Paths in Continuous-Time Markov Chains with Applications to Studies of Protein Tertiary Structure Evolution

Andreas Sand

31 December 2009

Contents

List of Figures

Chapter 1

Introduction

This report composes the compulsory project in the course *Research Topics in Computational Biology* held by Jotun Hein in autumn 2009 at Bioinformatics Research Centre, University of Aarhus. Part of the work presented in this report was cunducted in colaboration with Tomas Fabsic (tomasf167@hotmail.com) during the Summer Student Program 2009 at Department of Statistics, University of Oxford.

In the recent years Continuous-time Markov Chains (CTMCs) have been used successfully in the field of bioinformatics to model i.a. the evolution of DNA sequences and protein sequences. Until now little effort has been done to apply CTMCs to study the evolution of protein tertiary structures. In this report we use CTMCs to model the evolution of one tertiary protein structure A into another tertiary protein structure B. When studying the evolution of DNA and protein sequences one usually models each site in the sequences by a continuous-time Markov model, giving a state space of size four for each site. Compared to this, our use of CTMCs is very different: We use just one Markov chain with the state space containing two protein structures A and B together with a set of structures which are believed to be the most relevant structures for the evolution of A into B, and we study path from A to B via a subset of this set of relevant stepping stones.

In this report we focus on the development of techniques to study paths in CTMCs. Results for specific sets of proteins and stepping stones are therefor not presented, but the techniques are illustrated using a smaller an more intuitive example. In the protein structure case, the basis of our analysis is a distance matrix for the set of stepping stones and the two structures A and B. By this reason some of our analysis applies to the distances between the structures rather than to CTMCs.

The remainder of this report is structured as follows: Section 1.1 gives a description of a recurring example, The German Roadmap, that we will use to illustrate our techniques. Chapter 2 gives some tools for analysing distance matrices for path structure. In chapter 3 we describe statistical

Figure 1.1: The German roadmap

approaches to analyse paths in CTMCs from one state A to another state B. In chapter 4 we add an outside state to the model, modelling all protein structures not included in the set of stepping stones. In chapter 5 we discuss the results and suggest further work to be done.

1.1 Example: The German roadmap

In this section we describe our recurring example, The German Roadmap, which can be easilly visualised and is easy to reason about. We will use this example to illustrate the developed theory and to gain better insight into the importance of the different parameters of the model.

We use the highways and the greater cities in the western part of Germany as our model; constructing a continuous-time Markov chain $X(t)$ with the finite state space S consisting of the 17 numbered cities in figure 1.1. The transition rate matrix $Q = \{q_{ij}\}_{i,j \in S}$ is constructed from the distances between the cities in the following way: If there is a direct highway between the cities i and j, then $q_{ij} = 1/d_{ij}$, where d_{ij} is the distance between i and j. Otherwise $q_{ij} = 0$. That is: a transition between two cities is only possible, if there is a direct highway between the two cities. In that case the rate of the transition is a decreasing function of the distance between the two cities, such that transitions between cities close to each other is more likely than transitions between distant cities.

Chapter 2

Path Structure Analysis

In the case of analysing protein tertiary structures we are given a distance matrix for the set of stepping and the two structures A and B. In situations like this it would be useful to have some measure of structure in the distance matrix. This would be helpful in evaluating whether the model based on the distance matrix is likely to show some interesting behaviour. In our case the structure we are interested in is *path structure*; that is: does the distance matrix show any indication of the evolution going in any specific direction? E.g. from A to B.

2.1 Inbetweenness

One measure for path structure could be the amount of inbetweenness between the structures in our model. To use this we first need to define the *inbetweenness relation* on pairs of structures. This can be done in several different ways. If the distances between the structures in our model was a metric, it would be natural to define one structure b to be inbetween a and c if the distance traveled when going from a to c does not become much longer when going via b; that is if $d(a,c) + \epsilon \geq d(a,b) + d(b,c) \geq d(a,c)$ for some relatively small ϵ. However the distances provided to us does not compose a metric on the structures: they are e.g. not symmetric. By this reason we chose to define the inbetweenness relation a bit differently:

Definition 1 *Let S be the set of structures in the model, and let $a, b, c \in S$*

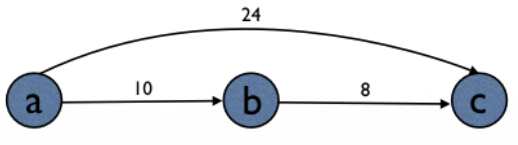


Figure 2.1: b is inbetween a and c for $\epsilon \geq 6$.

be three structures. b is said to be inbetween a and c if

$$d(a,c) + \epsilon \geq d(a,b) + d(b,c).$$

The inbetweenness score in_score is defined by:

$$in\_score = |\{(a,b,c) \in S \times S \times S | b \text{ is inbetween } a \text{ and } c\}|.$$

That is we allow the distance traveled when going from a to c via b to be smaller than the direct distance from a to c. This would not be possible in a metric.

In the case of minimum path structure in which the easiest way from any state x to any state y is the direct jump from x to y, the inbetweenness score will be 0. But in the case of maximal path structure between the start state A and the end state B, A would be placed at one end of a stretch containing all stepping stones and B would be placed at the other end. In this case we get the maximal inbetweenness score, since everything is inbetween everything. Thus we see intuitively that a low inbetweenness score follows from little path structure, while a higher inbetweenness score follows from more path structure.

