MODELING 5' REGIONS OF HISTONE GENES USING BAYESIAN NETWORKS

RAJESH CHOWDHARY

Knowledge Extraction Lab, Institute for Infocomm Research, 21 Heng Mui Keng Terrace Singapore 119613

R. AYESHA ALI

Department of Statistics and Applied Probability, 6 Science Drive, National University of Singapore, Singapore 117543

VLADIMIR B. BAJIC

Knowledge Extraction Lab, Institute for Infocomm Research, 21 Heng Mui Keng Terrace Singapore 119613

Histones constitute a rich protein family that is evolutionarily conserved across species. They play important roles in chromosomal functions in cell, such as chromosome condensation, recombination, replication, and transcription. We have modeled histone gene 5' end segments covering [-50,+500] relative to transcription start sites (TSSs). These segments contain parts of the coding regions in most of the genes that we studied. We determined characteristics of these segments for 116 mammalian (human, mouse, rat) histone genes based on distribution of DNA motifs obtained from MEME-MAST. We found that all five mammalian histone types (H1, H2A, H2B, H3, H4) have mutually distinct, prominent and strongly conserved properties downstream to the TSS reasonably well conserved across analyzed species. We then transformed the primary level motif data for each sequence into a higher order motif arrangement that involved only features such as presence of a motif, its position, its strand orientation, and mutual spacer length between motifs. We have built a Bayesian Network model based on these features and used the higher order motif arrangement data for its training and testing. When tested for classification between the five histone groups and using the leave-one-out cross-validation technique, the Bayesian model correctly classified 100% of histone H1 sequences, 100% of histone H2A sequences, 96.9% of histone H2B sequences, 94.4% of histone H3 sequences, and 95.8% of histone H4 sequences. Overall, the model correctly classified 97.4% of all histones sequences. Our Bayesian model has the advantage in having a small number of trainable parameters and it produces very few false positives. The model could be used to scan the genome for discovery of genes whose products are similar to histones.

1 Introduction

One important task after the completion of the human genome project has been to identify and characterize mutually similar genes in the genome. Similar genes tend to share similar DNA motif patterns in their promoter and coding regions. The task of identifying similar genes from the genome requires that some of the characteristic features of the genes be identified. One approach is to look at the similarities in the genes' 5' regions. This is also related to localization of the transcription start site (TSS) and using TSS as the reference point for approximate assessment of the promoter region

and its characteristics. Since identifying exact TSS location is difficult, researchers have focused on predicting the promoter region [1] of the gene. Previously, promoter prediction programs extensively used individual DNA motif occurrences in known promoters. Unfortunately, since DNA motifs have short lengths (6-20 bp) and can often occur in the genome purely by chance, these programs produce unacceptably high level of false positive. A possible solution that may reduce this problem to an extent is to consider motif modules. A motif module in a genomic segment can be defined as a collection of DNA motifs arranged in specific order, spacing, orientation and position with respect to each other. Lately, there have been programs based on Hidden Markov Models (HMMs) that try to model motif modules, such as COMET [2], MCAST [3], META-MEME [4] among others. All these programs have different pros and cons, such as limitation on the number of motif occurrence per sequence, higher number of trainable parameters, different feature sets based on motif spacing and order etc.

We propose a novel methodology to model *motif modules* using a Bayesian networks paradigm. The advantages of our proposed model are:

1. the model is simple and easy to implement,
2. the number of trainable parameters is linear in terms of the number of Bayesian network variables, and
3. the model is made specific by probabilistically combining information on an exhaustive set of features such as the motif, its strand, its absolute position from the TSS and mutual spacer distance.

In this paper we focus on a specific class of genes, the histone genes. Histones are basic proteins present in the eukaryotic cell nucleus. They are broadly divided into five types, namely H1, H2A, H2B, H3 and H4 [5]. Histones are evolutionarily conserved and have similar functions in all living organisms. This suggests that histone genes may share similar motif patterns in their 5' end region.

We collected 116 gene sequences comprising H1, H2A, H2B, H3, and H4 histone genes from three mammalian species of man, mouse and rat. We modelled motif modules present in the 5' end region covering [-50,+500] of the histone genes relative to the TSS and observed that this region covered the coding portion in most of the histone genes under study. Furthermore, this region contained strongly conserved motif patterns across species within different histone types, which enabled us to determine gene 5' end models for all five histone gene groups. We tested our model on the histone data using "leave-one-out cross-validation" technique and found that our system correctly classified 97.4% of the analyzed histone sequences to their respective histone groups. Our model can be used to classify an uncharacterized histone sequence into one of the five histone classes. Currently, we are working on extending the model to hunt for novel histone-like genes or genes co-regulated with them across mammalian genomes. Our model has a generic scope and can be scaled to include any type of motif definitions from a set of genes.

2 Methods

We used PromoSer tool release 3.0 [6] and FIE2.1, the latest version of FIE2.0 tool [7] (http://research.i2r.a-star.edu.sg/promoter/FIE2_1/), to extract histone gene sequences. We collected histone genes from three mammalian species of man, mouse and rat and extracted 5' end genomic region [-50,+500] relative to the TSS for each of the respective genes. We analyzed 116 histone gene sequences with 16 H1, 26 H2A, 32 H2B, 18 H3, and 24 H4 histone types.

Using MEME-MAST [8,9], we obtained DNA motif occurrences in the 116 histone sequences. Although, there are many software programs that discover DNA motifs with the *ab-initio* approach [see refs. 8,10-17], we chose MEME-MAST due to its relatively flexible parameter selection procedure. We built higher order motif definitions (HOMD) for each of these histone sequences using information on motif, its strand, its absolute position from the TSS, and mutual distance between motifs. We considered for HOMD a maximum of the first five motif occurrences per sequence. This way we had HOMD with 20 motif feature variables for each of the 116 histone sequences. HOMD data were used for training and testing the Bayesian model.

We used a naive Bayes network to develop a model for motif modules of the 5' end of histone genes and to classify a (known histone) gene into one of the five histone classes. Naive Bayes models are widely used for classification and are relatively easy to use. Figure 1 shows the model structure used to model the motif modules for histone genes. There is one parent class node representing the five histone classes and 19 child nodes representing tuples of four values (motif name, distance from TSS, motif strand and mutual distance) for the first five motifs occurring in a sequence. We used the Expectation Maximization (EM) algorithm [18] to train model parameters for each of the five histone classes, using uniform Dirichlet priors. We then used the junction-tree algorithm [19] for classification.

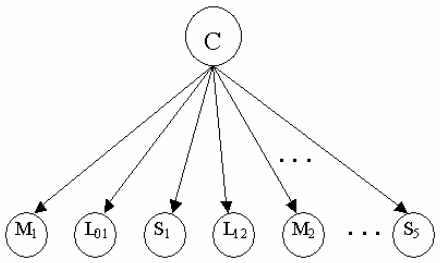

Figure 1. Naive Bayes model structure used for modeling motif modules in histone gene sequences in the region [-50,+500] relative to the TSS. The parent node C represents the class (H1, H2A, H2B, H3, H4) and the 19 child nodes represent features of each of the first five motifs occurring in a sequence (M_i - motif name, L_{0i} - length from start and S_i - strand (+/-) for i = 1, …, 5, and $L_{i(i+1)}$ - mutual spacer length between motifs for i = 1, …, 4).

The classification problem for our naive Bayes model can broadly be defined as follows:

Let $C=(C_1, C_2, \ldots, C_k)$ represent k histone classes and let $X = (X_1, X_2, \ldots, X_m)$ be a set of m features that characterize a test case. The Bayesian network assigns the test case to the class which has the highest posterior probability given the test sequence's feature set, $p(Ci \mid x_1, \ldots, x_m)$. Using Bayes' rule, one can re-write the posterior probability of belonging to class i up to a normalizing constant in terms of a) the likelihood of the features given the histone class, $p(x_1, \ldots, x_m \mid C_i)$, and b) a prior probability of membership to this histone class, $p(C_i)$. In other words,

$$\begin{aligned} p(C_i \mid x_1, \ldots, x_m) &= \frac{p(x_1, \ldots, x_m \mid C_i) \times p(C_i)}{p(x_1, \ldots, x_m)} \\ &\propto p(x_1, \ldots, x_m \mid C_i) \times p(C_i) \end{aligned} \tag{1}$$

Furthermore, since the naive Bayes approach assumes that the features are independent given the class membership, we can decompose the posterior probability in proportionality (1) to a product of the probability of each feature given the class membership:

$$\begin{aligned} p(C_i \mid x_1, \ldots x_m) &\propto p(x_1 \mid C_i)\, p(x_2 \mid C_i) \ldots p(x_m \mid C_i) \times p(C_i) \\ &\propto p(C_i) \prod_{j=1}^{m} p(x_j \mid C_i) \end{aligned} \tag{2}$$

Although the feature independence assumption may be generally not correct for histone genes, it does simplify the classification task because it allows the class-conditional probabilities $p(x_j \mid C_i)$ to be calculated separately for each feature variable, thereby reducing a multidimensional estimation task to a number of one dimensional estimation tasks. Furthermore, the naive Bayes model gives high classification rates despite the model simplification.

We validated our model using "leave-one-out cross-validation", in which each of the 116 histone sequences was reserved for testing and the remaining 115 sequences for training. We implemented our Bayesian network model in MATLAB's Bayesian Network Toolkit (BNT) libraries [20].

3 Results and Discussion

We found that all five mammalian histone gene groups (H1, H2A, H2B, H3, H4) have mutually distinct, prominent and strongly conserved regions with motif modules downstream of the TSS. These are also reasonably well conserved across the analyzed species. We observed that nearly all conserved motif segments were present in the coding portion of most of the histone genes under study. This may be because motifs found in

the coding regions were more strongly conserved in terms of statistical significance as compared to those in the non-coding regions. Most of the histone genes that we analyzed had very small 5' UTR segments and thus the genomic region [-50,+500] that we analyzed in these genes contained mostly the coding region.

Table 1 presents the classification results of our Bayesian network model for histone motif modules. In total, only 3 of the 116 histone sequences (2.6%) were misclassified, and none of the H1 and H2A histones were misclassified. The results show that the naive Bayes model, which exploits histone motif module information, performs well on classifying analyzed histone sequences to their respective histone groups.

Table 1: Classification results of Bayesian network model.

Histone Type	Correct	Incorrect	% Correct prediction
H1:	16	0	100
H2A:	26	0	100
H2B:	31	1	96.9
H3:	17	1	94.4
H4:	23	1	95.8
Total	113	3	97.4

We have proposed a new methodology to model motif modules. Our model probabilistically combines information on an exhaustive set of features including the motif, its strand, its absolute distance from the TSS and mutual spacer distance between motifs. Furthermore, it is composed of trainable parameters, the number of which is linear in the number of Bayesian network variables. Consequently, the time required for training our model parameters is also linear in the number of network variables.

The methodology looks promising and can be modified to search across genomes to recognize the 5' ends of both characterized and uncharacterized genes whose 5' end is similar to histone genes. This way we can search for those genes whose products potentially are similar to histones. We note that an alternative approach for discovering uncharacterized genes based on similarities with the known genes is at the protein level, but such an approach may not be efficient in case of genes with short exons. Our methodology is flexible and can be extended to model the promoter regions located upstream of the TSS. This work is currently underway.

References

1. T. Werner. The state of the art of mammalian promoter recognition. *Briefings in Bioinformatics*, 4(1):22-30, 2003.

2. M.C. Frith, J.L. Spouge, U. Hansen and Z. Weng. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucl. Acids Res.,* 30(14):3214-24, 2002.
3. T.L. Bailey and W.S. Noble. Searching for statistically significant regulatory modules. *Bioinformatics*. 19(2):II16-II25, 2003.
4. W.N. Grundy, T.L. Bailey, C. Elkan and M.E. Baker. Meta-MEME: Motif-based Hidden Markov Models of Biological Sequences. *Computer Appl. in Biosciences*, 13(4):397-406, 1997.
5. D. Doenecke, W. Albig, C. Bode, B. Drabent, K. Franke, K. Gavenis and O. Witt. Histones: genetic diversity and tissue-specific gene expression, a review. *Histochem. Cell Biol*., 107(1):1-10, 1997.
6. A.S. Halees and Z. Weng PromoSer: improvements to the algorithm, visualization and accessibility *Nucl. Acids. Res*., 32:W191-W194, 2004.
7. A. Chong, G. Zhang and V.B. Bajic. FIE2: A program for the extraction of genomic DNA sequences around the start and translation initiation site of human genes. *Nucl. Acids Res*., 31(13):3546-3553, 2003.
8. T.L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., 2:28-36, 1994.
9. T.L. Bailey and M. Gribskov. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 4:48-54, 1998.
10. J.D. Hughes, P.W. Estep, S. Tavazoie and G.M Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae, *J. Mol. Biol*., 296:1205-1214, 2000.
11. C.T. Workman and G.D. Stormo. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput*., pp. 467-478, 2000.
12. D. GuhaThakurta and G.D. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17:608-621, 2001.
13. J.S. Liu, A.F. Neuwald and C.E. Lawrence. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statist*. Assoc., 90:1156-1170, 1995.
14. L. Yang, E. Huang and V.B. Bajic. Some implementation issues of heuristic methods for motif extraction from DNA sequences. *Int.J.Comp.Syst.Signals* (accepted 2004).
15. S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites, *Proc. Int. Conf. Intell. Syst. Mol. Biol*., 8:344-354, 2000.
16. M.C. Frith, U. Hansen, J.L. Spouge and Z.Weng. Finding functional sequence elements by multiple local alignment. *Nucl. Acids Res*., 32(1):189-200, 2004.
17. M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*., 12(5):739-748, 2002.
18. A.P. Dempster, N.M. Laird and D.B. Rubin. Maximum Likelihood from incomplete data via the EM algorithm. *JRSSB*, 39:1-38, 1977.

19. C. Huang and A. Darwiche. Inference in Belief Networks: A Procedural Guide. *Intl. J. Approximate Reasoning*, 11:1-158, 1994.
20. http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html