

A Hidden Markov Model for Gene Function Prediction from Sequential Expression Data

Xutao Deng and Hesham Ali
College of Information Science and Technology
University of Nebraska at Omaha
{xdeng, hali}@mail.unomaha.edu

Abstract

Hidden Markov Models (HMMs) have demonstrated great successes in modeling noisy sequential data sets in the area of speech recognition and protein sequence profiling. Results from association test showed significant Markov dependency in time-series gene expression data, and therefore HMMs would be especially appropriate for modeling gene expressions. In this project, we developed a gene function prediction tool based on profile HMMs. Each function class is associated with a distinct HMM whose parameters are trained using yeast time-series gene expression data. The function annotations of the HMM training set were obtained from Munich Information Centre for Protein Sequences (MIPS) data base. We designed several structural variants of HMMs (single, double-split) and tested each of them on forty function classes each of which includes more than one hundred instances. The highest prediction sensitivity we achieved is 51% by using double-split HMM with 3-fold cross-validation.

1. Introduction

Genome scale sequencing and microarray projects provide a big picture of genome structure and behavior. The goal of this project is to apply machine learning techniques to infer budding yeast *Saccharomyces cerevisiae* gene functions from microarray expression data sets. The prediction is based on the hypothesis that genes with similar function often show similar expression patterns. Previous studies have shown relatively low prediction accuracy by using Support Vector Machines (SVMs) and K-Nearest Neighbors (KNNs) [1]. One main reason for the poor performance is that these methods fail to address the fact that expression

measurements are dependent from each other. Fig. 1 illustrates significant Markov dependency of expression data sets by showing distributions of next expression values conditioned on current expression values [1]. The value of expression is discretized into 10 categories with equal cumulative probability. HMMs are natural options for modeling noisy sequential data sets such as gene expressions. We designed several HMM variants to seek better prediction accuracy.

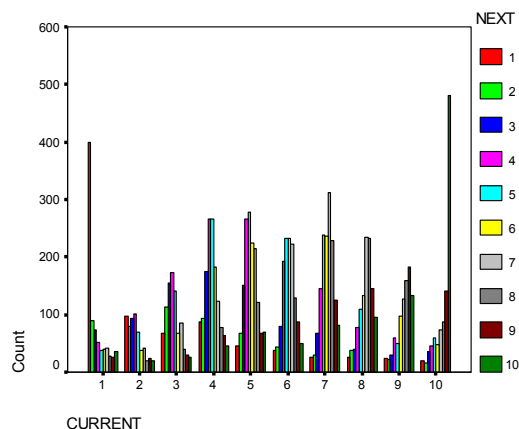


Fig. 1. Conditional Distribution of Gene Expression

2. Methods

The task of function prediction from expression data can be viewed as inference in a Bayesian network (Figure 2) where Class and Expression are random variables. The prior probability density function $P(\text{Class}=\text{class}_i)$ is set to discrete uniform distribution. According to Bayes' Theorem, the posterior probability can be easily computed:

$$P(\text{class}_i|\text{expression}) = \alpha \cdot P(\text{expression}|\text{class}_i) \cdot P(\text{class}_i)$$

Prediction is then made by picking the most likely class given the data. HMM is introduced here

