Available online at http://scik.org
J. Math. Comput. Sci. 5 (2015), No. 5, 716-727
ISSN: 1927-5307

SIMULATION OF NUCLEOTIDE SUBSTITUTION RATES IN A DNA USING JC69 MODEL

D.T. ACHAKU*, P.V. AYOO
Department of Mathematics, Federal University Lafia, P.M.B. 146 Lafia, Nigeria

Abstract: Exploring the exponential of a matrix, we derived the rate matrix by extending the Poisson distribution to Differential equation. The solution of which gave rise to the continuous time Markov transition probability matrix $P(t)$. From who estimated the MLE of JC69 substitution model as approximately equal to 0.2 we generated the probability transition matrix $P(t)$ with the four states as Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Using c^{++} and Excel with varying values for rate with a constant time of 1, the simulation showed that at the point when rate tends to infinity, over which a nucleotide sequence had been allowed to evolve, the proportion of nucleotides of each type A, G, C and T will reach $\frac{1}{4}$ for each. And this limiting distribution is called the stationary distribution and is maintained.

Keywords: Poisson distribution; Markov chain; nucleotide substitution; probability transition matrix.

2010 AMS Subject Classification: 60J27; 92D20

1. Introduction:

The information stored in a DNA and its ability to preserve it depends on its rate of nucleotide substitution. Tandem nucleotide repeat sequences are abundant in the human genome (Tautz and Renz, 1994; Smithies and Powers, 1986). Many tandem repeats are known to be polymorphic. Such repeat sequences can be classified as either simple, (one tandem repeating triplet unit representing at least $\frac{2}{3}$ of total length) or cryptic (Jacobson et. al, 1993). Simple and cryptic di-nucleotide repeats of alternating purines and pyrimidines of 26 or more base pairs in length $[\mathrm{RY}26^{+}]$ are highly enriched in human, mouse and yeast genomes (Sankar et. al., 1991). In

*Corresponding author

Received May 15, 2014

humans, both simple (Tautz and Renz, 1984; Weber, 1990) and cryptic (Sankar et. al., 1991) are polymorphic and occasionally hypervariable (Jacobson et. al, 1993).

The two RY (26^{+}) in the human factor IX gene are cryptic repeats. These repeats are polymorphic and certain alleles show racial specificity (Sankar et. al, 1991; Jacobson et. al, 1993; Sommer et. al, 1994). The RY(26^{+}) in intron 1 of the factor IX gene can contain as many as 216bp (probability of occurrence at random) of a melodic sequence in which there is no more than six consecutive repeats of any nucleotide (Jacobson et. al, 1993) seven polymorphic alleles have been found in one segment of this RY(26^{+}) (Sankar et. al, 1991) by screening 1800 human chromosomes. These alleles are of the form $A_{0-4} \quad B_1 \, and \, A_{1-3} \, B_2$ in which A = (GT)(AC)$_3$(AT)$_3$(GT)(AT)$_4$ and B = A with an additional $3'$ AT di-nucleotide. The sequence shows a novel type of hyper-variability characterized by many Tandem repeats of the form $(GT)_n(AC)_o(AT)_P(GT)_q(AT)_s$ where n, o, p, q and s are integers that range from 1 to 4. The Sequence suggests that the location of two RY (26^{+}) can be preserved during evolution, while the precise sequence varies substantially. Purine, purine, pyrimidine (RRY) repeats are also frequent and often polymorphic (Riggers et. al, 1992; Gostout et. al, 1993).

Instability of simple RRY has been recently implicated as the major source of mutations in multiple diseases (Yu et.al, 1991; Oberle et. al, 1991; Verkerk et. al, 1991). Nucleotide substitution has been found to follow Markov processes over time. In this work, we derived the Jukes and Cantor model and used C^{++} to simulate data using formulae which is capable of simulating evolutions of a nucleotide sequence over a given time.

Over time, nucleotides within a sequence can evolve through substitution. This process can cause a nucleotide $(T, C, A$ and $G)$ to change into another nucleotide. On average, this form of mutation only occurs once or twice every millions of years. However, in assessing the evolution of species over hundreds of millions of years, models are useful in evaluating how one sequence of nucleotides may have evolved from another.

This model is a phenomenological description of the evolution of DNA as a string of four discrete states. The Markov chain model does not explicitly depict neither the mechanism of mutation nor the action of natural selection, rather it describes the relative rates of different changes.

