Evaluating the Structure Properties of DNA by Using the Spanning Tree Invariant of the Topological Markov Chain Model

Jiunn-I Shieh* Kuei-Jen Lee** Hsiang-Chuan Liu** Hsing-Yuan Tseng**

*Department of Information Science and Applications, Asia University  
**Institute of Bioinformatics, Asia University

Abstract

In this study, four bacteria: Halobacterium sp. NRC-1, Haloarcula marismortui, Campylobacter jejuni, and Streptomyces coelicolor A3(2), were investigated using the spanning tree invariant. This study used a topological Markov chain model to establish a framework to evaluate the level of structural properties of DNA from the complete genomic DNA sequences. Based upon the spanning tree invariant, we can evaluate the randomness and compactness of the complete genomic structures. Briefly, we discuss the correlations between GC contents and the values of spanning tree invariants.

Keywords: topological Markov chain model, spanning tree invariant, DNA sequence, compactness
應用拓樸馬可夫鏈模式展開樹不變量評估 DNA 序列之結構性質

謝俊逸¹ 李桂仁² 劉湘川² 曾杏園²

¹亞洲大學資訊科學與應用學系
²亞洲大學生物資訊研究所

摘要

這項研究是使用一個拓樸馬可夫鏈模式從完全的基因組的去氧核糖核酸（DNA）序列順序中評估其架構性能的標準。在這項研究過程當中，我們選取 4 個細菌︰Halobacterium sp. NRC-1、Haloarcula marismortui、Campylobacter jejuni 及 Streptomyces coelicolor A3（2），並應用拓樸馬可夫鏈模式之展開樹不變量進行序列評估。我們使用一個馬可夫鏈模型，建立一種框架模式，並利用選取的四種細菌的 DNA 序列去進行評估，分析其架構性能的特性及標準。我們相信基於不變量生成樹原理，我們能從生成樹的不變量中評估出在完全的基因組 DNA 序列架構中的隨機性（randomness）和緊密性（compactness），而藉此種方法發現耐鹽性細菌的 DNA 序列特性。

關鍵字：拓樸馬可夫鏈模式、展開樹不變量、DNA 序列、緊密性
Introduction

The complete genomic sequences of four species: Halobacterium sp. NRC-1, Haloarcula marismortui, Campylobacter jejuni, and Streptomyces coelicolor A3(2), were conducted to investigate their structural properties of DNA by using the spanning tree invariant. One of the species in this work is Halobacterium sp. NRC-1, isolated from hypersaline environments. The genome of H. sp. NRC-1 (Ng et al., 2000) has been found with the size of 2,571,010 bp, composed of 3 circular replicons: a large chromosome of 2,014,239 bp and two smaller replicons (191,364 bp and 365,425 bp). These replicons have high GC contents, which are 67.9%, 57.9%, and 59.25, respectively. Proteins predicted by H. sp. NRC-1 are highly acidic. The acidity of proteins is associated with high surface negative charges, which would establish salting-out facilities to enhance solubility and activity in hyperosmotic cytoplasm (Kennedy, Ng, Salzberg, Hood, & DasSarma, 2001). The other species in this work is Haloarcula marismortui, isolated from the Dead Sea. The genome of H. marismortui has the size of 4275-kb (Baliga et al., 2004). The genome has been divided into nine circular replicons with GC contents in the range from 54% to 62% (Baliga et al., 2004). We also investigate two other bacteria: campylobacter jejuni (Parkhill et al., 2000) and Streptomyces coelicolor A3(2) (Bentley et al., 2002). C. jejuni is a microaerophilic, Gram-negative, flagellate, and spiral bacterium, and it has a circular chromosome of 1,641,481 bp with 30.6% GC content. C. jejuni is one of the food-borne pathogen microorganisms, and the main disease that it may cause is diarrhoeal (Parkhill et al., 2000). S. coelicolor, inhabiting in soil, is a filamentous bacterium. It has a linear chromosome with a size of 8,667,507 bp in which the GC content is 88.9%. S. coelicolor is responsible for the production of most natural antibiotics in current use (Bentley et al., 2002). Here we present the results of four complete bacterium genomes calculated using the spanning tree invariant.

In this work, from the ideas of the topological Markov chain and topological invariant we introduce an easy method to evaluate the level of structural properties of DNA. To demonstrate this method, we use four species as mentioned above. This study is outlined as follows: Section 2 briefly reviews a topological Markov chain model and the spanning tree invariant. A framework of applying a topological Markov chain model to evaluate the level of structural properties of DNA is established in Section 3. A brief illustrated example of using a topological Markov chain model in analyzing the level of structural properties for each complete genomic DNA sequence is depicted in Section 4. Finally, conclusions are drawn and summarized in Section 5.

