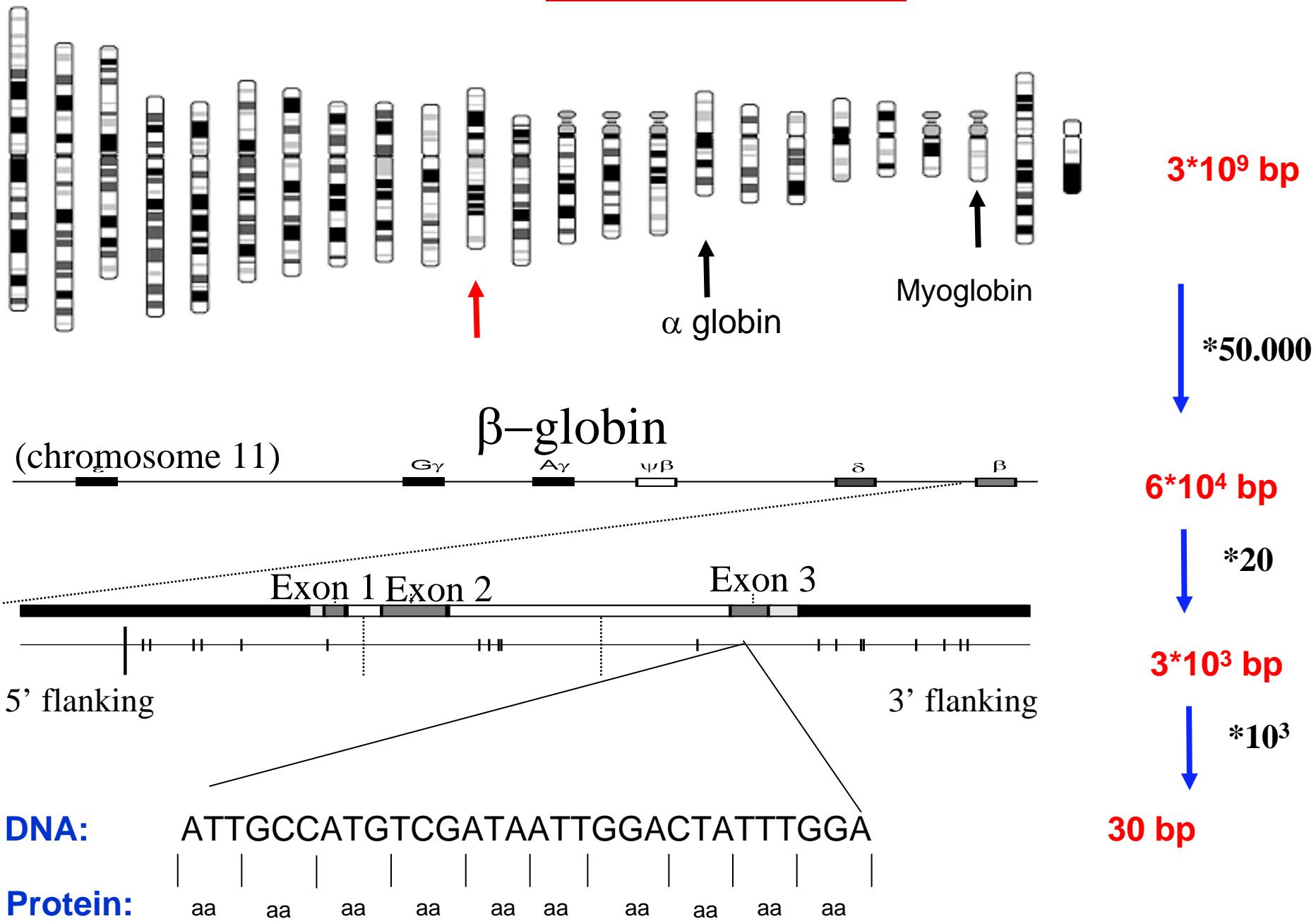


The Human Genome (Harding & Sanger)



Models of substitution I : Basic Models 15.10
Models of substitution II : Complex Models 16.10
Models of substitution III : Advanced Questions 22.10

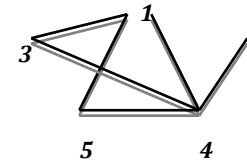
A
↓
T

Phylogenies I: Combinatorics 23.10

Phylogenies II: Parsimony 29.10
Phylogenies III: Likelihood 30.10



Phylogenies IV: Inference 5.11
Networks I: Dynamics 6.11



Networks II: Inference 12.11
Networks III: Evolution 13.11

Alignment Algorithms I Optimisation 19.11 (Rune)
Alignment Algorithms II Statistical Inference 20.11 (Rune)

