An Analysis of Continuous Time Markov Chains using Generator Matrices

G. Padma and Dr. C. Vijayalakshmi
Department of Mathematics
Sathyabama University, Chennai

ABSTRACT
This paper mainly analyzes the applications of the Generator matrices in a Continuous Time Markov Chain (CTMC). Hidden Markov models [HMMs] together with related probabilistic models such as Stochastic Context-Free Grammars [SCFGs] are the basis of many algorithms for the analysis of biological sequences. Combined with the continuous-time Markov chain theory of likelihood based phylogeny, stochastic grammar approaches are finding broad application in comparative sequence analysis, in particular the annotation of multiple alignments, simultaneous alignment. It was originally used to annotate individual sequences, then in later stages stochastic grammars were soon also combined with phylogenetic models to annotate the alignments. Thus, trees have been combined with HMMs to predict genes and conserved regions in DNA sequences, secondary structures and transmembrane topologies in protein sequences and base pairing structures in RNA sequences. The importance of Generator matrix is analysed in deriving the various properties of continuous time Markov chins with examples from the phylogenetic tree.

Keywords
Generator matrix, Continuous Time Markov Chains (CTMC), Embedded Markov Chain (EMC), Transition probability matrix, Stationary probabilities, Ergodicity, Jump process.

1. INTRODUCTION
A Markov process can have three important properties such as homogeneity, stationary and reversibility. A homogeneous process has an equilibrium distribution which is the limiting distribution when time approaches infinity. When a continuous time Markov chain is time homogenous then it satisfies the condition $P_{ij}(t, s) = P_{ij}(s)$ for all $t \geq 0$.

Since the set of states is discrete and the time parameter is continuous it is clearly not possible for the sample paths $X_t(w)$ to be continuous functions of 't'. At random times
$T_0 = 0$, T_1, T_2 … are called jump times (or transitions times) the process will change to a new state and the sequence of states constituting a discrete time process Y_0, Y_1… The jump time is given by $q_{ij} = \lim_{\Delta t \to 0} \frac{P_{ij}(\Delta t)}{\Delta t}$ and the exit rate from state 'i' is given by $q_{ii} = -\sum_{j \neq i} q_{ij} = \lim_{\Delta t \to 0} \frac{P_{ij}(\Delta t) - 1}{\Delta t}$. Hence a CTMC is a stochastic process $\{X(t) \mid t \geq 0, t \in \mathbb{R}\}$ such that for all $t_0,...,t_{n-1},t_n$, where $t \in \mathbb{R}$, $0 \leq t_0 < \ldots < t_{n-1} < t_n < t$, for all $n \in \mathbb{N}$. Alternatively it is defined as $P(X(t + s) = y / X(s) = x; X(t_n) = x_n, \ldots\ldots X(t_1) = x_1) = p(y; x; t)$.The Markov property and time homogeneity imply that if at time t the process is in state j, the time remaining in state j is independent of the time [13]. The time spent in state j is called as Sojourn times which are exponentially distributed. An amino acid substitution model which is discussed in this paper using the generator matrix can be used for the Likelihood calculation which is the simplest applications of Phylo-grammers [3] , [7].The aim is to find the maximum likelihood estimate of the various arm lengths in the tree, given some continuous-time evolutionary model. This is done by writing down the likelihood of the data in terms of these lengths as parameters and then maximizing this likelihood with respect to these lengths [18].

2. Q MATRICES AND THEIR EXPONENTIALS
The generator matrix is also called as an intensity matrix, rate matrix or 'Q" matrix of the Markov chain and is used to describe the process completely. It is also used to obtain the state transition probability matrix at discrete intervals of time. A regular discrete time Markov chain is called as an Embedded Markov Chain (EMC) or a jump process [11] Every element of the one step transition probability matrix of the EMC can be obtained from the Q matrix and hence the stationary probabilities are also calculated. The Ergodicity of the Markov chain can also be tested using the Q matrix.

