4-19-2011

Applications of Hidden Markov Models in Microarray Gene Expression Data

Huimin Geng
University of Nebraska at Omaha

Xutao Deng
University of Nebraska at Omaha

Hesham Ali
University of Nebraska at Omaha, hali@unomaha.edu

Recommended Citation

Huimin Geng, Xutao Deng and Hesham H Ali (2011). Applications of Hidden Markov Models in Microarray Gene Expression Data, Hidden Markov Models, Theory and Applications, Dr. Przemyslaw Dymarski (Ed.), InTech, DOI: 10.5772/15194. Available from: https://www.intechopen.com/books/hidden-markov-models-theory-and-applications/applications-of-hidden-markov-models-in-microarray-gene-expression-data

12

Applications of Hidden Markov Models in Microarray Gene Expression Data

Huimin Geng, Xutao Deng and Hesham H Ali
[1]*Department of Computer Science, University of Nebraska at Omaha, Omaha, NE 68182 USA*

1. Introduction

Hidden Markov models (HMMs) are well developed statistical models to capture hidden information from observable sequential symbols. They were first used in speech recognition in 1970s and have been successfully applied to the analysis of biological sequences since late 1980s as in finding protein secondary structure, CpG islands and families of related DNA or protein sequences [1]. In a HMM, the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. In this chapter, we described two applications using HMMs to predict gene functions in yeast and DNA copy number alternations in human tumor cells, based on gene expression microarray data.

The first application employed HMMs as a gene function prediction tool to infer budding yeast *Saccharomyces cerevisiae* gene function from time-series microarray gene expression data. The sequential observations in HMM were the discretized expression measurements at each time point for the genes from the time-series microarray experiments. Yeast is an excellent model organism which has reasonably simple genome structure, well characterized gene functions, and huge expression data sets. A wide variety of data mining methods have been applied for inferring yeast gene functions from gene expression data sets, such as Decision Tree, Artificial Neural Networks, Support Vector Machines (SVMs) and K-Nearest Neighbors (KNNs) [2-4]. However those methods achieved only about 40% prediction precision in function prediction of un-annotated genes [2-4]. Based on our observations, there are three main reasons for the low prediction performance. First, the computational models are too simple to address the systematic variations of biological systems. One assumption is that genes from the same function class will show a similar expression pattern. However, clustering results have shown that functions and clusters have many-to-many relationship and it is often difficult to assign a function to an expression pattern (Eisen et al., supplementary data) [5]. Second, the measurements of expression value are generally not very accurate and show experimental errors (or noise). The observed expression values may not reflect the real expression levels of genes. For example, a correlation as low as 60% was reported between measurements of the same sample hybridized to two slides [6]. Third, none of the above methods explicitly address the less obvious but significant correlation of gene expressions. Our results indicate that the expression value of a gene depends significantly on its previous expression value. Therefore, Markov property can be assumed to simplify the non-independence of gene

expressions. In this application, we developed a gene function classification tool based on HMMs from time-series gene expression profiles (GEP) in yeast and the goal was to provide a better tool for more accurate gene function prediction as compared to other existing tools. We studied 40 yeast gene function classes which have a sufficient number of open reading frames (ORFs) in Munich Information Centre for Protein Sequences (MIPS), for example greater than 100, for the training purpose of HMMs. Each function class was modeled as a distinct HMM and the set of 40 HMMs compose a discriminant model for unknown gene function prediction. Cross-validation showed our HMM-based method outperforms other existing methods by an overall prediction precision of 67%.

In the second application, we developed a DNA copy number alteration (CNA) prediction tool based on HMMs to infer genetic abnormalities in human tumor cells from microarray gene expression data. The sequential observations in this HMM application were the discretized expression measurements for the genes along the chromosomal locations for each chromosome. Instead of the temporal order of gene expression in the time-series experiments in the first application, in the second application we used the spatial order of the genes according to chromosomal locations as the sequential data in HMM. It is well known that chromosomal gains and losses play an important role in regulating gene expression and constitute a key mechanism in cancer development and progression [7, 8]. Comparative Genomic Hybridization (CGH) was developed as a molecular cytogenetic method for detecting and mapping such CNAs in tumor cells [9, 10]. However, in the post-genomic era, the majority of the genome-wide studies in cancer research have been focusing on gene expression but not CGH, and as a result, an enormous amount of GEP data have been accumulated in public databases for various tumor types [11-15], but few CGH studies have been performed in large series of tumor samples [16]. The vast amount of GEP data represents an important resource for cancer research, yet it has not been fully exploited. We hypothesized that with a well-designed computational model, GEP data can be readily used to derive functionally relevant genetic abnormalities in tumors. From the literature review, most studies including GEP and CGH have been focusing on the impact of one on the other or combining the two for identifying candidate tumor suppressor genes or oncogenes [8, 17-25]. In this application, we proposed a novel computational approach based on HMMs to predict CNAs from the GEP data. It would significantly reduce the cost of CNAs detection in tumor cells, and more importantly, it will reveal functionally more relevant CNAs as compared to those identified by CGH, since CGH in principle defines only the structural changes which may or may not reflect functional effects, but GEP-defined CNAs must have the functional effects reflected by changes of gene expression. HMMs have recently been applied in array CGH for segmentation, a procedure to divide the signal ratios of each clone on the array into states, where all of the clones in a state have the same underlying copy number [26, 27]. In this application, HMM was used for an integrative analysis of GEP-to-CGH prediction which intended to capture two primary sources of uncertainty embedded in genomic data: First, the significant but subtle correlations between GEP and CGH; Second, the sequential transitions of DNA copy number changes along a chromosome. The purpose was to enhance the limited CGH data with the wealth of GEP data and provide an integrative genomic-transcriptomic approach for identifying functionally relevant genetic abnormalities in cancer research. In this application, we studied 64 cases of Mantle Cell Lymphoma (MCL) which had both GEP and CGH data associated. Since chromosomal gains and losses occur on individual chromosomes, we developed and trained a separate HMM for each chromosome and the set of 24 HMMs compose a discriminant model for the human

tumor CNA prediction. Using cross validation, the training of the HMMs was done on the paired GEP and CGH data, and the prediction was made for a new tumor sample for its CNAs based on the trained HMMs from its GEP data. Our HMM method achieved 75% sensitivity, 90% specificity and 90% accuracy in predicting CNAs in 64 MCL samples when compared to the CNAs identified by experimental CGH on the same tumor samples.

2. Preliminary of HMM

Transition and emission probabilities

A HMM describes a doubly embedded stochastic process with one observable process {*Oi*} and one hidden process {*Hi*}. The hidden process {*Hi*} is an ordinary Markov model with state transition distribution defined as *aij*:

$$a_{ij} = P(H_n = j \mid H_0, H_1, \ldots, H_{n-1} = i) = P(H_n = j \mid H_{n-1} = i) \tag{1}$$

where ; a_{ij} is the transition probability from hidden state *i* to *j*.

The observable process $\{O_i\}$ is embedded upon the hidden process with a distinct probability distribution e_{ik} defined at each hidden state:

$$e_{ik} = P(O_n = k \mid H_1, H_2, \ldots, H_n = i) = P(O_n = k \mid H_n = i) \tag{2}$$

where e_{ik} is the emission probability of emitting symbol *k* at hidden state *i*.
Together with the initial hidden state distribution π, the HMM with discrete emission symbols can be readily represented as a 5-tuple (K, N, a_{ij}, e_{ik}, π), where K is the number of states of the hidden process and N is the number of observations at each hidden state. Given the observation sequence $O_1\ O_2\ \ldots O_T$, and a HMM, M = (K, N, a_{ij}, e_{ik}, π), we can efficiently compute $P(O_1\ O_2\ \ldots O_T \mid M)$, the probability of observing the sequence, by using the Viterbi or forward algorithms. In a HMM, the challenge is often to determine the hidden parameters from the observable parameters. To estimate the model parameters, when the state paths are known for the training datasets, Maximum Likelihood Estimation (MLE) or Bayesian Maximum A Posteriori (MAP) can be used; if the state paths are unknown, the Baum-Welch algorithm or Viterbi training can be used instead to learn the model parameters using a set of observation sequences. To estimate the hidden state path for a new case, two standard algorithms Viterbi and Forward-Backward can be used. Refer to Koski (2002), Durbin (1989) and Rabiner (1989) for details [1, 28, 29].

