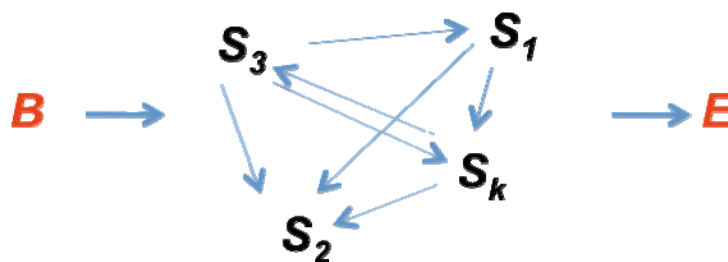


Integration over paths in Continuous Time Markov Chains

1.12.11

Motivation and Background. Continuous Time Markov Chain has been used to model the evolution of DNA and protein sequences in the late 60s (Jukes-Cantor, Neyman, Dayhoff). They are normally formulated in terms of a rate matrix, Q , and time between the two homologous sequences to be analyzed, t . Q has dimensions $k \times k$, where k is the size of the state space. q_{ij} is the rate with which the process jumps from i to j . These models have been extended seriously 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 rate matrices (Yang, 2006). If k is of moderate size, the transition probabilities, $p_{ij}(t)$ [in matrix form $P(t) = \{p_{ij}(t)\}$], can be calculated by matrix exponentiation, $P(t) = \exp(tQ)$.

Methods have also been extended to sets of sequences related by a tree or some other genealogical structure such as the ancestral recombination graph. Going from $P(t)$ (describing pair probabilities) to the probabilities of the observing the values on n leaves on a tree, can be done by Felsenstein's (1981) algorithm.



The state space has k elements, including B (beginning) and E (end). If there are any cycles on a path from B to E , with all positive q_{ij} 's, then there are infinitely many possible paths from B to E . Questions concerning probabilities and moments of paths are clearly of interest as they represent statements about the path of evolution. This illustration corresponds to the situation where B is the ancestor to E on a branch. This should clearly be generalized to the situation, where states have been observed on the leaves of a phylogeny.

In observing a pair of structures and making inference about parameters $P(s1, s2)$ is the key quantity, but it is not unproblematic to calculate. If the process is time reversible, then this is equal to $P(s1)P(s1 \rightarrow s2)$. If the process is not time reversible, then it is

$\sum_{sa} P(sa)P(sa \rightarrow s1)P(sa \rightarrow s2)$, which implies summing over all possible ancestors. If time reversible, then $P(s1)$ can be calculated by letting $t \rightarrow \text{infinity}$ and the equilibrium distribution can be found in the rows of $P(\text{infinity})$.

Besides $P(s1, s2)$ and $P(s1 \rightarrow s2)$, it is of interest to know properties the actual paths $s1 \rightarrow s2$ and some can be calculated easily, some only with difficulty. A key trick is the strong Markov property that allow the process stopped on a fixed time, but at the time defined by a stochastic event like reaching a specific state.

Quantities of interests are probabilities, densities, expectation and higher order moments and would in this situation be related to transition paths and different marginalisations [summing of a set with focus on specific properties] such as time spent in specific states and number of transitions taken. It can also be of interest to sample from a distribution without explicitly stating what the distribution is. Sampling paths $a \rightarrow b$ is not entirely straightforward as we condition on ending in b – the naïve approach of simulating paths unconditionally and then only keep the ones that actually are in b at time t , can be very inefficient if this event has low probability. Hobolth and Stone (2009) discuss three ways to sample a path (including jump times) from a to b . The most obvious is sampling paths from a to b and only keeping the ones that is in b at time t . One source of inefficiency is if $P_{ab}(t)$ is very small as then most simulations are discarded. This can be countered by *uniformisation*, but again at a computational cost. Uniformisation will increase all exits rates to the same level, by adding self-jumps with the appropriate rate. When exit rates are the same, the distribution of the jumps will follow a Poisson process. This can have several advantages. For path sampling, it allows sampling the jump number, n , independently of the states visited. The continuous MC can be discretized and given n , it is possible to calculate the probability of the next jump given we have been in b at jump n (an time t). This allows simulation without having to discard paths because we don't end in b . The computational cost of uniformisation is large is the rate augmentation is large.

Miklos, Lunter and Holmes (2004 – appendix A) present an algorithm that given a path $a \rightarrow s_1 \rightarrow s_2 \rightarrow \dots \rightarrow s_k \rightarrow b$ can calculate its probability (ie integrating out times) in k^2 time.

Expected number of transitions and time spent in a set of states can be calculated (see for instance Hobolth and Jensen, 2005, section 4.2 and Appendix B and C) by 1 dimensional integrals. Expectations of products of n transitional and m dwelling times can be valued by $n+m$ dimensional integration. However, by eigenvalue decompositions this can be substantially simplified. Minin and Suchard (2008a) present methods that can calculate the moments of the number of times a set of transitions has been taken.