The inbetweenness score is easy to compute by examining all triples of structures. The inbetweenness score in the German Road is 1209 when $\epsilon = 0.0$ and 1354 when $\epsilon = 20.0$. This is approximately 59% and 66% of the total number of triples, repectively. Unfortunately inbetweenness seems to be very hard to reason about: It is hard to evaluate a given inbetweenness score as significantly high or significantly low, since we do not know how high a score we can expect given the number of stepping stones. By this reason we will not use the inbetweenness score to evaluate the path structure of the given distances.

In the late phase of this project we have thought of a few additional measures of inbetweenness that might be of interest to study:

- How many structures are inbetween the starting state A and the ending state B; that is what is the number of states x such that there is a path from A to B via x shorter than $d(A,B)$?
- How many paths from A to B are shorter than $d(A,B)$?

These measures can be evaluated easily using a breath first search of the graph constructed from A, B, the stepping stones and the given distances, but we have not had time to persue this. Furthermore the same difficulties as above might appear when trying to evaluate the significance of these scores.

2.2 All pairs shortest paths

Another measure for the amount of path structure could be the difference between the given distances for all pairs of states and the lengths of the

shortest paths between all pairs of states. Given the distance matrix for a set of stepping stones and two structures A and B, it is not very likely to show any interesting behaviour for the evolution of A into B if the shortest path from A to B is the direct transition from A to B. On the contrary we wish the distance matrix to show a path structure, such that the shortest path from A to B are shorter than this tirect transition. This path structure measure together with a measure of path structure for all other pairs of structures in the model can be captured by the all pairs shortest paths matrix. The all pairs shortest paths matrix may be computed by the Floyd-Warshall algorithm.

2.2.1 The Floyd-Warshall algorithm

Consider a directed graph $G = (V, E)$ with distance matrix $D = \{d_{ij}\}_{i,j \in V}$. G may have negative negative weight edges but not negative weight cycles (as the length of the shortest path between any pair (i, j) of nodes for which such a cycle is reachable from i and j is reachable from the cycle would have lenght $-\infty$). Number the nodes of G such that $V = \{v_1, v_2, ..., v_{|V|}\}$, and let $dist(i, j, k)$ be the length of the shortest path from i to j visiting only nodes in $\{v_1, v_2, ..., v_k\}$ besides i and j. Then for $k = 0$

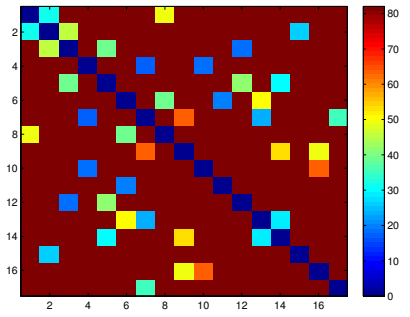
$$dist(i, j, 0) = d_{ij}, \tag{2.1}$$

and for $k = |V|$ $dist(i, j, |V|)$ is the length of the shortest path from i to j. Furthermore note that for $k \geq 1$

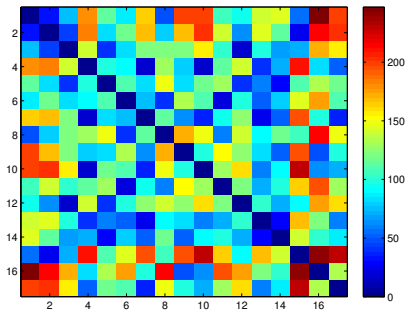
$$dist(i, j, k) = \min \begin{cases} dist(i, j, k-1) \\ dist(i, k, k-1) + dist(k, j, k-1), \end{cases} \tag{2.2}$$

since the shortest path from i to j with all intermediate nodes drawn from $\{v_1, v_2, ..., v_k\}$ either visits k or does not visit k.

Using the recursion described in (2.1) and (2.1) we can compute the all pairs shortest path distances in time $O(|V|^3)$ using $O(|V|^2)$ space. The all pairs shortest path distances in the German Roadmap is summarised in figure 2.2.



(a) Distances



(b) All pairs shortest path distances

Figure 2.2: All pairs shortest paths in The German Roadmap

Chapter 3

Statistical Analysis

In this chapter we study properties of paths in continuous-time Markov chains (CTMCs) starting in some state A at time 0 and ending in another state B at a specified point in time T. We will only consider the case of Markov chains evolving on a finite state space. The main quantities of interest are:

- The time spent in a state on the pah.
- The probability of visiting a particular state on the path.
- The probability distribution of the number of transitions on the path.
- The probability distribution of the number of transitions on the path.

In the remaining part of this chapter, we denote the continuous-time Markov chain $\{X(t) : 0 \leq t \leq T\}$ by $X(t)$ or X. S will denote the state space, and m will be the size of this. Finally Q will be the rate matrix of X and $P(t)$ will be the probability transition matrix of X. Using this notation we have:

$$P_{ij}(t) = P\left(X(t) = j | X(0) = i\right) \text{ and}$$
$$P(t) = \exp(Qt).$$

3.1 Probability of visiting a set of states

For any state $j \in S$ we would like to be able to compute the probability of j being visited during a path starting in state A and ending in state B at time T. For this purpose the concept of *taboo probabilities* as defined by Neuts[7] becomes very useful. We follow the exposition of taboo probabilities as it appears in [7].

