On the nucleotide distribution in bacterial DNA sequences

M. Sobottka* and A. G. Hart†

*sobottka@mtm.ufsc.br - Departamento de Matemática, Centro de Física e Matemática, Universidade Federal de Santa Catarina, 88040-900 Florianópolis - SC, Brazil.

†ahart@dim.uchile.cl - Centro de Modelamiento Matemático and Departamento de Ingeniería Matemática, UMI 2071 CNRS-UCHILE, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Casilla 170, Correo 3, Santiago, Chile.

It is probable that the distributional structure of DNA sequences arises from the accumulation of many successive stochastic events such as nucleotide deletions, insertions, substitutions and elongations [1, 2, 3, 4, 5, 6, 7]. Although the existence of long-range correlations in non-coding portions of DNA sequences is well established [8, 9, 10, 11], first order Markov chains might well capture aspects of their nucleotide distributions [12]. Here we propose a hidden Markov model based on a coupling of an urn process with a Markov chain to approximate the distributional structure of primitive DNA sequences. Then, by supposing that a bacterial DNA sequence can be derived from uniformly distributed mutations of some primitive DNA, we use the model to explain and predict some distributional properties of bacterial DNA sequences. The distributional properties intrinsic to the model were compared to statistical estimates from 1049 bacterial DNA sequences. In particular, the proposed model provides another possible theoretical explanation for Chargaff's second parity rule for short oligonucleotides [13, 14].

When studying a DNA sequence, the nucleotide frequencies and conditional dinucleotide frequencies are quantities often of interest to researchers. The nucleotide frequency π_i, for $i = A, C, G, T$, denotes the proportion of type i bases in the sequence while the conditional dinucleotide frequency P_{ij}, for $i, j = A, C, G, T$, is the proportion of type i bases that are immediately followed by a type j base. The matrix $P = (P_{ij})_{i=A,C,G,T}$ may be viewed as the transition matrix of a Markov chain, where P_{ij} gives the probability of making a transition from a type i base to a type j base, while the vector $\pi = (\pi_i)_{i=A,C,G,T}$ constitutes the stationary distribution for P, that is, $\pi P = \pi$. Given a two-stranded genome, let P and R be the matrices of conditional dinucleotide frequencies for the primary and complementary strands respectively according to the natural reading order on each strand. We shall denote their respective stationary distributions by π and ρ.

Let $\alpha : \{A, C, G, T\} \to \{A, C, G, T\}$ denote the permutation which maps each nucleotide to its complement ($\alpha(A) = T$, $\alpha(C) = G$, $\alpha(G) = C$ and $\alpha(T) = A$). This map is merely Chargaff's first parity rule, which encapsulates the fact that A always pairs with T and C always pairs with G in any DNA duplex. As a result, we have $\pi_i = \rho_{\alpha(i)}$ and $\pi_i P_{ij} = \rho_{\alpha(j)} R_{\alpha(j)\alpha(i)}$.

From here on, we shall use $(\pi(n), P(n))$ and $(\rho(n), R(n))$ to denote the pairs of mononucleotide and conditional dinucleotide frequencies for 1049 bacterial DNA sequences (see the Supplementary Information) estimated for the primary and complementary strands respectively of the n^{th} bacterium. The 1049 sequences are for complete genomes and were obtained from the GenBank repository.

Our first observation is that, for almost all bacterial DNA sequences, both estimated stationary distributions $\pi(n)$ and $\rho(n)$ obey Chargaff's second parity rule. Chargaff's second parity rule says that the frequency of any short nucleotide sequence $x_1 x_2 \ldots, x_k$ on a strand is approximately equal to the frequency of its reverse complement $\alpha(x_k)\alpha(x_{k-1}) \ldots \alpha(x_1)$ on the same strand. For the stationary distributions, Chargaff's second parity rule says that $\pi_i = \pi_{\alpha(i)}$ and $\rho_i = \rho_{\alpha(i)}$. For P and R, Chargaff's second parity rule means that $\pi_i P_{ij} = \pi_{\alpha(j)} P_{\alpha(j)\alpha(i)}$ and $\rho_i R_{ij} = \rho_{\alpha(j)} R_{\alpha(j)\alpha(i)}$. Thus, for two-stranded DNA, the first and second parity rules together are equivalent to $\pi = \rho$ and $P = R$ elementwise. The analyzed bacteria satisfy $\pi_i(n) \approx \rho_i(n)$. More precisely, the maximum absolute difference $|\pi_i(n) - \rho_i(n)|$ for all the genomes examined is 0.0825 while the average absolute difference is 0.0023. Further, the maximum absolute difference found between any position in a matrix $P(n)$ and the corresponding position in $R(n)$ is less than or equal to 0.1494, while the average of all the absolute differences $|P_{ij}(n) - R_{ij}(n)|$

taken over all 1049 DNA sequences is 0.0025. The matrices P and R for two bacterial genomes, together with their corresponding stationary distributions π and ρ, are given in Table 1.

