

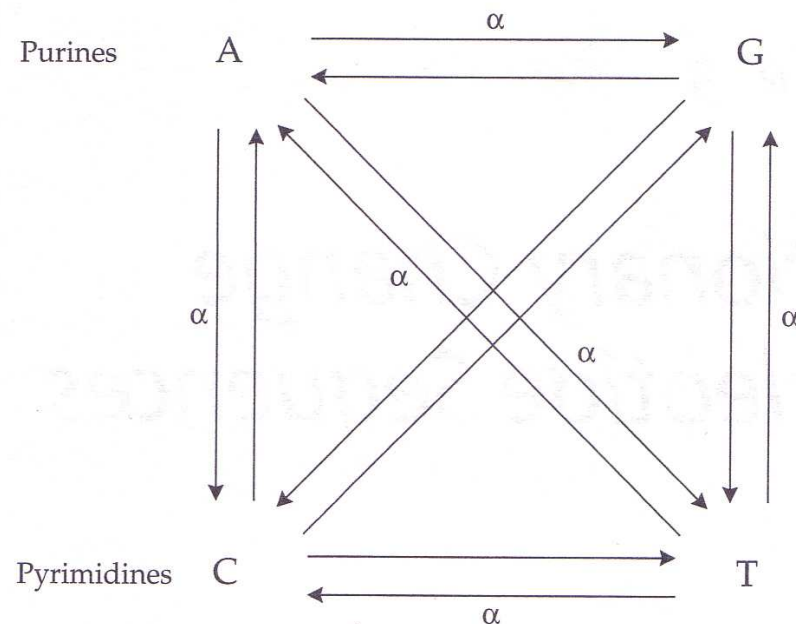
Nucleotide substitution models

Alexander Churbanov

University of Wyoming, Laramie

Jukes and Cantor's model [1]

- The simplest symmetrical model of DNA evolution
- All sites change independently
- All sites have the same stochastic process working at them



Probability for nucleotide (1)

- Let us assume nucleotide residing at certain site in DNA sequence is **A**,
- Consider probability that $p_{A(t)}$ that the site will be occupied by **A** at time t ,
- Since we start with **A**, $p_{A(0)} = 1$,
- At a time 1 probability of still having **A** is $p_{A(1)} = 1 - 3\alpha$.

Probability for nucleotide (2)

- At a time 2 $p_{A(2)} = (1 - 3\alpha)p_{A(1)} + \alpha(1 - p_{A(1)})$
 1. The nucleotide has remained unchanged with probability $1 - 3\alpha$
 2. The nucleotide did change to **T**, **C**, **G**, but subsequently reverted to **A** with probability α
- The following recurrence holds

$$p_{A(t+1)} = (1 - 3\alpha)p_{A(t)} + \alpha(1 - p_{A(t)}), \Rightarrow$$

$$p_{A(t+1)} - p_{A(t)} = 3\alpha p_{A(t)} + \alpha(1 - p_{A(t)}), \Rightarrow$$

$$\Delta p_{A(t)} = 3\alpha p_{A(t)} + \alpha(1 - p_{A(t)}) = -4\alpha p_{A(t)} + \alpha.$$

Probability for nucleotide (2)

- At a time 2 $p_{A(2)} = (1 - 3\alpha)p_{A(1)} + \alpha(1 - p_{A(1)})$
 1. The nucleotide has remained unchanged with probability $1 - 3\alpha$
 2. The nucleotide did change to **T**, **C**, **G**, but subsequently reverted to **A** with probability α
- The following recurrence holds

$$p_{A(t+1)} = (1 - 3\alpha)p_{A(t)} + \alpha(1 - p_{A(t)}), \Rightarrow$$

$$p_{A(t+1)} - p_{A(t)} = 3\alpha p_{A(t)} + \alpha(1 - p_{A(t)}), \Rightarrow$$

$$\Delta p_{A(t)} = 3\alpha p_{A(t)} + \alpha(1 - p_{A(t)}) = -4\alpha p_{A(t)} + \alpha.$$

