

IDENTIFYING FAMILIES OF COMPOSITIONALLY MODULATED MARKOV CHAINS

Robert H. Baran* and Hanseok Ko**

*Genostat, 4508 Cheltenham Drive, Bethesda, MD 20814 USA

**Department of Electronics and Computer Engineering, Korea University,
Anam-dong, Sungbuk-ku, Seoul 136-701, Korea

Abstract: Consider the family of first order Markov chains produced from a single chain by re-weighting its transition probabilities in a manner that depends only on the next state (not the present state). Given a collection of categorical time series, each of which is a realization of an ergodic Markov chain, we want to test the hypothesis that all chains belong to the same family. An iterative procedure based on the LMS algorithm is found to be expedient for finding least squares estimators of the parameters that join the observations in a family. Numerical experiments show that these estimators behave like maximum likelihood estimators in supporting chi-squared tests of the hypothesis. The methods are used to analyze bacterial genomic data of two kinds. Compositionally modulated Markov chains arise in the configurations of one-dimensional lattice systems where the hypothesis is implied by the assumption of homogeneous interactions between nearest neighbors.

Key words and phrases: Markov chain, maximum likelihood estimation, LMS algorithm, chi-squared tests, genomic data, nucleotide sequence, codon usage, Ising model.

1. Introduction

Feller (1957, Chapt.15) showed that every first order Markov chain is equivalent to an urn model. Specialized to finite chains, the argument proceeds as follows. The urns, indexed by $i \in \{1, \dots, N\}$, contain balls that are labeled with numbers j belonging to the same index set. Each time an urn is visited, a ball is sampled from it with replacement, and the numeric label on the ball is the index of the next urn to be visited. The probability θ_{ij} of drawing j from the i -th urn is called the *composition* of the urn. The composition of the t -th urn in the sequence evidently specifies the distribution of the $t+1$ st and thus the series of indices forms a Markov chain with $N \times N$ transition matrix $\Theta = [\theta_{ij}]$.

Let $\mathbf{C} = [c_{ij}]$ be an $N \times N$ matrix of bounded, positive real constants and let $\mathbf{b} = [b_j]$ be a row vector of N positive real components. The elements of the matrix $\mathbf{\Gamma} = [\gamma_{ij}]$, obtained by normalizing the rows of \mathbf{C} , can be regarded as the transition probabilities of an N -state Markov chain $\{X_t: t = 0, 1, 2, \dots\}$: $\gamma_{ij} = P(X_{t+1} = j | X_t = i)$ for $i, j \in \{1, \dots, N\}$. If the columns of

C are weighted by the components of \mathbf{b} prior to normalizing the rows, then a different Markov chain is obtained, and its transition probabilities are

$$\theta_{ij} = \frac{c_{ij} b_j}{\sum_{j=1}^N c_{ij} b_j}. \quad (1.1)$$

Defining an operator \mathfrak{r} that normalizes the rows of a matrix, we can write this as $\Theta = \mathfrak{r}[C\mathbf{D}(\mathbf{b})]$, using \mathbf{D} (instead of "diag") to create a diagonal matrix from its vector argument. It is easy to see that Θ is unchanged when C is replaced by Γ and \mathbf{b} is replaced by its normalized counterpart $\beta \equiv \mathbf{b}/\sum b_j$. Hence

$$\Theta = \mathfrak{r}[\Gamma\mathbf{D}(\beta)] \quad (1.2)$$

is implied by (1.1). Every row of Θ is "biased" towards the distribution β , and away from the corresponding row of Γ , in the same manner. In the urn model, β_j is a modulation factor that enhances (if $>1/N$) or diminishes (if $<1/N$) the proportion of j -type balls in all of the urns. Γ will be called the progenitor of the family.

Consider the family of chains generated by one stochastic matrix Γ together with all possible distributions β subject to the restriction that the components of Γ and β are *strictly* positive. This restriction guarantees that every family member is ergodic although milder conditions would be sufficient. Given a collection of N -valued sequences, each of which is seen as a realization of an ergodic Markov chain of order 1, we want to test the null hypothesis that all of them belong to the same family.

Before addressing the statistical issues, we can imagine that the length of every sequence in the collection tends to infinity, so that (by the ergodic assumption) any consistent estimator of the transition probabilities yields $\Theta(m)$ for the m -th sequence with arbitrarily small errors. To show that every chain in the collection belongs to the same family, it is sufficient to find a single Γ and a set $B_M = \{\beta(m) : m = 1, 2, \dots, M\}$ that satisfies equation (1.2) for every m . The problem can be posed as a homogeneous system of MN^2 linear equations

$$\theta_{ij}(m) \sum_{k=1}^N \gamma_{ik} \beta_k(m) - \gamma_{ij} \beta_j(m) = 0. \quad (1.3)$$

Suppose this solution exists and consider any $\beta(0)$ not necessarily in B_M but subject to the same restrictions. Then $\mathfrak{r}[\Gamma\mathbf{D}(\beta(0))] = \Theta(0)$ is the transition matrix of a Markov chain belonging to the family of Γ , and every chain in the collection also belongs to the family of $\Theta(0)$, since direct calculation shows that $\Theta(m) = \mathfrak{r}[\Theta(0)\mathbf{D}(\beta^o(m))]$ where

$$\beta_j^o(m) = \frac{\beta_j(0)\beta_j(m)}{\sum_{j=1}^N \beta_j(0)\beta_j(m)} \quad (1.4)$$

Thus the progenitor Γ is not identifiable but the family of Γ is identified by any one of its members.

Let $\mathbf{1}$ denote an $N \times N$ matrix of 1s and consider the family of compositionally modulated Markov chains derived from progenitor $\Gamma = \mathbf{1}/N$. Every row of $\Theta = \mathbf{r}[\mathbf{1D}(\beta)]$ is equal to β and, as the distribution of X_{t+1} does not depend on X_t , the first order Markov chain degenerates to zero order. This degenerate family includes all independent, identically distributed (i.i.d.) sequences of N -valued states.

