EE550
Computational Biology

Week 4 Course Notes

Instructor: Bilge Karaçalı, PhD

Topics

- Probabilistic amino acid sequence evolution models
  - PAM matrices
    - PAM1
    - Higher order PAM matrices
  - PAM distances
  - Scoring matrices
    - PAM
    - BLOSUM

Point Accepted Mutation Matrices for Amino Acid Substitutions

- Stochastic substitution models for amino acid sequences follow the same general philosophy as the nucleic acid substitution models
  - A sequence distance D is observed between any pair of amino acid sequences
  - This distance is linked to an evolutionary distance d
- The substitution (or co-occurrence) rates are determined using a set of aligned amino acid sequences after the gap regions have been eliminated
  - The obtained rate matrices inevitably reflect the characteristics of the studied protein families

Models of amino Acid Sequence Evolution

- The language of proteins have 20 letters
- The substitution rate matrix for amino acid sequences is therefore 20×20
  - Each spot on the matrix denotes the substitution rate ($r_{i,j}$) from one amino acid (the i'th amino acid in the list of 20) to another (the j'th amino acid) in unit time
- A substitution rate matrix can be calculated by
  1. Determining the substitutions in a group of closely related protein sequences
  2. Counting the number $A_{i,j}$ of times the i'th amino acid is replaced by the j'th amino acid between two sequences, $i, j = 1,2, \dots, 20$, $i \neq j$
- Note that
  - The value of $A_{i,j}$ is proportional to $\pi_i r_{i,j}$
  - Since $A_{i,j}$ is inevitably symmetrical, we have $A_{i,j} = A_{j,i}$, providing $\pi_i r_{i,j} = \pi_j r_{j,i}$

Point Accepted Mutations: The PAM Matrices

- PAM1: Probabilities with which the i'th amino acid is replaced by the j'th amino acid in Δt units of time form a matrix M with

$$M_{i,j} = P_{i,j}(\Delta t)$$

where Δt denotes the PAM unit time

- Note that for sufficiently small Δt,

$$M_{i,j} \propto \frac{A_{i,j}}{N_i}$$

where the number of occurrences of the i'th amino acid across the sequence data is given by

$$N_i = \pi_i N_{\text{tot}}$$

with N_{tot} denoting the total amino acid number

- Let λ be such that

$$M_{i,j} = \lambda \frac{A_{i,j}}{N_i}$$

Point Accepted Mutations: The PAM Matrices

- Now, supposing that Δt corresponds to a time interval during which only 1% of amino acids would change on the average,

$$0.01 = \sum_i \pi_i \sum_{j \neq i} M_{i,j} = \sum_i \pi_i \sum_{j \neq i} \lambda \frac{A_{i,j}}{N_i} = \sum_i \pi_i \sum_{j \neq i} \lambda \frac{A_{i,j}}{\pi_i N_{\text{tot}}} = \lambda \sum_i \sum_{j \neq i} \frac{A_{i,j}}{N_{\text{tot}}} = \lambda \frac{A_{\text{tot}}}{N_{\text{tot}}}$$

$$\Rightarrow \lambda = 0.01 \frac{N_{\text{tot}}}{A_{\text{tot}}}$$

- Finally, the matrix M is obtained as

$$M_{i,j} = \begin{cases} 0.01 \dfrac{N_{\text{tot}}}{N_i} \dfrac{A_{i,j}}{A_{\text{tot}}} & \text{if } i \neq j \\ 1 - \displaystyle\sum_{k \neq i} M_{i,k} & \text{if } i = j \end{cases}$$

- The degree at which the i’th amino acid is prone to substitutions can be evaluated by its relative mutability $m_i = (1 - M_{i,i})/0.01$

Higher Order PAM Matrices

- PAM1 measures the expected number of substitutions between amino acids in 1 PAM unit of time Δt
  - where the fraction (i.e. average probability) of substitutions is about 1%
- The expected substitutions after 2 units of time $2\Delta t$ can be computed from PAM1:
  - Since $M_{i,j} = P_{i,j}(\Delta t)$,

$$P_{i,j}(2\Delta t) = \sum_{k} P_{i,k}(\Delta t) P_{k,j}(\Delta t) = \sum_{k} M_{i,k} M_{k,j}$$

  - Thus, PAM2 is equal to M^2
  - Similarly, PAM100 is M^{100}, and PAM250 is M^{250}

