Searching Most Likely DNA Sequence Using HMM

S B Madagi

Abstract : **New DNA, RNA and Protein Sequences develop from pre- existing sequences rather than get invented by nature from scratch. This fact is the foundation of any sequence analysis through stochastic process. If we manage to relate a newly discovered sequence to a sequence about which something is already known, then chances are that the known information applies at least to some extent, to the new sequence as well. Further, all biological sequences are of evolutionary nature and may require techniques meant for evolutionary processes such as Markov models. The hidden Markov model is one among such useful tools. An example is used to illustrate the usefulness of HMM in searching most probable path.**

Keywords: **DNA, RNA, Protein, Sequence, Prokaryote, Eukaryote, Codon, Nucleotides, Site, Markov Chain**

I. Introduction

The studies in molecular biology sometimes require specific computational procedures on given sequence data, e.g., protein folding problem needs differential geometry and topology, evolution of biological sequences may be addressed using probabilistic evolutionary models etc. The procedure or the methods of identifying the prokaryotic genes were used to be, by extracting all ORFs from the DNA and analyzing each of them separately. It would be useful, however, if we can design an algorithm that could help us to analyze an unannotated DNA sequence directly, without making the preprocessing step of extracting all possible ORFs. In this paper an attempt is made to apply HMM for searching the most probable DNA sequence using hypothetical data.

II. Markov-Chain

Consider some finite set X of possible states G1, G2….. GN. At each of the time point t=1, 2,3…….…… a Markov chain occupies one of these states. In each time step t to t+1 the process either stays in the same state or moves to some other state in X. It does so in a probabilistic way, i.e. if at time t the process is in state Gi, then at time t+1 the process moves to any possible state Gj with a certain probability. This probability is assumed to depend only on i and j, not on t, or the states that the process occupied before state Gi. Now, P_{GiGj} or P_{ij}, is used to represent transition probability from state i to state j. The meaning is that, if we let the MC run freely, then, for every pair i, j, the proportion of observed transitions from Gi to G_j among all observed transitions from G_i tend to $P_{ij.}$ The probabilities Pij, i,j=1,2 ……. N

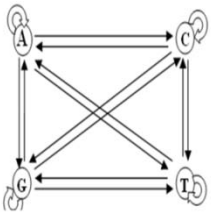
S. B. Madagi, Chairman, Dept. of Bioinformatics, Karnataka State Women University, Bijapur, Karnataka State, India (shivu_madagi@yahoo.com)

are called transition probabilities of the Markov chain and are arranged in the matrix form given below.

$$P = \begin{bmatrix} p_{11}\ p_{12} \ldots\ldots\ldots\ldots p_{1N} \\ p_{21}\ p_{22} \ldots\ldots\ldots\ldots p_{2N} \\ \vdots \quad \vdots \qquad\qquad \vdots \\ p_{N1}\ p_{N2} \ldots\ldots\ldots\ldots p_{NN} \end{bmatrix}$$

FIGURE I
Transition Probity Matrix

Suppose that in a set of prokaryotic DNA sequences n genes were experimentally identified. Then, we can set the connectivity diagram facilitating to the calculation of transition probabilities using the below mentioned formula.



$$P_{ab} = \frac{Hab}{\sum_{c \in Q} Hac} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

Where H_{ab} is the number of times nucleotide b follows a.

FIGURE II
Connectivity Diagram for Transition Probabilities

We may treat this Markov chain as a process of generating all possible sequences of any length L ≥2, ie, sequences of the form X_1, X_2….. X_L, where $X_j \in X$. In order to initiate the process , we need to fix in advance the probabilities $P(G_j)$ for all j = 1…….. N (called the initialization probabilities) that is N non negative numbers that sum up to 1.

For each j, $P(G_j)$ is the probability of the sequence generation process starting at state G_j. Choosing these values we mean that we are imposing the conditions like, number of run of the process goes to infinity, the proportion of sequences that starts with G_j among all generated sequences tends to $P(G_j)$ for all j = 1, 2…… N.

Next, if we generate a collection of sequences of length L and for every pair i, j determine the ratio of the number of times for which G_i in the generated sequences is immediately followed by G_j and the number of times for which G_i is immediately followed by any element from the set X. The resulting ratio is required to tend to P_{ij}. Then the frequency of a particular sequence $x=x_1\ x_2$ …….. x_L tends to

$$P(x) = P(x_1)\ P_{x1x2}\ P_{x2x3} \ldots\ldots\ldots P_{xl\text{-}1xl} \ldots\ldots\ldots (2)$$

We call this as the probability of sequence of x and $\sum_{j=1}^{N} Pj(x) = 1$ ……(3)

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Where summation is taken over all sequences of length L.

