OMICS A Journal of Integrative Biology
Volume 10, Number 3, 2006
© Mary Ann Liebert, Inc.

Hidden Markov Model for Defining Genomic Changes in Lung Cancer Using Gene Expression Data

CHIANG-CHING HUANG,[1] JEREMY M.G. TAYLOR,[2] DAVID G. BEER,[3] and SHARON L.R. KARDIA[4]

ABSTRACT

The study of gene expression patterns in relationship to chromosomal position, the "transcriptome map," has become an area of active research and has revealed unexpected chromosomal regions within which gene expression levels are highly correlated. In cancer research, these regional changes in gene expression that may result from alterations at the chromosome level such as gene amplification or loss. To facilitate the search for such regions utilizing gene expression data, we have developed a hidden Markov model (HMM). Maximum penalized likelihood is used to estimate the parameters in the model. This method is applied to a lung cancer microarray experiment, including 86 human lung adenocarcinomas. Several regions identified through the HMM are consistent with known recurrent regions of amplification or deletion in this cancer. We further demonstrate the association of these abnormal expression regions with measures of disease status, such as tumor stage, differentiation, and survival. These findings suggest that genes in these regions may play a major role in the process of carcinogenesis of the lung. Our proposed method provides a valuable tool to accurately pinpoint regions of abnormal expression for further investigation.

INTRODUCTION

Recent advances in DNA microarray technologies and the abundance of genomic information have provided unprecedented opportunities to decipher the underlying molecular mechanisms related to disease physiology. However, the exploding amount of gene expression data gathered poses great challenges to the scientific community and may provide misleading information without proper statistical interpretation. Consequently, the rich information provided in gene expression data and its broad biological implications will require new methods of analysis. These include employing sophisticated statistical and machine learning algorithms and intensive interaction with other genomic data sources to discover complex gene expression patterns and to correlate gene expression patterns with other biological processes. Recent examples of this include: combining of transcription factor binding and gene expression data to define transcriptional networks (Gao et al., 2004; Lee et al., 2002); utilizing DNA sequence, protein interaction, and microarray data to model genetic regulatory networks (Tamada et al., 2003; Nariai et al., 2004); and incorporating gene function data into microarray analysis for unknown gene functional inference (Zhou et al.,

[1]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois.
Departments of [2]Biostatistics, [3]Surgery, and [4]Epidemiology, University of Michigan, Ann Arbor, Michigan.

2002; Cui et al., 2004). The advantage of these approaches over the analysis of gene expression data alone is that, they can not only reduce false positive findings, but also render biological significance to the statistical significance of the analysis result.

One endeavor in this area that has received increasing attention is the study of the "transcriptome map," where the transcriptome map is the gene expression values as they relate to the position of the genes along the chromosome. These studies allow the visualization of global chromosomal expression patterns and have led to the discovery of many unexpected associations among genes (Cohen et al., 2000; Caron et al., 2001; Spellman and Rubin, 2002). The results of such analyses can also challenge the current notion of genetic regulation. For example, Cohen et al. (2000) used chromosome correlation maps from the *Saccharomyces cerevisiae* genome to demonstrate examples of adjacent pairs of genes with highly correlated expression patterns, in which the promoter of only one of the two genes contains an upstream activating sequence known to be associated with the expression pattern.

One application of the study of the transcriptome map is to identify chromosomal regions of genes demonstrating abnormal expression levels in cancer (i.e., regions in which all genes are either all highly under- or over-expressed). These genes have the potential for identifying areas where there has been gene amplification or deletion (Fujii et al., 2002; Bisognin et al., 2004). Such regions of gain and loss potentially harbor tumor oncogenes and suppressor genes. Hence, detection of these genetic events and the associated pattern of molecular architecture as evidenced by gene expression can provide a crucial step towards understanding genetic instability.

Current methods used to search for regions of increased or decreased expression rely mainly on averaging gene expression values from neighboring genes along the chromosome (Caron et al., 2001; Fujii et al., 2002; Pollack et al., 2002) or using scan statistics on a pre-defined length of gene sequence windows (Husing et al., 2003; Levin et al., 2005). These techniques could be effective at screening for possible regions of interest. However, they may lack precision and require ad-hoc choices, such as how many neighbors to include in this average. Another challenge is selecting a suitable threshold that categorizes a gene as having abnormal expression.

To address these issues, we developed a hidden Markov model (HMM) to identify regions of genes with abnormal expression levels from microarray gene expression data. A HMM is a type of stochastic model that has been used successfully in a variety of scientific applications (Rabiner, 1989; Koski, 2002). Such a model can be useful when there is an ordered sequence of entities and a measurement is taken on each entity. In our case the ordered entities are the genes along the chromosome and the measurements are the expression values. The model is particularly useful when there is thought to be a finite number of possible states each entity could be in, and there is a correlation between neighboring entities. The model has a probabilistic structure, which enhances the ability to make statistical inference. In particular the model allows the probability of different expression status for each gene can be calculated, and hence regions of abnormal expression identified.