PAM Distances

- The number of PAM unit of times n required (on the average) to transform one amino acid sequence into another is defined as the PAM distance
  - PAM1 is the substitution matrix for 1 PAM distance
  - PAM100 is the substitution matrix for 100 PAM distances
- Using the analogy of sequence distance D and evolutionary distance d,
  - $d(n) = 0.01n$ (by definition)
  - $D(n) = \left(1 - (M^n)_{1,1}\right) \cdot \pi_1 + \left(1 - (M^n)_{2,2}\right) \cdot \pi_2 + \left(1 - (M^n)_{3,3}\right) \cdot \pi_3 + \left(1 - (M^n)_{4,4}\right) \cdot \pi_4 + \ \ldots + \left(1 - (M^n)_{20,20}\right) \cdot \pi_{20}$

$$\Rightarrow D(n) = \sum_i \pi_i\left(1 - (M^n)_{i,i}\right)$$

- This defines a one-to-one relationship between d and D that can be plotted
  - Allowing to read d that corresponds to the observed D
- An alternative formula is to use

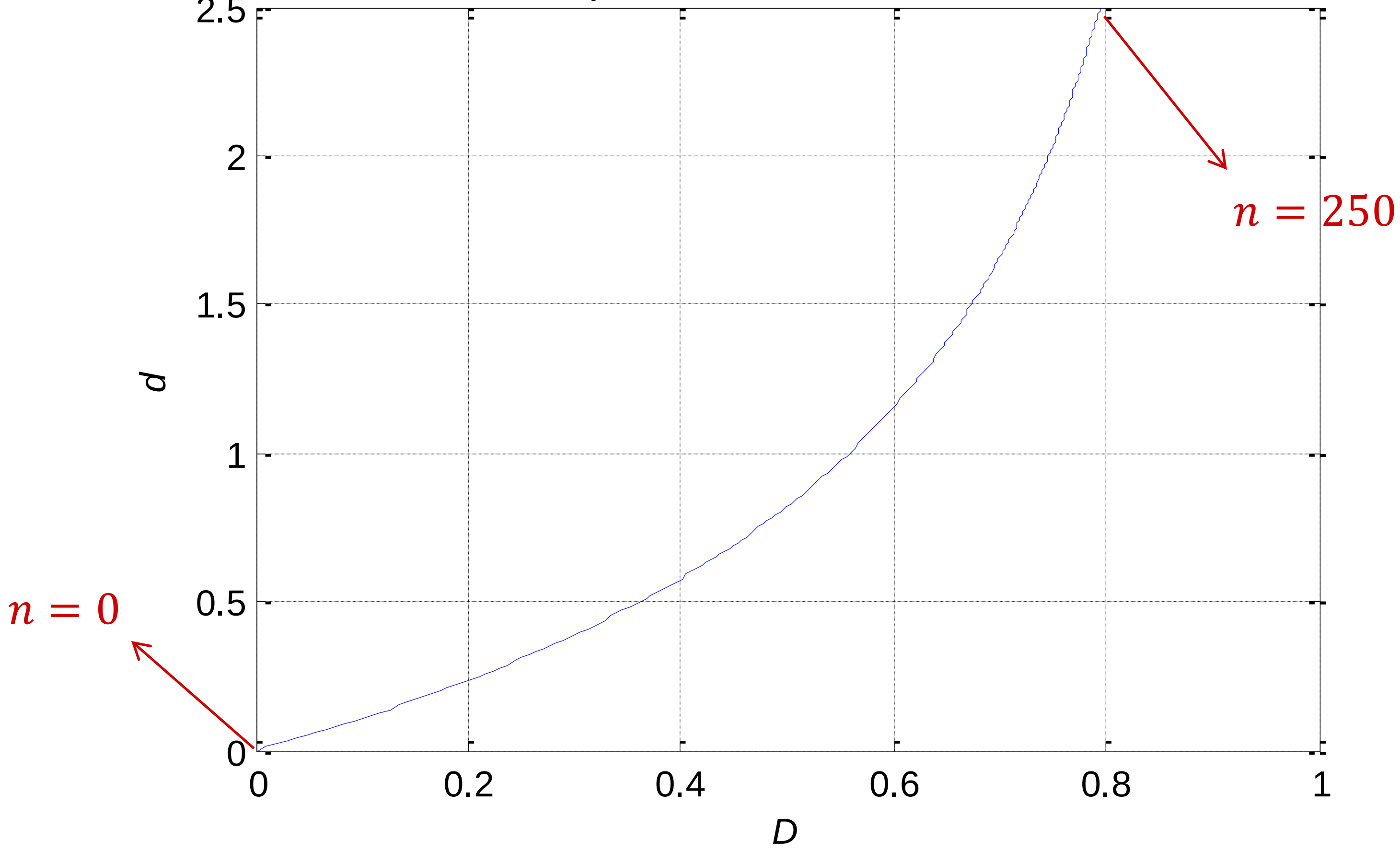
$$d = -\log(1 - D - D^2)$$

for $D < 0.7$.