Viterbi, Forward and Backward Algorithms

Viterbi decoding is a dynamic programming algorithm. Suppose the probability $v_k(i\text{-}1)$ of the most probable path ending in state k with observation $x_{i\text{-}1}$ is known for all the states k, then the probability $v_l(i)$ corresponding to the observation x_i with the state l can be calculated as in Eq. (3). The entire path π can be found recursively as in Algorithm (1).

$$v_l(i) = e_l(x_i) \max_k (v_k(i-1) a_{kl}) \tag{3}$$

where a_{kl} is the transition probability, $e_l(x_l)$ is the emission probability, k and l are states and x_i is an emission symbol.

Algorithm (1) Viterbi:

Initialization (i=0): $v_0(0)=1$, $v_k(0)=0$ for $k>0$.
Recursion (i=1…L): $v_l(i)=e_l(x_i)\max_k(v_k(i-1)a_{kl})$; $ptr(l)=\arg\max_k(v_k(i-1)a_{kl})$.
Termination: $P(x,\pi^*)=\max_k(v_k(L)a_{k0})$; $\pi_L^*=\arg\max_k(v_k(L)a_{k0})$.
Traceback (i=1…L): $\pi_{i-1}^*=ptr_i(\pi_i^*)$.

Posterior decoding is derived from Forward and Backward algorithms, which are similar dynamic programming procedures to Viterbi, but by replacing the maximization steps with sums to obtain the full probability for all possible paths. In Forward algorithm, $f_k(i)=P(x_1\ldots x_i,\pi_i=k)$ is the forward variable, representing the full probability for all the probable paths ending in state k with observation up to and including x_i. Then $f_l(i+1)$, which corresponds to the observation up to and including x_{i+1} and ending in state l, can be calculated by the recursion in Eq. (4). In Backward algorithm, the backward variable $b_k(i)=P(x_{i+1}\ldots x_L\mid\pi_i=k)$ is analogous to $f_k(i)$, but instead obtained by a backward recursion starting at the end of the sequence as in Eq. (5). The detailed Forward and Backward algorithms are shown in Algorithms (2) and (3).

$$f_l(i+1)=e_l(x_{i+1})\sum_k f_k(i)a_{kl} \tag{4}$$

$$b_k(i)=\sum_l e_l(x_{i+1})a_{kl}b_l(i+1) \tag{5}$$

where a_{kl} is the transition probability, $e_l(x_l)$ is the emission probability, k and l $\in\{H_+,H_-,H_o,L_+,L_-,L_o,M_+,M_-,M_o\}$ and $x_i\in\{H,L,M\}$.

Algorithm (2) Forward:

Initialization (i=0): $f_0(0)=1$, $f_k(0)=0$ for $k>0$.
Recursion (i=1…L): $f_l(i)=e_l(x_i)\sum_k f_k(i-1)a_{kl}$.
Termination: $P(x)=\sum_k f_k(L)a_{k0}$.

Algorithm (3) Backward:

Initialization (i=L): $b_k(L)=a_{k0}$ for $k>0$.
Recursion (i=L-1…1): : $b_k(i)=\sum_l e_l(x_{i+1})a_{kl}b_l(i+1)$.
Termination: $P(x)=\sum_l e_l(x_1)a_{0l}b_l(1)$.

Having $f_k(i)$ and $b_k(i)$, given the emitted sequence x, the posterior probability that observation x_i comes from a state k is shown in Eq. (6). The posterior probability that observation x_i comes from all possible states in the specific set is shown in Eq. (7), where $g(k)$ is a function defined on the states. Then we concatenate the most probable state at each position to form an entire state path.

$$P(\pi_i = k \mid x) = f_k(i)b_k(i) / P(x) \tag{6}$$

$$G(i \mid x) = \sum_k P(\pi_i = k \mid x)g(k) \tag{7}$$

3. Method

Application 1 – Yeast Gene Function Prediction

For a set of time-series expression data, we model the discretized expression measurements at each time point as observed emission symbols. Our goal is to train separate HMMs for each gene function class and use the set of HMMs as a whole discriminant model for unknown gene function prediction.

Data

The expression data set is obtained from http://rana.lbl.gov/EisenData.htm (Eisen, et al., 1998) [5]. The complete data set contains 2,467 genes with each gene having 79 experimental measurements recorded. Among the 2,467 genes, 2,432 have at least one function annotation at MIPS (http://mips.gsf.de/) [30]. For training purpose, we only include 40 function classes which have at least 100 open reading frames (ORFs) in MIPS.

The original data set is organized as a set of pairs $S = \{(\mathbf{X_i}, C_i) \mid 1 \le i \le n\}$ where $\mathbf{X_i}$= [X_{i1}, X_{i2},…, X_{iT}] is the expression vector and C_i is the class label for the ith training sample, T is the number of genes and n is the number of training samples. Expression vectors with the same class label are then grouped into the subset $S_j = \{(\mathbf{X_i}, C_i) \mid C_i = j,\ 1 \le j \le L,\ 1 \le i \le n\}$, where L is the number of classes. The entire prediction process can be performed in three steps: *discretization*, *training* and *inference*. *Discretization* is a data preprocessing step in which the continuous expression measurements are divided up into a fixed set of intervals (symbols). In the *training* step, the parameters of each distinct unit of HMM, M_i, are specified by using each training subset S_i. The whole model $M = \{M_i \mid 1 \le i \le L\}$ is a collection of unit models and used for *inference* of gene functions.

Model Structures

One advantage of HMMs is that designers can choose arbitrary transition and emission structures appropriate for modeling data from specific domains. In this application, two model structures were considered for modeling the sequential time-series microarray gene expression data (Figure 1). Model A defines states based on time sequence of experiments. This model is the backbone of the HMM for protein profiling [1, 31]. The advantage of model A is that the position-specific information is preserved. However, if we take a closer look, all transitions are of probability 1 and the HMM is actually degenerated into a weight matrix which doesn't consider the non-independence of data. To reflect the dependence of the expression data at different time points, we designed model B by combining the expression values and the experiment sequence as states. The state is defined as a pair,

(*value*, *experiment*). If an expression value is below a predefined threshold at time point *i*, the state is "low-*i*"; otherwise it is "high-*i*". Model B is a more complex model which is able to capture the Markov property of expression data. However, model B has more parameters than model A and hence requires a larger number of training samples. The model A is referred to "chain HMM" and model B as the "split HMM".

