alignment
EPSGFPAWVSTVHGQEQI
E----PAWVST-----QI
Score: (6 x 10) + (2 x 0 + 10 x 0)
Score = 60 + 0 = 60

Scoring Weights

- Matches: +10
- Mismatches: +4
- Gap Initiation: -10
- Gap Extension: -2
- No terminal gap penalty
### Different types of BLAST programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>nucleotide</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>blastp</td>
<td>protein</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>blastx</td>
<td>nucleotide</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>tblastn</td>
<td>protein</td>
<td>nucleotide</td>
<td>protein</td>
</tr>
<tr>
<td>tblastx</td>
<td>nucleotide</td>
<td>nucleotide</td>
<td>protein</td>
</tr>
<tr>
<td>megablast</td>
<td>nucleotide</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>PHI-Blast</td>
<td>protein</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>PSI-Blast</td>
<td>Protein-Profile</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>RPS-Blast</td>
<td>protein</td>
<td>Profiles</td>
<td>protein</td>
</tr>
</tbody>
</table>