Molekulare Evolution

Vorlesung *Molekulare Evolution* vom 15.11.2011

Mario Stanke
Institut für Mathematik und Informatik
Universität Greifswald
Models for protein sequences

GTR for amino acids

- state space \{ARNDCQEGHILKMFPS TWYV\}
- rate matrix

\[
Q = \begin{pmatrix}
- & q_{AR} & q_{AN} & q_{AD} & \cdots & q_{AV} \\
q_{RA} & - & q_{RN} & q_{RD} & \cdots & q_{RV} \\
& \vdots & \ddots & \vdots \\
q_{VA} & q_{VR} & q_{VN} & q_{VD} & \cdots & -
\end{pmatrix}
\]

where equilibrium frequencies satisfy \(P_i q_{ij} = P_j q_{ji}\).

- \(19 + 19 \cdot 20/2 = 209\) free parameters
  (equilibrium distribution + one factor per pair \(i \neq j\))
- requires large training set
PAM Matrix

PAM 001

- Dayhoff, Schwartz, Orcutt (1979)
- based on 1572 changes of amino acids in very closely related proteins
- chose small observed distance $p = 0.01$ so that double-mutations negligible and phylogenetic distance $d \approx p$
- table of $P(0.01) = (P_{ij}(0.01))_{ij}$
- PAM 250: $P(2.5) = P(0.01)^{250}$ = transition probabilities, when expected 2.5 mutations per site, has only 80% expected observed distance
**PAM (probability of accepted mutation)**

\[
\text{transpose of } 10000 \cdot P(0.01) = 10000 \cdot e^{Q \cdot 0.01}
\]

| A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
| Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| 9867 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 | 6 | 4 | 2 | 6 | 2 | 22 | 35 | 32 | 0 | 2 | 18 |
| 1 | 9913 | 1 | 0 | 1 | 10 | 0 | 0 | 0 | 10 | 3 | 1 | 19 | 4 | 1 | 4 | 6 | 1 | 8 | 0 | 1 |
| 4 | 1 | 9822 | 36 | 0 | 4 | 6 | 6 | 21 | 3 | 1 | 13 | 0 | 1 | 2 | 20 | 9 | 1 | 4 | 1 |
| 6 | 0 | 42 | 9859 | 0 | 6 | 53 | 6 | 4 | 1 | 0 | 3 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 9973 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 0 | 3 | 2 |
| 3 | 9 | 4 | 5 | 0 | 9876 | 27 | 1 | 23 | 1 | 3 | 6 | 4 | 0 | 6 | 2 | 2 | 0 | 0 | 1 |
| 10 | 0 | 7 | 56 | 0 | 35 | 9865 | 4 | 2 | 3 | 1 | 4 | 1 | 0 | 3 | 4 | 2 | 0 | 1 | 2 |
| 21 | 1 | 12 | 11 | 1 | 3 | 7 | 9935 | 1 | 0 | 1 | 2 | 1 | 1 | 3 | 21 | 3 | 0 | 0 | 5 |
| 1 | 8 | 18 | 3 | 1 | 20 | 1 | 0 | 9912 | 0 | 1 | 1 | 0 | 2 | 3 | 1 | 1 | 1 | 4 | 1 |
| 2 | 2 | 3 | 1 | 2 | 1 | 2 | 0 | 0 | 9872 | 9 | 2 | 12 | 7 | 0 | 1 | 7 | 0 | 1 | 33 |
| 3 | 1 | 3 | 0 | 0 | 6 | 1 | 1 | 4 | 22 | 9947 | 2 | 45 | 13 | 3 | 1 | 3 | 4 | 2 | 15 |
| 2 | 37 | 25 | 6 | 0 | 12 | 7 | 2 | 2 | 4 | 1 | 9926 | 20 | 0 | 3 | 8 | 11 | 0 | 1 | 1 |
| 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 5 | 8 | 4 | 9874 | 1 | 0 | 1 | 2 | 0 | 0 | 4 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 8 | 6 | 0 | 4 | 9946 | 0 | 2 | 1 | 3 | 28 |
| 13 | 5 | 2 | 1 | 1 | 8 | 3 | 2 | 5 | 1 | 2 | 2 | 1 | 1 | 9926 | 12 | 4 | 0 | 0 | 2 |
| 28 | 11 | 34 | 7 | 11 | 4 | 6 | 16 | 2 | 2 | 1 | 7 | 4 | 3 | 17 | 9840 | 38 | 5 | 2 | 2 |
| 22 | 2 | 13 | 4 | 1 | 3 | 2 | 2 | 1 | 11 | 2 | 8 | 6 | 1 | 5 | 32 | 9871 | 0 | 2 | 9 |
| 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 9976 | 1 | 0 |
| 1 | 0 | 3 | 0 | 3 | 0 | 1 | 0 | 4 | 1 | 1 | 0 | 0 | 21 | 0 | 1 | 1 | 2 | 9945 |
| 13 | 2 | 1 | 1 | 3 | 2 | 2 | 2 | 3 | 3 | 57 | 11 | 1 | 17 | 1 | 3 | 2 | 10 | 0 | 2 |

