Statistica Sinica **16**(2006), 29-43

STATIONARITY AND MIXING PROPERTIES OF REPLICATING CHARACTER STRINGS

Probal Chaudhuri and Amites Dasgupta

Indian Statistical Institute

Abstract: In this article, some models for random replication of character strings are considered that involve random mutations, deletions and insertions of characters. We derive some sufficient conditions on the replication process and the ancestor chain that ensure stationarity and mixing properties of the replicated chain. We also give examples of replication processes which lead to descendant chains not having any mixing properties even if the ancestor chain is i.i.d. in nature. Stationarity and mixing properties are two properties of dependent processes that are of fundamental importance and well studied in the literature. These properties are quite useful in generalizing many asymptotic results for i.i.d. processes to dependent processes and, in many situations, they are useful in justifying statistical estimation and inference based on dependent data. The presence of random deletions and insertions makes our stochastic replication model considerably different from simpler models that involve only mutations, and it leads to some interesting theoretical problems.

Key words and phrases: α-mixing property, exchangeable processes, hidden Markov processes, Markov chains, stationary processes.

1. Introduction

Suppose that we have observed a random character string $\{Y_1, Y_2, \ldots\}$, where $Y_i \in \mathcal{A}$, a finite alphabet of symbols ($= \{\alpha_1, \ldots, \alpha_k\}$, say). We assume that this observed sequence is generated by a random replication process operating on a (possibly unobserved) ancestor string $\{X_1, X_2, \ldots\}$ of characters from the same alphabet. Such replication of character strings arises in molecular evolution of nucleic acids and protein sequences. Several stochastic models for biological sequences (i.e., DNA, RNA and protein sequences) have been considered in the literature, and their biological significance has been investigated by several authors (see e.g., Churchill (1989), Durbin, Eddy, Krogh and Mitchison (1998), Ewens and Grant (2001), Krogh, Brown, Mian, Solander and Hausler (1994), Pevzner, Borodovsky and Mironov (1989a, 1989b), Schbath, Prum and Turckheim (1995) and Waterman (1995)). Among those models, Markov and hidden Markov models are possibly the most extensively studied for biological sequences. It is well known that a stationary Markov chain on a finite state space satisfies the α-mixing property with a geometric rate of decay (see e.g., Billingsley (1986,

1999)) if the chain is irreducible and aperiodic in nature. On the other hand, for a hidden Markov process, where both the output process and the underlying Markov process have finite state spaces, the output chain can be viewed as a function of a Markov chain on a finite state space which can be taken to be the Cartesian product of the state spaces of the output chain and the underlying Markov chain (see e.g., Chaudhuri and Dasgupta (2005) for more details and related results). Consequently, the stationarity and the mixing properties of the output process will be a simple consequence of those properties of the Markov chain on that product state space, and those can be ensured by appropriate conditions on the output distributions and the distribution of the hidden chain.

Let us consider a situation where $\{Y_1, Y_2, Y_3, \ldots\}$ is obtained by replicating $\{X_1, X_2, \ldots\}$, and the replication process is subject to random *mutations*, *insertions* and *deletions* of characters at various positions. The main question that we intend to address in this article is whether the stationarity and the mixing properties hold for the descendant Y-sequence when similar properties are known to hold for the ancestor X-sequence. Clearly, the answer to this question will depend on the nature of the random replication process. We derive some sufficient conditions for the replication process that will ensure a positive answer, and present some interesting examples to show that the answer might be negative in some simple yet important special cases.

If we view random replication as a stochastic transformation on the space of sequences equipped with a probability measure governing the probability law for the ancestor chain, the question raised in the preceding paragraph translates into the problem of invariance or non-invariance of stationarity and mixing properties under such random transformations. We have specifically chosen to study stationarity and mixing properties in this paper because these are of fundamental importance for the asymptotic results related to dependent sequences. As is well known many asymptotic results, such as laws of large numbers and central limit thorems for averages related to simple i.i.d. sequences, can be generalized to stationary and mixing processes. In particular, as has been discussed in detail in Waterman (1995), empirical distributions of words of finite length obtained from finite state space stationary processes satisfying appropriate mixing conditions lead to asymptotically consistent estimates of finite dimensional probability distributions of that process. This is of critical importance in justifying various asymptotic statistical estimation and inference techniques when they are applied to dependent sequences.

Surprisingly, though the statements of the main theorems that we formulate are simple and intuitive, the proofs require a combination of ideas related to properties of finite Markov chains, including their large deviation properties. The proofs also show the special role played by the Markov nature of the replication process, and we have not been able to generalize the results without it. The organization of the paper is as follows. Section 2 introduces the necessary

notation for the description of the replication model. The main mathematical challenge lies in dealing appropriately with the randomness of positions introduced by random insertions or deletions. Section 3 contains the main theorems that give some sufficient conditions for the invariance of stationarity and mixing properties. Section 3 also presents an example of an exchangeable replication process to demonstrate lack of invariance of mixing properties in a special situation. The concluding section indicates some unresolved issues for further research related to our proposed model for stochastic replication of character strings. All proofs are in the Appendix.