A Topological Markov Chain Model

From a topological point of view, the spectrum which is the set of graph eigenvalues helps to uncover the hidden topological structures of a complex interaction system. It was found that for each eigenvector with a positive eigenvalue, the nucleotides corresponding to absolutely larger components tend form a quasi-clique (i.e. every two of them tend to interact with each other) (Bu et al., 2003), whereas for each eigenvector with a negative eigenvalue, such nucleotides tend to form a quasi-bipartite (i.e. the nucleotides in which two disjoint subsets express high level connectivity between sets rather than within sets) (Bu et al., 2003). For using the spectrum to uncover the hidden topological structures of a complex interaction system, a topological invariant of a dynamical system which is related to the spectrum is introduced. An invariant of a dy-
namical system, the spanning tree invariant in the system, is applied to describe the
dynamical behavior of such a dynamical system.

Given a Markov chain, there is an associated graph, $G = (v, E)$ where $v$ is a
vertex set and $E$ is an edge set, defined as the vertices of $G$ are the states of $S$ with
positive probabilities and the edges of $G$ are the transitions with positive probabilities.
The definition of a Markov chain model (MCM) is as follows (Baliga et al., 2004); (Jarvis, & Shier, 1999); (Kennedy et al., 2001). Let a finite set $S = \{O_i \mid j = 1, 2, \ldots, m\}$
represents the exhaustive and mutually exclusive states of a system at any time. Initially
at time $t_0$, the system may be in any of these states. Let $a_j^{(0)}$ be the absolute probability
that the system is in state $O_j$ at time $t_0$. If the system is Markovian, then define

$$p_{ij} = P\{X_{t_0} = j \mid X_{t_n} = i\}$$  \hspace{1cm} (1)

where $p_{ij}$ is the transition probability of going from state $i$ at time $t_{n-1}$ to state $j$ at
time $t_n$, and assume these probabilities are stationary over time. Moreover, $X_{t_n}$ and
$X_{t_{n-1}}$ in Equation (1) are random variables. The transition probabilities from state $O_i$
to state $O_j$ can be further expressed in a matrix form:

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & \cdots \\
p_{21} & p_{22} & p_{23} & p_{24} & \cdots \\
p_{31} & p_{32} & p_{33} & p_{34} & \cdots \\
p_{41} & p_{42} & p_{43} & p_{44} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$  \hspace{1cm} (2)

where individual $p_{ij}$ values are usually determined empirically, and $\sum_{j} p_{ij} = 1$ for every
$i$ since all entries in the matrix $P$ are nonnegative and the entries in each row must
sum to 1. In summary, a MCM is defined as a transition matrix $P$, assuming irre-
ducible, along with the initial probability $a_j^{(0)}$ associated with the states $O_j$ (Parkhill et
al., 2000). Let $G(P)$ be the directed graph with a vertex set $v = \{1, 2, \ldots, v\}$. If $p_{ij} > 0$ ,
the edge from state $i$ to state $j$ is attached to $G(P)$. In contrast, if $p_{ij} = 0$, no edge is
attached to $G(P)$. Besides, let $E$ denote the resulting edge set for $G(P)$.

Topological Markov chains capture the structure of the possible state transitions,
but the actual numerical values of the initial state probabilities are ignored. Thus, the
state classification is an essential part of the analysis of topological Markov chains
(Jarvis & Shier, 1999; Kennedy et al., 2001). To develop an understanding of the long
term behavior of a Markov chain, the efforts involve determining the structures of a
time-homogeneous finite Markov chain by a transition matrix. Some mathematical
definitions needed in this study are provided as follows (Hillier & Lieberman, 2005):

Definition 1: State $j$ is said to be accessible from state $i$ if $p_{ij}^{(r)} > 0$, for some $r > 0$.
Definition 2: A state $i$ is accessible from state $j$ and state $j$ is accessible from state $i$.

Then two states $i$ and $j$ are said to communicate.

Definition 3: If all states of a Markov chain communicate, the chain is called irreducible.

That is, for every pair of states $i$ and $j$ the property is $p_{ij}^{(k)} > 0$ for some
$k > 0$, where $k$ depends on both $i$ and $j$.

56
Definition 4: A state $j$ is called periodic with period $d$ if $p^{(n)}_{ij} = 0$ unless $n = md$ for some positive integer $m$, and $d$ is the largest integer with this property. That is, the period of state $i$ is $\gcd(n > 0\{p^{(n)}_{ii} > 0\})$, where $\gcd$ is the largest common divisor. A state that is not periodic is called aperiodic.