Let the Q matrix of a continuous time Markov chain on a countable set I be defined as $Q = (q_{ij})$ for $i, j \in I$ has the following properties:

(i) $0 \leq -q_{ii} < \infty \quad \forall\, i$
(ii) $q_{ij} \geq 0 \quad \forall\, i \neq j$
(iii) $\sum_{j \in I} q_{ij} = 0 \quad \forall\, i.$

Thus in each row of Q the off-diagonal entries are to be any non-negative real numbers subject only to the constraint that the off-diagonal row sum is finite.

$$q_i = \sum_{j \neq i} q_{ij} < \infty.$$

The diagonal entry q_{ii} is defined as $-q_i$, making the row sum as zero.

Consider a Q matrix as $Q = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -1 & 0 \\ 2 & 1 & -3 \end{pmatrix}$

From the above Q matrix the continuous time Markov chain can be represented by the following diagram as

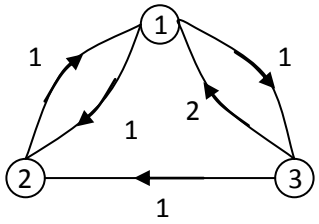

Fig 1: Markov Graph for the given Q matrix

Thus each off-diagonal entry q_{ij} assumes the value attached to the (i, j) arrow on the diagram which is the rate of going from i to j. The numbers q_i are called the rate of leaving i which are not shown in the diagram. For any matrix Q = (q_{ij}, i, j ∈ I) the series $\sum_{K=0}^{\infty} \frac{Q^K}{K!}$ converges to e^Q then let P (t) = e^{tQ}. Then P(t) is the unique solution to the forward equation

$$\frac{d}{dt}P(t) = P(t)Q, \quad \text{with} \quad P(0) = 1.$$

Also it is the unique solution of the backward equation as

$$\left[\left(\frac{d}{dt}\right)^K P(t)\right]_{t=0} = Q^K.$$

Let P is a Stochastic Matrix of the form e^Q then define a process indexed by {n | m : n = 0, 1, 2, …} as $X_{n/m} = X_n^m$ then X_n^m be the discrete time Markov process. Thus discrete time Markov chains with arbitrarily fine grids {n | m: n = 0, 1 …} as time parameter sets give rise to Markov processes when sampled at integer time [4].

2.1 Calculation of P from Q Matrix

Consider a Q matrix defined by

$$Q = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -1 & 0 \\ 2 & 1 & -3 \end{pmatrix} \text{ With } Q^2 = \begin{pmatrix} 7 & -2 & -5 \\ -3 & 3 & 1 \\ -9 & -2 & 11 \end{pmatrix}$$

The characteristic equation of the above matrix is det (Q−x) = $x(x+2)(x+4) = 0$ Hence the eigen values are 0, −2, −4.

Let P_{ij} (t) = a + be^{-2t} + ce^{-4t} for some constants a, b and c with i, j = 1, 2, 3, t ≥ 0.

These constants then obey the following equations as

a + b + c = δ_{ij} with P (0) = 1.

b (−2) + c (−4) = q_{ij} as $\frac{d}{dt}P(0) = Q$

b $(-2)^2$ + c $(-4)^2$ = $q_{ij}^{(2)}$ as $\frac{d^2}{dt^2}P(0) = Q^{(2)}$

Also

$$e^{tQ} = \sum_{K=0}^{\infty} \frac{(tQ)^K}{K!}$$

$$= U\sum_{K=0}^{\infty} \frac{1}{K!}\begin{pmatrix} 0^K & 0 & 0 \\ 0 & (-2t)^K & 0 \\ 0 & 0 & (-4t)^K \end{pmatrix} U^{-1}$$

Where U is an invertible matrix used to diagonalise Q. The constants can also be determined by

$$P_{11}(0) = 1 = a + b + c \quad (1)$$

$$P'_{11}(0) = q_{11} = -2 = -2b - 4c \quad (2)$$

$$P''_{11}(0) = q_{11}^{(2)} = 7 = 4b + 16c \quad (3)$$

$$(2) \Rightarrow -2b - 4(3/2) = -2$$

$$-4b - 3 = -4$$

$$-4b = -4 + 3 = -1$$

$$b = \frac{1}{4}$$

$$(1) \Rightarrow a + \frac{1}{4} + \frac{3}{8} = 1$$

$$a + \frac{2+3}{8} = 1$$

$$a = 1 - \frac{5}{8}$$

$$= 3/8$$

$\therefore$ P_{11} (t) = 3/8 + 1/4 e^{-2t} + 3/8 e^{-4t}. For any time‘t’ the values of p_{ij} can be calculated after evaluating the constants a, b, c.

