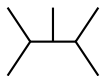


Computational Biology and Bioinformatics

- 10.10 Models of substitution I: Basic Models
- 12.10 Models of substitution II: Complex Models
- 17.10 Phylogenies I: Combinatorics
- 19.10 Phylogenies II: Distance, parsimony & likelihood
- 24.10 Ancestral recombination graphs & pedigrees
- 26.10 Alignment algorithms I: Optimisation
- 31.10 Alignment algorithms II: Statistical
 - 2.11 Hidden Markov models & stochastic grammars
 - 7.11 RNA structures
 - 9.11 Finding signals in sequences
- 14.11 Challenges in genome annotation
- 16.11 Networks: dynamics & inference
- 21.11 Networks: evolution
- 23.11 Models of evolution of structure & movements & shapes & grammars
- 28.11 Integrative genomics: the omics
- 30.11 Integrative genomics: mapping

A

↓
T



ACT-T
-GTCT



Course

Teaching

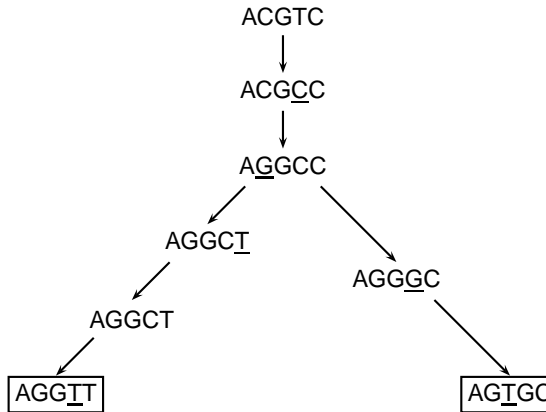
- Self contained
- Contains probability theory, combinatorics, algorithmics & mathematics
- Two lectures and one class each week
- Independence encouraged
- Research oriented

Mini Project Examination

- Expected to be 3 days' worth of work & approximately 7–10 pages
- Will be given in week 8
- Choice between a set of topics related to taught material
- Topic assignment will contain 2–4 key references and a set of guiding questions
- **Examples:** Comparison of Networks, Grammar Evolution Model, Automated Annotation of Genes, Stochastic Context Free Grammars in RNA Secondary Structure Prediction, Substitution Models with Rate Heterogeneity, Probability Theory of Networks, Inference of Gene Regulatory Networks using Differential Equations, Network flow and its applications to the analysis of metabolism, Identification of Regulatory Elements – Phylogenetic Footprinting, Phylogeny Reconstruction – Distance-based Methods, Probable and Improbable Paths in Sequence Evolution & Probabilistic Methods for DNA Sequence Alignment

Central Problem: History cannot be observed, only end products

ACGTC
↓
ACGCC
↓
AGGCC
↓
AGGCT
↓
AGGCT
↓
AGGIT



Even if history could be observed, the underlying process couldn't

Simplifying Assumptions

Probability of data

$$P = \mathbb{P} \left(\begin{array}{c} \text{Common ancestor } a \text{ (unknown)} \\ \diagdown \quad \diagup \\ \text{TCGGTA} \quad \text{TGGTT} \end{array} \right)$$

1) Independent lineages

$$P = \sum_a \mathbb{P}(a) \cdot \mathbb{P}(a \rightarrow \text{TCGGTA}) \cdot \mathbb{P}(a \rightarrow \text{TGGTT})$$

2) Only substitutions

$$\begin{array}{l} s_1 \quad \text{TCGGTA} \\ s_2 \quad \text{TGGT-T} \end{array} \quad \rightarrow \quad \begin{array}{l} s_1 \quad \text{TCGGA} \\ s_2 \quad \text{TGGTT} \end{array}$$

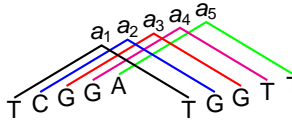
$$P = \sum_a \mathbb{P}(a) \cdot \mathbb{P}(a \rightarrow \text{TCGGA}) \cdot \mathbb{P}(a \rightarrow \text{TGGTT})$$