2. A Model for Random Replication

We begin by describing a model for the random replication or copying mechanism that operates on the (possibly unobserved) ancestor sequence of the X's to produce the observed sequence of the Y's, using a stochastic process $\{Z_1, Z_2, Z_3, \ldots\}$. We assume that the Z-process has state space $\{D, I, M\}$. In state D, the replication process Z will *delete* the character in the X-sequence that it encounters. In state I, the process Z will *insert* one letter from $\mathcal{A}$ into a position in the X-sequence that it encounters by randomly selecting that letter from $\mathcal{A}$ according to the probability distribution $P(\text{“Inserted letter is } \alpha_i\text{”}) = \pi_i$ $(1 \le i \le k, \pi_i > 0, \sum_{i=1}^{k} \pi_i = 1)$ *that depends neither on the X-sequence nor on the Z-process.* In state M, the process Z will *mutate* the character in the X-sequence that it encounters according to a $k \times k$ transition probability matrix $((\theta_{i,j}))$, *which is assumed to be independent of the Z-process.* Here for $1 \le i, j \le k$, $\theta_{i,j} =$ the conditional probability $P(\text{“The letter is mutated into } \alpha_j \text{ in the descendent chain”} \mid \text{“The letter was } \alpha_i \text{ in the ancestor chain”})$, and $\sum_{j=1}^{k} \theta_{ij} = 1$ for all $1 \le i \le k$. We assume that $\theta_{ii} > 0$, so that even if the Z-process is in state M, the corresponding character in the X-sequence may remain unchanged with a positive probability.

Let T_i be the time of the ith visit of the Z-process to the state I or the state M. To keep track of the index (i.e., the position of a letter) in the X-sequence on which Z_i operates, let

$$\chi_i = \begin{cases} 1, & \text{if } Z_i \in \{M, D\}, \\ 0, & \text{if } Z_i = I. \end{cases} \tag{1}$$

Also, let us define $S_n = \sum_{k=1}^{n} \chi_k$. Then a letter in the observed Y-chain, which is obtained by copying the ancestor X-chain using the Z-process, can be written as

$$Y_i = \begin{cases} \alpha_s \text{ with probility } \pi_s, & \text{if } Z_{T_i} = I\ , \\ \alpha_s \text{ with probability } \theta_{rs}, & \text{if } Z_{T_i} = M \text{ and } X_{S_{T_i}} = \alpha_r. \end{cases} \tag{2}$$

Remark. In the above notation if $Z_1 = Z_2 = \cdots = Z_m = I$, we have $S_{T_1} = S_{T_2} = \cdots = S_{T_m} = 0$. Since the X-chain starts from X_1, we need to

define $X_{S_{T_1}} = \cdots = X_{S_{T_m}} = X_0$ by introducing X_0. This X_0 is never operated upon by the Z-sequence because the effect of insertions at the beginning is to shift the $\{X_1, X_2, \ldots\}$ part to the right, and as soon as we have the first D or M according to the above notation, that acts on X_1. Thus we consider the $\{X_0, X_1, \ldots\}$ sequence, whose $\{X_1, X_2, \ldots\}$ part is to be replicated by the $\{Z_1, Z_2, \ldots\}$. Without loss of generality for all our subsequent mathematical results, we assume that $\{X_0, X_1, \ldots\}$ is a stationary sequence with the same finite dimensional distributions as the stationary sequence $\{X_1, X_2, \ldots\}$.

3. Invariance of Stationarity and Mixing Properties

Throughout this section the Z-process will be same as before, and we now assume that it is an irreducible, aperiodic Markov chain with stationary transition probabilities. We also assume that the Z-sequence and the X-sequence are independent.

Theorem 3.1. *If the X-sequence and the sequence $\{Z_{T_1}, Z_{T_2}, Z_{T_3}, \ldots, \}$ are both stationary, the sequence $\{Y_2, Y_3, Y_4, \ldots, \}$ is also stationary.*

Note that the stationarity of the Z-process does not necessarily imply that the Z_T-process will be stationary, though it implies that the Z_T-process will be Markov with stationary transition probabilities and vice versa. However, if the Z-process is i.i.d., so is the Z_T-process. It is important to note that an ergodic Z-process does not imply an ergodic Z_T-process. For this reason, our next theorem has been stated with an extra assumption. For a Markov chain, such an assumption is satisfied if the entries of the transition probability matrix are all positive, which in practical problems is not too restrictive a condition.

We now address the mixing properties of the ancestor X-chain and the descendant Y-chain produced by the replication process. We follow the notation and the terminology used in Billingsley (1986, 1999) and, in our next theorem, we establish that under appropriate regularity conditions on the Z-process, α-mixing of the ancestor chain is passed down to the descendant chain.

Theorem 3.2. *Let the X-sequence be stationary and have the α-mixing property with geometric decay, and let the Z-sequence be Markov such that $\{Z_{T_1}, Z_{T_2}, \ldots\}$ is stationary with an irreducible and aperiodic transition matrix. Then the Y-sequence is also stationary, and it possesses the α-mixing property with geometric decay.*

3.1. Exchangeable replication process and lack of invariance of mixing properties

Consider an exchangeable Z-sequence, i.e., a sequence that is distributionally invariant under any finite permutation of the Z_i's. It is a well known result (see e.g., Feller (1971)) that any such exchangeable process with a finite state

space is a mixture of i.i.d. processes with the same state space. Clearly, an exchangeable process is stationary in nature. However, it is fairly easy to see that an exchangeable sequence with a finite state space will not satisfy the α-mixing property, nor any other standard mixing properties, unless the sequence is actually i.i.d..

Suppose now that the ancestor X-sequence is i.i.d.. In this case, the descendant Y-sequence will be an exchangeable sequence if the Z-sequence is. This follows easily from the fact that a random sequence with a finite state space is exchangeable in nature if and only if it is conditionally an i.i.d. sequences with the same state space given the tail σ-field of the exchangeable process (see e.g. Feller (1971)). Consequently, given the tail σ-field of the Z-process, the X-, the Z- and hence the Y-processes will be conditionally i.i.d.. This implies that, unconditionally, the Y-sequence will be exchangeable in nature. In other words, even if the ancestor sequence is a purely i.i.d. sequence, the descendant sequence, though stationary in nature, may fail to satisfy any standard mixing properties, including the α-mixing property, when the replication process is exchangeable but not i.i.d..