2.0 Continuous- time Markov chains

For stochastic processes refer (Choji, D.N. and Achaku, D.T., 2007; Choji and Oduwole, 1998; Korve et. al, 2006; Taylor and Karlin, 1994). But, a continuous- time Markov chain has the usual transition matrix which is in addition parameterized by time t. Specifically, if e_1,e_2, e_3, e_4 are the states, then the transition matrix

$$P(t) \ = \ P_{ij}\,(t).$$

From Jukes and Cantor (1969), Tan Wai-Yuan (2002) this substitution process in DNA sequence has a corresponding matrix as,

$$P(t) = \begin{pmatrix} P_{AA}(t) & P_{GA}(t) & P_{CA}(t) & P_{TA}(t) \\ P_{AG}(t) & P_{GG}(t) & P_{CG}(t) & P_{TG}(t) \\ P_{AG}(t) & P_{GC}(t) & P_{CC}(t) & P_{TC}(t) \\ P_{AT}(t) & P_{GT}(t) & P_{CT}(t) & P_{TT}(t) \end{pmatrix}$$

Where the top left and bottom right 2× 2 blocks correspond to transition probabilities and top – right and bottom-left 2× 2 blocks correspond to transversion probabilities.

Theorem 2.1: *A Continuous time Markov chain satisfies* $P(t \ + \ r) \ = \ P(t) \ P(r)$.

Consider a DNA sequence of fixed length m evolving in time by base replacement. Assume that the processes followed by the m sites are Markovian independent, identically distributed and constant in time, so for a fixed site let,

P(t) = $[P_A(t),\ P_G(t),\ P_C(t),\ P_T(t)]^T$

Be the column vector probabilities of states A, G, C and T at time t. With S = {A, G, C, T}be the state space. For any two distinct states $, G \ \in \ S$, let μ_{AG} be the transition rate from state A to state G. Similarly, for any A, let

$$\mu_A = \sum_{G \neq A} \mu_{AG}$$

The changes in the probability distribution $P_A(t)$ for small increments of time Δt are given by,

$P_A(t + \ \Delta t)$ = $P_A(t) - P_A(t)\mu_A \Delta t$ ………… (1)

In other words the frequency of A'^S at time $t + \ \Delta t$ is equal to the frequency at time t minus the frequency of the lost A's plus the frequency of the newly created A'^S. And similarly for $P_G(t)$, $P_C(t)$ and $P_T(t)$. So, write (1) compactly as or alternately,

$$P'(t) = QP(t)$$

With,

‘

$$Q = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_A & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_A & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_A \end{pmatrix}$$

2.1 Ergodicity

If all states A, G $\epsilon\ S$ communicate, then the Markov chain has a stationary distribution $\prod S$ = $\{\prod A\ , A \in S\}$, where each $\prod A$ is the proportion of time spent in state A after the Markov chain has run for infinite time and this probability does not depend upon the initial state of the process. This type of chain is called ergodic. In DNA evolution, under the assumption of a common process for each site, the stationary frequencies are: Π_A, Π_G, Π_C, Π_T corresponds to equilibrium base compositions.

Definition 2.1: A Markov process is stationary if its current distribution is the stationary distribution, i.e. $P(t)$ = Π. From the differential equation above:

$$\prod{}'(t) = Q\Pi = 0.$$

Definition 2.2: *Time Reversibility*

A stationary Markov process is time reversible if the amount of change from state A to G is equal to the amount of change from G to A, though frequency may be different. This means that;

$$\Pi_A\ \mu_{AG} = \Pi_G\ \mu_{GA}$$

Though not all stationary processes are reversible, however, almost all DNA evolution models assume time reversibility. Hence we can write this as

$$S_{AG} = \frac{\mu_{AG}}{\Pi_G} \text{ then } S_{AG} = S_{GA}$$

This symmetric term S_{AG} is called the exchangeability between A and G.

3.0 Derivation of Juke's-Cantor Model

Assumption: It assumes equal frequencies (are Π_A= Π_G= Π_C= $\Pi_T = \frac{1}{4}$) and equal mutation rates. The only parameter is ⋋, the substitution rate.