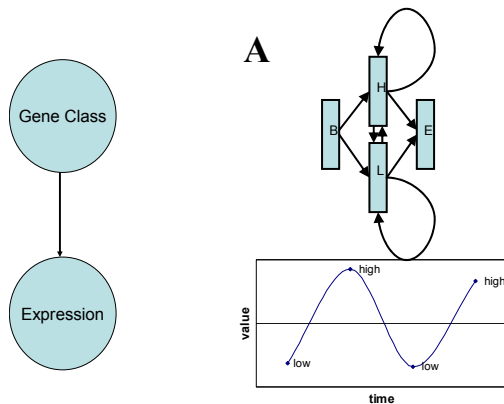


Fig. 2.
Bayesian
Inference

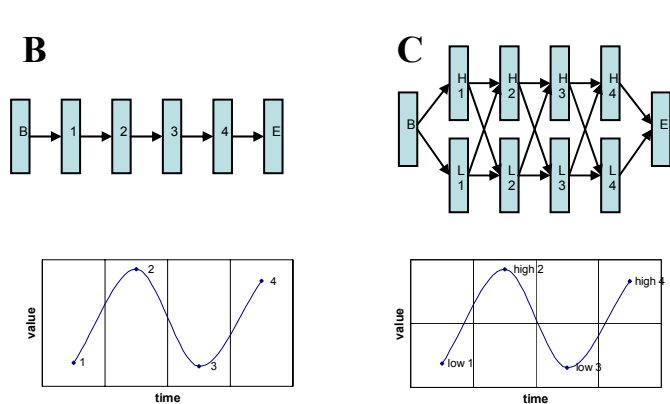


Fig. 3. Three HMMs for Modeling Expression Data. **A.** States defined according to expression value. **B.** States defined based on experiment order. **C.** States define according to both expression value and experiment order. Legends: B: Begin E: End H: High L: Low.

to efficiently generate and retrieve conditional probability $P(expression|class_i)$. Once the parameters of an HMM has been determined from learning data, the probability $P(expression|class_i)$ can then be efficiently computed for each of the $class_i$.

We consider three model structures illustrated in Figure 3. The problem of model A is its ignorance of position specific information which is certainly important. Model B is equivalent to a weight matrix which doesn't consider the non-independence of data. We prefer model C because it avoids the problems of model A and B by combining the value and order as state at the cost of excessive number of parameters to train.

3. Results and Conclusions

We tested 2467 ORFs coming from 40 classes. We obtained gene class label from MIPS [2] as the true class assignment of each gene in [1]. We implemented HMMs Model B (Single) and Model C (Double). Since a gene can have multiple functions, we can make two or three predictions for a single ORF. We refer to this process by double dip or triple dip. We split the data into training and testing sets. The sizes of training sets vary from 20% to 99%. The overall prediction results are showed in Fig. 4. The upper figure shows that double HMM generates higher precision (0.67 maximum) than single HMM. The lower figure shows double HMM generates higher precision than single HMM at the same level of recall but single HMM showed higher recall in general. Specifically, 19 functional classes showed precision greater than 60%. As we expected, the overall prediction accuracy is significantly higher than that of SVMs and KNNs.

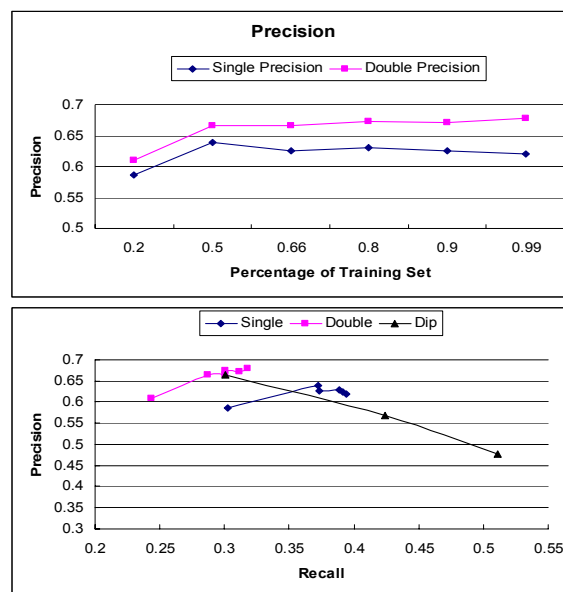


Fig. 4. Testing results of various experiment settings. Note: Precision = True P/TP+FP, Recall=TP/TP+FN

4. References

- [1] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C. Sugnet, T.S. Furey, Jr. M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *PNAS*, 97(1):262–267, 2000.
- [2] Mewes HW, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Münsterkoetter M, Rudd S, Weil B MIPS: a database for genomes and protein sequences. *Nucleic Acids Research* Jan 1;30(1):31-4, 2002.