Continuous time

$$\frac{dp_{A(t)}}{dt} = -4\alpha p_{A(t)} + \alpha.$$

This first-order linear differential equation has solution

$$p_{A(t)} = \frac{1}{4} + \left(p_{A(0)} - \frac{1}{4} \right) e^{-4\alpha t}$$

Initial probability is $p_{A(0)} = 1$, therefore

$$p_{A(t)} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$$

Probabilities

If the initial nucleotide is not **A**, then $p_{A(0)} = 0$ and

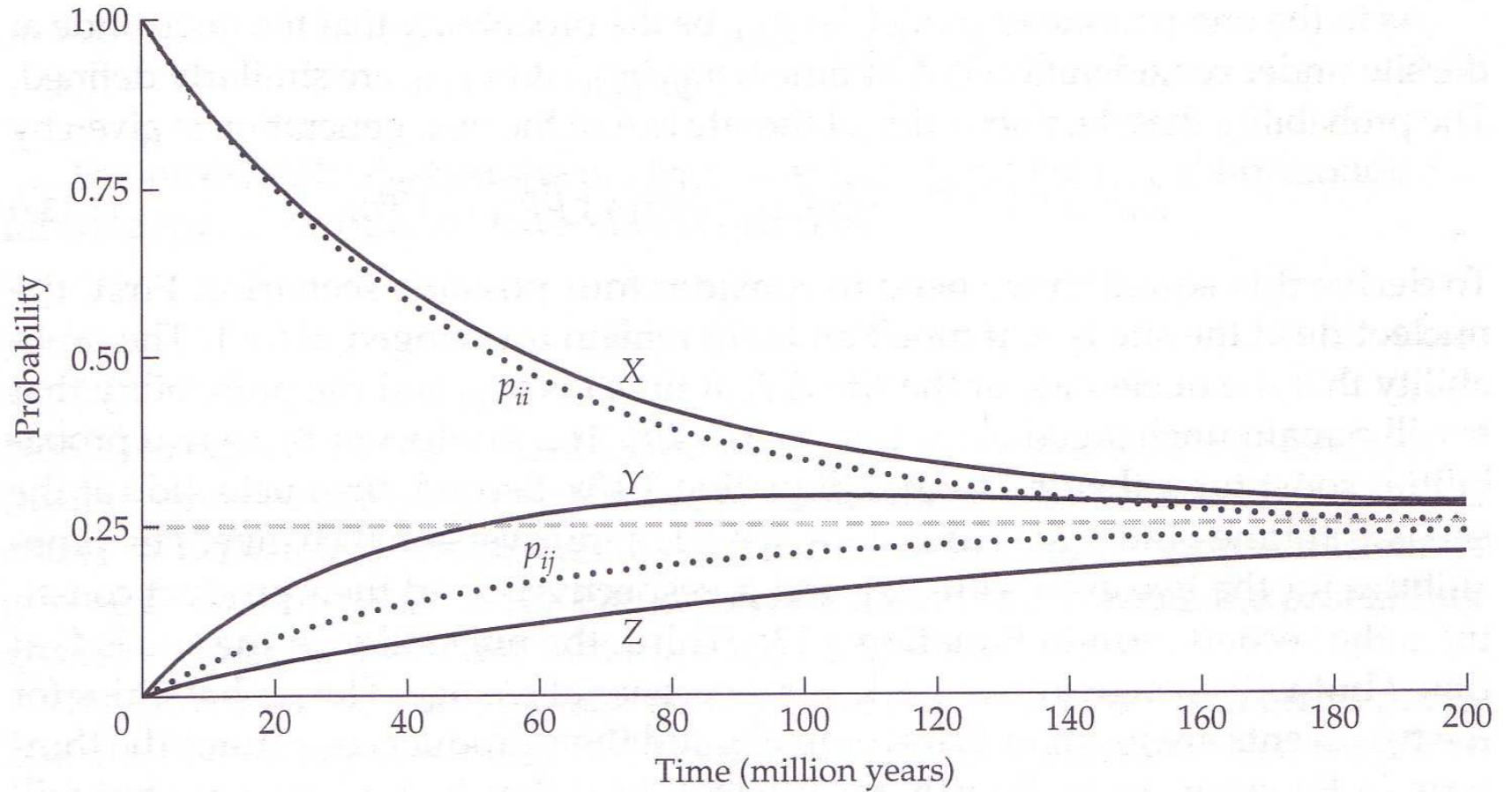
$$p_{A(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

