Mathematical Forum
Vol.XXI, 2008-2009

MODELING DNA MUTATION USING FUZZY RELATIONS

Tazid Ali and Chandra Kanta Phukan
Dept. of Mathematics, Dibrugarh University, Assam-786004, India.
E-mail: tazidali@yahoo.com,
chandra_kanta25@yahoo.com

Abstract

DNA mutation processes includes various types of mutations namely substitution, deletion, insertion and inversion of a section of the sequence. The processes have been modeled using the language of probability. In this paper we have attempted to generalize the existing probability model using fuzzy relations.

Key words : DNA mutation, Substitution, deletion, insertion, fuzzy transition matrix, and membership matrix.

1. Introduction:

Molecular evolution is a wide range of biological processes. Depending on the Nature, species accumulate and lose external as well as internal characters, behaviors etc. from generation to generation in the processes of evolution. In molecular level also species carry some underlying variability in their genetic set

up, especially in DNA molecules of cells. Emphasis on DNA level mutation processes, a probability model has already been developed to deduce the effect genetic variability of ancestral generation on future generation.

In our study, we have developed a model to generalize the existing model, by removing some kind of restrictions from it, and have created fuzzy relations among the sites of the DNA sequences of two successive generations. In addition to this, the model has been extended to study three different types of mutation namely, deletion, insertion and both together respectively.

2. DNA background:

DNA (Deoxyribonucleic acid) is found in the cells of all living organisms except plant virus. The main role of DNA in cell is long-term storage of genetic information, which is copied during cell division. DNA is composed of three types of compounds such as Sugar molecules (Pentose Sugar), Phosphoric acid and Four nitrogenous bases namely Adenine, Guanine, Cytosine, and Thymine represented by letters A, G, C and T respectively. Because of chemical similarity, Adenine and Guanine are contained in purines and that of Cytosine and Thymine in pyrimidines.

A DNA molecule forms a double helix i.e., a twisted ladder- like structure. Each pole of the ladder is composed of sugar molecules and each sugar molecule attaches one phosphate group at c-3 and c-5 carbon atom of it alternatively. While c-1 carbon atom of sugar is attached by one of the four bases A, G, C, & T. The attached bases of these two poles can also be joined through hydrogen bond and these form the rungs of the ladder. The unique feature of base paring is $A \rightarrow T$, $T \rightarrow A$, $G \rightarrow C$ and $G \rightarrow C$. Therefore the sequence of bases of one side is enough to deduce the other. For e.g. if bases along one pole is AGTCGCTA then the other have the complementary sequence TCAGCGAT. Some segments of DNA sequence that carry genetic information are called genes. Genes encodes instruction for the manufacturing of protein through messenger RNA. Three consecutive bases of genes forms codons, each codon specify a particular amino acid say GCT-Alamine.

3. Fuzzy sets and fuzzy relation:

A fuzzy set of a universal set X is a function $\mu : X \to [0, 1]$. For $x \in X$, $\mu(x)$ is called grade of membership of x. The height $h(\mu)$ of a fuzzy set A is the largest membership grade obtained by an element of that set. Formally, $h(\mu) = sup_{x \in X} \mu(x)$ the fuzzy μ is said to be normal if $h(\mu) = 1$. If we divide each grade of membership of $x \in X$ by $h(\mu)$, then we get a normal fuzzy set. This process of obtaining a normal fuzzy set from a given fuzzy set is called normalization and the new fuzzy set obtained is called normalized fuzzy set. If we are concerned with only the relative grade of membership of different elements of the universal set, then we can replace a given fuzzy set by its normalized form.

For two universal sets X, Y a fuzzy relation from X to Y is a fuzzy subset of $X \times Y$ i.e., a function $\mu : X \times Y \to [0,1]$. $\mu(x, y)$ will represent the degree of association of x to y. A convenient representation of fuzzy relation is by membership matrices $M = [r_{xy}]$, where $r_{xy} = \mu(x, y)$. Another useful representation of fuzzy relation is a sagittal diagram. For a fuzzy relation μ from $X \to Y$, the inverse fuzzy relation $\mu^{-1} : Y \times X \to [0,1]$ is defined as $\mu^{-1}(y, x) = \mu(x, y)$. Consequently the membership matrix of μ^{-1} is the transpose of the membership matrix of μ. Let γ be a fuzzy relation from $X \to Y$ and ν be a fuzzy relation from $Y \to Z$ then the composite relation from $X \to Z$ is given by $(\gamma \circ \nu)(x, z) = \max [\{\min \gamma(x, y), (y, z)\} : y \in Y]$. Other types of composition can also be used in fuzzy relation such as max-max, min-min, max-product, and max-average etc. but the max-min composition has become the best known and the most frequently used one. The physical interpretation of the composition can be given as the strength of a set of chains linking x to z. Each chain has the form x-y-z. The strength of such a chain is that of the weakest link. The strength of the relation between x and z is that of the strongest chain between x and z.