III. Hidden Markov Model:

An approach to look for new prokaryotic genes requires an extract of all ORFs from the DNA sequence in question and need to analyse each ORF separately. It would be useful, if we could analyze unannotated DNA, sequences directly, without making the preprocessing step of extracting all possible ORFs. The need for such an algorithm becomes even more clear, if instead of the problem of searching for prokaryotic genes, we consider the problem of searching for some other DNA features that don't have such well-defined boundaries as genes (start and stop codons).

Now, to construct such an algorithm, we have to model both the sequence composition of genes and that of intergenic regions. One possibility that can be tried is to use a model with conectivity shown in Fig. II for genes, another model with the same connectivity for intergenic regions, allow all possible transitions between the states of two models, and then add the "begin " and "end" states.

This resulting new model is called two block model. Here let A_g, C_g, G_g, T_g be the states of first model and A_{ig}, C_{ig}, G_{ig}, T_{ig}, be the states of second model. Since one has to allow for all possible transitions, the one to one correspondence doesn't remain longer and its interdeterminancy arises because of non-availability of prior information of the sequence X. In other words, the state sequence becomes hidden. Under such situation the application of Markov chain becomes a tool for searching most probable sequence and it is called Hidden Markov Model.

Definition of HMM:

An HMM is an ordinary discrete time Markov Chain with states G_1, G_2 …… G_N , transition probabilities P_{0j} , P_{ij} , P_{j0} , i,j= 1, 2, ……N, that in addition, at each state emits symbols from an alphabet Q (DNA alphabets, A, C, G, T). For each state G_k and each symbol A ∈ Q on emission probability $q_k(a)$ is specified and for each k=1….N, the probability sum up to 1 over all A ∈Q.

IV. Model Construction

Let x = x_1 x_2 ……. X_l be a sequence of letters from Q and π = π1 π2 ……..πl be the path of the same length. We will now define the probability P(x, π) as follows

$$P(x, \pi) = P_0\, \pi_1\, q\pi 1\,(x_1)\, P_{\pi 1\, \pi 2}\, q_{\pi 2}\,(x_2) \ldots P_{\pi l\text{-}1 \pi l}\, q_{\pi l}\,(x_l)\, P_{\pi l 0} \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Which gives the probability of the sequence x being generated along the path π. Since the path of biological sequences are not generally known, we need to develop an algorithm which will maximize P(x π) over all paths π of length equal to the length of x. By doing so, one of the most probable path is often generated as the path along which x is generated by the model i.e.

$$P(x) = \sum_{\text{all } \pi \text{ of length l}} P(x\, \pi) \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

$$\Longrightarrow \sum_x P(x) = 1$$

Based on this, the concept of Viterbi algorithm is developed.

Viterbi Algorithm- It is nothing but a dynamic programming algorithm designed to determine the most probable paths of nucleotides(Nucleic acids of DNA).

Consider an HMM whose underlying Markov Chain has a state set X=[G_1……. G_N], end state and transition probabilities P_{0j}, P_{ij},……P_{j0}, I,j=1……..N, Let X=X_1……X_L be a sequence of letters representing the nucleotides of DNA, then define

$$V_K(1) = P_{0k} q_k(X_1) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

for k=1,2…N

$$\text{And } V_k(i) = \text{Max } P_{0\pi 1}\, q_{\pi 1}(X_1)\, p_{\pi 1 \pi 2}\, q_{\pi 2}\,(X_2) \ldots\ldots. p_{\pi i\text{-}2\, \pi i\text{-}1}\, q_{\pi i\text{-}1}\,(X_{i\text{-}1})\; q_k(X_i)\, P_{\pi\text{-}1} G_k q_k(X_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

From i=2 ……. L and k=1……N. Then using eqn. (6) and (7) the recursion formula that emerges is

$$V_k\,(i+1) = q_k(X_{i+1})\, \text{Max}_{L=1}\, (V_L(i)\, P_{0k} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Hence, $V_k(i)$ can be calculated using initial condition ie eqn.(6) and recursion formula i.e. eqn. (7).

Further, using the set $V_k(i)$ and P(X π*)=Max (all π of length L) P(x, π)

$$\text{Whre } P(x,\pi) = P_{0\pi 1}\, q_{\pi 1}\,(X_1)\, P_{\pi 1\, \pi 2}\, q_{\pi 2}\,(X_2) \ldots\ldots\ldots P_{\pi L\text{-}1\, \pi L}\, q_{\pi L}\,(X_L)\, P_{\pi L 0} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)$$

The above described algorithm is illustrated using hypothetical data to identify the most probable path of genes.