Simplifying Assumptions

Probability of data

$$P = \mathbb{P} \left(\begin{array}{c} \text{Common ancestor } a \text{ (unknown)} \\ \swarrow \quad \searrow \\ \text{TCGGA} \quad \text{TGGTT} \end{array} \right)$$

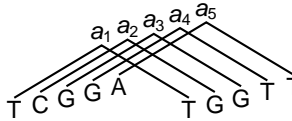
3) Processes in different positions are independent



A phylogenetic tree with a root node labeled 'Common ancestor a (unknown)'. The tree branches into two lineages. The left lineage has nodes labeled a_1, a_2, a_3, a_4 and ends with the sequence TCGGA. The right lineage has nodes labeled a_5 and ends with the sequence TGGTT. Colored lines connect the nodes to their respective positions in the sequences: blue for the first position (T), red for the second (C), green for the third (G), pink for the fourth (G), and black for the fifth (A/T).

$$P = \prod_{i=1}^5 \sum_{a_i} \mathbb{P}_i(a_i) \cdot \mathbb{P}_i(a_i \rightarrow s_{1,i}) \cdot \mathbb{P}_i(a_i \rightarrow s_{2,i})$$

4) Processes in different positions are identical



A phylogenetic tree with a root node labeled 'Common ancestor a (unknown)'. The tree branches into two lineages. The left lineage has nodes labeled a_1, a_2, a_3, a_4 and ends with the sequence TCGGA. The right lineage has nodes labeled a_5 and ends with the sequence TGGTT. Lines connect the nodes to their respective positions in the sequences, with all lines being black.

$$P = \prod_{i=1}^5 \sum_{a_i} \mathbb{P}(a_i) \cdot \mathbb{P}(a_i \rightarrow s_{1,i}) \cdot \mathbb{P}(a_i \rightarrow s_{2,i})$$

Simplifying Assumptions

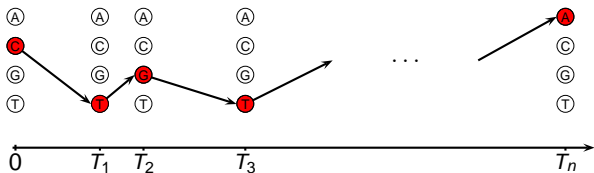
Probability of data

$$P = \mathbb{P} \left(\begin{array}{c} \text{Common ancestor } a \text{ (unknown)} \\ \swarrow \quad \searrow \\ \text{TCGGA} \quad \text{TGGTT} \end{array} \right)$$

5) Time reversibility ($\forall i, j, t : \pi_i \cdot \mathbb{P}_t(i \rightarrow j) = \pi_j \cdot \mathbb{P}_t(j \rightarrow i)$) where π_i is stationary distribution and P_t probability of change in time t)

$$\text{TCGGA} \xrightarrow{t_1} \text{TGGTT} \quad P = \prod_{i=1}^5 \mathbb{P}(s_{1,i}) \cdot \mathbb{P}_{t_1+t_2}(s_{1,i} \rightarrow s_{2,i})$$

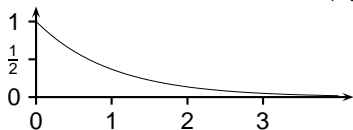
6) Process is time-homogeneous continuous time Markov chain with exponential waiting time



Exponential Distribution and Jump Probabilities

Exponential distribution

$$X \sim \text{Exp}(a) : \mathbb{P}(X > t) = e^{-at} = \sum_{i=0}^{\infty} \frac{(-at)^i}{i!} = 1 - at + o((at)^2)$$



Important properties

$\{X_i \sim \text{Exp}(a_i)\}_{i=1}^n$ independent

- $\mathbb{P}(X > T_2 \mid X > T_1) = \mathbb{P}(X > T_2 - T_1)$ if $T_2 \geq T_1$
- $\min_{i \in \{1, \dots, n\}} \{X_i\} \sim \text{Exp}(\sum_{i=1}^n a_i)$
- $\mathbb{P}(X_i = \min_{j \in \{1, \dots, n\}} \{X_j\}) = a_i / \sum_{j=1}^n a_j$