ACT-T
-GTCT

Stochastic Grammars and their Biological Applications: Hidden Markov Models 26.11
Finding Signals in Sequences 27.11



RNA structures (Rune) 3.12

Finding Recombinations in Sequences 4.12

Projects in substitution models, phylogenies, networks, grammars, RNA structure, signals, your choice

Course should appeal to combinatorics, probability theory, statistics, algorithmics, software design,

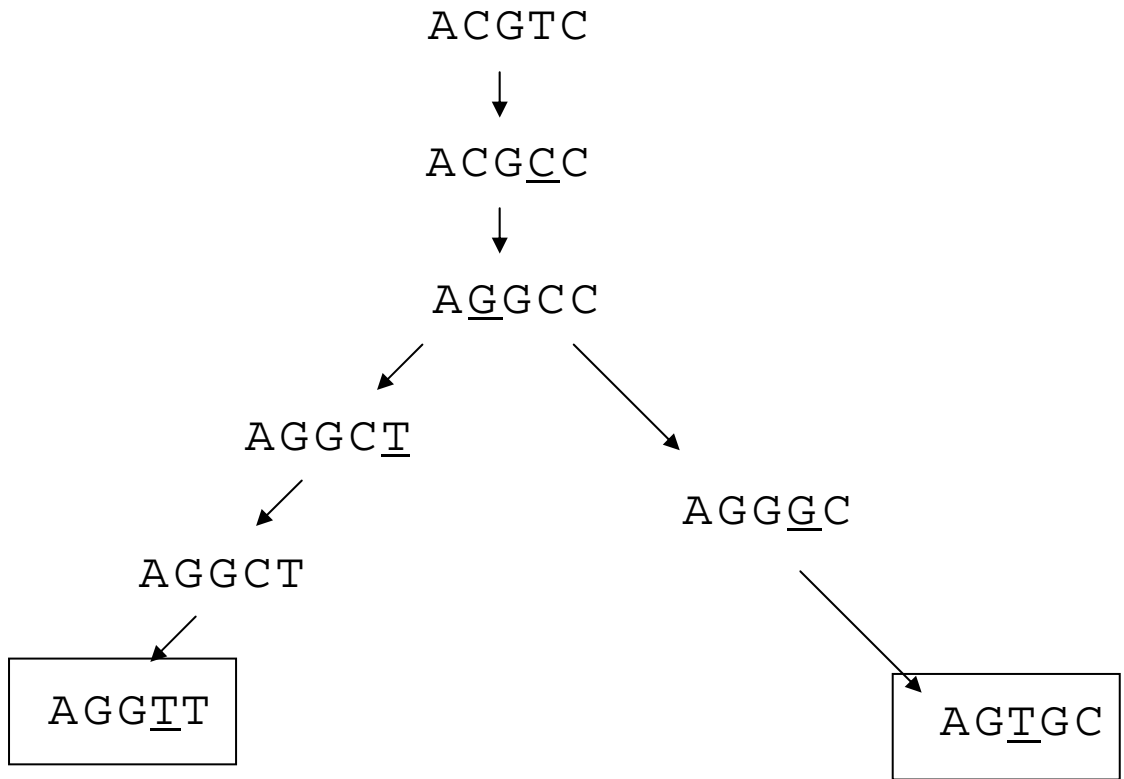
Summer projects: <http://www.stats.ox.ac.uk/research/genome/projects>

EMAILS: lyngsoe@stats.ox.ac.uk hein@stats.ox.ac.uk

Schedule

Central Problems: History cannot be observed, only end products.

ACGTC
↓
ACGCC
↓
AGGCC
↓
AGGCT
↓
AGGCT
↓
AGGTT



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

Some Definitions

*State space – a set often corresponding of possible observations
ie $\{A,C,G,T\}$.*

*A random variable, X can take values in the state space with probabilities
ie $P\{X=A\} = 1/4$. The value taken often indicated by small letters - x*

*Stochastic Process is a set of time labeled stochastic variables X_t
ie $P\{X_0=A, X_1=C, \dots, X_5=G\} = .00122$*

*Time can be discrete or continuous, in our context it will almost always
mean natural numbers, $N \{0,1,2,3,4..\}$, or an interval on the real line, R .*

*Markov Property: $P\{X_i|X_{i-1},\dots,X_0\} = P\{X_i|X_{i-1}\}$
ie $P\{X_i,X_{i-1},\dots,X_0\} = P\{X_0\}P\{X_1|X_0\}\dots P\{X_i|X_{i-1}\}$*

Time Homogeneity – the process is the same for all t .