- A more precise way to measure evolutionary distance uses maximum likelihood estimation for the most probable PAM distance
  - See literature

PAM Distances

$$(D,d) = \left(\sum_i \pi_i\big(1-(M^n)_{i,i}\big), 0.01n\right)$$



Scoring Matrices

- Scoring matrices are used to determine the goodness of a correspondence between amino acid sequences
  - Matches and mismatches are evaluated in terms of the likelihood of having resulted from amino acid substitutions
  - The correspondence is established in a way to maximize the likelihoods of the required substitutions
  - Contains essentially the same information as a substitution rate matrix, but in a more computationally desirable format

PAM Scoring Matrices

- For two amino acid sequences n PAM distances apart, the relative frequency

$$R_{i,j} = \frac{\pi_i (M^n)_{i,j}}{\pi_i \pi_j} = \frac{(M^n)_{i,j}}{\pi_j}$$

calculates the average number of times the i'th amino acid is aligned with the j'th, normalized against the number of times they would be aligned by chance.

- Given two sequences $\boldsymbol{A}$ and $\boldsymbol{B}$, both of length L, with $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ representing the indices of the amino acids at the k'th position in the respective sequences, the relative likelihood is defined by

$$R(\boldsymbol{A}, \boldsymbol{B}) = \prod_{k=1}^{L} R_{\boldsymbol{A}_k, \boldsymbol{B}_k}$$

where k runs from 1 to L.

  - The more similar the sequences, the higher the relative likelihood

- An additive alternative to relative frequency is the log-odds matrix given by

$$S_{i,j} = \left[C_0 \log_{C_1} R_{i,j} \right]$$

where C_0 and C_1 are chosen to produce conveniently scaled values $S_{i,j}$rounded to the nearest integer.

Example: Computation of the Point Accepted Mutation Matrices

- Data:
  - A synthetic dataset of 1000 amino acid sequences
    - Using only cysteine (C), serine (S), threonine (T), proline (P), alanine (A), and glycine (G)
    - A single original sequence of length $N = 100$ is independently evolved for 100 epochs
    - Other sequences are obtained by repeating this procedure independently 999 more times
    - The mutation probabilities for one epoch governed by a pre-determined substitution probability matrix

Example: Computation of the Point Accepted Mutation Matrices

- Procedure:
  - Substitution numbers $A_{i,j}$ will be computed for all amino acid pairs for $i, j = 1,2, \dots, 6$
    - Initially, $A_{i,j} = 0$ for all i, j
    - For each sequence pair $\boldsymbol{SQ}_k$ and $\boldsymbol{SQ}_\ell$, for all sites $n = 1,2, \dots, N$
      - Let i and j denote the indices of the amino acids $\boldsymbol{SQ}_k(n)$ and $\boldsymbol{SQ}_\ell(n)$
      - If $i \neq j$, increment $A_{i,j} \leftarrow A_{i,j} + 1$ and $A_{j,i} \leftarrow A_{j,i} + 1$
  - The PAM-1 matrix will be computed using

$$M_{i,j} = \lambda \frac{A_{i,j}}{N_i}$$

using $\lambda = 0.01 \frac{N_{\text{tot}}}{A_{\text{tot}}}$.

  - The graph of evolutionary distance d versus the sequence distance D will be determined for increasing n
    - $d = 0.01n$
    - $D = \sum_i \pi_i \left(1 - (M^n)_{i,i}\right)$