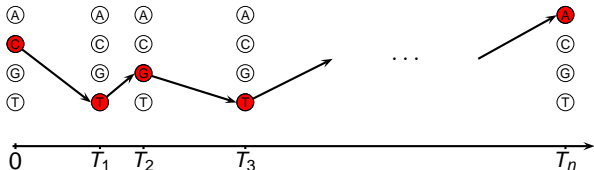
Exponential Distribution and Jump Probabilities

Important properties

$\{X_i \sim \text{Exp}(a_i)\}_{i=1}^n$ independent

- $\mathbb{P}(X > T_2 \mid X > T_1) = \mathbb{P}(X > T_2 - T_1)$ if $T_2 \geq T_1$
- $\min_{i \in \{1, \dots, n\}} \{X_i\} \sim \text{Exp}(\sum_{i=1}^n a_i)$
- $\mathbb{P}(X_i = \min_{j \in \{1, \dots, n\}} \{X_j\}) = a_i / \sum_{j=1}^n a_j$

Jump probabilities and simulation



If q_{ij} is rate of going from state i to state j and are in state i ,

- $t' = t + \text{Exp}(\sum_{j \neq X(t)} q_{X(t),j})$
- $X(t') = [q_{X(t),1}, q_{X(t),2}, \dots, q_{X(t),k}] / \sum_{j \neq i} q_{ij}$

The Rate Matrix

Let $P_{ij}(t) = \mathbb{P}(X(t) = j \mid X(0) = i)$ be transition probability after time t

$$P(t) \approx \mathbf{I} + \mathbf{Q}t = \mathbf{I} + \begin{bmatrix} -\sum_{j \neq 1} q_{1j} & q_{12} & \cdots & q_{1k} \\ q_{21} & -\sum_{j \neq 2} q_{2j} & \cdots & q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k1} & q_{k2} & \cdots & -\sum_{j \neq k} q_{kj} \end{bmatrix} \cdot t \text{ for small } t$$

Transition probability matrix derivative

$$\begin{aligned} P'(t) &= \lim_{h \rightarrow 0} \frac{P(t+h) - P(t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{P(t)P(h) - P(t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{P(t)(R(h) - \mathbf{I})}{h} \\ &= \mathbf{Q}P(t) \end{aligned}$$

From Q to $P(t)$

Differential equation approach

Only solution to $P'(t) = QP(t)$ is

$$P(t) = e^{Qt} = \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!} = \mathbf{I} + Qt + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \dots$$

Limit process approach

If $\frac{t}{n}$ is small

$$P(t) = P(t/n)^n \approx \left(\mathbf{I} + Q \frac{t}{n} \right)^n = \sum_{i=0}^n \binom{n}{i} \frac{(Qt)^i}{n^i} = \sum_{i=0}^n \prod_{j=1}^{i-1} \frac{n-j}{n} \frac{(Qt)^i}{i!}$$

Boundary conditions & important properties

- $P(0) = \mathbf{I} \Rightarrow P'(0) = Q$
- Equilibrium frequency $\pi_i = \lim_{t \rightarrow \infty} P(t)_{ji}$ (and any normal 0 left eigenvector of Q will be in equilibrium)
- Time and rates are inseparable with no further information ($Qt = \frac{Q}{r}(rt)$)

Jukes–Cantor (JC69): Total symmetry

Rate matrix

$$Q = \begin{array}{c|cccc} & \text{A} & \text{C} & \text{G} & \text{T} \\ \hline \text{A} & -3\alpha & \alpha & \alpha & \alpha \\ \text{C} & \alpha & -3\alpha & \alpha & \alpha \\ \text{G} & \alpha & \alpha & -3\alpha & \alpha \\ \text{T} & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

Transition probabilities

$$P(t) = \frac{1}{4} \begin{array}{c|cccc} & \text{A} & \text{C} & \text{G} & \text{T} \\ \hline \text{A} & 1 + 3e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} \\ \text{C} & 1 - e^{-4\alpha t} & 1 + 3e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} \\ \text{G} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 + 3e^{-4\alpha t} & 1 - e^{-4\alpha t} \\ \text{T} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 + 3e^{-4\alpha t} \end{array}$$

