Markov Chains, substitution matrices, ..

Andrew Torda wintersemester 2007/08

- Aim : make the best possible alignments
- What is the philosophy underlying substitution matrices ?

- What do substitution matrices do ? proteins

```
. . . D A F A R A D C D M A . .
. . A D C F A - G D Q R M A .
```

- how similar
  - are **C** and **A** ?
  - the **F** / **F** match ?
- this can be quantified
- how important are alignments ?

much from : Ewens, W.J. and Grant, G.R. 2001, Statistical Methods in Bioinformatics, Springer-Verlag, New York

Importance of correct alignments

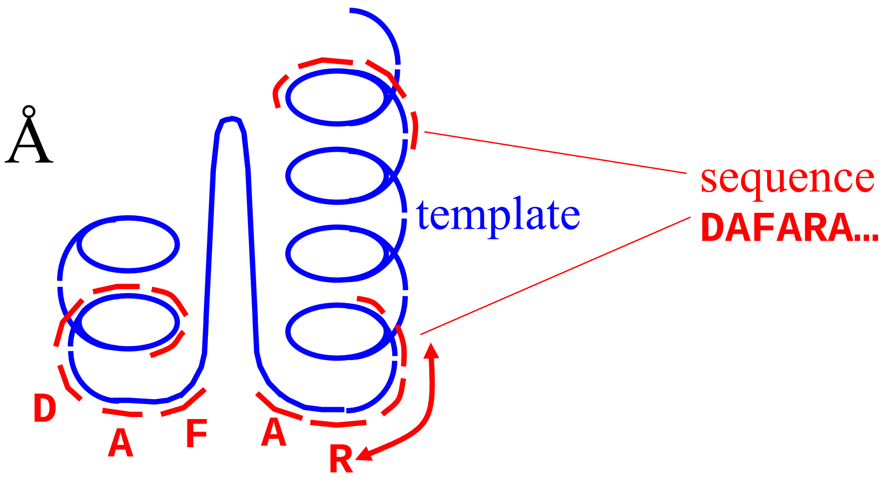
- As sequences:

  . . . D A F A R A D C D M A . .

  . . A D C F A - G D Q R M A .

- In structural terms:
  - moving one residue is 3.8 Å

- When else do we care ?
  - you have a protein
    - vital to your chemistry
    - no idea about fold …



What do we know from nucleotides ?

- Typical nucleotide matrix
  - boring
  - no knowledge of specific mutations

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

- why is the idea obviously bad for proteins ?
  - example
  - D (asp, small, acidic)
    - does it mutate to W (trp, large hydrophobic) ?
    - does it mutate to E (glu, small, acidic) ? yes
  - imagine …

- what does a full matrix look like ?

	D	E	W	...
D	10	5	-5	
E	5	10	-5	
W	-5	-5	15	
...				

A serious protein similarity matrix

- blosum62:

	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

- some features
  - diagonal
  - similar
  - different

Model for mutation

- A series of evolutionary steps

```
x x x E x x x
x x x E x x x
x x x N x x x
x x x N x x x
x x x K x x x
x x x L x x x
x x x W x x x
```

- different protein sample

```
x x x E x x x
x x x L x x x
x x x L x x x
x x x W x x x
```

- a table that tells us about direct mutations
  - **A** → **E**
- but also indirect
  - not **A** → **S** → **T** → **A** → **D** → **E**
- other terminology.. Markov chains / matrices

Markov chains / matrices / nomenclature

- nomenclature
  - time t
  - a set of possible states E_1, E_2, E_3, …
- Markov chain
  - series of steps from $E(t)$, $E(t+\delta t)$, $E(t+2\delta t)$, ..
- rule
  - state at $t+\delta t$ depends on now, t, not $t-\delta t$
  - no memory / inertia / history
- in state E_j now,
  - probability of being in state E_k at $t+ \delta t$ is p_{jk}

Markov Chains

- From each state, system can move to another state with a certain probability p_{jk}
- my system may not disappear
  - at each step, my total population must remain the same

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \cdots & \cdots & \cdots & \cdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{bmatrix}$$

A markov transition matrix ?