Example: Amino Acid Sequence Data

```
CCPSTASTTATAPSSTGAAGTAPGGAPGAGSGPSSCPGSSAPTSTSSASTASTTCGCTTCCGGASGSPPPSGTSCAPTASGCCGAGAPTSAGTGSGTTPS.........
CGSSTASTTATAPSSPGAAGTACGGAPGAGSGPSSCPASPAPSSTTSASTTSSGCASGACTGGGAGSPPPPGTSCAPTASGCCPGGASTPTTTTPGCTPS.........
CTPSTTSTTAACPSSPPAGGTACGGAPGAGSTPSGTPASPTPSTTSSASTTSSACAPCACCGGAGGSPPSPGTACAPTTSGCCGAAGPTPATTTPPPAPS.........
CTPAATSTTATGPTTPGAAATACGGAPGAGTACSSCPASPAPSSTSSATTTTSTCAPGTCCGGAAGSPPPPGTSCAPTASGTCGAGACTPATTTCGCTPS.........
CTPTTASTTATAPSSPGAAGTATGCAPGAGSAPTSCPASPACSSCTSGSTTSSTTASGACCGGAAGSPPCPGTTCTPTASGACGASAPGPATATPGCTPT.........
CTPSTATTTATAPSSPGAAGTTCGGAPGTGAAPSSCPASPAPSSTTSTSTTSSCCAPGACCAGSAGTPPPPGTSCGSTGSGCTCAGAPTPATTTPGCTPS.........
CTPSTTTTTTTGCSSPGAAGTACGGAPGGGSGPSSTPASSGTSSTSSASTTSSTCGPGACCGGTAGSPPPCGTSCAPTASGCCATGGCTPGTTTPGCTPS.........
CTCATASTTTTASSAPGAAGTACGGTSGAGSAPSATPASSAPSSTSSASTTSSTCAPGTCCGGAAGSSPPPGTSCASTASGCCGTGAPTPATATCPCTPS.........
STSSTTSTGATGGSSPGTGGTACGGAPGAGSAPSTCPSSPASSSTSSGSTTSATCGTGGCCGSAAGSPSPPGTSTGPTATCCCGAGASCPATCTPGCTSS.........
CTPATASTTSTAPSSSGGAGTTTGGTCGTGSAPSSCPASPAPSSTSSASTTTSTPASGACCGGAAGSPPSPGTSCGSTASCCCGGGAPTPATTTPGCTPS.........
CTPTTASTTGGAPSSPGGAGTACGGAPGACSAPSSTPASPAPSTTSTAATCSSTCAPGATCGGASGSPPPPGTSTAPTTSGCCGAGAPTPGTTTPGTTPS.........
CTCSTASTTATAPSSSGGAGTACGGAPGAGSTPSSCCASPAPTSTSSASTTSTTCAPGGCCGGAASSPAPPGTACAPTAAGCTGAGAPTPATTTCGTGPT.........
CTPATASTAACAPSSPGTGGTATSGAPGAGSAPSSCPTSPAPASCSSTSTTSSTCGPGACTGGAAPSPPPPCTSTACTASGCCGAGAPTPATTTPGTTSS.........
CTPSTATATATAPSSSGAAGTACGGAPGTGSAPSSCPGSPGPSSTSSASATSSAPAPGGCCAGAAGSPPSPGTSCASTASGCCGAGAPTPATTTPGCTPS.........
CTPTTTSTAATAPSSPGAAGAGCCGASGAGSAPSSCSTSPAPTCSSSGSTTSSTCAPGACCGGAAGSPCPCGTSPAPTATGCTGTGAPTGGCTTPGCGPS.........
CTPSTAATTACTPSSPGTAGTACGGAPGAGSACSSCPASSAPSSTATASTTSSTCAPGGCCGGTGGASPPSGTTCGTTASGCCGAGAPTSGTTTSGCGSS.........
TTPSTASTTATTPSSSGAAGTACGGASGAGSASAGCPCSCGPSATSSASTTSATCAPSACTGPAAGSPPPPGTSCAPTASGTCGAGAPAPATTTPGTTSS.........
CTPSTTSTTATAPSSPGAAATACGGAPSAGSASSSCPASPAPASATSATTTSSATAPGACCGGATGSPPPPGTSCAPTASCCCGAGGPTCTTTTSGCTPS.........
CTPSTTSTTGTAPSTSGTAGTATSGGPGAGTAPASCPASPAPSSTTSGCATATTCAPGACTGGTAGSPPPPGTSCAPTAAGCCCAGAPCPATTTPGSTPS.........
CTPSTAATTATAPSSCGAGGTACPGAPGAGAAPSSPPASPAPSTTSSASTTSTTCAPGGCCGGAAGSPPPTGTSCAPTASGCCTAGAPAPATTTPTCTPS.........
CTSSTASATATTPSSPGTAGTTCGGAPCAGSAPSSCPASPAPTSTSSASTTSTTCAPGTTCGGAAGSPSPPGTSCAPTASGCCGAGTPTPATTAPGCAPS.........
CTPSTAATTGTGPSSCGATGTGCGGGPGGGSAPSSCPASPAPSSTSAASTTSSTCAPGACPGGGAGSPPSPGTSCTSTTSGCCGAGAPTPATTATGCTPS.........
CAPSTASTAATAPSSSGAAGTACGGAPGAGSAPSSCPGTPGPSSTSSASTTSSTCAPGACTGGATGSPPPPGASCAPTASGCTAAGSPTPGCSTPGCTPS.........
CTCSTASTGATAPSSPAAAGAACGGAPGAGTAPSCTTASPAPSSASSASTTSSTCAPGAPCGGAAGAPPPPGTACAPTCSGCCGAGAPTPAATTPGCTPT.........
PTCSTASTTATAPSSPGAGGTATGGAPGGGSAPGSCPASPACSSTSSASTTTSTCAPGACCGGTAGSPPSPGTSCGPTASGCPGTGAPTSATTTPCCTPT.........
CTPSGTSTTATAPASPGAGGTTCGAAPGAGSAPSSCPGSPAPTSTSSGSTTSGACGPSGCCGGGCGSSSPCGTSCSPTATGCCGAGAPTPATTGPGCTPT.........
CTPTTASTTATAPSTPGAGGSGCGGAPGAGSGPSSCPASPAPSSTSSASTTSATCTPPATCGGPAGTPPPPGTSCAPTASGCCGAGGPTPATTTPGCTCS.........
CTCTTTSTTAGAPSSPGACGTTCCGAPGTGSAPSSCPASSASSTTSSASTGSGGCAPGTCPGGAAASSPPPGTSCAPTASGCCGTGAPCPTTTTPGCTPA.........
CTPSTGTTTATAPSSPGTAGTTTGGAPGAASAPSACPASPAPTSTSSATTTSSACAPGACCGGTAGTPPCPGTCCAPTASGCCGAGASTPATTTSCCTPS.........
.........
```

