UNIVERSITÄT
DES
SAARLANDES

ZBI ZENTRUM FÜR
BIOINFORMATIK

Score Matrices

Algorithms for Sequence Analysis

Sven Rahmann
(with material from Tobias Müller, Würzburg)

Summer 2021

Overview

Previously: Pairwise Sequence Alignment

- score maximization with general scoring schemes,
- Four variants: global, semiglobal, overlapping, local

Overview

Previously: Pairwise Sequence Alignment

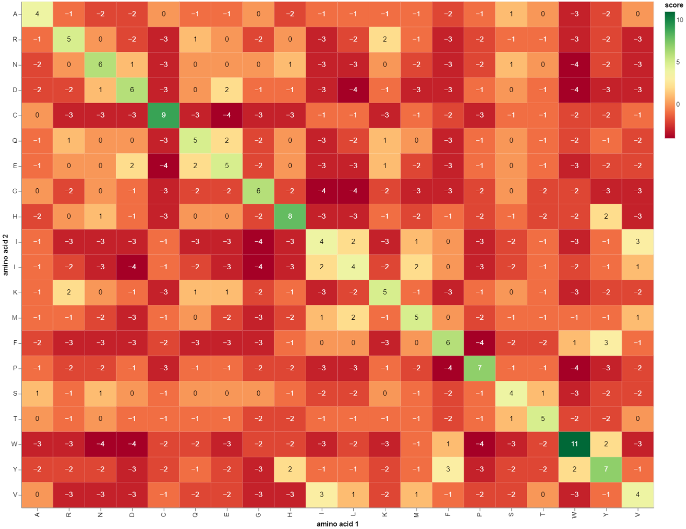
- score maximization with general scoring schemes,
- Four variants: global, semiglobal, overlapping, local

Today's Lecture: Score Matrices

- Where do (families of) score matrices (like BLOSUM62) come from?
- Evolutionary distances (units PAM, PEM)
- Excursion: Time-continuous Markov processes, matrix exponentials
- Estimation of score matrices from alignments of different divergence times
- General vs. special purpose score matrices

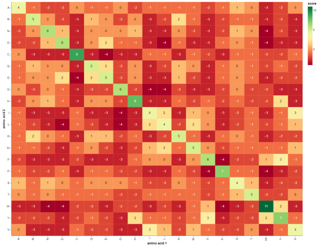
Score Matrices for Comparing Proteins

Example: BLOSUM62 Scoring Matrix for Amino Acids



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	−1	−2	−2	0	−1	−1	0	−2	−1	−1	−1	−1	−2	−1	1	0	−3	−2	0
R	−1	5	0	−2	−3	1	0	−2	0	−3	−2	2	−1	−3	−2	−1	−1	−3	−2	−3
N	−2	0	6	1	−3	0	0	0	1	−3	−3	0	−2	−3	−2	1	0	−4	−2	−3
D	−2	−2	1	6	−3	0	2	−1	−1	−3	−4	−1	−3	−3	−1	0	−1	−4	−3	−3
C	0	−3	−3	−3	9	−3	−4	−3	−3	−1	−1	−3	−1	−2	−3	−1	−1	−2	−2	−1
Q	−1	1	0	0	−3	5	2	−2	0	−3	−2	1	0	−3	−1	0	−1	−2	−1	−2
E	−1	0	0	2	−4	2	5	−2	0	−3	−3	1	−2	−3	−1	0	−1	−3	−2	−2
G	0	−2	0	−1	−3	−2	−2	6	−2	−4	−4	−2	−3	−3	−2	0	−2	−2	−3	−3
H	−2	0	1	−1	−3	0	0	−2	8	−3	−3	−1	−2	−1	−2	−1	−2	−2	2	−3
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4	2	−3	1	0	−3	−2	−1	−3	−1	3
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4	−2	2	0	−3	−2	−1	−2	−1	1
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5	−1	−3	−1	0	−1	−3	−2	−2
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5	0	−2	−1	−1	−1	−1	1
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6	−4	−2	−2	1	3	−1
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7	−1	−1	−4	−3	−2
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4	1	−3	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5	−2	−2	0
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11	2	−3
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	−1
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

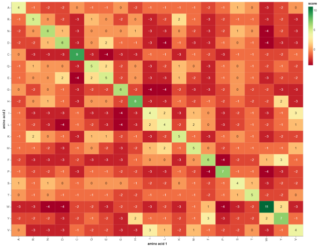
How Score Matrices are Obtained



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

- **Idea:** Physically and chemically similar pairs of amino acids have positive score, dissimilar pairs have negative score. Zero is a neutral value.
- However, who is to quantify “similarity”? Experts ?

