**Bioinformatics**

“Nothing in Biology makes sense except in the light of evolution” (Theodosius Dobzhansky (1900-1975))

“Nothing in bioinformatics makes sense except in the light of Biology (and thus evolution)”

---

**Phylogenetics**

- **Phylogenetics is the study of the evolutionary history of living organisms**
- This is mostly done using tree-like diagrams to represent pedigrees of these organisms.
- The tree branching patterns representing the evolutionary divergence are referred to as phylogeny.
- Knowing the evolutionary patterns provides a means to predict the future (e.g. flu).

**Three kingdoms of Life**

Finding relationships between genes and gene products of different species, including those at large evolutionary distances

---

**Archaea**

- Domain Archaea is mostly composed of cells that live in extreme environments. While they are able to live elsewhere, they are usually not found there because outside of extreme environments they are competitively excluded by other organisms.
- Species of the domain Archaea are
  1. not inhibited by antibiotics (unlike bacteria, which are inhibited)
  2. lack peptidoglycan in their cell wall (unlike bacteria, which have this sugar/polypeptide compound),
  3. and can have branched carbon chains in the lipids of the phospholipid bilayer, which is a constituent of the cell membrane.

**Archaea (Cnt.)**

- Archaea are very similar to prokaryotes (i.e. bacteria) that inhabited the earth billions of years ago. It is believed that eukaryotes evolved from Archaea, because they share many mRNA sequences, have similar RNA polymerases, and both have introns.
- Probably, Archaea and Bacteria branched from each other very early in history, after which membrane infolding produced eukaryotic cells in the archaean branch approximately 1.7 billion years ago.

There are three main groups of Archaea:
1. extreme halophiles (salt),
2. methanogens (methane producing anaerobes),
3. and hyperthermophiles (e.g. living at temperatures >100º C).

*Membrane infolding is believed to have led to the nucleus of eukaryotic cells, which is a membrane-enclosed organelle that holds the cellular DNA. Prokaryotic cells are more primitive and do not have a nucleus.*
### Darwinian Evolution

**What is needed:**

1. Template (DNA)
2. Copying mechanism (meiosis/fertilisation)
3. Variation (e.g. resulting from copying errors, gene conversion, crossing over, genetic drift, etc.)
4. Selection

---

### DNA evolution

- **Gene nucleotide substitutions**: can be synonymous (i.e. not changing the encoded amino acid) or nonsynonymous (i.e. changing the a.a.).
- **Rates of evolution**: vary tremendously among protein-coding genes. Molecular evolutionary studies have revealed an ~1000-fold range of nonsynonymous substitution rates (Li and Graur 1991).
- **The strength of negative (purifying) selection** is thought to be the most important factor in determining the rate of evolution for the protein-coding regions of a gene (Kimura 1983; Ohta 1992; Li 1997).

---

### Ka/Ks Ratios

- **Ks** is defined as the number of synonymous nucleotide substitutions per synonymous site.
- **Ka** is defined as the number of nonsynonymous nucleotide substitutions per nonsynonymous site.
- The **Ka/Ks ratio** is used to estimate the type of selection exerted on a given gene or DNA fragment.
- **Need orthologous nucleotide sequence alignments**.
- **Observe nucleotide substitution patterns** at given sites and consider numbers using, for example, the Pamilo-Bianchi-Li method (Li 1993; Pamilo and Bianchi 1993).
- **Correction is needed because of the following:**
  - Consider the codons specifying aspartic acid and lysine: both start AA, lysine ends A or G, and aspartic acid ends T or C. So, if the rate at which C changes to T is higher than the rate at which C changes to G or A (as is often the case), then more of the changes at the third position will be synonymous than might be expected. Many of the methods to calculate Ka and Ks differ in the way they make the correction needed to account for this bias.
  - Lysine (K) - AA<sub>G</sub> C → T
  - Aspartic acid (D) - AA<sub>C</sub> C → G
  - C → A
Ka/Ks ratios