In addition, bacteria having similar CG-contents seem to have similar matrices $P(n)$. Next fixing $i, j \in \{A, C, G, T\}$, the points $(\pi_j(n), P_{ij}(n))$ for $n = 1, \ldots, 1049$ appear to be systematically distributed around some curve (Supplementary Fig. 1).

To explain the features mentioned above, we consider a model in which DNA is obtained from uniformly distributed mutations on some primitive DNA sequence which is constructed as follows: A new nucleotide is sporadically selected as a candidate for being attached to one of the extremities of some strand of the sequence duplex; The candidate nucleotide then joins the strand according to some fixed probability. We make two assumptions about this construction process: (A1) At each moment, each type of nucleotide has some probability of being selected as a candidate for joining a strand (such probabilities are supposed to be constant throughout the construction of each primitive DNA sequence and could be interpreted as the availability of each nucleotide type in the environment); (A2) The probability of a candidate nucleotide actually being joined to the sequence depends on the type of the candidate nucleotide and the type of the last nucleotide in the sequence (these probabilities are assumed to be constant and the same for all primitive DNA sequences, and could intuitively be thought of as resulting from chemical and other physical properties of the bases).

Let us represent a DNA duplex of length $L = M + N + 1$ by the finite sequence $\binom{x_\ell}{y_{-\ell}}_{-M \le \ell \le N}$, where $x_\ell, y_{-\ell} \in \{A, C, G, T\}$, and M and N are two positive integers. In such a representation, $(x_\ell)_{-M \le \ell \le N}$ corresponds to the primary strand while $(y_\ell)_{-N \le \ell \le M}$ corresponds to the complementary strand. The natural reading order for both strands is in the direction from $-$ to $+$, so that the strands are read in opposite directions within the duplex. We want to describe the construction of a sequence $\binom{x_\ell}{y_{-\ell}}_{-M \le \ell \le N}$ around an initial nucleotide pair $\binom{x_0}{y_0}$.

Observe that $x_\ell = \alpha(y_{-\ell})$. In particular, the initial nucleotide pair $\binom{x_0}{y_0}$ satisfies $x_0 = \alpha(y_0)$. Consequently, knowledge of one strand is sufficient to reconstruct the entire duplex. In a similar way, knowledge of $(x_\ell)_{1 \le \ell \le N}$ and $(y_\ell)_{1 \le \ell \le M}$, together with either x_0 or y_0, also suffices to specify the duplex, in which case it may be represented as $\left(\binom{\alpha(y_{-\ell})}{y_\ell}_{0<\ell\le M}, \binom{x_0}{y_0}, \binom{x_\ell}{\alpha(x_{-\ell})}_{0<\ell\le N}\right)$.

Enterococcus faecalis V583:

P	A	C	G	T
A	0.3900	0.1632	0.1676	0.2791
C	0.3393	0.1947	0.1859	0.2801
G	0.2985	0.2325	0.1950	0.2741
T	0.2235	0.1827	0.2024	0.3914

$\pi = (0.3112, 0.1882, 0.1871, 0.3136)$

R	A	C	G	T
A	0.3914	0.1635	0.1681	0.2770
C	0.3393	0.1950	0.1870	0.2787
G	0.3044	0.2311	0.1947	0.2699
T	0.2253	0.1795	0.2052	0.3900

$\rho = (0.3136, 0.1871, 0.1882, 0.3112)$

Helicobacter pylori J99:

P	A	C	G	T
A	0.4171	0.1330	0.1871	0.2628
C	0.2925	0.2289	0.1842	0.2944
G	0.2613	0.3036	0.2314	0.2036
T	0.2238	0.1717	0.1863	0.4182

$\pi = (0.3033, 0.1970, 0.1949, 0.3048)$

R	A	C	G	T
A	0.4182	0.1302	0.1902	0.2614
C	0.2914	0.2314	0.1861	0.2911
G	0.2658	0.3004	0.2289	0.2049
T	0.2250	0.1680	0.1900	0.4171

$\rho = (0.3048, 0.1949, 0.1970, 0.3033)$

Table 1: The mononucleotide frequencies, π and ρ, and conditional dinucleotide frequencies, P and R, estimated for the primary and complementary strands respectively of two bacterial genomes. For example, having seen a G on the primary strand of Helicobacter pylori J99, the probability that it is followed by a T is 0.2036, while the same probability on the complementary strand is 0.2049.

The model comprises two parts. Let the vector of probabilities $\mu = (\mu_A, \mu_C, \mu_G, \mu_T)$ (assumed in (A1)) be the relative abundance of each nucleotide type in the environment. Candidate nucleotide types are selected at random according to these probabilities. Candidates that are rejected remain in the environment with the possibility of being selected again in the future. Next, a primitive sequence of nucleotides is generated using a matrix of probabilities $\aleph = (a_{ij})_{i,j=A,C,G,T}$ which determines the suitability of candidate nucleotides for extending a strand. Here, a_{ij} represents the probability of a candidate nucleotide of type j being accepted and attached to a nucleotide of type i at the end of the strand. The construction of $(x_\ell)_{0 \le \ell \le N}$ and $(y_\ell)_{0 \le \ell \le M}$ proceeds according to (A1) and (A2) as follows: Suppose the primary strand has been extended from the origin 0 up to the point ℓ, which is a nucleotide of type i. With probability μ_j, a nucleotide of type j is selected as a candidate to join the strand. The probability of the candidate being accepted as the next nucleotide in the sequence, given that it is of type j, is a_{ij}. This process

corresponds to the coupling of an urn process characterized by μ with a Markov chain whose transition matrix is obtained by normalizing the rows of $\aleph$.