At the simplest level, the proportion of different nucleotides p can be used to measure the evolutionary divergence between two aligned sequences.

$$P = \frac{n_d}{n}$$

Where n is the total number of nucleotides in the sequences and n_d is the number of different nucleotides for the pair. Suppose the distribution of the number of substitutions s is a poisson random variable with mean ⋋t. The rate of substitutions (relative to the unit of time) at a given site is ⋋. The probability of $s > 0$ at a site in a time period t is,

$$P_r(\mathrm{s}) = \frac{e^{\leftthreetimes \mathrm{t.}} \leftthreetimes \mathrm{t.}^{s}}{s!}$$

If the mean number of substitutions during t units of time is ⋋t, then the probability of no substitutions occurring at a site is

$$P_r(\mathrm{s} = 0) = e^{-\leftthreetimes \mathrm{t}}$$

And probability of at least one substitution is

$$P_r(\mathrm{s} = 0) = 1 - e^{-\leftthreetimes \mathrm{t}}$$

For small t, these probabilities can be approximated as

$$p_r(s = 0) \approx 1 - \lambda t.$$

And

$$p_r(s \neq 0) \approx \lambda t$$

These probabilities can be seen as the infinitesimal probabilities relating to a Markov process. Thus for a DNA sequence, this process is a 4-state chain. Let p_{ij} be the transition probability that the next state (nucleotide) is j given that the current state is i,

$p_{ij} = p_r${Next state s_j/ current state s_i}

Let $p = p_{ij}$ denote the matrix of transition probabilities for the Markov process, then this property is true,

$$p(t + h) = p(t)p(h) \qquad \ldots\ldots\ldots\ldots\ldots.. \quad (2)$$

Extending the results from the Poisson model to the 4-state nucleotide case, for h small, the probabilities are approximated to

$$p(h) \approx I + Qh \qquad \ldots\ldots\ldots\ldots\ldots.. \quad (3)$$

Where Q is the rate of substitution matrix. Substituting () in () and taking limit as $h \to 0, p$ solves

$$p' = PQ$$

With initial condition, $p(0) = I$, the solution of the first order differential equation is

$$\frac{dp}{dt} = PQ$$

$$\int \frac{dp}{p} = \int Qdt$$

$$\ln p = Qt + c$$

$$p = \frac{Q}{\lambda} e^{-\lambda t}$$

And from (1)

$$p = \frac{1}{4}\left(1 - \frac{Q}{\lambda} e^{-\lambda t}\right)$$

Hence the chance of a nucleotide i changing to a nucleotide j in time t is

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

And since there are 12 ways this can occur, the chance of a nucleotide staying the same is

$$p_{ij} = \frac{1}{4} + \frac{3}{4} e^{-4\lambda t}$$

With the rate matrix as

$$Q = \begin{pmatrix} * & \frac{\lambda}{4} & \frac{\lambda}{4} & \frac{\lambda}{4} \\ \frac{\lambda}{4} & * & \frac{\lambda}{4} & \frac{\lambda}{4} \\ \frac{\lambda}{4} & \frac{\lambda}{4} & * & \frac{\lambda}{4} \\ \frac{\lambda}{4} & \frac{\lambda}{4} & \frac{\lambda}{4} & * \end{pmatrix}$$

And transition probability as

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4} e^{-4\lambda t} & \frac{1}{4} - \frac{1}{4} e^{-\lambda t} & \frac{1}{4} - \frac{1}{4} e^{-\lambda t} & \frac{1}{4} - \frac{1}{4} e^{-\lambda t} \\ \frac{1}{4} - \frac{1}{4} e^{-4\lambda t} & \frac{1}{4} + \frac{3}{4} e^{-\lambda t} & \frac{1}{4} - \frac{1}{4} e^{-\lambda t} & \frac{1}{4} - \frac{1}{4} e^{-\lambda t} \\ \frac{1}{4} - \frac{1}{4} e^{-\lambda t} & \frac{1}{4} - \frac{1}{4} e^{-\lambda t} & \frac{1}{4} + \frac{3}{4} e^{-\lambda t} & \frac{1}{4} - \frac{1}{4} e^{-\lambda t} \\ \frac{1}{4} - \frac{1}{4} e^{-\lambda t} & \frac{1}{4} - \frac{1}{4} e^{-\lambda t} & \frac{1}{4} - \frac{1}{4} e^{-\lambda t} & \frac{1}{4} + \frac{3}{4} e^{-\lambda t} \end{pmatrix}$$

3.1 Maximum Likelihood Estimates (MLE) - JC69 model

Maximum likelihood estimates are used to estimate parameter values for a statistical model when applying that model to a dataset. In the case of nucleotide substitutions, the statistical models fitted to data are the models of nucleotide substitution and the parameter estimated is the value for rate and time. And of course they cannot be distinguished from one another; the single value λt can be produced by a combination of values of alpha or time. Likelihood methods for phylogenies were first introduced by Edwards and Cavalli-sforza (1964) for gene frequency data. Neyman (1971) applied likelihood to molecular sequences and this work was extended by

Kashyap and Subas (1974). Felsenstein (1973, 1981) brought the maximum likelihood framework to nucleotide-based phylogenetic inference. The following is the probability mass function of the binomial distribution:

$\binom{n}{k} p^k (1-p)^{nk}$. Where $n -$ total length of a sequence and $k -$ the number of nucleotides which differ between each sequence.