2. Parameter Estimation Under the Null Hypothesis

Let $\Theta(1), \dots, \Theta(M)$ be the transition matrices of ergodic Markov chains belonging to the family of Γ and having $B_M = \{\beta(m) : m = 1, 2, \dots, M\}$ for their compositional modulation vectors. The corresponding stationary distributions are the row vectors that solve $\pi(m) = \pi(m)\Theta(m)$ subject to $\sum \pi_j(m) = 1$ for each m . Let $[\tilde{\gamma}_{ij}]$ and $[\tilde{\beta}_j(m)]$ be trial values (estimates or guesses) of Γ and $\beta(m)$, respectively, producing

$$\tilde{\theta}_{ij}(m) = \tilde{\gamma}_{ij} \tilde{\beta}_j(m) / \sum_{j=1}^N \tilde{\gamma}_{ij} \tilde{\beta}_j(m) \quad (2.1)$$

for the elements of $\tilde{\Theta}(m)$. The resulting error in the transition probability is

$$e_{ij}(m) = \theta_{ij}(m) - \tilde{\theta}_{ij}(m) \quad (2.2)$$

and the corresponding error in the stationary distribution is

$$d_j(m) = \pi_j(m) - \sum_{i=1}^N \pi_i(m) \tilde{\theta}_{ij}(m). \quad (2.3)$$

The Least Mean Square (LMS) algorithm of Widrow and Stearns (1985, Chapt.6) can be used to drive these errors toward zero by making incremental changes

$$\Delta \tilde{\gamma}_{ij} = \eta_1 \sum_{m=1}^M e_{ij}(m) \quad (2.4)$$

to the elements of the progenitor matrix while simultaneously updating the components of the compositional modulation vectors according to the rule

$$\Delta \tilde{\beta}_j(m) = \eta_2 d_j(m), \quad (2.5)$$

re-normalizing its rows ($\tilde{\gamma}_{ij} \leftarrow \tilde{\gamma}_{ij} / \tilde{\gamma}_{i\cdot}$) after each update, for suitably small (positive) choices of the learning rate parameters η_1 and η_2 . The LMS algorithm, which finds extensive use in adaptive signal processing and control, is equivalent to the "delta rule" for incrementally modifying the weights of a feed-forward neural network designed for pattern recognition

(Warner and Misra (1996)). Iterating equations (2.4) and (2.5) leads downhill on a quadratic surface defined by the Sum Of Squared Errors,

$$\text{SOSE} = \sum_{m=1}^M \sum_{j=1}^N \sum_{i=1}^N e_{ij}^2(m),$$

similar to the method of steepest descent, though generally much slower, but without much sensitivity to the initial conditions. The graph of SOSE versus iteration number is called a learning curve.

Numerical experiments show that this procedure is typically reliable with $\eta_1 = \eta_2 = \eta \leq 0.1$. For example, let $c_{ij} \sim a + U$ and $b_j(m) \sim a + U$ where $a \geq 0$ is an arbitrary constant and each U is an independent random number, uniform on $[0, 1]$, for every (i, j) and m . Next compute $\Theta(m) = \mathbf{r}[\mathbf{CD}(\mathbf{b}(m))]$ for each m and find the stationary distributions by numerical iteration of $\pi(m) = \pi(m)\Theta(m)$, taking π to be a row of $\mathbf{1}/N$ initially. As initial guesses of Γ and B_M , take $\tilde{\gamma}_{ij} = 1/N$ and $\tilde{\beta}_j(m) = \pi_j(m)$. The square root of SOSE declines exponentially with the number of iterations of the LMS algorithm. This is illustrated for $N = 3$, $M = 5$, and $a = 1/2$ in Figure 1 where the downward slope of $1/2 \log(\text{SOSE})$ approximately doubles when η is multiplied by 2. Still larger choices of η may produce faster convergence at the risk of failure to converge which occurs consistently with $\eta = 1$.

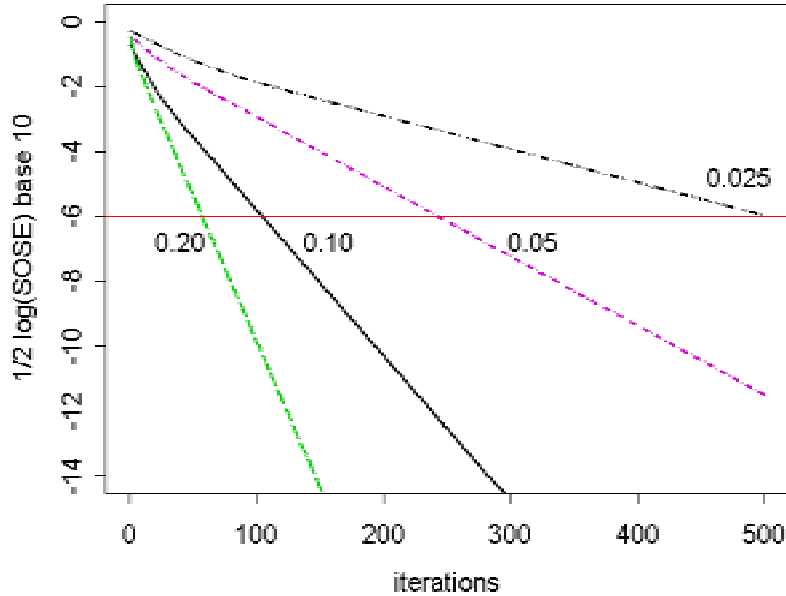


Figure 1. Learning curves for estimation of a family of five Markov chains on three states when the transition probabilities are given ($T = \infty$). Each log-linear curve is produced by application of the LMS algorithm with the indicated value of the learning rate parameters.

Now reconsider the same problem beginning with a collection of finite length sequences. A T -long sequence \mathbf{x} in the collection is scanned from beginning to end, and there

are n_{ij} transitions from state i to state j . The transition count matrix $\mathbf{n} = [n_{ij}]$ is a sufficient statistic for the transition probabilities of a first order Markov model that is fitted to the data by the method of maximum likelihood as $\hat{\theta}_{ij} = n_{ij}/n_{i\cdot}$ with the dot indicating summation over the subscript (Billingsley (1961)). The row and column margins of \mathbf{n} are normalized to give $\hat{\pi}_{i\cdot} = n_{i\cdot}/n_{\cdot\cdot}$ and $\hat{\pi}_{\cdot j} = n_{\cdot j}/n_{\cdot\cdot}$ where $n_{\cdot\cdot} = T - 1$. This procedure applies separately to each $\mathbf{x}(m)$ in the collection.

A tedious derivation might lead from simultaneous equations like (1.3) to the maximum likelihood estimators (MLEs) of $\Theta(1), \dots, \Theta(M)$ subject to constraints that unify them in a single family of compositionally modulated Markov chains. Here we take the more expedient approach of substituting the unconstrained MLEs for the parameters in equations (2.2) and (2.3) and then using the LMS algorithm to estimate Γ and B_M . In other words, the unconstrained MLEs are treated as if they were exact and the constraints of family membership are imposed by subsequent minimization of the SOSE. The formal substitutions $\pi_i \leftarrow \hat{\pi}_{i\cdot}$ and $\pi_j \leftarrow \hat{\pi}_{\cdot j}$ in equation (2.3) are needed because the normalized row and column margins are generally unequal for finite T .

