Global vs. local alignments

Global alignment (Semi-global alignment)

\[ s(a, b) = \begin{cases} 2 & \text{if } a = b, \\ 0 & \text{if } |a - b| = 1, \\ -2 & \text{otherwise} \end{cases} \]

Local alignment

Match/Mismatch scores

- Gap penalty

Indel Evolution and Gap Penalty

- A gap of length \( k \neq k \) gaps of length 1

\[ \text{TTC} \quad \text{ACG} \]

\[ \text{ATTCCG} \]

\[ \text{deletion} \]

\[ \text{deletion} \]

\[ \text{deletion} \]

\[ \text{ACG} \]

Which is more likely?

Which is biologically easier?

- Single insertion/deletion event

- Multiple insertion/deletion events

Indel Evolution and Gap Penalty

- Indel mutations are often strongly deleterious
- Indel events are rare (less common than point mutations)
- Multi-residue indels are not uncommon (e.g., hotspot, repetitive DNA)

Unequal crossover

- Replication slippage

From Human Molecular Genetics 2 (available in NCBI Bookshelf)

Indel Evolution and Gap Penalty

- Indel mutations are often strongly deleterious
- Indel events are rare (less common than point mutations)
- Multi-residue indels are not uncommon
- Fewest number of unlikely events → most likely evolutionary hypothesis

Maximum parsimony

- AATCTATA 2 indels
- AATCTATA 1 indel (more likely than 2 indel events)
**Indel Evolution and Gap Penalty**

- Fewer, but longer, indel event is more likely than too many small indels

During alignments

**Gap Penalty Functions**

- **Linear (length-proportional) gap penalty:**
  \[ w(x) = gx \]
  
  - \( g \): gap penalty
  - \( x \): length of a gap

- **Affine gap penalty:**
  \[ w(x) = \begin{cases} 
  g_o + g_e (x - 1) & \text{when } x > 0 \\
  0 & \text{when } x = 0 
  \end{cases} \]

  - \( g_o \): gap opening penalty
  - \( g_e \): gap extension penalty (usually \( g_o > g_e \))
  - \( x \): length of a gap

**Simple Alignments**

- **Varied length & gaps considered**

  - Alignment Score = 
    \[ \begin{align*}
    & \text{(match score)} \times \text{(the number of matched pairs)} + \\
    & \text{(mismatch score)} \times \text{(the number of mismatched pairs)} + \\
    & \sum \left( \text{(gap opening penalty)} \times \text{(gap length - 1)} \right) 
    \end{align*} \]

  - If match score = 1, mismatch score = 0, gap penalty = -1
  - If using linear gap penalty.

  - Example:
    - \( \text{AACTATA} \)
    - \( \text{AA-GATA} \)
    - \( \text{AAG-AT-A} \)

    - Alignment Score = 2

  - Example:
    - \( \text{AACTATA} \)
    - \( \text{AA-GATA} \)
    - \( \text{AAG-AT-A} \)

    - Alignment Score = 3

  - Example:
    - \( \text{AACTATA} \)
    - \( \text{AA--GATA} \)
    - \( \text{AAG-AT-A} \)

    - Alignment Score = 1

- **Varied length & gaps considered**

  - Alignment Score = 
    \[ \begin{align*}
    & \text{(match score)} \times \text{(the number of matched pairs)} + \\
    & \text{(mismatch score)} \times \text{(the number of mismatched pairs)} + \\
    & \sum \left( \text{(gap opening penalty)} \times \text{(gap length - 1)} \right) 
    \end{align*} \]

  - If match score = 1, mismatch score = 0, gap penalty = -1
  - If using affine gap penalty.

  - Example:
    - \( \text{AACTATA} \)
    - \( \text{AA-GATA} \)
    - \( \text{AAG-AT-A} \)

    - Alignment Score = 2

  - Example:
    - \( \text{AACTATA} \)
    - \( \text{AA--GATA} \)
    - \( \text{AAG-AT-A} \)

    - Alignment Score = -1
Empirical Indel Distribution: DNA

Based on the comparisons of 78 processed pseudogenes against their functional homologues in the human genome.


Deletions are more frequent than insertions.


Empirical Indel Distribution: DNA

Based on the comparisons of 23 noncoding region sequences between Drosophila simulans and D. sechellia.