Thus using the data set of two sequences of nucleotides of equal length as $n = 100$ and $k = 40$. Thus approximating the Binomial distribution to a Poisson with the probability mass function equal 1, we have

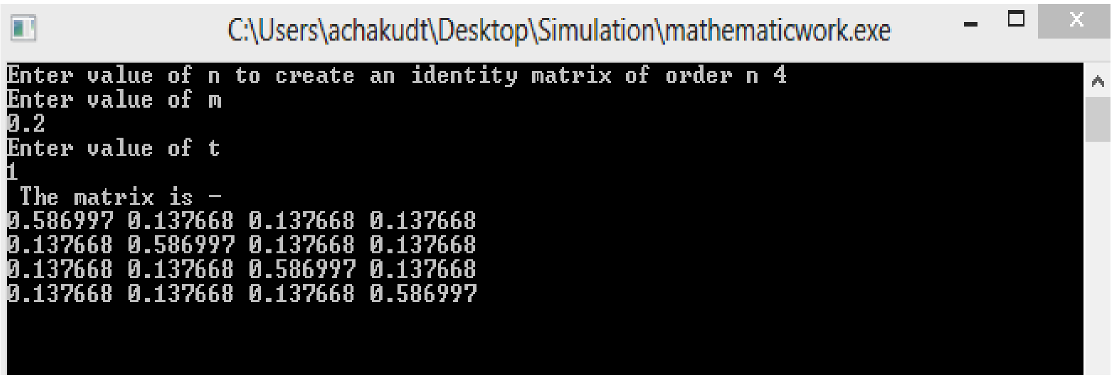
$$1 = pow\left(\frac{1}{4} + \frac{3}{4} * \exp(-4 * m), 60\right) * pow\left(\frac{1}{4} - \frac{1}{4} * \exp(-4 * m), 40\right)$$

Where, $m = \lambda t$. Goldman and Whelan (2001) estimated the maximum likelihood estimate for Jukes Cantor model as $0.19 \cong 0.2$

3.2 Simulation of nucleotide Sequence

The discussed model and other models of nucleotide substitution, all allow for the generation of probabilities that determine how a nucleotide sequence will or have evolved based on likelihood. Thus for the JC69 model, we can say that this probabilities are equal.

The time intervals in which mutations will occur are taken simply as $t = 0$ to $t = 1$ as used often (Timex). Before the mutation, a nucleotide sequence of length 100 was generated (genseq) with a rate estimated above as $\simeq 0.2$, that is parameter, to generate the matrix $P(t)$. A code written in C^{++} emulates the matrix $P(t)$.



```
Enter value of n to create an identity matrix of order n 4
Enter value of m
0.2
Enter value of t
1
 The matrix is -
0.586997 0.137668 0.137668 0.137668
0.137668 0.586997 0.137668 0.137668
0.137668 0.137668 0.586997 0.137668
0.137668 0.137668 0.137668 0.586997
```

3.3 Illustration

The following is a sequence of nucleotides before and after mutation

Before genseq (10).