Let $H \subseteq S$ be a subset of the state space of X, and let $P_{ij}^{H}(t)$ be the probability that X is in state j at time t and $X(t') \notin H$ for any $t' \in [0; t]$ given X starte in state i. That is:

$$P_{ij}^{H}(t) = P\left(X(t) = j, \forall t' \in [0; t] X(t') \notin H | X(0) = i\right).$$

Let D be the $(m - |H|) \times (m - |H|)$ matrix derived from Q by for each $k \in H$ deleting the kth row and kth column. Then

$$P_{ij}^{H}(t) = \exp(Dt)_{ij}. \tag{3.1}$$

Now let $F(k)$ be the probability that the state k is visited during a path starting in state A and ending in state B at time T. Then $F(k)$ may be computed by

$$\begin{aligned} F(k) &= 1 - \frac{P_{AB}^{\{k\}}(T)}{P\left(X(T) = B | X(0) = A\right)} \\ &= 1 - \frac{P_{AB}^{\{k\}}(T)}{\exp(Q \cdot T)_{AB}}. \end{aligned} \tag{3.2}$$

3.2 Uniformization

An important technique for analysing CTMCs is *uniformization*. We introduce it here, and we use it for deriving a formula for the expected time spent in a state on a path of X from state A to state B, and to derive recursions for the probability mass distributions of the total number of transitions on paths from A to B and the number of transitions between two specified states on paths from A to B.

We recall that we are analysing a CTMC $\{X(t) :) \leq t \leq T\}$ with rate matrix Q, state space S of size m, and $X(0) = A$, $X(T) = B$. In this setting the transition probability matrix is given by $P(t) = \exp(Qt)$. Let $q_i = -q_{ii}$ be the rate for leaving state i and define $\mu = max_{i \in S} q_i$. That is μ is the highest rate with which $X(t)$ will leave a state. We observe that

$$\begin{aligned} Q &= \mu \left(Q/\mu + I - I \right) \\ &= \mu \left(R - I \right), \end{aligned} \tag{3.3}$$

where $R = Q/\mu + I$. Note that R is a probability matrix since the rows of Q sums to zero. Using (3.3) we get the probability matrix for the uniformized CTMC:

$$\begin{aligned} P(t) = e^{Qt} &= e^{\mu(R-I)t} = e^{-\mu t} e^{\mu R t} \\ &= e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\mu t R)^n}{n!} \\ &= \sum_{n=0}^{\infty} e^{-\mu t} \frac{(\mu t)^n}{n!} R^n \\ &= \sum_{n=0}^{\infty} Pois(n; \mu t) R^n. \end{aligned} \tag{3.4}$$

Thus the uniformized CTMC is a process where the epochs of state changes are determined by a homogeneous Poisson process in which the rate of transitions is μ, and the transitions are determined by a discrete-time Markov

chain with transition matrix $R = Q/\mu + I$. Let this process be denoted by $Y(t)$. We assume that $X(t)$ only makes transitions when $Y(t)$ does. If the diagonal entries of Q are not all identical, $Y(t)$ may have virtual transitions in which a transition occurs but the state does not change (a transition in $X(t)$ does not occur). We will let J denote the number of transitions of $Y(t)$ (including virtual transitions) and let N denote the number of (real) transitions of $X(t)$ (excluding virtual transitions). J and N are then stochastic variables and N is always less than or equal to J.

In the succeeding three sections we follow the lines of the exposition in [3].

3.3 Expected time spent in a state

For any state $j \in S$ we would like to be able to compute the expected time spent in j on a path of X starting in state A at time 0 and ending in state B at time T. That is, we would like to compute $E\left(T(j)|X(0) = A, X(T) = B\right)$. We write $E\left(T(j)|Q, a, b\right)$ for this expectation. Hobolth and Jensen gives an expression for this expectation[4]:

$$E\left(T(j)|a, b\right) = \frac{\int_0^T P(t)_{Aj} P(T-t)_{jB} dt}{P(T)_{AB}}. \tag{3.5}$$

Using (3.4) we get the denominator:

$$P(T)_{AB} = \sum_{n=0}^{\infty} Pois(n; \mu T)(R^n)_{AB}, \tag{3.6}$$

and we derive the nominator:

$$\begin{aligned}
&\int_0^T P(t)_{Aj} P(T-t)_{jB} \\
&\quad = \int_0^T \left[\sum_{k=0}^{\infty} \frac{(\mu t)^k}{k!} e^{-\mu t} (R^k)_{Aj}\right] \left[\sum_{l=0}^{\infty} \frac{(\mu(T-t))^l}{l!} e^{-\mu(T-t)} (R^l)_{jB}\right] dt \\
&\quad = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} (R^k)_{Aj} (R^l)_{jB} \frac{T}{(k+l+1)!} e^{-\mu T} (\mu T)^{k+l} \\
&\quad = \sum_{n=0}^{\infty} \frac{T}{n+1} e^{-\mu T} \frac{(\mu T)^n}{n!} \sum_{m=0}^{n} (R^m)_{Aj} (R^{n-m})_{jB}.
\end{aligned} \tag{3.7}$$

3.4 Probability distribution for the total number of transitions

We would also like to be able to compute the probability mass distribution of the total number of transitions on a path of $X(t)$ starting in state A at

time 0 and ending in state B at time T. The recursion for this is based on the derivation by Siepel et al. in [11]. Recall that N is the stochastic variable capturing the number of transitions by the uniformized CTMC Y excluding virtual transitions, and J is the stochastic variable capturing the total number of transitions by the uniformized CTMC including virtual transitions. We observe that

$$\begin{aligned}
&P(N(T) = n, X(T) = B | X(0) = A) \\
&\quad = \sum_{j=n}^{\infty} P(N = n, X(T) = B, J(T) = j | X(0) = A) \\
&\quad = \sum_{j=n}^{\infty} P(N = n, Y(j) = B | J = j, Y(0) = A) P(J(T) = j | Y(0) = A) \\
&\quad = \sum_{j=n}^{\infty} P(n, B | j, A) Pois(j; \mu T),
\end{aligned} \tag{3.8}$$

where $P(n, B|j, A)$ is the probability of that $X(t)$ makes n transitions (excluding virtual transitions) and ends in B given that it started in A and the uniformized CTMC makes j transitions. $Pois(j; \mu T)$ is the probability that the uniformized CTMC makes j transitions (including virtual transitions) in time T.