How Score Matrices are Obtained



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

- **Idea:** Physically and chemically similar pairs of amino acids have positive score, dissimilar pairs have negative score. Zero is a neutral value.
- However, who is to quantify "similarity"? Experts ?
- Instead, use data-driven approach. (Today, you call this machine learning.)
- Observe the (relative) frequencies of amino acids in proteins.
- Observe the joint frequencies of amino acids in confirmed alignments: Similar amino acids more often replace each other than dissimilar amino acids.

Score Matrices from Observed Alignments

Counting amino acid replacements in confirmed alignments

```
24.7% identity in 97 aa overlap;  score: 109

ref   IFLHDNAPSHTARAVRDTLETLNWEVLPHAAYSPDLAPSDYHLFASMGHALAEQRFDSYESVKKWLDEWFAAKDDEFYWRGIHKLPERWEKCVASDG
       .: .:: :.::.  ::. ..    ..:   . :::: : . ::.  . . :.  : .. ..  . :. . :       . : ..:.: .  . ..:
query VFQQDNDPKHTSLHVRSWFDRRFVDLLDWPSQSPDLNPIE-HLWEELERRLGGIRASNADAKFNQLPNAWKAIPMSVIHKLIDSMPRRCQAVIDANG
```

Score Matrices from Observed Alignments

Counting amino acid replacements in confirmed alignments

```
24.7% identity in 97 aa overlap;  score: 109

ref   IFLHDNAPSHTARAVRDTLETLNWEVLPHAAYSPDLAPSDYHLFASMGHALAEQRFDSYESVKKWLDEWFAAKDDEFYWRGIHKLPERWEKCVASDG
       .: .:: :.::.  ::. ..    ..:   . :::: : . ::.  . . :.  : .. ..  . :. . :       . : ..:.: .  . ..:
query VFQQDNDPKHTSLHVRSWFDRRFVDLLDWPSQSPDLNPIE-HLWEELERRLGGIRASNADAKFNQLPNAWKAIPMSVIHKLIDSMPRRCQAVIDANG
```

Examples of pair counts

$$\#(D, E) = 4$$

$$\#(N, F) = 1$$

Score Matrices from Observed Alignments

Counting amino acid replacements in confirmed alignments

```
24.7% identity in 97 aa overlap;  score: 109

ref   IFLHDNAPSHTARAVRDTLETLNWEVLPHAAYSPDLAPSDYHLFASMGHALAEQRFDSYESVKKWLDEWFAAKDDEFYWRGIHKLPERWEKCVASDG
       .: .:: :.::.  ::. ..    ..:   . :::: : . ::.  . . :.  : .. ..  . :. . :       . : ..:.: .  . ..:
query VFQQDNDPKHTSLHVRSWFDRRFVDLLDWPSQSPDLNPIE-HLWEELERRLGGIRASNADAKFNQLPNAWKAIPMSVIHKLIDSMPRRCQAVIDANG
```

Examples of pair counts

$$\#(D, E) = 4$$
$$\#(N, F) = 1$$

Assumption (for now)

All alignments used for counting have the same **degree of divergence** (evolutionary distance). Otherwise, they are not be comparable.

Markov Model of Protein Evolution

$$\begin{array}{ccccccccc}
\cdots & \mathsf{A} & \mathsf{S} & \mathsf{A} & \mathsf{R} & \mathsf{D} & \mathsf{S} & \mathsf{D} & \cdots \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\
\cdots & \mathsf{D} & \mathsf{S} & \mathsf{D} & \mathsf{A} & \mathsf{A} & \mathsf{S} & \mathsf{D} & \cdots \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\
\cdots & \mathsf{D} & \mathsf{S} & \mathsf{D} & \mathsf{R} & \mathsf{A} & \mathsf{S} & \mathsf{D} & \cdots \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\
\cdots & \mathsf{A} & \mathsf{E} & \mathsf{D} & \mathsf{A} & \mathsf{D} & \mathsf{S} & \mathsf{D} & \cdots
\end{array}$$

Assumptions

- Replacement probabilities at any site depend only on the amino acid at that site and on transition probabilities, but not on the history (past) at that site.
- Sequences have average (typical) amino acid composition.
- Time-reversible process (direction of time arrow is irrelevant)

Markov Model of Protein Evolution

Model Parameters

- Amino acid frequencies $\pi = (\pi_i)_{i=1}^{20}$ (row vector)
- Conditional transition probabilities for a fixed time unit ("1 step"):

$$P = (P_{ij}) \qquad (i = 1, \dots, 20,\ j = 1, \dots, 20)$$

  with $\sum_{j=1}^{20} P_{ij} = 1$ for all i (rows sum to 1).

