RAPPORT BIBLIOGRAPHIQUE
(BIBLIOGRAPHICAL WORK)

Robert Schmidt

D.E.A. Analyse et Modélisation des Systèmes Biologiques

Année Universitaire 2000/2001

Université Claude Bernard – Lyon I

Directeur de Recherches: Jean Lobry

Parrains: Daniel Chessel
Alain Viari
Christophe Lett

Thème: Les modèles de l’evolution des fréquences des bases de l’ADN, typologie, propriétés connues, et statut par rapport à l’hypothèse de symmétrie (PR1-hypothesis)
(The models of the evolution of the DNA- base frequencies, typology, known properties, and their position in respect to the assumption of symmetry (PR1-hypothesis))

A. Introduction and the general model:

The goal of this work is to describe and discuss major models and methods that allow a mathematical description of the evolutionary process of DNA base substitution, in particular Markov models that use matrices of base substitution probabilities or base substitution rates for this description, and, in addition, to show the relationships between these models.

Before considering mathematical models built to describe the molecular evolution, we should first take a look at major molecular processes underlying it: the process of molecular evolution is governed by two major “forces“: mutation and selection (Sueoka, 1995). Mutations (substitutions, recombinations, deletions, insertions, inversions) occur at the individual level and should not be confused with the evolutionary change of DNA sequences. This change occurs in respect to assumed representative (average) sequences of species the evolutionary changes of which do not simply reflect mutation rates. Selection causes that certain mutations may have a higher or lower probability to become “fixed“ in a population. Furthermore, the molecular evolution represents a complex statistical process and, therefore, random genetic drift certainly also plays a major role (Graur and Li, 2000).

From now on, in this work, “mutations“ are regarded as the changes that occur in sequences representing species, rather than individuals. To allow a mathematical description, many models share simplifying assumptions. The most important assumptions generally made are:

- the evolutionary process consists only of substitutions; other mutations are neglected
- the sites evolve independently and according to the same probabilistic process
- the evolutionary process can be described as a Markov process: $P_{ij}(t) = \Pr[X(s+t) = j \mid X(s) = i]$ is the probability that base i will change to j during the time t. Markov assumption: $P_{ij}(t)$ is independent of the time s $(\mathrm{s} \geq 0)$ when the process starts, that means, the process is “memoryless“ (Lio and Goldman, 1998)
- the Markov process is homogeneous: the rate matrix (see below) is constant in time,
- and it is stationary: the process is at equilibrium (the base frequencies remain constant).

Many violations of these assumptions can be detected in actual sequences, for example:

- many mutations are not substitutions (see above)
- mutation rates are affected by chromosomal position, codon position (the three codon positions can be analysed separately, though), and they are also influenced by nearest neighbor bases (Lio and Goldman, 1998)
- the evolutionary process is not homogeneous in time; population sizes, geological changes etc. can alter selection pressures and the statistical properties of this process, and finally,
- the stationarity is clearly violated when the base frequencies of the compared species are quite different (Lio and Goldman, 1998) .

<u>overview</u>:

purines: A, G ; pyrimidines: C, T

DNA: A („X “) binds T („$\bar{X}$ “) , C („Y “) binds G („$\bar{Y}$ “) ; $\bar{X}$ and $\bar{Y}$ are the complementary bases (nucleotides) of X and Y , respectively.

<u>G12: The general model (12 free parameters)</u>:

The nucleotide substitution rates per site r_{ij} form the rate matrix **R:**

$$\begin{array}{cccc|cccc|}
 & & & & j = 1 & 2 & 3 & 4 \\
 & & & \text{to:} & X & \bar{X} & Y & \bar{Y} \\
 & & & & \mathrm{A} & \mathrm{T} & \mathrm{C} & \mathrm{G} \\
\cline{5-8}
 & i = 1 & X & \mathrm{A} & - & r_{AT} & r_{AC} & r_{AG} \\
\text{from:} & 2 & \bar{X} & \mathrm{T} & r_{TA} & - & r_{TC} & r_{TG} \\
 & 3 & Y & \mathrm{C} & r_{CA} & r_{CT} & - & r_{CG} \\
 & 4 & \bar{Y} & \mathrm{G} & r_{GA} & r_{GT} & r_{GC} & - \\
\cline{5-8}
\end{array}$$

or with more convenient names for the parameters:

$$\begin{array}{|cccc|}
\hline
- & \mathrm{t1} & \mathrm{c1} & \mathrm{g1} \\
\mathrm{a2} & - & \mathrm{c2} & \mathrm{g2} \\
\mathrm{a3} & \mathrm{t3} & - & \mathrm{g3} \\
\mathrm{a4} & \mathrm{t4} & \mathrm{c4} & - \\
\hline
\end{array}$$

The diagonal elements (- l_i) of the matrix are always the negative sum of the other elements in the same row (l_A is, for example, $r_{AT}+r_{AC}+r_{AG}$) (Lio and Goldman, 1998). Therefore, the sum of each row is zero. For simplicity, in this work the diagonal elements are always symbolized by a “-“ .

If we assume infinitely long sequences and consider the evolutionary process a continous process in time, we can calculate the changes of the frequencies A(t), T(t), C(t), and G(t) of the four bases in an intervall dt. For A, for instance, we obtain the expression: $A(t+dt)=A(t)(1-l_A dt)+T(t)r_{T,A}dt+C(t)r_{C,A}dt+G(t)r_{G,A}dt$. l_A is the rate by which a given nucleotide A changes (towards T, C, or G). $(1-l_A dt)$ is, therefore, the probabilty that a nucleotide A remains unchanged during dt. $r_{T,A}dt$ is the probability that a T is exchanged for A in dt. The expressions for T(t+dt), C(t+dt), and G(t+dt) can be obtained in the same fashion. If X(t) denotes the vector of A(t), T(t), C(t), and G(t), we can write these four equations with the help of the rate matrix **R**: X(t+dt)=**R‘**X(t)dt+X(t) . **R‘** be the transpose of the matrix **R**. Some authors use a rate matrix where the column defines the original nucleotide and the row the one to which it changes with the respective rate. In that case, instead of **R‘** , one must write **R** in that equation. We can easily derive: $\frac{dX(t)}{dt}=\lim_{dt\to 0}\frac{X(t+dt)-X(t)}{dt}$ =**R’**X(t) . In matrix notation, the solution of this system of differential equations reads: X(t)=exp(**R’**t)X0 where X0 be the vector X(t) at t=0. If one can calculate the diagonalization of **R‘** : **R‘**= $P^{-1}DP$ (D being a diagonal matrix with the elements d1, d2, d3, d4), exp(**R‘t**) can be expressed as $P^{-1}\exp(Dt)P$ where exp(Dt) is the diagonal matrix with the elements exp(d1t), exp(d2t), exp(d3t), exp(d4t) (see Lio and Goldman, 1998).

For the system $\frac{dX(t)}{dt}$ =**R’**X(t) , there exists an equilibrium frequency for the vector X, that means, for the bases A, T, C, and G, denoted a, t, c, and g, respectively. It is defined by the equations: $\frac{dX(t)}{dt}=0$ → **R’**X = 0. This is a homogenous system of linear equations and can, consequently, be solved (for instance by the Gauss-algorithm) for any matrix **R‘** the determinant of which equals zero. This requirement is always fulfilled for the type of matrix-construction where the elements $-l_i$ in the diagonal are the negative sum of the other elements in the i’th row. I verified this using the program “mathematica“ with the commands:

```
l1=      t1+c1+g1;
l2=a2     +c2+g2;
l3=a3+t3      +g3;
l4=a4+t4+c4     ;
tm:={{-l1, t1,c1,g1},{a2,-l2,c2,g2},{a3,t3,-l3,g3},{a4,t4,c4,-l4}};
Det[tm]   .
```

The result is zero and, therefore, one can obtain three equations, each containing, for instance, one of the variables a, t, and c as a function of g. Using the requirement: a+t+c+g=1, one can obtain the equilibrium frequencies. Though possible, it is not worthwhile to give here the general expressions for the equilibrium frequencies a, t, c, and g in terms of the 12 parameters of the gereral model (this will be an issue of the “rapport technique“), but the ones that are given for the other models in the table in the classification part of this work (see below) may easily be verified. For example, for the TN model (Tamura and Nei, 1993) one can verify the first equation of the system: **R’**X = 0 as follows (the reader may look at the TN matrix): the equation reads $-l_A a + abt + abc + aa_1 g \overset{?}{=} 0$. This is the first column of the rate matrix (because one has to use the transposed matrix) times the equilibrium frequencies a, t, c, and g, as a vector: $\overline{X}$. When l_A is substituted by: $tb + cb + ga_1$ (the sum of the elements in the first row besides l_A), one finds 0=0, as expected.