	G	G	A	T	C	C	G	C	A	A
	↓	↓		↓		↓			↓	
After	G	A	G	T	A	C	C	C	A	G

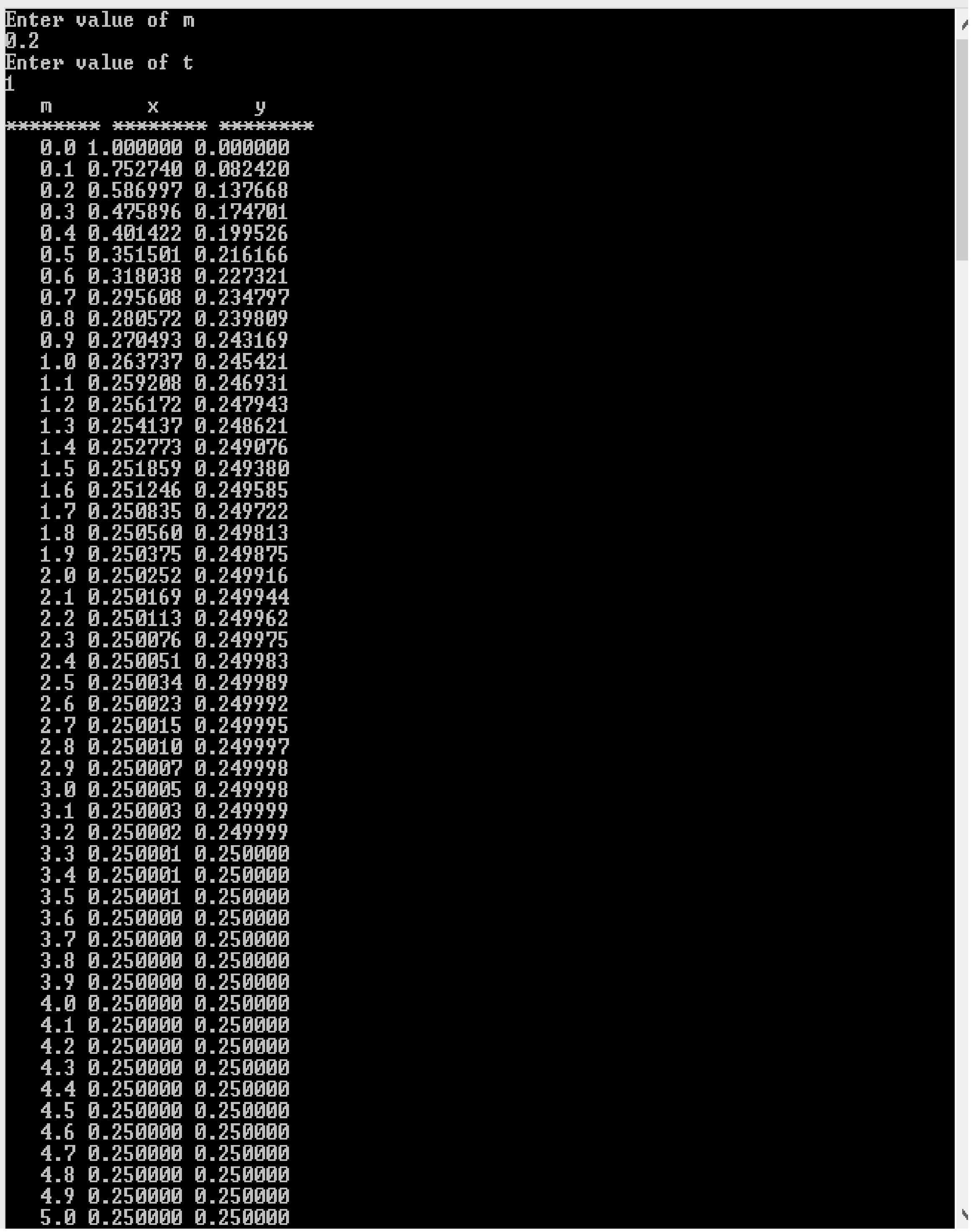
Although 5 differences are visible from the initial sequence to the sequence after mutation, 7 actual mutations had occurred with two of the mutations acting on the same starting nucleotide, the 8th, with the second mutation returning the 8th nucleotide back to its starting state that is nucleotide C.

From the formulae used to calculate the transition probabilities namely:

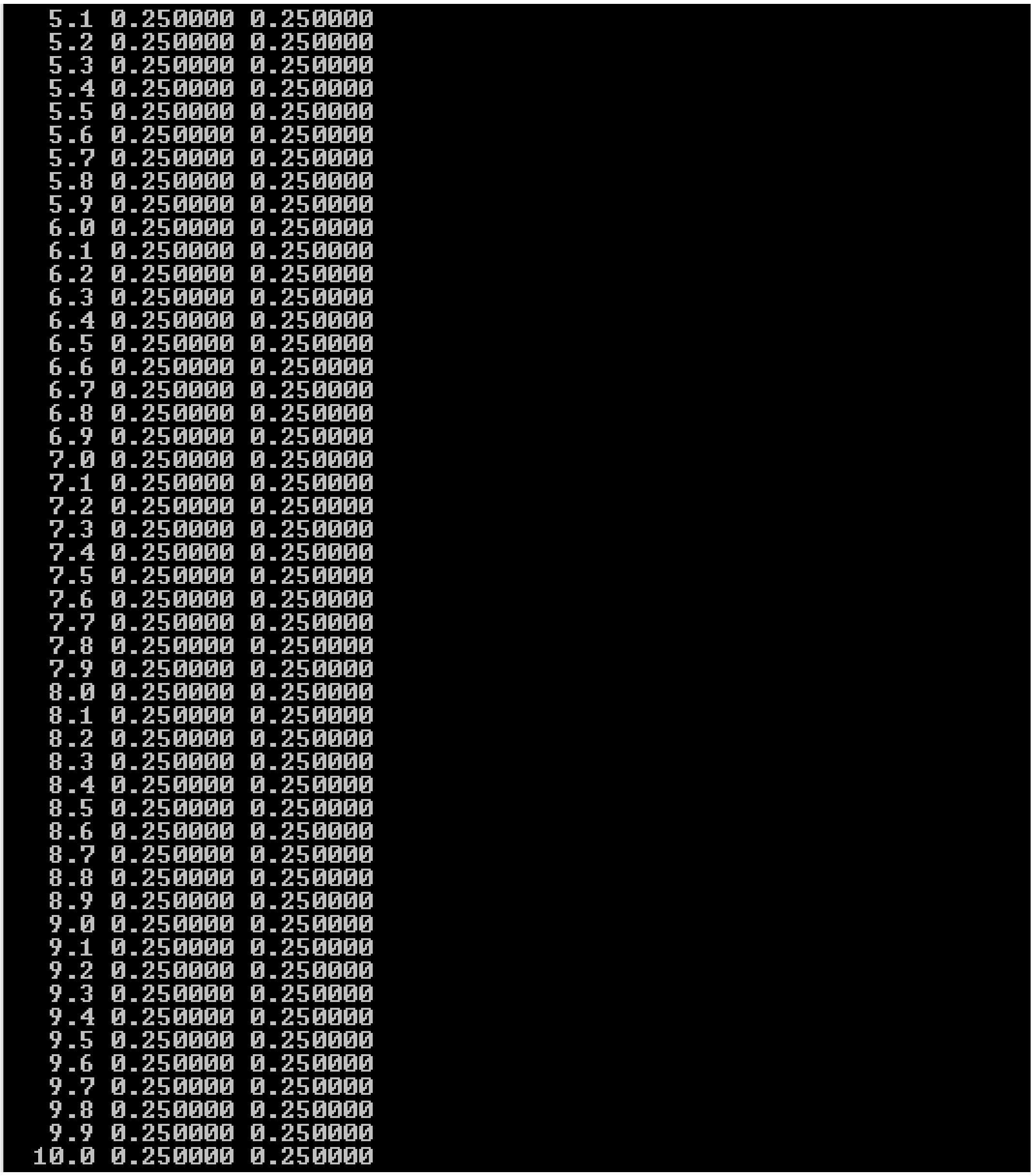
$$x = \frac{1}{4} + \frac{3}{4} * exp^{-4\lambda t}$$