Definition 5: A state is said to be a recurrent state if the process definitely will return to this state again.

The definition of the spanning tree invariant of a graph is as follows. For any subgraph $H$ of $G$, we define $P$-weight of $H$ to be $W_p = \prod_{e\in H} p(e)$, where $e \in E$ and $p(e) = p_{ij}$ if $e$ goes from state $i$ to state $j$. Let $P$ be a transition matrix, and let $\mathcal{W}$ denote the set of spanning tree for $G(P)$. Define the spanning tree invariant in the system of $P$ to be $\tau(p) = \sum_{S \in \mathcal{W}} W_p(S)$, where $S$ is a spanning tree. A tree in a finite directed graph $G = (v,E)$ rooted at $r \in v$ is a subgraph $T$ of $G$ with the two following properties:

1. Every vertex in $T$ except $r$ has a unique outgoing edge in $T$ and there is no outgoing edge at $r$.

2. From every vertex in $T$ except $r$, there is a unique path ending at $r$ (Kennedy et al., 2001). Figure 1(a) provides a tree rooted at the vertex $a$, while Figure 1(b) represents a tree which is not rooted at the vertex $b$.

A tree is spanning if it contains every state. Let $S_r$ denote the set of spanning trees rooted at $r$, and $S = \bigcup_r S_r$ be the set of all spanning trees in $G$. Since spanning trees are the maximum subgraph without loops, this means that an operation is orthogonal to recurrent behavior in some sense (Kennedy et al., 2001). The larger value of the spanning tree invariant $\tau$ is, the less recurrent behavior is. Assume that the topological Markov chain is connected and aperiodic. Then, $\tau(P)$ can be easily computed from the eigenvalues of $P$. That is, let $\{\lambda_i\}$ be the eigenvalue of $P$. The spanning tree invariant in the system is computed by taking the product of $1 - \lambda_i$ expect $\lambda_i = 1$ for some $i$.

Based on computer generated random numbers, we have generated 60 random sequences and each sequence has $1.0 \times 10^7$ nucleotides, the simulated results have an agreement with the value of the spanning tree invariant, which is near one. In the extreme case (perfect random), we believe the value must be one. Figure 2 provides the simulated values of the spanning tree invariant vs. sample points.
Evaluating the Structure Properties of DNA by Using the Spanning Tree Invariant of the Topological Markov Chain Model

Fig. 2 The simulated values of the spanning tree invariant vs. sample points

So the larger value of the spanning tree invariant away from the value of 1 is, the less the randomness is. That is, there exist some patterns in a complete genomic sequence. Moreover, the larger value of the spanning tree invariant is, the higher level of connectivity between sets rather than within sets is. We have found an empirical rule for the value of spanning tree invariant is greater than 1, and the $G+C$ content is greater than 60% (data not shown). If the value of the spanning tree invariant is less than 1, then the $G+C$ content is less than 60%.

More information inferred from the spanning tree invariant is the compactness of the structure of DNA. Let $\chi = \min_{i} \{1 - \lambda_{i} | \lambda_{i} \neq 1\}$. Then, we have $(\chi)^{n-1} \leq \tau(P)$, where $n$ is the number of vertices in the topological Markov chain. That is, the smaller the value of the spanning tree invariant is, the smaller the value of $X$. Note that the smaller the value of $X$ is, the less compactness of the structure of DNA (Mohar, 1991). So we have that the smaller the value of the spanning tree invariant is, the less compactness of the structure of DNA.

A Framework of Using the Topological Markov Chain to Evaluate the Level of Structural Properties

The purpose of this study is to analyze how to predict the structural properties from complete genomic DNA sequences. The advantage of the spanning tree invariant is that we can obtain structural properties for DNA, but without knowing the whole structures of DNA. DNA macromolecules are composed of nucleotide subunits. Each is composed of a phosphate group, a five-carbon sugar, and a cyclic nitrogen-containing compound called the base. There are just 4 kinds of bases found in DNA. For DNA: adenine (A), guanine (G), thymine (T) and cytosine (C). To compute the invariant of a dynamical system, a three-step procedure based on Lind and Tuncel (Kennedy et al., 2001) is as follows.