2.2 Applications of Q Matrix

The stationary distribution $\boldsymbol{\pi}$ is a normalized (meaning that the sum of its entries is 1) left eigenvector of the transition matrix associated with the eigenvalue 1 [9]. The Kolmogorov differential equation can also be defined using the Q matrix as $\frac{d}{dt}\pi(t) = \pi(t)Q$. If $\lim_{t\to\infty} \pi(t)$ exists then taking the limit of the Kolmogorov differential equation[14]. The steady state probabilities can be obtained as πQ = 0 and πe = 1 with Σ π_j = 1. Also

$$\frac{d}{dt}\pi_j(t) = q_{jj}\pi_j(t) + \sum_{i\neq j} q_{ij}\pi_i(t) \quad (j = 1, 2 \ldots)$$

$$\frac{d}{dt}\pi_0(t) = q_{00}\pi_0(t) \quad (j = 0)$$

These above equations can be written as $\frac{d}{dt}f(t) = \beta f(t)$ where β is called the decaying rate if (β < 0) or growth rate (if β > 0) of the entity that f (t) represents. Also the state transition probabilities for discrete continuous-time Markov chains be expressed as P_{ij} = P [x_{K+1} = j | X_k= i] are correlated to the Q matrix as

$$P_{ij} = \frac{q_{ii}}{-q_{ii}} \quad (j \neq i)$$

Also

$$P(t) = e^{Qt} = \lim_{n\to\infty}\left(I + \frac{Qt}{n}\right)^n.$$

2.3 Numerical Example

Evolutionary analyses of sequences are conducted on a wide variety of time scales. A subset of the class of phylo-grammars is the class of homogeneous substitution models, where the mutation rate is not a function of position but it is identical for every site. Such models can be represented as a single-state phylo - HMM [6].Thus, it is convenient to express these models in terms of the instantaneous rates of change between different states [18]. Given a starting (ancestral) state at one position, and

a branch length expressing the expected number of changes to have occurred since the ancestor, then we can derive the probability of the descendant sequence having each of the four states. By expressing models in terms of the instantaneous rates of change we can avoid estimating a large numbers of parameters for each branch on a phylogenetic tree (or each comparison if the analysis involves many pair wise sequence comparisons). Branch lengths (and path lengths) in phylogenetic analyses are usually expressed in the expected number of changes per site. The path length is the product of the duration of the path in time and the mean rate of substitutions.

Phylogenetic methods, particularly for molecular sequence data, have become the primary tool for the determination of evolutionary relationships. These tools have been used to confirm expected relationships such as the chimpanzees are the closest living relative to humans. Phylogenetic trees have been used for various instances to provide evidence about the likely transmission of HIV. Phylogenetics is being used increasingly in comparative genomics and study of gene function. By comparing exact sequences, the amount of sequence divergence can be determined. This measurement of divergence provides information about the number of changes that have occurred along the path separating the sequences. The simple count of differences between sequences will often underestimate the number of substitution because of multiple hits. Since it is very much difficult to estimate the exact number of changes that have occurred, branch lengths (and path lengths) in phylogenetic analyses are usually expressed in the expected number of changes per site. The path length is the product of the duration of the path in time and the mean rate of substitutions. While their product can be estimated, the rate and time are not identifiable from sequence divergence. The phylo-grammar approaches that have been used to day have often used approximate and inefficient versions of EM to estimate parameters [10],[15], or have been limited to particular subclasses of model, e.g. reversible or otherwise constrained models [2],[5] also showed how to apply the EM algorithm to estimate substitution rates in a phylogenetic reversible continuous-time Markov chain model.