- One step of transitions must not change overall frequencies, i.e., π must be the/a **stationary distribution** for P:

$$\sum_{i=1}^{20} \pi_i \cdot P_{ij} = \pi_j \qquad \text{or} \qquad \pi \cdot P = \pi \,.$$

- Symmetric **joint** (or pair) frequencies $J_{ij} = \pi_i \cdot P_{ij} = \pi_j \cdot P_{ji} = J_{ji}$ (symmetry of J: **time-reversibility**)

Parameter Estimation

Procedure

- Estimate $J = (J_{ij})$ symmetrically by counting pairs of amino acids in alignments. Normalize, such that $\sum_{i,j} J_{ij} = 1$ (probability distribution over pairs), i.e., divide by total number N of observed pairs.
- Obtain π as marginals of J, i.e., $\pi_i = \sum_j J_{ij}$
- Obtain P by normalizing row sums of J to 1.

Parameter Estimation

Procedure

- Estimate $J = (J_{ij})$ symmetrically by counting pairs of amino acids in alignments. Normalize, such that $\sum_{i,j} J_{ij} = 1$ (probability distribution over pairs), i.e., divide by total number N of observed pairs.
- Obtain π as marginals of J, i.e., $\pi_i = \sum_j J_{ij}$
- Obtain P by normalizing row sums of J to 1.

Problem

- Procedure above assumes that all observed alignments have the same divergence time ("one step").
- We will generalize this in a moment; for now, stick with the assumption.

Derivation of Score Matrix

Log-Odds Scores

- Compare the **observed** joint frequencies in real alignments with the **expected** pair frequencies based on amino acid frequencies:

$$\text{Quotient or enrichment or odds:} \quad J_{ij}/(\pi_i \cdot \pi_j)$$

Derivation of Score Matrix

Log-Odds Scores

- Compare the **observed** joint frequencies in real alignments with the **expected** pair frequencies based on amino acid frequencies:

$$\text{Quotient or enrichment or odds:} \quad J_{ij}/(\pi_i \cdot \pi_j)$$

- Bring to additive scale by taking logarithm (e.g., base 2):

$$\text{Log-odds score} \quad \tilde{S}_{ij} = \log_2\left(\frac{J_{ij}}{\pi_i \cdot \pi_j}\right) \quad \text{[bits]}$$

Derivation of Score Matrix

Log-Odds Scores

- Compare the **observed** joint frequencies in real alignments with the **expected** pair frequencies based on amino acid frequencies:

$$\text{Quotient or enrichment or odds:} \quad J_{ij}/(\pi_i \cdot \pi_j)$$

- Bring to additive scale by taking logarithm (e.g., base 2):

$$\text{Log-odds score} \quad \tilde{S}_{ij} = \log_2\left(\frac{J_{ij}}{\pi_i \cdot \pi_j}\right) \quad \text{[bits]}$$

- Scale and round to integer:

$$\text{Score} \quad S_{ij} = \mathrm{rd}\left(2\log_2\left(\frac{J_{ij}}{\pi_i \cdot \pi_j}\right)\right) \quad \text{[half-bits]}$$

Score Matrices of Different Divergence Times

Remember

Amino acid score matrices are rounded scaled log-odds scores,
comparing observed joint frequencies with expected pair frequencies from marginals.

$$\text{Score} \quad S_{ij} = \mathrm{rd}\Big(3 \log_2 \Big(\frac{J_{ij}}{\pi_i \cdot \pi_j}\Big)\Big) \quad \text{[third-bits]}$$

Score Matrices of Different Divergence Times

Remember

Amino acid score matrices are rounded scaled log-odds scores,
comparing observed joint frequencies with expected pair frequencies from marginals.

$$\text{Score} \quad S_{ij} = \mathrm{rd}\left(3\log_2\left(\frac{J_{ij}}{\pi_i \cdot \pi_j}\right)\right) \quad \text{[third-bits]}$$