generalizing for nucleotides i and j , where $i \neq j$

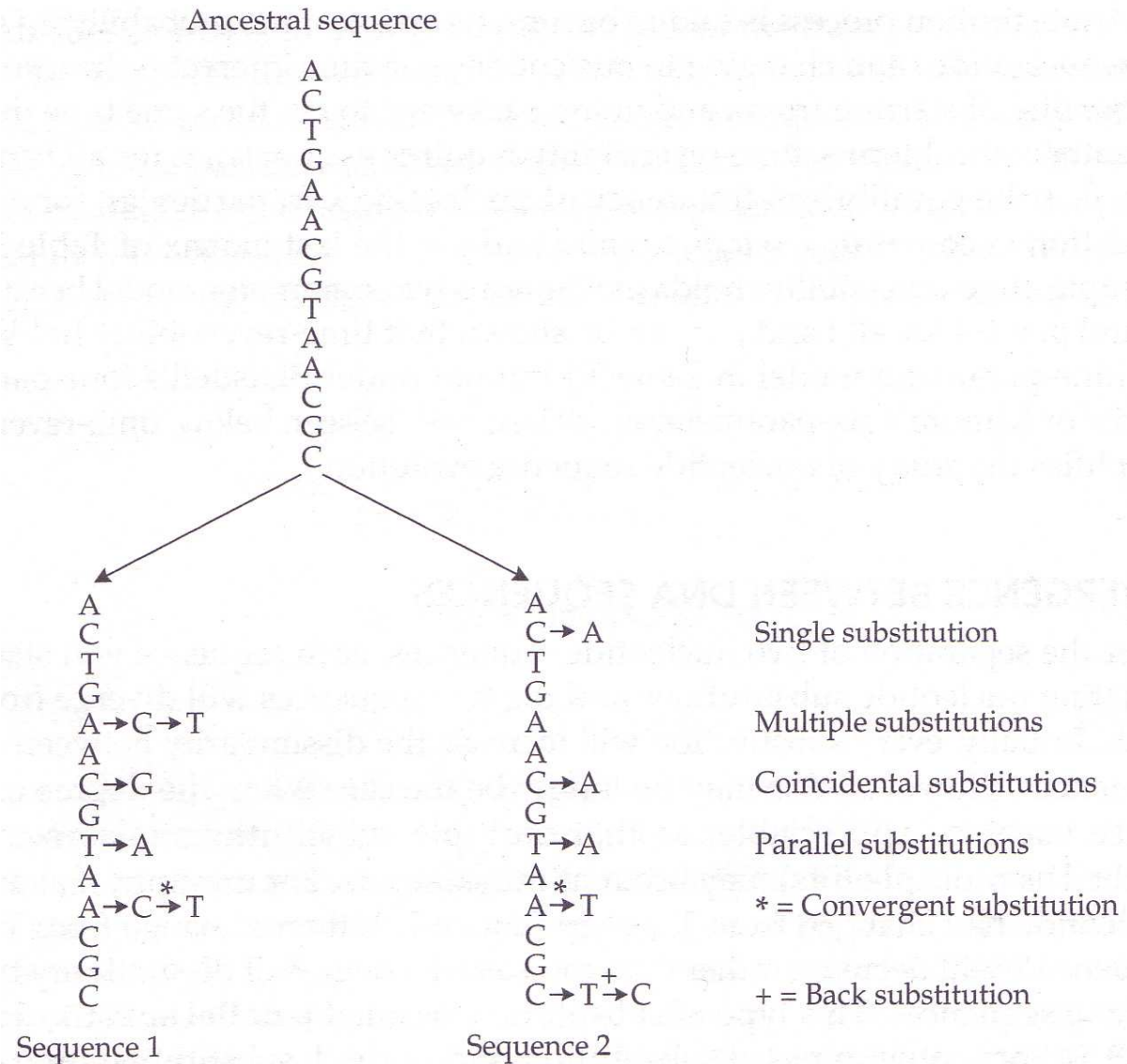
$$p_{ii(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

$$p_{ij(t)} = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$

Graphical interpretation



Sequence similarity (1)



