Journal of Theoretical and Applied Information Technology
31 August 2012. Vol. 42 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EVOLUTIONARY DISTANCE MODEL BASED ON DIFFERENTIAL EQUATION AND MARKOV PROCESS

XIAOFENG WANG

College of Mathematical and Physical Science, Chongqing University of Science and Technology, Chongqing 401331, China

ABSTRACT

Evolutionary distance is a measure of evolutionary divergence time between two homologous sequences. It is fundamental for the study of molecular evolution. In this paper, we exploit one-parameter model and two-substitution model using the differential equation and the Markov process in stochastic process. Through them, we can acquire the evolutionary distance between two sequences as estimated number of changes that have occurred per site. Experimental results are evaluated through Monte Carlo experiments demonstrating the usefulness of one-parameter model and two-substitution model, and comparing two models, the two-substitution model may get better accuracy than one-parameter model.

Keywords: *Evolutionary Distance, Differential Equation, Two-substitution Model, Markov Process*

1. INTRODUCTION

Evolutionary distance is a measure of evolutionary divergence between two homologous sequences, which is fundamental for the study of molecular evolution and useful for phylogenetic reconstructions and the estimation of divergence times [1, 2, 3]. More precisely, evolutionary distance is the number of residue substitutions which have occurred between two sequences, as they diverged from their common ancestor. We know, the evolutionary change of DNA sequences occurs by nucleotide substitution, deletion, and insertion, which is measured in terms of the number of nucleotide substitutions per site between two homologous DNA sequences [11]. Estimation of evolutionary distances between protein and DNA sequences is important for constructing phylogenetic trees, acquiring species's divergence time and under-standing the mechanisms of evolution of genes, proteins, populations .Therefore, estimating the evolutionary distances between homologous sequences in terms of the number of base substitution is essential.

In probability theory and statistics, a Markov process is a time-varying random phenomenon for which a specific property. Using this analysis, we can generate a new sequence of random but related events, which will look similar to the original. So, the Markov process is useful to analyze dependent random events, that is, events whose likelihood depend on what happened last. The process of evolutionary in DNA sequences has been modeled as a Markov process. The Markov transition matrices are estimated by discrete time matrix methods. In 1962, Zuckerandl and Pauling [4] first suggested that the evolutionary distance between two protein (or DNA) sequences X and Y can be inferred from the observed divergence matrix from counts of the occurrences of amino acids. Therefore, we present the evolutionary distance of the DNA sequences by the Markov methods. A number of different Markov models of DNA sequence evolution have been proposed [5, 6, 7]. A continuous Markov model describes the probabilities of changing from a state to another state after some time. In the case of sequence evolution, the set of states are the DNA and the Markov model describes the probabilities of their substitution over time. This means, that all sites of a DNA sequence are treated independently from each other.

The differential equation [10] is a mathematical equation for an unknown function of one or several variables that relates the values of the function itself and its derivatives of various orders. For the analysis of evolutionary distance, we need to derive rates of change in the substitution as time change, which are deduced from differential equation. So it is better to acquire results combining the differential equation and the Markov process.

In this paper, we acquire one-parameter model and two-substitution model using the differential equation and the Markov process. The simulation results through Monte Carlo experiments indicate that the presented approaches perform well. Comparing two models, the two-substitution model may get better accuracy than one-parameter model.

Journal of Theoretical and Applied Information Technology
31 August 2012. Vol. 42 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The paper is organized as follows: Section 2 gives one-parameter model through differential equation and the Markov process. We present two-substitution models through differential equation and the Markov process in Section 3, experimental results show the efficacy of our models by Monte Carlo experiments in Section 4. Section 5 concludes.

2. ONE-PARAMETER MODEL

One-parameter model is the simpler model of DNA sequence evolution [8, 9]. We may assume that every base (i.e. the purines A, G and the pyrimidines C, T) has a constant probability per unit time T changing into each of the others bases. If we denote the probability $Q_i(t)$ at base i at time t and we get

$$Q(t) = (Q_A(t), Q_C(t), Q_G(t), Q_T(t)),$$

and then we acquire the following differential equation and the Markov process:

$$\frac{dQ}{dt} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} Q \qquad (1)$$

We can present the deduction process below in detail. We have developed the following distance measure, under the assumption that the substitution rate is the same between any pair of nucleotides.

We assume that the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G. Nucleotide substitution occur at any nucleotide site with equal frequency and at each site a nucleotide changes to one of the other nucleotides with a probability of α per unit time T. Therefore, the probability of change of a nucleotide to any of the three nucleotides is $\gamma = 3\alpha$, where γ is equal.