$$y = \frac{1}{4} - \frac{1}{4} * \exp(-4 * \lambda * t).$$

We noticed that the exponential of this negative value tends to zero (0) as the negative value tends to infinity. This implies that x and y above tends to $\frac{1}{4}$ for each nucleotide substitution. For $t = 1$ and $m = 0, 0.1, \ldots$ as can be seen in the simulation below.



```
Enter value of m
0.2
Enter value of t
1
   m         x         y
******** ******** ********
   0.0 1.000000 0.000000
   0.1 0.752740 0.082420
   0.2 0.586997 0.137668
   0.3 0.475896 0.174701
   0.4 0.401422 0.199526
   0.5 0.351501 0.216166
   0.6 0.318038 0.227321
   0.7 0.295608 0.234797
   0.8 0.280572 0.239809
   0.9 0.270493 0.243169
   1.0 0.263737 0.245421
   1.1 0.259208 0.246931
   1.2 0.256172 0.247943
   1.3 0.254137 0.248621
   1.4 0.252773 0.249076
   1.5 0.251859 0.249380
   1.6 0.251246 0.249585
   1.7 0.250835 0.249722
   1.8 0.250560 0.249813
   1.9 0.250375 0.249875
   2.0 0.250252 0.249916
   2.1 0.250169 0.249944
   2.2 0.250113 0.249962
   2.3 0.250076 0.249975
   2.4 0.250051 0.249983
   2.5 0.250034 0.249989
   2.6 0.250023 0.249992
   2.7 0.250015 0.249995
   2.8 0.250010 0.249997
   2.9 0.250007 0.249998
   3.0 0.250005 0.249998
   3.1 0.250003 0.249999
   3.2 0.250002 0.249999
   3.3 0.250001 0.250000
   3.4 0.250001 0.250000
   3.5 0.250001 0.250000
   3.6 0.250000 0.250000
   3.7 0.250000 0.250000
   3.8 0.250000 0.250000
   3.9 0.250000 0.250000
   4.0 0.250000 0.250000
   4.1 0.250000 0.250000
   4.2 0.250000 0.250000
   4.3 0.250000 0.250000
   4.4 0.250000 0.250000
   4.5 0.250000 0.250000
   4.6 0.250000 0.250000
   4.7 0.250000 0.250000
   4.8 0.250000 0.250000
   4.9 0.250000 0.250000
   5.0 0.250000 0.250000
```



```
   5.1 0.250000 0.250000
   5.2 0.250000 0.250000
   5.3 0.250000 0.250000
   5.4 0.250000 0.250000
   5.5 0.250000 0.250000
   5.6 0.250000 0.250000
   5.7 0.250000 0.250000
   5.8 0.250000 0.250000
   5.9 0.250000 0.250000
   6.0 0.250000 0.250000
   6.1 0.250000 0.250000
   6.2 0.250000 0.250000
   6.3 0.250000 0.250000
   6.4 0.250000 0.250000
   6.5 0.250000 0.250000
   6.6 0.250000 0.250000
   6.7 0.250000 0.250000
   6.8 0.250000 0.250000
   6.9 0.250000 0.250000
   7.0 0.250000 0.250000
   7.1 0.250000 0.250000
   7.2 0.250000 0.250000
   7.3 0.250000 0.250000
   7.4 0.250000 0.250000
   7.5 0.250000 0.250000
   7.6 0.250000 0.250000
   7.7 0.250000 0.250000
   7.8 0.250000 0.250000
   7.9 0.250000 0.250000
   8.0 0.250000 0.250000
   8.1 0.250000 0.250000
   8.2 0.250000 0.250000
   8.3 0.250000 0.250000
   8.4 0.250000 0.250000
   8.5 0.250000 0.250000
   8.6 0.250000 0.250000
   8.7 0.250000 0.250000
   8.8 0.250000 0.250000
   8.9 0.250000 0.250000
   9.0 0.250000 0.250000
   9.1 0.250000 0.250000
   9.2 0.250000 0.250000
   9.3 0.250000 0.250000
   9.4 0.250000 0.250000
   9.5 0.250000 0.250000
   9.6 0.250000 0.250000
   9.7 0.250000 0.250000
   9.8 0.250000 0.250000
   9.9 0.250000 0.250000
  10.0 0.250000 0.250000
```

Using the values in the table above, we draw the graph for the rates of substitution against m.