A cross section of the SOSE with respect to two of the parameters can be visualized as a bowl-shaped surface on which the algorithm seeks a minimum. Substituting MLEs for the parameters has the effect of flattening the bowl and hence the learning curve attains a shallower minimum in fewer iterations. Figure 2 shows the learning curves resulting from repeating the experiments described above with finite samples of length $T = 2000$. Each $\Theta(m)$ is computed as before but this time it is used to generate a realization $\mathbf{x}(m)$ from which the unconstrained MLEs are obtained as noted.

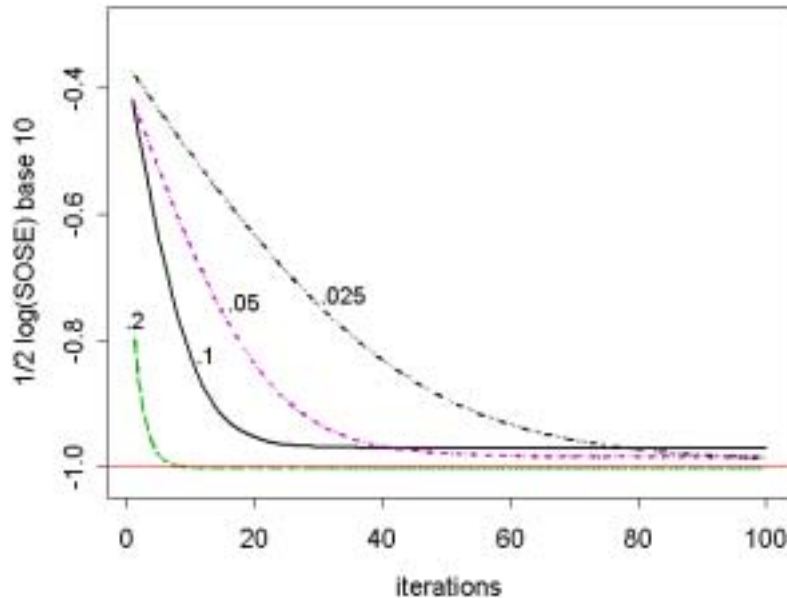


Figure 2. Learning curves for estimation of a family of five Markov chains on three states following the same procedure as for Figure 1 except that each chain is specified by a realization of length $T = 2000$.

3. Parameter Estimation with a Fixed Progenitor

The null hypothesis is negated if just one sequence in the collection is foreign to the family of all the others. Using the LMS algorithm in this situation may result in convergence to $\tilde{\Gamma}$ that does not belong to the family of the true progenitor. This reflects a fundamental difficulty of simultaneous inference in the presence of outliers and there would seem to be no easy, practical, and guaranteed effective solution. Yet there may be settings where a good estimator of Γ is given as the starting point of the analysis and the problem then is to decide which sequences belong to its family. Let G denote this progenitor matrix which may be a Good guess, a consistent Global estimate, or Given by a specific model. In such cases we set $\eta_1 = 0$ but keep $\eta_2 > 0$ so as to estimate the compositional modulation vectors. The model fitting procedure is formally the same as before except that $\eta_1 = 0$ causes $\tilde{\Gamma} = G$ to remain fixed while only \tilde{B}_M changes as the LMS algorithm is iterated with $\eta_2 > 0$.

One such case arises if any sequence (say the first) in the collection is semi-infinite, since the unconstrained MLE of its transition matrix is a consistent estimator of $\Theta(1) = \mathbf{r}[\mathcal{D}(\beta(1))]$, and any member of the family of Γ also belongs to the family of $G = \Theta(1)$.

Because the progenitor of a family of chains defined by (1.2) is not unique, we cannot identify the parameters (C, b) from which the family was derived in (1.1), unless more specific structural assumptions are imposed on them. Examples of such structure are found in one-dimensional Ising models as explained in Section 5 below. When C is specified by a parametric model, it may be possible to obtain a consistent global estimator of $\mathbf{r}[C]$ from all M sequences in the collection.

Unfortunately, the global estimators formed by averaging the data over all M sequences are usually inconsistent in the sense that they converge to the transition probabilities of Markov chains that do not belong to the family of their progenitor. If $\alpha_i \equiv 1/\sum_j \gamma_{ij} \beta_j$ and $\alpha = [\alpha_i]$ is a row vector, then (1.2) implies

$$\Theta = \Gamma \otimes (\alpha \beta) \quad (3.1)$$

where \otimes denotes an element-by-element product of two matrices and $\alpha \beta$ is the outer product of two vectors. Let $\bar{\Theta} = \sum \omega_m \Theta(m)$ for any set $\{\omega_m\}$ of normalized weights. Then

$$\bar{\theta}_{ij} = \sum_m \omega_m \gamma_{ij} \alpha_i(m) \beta_j(m) = \gamma_{ij} \sum_m \omega_m \alpha_i(m) \beta_j(m) \quad (3.2)$$

has the form of (3.1) only if the last sum is separable into the outer product of two vectors. This separability requirement will be satisfied in the family of $\mathbf{r}[\mathbf{1}]$, where $\alpha_i = 1$ for each i , but not in general. Alternatively, let $\bar{\Theta}$ be obtained by forming the weighted sum of joint probabilities $[\pi_i(m) \theta_{ij}(m)]$ and then normalizing the rows. The result

$$\bar{\theta}_{ij} = \gamma_{ij} \frac{\sum_m \omega_m \pi_i(m) \alpha_i(m) \beta_j(m)}{\sum_m \omega_m \pi_i(m)} \quad (3.3)$$

is not even separable when the M chains all belong to the family of $\mathfrak{r}[\mathbf{1}]$, a fact that confounds tests of serial independence in the presence of compositional inhomogeneity. Therefore, when consistent estimators are substituted for the transition probabilities and stationary probabilities, the resulting estimates of $\bar{\Theta}$ are not consistent in the present sense.

Of course, if all sequences in the collection are realizations of the same chain, then (3.2) and (3.3) are both consistent in the present sense, as all members of B_M are then identical. This possibility—that the collection is homogeneous—should not be overlooked in model fitting and hence next section begins by recalling the standard test of that hypothesis.