Problem of Different Divergence Times: J_{ij} are mixed

```
VCKITPHSSNKSYPDGVYGTSGSANDDKQDAPHYIGTLDMTAFGSLFHEDDFELNFGTAK ...
:::::::. .::.::::::: :::: .:::::.:::::::::::::::::.:::.:::.:
VCKITPHAPHKSHPDGVYGTPGSANADRQDAPNYIGTLDMTAFGSLFHEDEFELTFGTTK ...
```

$$\#(D,E) = 1$$
$$\#(N,F) = 0$$

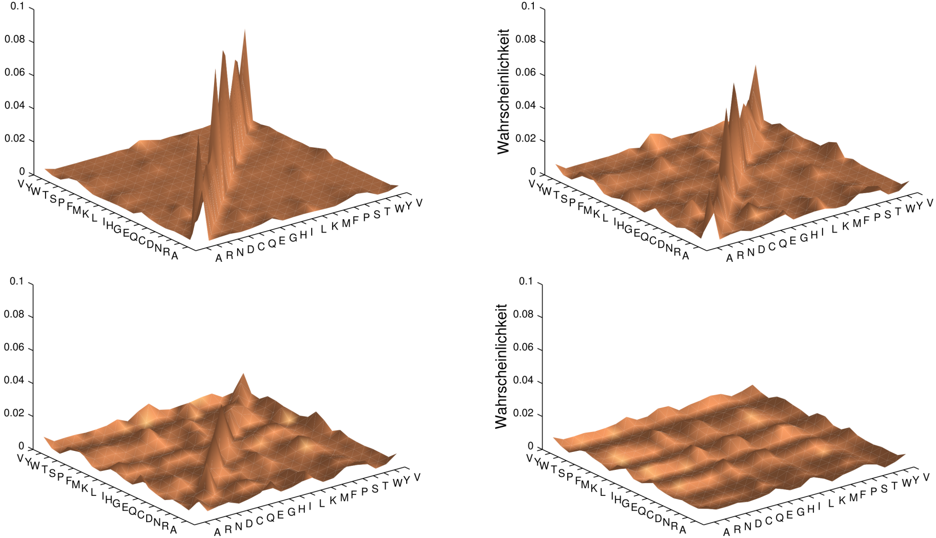
```
KLNELIPTRLDRKGLQSGGKVDRYQDEKYRKVGSPYFKKSHARKLAGSLTSDAITTLVRA ...
....:   ::.: :::::::. :: . .    :  :   :..: : :. :.::   ::..
RVSDLYGIRLERAGLQSGGKLARYVEASLTTHGLAYNMASRTRLLQGAHTGDASDGLVKT ...
```

$$\#(D,E) = 3$$
$$\#(N,F) = 1$$

```
PKNDSHTQVKEGTEQTFVLPKAHAASKLVEDLLGAGVDSKPNGAYTQESDPSSVPEGVTD ...
:. .. :  :.:.  . :  . ::.  ..  . : .:..:  :   ...::  . .:. .
PQFEGFTTGKDGAPLAAVQKQYHATVMFIVMMGGFAVEQKGFGFRGSDKDPCHTSHGLLE ...
```

$$\#(D,E) = 5$$
$$\#(N,F) = 2$$

Example: Joint Frequencies at Different Divergence Times



Dealing with Different Divergence Times

BLOSUM way: Accept, limit and mix

- Pool confirmed alignments with relatively high identity (e.g., BLOSUM62: at least 62% sequence identity in alignment)
- Mix alignments with identity above threshold to estimate J, P, π, S as above.

Dealing with Different Divergence Times

BLOSUM way: Accept, limit and mix

- Pool confirmed alignments with relatively high identity (e.g., BLOSUM62: at least 62% sequence identity in alignment)
- Mix alignments with identity above threshold to estimate J, P, π, S as above.

PAM way (Dayhoff): Strictly limit, slightly mix, normalize, extrapolate

- Pool confirmed alignments with very high identity (e.g., $> 85\%$)
- Estimate transition probability matrix $P = (P_{ij})$
- Normalize P such that the overall amount of change is 1%: $\sum_{i \neq j} J_{ij} = \sum_i \pi_i \sum_{j \neq i} P_{ij} \overset{!}{=} 1/100$ (1 PAM = percent accepted mutations)
- Set $P_{ii} := 1 - \sum_{j \neq i} P_{ij}$
- Defines a convenient PAM “unit of evolution”, extrapolate to longer times.

