Quantifying sequence similarity

After this lecture, you can...

... define homology, similarity, and identity

... give two examples of DNA substitution matrices

... list four properties of amino acids that might be important in determining their physico-chemical similarity

... derive BLOSUM scores from counts of aligned residues

... interpret BLOSUM scores and BLOSUM-based alignment scores by performing calculations with log-odds values

... explain when to use matrices based on high/low BLOSUM numbers (e.g. 45/62/80) and why they are suitable

... list some of the elements in a model of evolution, explain what it is and why we use it

... convert mutational distance into evolutionary distance by correcting for mutational saturation with Jukes-Cantor
Homology

- Phylogenetic trees and sequence alignments contain a lot of information about function and evolution of sequences
- But they only make sense if the sequences descend from a common ancestor, i.e. they are evolutionarily related

**We are NEVER allowed to interpret trees or alignments, UNLESS they are made with homologous sequences!**

... so how do we determine if sequences share a common ancestor? (i.e. are homologs?)

Similarity

- Things that look really similar probably share a common ancestor
Identity

• Alignment means: adding gaps in one and/or the other sequence until they are both equally long:

```
-AGGCTATCACGTACCTCCAGGCAGGCCA--TGCCC--
TAG-CTATCAC--GACCCG--GCTCGATTGCCCAGAC
```

• Before alignment:

```
seq1 MAPFAFSLRYFMLAAPKWLTKMNVWEGHQQHSCV
seq2 MAPGAFSLYPFMLAPRWLTHMVWEGHQQHSCVAR
```

• After alignment:

```
seq1 MAPFAFSLRYFMLAAPKWLTKMNVWEGHQQHSCV--
seq2 MAPGAFSLYPFMLAPRWLTHMVWEGHQQHSCVAR
```

• Aligned sequences allow us to calculate percent identity:
  - 8 different positions, and 31 identical positions
  - The two sequences are 100 * \(\frac{31}{39}\) = 79% identical

• Identity can be quantified, but homology cannot!
  - Saying “79% homologous” is wrong (even though too many people say it)

Definitions

• Homology
  - Property of two sequences that have a shared ancestor
  - Homology is TRUE or FALSE: either you’re family or you’re not

• Identity
  - Percentage of identical residues in an alignment
  - Used for amino acids or nucleotides

• Similarity
  - Percentage of amino acid residues in an alignment with a positive substitution score
  - Not used for DNA
Evolutionary distance

- The evolutionary distance between two sequences is often expressed as the frequency of mismatches
  - The distance between virus3 and virus7: 0.77 mutations/site
- Assumption: positions evolve randomly and independently

The branch length represents the evolutionary time between two nodes. Unit: substitutions per sequence site.

DNA substitution matrices

- Similarity is quantified with a substitution matrix
  - Scores for matches in a sequence alignment
  - Penalties for mismatches in a sequence alignment
- The identity matrix is the most often used for scoring DNA sequence alignments
- Not all nucleotide substitutions are equally likely
  - Transitions occur about twice as often as transversions

Transitions:
- A ↔ G
- C ↔ T

Transversions:
- A ↔ C
- A ↔ T
- C ↔ G
- G ↔ T
Identity and similarity

- These three sequences are all 66.7% identical...
- ...but what are those colors? *
  - Some amino acids are more similar than others

(* hint)

Cysteine (C)  Aspartic acid (D)  Glutamic acid (E)

Building Blocks of Life

Amino acid properties

http://swift.cmbi.ru.nl/teach/B1B/HTML/PosterA4_nbic_new.pdf
The similarity signal

• What is the most relevant definition of a “similarity signal” between amino acids?
  – Size?
  – Polarity?
  – Hydrophobicity?
  – Preferred protein fold?
  – ...any other measures that might be important?

• Similarity signal of what, exactly? *
  – “Of the function of the amino acid in the protein”
  – “Of an evolutionary relationship”

• Evolution has tested this for us
  – So let’s use evolved sequences to score similarity!

Which amino acids mutate into each other?