The descriptions of rate matrices reflect the relative magnitude of different substitutions. The scaling of these matrices can be done by multiplying every element of the matrix by the same factor, or simply by scaling the branch lengths. If we use β to denote the scaling factor, and ν to denote the branch length measured in the expected number of substitutions(changes) per site then $\beta\nu$ is used the transition probability formulae below in place of μt. Note that ν is a parameter to be estimated from data, and is referred to as the branch length, while β is simply a number that can be calculated from the rate matrix (it is not a separate free parameter).

Phylogenies are usually estimated from aligned DNA sequence data. In phylogenetics, sequences are often obtained by finding a nucleotide or protein sequence alignment, and then taking the bases or amino acids at corresponding positions in the alignment as the characters. Sequences achieved by this might look like AGCGGAGCTTA and GTAGACGC. Commonly used models of molecular evolution treat sites as independent. These common models just need to describe the substitutions among four bases A, C, G, and T at a single site over the time. This substitution process is modeled as a continuous-time Markov chain.

Let X (t) represents the base at time t. Hence X (t) $\in$ {A, C, G, T} for DNA. The variables used in the model are obtained from the results of the phase one phylogenetic analysis. Time t and s are the observed sampling times for an inferred ancestor/descendant transition. States i and j represent the ancestor and descendant genotypes. The number of transitions from i to j were aggregated for all such ancestor/descendant relationships among all trees and are represented by N (i,j). [1] presented parameter estimates for a continuous time Markov process that are analogous to estimates for a discrete time Markov chain. Let P(t) be the transition probability matrix for a continuous Markov model. In a time interval t, the system undergoes a change of state (or stays in the same state, a repetition) according to a set of probabilities associated with the state. P(t) can be expressed in the form [12]

$P(t)= e^{Q t} = \lim_{n\to\infty}\left(I+\frac{Q t}{n}\right)^{n}$ where Q is the infinitesimal generator of the continuous Markov process is an m×m matrix encoding the time independent transition rates for a set of 'm' states.

Consider the Rate matrix

$$Q = \{q_{ij}\} = \begin{pmatrix} -1.1 & 0.3 & 0.6 & 0.2 \\ 0.2 & -1.1 & 0.3 & 0.6 \\ 0.4 & 0.3 & -0.9 & 0.2 \\ 0.2 & 0.9 & 0.3 & -1.4 \end{pmatrix}$$ then

The time to the next transition is ~ $e^{(qi)}$, Where $q_i = -q_{ii}$ .The transition is to state j with probability $\frac{q_{ij}}{\sum_{k\neq i} q_{ik}}$. Let the beginning is at A, change to G at time 0.3, change to C at time 0.8, and then no more changes before time t = 1.

2.4 Transition Probability Matrices

The transition matrix is $P(t) = e^{Q t}$. The matrix Q can be factored as A D A^{-1} where D is a diagonal matrix of the Eigen values and A is the matrix whose columns are corresponding Eigen vectors. All rate matrices Q will have an Eigen value 0 with an eigenvector of all 1s as the rows sum to 0 by construction. For the example rate matrix Q has Eigen values 0, −1, −1.5, and −2.

Hence

$$Q = \begin{pmatrix} 1 & 1 & 3 & 0 \\ 1 & -1 & 0 & 2 \\ 1 & 1 & -2 & 0 \\ 1 & -1 & 0 & -3 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1.5 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix}$$

$$\begin{pmatrix} 0.2 & 0.3 & 0.3 & 0.2 \\ 0.2 & -0.3 & 0.3 & -0.2 \\ 0.2 & 0.0 & -0.2 & 0.0 \\ 0.0 & 0.2 & 0.0 & -0.2 \end{pmatrix}$$

The probability transition matrix is of the form P (t) = A e^{Dt} A^{-1}.

