Global vs. local alignments

Global alignment (Semi-global alignment)

\[ s(a, b) = \begin{cases} 2 & \text{if } a = b \\ -1 & \text{if } a \neq b \\ -2 & \text{for free end gap} \end{cases} \]

Local alignment

Match/Mismatch scores

Gap penalty

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)

Indel Evolution and Gap Penalty

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

\[ \begin{align*}
\text{ATTCCG} & \quad \text{deletion} \\
\text{ATCCG} & \quad \text{deletion} \\
\text{ACCG} & \quad \text{deletion} \\
\text{ACG} & \quad \text{deletion}
\end{align*} \]

Which is more likely?
Which is biologically easier?

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 \( \rightarrow \) most likely evolutionary hypothesis
  Maximum parsimony

Replication slippage
Unequal crossover

From Human Molecular Genetics 2 (available in NCBI Bookshelf)
Indel Evolution and Gap Penalty

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

![Diagram showing the difference between single and multiple indel events.]

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**

  - If match score = 1, mismatch score = 0, gap penalty = -1
  - If using linear gap penalty, gap opening penalty = -2, gap extension penalty = -1
  - If using affine gap penalty, gap opening penalty = -2, gap extension penalty = -1
Empirical Indel Distribution: DNA

Based on the comparisons of >1700 processed pseudogenes against their functional homologues in the human genome.

Deletions are more frequent than insertions.

Zhang and Gerstein (2003)

Gap Length


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

Sequences were aligned by a dynamic programming method.

Gap Penalty Function (more realistic)

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

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

- Gap extension penalty is proportional to the logarithm of gap length \( k \) (logarithmic gap penalty system)
  \[ w = a + b \ln(k) \]

- Increases more slowly with gap length than in the affine gap penalty system (easier long gaps)
  \( (e.g., \text{Cartwright} 2006; \text{Cartwright} 2009; \text{Loewenthal et al.} 2021) \)

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.

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 MCALIGN2: Wang et al. (2006) BMC Bioinformatics 7: 202.

Empirical Indel Distribution: DNA

Based on the comparisons of >1700 processed pseudogenes against their functional homologues in the human genome.

Deletions are more frequent than insertions.

Zhang and Gerstein (2003)

Gap Length


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

Sequences were aligned by a dynamic programming method.

Gap Penalty Function (more realistic)

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

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

- Gap extension penalty is proportional to the logarithm of gap length \( k \) (logarithmic gap penalty system)
  \[ w = a + b \ln(k) \]

- Increases more slowly with gap length than in the affine gap penalty system (easier long gaps)
  \( (e.g., \text{Cartwright} 2006; \text{Cartwright} 2009; \text{Loewenthal et al.} 2021) \)

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.

Empirical Indel Distribution: Protein

Based on the comparisons of >1700 processed pseudogenes against their functional homologues in the human genome.

Deletions are more frequent than insertions.

Zhang and Gerstein (2003)

Gap Length


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

Sequences were aligned by a dynamic programming method.

Gap Penalty Function (more realistic)

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

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

- Gap extension penalty is proportional to the logarithm of gap length \( k \) (logarithmic gap penalty system)
  \[ w = a + b \ln(k) \]

- Increases more slowly with gap length than in the affine gap penalty system (easier long gaps)
  \( (e.g., \text{Cartwright} 2006; \text{Cartwright} 2009; \text{Loewenthal et al.} 2021) \)

Empirical Indel Distribution: Protein

Based on the comparisons of >1700 processed pseudogenes against their functional homologues in the human genome.

Deletions are more frequent than insertions.

Zhang and Gerstein (2003)
Scoring (Substitution) Matrix: DNA

- Match/mismatch scores can be expressed in a matrix format

### DNA Identity Matrix

<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>

- Match score = 1
- Mismatch score = 0

$s(i, j)$: the similarity score between nucleotides $i$ and $j$

$n(x)$: gap penalty

DNA Substitution Types

- **Purines**: Adenine (A), Guanine (G)
- **Pyrimidines**: Thymine (T), Cytosine (C)

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

Scoring (Substitution) Matrix: DNA

- Match/mismatch scores can be expressed in a matrix format

### Transition/Transversion Matrix

<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>?</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>

- Match score = 1
- Mismatch score:
  - transition = ?
  - transversion = ?

DNA Substitution Types

Which substitutions happen more often?

Zhang and Gerstein (2003)

Based on the comparisons of >1700 processed pseudogenes against their functional homologues in the human genome

Transition $\gg$ Transversion

### Transition/Transversion Matrix

- Transition score = 1
- Mismatch score:
  - transition = -1 (more allowed $\Rightarrow$ smaller penalty)
  - transversion = -5 (fewer allowed $\Rightarrow$ larger penalty)
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 ...
  - AAMindex
  - Matrices based on empirical data
  - Matrices based on AA properties
  - Genetic code matrix

Amino acid physico-chemical properties

- Major factors in Protein folding
  - Protein functions
  - Substitutions between similar amino acids are more common
  - In functionally or structurally important regions, substitutions are limited between similar amino acids

Amino acid substitutions
Amino acid substitutions

- Blue-sensitive opsin proteins

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

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 …

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


Amino acid substitution matrices

- 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

- Blue-sensitive vs. photon-sensitive opsins

Amino acid substitution matrices

- Matrices based on empirical data
  - Dayhoff matrices (PAM120 etc.)
  - BLOSUM matrices (BLOSUM62 etc.)
  - Gonnet matrices (Gonnet 250 etc.)
  - JTT matrices
  - and more …
Blue-sensitive opsin proteins

Amino acid substitution matrices

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

Amino acid substitution matrices

Substitution matrices based on empirical data

- PAM matrices
  - Dayhoff, Schwartz, and Orcutt (1978)

- BLOSUM matrices
  - Henikoff and Henikoff (1992)


Margaret O. Dayhoff (1925-1983)

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

Read the Smithsonian website; also in Strasser (2010)