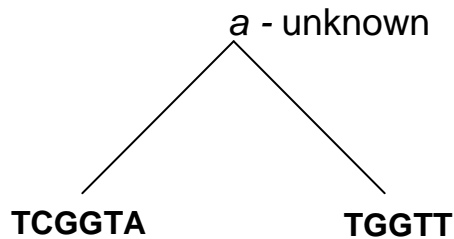
Simplifying Assumptions I

Data: $s_1=TCGGTA, s_2=TGGTT$

Probability of Data

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

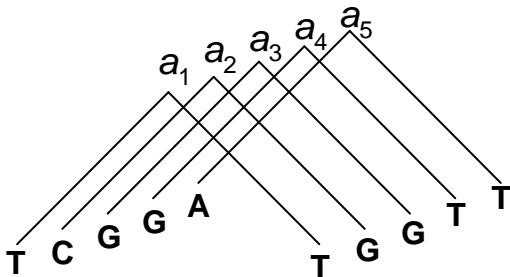
Biological setup



1) Only substitutions. s_1 TCGGTA s_1 TCGGA →
 s_2 TGGT-T s_2 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.



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

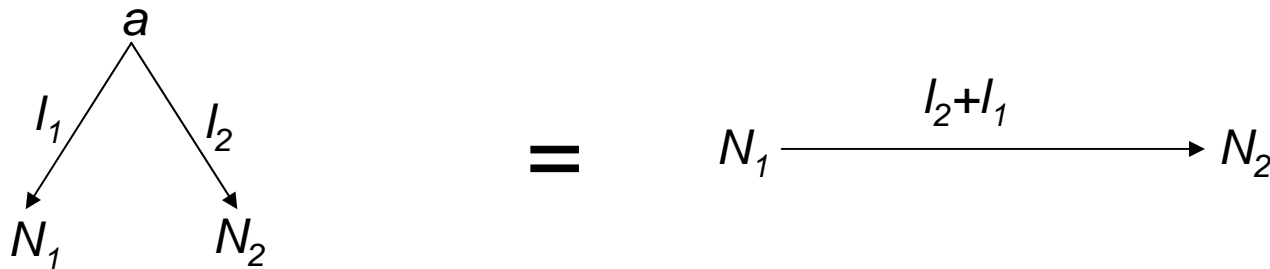
Simplifying Assumptions II

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

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



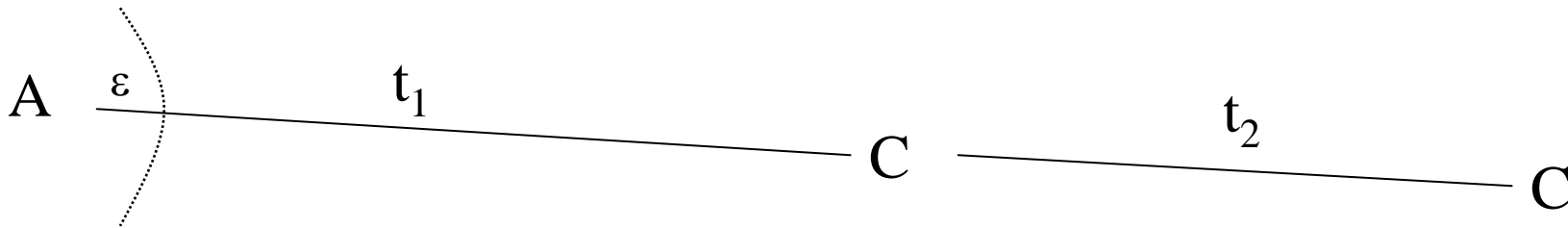
$$P = \prod_{i=1}^5 P(s1_i) P(s1_i \rightarrow s2_i)$$

Simplifying assumptions III

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}.

$$\lim_{\varepsilon \rightarrow 0} \frac{P_{i,j}(\varepsilon)}{\varepsilon} = q_{ij} \qquad \lim_{\varepsilon \rightarrow 0} \frac{P_{i,i}(\varepsilon) - 1}{\varepsilon} = -q_{ii}$$



Q - rate matrix:

		A	C	G	T
F	A	$-(q_{A,C} + q_{A,G} + q_{A,T})$	$q_{A,C}$	$q_{A,G}$	$q_{A,T}$
R	C	$q_{C,A}$	$-(q_{C,A} + q_{C,G} + q_{C,T})$	$q_{C,G}$	$q_{C,T}$
O	G	$q_{G,A}$	$q_{G,C}$	$-(q_{G,A} + q_{G,C} + q_{G,T})$	$q_{G,T}$
M	T	$q_{T,A}$	$q_{T,C}$	$q_{T,G}$	$-(q_{T,A} + q_{T,C} + q_{T,G})$

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 $r_i * t$.