Let the stationary distribution be $\pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix}$. For finite-state-space chains, irreducibility is sufficient.

When the time t is large enough, the probability $P_{ij}(t)$ will be close to $\pi_{\text{ j for}}$ each j.

Also the stationary distribution satisfies $\pi^T Q = 0$.

i.e. $(\pi_1\ \pi_2\ \pi_3\ \pi_4)\begin{pmatrix} -1.1 & 0.3 & 0.6 & 0.2 \\ 0.2 & -1.1 & 0.3 & 0.6 \\ 0.4 & 0.3 & -0.9 & 0.2 \\ 0.2 & 0.9 & 0.3 & -1.4 \end{pmatrix}$

$= (0 \quad 0 \quad 0 \quad 0)$

i.e. $-1.1\,\pi_1 + .2\,\pi_2 + .4\,\pi_3 + .2\,\pi_4 = 0$

$.3\,\pi_1 - 1.1\,\pi_2 + .3\,\pi_3 + .9\,\pi_4 = 0$

$.6\,\pi_1 + .3\,\pi_2 - .9\,\pi_3 + .3\,\pi_4 = 0$

$.2\,\pi_1 + .6\,\pi_2 + .2\,\pi_3 - 1.4\,\pi_4 = 0$

Also $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1.$

Solving the above System the Stationary Probabilities were $\pi = (.2 \quad .3 \quad .3 \quad .2)$.

If β denotes the scaling factor, the value of β can be found by forcing the expected rate of flux of states to 1. The diagonal entries of the rate-matrix (the Q matrix) represent -1 times the rate of leaving each state. Thus we can find the expected rate of change by calculating the sum of flux out of each state weighted by the proportion of sites that are expected to be in that class. Setting β to be the reciprocal of this sum will guarantee that scaled process has an expected flux of 1, i.e. $\beta = 1 \Big/ \left(-\sum_i \pi_i \mu_{ii} \right)$. For DNA substitution models, mainly mechanistic models (as described above) are employed. The small number of parameters to estimate makes this feasible, but also DNA is often highly optimized for specific purposes (e.g. fast expression or stability) depending on the organism and the type of gene, making it necessary to adjust the model to these circumstances. Unlike the DNA models, amino acid models traditionally are empirical models. To evaluate the statistical significance of the rate classes found by the exploratory CVA analyses, we use the following likelihood-based approach. The hidden-state model allows a variety of different substitution rate matrices to be used, depending on a hidden state variable that specifies the structural context of the site [8]. We assume initially that the substitution matrix at site i has the form m_iQ, where the rate matrix *Q* is common to each site, and $m_i \geq 0$ specifies the relative rate at site i [16].

2.5 Comparative Study

The inference of generators for Markov jump processes from discretely sampled time-series is a crucial problem in various field of science.In most practical cases of the embedding problem for Markov chains the generator of a particular transitional probability matrix does not exist.To find the transitional probability matrix the condition $P(t) = e^{Q t}$ is used.

The computation of Q proceeds by various methods such as Diagonal adjustment (DA), weighted adjustment (WA) and MCMC methods. In this paper the using the Diagonal Adjustment method the transition probability matrix along with the Stationary Probabilities were calculated.[19] described in the estimation of portfolio credit risk and pricing credit risk securities. They concluded the method WA performs well by calculating the transitional probability matrix and comparing it with actual transitional probability matrix.[20] describes in credit risk management MCMC method gives the diffence in both transitional probability matrix is very less when compared with DA,WA but the stationary distribution was not calculated. Hence when it is required to calculate the long run stationary distribution for many the inference problems the method discussed in this paper(DA) is Parsimonious.

3. CONCLUSIONS

Modeling in terms of the rate matrix is useful in various ways. For example, a hydrophobic ally-inclined generator matrix might be used for buried amino acids and a hydrophilic matrix for exposed amino acids. An extension to the hidden-state model allows the hidden state variable itself to change over time at some slow rate, modeling rare changes in structural context. An alternative extension allows correlations between hidden state variables at adjacent sites. In this paper a sequence of alignment of nucleotide from the Substitution model with four basics is expressed as a continuous time Markov Chain in terms of its generator matrix. Using the Generator matrix the Stationary probabilities and the Transition Probabilities are calculated. The rate of leaving the state, the maximum likelihood of the rate classes can also be calculated using the generator matrix.