4. Overview of existing probability model [1]:

The most common mutation that occurs in the coping of DNA sequences is base substitutions. This is nothing but the replacement of a base by another at a

certain site in the sequence. For example if AGATC in an ancestral sequence becomes ACATC in descendent, the only substitution occurs is $G \rightarrow C$. Focusing on base substitution (no deletion, insertion and inversion occurs) and ignoring the effect of codons in protein, a probability model has been carried out. Given an ancestral sequence, the probability of each of the four bases A, G, C and T to be present in the sequences are P_A, P_G, P_C, & P_T respectively and denoted by a vector P_o that is $P_o = (P_A, P_G, P_C, P_T)$, where $P_A + P_G + P_C + P_T = 1$ (sum of the total probability). This vector describes ancestral probability distribution of bases. It is observed that 16 conditional probabilities have been specified through out the processes. These are $P(S_1 = i/S_0 = j)$ for i, j = A, G, C & T. Now these can be arranged in a 4 × 4 matrix of vectors to make it mathematically tractable i.e.,

$$S_o \rightarrow \quad \begin{array}{c} \\ A \\ \uparrow\ G \\ S_1\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} P_{A/A} & P_{A/G} & P_{A/C} & P_{A/T} \\ P_{G/A} & P_{G/G} & P_{G/C} & P_{G/T} \\ P_{C/A} & P_{C/G} & P_{C/C} & P_{C/T} \\ P_{T/A} & P_{T/G} & P_{T/C} & P_{T/T} \end{bmatrix} \end{array} = \mathbf{M\ (say)}$$

Where, $P_{x/y} = P(S_1 = x/S_o = y) = \dfrac{P(S_1 = x \text{ and } S_o = y)}{P(S_o = y)}$

The matrix $M_{4\times4}$ is considered as transitional matrix. On ordinary multiplication of matrix $M_{4\times4}$ with $P_{o4\times1}$ (the column matrix of initial probability), one 4×1 matrix can be obtained and denoted it by P_1, which gives the probabilities of various bases occurring in the sequences S_1.

5. Generalized model for DNA mutation:

Suppose we are given a DNA sequence having 'm' sites in S_o generation as

S_o = AGTCGATCCGTAGAGG.....(m site),

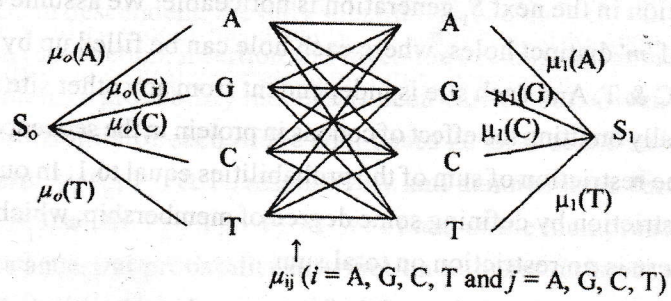
where mutation will occur during molecular evolution. Now the effect of mutations of S_0 generation in the next S_1 generation is noticeable. We assume the sequence to be a series of 'm' distinct holes, where each hole can be filled up by one of the four bases A, G, C & T. And each site is independent from any other site in the sequence i.e., biologically omitting the effect of codons in protein in the sequences. The previous model has the restriction of sum of the probabilities equal to 1. In our model we will relax this restriction by defining some degree of membership, which takes values in [0, 1] but there is no restriction on total sum.

