Lecture outline

- Sequence alignment
  - Why do we need to align sequences?
  - Evolutionary relationships
- Gaps and scoring matrices
- Dynamic programming
  - Global alignment (Needleman & Wunsch)
  - Local alignment (Smith & Waterman)
- Database searches
  - BLAST
  - FASTA

Alignment

Procedure of comparing two (pairwise) or more (multiple) sequences by searching for a series of individual characters that are in the same order in the sequences

```
GCTAGTCAGATCTGACGCTA
| | | | | | | | | |
TGGTACATCTGCGC
```

Sequence alignment

- Comparing DNA/protein sequences for
  - Similarity
  - Homology
- Prediction of function
- Construction of phylogeny
- Shotgun assembly
  - End-space-free alignment / overlap alignment
- Finding motifs

Homology

- Orthologs
  - Divergence follows speciation
  - Similarity can be used to construct phylogeny between species
- Paralogs
  - Divergence follows duplication
- Xenologs
- Article on terminology
- ISMB tutorial on protein sequence comparison

Orthologs and paralogs
Sources of variation

- Nucleotide substitution
  - Replication error
  - Chemical reaction
- Insertions or deletions (indels)
  - Unequal crossing over
  - Replication slippage
- Duplication
  - a single gene (complete gene duplication)
  - part of a gene (internal or partial gene duplication)
    - Domain duplication
    - Exon shuffling
  - part of a chromosome (partial polysomy)
  - an entire chromosome (aneuploidy or polysomy)
  - the whole genome (polyploidy)

Differing rates of DNA evolution

- Functional/selective constraints (particular features of coding regions, particular features in 5' untranslated regions)
- Variation among different gene regions with different functions (different parts of a protein may evolve at different rates).
- Within proteins, variations are observed between
  - surface and interior amino acids in proteins (order of magnitude difference in rates in haemoglobins)
  - charged and non-charged amino acids
  - protein domains with different functions
  - regions which are strongly constrained to preserve particular functions and regions which are not
  - different types of proteins — those with constrained interaction surfaces and those without

Common assumptions

- All nucleotide sites change independently
- The substitution rate is constant over time and in different lineages
- The base composition is at equilibrium
- The conditional probabilities of nucleotide substitutions are the same for all sites, and do not change over time
- Most of these are not true in many cases…

Lecture outline

- Sequence alignment
  - Why do we need to align sequences?
  - Evolutionary relationships
- Gaps and scoring matrices
- Dynamic programming
  - Global alignment (Needleman & Wunsch)
  - Local alignment (Smith & Waterman)
- Database searches
  - BLAST
  - FASTA

A simple alignment

- Let us try to align two short nucleotide sequences:
  - AATCTATA and AAGATA
- Without considering any gaps (insertions/deletions) there are 3 possible ways to align these sequences

<table>
<thead>
<tr>
<th>AATCTATA</th>
<th>AATCTATA</th>
<th>AATCTATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAGATA</td>
<td>AAGATA</td>
<td>AAGATA</td>
</tr>
</tbody>
</table>

- Which one is better?

Scoring the alignments

- We need to have a scoring mechanism to evaluate alignments
  - match score
  - mismatch score
- We can have the total score as:
\[ \sum_{i=1}^{n} \text{match or mismatch score at position } i \]
- For the simple example, assume a match score of 1 and a mismatch score of 0:

<table>
<thead>
<tr>
<th>AATCTATA</th>
<th>AATCTATA</th>
<th>AATCTATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAGATA</td>
<td>AAGATA</td>
<td>AAGATA</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
Good alignments require gaps

- Maximal consecutive run of spaces in alignment
  - Matching mRNA (cDNA) to DNA
  - Shortening of DNA/protein sequences
  - Slippage during replication
  - Unequal crossing-over during meiosis
  - ...
- We need to have a scoring function that considers gaps also

More complicated gap penalties

- Nature favors small number of long gaps compared to large number of short gaps.
- How do we adjust our scoring scheme to account for this fact above?
  - By having different gap opening and gap extension penalties.
- Choices of gap penalties
  - Linear
  - Affine
    - Gap open penalty
    - Gap extension penalty
  - Arbitrary

Simple alignment with gaps

- Considering gapped alignments vastly increases the number of possible alignments:

  \[
  \begin{align*}
  \text{AATCTATA} & \quad \text{AACTATA} & \quad \text{AATCTATA} \\
  \text{AAG~AT~A} & \quad \text{AA~G~ATA} & \quad \text{AA~G~ATA} \\
  1 & \quad 3 & \quad 3
  \end{align*}
  \]

- If gap penalty is -1 what will be the new scores?

Score matrix

- Instead of having a single match/mismatch score for every pair of nucleotides or amino acids, consider chemical, physical, evolutionary relationships:
  - E.g.
    - Alanine vs. valine or alanine vs. lysine? Alanine and valine are both small and hydrophobic, but lysine is large and charged.
    - Which substitutions occur more in nature?
- Assign scores to each pair of symbol
  - Higher score means more similarity
Major Differences between PAM and BLOSUM

<table>
<thead>
<tr>
<th></th>
<th>PAM</th>
<th>BLOSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Built from</td>
<td>Built from local alignments</td>
<td>Built from vast amount of data</td>
</tr>
<tr>
<td>data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td>Counting based on minimum</td>
<td>Counting based on groups of related</td>
</tr>
<tr>
<td></td>
<td>replacement or maximum parsimony</td>
<td>sequences counted as one</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perform</td>
<td>Better for finding local</td>
<td>Better for finding local alignments</td>
</tr>
<tr>
<td>better</td>
<td>alignments</td>
<td></td>
</tr>
<tr>
<td>Higher PAM</td>
<td>Higher PAM series means more</td>
<td>Lower BLOSUM series means more</td>
</tr>
<tr>
<td>series</td>
<td>divergence</td>
<td>divergence</td>
</tr>
</tbody>
</table>

Typical score matrix

- **DNA**
  - Match = +1
  - Mismatch = -3
  - Gap penalty = -5
  - Gap extension penalty = -2

- **Protein sequences**
  - BLOSUM62 matrix
  - Gap open penalty = -11
  - Gap extension = -1