More on Markov Processes

Long-term transition probabilities

- $P = P^{(1)}$: one-step (1 PAM) transition probabilities
- What happens in 120 PAM (time period 120× longer than 1 PAM)?

More on Markov Processes

Long-term transition probabilities

- $P = P^{(1)}$: one-step (1 PAM) transition probabilities
- What happens in 120 PAM (time period 120× longer than 1 PAM)?
- **Not** 120% of accepted mutations:
  Multiple substitutions, back-substitutions: I → L → F → I

More on Markov Processes

Long-term transition probabilities

- $P = P^{(1)}$: one-step (1 PAM) transition probabilities
- What happens in 120 PAM (time period 120× longer than 1 PAM)?
- **Not** 120% of accepted mutations:
  Multiple substitutions, back-substitutions: I → L → F → I

Markov Chain evolution: Chapman-Kolmogorov Equations

- Given $P = P^{(1)}$, what is $P^{(t)}$, or especially $P^{(2)}$?

More on Markov Processes

Long-term transition probabilities

- $P = P^{(1)}$: one-step (1 PAM) transition probabilities
- What happens in 120 PAM (time period $120\times$ longer than 1 PAM)?
- **Not** 120% of accepted mutations:
  Multiple substitutions, back-substitutions: I $\rightarrow$ L $\rightarrow$ F $\rightarrow$ I

Markov Chain evolution: Chapman-Kolmogorov Equations

- Given $P = P^{(1)}$, what is $P^{(t)}$, or especially $P^{(2)}$?
- To move $i \rightarrow j$ in 2 PAM, move $i \mapsto k \mapsto j$ for some k:
  $P^{(2)}_{ij} = \sum_k P^{(1)}_{ik} \cdot P^{(1)}_{kj}$ or $P^{(2)} = P^{(1)} \cdot P^{(1)}$.

More on Markov Processes

Long-term transition probabilities

- $P = P^{(1)}$: one-step (1 PAM) transition probabilities
- What happens in 120 PAM (time period 120× longer than 1 PAM)?
- **Not** 120% of accepted mutations:
  Multiple substitutions, back-substitutions: I $\to$ L $\to$ F $\to$ I

Markov Chain evolution: Chapman-Kolmogorov Equations

- Given $P = P^{(1)}$, what is $P^{(t)}$, or especially $P^{(2)}$?
- To move $i \to j$ in 2 PAM, move $i \mapsto k \mapsto j$ for some k:
  $P^{(2)}_{ij} = \sum_k P^{(1)}_{ik} \cdot P^{(1)}_{kj}$ or $P^{(2)} = P^{(1)} \cdot P^{(1)}$.
- In general, $P^{(s+t)} = P^{(s)} \cdot P^{(t)}$ (Chapman-Kolmogorov)
- Therefore, $P^{(t)} = P^t$, the t-th power of P ($t \in \mathbb{N}$), and $P^{(0)} = P^0 = \mathrm{Id}$.

Dayhoff's Extrapolation Method

PAM Matrix Family

- Remember: $P = P^{(1)}$ was created from mixed closely related alignments, artificially normalized to 1 PAM.
- Now $P^{(120)} = P^{120}$ extrapolates to time span 120 PAM.
- Limits: $P^{(0)} = \mathrm{Id}$ (no change), and $P_{ij}^{(\infty)} = \pi_i \cdot \pi_j$.

Dayhoff's Extrapolation Method

PAM Matrix Family

- Remember: $P = P^{(1)}$ was created from mixed closely related alignments, artificially normalized to 1 PAM.
- Now $P^{(120)} = P^{120}$ extrapolates to time span 120 PAM.
- Limits: $P^{(0)} = \mathrm{Id}$ (no change), and $P_{ij}^{(\infty)} = \pi_i \cdot \pi_j$.

Disadvantages

- The estimation procedure cannot utilize distant alignments.
- Rare replacements are too infrequent to resolve transition probabilities accurately.
- Errors in 1 PAM matrix are magnified in the extrapolation to 120 PAM.

Dayhoff's Extrapolation Method

PAM Matrix Family

- Remember: $P = P^{(1)}$ was created from mixed closely related alignments, artificially normalized to 1 PAM.
- Now $P^{(120)} = P^{120}$ extrapolates to time span 120 PAM.
- Limits: $P^{(0)} = \mathrm{Id}$ (no change), and $P_{ij}^{(\infty)} = \pi_i \cdot \pi_j$.

