1-Month Practical Master Course
Genome Analysis (Integrative Bioinformatics & Genomics)

Jaap Heringa
Centre for Integrative Bioinformatics VU (IBIVU)
Vrije Universiteit Amsterdam
The Netherlands

www.ibivu.cs.vu.nl
heringa@cs.vu.nl

Biological Sequence Analysis

Pair-wise sequence alignment
Residue exchange matrices
Multiple sequence alignment
Phylogeny

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>φX-174 virus</td>
<td>5,386</td>
</tr>
<tr>
<td>Epstein Bar Virus</td>
<td>172,282</td>
</tr>
<tr>
<td>Mycoplasma genitalium</td>
<td>580,000</td>
</tr>
<tr>
<td>Hemophilus Influenza</td>
<td>$1.8 \times 10^6$</td>
</tr>
<tr>
<td>Yeast (S. Cerevisiae)</td>
<td>$12.1 \times 10^6$</td>
</tr>
<tr>
<td>Human</td>
<td>$3.2 \times 10^9$</td>
</tr>
<tr>
<td>Wheat</td>
<td>$16 \times 10^9$</td>
</tr>
<tr>
<td>Lilium longiflorum</td>
<td>$90 \times 10^9$</td>
</tr>
<tr>
<td>Salamander</td>
<td>$100 \times 10^9$</td>
</tr>
<tr>
<td>Amoeba dubia</td>
<td>$670 \times 10^9$</td>
</tr>
</tbody>
</table>

Three main principles

- DNA makes RNA makes Protein
- Structure more conserved than sequence
- Sequence $\rightarrow$ Structure $\rightarrow$ Function
How to go from DNA to protein sequence

6-frame translation using the codon table (last lecture):

5’ attcggtgcaaatgccccctatccggt 3’

3’ taagcaaacggtttacgccggtagcgc 5’

DNA direction is from 5’ to 3’

Evolution and three-dimensional protein structure information

Isocitrate dehydrogenase:
The distance from the active site (in yellow) determines the rate of evolution (red = fast evolution, blue = slow evolution)


Widely used tool for homology detection: PSI-BLAST

- Heuristic tool to cut down computations required for database searching (~1M sequences in DB)
- Sensitivity gained by iteratively finding hits (local alignments) and repeating search
“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”
**Mutations under divergent evolution**

- **Ancestral sequence**
- **Sequence 1:** ACCTGTAATC
- **Sequence 2:** ACCTGCACATC
- \( D = \frac{3}{10} \) (fraction different sites (nucleotides))

**Convergent evolution**

- Often with shorter motifs (e.g., active sites)
- Motif (function) has evolved more than once independently, e.g., starting with two very different sequences adopting different folds
- Sequences and associated structures remain different, but (functional) motif can become identical
- Classical example: serine proteinase and chymotrypsin

**Serine proteinase (subtilisin) and chymotrypsin**

- Different evolutionary origins, no sequence similarity
- Similarities in the reaction mechanisms. Chymotrypsin, subtilisin and carboxypeptidase C have a catalytic triad of serine, aspartate and histidine in common: serine acts as a nucleophile, aspartate as an electrophile, and histidine as a base.
- The geometric orientations of the catalytic residues are similar between families, despite different protein folds.
- The linear arrangements of the catalytic residues reflect different family relationships. For example the catalytic triad in the chymotrypsin clan (SA) is ordered DHS, but is ordered DHS in the subtilisin clan (SB) and SDH in the carboxypeptidase clan (SC).

**A protein sequence alignment**

MSTGAVLY--TSILIKECHAMPAGNE-----
GILLFHRTHELIKESHAMANDEGGSNNS
* * * * * * * * * * *

**A DNA sequence alignment**

attcgtggcaaatcgcccttatccggctttaa
att---tggcggatcg-cctctacgagcc---
*** *** *** *** *** *** *** ***

**What can sequence tell us about structure**

(HSSP) 
Sander & Schneider, 1991

**Searching for similarities**

**What is the function of the new gene?**

The “lazy” investigation (i.e., no biological experiments, just bioinformatics techniques):

- Find a set of similar protein sequences to the unknown sequence
- Identify similarities and differences
- For long proteins: identify domains first
Evolutionary and functional relationships

Reconstruct evolutionary relation:
• Based on sequence
  - Identity (simplest method)
  - Similarity
• Homology (common ancestry: the ultimate goal)
• Other (e.g., 3D structure)

Functional relation:
Sequence → Structure → Function

Searching for similarities

Common ancestry is more interesting:
Makes it more likely that genes share the same function

Homology: sharing a common ancestor
– a binary property (yes/no)
– it is a very useful property:
When (an unknown) gene X is homologous to (a known) gene G it means that we gain a lot of information on X: what we know about G can be transferred to X as a good suggestion.

Biological definitions for related sequences

- **Homologues** are similar sequences in two different organisms that have been derived from a common ancestor sequence. Homologues can be described as either orthologues or paralogues.
- **Orthologues** are similar sequences in two different organisms that have arisen due to a speciation event. Orthologues typically retain identical or similar functionality throughout evolution.
- **Paralogues** are similar sequences within a single organism that have arisen due to a gene duplication event.
- **Xenologues** are similar sequences that do not share the same ancestry origin, but rather have arisen out of horizontal transfer events through symbiosis, viruses, etc.

How to evolve

Important distinction:
• **Orthologues**: homologous proteins in different species (all deriving from same ancestor)
• **Paralogues**: homologous proteins in same species (internal gene duplication)

• In practice: to recognize orthology, bi-directional best hit is used in conjunction with database search program (this is called an operational definition)

So this means ...

Pairwise sequence alignment needs sense of evolution
Global dynamic programming

How to determine similarity
Frequent evolutionary events at the DNA level:
1. Substitution
2. Insertion, deletion
3. Duplication
4. Inversion

We will restrict ourselves to these events

A DNA sequence alignment
atcgttgccatcgtatccggaccttaa
att---tgqcgagctcctacggtgccc----
*** **** *** *** ******

A protein sequence alignment
MSTGAVLIY-TSILEKEMANDEGGSNS
---GGILLFHRTHESILSHEAMANDEGGSNS
* * * **** ***

nucleotide one-letter code
amino acid one-letter code