4. Hypothesis Tests

To test the hypothesis that $\mathbf{x}(m)$ of is realization of the ergodic Markov chain defined by $\Theta = [\theta_{ij}]$, its transition count matrix $\mathbf{n}(m) = [n_{ij}(m)]$ is used to compute the statistic

$$X_0^2(m) = \sum_{i,j=1}^N \frac{[n_{ij}(m) - n_{i\bullet}(m)\theta_{ij}]^2}{n_{i\bullet}(m)\theta_{ij}} = \sum_{i,j=1}^N n_{i\bullet}(m) [\hat{\theta}_{ij}(m) - \theta_{ij}]^2 / \theta_{ij} \quad (4.1)$$

which is asymptotically distributed as $\chi_{N(N-1)}^2$ when the hypothesis is true. The usual proviso is that every expected cell count $n_{i\bullet}\theta_{ij}$ should exceed 4 because smaller counts will tend to inflate the true significance level (Agresti (1990, Chapt.3)). The second equation (4.1) expresses the statistic in terms of the unconstrained MLE, $\hat{\theta}_{ij}(m) = n_{ij}(m)/n_{i\bullet}(m)$. Summing over M independent sequences in the collection, $\Sigma X_0^2(m)$ is chi-squared with $MN(N-1)$ degrees of freedom when all of them are realizations of the chain defined by Θ . When Θ is not given a priori, but estimated by pooling the transition counts, then $\mathfrak{r}[\Sigma \mathbf{n}(m)]$ replaces Θ in (4.1), and the degrees of freedom are reduced to $(M-1)N(N-1)$ because the expected counts for any single m are determined by the others (Anderson and Goodman (1957)).

To test the hypothesis that $\mathbf{x}(m)$ is a realization of a chain belonging to the family of Γ , we obtain a constrained estimator $\tilde{\theta}_{ij}(m)$ by stopping the LMS algorithm after some number of iterations. Substituting it for the true parameters in (4.1) produces a test statistic

$$X_1^2(m) = \sum_{i,j=1}^N n_{i\bullet}(m) \frac{[\hat{\theta}_{ij}(m) - \tilde{\theta}_{ij}(m)]^2}{\tilde{\theta}_{ij}(m)} = \sum_{i,j=1}^N n_{i\bullet}(m) e_{ij}^2(m) / \tilde{\theta}_{ij}(m). \quad (4.2)$$

The second equation shows, with reference to (2.2), that minimizing SOSE may not minimize this statistic, since the latter is a weighted sum of squared errors. Assuming that the scaled errors $(n_{i\bullet})^{1/2} e_{ij}$ have a limiting normal distribution with means zero and variances and covariances depending on $\hat{\theta}_{ij}$ in the same way as obtains for multinomial estimates, this statistic will have a limiting chi-squared distribution; but the degrees of freedom will differ from the standard problem.

Suppose (first) that the matrix $\Gamma = \mathbf{G}$ is fixed and that a compositional modulation vector is fitted to the data by iteration of (2.5) with

$$d_j(m) = \hat{\pi}_{\bullet j}(m) - \sum_{i=1}^N \hat{\pi}_{i\bullet}(m) \tilde{\theta}_{ij}(m) \quad (4.3)$$

where $\tilde{\Theta} = \mathbf{r}[\mathbf{GD}(\tilde{\beta})]$ with the m -dependence suppressed. The bottom margin of the matrix $n_{i\bullet} \tilde{\theta}_{ij}$ converges to the vector of observed counts $n_{\bullet j}$ as $d_j \rightarrow 0$. In the limit, both the row and column margins of the expected count matrix equal the observed counts. Since $\sum_i n_{i\bullet} e_{ij} = \sum_j n_{i\bullet} e_{ij} = 0$, we have $2N$ linear equality constraints, one of which is redundant since the total count is fixed. Thus the asymptotic distribution of (4.2) is chi-squared with $(N-1)^2$ degrees of freedom and hence, for the whole collection of independent sequences, $\Sigma X_1^2(m)$ is chi-squared with $M(N-1)^2$ degrees of freedom under the null hypothesis of common membership in the family of \mathbf{G} .

Suppose (second) that \mathbf{I} is not given but estimated simultaneously with the compositional modulation vectors by iteration of (2.4) and (2.5) as in section 2 above. The marginal differences vanish as before for every m but now the total sum of errors $\sum_m \sum_{i,j} n_{i\bullet} e_{ij}$ tends to zero if the LMS estimator is unbiased. Then $n_{i\bullet} e_{ij}(M)$ is determined by $n_{i\bullet} e_{ij}(1), \dots, n_{i\bullet} e_{ij}(M-1)$ and hence $\Sigma X_1^2(m)$ will have only $(M-1)(N-1)^2$ degrees of freedom under the null hypothesis of common membership in the family of $\tilde{\mathbf{I}}$.

To verify these heuristic calculations, we repeat the same kind of numerical experiment that produced the learning curves of Figure 2, but assess the distributions of $\Sigma X_1^2(m)$ for various choices of N and M . Each experimental trial starts by generating \mathbf{I} and B_M as before. Simulating a collection of realizations of different chains belonging to the family of a randomly generated \mathbf{I} , the parameters are estimated under two models:

- (Model 1:) The correct \mathbf{I} is assumed and the LMS algorithm is iterated 400 times with $(\eta_1, \eta_2) = (0, 0.025)$. The test statistic is $\Sigma X_1^2(m) \equiv Y_1$.
- (Model 2:) $\mathbf{I} = \mathbf{1}/N$ initially and the LMS algorithm is iterated 400 times with $\eta_1 = \eta_2 = 0.025$. The test statistic is $\Sigma X_1^2(m) \equiv Y_2$.

After L independent trials, the means of Y_1 and Y_2 are unbiased estimators of the true degrees of freedom (df) when they are indeed chi-squared. These means $\mu(Y)$ are approximately normal, and a 95% confidence interval for the true degrees of freedom (df) will be contained in $[\mu(Y) \pm \frac{1}{2}]$ when $L \geq 22\text{df}$. Then rounding to the nearest integer will give the true df with a 5% probability of error. Similarly $L = 72\text{df}$ will produce a 99.9% confidence interval (Jernigan and Baran (2003)). Table 1 shows the results of spot-checks that are in good agreement with the predictions. Using the predictions to compute P-values from the statistics Y_1 and Y_2 in the individual trials, the distributions of the P-values are uniform according to the one sample Kolmogorov-Smirnov test performed at the 5% significance level by S-Plus version 4.5.

5. Relationship to Ising Models

It may be possible to identify the parameters (\mathbf{C}, \mathbf{b}) from which a family of compositionally modulated chains is derived in (1.1) when specific structural assumptions are imposed on them. Examples of such structure are found in the one-dimensional versions of

Table 1. Comparing Monte Carlo estimates ("est.") of the degrees of freedom (df) to predictions ("true") for selected values of N and M using T -long realizations of the Markov chains. Each df estimate is the average over L independent trials.