Let us now consider two nucleotide sequences, X and Y, which diverged from the common ancestral sequence t unit time ago.

Therefore, we may acquire the following difference equation.

$$Q(T+\Delta T) = (1-2\gamma)Q(T) + (2/3)\gamma(1-Q(T)) \qquad (2)$$

Then may be written as

$$Q(T+\Delta T) - Q(T) = (2\gamma/3) - (8\gamma/3)Q(T) \qquad (3)$$

Let us now use a continues time model and represent $Q(T+\Delta T) - Q(T)$ by dQ/dt, dropping the subscript t of $Q(T)$.

We then have the following differential equation:

$$dQ/dT = (2\gamma/3) - (8\gamma/3)Q \qquad (4)$$

Usually, we assume that we know the state of a site at time $t = 0$ and $Q_i(0) = 1$, So the solution of this equation with the initial condition

$$Q = 1 - (3/4)(1 - e^{-8rt/3}) \qquad (5)$$

Under the present model, the expected number of nucleotide substitutions per site for the two sequences is 2rt. Therefore, d is given:

$$d = -\frac{3}{4} In\left(1 - \frac{4}{3}p\right) \qquad (6)$$

where p is the proportion of sites that differ between the two sequences.

The variance of this distance is given by

$$V(d) = \frac{9p(1-p)}{(3-4p)^2 n} \qquad (7)$$

So we acquire the evolutionary distance between two sequences as estimated number of changes that have occurred per site.

3. TWO- SUBSTITUTION MODEL

In the above model, the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G [10]. Although the evolutionary distance between two protein (or DNA) sequences X and Y can be inferred from One-parameter model, which depends on assuming that all kinds of base substitutions were equally likely. Some examples were worked out using reported globins sequences to show that synonymous substitutions occur at much higher rates than amino acid-altering substitutions in evolution. Since there are 4 possibilities (A, T, C, and G) at each site, there are 16 combinations and three types when the homologous sites in two species are compared. These are listed in Table I.

We can get the following differential equation and the Markov process :

Journal of Theoretical and Applied Information Technology
31 August 2012. Vol. 42 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$\frac{dP}{dt}=\begin{pmatrix} -2\beta-\alpha & \beta & \alpha & \beta \\ \beta & -2\beta-\alpha & \beta & \alpha \\ \alpha & \beta & -2\beta-\alpha & \alpha \\ \beta & \alpha & \beta & -2\beta-\alpha \end{pmatrix} P \quad (8)$$

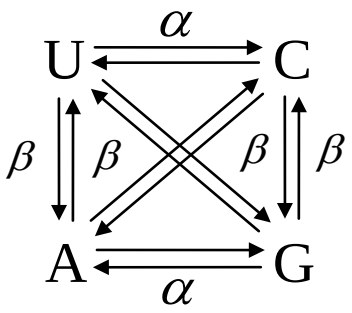

Figure 1: Scheme of evolutionary base substitutions and their rates per unit time.

Table I

Types of nucleotide base pairs occupied at homologous sites in two species. There are three types including same, difference I, difference II. Difference I includes four cases in which both are purines or both are pyrimidines. Difference II consists of eight cases in which one of the bases is a purine and the other is a pyrimidine.

	Nucleotide base pairs (Frequency)				
Same	UU	CC	AA	GG	SUM
	R_1	R_2	R_3	R_4	R
Difference I	UC	CU	AG	GA	SUM
	P_1	P_1	P_2	P_2	P
Difference II	UA	AU	UG	GU	SUM
	Q_1	Q_1	Q_2	Q_2	Q
	CA	AC	CG	GC	
	Q_3	Q_3	Q_4	Q_4	

We can derive the present differential equation for P and Q at time $T+\Delta T$ using Markov, and the deduction process below in detail .Let T be the time since divergence of the two species (measured in years) and α the rate of transition type substitutions per site per unit time(year) , 2β transversion type substitutions per site per unit time(year). The total rate of substitutions per site per unit time (year) is $k=\alpha+2\beta$. P is the probability of homologous sites showing a transition type substitution, Q is the probability of homologous sites showing a transversion type substitution, and ΔT stands for the length of a short time interval.

It is discussed with the example of UA to deduction process at time $T+\Delta T$ which is derived from pairs at time T . There are three different situations:

1) UA is derived from UA when U, A remain unchanged. We know that the probability of substitution per during ΔT is $(\alpha+2\beta)\Delta T$, while probability of no change is $[1-(\alpha+2\beta)\Delta T]^2$,since the order $(\Delta T)^2$ is small terms, we neglect one and get $[1-2(\alpha+\beta)\Delta T]P_1(T)$.

