# On some stochastic models for replication of character strings

Probal Chaudhuri, Amites Dasgupta*

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

## Abstract

In this note, some probabilistic models for replication of character strings are considered. These replication processes involve random mutations, deletions and insertions of characters. We investigate invariance of certain probabilistic properties of replicating character strings under the proposed stochastic models for the replication process. It is shown that some well-known types of hidden Markov models with finite state spaces arise as special cases of our stochastic replication models. We also introduce the notion of a hidden mixed Markov model for a character string that arises in a situation where the replication process satisfies exchangeability conditions.

*Keywords:* Exchangeability; Hidden Markov process; Markov process; Mixed Markov process

## 1. Introduction

Suppose that we have observed a random character string $\{Y_1, Y_2, \ldots\}$, where $Y_i \in \mathscr{A} = a$ finite alphabet of symbols ($= \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$, say). We assume that this observed sequence is generated by a random replication process operating on an (possibly unobserved) ancestor string $\{X_1, X_2, \ldots\}$ of characters from the same alphabet $\mathscr{A}$. Such replication of character strings arise in molecular evolution of DNA, RNA 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.,

*Corresponding author.

*E-mail addresses:* probal@isical.ac.in (P. Chaudhuri), amites@isical.ac.in (A. Dasgupta).

Churchill, 1989; Durbin et al., 1998; Ewens and Grant, 2001; Krogh et al., 1994; Pevzner, 1992; Pevzner et al., 1989a, b; Phillips et al., 1987a, b; Schbath et al., 1995; Waterman, 1995). Among those models, Markov and hidden Markov models are possibly the most extensively studied models for biological sequences. Let us assume that $\{Y_1, Y_2, Y_3, \ldots\}$ is obtained by replicating $\{X_1, X_2, \ldots\}$, where the replication process is subject to random *mutations*, *insertions* and *deletions* of characters at various positions. A natural question that arises now is what are the properties or features (e.g., Markov property, hidden Markov property, exchangeability) of a stochastic sequence that one can expect to hold for the descendant $Y$-sequence given similar properties of the ancestor $X$-sequence under appropriate random replication models.

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 questions raised in the preceding paragraph translate into invariance or non-invariance of that probability measure or some specific features of that probability measure under such a random transformation. These fundamental issues for character strings that undergo replication subject to random alterations will be investigated in detail in this paper. In course of this investigation, it will be shown that the random replication model considered here is more general than some standard hidden Markov models with finite state spaces considered in the literature (see e.g., Churchill, 1989; Ewens and Grant, 2001; Krogh et al., 1994; Waterman, 1995) in the sense that those latter models can be derived under special probabilistic conditions imposed on the replication process.

## 2. A model for random replication

We begin by describing a model for the random replication or copying mechanism, which 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 will 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 the alphabet $\mathscr{A}$ into a position in the $X$-sequence that it encounters by randomly selecting that letter from $\mathscr{A}$ according to the probability distribution $P$(''Inserted letter is $\alpha_i$'')$= \pi_i$ ($\pi_i \geqslant 0, 1 \leqslant i \leqslant k, \sum_{i=1}^{k} \pi_i = 1$), which 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 $k \times k$ stochastic matrix $((\theta_{i,j}))$, which will be assumed to be independent of the $Z$-process. Here for $1 \leqslant i, j \leqslant k$, $\theta_{i,j} =$ the conditional probability $P$(''The letter is mutated into $\alpha_j$ in the descendent chain'' | ''The letter was $\alpha_i$ in the ancestor chain''), and $\sum_{j=1}^{k} \theta_{ij} = 1$ for all $1 \leqslant i \leqslant k$. Note that the term ''mutation'' here does not necessarily mean alteration of a character in the $X$-sequence as we do allow for the possibility of $\theta_{ii} > 0$ for some or all the $i$'s.

Let $T_i$ be the time of the $i$th visit of the $Z$-process to the state $I$ or the state $M$. The following random variable can be used to keep track of the index (i.e. the position of a letter) in the $X$-sequence on which $Z_i$ operates.

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

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

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

As we will see in the forthcoming section, if the $X$-process is Markov, the $Y$-sequence may not be Markov even if the $Z$-process is i.i.d in nature. However, in this case, the $Y$-sequence will be a stochastic function of a Markov sequence.