Step 1: Given the complete genomic DNA sequences, we want to compute the corresponding transition matrix $P$ for each complete genomic DNA sequence. By computing the pattern frequency of each complete genomic DNA sequence, the transition probabilities can be calculated from the joint probabilities of AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT, where, for instance, AA represents two consecutive nucleotides in a complete genomic DNA sequence. The initial probabilities are


$$P(C) = P(C,A) + P(C,C) + P(C,G) + P(C,T),$$

$$P(G) = P(G,A) + P(G,C) + P(G,G) + P(G,T),$$

and


58
Then calculate the follows:
\[
p_{11} = P(A, A)/P(A), \quad p_{12} = P(C, A)/P(A), \\
p_{13} = P(G, A)/P(A), \quad p_{14} = P(T, A)/P(A), \\
p_{21} = P(A, C)/P(C), \quad p_{22} = P(C, C)/P(C), \\
p_{23} = P(G, C)/P(C), \quad p_{24} = P(T, C)/P(C), \\
p_{31} = P(A, G)/P(G), \quad p_{32} = P(C, G)/P(G), \\
p_{33} = P(G, G)/P(G), \quad p_{34} = P(T, G)/P(G), \\
p_{41} = P(A, T)/P(T), \quad p_{42} = P(C, T)/P(T), \\
p_{43} = P(G, T)/P(T), \quad \text{and} \quad p_{44} = P(T, T)/P(T).
\]

Let \( P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \)

For example, the complete genomic DNA sequence of Streptomyces coelicolor A3(2) in GenBank is under the accession number of AL645882, and the correspondent transition matrix \( P \) is \( P_1 \).

Step 2: For each transition matrix \( P \) for each complete genomic DNA sequence, calculate the eigenvalues of the correspondent transition matrix \( P \). For example, let \( \lambda_1, \lambda_2, \lambda_3, \text{and} \lambda_4 \) be the eigenvalues of \( P_1 \).

Step 3: For each transition matrix \( P \), calculate the spanning tree invariant by taking the product of \( 1 - \lambda_i \) for \( \lambda_i \neq 1 \).

**A Brief Illustrated Example**

To demonstrate the usefulness of a topological Markov chain model, a sample of four complete genomic DNA sequences was studied. The relationship between the value of spanning tree invariant and the level of structural properties is as follows: the larger value of the spanning tree invariant away from the value of 1 is, the less the randomness is. That is, there exist some patterns in a complete genomic sequence. The larger value of the spanning tree invariant is, the higher level of connectivity between sets rather than within sets is. In addition, the smaller the value of the spanning tree invariant is, the less compactness of the structure of DNA.

To analyze the structural properties of DNA, the computation of the spanning tree invariant is needed for each complete genomic DNA sequence. The first step described in Section 3 is to compute the correspondent transition matrix \( P \) given the complete genomic DNA sequence. In this case, four complete genomic DNA sequence were given, and the correspondent transition matrices are depicted in Tables 1 to Table 4.
Evaluating the Structure Properties of DNA by Using the Spanning Tree Invariant of the Topological Markov Chain Model

### Table 1 The transition matrix for *Streptomyces coelicolor A3*(2)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.11395</td>
<td>.41302</td>
<td>.34114</td>
<td>.13189</td>
</tr>
<tr>
<td>C</td>
<td>.1392</td>
<td>.3164</td>
<td>.41054</td>
<td>.13387</td>
</tr>
<tr>
<td>G</td>
<td>.17493</td>
<td>.3481</td>
<td>.31765</td>
<td>.15932</td>
</tr>
<tr>
<td>T</td>
<td>.06965</td>
<td>.45108</td>
<td>.36562</td>
<td>.11366</td>
</tr>
</tbody>
</table>

### Table 2 The transition matrix for *Haloarcula marismortui*

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.18635</td>
<td>.34915</td>
<td>.27258</td>
<td>.19192</td>
</tr>
<tr>
<td>C</td>
<td>.17981</td>
<td>.24662</td>
<td>.40009</td>
<td>.17349</td>
</tr>
<tr>
<td>G</td>
<td>.25679</td>
<td>.27556</td>
<td>.24623</td>
<td>.22142</td>
</tr>
<tr>
<td>T</td>
<td>.12839</td>
<td>.40217</td>
<td>.28322</td>
<td>.18622</td>
</tr>
</tbody>
</table>

### Table 3 The transition matrix for *Campylobacter jejuni*

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.43714</td>
<td>.10846</td>
<td>.1653</td>
<td>.28909</td>
</tr>
<tr>
<td>C</td>
<td>.35964</td>
<td>.16907</td>
<td>.09379</td>
<td>.37749</td>
</tr>
<tr>
<td>G</td>
<td>.31739</td>
<td>.2678</td>
<td>.16941</td>
<td>.2454</td>
</tr>
<tr>
<td>T</td>
<td>.26742</td>
<td>.14154</td>
<td>.1569</td>
<td>.43414</td>
</tr>
</tbody>
</table>