The TSP algorithm and Markov Chain Monte Carlo (MCMC) The above methods are deterministic in the sense that quantities in questions are calculated without use of randomness. When the state space become large ($\gg 1000$) these involved matrix calculations cannot be done and other techniques must be used. The method of choice would be MCMC. These methods have been studied by statisticians for decades now (Liu, 2001). Highly similar methods have been developed and used under the name Transition Path Sampling (TSP) (Bolhuis et al, 2002). Applications often involve dynamic paths, where both start configuration and end configuration of the system is known – for instance the catalysis of a substrate into a product. These algorithms and extensions are now being used large scale in Molecular Dynamics (MD – Fu et al., 2007; Rogal and Bolhuis, 2008) (The modelling problem described here is very similar to the TPS problem and can be explored using the same algorithm. Let $x(t)$ be the configuration at time t . $x(0)$ will be the configuration of the first protein and $x(T)$ the configuration of the second protein. For each $x(t)$ there is a set of possible neighbours, $N(x(t))$, defined by applying to the sequence of $x(t)$ and predicting its structure. Let $S=[x(0),x(\delta),x(2\delta),\dots,x(k\delta)=x(T)]$ be a proposed evolutionary trajectory. Let $F(x)$ be the fitness of x , then $P[x(i\delta) \rightarrow x((i+1)\delta)] = F(x((i+1)\delta)) / \sum F(x')$, where the summation is over all neighbours to $x(i\delta)$. This summation must be done stochastically. The $P[S]$ is the product of the probabilities of all the individual steps. Given a way to define a neighbourhood to a path, then an alternative path S' can be proposed and this path can be chosen as current with probability $\max[1, P[S']/P[S]]$. This defines a random walk in path space that will eventually visit all paths according to their probability.

Possible Contents of report/presentation:

- Introduction/Motivation
- Technical background: CTMC, path sampling,
- Comparison of the basic path sampling strategies
- Extensions and Specializations [more end points, sparse matrices,...]
- Application and Examples Sequence Evolution
 - Structure Evolution
 - Molecular Dynamics
 -
- Discussion

Comment: The presentation/report should have a strong bias towards recent statistical models, evolutionary models and evolutionary data analysis. This is mainly because this is where the field is moving and there is such an enormous number of earlier methods that a long report [like the Brown et al (1996) paper] could be written only dedicated to work up to 2000, which would not lead to exciting new ideas.

References.

- Bolhuis, Chandler, Dellago and Geissler "TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark" An Rev of Physical Chemistry Vol. 53: 291-318
- Felsenstein, J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach *J Mol Evol.* 1981;17(6):368-76.
- Fu, Yang and Gao (2007) Selective Sampling of transition paths *J. Chem Phys.* 127.154107/7
- Hill, T (1988) Discrete-time random walks on diagrams (graphs) with cycles *PNAS* 85.5345-5349
- Hill, T (1988) Further properties of random walks on diagrams (graphs) with and without cycles *PNAS* 85.3271-3275
- Hill, T (1989) Free Energy Transduction and Biochemical Cycle Kinetics Dover
- Hobolth, A. (2008). A Markov Chain Monte Carlo Expectation Maximization algorithm for statistical analysis of DNA *J Compu Graph Stat*, 17, 138-164
- Hobolth and Stone (2009) EFFICIENT SIMULATION FROM FINITE-STATE, CONTINUOUS-TIME MARKOV CHAINS WITH INCOMPLETE OBSERVATIONS
- Hobolth, A. and Jensen, J.L. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Stat appl Genet Mol. Biol.* 4, 18
- Hobolth and Stone (2009) Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution** *Ann Appl. Stats.* 3.3.1204-1231.
- O'Brien, Minin, and Suchard (2009) Learning to Count: Robust Estimates for Labelled Distances between Molecular Sequences *Mol.Biol.Evol.* 26(4):801–814.
- Liu (2001) Monte Carlo Strategies for Scientific Computing Springer
- Minin and Suchard (2008) Fast, accurate and simulation-free stochastic mapping 3995 363 *Phil. Trans. R. Soc. B*
- Minin and Suchard (2008) Counting labeled transitions in continuous-time Markov models of evolution *J. Math. Biol.* 56:391–412
- MOLER & VAN LOAN 2003 Nineteen Dubious Ways to Compute the Exponential of a. Matrix, Twenty-Five Years later. *SIAM.* 45.1.3–000.
- Rogal and Bolhuis (2008) Multistate transition path sampling *J Chem Phys* 129.224107/9
- Yang, Z. (2006) Computational Molecular Evolution OUP