This distribution was later used in MCALIGN:


Empirical Indel Distribution: Protein

Based on the comparisons of 4,952 protein pairs from human, mouse, and rat.

Sequences were aligned by a dynamic programming method.


Empirical Indel Distribution: Protein

Based on the comparisons of 1,310 orthologous families from 22 fungal species.


Loop regions have more indels compared to the regions with secondary structures.


Gap Penalty Function (more realistic)

- Empirical indel size distributions (both for DNA and proteins) can be described by a power law:
  \[ f_k = C k^{-b} \]  
  \[ k: \text{indel size}, \ b: \text{the power parameter} \]

- Corresponding gap penalty function
  \[ w = a + b \ln(k) \]
  \[ a: \text{gap opening penalty} \]
  \[ b: \text{gap extension penalty} \]

- Gap extension penalty is proportional to the logarithm of gap length \( k \) (logarithmic gap penalty system)

- Increases more slowly with gap length than in the affine gap penalty system (long gaps easier)

\( (\text{e.g., Cartwright 2009, MBE 26:473 and the cited refs}) \)

Number of substitutions (aa)

Indel rates/lengths are affected by different alignment methods!
### Scoring (Substitution) Matrix: DNA

- **DNA Identity Matrix**
  - Match score = 1
  - Mismatch score = 0
  - Match/mismatch scores can be expressed in a matrix format

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ S = \sum (s(i, j) + w(x)) \]

- \( s(i, j) \): the similarity score between nucleotides \( i \) and \( j \)
- \( w(x) \): gap penalty

### Transition/Transversion Matrix

- Match score = 1
- Mismatch score:
  - transition = 1 (more allowed → smaller penalty)
  - transversion = -5 (fewer allowed → larger penalty)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>T</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>C</td>
<td>?</td>
<td>?</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>G</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>

### DNA Substitution Types

- **Pyrimidines**: T, C, A, G
- **Purines**: A, T

- **Transition substitutions**: between pyrimidines, between purines
- **Transversion substitutions**: between pyrimidines and purines

- **Transition**
- **Transversion**

### Amino acid substitution matrices

- **Identity matrix**
- **Genetic code matrix**
- **Matrices based on AA properties**
- **Matrices based on empirical data**
  - Dayhoff matrices (PAM120 *etc.*)
  - BLOSUM matrices (BLOSUM62 *etc.*)
  - Gonnet matrices (Gonnet 250 *etc.*)
  - JTT matrices
  - and more ...

Amino acid substitution matrices

- Matrices based on various amino acid properties (hydrophobicity, charge, electronegativity, size, etc.)
  - Biologically meaningful matrix can be obtained by combining all of these matrices (including genetic code matrix). Not easy!
- Matrices based on empirical data
  - Alignments show the results of experiments done by the Nature
  - Capture the relative substitutability of amino acid pairs in the context of evolution
  - The model of protein evolution

Substitution matrices based on empirical data

- PAM matrices
  - Dayhoff, Schwartz, and Orcutt (1978)
- BLOSUM matrices
  - Henikoff and Henikoff (1992)


Margaret O. Dayhoff (1925-1983)

- Founder of the field of Bioinformatics
- The first woman in the field
- Collection of all known protein sequences
- 1st Atlas contained 65 proteins
- Developed into PIR (Protein Information Resource), a brain-child of Dayhoff
- Dayhoff developed a single letter code for the amino acids


Blue-sensitive opsin proteins

- Aliphatic
- Non-polar
- Tiny
- Small
- Polar
- Charged
- Aromatic
- Proline
- Positive
- Negative

E

D

Amino acid substitution matrices

Identity matrix
Genetic code matrix
Matrices based on AA properties
Matrices based on empirical data
- Dayhoff matrices (PAM120 etc.)
- BLOSUM matrices (BLOSUM62 etc.)
- Gonnet matrices (Gonnet 250 etc.)
- JTT matrices
- and more …

**PAM matrices (Dayhoff et al. 1978)**

- **Accepted point mutations** (point accepted mutations, percent accepted mutations)
  - **accepted by selection**: no (or very weak) deleterious effect, maintaining the function
- **Based on 1,572 changes in 71 groups of closely related proteins** (34 protein families)
  - at least 85% identical
  - no ambiguity in alignments, no gap
  - most likely observed substitutions do not affect protein functions (accepted by selection, close to neutral)
  - successive (multiple) substitutions at one site are minimal (no hidden substitution)