Recall that the sequences $(y_\ell)_{-N\le\ell\le 0} = (\alpha(x_{-\ell}))_{-N\le\ell\le 0}$ and $(x_\ell)_{-M\le\ell\le 0} = (\alpha(y_{-\ell}))_{-M\le\ell\le 0}$ do not necessarily have the same distribution as $(x_\ell)_{0\le\ell\le N}$ and $(y_\ell)_{0\le\ell\le M}$. In fact, the observable mononucleotide and conditional dinucleotide frequencies of $(x_\ell)_{0\le\ell\le N}$ and $(y_\ell)_{0\le\ell\le M}$ are given by the matrix $Q = (Q_{ij})_{i,j=A,C,G,T}$, while the observable mononucleotide and conditional dinucleotide frequencies of $(y_\ell)_{-N\le\ell\le 0}$ and $(x_\ell)_{-M\le\ell\le 0}$ are determined by the matrix $\bar{Q} = (\bar{Q}_{ij})_{i,j=A,C,G,T}$, where

$$Q_{ij} = \frac{a_{ij}\mu_j}{\sum_k a_{ik}\mu_k} \quad \text{and} \quad \bar{Q}_{ij} = \frac{\nu_{\alpha(j)}}{\nu_i} Q_{\alpha(j)\alpha(i)}, \tag{1}$$

with $\nu = (\nu_A, \nu_C, \nu_G, \nu_T)$ the stationary distribution of Q. Furthermore, the stationary distribution of $\bar{Q}$ is $\bar{\nu} = (\bar{\nu}_A, \bar{\nu}_C, \bar{\nu}_G, \bar{\nu}_T)$, where $\bar{\nu}_i = \nu_{\alpha(i)}$.

Notice that each of the sequences $(x_\ell)_{-M\le\ell\le N}$ and $(y_\ell)_{-N\le\ell\le M}$ produced by the model is a concatenation of the origin with 2 Markovian sequences, one of length N and the other of length M. In fact, as L increases, then $t := N/L$ tends to the proportion of the primary strand generated from Q while $1-t$ approaches the proportion of the primary strand generated from $\bar{Q}$. These proportions are $1-t$ and t, respectively, in the complementary strand. Thus, if L is large, then the transition matrices and stationary distributions of $(x_\ell)_{-M\le\ell\le N}$ and $(y_\ell)_{-N\le\ell\le M}$ are respectively approximated by

$$\begin{aligned} P_{ij} &= \frac{t\nu_i Q_{ij} + (1-t)\nu_{\alpha(j)} Q_{\alpha(j)\alpha(i)}}{t\nu_i + (1-t)\nu_{\alpha(i)}}, \\ \pi_i &= t\nu_i + (1-t)\nu_{\alpha(i)} \end{aligned} \tag{2}$$

and

$$\begin{aligned} R_{ij} &= \frac{(1-t)\nu_i Q_{ij} + t\nu_{\alpha(j)} Q_{\alpha(j)\alpha(i)}}{(1-t)\nu_i + t\nu_{\alpha(i)}}, \\ \rho_i &= (1-t)\nu_i + t\nu_{\alpha(i)}. \end{aligned} \tag{3}$$

From assumption (A1), if a nucleotide is proposed as a candidate for joining the end of one of the strands, then its complement is automatically a candidate for joining the beginning of the other strand. Therefore, the probability of a nucleotide being selected as candidate for joining one strand is the same as the probability of its complement also being one, that is,

($\mathcal{P}1$) $\mu = \mu(m) = (m, 0.5-m, 0.5-m, m)$, $0 \le m \le 0.5$.

Now, since the matrix $\aleph$ is assumed to be the same for all genomes, any sequence produced by the model is a realization of a Markov chain belonging to a family of Markov chains parameterized by m and t, where $0 < m < 0.5$ and $0 \le t \le 1$.

Further, from (A2) we have that the probability of a candidate nucleotide of type j being accepted to follow a nucleotide of type i at the end of one strand is equal to the probability of a nucleotide of type $\alpha(i)$ preceding a nucleotide of type $\alpha(j)$ at the beginning of the other strand, that is, the matrix $\aleph$ has the form

($\mathcal{P}2$) $a_{ij} = a_{\alpha(j)\alpha(i)}$.

Now, note that if $t = 0.5$, then equations (2) and (3) imply $P_{ij} = \frac{\pi_{\alpha(j)}}{\pi_i} P_{\alpha(j)\alpha(i)}$ and $\pi_i = \pi_{\alpha(i)}$, whence $\pi = \rho$ and $P = R$. Apart from explaining the structure observed in the analyzed bacterial genomes, this could possibly also explain why many genome sequences comply with Chargaff's second parity rule. In fact, if a primitive DNA duplex were constructed according to the model without one strand being favored over the other (that is, N significantly smaller or larger than M), then in general we would have $t \approx 0.5$. Furthermore, if a genome resulted from relatively few mutations distributed uniformly throughout some primitive DNA sequence, then we could hope to see the original distributional structure preserved along large segments of the sequence. In particular, although elongations (repetitions of parts of the sequence separated by arbitrarily long distances) would generate long-range correlations, the resulting sequence would have similar distributions in both the original and the new part, while only a small variation would appear in the position where the new part was concatenated with the original one. We remark that

$\bar{d}_{rs}$	1	2	3	4
1	-	-	-	-
2	0.0735	-	-	-
3	0.0748	0.0169	-	-
4	0.0168	0.0748	0.0743	-

$\bar{D}_{rs}$	1	2	3	4
1	-	0.0050	0.0050	0.0003
2	0.0050	-	0.0003	0.0050
3	0.0050	0.0003	-	0.0051
4	0.0003	0.0050	0.0051	-

Table 2: The entries in the r^{th} row and s^{th} column in the left and right tables above give the means $\bar{d}_{rs} = \frac{1}{1049}\sum_{n=1}^{1049} d_{rs}(n)$ and $\bar{D}_{rs} = \frac{1}{1049}\sum_{n=1}^{1049} D_{rs}(n)$ respectively.

such a model is consistent with Chargaff's second parity rule holding for many kinds of double-stranded DNA, but not for single-stranded RNA/DNA (see [15]).

Since bacterial DNA is generally circular, it is necessary to break the circle at some arbitrary position in order to present it as a linear sequence. This does not affect the computation of the mononucleotide and conditional dinucleotide frequencies, but if the DNA sequence were produced by the proposed model, then each such (linearized) sequence is a translation of some sequence produced by the model. Thus, $t \approx 0.5$ with mutations distributed uniformly throughout the sequence would imply that the nucleotide distributions for the first (second) half of each strand are closer to each other than to the distribution over any other half (see Supplementary Fig. 2). We have examined this property in bacterial genomes. Each strand of the 1049 bacterial DNA sequences was partitioned into two parts of equal length. For the n^{th} bacterium, we use $(\pi^1(n), P^1(n))$ and $(\pi^4(n), P^4(n))$ to denote the mononucleotide and conditional dinucleotide frequencies computed from the first half of the primary and complementary strands, respectively, and we let $(\pi^3(n), P^3(n))$ and $(\pi^2(n), P^2(n))$ be the mononucleotide and conditional dinucleotide frequencies computed from the second half of the primary and complementary strands, respectively. Then, we compared $P^r(n)$ and $P^s(n)$ using two distinct criteria: The Euclidian distance $d_{rs}(n) := \left(\sum_{i,j=A,C,G,T}\left(P^r_{ij}(n) - P^s_{ij}(n)\right)^2\right)^{1/2}$ and the Kullback-Leibler divergence $D_{rs}(n) := \sum_{i,j=A,C,G,T} \pi^r_i(n) P^r_{ij}(n) \log \frac{P^r_{ij}(n)}{P^s_{ij}(n)}$.

Note that while $d_{rs}(n)$ measures how close the transition matrices estimated for the halves designated r and s of the n^{th} DNA duplex are to each other in the Euclidean sense, the quantity $D_{rs}(n)$ is a premetric (it is not symmetric) and measures the relative entropy between the two specified half-strands within the duplex.

In 865 DNA sequences, we found that the first half of each strand is closer to the first half of the other strand than any other part of the sequence duplex in terms of the Euclidean distance or the Kullback-Liebler divergence, while the second part of each strand is closer to the second part of the other strand than to any other part of the duplex. More specifically, 857 DNA sequences exhibit the indicated property for the Euclidean metric, while 851 DNA sequences comply for the Kullback-Leibler divergence. Table 2 displays the average value of each distance measurement between each pair of half-strands.

Using the optimization toolbox in MATLAB, we estimated the matrices $\aleph(n)$, vectors $\mu(n)$ and parameters $t(n)$ corresponding to a matrix which best fits each $P(n)$. Firstly, we carried out a free estimation without the *a priori* assumption of the properties $(\mathcal{P}1)$ and $(\mathcal{P}2)$.

The solution to the optimization problem is sensitive to the value of $t(n)$ chosen to initialize the optimization algorithm. Since previous tests suggest that neither strand is significantly favored over the other during the construction of the sequence, we used $t(n) = 0.5$ as the initial value. Furthermore, the optimization process only permits the estimation of the probability of a nucleotide of type j joining a nucleotide of type i relative to the probability of a nucleotide of type k joining a nucleotide of type i.