L/df_1	T	N	M	Model 1		Model 2	
				df ₁ true	df ₁ est.	df ₂ true	df ₂ est.
72	100	2	2	2	2.07	1	1.06
72	100	2	3	3	3.01	2	1.94
22	100	2	10	10	9.86	9	9.10
22	200	3	5	20	19.47	16	16.37
10	200	3	10	40	40.92	36	36.97
5	300	4	7	63	63.54	54	53.20
5	300	4	10	90	90.73	81	82.18
5	500	6	8	200	201.66	175	180.92

the Ising model and its many generalizations used by statistical physics to explain cooperative phenomena in nearest neighbor systems. Models of this kind have been adapted to the statistical analysis of spatial data by Besag (1974) and others.

In the simplest case, treated by Kramers and Wannier (1941), each site $t = 0, 1, 2, \dots, T$ in a one-dimensional lattice is occupied by an up/down spin denoted by the random variable $S_t \in \{1, -1\}$. Adjacent sites are linked by bonds of strength J and all sites are influenced by an applied magnetic field of strength H . When $S_0 = s_0$ is held fixed, a spin configuration $(s_1, \dots, s_T) = \mathbf{s} \in \{1, -1\}^T$ has energy

$$E(\mathbf{s}) = - \sum_{t=0}^{T-1} (Js_t s_{t+1} + Hs_{t+1}) \quad (5.1)$$

and its probability is given by a Gibbs distribution, which is

$$p(\mathbf{s}) = Z^{-1} \exp[-E(\mathbf{s})] \quad (5.2)$$

at unit temperature. The denominator Z , called the partition function, is the sum of the Boltzmann exponentials over all 2^T possible configurations, which normalizes the distribution.

As $E(\mathbf{s})$ is found by summing nearest neighbor interactions, so $p(\mathbf{s})$ can be factored as a product of one-step transition probabilities

$$P(S_{t+1} = s' | S_t = s) = Z^{-1/T} \exp(Jss' + Hs'). \quad (5.3)$$

(Here the prime does *not* denote a transpose.) Each configuration can be regarded as a realization of a Markov chain on $N = 2$ states. The transformation $X_t = \frac{1}{2}(3 - S_t)$ converts the

spin variable to a state index $X_t \in \{1, 2\}$. Then $(i, j) = \frac{1}{2}[(3, 3) - (s, s')]$ maps $\{1, -1\}^2$ onto $\{1, 2\}^2$ and the transition probabilities (5.3) can be expressed in terms of coefficients $c_{ij} = \exp(Jss')$ and $b_j = \exp(Hs')$ by normalizing the rows of

$$\mathbf{CD}(\mathbf{b}) = \exp \begin{bmatrix} J + H & -J - H \\ -J + H & J - H \end{bmatrix} \quad (5.4)$$

as prescribed in (1.1). To normalize the i -th row, divide it by $\frac{1}{2}\cosh(J + (-1)^{1+i}H)$. But this step can be omitted in relating the transition probabilities to model parameters as

$$\frac{1}{4} \ln \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}} = J \quad \text{and} \quad \frac{1}{4} \ln \frac{\theta_{11}\theta_{21}}{\theta_{12}\theta_{22}} = H \quad (5.5)$$

from (5.4) since the normalizing factors cancel in both quotients. Thus we can identify the progenitor

$$\mathbf{\Gamma} = \mathbf{r}(\mathbf{C}) = \frac{1}{2\cosh(J)} \exp \begin{bmatrix} J & -J \\ -J & J \end{bmatrix} \quad (5.6)$$

and the compositional modulation vector

$$\boldsymbol{\beta}(H) = \frac{1}{2\cosh(H)} \exp \begin{bmatrix} H & -H \end{bmatrix} \quad (5.7)$$

in the family of chains described by this model.

Consider the proportion of up-oriented sites, which is $\pi_1 = P(S_t = +1)$ when position t is treated as a random variable. The row vector $\boldsymbol{\pi} = [\pi_1 \quad 1 - \pi_1]$ satisfies the equation $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{\Theta}$ for the stationary distribution of the Markov chain. Direct calculation shows that

$$P(S_t = s) = \frac{e^{2sH} + e^{-J}}{2\cosh(2H) + 2e^{-J}}, \quad (5.8)$$

$s = \pm 1$, which coincides with the corresponding component of $\boldsymbol{\beta}(H)$ only if $J = 0$ or $H = 0$. At the other extreme, $\boldsymbol{\pi} \rightarrow \boldsymbol{\beta}(2H)$ as $J \rightarrow \infty$, so the effect of nearest neighbor interaction on the stationary distribution is to amplify the influence of the applied field (as in the physical phenomenon of paramagnetism).

The one-dimensional Ising model can be generalized by imagining a nearest neighbor system in which the each site is N -valued and $J_{ij} \geq 0$ is the strength of interaction between adjacent sites that take values i and j . The applied field is generalized as a state-dependent influence $H_j \geq 0$ that biases the distribution of the states toward a preferred composition. Invoking the Gibbs-Markov equivalence, the distribution of the configurations is a generalization of (5.1) and (5.2), and (5.3) carries over to

$$P(X_{t+1} = x' | X_t = x) \propto \exp(J_{xx'} + H_{x'})$$

for $x, x' \in \{1, \dots, N\}$. Substituting (i, j) for (x, x') , this transition probability (5.9) is θ_{ij} as defined in (1.1) and (1.2), with $\mathbf{C} = \exp(\mathbf{J})$ and $\mathbf{b} = \exp(\mathbf{H})$.

Given a sample of equilibrium configurations $\{\mathbf{x}(m): m = 1, \dots, M\}$ for a single lattice of length T , or a collection of configurations from different lattices in equilibrium at the same temperature, we might hypothesize that the interaction matrix \mathbf{J} is constant throughout. This hypothesis implies that \mathbf{C} is constant and hence that every $\mathbf{x}(m)$ in the sample is a realization of a Markov chain belonging to a single family with progenitor $\mathbf{F} = \mathbf{r}[\exp(\mathbf{J})]$. Note that the degenerate family of $\mathbf{r}[\mathbf{1}]$ arises when $\mathbf{J} = \mathbf{0}$.

6. Applications to Genomic Data

This section describes two applications of the theory to bacterial genomics using public data from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov).

6.1. Ising model of transcription polarity

The circular bacterial chromosome can be viewed as a closed chain of simple protein-coding genes each of which has a polarity (+ or -) given by the sign of its open reading frame (Baran, Ko and Jernigan (2003)). Polarity is synonymous with orientation and the sign is positive if the coding sequence is read from left to right on the published strand, negative if read from right to left on the complementary strand. The chromosome is partitioned into two parts called replichores, designated 1 and 2, which are roughly equal in length, and which differ in respect to whether the leading strand of replication is (1) the published strand or (2) the reverse complement. Polarity is more frequently positive in replichore 1 and more frequently negative in replichore 2. This preference of genes to be coded on the leading strand of replication was called gene-strand bias by Rocha (2003). Adjacent genes also tend to be co-oriented, forming runs (clusters) of like polarity, especially when polarity coincides with the direction of replication.