Sequence similarity (2)

- A common measure for sequence similarity is the proportion of identical nucleotides between the two sequences under study.
- The expected value of this proportion is equal to the probability $I(t)$ that the nucleotide at a given site at a time t is the same in both sequences. Cases include nucleotide conservation $p_{ii(t)}^2$ and parallel substitutions $p_{ij(t)}^2$.

$$I(t) = p_{AA(t)}^2 + p_{AT(t)}^2 + p_{AC(t)}^2 + p_{AG(t)}^2, \Rightarrow$$

$$I(t) = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}.$$

Estimating substitutions (1)

- The probability that the two sequences are different at a site at time t is $p = 1 - I_{(t)}$

$$p = \frac{3}{4} (1 - e^{-8\alpha t}), \Rightarrow$$

$$8\alpha t = -\ln \left(1 - \frac{4}{3}p \right).$$

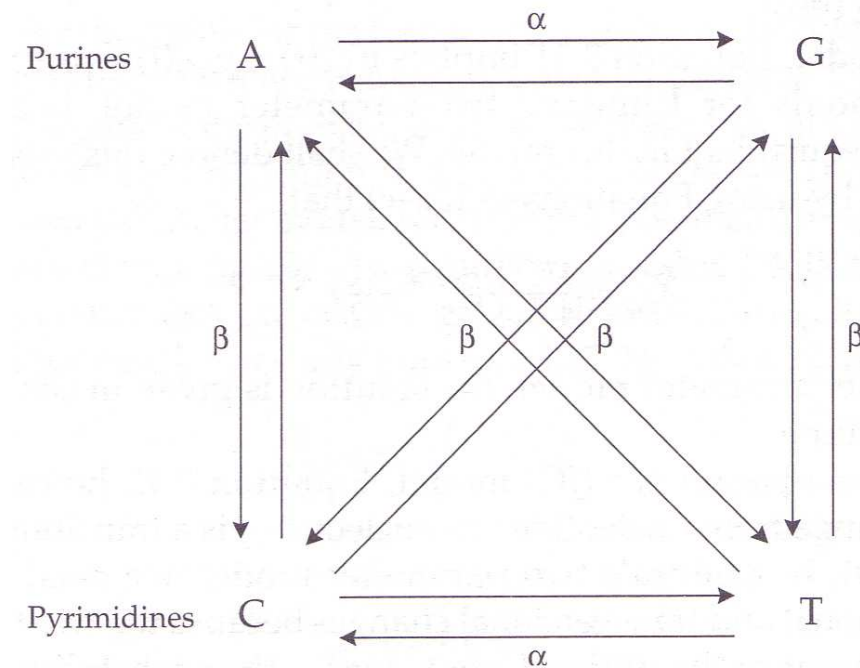
Estimating substitutions (2)

- We estimate K , the actual number of substitutions per site since the divergence between the two sequences.
- In the one parameter model, $K = 2(3\alpha t)$, where $3\alpha t$ is the expected number of substitutions per site in one lineage.

$$K = - \binom{3}{4} \ln \left(1 - \frac{4}{3}p \right)$$

Kimura model [2]

- The method has the merit of incorporating the possibility that sometimes “transition” type substitutions (with rate α) may occur more frequently than “transversion” type substitutions (with rate β).



Kimura model [2]

Same	UU	CC	AA	GG	Total
(Frequency)	(R_1)	(R_2)	(R_3)	(R_4)	(R)

Different, Type I	UC	CU	AG	GA	Total
(Frequency)	(P_1)	(P_1)	(P_2)	(P_2)	(P)

Different, Type II	UA	AU	UG	GU	Total
(Frequency)	(Q_1)	(Q_1)	(Q_2)	(Q_2)	(Q)
	CA	AC	CG	GC	
(Frequency)	(Q_3)	(Q_3)	(Q_4)	(Q_4)	

Kimura model (1)

- Total rate of substitutions per site per year is
 $k = \alpha + 2\beta$
- P is the probability of homologous sites showing a type I difference
- Q is the probability of homologous sites showing a type II difference
- R is the probability of homologous sites to be the same
- We denote probability of identity at homologous sites at time T as $R(T) = 1 - P(T) - Q(T)$