	D	E	…	W
D	p_{DD}	p_{DE}	…	p_{DW}
E	p_{ED}	p_{EE}	…	p_{EW}
…	…	…	…	…
W	p_{WD}	p_{WE}	…	p_{WW}

- a simple / initial substitution matrix is a true transition probability matrix
- this places restrictions on relevant data
  - **D** → **E**
  - not **D** → **S** → **T** → **A** → **D** → **E**

- rows should sum to 1 $\sum_j p_{ij}$

Applying a matrix

- three types of amino acid E, D, W
- population E, D, W = 0.4, 0.4, 0.2

$$\mathbf{P} = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.3 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

- at time $t+\delta t$

$$\begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.3 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \begin{bmatrix} 0.4 \\ 0.4 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 0.6 \times 0.4 + 0.3 \times 0.4 + 0.1 \times 0.2 \\ 0.3 \times 0.4 + 0.6 \times 0.4 + 0.1 \times 0.2 \\ 0.1 \times 0.4 + 0.1 \times 0.4 + 0.8 \times 0.2 \end{bmatrix}$$

Properties and definitions

- What happens if we have two steps ? $\begin{bmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{bmatrix}$

$$\begin{bmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 9/16 + 1/8 & 3/16 + 1/8 \\ 3/8 + 1/4 & 1/8 + 1/4 \end{bmatrix}$$

$$= \begin{bmatrix} 11/16 & 5/16 \\ 5/8 & 3/8 \end{bmatrix}$$

- the rows still sum to 1 $\sum_j p_{ij}$

Stationary Distribution

- Apply matrix multiplication infinitely
  - what would happen ? (biological case - aperiodic)

- Informal arguments
  - whatever you are (A, C, G, T or A, C, D, E, G, H… W, Y)
  - add up all the probabilities which lead to "A"
  - eventually the system will stop changing
  - can be argued (and solved) formally

Stationary Distribution

- argument similar to detailed balance
  - there is a set of probabilities for leaving state *i*, p_{ix}
  - a set of probabilities for entering state *i*, p_{xi}
  - a population in state *i*, π_i
  - the decrease in population depends on $p_{ix}\pi_i$
  - if π_i were big, $p_{ix}\pi_i$ is big
    - π_i decreases until $p_{ix}\pi_i = p_{xi}\pi_{(not\ i)}$
- nomenclature.. $\mathbf{P}^n$ where $n \rightarrow \infty$
- biological sense ?

Stationary Distribution

- biological sense
  - we survey all proteins and find gly = 5%, trp=2%, …
  - this is the stationary distribution
- I start with one protein (not near stationary distribution)
  - it evolves forever - becomes a pure random sequence
  - will not happen
  - for real proteins *n* (time / generations) is finite or
    - you die

- More properties of these processes

Broken Matrices

- What if rows do not sum to one ?

$$\mathbf{P} = \begin{bmatrix} 3/4 & 1/8 \\ 1/2 & 1/2 \end{bmatrix}$$

$$\begin{bmatrix} 3/4 & 1/8 \\ 1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 3/4 & 1/8 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 9/16 + 1/16 & 3/32 + 1/16 \\ 3/8 + 1/4 & 1/16 + 1/4 \end{bmatrix}$$

$$= \begin{bmatrix} 5/8 & 5/32 \\ 5/8 & 5/16 \end{bmatrix}$$

- the p_{ij} values will get smaller and smaller
  - the sequence will disappear

    - could have made a version which increases

Unlikely matrices

- rows all sum to 1

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.3 & 0.7 \\ 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \end{bmatrix}$$

- if I am in state 1 or 2
  - will move to 3 or 4 (and vice versa)
- this is a periodic Markov matrix
- does not happen in sequences (or statistical mechanics (usually))
- we believe
  - transition matrices for sequences are "aperiodic"

Absorbing states

- I start in state i
- eventually reach state 2
  - cannot escape
- state 2 is an absorbing state
- what is stationary distribution ?

$$\mathbf{P} = \begin{bmatrix} 0.2 & 0.3 & 0.25 & 0.25 \\ 0 & 1 & 0 & 0 \\ 0.3 & 0.1 & 0.1 & 0.5 \\ 0.2 & 0.3 & 0.2 & 0.3 \end{bmatrix}$$

Summary of properties

- rows sum to 1 $\sum_j p_{ij}$
- processes are not periodic
- there are no absorbing states
- infinite number of mutations either
  - does not occur or
  - you die
- DNA world: small 4×4 matrix
- proteins 20×20

Applications

- basis of calculating evolutionary distances
- philosophy of substitution matrices
- chemistry

Stationary distribution in chemistry

- who really invented Markov chains ?
- stationary distribution ? easy

$$\pi_i = \frac{e^{-E_i/kT}}{\sum_{j=0}^{N_{states}} e^{-E_j/kT}}$$

- transition matrix
  - not uniquely determined
  - sometimes estimated (simulations)

$$= \frac{e^{-E_i/kT}}{Z}$$

Applications / Summary

- chemistry / physics
- evolutionary models - phylogeny
  - state (residue / base) now depends on previous generation
- substitution matrices

```
. . . D A F A R A D C D M A . .
  . . A D C F A - G D Q R M A .
```

- **C**→**D** probability in one generation ? 100 generations ?

- Restrictions
  - periodicity / absorbing states

- Differences to sequence analysis people