Baran and Ko (2006) developed an Ising model to simultaneously quantify gene strand bias and clustering. According to this model, genes behave like magnetic dipoles in a one-dimensional lattice and the equilibrium configurations of a replichore, reduced to the analogous spin configuration, follow the statistical description of the previous section. The nearest neighbor interactions are of uniform strength J but the polarity entraining force, corresponding to applied field H , reverses sign at replichore boundaries. Substituting unconstrained MLEs for the transition probabilities in (5.5) produces MLEs of the replichore-specific parameters.

Since the designation of the published strand is arbitrary, it may be expected that gene strand bias is the same in both replichores. If spin configurations $s(1)$ and $s(2)$ in the replichores are independent realizations of oppositely biased Markov chains then $s(1)$ and $-s(2)$ are realizations of the same chain. After converting spins $\{1, -1\}$ to states $\{1, 2\}$, the transition counts are tallied, and reversing the polarity of $s(2)$ carries $n_{ij}(2)$ into $n_{ji}(2)$. Hence the pooled counts $n_{ij}(1) + n_{ji}(2)$ are sufficient for the MLEs of J and $|H|$ under these additional constraints. The model was tested non-parametrically, using the χ^2 -test of homogeneity that was recalled in Section 4, and accepted at the 5% significance level in 9 of 10 cases described by the observed transition counts in Table 2. These 10 cases, identified by species name and NCBI reference sequence number, are (i) *Bacillus subtilis* NC_000964, (ii) *Borrelia burgdorferi* NC_001318, (iii) *Campylobacter jejuni* NC_002163, (iv) *Chlamydomophila*

pneumoniae NC_002491, (v) *Escherichia coli* NC_000913, (vi) *Haemophilus influenzae* NC_000907, (vii) *Helicobacter pylori* NC_000921, (viii) *Mycoplasma pneumoniae* NC_000908, *Staphylococcus aureus* NC_002745, and (x) *Vibrio cholerae* NC_002505, chromosome 1. Table 2 also lists the transition counts and global estimates of J under the model.

Table 2. Replichore-specific transition counts (n_{ij}) and estimates of bond strength (J) in Ising models of ten bacterial chromosomes identified by case number in the text. The right-most column lists P-values from the χ^2 tests of the hypothesis that J is invariant between replichores.

case	Replichore 1				Replichore 2				J	P
	n_{11}	n_{21}	n_{12}	n_{22}	n_{11}	n_{21}	n_{12}	n_{22}		
i	1198	232	232	279	333	286	286	1364	.441	.401
ii	245	51	51	108	114	50	50	242	.590	.772
iii	422	82	82	227	213	84	84	447	.657	.762
iv	263	67	68	203	233	65	64	266	.646	.387
v	810	375	376	713	749	380	379	858	.363	.495
vi	312	75	75	257	393	129	129	460	.622	.242
vii	375	79	79	288	274	81	81	387	.704	.770
viii	226	24	24	20	36	28	28	194	.541	.613
ix	898	133	133	218	195	164	165	907	.533	.007
x	641	191	191	430	446	193	193	676	.514	.674

Homogeneity was rejected only in the case of (ix) *S. aureus*, where inserting the pooled counts in (5.5) gives $J = 0.5327$ and $|H| = 0.3696$ for the MLEs. Rejecting homogeneity, we can suppose that either J or $|H|$ is not the same in the two replichores. The hypothesis of constant J implies that both replichores belong to one family of compositionally modulated chains. Hence we estimate the modulation vectors under the assumption that the fixed progenitor G is given by substituting the MLE of J in (5.6). The modulation vectors are assumed converged after 400 iterations of the LMS algorithm with $\eta_1 = 0$ and $\eta_2 = 0.025$. Noting that $df_1 = M(N-1)^2 = 1$ in this problem, the test statistics $Y_1 = \sum X_1^2(m)$ are converted to P-values $P\{Y_1 > \chi_1^2\}$ in the last column of Table 2. In the 9 cases where homogeneity was accepted at the 5% level, the less restrictive hypothesis of constant J is confirmed. In case (ix), however, the hypothesis is clearly rejected, implying that bond strength differs between replichores of the *S. aureus* chromosome.

6.2. Markovian model of coding sequences

Markovian models serve a variety of purposes in analyzing sequences of nucleotides (A, C, G, and T). For instance, Cortez, Lascano and Becerra (2005) found that a first order Markov model was better for detecting horizontally transferred genes (acquired from another microbial species) than competing methods based on codon usage statistics and G+C percentage. The nucleotide sequence of a simple gene encodes a protein as a series of ordered triples called codons. The genetic code, as a mapping from the $4^3 = 64$ codons to a set of 21 elements (20 amino acids and the translation stop), is multiply degenerate, and there are usually many ways to encode a given sequence of amino acids. Preferential codon usage

varies between species, as does nucleotide composition, since these two sets of variables are algebraically related. A number of investigations, including Wan, Xu, Kleinhofs and Zhou (2004), have concluded that evolutionary pressures acting on the nucleotides cause variation in codon and amino acid usage (instead of the reverse causality).

If xyz is an ordered triple of nucleotides, and the codon counts are given, it is easy to compute the joint distribution $P_{123}(xyz) = p_{xyz}$ and hence the position-specific compositions $P_1(x) = p_{x\bullet\bullet}$, $P_2(y) = p_{\bullet y\bullet}$, and $P_3(z) = p_{\bullet\bullet z}$, which can be averaged to obtain the gross composition. Dinucleotide distributions $P_{12}(xy) = p_{xy\bullet}$ and $P_{23}(yz) = p_{\bullet yz}$ are also functions of the codon counts which have been tabulated for the complete annotated genomes of many bacterial species and strains (Nakamura, Gojobori and Ikemura (2000)).