Example: Counting the Substitutions

CCPSTASTTATAPSSTGAAGTAPGGAPGAGSGPSSCPGSSAPTSTSSASTASTTCGCTTC......
CGSSTASTTATAPSSPGAAGTACGGAPGAGSGPSSCPASPAPSSTTSASTTSSGCASGAC......

$A_{1,6} \leftarrow A_{1,6} + 1,\ A_{6,1} \leftarrow A_{6,1} + 1$

CCPSTASTTATAPSSTGAAGTAPGGAPGAGSGPSSCPGSSAPTSTSSASTASTTCGCTTC......
CGSSTASTTATAPSSPGAAGTACGGAPGAGSGPSSCPASPAPSSTTSASTTSSGCASGAC......

$A_{4,2} \leftarrow A_{4,2} + 1,\ A_{2,4} \leftarrow A_{2,4} + 1$

CCPSTASTTATAPSSTGAAGTAPGGAPGAGSGPSSCPGSSAPTSTSSASTASTTCGCTTC......
CGSSTASTTATAPSSPGAAGTACGGAPGAGSGPSSCPASPAPSSTTSASTTSSGCASGAC......

$A_{3,4} \leftarrow A_{3,4} + 1,\ A_{4,3} \leftarrow A_{4,3} + 1$

CCPSTASTTATAPSSTGAAGTAPGGAPGAGSGPSSCPGSSAPTSTSSASTASTTCGCTTC......
CGSSTASTTATAPSSPGAAGTACGGAPGAGSGPSSCPASPAPSSTTSASTTSSGCASGAC......

$A_{4,1} \leftarrow A_{4,1} + 1,\ A_{1,4} \leftarrow A_{1,4} + 1$

Example: The Substitution Matrix

- $N_{\text{tot}} = 100 \cdot 1000 = 10^5$
- $A_{\text{tot}} = 30697512$
- Therefore,

$$\lambda = 0.01 \frac{N_{\text{tot}}}{A_{\text{tot}}} = 3.2576\ 10^{-5}$$

- Finally, the PAM-1 matrix M is obtained using

$$M_{i,j} = \lambda \frac{A_{i,j}}{N_i}$$

with

- $N_1 = 9722$
- $N_2 = 17864$
- $N_3 = 21011$
- $N_4 = 15988$
- $N_5 = 18975$
- $N_6 = 16440$

$$A = \begin{bmatrix} 0 & 175443 & 1528179 & 950503 & 139540 & 464977 \\ 175443 & 0 & 1943681 & 1606926 & 1068301 & 528113 \\ 1528179 & 1943681 & 0 & 441658 & 2777614 & 923035 \\ 950503 & 1606926 & 441658 & 0 & 121269 & 360350 \\ 139540 & 1068301 & 2777614 & 121269 & 0 & 2319167 \\ 464977 & 528113 & 923035 & 360350 & 2319167 & 0 \end{bmatrix}$$

$$10^4 \cdot M = \begin{bmatrix} 9891 & 6 & 51 & 32 & 5 & 16 \\ 3 & 9903 & 35 & 29 & 19 & 10 \\ 24 & 30 & 9882 & 7 & 43 & 14 \\ 19 & 33 & 9 & 9929 & 2 & 7 \\ 2 & 18 & 48 & 2 & 9890 & 40 \\ 9 & 10 & 18 & 7 & 46 & 9909 \end{bmatrix}$$