Disadvantages

- The estimation procedure cannot utilize distant alignments.
- Rare replacements are too infrequent to resolve transition probabilities accurately.
- Errors in 1 PAM matrix are magnified in the extrapolation to 120 PAM.

$\Rightarrow$ Design method to utilize alignments of varying divergence (Tobias Müller, ca. 2000)

Continuous-Time Markov Processes

Rate matrices

- For real numbers p, we have $p^n = p \cdot \dots \cdot p$ (n times). Generalize to real exponents $t \in \mathbb{R}$ by $p^t = \exp(t \cdot \log p)$.
- Can it be done for matrices, too?

Continuous-Time Markov Processes

Rate matrices

- For real numbers p, we have $p^n = p \cdot \dots \cdot p$ (n times). Generalize to real exponents $t \in \mathbb{R}$ by $p^t = \exp(t \cdot \log p)$.
- Can it be done for matrices, too?
- **Definition**: A matrix Q such that $\exp(Q) = P$ is called **rate matrix** or **infinitesimal generator** of the Markov process, or **matrix logarithm**.
- Does not exist for every matrix, but does for positive definite matrices. Compute by diagonalization and taking logarithm of each (positive) eigenvalue.
- Matrix exponential for square matrices defined by power series

$$\exp(Q) := \mathrm{Id} + Q + Q^2/2 + \dots + Q^k/k! + \dots$$

- Property: $\exp(tQ) \cdot \exp(sQ) = \exp((s+t)Q)$

Understanding the Rate Matrix

Linear approximation for small $t > 0$

- $P^t = \exp(tQ) = \mathrm{Id} + tQ + t^2 Q^2/2 + \cdots + t^k Q^k/k! + \cdots \approx \mathrm{Id} + tQ$
- Also, $Q = \lim_{t \searrow 0} \left(P^{(t)} - \mathrm{Id}\right)/t = P'^{(0)}$
- Therefore, Q contains the rates describing how fast P^t_{ij} changes near $t = 0$.

Understanding the Rate Matrix

Linear approximation for small $t > 0$

- $P^t = \exp(tQ) = \mathrm{Id} + tQ + t^2Q^2/2 + \cdots + t^kQ^k/k! + \cdots \approx \mathrm{Id} + tQ$
- Also, $Q = \lim_{t \searrow 0} \left(P^{(t)} - \mathrm{Id}\right)/t = P'^{(0)}$
- Therefore, Q contains the rates describing how fast P^t_{ij} changes near $t = 0$.

Some Properties

- Valid rate matrix has $Q_{ij} > 0$ for $i \neq j$ and $Q_{ii} < 0$ for all i.
- Zero row sums: $\sum_j Q_{ij} = 0$ or $Q_{ii} = -\sum_{j \neq i} Q_{ij} < 0$

Understanding the Rate Matrix

Linear approximation for small $t > 0$

- $P^t = \exp(tQ) = \mathrm{Id} + tQ + t^2 Q^2/2 + \cdots + t^k Q^k/k! + \cdots \approx \mathrm{Id} + tQ$
- Also, $Q = \lim_{t \searrow 0} \left(P^{(t)} - \mathrm{Id}\right)/t = P'^{(0)}$
- Therefore, Q contains the rates describing how fast P^t_{ij} changes near $t = 0$.

Some Properties

- Valid rate matrix has $Q_{ij} > 0$ for $i \neq j$ and $Q_{ii} < 0$ for all i.
- Zero row sums: $\sum_j Q_{ij} = 0$ or $Q_{ii} = -\sum_{j \neq i} Q_{ij} < 0$

Alternative scaling or calibration

- So far: Scale Q such that $P = \exp(Q)$ has 1 PAM (amount of change)
- Alternative: Scale Q such that $\sum_i \pi_i \sum_{j \neq i} Q_{ij} = 1/100$, unit of 1 PEM (percent expected mutation events)

Different Expressions for the Rate Matrix

How to obtain Q from the family $P(t)$?

$$Q = \log(P^{(t)})/t \qquad \text{(any } t > 0\text{)}$$

$$Q = \alpha \cdot \mathrm{Id} - \left(\int_0^\infty \mathrm{e}^{-\alpha t}\, P^{(t)}\, \mathrm{d}t \right)^{-1} \qquad \text{(any } \alpha > 0\text{)}$$

The integral is called the **resolvent** (or **Laplace transform**) of $P^{(t)}$, $t > 0$.