Felsenstein, “Inferring Phylogenies”

- have used rooted tree: not time-reversible
Rate and substitution matrices

<table>
<thead>
<tr>
<th>Rate matrix</th>
<th>Substitution matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>• $Q$, such that $P_{ij}(t) = e^{Qt}[i,j]$</td>
<td>• $S = (s_{ij})$, used to score alignment of amino acids $i$ and $j$</td>
</tr>
<tr>
<td>• used for trees and distances</td>
<td>• different matrices for different degrees of divergence (e.g. BLOSUM{62,45,80})</td>
</tr>
</tbody>
</table>

Converting a rate matrix to a substitution matrix

$Q$ determines equilibrium probabilities $P_i$. Choose a $t$. Then

$$s_{ij} := \log \frac{\text{prob. of } i, j \text{ related}}{\text{prob. of } i, j \text{ unrelated}} = \log \frac{P_i P_{ij}(t)}{P_i P_j} = \log \frac{P_{ij}(t)}{P_j} \tag{1}$$

can define a scoring matrix. The log-odds score (1) is often also used directly to construct a substitution matrix from MSAs. In general, a substitution matrix cannot be turned into a rate matrix.
Protein Evolution
Amino Acid Substitution Models
Codon Substitution Models
Estimating $\omega = d_N / d_S$
### Standard Genetic Code

AA = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRIIIMTTTTNKKSSRRVVVVAAAAADDEEGGGG
base1 = TTTTTTTTTTTTTTTTTCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGGGGG
base2 = TTTTCCCAAAAAAGGGGTTTTTCCCCAAAAGGGGTGTTTTCCCCAAAAAGGGGGTTTTTCCCCAAAAGGGGG
base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG

One-letter-codes for amino acids, * = stop codon

### Synonymous/nonsynonymous

- **synonymous** mutation, e.g. TTT → TTC
- **nonsynonymous** mutation, e.g. TTT → ATT
- transitions overrepresented among synonymous mutations
- Jukes-Cantor model ⇒ 25.5% of mutations are synonymous
- transition/transversion ratio $\kappa = 5$ ⇒ 30.9% of mutations are synonymous
Amino acid versus codon model

- **synonymous** mutations not visible in amino acid sequence
- codon substitution model can be chosen to
  - have fewer parameters
  - account for transition/transversion ratio
  - equilibrium nucleotide frequencies
- closely related $\Rightarrow$ nucleotide sequence better
- remotely related $\Rightarrow$ may choose amino acid model or parameter-rich codon model
try both
Codon Substitution Models

Model distinguishing transitions/transversions and synonymous/nonsynonymous substitutions

GY94 (Goldman, Yang), Rate matrix $Q$:

\[
q_{ij} = \begin{cases} 
0, & \text{if } i \text{ and } j \text{ differ at more than one position} \\
P_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion} \\
\kappa P_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition} \\
\omega P_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion} \\
\omega \kappa P_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition}
\end{cases}
\]

$i \neq j$ sense codons (state space has 61 states)

$P_j$: equilibrium probability of codon $j$

By construction, this model is time-reversible: $P_i q_{ij} = P_j q_{ji}$

Typically, $\kappa > 1$, $\omega < 1$.