Let us note at this point that if $Z_1 = Z_2 = \cdots = Z_m = I$, we will have $S_{T_1} = S_{T_2} = \cdots = S_{T_m} = 0$. Since the $X$-chain starts from $X_1$, one has 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 $\{X_1, X_2, \cdots\}$ part to the right, and as soon as we have the first $D$ or $M$ according to the above notation, that acts on $X_1$. Consequently, we may consider the sequence $\{X_0, X_1, \cdots\}$, whose $\{X_1, X_2, \cdots\}$ part is to be replicated by the $\{Z_1, Z_2, \cdots\}$. Without loss of generality for all our subsequent mathematical results, we will assume that $\{X_0, X_1, \cdots\}$ is a sequence with the same probabilistic features (to be imposed in the following sections) as the sequence $\{X_1, X_2, \cdots\}$.

### 2.1. Hidden Markov models

The following results demonstrate how some standard hidden Markov models with all state spaces finite (see e.g., Churchill, 1989; Ewens and Grant, 2001; Krogh et al., 1994; Waterman, 1995) may arise as special cases of our random replication model described above.

**Theorem 2.1.** *Suppose that $\{X_1, X_2, X_3, \ldots\}$ and $\{Z_1, Z_2, Z_3, \ldots\}$ are independent Markov chains with stationary transition probabilities, and define $(X_{S_{T_i}}, Z_{T_i}) = H_i$. Then $\{H_1, H_2, H_3, \ldots\}$ is a Markov chain with stationary transition probability. In this case, the Y-sequence satisfies the hidden Markov model in the sense that the $Y_i$'s are conditionally independent given the $H_i$'s, and the conditional distribution of $Y_i$ depends only on $H_i$ for all $i \geqslant 1$.*

**Proof.** We want to show that the conditional distribution of

$$(X_{S_{T_{n+1}}}, Z_{T_{n+1}}) \text{ given } (X_{S_{T_n}}, Z_{T_n}, \ldots, X_{S_{T_1}}, Z_{T_1})$$

is same as the conditional distribution of

$$(X_{S_{T_{n+1}}}, Z_{T_{n+1}}) \text{ given } (X_{S_{T_n}}, Z_{T_n}).$$

In the following, the summations ($\sum$) are over the values of $T_i$'s and $S_{T_i}$'s, and instead of writing $X_{S_{T_i}} = \alpha_j, Z_{T_i} = M$ (or $I$) etc., we simply write $X_{S_{T_i}}, Z_{T_i}$. Then, we have the following by the independence of the $X$-process and the $Z$-process,

$$\begin{aligned} &P(X_{S_{T_1}}, \ldots, X_{S_{T_{n+1}}}, Z_{T_1}, \ldots, Z_{T_{n+1}}) \\ &\quad = \sum P(X_{S_{T_1}}, \ldots, X_{S_{T_{n+1}}}) P(T_1, \ldots, T_{n+1}, S_{T_1}, \ldots, S_{T_{n+1}}, Z_{T_1}, \ldots, Z_{T_{n+1}}). \end{aligned}$$

Notice that given $Z_{T_n}$, the chain $\{Z_{T_n+1}, Z_{T_n+2}, \ldots\}$ is an independent Markov chain (depending on the initial state $Z_{T_n}$), with the same transition probabilities as the original $\{Z_1, Z_2, \ldots\}$ chain with state space $\{I, M, D\}$, which we notationally distinguish by using *primes*. Further, using the Markov property of the $X$-chain and the $Z$-chain, the above can be written as

$$\sum P(T_1, \ldots, T_n, S_{T_1}, \ldots, S_{T_n}, Z_{T_1}, \ldots, Z_{T_n}) P(X_{S_{T_1}}, \ldots, X_{S_{T_n}})$$
$$\times \left\{ \sum P_{Z_{T_n}}(T'_1, S'_{T'_1}, Z'_{T'_1}) P_{X_{S_{T_n}}}(X_{S'_{T'_1}}) \right\},$$

where the second sum is over the values of the *primed* random variables. Since $Z_{T_n}$ and $X_{S_{T_n}}$ are fixed, the inner sum is independent of $Z_{T_1}, \ldots, Z_{T_{n-1}}$ as well as $X_{S_{T_1}}, \ldots, X_{S_{T_{n-1}}}$, and the outer sum is equal to

$$P(X_{S_{T_1}}, \ldots, X_{S_{T_n}}, Z_{T_1}, \ldots, Z_{T_n}).$$

This completes the proof of the Markov property.

The proof of the Theorem is now complete in view of the description of the $Y$-sequence in terms of the $X$-sequence and the $Z$-sequence given in (2). □

**Theorem 2.2.** *Suppose that the ancestor sequence $\{X_1, X_2, X_3, \ldots\}$ itself satisfies a hidden Markov model such that there exists a Markov chain denoted by $\{Q_1, Q_2, Q_3, \ldots\}$ with a finite state space and stationary transition probability and the $X_i$'s are conditionally independent given the $Q$-process with the conditional distribution of $X_i$ depending only on $Q_i$. Assume also that the replication process $\{Z_1, Z_2, Z_3, \ldots\}$ is a Markov chain, with stationary transition probability, which is independent of the $\{Q_i\}$- and the $\{X_i\}$-chains. Then the observed $Y$-chain satisfies a hidden Markov model with an underlying Markov chain that too has a finite state space and stationary transition probability matrix.*

**Proof.** Using the arguments used in the proof of the preceding proposition, we get that the sequence $\{(Q_{S_{T_i}}, Z_{T_i}) : i = 1, 2, 3, \ldots\}$ is a Markov chain with stationary transition probability. Further, in the case when both of the $Q$- and $Z$-sequences have transition probability matrices with all their entries positive, the same is true for the Markov chain underlying the $Y$-sequence. The proof of the Theorem is now complete by observing that the $Y_i$'s are conditionally independent given this Markov chain, and the conditional distribution of $Y_i$ depends only on $(Q_{S_{T_i}}, Z_{T_i})$ as follows. If $Z_{T_i} = I$, we have $P(Y_i = \alpha_s | (Q_{S_{T_i}}, Z_{T_i})) = \pi_s$. On the other hand, if $Z_{T_i} = M$, we have $P(Y_i = \alpha_s | (Q_{S_{T_i}}, Z_{T_i})) = \sum_{r=1}^{k} P(X_{S_{T_i}} = \alpha_r | Q_{S_{T_i}}) \theta_{rs}$. □

An important implication of the preceding Theorem is that when the ancestor chain satisfies a hidden Markov model with an underlying Markov chain having finite state space and stationary transition probability and the replication process is Markov, the descendant sequence also satisfies a similar hidden Markov model. In other words, such hidden Markov property for a character sequence is preserved when it is subject to replication by a Markov replication process. However, the state space of the underlying (i.e., the hidden) chain for the ancestor string of characters might be smaller than the state space of the underlying Markov chain for the descendant string though both of the ancestor and the descendant strings are generated from the same alphabet of characters. Hence, the number of parameters needed for the hidden Markov model describing the $Y$-sequence may be more than the number of parameters needed for the hidden Markov model describing the $X$-sequence. It also follows from Theorem 2.2 that after repeated replication we still have a hidden Markov model, i.e. if the same chain is copied $n$ times by independent (across copies) $Z$-chains, the $n$th copy satisfies a hidden Markov model, and the number of hidden states in the model in this formulation will grow geometrically with the number of replications $n$.

### 2.2. Exchangeable processes and hidden mixed Markov models

Let us now consider a situation where the $Z$-chain is an exchangeable sequence. This is equivalent to assuming that different positions of the replication chain are stochastically indistinguishable. In such a situation, it can be shown that the $(X_{S_{T_i}}, Z_{T_i})$ chain is not necessarily Markov, but it will be conditionally Markov given some appropriate $\sigma$-field. This now motivates us to introduce another family of models for character strings and explore how it remains invariant under exchangeable replication processes. We will call a stochastic process a mixed Markov chain if it is a Markov chain with a random stationary transition probability matrix. In other words, the process is Markov given that transition probability matrix, and the term 'mixed' here refers to the integration with respect to the distribution of the random transition probability matrix in obtaining the unconditional distribution of the chain. Note that a usual Markov chain with a stationary transition probability matrix is a special type of mixed Markov process.

Then a character sequence (e.g., the $X$-chain) will be said to satisfy a hidden mixed Markov model if it is conditionally independent given an underlying mixed Markov process (e.g., the $Q$-chain—here the $X_i$'s will be independently distributed given the $Q_i$'s with the conditional distribution of $X_i$ depending only on $Q_i$ as before). The following theorem then follows.

