Dealing with large, sparse continuous time Markov chains

Paula Tataru
Bioinformatics Research Center
Aarhus University, Denmark

December 8, 2011

Motivation and background

The evolution of DNA and protein sequences has been modeled using Continuous Time Markov Chains (CTMCs) since the late 60s (Jukes-Cantor [1], Dayhoff [2]). Newer models have been extended to incorporate a long series of observed biological factors, such as rate heterogeneity, context dependency, insertion-deletion events, hidden structures (annotation by genes for instance) and more general matrices [3]. Extending the basic DNA, coding DNA or amino acid models, CTMCs have been developed where the sequence evolution is context-dependent [4], [5]. CTMCs have also been receiving attention in ecological studies, where the tendency is to turn from deterministic models towards stochastic ones [8]. In this case, the CTMC describes the evolution of populations, by keeping track of the population size.

A CTMC is a stochastic process $\{X(t) : 0 \leq t\}$ that takes values from a set called the state space, and that satisfies the Markov property: at any times $s > t > 0$, the conditional probability distribution of the process at time s given the whole history of the process up to and including time t, depends only on the state of the process at time t. A rate matrix Q of dimensions $n \times n$ is used to describe a CTMC, where n is the size of the state space. q_{ij} represents the rate with which the process jumps from i to j and the diagonal elements are defined as $q_{ii} = -\sum_{j \neq i} q_{ij}$ such that rows add to zero. In this framework, the transition probability matrix over time t is obtained by $P(t) = e^{Qt}$, giving the probability of starting in i and ending in j after time t: $\mathbb{P}(X(t) = j | X(0) = i) = p_{ij}(t)$.

The main purpose of using CTMCs is to do inference on observed data. In the ideal case, given continuously observed sample paths, statistical inference is straightforward: the required statistics are simply the number of transitions between any two states and the total time spent in a state. Even when the process is not continuously observed, expectations of these statistics can still be computed [6], [7]. Other approaches [4], [5] are based on maximum likelihood estimators (MLE): they evaluate the likelihood of observing the data and then estimate the underlying parameters by numerical maximization. The difficult part in computing the likelihood is computing the transition probabilities $P(t)$.

Dealing with large, sparse rate matrices

When the CTMC used is of small size, computing transition probabilities and expectations of statistics can easily be accomplished [7]. However models of exponential size are employed in [4], [5], while in [8] the size represents the maximum population size. Therefore, there is an increasing need for extending the methodology to deal with large matrices. Standard numerical methods for exponentiating matrices could be used [10], but as the rate matrix that needs to be exponentiated has special properties, it is more efficient to employ approaches that take advantage of it.

Next to being very big, the referenced large models have the extra property of being sparse. The context-dependent models allow a non-zero rate only between sequences that differ at one position, while in the population models in [8] at most half of the rates are different from zero.

[4] presents a probabilistic, process-based model for the evolution of promoter regions, by employing a basic DNA evolution model. The evolution of an entire promoter sequence of length l is modeled using a CTMC with a state space of size 4^l which contains all sequences of length l. To approach the problem of computing the likelihood, [4] reduces the model by only considering parsimonious paths between two extant sequences, reducing the space drastically.

[5] concentrates on site-interdependent Markovian codon substitution processes as means of mechanistically accounting for selective features over long-range evolutionary scales. For a codon sequence of length l, the rate matrix is defined by a $61^l \times 61^l$ matrix (assuming the universal genetic code), in principle allowing the possibility of a total dependence between all codons, although still based on a point mutation process. To evaluate the likelihood, [5] relies on various MCMC computational devices.