Different Expressions for the Rate Matrix

How to obtain Q from the family $P(t)$?

$$Q = \log(P^{(t)})/t \qquad \text{(any } t > 0\text{)}$$

$$Q = \alpha \cdot \mathrm{Id} - \left(\int_0^\infty \mathrm{e}^{-\alpha t} \, P^{(t)} \, \mathrm{d}t \right)^{-1} \qquad \text{(any } \alpha > 0\text{)}$$

The integral is called the **resolvent** (or **Laplace transform**) of $P^{(t)}$, $t > 0$.

Why is the resolvent representation useful?

The resolvent expression integrates (in a weighted manner) over **all** times t.
We can use alignments of different degrees of divergence and adjust weighting via α.

Estimation of Q based on the Resolvent Expression

1. Start with an initial rate matrix Q and pairwise alignments (A_i)
2. Calculate empirical transition matrix $P_{(i)}$ from A_i for all i
3. Estimate divergence time t_i for A_i using existing rates Q
4. Combine different $P_{(i)}^{(t_i)}$ with approximately equal t_i
   (fewer time points, but better $P^{(t)}$ estimates at each time point t)
5. Estimate the resolvent $R_\alpha = \int_0^\infty \mathrm{e}^{-\alpha t} P^{(t)} \,\mathrm{d}t \approx \sum_t \mathrm{e}^{-\alpha t} P^{(t)}$
   for different $\alpha > 0$
6. Select "best" parameter α^* by Maximum-Likelihood-like procedure
7. Set $Q := \alpha^* \cdot \mathrm{Id} - R_{\alpha^*}^{-1}$
8. Iterate steps 3 – 7 until Q converges

Estimating the Divergence Time

Problem

Given alignment A (yielding empirical transition matrix $\hat{P} = \hat{P}^{(t)}$); rate matrix Q, **estimate** divergence time t, such that $\exp(tQ) \approx \hat{P}$

Estimating the Divergence Time

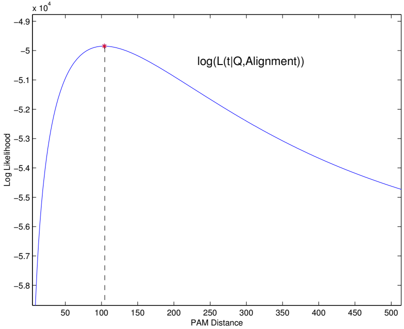
Problem

Given alignment A (yielding empirical transition matrix $\hat{P} = \hat{P}^{(t)}$); rate matrix Q, **estimate** divergence time t, such that $\exp(tQ) \approx \hat{P}$

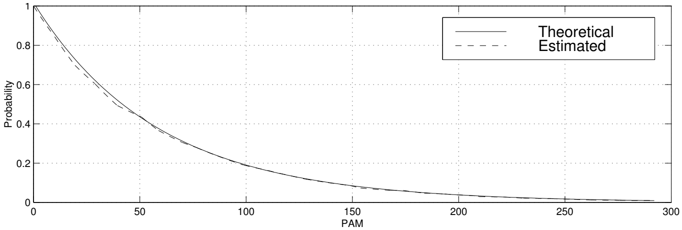
Approach: Maximum Likelihood

Probability of observing A, given Q and time t:
$\mathbf{P}(A \mid Q, t) = \prod_{i,j} (\exp(tQ)_{ij})^{\hat{P}_{ij}}$.

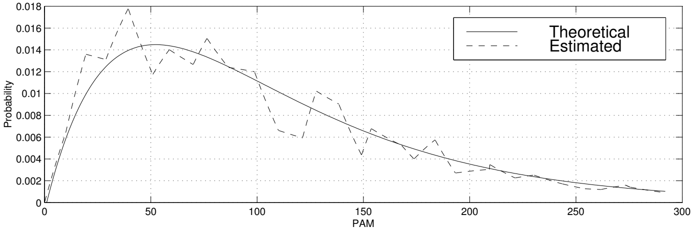
Log-likelihood of t:
$\log L(t \mid Q, A) = \sum_{i,j} \hat{P}_{ij} \cdot \log[\exp(tQ)_{ij}]$
→ maximize



Integrating Different $P^{(t)}$ with the Resolvent $\int e^{-\alpha t} P_{ij}^{(t)} \, dt$



i = j



i ≠ j

Picking Parameter α and Other Details

Interpretation of α

- Controls speed of exponential decay of weighting: $\sum_t e^{-\alpha t} P^{(t)}$
- Small α: High weight to large divergence times
  Large α: Small weight to large divergence times
- Pick α to let $e^{-\alpha t}$ fit amount of alignment data at each time t.
- Can use a maximum-likelihood-like approach or curve fitting