4. Concluding Remarks

We have modeled the replication of a biological sequence and the changes that gradually occur, leading to biological evolution, by a Z-process involving random mutations, deletions and insertions. Under this model, we have investigated the invariance of certain key properties of the family of probability models for character strings. This invariance is important as it ensures that the ancestor and the descendant sequences are driven by the same type of probability laws, which in turn ensures certain consistency of the probability models involved as the ancestor string itself is created by replication of its predecessor. Further, distributions of DNA words lead to useful statistical tools for analysis of DNA sequences resulting in biologically significant discoveries (see e.g., Basu, Burma and Chaudhuri (2003), Chaudhuri and Das (2001), Karlin and Ladunga (1994), Karlin, Ladunga and Blaisdell (1994) and Nussinov (1984a, 1984b)), and asymptotic distributions of word frequencies obtained from large DNA sequences can be conveniently derived when stationarity and mixing properties of such large sequences hold.

We conclude by making an interesting observation. When there is a single ancestor chain (the X-chain), as in our stochastic replication model, and it is replicated by two independent replication processes (say, the $Z^{(1)}$-process and the $Z^{(2)}$-process), the paired descendant chain (say, the $(Y^{(1)}, Y^{(2)})$-chain) is *in general not stationary* as a bivariate process even if all the X- and the Z-processes are i.i.d. in nature. For example, consider i.i.d. X's as well as two i.i.d. copying

processes, all mutually independent. Assume that there is no insertion, so that $S_{T_i} = T_i$ for all $i \geq 1$. Let μ_1 and μ_2 denote the mutation transformations that operate on characters in the X-sequence replicated by the $Z^{(1)}$-process and the $Z^{(2)}$-process in their mutation states, respectively. Then, using the i.i.d. nature of the X's, it can be verified that

$$\begin{aligned}
&P\Big\{Y^{(1)}_{1+k} = \alpha_1, Y^{(2)}_{1+k} = \alpha_2\Big\} \\
&= P\Big\{Z^{(1)}_{T^{(1)}_{1+k}} = M, Z^{(2)}_{T^{(2)}_{1+k}} = M, T^{(1)}_{1+k} \neq T^{(2)}_{1+k}\Big\} P\Big\{\mu_1(X_1) = \alpha_1\Big\} P\Big\{\mu_2(X_2) = \alpha_2\Big\} \\
&\quad + P\Big\{Z^{(1)}_{T^{(1)}_{1+k}} = M, Z^{(2)}_{T^{(2)}_{1+k}} = M, T^{(1)}_{1+k} = T^{(2)}_{1+k}\Big\} P\left\{\mu_1(X_1) = \alpha_1, \mu_2(X_1) = \alpha_2\right\} \\
&= P\left\{\mu_1(X_1) = \alpha_1\right\} P\left\{\mu_2(X_2) = \alpha_2\right\} + P\left\{T^{(1)}_{1+k} = T^{(2)}_{1+k}\right\} \\
&\quad \times \left[P\left\{\mu_1(X_1) = \alpha_1, \mu_2(X_1) = \alpha_2\right\} - P\left\{\mu_1(X_1) = \alpha_1\right\} P\left\{\mu_2(X_2) = \alpha_2\right\}\right].
\end{aligned}$$

As all the Z's are i.i.d., the renewal times of the state M are sums of i.i.d. geometric random variables. In other words, $T^{(1)}_{1+k}$ and $T^{(2)}_{1+k}$ are two i.i.d. negative binomial random variables each with index $1 + k$. Hence, the above probability depends on k, violating stationarity whenever

$$P\left\{\mu_1(X_1) = \alpha_1, \mu_2(X_1) = \alpha_2\right\} - P\left\{\mu_1(X_1) = \alpha_1\right\} P\left\{\mu_2(X_2) = \alpha_2\right\} \neq 0 .$$

This is a problem that needs to be tackled in our probabilistic modeling before we can study multiple descendants of a common ancestor. We intend to pursue this in a future paper.

Acknowledgement

We are indebted to Professor R. L. Karandikar for some helpful suggestions. An anonymous referee read an earlier version of the paper with great care and provided several useful comments that led to significantly improved presentation of various mathematical details. Research presented in this article was supported in part by a grant from the Council for Scientific and Industrial Research (CSIR), Government of India.

Appendix. Mathematical Details and Proofs of Theorems

From now on we use the i.i.d. sequence $\{J_1, J_2, \ldots\}$ to describe the characters inserted during the replication process. To be more precise, J_i is the letter inserted if $Z_{T_i} = I$ (i.e., $Y_i = J_i$), and $J_i = \alpha_j$ with probability π_j. Further, for $1 \leq i \leq k$, we define an i.i.d. sequence of random characters R^i_n's generated according to the probability distribution of the ith row of the mutation matrix $((\theta_{ij}))$ described in Section 2.

Using the above notation, we next observe that

$$Y_n = g(Z_{T_n}, X_{S_{T_n}}, U_n), \tag{3}$$

where the U_n's are i.i.d. and completely independent of $\{(Z_{T_n}, X_{S_{T_n}})\}$, and g is a deterministic function of its arguments. Here, U_n consists of $(k+1)$ independent components (recall that k is the number of characters in the alphabet $\mathcal{A}$). The i-th component of U_n is $R_n^i, 1 \le i \le k$, and the $(k+1)$-st component of U_n is J_n. The g function in this case takes the value J_n if $Z_{T_n} = I$, and takes the value R_n^i if $Z_{T_n} = M$ and $X_{S_{T_n}} = \alpha_i$. This representation reduces the proofs of stationarity and mixing of $\{Y_n\}$ to the proofs of stationarity and mixing of $\{(Z_{T_n}, X_{S_{T_n}})\}$.