We found that the vectors $\mu(n)$ estimated for each matrix $P(n)$ generally satisfy the property $(\mathcal{P}1)$. The mean absolute difference between $\mu_i(n)$ and $\mu_{\alpha(i)}(n)$ was found to be 0.0254, while the median absolute difference was 0.0198. Furthermore, we computed the average

$$\bar{\aleph} = \begin{pmatrix} 0.5762 & 0.4180 & 0.4716 & 0.5326 \\ 0.5440 & 0.4834 & 0.4900 & 0.4778 \\ 0.5194 & 0.5738 & 0.4817 & 0.4235 \\ 0.3631 & 0.5124 & 0.5395 & 0.5759 \end{pmatrix}$$

of all 1049 estimated matrices $\aleph(n)$ (see Supplementary Fig. 3). To assess whether or not we can interpret $\bar{\aleph}$ as having Property $(\mathcal{P}2)$, we estimated $\aleph(n)$, $\mu(n)$ and $t(n)$ again, this time imposing the restrictions stipulated by $(\mathcal{P}1)$

and ($\mathcal{P}2$). In this case we obtained the following average:

$$\bar{\bar{\aleph}} = \begin{pmatrix} 0.8375 & 0.5404 & 0.6315 & 0.7709 \\ 0.7837 & 0.6317 & 0.6936 & 0.6315 \\ 0.7712 & 0.8521 & 0.6317 & 0.5404 \\ 0.6088 & 0.7712 & 0.7837 & 0.8375 \end{pmatrix}.$$

Dividing each line of $\bar{\aleph}$ and $\bar{\bar{\aleph}}$ by their respective sums, we obtained the stochastic matrices $N(\bar{\aleph})$ and $N(\bar{\bar{\aleph}})$, respectively. The Euclidean distance between $N(\bar{\aleph})$ and $N(\bar{\bar{\aleph}})$ is 0.0404, while the Kullback- Leibler divergence between them is 0.0015. Furthermore, we can use $\bar{\aleph}$ and $\bar{\bar{\aleph}}$ with distinct values of $m \in (0,\ 0.5)$ to produce mononucleotide and conditional dinucleotide frequencies $(\bar{\pi}(m), \bar{P}(m))$ and $(\bar{\bar{\pi}}(m), \bar{\bar{P}}(m))$, respectively, according to equations (1) and (2) with $t = 0.5$. Then, plotting the points $(\bar{\pi}_j(m), \bar{P}_{ij}(m))$ and $(\bar{\bar{\pi}}_j(m), \bar{\bar{P}}_{ij}(m))$, we obtained curves that were very close to each other. Furthermore, the majority of the points $(\pi_j(n), P_{ij}(n))$ are distributed around those curves (see Supplementary Fig. 4). This suggests that the construction process defined by $\bar{\aleph}$ is close to one defined using a matrix which satisfies ($\mathcal{P}2$).

Note that the estimation carried out above was blind to the fact that we are describing the construction of the DNA sequence in terms of $(x_\ell)_{0\leq\ell\leq N}$ and $(y_\ell)_{0\leq\ell\leq M}$ instead of $(x_\ell)_{-M\leq\ell\leq 0}$ and $(y_\ell)_{-N\leq\ell\leq 0}$. This should explain why there are variations in the estimated matrices $\aleph(n)$, but they exhibit property ($\mathcal{P}2$) that we observe in $\bar{\aleph}$ on average.

We remark that Chargaff's second parity rule, together with the property that the nucleotide distributions for the first (second) half of each strand are closer to each other than to the distribution over any other half, can be derived from any stochastic construction around an initial nucleotide pair, that is, neither phenomenon requires such construction to be Markovian in nature. In fact, if $(x_\ell)_{0\leq\ell\leq N}$ and $(y_\ell)_{0\leq\ell\leq M}$ were generated using the same non-Markovian law, we would observe those properties. On the other hand, the fact that we found good agreement with Properties ($\mathcal{P}1$) and ($\mathcal{P}2$) in the estimation of $\aleph$ and μ suggests that the Markovian model for the construction of primitive DNA sequences succeeds in capturing gross structure at the level of dinucleotides.

We could also extend the proposed model to the case where μ varies during the construction process. In such a case, we could derive equations like (1), (2) and (3), and use them to explain Chargaff's second parity rule. However, if μ varies with time, then we should not observe $P^1(n)$ closer to $P^4(n)$ than to $P^2(n)$ or $P^3(n)$. This, together with the fact that our estimation returned a $\bar{\aleph}$ which provides a reasonable approximation to the mononucleotide and conditional dinucleotide frequencies $(\pi(n), P(n))$, seems to indicate that μ does not vary much during the construction process.

References

[1] T. H. Jukes and C. R. Cantor. *Evolution of Protein Molecules*. Academy Press, (1969).

[2] M. Kimura. *Journal of Molecular Evolution* **16**, 111–120 (1980).

[3] J. Felsenstein. *Journal of Molecular Evolution* **17**, 368–376 (1981).

[4] S. Tavare and B. W. Giddings. In *Mathematical methods for DNA sequences*, 117–131. CRC Press, Boca Raton (1989).