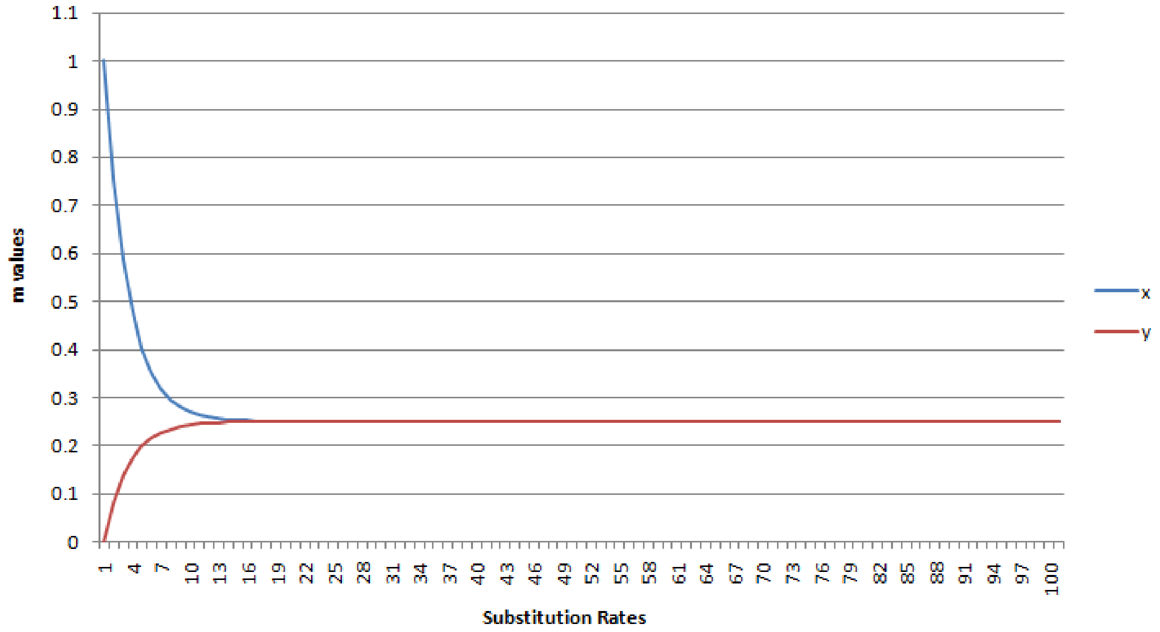


Figure 3.1: Graph showing Substitution Rates against Maximum Likelihood Estimate

4. Conclusion

This work is not an experiment from the laboratory but rather the emulation of JC69 nucleotide substitution model onto the C^{++} and excel software so as to use this in a more practical situation. Therefore, our conclusion is that, the code that we used to emulate this statistical model has been successful and maybe applied to practical dataset. That is, by assessing the same length, the number of differences may be recorded and then used to estimate a time using the maximum likelihood method.

Conflict of Interests

The authors declare that there is no conflict of interests.

REFERENCES

[1] D.N. Choji, and D.T. Achaku, Application of Markov chain Model to Partial Colour Blindness, Journal Science of researchers. 4(1 and 2), (2007).

[2] D. D. Bigner, et al, Non-random patterns of simple and cryptic repeats in coding and non-coding

sequences, Genomics. 26(3) (1995), 510 – 520.

[3] A. Edwards, and L. Cavalli – Sforza, Reconstruction of evolutionary trees. Systematic association Publication. 6 (1964).

[4] D.P. Jacobson. et al, characterization of the patterns of polymorphism in a cryptic repeat, reveals a novel type of hypervariable sequence; American Journal of Genetics. 53 (1993), 555 – 561.

[5] J. Felsenstein, Evolutionary trees from DNA sequences; A maximum likelihood approach. Journal of molecular evolution. 16(2) (1981), 368 – 376.

[6] K. N. Korve, et al, Markov chain properties of Daily rainfall occurrence in Northern Nigeria, The Journal of tropical Geography. 1(1) (2006), 27-34.

[7] T. Jukes, and Cantor, Evolution of protein molecules: mammalian protein metabolism. Academic Press, New York. (1969).

[8] W. Li, Molecular Evolution. Sinauer Associates Massachusetts. (1997).

[9] I. Oberle, et al, Inability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. Science. 252 (1991), 1097 – 1102.

[10] G. Sarkar, et al, Segments containing alternating purines and pyrimidines dinucleotide; Patterns of polymorphism in humans. Nucleic acid Res. 19 (1991), 631 – 636.

[11] O. Smithes, and P.A. Powers, Gene conversions and their relation to homologous pairing. Philes Trans, Society London B 312 (1986), 291 – 302.

[12] T.P. Speed, et al, Molecular Evolution, substitution models and phylogenes. (2000).

[13] T. Wai-Yuan, Application of Markov chain to Cancers, AIDS and other Biomedical. Scientific World Press, New Jersey. 4 (2002).

[14] H.M.Taylor, and S. Kelvin, An introduction to Stochastic modeling, Academic Press, San Diego. (1994).

[15] D. Tautz, and M. Benz, Simple Sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic acids res. 12 (1994), 4129 – 4138.

[16] A.J.M.H. Verkerk, et al, Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65 (1991), 905 – 914.

[17] J. L. Weber, et al, Characterization of a cosmid library from flow-sorted chromosomes. Chromosome Research. 2 (1994), 201 – 207.

[18] Y. Xua et al, Human Y- chromosome Base-Substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Current Biology. 19 (2009), 1453 – 1457.

[19] S. Yu, et al, Fragile X syndrome characterized by an unstable region of DNA. Science Journal. 252 (1991), 1179 -1181.

[20] N. Goldman and S. Whelan, A general empirical model of protein evolution derived from multiple protein families using a Maximum likelihood approach, Molecular Biology Evolution, 18(5), (2001), 691 – 699.