A.1. Proof of stationarity

Denote $(a_p, a_{p+1}, \ldots, a_q)$ by $\{a\}_p^q$. Further, we write

$$\{T\}_p^q = (T_p, \ldots, T_q), \{Z_T\}_p^q = (Z_{T_p}, \ldots, Z_{T_q}), \{X_{S_T}\}_p^q = (X_{S_{T_p}}, \ldots, X_{S_{T_q}}), \text{ etc.},$$

and this notation is used throughout.

Proof of Theorem 3.1. We want to show that for any given $k \ge 1$, the distribution of $(Y_{n+1}, \ldots, Y_{n+k})$ is the same for all $n \ge 1$. Using the representation (3), it is enough to show that for each fixed k, the distribution of

$$(Z_{T_{n+1}}, \ldots, Z_{T_{n+k}}, X_{S_{T_{n+1}}}, \ldots, X_{S_{T_{n+k}}})$$

is the same for any $n \ge 1$. We do this next.

Condition on $T_n = t_n$, $S_{T_n} = s_{t_n}$, and $Z_{T_n} = I$ (or $Z_{T_n} = M$). Take $Z_{T_n} = I$ ($Z_{T_n} = M$). From index $t_n + 1$, the Z-chain evolves as an independent Markov chain with the same transition probabilities as the original Z-chain, tempered by the initial condition $Z_{T_n} = I$ ($Z_{T_n} = M$). We denote the evolution of this chain by *primed* variables, e.g., $Z'_1, Z'_{T'_1}$, etc. Corresponding probabilities are denoted by P_I (P_M). Also note that $S_{T_{n+1}} = s_{t_n} + S'_{T'_1}$, $S_{T_{n+2}} = s_{t_n} + S'_{T'_2}$, etc. Then, we have

$$\begin{aligned}
&P(\{Z_T\}_{n+1}^{n+k} = \{z\}_1^k, \{X_{S_T}\}_{n+1}^{n+k} = \{x\}_1^k) \\
=&\textstyle\sum_1 P(T_n = t_n, S_{T_n} = s_{t_n}, Z_{T_n} = I) \\
&\textstyle\sum_2 P_I(\{T'\}_1^k = \{t'\}_1^k, \{S'_{T'}\} = \{s'_{t'}\}_1^k, \{Z'_{T'}\}_1^k = \{z\}_1^k) P(\{X_{s_{t_n}+s'_{t'}}\}_1^k = \{x\}_1^k) \\
&+\textstyle\sum_1 P(T_n = t_n, S_{T_n} = s_{t_n}, Z_{T_n} = M) \\
&\textstyle\sum_2 P_M(\{T'\}_1^k = \{t'\}_1^k, \{S'_{T'}\} = \{s'_{t'}\}_1^k, \{Z'_{T'}\}_1^k = \{z\}_1^k) P(\{X_{s_{t_n}+s'_{t'}}\}_1^k = \{x\}_1^k), \quad (4)
\end{aligned}$$

where the sum $\sum_1$ is over (t_n, s_{t_n}) and $\sum_2$ is over the *primed* variables $(t'_n, s'_{t'_n})$. Since s_{t_n} is fixed, we use the stationarity of the X-sequence to replace

$P(\{X_{s_{t_n}+s'_{t'}}\}_1^k = \{x\}_1^k)$ by $P(\{X_{s'_{t'}}\}_1^k = \{x\}_1^k)$. Now, consider the independent Z'-chain $= \{Z'_0 = Z_{t_n}, Z'_1 = Z_{t_n+1}, Z'_2 = Z_{t_n+2} \cdots\}$, which has the same stationary transition probabilities as our Z-chain, and Z'_0 is restricted to be I or M. The chain $\{Z'_1, Z'_2, \ldots\}$ operates on $\{X_{s_{t_n}}, X_{s_{t_n}+1}, \ldots\}$, which has the same distribution as $\{X_0, X_1, X_2, \ldots\}$, in view of the assumed stationarity of the X-process. It is then clear that the inside sum $\sum_2$ reduces to respective probabilities of events involving $(Z'_{T'_1}, \ldots, Z'_{T'_k}, X_{S'_{T'_1}}, \ldots, X_{S'_{T'_k}})$ with the additional condition that $Z'_0 = I$ (M). We denote these probabilities with appropriate suffices. Summation over (t_n, s_{t_n}) leads to

$$\begin{aligned} & P(\{Z_T\}_{n+1}^{n+k} = \{z\}_1^k, \{X_{S_T}\}_{n+1}^{n+k} = \{x\}_1^k) \\ = & P(Z_{T_n} = I) P_I(\{Z'_{T'}\}_1^k = \{z\}_1^k, \{X_{S'_{T'}}\}_1^k = \{x\}_1^k) \\ & + P(Z_{T_n} = M) P_M(\{Z'_{T'}\}_1^k = \{z\}_1^k, \{X_{S'_{T'}}\}_1^k = \{x\}_1^k). \end{aligned} \tag{5}$$

The assumed stationarity of Z_{T_n} concludes the proof of stationarity for *simple events* of the above form. For a general event B involving $(Z_{T_{n+1}}, \ldots, Z_{T_{n+k}}, X_{S_{T_{n+1}}}, \ldots, X_{S_{T_{n+k}}})$, decomposition into disjoint union of such *simple events* and then summation leads to

$$P(B) = P(Z_{T_n} = I) P_I(B) + P(Z_{T_n} = M) P_M(B). \tag{6}$$

This concludes the proof of stationarity of $(Z_{T_n}, X_{S_{T_n}})$.

A.2. Proof of mixing property

Since the arguments required to prove Theorem 3.2 are somewhat complex (primarily due to the notation needed to write such a proof rigorously with all relevant details), we will first prove a simpler result (Proposition 3.1), which is a result on the α-mixing property of the ancestor and the descendant chains, when the Z-process is assumed to be i.i.d.. The main idea of the proof is to use α-mixing property of the X-process when $S_{T_{n+k}} - S_{T_k}$ is large, and to use large deviation inequalities when $S_{T_{n+k}} - S_{T_k}$ is small. After the reader sees the main ideas in a simpler setting, Theorem 3.2 will be presented as an extension of Proposition 3.1, and the proof of Theorem 3.2 will provide the necessary modification of the arguments and the mathematical results used in the proof of the Proposition 3.1 when the Z-sequence is Markovian.