### Table 4 The transition matrix for *Halobacterium sp. NRC-1*

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.1607</td>
<td>.40093</td>
<td>.26831</td>
<td>.17007</td>
</tr>
<tr>
<td>C</td>
<td>.15557</td>
<td>.25779</td>
<td>.44726</td>
<td>.13939</td>
</tr>
<tr>
<td>G</td>
<td>.22578</td>
<td>.30855</td>
<td>.25824</td>
<td>.20743</td>
</tr>
<tr>
<td>T</td>
<td>.09862</td>
<td>.43832</td>
<td>.30157</td>
<td>.16149</td>
</tr>
</tbody>
</table>

The second step is to calculate the eigenvalues for each transition matrix, summarized in Table 5, where \( P_1, P_2, P_3, \) and \( P_4 \) represent the matrices for each complete genomic DNA sequences. The detailed calculations are as follows:

Let \( X = (x_1, x_2, x_3, x_4) \), then solve the equation \( XP = \lambda X \), expressed as follows:

\[
(x_{1}, x_{2}, x_{3}, x_{4}) \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} = \lambda \begin{bmatrix} x_{1} \\ x_{2} \\ x_{3} \\ x_{4} \end{bmatrix}
\]  

(3)

For further computation, Equation (3) becomes

\[
\begin{vmatrix} p_{11} - \lambda & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} - \lambda & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} - \lambda & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} - \lambda \end{vmatrix} = 0
\]  

(4)

Then, Equation (4) can be resolved by MATLAB.
The third step is to compute the spanning tree invariant $\tau$, where $\tau$ is computed by taking the product of $1 - \lambda_i$ for $\lambda_i \neq 1$. The different numerical figures in the spanning tree invariant column depicted in Table 6 represent different meanings.

Table 5 The eigenvalues of the four transition matrices

<table>
<thead>
<tr>
<th></th>
<th>Eigenvalues</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Streptomyces coelicolor A3(2)</td>
<td>1</td>
<td>0.01471</td>
<td>-0.06991</td>
</tr>
<tr>
<td>Haloarcula marismortui</td>
<td>1</td>
<td>0.010129</td>
<td>-0.07236</td>
</tr>
<tr>
<td>Campylobacter jejuni</td>
<td>1</td>
<td>0.162676</td>
<td>0.023547</td>
</tr>
<tr>
<td>Halobacterium sp. NRC-1</td>
<td>1</td>
<td>0.016153</td>
<td>-0.08897</td>
</tr>
</tbody>
</table>

For Campylobacter jejuni, the value of the spanning tree invariant is 0.806331 which is less than one and it means the GC content is smaller than 60%. For Streptomyces coelicolor A3(2), Haloarcula marismortui, and Halobacterium sp. NRC-1, the values are greater than one, which implies the GC content is greater than 60%. Moreover, the value of spanning tree invariant for Campylobacter jejuni is far away from one than those of the others. Thus, there might have some patterns in the complete genome of Campylobacter jejuni. Also, the compactness can be inferred from Table 6.

Table 6 The spanning tree invariant for each transition matrix

<table>
<thead>
<tr>
<th></th>
<th>$\tau(P)$</th>
<th>GCcontent</th>
<th>Salt tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Streptomyces coelicolor A3(2)</td>
<td>1.146553</td>
<td>0.721187</td>
<td>High</td>
</tr>
<tr>
<td>Haloarcula marismortui</td>
<td>1.147916</td>
<td>0.611188</td>
<td>High</td>
</tr>
<tr>
<td>Campylobacter jejuni</td>
<td>0.806331</td>
<td>0.305486</td>
<td>Low</td>
</tr>
<tr>
<td>Halobacterium sp. NRC-1</td>
<td>1.180618</td>
<td>0.659335</td>
<td>High</td>
</tr>
</tbody>
</table>

Conclusions

This study has applied a topological Markov chain model to evaluate the level of structural properties from complete genomic DNA sequences. From the spanning tree invariant we can evaluate the randomness of the complete genomic structures. The larger value of the spanning tree invariant away from the value of 1 is, the less the randomness is. Furthermore, as shown in Table 6, the larger values of the spanning tree suggest that the salt tolerance is higher. Moreover, the smaller the value of the spanning tree is, the less compactness of the structure of DNA. In the future works, the correlations between GC contents and the values of the spanning tree invariant will be proved rigorously.

References


Evaluating the Structure Properties of DNA by Using the Spanning Tree Invariant of the Topological Markov Chain Model


