Pairwise Sequence Alignment and Database Search

Zhongming Zhao, PhD
Email: zhongming.zhao@vanderbilt.edu
Sequence Similarity

- **match**
- **mismatch**
- **gap**
Why make sequence alignments?

1. The sequences *may* share a common origin - a common ancestor sequence. If the similarity is sufficiently convincing or if we have additional evidence for an evolutionary relationship, then we say that the sequences are homologous.

2. The sequences *may* have the same or related structure and function.

3. The difference in the alignments may be linked to the functional changes/diseases.
Approaches in Pairwise Sequence Alignment

1. Dot Matrix
2. Global Alignment
3. Local Alignment
**Visualization: Dot matrix**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dot matrix II
Dot matrix III
Dot matrix -- *Staphylococcus epidermidis* RP62A and ATCC12228
Alignment -- *Staphylococcus epidermidis* RP62A and ATCC12228
A high-quality alignment?

- For DNA sequences
  - Long runs of identity
  - Few gaps in the aligned regions
  - An overall high degree of identity (>80%)

- For protein sequences
  - Includes most of each sequence
  - A significant proportion of identities throughout the alignment
  - Multiple examples of conservative substitutions
  - Relatively few gaps
  - 50% is very good
The alternative pathways that could form the maximum match are illustrated. The maximum match terminates at the largest number in the first row or first column, 8 in this case.
Local Alignment: Smith-Waterman Algorithm (1981)

Match: 1.0
Mismatch: -1/3
Gap $w_k = 1.0 + 1/3 \times k$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>C</th>
<th>U</th>
<th>C</th>
<th>G</th>
<th>C</th>
<th>U</th>
<th>U</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>0.3</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.7</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>0.7</td>
<td>1.0</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
<td>1.3</td>
<td>0.3</td>
<td>1.0</td>
<td>0.3</td>
<td>2.0</td>
<td>0.7</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>1.0</td>
<td>0.7</td>
<td>0.0</td>
<td>3.0</td>
<td>1.7</td>
<td>1.3</td>
<td>1.0</td>
<td>1.3</td>
<td>1.7</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
<td>0.7</td>
<td>0.3</td>
<td>1.7</td>
<td>2.7</td>
<td>1.3</td>
<td>1.0</td>
<td>0.7</td>
<td>1.0</td>
<td>1.3</td>
<td>1.3</td>
<td>0.0</td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.7</td>
<td>0.3</td>
<td>1.3</td>
<td>2.7</td>
<td>2.3</td>
<td>1.0</td>
<td>0.7</td>
<td>1.7</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.3</td>
<td>1.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.0</td>
<td>0.7</td>
<td>1.7</td>
<td>2.7</td>
<td>1.7</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.3</td>
<td>0.0</td>
<td>1.0</td>
<td>2.0</td>
<td>3.3</td>
<td>2.0</td>
<td>1.7</td>
<td>1.3</td>
<td>2.3</td>
<td>2.7</td>
<td>2.7</td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.7</td>
<td>0.7</td>
<td>2.0</td>
<td>3.0</td>
<td>1.7</td>
<td>1.3</td>
<td>2.3</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>C</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.0</td>
<td>0.7</td>
<td>1.7</td>
<td>1.7</td>
<td>3.0</td>
<td>2.7</td>
<td>1.3</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>1.0</td>
<td>0.3</td>
<td>0.7</td>
<td>1.7</td>
<td>0.3</td>
<td>2.7</td>
<td>1.7</td>
<td>2.7</td>
<td>2.3</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>G</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.3</td>
<td>1.3</td>
<td>1.3</td>
<td>2.3</td>
<td>1.3</td>
<td>2.3</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Smith-Waterman Algorithm vs Needleman-Wunsch Algorithm

A. needle (Needleman Wunsch) global alignment.
Percent Identity: 75/238 (31.5%)
Percent Similarity: 118/238 (49.6%)

B. water (Waterman Smith) local alignment.
Percent Identity: 60/125 (48.0%)
Percent similarity: 82/125 (65.6%)
Database Searching

- Similarity searches in sequence databases have become a mainstay of bioinformatics.

- A sequence by itself is not information. Comparison can help find the important biological information, e.g. function of unknown genes, structure of query sequences, duplicated genes.

- Similar scores: allowing substitutions or residues with similar characteristics (e.g. BLOSUM62, PAM250).

- Two programs, which greatly facilitated the similarity search, were developed: FASTA (Pearson and Lipman 1988) and BLAST (Altschul et al. 1990). Many programs have been further developed from them.

- Sequence databases, e.g. NCBI.
Basic Local Alignment Search Tool (BLAST)

- Basic Local Alignment Search Tool (BLAST) was developed as a new way to perform sequence similarity search.
- It is a string pattern search.
What BLAST Tells You

- BLAST reports surprising alignments
  - Different than chance

Assumptions
- Random sequences
- Constant composition

Conclusions
- Surprising similarities imply evolutionary homology
Basic Local Alignment Search Tool (BLAST)

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database.
  - DNA vs DNA (BLASTN)
  - DNA translation vs Protein (BLASTX)
  - Protein vs Protein (BLASTP)
  - Protein vs DNA translation (TBLASTN)
  - DNA translation vs DNA translation (TBLASTX)
- www, standalone, and network clients
BLAST’s Short Cut: Word Hits

Query: \textbf{GTACTGGACATGGGACCTACAGGAA}

Word Size = 11

\textbf{GTACTGGACAT}

Minimum word size = 7

\textbf{TACTGGACATG}

blastn default = 11

\textbf{ACTGGACATGG}

megablast default = 28

\textbf{CTGGACATGGGA}

Make a lookup table of words

\textbf{TGGACATGGGAC}

\textbf{GGACATGGGACC}

\textbf{GACATGGGACCC}

\textbf{ACATGGGACCCT}
Online BLAST Search

BLAST Assembled RefSeq Genomes
Choose a species genome to search, or list all genomic BLAST databases.
- Human
- Mouse
- Rat
- Arabidopsis thaliana
- Oryza sativa
- Bos taurus
- Danio rerio
- Drosophila melanogaster
- Gallus gallus
- Pan troglodytes
- Microbes
- Apis mellifera

Basic BLAST
Choose a BLAST program to run.
- Nucleotide blast
  - Search a nucleotide database using a nucleotide query
    - Algorithms: blastn, megablast, discontinuous megablast
  - Search protein database using a protein query
    - Algorithms: blastp, psi-blast, db-blast
- Protein blast
  - Search protein database using a translated nucleotide query
  - Search translated nucleotide database using a protein query
- Blast
  - Search translated nucleotide database using a translated nucleotide query

Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)
- Make specific primers with Primer BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdat)
- Search sequences that have gene expression profiles (GED)
- Search immunoglobulins (igBLAST)
- Search using SNP Tracks
- Screen sequence for vector contamination (vecscreen)
>gi|127552|sp|P23367|MUTL_ECOLI  DNA mismatch repair protein mutL
Length = 615

Score = 42.0 bits (97),  Expect = 3e-04
Identities = 26/59 (44%), Positives = 33/59 (55%), Gaps = 9/59 (15%)

Query   9    LPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILEV-QQHIESKL  58
         +    L   +    P   L LEI P  VDVNVHP KHEV F   +H+   +    +L V QQ   +E+ L
Sbjct  280  LGADQQPAFVLYLEIDPHQVDVNHFAKHEVRFHQSRLVHDFIYQGVLSVLOQQLETPL  338

Identical match
positive score (conservative)
negative substitution
gap

From NCBI training tutorial
Exercise

• Perform Blast search of the following sequence.
  • In which gene?
  • In the coding region?

• Translate it into aa sequence, and perform Blastp search

GGCCGTGCCT GGGGATCCAA GTTCCCCTCT CTCCACCTGT GCTCACCTCT CCTCCGTCCC CAAACCCTGCA CAGGCAAGAT CGTGGACGCC GTGATTCAAG AGCACCAGCC CTCCGTGCTG CTGGAGCTGG GGGCCTACTG TGGCTACTCA GCTGTGCGCA TGGCCGCCT GCTGTCACCA GGGGCGAGGC TCATCACCAT CGAGATCAAC CCCGACTGTG CCGCCATCAC CCAGCGGATG GTGGATTTCG CTGGC