5.1 Model for base substitution:

As already mentioned a DNA sequence can be considered a sequence of 'm' distinct holes. The n^{th} site of So generation will be denoted by N_0. Concentrating on an arbitrary n^{th} hole that is n^{th} site of the sequence, we can fill up this by any one of four bases A, G, C, & T with some grade of membership say, $\mu_0(A)$, $\mu_0(G)$, $\mu_0(C)$, $\mu_0(T)$ i.e., we can consider the n^{th} site as fuzzy subset of {A, G, C, T}. This can also be looked upon as fuzzy relation $\mu_0 : \{N_0\} \rightarrow \{A, G, C, T\}$. Similarly the n^{th} site of S_1 generation will be denoted by N_1. There the membership of A, G, C, T will be denoted by $\mu_1(A)$, $\mu_1(G)$, $\mu_1(C)$, $\mu_1(T)$ which also gives a fuzzy relation say $\mu_1 : \{N_1\} \rightarrow \{A, G, C, T\}$. Let

$$\begin{bmatrix} \mu_0(A) \\ \mu_0(G) \\ \mu_0(C) \\ \mu_0(T) \end{bmatrix} = M_0 \qquad \begin{bmatrix} \mu_1(A) \\ \mu_1(G) \\ \mu_1(C) \\ \mu_1(T) \end{bmatrix} = M_1$$

M_0 and M_1 will be called membership matrix of S_0 generation and S_1 generation respectively. When the mutation takes place, any of the bases A, G, C, T can mutate to any one of A, G, C, T. There will be 16 possible mutations. This can be represented by a fuzzy relation $\mu_{ij} : \{A, G, C, T\} \times \{A, G, C, T\} \rightarrow [0,1]$ i = A, G, C, T j = A, G, C, T. Now the fuzzy relation μ_1 can be obtained as a composite of fuzzy relations μ_0 and μ_{ij}, which can be express by a sagittal diagram as below



The fuzzy relation μ_{ij} can be represented by a 4 × 4 matrix as shown below. We call this matrix a fuzzy transition matrix (FTM).

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} \mu_{AA} & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & \mu_{GG} & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & \mu_{CC} & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & \mu_{TT} \end{bmatrix} \end{array} = F$$

On operating the FTM F and the membership matrix M_o, applying max.min composition, we have obtained one 4 × 1 matrix M_1. And each entry of M_1 represents the membership grade for n^{th} site (N_1) of S_1 generation. This is given by

$\mu_1(j) = \max [\min\{(\mu_{Aj}, \mu_o(A)), (\mu_{Gj}, \mu_o(G)), (\mu_{Cj}, \mu_o(C)), (\mu_{Tj}, \mu_o(T))\}], j$ = A, G, C, T.

5.2 Model for substitution and deletion:

This model is an extension of 5.1. In this case we assume the n^{th} site of S_1 generation may disappear in the process of mutation although it is already being filled by one of A, G, C and T in S_o. We introduce a letter B as one more site in S_1 to represent the disappearances. In a similar manner of 5.1 the FTM is defined as relation from the (A, G, C, T) to (A, G, C, T, B). The corresponding matrix is shown below

$$\begin{array}{c} \\ A \\ G \\ C \\ T \\ B \end{array}\begin{bmatrix} A & G & C & T \\ \mu_{AA} & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & \mu_{GG} & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & \mu_{CC} & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & \mu_{TT} \\ \mu_{AB} & \mu_{GB} & \mu_{CB} & \mu_{TB} \end{bmatrix} = F \quad \text{and}$$

Now applying max.min composition between F & M_o we will get a 5 × 1 membership matrix whose entries will be

$\mu_1(j) = \max [\min\{(\mu_{Aj}, \mu_o(A)), (\mu_{Gj}, \mu_o(G)), (\mu_{Cj}, \mu_o(C)), (\mu_{Tj}, \mu_o(T))\}]$, j = A, G, C, T, B.

And they represent the membership grades of the bases A, G, C, and T in the n^{th} site of S_1.

5.3 Model for substitution and insertion:

To model this situation we have introduced a new membership grade in an arbitrary site of S_o to be a blank site, where the insertion may take place. Here membership matrix M_o of S_o will have five components (μ_o(A), μ_o(G), μ_o(C), μ_o(T), μ_o(B) and the FTM is of 4 × 5 order i.e.

$$\begin{array}{c} \\ A \\ G \\ C \\ T \end{array}\begin{bmatrix} A & G & C & T & B \\ \mu_{AA} & \mu_{GA} & \mu_{CA} & \mu_{TA} & \mu_{BA} \\ \mu_{AG} & \mu_{GG} & \mu_{CG} & \mu_{TG} & \mu_{BG} \\ \mu_{AC} & \mu_{GC} & \mu_{CC} & \mu_{TC} & \mu_{BC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & \mu_{TT} & \mu_{BT} \end{bmatrix} = F \quad \text{Here} \quad M_o = \begin{bmatrix} \mu_o(A) \\ \mu_o(G) \\ \mu_o(C) \\ \mu_o(T) \\ \mu_o(B) \end{bmatrix}$$

Similarly applying max.min composition we will get a 4 × 1 membership matrix of S_1 with entries $\mu_1(j) = \max [\min\{(\mu_{Aj}, \mu_o(A)), (\mu_{Gj}, \mu_o(G)), (\mu_{Cj}, \mu_o(C)), (\mu_{Tj}, \mu_o(T)), (\mu_{Bj}, \mu_o(B))\}]$, j = A, G, C, T.