Calculating mono- and di-nucleotide frequencies from the codon counts is straightforward but the reverse operation--predicting codon frequency from nucleotide composition--requires a stochastic model. The tautological model,

$$P_{123}(xyz) = P_1(x)P_{2|1}(y|x)P_{3|12}(z/xy) = p_{x\bullet\bullet} \left(\frac{p_{xy\bullet}}{p_{x\bullet\bullet}} \right) \left(\frac{p_{xyz}}{p_{xy\bullet}} \right), \quad (6.1)$$

follows the definition of conditional probability. The three terms of the factorization represent $3 + 12 + 48 = 63$ parameters, the same as for the codon distribution. If we can neglect the relatively weak dependence between the first and third positions, assuming

$$P_{3|12}(z/xy) = P_{3|2}(z/y) = \frac{p_{\bullet yz}}{p_{\bullet y\bullet}}, \quad (6.2)$$

then the codon reduces to a Markov chain,

$$P_{123}(xyz) = P_1(x)P_{2|1}(y|x)P_{3|2}(z/y) \quad (6.3)$$

with $3 + 12 + 12 = 27$ parameters. A more drastic simplification asserts the mutual independence of all three positions to obtain the 9-parameter model

$$P_{123}(xyz) = P_1(x)P_2(y)P_3(z) \quad (6.4)$$

of Knight, Freeland, and Landweber (2001) who found that it explained 80% of the variance in the regression of codon frequency on total G+C content in prokaryotes.

It would be possible to reduce the number of parameters in the Markov model of the codons in a single gene if most genes in the given species belong to a few families of compositionally modulated chains. Writing $\pi_x^{(k)}$ for the composition at the k th position ($k = 1, 2, \text{ or } 3$) and $\theta_{xy}^{(k)}$ for the transition probability from position k to $k+1$ ($k = 1 \text{ or } 2$), we have $\pi_x^{(1)}\theta_{xy}^{(1)}\theta_{yz}^{(2)}$ for the (unknown) joint probability of the nucleotides in a single gene under the Markov model (6.3). The unconstrained MLEs of the 27 free parameters can be expressed in terms of two 4×4 matrices of transition counts $\mathbf{n}^{(k)}$ which give the numbers of dinucleotides starting at the superscripted position as

$$\hat{\pi}_i^{(1)} \hat{\theta}_{ij}^{(1)} \hat{\theta}_{jk}^{(2)} = \left(\frac{n_{i\bullet}^{(1)}}{n} \right) \left(\frac{n_{ij}^{(1)}}{n_{i\bullet}^{(1)}} \right) \left(\frac{n_{jk}^{(2)}}{n_{j\bullet}^{(2)}} \right) \quad (6.5)$$

where n is one less than the number of codons in the sequence. We hypothesize the existence of progenitors $\Gamma^{(k)}$ and compositional modulation vectors $\beta^{(k)}$ such that $\Theta^{(k)} = \mathbf{r}[\Gamma^{(k)}\mathbf{D}(\beta^{(k)})]$ and $\pi^{(k+1)} = \pi^{(k)}\Theta^{(k)}$. Then the gene is modeled by the nine locally variable compositional parameters of $\{\pi^{(1)}, \beta^{(1)}, \beta^{(2)}\}$, since the 12 parameters of each progenitor are regarded as global constants that may vary between species but not within them.

Let $m = 1, 2, \dots, M$ index the members of a collection of coding sequences having dinucleotide counts $\mathbf{n}^{(k)}(m)$ starting at position $k \in \{1, 2\}$. We attempt to fit these counts to two families of compositionally modulated Markov chains by application of the LMS algorithm. For initial estimates of the parameters we use the position-specific dinucleotide frequencies $p_{xy\bullet}$ and $p_{\bullet yz}$ from whole genome codon counts, defining $\mathbf{G}^{(1)} = \mathbf{r}[p_{xy\bullet}]$ and $\mathbf{G}^{(2)} = \mathbf{r}[p_{\bullet yz}]$.

The preliminary step, in keeping with last point of Section 3, is to check the possibility that all counts at position k describe the same chain in every gene. This step involves fitting Model 0: $\Theta^{(k)}(m) = \mathbf{G}^{(k)}$ for every m . This model implies $H_0: Y_0 = \Sigma X_0^2(m) \sim \chi^2$ with degrees of freedom $df_0 = MN(N-1) = 12M$ with reference to (4.1); and each individual gene can be used to compute a P-value $P\{X_0^2(m) > \chi_{12}^2\}$ for each k .

If Model 0 is rejected, the next step is to fit Model 1: $\Theta^{(k)}(m) = \mathbf{r}[\mathbf{G}^{(k)}\mathbf{D}(\beta^{(k)}(m))]$. This is accomplished by using the LMS algorithm (with $\eta_1 = 0$) to estimate the M compositional modulation vectors for each k . Recalling (4.2), Model 1 implies $H_1: Y_1 = \Sigma X_1^2(m) \sim \chi^2$ with $df_1 = MN(N-1) = 9M$ degrees of freedom. Each individual gene can be used to compute a P-value $P\{X_1^2(m) > \chi_9^2\}$ for each k .

If Model 1 is rejected, the next step is to fit Model 2: $\Theta^{(k)}(m) = \mathbf{r}[\Gamma^{(k)}\mathbf{D}(\beta^{(k)}(m))]$. This is accomplished by using the by LMS algorithm (with $\eta_1 = \eta_2 > 0$) to estimate one progenitor and M modulation vectors for each k . Model 2 implies $H_2: Y_2 = \Sigma X_1^2(m) \sim \chi^2$ with $df_2 = (M-1)N(N-1) = 9(M-1)$ degrees of freedom. Gene-specific P-values $P\{X_1^2(m) > \chi_9^2\}$ will be liberal since each $X_1^2(m)$ is stochastically less than χ_9^2 .

Table 3 presents results from application of this test procedure to the first ten genes of the *Bacillus anthracis* (Ames strain) sequence, NC_003997. All genes except the third are longer than $3T = 300$ codons and all but two are longer than 1000 codons. The columns headed by H_0 show gene-specific P-values from the 12 degree of freedom tests of Model 0. The last row of the table indicates that this hypothesis is rejected at the 5% level in 7 of 10 genes at each position k . The columns headed by H_1 show P-values from the 9 degree of freedom tests of Model 1 which was fitted to the data via 400 iterations of the LMS algorithm with $\eta_1 = 0$ and $\eta_2 = 0.025$. The fit is satisfactory in just two instances where Model 0 was rejected (namely $m = 4$ and 5 with $k = 1$) as the P-value rises above the 5% level. The columns headed by H_2 list the liberal P-values from fitting Model 2 via 400 iterations of the LMS algorithm with $\eta_1 = \eta_2 = 0.025$. Summing over both positions, Model 2 is rejected in 8 of 20 cases, and all of the summed statistics (Y_0 , Y_1 , and Y_2) are significant at each position.