2) UA is derived from UU when U in the second position is replaced by A, and from AA when A in the first position is replaced by U. UU, AA occurs with frequencies $R_1(T)$, $R_3(T)$, respectively. UA at $T+\Delta T$ is $\alpha\Delta T[R_1(T)+R_3(T)]$.

3) UA is derived from UC, UG, CA and GA, while each occurs with respective frequencies $Q_2(T)$, $Q_3(T)$, $P_1(T)$, $P_4(T)$. The change to UA at $T+\Delta T$ is $\alpha\Delta T[Q_2(T)+Q_3(T)]+\beta\Delta T[P_1(T)+P_4(T)]$.

Combining all above results, we get

$$(T+\Delta T)Q_1(T)=[1-(2\alpha+4\beta)\Delta T]Q_1(T)+\beta\Delta T[R_1(T)+R_3(T)]+\alpha\Delta T[Q_2(T)+Q_3(T)]+\beta\Delta T[P_1(T)+P_4(T)] \quad (9)$$

Similarly, we can get for the base pair UG, CA, CG in difference II, for example UG below:

$$(T+\Delta T)Q_2(T)=[1-(2\alpha+4\beta)\Delta T]Q_2(T)+\beta\Delta T[R_1(T)+R_4(T)]+\alpha\Delta T[Q_2(T)+Q_4(T)]+\beta\Delta T[P_1(T)+P_3(T)] \quad (10)$$

Summing all equations for UA, UG, CA and CG, and getting

$$\Delta Q(T)/\Delta T=4\beta-8\beta Q(T) \quad (11)$$

Carrying out a similar series of calculations in difference I, we obtain

$$\Delta P(T)/\Delta T=2\alpha-4(\alpha+\beta)P(T)-2(\alpha-\beta)Q(T) \quad (12)$$

The solution at initial condition:

$$P(0)=Q(0)=0 \quad (13)$$

Calculate the differential equations above, and get

Journal of Theoretical and Applied Information Technology
31 August 2012. Vol. 42 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$P(T)=\frac{1}{4}-\frac{1}{2}e^{-4(\alpha+\beta)T}+\frac{1}{4}e^{-8\beta T} \quad (14)$$

$$Q(T)=\frac{1}{2}-\frac{1}{2}e^{-8\beta T} \quad (15)$$

Therefore, the expected number of nucleotide substitutions per site between X and Y is given by

$$d=\alpha+2\beta \quad (16)$$

The total number of substitutions per site which involve two branches each with length T is

$$D=2Td=-\frac{1}{2}\log_e\left\{(1-2P-Q)\sqrt{1-2Q}\right\} \quad (17)$$

and the variance of d is given by

$$V(d)=\frac{1}{n}\left[c_1^2P+c_2^2Q-(c_1P+c_2Q)^2\right] \quad (18)$$

Where

$$c_1=\frac{1}{1-2P-Q} \quad (19)$$

And

$$c_2=\frac{1}{2}\left(\frac{1}{1-2P-Q}+\frac{1}{1-2Q}\right) \quad (20)$$

4. EXPERIMENTS

In this section, we perform Monte Carlo experiments to check our one-parameter model and two-substitution model through MATLAB.

First, we acquire a random nucleotide sequence, including A, G, C, T, as a common ancestor using Monte Carlo algorithm. In simulation experiment, we assign values of substitution rates, respectively. We provide all of the parameter settings used in our algorithm in Table 2. For one-parameter model, we assign values of substitution rate $\gamma=0.008$ for every gene every times. For two-substitution model, we assign values of substitution rate $\alpha=0.03$, $\beta=0.02$ for every gene every times. The experiments are continued until one substitution per site has occurred on the average from time T =0.1 to 1.5.

And then we compare the two sequences and compute actual number of nucleotide substitutions. The total number of nucleotide substitutions, d using equation of one-parameter model and two-substitution model, was monitored by summing the actual numbers of substitutions observed until a given time T . The results are illustrated through Figure 2.

The simulation experiments show that above methods is useful, especially two-substitution model. From Fig 1, we see that equation provide good estimates of the actual evolutionary distance. For d <1, the estimated evolutionary distance gives a relatively good underestimate for d , while For 1< d <1.5 the estimated evolutionary distance gives an underestimate for d . That is because there are a great many repeated substitutions happening as time T go on. Comparing one-parameter model and two-substitution model, we see that two-substitution model is more accurate than one-parameter model from Figure 1, because two-substitution model consider transversion and transition in more detail.