V. Illustration

Consider the HMM for which Q is the two letter alphabet [AB] and CM is given in the fig-1 and the emission probabilities are:

Q1 (A) = 0.5 Q1 (B) =0.5

Q2 (A) = 0.1 Q2 (B) = 0.9

Q3 (A) = 0.9 Q3 (B) = 0.1

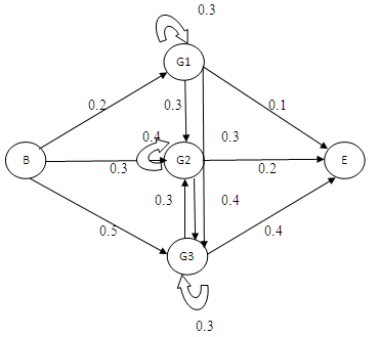


FIGURE III
CONNECTIVITY DIAGRAM FOR GENE SEQUENCES

Let x= BAB, and π_1= G1=BAB, π_2=G2= ABB, π3=G3= BBA.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Now considering the formula

$V_k(1) = V_{ok}\, qk\,(x1)$ and
$V_k(i+1) = q_k(x_{i+1})\, l^{Max}\,(V_l(i)P_{lk})$

We have the following calculations,

$V_1(1) = 0.2x\ 0.5 = 0.1 = p_{01}q_1$ (B)
$V_2(1) = 0.3 \times 0.9 = 0.27 = p_{02}q_2$ (B)
$V_3(1) = 0.5 \times 0.1 = 0.05 = p_{03}q_3$ (B)

$V_1(2) = q_1(x2)$ Max $[v_1(1)\, p_{11,}\, v_2(1)\, p_{21,}\, v_3(1)\, p_{31}]$
$= 0.5 \times$ Max [0.1 x .3, 0.27 x 0.4, 0.5 x0]
$= 0.5 \times 0.1 \times 0.3 = 0.015$

$V_2(2) = q_2(x2)$ Max $[v_1(1)\, p_{12,}\, v_2(1)\, p_{22,}\, v_3(1)\, p_{32}]$
$= 0.1x$ Max [0.1 x .3, 0.27 x 0.4, 0.5 x0]
$= 0.1 \times 0.27x\ 0.4 = 0.0108$

$V_3(2) = q_1$ (B) Max $[v_1(2)\, p_{11,}\, v_2(2)\, p_{21,}\, v_3(2)\, p_{31}]$
$= 0.5x$ Max [0.015 x .3, 0.0108 x 0.4, 0.0972x0]
$= 0.5 \times 0.015 \times 0.3$
$= 0.00225$

$V_1(3) = q_1$ (B) Max $[v_1(2)\, p_{11,}\, v_2(2)\, p_{21,}\, v_3(2)\, p_{31}]$
$= 0.5x$ Max [0.015 x 0.3, 0.0108 x 0.4, 0.0972x0]
$= 0.5 \times 0.015 \times 0.3$
$= 0.00225$

$V_2(3) = q_2$ (B) Max $[v_1(2)\, p_{12,}\, v_2(2)\, p_{22,}\, v_3(2)\, p_{32}]$
$= 0.9 \times$ Max [0.015 x 0 .3, 0.0108 x 0.4, 0.0972 x 0.3]
$= 0.9 \times 0.0972 \times 0.3$
$= 0.026244$

$V_3(3) = q_3$ (B) Max $[v_1(2)\, p_{13,}\, v_2(2)\, p_{23,}\, v_3(2)\, p_{33}]$
$= 0.1 \times$ Max [0.015 x 0.3, 0.0108 x 0.4, 0.0972 x 0.3]
$= 0.1 \times 0.02916$
$= 0.002916$

This gives Max $p(x, \pi) = 0.026244 \times 0.2$
$= 0.0052488$

If we look at $V_k(i)$ values we find the most probable path

i.e
$\Pi^* =$ $G_2G_3G_2$ called Viterbi Path.

VI. Conclusion

The above illustration indicates that the concept of HMM would help us in searching most probable path of a gene sequence.

References

[1] Burge, C., Karlin, S. : Prediction of Complete Gene Structure in Human Genomic DNA. J. Mol. Biol.268,78-94(1997)
[2] Durbin, R., Eddy, S., Crogh, A., Mitchison, G., : Biological Sequence
[3] Durbin, R., Eddy, S., Crogh, A., Mitchison, G., : Biological Sequence Analysis Cambridge University Press (1998)
[4] Ewens, W.J., Grant, G. R. :Statistcal methods in Bioinformatics. Springer – verlag(2001)
[5] Karlin, S : A First course in stochastic process. Academic press N.y, London (1968

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)