Bioinformatics
Questions
– For some gene, are there similar genes in a given genome?
– Given a piece of a protein structure, what other proteins have similar chunks?
– For some drug or metabolite, what proteins might bind it?
– Where is a gene located in a genome?

Target problems

Methods
Databases can tell us something about unknown sequences
Bioinformatics

• Questions

• Target problems
  – Drug discovery
  – Drawing phylogenetic trees
  – Understanding biological mechanism

• Methods
Bioinformatics

• Questions
• Target problems
• Methods
  – Search for strings
  – Compare two strings
  – Align strings
  – Analyze swap frequencies and conserved positions
Understanding protein evolution can help us compare sequences

• Types of mutations
  
  – Insertion
    
    -xxAxx- (mother)
    -xxABxx- (daughter)
  
  – Deletion
    
    -xxAxx- (mother)
    -xxxx- (daughter)
  
  – Substitution
    
    -xxAxx- (mother)
    -xxBxx- (daughter)
Sequence comparison

• Which pair of sequences is more closely related?
  – Depends on scoring function
  – Similarity vs. homology
  – Ambiguity in mechanism, parsimony

I AM MOM (mother)
I AM TOM (daughter)
(1 substitution)

I AM MOM (mother)
I _ M MORT (daughter)
(1 deletion, 1 insertion, 1 substitution)
Similarity scoring

- Score components
  - Match/mismatch (Hamming distance, etc.)
  - Gap penalties (indels)
  - Mutation weights may depend on physical properties or substitution frequencies

- Substitution matrices
  - PAM, BLOSUM, etc.

- Total score is a sum over all positions
Substitution matrices

- How often is residue x swapped with residue y?

![BLOSUM62 matrix](image)

\[
\text{LogOddRatio} = 2 \log_2 \left( \frac{P(O)}{P(E)} \right)
\]
Sequence conservation can be described by entropy

- Different monomer preferences at different positions

\[
\begin{align*}
\text{(1) Different monomer preferences at different seq posns.} \\
\text{ACCBCD...z } \{ \text{a sequence}\} \\
\text{Look at position } j. \text{ Frequencies of monomer types?} \\
\text{Pi} \\
\text{ala gly trp ... amino acid type } i
\end{align*}
\]
Sequence conservation can be described by entropy

• Different monomer preferences at different positions

• Sequence entropy at position $j$
  – Kullback-Leibler divergence
    $$ S = -\sum_i p_i \ln(p_i/p_j) $$

(3) $S = 0$ if $\ldots$ \{Highly conserved\}

$S = \ln n$ if $\ldots$ \{Not conserved\}
Sequence conservation can be described by entropy
Sequence conservation is affected by physical and biological factors

- The location, structural relevance, and function of residues all contribute to evolutionary conservation
Sequence conservation is affected by physical and biological factors

- The location, structural relevance, and function of residues all contribute to evolutionary conservation
Sequence conservation is affected by physical and biological factors

- The location, structural relevance, and function of residues all contribute to evolutionary conservation
Sequence alignments

- **Global alignment**
  - Needleman-Wunsch algorithm

- **Local alignment**
  - Smith Waterman algorithm
  - BLAST

- **Multiple sequence alignments**
  - PSI-BLAST
  - CLUSTALW
Needleman-Wunsch

• Align two sequences, GCATGCU ; GATTACA
• Scoring method
  – Match: +1
  – Mismatch: -1
  – Gap: -1
• Align
  – Make a matrix
  – Start at top left, score first row and column
  – Iterate over rows and columns. The score for position (i,j) is the max of the score from (i-1,j), (i-1,j-1), and (i,j-1) using the scores given above
  – Traceback to find alignment
### Needleman-Wunsch

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</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>
</tr>
<tr>
<td>A</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>
</tr>
<tr>
<td>T</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>
</tr>
<tr>
<td>C</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>
</tr>
</tbody>
</table>
Needleman-Wunsch

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-8</td>
</tr>
<tr>
<td>A</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>
</tr>
<tr>
<td>T</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>
</tr>
<tr>
<td>C</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>
</tr>
</tbody>
</table>
## Needleman-Wunsch

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Needleman-Wunsch

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Needleman-Wunsch

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td></td>
<td></td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Needleman-Wunsch

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Needleman-Wunsch

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>T</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>(-2,-1,0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-5</td>
<td>-3</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-6</td>
<td>-4</td>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-7</td>
<td>-5</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Needleman-Wunsch

