Promoter Database Search using Hidden Markov Model

Meera.A
BMS College of Engineering
Basavanagudi
Bengaluru

Lalitha Rangarajan
University of Mysore
Manasagangotri
Mysore

ABSTRACT
A common task in bioinformatics is the comparison of biological sequences to probabilistic models in order to evaluate their similarity. Completion of genomes of most of the organisms lead to profitable comparative analyses, providing insights into non-coding regions as well as into protein coding regions of DNA. In the present work we propose a method for finding similar sequence in a database of upstream sequences of DNA. For testing purpose, we have extracted upstream sequences of different mammals of citrate synthase and actin genes and also that of cab gene in different plants. The promoter sequences are extracted from NCBI database. Motifs/ TFBS of the upstream sequences are extracted using the software tool 'TF search'. Then probabilistic models are obtained for motif sequences by HMM method. Query motif sequence can be compared with all the motif sequences in the data base and based on maximum likelihood procedure, degree of similarity between query and all the motif sequences is obtained.

General Terms
Pattern matching, information retrieval

Keywords
Database, Hidden Markov Model, Promoter sequence, pattern matching, Transcription factors (TFs), Transcription factor binding sites (TFBS), Similarity measure

1. INTRODUCTION
Describing and modeling biological features of eukaryotic promoters remains an important and challenging problem in computational biology.

DNA is the molecule in which life organisms store information for their biological processes. The analysis of DNA sequences involves identifying the various patterns and understanding their functional roles. The order of occurrence of four alphabets in a DNA sequence is not completely random (else the percentage of occurrence of each alphabet would be 0.25). Different regions of the genome exhibit different patterns of these alphabets, A, T, G, C, e.g., protein coding regions, regulatory regions, which govern the production of proteins and enzymes, regulatory regions, repeat regions, intron/exon boundaries, etc.

A gene consists of a sequence of nucleotides from which RNAs (Ribo Nucleic Acid) can be transcribed to give either non coding regulatory RNAs or mRNAs or rRNAs. The information stored in the mRNA is used in the biosynthesis of proteins. The temporal, spatial and quantitative expression of RNA from genes is decided by the sequence existing before the sequence of a gene. This region is the promoter region. Transcription factors are those that bind to specific sequences in the promoter called transcription factor binding sites (TFBS) and regulate expression of the gene.

However, TFBSs associated to the same TF are known to tolerate sequence substitutions without losing functionality, and are often not conserved. Consequently, promoter regions of genes with similar expression patterns may not show sequence similarity, even though they may be regulated by similar configurations of TFs. Despite the recent progress due to the development of techniques based on so-called phylogenetic footprinting [Wasserman W. W. et. al., 2004, Meera A. et.al., 2009], lack of nucleotide sequence conservation between functionally related promoter regions may partially explain the still limited success of currently available computational methods for promoter characterization [Ficket J. W. et.al., 1997] and [Tompa M. et.al., 2005].

In recent years, the number of sequences and therefore, the size of databases available for comparison have grown exponentially. This growth has prompted scientists to develop faster and more sophisticated algorithms to keep pace with the increasing size of the databases. Software improvements combined with state-of-the-art hardware have allowed computational biologists to enter a new era of comparative genomics [S.F. Altschul et.al., 1990], [S. Altschul et.al., 1997], [W.R. Pearson et.al., 1988]. One example of this new growth may be seen in the use of probabilistic methods in bioinformatics, in particular, in the database searches. Although hidden Markov models (HMMs) were initially introduced for pattern recognition in digitized acoustics of the human voice [L.R. Rabiner et.al, 1989], they have become popular in bioinformatics. Current efforts in this area (including software) have been reviewed by [S.R. Eddy et.al, 1998]. HMMs in bioinformatics have been used in multiple-sequence alignments [Krogh et al, 1994]. They are also used in sequence analysis to produce an HMM that represents a sequence profile to study sequence composition and patterns [Durbin R. et al., 1999], to locate genes, and to predict protein structures. Searching a database against a given query is a fundamental process in bioinformatics.