[5] G. I. Bell and J. Jurka. *Journal of Molecular Evolution* **44**, 414–421 (1997).

[6] P. Baldi and P.-F. Baisnee. *Bioinformatics* **16**(10), 865–889 (2000).

[7] J. C. Whittaker, R. M. Harbord, N. Boxall, I. Mackay, G. Dawson and R. M. Sibly. *Genetics* **164**(2), 781–787 (2003).

[8] Li, W. and Kaneko, K. *EPL (Europhysics Letters)* **17**(7), 655 (1992).

[9] C. K. Peng, S. Buldyrev, A. L. Goldberg, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley. *Nature* **356**, 168–170 (1992).

[10] A. Fukushima, M. Kinouchi, S. Kanaya, Y. Kudo and T. Ikemura. Technical report, Genome Inform, (2000).

[11] Z.-G. Yu, V. V. A. and Wang, B. *Phys. Rev. E* **63**(1), 011903 Dec (2000).

[12] J. Gao, Z.-Y. Xu and L.-T. Zhang. *Physica A: Statistical Mechanics and its Applications* **388**(17), 3475 – 3485 (2009).

[13] E. Chargaff. *Federation Proceedings* **3**(10), 654–659 (1951).

[14] R. Rudner, J.D. Karkas and E. Chargaff. *Proceedings of the National Academy of Sciences of United States of America* **3**(60), 921–922 (1968).

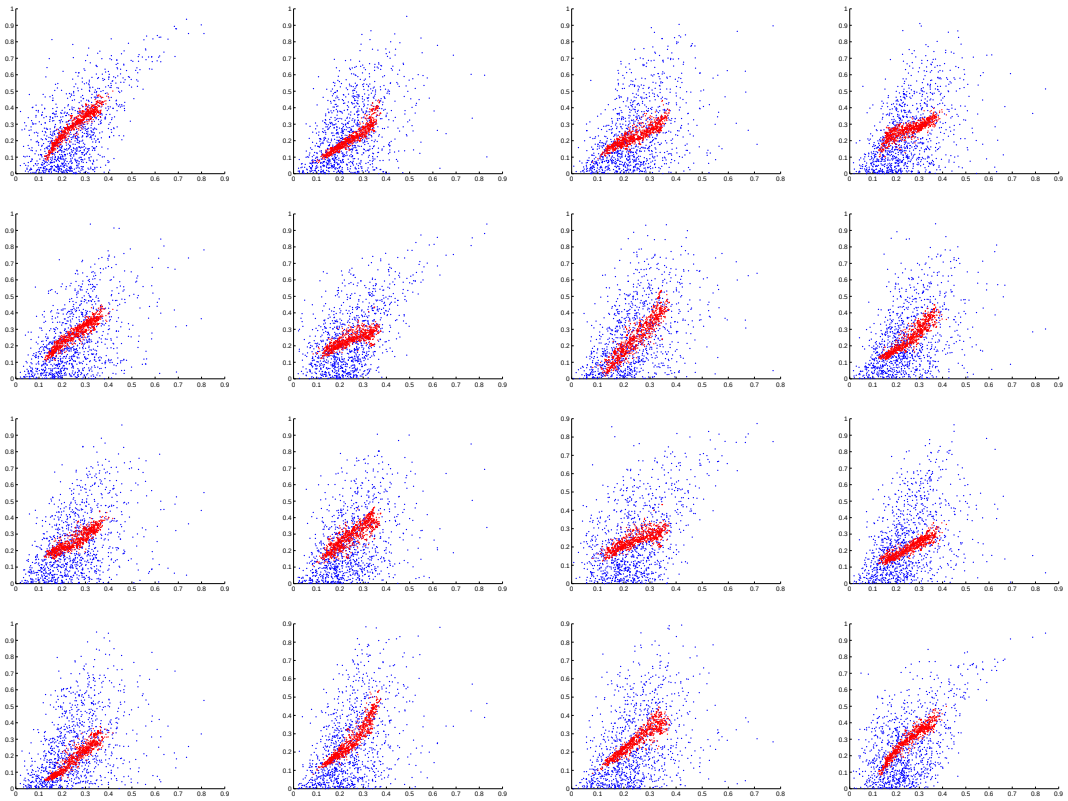

Figure 1: The graph in the i^{th} row and j^{th} column plots the points $(\pi_j(n), P_{ij}(n))$ for $n = 1, \ldots, 1049$. The points $(\pi_j(n), P_{ij}(n))$ for each of the 1049 bacteria are plotted in red. In addition, 1049 stochastic matrices $Q(n)$ were randomly generated and their stationary distributions $\xi(n)$ computed. The points $(\xi_j(n), Q_{ij}(n))$ are shown in blue on the appropriate graph (that is, on the graph in the i^{th} row and j^{th} column).

[15] D. Mitchell and R. Bridge. *Biochemical and Biophysical Research Communications* **340**(1), 90–94 (2006).