• We use alignments to discover which amino acids are more similar than others, according to evolution
• We quantify this by counting how often two amino acids really mutated into each other during evolution
• The most relevant signal to detect homology is based on reliable, well-aligned homologs

Well-aligned homologs
BLOcks SUbstitution Matrix (BLOSUM)

1. Take some aligned homologous sequences

2. Group highly identical sequences (e.g. >62% identity)
   – To remove redundancy biases in the sequence database

3. Identify well-aligned blocks
   – So that only real mutations are compared (reliable, well-aligned homologs)

4. Count how often each pair of two amino acids mutated into each other (i.e. how often they are aligned)

---

How to quantify this similarity?

- BLOSUM defines similarity between amino acids based on the statistical probability that they align in well-aligned homologs
  - Observed/expected ratio (odds ratio)
    - An observed/expected ratio of 2 means something happens twice as often as you expected by random chance
  - Observed: aligned in well-aligned homologs
  - Expected: “aligned” in unaligned sequences

BLOSUM score:

\[
S_{ij} = 2 \cdot \log_2 \left( \frac{F_i \cdot F_j}{F_{ij}} \right)
\]

- The observed/expected ratio for a whole alignment can be calculated by multiplying the ratios for all aligned amino acids
  - This is a lot of multiplying for long sequences
- Logarithms allow us to sum the scores of aligned residues
  - Much easier to use for a scoring system
- Hence: the BLOSUM score
An example

- Let’s say that we have a well-aligned block of:
  - 100 amino acids long
  - 1,000 proteins “deep”
  - With no gaps
- 7.4% are alanine (A) → $F_A = \frac{7,400}{100 \cdot 1,000} = 0.074$
- 1.3% are tryptophan (W) → $F_W = 0.013$
- Randomly, we expect a fraction of A-W alignments of:
  \[ F_A \cdot F_W = 0.074 \cdot 0.013 = 0.000962 \]
- In reality, we observe a fraction of A-W alignments of: 0.00034
  - The A-W mutation occurred less frequently in evolution than expected!
  - … so the substitution score must be negative
- The A-W substitution score is:
  \[ S_{i,j} = 2 \cdot \log_2 \left( \frac{F_{i,j}}{F_i \cdot F_j} \right) \]
  \[ S_{A,W} = 2 \cdot \log_2 \left( \frac{0.00034}{0.000962} \right) = -3 \]

BLOSUM62

- Why are the highest numbers in each row/column on the diagonal?
  - Because in well-aligned homologs, every amino acid is most often aligned to itself (i.e. it is not mutated)
So what do the numbers mean?

\[ S_{ij} = 2 \cdot \log_2 \left( \frac{F_{ij}}{F_i \cdot F_j} \right) \rightarrow 2^{(S_{ij}/2)} = \frac{F_{ij}}{F_i \cdot F_j} \]

- A score of -3 for the substitution A-W means that this substitution is observed in well-aligned homologs 0.35 times as often as expected:
  \[ 2^{(-3/2)} = 0.35 = \frac{0.00034}{0.000962} \]
  - ... or: \( \frac{1}{0.35} = 2.83 \) times less often than expected

- A score of 2 for the substitution D-E means that this substitution is observed in well-aligned homologs twice as often as expected:
  \[ 2^{(2/2)} = 2 \]

- A score of 9 for the substitution C-C means that this substitution is observed in well-aligned homologs 22.6 times as often as expected:
  \[ 2^{(9/2)} = 22.6 \]

Exercise

- Calculate the similarity score between these sequences:
  a) KAWSADV
  -1 + 4 -1 + 1 - 3 + 2 - 2 = 0
  b) HAMNWEQ
  -2 + 11 + 4 + 4 - 3 + 4 = 20
  c) RDWSACV
  -2 - 2 - 1 + 1 - 1 - 3 + 5 - 2 = -2
Odds ratio that sequences are well-aligned homologs

- How likely is it that two sequences are well-aligned homologs?

\[
\begin{align*}
\text{seqA} - \text{seqB} &: -1 + -3 + -2 = -6 \\
\text{seqA} - \text{seqC} &: -1 + -2 + -2 = -5 \\
\text{seqB} - \text{seqC} &: 4 + 2 + 4 = 10
\end{align*}
\]

\[
2^{-6/2} = 0.125 \quad 2^{-5/2} = 0.177 \quad 2^{10/2} = 32
\]

- When you observe RYD aligned to SDA, these sequences are 0.125 times more likely to be well-aligned homologs than expected
- ... or 8 times less likely to be well-aligned homologs than expected
- When you observe RYD aligned to SEA, these sequences are 0.177 times more likely to be well-aligned homologs than expected
- ... or 5.6 times less likely to be well-aligned homologs than expected
- When you observe SDA aligned to SEA, these sequences are 32 times more likely to be well-aligned homologs than expected

Length of the alignment

- As you see a more and more length of two sequences...

```plaintext
MGAAVLNRKIAPGIDRDKNLGFNPYEYNRAVF
MGAAVLNRKIAPGIDRDKNLGFNPYEYNRAVF
```

- If two sequences are well-aligned homologs, most of the aligned amino acids will have positive scores
- ... so the alignment score keeps increasing as you see more length of the sequences
- ... so the likelihood that they are well-aligned homologs keeps increasing

```plaintext
KLPFSGVALARVCCSKVIGVTFSSLNSVVEDASKRQIAAEAVC
KLPFSGVALARVCCSKVIGVTFSSLNSVVEDASKRQIAAEAVC
```

- If two sequences are not homologous, most of the aligned amino acids will have negative scores
- ... so the score keeps decreasing as you see more sequence
- ... so the likelihood that they are well-aligned homologs keeps decreasing
BLOSUM numbers

- BLOSUM removes redundancy biases from the blocks of well-aligned homologs
  - The sequence selection in the database is biased
  - Clusters of highly identical sequences are collapsed
  - Per cluster, one consensus sequence is used to calculate the BLOSUM matrix
- High numbers collapse only very identical homologs
  - Used to compare and detect closely related proteins
- Low numbers also collapse more divergent homologs
  - Used to compare and detect more distant proteins
- The most used matrices:
  - BLOSUM80: closely related sequences collapsed (>80% id)
  - BLOSUM62: default, sequences collapsed if they have >62% identity
  - BLOSUM45: distantly related sequences collapsed (>45% id)

Substitution matrices

- Substitution matrices...
  ...are used in (almost) all sequence analyses
  ...strongly influence your results

- Other amino acid substitution matrices
  - Point Accepted Mutations (PAM)
    - Developed by Margaret Dayhoff in 1978
    - Based on counting 1,572 real mutations that occurred during the evolution of 71 proteins
  - Gonnet (1992)
    - Similar to PAM, but based on more sequences
    - Default in popular alignment program Clustal Omega

- Bioinformatic programs often have default parameter values set
  - For example the choice of a substitution matrix
  - Default settings are not always the most optimal for your questions
  - If you have no idea, check the literature in your biological field what is considered reliable
Models of evolution

- We use models to check if we really understand something
  - If our observations agree with the model, then we understand that something well (and vice versa)

- In a model of evolution, we quantitatively formalize the evolutionary processes that we think are at play
  - A substitution matrix is (part of) a model of evolution, because it formalizes all the relative substitution rates between residues
  - If other types of mutations are included, the model often agrees better with genome sequence data

Absolute versus evolutionary distance

- Consider a recently diverged sequence

- The two DNA sequences will diverge, but how?
  - The divergence saturates at ~25% identity
    - Mutational saturation: if the same position mutates again
    - The identity between two random DNA sequences is 25%
  - We can correct for it with the Jukes-Cantor formula:
    
    \[ d = \frac{3}{4} \ln \left( 1 - \frac{4}{3}D \right) \]

- Time à 

<table>
<thead>
<tr>
<th>Sequence identity</th>
<th>Time →</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>% id for random seqs</th>
<th>%GC of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td>1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

- Absolute versus evolutionary distance