To compute $P(n, B|j, A)$, we note that

$$P(n, B | j = 0, A) = \begin{cases} 1 & \text{if } a = b \text{ and } n = 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

and

$$\begin{aligned}
P(n, B | j, A) =& P(N = n, Y(j) = B | J = j, Y(0) = A) \\
=& P(N = n, Y(j) = B, Y(j-1) = B | J = j, Y(0) = A) \\
&+ \sum_{C \in S \setminus \{B\}} P(N = n, Y(j) = B, Y(j-1) = C | J = j, Y(0) = A) \\
=& P(Y(j) = B | Y(j-1) = B) \\
&\cdot P(N = n, Y(j-1) = B | J = j-1, Y(0) = A) \\
&+ \sum_{C \in S \setminus \{B\}} P(Y(j) = B | Y(j-1) = C) \\
&\cdot P(N = n-1, Y(j-1) = C | J = j-1, Y(0) = A) \\
=& R_{BB} P(n, b | j-1, A) + \sum_{C \in S \setminus \{B\}} R_{CB} P(n-1, C | j-1, a).
\end{aligned} \tag{3.10}$$

Using this recursion we compute we compute $P(n, B|j, A)$ for $j = 0, 1, ...j_{max}$, where j_{max} is the smallest j such that $Pois(j + 1, \mu T)$ is smaller than the

machine precision. This ensures that we can compute $P(N(T) = n, X(T) = B|X(0) = A)$ to the highest possible precision using (3.8) summing from n to j_{max}.

Finally, having computed $P(N(T) = n, X(T) = B|X(0) = A)$ we get $P(N(T) = n|X(0) = A, X(T) = B)$ by

$$\begin{aligned} P(N(T) = n|X(0) = A, X(T) = B) &= \frac{P(N(T) = n, X(T) = B|X(0) = A)}{P(X(T) = B|X(0) = A)} \\ &= \frac{P(N(T) = n, X(T) = B|X(0) = A)}{\exp(QT)_{AB}}. \end{aligned}$$

Alternatively the denominator can be computed by

$$P(X(T) = B|X(0) = A) = \sum_{n=0}^{\infty} P(N(T = n, X(T) = B|X(0) = A)).$$

3.5 Probability distribution for the number of transitions between specific states

Let $i, j \in S$ be two given distinct states of $X(t)$ and let N_{ij} denote the number of transitions between i and j. We will show how to compute the probability mass distribution of N_{ij} given that $X(0) = A$ and $X(T) = B$. That is, we will show how to compute $P(N_{ij} = n_{ij}|X(0) = A, X(T) = B)$.

Now note that

$$\begin{aligned} &P(N_{ij} = n_{ij}, X(T) = B|N(T) = n, X(0) = A) \\ &= P(N_{ij} = n_{ij}, Z(n) = B|N(T) = n, Z(0) = A), \end{aligned} \tag{3.11}$$

where Z is the skeleton process capturing state change. The matrix $S = \{s_{ij}\}_{i,j\in S} = \{q_{ij}/q_{ii}\}_{i,j\in S}$ is then the transition probability matrix of Z, in which s_{ij} is the probability of jumping from state i to state j given a transition occurs.

To compute $P(N_{ij} = n_{ij}, X(T) = B|N(T) = n, X(0) = A)$ we observe

$$\begin{aligned} &P(N_{ij} = n_{ij}, Z(1) = B|N(T) = 1, Z(0) = A) \\ &= \begin{cases} s_{AB} & \text{if } A = i, B \neq j \text{ and } n_{ij} = 0 \\ s_{ij} & \text{if } A = i, B = j \text{ and } n_{ij} = 1 \\ s_{AB} & \text{if } a \neq i \text{ and } n_{ij} = 0 \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{3.12}$$

and for $n \geq 2$ we have

$$\begin{aligned} &P(N_{ij} = n_{ij}, Z(n) = B|N(T) = n, Z(0) = A) \\ &= s_{ij} P(N_{ij} = n_{ij} - 1, Z(n-1) = i|N(T) = n-1, Z(0) = A) \\ &\quad + \sum_{C\in S\setminus\{i\}} s_{Cj} P(N_{ij} = n_{ij}, Z(n-1) = C|N(T) = n-1, Z(0) = A), \end{aligned} \tag{3.13}$$

if $j = B$, and

$$\begin{aligned} &P(N_{ij} = n_{ij}, X(n) = B | N(T) = n, X(0) = A) \\ &= \sum_{C \in S} s_{CB} P(N_{ij} = n_{ij}, Z(n) = C | N(T) = n-1, Z(0) = A), \end{aligned} \tag{3.14}$$

if $j \neq B$. Now the probability of the number of transitions between state i and j being n_{ij} can be computed by

$$\begin{aligned} &P(N_{ij} = n_{ij} | X(n) = B, X(0) = A) \\ &\quad = \sum_{n=0}^{\infty} P(N_{ij} = n_i j | N(T) = n, X(n) = B, X(0) = A) \\ &\quad \cdot P(N(T) = n | X(n) = B, X(0) = A). \end{aligned} \tag{3.15}$$

The last term in this expression is computed by the recursion in section 3.4, and the first term is computed by

$$\begin{aligned} &P(N_{ij} = n_{ij} | N(T) = n, X(T) = B, X(0) = A) \\ &\quad = \frac{P(N_{ij} = n_{ij}, Z(n) = B | N(T) = n, Z(0) = A)}{P(Z(n) = B | N(T) = n, Z(0) = A)} \end{aligned}$$

3.6 Approximating T

To perform the analysis described in the previous sections, we need to estimate the time T in which the protein structure A evolves into the protein structure B. For our estimate (and in lack of better alternatives) we choose the expected hitting time of B given that the chain starts in A. A method for calculating this is presented by Norris in [8], and we follow his exposition here.

Let $J \subseteq S$ be a subset of the state space of X, and let $D^j = \inf\{t \geq 0 : X(t) \in J\}$. Then D^j is called the hitting time of the set J; D^j is the time elapsed before X is in a state in J. Let $k_i^J = E(D^J | X(0) = i)$ be the expected hitting time of J when X is in state i at time 0. Norris then proves the following theorem:

Theorem 2 *Assume $q_{ii} < 0$ for all $i \in J$. The vector of expected hitting times $k^J = (k_i^J)_{i \in S}$ is the minimal non-negative solution to the system of linear equations*

$$\begin{aligned} k_i^J &= 0 \quad \textit{for } i \in J \\ -\sum_{l \in S} q_{ii} k_l^J &= 1 \quad \textit{for } i \notin J. \end{aligned} \tag{3.16}$$

Using theorem 2 we can easilly compute k^J as the solution of the linear program corresponding to (3.16) minimizing $f(K^J) = k_1^J + k_2^J + ... + k_m^J$. Letting $J = \{B\}$, the expected hitting time of B when the chain starts in A is k_A^J.

3.7 Example: The German roadmap

To use the methods from the previous sections we chose the initial (A) state to be the one corresponding to München (state 16) and the ending state (B) to be the one corresponding to Saarbrücken (state 17). We chose the time T to be the expected hitting time of Saarbrücken when traveling from München with rates being the reciprocal of the distances. We us the method described in section 3.6 to find this. Hereby the setting is:

$$X(0) = 16 \quad X(T) = 17 \quad T \approx 735.78$$

Our analysis is summarised in figure 3.1. The probability of visiting each city on the travel from München to Saarbrücken is plotted in figure 3.3(a). We notice that Mannheim (state 7) has probability 1 of being visited and Kiel (state 15) has the smallest probability of being visited. This is what we expected since you have to go through Mannheim to go to Saarbrücken no matter where you come from, and Kiel is very far away from the direct path from München to Saarbrücken. We also not that Nürnberg (state 9) has higher probability of being visited that Stuttgart (state 10). This is a result of the distance from München to Nürnberg being shorter than the distance to Stuttgart, and that both ways lead to a direct path to Saarbrücken. Finally note that the probability of Karlsruhe (state 4) being visited is greater than the probability of Stuttgart (state 10) being visited even though the direct path from München through Karlsruhe to Saarbrücken leads through Stuttgart. This is a result of Karlsruhe being bery close to Mannheim and Mannheim having very high probability (1) of being visited. It also tells us that T might be approximated to be too big, as we would expect Karlsruhe to have approximately the same probability of being visited as Stuttgart if the time was scarce. This may also be seen from the fact that e.g. Kiel is has rather high probability of being visited even though it is very far away from the direct path to Saarbrücken.

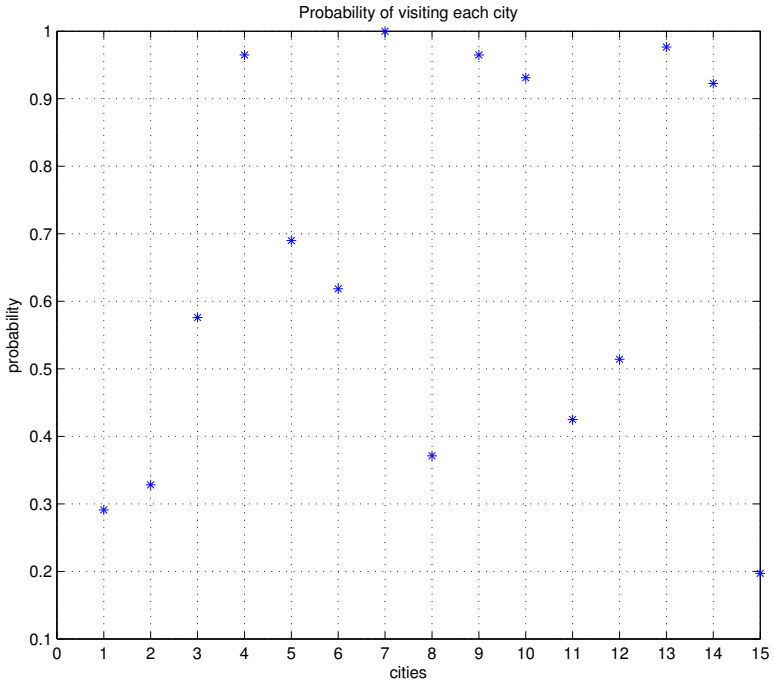
In figure 3.3(b) we note that $X(t)$ spends most time in the starting state and in the ending state reflecting that it is guaranteed to visit these two states and that both have relatively low exit rates. $X(t)$ is also guaranteed to visit Mannheim (state 7), but contrary to München and Saarbrücken , $X(t)$ has very high exit rate in Mannheim as it can leave to four other cities with high rates for at least two of them. A a result of this we see that $X(t)$ does not spend as much time in Mannheim as in München and Saarbrücken. We also not that we expect $X(t)$ to spend very little time in Kiel, Brem and Münster as the probability of visiting these cities are relatively low.