In this paper, we employ Hidden Markov Model (HMM), an important tool in statistical modeling that has enjoyed great success in speech recognition. The basic theory behind HMMs was described by Baum and colleagues in a series of classic papers [Baum L.E**.** et al., 1970][Baum L.E. and J. Eagon, 1967] and was implemented by Baker [J.Baker et al., 1975] in the 1970s for speech processing. In recent years, HMMs have found wide application in computational biology.

2. METHODOLOGY

The promoter sequences of genes are extracted from NCBI database and subjected to TF search using 'TF search' tool. Thus the promoter sequences are converted into motif sequences.

2.1 Steps for HMM Modeling

Step1: Motif sequences are converted into numerical sequences. We have assumed N=Number of states in the model=5

Step2: Training the motif sequence:
Initial estimate for $\lambda = (\bar{A}, \bar{B}, \bar{\pi})$ are assumed.
A=a_{ij} is state transition matrix
B is output/ Observation / emission matrix.
Π is initial state distribution matrix.
Initially, we assume

$$A = a_{ij} = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix} \quad \{\text{NXN ie. 5X5}\}$$

$$\pi = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \{\text{NX1 ie. 5X1}\}$$

$$\text{B}=b_j(o_t) = \begin{bmatrix} 5/210 & \cdots & 5/210 \\ \vdots & \ddots & \vdots \\ 5/210 & \cdots & 5/210 \end{bmatrix} \quad \text{NXM ie. 5X210}\}$$

Where M=210(for example), is the number of observables (motifs) in the given motif sequence.

Step 3: Using Equation

$$\alpha_1(i) = \pi_i b_i(o_1) \quad 1 \le j \le N. \qquad (1)$$

the initial value of forward variable α_1, is calculated

$\alpha_1(i) = (1)(5/210)=5/210$

Then we obtain forward variable $\alpha_t(i)$, (probability of partial observation of sequence) using Equation

$$\alpha_{t+1}(j) = b_j(O_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij}, \quad 1 \le j \le N \qquad (2)$$

Step 4 : The backward variable $\beta_{t\,(i)}$ (the probability of being in state S_i, given the partial observation $o_{t+1},\ldots,o_T$) is calculated using equation

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(i) a_{ij} o_j(O_{t+1}), \quad 1 \le i \le N \qquad (3)$$

$$t = T-1 \ldots . N$$

e.g. $\beta_{1\,(1)} = a_{1,1} * b_1(O_2) * \beta_2(1)$
[at i=1,j=1,t=1] = (0.2)(5/210)(1)=1.0/210

Step 5: Using an initial parameter instantiation λ=(A,B,π), the Baum-welch algorithm or Expectation Maximization[(EM method) re-estimates three parameters π_i, $a_{i,j}$ and $b_i(o_t)$ where π_i = initial state distribution

$a_{i,j}$ =Transition probabilities
$b_i(o_t)$=Emission probabilities

Here we re-estimate1) Transition Probabilities:

$\xi_t(i,j)$ is the probability of being in state s_i at time t and going to state s_j, given the current model and parameters

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j \mid O, \lambda) \qquad (4)$$

$$= \frac{P(q_t = i, q_{t+1} = j \mid O, \lambda)}{P(O \mid \lambda)}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta t+1(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j)}$$

$$a_{ij} = \frac{expected\ number\ of\ transitions\ from\ state\ i\ to\ state\ j}{expected\ number\ of\ transitions\ from\ state\ i}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad (5)$$

$$= \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(i)}$$

2) Emission probabilities $b_i(k)$ =

$$\frac{expected\ number\ of\ times\ in\ s_{ii}\ and\ observed\ symbol\ vk}{expected\ number\ of\ times\ of\ times\ in\ state\ si}$$

^

$$b_j(k) = \sum_{t=1}^{t} \delta(o_t, v_k)\ \gamma_t(i) / \sum_{t=1}^{T} \gamma_t(i)\ b_i(k) \qquad (6)$$

Where $\delta(o_t, v_k) = 1\ if o_t = v_k\ otherwise\ 0$

and $\gamma_t(i)$ (state probability), the probability of being in state s_i, given the complete observation o_1……o_T.