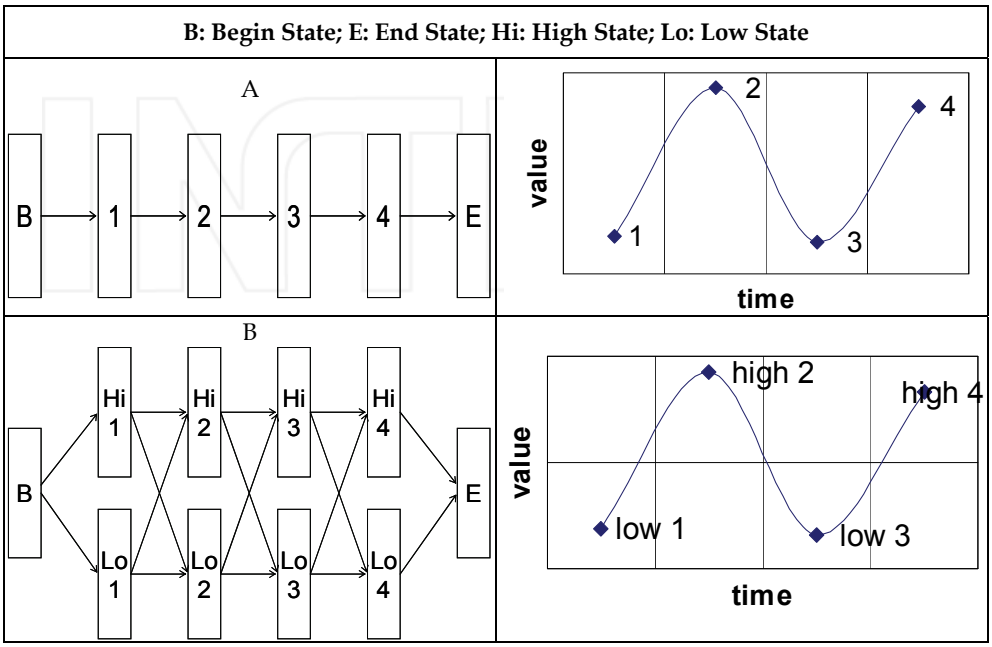


Fig. 1. Two HMM model structures for modeling yeast time-series microarray gene expression data. Left panels are state transition diagrams and right panels are expression patterns. A. States defined based on experiment order (i.e. time points). B. States definition according to both expression value and experiment order. Note that "B" and "E" are special dummy states without emission.

Training

In this section, we describe the training of prior distribution $P(M_i)$, transition distribution (a_{ij}) and emission distribution (e_{ik}) for each gene function class. All the distributions are considered as a multinomial distribution with certain unknown parameters. We will use Bayes' learning to determine the point estimator of all unknown parameters from the data. The prior probability density function $P(M_i)$ is estimated by the Maximum Likelihood Estimator (MLE) of a multinomial distribution as in Eq. (8):

$$\hat{P}(M_i) = \frac{|S_i|}{\sum_{i'=1}^{L} |S_{i'}|} \tag{8}$$

where $|S_i|$ is the number of training samples in the *i*th class and *L* is the number of classes.

The transition probability distribution can be modeled as a multinomial distribution $Mul(a_{i1}, a_{i2}, ..., a_{iK_i})$. We use a Dirichlet distribution $Dir(1, 1, \ldots, 1)$ as the conjugate prior to avoid the problem of zero probability. Therefore, the Mean Posterior Estimator for the transition probability a_{ij} is defined as in Eq. (9):

$$\hat{a}_{ij} = \frac{A_{ij} + 1}{\sum_{j'=1}^{K_i} A_{ij'} + K_i}, \; 1 \le i \le K, \; 1 \le j \le K_i \tag{9}$$

where A_{ij} is the count of transitions from state i to j in the training samples, K is the number of states defined in the HMM, and K_i is the out-degree of state i (i.e., K_i=1 in the chain HMM, , and K_i =2 in the split HMM). Similar to the transition probabilities, emission probabilities can be estimated by counting the frequencies as well in Eq. (10),

$$\hat{e}_{ij} = \frac{E_{ij} + 1}{\sum_{j'=1}^{N} E_{ij'} + N}, \; 1 \le i \le K, \; 1 \le j \le N \tag{10}$$

where E_{ij} is the count of emissions of symbol j at state i in the training samples, K is the number of states defined in the HMM, and N is the total number of symbols at each state. The training was performed for each function class.

Inference

Given an expression vector **X** and the model M (a collection of each unit model M_i), inferring the function of **X** is performed based on Bayes' theorem:

$$P(M_i \mid \mathbf{X}) \propto P(\mathbf{X} \mid M_i) P(M_i) \tag{11}$$

In order to calculate the posterior probability $P(M_i \mid X)$, we need the prior probability of each $P(M_i)$ and the likelihood $P(\mathbf{X} \mid M_i)$. $P(M_i)$ is specified in Eq. (8). The likelihood $P(\mathbf{X} \mid M_i)$ can be considered as the marginal probability of **X** across all hidden state paths π.

$$P(\mathbf{X} \mid M_i) = \sum_{\pi} P(\mathbf{X}, \pi \mid M_i) = \sum_{\pi} a_{0\pi 1} \prod_{l=1}^{T} a_{\pi(l)\pi(l+1)} e_{\pi(l)X(l)} \tag{12}$$

where $\pi(l)$ is the lth hidden state in the path π and $X(l)$ is the symbol at the lth state. Once the parameters of the HMM M_i have been determined from training data, the likelihood term can be efficiently computed by the forward algorithm, as described previous section *Preliminary of HMM*. The inference of function class is then made by choosing the most likely class given the data, as in Eq. (13).

$$M^* = \arg\max_{M_i} P(M_i \mid \mathbf{X}) = \arg\max_{M_i} P(\mathbf{X} \mid M_i) P(M_i) \tag{13}$$

Application 2—Human Tumor CNA Prediction

In the second application, HMMs are used to address the following question: "Given a sequence of gene expression data along chromosomal locations as observations, predict the hidden CGH status of the chromosomal gains or losses."

Model Structure

In the HMM-CNA prediction, the observable process $\{O_i\}$ describes discretized gene expression values of genes along a chromosome, where O_i ="*H*", "*L*" or "*M*" for *high*, *low* or *medium* expression, respectively; the hidden process $\{H_i\}$ describes the underlying CNAs, where H_i ="+", "-" or "*o*" for *gain*, *loss* or *normal* copy number status of a gene, respectively. In Figure 2A, the HMM model was illustrated as a Bayesian network, where the shaded nodes S_1, S_2, …, S_n represent hidden state variables, and the visible nodes E_1, E_2, …, E_n represent the observations for the variables, for the genes along a chromosome. The emission space consists of three symbols {*H*, *L*, *M*} and the hidden state space consists of nine states that the gene expression values superimposed on the CNAs $\{H_+, L_+, M_+, H_-, L_-, M_-, H_o, L_o, M_o\}$, where E_α emits E, $E \in \{H, L, M\}$ and $\alpha \in \{+, -, o\}$. Figure 2B showed the state transition diagram. The model is a single chain incorporating three Markov sub-chains. In each sub-chain, there is a complete set of state transitions, describing the elongation of a DNA segments with a gain, loss or normal copy number. The state transitions between sub-chains are also allowed to describe the state change of a gain, loss or normal CNA. This design of intra- and inter- sub-chain transitions in HMM makes it possible to identify alterative gain, loss and normal regions of variable length automatically.