A.2.1. Some useful notation and representation

In the proof of the mixing property one needs an expression for $P(A \cap B) - P(A)P(B)$, where A is an event that is describable in terms of $Z_{T_1}, \ldots, Z_{T_k}, X_{S_{T_1}}, \ldots, X_{S_{T_k}}$, and B is an event that is describable in terms of finitely many of the

variables $Z_{T_{n+k+1}}, \cdots, X_{S_{T_{n+k+1}}}, \cdots$. We first write $A = \{(Z_{T_1}, \ldots, Z_{T_k}, X_{S_{T_1}}, \ldots, X_{S_{T_k}}) \in C\}$, where C is a set of $2k$-tuples, of which the first k entries are letters from the set $\{I, M\}$, and the latter k entries are letters from the alphabet $\mathcal{A}$. We write C as a disjoint union over the first k coordinates $\{c\}_1^k$ times its $\{c\}_1^k$ section given the first k coordinates $C_{\{c\}_1^k}$ (a set of k-tuples). In other words,

$$C = \cup\{\{c\}_1^k\} \times C_{\{c\}_1^k}, \tag{7}$$

where the union is over disjoint sets, each of which is the Cartesian product of a singleton set (i.e., $\{\{c\}_1^k\}$) and a set of k-tuples of letters from the alphabet $\mathcal{A}$ (i.e., $C_{\{c\}_1^k}$).

The event A is broken up in terms of values of Z_{T_i}, T_i, S_{T_i}. For this, we select $1 \leq t_1 < \cdots < t_k$, a permissible k-tuple from $\{M, I\}^k$ (allowed by the representation (7) of A), say $\{c\}_1^k$, and permissible values for $0 \leq s_{t_1} \leq \cdots \leq s_{t_k}$ (governed by the two previous constraints and following the definition of the S_n-sequence). Clearly, for the event A, given $\{Z_t\}_1^k = \{c\}_1^k$, the sequence $\{X_{s_t}\}_1^k$ is forced to lie in the section $C_{\{c\}_1^k}$, which is a finite set as the alphabet is finite. Then, using the independence of the X-and the Z-sequences, we have

$$P(A) = \sum P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k) P(\{X_{s_t}\}_1^k \in C_{\{c\}_1^k}), \tag{8}$$

where the summation $\sum$ is over the $3k$-tuples $(\{t\}_1^k, \{s_t\}_1^k, \{c\}_1^k)$'s allowed by A.

Now, let B denote an event that is describable in terms of

$$Z_{T_{n+k+1}}, \ldots, Z_{T_{n+k+l}}, X_{S_{T_{n+k+l}}}, \ldots, X_{S_{T_{n+k+l}}}.$$

With a representation analogous to (7) given as $E = \cup\{\{e\}_1^l\} \times E_{\{e\}_1^l}$, and following the same notational convention, we can write

$$\begin{aligned} P(B) = \sum P(\{T\}_{n+k+1}^{n+k+l} = \{t\}_{n+k+1}^{n+k+l}, \{S_T\}_{n+k+1}^{n+k+l} = \{s_t\}_{n+k+1}^{n+k+l}, \{Z_T\}_{n+k+1}^{n+k+l} = \{e\}_1^l) \\ \times P(\{X_{s_t}\}_{n+k+1}^{n+k+l} \in E_{\{e\}_1^l}). \end{aligned} \tag{9}$$

Assume the Z-sequence to be i.i.d. in nature and, as in the proof of Theorem 3.1, *primes* will denote an independent chain after a visit of the Z-chain to the state of mutation (M) or insertion (I). Then, using stationarity, $P(B)$ is also obtained from (9) as

$$P(B) = \textstyle\sum_2 P(\{T'\}_1^l = \{t'\}_1^l, \{S'_{T'}\}_1^l = \{s'_{t'}\}_1^l, \{Z'_{T'}\}_1^l = \{e\}_1^l) P(\{X_{s'_{t'}}\}_1^l \in E_{\{e\}_1^l}). \tag{10}$$

Also, we can write $(T_{n+k+j}, S_{T_{n+k+j}}) = (T_{n+k}, S_{T_{n+k}}) + (T'_j, S'_{T'_j})$. Now, from the descriptions of A and B combined with the previous observation,

$$\begin{aligned} & P(A \cap B) \\ = & \textstyle\sum_1 \sum_3 \sum_2 P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, T_{n+k} = t_{n+k}, \\ & S_{T_{n+k}} = s_{t_{n+k}}) \times P(\{T'\}_1^l = \{t'\}_1^l, \{S'_{T'}\}_1^l = \{s'_{t'}\}_1^l, \{Z'_{T'}\}_1^l = \{e\}_1^l) \\ & \times P((\{X_{s_t}\}_1^k \in C_{\{c\}_1^k}) \cap (\{X_{s_t}\}_{n+k+1}^{n+k+l} \in E_{\{e\}_1^l})), \end{aligned} \tag{11}$$

where (i) the sum $\sum_1$ needs to be taken over the $3k$-tuples $(\{t\}_1^k, \{s_t\}_1^k, \{c\}_1^k)$ allowed by the event A; (ii) the sum $\sum_2$ needs to be taken over the $3l$-tuples $(\{t'\}_1^l, \{s'_{t'}\}_1^l, \{e\}_1^l)$ allowed by the event B; and (iii) the third summation $\sum_3$ needs to be taken over $(t_{n+k}, s_{t_{n+k}})$ (with the notational convention $(t_{n+k+j}, s_{t_{n+k+j}}) = (t_{n+k}, s_{t_{n+k}}) + (t'_j, s'_{t'_j})$).