Suppose, the system is at equilibrium since t=0 (divergence of two species, for instance), or was even before at equilibrium, so that the base frequencies remain constant ($\overline{X}$). Then, when the molecular clock hypothesis holds, that is, when **R** is constant in time and the evolution approximately follows this model (Lio and Goldman, 1998), evolutionary distances in terms of an estimated number of nucleotide substitutions (including unobserved substitutions like parallel, reverse, and multiple changes) between sequences can be calculated (this is, in fact, one of the major purposes of that type of models). n(dt) denotes the number of substitutions in one linage during dt and can be calculated as: $n(dt) = \sum_i X_i(t) l_i dt$ (i=A, T, C, G).

Consequently, when the two sequences evolve according to the same law, we find for the number of substitutions until T (the distance): d= $2\int_0^T n(dt)dt = 2\int_0^T \sum_i X_i(t) l_i dt$. As $X_i(t)$ remains constant: $\overline{X}$, this simplifies to: d= $2\left(\sum_i \overline{X}_i l_i\right)T$ =2(a l_A +t l_T +c l_C +g l_G)T .

k=a l_A +t l_T +c l_C +g l_G is the average rate of nucleotide substitution per site. This is also intuitively clear: in a (infinitely long) sequence, a, t, c, and g are, respectively, the probabilties to find an A, T, C, or G in a certain position of the sequence, and l_A, l_T, l_C, and l_G are, respectively, the rates by which these nucleotides change (to any of the other nucleotides). The sum of these products, clearly, gives the average substitution rate per site.

B. Classification of the models:

I used four different criteria to classify the models: symmetry, reversibility, no-strand-bias condition, and transition/transversion (in the table: “sym“, “rev“, “no strand bias“, and “TR/TV“, respectively; the fulfillment of these criteria is indicated by a “+“ sign).

Symmetry means, that the substitution rates of the direct and reverse substitutions are equal (Zharkikh, 1994). Therefore, this condition can be expressed as: $r_{ij} = r_{ji}$ (for any i, j). Thus, a model is symmetrical when the rate matrix is symmetrical. An example: each of the parameters in the matrix of the HKY model contains one of the factors a and b that are symmetrically distributed as in the K2 model. But they also contain the equilibrium frequencies of the respective mutant nucleotide as factors; for example: $r_{21} = ab$, whereas $r_{12} = tb$. Since a is not assumed to equal t, $r_{21} \neq r_{12}$. The model is not symmetrical.

Reversibility means, that the probabilistic substitution process, described as a markov process, is theoretically indistinguishable from the same process watched in reverse (Liò and Goldman, 1998). The condition for reversibility can be expressed as: $\bar{X}_i r_{ij} = \bar{X}_j r_{ji}$ (for any i, j) where $\bar{X}_i$ is the equilibrium frequency of the nucleotide i. Reversibility leads to significant simplifications in the mathematical treatment of a model (Tavaré, 1986).

The condition can be checked by multiplying each rate below the diagonal of the matrix (that consists of the “-“ signs) with the equilibrium frequency that can be found in the same row, and comparing this expression with the opposite rate on the upper side of the matrix times the equilibrium frequency of the respective row. An example: for the T3 model, for i=2 and j=1, we find the rate q b and the equilibrium frequency t=q/2 . On the opposite side we find the same rate q b , and the same equilibrium frequency: a=q/2. For i=3 and j=1 we find: q b times p/2, and on the opposite side: p b times q/2 which gives the same expression. The equality also holds for the other pairs of i and j, and, consequently, T3 is reversible.

The no-strand-bias condition means that the equalities corresponding to the parity rule 1 (PR1, see discussion of the NSB model, below) are fulfilled: the condition: $r_{XY} = r_{\overline{XY}}$ (Lobry, 1995) yields the following equalities: $r_{TA} = r_{AT}$, $r_{CA} = r_{GT}$, $r_{GA} = r_{CT}$, $r_{AC} = r_{TG}$, $r_{TC} = r_{AG}$, and $r_{GC} = r_{CG}$ (or for the more convenient parameter names: a3=t4; a4=t3; c2=g1; c1=g2; a2=t1; c4=g3). It may be helpful to look at the NSB rate matrix to find these equalities rapidly. They can easily be checked for any model. An example: for the B4 model, we find: $r_{GA} = r_{CT}$, but already $r_{TA} \neq r_{AT}$ is sufficiant to determine that B4 does not fulfill PR1.

The transition/transversion criterion was not inspired by literature. It is supposed to indicate weather a model allows to distinguish between the rates of transitions and transversions (see discussion of the K2 model, below). When the nucleotides are arranged in the order: A, T, C, G, the transition rates are to be found on the diagonal that goes from the lower left to the upper right “corner“ of the matrix.

I tried to define the criterion in a way that leaves no ambigous cases: the TR/TV criterion is fulfilled when every transition rate term contains a factor that is not contained in any of the transversion rate terms, or every transversion rate term contains a factor that is not contained in any of the transition rate terms or both. Then the parameters for transitions and transversions can be adjusted separately. A formalization of that criterion like for the other criteria above does not seem to be possible.

An example: the SYM model contains only the parameters $a1$ and $a2$ in the diagonal, none of which can be found outside of the diagonal. Therefore, the SYM model fulfills the TR/TV criterion and is said to be able to distinguish between transitions and transversions.

More diffucult is the situation for the models that contain products as rate parameters. In the HKY model, all terms on the diagonal contain the transion rate factor a (like in the K2 model) that can not be found outside of the diagonal. Therefore, the transition rates can be altered separetely from the transversion rates, even when it is not possible to put them all on the same value (like in the K2 model), unless the equilibrium frequencies a, t, c, and g are all equal. But this is when the HKY model actually becomes the K2 model. The most “unpleasent“ model in this respect is the TK4 model, because the transition rates are a and b , but both of these parameters are also to be found as factors in transversion rates. But according to the above definition of the TR/TV criterion, the model is still said to be able to distinguish between transitions and transversions, because every transversion rate contains a factor (g or p) that is not contained in any transition rate, and, therefore, the transversion rates can be altered independently of the transition rates. The results are summarized in the table:

model	sym	rev	no strand bias	TR / TV	n.of free param.	parameters	restrictions	equilibrium frequency (remarks)
JC $\begin{matrix} - & a & a & a \\ a & - & a & a \\ a & a & - & a \\ a & a & a & - \end{matrix}$	+	+	+	-	1	a		0.25 0.25 0.25 0.25
K2 $\begin{matrix} - & b & b & a \\ b & - & a & b \\ b & a & - & b \\ a & b & b & - \end{matrix}$	+	+	+	+	2	a, b		0.25 0.25 0.25 0.25
3 ST $\begin{matrix} - & b & g & a \\ b & - & a & g \\ g & a & - & b \\ a & g & b & - \end{matrix}$	+	+	+	+	3	$a, b,$ g		0.25 0.25 0.25 0.25

model	sym	rev	no str-and bias	TR / TV	n.of free par-am.	parameters	restrictions	equilibrium frequency (remarks)
SYM - $b1$ $g1$ $a1$ $b1$ - $a2$ $g2$ $g1$ $a2$ - $b2$ $a1$ $g2$ $b2$ -	+	+	-	+	6	$a1, a2$ $b1, b2$ $g1, g2$		0.25 0.25 0.25 0.25
T3 - qb pb pa qb - pa pb qb qa - pb qa qb pb -	-	+	+	+	3	a, b p	q=1-p	q/2 q/2 p: G+C-content (p=q) p/2 p/2
EI - tf cf gf af - cf gf af tf - gf af tf cf -	-	+	-	-	4	a, t, c, f	g=1-(a+t+c)	a estimate: observed t equilibrium c frequencies g
HKY - tb cb ga ab - ca gb ab ta - gb aa tb cb -	-	+	-	+	5	a, b, a, t, c	g=1-(a+t+c)	a estimate: observed t equilibrium c frequencies g
TN - tb cb g$a1$ ab - c$a2$ gb ab t$a2$ - gb a$a1$ tb cb -	-	+	-	+	6	$a1, a2$, b , a, t, c	g=1-(a+t+c)	a estimate: observed t equilibrium c frequencies g
REV - t$b1$ c$g1$ g$a1$ a$b1$ - c$a2$ g$g2$ a$g1$ t$a2$ - g$b2$ a$a1$ t$g2$ c$b2$ -	-	+	-	+	9	$a1, a2$, $b1, b2$, $g1, g2$, a, t, c	g=1-(a+t+c)	a estimate: observed t equilibrium c frequencies g
TK4 - g pa a g - a pa pb b - g b pb g -	-	+	+	+	4	a, b, g , p		b /h b /h h=2($a + b$) a /h a /h