Picking Parameter α and Other Details

Interpretation of α

- Controls speed of exponential decay of weighting: $\sum_t e^{-\alpha t} P^{(t)}$
- Small α: High weight to large divergence times
  Large α: Small weight to large divergence times
- Pick α to let $e^{-\alpha t}$ fit amount of alignment data at each time t.
- Can use a maximum-likelihood-like approach or curve fitting

Starting with an initial Q

- Initially, choose Q such that all rates are equal,
  calibrate to 1 PEM or or 1 PAM.
- Gives approximate divergence time estimates for first iteration.

Again: Resolvent Estimation of Q

1. Start with an initial rate matrix Q and pairwise alignments (A_i)
2. Calculate empirical transition matrix $P_{(i)}$ from A_i, for all i
3. Estimate divergence time t_i for A_i using existing rates Q
4. Combine different $P_{(i)}^{(t_i)}$ with approximately equal t_i
   (fewer time points, but better $P^{(t)}$ estimates at each time point t)
5. Estimate the resolvent $R_\alpha = \int_0^\infty \mathrm{e}^{-\alpha t}\, P^{(t)}\, \mathrm{d}t \approx \sum_t \mathrm{e}^{-\alpha t}\, P^{(t)}$
   for different $\alpha > 0$
6. Select "best" parameter α^* by Maximum-Likelihood-like procedure
7. Set $Q := \alpha^* \cdot \mathrm{Id} - R_{\alpha^*}^{-1}$
8. Iterate steps 3 – 7 until Q converges

Families of Score Matrices

Score matrices have a time parameter t

$$\tilde{S}_{ij}^{(t)} = \log_2 \frac{J_{ij}^{(t)}}{\pi_i \cdot \pi_j} = \log_2 \frac{\pi_i P_{ij}^{(t)}}{\pi_i \cdot \pi_j} = \log_2 \frac{\exp(tQ)_{ij}}{\pi_j} \quad \text{[bits]}$$

- PAM family indexed by t (in PAM units), Dayhoff method
- VTML family indexed by t (in PAM units), resolvent method
- BLOSUM family indexed by percent identity (no rate matrix)

Families of Score Matrices

Score matrices have a time parameter t

$$\tilde{S}_{ij}^{(t)} = \log_2 \frac{J_{ij}^{(t)}}{\pi_i \cdot \pi_j} = \log_2 \frac{\pi_i P_{ij}^{(t)}}{\pi_i \cdot \pi_j} = \log_2 \frac{\exp(tQ)_{ij}}{\pi_j} \quad \text{[bits]}$$

- PAM family indexed by t (in PAM units), Dayhoff method
- VTML family indexed by t (in PAM units), resolvent method
- BLOSUM family indexed by percent identity (no rate matrix)

Non-symmetric score matrices?

- So far, score matrices were symmetric, but sequence roles in alignments may differ.
- Search with a transmembrane domain (τ) in a general protein database (π):
  may want $\tilde{S}_{ij}^{(t)} = \log_2 \frac{J_{ij}^{(t)}}{\tau_i \cdot \pi_j}$ (e.g., SLIM matrix family, Müller et al., 2001)

Summary

Score Matrices

- Joint frequencies J, transition probabilities P, marginal probabilities π
- (Scaled, rounded) Log-odds scores S
- Evolutionary time units: PAM, PEM
- Markov processes, Chapman-Kolmogorov equation $P^{(s+t)} = P^{(s)} \cdot P^{(t)}$
- Rate matrix Q and time-t transition matrices: $P^{(t)} = P^t = \exp(tQ)$
- Resolvent (Laplace transform) method allows estimation of Q from alignments of varying divergence.
- Symmetric general-purpose vs. (perhaps non-symmetric) special-purpose matrices

Possible Exam Questions

- Define joint frequencies J, transition probabilities P, marginal probabilities π.
- Describe how to compute log-odds scores.
- Explain how the BLOSUM and PAM matrix families were constructed.
- What is a rate matrix?
- Define the evolutionary time units 1 PAM and 1 PEM.
- How can Q be expressed in terms of P or all $P^{(t)}$?
- How can you estimate the divergence time of an observed alignment?
- What are the advantages of the resolvent method for estimating Q?
- Are score matrices symmetric? Why? When not?