G4120: Introduction to Computational Biology

Oliver Jovanovic, Ph.D.
Columbia University
Department of Microbiology

Copyright © 2007 Oliver Jovanovic, All Rights Reserved.
Algorithm
- An algorithm is simply a series of steps used to solve a problem. One of a computer’s great strengths is its ability to rapidly and accurately repeat recursive steps in an algorithm.

Consensus
- Early algorithms for searching sequence data depended on consensus sequences. Thus, to find a prokaryotic promoter, one would try to find something that matched a consensus -10 sequence (TATAAT), not too far downstream of a consensus -35 sequence (TTGACA).
- It rapidly became clear that biologically significant sequences rarely perfectly matched a consensus, and more sophisticated approaches were adopted, including the use of matrices, Markov chains and hidden Markov models.

Matrices
- Matrices take into account the distribution of every possible nucleotide (or amino acid) at a position in a set of known sequences. Searching with a matrix is therefore more sensitive than searching with a consensus, and can find biological features that a strict consensus approach would miss.

Markov chains and hidden Markov models (HMMs)
- Markov chains and hidden Markov models are probabilistic models of sequences, and have proven useful in database searching, gene finding and multiple sequence alignment.
- A first-order Markov chain is a finite state automaton (a restricted Turing machine which only moves left to right) with probabilities for each transition to a new state (symbol) based on its current state. Higher order Markov chains take into account one or more previous states.
- A hidden Markov model is a Markov chain in which only the output can be observed (its current state is hidden).
## E. coli Promoter Consensus

<table>
<thead>
<tr>
<th>-35 Region</th>
<th>-10 Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTGACA.................TATAAT</td>
<td></td>
</tr>
</tbody>
</table>

## E. coli Promoter Matrix

### -35 Region

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>T</th>
<th>G</th>
<th>A</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

### Spacer Region

<table>
<thead>
<tr>
<th>Length</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>6</td>
<td>14</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### -10 Region

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>A</th>
<th>T</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>T</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>
Global Alignment Search

- **Needleman-Wunsch** algorithm, Needleman & Wunsch, 1970
- Finds the best complete alignment of two sequences that maximizes the number of matches and minimizes the number of gaps.

Local Alignment Search

- **Smith-Waterman** algorithm, Smith & Waterman, 1981
- Makes an optimal alignment of the best segment of similarity between two sequences.
- Often better for comparing sequences of different lengths, or when looking at a particular region of interest.

Heuristic Approximations to Smith-Waterman

- **FASTA**, Pearson, 1988
- **BLAST**, Altschul, 1990
- **BLAST 2** (aka Gapped BLAST), Altschul, 1997
- **BLAT**, Kent, 2002
Global Alignment (Needleman-Wunsch)

Gap, from the GCG Wisconsin Package, uses the algorithm of Needleman and Wunsch to find the alignment of two complete sequences that maximizes the number of matches and minimizes the number of gaps.

GAP RK2_ssb x Ecoli_ssb January 29, 2003 00:07

1 ..MSHNQFQFGNLTRDTEVRHGNSNKPFMAIFDIAVNEWRNDAGDKQE 47

Matrix: blosum62
Gap Penalties: default
Length: 177
Percent Similarity: 45.690
Percent Identity: 32.759
Local Alignment
(Smith-Waterman)

**BestFit**, from the GCG Wisconsin Package, makes an optimal alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maximize the number of matches using the local homology algorithm of Smith and Waterman.

```
BESTFIT RK2_ssb x Ecoli_ssb       January 29, 2003 00:08

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4   NQFQFIGNLTRDTEVRHGNNSNKPQAIFDIAVNEEWRNDA.GDKQERTDF</td>
<td>52   NKFVILVGNLQDPEVRYMPNGAVANITLATSESWRDKATGEMKQTEWH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6   NKFVILVGNLQDPEVRYMPNGAVANITLATSESWRDKATGEMKQTEWH</td>
<td>55   RIKCFGSQAEAHGKYLGKYGSLFVQGKIRNTKY.EKDGQTVMGTDFIAD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53  RIKCFGSQAEAHGKYLGKYGSLFVQGKIRNTKY.EKDGQTVMGTDFIAD</td>
<td>100  RIKCFGSQAEAHGKYLGKYGSLFVQGKIRNTKY.EKDGQTVMGTDFIAD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56  RIKCFGSQAEAHGKYLGKYGSLFVQGKIRNTKY.EKDGQTVMGTDFIAD</td>
<td>104  RIKCFGSQAEAHGKYLGKYGSLFVQGKIRNTKY.EKDGQTVMGTDFIAD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Matrix:  blosum62
Gap Penalties:  default
Length:  99
Percent Similarity:  50.515
Percent Identity:  36.082
```
Global (Needleman-Wunsch) versus
Local (Smith-Waterman) Alignment

1  ..MSHNQFQFIGNLTRDTEVRHGNSNKPAIFDIAVNEEWNRDA.GDKQE  47
1  ASRGVNKVIILVGLGQDPEVRYMPNGGAVANITLATESWRDKatGEMKE  50

48  RTDFFIKCFSQAEAHGKYLGKSLVFVQGKIRNTKY.EKDQTVYGTD  96
51  QTEWHRVVLFGKLAEVEYLRKGSQVYIEQQLRTRKWTQGSGQRDRYUTE  100

97  FIAD...KVDYLDTKAPGGSNQE.......................... 116
  : . . | : ||
101  VVNVGGTMQMLGGRQGGGAPAGNNIGGGQPGGWQPQPQQPSGNGQFSGG  150

..................