A.2.2. Some preliminary results

With the preceding results at hand, we can state and prove the following.

Lemma 3.1. *Assume that the X-sequence is stationary, the Z-sequence is i.i.d., and the events A and B are as before. Then we have*

$$\begin{aligned} & P(A \cap B) - P(A)P(B) \\ = & \textstyle\sum_1 \sum_2 \sum_3 P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, T_{n+k} = t_{n+k}, \\ & S_{T_{n+k}} = s_{t_{n+k}}) \times P(\{T'\}_1^l = \{t'\}_1^l, \{S'_{T'}\}_1^l = \{s'_{t'}\}_1^l, \{Z'_{T'}\}_1^l = \{e\}_1^l) \\ & \times \Big\{ P((\{X_{s_t}\}_1^k \in C_{\{c\}_1^k}) \cap (\{X_{s_t}\}_{n+k+1}^{n+k+l} \in E_{\{e\}_1^l})) \\ & - P((\{X_{s_t}\}_1^k \in C_{\{c\}_1^k}) P(\{X_{s_t}\}_{n+k+1}^{n+k+l} \in E_{\{e\}_1^l})) \Big\}, \end{aligned} \tag{12}$$

where the sums $\sum_1, \sum_2, \sum_3$ are taken over the respective items as listed after (11).

Proof. The proof of Theorem 3.1 shows that the sequence $\{(Z_{T_n}, X_{S_{T_n}})\}$ is stationary (note that in the present case the Z_T's are actually i.i.d.). Then, from the expressions for $P(A)$ and $P(B)$ as in (8) and (10), we have

$$\begin{aligned} & P(A)P(B) \\ = & \textstyle\sum_1 P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k) P(\{X_{s_t}\}_1^k \in C_{\{c\}_1^k}) \\ & \times \textstyle\sum_2 P(\{T'\}_1^l = \{t'\}_1^l, \{S'_{T'}\}_1^l = \{s'_{t'}\}_1^l, \{Z'_{T'}\}_1^l = \{e\}_1^l) P(\{X_{s'_{t'}}\}_1^l \in E_{\{e\}_1^l}). \end{aligned} \tag{13}$$

We now proceed as follows in order to compare (11) with (13). Notice that $(T_{n+k} - T_k, S_{T_{n+k}} - S_{T_k})$ is independent of (T_k, S_{T_k}) and has the same distribution

as (T_n, S_{T_n}). Suppose that $(T''_n, S''_{T''_n})$ is an independent copy of (T_n, S_{T_n}). Since $\sum P(T''_n = t''_n, S''_{T''_n} = s''_{t''_n}) = 1$, where the sum $\sum$ is over all possible $(t''_n, s''_{t''_n})$, we have

$$\begin{aligned}
&P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k) \\
&= P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k) \times \sum P(T''_n = t''_n, S''_{T''_n} = s''_{t''_n}) \\
&= \textstyle\sum_3 P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, T_{n+k} = t_{n+k}, S_{T_{n+k}} = s_{t_{n+k}}),
\end{aligned}$$

where we have assumed $t_k + t''_n = t_{n+k}$ and $s_{t_k} + s''_{t''_n} = s_{t_{n+k}}$. This shows that (13) can be written as

$$\begin{aligned}
P(A)P(B) = \textstyle\sum_1\sum_3 P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, \\
T_{n+k} = t_{n+k}, S_{T_{n+k}} = s_{t_{n+k}}) \times P(\{X_{s_t}\}_1^k \in C_{\{c\}_1^k}) \\
\times \textstyle\sum_2 P(\{T'\}_1^l = \{t'\}_1^l, \{S'_{T'}\}_1^l = \{s'_{t'}\}_1^l, \{Z'_{T'}\}_1^l = \{e\}_1^l) \\
P(\{X_{s'_{t'}}\}_1^l \in E_{\{e\}_1^l}). \qquad (14)
\end{aligned}$$

Given $(T_{n+k}, S_{T_{n+k}}) = (t_{n+k}, s_{t_{n+k}})$ and $(T_{n+k+j}, S_{T_{n+k+j}}) = (t_{n+k+j}, s_{t_{n+k+j}})$, consider $(t'_j, s'_{t'_j})$ obtained from $(t_{n+k+j}, s_{t_{n+k+j}}) = (t_{n+k}, s_{t_{n+k}}) + (t'_j, s'_{t'_j})$. Using the stationarity of the X-sequence, we have

$$P(\{X_{s'_{t'}}\}_1^l \in E_{\{e\}_1^l}) = P(\{X_{s_t}\}_{n+k+1}^{n+k+l} \in E_{\{e\}_1^l}). \qquad (15)$$

We use (15) to replace $P(\{X_{s'_{t'}}\}_1^l \in E_{\{e\}_1^l})$ in (14). Combining (11), (13), (14) and (15), we get the required (12).

Proposition 3.1. *Let the X-sequence be stationary and satisfy the α-mixing property with geometric decay, and let the Z-sequence be i.i.d.. Then the Y-sequence is stationary and possesses the α-mixing property with geometric decay.*

Proof of Proposition 3.1. Stationarity of the Y-sequence follows from the previous Theorem 3.1. We prove the appropriate mixing property of the sequence $\{(Z_{T_n}, X_{S_{T_n}})\}$. Then the desired mixing property of the sequence $\{Y_n\}$ follows from (3).