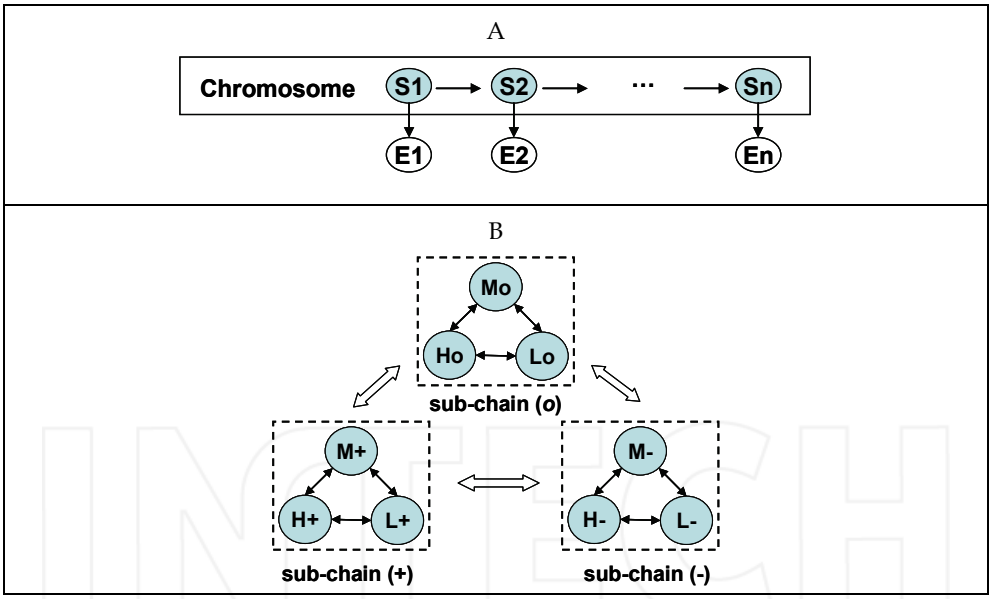


Fig. 2. HMM model structure for human CNA prediction. (A) HMM model presented as a Bayesian network. The shaded nodes S_1, S_2, …, S_n represent hidden state variables and the white nodes E_1, E_2, …, E_n represent the observations for the variables. (B) State transition diagram of HMM-CNA model. The model is a single HMM chain integrating three Markov sub-chains: (+), (-) and (o). In each sub-chain, a Markov chain is graphically shown as a collection of states, with arrows between them describing the state transitions within a gain, loss or normal CNA. There are also arrows between sub-chains, describing the state transitions from a gain, loss or normal CNA to another.

Training

Since chromosomal gains and losses are based on individual chromosomes, we developed and trained a separate HMM for each chromosome accordingly. The whole HMM-CNA prediction model was made of individual HMMs. In our training dataset, given the paired GEP and CGH data, the hidden state path for each observation sequence is known. Therefore, the transition and emission probabilities can be estimated using MLE as in Eq. (9) and (10).

Inference

Having the model parameters trained by training data, we used Viterbi or Posterior (also called Forward and Backward) decoding algorithms [1] to infer hidden CNA states for a new tumor sample based on its GEP observations. Viterbi algorithm works by finding the highest probability path as a whole as a hidden state path, while alternatively, Posterior algorithm finds the most likely state for each position and then concatenate those states as a hidden state path. The detailed algorithms of Viterbi and Posterior decoding were shown in *Preliminary of HMM* section.

An alternative inference method for HMM when given only emissions (i.e. GEP observations) as training data is the Baum-Welch algorithm [1], which estimates the model parameters (transition and emission probabilities) together with unknown CGH states by an iterative procedure. We chose not to use this algorithm as there are many parameters in the model, but relatively few data points at each gene position to estimate these parameters. Instead, we use the true CGH states to guild the HMM prediction using the Viterbi or Posterior algorithms.

Smoothing Algorithm

Since gains and losses identified by our experimental CGH had the resolution on cytobands, we determined the gains and losses on cytoband resolution as well by applying the following smoothing method. Basically, a multinomial probability was used to measure the likelihood of a cytoband harboring a gain/loss or not. In Eq. (14), L is the likelihood under a hypothesis H, where H_1 for the alternative hypothesis that "a cytoband is harbouring a gain or loss", and H_0 for the null hypothesis that "a cytoband is not harbouring a gain or loss"; n_+, n_-, and n_o are the numbers of genes in the gain, loss and normal status within this cytoband, and n is the total number of genes on this cytoband, $n=n_+ + n_- + n_o$; θ_+, θ_- and θ_o are the corresponding multinomial parameters which can be estimated using MLE in Eq.(15). Under H_1, $\theta_{1,+}$, $\theta_{1,-}$ and $\theta_{1,o}$ are estimated by the number of genes n_+, n_- and n_o on a cytoband ($n=n_+ + n_- + n_o$), while under H_0, $\theta_{0,+}$, $\theta_{0,-}$ and $\theta_{0,o}$ are estimated by the number of genes N_+, N_- and N_o on the whole genome as background ($N=N_+ + N_- + N_o$). Log-of-odds (LOD), which is $Log10$ of the ratio of the two likelihoods, was used to measure how likely that the cytoband harbors a gain or loss in Eq. (16). The higher the LOD score, the more likely this cytoband harbors a genomic gain or loss.

$$L(n_+, n_-, n_o \mid H) = \frac{n!}{n_+! n_-! n_o!} \theta_+^{n_+} \theta_-^{n_-} \theta_o^{n_o} , \tag{14}$$

$$\theta_{1,+} = \frac{n_+}{n},\ \theta_{1,-} = \frac{n_-}{n},\ \theta_{1,o} = \frac{n_o}{n} \text{ and } \theta_{0,+} = \frac{N_+}{N},\ \theta_{0,-} = \frac{N_-}{N},\ \theta_{0,o} = \frac{N_o}{N} \tag{15}$$

$$LOD = \log_{10} \frac{L(n_+, n_-, n_o \mid H_1)}{L(n_+, n_-, n_o \mid H_0)} = \log_{10} \frac{\theta_{1,+}^{n_+} \theta_{1,-}^{n_-} \theta_{1,o}^{n_o}}{\theta_{0,+}^{n_+} \theta_{0,-}^{n_-} \theta_{0,o}^{n_o}} \tag{16}$$

Other Simple Methods

To compare with HMM, we also made two other simple methods, rGEP (raw GEP) and sGEP (smoothing GEP), to map the GEP status to CGH status without a sophisticated learning and inference process. By rGEP, we mean that a high expression status of a gene is mapped to a gain (i.e. "*H*" → "+"), low expression to a loss (i.e. "*L*" → "-"), and medium expression to a normal (i.e. "*M*" → "*o*") status. In sGEP, a smoothing method (a multinomial model, as described above) was applied after rGEP to get a gain or loss status for a cytoband across a number of consecutive genes.

Data

The data we used in this application include 64 MCL samples performed with both GEP and CGH experiments [22]. The GEP data were obtained using Affymetrix HG-U133 plus2 arrays and normalized (global median normalization) using BRB-Array Tool [32]. 1.5-fold change was adopted to determine high (>1.5 fold increase), low (>1.5 fold decrease) or medium (<1.5 fold change) expression of each gene in a tumor case as compared to the median expression of this gene across all tumor cases. The CGH experiments were performed by Vysis CGH kits (Downers Grove, IL). aCGH-Smooth [33] was used to determine breakpoints and relative levels of DNA copy number. The company recommended 1.25 and o.75 signal ratio of tumor to normal cells was used to segregate gain (>1.25), loss (<0.75) and normal (between 0.75 and 1.25) regions. Small-sized chromosomes and sex chromosomes were excluded from the study due to technical limitation and lack of gender data, including chromosomes 19-22, X and Y. The chromosomal locations of genes and cytobands were obtained by Affymetrix probesets alignments and NCBI Human Genome database Build 36.1. The LOD score of 2 was used as the cutoff to call a gain or loss for a cytoband after the smoothing algorithm.

4. Results

Evaluation criteria

The prediction performs were evaluated using cross validation, in which the set of samples was divided up into the training set and the test set. The total *n* samples were randomly split into *k* subsets of equal size. We use each subset as the test set and the other *k*-1 subsets as the training set (which is called *k*-fold cross validation). In the extreme case where one case was used in the test set and all the other *n*-1 cases were used in training, it is called leave-one-out cross validation (LOOCV). *K*-fold cross validation or LOOCV were used to validate the HMMs in the two studies.

The performance of predictions was evaluated using the criteria of *precision*, *recall, sensitivity*, *specificity* and *accuracy*, defined as below:

$$precision \equiv \frac{|TP|}{|TP| + |FP|} \tag{17}$$

$$recall \equiv sensitivity \equiv \frac{|TP|}{|TP| + |FN|} \tag{18}$$

$$specificity \equiv \frac{|TN|}{|TN|+|FP|} \tag{19}$$

$$accuracy \equiv \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|} \tag{20}$$

where $|TP|$ means the number of true positive, $|FP|$ for false positive, $|TN|$ for true negative, and $|FN|$ for false negative.

Application 1—Yeast Gene Function Prediction

Many factors affect prediction performance, such as the model structure, the size of training data and the number of predictions made. We performed experiments for different settings and search for the setting at which the best performance was achieved. A graphical output of a trained split HMM model is shown in Figure 3 for function class 32 (cell rescue, defense and virulence in Table 1). Only the first eight expression measurements are shown. Six emission symbols plus a missing symbol are color-coded corresponding to the relative expression level in the microarray image. The thickness of each edge represents the transition probability. The width of each vertical bar represents the probability of each symbol at a specific state. It is obvious that in the high expression value state (at top of the chain), the observed expression measurements are also relatively high (the bar widths are wider at 3 to 6); while in the low expression value state (at bottom of the chain), the observed expression measurements are also relatively low (the bar widths are wider at 0 to 2).

Since a single gene may have multiple functions, we can make multiple predictions for a testing gene. Besides choosing the function class with the highest posterior probability, we

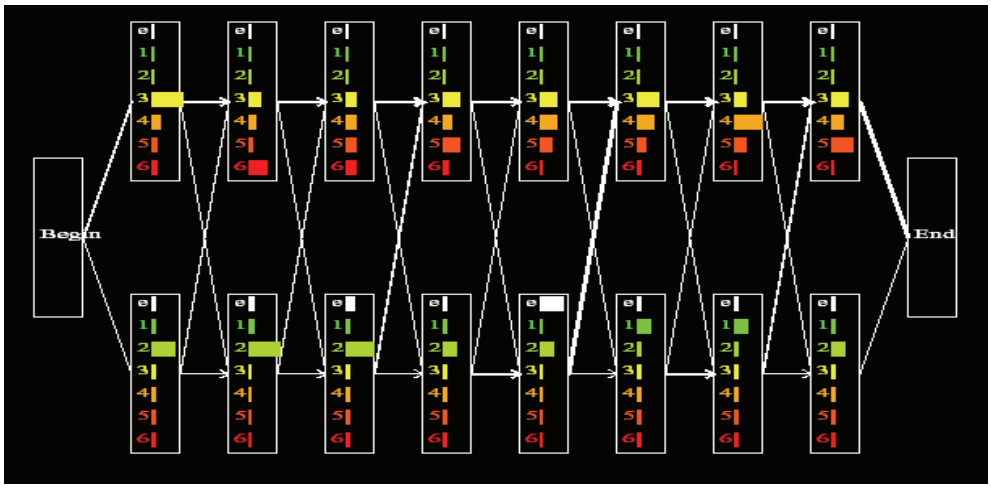


Fig. 3. HMM training results for class 32 (cell rescue, defense and virulence). Only the first eight expression measurements are showed here. Six emission symbols plus a missing symbol are color-coded corresponding to the relative expression level in the microarray image (red high expression and green low expression). The thickness of each edge represents the transition probability. The width of each vertical bar represents the probability of each symbol at a specific state.

can choose the second highest, the third highest, and so on. This procedure is termed single-dip, double-dip, and trip-dip, etc. The prediction performance at different dips was shown in Figure 4c. The *precisions* achieved for single- and double- dip settings are about 60% or higher and for triple-dip is about 50%. The *recalls* are around 30% for single-dip, but can reach 51% when multiple predictions are made (triple-dip). The overall prediction accuracy is significantly higher than that of SVMs and KNNs (40% precisions, 30% recalls). From Figure 4c, we can also see that *precision* and *recall* are inherently conflicting measures such that improving one will often be at a cost of the decrease of the other if other conditions are same.

From Figure 4a and Figure 4b, the number of *TP*s, the number of predictions (*TP*s+*FP*s), *precision* and *recall* generally increase as the size of training set increases. At the *n*-fold cross validation (i.e. LOOCV), HMM-based method achieved an overall prediction *precision* of 67%, which outperforms other existing methods (40% *precisions* in SVMs and KNNs). From Figure 4a and 4b, we also observed that the split HMMs seem to be a more conservative method than the chain HMMs. At the same level of fold, the chain HMMs tend to have higher *TP*s and *recalls* than the split HMMs, but lower *precisions* than the split HMM.

Precision is the most important evaluation measurement for prediction, because it tells directly how likely the prediction is correct. *Recall*, on the other hand, tells how sensitive the prediction is. Figure 4d shows that the split HMMs generate higher *precisions* than the chain HMMs at the same level of *recall*. However, in general the chain HMMs show higher *recalls* than the split HMMs. Table 1 shows the detailed prediction results for each function class of genes. Only five classes have a precision less than 60%.

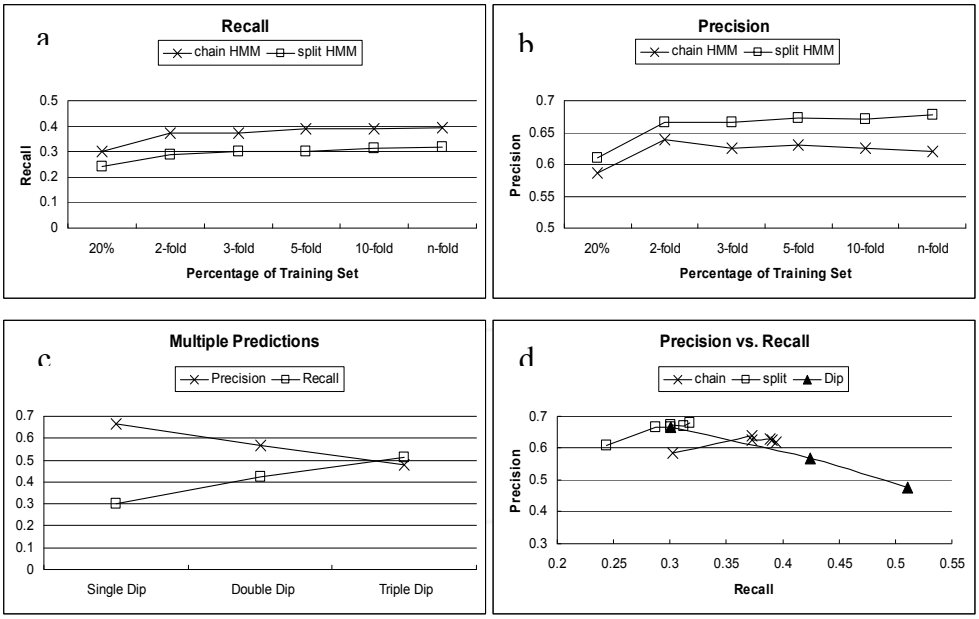


Fig. 4. Testing results of various experiment settings. ***a***. comparison of *recalls* between chain HMM and split HMM. ***b***. comparison of *precisions* between chain HMM and split HMM. ***c***. pattern of *precision*s and *recalls* for multiple predictions. ***d***. relationship between *precisions* and *recalls*.

Function Entry	Annotation	|TP|	|TP|+|FP|	|TP|+|FN|	Precision	Recall
01	metabolism	199	295	631	67.5%	31.5%
01.01	amino acid metabolism	24	29	137	82.8%	17.5%
01.03	nucleotide metabolism	15	24	94	62.5%	16.0%
02	energy	72	105	159	68.6%	45.3%
10	cell cycle and DNA processing	203	342	388	59.4%	52.3%
10.01	DNA processing	42	47	186	89.4%	22.6%
10.01.03	DNA synthesis and replication	3	3	77	100%	3.9%
10.03	cell cycle	44	64	265	68.8%	16.6%
10.03.01	mitotic cell cycle, cell cycle control	16	17	217	94.1%	7.4%
10.03.02	meiosis	1	1	44	100%	2.3%
11	transcription	346	623	553	55.5%	62.6%
11.02	RNA synthesis	92	157	354	58.6%	26.0%
11.04	RNA processing	56	76	170	73.7%	32.9%
11.04.01	RNA processing	21	34	54	61.8%	38.9%
11.04.03	mRNA processing (splicing, 5'-, 3'-end proc)	16	16	100	100%	16.0%
12	protein synthesis	162	204	300	79.4%	54.0%
12.01	ribosome biogenesis	105	108	176	97.2%	59.7%
14	protein fate (folding, modification, destination)	176	285	448	61.8%	39.3%
14.07	protein modification	16	16	139	100%	11.5%
14.13	protein degradation	40	43	116	93.0%	34.5%
14.13.01	cytoplasmic and nuclear protein degradation	29	29	83	100%	34.9%
20	cellular transport, transport facilitation and routes	207	392	431	52.8%	48.0%
20.01	transported compounds (substrates)	36	42	151	85.7%	23.8%
20.03	transport facilitation	7	7	79	100%	8.9%
20.09	transport routes	64	112	367	57.1%	17.4%
20.09.18	cellular import	5	5	95	100%	5.3%
32	cell rescue, defense and virulence	43	56	143	76.8%	30.1%
32.01	stress response	21	27	92	77.8%	22.8%

Function Entry	Annotation	|TP|	|TP|+|FP|	|TP|+|FN|	Precision	Recall
32.07	detoxification	5	6	49	83.3%	10.2%
34	interaction with the cellular environment	42	45	212	93.3%	19.8%
34.01	ionic homeostasis	9	9	89	100%	10.1%
34.01.01	homeostasis of cations	3	3	83	100%	3.6%
34.11	cellular sensing and response	21	22	125	95.5%	16.8%
34.11.03	chemoperception and response	21	22	125	95.5%	16.8%
42	biogenesis of cellular components	72	108	263	66.7%	27.4%
42.01	cell wall	7	7	84	100%	8.3%
42.04	cytoskeleton	6	7	71	85.7%	8.5%
42.16	mitochondrion	36	46	71	78.3%	50.7%
43	cell type differentiation	12	12	193	100%	6.2%
43.01	fungal/microorganismic cell type differentiation	12	12	193	100%	6.2%
Total		2307	3458	7607	66.7%	30.3%

Table 1. Detailed prediction results on 40 gene classes in yeast time-series gene expression dataset [5]. For training purposes in HMM, only 40 gene classes were included which have at least 100 ORFs in MIPS. The results were based on the split-HMMs using 3-fold cross validation.

Application 2 – Human Tumor CNA Prediction

Using cross validation, HMM-CNA was applied to 64 MCLs, on which both GEP and CGH experiments were performed [22]. The entire dataset was split into training and testing datasets. In the training dataset, the HMM model was trained by the paired GEP and CGH data on the same tumor samples, and in the testing dataset, the specified HMM model was applied to the GEP data for a new tumor sample to predict its CNAs. The predicted gains and losses were compared with those identified by experimental CGH on the gene level for the sensitivities and specificities, and on the cytoband level for the recurrent genetic abnormalities.

Gene-Level Validation

We first evaluated HMM Viterbi decoding method by *sensitivity*, *specificity* and *accuracy* against experimental CGH in predicting gain and loss of each gene for all the samples using LOOCV. For comparison purpose, the performance of rGEP and sGEP methods were also included. Table 2 summarized the average *sensitivity*, *specificity* and *accuracy* for all chromosomes on 64 MCL samples. Figures 5 showed the performance on individual chromosomes for *sensitivity* (A) and *specificity* (B). In general, *sensitivity* was improved from 40% in GEP to 45% in sGEP and to 75% in HMM, and *specificity* from 70% in GEP to 85% in sGEP and to 90% in HMM, in predicting gain; in predicting loss, *sensitivity* from 30% in GEP to 50% in sGEP and to 60% in HMM, and *specificity* from 80% in GEP to 90% in sGEP and

HMM. These results suggested that the HMM were able to capture the hidden genomic CNA information buried in the GEP data; while directly mapping GEP status to CGH status without any learning process, such as rGEP and sGEP methods, could not predict well.

		Sensitivity (%)	Specificity (%)	Accuracy (%)
Gain	rGEP	38.45± 2.93	71.43± 0.58	69.46± 1.50
	sGEP	42.77± 10.42	86.33± 2.97	83.59± 3.48
	HMM	**74.49± 17.77**	**88.56± 4.78**	**87.50± 5.59**
Loss	rGEP	28.26± 3.70	80.94± 0.70	77.71± 3.47
	sGEP	50.66± 24.13	**92.71± 2.00***	89.19± 3.63
	HMM	**59.63± 17.22**	90.63± 4.91	**89.30± 5.69**

Table 2. Sensitivity, specificity and accuracy of HMM, rGEP and sGEP as compared to experimental CGH.

The HMM prediction were good for the majority of chromosomes, but we noticed that on some chromosomes HMM prediction was not good, such as chromosomes 1, 6, 9, 10 and 13 for gain and chromosomes 4, 5, 15 and 18 for loss. This is due to infrequent aberrations and hence insufficient training data for the gains or losses on those chromosomes. For example, in the CGH data of the 64 MCL cases, only one, three, one, two and one cases were observed with gain on chromosomes 1, 6, 9, 10 and 13, respectively, and two, one, one and two cases with loss on chromosomes 4, 5, 15 and 18, respectively.

Cytoband-Level Validation

Cytobands are defined as the chromosomal areas distinguishable from other segments by appearing darker or lighter by one or more banding techniques for karyotype description. To compare HMM prediction with the "gold standard" experimental CGH on the same resolution (our experimental CGH detected gains and losses on cytobands), we also determined cytoband-level gains and losses from HMM by applying a smoothing algorithm as described in *Method* section.

Figures 6 showed the results of cytoband level gains and losses on MCL dataset. The two HMM decoding methods, Viterbi and Posterior, were shown in panels A and B, respectively, where loss frequencies for cytobands (i.e. the number of cases harboring a loss on a cytoband) were shown on left-sided bars and gain frequencies on right-sided bars. In Posterior decoding (Figure 6B), as expected, the frequencies of gains and losses decrease as posterior probability increases (p=0.5, 0.6, 0.7, 0.8 and 0.9), and those frequencies are highly correlated (Pearson's correlation coefficients around 0.99, Table 3). Comparing the results from Viterbi (panel A) and Posterior (panel B), a high concordance was also observed (Pearson's correlation coefficients around 0.98, Table 3). Therefore, the Viterbi method was used to represent HMM in comparison of the experimental CGH side by side in panel C.

Table 3 showed that Pearson's correlation coefficients between HMM and CGH are around 0.8 for gains and losses. In Figure 6C, gains and losses were shown separately with CGH results colored yellow above X axis and HMM results colored red below X axis. Apparently, the majority of the frequent gains and losses predicted by HMM are in good concordance with those identified by experimental CGH, such as gains of 3q, 7, 8q, 15q and 18 and losses