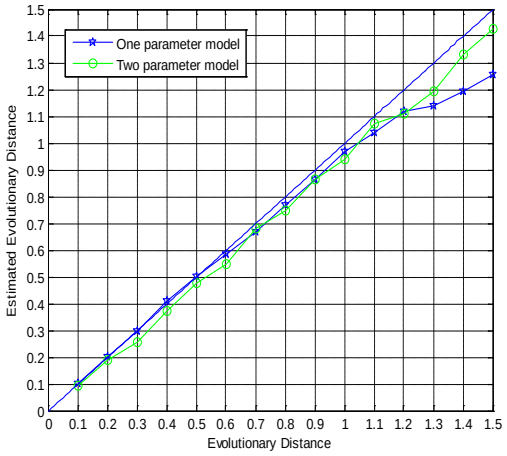

Figure 2: Relation between the actual evolutionary distance and the estimated evolutionary distance based on one-parameter model and two-substitution model. The solid lines represent that both above is equal. One marked by open circles represents one-parameter model while by asterisk represents two-substitution model.

Table II
Parameter settings for one-parameter model and two-substitution model in our algorithm

γ	α	β
0.008	0.03	0.02

.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. DISCUSSION

Evolutionary distance is very important to the protein evolution of the species, which may contribute to construct phylogenetic trees, estimate species's divergence time. to acquire evolutionary distance of the gene, we have exploited differential equation and the markov process in stochastic process. we may be acquainted with evolutionary distance from the viewpoint of a new angle. through them, we have got one-parameter model and two-substitution model, which have their own advantages and disadvantages. we can estimate evolutionary distance of the gene using nice property of differential equation and the markov process, which makes the two models applicable to a wider condition. experimental results have been evaluated through monte carlo demonstrating the usefulness of one-parameter model and two substitution model. comparing two models, the two-substitution model have got better accuracy than one-parameter model.

ACKNOWLEDGEMENTS

This work is supported by the Science and Technology Foundation of the Education Department of Chongqing (KJ121404), Innovation team of Chongqing University of Science & Technology (1809013) and Research Foundation of Chongqing University of Science & Technology (CK2011Z15).

REFRENCES:

[1] Doğan, T, "Evolutionary relationships between gene sequences via nonlinear embedding", *Proceedings of 15th National Biomedical Engineering Meeting*, Vol. 1, 2010, pp.1-4.

[2] Jukes T H, Cantor C R, "Evolution of protein molecules", In Mammalian Protein Metabolism, 21-132,1969.

[3] Kimura, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences", *J Mol Evol.*, Vol.16, No.2, 1980, pp.111-120.

[4] Zuckerandl E and Pauling L, "Molecular disease, evolution, and generic heterogeneity", Horizons in biochemistry, 1962, pp. 189-225.

[5] Motoo Kimura, "Estimation of evolutionary distances between homologous nucleotide sequences", *Proc. NatL. Acad.*, Vol.78, No 1, 1981, pp.454-458.

[6] Naoyuki T, Kimura M, "A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes", *Genetics*, Vol. 98, No 3, 1981, pp. 641-657.

[7] Burczynski, T. Gliwice Kus, W. Majchrzak, E. Orantek, P., Dziewonski, M, "Evolutionary computation in identification of a tumor", *Proceedings of the 2002 Congress on CEC '02*, Vol. 2, 2002, pp. 1250-1254.

[8] Masatoshi Nei, Sudhir Kumar, "Molecular Evolution and Phylogenetics", Oxford University Press, 2000 .

[9] Carl D. Soulsbury, Graziella Iossa and Keith J. Edwards ,"The influence of evolutionary distance between cross-species microsatellites and primer base-pair composition on allelic dropout rates", *Conserv Genet*, Vol.98, No 10, 2009, pp. 797-802.

[10] Yu Lin, et al., "Estimating true evolutionary distances under rearrangements, duplications, and losses", *BMC Bioinformatics*, Vol. 11, No. 54, 2010, pp. 112-121.

[11] Ashlock, D, Cottenie, K., Carson, L., Bryden, K.M. , Corns, S. "An Evolutionary Algorithm for the Selection of Geographically Informative Species", *Proceedings of the Computational Intelligence and Bioinformatics and Computational Biology*, Vol.1, 2006, pp. 1-7.