To prove the mixing property of the sequence $\{(Z_{T_n}, X_{S_{T_n}})\}$, consider (12) for $P(A \cap B) - P(A)P(B)$ as derived in Lemma 3.1. We divide this expression into two parts by intersecting the event $(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, T_{n+k} = t_{n+k}, S_{T_{n+k}} = s_{t_{n+k}})$ with the events $\{S_{T_{n+k}} - S_{T_k} > n\beta\}$ and $\{S_{T_{n+k}} - S_{T_k} \le n\beta\}$ respectively, where $0 < \beta < 1$ remains to be chosen. Using the stationarity and the α-mixing property with geometric decay of the X-process, the absolute value of the expression that involves intersection with

$\{S_{T_{n+k}} - S_{T_k} > n\beta\}$ will be bounded above by

$$\begin{aligned}
&\phi^{[n\beta]} \times \textstyle\sum_1 \sum_3 \sum_2 P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, \\
&T_{n+k} = t_{n+k}, S_{T_{n+k}} = s_{t_{n+k}}, S_{T_{n+k}} - S_{T_k} > n\beta) \\
&\times P(\{T'\}_1^l = \{t'\}_1^l, \{S'_{T'}\}_1^l = \{s'_{t'}\}_1^l, \{Z'_{T'}\}_1^l = \{e\}_1^l) \\
&\leq \phi^{[n\beta]} P(S_{T_{n+k}} - S_{T_k} > n\beta) \qquad (16)
\end{aligned}$$

for sufficiently large n. Here $\phi \in (0,1)$ is such that for each $n \geq 1$, the X-process satisfies $\alpha(n) \leq \phi^n$, and $[n\beta]$ is the greatest integer smaller than or equal to $n\beta$. Here, for the summation convention, the reader may check (8) and (11). The absolute value of the other expression that involves intersection of $(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, T_{n+k} = t_{n+k}, S_{T_{n+k}} = s_{t_{n+k}})$ with $\{S_{T_{n+k}} - S_{T_k} \leq n\beta\}$ is obviously bounded by

$$P(S_{T_{n+k}} - S_{T_k} \leq n\beta) \times 2. \qquad (17)$$

Combining (12), (16) and (17), we get

$$|P(A \cap B) - P(A)P(B)| \leq \phi^{[n\beta]} + 2P(S_{T_{n+k}} - S_{T_k} \leq n\beta)\ . \qquad (18)$$

Notice that

$$\begin{aligned}
P(S_{T_{n+k}} - S_{T_k} \leq n\beta) &= P(\chi_{T_k+1} + \chi_{T_k+2} + \cdots + \chi_{T_{n+k}} \leq n\beta) \\
&\leq P(\chi_{T_{k+1}} + \chi_{T_{k+2}} + \cdots + \chi_{T_{n+k}} \leq n\beta). \qquad (19)
\end{aligned}$$

Since the χ_{T_i}'s are independent $0-1$ valued random variables, by Hoeffding's inequality (see Hoeffding (1963)), there exist $\beta > 0$ and $0 < \theta < 1$ such that

$$P(\chi_{T_{k+1}} + \chi_{T_{k+2}} + \cdots + \chi_{T_{n+k}} \leq n\beta) \leq \theta^n, \qquad (20)$$

whenever $0 < P(\chi_{T_i} = 1) < 1$. Combining (18), (19) and (20), the proof of α-mixing with geometric decay is complete.

5.2.3. Proof of Theorem 3.2

We now consider the case where the X-sequence is stationary and possesses the α-mixing property with geometric decay, and the Z_T-sequence is stationary and Markov. Since we use some of the main ideas of the previous proof and their necessary modifications for handling a Markov replication process, the next proof refers to the previous proof for the analogous parts.

We first derive a modification of the expression for $P(A \cap B) - P(A)P(B)$ as in Lemma 3.1 when the Z-process is Markov instead of being i.i.d.. The main

difference is that we have to condition on the event $Z_{T_{n+k}} = I$ (respectively M). First, we have, using the representation (6) from Theorem 3.1,

$$P(B) = P(Z_{T_{n+k}} = I)P_I(B) + P(Z_{T_{n+k}} = M)P_M(B).$$

Using the fact that the X-sequence is independent of the Z-sequence, we can conclude that

$$P(A \cap \{Z_{T_{n+k}} = I\} \cap B) - P(A)P(Z_{T_{n+k}} = I)P_I(B) \tag{21}$$

can be written as the sum of two terms:

$$\begin{aligned}
&\textstyle\sum_1\sum_3\sum_2 P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k,\\
&T_{n+k} = t_{n+k}, S_{T_{n+k}} = s_{t_{n+k}}, Z_{T_{n+k}} = I)\\
&\times P_I(\{T'\}_1^l = \{t'\}_1^l, \{S'_{T'}\}_1^l = \{s'_{t'}\}_1^l, \{Z'_{T'}\}_1^l = \{e\}_1^l)\\
&\times \Big\{P((\{X_{s_t}\}_1^k \in C_{\{c\}_1^k}) \cap (\{X_{s_t}\}_{n+k+1}^{n+k+l} \in E_{\{e\}_1^l}))\\
&-P((\{X_{s_t}\}_1^k \in C_{\{c\}_1^k})P(\{X_{s_t}\}_{n+k+1}^{n+k+l} \in E_{\{e\}_1^l}))\Big\}; \qquad (22)\\
&\textstyle\sum_1\sum_3\Big\{P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k,\\
&T_{n+k} = t_{n+k}, S_{T_{n+k}} = s_{t_{n+k}}, Z_{T_{n+k}} = I)\\
&- P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, T_{n+k} = t_{n+k}, S_{T_{n+k}} = s_{t_{n+k}})\\
&\times P(Z_{T_{n+k}} = I)\Big\} \times P(\{X_{s_t}\}_1^k \in C_{\{c\}_1^k})\\
&\times P_I(B). \qquad (23)
\end{aligned}$$