[8] argues that stochastic ecological and epidemiological models are now routinely used in practice, but the difficulty with such models is the need to resort to simulation methods when the population size (and hence the size of the state space) becomes large. [8] presents two methods for evaluating the transition matrices for CTMCs with large state spaces: the interpolating polynomial method (IP) and the exact probabilities and exact time method (EPET). [8] illustrates these methods using two models: the Susceptible-Infectious-Susceptible disease dynamics (SIS) and studying species that are affected by catastrophic events (BDC). SIS is a special type of Markov chain, a pure birth-death process, with a highly sparse matrix. The rate matrix involved in BDC in almost triangular. The EPET method reduces the size of the state space by grouping states into contiguous bins and approximating transitions between them as Markovian. The IP method also reduces the state space, but via use of interpolating polynomials to infer intermediate, omitted state probabilities.

In all cases, a more direct solution would be preferable to reducing the state space or using sampling strategies.

In [7] three methods are presented which can be used for computing the transition probabilities but also expectations of statistics. One of the methods, called uniformization, exploits the fact that the exponentiation is done on a rate matrix, rather than a general one. The method involves an auxiliary stochastic process $Y(t)$ dependent on the uniformization rate $\mu = \max_i(-q_{ii})$. The calculations involve only basic operations (addition, multiplication) on matrices.

[9] presents an inexact version of the uniformization method. Inexact algorithms allow certain operations, in particular matrix exponentiation, to be performed inexactly in view of trading accuracy for speed. [9] finds that the inexact approach can retain the desired accuracy and that, the larger the problem size, the more effective the approach is. [9] also argues that it is important to use appropriate sparse storage formats to realize the potential of the method.

Research directions

As dealing with large, sparse matrices is of obvious interest, a literature review would be required for an appropriate coverage of the field. Extensions of the presented methods should be considered, such as adapting uniformization to using sparse matrices.

Uniformization can be, in some cases, computationally very intensive. Adaptive uniformization has been introduced in [11] to overcome this problem. The basic idea is that the general uniformization rate is replaced by one that adapts depending on the set of states that the process can be in after a particular number of jumps. [12] presents an on-the-fly variant of adaptive uniformization, where the speed of the original algorithm is improved at the cost of a small approximation error. Therefore, the uniformization method could be changed to consider the adaptive approach.

It is unclear which of the three presented approaches in [8] and [9] is more effective and a comparison of their performance, as well as of the uniformization-based methods, is of interest.

The final question that arises is if the direct calculations are more efficient than the approaches presented in [4], [5].

References

[1] Jukes TH, Cantor CR (1969): **Evolution of protein molecules.** *Mammalian protein metabolism*, **3**:21–132.

[2] Dayhoff MO, Schwartz R, Orcutt BC (1978): **A model of evolutionary change in proteins.** *Atlas of protein sequence and structure*, **5**:345–358.

[3] Yang Z (2006): **Computational molecular evolution.** *Oxford University Press.*

[4] Raijman D (2007): **A probabilistic model for the evolution of promoters.** *M.Sc. thesis*, Tel-Aviv University, Faculty of Exact Sciences.

[5] Rodrigue N et al. (2009): **Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons.** *Molecular Biology and Evolution*, **26(7)**:1663–1676.

[6] Hobolth A, Jensen JL (2011): **Summary statistics for endpoint conditioned continuoustime Markov chains.** *Journal of Applied Probabilities*, **48**:1–14.

[7] Tataru P, Hobolth A (2011): **Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time Markov chains.** *BMC Bioinformatics*, **12**:465.

[8] Keeling Mj, Ross JV (2009): **Efficient methods for studying stochastic disease and population dynamics.** *Theoretical Population Biology*, **75**:133–141.

[9] Sidje RB (2011): **Inexact uniformization and GMRES methods for large Markov chains.** *Numerical Linear Algebra with Applications*, **18**:947–960.

[10] Moler C, Van Loan CF (2003): **Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later.** *SIAM Review*, **45(1)**:3.

[11] Van Moorsel APA, Sanders WH (1994): **Adaptive uniformization.** *Communications in Statistics: Stochastic Models*, **10**:619-648.

[12] Mateescu et al. (2010): **Fast adaptive uniformization of the chemical master equation.** *IET Systems Biology*, **6**:441-452.