Example: Higher Order Substitution Matrices

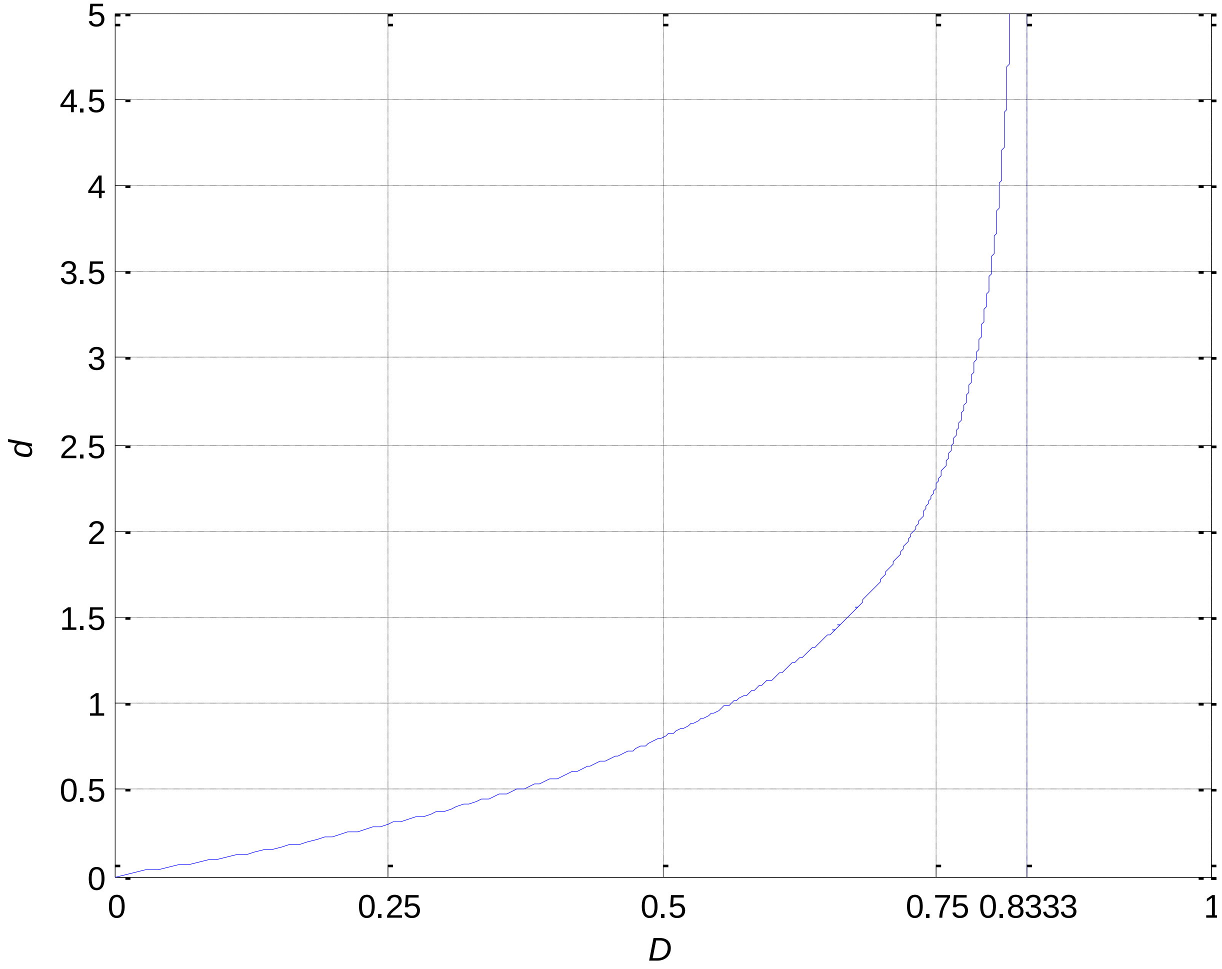
$$10^4 \cdot M^2 = \begin{bmatrix} 9783 & 12 & 101 & 63 & 10 & 31 \\ 6 & 9807 & 70 & 58 & 39 & 19 \\ 47 & 60 & 9766 & 14 & 85 & 29 \\ 38 & 65 & 18 & 9859 & 5 & 15 \\ 5 & 36 & 94 & 4 & 9781 & 79 \\ 18 & 21 & 37 & 14 & 91 & 9819 \end{bmatrix}$$

$$10^4 \cdot M^5 = \begin{bmatrix} 9468 & 31 & 246 & 154 & 25 & 76 \\ 17 & 9526 & 171 & 142 & 95 & 48 \\ 114 & 146 & 9428 & 35 & 207 & 71 \\ 94 & 159 & 46 & 9652 & 13 & 36 \\ 13 & 90 & 229 & 11 & 9465 & 192 \\ 45 & 52 & 91 & 35 & 222 & 9555 \end{bmatrix}$$

$$10^4 \cdot M^{10} = \begin{bmatrix} 8969 & 65 & 467 & 296 & 55 & 148 \\ 36 & 9081 & 328 & 274 & 186 & 95 \\ 216 & 279 & 8900 & 71 & 394 & 140 \\ 180 & 306 & 93 & 9320 & 29 & 72 \\ 28 & 175 & 437 & 24 & 8968 & 368 \\ 87 & 103 & 179 & 70 & 424 & 9136 \end{bmatrix}$$

$$10^4 \cdot M^{20} = \begin{bmatrix} 8061 & 143 & 844 & 548 & 126 & 278 \\ 78 & 8269 & 604 & 508 & 354 & 187 \\ 390 & 514 & 7960 & 145 & 717 & 274 \\ 334 & 568 & 190 & 8702 & 66 & 141 \\ 65 & 333 & 794 & 56 & 8078 & 674 \\ 165 & 203 & 350 & 137 & 778 & 8367 \end{bmatrix}$$