In order to verify the above assertions, we need to show that the negative term in (22) is the positive term in (23), so that (22) plus (23) is (21). For this, after removal of the curly brackets in (22), we use the stationarity of the X-sequence (as we did in (15)), to write $P(\{X_{s_t}\}_{n+k+1}^{n+k+l} \in E_{\{e\}_1^l}) = P(\{X_{s'_{t'}}\}_1^l \in E_{\{e\}_1^l})$, and perform the $\sum_2$ summation in the negative term of (22) to get

$$\textstyle\sum_2 P_I(\{T'\}_1^l = \{t'\}_1^l, \{S'_{T'}\}_1^l = \{s'_{t'}\}_1^l, \{Z'_{T'}\}_1^l = \{e\}_1^l) \times P(\{X_{s'_{t'}}\}_1^l \in E_{\{e\}_1^l}) = P_I(B).$$

Here as before, we use $(t_{n+k+j}, s_{t_{n+k+j}}) = (t_{n+k}, s_{t_{n+k}}) + (t'_j, s'_{t'_j}), j = 1, \ldots, l$. Also, in this case, because of the conditioning on the event $Z_{T_{n+k}} = I$, summation over the *primed variables* leads to $P_I(B)$ and not $P(B)$. This completes the proof that (22) plus (23) is (21).

If we sum over $(t_{n+k}, s_{t_{n+k}})$ while keeping $Z_{T_{n+k}} = I$, the term within curly

brackets (after $\sum_3$) of (23) becomes

$$
\begin{aligned}
&= \sum\nolimits_1 \Big\{ P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, Z_{T_{n+k}} = I) \\
&\quad - P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k) P(Z_{T_{n+k}} = I) \Big\} \\
&= \sum\nolimits_1 P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, Z_{T_k} = I) \\
&\quad \{P(Z_{T_{n+k}} = I | Z_{T_k} = I) - \Pi_I\} \\
&\quad + \sum\nolimits_1 P(\{T\}_1^k = \{t\}_1^k, \{S_T\}_1^k = \{s_t\}_1^k, \{Z_T\}_1^k = \{c\}_1^k, Z_{T_k} = M) \\
&\quad \{P(Z_{T_{n+k}} = I | Z_{T_k} = M) - \Pi_I\}. \qquad (24)
\end{aligned}
$$

Here we use the assumed stationarity of $Z_{T_{n+k}}$ to denote $P(Z_{T_{n+k}} = I)$ by Π_I, which does not depend on n or k. The expression (22) can be handled in the same way as in Proposition 3.1 using large deviation bounds (see Theorem 3.1 of Ellis (1984)) for the crucial observation that the Z_T's (and hence the χ_T's, which are one-one functions of the Z_T's) now form a Markov chain.

If the Z_T's form an irreducible, aperiodic and stationary Markov chain, the absolute values of the terms enclosed within curly brackets in (24) are less than or equal to $K\gamma^n$ (see Billingsley (1986), p.128), where $K > 0$ and $0 < \gamma < 1$ are constants related to the transition probability matrix of the Z_T-chain. Applying this to (24) and combining it with (23), we get that (23) is smaller in absolute value than $K\gamma^n P(A) P_I(B)$. This, and the previous paragraph, give a geometrically decreasing bound for (21). The intersection with $Z_{T_{n+k}} = M$ can be handled similarly. This completes the proof.

References

Basu, S., Burma, D. P. and Chaudhuri, P. (2003). Words in DNA sequences: some case studies based on their frequency statistics. *J. Math. Biol.* **46**, 479-503.

Billingsley, P. (1986). *Probability and Measure*. John Wiley, New York.

Billingsley, P. (1999). *Convergence of Probability Measures*. John Wiley, New York.

Chaudhuri, P. and Das, S. (2001). Statistical analysis of large DNA sequences using distribution of DNA words. *Current Sci.* **80**, 1161-1166.

Chaudhuri, P. and Dasgupta, A. (2005). On some stochastic models for replication of character strings. To appear in *Statist. Probab. Lett.*

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79-94.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Ellis, R. S. (1984). Large deviations for a general class of random vectors. *Ann. Probab.* **12**, 1-12.

Ewens, W. J. and Grant, G. R. (2001). *Statistical Methods in Bioinformatics*. Springer, New York.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. II. John Wiley, New York.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13-30.

Karlin, S. and Ladunga, I. (1994). Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci. USA* **91**, 12832-12836.

Karlin, S., Ladunga, I. and Blaisdell, B. E. (1994). Heterogeneity of genomes: measures and values. *Proc. Natl. Acad. Sci. USA* **91**, 12837-12841.

Krogh, A., Brown, M., Mian, I. S., Solander, K. and Hausler, K. (1994). Hidden Markov models in computational biology : applications to protein modeling. *J. Molecular Biol.* **235**, 1501-1531.

Nussinov, R. (1984a). Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Res.* **12**, 1749-1763.

Nussinov, R. (1984b). Strong doublet preferences in nucleotide sequences and DNA geometry. *J. Molecular Evolution* **20**, 111–119.

Pevzner, P. A., Borodovsky, M. Y. and Mironov, A. A. (1989a). Linguistics of nucleotide sequences I : the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. Biomol. Struct. Dyn.* **6**, 1013-1026.

Pevzner, P. A., Borodovsky, M. Y. and Mironov, A. A. (1989b). Linguistics of nucleotide sequences II: stationary words in genetic texts and the zonal structure of DNA. *J. Biomol. Struct. Dyn.* **6**, 1027-1038.

Schbath, S., Prum, B. and de Turckheim, E. (1995). Exceptional motifs in different Markov chain models for statistical analysis of DNA sequences. *J. Comput. Biol.* **2**, 417-437.

Waterman, M. S. (1995). *Introduction to Computational Biology.* Chapman and Hall, New York.

Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India.

E-mail: probal@isical.ac.in

Theoretical Statistics & Mathematics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India.

E-mail: amites@isical.ac.in

(Received January 2004; accepted January 2005)