Kimura model (2)

- We can derive the equation for P and Q at time $T + \Delta T$ in terms of P , Q and R at time T
- We can distinguish three ways by which **UC** (**U** at homologous position of organism 1 corresponding to **C** in organism 2) at time $T + \Delta T$ is derived from various base pairs at time T .
 1. Pair **UC** is derived from **UC**. Since probability of substitution in short time interval is ΔT is $(\alpha + 2\beta)\Delta T$. Thus the probability of no change occurring in both homologous sites is $[1 - (\alpha + 2\beta)\Delta T]^2$, so this case contribution is $[1 - (\alpha + 2\beta)\Delta T]^2 P_1(T)$

Kimura model (3)

2. Pair **UC** is derived either from **UU** or from **CC** with probability

$$\alpha \Delta T [R_1(T) + R_2(T)]$$

3. Pair **UC** could be derived from **UA**, **UG**, **AC** and **GC** with probability

$$\beta \Delta T [Q_1(T) + Q_2(T) + Q_3(T) + Q_4(T)] = \beta \Delta T \frac{Q(T)}{2}$$

Kimura model (4)

Combining contributions coming from different classes resulting in **UC** and disregarding terms with $(\Delta T)^2$, we get

$$(1) \quad P_1(T + \Delta T) = [1 - (2\alpha + 4\beta)\Delta T] P_1(T) + \alpha\Delta T [R_1(T) + R_2(T)] + \beta\Delta T \frac{Q(T)}{2}$$

Similarly, for the base pair **AG** we get

$$(2) \quad P_2(T + \Delta T) = [1 - (2\alpha + 4\beta)\Delta T] P_2(T) + \alpha\Delta T [R_3(T) + R_4(T)] + \beta\Delta T \frac{Q(T)}{2}$$

Kimura model (5)

Summing equations (1) and (2), and noting $P(T) = 2P_1(T) + 2P_2(T)$, we get

$$(3) \quad \frac{\Delta P(T)}{\Delta T} = 2\alpha - 4(\alpha + \beta)P(T) - 2(\alpha - \beta)Q(T)$$

$$(4) \quad \frac{\Delta Q(T)}{\Delta T} = 4\beta - 8\beta Q(T)$$

Converting (3) and (4) to continuous case, we get

$$\frac{dP(T)}{dT} = 2\alpha - 4(\alpha + \beta)P(T) - 2(\alpha - \beta)Q(T)$$

$$\frac{dQ(T)}{dT} = 4\beta - 8\beta Q(T)$$

Kimura model (5)

The solution for these equations satisfy the initial condition $P(0) = Q(0) = 0$, i.e. no base difference exists at $T = 0$

$$P(T) = \frac{1}{4} - \frac{1}{2}e^{-4(\alpha+\beta)T} + \frac{1}{4}e^{-8\beta T}$$

$$Q(T) = \frac{1}{2} - \frac{1}{2}e^{-8\beta T}$$

Kimura model (6)

It follows that

$$(5) \quad 4(\alpha + \beta)T = -\ln(1 - 2P(T) - Q(T))$$

and

$$(6) \quad 8\beta T = -\ln(1 - 2Q(T))$$

so that

$$(7) \quad 4\alpha T = -\ln(1 - 2P(T) - Q(T)) \\ + \frac{1}{2} \ln(1 - 2Q(T))$$

Kimura model (7)

Since evolutionary rate is $k = \alpha + 2\beta$, the total number of substitutions per two diverged sequences is

$$K = 2Tk = 2\alpha T + 4\beta T$$

By omitting index T and following equations (6) and (7) we obtain

$$K = -\frac{1}{2} \ln \left\{ (1 - 2P - Q) \sqrt{1 - 2Q} \right\}$$

References

- [1] T.H. Jukes and C.R. Cantor, *Evolution of protein molecules*, Mammalian protein metabolism (H.N. Munro, ed.), Academic Press, New York, 1969, pp. 21–132.
- [2] M. Kimura, *A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences*, *Journal of Molecular Evolution* **16** (1980), 111–120.