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>T</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-5</td>
<td>-3</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-6</td>
<td>-4</td>
<td>-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-7</td>
<td>-5</td>
<td>-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Needleman-Wunsch

\[
\begin{array}{|c|c|c|c|c|c|c|c|}
\hline
 & G & C & A & T & G & C & U \\
\hline
0 & -1 & -2 & -3 & -4 & -5 & -6 & -7 \\
\hline
G & -1 & 1 & 0 & -1 & -2 & -3 & -4 & -5 \\
\hline
A & -2 & 0 & 0 & 1 & 0 & -1 & -2 & -3 \\
\hline
T & -3 & -1 & -1 & 0 & 2 & 1 & 0 & -1 \\
\hline
T & -4 & -2 & -2 & -1 & 1 & 1 & 0 & -1 \\
\hline
A & -5 & -3 & -3 & -1 & 0 & 0 & 0 & -1 \\
\hline
C & -6 & -4 & -2 & -2 & -1 & -1 & 1 & 0 \\
\hline
A & -7 & -5 & -3 & -1 & -2 & -2 & 0 & 0 \\
\hline
\end{array}
\]
Needleman-Wunsch traceback

• Traceback finds the optimal alignment(s)

• Start at bottom right

• Move up and left through matrix, move along path of highest score
Needleman-Wunsch traceback
Needleman-Wunsch traceback

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>G</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>T</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>-5</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>-6</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>-7</td>
<td>-5</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
</tr>
</tbody>
</table>

Best alignments:
- G–ATTACA GCATG–CU
- G–ATTACA GCA–TGCU
- G–ATTACA GCAT–GCU

Best alignments
Local alignments using BLAST

• BLAST is a fast tool used to search large databases for alignments to your query sequence

• Break QUERY into 3-letter words
  • QUE, UER, ERY

<table>
<thead>
<tr>
<th>Your query sequence:</th>
<th>A database:</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUERY</td>
<td>{ABC ... XYZ</td>
</tr>
<tr>
<td></td>
<td>CZD UER ... PL</td>
</tr>
<tr>
<td></td>
<td>XERY ... MNO</td>
</tr>
</tbody>
</table>
Local alignments using BLAST

- BLAST is a fast tool used to search large databases for alignments to your query sequence
- Break QUERY into 3-letter words
  - QUE, UER, ERY
- Search the database for these words
- Score the matches
- Extend ‘seed’ sequences into high-scoring sequence pairs
BLAST options

- Database (GenBank, PDB, UnkProtKB, etc.)
- What organisms?
- Threshold value
- Other parameters (scoring parameters, etc.)
BLAST

• Input FASTA format

• Alignments and scores
Drawing phylogenetic trees

- Collect distances among sequences
- Locate shortest distance
Drawing phylogenetic trees

• Connect closest sequences

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

![Phylogenetic tree diagram]
Drawing phylogenetic trees

- Remove A and D from matrix, add the midpoint a
- Repeat
Drawing phylogenetic trees

- Remove $a$ and $E$ from matrix, add the midpoint $b$
- Repeat
Drawing phylogenetic trees

• Connect final branches using full matrix
Protein structure prediction

• Why?
  – Understanding biological mechanism
  – Drug discovery

• Approaches (Sample, Score)
  – Search PDB for sequences
    • Align
    • Generate new configurations
    • Score
  – Sample using molecular dynamics or other physics-based methods
    • Analyze populations
  – Test in CASP (Critical Assessment of protein Structure Prediction)
Tools

• Sequence similarities
  – BLAST
  – Psi-blast
  – HHblits

• Secondary structure servers
  – Psipred
  – Neural net

• 3D structures
  – Modeller

• Structure prediction
  – ROSETTA
  – MELD
  – Brute force
Statistical potentials

• Infer energy-like quantities from databases
• Observe a probability distribution from a database

- Secondary structures
- Bond angles
- Sidechain conformations
- Amino acid contacts

• Turn it into a pseudo-energy:

\[
\frac{\Delta E}{RT} = -\ln\left(\frac{P_i}{P_0}\right)
\]

• Example: contact potential

- Proteins in PDB
- Amino acid contacts
Example 4 body potential

Correlated mutations

• Mutual information (MI)

\[ I(i, j) = \sum_{x_i=1}^{21} \sum_{y_j=1}^{21} P(x_i, y_j) \log \left( \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \right) \]

• Correlated mutations may suggest contact pairs or allosteric pathways
Correlated mutations

- Correlated mutations may suggest contact pairs or allosteric pathways

Communication network through bovine rhodopsin
Structure prediction: where are we?

• Database methods are much better than physics-based methods

• Not much improvement in 20 years

• Not yet good enough for drug discovery

K A Dill, and J L MacCallum Science 2012;338:1042-1046
Structure prediction: where are we?

All-atom MD from DESRES

Structure prediction: where are we?

All-atom MD from MELD