Combination of all the above three cases gives us an extended model, where membership matrix of S_o has five components μ_o(A), μ_o(G), μ_o(C), μ_o(T), μ_o(B) with a 5 × 5 order FTM as

$$
\begin{array}{c} \\ A \\ G \\ C \\ T \\ B \end{array}
\begin{array}{c}
\begin{array}{ccccc} A & G & C & T & B \end{array} \\
\begin{bmatrix}
\mu_{AA} & \mu_{GA} & \mu_{CA} & \mu_{TA} & \mu_{BA} \\
\mu_{AG} & \mu_{GG} & \mu_{CG} & \mu_{TG} & \mu_{BG} \\
\mu_{AC} & \mu_{GC} & \mu_{CC} & \mu_{TC} & \mu_{BC} \\
\mu_{AT} & \mu_{GT} & \mu_{CT} & \mu_{TT} & \mu_{BT} \\
\mu_{AB} & \mu_{GB} & \mu_{CB} & \mu_{TB} & \mu_{BB}
\end{bmatrix}
\end{array} = F
$$

$\mu_1(j) = \max [\min\{(\mu_{Aj}, \mu_o(A)), (\mu_{Gj}, \mu_o(G)), (\mu_{Cj}, \mu_o(C)), (\mu_{Tj}, \mu_o(T)), (\mu_{Bj}, \mu_o(B)\}],$

j = A, G, C, T, B.

The entry μ_{BB} of FTM represents the grade of association of a blank site of S_o to a blank site in the sequence of S_1. Finally this model becomes the most generalized model in our study.

Conclusion:

In this paper we have attempted to construct a model by which we can predict the membership matrix of the next generation, if we know the membership matrix of the previous generation. The model involves many membership values e.g., μ_{AA}, μ_{GA}, μ_{CA} etc. These values will vary from species to species. Even for the same species these values will be influenced by various biological factors. So these values will have to be supplied by experts in the field. However we can make certain assumption to make the model a bit simpler. For example it is usually observed that there is a natural tendency for a base to remain unchanged. So we can assume that $\mu_{AA} \geq \{\mu_{AG}, \mu_{AC}, \mu_{AT}, \mu_{AB}\}$. Another assumption that we can make is that $\mu_{AA} = \mu_{GG} = \mu_{CC} = \mu_{TT} = \mu_{BB}$. Again we are more concerned with relative membership values in the membership matrix so we can divide each entry of the transition matrix by μ_{AA}. The third assumption we can make is that $\mu_{AG} = \mu_{GA}$ and etc. This assumption is in fact essential since it is inherent in the very definition of inverse of a fuzzy relation. With the above assumptions our transition matrix will be a symmetric matrix with 1 in each entry of its diagonal. Our model will in fact predict membership

matrix of the previous generation if we are provided with information on the membership matrix of the new generation. Further we note that if we replace entries of the membership matrix by probability and conditional probabilities as entries of the fuzzy transition matrix i.e., the grade of associations as conditional probability, then we get the probability model as a particular case of our generalized model. In our model we have used the max-min operator as composition of fuzzy relation. There are other versions of fuzzy relation composition such as max-avg, max-product etc. Whenever data is available one can look for the suitable composition for the model.

Acknowledgement: This work is partially supported by SAP(DRS-I), Department of Mathematics, Dibrugarh University

REFERENCES:

[1] **Elizabeth S. Allman, Jhon A. Rhodes,** (2004), Mathematical Modeling in Biology, an Introduction, Cambridge University Press.

[2] **George J. Klir, Tina A. Floger** (1993), Fuzzy set, Uncertainty and Information, Prentice-Hall of India Privet Limited New Delhi-110001.

[3] **Richard Durrett,** Probability and its Application, Probability Model for DNA Sequence Evolution, Second Edition, Springer Science + Business Media, LLC233 Springer Street, New York, NY 10013, USA.

[4] **Veer Bala Rastogi,** Fundementals of Molecular Biology, Ane Book India, 4821, Parwana Bhawan,1st Floor, 24 Ansari Road, Darya Ganj, New Delhi-100012, India.