Figure 3.3(c) shows the distribution of the total number of transition on the path from München to Saarbrücken as it would be in 2000 paths from München to Saarbrücken. We see that the most probable number of transitions is in the interval between 52 and 57. This is a very high number of transitions as the number of transitions required on the direct

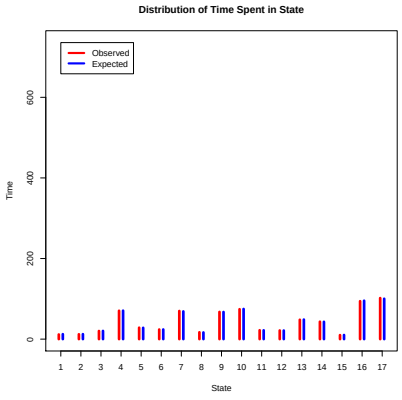
path is 4. This is another evidens that T may be approximated too big. To study this we compute the distribution of the number of transitions between *Mannheim* (state 7) and each of its neighbours. The result of this analysis is summarised in figure 3.2. If the time was scarce we would not expect the number of transitions to Frankfurt, Karlsruhe and Rurnberg to be significant. This is clearly seen not to be the case, and we conclude that T is approximated too big. An analysis of The German Roadmap with smaller T ($T \approx 735.78$) is summarised in figure 3.3. This analysis does not have the flaws of the analysis in figure 3.1.

The red bars in figure 3.3(b) and 3.3(c) show the summary of 2000 sample paths generated using modified rejection sampling[5], and the blue bars show the results from the methods in the previous sections. Both plots show an excellent agreement between the observed and the expected values. We therefor trust our implementations of the methods to be reliable.

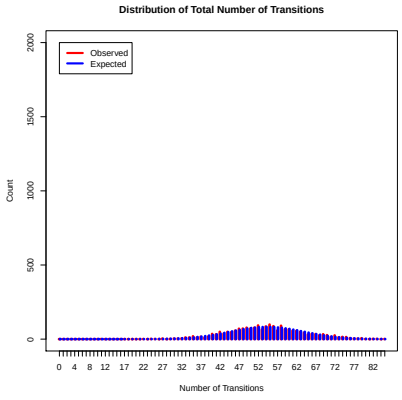
Finally we notice that a good approximation of T is crucial for our analysis. A too genererous approximation will make the effect of conditioning an the CTMC ending in some state at time T insignificant, while a too small approximation will make it too significant.



(a) The probability for each of the cities being visited

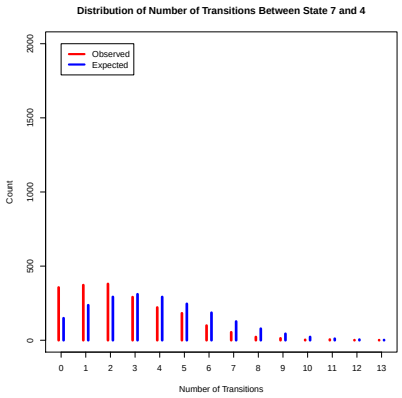


(b) The expected time spent in each city

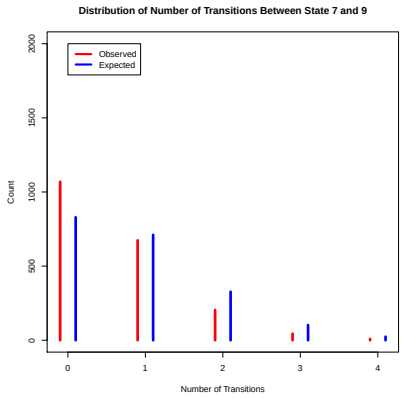


(c) The distribution of the total number of transitions

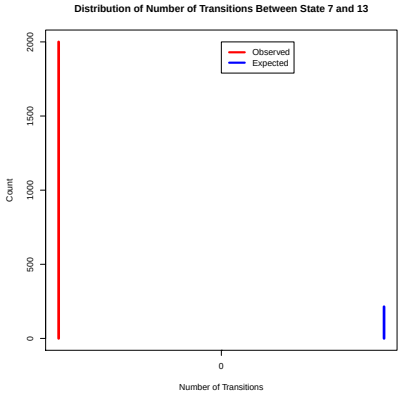
Figure 3.1: Analysis of The German Roadmap



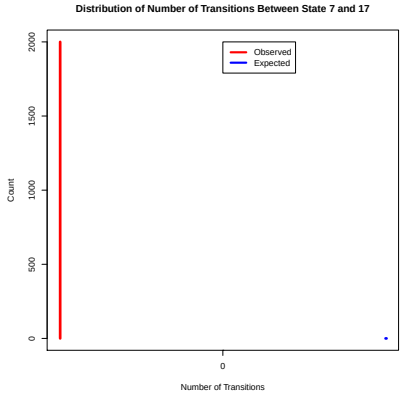
(a) Distribution of transition to Karlsruhe



(b) Distribution of transition to Nürnberg

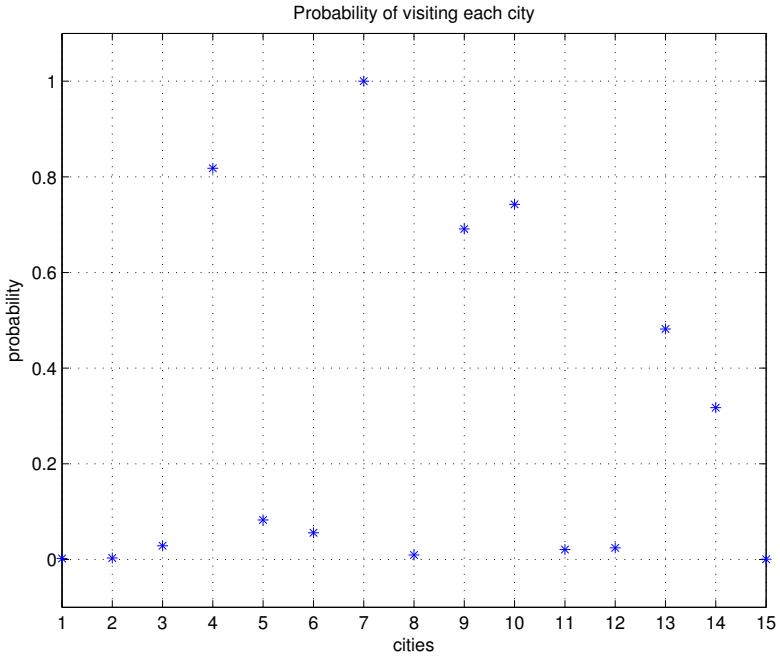


(c) Distribution of transition to Frankfurt

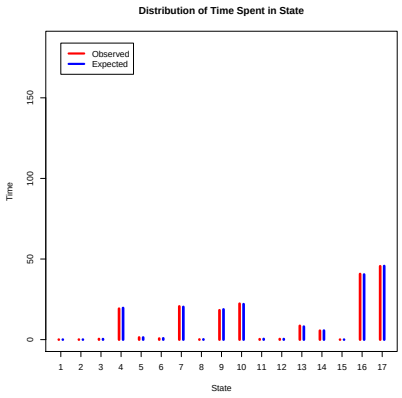


(d) Distribution of transition to Saarbrücken

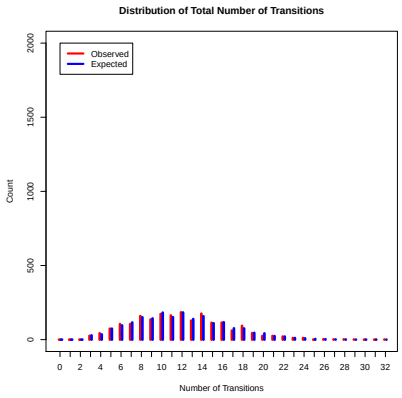
Figure 3.2: The distribution of transitions from Mannheim to each of its neighbours



(a) The probability for each of the cities being visited



(b) The expected time spent in each city



(c) The distribution of the total number of transitions

Figure 3.3: Analysis of The German Roadmap, $T = 735.78/4$

Chapter 4

Adding The Outside State

Our model consists for studying the evolution of protein tertiary structures consists of the starting structure A, the ending structure B and a set of stepping stones, that are believed to be the most relevant protein structures with respect to the evolution of A into B. But we also need to model the (gigantic) set of structures that are not included in the set of stepping stones: it is unlikely that we have included all relevant structures in the set of stepping stones, and even though all relevant structures are included, evolution might just by chance leave the set of stepping stones. In this chapter we describe how we add a state, called *The Outside State*, to the CTMC, modelling the set of all protein structures not included in the set of stepping stones.

4.1 Setting the rates for The Outside State

We add a single state, The Outside State, modelling the set of all states not being A or B or any of the stepping stones. We will call this set *Outside*. To do this, we need to decide how to set the rates for transitions *to* The Outside State from any other state and the rates for transitions *from* The Outside State to any other state. This can be done in many ways, but especially two ways seem intuitive to us. In both cases the rates for leaving The Outside State should be small, since The Outside State models a gigantic set of stepping stones, and transitions to the set of stepping stones from Outside are not more likely than transitions internal in Outside. If we believe that all relevant stepping stones are included and Outside is unlikely to be visited because it is 'distant' or the barrier height to structures in Outside, the rate for transitions *to* The Outside State should be small. On the other hand, if we believe that it is unlikely that we have included all relevant stepping stones, and we have no reason to belive that transitions internal in the set of stepping stones will be prefered to transitions to Outside, the rates for transitions to The Outside State should be high, since Outside models a

gigantic set of structures. We chose this second, more self-critical approach, and let the rate for entering The Outside State from another state $i \in S$ be 10000 times the maximum rate for leaving i and entering any other state. That is:

$$q_{i,out} = 10000 \cdot \max_{j \in S \setminus \{i, s_{out}\}} \{q_{i,j}\} \quad \forall i \in S \setminus \{s_{out}\}.$$

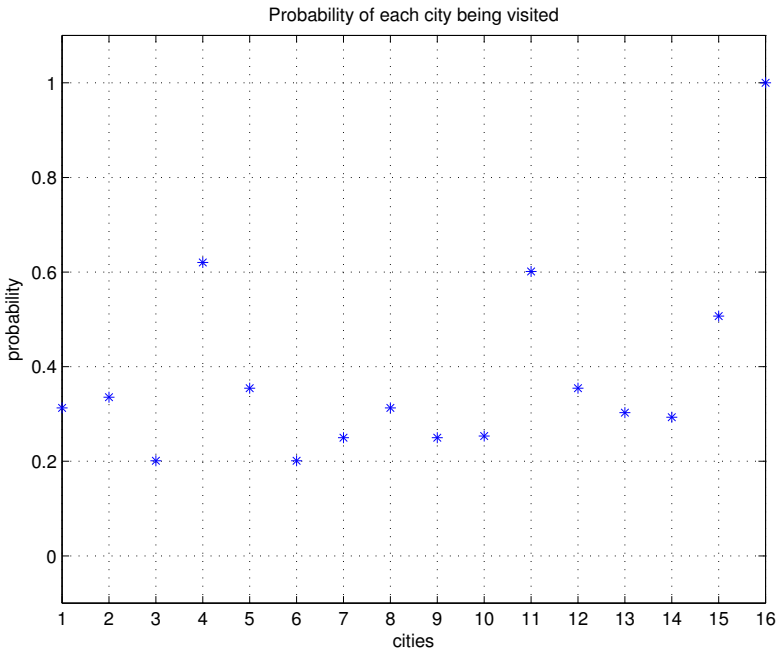
And we let the rate for leaving The Outside State and entering another state i be the minimum rate for entering i from any other state divided by 10:

$$q_{out,i} = \frac{\min_{j \in S \setminus \{i, out\}} \{q_{j,i}\}}{10} \quad \forall i \in S \setminus \{s_{out}\}.$$

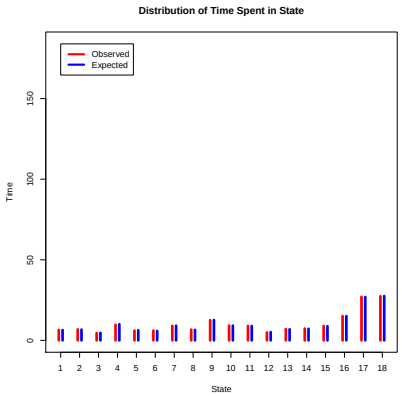
The constants 10000 and 10 are chosen rather arbitrarily.

4.2 Example: The German roadmap

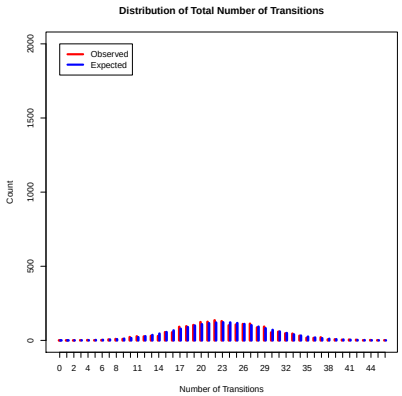
Using the setup described in the previous section, we analyse the german roadmap example. Since the analysis in 3.7 with T set to one fourth of the expected hitting time of Saarbrücken seemed most reliable, the analysis is based on $T \approx 735.78/4$. Figure 4.2 shows a summary of this analysis.



(a) The expected time spent in each city



(b) The expected time spent in each city



(c) The distribution of the total number of transitions

Figure 4.1: Analysis of The German Roadmap including The Outside State, $T = 735.78/4$

Chapter 5

Conclusion and Outlook

In this report we have described methods to evaluate the amount of path structure in a given distance matrix for a set of protein structures, we have developed methods to analyse paths in CTMCs conditioned on starting in a specified state and ending in another specified state, and finally we have described how we add a state to the CTMC derived from the distance matrix, modelling all protein structures not already included in the model.

The methods for statistical analysis of paths in CTMCs conditioned on starting and ending states are highly useful. Together with the methods described in [10] and the sampling methods described in [5] they constitute a comprehensive toolset for analysing path like this.

On the other hand, the methods for analysing distance matrices for path structure need more work. More refined measurements must be developed and tested. Some measurements are listed in section 2.1. Another suggestion would be to study the graph derived from the distance matrix as a flow network with the starting state being the source and the ending state being the sink.

Adding The Outside State is intuitive, but more experiments need to be performed to understand the dynamics added to the model by adding this state and to understand the dynamics of the added parameters.

Finally the methods must be applied to small examples of real protein tertiary structure evolution - preferably examples in which the evolution is already well studied such that our methods can be evaluated.

Bibliography

[1] F. Ball and R.K. Milne. Simple derivations of properties of counting processes associated with Markov renewal processes. *Journal of Applied Probability*, pages 1031–1043, 2005.

[2] P. Guttorp. *Stochastic modeling of scientific data*. Chapman & Hall/CRC, 1995.

[3] A. Hobolth. Recursions for the time spent in a state and the number of transitions between states in end-point conditioned continuous-time Markov chains. 2009.

[4] A. Hobolth and J.L. Jensen. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical applications in genetics and molecular biology*, 4, 2005.

[5] A. Hobolth and E.A. Stone. Efficient simulation from finite-state, continuous-time markov chains with incomplete observations.

[6] V.N. Minin and M.A. Suchard. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology*, 56(3):391–412, 2008.

[7] M.F. Neuts. *Algorithmic probability: a collection of problems*. Chapman & Hall/CRC, 1995.

[8] J.R. Norris. *Markov chains*. Cambridge Univ Pr, 1998.

[9] L. Pachter and B. Sturmfels. *Algebraic statistics for computational biology*. Cambridge Univ Pr, 2005.

[10] A. Sand and T. Fabsic. Paths in Markov Chains with Application to Protein Structure Evoluton.

[11] A. Siepel, K.S. Pollard, and D. Haussler. New methods for detecting lineage-specific selection. *Lecture Notes in Computer Science*, 3909:190–205, 2006.

[12] R. Syski. *Passage times for Markov chains*. Ios Pr Inc, 1992.

[13] W.R. Taylor and A. Aszódi. *Protein geometry, classification, topology and symmetry: a computational analysis of structure.* Taylor & Francis, 2004.

[14] Z. Yang. *Computational molecular evolution.* Oxford University Press, USA, 2006.