

Computational Biology and Bioinformatics

<http://www.stats.ox.ac.uk/research/genome/projects>

11.10 Models of substitution I : Basic Models

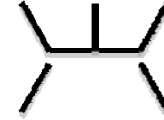
12.10 Models of substitution II : Complex Models

A
↓
T

18.10 Phylogenies I: Combinatorics

19.10 Phylogenies II: Distance, Parsimony & Likelihood

25.10 The Ancestral Recombination Graph & Pedigrees



26.10 Alignment Algorithms I Optimisation Alignment

01.11 Alignment Algorithms II Statistical Alignment

ACT-T
-GTCT

02.11 Stochastic Grammars and their Biological Applications: Hidden Markov Models

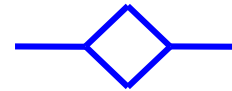
08.11 RNA structures

09.11 Finding Signals in Sequences



15.11 Challenges in Genome Annotation

16.11 Networks: Dynamics and Inference

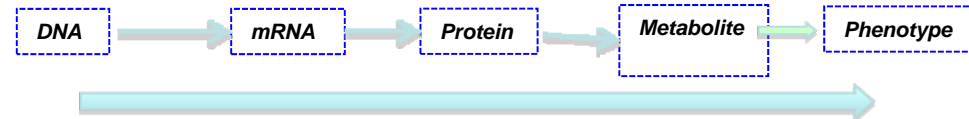


22.11 Networks: Evolution

23.11 Models of Evolution of Structures & Movements & Shapes & Grammars

29.11 Integrative Genomics: The Omics

30.11 Integrative Genomics: Mapping



The course

- *Self contained*
- *Contains Probability Theory, Combinatorics, Algorithmics and Mathematics*
- *Independence is encouraged*
- *Research Oriented*

Mini project-examination

- *It is expected to be 3 days worth of work.*
- *You will be given this in week 8*
- *I would expect 7-10 pages*
- *You will be given 2-4 key references*
- *A set of guiding questions that might help you in your writing*
- *You can chose between a set of topics broadly covering the taught material*

"Where a topic is assessed by a mini-project, the mini-project should be designed to take a typical student about three days. You are not permitted to withdraw from being examined on a topic once you have submitted your mini-project to the Examination Schools."

Mini chosen so far

2008 Computational Biology and Bioinformatics

Comparison of Networks	
Grammar Evolution Model	42

2009 Computational Biology of Sequences

Automated Annotation of Genes	87
Stochastic Context Free Grammars in RNA Secondary Structure Prediction	81
Substitution Models with Rate Heterogeneity	70

2009 Computational Biology of Networks

Probability Theory of Networks	93
Inference of Gene Regulatory Networks using Differential Equations	84
Network flow and its applications to the analysis of metabolism	57

2010 Computational Biology and Bioinformatics

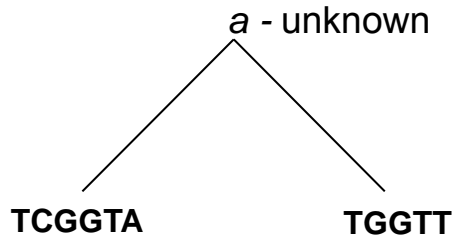
Identification of Regulatory Elements – Phylogenetic Footprinting	77
Phylogeny Reconstruction – Distance-based Methods	57
Probable and Improbable Paths in Sequence Evolution	79
Probabilistic Methods for DNA Sequence Alignment	69

Simplifying Assumptions I

Data: $s1=TCGGTA, s2=TGGTT$

Probability of Data $P = P(TCGGTA, TGGTT)$

Biological setup



0) *Independent Lineages:*

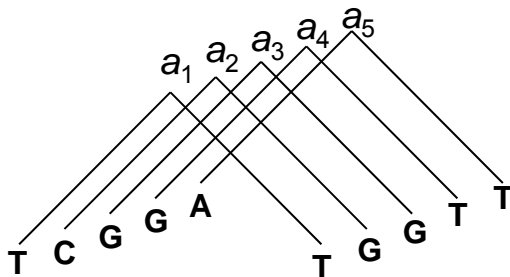
$$P = \sum_a P(a) * P(a \rightarrow TCGGTA) P(a \rightarrow TGGTT)$$

1) *Only substitutions.*

s1	TCGGTA	→	s1	TCGGA
s2	TGGT - T		s2	TGGTT

$$P = \sum_a P(a) * P(a \rightarrow TCGGA) P(a \rightarrow TGGTT)$$

2) Processes in different positions of the molecule are independent, so the probability for the whole alignment will be the product of the probabilities of the individual patterns.