ACKNOWLEDGMENTS: This work was supported by the Basal CONICYT program PFB 03 hosted by the Center for Mathematical Modeling (CMM) at the University of Chile. M. Sobottka was supported by CNPq-Brazil grant and by FUNPESQUISA/UFSC. The authors thank the Laboratory of Bioinformatics and Mathematics of the Genome in the CMM for assisting in the acquisition of data and providing advice.

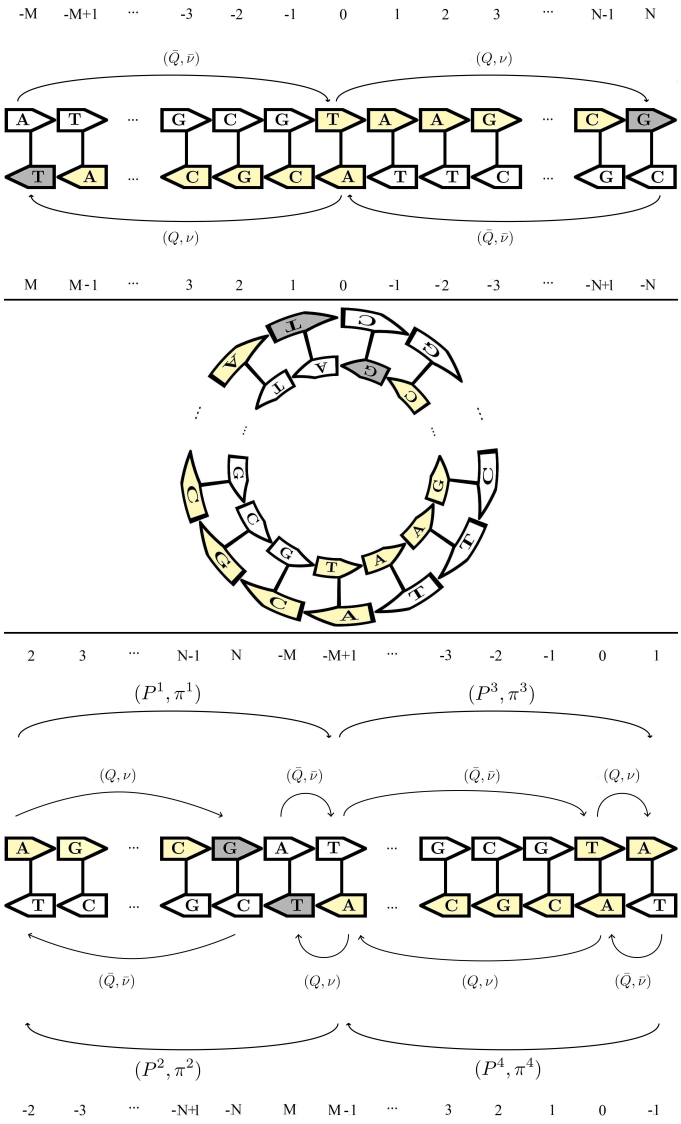

Figure 2: The top diagram shows a schematic representation of a sequence produced by the model around an initial nucleotide pair at position 0. The illustration in the middle shows the circular DNA sequence obtained by joining the extremities of the sequence. In the bottom diagram, a new sequence is obtained by cutting the circular sequence at an arbitrary point. This final sequence is a translation (cyclic shift) of the original one. If $M \approx N$ then the nucleotide distributions in the first (second) half of each strand are very close to each other.