model	sym	rev	no str-and bias	TR / TV	n.of free par-am.	parameters	restrictions	equilibrium frequency (remarks)
TK5 $\begin{matrix} - & g & d & a \\ g & - & a & d \\ e & b & - & g \\ b & e & g & - \end{matrix}$	-	-	+	+	5	$a, b,$ $g, d,$ e		$(e+b)$/h $(e+b)$/h h=2$(a+b+d+e)$ $(a+d)$/h $(a+d)$/h
NSB $\begin{matrix} - & g1 & d & a \\ g1 & - & a & d \\ e & b & - & g2 \\ b & e & g2 & - \end{matrix}$	-	-	+	+	6	$a, b,$ $g1, g2,$ d, e		$(e+b)$/h $(e+b)$/h h=2$(a+b+d+e)$ $(a+d)$/h $(a+d)$/h
B4 $\begin{matrix} - & b & b & a \\ d & - & a & d \\ d & g & - & d \\ g & b & b & - \end{matrix}$	-	-	-	+	4	$a, b,$ g, d		$d(b+g)$/h $b(b+g)$/h h=$(a+g)(b+d)+4bd$ $b(a+b)$/h $d(a+b)$/h
GIN $\begin{matrix} - & a1 & a & a \\ b1 & - & a & a \\ b & b & - & a2 \\ b & b & b2 & - \end{matrix}$	-	-	-	-	6	$a, a1,$ $a2, b,$ $b1, b2$		$b\ (a+b1)$/h1 h1=(2$a+b1+a1$)($a+b$) $b\ (a+a1)$/h1 $a\ (b+b2)$/h2 h2=(2$b+a2+b2$)($a+b$) $a\ (b+a2)$/h2
8P $\begin{matrix} - & b2 & b3 & a4 \\ b1 & - & a3 & b4 \\ b1 & a2 & - & b4 \\ a1 & b2 & b3 & - \end{matrix}$	-	- (*)	-	+	8	$a1, a2,$ $a3, a4,$ $b1, b2,$ $b3, b4$		(*) *: see 8P model in the discussion section of the models
G12 $\begin{matrix} - & t1 & c1 & g1 \\ a2 & - & c2 & g2 \\ a3 & t3 & - & g3 \\ a4 & t4 & c4 & - \end{matrix}$	-	-	-	+	12	a2, a3, a4, t1, t3, t4, c1, c2, c4, g1, g2, g3		issue of the “rapport technique“

The relations between the models:

By equalizing parameters, models can be simplified to become identical to other models. This yields a hierarchy of models (Zharkikh, 1994). For the models presented in this work, I built such a hierarchy that can be seen in the graphic (below). The hierarchy is, effectively, an extension of the one presented by Zharkikih (though I found an error in the latter one; see below). The hierarchy also contains all the information that can be found in the four classification columns of the table (see above): a model that is placed below the REV, SYM, or NSB model and that is connected by arrow(s) to it fulfills the respective criterion, that is, reversibility, symmetry, or the no-strand-bias condition. The models that fulfill the transition/transversion criterion (TR/TV) are indicated by a "*" . Neither these criteria, nor the hierarchy that can be built by equalizing parameters, allow a strict order of the models, and that is why the order in the table above is, in fact, somewhat arbitrary. Nevertheless, it is inspired by the hierarchy in the graphic below (when one follows the table from the end to the beginning, one goes down the hierachy). The order in the discussion part of the models (below) was also inspired by historical considerations, that is, more complex models are usually built on the basis of existing, simpler models, and, therefore, when one works through the literature, it might be advantageous to, at least partially, follow the historical order of the models.

In the following, I will present the equalities that simplify models to become identical to other models. The symbols (a1, b1...) can be found at the side of the respective arrows in the graphic below. Note that one "=" sign goes along with the loss of one free parameter:

a	a2=a3; t1=t4; c1=c4; g2=g3 (note that the 8P model is depicted twice in the graphic for a better overview)	b	a3=t4; a4=t3; c2=g1; c1=g2; a2=t1; c4=g3
c	a3=a4=t3=t4; c1=c2=g1=g2	d	$b1 = b2 = g1 = g2$
e	a=t=c=g	f	$a1 = b1; a2 = b2; a3 = b3; a4 = b4$ (one obtains the EI rate matrix as in the article of Tajima and Nei (1982); see the discussion of the EI model)
g	$a1 = a2; a3 = a4; b1 = b4; b2 = b3$	h	$a1 = a2$
i	$a1 = a2; b1 = b2; g1 = g2$	j	$g1 = g2$
k	$a = a1 = a2 = b = b1 = b2$	l	$a = b$

m	a=t; c=g	n	$e = pb; d = pa$ (only one free parameter is lost since p is introduced as a new parameter)
o	a=t=c=g	p	$a = b$ (pa can then be renamed)
q	$a = g; b = d$	r	p=q
s	$b = g$	t	$a = b$

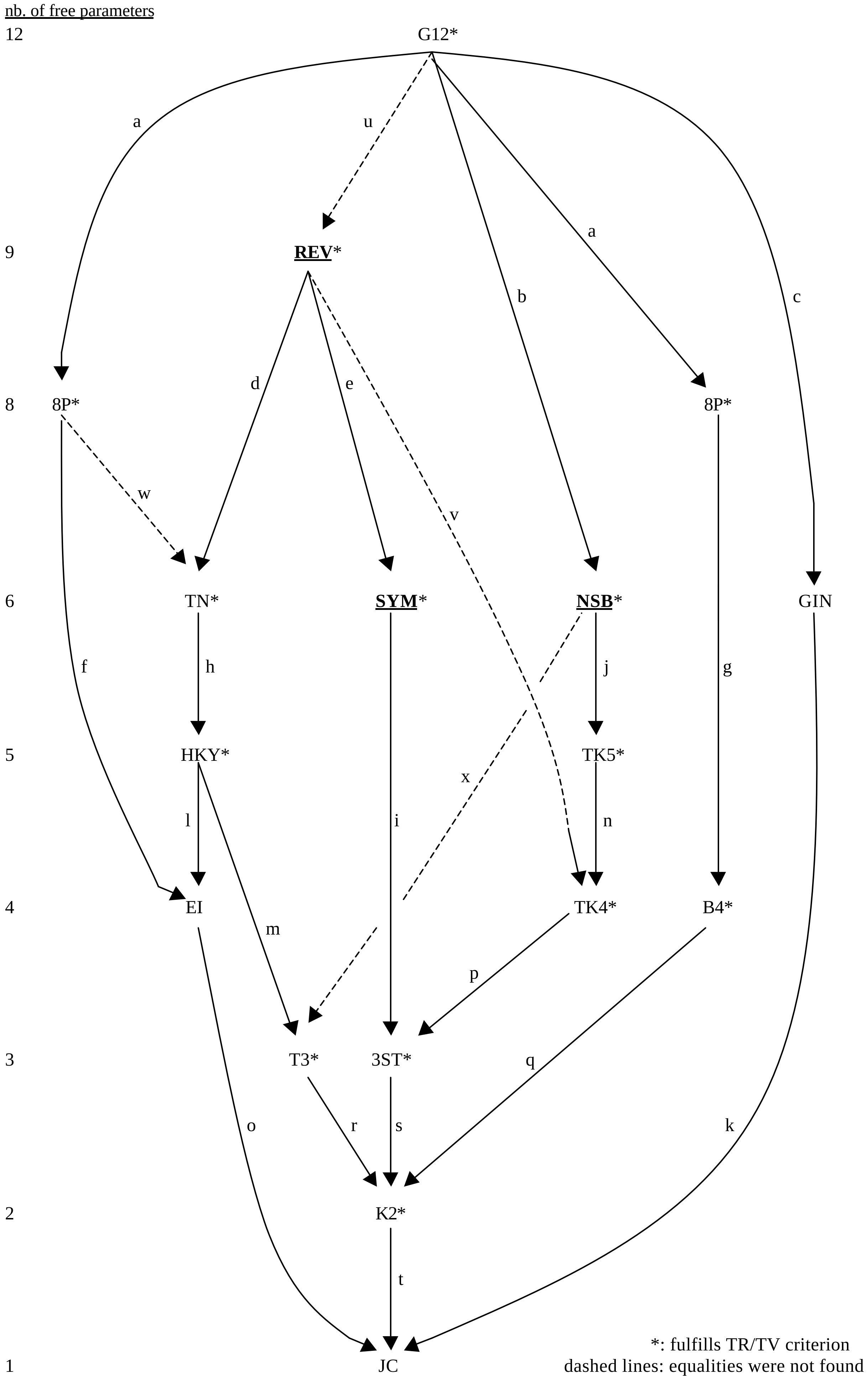
u: all 12 parameters are replaced by products so the criterion: $\overline{X}_i r_{ij} = \overline{X}_j r_{ji}$ is generally fulfilled; this reduces the number of free parameters to 9; see also explanation for points v and x (below)

v: though possible (because TK4 is, unlike TK5, reversible), there is no direct way of equalizing parameters to simplify REV to become TK4; TK4 is an exception because it contains products but it is not constructed like the other models that contain products in that it does not contain the equilibrium frequencies in the products (like does, for instance, REV); it could be transformed into such a form but it would then probably become quite complicated; see also Rzhetsky and Nei (1995)

w: see discussion of the 8P model (below)

x: here we have the reverse situation as under v: the more complex NSB model is not constructed of products that contain the equilibrium frequencies, while T3 is; in principle, it must be possible to simplify the NSB model to become T3, though, because T3 fulfills the no-strand-bias condition; a further study of this problem could be fruitful in respect to the “rapport technique“ .