$$P = \prod_{i=1}^5 \sum_a P_i(a_i) * P_i(a_i \rightarrow s1_i) P_i(a_i \rightarrow s2_i)$$

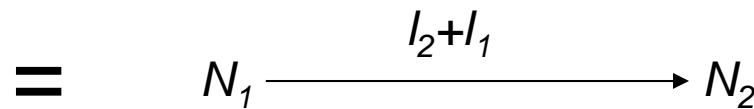
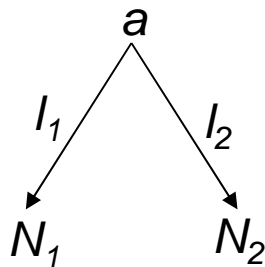
Simplifying Assumptions II

3) The evolutionary process is the same in all positions

$$P = \prod_{i=1}^5 \sum_a P(a_i) * P(a_i \rightarrow s1_i) P(a_i \rightarrow s2_i)$$

4) Time reversibility: Virtually all models of sequence evolution are time reversible. I.e. $\pi_i P_{i,j}(t) = \pi_j P_{j,i}(t)$, where π_i is the stationary distribution of i and $P_t(i \rightarrow j)$ the probability that state i has changed into state j after t time. This implies that

$$\sum_a P(a) * P_{a,N1}(l_1) * P_{a,N2}(l_2) = P_{N1} * P_{N1,N2}(l_1 + l_2)$$

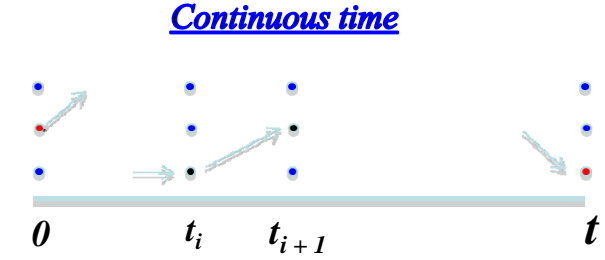
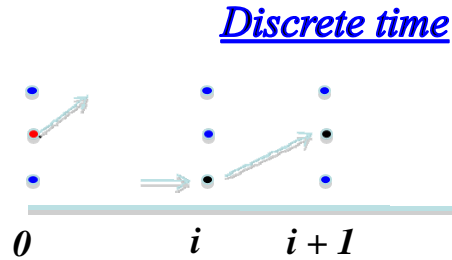
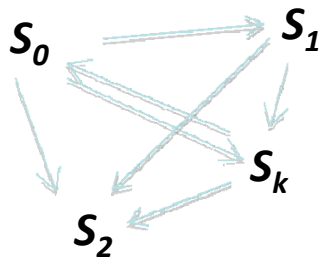


$$P = \prod_{i=1}^5 P(s1_i) P(s1_i \rightarrow s2_i)$$

Simplifying assumptions III: Continuous time markov chains (CTMC)

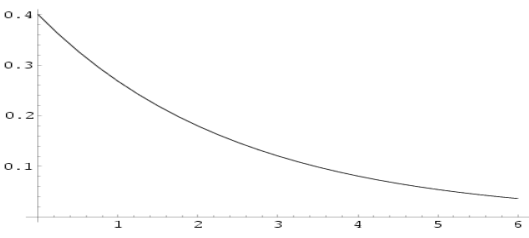
5) The nucleotide at any position evolves following a continuous time Markov Chain.

$P_{i,j}(t)$ continuous time markov chain on the state space $\{A,C,G,T\}$.



Exponential Distribution:

$$P(\mathbf{X} > t) = e^{-at}$$



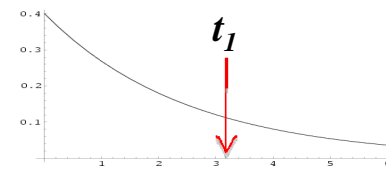
Properties: $X \sim \text{Exp}(a)$ $Y \sim \text{Exp}(b)$ independent

$$P(\mathbf{X} > t_2 | \mathbf{X} > t_1) = P(\mathbf{X} > t_2 - t_1) \quad (t_2 > t_1)$$