The frequency of different values of Ka/Ks for 835 mouse–rat orthologous genes. Figures on the x-axis represent the middle figure of each bin; that is, the 0.05 bin collects data from 0 to 0.1

Ka/Ks ratios

Three types of selection:
1. Negative (purifying) selection \( \rightarrow Ka/Ks < 1 \)
2. Neutral selection (Kimura) \( \rightarrow Ka/Ks \sim 1 \)
3. Positive selection \( \rightarrow Ka/Ks > 1 \)

Orthology/paralogy

Orthologous genes are homologous (corresponding) genes in different species
Paralogous genes are homologous genes within the same species (genome)

Orthology/paralogy

Operational definition of orthology:

**Bi-directional best hit:**
- Blast gene A in genome 1 against genome 2: gene B is best hit
- Blast gene B against genome 1: if gene A is best hit
  \( \rightarrow A \) and B are orthologous

A number of other criteria is also in use (part of which is based on phylogeny)

Xenology

- **Xenologs** result from the horizontal transfer of a gene between two organisms.
- The function of xenologs can be variable, depending on how significant the change in context was for the horizontally moving gene. In general, though, the function tends to be similar (before and after horizontal transfer)

Multivariate statistics – Cluster analysis

Raw table

Similarity matrix

Cluster criterion

Dendrogram
Multivariate statistics – Cluster analysis

Why do it?
• Finding a true typology
• Model fitting
• Prediction based on groups
• Hypothesis testing
• Data exploration
• Data reduction
• Hypothesis generation

But you can never prove a classification/typology!

Cluster analysis – data normalisation

<table>
<thead>
<tr>
<th>Raw table</th>
<th>Normalised table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column normalisation</td>
<td>$x_{i}/\text{max}$</td>
</tr>
<tr>
<td>Column range normalise</td>
<td>$(x_{i}-\text{min})/(\text{max}-\text{min})$</td>
</tr>
</tbody>
</table>

Cluster analysis – (dis)similarity matrix

$D_{ij} = (\sum_{k} |x_{ik} - x_{jk}|^{r})^{1/r}$ \textit{Minkowski metrics}

$r = 2$ \textit{Euclidean distance}
$r = 1$ \textit{City block distance}

Example:

<table>
<thead>
<tr>
<th></th>
<th>length</th>
<th>height</th>
<th>width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cow 1</td>
<td>11</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Cow 2</td>
<td>7</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Euclidean dist. $= \sqrt{4^2 + 3^2 + (-2)^2} = \sqrt{29} = 5.39$
City Block dist. $= |4| + |3| + |-2| = 9$

Cluster analysis – Clustering criteria

1. Start with $N$ clusters of 1 object each
2. Apply clustering distance criterion iteratively until you have 1 cluster of $N$ objects
3. Most interesting clustering somewhere in between

Note: a dendrogram can be rotated along branch points (like mobile in baby room) – distances between objects are defined along branches
Comparing sequences - Similarity Score -

Many properties can be used:
- Nucleotide or amino acid composition
- Isoelectric point
- Molecular weight
- Morphological characters
- But: molecular evolution through sequence alignment

Why phylogenetic trees?
- Most of bioinformatics is comparative biology
- Comparative biology is based upon evolutionary relationships between compared entities
- Evolutionary relationships are normally depicted in a *phylogenetic tree*, which shows the relation of each of the individual entities relative to all others

Where can phylogeny be used
- For example, finding out about orthology *versus* paralogy
- Predicting secondary structure of RNA
- Studying host-parasite relationships
- Predicting protein-protein interactions
  - Mapping membrane-bound receptors onto their binding ligands
- Multiple sequence alignment (e.g. Clustal)

Sequence similarity criterion for phylogeny
- There are various models to correct for the fact that the true rate of evolution cannot be observed through nucleotide (or amino acid) exchange patterns (e.g. back mutations)
- Saturation level is ~94%, higher real mutations are no longer observable; e.g. observed A→T through A→C→G→T or observed G→G through G→C→G

Similarity criterion for phylogeny
- There are various models to correct for the fact that the true rate of evolution cannot be observed through nucleotide (or amino acid) exchange patterns (e.g. back mutations)
- Saturation level is ~94%, higher real mutations are no longer observable; e.g. observed A→T through A→C→G→T or observed G→G through G→C→G
Sequence Identity and Sequence-based Evolutionary Distance

Lactate dehydrogenase multiple alignment

How can you see that this is a distance matrix?