As mentioned above, there seems to be an error in the hierarchy graphic in the article of Zharkikh (1994). He depicted the B4 model as a simplified model of GIN. This transition, though, is not possible as can easily be verified when one compares the rate matrices in the table. I, therefore, placed them as independent “descendants“ of the G12 model in the graphic. I found further errors in this article concerning the equilibrium frequencies. For the TK5 model, the same expressions are given as for the simpler TK4 model. I found that they are correct for the latter one, but wrong for the former one. For the TK5 model, the equilibrium frequencies can be derived from the ones of the NSB model since TK5 is a special case of this model. In fact, they are identical because, interestingly, the parameters $g1$ and $g2$, that are equalized to simplify NSB to become TK5, are not in the expressions for the equilibrium frequencies of the NSB model. Further, in the article of Zharkikh, in the expressions for the equilibrium frequencies of the B4 model, b is twice exchanged by d .



*: fulfills TR/TV criterion
dashed lines: equalities were not found

C: Discussion of major models* and methods:

*: see the table in the classification part of this work

K2: Kimura’s two-parameter model (Kimura 1980):

This model was built in order to estimate the rates of base substitutions for homologous sequences. It distinguishes between the evolutionary base substitution rates for transitions (a) and transversions (b). A transition is a difference in a position in two homologous sequences that are compared where the two bases are either both purines (A,G) or both pyrimidines (C,T). A transversion type difference is a position with a purin in one and a pyrimidin in the other sequence. This yields the rate matrix (see table in classification part). Setting the four equations $\frac{dX}{dt} = R'X$ equal to zero yields the equilibrium frequencies ¼ for the four bases.

P(t) be the fraction of transitions in the sequence (number of transition type differences divided by number of positions) at time t and Q(t) the fraction of transversions. (1-P(t)-Q(t)) is the fraction of positions with identical bases being observed.

Kimura derived differential equations for P(t) and Q(t) containing a and b as parameters. Setting them equal to zero yields the relation: 2P=Q=1/2 at equilibrium for any a and b . Equilibrium is not necessary for the application of the model, though.

The differential equations can be solved and inverted to yield expressions of the form: a T=… and b T=… where t was replaced by T, the divergence time of the two sequences. The total number of substitutions since divergence can be calculated as d=2kT (see introduction, above), where k denotes the rate of evolutionary base substitution per site. For this model k= a +2 b because all four l_i equal a +2 b , they are multiplied by ¼ each, and these products sum up to a +2 b . Therefore, d=2(a +2 b)T can be calculated using the expressions for a T and b T: $d = -\frac{1}{2}\ln(1-2P(T)-Q(T)) - \frac{1}{4}\ln(1-2Q(T))$. This equation gives an estimation of the total number of substitutions and is a correction for unobserved substitutions. For small P and Q, d converges to P+Q for any a and b .

In the special case: $b = a$, the model becomes the Jukes and Cantor (JC) model. d simplifies to: $-\frac{3}{4}\ln(1-\frac{4}{3}l)$ where l =P+Q is the fraction of sites with differing bases.

3 ST: Kimura’s three-substitution-type model (Kimura 1981):

Like Kimura’s two parameter model, this model makes the distinction between transition and transversion type differences between the bases in the compared sequences. a be the rate of transition type evolutionary base substitutions, again. Without biological explanation, in this model the transversion rate is separated into two different rates: b and g .

In all columns of the rate matrix **R** we find the same sums: $a + b + g$. Since all equilibrium frequencies are $\frac{1}{4}$, k equals $a + b + g$. The total number of base substitutions per site is, therefore, estimated by d=2kT= $2(a + b + g)T$.

P be fraction of sites showing transition type differences, again. Q be the fraction of sites with differing bases between which we find the rate b in the matrix, and R be the fraction of sites with differing bases between which we find the rate g . Q+R is the fraction of sites with transversion type differences.

In a similar fashion as in the article of Kimura (1980), formulae that permit a simple estimation of evolutionary distances are derived. Three differential equations for P, Q, and R are given, as well as their solution for the initial condition P=Q=R=0 at t=0, because at the divergence point we assume only one ancestor species with one hypothetical representative sequence. The equations of the solution can be inverted and combined to yield an expression for ($a + b + g$)T. This gives: $d = -\frac{1}{4}\ln\left[(1-2P-2Q)(1-2P-2R)(1-2Q-2R)\right]$ as an estimate for the number of base substitutions per site including unobserved substitutions.

2 FC: Kimura’s two-frequency-class model (Kimura 1981):

This model was designed in order to take into account the elevated frequencies of the bases G and C in the third codon positions of mammalian mRNA’s.

U and A (U instead of T when mRNA sequences are considered) are grouped into one group: A1, G and C into another: A2. a and b be, respectively, the substitution rate from A2 to A1, and from A1 to A2. In figure 1 (b) of Kimura’s article, the other substitution rates are named as follows: T→A: a_1; A→T: b_1; C→G: a_2; G→C: b_2. The latter parameters are not used for the calculations that are presented in the article.

The grouping of G+C and U+A is, for my understanding, a loss of information because one only considers weather a base belongs to group A1 or A2. As far as the model is discussed in the article, it does not seem worthwhile to present it in the form of a substitution rate matrix, but for completeness I will “force it” into such a matrix:

$$\begin{array}{cccc} - & b1 & \tilde{a}1 & \tilde{a}2 \\ a1 & - & \hat{a}1 & \hat{a}2 \\ \tilde{b}1 & \tilde{b}2 & - & a2 \\ \hat{b}1 & \hat{b}2 & b2 & - \end{array}$$

with the restrictions: $\tilde{a}1+\tilde{a}2=\hat{a}1+\hat{a}2=a$, and $\tilde{b}1+\tilde{b}2 = \hat{b}1+\hat{b}2 = b$. This appears, in fact, as a model with eight free parameters (degrees of freedom), though Kimura introduced only six parameters (two of which he only used in his actual calculations). The actual values in the matrix seem to be of little importance, as long as the restrictions are fulfilled.

For these reasons, the matrix is not presented in the classification part of this work. If we choose: $\tilde{a}1=\tilde{a}2=\hat{a}1=\hat{a}2=a$ /2 and $\tilde{b}1=\tilde{b}2 = \hat{b}1=\hat{b}2 = b$ /2 (six degrees of freedom), though, we obtain the matrix:

$$\begin{array}{cccc} - & b1 & a/2 & a/2 \\ a1 & - & a/2 & a/2 \\ b/2 & b/2 & - & a2 \\ b/2 & b/2 & b2 & - \end{array}$$

which is in fact the matrix of the GIN-model, only with different names for the parameters (see below: discussion of the GIN-model).

Kimura denotes by X the frequency of base pairs where both bases are in group A1, by Y : both in A2, and by Z: one base in A1 and the other in A2. Therefore: X+Y+Z=1. He gives differential equations for X,Y, and Z. He further assumes equilibrium for the frequencies of A1 and A2. Therefore, p=freq.(A1)= $b/(a+b)$, and q=freq.(A2)=1-p . Finally, he finds: $d=-q\ln(1-Z/q)$ where $q=2pq$. As Z approaches to zero, d converges to Z for any q . For q =3/4 the formula is equivalent to the formula in the work of Jukes and Cantor (see discussion of the K2 model, above).

Kimura gives examples for calculations and estimations of the standard deviations. Probably because this model includes a loss of information, the standard deviations are always larger than that of the K3-model. He argues, though, that the estimated distances are more accurate than the ones found with K3 for large C+G-content biases and large divergence times T.

TK5 and TK4: The models of Takahata and Kimura (Takahata and Kimura 1981):

Similar to Kimura's 2FC-model, TK5 allows for different equilibrium frequencies of the bases A and U in comparison with G and C. The equilibrium frequencies of A and U are equal, though, as for G and C. TK5 does not have the loss of information as was described for the 2FC-model. It has five degrees of freedom and includes a different parametrization than 2FC, again without any biological explanation (other than that it allows for a G+C-content different from equality). a , for example, is the transition rate from U to C and A to G, b the rate for the reverse transitions. The transversion rates are g ,d , and e . P, Q, and R are defined as for Kimura's 3ST model. In addition to that, the frequencies of the following couples (bases at the same position in the two sequences) are defined: UC: P1 ; AG: P2 ; UA: Q1 ; CG: Q2 ; UG: R1 ; AC: R2 . For the reverse couples (CU…) the authors imply the same set of frequencies (P1…). (For this model, all practical values that are used for the calculations be the arithmetical means.) Further, the frequencies of the couples UU, CC, AA, and GG are denoted S1, S2, S3, and S4, respectively. S equals S1+S2+S3+S4, P=2P1+2P2, the same for Q and R.