Example 1

$q_{\text{TTT,TTC}} = \kappa P_{\text{TTC}}$ (syn. transition), $q_{\text{AGA,CGA}} = P_{\text{CGA}}$ (syn. transversion)
$q_{\text{TTT,CTT}} = \omega \kappa P_{\text{CTT}}$ (nonsyn. transition), $q_{\text{TCT,AGT}} = 0$ (>1 mutation)
**Codon Substitution Models**

**Choices for equilibrium distribution**

1. \( P_j = 1/61 \) for all non-stop codons \( j \) (2 free parameters)

2. Codon \( xyz \) has equilibrium probability

\[
P_{xyz} = \frac{q_x q_y q_z}{\sum_{\text{non-stop codons } abc} q_a q_b q_c}
\]

where e.g. \( q_a \) is the equilibrium probability of nucleotide \( a \) (5 free parameters)

3. Codon \( xyz \) has equilibrium probability

\[
P_{xyz} = \frac{q_x^0 q_y^1 q_z^2}{\sum_{\text{non-stop codons } abc} q_a^0 q_b^1 q_c^2}
\]

where \( q_a^i \) is the equilibrium probability of nucleotide \( a \) at frame position \( i \) (11 free parameters)

4. All \( P_j \) are parameters (62 free parameters)
## Codon Substitution Models

### More general rate matrices

- above codon substitution model does not account for amino acid similarities
- the GTR for codons has \( 60 + 60 \cdot 61/2 = 1890 \) free parameters
- usually too many to estimate independently for specific data set
- dimensionality reduction through principal component analysis (PCA)
Codon Substitution Models

Selective pressure depends on the site $s$ and on the branch $b$

$$q_{ij} = \begin{cases} 
0 & \text{, if } i \text{ and } j \text{ differ at more than one position} \\
\kappa P_j & \text{, if } i \text{ and } j \text{ differ by a synonymous transversion} \\
\omega_s P_j & \text{, if } i \text{ and } j \text{ differ by a synonymous transition} \\
\omega_s^b P_j & \text{, if } i \text{ and } j \text{ differ by a nonsynonymous transversion} \\
\omega_s^b \kappa P_j & \text{, if } i \text{ and } j \text{ differ by a nonsynonymous transition} 
\end{cases}$$

For example,

- Some sites $s$ may be under positive selection while most sites of the same gene are under negative selection.
- Influenza from different hosts: may be higher selective pressure on the branch separating sequences from different hosts (adaptation required).

Can’t estimate $\omega_s^b$ for all site/branch combinations individually.

- One may want to consider branch classes for $b$, e.g., using different habitats.
- Interest in (small number of) sites $s$ with positive selection.
### Selection

#### Positive/negative selection

- **negative selection**, purifying selection
  - $\omega < 1$
  - (amino acid) changes reduce fitness and are selected against

- **positive selection**
  - $\omega > 1$
  - **example**: directional selection, allele frequency in a population shifts continuously in one direction
    - HIV-1 in a patient with antiretroviral drugs
    - size of black bears responded to climate changes
  - **example**: balancing selection
    - heterozygote advantage: sickle cell anemia/malaria
    - negative frequency-dependent selection: prey switching

- **neutral evolution**
  - $\omega \approx 1$
  - mutations not effecting fitness, e.g. “gene” not functional
Estimating $\omega = d_N/d_S$

Alignment

1. align sequences on protein level (translate nuc. seqs)
2. obtain induced alignment of nucleotide sequences
3. for each codon, ignore sequences with a gap (to avoid complications)

2 ways

1. ML-estimation of evolutionary codon substitution model:

$$\ln L = \sum_i \sum_j n_{ij} \ln P_i P_{ij}(t) \rightarrow \text{max}$$

(summation over $i$ and $j$ goes over the 61 sense codons $n_{ij}$: number of times, codons $i$ and $j$ are aligned.
Also works for multiple alignment and evolution along a tree.
The function in (2) has 3 parameters ($\omega, \tau, t$) and can be optimized numerically ($P_i$’s estimated directly from observed counts).