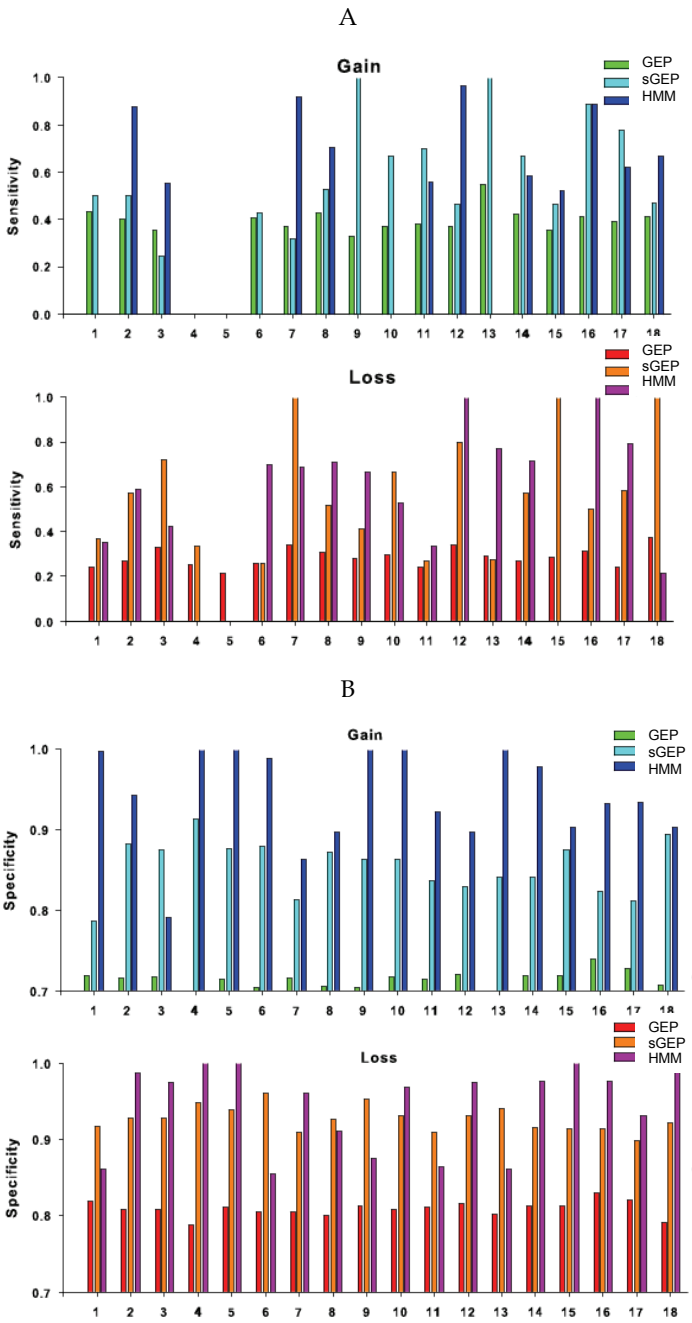


Fig. 5. Sensitivity (A) and specificity (B) in predicting gain and loss regions by GEP, sGEP and HMM in 64 MCL.

of 1p21-p31, 6q, 8p, 9p, 9q, 11q21-q23, 13 and 17p13-pter. Those regions have also been revealed as high-frequency chromosomal alteration regions in various studies using conventional cytogenetics, CGH and array CGH {Bea, 2005 #38}.

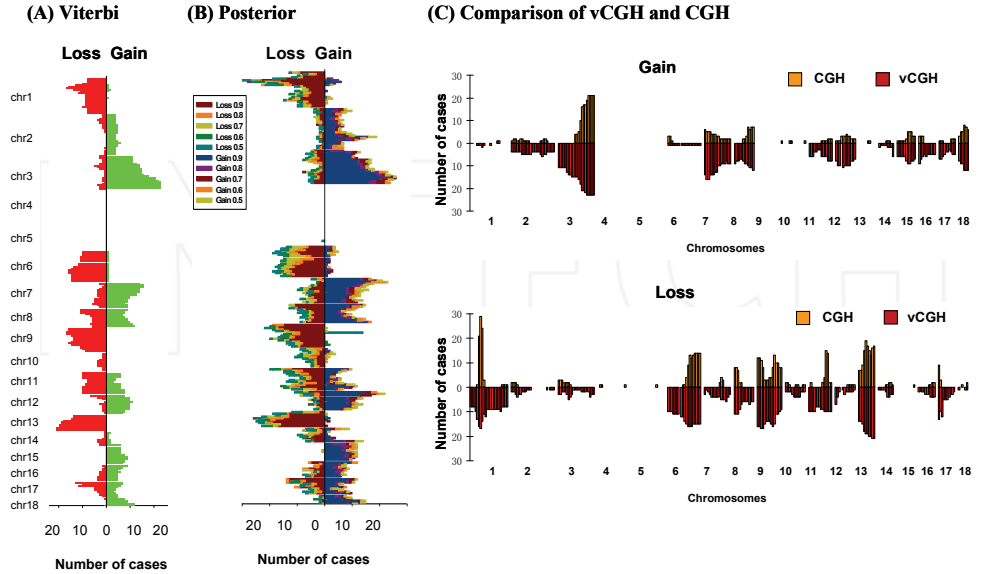


Fig. 6. Cytoband-level gains and losses by HMM and CGH on 64 MCLs. (A) HMM Viterbi method. (B) HMM Posterior method with a series of probability cutoffs (p=0.5, 0.6, 0.7, 0.8 and 0.9). In (A) and (B), left-sided bars correspond to losses whereas right-sided bars to gains. On Y axis are the cytobands ordered from pter to qter (from top to bottom) for each chromosome and on X axis, the length of each bar indicates the gain and loss frequencies, i.e. the number of cases harboring a gain or loss on a cytoband. (C) Comparison of HMM and CGH where Viterbi method was used to represent HMM. Gain and loss were shown separately. CGH results were shown in yellow (above X axis) and HMM prediction were shown in red (below X axis). On X axis, each bar represents a cytoband ordered from pter to qter for chr1 to chr18. On Y axis, the height of each bar indicates frequency, i.e. the number of cases horboring a gain or loss on a cytoband.

There are also some limitations of the approach due to utilization of transcripts-based GEP data. For example, it may not predict well for regions with few genes (also called "gene desert"), or if the genes in a region are not expressed at a sufficiently high level for GEP. The HMM approach is also limited by the design of the GEP arrays. For example, on Affymetrix HG-U133 plus 2 platform, there are no probes distributed on the p arms of chromosomes 13, 14, 15, 21 and 22, and hence those regions are unpredictable for gains or losses by HMM.

5. Conclusions

In this chapter, we demonstrated two applications using HMMs to predict gene functions and DNA copy number alternations from microarray gene expression data. In the first application of HMM on yeast time-series expression data, the overall prediction accuracy of

		CGH	HMM Viterbi	HMM Posterior				
				p_0.5	p_0.6	p_0.7	p_0.8	p_0.9
CGH		1	0.766	0.734	0.745	0.744	0.752	0.756
HMM Viterbi		0.828	1	0.970	0.978	0.978	0.978	0.973
HMM Posterior	p_0.5	0.831	0.978	1	0.990	0.986	0.982	0.969
	p_0.6	0.828	0.980	0.996	1	0.996	0.991	0.978
	p_0.7	0.828	0.983	0.992	0.995	1	0.993	0.981
	p_0.8	0.827	0.985	0.988	0.992	0.996	1	0.990
	p_0.9	0.820	0.981	0.983	0.986	0.991	0.993	1

Table 3. Pearson correlation of cytoband-level gain and loss frequencies between CGH, HMM Viterbi, and HMM Posterior. Gain were shown in the bottom, and loss in the top triangles.

our model is significantly higher than that of SVMs and KNNs. A properly designed HMM is capable of modeling the three features of expression data: pattern variation within a class, experimental noise, and Markov property. The extraordinary flexibility of HMMs allows us to extend the current model in several directions. One of the current developments is to extend the HMM model to Bayes nets for functional prediction from heterogeneous data sets. With appropriately constructed conditional distribution, all kinds of available data can be applied for training and inference using Bayesian network model. Besides the functional prediction of unknown ORFs, one potential application of this method is to search for ORFs of functional homology in databases. To do this, a proper cut-off likelihood value (or equivalently a LOD score) must be specified. Whenever a homolog is found (i.e., beyond the cut-off value), it may be included as a training set to reinforce the training process.