In concluding that these ten genes do not belong to the same family, we could entertain the possibility that the sample contains one gene that was horizontally transferred from another species. The P-values for the seventh gene are particularly low. Hence Model 2

is re-fitted after deleting it ($m = 7$) from the collection and the results are shown in the columns headed by H_3 . The total fraction of rejections drops from 8/20 to 5/18 but these improved results leave the main conclusion unchanged.

Table 3. P-values for assessing lack of fit in modeling the first 10 genes of the *B. anthracis* genome sequence.

m	$3T$	$k = 1$				$k = 2$			
		H_0	H_1	H_2	H_3	H_0	H_1	H_2	H_3
1	1341	0.005	0.047	0.242	0.414	0.000	0.002	0.084	0.014
2	1139	0.020	0.006	0.038	0.025	0.093	0.708	0.245	0.175
3	213	0.920	0.974	0.972	0.947	0.841	0.652	0.755	0.673
4	1128	0.031	0.098	0.324	0.298	0.002	0.002	0.008	0.015
5	1923	0.024	0.051	0.514	0.808	0.001	0.003	0.093	0.147
6	2472	0.000	0.000	0.001	0.013	0.000	0.003	0.033	0.130
7	1000	0.000	0.000	0.000	-	0.000	0.000	0.000	-
8	1464	0.375	0.392	0.413	0.618	0.000	0.022	0.048	0.109
9	1308	0.149	0.627	0.157	0.130	0.087	0.433	0.643	0.836
10	888	0.000	0.004	0.096	0.141	0.000	0.000	0.010	0.047
Reject at 5%		7	5	3	2	7	7	5	3

7. Compositionally Modulated Markov Chains of Higher Order

The notion of a compositionally modulated Markov chain is easily generalized to higher order. The matrix of transition probabilities describing an N -state Markov chain of order $\nu \geq 1$ has N^ν rows but only N columns as before. With this understanding, we can again write Γ for the progenitor and Θ for the family member, and equation (1.2) needs no formal revision.

Let h denote the $\nu - 1$ states visited before $t - 1$ so that θ_{hij} is an element of Θ and the stationary distribution of any $\nu + 1$ consecutive states is

$$\lim P\{(X_{\nu-1}, \dots, X_{t-1}, X_t) = (h, i, j)\} = \pi_{hij} = \pi_{hi\bullet} \theta_{hij}$$

as $t \rightarrow \infty$. The joint probability of $X_{t-1} = i$ and $X_t = j$ is $\pi_{\bullet ij}$ and the conditional probability is

$$q_{ij} \equiv P\{X_t = j \mid X_{t-1} = i\} = \frac{\pi_{\bullet ij}}{\pi_{\bullet i\bullet}} = \left(\sum_h \pi_{hi\bullet} \frac{\gamma_{hij} \beta_j}{\sum_{j=1}^N \gamma_{hij} \beta_j} \right) / \sum_h \pi_{hi\bullet}$$

after expressing θ_{hij} in terms of the parameters (Γ, β) and summing over j first in the denominator. These conditional probabilities can be collected in an $N \times N$ array $Q = [q_{ij}]$ which might be taken as the transition matrix a first order Markov chain. This first order chain could serve as a crude approximation of the higher order process. Under what conditions will the higher order family of Γ produce first order transition matrices Q that belong to a single

family? Viewing the last equation in light of (3.1), it is necessary to produce a factorization of the form

$$\sum_h \frac{\pi_{hi} \cdot \gamma_{hij}}{\sum_{j=1}^N \gamma_{hij} \beta_j} = a_i g_{ij}$$

for every i and j , which is trivially satisfied when $\gamma_{hij} = (1/N)g_{ij}$ does not depend on h . But this contradicts the assumption of higher order dependence. It seems that the desired conditions are at least highly restrictive.

Our attempt to model dinucleotide frequencies in Section 6.2 may have been doomed for this reason and the lack of fit can be attributed to the false assumption of independence in (6.2). A better model of the codons in a gene might be formulated as a Markov field of three interacting lattice sites with energy function

$$-E(xyz) = J_{xy}^{(1)} + J_{yz}^{(2)} + J_{zx}^{(3)} + H_x^{(1)} + H_y^{(2)} + H_z^{(3)}$$

analogous to (5.1). The $3(4^2 + 4) = 60$ parameters would be effectively reduced to 12 if each 4×4 interaction matrix $\mathbf{J}^{(k)}$ is globally constant. Equation (6.2) was tantamount to assuming $\mathbf{J}^{(3)} = \mathbf{0}$. Correcting this mistake, the distribution of the state of one site will be determined by the states of the other two, and the homogeneity of pair-wise interactions within genomes could be tested by extension the present methods to the second order problem.

References

- Agresti, A. (1990). *Categorical Data Analysis*. John Wiley, New York.
- Anderson, T.W. and Goodman, L.A. (1957). Statistical inference about Markov chains. *Ann. Math. Stat.* 28, 89-110.
- Baran, R.H. and Ko, H. (2006). An Ising model of transcription polarity in bacterial chromosomes. *Physica A* 362, 403-422.
- Baran R.H., Ko, H. and Jernigan, R.W. (2003). Methods for comparing sources of strand compositional asymmetry in microbial chromosomes. *DNA Research* 10, 85-95.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Royal Stat. Soc. B* 36, 192-225.
- Billingsley, P. (1961). Statistical methods in Markov chains. *Ann. Math. Stat.* 32, 488-497.
- Cortes, D.Q., Lazcano, A. and Becerra, A. (2005). Comparative analysis of methodologies for the detection of horizontally transferred genes. *In Silico Biology* 5.
- Nakamura, Y., Gojobori, T. and Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases. *DNA Research* 28, 292.
- Feller, W. (1957). *Intro. to Probability Theory and Its Applications*, Vol. 1, 2nd edn, John Wiley, New York.
- Jernigan, R.W. and Baran, R.H., (2003). Testing lumpability in Markov chains. *Statistics & Probability Ltrs.* 64, 17-23.
- Knight, R.D., Freeland, S.J. and Landweber, L.F. (2001). A simple model based on mutation and selection explains trends in codon and amino acid usage and GC composition within and across genomes. *Genome Biology* 2.

- Kramers H.A. and Wannier, G.H. (1941). Statistics of the two-dimensional ferromagnet. *Phys. Rev.* 60, 252-267.
- Rocha E.P.C. (2003). Essentiality drives gene-strand bias in bacteria. *Nature Genetics* 34.
- Wan, X-F., Xu, D., Kleinhofs, A. and Zhou, J. (2004). Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evolutionary Biology* 4, 19.
- Warner, B. and Misra, M. (1996). Understanding neural networks as statistical tools, *American Statistician* 50, 284-293.
- Widrow, B. and Stearns, S.D. (1985). *Adaptive Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ.