Mutations

---

Mutations vs. Substitutions

- **Mutations** are changes in DNA
- **Substitutions** are mutations that evolution has tolerated

Which rate is greater?

Replicative proofreading and DNA repair constrain the mutation rate
Selectionist Evolution

- Most mutations are deleterious; removed via negative selection
- Advantageous mutations positively selected
- Variability arises via selection

Why Are Mutations Important?

Mutations can be deleterious
Mutations drive evolution

UV Damage to DNA

UV

Thymine dimers

What happens if damage is not repaired?

Radiation Resistant

- 10 Gray will kill a human
- 60 Gray will kill an E. coli culture
- Deinococcus can survive 5000 Gray
A Sequence Mutating at Random

Multiple substitutions at one site can cause underestimation of number of actual substitutions

9 actual substitutions

p=5 observed substitutions

Simulating Random Mutations

Measuring Sequence Divergence: Why Do We Care?

- Inferring phylogenetic relationships
- Dating divergence, correlating with fossil record
- Use in sequence alignments and homology searches of databases*

DNA Structure

G-C: 3 hydrogen bonds
A-T: 2 hydrogen bonds

Two base types:
- Purines (A, G)
- Pyrimidines (T, C)

* Comparative genomics is an important field. Determining not only how many substitutions exist between two sequences but how similar two sequences are.
Not All Base Substitutions Are Created Equal

- Transitions
  - Purine to purine (A → G or G → A)
  - Pyrimidine to pyrimidine (C → T or T → C)

- Transversions
  - Purine to pyrimidine (A → C or T; G → C or T)
  - Pyrimidine to purine (C → A or G; T → A or G)

Transition rate ~2x transversion rate

Substitution Rates Differ Across Genomes

Alignment of 3,165 human-mouse pairs

The PAM Model of Protein Sequence Evolution

- Empirical data-based substitution matrix
- Global alignments of 71 families of closely related proteins.
- Constructed hypothetical evolutionary trees
- Built matrix of 1572 amino acid point accepted mutations

Original PAM Substitution Matrix

Count number of times residue $i$ was replaced with residue $j$
Deriving PAM Matrices

For each amino acid, calculate its relative mutability, i.e., the likelihood that the amino acid will mutate:

\[ m_j = \frac{\# \text{ times amino acid } j \text{ mutated}}{\text{total occurrences of amino acid } j} \]

### Table 3.1: Relative Mutabilities of Amino Acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Relative Mutability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn</td>
<td>134</td>
</tr>
<tr>
<td>Ser</td>
<td>120</td>
</tr>
<tr>
<td>Asp</td>
<td>106</td>
</tr>
<tr>
<td>Glu</td>
<td>102</td>
</tr>
<tr>
<td>Ala</td>
<td>100</td>
</tr>
<tr>
<td>Thr</td>
<td>97</td>
</tr>
<tr>
<td>Ile</td>
<td>96</td>
</tr>
<tr>
<td>Met</td>
<td>94</td>
</tr>
<tr>
<td>Gln</td>
<td>93</td>
</tr>
<tr>
<td>Val</td>
<td>74</td>
</tr>
</tbody>
</table>

The value of alanine is arbitrarily set to 100.


Deriving PAM Matrices

Calculate mutation probabilities for each possible substitution

\[ M_{i,j} = \frac{m_j \times A_{i,j}}{\sum_i A_{i,j}} \]

PAM1 Mutation Probability Matrix

Calculate log odds ratio to convert mutation probability to substitution score

\[ S_{i,j} = 10 \times \log_{10} \left( \frac{M_{i,j}}{f_i} \right) \]

(where \( f_i \) is an accepted mutation)

Dayhoff, 1978
Deriving PAM Matrices

Scoring in log odds ratio:
- Allows addition of scores for residues in alignments

Interpretation of score:
- Positive: non-random (accepted mutation) favored
- Negative: random model favored

Using PAM Scoring Matrices

PAM1: 1% difference (99% identity)

Can “evolve” the mutation probability matrix by multiplying it by itself, then take log odds ratio

(PAMn = PAM matrix multiplied by itself n times)
BLOSUM Uses Clustering To Reduce Sequence Bias

- Cluster the most similar sequences together
- Reduce weight of contribution of clustered sequences
- BLOSUM number refers to clustering threshold used (e.g. 62% for BLOSUM 62 matrix)

BLOSUM and PAM Substitution Matrices

<table>
<thead>
<tr>
<th>BLOSUM</th>
<th>PAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>250 (80)</td>
</tr>
<tr>
<td>62</td>
<td>120 (66)</td>
</tr>
<tr>
<td>90</td>
<td>90 (50)</td>
</tr>
</tbody>
</table>

% identity % change

BLAST algorithm uses BLOSUM 62 matrix

PAM and BLOSUM

- Smaller set of closely related proteins - short evolutionary period
- Use global alignment
- More divergent matrices extrapolated
- Errors arise from extrapolation

BLOSUM

- Larger set of more divergent proteins - longer evolutionary period
- Use local alignment
- Each matrix calculated separately
- Clustering to avoid bias
- Errors arise from alignment errors

Importance of Scoring Matrices

- Scoring matrices appear in all analyses involving sequence comparison
- The choice of matrix can strongly influence the outcome of the analysis