Cluster analysis – Clustering criteria

Note: these are all agglomerative cluster techniques; i.e. they proceed by merging clusters as opposed to techniques that are divisive and proceed by cutting clusters

General agglomerative cluster protocol

1. Start with $N$ clusters of 1 object each
2. Apply clustering distance criterion and merge clusters iteratively until you have 1 cluster of $N$ objects
3. Most interesting clustering somewhere in between

Dendrogram (tree)

Note: a dendrogram can be rotated along branch points (like mobile in baby room) – distances between objects are defined along branches

Single linkage clustering (nearest neighbour)
Distances from point to cluster is defined as the smallest distance between that point and any point in the cluster.
Single linkage clustering (nearest neighbour)

Let \( C_i \) and \( C_j \) be two disjoint clusters:

\[
d_{ij} = \min(d_{pq}), \text{ where } p \in C_i \text{ and } q \in C_j
\]

Single linkage dendrograms typically show **chaining** behaviour (i.e., all the time a single object is added to existing cluster)

Complete linkage clustering (furthest neighbour)
Complete linkage clustering (furthest neighbour)

Distance from point to cluster is defined as the largest distance between that point and any point in the cluster.

Let $C_i$ and $C_j$ be two disjoint clusters:
$$d_{ij} = \text{Max}(d_{pq})$$
where $p \in C_i$ and $q \in C_j$

More ‘structured’ clusters than with single linkage clustering

Clustering algorithm

1. Initialise (dis)similarity matrix
2. Take two points with smallest distance as first cluster (later, points can be clusters)
3. Merge corresponding rows/columns in (dis)similarity matrix
4. Repeat steps 2. and 3. using appropriate cluster measure when you need to calculate new point-to-cluster or cluster-to-cluster distances until last two clusters are merged
Single Linkage example

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Complete Linkage example

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Phylogenetic trees

- Multiple sequence alignment (MSA)
- Similarity/Distance matrix
- Cluster criteria

Phylogenetic tree (unrooted)

- Clade - group of two or more taxa that includes both their common ancestor and all of their descendants.

Phylogenetic tree (rooted)
How to root a tree

- Outgroup – place root between distant sequence and rest group
- Midpoint – place root at midpoint of longest path (sum of branches between any two OTUs)
- Gene duplication – place root between paralogous gene copies (see earlier globin example)

Phylogenetic tree terminology

- A cladogram has unscaled branch lengths (i.e. external taxa are normally neatly lined up and only the topology matters) whilst in a phylograms these are scaled

How many trees?

- Number of unrooted trees = \((2n-5)! / 2^{n-3} (n-3)!\)
- Number of rooted trees = \((2n-3)! / 2^{n-3}2(n-2)!\)

Combinatoric explosion

<table>
<thead>
<tr>
<th># sequences</th>
<th># unrooted trees</th>
<th># rooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
<td>945</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
<td>10,395</td>
</tr>
<tr>
<td>8</td>
<td>10,395</td>
<td>135,135</td>
</tr>
<tr>
<td>9</td>
<td>135,135</td>
<td>2,027,025</td>
</tr>
<tr>
<td>10</td>
<td>2,027,025</td>
<td>34,459,425</td>
</tr>
</tbody>
</table>

Wrapping up

- Evolution
  - requirements
  - negative/positive selection on genes (e.g. Ka/Ks)
  - homology/paralogy/orthology (operational definition ‘bi-directional best hit’)
- Clustering
  - single linkage
  - complete linkage
- Tree characteristics
  - Unrooted/rooted trees
  - How to root a tree
  - Combinatorial explosion of number of possible trees