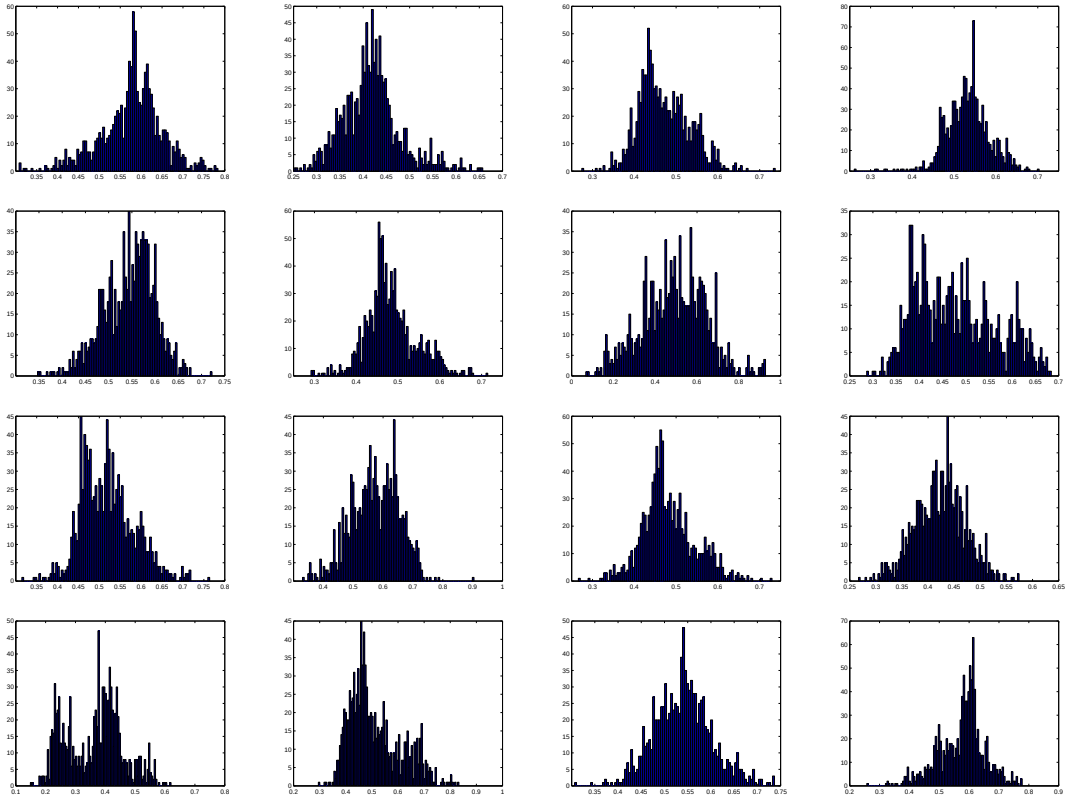

Figure 3: The figure in the i^{th} row and j^{th} column displays a histogram of the values found for $\aleph_{ij}(n)$, $n = 1, \ldots, 1049$.

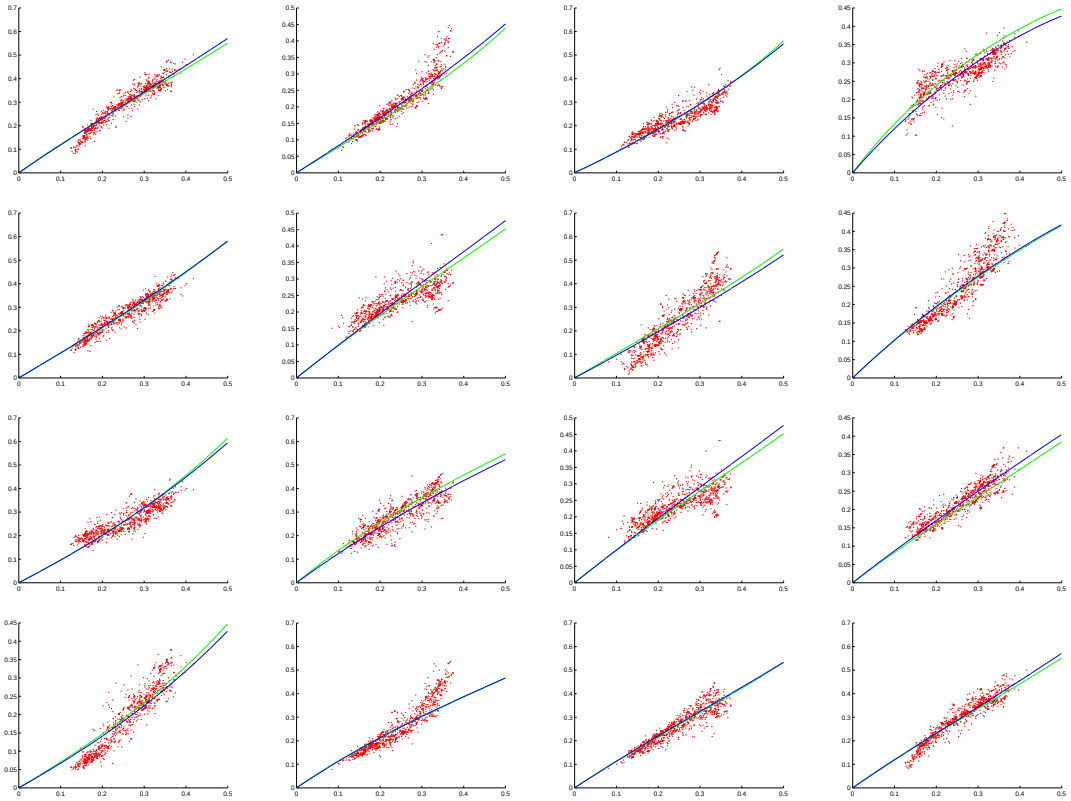

Figure 4: The graph in the i^{th} row and j^{th} column plots the points $(\pi_j(n), P_{ij}(n))$ for the 1049 bacteria in red. Then, fixing $t = 0.5$ and using the matrices $\bar{\aleph}$ and $\bar{\bar{\aleph}}$, we computed the matrices $\bar{P}(m)$ and $\bar{\bar{P}}(m)$ respectively for many distinct values of $\mu(m) = (m, 0.5 - m, 0.5 - m, m)$. The points $(\bar{\pi}_j(m), \bar{P}_{ij}(m))$ (green) and $(\bar{\bar{\pi}}_j(m), \bar{\bar{P}}_{ij}(m))$ (blue) have been plotted on the appropriate graph in the i^{th} row and j^{th} column. Here, $\bar{\pi}(m)$ and $\bar{\bar{\pi}}(m)$ denote the stationary distributions for $\bar{P}(m)$ and $\bar{\bar{P}}(m)$ respectively.