4. REFERENCES

[1] Albert, A ,1962, Estimating the infinitesimal generator of a continuous time, finite state markov process, Ann. Mathemat. Stat., 33, pp (727–753).

[2] Bruno WJ, 1996, "Modelling residue usage in aligned protein sequences via maximum likelihood". Molecular Biology and Evolution, 13(10), pp(1368-1374).

[3] Dayhoff .MO, Eck. RV, Park .CM ,1972, A model of evolutionary changein proteins"**.** In Atlas of Protein Sequence and Structure, Volume 5. Edited by Dayhoff MO, National Biomedical Research Foundation, Washington, DC, pp. 89–99.

[4] Darren.J,2006, Wilkinson. Stochastic Modeling for Systems Biology, Chapman & Hall/CRC.

[5] Holmes I, Rubin GM,2002, "An Expectation Maximization algorithm for training hidden substitution models**,** Journal of Molecular Biology, 317(5), pp. 757–768.

[6] Jukes TH,Cantor C,1969, Evolution of protein molecules, In Mammalian Protein Metabolism Academic Press, New York, pp 121-132.

[7] Kall .L, Krogh .A, Sonnhammer. EL, 2004, A combined transmembrane topology and signal peptide prediction method,Journal ofMolecular Biology, 338(5),pp1027-1036.

[8] Koshi JM, Goldstein RA ,1995, Context-dependent optimal substitution matrices, Protein Engineering, 8, pp. 641–645.

[9] Meyn, S.P. and Tweedie, R.L, 1993, Markov Chains and Stochastic Stability. London: Springer-Verlag. ISBN 0-387-19832-6, Cambridge University Press.

[10] Michalek S, Timmer J, 1999, "Estimating rate constants in hidden Markov models by the EM algorithm", IEEE Transactions in Signal Processing, 47,p.p(226-228).

[11] Norris, J.R., 1999, Markov chains, volume 2 of Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, ISBN0-521-3966, pages (60-70).

[12] Patricia Buendia, Brice Cadwallader and Victor DeGruttola, 2009, "A phylogenetic and Markov model approach for the reconstruction of mutational pathways of drug resistance Bioinformatics, Vol. 25, No. 19, pp. 2522–2529.

[13] Padma.G, Dr.C.Vijayalakshmi,2010, "Analysis of Stochastic Differential equations using Propagators", Publications of Manonmaniam Sundaranar university, Tirunelveli, India, ISBN- 978-81-908352-8-2, pp(43–47).

[14] Padma.G, Dr.C.Vijayalakshmi,2010, "Propagators of Markov Processes" ,Proceedings of the National Conference at B.S. Abdur Rahman University, Chennai, Vol. No1,pp(171-175).

[15] Siepel A, Haussler D, 2004, "Phylogenetic estimation of context dependent substitution rates by maximum likelihood", Molecular Biology and Evolution, 21(3),pp(468-488).

[16] Simon Tavaré, C. Adams, Olivier Fedrigo, Gavin j. P. Naylor ,2001, A model for phylogenetic inference using Structural and chemical covariates, Pacific Symposium on Bio Computing, 6, pp(215–225).

[17] Yang, Z, 1993, "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites", Mol. Biol. Evol., 10, pp(1396–1401).

[18] Yang Z,1994, "Estimating the pattern of nucleotide substitution",Journal of Molecular Evolution, 39,pp(105-111).

[19] Alexander Kreinin and Marina Sidelnikova,2001, "Regularization Algorithms forTransition Matrices" Journal of Algo research quarterly, pp (23-40).

[20] Yasunari Inamura, 2006, "Estimating Continuous Time TransitionMatrices From Discretely Observed Data" Bank of Japan Working Paper Series,pp (1-40).