One possibility to derive differential equations for S1, S2, S3, S4, P1, P2, Q1, Q2, R1, and R2 that is explained in the article can be summarized as follows: for the four bases, difference equations can be written down in a straight forward fashion, for U(T), for example, we obtain: $U(T+\Delta T)=\{1-(a+d+g)\Delta T\}U(T)+b\Delta TC(T)+e\Delta TC(T)+g\Delta TA(T)$. The first term, for instance, represents the probability for U to remain unchanged during ΔT . These equations can be transformed into differential equations which in turn can be used to replace the derivatives in relationships as: $\frac{dP1(T)}{dT}=U(T)\frac{dC(T)}{dT}+C(T)\frac{dU(T)}{dT}$. This is the differential of: $P1(T)=U(T)C(T)$ which follows directly from the assumption that the two linages evolve independantly. One obtains the differential equations that were sought.

The treatment of this system of differential equations is rather complicated and will not be discussed here. The goal is to replace the parameters by observables in the following expression for the estimated distance: $d=2kT=2(g+2w(1-w)(a+b+d+e))T$ where w is the frequency of U+A in the sequences and that can be expressed as: $w=(b+e)/(a+b+d+e)$. The authors use the simplification that w be constant in time. Further use the authors the simplification: $d=pa$ and $e=pb$, where p is a constant that can be estimated (a formula is given in the article), to facilitate the mathematical treatment of the system. This reduces the degrees of freedom to four independant parameters. w simplifies to $\frac{b}{a+b}$. This simplified model is denoted TK4. For the special case of $b=a$, TK4 becomes the 3 ST model of Kimura ($b=a$ causes $e=d$). For the TK4 model, the authors succeeded in finding an expression for d that contains only observables (for instance P, Q1, S1, w …). Interestingly, the somewhat critical parameter p is not in this expression.

The authors performed a great deal of Monte-Carlo-simulations to test the accuracy of the formula and, therefore, the TK4 model (in comparison to the 3ST and the Jukes and Cantor model; the 2 FC model was not mentioned any more). They found that the estimated distances were almost identical if the substitution rates were equal in all directions. The TK4 model performed better, though, when the transversion rates were low and when there was a strong G+C- content bias. Unfortunately, there are cases when these formulae are inapplicable because arguments of logarithms can become negative, especially when $d \geq 1$ and for small sample sizes, for instance less than 100 positions compared.

GIN: A further analysis of Kimura's 2 FC model by Gojobori, Ishii, and Nei (1982):

The authors provided a re-consideration of Kimura's 2 FC model (see above) with a further mathematical analysis by which they obtained somewhat different formulae than Kimura in his work. They also performed computer simulations in order to compare the performance of this model with the ones of Takahata and Kimura's TK4 model and Jukes and Cantor's model.

The rate matrix shows that the authors used a different parametrization than Kimura originally introduced. The differences are not explained, though, and might possibly cause some confusion since the same names of parameters are used for different rates. The mathematical analysis in Gojobori et al's paper seems more accurate, though, and the use of parameters within this work is consistent, so I used their rate matrix in the classification part. The mathematical analysis in this paper is more complex than the one performed by Kimura and will not be presented here. Essentially, the authors built their analysis on the different frequencies x_{ij}, i and j representing one of the four bases in a homologous position of the two sequences. In that formulation, the expression that the authors found for the distance d gets quite complex, as does the expression suggested by Kimura when transformed into that formulation. The authors show, though, that Kimura's formula is in fact wrong and that it gives slight overestimates for d. The extend of the overestimate is small, though, when the number of nucleotide substitutions is small.

The computer simulations gave results that can be summarized as follows: the Jukes and Cantor- and the TK4 model often give underestimates when the number of nucleotide substitutions is large, whereas the GIN model gives good estimates for the examined substitution schemes. (The performance of a model, naturally, depends on how well it reproduces the real substitution process. The more realistic the assumptions made in the model are, the better its performance.) However, the latter two models produce many inapplicable cases as the number of nucleotide substitutions gets large, unless the compared sequences are very long.

EI: The Equal Input model and the equal output model (Tajima and Nei, 1982; Tajima and Nei, 1984):

In the first article, the authors present a formula for the calculation of the average number of nucleotide substitutions per site for data obtained by the restriction enzyme technique and a discussion of its bias. That technique and its theory is not issue of this work since nowadays sequence data is readily available. The authors introduce two substitution models that are of interest, though, even when they considered them from a different point of view.

Generally, the mean number of nucleotide substitutions per site since divergence T years back can be calculated as: d=2kT where k=$a\,l_1 + t\,l_2 + c\,l_3 + g\,l_4$ (see introduction, above).

The equal input model was inspired by the observation that base frequencies often are neither the same for A, T, C, and G (at equilibrium: a, t, c, g), nor they fulfill the assumption: a=t and c=g as predicted by models like TK5 or GIN. According to the EI model, the substitution rate to the i-th nucleotide is the same regardless of the original nucleotide. From that assumption, one obtains the substitution rate matrix as in the paper of Tajima and Nei:

$$\begin{array}{cccc} - & a2 & a3 & a4 \\ a1 & - & a3 & a4 \\ a1 & a2 & - & a4 \\ a1 & a2 & a3 & - \end{array} \quad .$$

The equilibrium frequencies are: $\overline{X}_i = ai/(a1 + a2 + a3 + a4)$ =ai/f. In the classification part, I chose a more convenient presentation of the same model: from $\overline{X}_i$=ai/f we get: ai= $\overline{X}_i$ f , where f be a factor that serves as a free parameter in the model. Three of the four equilibrium frequencies ($\overline{X}_i (i = 1...4) \equiv a, t, c, g$) can consequently be directly considered free parameters and their estimates are observables in the sequences under study; the four'th is determined by: a+t+c+g=1.

In the second article, the authors developed a formula for the estimation of the distance d, based on the EI model. With an empirical correction, they improved the formula to give quite reliable estimates for many different substitution schemes, that means schemes that do not follow the EI model:

They started from the formula in the JC model: d=$-\frac{3}{4}\ln(1-\frac{4}{3}l)$ (l being the fraction of sites with differing bases; see above: K2), that can be expressed as: d=$-b_1 \ln(1 - l/b_1)$, where

$b_1 = 1 - \sum \overline{X}_i^{\,2}$. For any scheme of nucleotide substitution, b_1 represents the value of l for $t \to \infty$ (then, $d \to \infty$). The two formulae for d are equivalent for a=t=c=g ($b_1 = 3/4$). Using a discrete time model, the authors showed that the second formula for d is a non-biased estimator for the EI model. When the substitution scheme does not follow this model, though, simulations suggested that it gives underestimates. The authors derived another formula for d for the EI model that they showed to give overestimates when the substitution scheme differs from EI. They concluded that the combination of the two formulae may give reliable estimates for any substitution scheme: d=$-b\ln(1-l\ /b)$, where b=$\left(1-\sum \overline{X}_i^{\,2} + l^2/h\right)/2$, where $h = \sum_{i=1}^{3}\sum_{j=i+1}^{4} x_{ij}^{\,2}/2(\overline{X}_i \overline{X}_j)$, where x_{ij} is the fraction of positions with nucleotide i in one of the sequences and j in the other one. This equation holds exactly for the EI model, but the authors showed that for a variety of observed and by random synthesized data, it gave considerably better estimates than the methods: JC, K2, 3ST, TK4, and GIN, when the distances were small, say $d \le 1$. For larger distances, the GIN method was shown to give the best results (when the inapplicable cases were excluded).

In the article of 1982, the authors also considered another model that they named "equal output model". Though it found less attention in the literature, for completeness I will show its substitution rate matrix (in a somewhat modified way):

- a1 a1 a1

a2 - a2 a2

a3 a3 - a3

a4 a4 a4 - .

This can be interpreted as a model according to which the substitution rate depends on the original nucleotide, but when there is a change, it is substituted by any of the other nucleotides with the same probability. The equilibrium frequencies of this model are:

$$\overline{X}_i = \frac{1}{ai(\frac{1}{a1}+\frac{1}{a2}+\frac{1}{a3}+\frac{1}{a4})}$$

=g/ai , g being another factor. An alternative presentation as for the EI model would be possible.

T3: A model for (strong) transition/transversion and G+C-content biases (Tamura 1992):

This model was designed to allow a better estimation of the number of nucleotide substitutions per site (d) for the special case of sequences with a strong transition/transversion bias and a strong G+C-content bias, a situation that can be found, for instance, in the third codon positions of mitochondrial DNA in Drosophila.

The model is an extension of Kimura's model K2 to the case where the G+C-content p (in the paper:q) differs (considerably) from 0.5 . (q=1-p is the A+T-content.) The matrix of the model can be seen in the classification part. It is a special case of the HKY-model by Hasegawa et al. (see below) where a=t and c=g: the equilibrium frequencies of A and T, and of C and G, are assumed to be equal.