3) Initial distribution probability

$$\hat{\pi} = \gamma_1(i)$$

The Updated Model is

λ' = $(\hat{A}, \hat{B}, \hat{\pi})$ by the following update rules:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\hat{b}i(k) = \sum_{t=1}^{t} \delta(o_t, v_k)\ \gamma_t(i) / \sum_{t=1}^{T} \gamma_t(i)$$

$\hat{\pi} = \gamma_1(i)$ Where $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$.

For each values of updated model λ, P(O/ λ) is calculated and above step is repeated until P(O/ λ) reaches local maxima.

Step 6: Testing the motif sequence (query sequence): Using Viterbi algorithm, we find P(O/λ) for optimal sequence.

1. Initialization

$$\delta_1(i) = \pi_i b_i(\boldsymbol{o}1), \quad 1 \le i \le N$$

2. Induction

$$\delta_t(j) = b_j(\boldsymbol{o}_t) \max_{0 \le x \le 1} \delta_{t-1}(i) a_{ij}$$
$$1 \le j \le N$$

$$P^* = \max_{0 \le x \le N} [\delta t - 1(i) a_{ij}],$$
$$1 \le j \le N$$

The probability value is generated for all models.

Fast Viterbi algorithm or Altenate Viterbi algorithm uses log probabilities rather than normal probabilities. This is done to replace the multiplication with addition and also to increase numeric fidelity as multiplications of the probabilities (which are less than 1) would demand high precision maintenance. The obtained probability P(O/λ) is logarithmic probability value.

2.2 Need for Threshold calculation for preventing false acceptance

In order to reject the sequences which do not have significant similarity with any of the sequences in the database, the following steps are used.

Let M be the number of such sequences to be trained to develop M statistical models given by $\lambda_1, \lambda_2, \ldots\ldots \lambda_M$ for the sequences $S_1, S_2, \ldots S_M$.

Let M be 5.

In the probability matrix given below,

$$D = \begin{bmatrix} p(S_1/\lambda_1) & p(S_1/\lambda_2) & p(S_1/\lambda_3) & p(S_1/\lambda_4) & p(S_1/\lambda_5) \\ p(S_2/\lambda_1) & p(S_2/\lambda_2) & p(S_2/\lambda_3) & p(S_2/\lambda_4) & p(S_2/\lambda_5) \\ p(S_3/\lambda_1) & p(S_3/\lambda_2) & p(S_3/\lambda_3) & p(S_3/\lambda_4) & p(S_3/\lambda_5) \\ p(S_4/\lambda_1) & p(S_4/\lambda_2) & p(S_4/\lambda_3) & p(S_4/\lambda_4) & p(S_4/\lambda_5) \\ p(S_5/\lambda_1) & p(S_5/\lambda_2) & p(S_5/\lambda_3) & p(S_5/\lambda_4) & p(S_5/\lambda_5) \end{bmatrix}$$

In any row, the diagonal element of the matrix is said to be having maximum value.

For example among the third row elements,

$$p(S_3/\lambda_1), p(S_3/\lambda_2), p(S_3/\lambda_3), p(S_3/\lambda_4), p(S_3/\lambda_5)$$

the third element being one of the diagonal elements of the entire matrix, is maximum among all the elements in the row.

But this may not be true for the elements present in the third column of the matrix D.ie., in the elements mentioned,

$$p(S_1/\lambda_3), p(S_2/\lambda_3), p(S_3/\lambda_3), p(S_4/\lambda_3), p(S_5/\lambda_3)$$

If an unknown test sequence which is not one among the trained sequences is used in the testing procedure, say S_{15}.

Then in testing procedure, by using Viterbi algorithm for every combination of the test sequence and statistical model, the following sequence of probabilities is evaluated.

$$p(S_{15}/\lambda_1), p(S_{15}/\lambda_2), p(S_{15}/\lambda_3), p(S_{15}/\lambda_4), p(S_{15}/\lambda_5)$$

Assuming $p(S15/\lambda 3)$ is maximum among all the probabilities given above, the S_{15} will be misjudged as S_3 due to the maximum likelihood principle.

So there is need for fixing threshold for each statistical model λ.

Then $p(S_{15}/\lambda_3)$ is to be compared with the threshold T_3. The threshold for the model λ_3 must be calculated in such way that the expression $p(S_{15}/\lambda 3) < T_3$ must be true.

2.3 Threshold Calculation for each statistical model λ

Let every sequence 'S_i' has certain number of variants of it given by S_{i1}, S_{i2}, S_{i3}, S_{i4}, S_{i5} and S_{i6}.

For example for the sequence S_1, the variants are given by S_{11}, S_{12}, S_{13}, S_{14}, S_{15} and S_{16}.

The probability matrix D for every set of variants for sequences S_1 to S_5 is evaluated.

The probability matrices are given by D_1, D_2, D_3, D_4, D_5 and D_6.The in-phase probabilities for each model λ_i are given by,

$$P(Si1/\lambda i), P(Si2/\lambda i), P(Si3/\lambda i), P(Si4/\lambda i), P(Si5/\lambda i), P(Si6/\lambda i)$$

The out-phase probabilities for every λi are given by,

$$P(Sj1/\lambda i), P(Sj2/\lambda i), P(Sj3/\lambda i), P(Sj4/\lambda i), P(Sj5/\lambda i), P(Sj6/\lambda i)$$

With $1 \le j \le 6$

The mean of the in-phase components for the model λ_i is evaluated and termed as $\mu_{in\text{-}i}$.

The mean of the out-phase components for the model λ_i is evaluated and termed as $\mu_{out\text{-}i}$.

The standard deviation of the in-phase components for the model λ_i is evaluated and termed as $\sigma_{in\text{-}i}$.

The standard deviation of the out-phase components for the model λ_i is evaluated and termed as $\sigma_{out\text{-}i}$.

The threshold Ti for the statistical model λi, is calculated by

$$T_i = \frac{\mu_{in-i}.\sigma_{out-i} + \mu_{out-i}.\sigma_{in-i}}{\sigma_{out-i} + \sigma_{in-i}}$$

Thus calculated T_i s will be more or less around the diagonal element $P(Si/\lambda i)$.Hence during the testing of a query sequence S_{15}, there is a chance that $P(S_{15}/\lambda i)$ be more than the $P(Si/\lambda i)$ and hence more than the threshold value T_i.

To prevent this, an array of variance sequence is calculated during the threshold calculation itself.

For every sequence Si with its variant sequence set say S_{i1}, S_{i2}, S_{i3}, S_{i4}, S_{i5} and S_{i6}, the variance V_i is calculated by finding the sum of the square of distances between the corresponding threshold T_i and the probability value, $P(S_{ik}/\lambda i)$, where $1 \le k \le 6$

$$V_i = \sum_{k=1}^{6} (Ti - p(S_{ik}/\lambda_i))^2$$
$$1 \le i \le M$$

while testing the query sequence say S_{15}, which is not used during the training stage, The probability sequence containing probabilities $p(S15/\lambda 1)$ *to* $p(S15/\lambda 5)$ are evaluated.

With maximum likelihood principle, the maximum value among these probabilities is determined.

Assuming $p(S_{15}/\lambda 3)$ to be maximum among all the other values, the square of the difference between this probability and the corresponding threshold value T_3 is determined.

(T3- $p(S_{15}/\lambda 3)$)^2 is compared with the variance V_3 available in the variance sequence set V.

S_{15} being 'foreign' sequence, the value determined from the expression (T3- $p(S_{15}/\lambda 3)$)^2 will be greater than V_3.

Thus S_{15} is considered as sequence which cannot be classified to be among the trained sequences.

3. RESULT

The promoter sequences of genes coding for citrate synthase and actin gene of different mammals and promoters of different plants of cab gene are retrieved from NCBI database and subjected to TF search using 'TF search' tool. Thus the primary promoter sequences are converted to motif sequences.

The query promoter sequence which is in the form of motif sequence is compared with all the sequences in the database and the log probability score with each sequence is displayed. The sequence in the database which has highest similarity with the query sequence is highlighted (in blue and underlined). Here, a section of the result has been shown.

pathway- CMP(citrate synthase) and Organism - BOS-5: -2571.002094
pathway- CMP(citrate synthase) and Organism - BOS-10: -2604.309499
pathway-CMP(citrate synthase) and Organism - Can-10: -3319.908145
pathway- CMP(citrate synthase) and Organism - HS-19: -1365.157139
pathway -CMP(citrate synthase) and Organism - HS-12: -4079.434347
pathway- CMP(citrate synthase) and Organism - HS-2: -2743.449172
pathway- CMP(citrate synthase) and Organism - HS-3: -2527.987613

The result reveals that the query motif sequence has highest similarity with Homosapien chromosome 19 of citrate synthase gene in Central Metabolic Pathway.

4. DISCUSSION AND CONCLUSION

The program has been tested for a number of query sequences and it is found that it is working efficiently. False acceptance rate of sequences which do not have significant similarity with any of the sequences in the database is zero.

Hidden Markov Modeling (HMM), is an important tool in statistical modeling that has enjoyed great success in Bioinformatics. HMMs have been used in classification [Georgina Mirceva and Danco Davcev, 2009, Denis F Wolf, et. al, 2005].

5. REFERENCES

[1] Baum L. E., Ted petrie, George Soules, and Normal Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Statist.; 41:pp164—171.

[2] Brutlag.D. 2002. Multiple sequence alignment and Motifs. Bioinformatics methods and Techniques. Stanford University, Stanford center for Professional development.

[3] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L.Wheeler, 2006. GenBank. Nucleic Acids Research, vol. 34, pp. D16--D20

[4] Denis F Wolf, Gaurav S Sukhatme, Dieter Fox, Wolfram Burgard, 2005. Autonomous Terrain Mapping and Classification Using Hidden Markov Models. Proceedings of the 2005 IEEE International Conference on Robotics and Automation Volume: 2005, Issue: April, Publisher: Ieee, pp 2026—2031.

[5] E. Lawrence and A. Reilly, 1990. An expectation maximization algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins, vol. 7, pp. 41—51.

[6] Eddy SR. 1998. Profile hidden Markov models. Bioinformatics.; 14:755

[7] Georgina Mirceva and Danco Davcev, 2009. HMM based approach for classifying protein structures. International Journal of Bio- Science and Bio-Technology Vol. 1, No. 1, December.

[8] J. Baker, 1975. The DRAGON system-an overview, IEEE Trans. Acoustics Speech and Signal Processing, vol. ASSP-23, pp. 24-9.

[9] Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler, 1994. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. J. Molecular Biology, vol. 235, pp. 1501-1531.

[10] Meera A, Lalitha Rangarajan, Savithri Bhat:, 2009. Computational Approach Towards Finding Evolutionary Distance And Gene order Using Promoter Sequences Of Central Metabolic Pathway" Inter disciplinary sciences-computational life sciences DOI: 0.1007/s12539-009-0017-3 [Spriger link]

[11] Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77:257-286.

[12] S. Altschul, T.L. Madden, A.A. Schaffer, J. Zheng, Z. Zhang, M.Miller, and D.L. Lipman, 1997. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs, Nucleic Acid Research, vol. 25, pp. 3389—3402

[13] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, 1990. Basic Local Alignment Search Tool. J. Molecular Biology, vol. 215,pp. 403--410.

[14] Wasserman WW, Sandelin A, 2004. Applied bioinformatics for the identification of regulatory elements. Nat. Rev Genet. 5, 276–286.

[15] W.R. Pearson and D.J. Lipman, 1988. Improved Tools for Biological Sequence Comparison, Proc. Nat'l Academy of Science, vol. 85,pp. 2444-2448.

[16] L. E. Baum, J. A. Eagon, 1967. An inequality with application to statistical estimation for probabilistic functions of Markov processes and to a model of ecology, Bull American Society, Vol, 73, pp. 360-363.