2. heuristic counting method for pairwise sequence comparison
Estimating $\omega = d_N/d_S$

**Heuristic counting method for pairwise sequence comparison**

- $d_S$: estimate of the number of **synonymous** substitution per “synonymous sites”
- $d_N$: estimate of the number of **nonsynonymous** substitution per “nonsynonymous sites”

1. count synonymous and nonsynonymous “sites” ($S$, $N$)
2. count synonymous and nonsynonymous differences ($D_S$, $D_N$)
3. correct for difference between observed and genetic distance ($d_N$, $d_S$)
Estimating $\omega = d_N/d_S$

**Example 2**

<table>
<thead>
<tr>
<th>human</th>
<th>H</th>
<th>G</th>
<th>G</th>
<th>P</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>CAC</td>
<td>GGT</td>
<td>GGG</td>
<td>CCA</td>
<td>AAG</td>
</tr>
<tr>
<td>mouse</td>
<td>CAT</td>
<td>GGT</td>
<td>GGC</td>
<td>CCA</td>
<td>GCG</td>
</tr>
<tr>
<td>mouse</td>
<td>H</td>
<td>G</td>
<td>G</td>
<td>P</td>
<td>A</td>
</tr>
</tbody>
</table>

**Counting sites:**

Of the 9 single-base mutations of CAC 1 is synonymous (to CAT) and 8 are nonsynonymous. Count $3 \cdot \frac{1}{9}$ synonymous and $3 \cdot \frac{8}{9}$ nonsynonymous sites.

<table>
<thead>
<tr>
<th>sites</th>
<th>CAC</th>
<th>GGT</th>
<th>GGG</th>
<th>CCA</th>
<th>AAG</th>
<th>CAT</th>
<th>GGC</th>
<th>GCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$ synonymous</td>
<td>$\frac{3}{9}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$\frac{3}{8}$</td>
<td>$\frac{3}{9}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$N$ nonsynonymous</td>
<td>$\frac{24}{9}$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>$\frac{21}{8}$</td>
<td>$\frac{24}{9}$</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Beware:** stop codons do not count. Sum over codons in sequence, average over sequences and scale so that the total number of sites is $N + S =$ sequence length.

$N = \frac{527}{48} \approx 10.98$, $S = \frac{193}{48} \approx 4.02$, $N + S = 15$
Estimating $\omega = d_N/d_S$

**Example 3**

<table>
<thead>
<tr>
<th></th>
<th>human</th>
<th>human</th>
<th>mouse</th>
<th>mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H G G P K</td>
<td>CAC GGT GGG CCA AAG</td>
<td>CAT GGT GCC CCA GGC</td>
<td>H G G P A</td>
</tr>
</tbody>
</table>

**Counting differences:**

2 codons did not change at all (GGT).
2 synonymous one-base changes (CAC - CAT, GGG - GGC).
1 two-base change (AAG - GCG), explained by 2 possible pathways of one-base changes:

<table>
<thead>
<tr>
<th>pathway</th>
<th>syn.</th>
<th>nonsyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAG (K) - GAG (E) - GCG (A)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>AAG (K) - ACG (T) - GCG (A)</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Weight all paths equally (here $1/2$, $1/2$) and average $\Rightarrow$ 2 nonsynonymous one-base changes.

**total:** $D_S = 2$ synonymous differences

$D_N = 2$ nonsynonymous differences.
**Estimating** $\omega = d_N/d_S$

**Example 4**

<table>
<thead>
<tr>
<th>Human</th>
<th>H</th>
<th>G</th>
<th>G</th>
<th>P</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>CAC</td>
<td>GGT</td>
<td>GGG</td>
<td>CCA</td>
<td>AAG</td>
</tr>
<tr>
<td>Mouse</td>
<td>CAT</td>
<td>GGT</td>
<td>GGC</td>
<td>CCA</td>
<td>GCG</td>
</tr>
<tr>
<td>Mouse</td>
<td>H</td>
<td>G</td>
<td>G</td>
<td>P</td>
<td>A</td>
</tr>
</tbody>
</table>

**Correct for difference between observed and genetic distance:**

**Observed distance**

\[
p_S = D_S/S \approx 0.50
\]
\[
p_N = D_N/N \approx 0.18
\]

In reality, some mutations may not manifest in differences (e.g. back-mutations). Use correction of JC69 model.

\[
d_S = -\frac{3}{4} \ln(1 - \frac{4}{3} p_S) \approx 0.82
\]
\[
d_N = -\frac{3}{4} \ln(1 - \frac{4}{3} p_N) \approx 0.21
\]

$\hat{\omega} := d_N/d_S = 0.26$ (negative selection)