151  AQSRRPQQSAPAAPSNEPPMDFDIIPF  177
Optimal versus Heuristic Algorithms

**Optimal Algorithm**

- Finds the optimal solution to a problem.
- Often uses an approach called dynamic programming, which solves the problem by breaking it into smaller subproblems, which are separately solved, then sequentially reassembled to solve the entire problem.
- This approach was first applied to solving biological sequence comparison problems by Saul Needleman and Christian Wunsch in 1970, and can be used to solve either a global comparison problem (Needleman-Wunsch) or a local comparison problem (Smith-Waterman).

**Heuristic Algorithm**

- Solves a problem by using rules of thumb to reach a solution. The solution is not guaranteed to be an optimal solution, but is generally arrived at far faster than using an optimal solution approach such as dynamic programming.
- In sequence comparison, well known heuristic approaches include the BLAST algorithm, developed by Stephen Altschul in 1990, and the FASTA algorithm, developed by William Pearson in 1988.
- Although a heuristic algorithm such as BLAST may be 100 to 1,000 times faster than an optimal algorithm such as Smith-Waterman, it may miss matches found by the optimal algorithm.
BLAST (Heuristic) versus Smith-Waterman (Optimal)
Gaps

- When scoring an alignment, penalties are assigned for creating and extending gaps. The longer the gap, the greater the penalty.
- Can vary the gap penalties. This will likely change your results, so take note.

Dayhoff Substitution Matrices

- For protein comparison, all of these search algorithms use Dayhoff substitution matrices which encode likelihoods of an amino acid substitution.
- When scoring an alignment, penalties are assigned for what the substitution matrix considers poor substitutions, the worse the substitution, the greater the penalty.
- Can get different results depending on which substitution matrix you use.

Blosum 62

Blosum 62 is the default matrix for NCBI BLAST protein comparison. It is optimized for known close homologies.

PAM 250

PAM 250 is a matrix which is optimized for known distant homologies.
Blosum 62 (Close) versus PAM 250 (Distant)
NCBI


- PubMed, PubMed Central (over 1.1 million articles), Books and other reference material
- GenBank, RefSeq, CDD, MMDB and other sequence and structure databases
- Over 130 billion base pairs (130 gigabases) from over 240,000 organisms
- Prokaryotic genome data and browsers (over 600 microbial genomes, over 2,500 viruses, over 1,100 plasmids, environmental samples and additional sequences)
- Eukaryotic genome data and browsers (24 complete eukaryotic genomes, 400 additional eukaryotic genomes in progress, maps and partial sequences)
- BLAST, PSI-BLAST and VAST search tools

NCBI BLAST


BLAST for Beginners

http://www.geospiza.com/outreach/BLAST/index.html
http://www.geospiza.com/outreach/organelles/index.html
http://www.geospiza.com/outreach/snp_hunting/index.html

NCBI BLAST Tutorial

PubMed


EndNote and PubMed

• EndNote can directly connect to PubMed and search and retrieve references from it:
  Edit > Connection Files > Open Connection Manager... > PubMed (NLM).enz > click in Favorites and close window, then Tools > Connect > PubMed (NLM).enz

• Alternately, the PubMed Clipboard (use the entrez URL above for PubMed) can be used to collect references, export them as a text file in MEDLINE format, then import them into EndNote:
  In PubMed, select Clipboard from Send to > Clipboard tab > select Display MEDLINE and File from Send to
  In EndNote, Edit > Import Filters > Open Filter Manager... > PubMed (NLM).enf > click in Favorites and close window, then with a library open, File > Import... > Import Options: PubMed (NLM).enf > select file to import > Import (the MEDLINE format PubMed text file will typically be named pubmed-result.txt and be saved on your Desktop)

EndNote and PDF Files

• Can also use EndNote to organize PDF files, which are otherwise easy to lose track of:
  With a reference open, References > File Attachments > Link to File inserts a link to the PDF file you select. Use References > File Attachments > Open File to open the linked PDF (or double click the file in the File Attachments field). To remove, just select and delete the linked file.
Bioinformatics Resources

Books

*Fundamental Concepts of Bioinformatics* by Dan E. Krane & Michael L. Rayme
*Developing Bioinformatics Computer Skills* by Cynthia Gibas & Per Jambeck
*BLAST: An Essential Guide to the BASIC Local Alignment Search Tool* by Ian Korf, Mark Yandell & Joseph Bedell
*Sequence Analysis in a Nutshell: A Guide to Common Tools and Databases* by Scott Markel & Darryl Leon
*Introduction to Computational Biology: Maps, Sequences and Genomes* by Michael S. Waterman
*Bioinformatics: Sequence and Genome Analysis* by David W. Mount
*Biological Sequence Analysis* by Richard Durbin, et al.

Journals

*Bioinformatics* (formerly CABIOS)
*Journal of Computational Biology*
*Comparative and Functional Genomics*
*Biotechnology Software & Internet Journal*
*Nucleic Acids Research* (first annual web-server issue at [http://nar.oupjournals.org/content/vol31/issue13/](http://nar.oupjournals.org/content/vol31/issue13/))

Websites

[http://www.sanger.ac.uk](http://www.sanger.ac.uk) Wellcome Trust Sanger Institute
[http://www.ebi.ac.uk](http://www.ebi.ac.uk) European Bioinformatics Institute
[http://www.google.com](http://www.google.com) Google