Example: Evolutionary Distance versus Sequence Distance



Example: PAM Scoring Matrices

- For two amino acid sequences that are n PAM distances apart,

$$R_{i,j} = \frac{(M^n)_{i,j}}{\pi_j}$$

- An additive alternative is the following log-odds matrix:

$$S_{i,j} = \left[10 \log_{10} R_{i,j}\right]$$

- The relative likelihood of two sequences $\boldsymbol{A}$ and $\boldsymbol{B}$ computed by

$$R(\boldsymbol{A}, \boldsymbol{B}) = \sum_{k} S_{\boldsymbol{A}_k, \boldsymbol{B}_k}$$

$$R\Big|_{n=20} = \begin{bmatrix} 8.2911 & 0.0799 & 0.4015 & 0.3430 & 0.0664 & 0.1694 \\ 0.0799 & 4.6289 & 0.2876 & 0.3178 & 0.1865 & 0.1138 \\ 0.4015 & 0.2876 & 3.7885 & 0.0905 & 0.3780 & 0.1667 \\ 0.3430 & 0.3178 & 0.0905 & 5.4426 & 0.0350 & 0.0856 \\ 0.0664 & 0.1865 & 0.3780 & 0.0350 & 4.2574 & 0.4098 \\ 0.1694 & 0.1138 & 0.1667 & 0.0856 & 0.4098 & 5.0896 \end{bmatrix} \qquad S = \begin{bmatrix} 9 & -11 & -4 & -5 & -12 & -8 \\ -11 & 7 & -5 & -5 & -7 & -9 \\ -4 & -5 & 6 & -10 & -4 & -8 \\ -5 & -5 & -10 & 7 & -15 & -11 \\ -12 & -7 & -4 & -15 & 6 & -4 \\ -8 & -9 & -8 & -11 & -4 & 7 \end{bmatrix}$$

BLOSUM Scoring Matrices

- Provides log-odds matrices similar to PAM
  - directly from alignment data by counting the number of aligned amino acids, and
  - without invoking an evolutionary model (much like $(M^n)_{i,j}$ in PAM)
- Given a set of aligned sequences,
  - Obtain the priors π_i by counting the number of times each amino acid is observed divided by the total number of amino acids in the sequence data
  - Determine the total number of times the i’th amino acid is aligned with the j’th amino acid as $A_{i,j}$, including cases where $i = j$
  - Compute the fraction of i-j alignments using

$$q_{i,j} = \frac{A_{i,j}}{A_{\text{tot}}}$$

  where $A_{\text{tot}} = \sum_{i,j} A_{i,j}$.

  - Compute the relative frequencies via

$$R_{i,j} = \frac{q_{i,j}}{\pi_i \pi_j}$$

Example: BLOSUM Scoring Matrices

- Procedure:
  - Determine the total number of times the i'th amino acid is aligned with the j'th amino acid as $A_{i,j}$
  - Compute the fraction of instances where the i'th and the j'th amino acids are aligned

$$q_{i,j} = \frac{A_{i,j}}{A_{\text{tot}}}$$

  where $A_{\text{tot}} = \sum_{i,j} A_{i,j}$.

  - Compute the relative frequencies

$$R_{i,j} = \frac{q_{i,j}}{\pi_i \pi_j}$$

  - Compute the log-odds matrix *S* using

$$S_{i,j} = \left[2 \log_2 R_{i,j}\right]$$

Example: BLOSUM Scoring Matrices

CCPSTASTTATAPSSTGAAGTAPGGAPGAGSGPSSCPGSSAPTSTSSASTASTTCGCTTC......
CGSSTASTTATAPSSPGAAGTACGGAPGAGSGPSSCPASPAPSSTTSASTTSSGCASGAC......

$A_{1,1} \leftarrow A_{1,1} + 2$

CCPSTASTTATAPSSTGAAGTAPGGAPGAGSGPSSCPGSSAPTSTSSASTASTTCGCTTC......
CGSSTASTTATAPSSPGAAGTACGGAPGAGSGPSSCPASPAPSSTTSASTTSSGCASGAC......

$A_{1,6} \leftarrow A_{1,6} + 1,\ A_{6,1} \leftarrow A_{6,1} + 1$

CCPSTASTTATAPSSTGAAGTAPGGAPGAGSGPSSCPGSSAPTSTSSASTASTTCGCTTC......
CGSSTASTTATAPSSPGAAGTACGGAPGAGSGPSSCPASPAPSSTTSASTTSSGCASGAC......