$$E(\mathbf{X}) = 1/a.$$

$$P(\mathbf{X} < \mathbf{Y}) = a / (a + b).$$

$$\min(\mathbf{X}, \mathbf{Y}) \sim \text{Exp}(a + b).$$



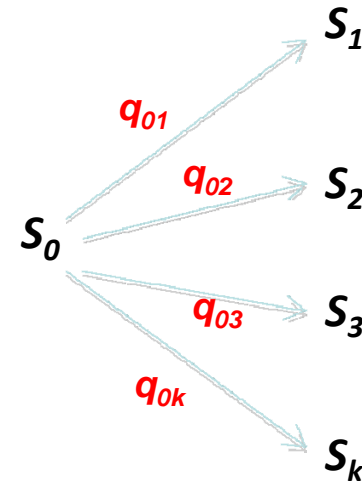
Rate Matrix: $Q := \lim_{h \rightarrow 0} \frac{P(h) - I}{h}$

6) The rate matrix, Q , for the continuous time Markov Chain is the same at all times (and often all positions). However, it is possible to let the rate of events, r_i , vary from site to site, then the term for passed time, t , will be substituted by $r_i * t$.

From Q to waiting times and jump probabilities

Let $X(t)$ be stochastic process governed by Q .

$$q_0 = q_{01} + q_{02} + q_{03} + \dots + q_{0k}$$



The waiting time in S_i , $T = \min_t \{X(t) \neq S_i \mid X(0) = S_i\}$ is exponentially distributed with intensity q_0 .

Jump probabilities : $P(X(T) = S_j \mid X(0) = S_i) = q_{ij}/q_i$.

How to simulate a random path starting in A : $A \rightarrow X(0) \quad t=0.0$

1) $t_+ \sim \text{Exp}[q_{X(t),X(t)}]$

2) $X(t) := [q_{X(t),1}, q_{X(t),2}, \dots, q_{X(t),k}] / q_{X(t),X(t)}$ (many $q_{X(t),j}$ can be 0.0 and be ignored)

From Q to P(t)

What is the probability of going from i (C?) to j (G?) in time t with rate matrix Q?

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

- i.** $P(0) = I$
- ii.** $P(\varepsilon)$ close to $I + \varepsilon Q$ for ε small
- iii.** $P'(0) = Q$.
- iv.** $\lim P(t)$ has the equilibrium frequencies of the 4 nucleotides in each row
- v.** Waiting time in state j, T_j , $P(T_j > t) = e^{q_{jj}t}$

Expected number of events at equilibrium

$$t \sum_{\text{nucleotides}} -q_{ii} \pi_i$$

Jukes-Cantor (JC69): Total Symmetry

Rate-matrix, R:

T O

		A	C	G	T
FROM	A	$-3*\alpha$	α	α	α
	C	α	$-3*\alpha$	α	α
	G	α	α	$-3*\alpha$	α
	T	α	α	α	$-3*\alpha$

Transition prob. after time t, $a = \alpha*t$:

$$P(\text{equal}) = \frac{1}{4}(1 + 3e^{-4a}) \sim 1 - 3a$$

$$P(\text{specific difference}) = \frac{1}{4}(1 - e^{-4a}) \sim a$$

Stationary Distribution: (1,1,1,1)/4.

$$\begin{aligned}
 P &= P(s1) \prod_{i=1}^5 P(s1_i \rightarrow s2_i) = \left(\frac{1}{4}\right)^5 P(T \rightarrow T)P(C \rightarrow G)P(G \rightarrow G)P(G \rightarrow T)P(A \rightarrow T) \\
 &= \left(\frac{1}{4}\right)^5 \left(\frac{1}{4}\right)^5 (1 + 3e^{-4a})^2 (1 - e^{-4a})^3
 \end{aligned}$$

Exponentiation/Powering of Matrices

By eigen values:

If $Q = B\Lambda B^{-1}$ where $\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix}$ then $Q^i = B\Lambda B^{-1}B\Lambda B^{-1} \dots B\Lambda B^{-1} = B\Lambda^i B^{-1}$

and $\sum_{i=0}^{\infty} \frac{(tQ)^i}{i!} = \sum_{i=0}^{\infty} \frac{(tB\Lambda B^{-1})^i}{i!} = B \left[\sum_{i=0}^{\infty} \frac{(t\Lambda)^i}{i!} \right] B^{-1} = B \begin{pmatrix} \exp t\lambda_1 & 0 & 0 & 0 \\ 0 & \exp t\lambda_2 & 0 & 0 \\ 0 & 0 & \exp t\lambda_3 & 0 \\ 0 & 0 & 0 & \exp t\lambda_4 \end{pmatrix} B^{-1}$