**Theorem 2.3.** *Suppose that the $X$-process is hidden mixed Markov in nature with the underlying $Q$-chain having a finite state space. Assume that the $Z$-process is exchangeable, and it is independent of the $\{Q_i\}$- and the $\{X_i\}$-chains. Then the $Y$-process is also hidden mixed Markov in nature with an underlying chain having a finite state space.*

**Proof.** Since exchangeability is equivalent to being conditionally i.i.d given the tail $\sigma$-fields of the respective processes (see e.g., Feller, 1971), the $Z$-sequence is conditionally i.i.d. Also, the $Q$-sequence is conditionally Markov given its transition probability matrix. Then, in view of Theorem 2.1, $(Q_{S_{T_i}}, Z_{T_i})$ is conditionally Markov given the $\sigma$-field that is generated by the tail $\sigma$-field of the $Z$-process and the random transition probability matrix of the $Q$-chain. Hence, the $(Q_{S_{T_i}}, Z_{T_i})$ sequence will be unconditionally mixed Markov with the desired properties. Clearly, the $Y$-sequence is conditionally independent given the $(Q_{S_{T_i}}, Z_{T_i})$ sequence, and the conditional distribution of $Y_i$ depends only on $(Q_{S_{T_i}}, Z_{T_i})$. This completes the proof. □

A straight-forward implication of Theorem 2.3 is that after repeated replication of a hidden mixed Markov sequence by independent exchangeable replication processes, we still have a hidden mixed Markov process. In other words, if the same hidden Markov sequence $X$ is copied $n$ times by independent (across copies) exchangeable $Z$-chains, the $n$th copy satisfies a hidden mixed Markov model, and the number of hidden states in the model in this formulation will grow geometrically with the number of replications $n$.

## 3. Concluding remarks

Let us now try to summarise our main findings in this paper. We have observed at the beginning that simple Markov property of a character string is not necessarily preserved under stochastic replication even if the replication process is i.i.d. in nature. Subsequently we have established that when the replication process is Markov in nature, certain hidden Markov properties of the ancestor string with an underlying finite stationary Markov chain is preserved in the descendant string. We have also introduced the notions of mixed Markov and hidden mixed Markov processes and indicated how these properties can be preserved under exchangeable stochastic replication.

Of course, the actual replication of a biological sequence and the changes that gradually occur there leading to biological evolution are extremely complex in nature. It is a simplistic approach to model such replications by our $Z$-process involving random mutations, deletions and insertions. Nevertheless, when a family of stochastic models (e.g., Markov or hidden Markov) is used for character strings that undergo random replication, the invariance of that family of models under appropriate probabilistic conditions on the replication process is a fundamental requirement. Unless such an invariance holds, there will be an intrinsic inconsistency in view of the fact that the ancestor string itself is created by replication of its predecessor. Hence, the ancestor and the descendant sequences cannot be driven by two completely different types of probability laws.

It is easy to see that in Theorem 2.2, if the stationary transition probability matrices of the $Q$-sequence and the $Z$-sequence have all their entries positive, the same will hold for the Markov chain underlying the $Y$-sequence. In this the underlying Markov chain will be ergodic and the $Y$-sequence will possess the $\alpha$-mixing property with geometric decay. Readers are referred to Chaudhuri and Dasgupta (2005) for an extensive study of stationarity and mixing properties of replicating character strings. In the case of Theorem 2.3, if the mixed Markov chain $Q$ underlying the $X$-sequence has a random stationary transition probability matrix with all its entries positive with probability one, the mixed Markov chain underlying the $Y$-sequence will also have the same property. This would imply conditional ergodicity and conditional *alpha*-mixing properties when one conditions on appropriate transition probability matrices.

## Acknowledgements

An anonymous referee read the original version of the paper with great care and made several useful comments, which have helped in improving the original presentation. Research presented in this article was supported in part by a grant from the Council for Scientific and Industrial Research (CSIR), Government of India.

## References

Chaudhuri, P., Dasgupta, A., 2005. Stationarity and mixing properties of replicating character strings. Statistica Sinica, in press.

Churchill, G.A., 1989. Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. 51, 79–94.

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

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

Feller, W., 1971. An Introduction to Probability Theory and Its Applications, vol. II. Wiley, New York.

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

Pevzner, P.A., 1992. Nucleotide sequences versus Markov models. Comput. Chem. 16, 103–106.

Pevzner, P.A., Borodovsky, M.Y., 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 (5), 1013–1026.

Pevzner, P.A., Borodovsky, M.Y., 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 (5), 1027–1038.

Phillips, G., Arnold, J., Ivarie, R., 1987a. Mono- through hexa-nucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. Nucleic Acids Res. 15, 2611–2626.

Phillips, G., Arnold, J., Ivarie, R., 1987b. The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over- and underrepresented sequences by Markov chain analysis. Nucleic Acids Res. 15, 2627–2638.

Schbath, S., Prum, B., 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.