For this model we find $l_A = l_T = pa + (p+q)b = \mathrm{p}a + b$ and $l_C = l_G = qa + (p+q)b$ $=\mathrm{q}a + b$. The equilibrium frequencies are q/2 for A and T and p/2 for C and G. Therefore, one obtains: $k = \sum_i l_i \overline{X}_i = (pa + b)q + (qa + b)p = 2pqa + b$. d is, then, 2kT.

In a similar fashion as in the works of Kimura (1981) or Takahata and Kimura (1981), Tamura derives a formula for an estimate of d that contains only observables. Since p (and, therefore, q) are observables, themselves (to be precise: their estimates are observables in the sequences), only a and b remain to be replaced in the expression for d. He finds the following formula: $d = -2pq\ln\left(1 - \frac{1}{2pq}P(T) - Q(T)\right) + \left(pq - \frac{1}{2}\right)\ln\left(1 - 2Q(T)\right)$, where, as in the work of Kimura (K2: 1980), P(T) and Q(T) are, respectively, the fractions of transitions and transversions in the sequences. For the case that the G+C-content p is not the same in the two sequences, Tamura gives a corrected formula for d (not shown).

In a numerical example for distance calculations between drosophila species and in computer simulations, Tamura shows that his formula apparently performs better than the ones of the models: JC, K2, EI, TK4, and GIN when there are strong transition/transversion and G+C-content biases (these models give underestimates of d). This does not seem to be surprising, though, since the sample sequences were obtained according to the same rate matrix that was used to derive the formula for d. On the other hand, the standard deviations of d get large, as does the number of inapplicable cases (even for long sequences under study), when in addition to the large biases, d itself also gets large ($\mathrm{d} \geq 1$) .

HKY: A model for transition/transversion and equilibrium frequency bias (Hasegawa, Kishino, and Yano 1985):

This model was designed for the estimation of divergence dates, especially for the separation of mouse, gibbon, orangutan, gorilla, chimpanzee, and human. For the model, a new statistical method was developed.

The model is constructed as a combination of the EI and the K2 models. Like in Kimura's K2 model, it involves a common factor for the transition rates (a) and one for the transversion rates (b). On the other hand, like in Tajima and Nei's EI model, it introduces as common factors the equilibrium frequencies of the mutant nucleotides for the rates in the columns. Like this, it allows for any (biased, that means unequal to $\frac{1}{4}$) equilibrium frequencies and for a (strong) transition/transversion bias.

The statistical method designed to estimate the divergence times is not exactly straight forward and rather complicated. It is based on an a priori existing tree topology that is inferred by other methods and data, for instance paleontological knowledge. It involves a specific value decomposition of exp(**R**t) in the form: $\mathbf{R}=\sum_i^4 \Lambda_i \vec{u}_i \vec{v}_i\,'$ from where follows:

$\exp(Rt) = \sum_i^4 \exp(t\Lambda_i)\vec{u}_i \vec{v}_i\,'$. The vectors $\vec{u}_i$ and $\vec{v}_i$ are, respectively, the eigenvectors of **R** and **R'** , and they are given in the article. Further, the transition- and transversion type differences are expressed in terms of the equilibrium frequencies. Instead of working out formulae for a pair of sequences under study, the method involves a least-squares fitting of the whole set of data, that means for all the sequences at once. The authors pointed out that only two sequences were not sufficient to give a reliable estimate for a, b, and f, the fraction of variable sites (another extension of the models discussed so far; not all positions are supposed to be variable in the evolution).

As described in the article, an interesting property of the number of differences between two sequences that are transitions and transversions is that, while the number for the transversions increases monotonously in time, the number for the transitions reaches a maximum, and, thereafter, decreases. This is always the case when $a > b$.

B4: A method of estimating the ancestral composition and rates of substitution for two aligned sequences using discrete time matrix methods (Blasidell 1985):

Unlike the models and methods discussed so far, Blasidell's method is independant of the rate matrix used. It would deserve its own chapter, but here I will only present its main ideas. Though the method is independant of the rate matrix used, Blasidell introduces a rate matrix that proved useful when applied in his method (he compared its performance with the ones for: JK, K2, and TK4). The matrix can be found in the classification part. Blasidell does not give any (biological) explanation for the design of that matrix other than that it allows the distinction between transition and transversion rates and that these may be different in the forward and backward directions. The equilibrium frequencies of purines (and pyrimidines)

need not be the same, and the number of positions where base i is in sequence 1 and j in sequence 2 need not be the same as the number of positions where j is in 1 and i is in 2.
Like most of the other methods described, this method assumes a Markov process as model for the molecular evolution and equal rates of substitution for all positions. Unlike the other methods, though, it does not assume that the base composition does not change with time, that the substitution rates are equal for the two linages, and that the present differences between the two sequences (the matrix of these differences be the “divergence“-matrix) equal twice the difference between each sequence and the common ancestor (the matrix of these differences be the “substitution“-matrix). A divergence matrix consists of elements that are similar to the different fractions x_{ij} in the GIN model (see above), i and j representing one of the four bases in a homologous position of the two sequences. In the divergence matrix, though, one uses directly the numbers of counts instead of the fractions. The element in row “A“ and column “T“, for instance, represents the number of positions where there is an A in the first and a T in the second sequence, whereas the element in row “T“ and column “A“ is the number for T in the first and A in the second sequence, the sequences being well distinguished from each other.
The Markov process is treated as a Markov chain with a finite number of states (four bases) and is developed through successive epochs (steps) by matrix multiplication. The substitution rates are not necessarily constant in time because the assumption of constant probability of substitution in each epoch does not imply that epochs correspond to constant intervals of time. **S**={s(i,j)} (i, j = A, T, C, G) be the first-order Markov transition matrix for the first linage. s(i,j) is the probability that the base at a given position is j at epoch k when it was i at epoch (k-1) (independant of the epochs before k-1). **S**(k)={s(i,j,k)} (k=0,1,...) be the transition matrix for the k epochs. s(i,j,k) denotes the probability that the base is j at epoch k when it was i at epoch 1. Note that, unlike in substitution rate matrices, in transition matrices the sum of each row must equal 1. $\mathbf{S}$(k) is obtained by matrix multiplication: **S**(k)=$\mathbf{S}\,(1)^k$.
T(k)={t(i,j,k)} be the transition matrix for the second linage. Then, the divergence probability matrix for the two linages is: **D**(k)={d(i,j,k)}. d(i,j,k) is the fraction of positions where there is base i in the first and j in the second sequence at epoch k. Therefore, $\sum_{i,j} d(i,j,k)=1$. Note that d(i,j,k) does not necessarily equal d(j,i,k) (the sequences are distinguished). c(l, 0) be the fraction of base l at epoch 0 (the ancestral sequence). c(G, 0) can be calculated as 1-(c(A, 0)+c(T, 0)+c(C, 0)) . That gives already three free parameters to be estimated by the discrete matrix method. Now, d(i,j,k) is calculated as follows: d(i,j,k)= $\sum_{l} c(l,0)s(l,i,k)t(l,j,k)$.

The estimation of the parameters is performed by nonlinear least-squares minimizing of the squares of the differences between the calculated divergence matrix **D**(k) and the observed divergence matrix (this process is not explicitly described in the article, but Blasidell refers to the iterative “complex“ optimization of Box from 1965). One can use the average values for the two sequences in the observed divergence matrix as initial base composition in the ancestor sequence. For initial rates, one can use the average values (for the parameters used in the actual rate matrix model) in the observed divergence matrix divided by the chosen number of epochs (for instance 4 or 16).

The parameters to be estimated are three parameters defining the base composition in the ancestral sequence, as well as all the parameters in the transition matrices **S** and **T**. As each of the rows of these matrices sums to 1, there are 12 free parameters left for each matrix. That gives a total of 27 possible parameters that determine the Markov process. As $\sum_{i,j} d(i,j,k) = 1$, though, one has only 16-1=15 independant observables in the divergence matrix $\mathbf{D}$. Therefore, the number of parameters used for the matrices may not exceed 12, altogether. On the other hand, the speed and reliability of the optimization process decreases rapidly as the number of parameters to be estimated increases. Therefore, one should keep this number small. If one uses, for instance, Blasidell’s rate matrix with 4 free parameters and one assumes the same matrix for both linages, there are 7 parameters to be estimated.

Blasidell used several matrices for his method, and he compared the results with the ones obtained by using methods that give an estimate of the distance in a closed form, a formula. Using his results, he argues that the other methods give the correct result only for the distance between the observed present-day sequence and its inaccessible ancestral sequence. He states: *“When using the present-day divergence between two sequences to infer the evolutionary distance between them, all these methods make the obviously incorrect assumption that the divergence between two sequences is twice the divergence of each from their common ancestral sequence. It is clear that the divergence between two sequences will on the average be less than twice the divergence of each from the common ancestral sequence, because some of the net substitutions in the two linages may be the same.“* I admit that I may well have misunderstood the author’s argumentation, but it appears to me that Blasidell was wrong in his statement. When using the formula: d=2kT for the evolutionary distance, and that is what he seems to refer to in his statement, one calculates the estimate of the total number of substitutions that occured in the course of the evolution of the two diverging sequences, corrected for multiple and parallel substitutions. Blasidell, on the other hand, aimed to calculate the fraction of positions that differ between two sequences, although this also involves a correction for multiple and parallel substitutions. This is a different definition of a distance (“divergence“). If my impression is correct, there should have been some discours in the scientific literature, after Blasidells article. In the course of my bibliographical work, so far, I did not find any evidence that Blasidells argumentation had been criticized.