Finding Λ : $\det(Q - \lambda I) = 0$

Finding B : $(Q - \lambda_i I)b_i = 0$

JC69:

$$P(t) = \begin{pmatrix} 1 & 1/4 & 0 & 1 \\ 1 & 1/4 & 0 & -1 \\ 1 & -1/4 & 1 & 0 \\ 1 & -1/4 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \exp -4t\alpha & 0 & 0 \\ 0 & 0 & \exp -4t\alpha & 0 \\ 0 & 0 & 0 & \exp -4t\alpha \end{pmatrix} \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ 0 & 0 & 1 & -1 \\ 1 & -1 & 0 & 0 \end{pmatrix}$$

Numerically:

$$\sum_{i=0}^{\infty} \frac{(tQ)^i}{i!} \sim \sum_{i=0}^k \frac{(tQ)^i}{i!} \quad \text{where } k \sim 6-10$$

From Q to P for Jukes-Cantor

$$\begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix} = \alpha \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix} = 4^{i-1} \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix}$$

$$\sum_{i=0}^{\infty} \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}^i \frac{t^i}{i!} = 1/4 \left[I - \sum_{i=1}^{\infty} (-4\alpha t)^i \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix} \frac{1}{i!} \right] =$$

$$1/4 \left[I + \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix} e^{-4\alpha t} \right]$$

Principle of Inference: Likelihood

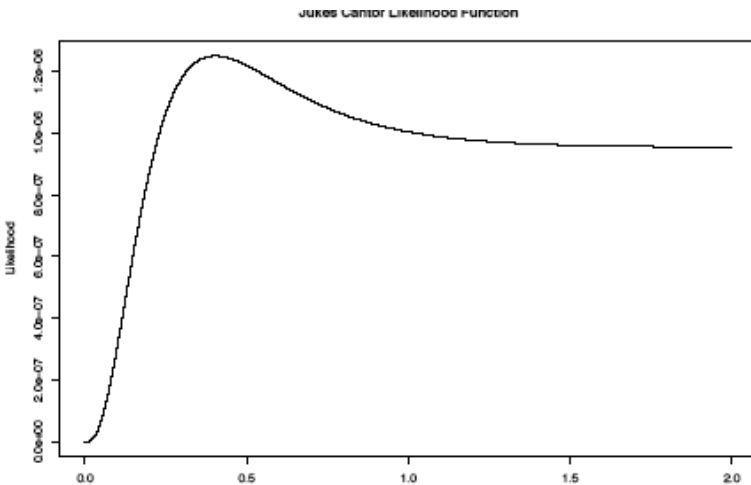
Likelihood function $L()$ – the probability of data as function of parameters: $L(\Theta, D)$

LogLikelihood Function – $l()$: $\ln(L(\Theta, D))$

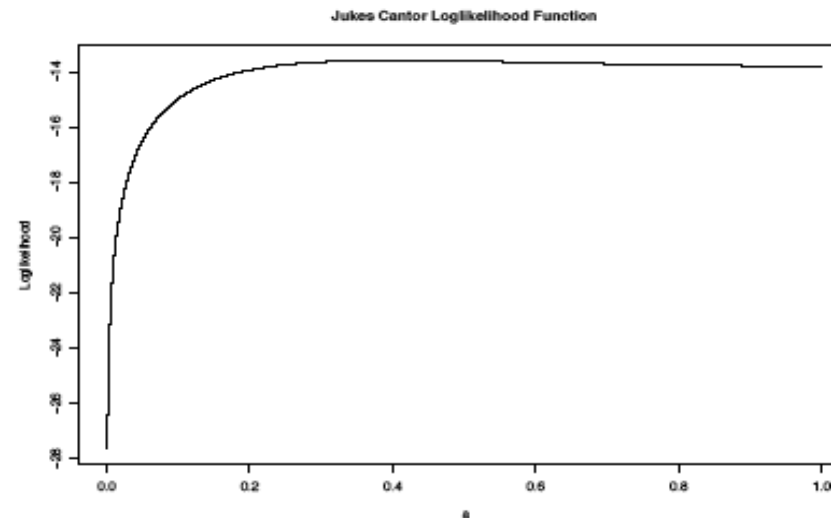
If the data is a series of independent experiments $L()$ will become a product of Likelihoods of each experiment, $l()$ will become the sum of LogLikelihoods of each experiment

Consistency : $\hat{\Theta}(D) \rightarrow \Theta_{true}$ as data increases.

Likelihood



LogLikelihood



In Likelihood analysis parameter is not viewed as a random variable.