In the second application of HMM on the human tumor microarray gene expression data, we proposed a novel computational approach based on HMMs to predict genetic abnormalities in tumor cells. Taking advantage of the rich GEP data already publicly available, HMM may significantly enhance the identification of genetic abnormalities in cancer research. Our model is among the first which employed HMM for the purpose of GEP-to-CGH prediction. We expected the HMM to capture two primary sources of uncertainty embedded in genomic data: the significant but subtle correlations between GEP and CGH, and the sequential transitions of CNAs along chromosomes. We applied the HMM model to 64 MCL samples and using cross validation, HMM achieved 80% sensitivity, 90% specificity and 90% accuracy in predicting gains and losses as compared to the experimental CGH on the same tumor cases. The recurrent gains and losses predicted by HMM on cytobands were concordant with those identified by CGH. In addition, our model does not only highlight DNA CNA regions but also served as an integrative tool cross-linking genomic and transcriptiomic data for functionally relavent genomic abnormal regions. As this HMM-based method is a general computational tool which can be applied to any types of tumors, it may significantly enhance the identification of genetic abnormalities in cancer research. To improve the model, we plan to add relevant biological parameters to preprocess or filter the data in prediction. We will consider gene densities, transcriptional units, regional epigenomic silencing, genes that not expressed in normal samples, common "genomic aberrations", and human genomic copy number polymorphism in the model.

6. References

[1] Durbin R, Eddy S, Krogh A, Mitchison G: Biological sequence analysis: probabilistic models of proteins and necleic acids. New York: Cambridge Unisersity Press; 1998.

[2] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., Haussler D: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America* 2000, 97(1):262-267.

[3] Kuramochi M, Karypis G: Gene Classification Using Expression Profiles: A Feasibility Study. In: *Proceedings of the 2nd IEEE International Symposium on Bioinformatics & Bioengineering (BIBE 2001).* 2001: 191.

[4] Pavlidis P, Weston J, Cai J, Grundy WN: Gene function classification from heterogeneous data. In: *Proceedings of the 5th International Conference on Computational Molecular Biology (RECOMB 2001).* 2001: 242-248

[5] Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 1998, 95(25):14863-14868.

[6] Churchill GA: Fundamentals of experimental design for cDNA microarrays. *Nature genetics* 2002, 32 Suppl:490-495.

[7] Cahill DP, Kinzler KW, Vogelstein B, Lengauer C: Genetic instability and darwinian selection in tumours. *Trends Cell Biol* 1999, 9(12):M57-60.

[8] Phillips JL, Hayward SW, Wang Y, Vasselli J, Pavlovich C, Padilla-Nash H, Pezullo JR, Ghadimi BM, Grossfeld GD, Rivera A *et al*: The consequences of chromosomal aneuploidy on gene expression profiles in a cell line model for prostate carcinogenesis. *Cancer Res* 2001, 61(22):8143-8149.

[9] du Manoir S, Speicher MR, Joos S, Schrock E, Popp S, Dohner H, Kovacs G, Robert-Nicoud M, Lichter P, Cremer T: Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. *Human genetics* 1993, 90(6):590-610.

[10] Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (New York, NY* 1992, 258(5083):818-821.

[11] Stanford Microarray Database [http://smd.stanford.edu/]

[12] Gene Expression Omnibus, NCBI [http://www.ncbi.nlm.nih.gov/geo/]

[13] UPenn RAD database [http://www.cbil.upenn.edu/RAD/php/index.php]

[14] caArray, NCI [https://cabig.nci.nih.gov/tools/caArray]

[15] ArrayExpress at EBI [http://www.ebi.ac.uk/microarray-as/ae/]

[16] SKY/M-FISH & CGH Database at NCBI [http://www.ncbi.nlm.nih.gov/sky/]

[17] Bea S, Zettl A, Wright G, Salaverria I, Jehn P, Moreno V, Burek C, Ott G, Puig X, Yang L *et al*: Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood* 2005, 106(9):3183-3190.

[18] Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, Ringner M, Sauter G, Monni O, Elkahloun A *et al*: Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res* 2002, 62(21):6240-6245.

[19] Iqbal J, Kucuk C, Deleeuw RJ, Srivastava G, Tam W, Geng H, Klinkebiel D, Christman JK, Patel K, Cao K *et al*: Genomic analyses reveal global functional alterations that

promote tumor growth and novel tumor suppressor genes in natural killer-cell malignancies. *Leukemia* 2009, 23(6):1139-1151.

[20] Lenz G, Wright GW, Emre NC, Kohlhammer H, Dave SS, Davis RE, Carty S, Lam LT, Shaffer AL, Xiao W *et al*: Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proceedings of the National Academy of Sciences of the United States of America* 2008, 105(36):13520-13525.

[21] Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America* 2002, 99(20):12963-12968.

[22] Salaverria I, Zettl A, Bea S, Moreno V, Valls J, Hartmann E, Ott G, Wright G, Lopez-Guillermo A, Chan WC *et al*: Specific secondary genetic alterations in mantle cell lymphoma provide prognostic information independent of the gene expression-based proliferation signature. *J Clin Oncol* 2007, 25(10):1216-1222.

[23] Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle A, Krahe R: Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98(3):1124-1129.

[24] Linn SC, West RB, Pollack JR, Zhu S, Hernandez-Boussard T, Nielsen TO, Rubin BP, Patel R, Goldblum JR, Siegmund D *et al*: Gene expression patterns and gene copy number changes in dermatofibrosarcoma protuberans. *The American journal of pathology* 2003, 163(6):2383-2395.

[25] Varis A, Wolf M, Monni O, Vakkari ML, Kokkola A, Moskaluk C, Frierson H, Jr., Powell SM, Knuutila S, Kallioniemi A *et al*: Targets of gene amplification and overexpression at 17q in gastric cancer. *Cancer Res* 2002, 62(9):2625-2629.

[26] Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN: Hidden Markov Models Approach to the Analysis of Array CGH Data. *J Multivariate Anal* 2004, 90:132-153.

[27] Marioni JC, Thorne NP, Tavare S: BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics (Oxford, England)* 2006, 22(9):1144-1146.

[28] Koski T: Hidden Markov Models of Bioinformatics: Kluwer Academic Pub.; 2002.

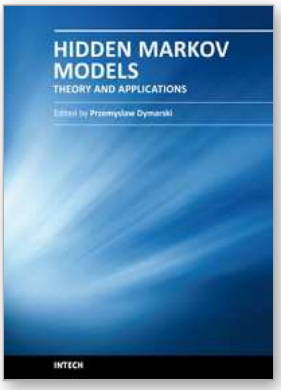
[29] Rabiner L: A tutorial on hidden Markov models and selected applications in speech recognition. In: *Proc IEEE.* vol. 77; 1989: 257-286.

[30] Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 2002, 30(1):31-34.

[31] Eddy SR: Profile hidden Markov models. *Bioinformatics (Oxford, England)* 1998, 14(9):755-763.

[32] BRB-Array Tool [http://linus.nci.nih.gov/BRB-ArrayTools.html]

[33] Jong K, Marchiori E, Meijer G, Vaart AV, Ylstra B: Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics (Oxford, England)* 2004, 20(18):3636-3637.



Hidden Markov Models, Theory and Applications

Edited by Dr. Przemyslaw Dymarski

ISBN 978-953-307-208-1
Hard cover, 314 pages
Publisher InTech
Published online 19, April, 2011
Published in print edition April, 2011

Hidden Markov Models (HMMs), although known for decades, have made a big career nowadays and are still in state of development. This book presents theoretical issues and a variety of HMMs applications in speech recognition and synthesis, medicine, neurosciences, computational biology, bioinformatics, seismology, environment protection and engineering. I hope that the reader will find this book useful and helpful for their own research.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Huimin Geng, Xutao Deng and Hesham H Ali (2011). Applications of Hidden Markov Models in Microarray Gene Expression Data, Hidden Markov Models, Theory and Applications, Dr. Przemyslaw Dymarski (Ed.), ISBN: 978-953-307-208-1, InTech, Available from: http://www.intechopen.com/books/hidden-markov-models-theory-and-applications/applications-of-hidden-markov-models-in-microarray-gene-expression-data

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821