TN: A method allowing for a transition/transversion bias, unequal nucleotide frequencies, a purine/pyrimidine bias among the transitions, and substitution rate variation among different sites (Tamura, Nei: 1993):

The model was designed to provide a good estimation of the evolutionary distance and the number of transitional and transversional changes per site for sequence data of the control region of mitochondrial DNA in humans and apes that evolves at a high rate, with a strong transition/transversion bias, and where the rate varies extensively among sites.
To estimate the relative frequencies of substitutions between the four bases, the authors first constructed a phylogenetic tree using the neighbor-joining method for sequence data from humans from different parts of the earth, and from apes. Then, they applied a parsimony method to infer the ancestral sequences in the tree. The directional changes of nucleotides could then be counted comparing each sequence with its immediate ancestral sequence. The observed numbers for the 12 possible directional changes between the four nucleotides were then divided by the observed nucleotide frequencies of the original nucleotide (calculated from all sequences together), to yield the relative frequency of each class of directional substitution. Finally, these were scaled to be 100% in the sum and presented in the form of a (observed) substitution rate matrix. The values in the matrix prompted the authors to assume that the rates would depend on the mutant nucleotide, so they introduced the equilibrium frequencies a, t, c, and g as factors into their model (as in the EI model), that there was a strong transition/transversion bias, so they used the factors α and β (as in the K2 model), and that the rates among the pyrimidines ($T \leftrightarrow C$) were higher than among the purines ($A \leftrightarrow G$) for the transitional changes, so instead of α, they allowed for two factors α_1 and α_2. The resulting rate matrix can be found in the classification part. When $\alpha_1 = \alpha_2$, the model becomes the HKY model.
To develop the method, the authors first derived a formula for the estimated number of nucleotide substitutions per site, in a similar fashion like in the work of Kimura (K2: 1980). The average rate of nucleotide substitution per site is: k= $2ag\alpha_1 + 2tc\alpha_2 + 2(a+g)(t+c)\beta$, as can easily be verified. This is used in the formula: d=2kT, and, since a, t, c, and g can be replaced by the observed equilibrium frequencies, only α_1, α_2, and β remain to be replaced. The author refers to a method described by Tamura, that involves the expression of the number of positions with transitional differences between purines (P1), between pyrimidines

(P2), and with transversional differences (Q), in terms of a_1, a_2, b , a, t, c, g, and T. With these three equations, the three parameters can be replaced. One obtains an equation for the estimation of d that contains only observables. For d$\geq$1, the formula produces many inapplicable cases. The large-sample variance of d can be estimated by evaluating the expression: V(d)=[$(c1^2 P1 + c2^2 P2 + c3^2 Q) - (c1P1 + c2P2 + c3Q)^2$]/n , where c1=$\frac{\partial d}{\partial P1}$, c2=$\frac{\partial d}{\partial P2}$, c3=$\frac{\partial d}{\partial Q}$, and n is the number of positions.

For the case that the rates are allowed to vary with position, the authors gave corrected formulae for P1, P2, and Q, that yield a corrected formula for d. For the estimated rate per site, k, they assumed the negative binomial distribution which is generated when k follows the gamma distribution. This introduces another parameter (in the paper: “a“) to be estimated into the method. For its estimation, one can use the mean and variance of the number of nucleotide substitutions per site.

Without derivation, the authors further give formulae for the estimation of the average numbers of transitional and transversional substitutions.

When applying this method to the actual sequences, the authors found for the human-chimpanzee splitting a value for d that was substantially larger than that obtained by the K2 method, which, they argued, was an underestimate due to the fact that the model was not appropriate for this type of data.

NSB: The general model under No-Strand-Bias Condition (Sueoka 1995; Lobry 1995):

This model is built on the biological assumption that there are no biases in the substitution rates between the two DNA strands, that is mutation and selection follow the same rules in both strands.

According to the base pairing rule of Watson and Crick, when a base changes in the first strand, the base in the same position of the second strand changes to the complementary base of the mutant base in the first strand. The partial substitution rates caused by changes that affect the first strand, only, be named s, the substitution rates caused by changes that affect the second strand be named t. The partial substitution rate of the base X in the first strand towards the mutant base Y, caused by a change that affects only the first strand, is s_{XY} . Consequently, this causes a secondary substitution rate $s_{\overline{XY}}$ on the second strand from the complementary base of X: $\overline{X}$ towards the complementary base of Y:$\overline{Y}$. On the other hand causes the partial rate t_{XY} on the second strand the secondary rate $t_{\overline{XY}}$ on the first strand. One obtains the total substitution rates as sums of the partial rates: $r_{XY} = s_{XY} + t_{\overline{XY}}$, and $r_{\overline{XY}} = s_{\overline{XY}} + t_{XY}$, because a substitution in one strand is caused either by a change that directly affects this strand, or by a substitution that affects the other strand when the complementary nucleotide of the original one is substituted by the complementary nucleotide of the mutant nucleotide. Under no-strand-bias condition, the partial substitution rates s_{XY} and t_{XY} are equal, and $s_{\overline{X}\overline{Y}}$ and $t_{\overline{X}\overline{Y}}$ are equal, since one assumes the same evolutionary processes for both strands. By substitution, r_{XY} can be expressed as $s_{XY} + s_{\overline{X}\overline{Y}}$, as can $r_{\overline{X}\overline{Y}}$. r_{XY} and $r_{\overline{X}\overline{Y}}$ are, therefore, under no-strand-bias condition, equal (PR1-hypothesis). (An example: $r_{AC} = r_{TG}$.) The relation $r_{XY} = r_{\overline{X}\overline{Y}}$ provides, altogether, six equalities of parameters in the rate matrix and, consequently, reduces the number of free parameters to six. The matrix can be found in the classification part.

For this model, the equilibrium frequencies a and t become equal, and so do c and g (PR2-hypothesis). Therefore, one can test the no-strand-bias condition by testing the equalities a=t and c=g in sequence data (for either one of the two DNA-strands).

These equalities are well fulfilled for a great variety of species when considering long sequences, even when one reorganizes the data in order to delete the possibly neutralizing effect of genes organized in opposite orientations. In short sequences, the no-strand-bias condition is generally not fulfilled, probably because of unequal codon usage. Codons that contain a fourfold degenerate third position are not used with equal probability though they code for the same nucleotide. The unequal usage seems to depend especially on the amino acid that is being coded (and also on the G-C content of the region where the gene is located). An explanation for this bias could be the relative abundance of tRNA for the different codons that code for the same amino acid. That would mean that selectional pressure rather than mutational bias causes the strand-bias. Since among the codons that, among fourfold degenerate codons, are preferably found in coding sequences (especially in frequently expressed genes) there seems to be no general tendency to contain certain bases rather than others in the third position (for instance, it is not that in these codons T, for example, is found more frequently in the degenerate position), on the average when considering longer sequences, the strand-bias effect declines, that is, the frequencies of A and T are about equal, as for C and G.

<u>REV: The general reversible model (Yang, 1994)</u>:

First introduced by Tavaré (1986), Yang studied the REV model using a maximum likelihood approach in order to estimate substitution patterns in real sequences, and he compared its performance with the ones of the HKY and the general model (12 parameters). His approach is an alternative to the method of parsimony analysis that was described for the TN model (see above).

The REV model is a combination of the general symmetric (SYM) and the EI model. The construction of its rate matrix (classification part) is not biologically justified, but it is mathematically convenient. Like the EI model, it allows for any equilibrium frequencies. The used method is not described in the article and goes beyond the scope of this work, so I will only summarize some major results: for a dataset of pseudogenes, the REV model fitted the data much better than HKY. For mtDNA data, the differences were not statistically significant. The author generally recommended the usage of the REV model, especially for large data sets or sequences with “extreme substitution patterns“. The general model was not found to improve the analyses and its use “does not appear to be worthwhile“.

8P: The Eight-Parameter model (Rzhetsky and Nei, 1995):

In the article, the authors present model-specific test statistics for the applicability of the models: JC, K2, EI, HKY, T3, TN, and a new model that they call “Eight-Parameter model“. Further, they developed analytical formulae for the estimation of evolutionary distances for the HKY and their 8P model. Here, I will only discuss the 8P model.
The substitution rate matrix **R** can be found in the table (above). The equilibrium frequencies of the model are:

$$a = [(b_2 + b_3)b_1 + (b_1 + b_4)a_1]/m2m4$$

$$t = [(b_2 + b_3)a_2 + (b_1 + b_4)b_2]/m3m4$$

$$c = [(b_2 + b_3)a_3 + (b_1 + b_4)b_3]/m3m4$$

$$g = [(b_2 + b_3)b_4 + (b_1 + b_4)a_4]/m2m4$$

where mi denote the eigenvalues of **R** (and **R‘**):

m2=$-(a_1 + a_4 + b_2 + b_3)$

m3=$-(a_2 + a_3 + b_1 + b_4)$

m4=$-(b_1 + b_2 + b_3 + b_4)$.