$$\pi = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right]$$

Jukes–Cantor (JC69): Total symmetry

Transition probabilities

$$P(t) = \frac{1}{4} \begin{bmatrix} & \text{A} & \text{C} & \text{G} & \text{T} \\ \text{A} & 1 + 3e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} \\ \text{C} & 1 - e^{-4\alpha t} & 1 + 3e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} \\ \text{G} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 + 3e^{-4\alpha t} & 1 - e^{-4\alpha t} \\ \text{T} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 + 3e^{-4\alpha t} \end{bmatrix}$$

$$\pi = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right]$$

Data example

$$P = \mathbb{P} \left(\begin{array}{c} \text{Common ancestor } a \text{ (unknown)} \\ \swarrow \quad \searrow \\ \text{TCGGA} \quad \text{TGGTT} \end{array} \right) = \left(\frac{1}{4} \right)^5 \left(\frac{1}{4} \right)^5 \left(1 + 3e^{-4\alpha t} \right)^2 \left(1 - e^{-4\alpha t} \right)^3$$

Matrix Exponentiation

Diagonalisation

If $Q = BLB^{-1}$ with $L = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_k \end{bmatrix}$ then $Q^i = BLB^{-1}BLB^{-1} \dots BLB^{-1} = BL^iB^{-1}$

$$\text{Hence } \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!} = \sum_{i=0}^{\infty} \frac{(BLB^{-1}t)^i}{i!} = B \sum_{i=0}^{\infty} \frac{(Lt)^i}{i!} B^{-1} = B \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & e^{\lambda_k t} \end{bmatrix}$$

Jukes-Cantor

$$P(t) = \begin{bmatrix} 1 & \frac{1}{4} & 0 & 1 \\ 1 & \frac{1}{4} & 0 & -1 \\ 1 & -\frac{1}{4} & 1 & 0 \\ 1 & -\frac{1}{4} & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{-4\alpha t} & 0 & 0 \\ 0 & 0 & e^{-4\alpha t} & 0 \\ 0 & 0 & 0 & e^{-4\alpha t} \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{1}{8} & -\frac{1}{8} \\ 0 & 0 & 1 & -1 \\ 1 & -1 & 0 & 0 \end{bmatrix}$$

Matrix Exponentiation

Numerical approximation

$$\sum_{i=0}^{\infty} \frac{(Qt)^i}{i!} \approx \sum_{i=0}^k \frac{(Qt)^i}{i!} \text{ with } k \sim 6 - 10$$

Matrix specific analysis (Jukes–Cantor)

$$Q = \alpha \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix}, \forall i > 0 : Q^i = (-4\alpha)^{i-1} Q$$

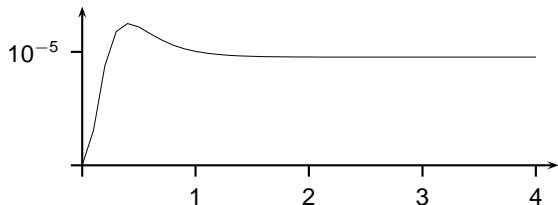
$$\begin{aligned} \sum_{i=0}^{\infty} \frac{(Qt)^i}{i!} &= \mathbf{I} + \sum_{i=1}^{\infty} \frac{1}{-4\alpha} \frac{(-4\alpha t)^i}{i!} Q = \mathbf{I} - \frac{1}{4\alpha} \left(Q \sum_{i=0}^{\infty} \frac{(-4\alpha t)^i}{i!} - Q \right) \\ &= \mathbf{I} - \frac{1}{4\alpha} Q \left(e^{-4\alpha t} - 1 \right) = \frac{1}{4} \begin{bmatrix} 1 + 3e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} \\ 1 - e^{-4\alpha t} & 1 + 3e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} \\ 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 + 3e^{-4\alpha t} & 1 - e^{-4\alpha t} \\ 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 - e^{-4\alpha t} & 1 + 3e^{-4\alpha t} \end{bmatrix} \end{aligned}$$

Principle of Inference: Likelihood

Likelihood function

Likelihood function $L(\Theta, D)$ is probability of data D as function of parameters Θ

Maximum likelihood inference



Choose parameters $\hat{\Theta}$ maximising likelihood

- Consistent: $\hat{\Theta} \rightarrow \Theta_{\text{True}}$ for increasing sample size
- Parameters are viewed as fixed constants rather than random variables