$A_{4,2} \leftarrow A_{4,2} + 1,\ A_{2,4} \leftarrow A_{2,4} + 1$

CCPSTASTTATAPSSTGAAGTAPGGAPGAGSGPSSCPGSSAPTSTSSASTASTTCGCTTC......
CGSSTASTTATAPSSPGAAGTACGGAPGAGSGPSSCPASPAPSSTTSASTTSSGCASGAC......

$A_{2,2} \leftarrow A_{2,2} + 2$

Example: BLOSUM Scoring Matrices

$$A = \begin{bmatrix} 6453636 & 175443 & 1528179 & 950503 & 139540 & 464977 \\ 175443 & 12523672 & 1943681 & 1606926 & 1068301 & 528113 \\ 1528179 & 1943681 & 13375822 & 441658 & 2777614 & 923035 \\ 950503 & 1606926 & 441658 & 12491306 & 121269 & 360350 \\ 139540 & 1068301 & 2777614 & 121269 & 12530134 & 2319167 \\ 464977 & 528113 & 923035 & 360350 & 2319167 & 11827918 \end{bmatrix}$$

$$q = \begin{bmatrix} 0.0646 & 0.0018 & 0.0153 & 0.0095 & 0.0014 & 0.0047 \\ 0.0018 & 0.1254 & 0.0195 & 0.0161 & 0.0107 & 0.0053 \\ 0.0153 & 0.0195 & 0.1339 & 0.0044 & 0.0278 & 0.0092 \\ 0.0095 & 0.0161 & 0.0044 & 0.1250 & 0.0012 & 0.0036 \\ 0.0014 & 0.0107 & 0.0278 & 0.0012 & 0.1254 & 0.0232 \\ 0.0047 & 0.0053 & 0.0092 & 0.0036 & 0.0232 & 0.1184 \end{bmatrix}$$

$$R = \begin{bmatrix} 6.8348 & 0.1011 & 0.7489 & 0.6121 & 0.0757 & 0.2912 \\ 0.1011 & 3.9283 & 0.5184 & 0.5632 & 0.3155 & 0.1800 \\ 0.7489 & 0.5184 & 3.0329 & 0.1316 & 0.6974 & 0.2675 \\ 0.6121 & 0.5632 & 0.1316 & 4.8916 & 0.0400 & 0.1372 \\ 0.0757 & 0.3155 & 0.6974 & 0.0400 & 3.4836 & 0.7442 \\ 0.2912 & 0.1800 & 0.2675 & 0.1372 & 0.7442 & 4.3807 \end{bmatrix}$$

$$S = \begin{bmatrix} 6 & -7 & -1 & -1 & -7 & -4 \\ -7 & 4 & -2 & -2 & -3 & -5 \\ -1 & -2 & 3 & -6 & -1 & -4 \\ -1 & -2 & -6 & 5 & -9 & -6 \\ -7 & -3 & -1 & -9 & 4 & -1 \\ -4 & -5 & -4 & -6 & -1 & 4 \end{bmatrix}$$

Reading the Scoring Matrices

- Both PAM and BLOSUM log-odds scoring matrices quantify the likelihood of substituting one amino acid with another one
- These matrices can thus be used to identify the amino acid substitution characteristics in the studied sequence dataset
  - Conservation:

    PAM : (most conserved) C, P, G, S, A, T (least conserved)

    BLOSUM : (most conserved) C ,P, G, S, A, T (least conserved)
  - Substitution:

    PAM : (most common) A-G, G-A, C-T, T-C, T-A, A-T, C-P, P-C, S-P, P-S, S-T, T-S, S-A, A-S, C-G, G-C, T-G, G-T, G-S, S-G, P-T, T-P, P-G, G-P, S-C, C-S, C-A, A-C, P-A, A-P (least common)

    BLOSUM: (most common) C-T, T-C, A-G, G-A, T-A, A-T, C-P, P-C, S-P, P-S, S-T, T-S, S-A, A-S, C-G, G-C, T-G, G-T, S-G, G-S, P-G, G-P, T-P, P-T, C-S, S-C, C-A, A-C, P-A, A-P (least common)
- This information is extremely useful for sequence alignment!!
  - aligning amino acids that rarely occur at the same site is strongly avoided to improve the overall alignment score

Summary

- Sequence evolution models try to capture the mechanism of substitution mutations
  - Models of nucleic acid substitution are derived from a statistical evolution model
  - Amino acid sequence substitution models are derived from a given set of aligned sequence data
- Generally, the models put similar sequences in close evolutionary proximity
- The estimates of evolutionary distances, however, are subject to errors