The eigenvalue m1 equals zero, and, therefore, its associated eigenvector of **R‘** is a basis of the subspace of equilibrium points of the system $\frac{dX}{dt} = R'X$ (see introduction) (Lobry, 1995).
The authors rewrote the matrix **R** in terms of the equilibrium frequencies a, t, c, and g:

- tb2 ck2 ga4

ak1 - ca3 gk3

ak1 ta2 - gk3

aa1 tb2 ck2 - .

$b2$ and $a1...4$ are in this formulation not the same parameters as in the original formulation (even though they can be found in the same position of the matrix) because, evidently, for instance: $a1$=a $a1$ does not hold. For better clearity, the authors should have used new names for these parameters, for example: $a1'$, that could then be defined as: $a1' = a1$/a.

$k1, k2,$ and $k3$ are rather complicated expressions that contain all parameters of the original matrix ($a1...4, b1...4$) and the equilibrium frequencies a, c, and g (but not t). Note that this formulation of **R** is not equivalent to that of, for instance, the EI or the REV model, because $k1$ and $k3$ still contain the equilibrium frequencies a, c, and g, and, therefore, they are not independant parameters. The authors showed this formulation of **R** to argue that the 8P model is not reversible. The authors further stated that the TN model (and, therefore, also JC, K2, EI, HKY, and T3) is a "decendant" of the 8P model. For this transition, the equalities $a3 = a2, a4 = a1,$ and $k1 = k2 = k3 = b2$ would have to be fulfilled. In fact, $k1, k2,$ and $k3$ contain the expressions: $(a4 - a1)$ and $(a2 - a3)$ as factors such that $k1 = k2 = k3 = b2$ is "automatically" fulfilled when $a3 = a2$ and $a4 = a1$. Note, again, that these two equalities refer to the "new" parameters $a1...4$, not the original ones that can be found in the rate matrix in the table (above). This is also why the arrow for this transition (w) in the graphic (above) is depicted as a dashed line. To express these equalities in terms of the original parameters, one would have to evaluate the equations: $a4/g = a1/a$ and $a3/c = a2/t$.

D: Discussion (in respect to the "rapport technique"):

In the introduction, examples for violations of the assumptions that are made for many models and methods were given. One of them was that all sites under study evolve according to the same probabilistic process, that means, with the same substitution rate. Several models were found to fit the data considerably better when one allows for different rates at different sites (for instance: Tamura and Nei, 1993). Especially the flexible gamma-distribution was found to be appropriate to describe the variation of substitution rate among sites, even when it was replaced by its discrete representation with only four states (Lio and Goldman, 1998). But another source of systematic error in the calculations may simply result from substitution models used that do not reflect the real substitution patterns in an appropriate manner (Lio and Goldman, 1998). As many models as there are, as many attempts to improve the representation of the real substitution processes were made. The K2 model, for example, allows for a transition/transversion bias, the EI model allows for unequal equilibrium frequencies etc.. Most of the articles have one thing in common: the authors claim that their models and methods are superior to the others existing so far (which may well be the case for certain types of data), and sometimes it seems difficult to judge weather the methods used to show the superiority of their models (usually simulations) were convincing. When the authors synthesized data according to the same substitution pattern that is represented in the rate matrix of their model, it may not be surprising when their method, developed on the basis of that same model, estimates the distances more accurately than other methods. One the other hand, large comparative studies have been performed to test a variety of models (for instance: Zharkikh, 1994), but even these do not seem to be able to point out a model that is superior to all the others under any condition. Rzhetsky and Nei (1995) argued that models with a larger number of parameters may fit various sets of sequence data, but the estimates of evolutionary distances have larger variances than the ones obtained when using a simpler model that fits the data equally well. The larger variances can cause errors in the estimates. Therefore, the most appropriate model will be the simplest model that is still able to fit the sequence data well.

The key for future improvements seems to be to introduce more biological realism into the models. This work included a classification of major models that may be a contribution to the attempt of evaluating models in respect to the biological realism of their assumptions. It included four criteria, two of which are purely of mathematical interest (symmetry, reversibility), but the other two of which reflect biologically relevant assumptions. It seems to be an important property to allow for a distinction between transitions and transversions, since transitions are generally found to occur more frequently than transversions (Kimura, 1980; Tamura, 1992). The no-strand-bias condition comes from the biologically relevant assumption that the two DNA strands are affected by the same evolutionary processes. It has been shown that the PR2-hypothesis, which is an indicator for the fulfillment of the NSB condition, is well justified for a great variety of species when considering long sequences (so it may be considered an asymptotic property) (Lobry, 1995), and even when it was shown that PR2 does not hold when considering shorter sequences, the NSB condition may still be a good approximation of the real biological processes. That is why the general no-strand-bias model (that also allows for a distinction between transition and transversion rates) may be a good model to start from in respect to the “rapport technique“ , the goal of which it will be to build and study a model that allows for a violation of the NSB condition.

In the course of this work, I found several aspects that may possibly deserve a more in-depth study in the “rapport technique“, for instance:

- calculate and discuss the expressions for the equilibrium frequencies of the general model (G12)
- express the parameters of the NSB model in terms of the equilibrium frequencies, that is, the A+T-, and the G+C- content (since a=t and c=g (PR2)), and rewrite the rate matrix (see Annex, below)

- combine properties of existing models to build a new model, for instance, try to introduce further parameters into the NSB model to allow for equilibrium frequencies where the equalities a=t and c=g may not be fulfilled (a model allowing for a strand bias), especially parameters that reflect further biological assumptions of the evolutionary process
- apply the NSB model and new model(s) to actual sequence data, for instance by using the B4 method that is readily available and is independant of the model used.

Annex:

I already suceeded in finding a formulation of the NSB model that contains the equilibrium frequencies as factors in the rate matrix:

$$\begin{array}{cccc} - & g1\mathrm{T} & (s-a)\mathrm{G} & a\,\mathrm{G} \\ g1\mathrm{T} & - & a\,\mathrm{G} & (s-a)\mathrm{G} \\ (s-b)\mathrm{T} & b\,\mathrm{T} & - & g2\mathrm{G} \\ b\,\mathrm{T} & (s-b)\mathrm{T} & g2\mathrm{G} & - \end{array} \quad .$$

Here, T and G are the equilibrium frequencies a=t, and c=g, respectively. Note that we have the parameters: $a, b, g1, g2, s$, and T represents the six‘th free parameter since 2T+2G=1 $\rightarrow$ G=1/2-T.

The fulfillment of the no-strand-bias condition and the equilibrium frequencies can easily be checked. A derivation of this formulation of the NSB model will be provided in the “rapport technique“.

my references:

- Blasidel (1985) A method of estimating from two aligned present-day DNA sequences their ancestral composition and subsequent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitutions at the same site.J Mol Evol 22:69-81
- Gojobori, Ishii, Nei (1982) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J Mol Evol 18:414-422
- Graur and Li (2000) Fundamentals of Molecular Evolution. Sinauer Associates; ISBN 0-87893-266-6
- Hasegawa, Kishino, Yano (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174
- Kimura (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120
- Kimura (1981) Estimation of evolutionary differences between homofogous nucleotide sequences. Proc Natl Acad Sci USA 78:454-458
- Lio and Goldman (1998) Models of Molecular Evolution and Phylogeny. Cold Spring Harbor Laboratory Press ISSN 1054-9803/98: 8:1233-1244
- Lobry (1995) Properties of a Gereral Model of DNA Evolution Under No-Strand-Bias Conditions. J Mol Evol 40:326-330; 41:680
- Rzhetsky and Nei (1995) Tests of Applicability of Several Substitution Models for DNA Sequence Data. Mol Biol Evol 12(1): 131-151
- Sueoka (1995) Intrastrand Parity Rules of DNA Base Composition and Usage Biases of Synonymous Codons. J Mol Evol 40:318-325; 42:323
- Tajima, Nei (1982) Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J Mol Evol 18:115-120
- Tajima, Nei (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269-285
- Takahata, Kimura (1981) A model of evolutionary base substitution and ist application with special reference to rapid change of pseudo-genes. Genetics 98:641-657
- Tamura (1992) Estimation of the number of nucleotide substitutions when there are stong transition-transversion and G+C content biases. Mol Biol Evol 9:678-687
- Tamura, Nei (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512-526
- Tavaré (1986) Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences Vol 17:57-86
- Yang (1994) Estimating the Pattern of Nucleotide Substitution. J Mol Evol 39:105-111
- Zharkikh (1994) Estimation of Evolutionary Distances Between Nucleotide Sequences. J Mol Evol 39: 315-329