METHODS

The genes along each chromosome are ordered in the sequence 1, 2, 3, 4, etc. The observed gene expression data is denoted by $\{x_1, x_2, x_3, \ldots\}$. It is assumed that each gene can be in one of a finite number of "hidden" states, and that the distribution of possible expression values for a gene is determined by its "hidden" state. The purpose of a HMM is, through a probabilistic structure, to infer the "hidden" state sequence $\{S_1, S_2, S_3, \ldots\}$ from the corresponding observed data $\{x_1, x_2, x_3, \ldots\}$. In our problem of the identification of regions of over- or under-expression, each gene can be in one of 5 possible states, defined as follows:

$$S_t = \begin{cases} 1, & \text{region of underexpression} \\ 2, & \text{singleton of underexpression} \\ 3, & \text{normal expression} \\ 4, & \text{region of overexpression} \\ 5, & \text{singleton of overexpression} \end{cases}$$

Regions 1 and 4 are the ones of primary interest to us, as they might represent genomic changes however as some abnormal expression genes may not reside in those regions we also include states 2 and 5. State 3 represents genes that are expressing at a level typical of that seen in normal samples. In the over and under expressed regions, expression values from neighboring genes are likely to be highly correlated. This is incorporated in a HMM by considering the transition probabilities between neighboring genes.

Let $P_{ij} = P(S_{t+1} = j| S_t = i)$ denote the transition probability of gene $t + 1$ being in the state j given gene t is in state i. Here, gene t indicates the tth gene in the ordered position from the direction of the p arm to the q arm in a chromosome. The distribution of the gene expression value for a gene is determined by its hidden state. We assume a parametric form for the density of the gene expression level for each state and denote these densities by $f(x| \theta_1)$, $f(x| \theta_2)$, $f(x| \theta_3)$, $f(x| \theta_4)$, and $f(x| \theta_5)$, where x is the observed data and $\theta_j = (\mu_j, \sigma_j)$ are the parameters in each density function. Fitting a HMM requires estimation of the parameters P_{ij} and $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$, from which one can infer the probability that the gene is in a particular state. The fit also provides limiting probabilities π_j's for the proportion of genes in each of the 5 states.

Order reversibility

We chose to order the genes from the p arm to the q arm. For the purpose of analyzing gene expression data we could equally well have reversed the order of the genes along the chromosome. Thus it is appropriate in this context to impose an order-reversible constraint on the transition probabilities. In other words, suppose the expression state sequence from p arm to q arm in a chromosome is a stationary Markov chain having transition probabilities P_{ij}'s and limiting probabilities π_j's, then the reversed state sequence is also a Markov chain with transition probabilities given by $Q_{ij} = \pi_j P_{ji}/\pi_i$ (Ross, 1993). It is well known that if the state space is finite (as in our case) and $P_{ij} > 0$ for all i, j, the necessary and sufficient condition for order reversibility is

$$P_{ij}P_{jk}P_{ki} = P_{ik}P_{kj}P_{ji} \qquad \forall \ i, j, k \tag{1}$$

Using simple combinatorial mathematics, it can be shown that 6 non-redundant equations of the form (1) can be constructed for a 5-state order reversible Markov chain. Specifically, the 6 nonlinear equations of P_{ij} can be written as

$$P_{12}P_{23}P_{31} = P_{13}P_{32}P_{21},\ P_{12}P_{24}P_{41} = P_{14}P_{42}P_{21},$$
$$P_{12}P_{25}P_{51} = P_{15}P_{52}P_{21},\ P_{13}P_{34}P_{41} = P_{14}P_{43}P_{31},$$
$$P_{13}P_{35}P_{51} = P_{15}P_{53}P_{31},\ P_{14}P_{45}P_{51} = P_{15}P_{54}P_{41}.$$

Thus, the order reversibility condition reduces the number of free parameters that need to be estimated.

Identifiability of mixture distribution

A HMM is a type of mixture model since an independent mixture model can be considered as a special case of HMM when the transition probabilities P_{ij} do not depend on i. A well known general problem with mixture models is identifiability of the parameters. In our model, because the support of the distributions for state 1 and state 2 (or state 4 and state 5) are negative (positive) and their distributions are assumed to have the same parametric form, we can exchange the indices of state 1 and state 2 (or state 4 and state 5) of the Markov chain without changing the mixture distribution. Without further restrictions or constraints there will not be a unique maximizing point of the likelihood, and numerical methods for estimating the parameters are likely to have problems.

One way to overcome the problems associated with the lack of identifiability of the parameters is by the imposition of an appropriate constraint on the parameters (McLachlan and Peel, 2000). Since states 1 and 4 are regions of under- or over-expression it is appropriate to assume that the transition probability from state 1 (4) to state 1 (4) should be large while the transition probability within singleton states, i.e. from state 2 (5) to state 2 (5) should be small. One way to ensure this in the estimation method is to impose constraints such as $P_{11} > P_{22}$ and $P_{44} > P_{55}$. Alternatively, we may impose prior distributions on these particular parameters, we do this by adding a corresponding penalty functions to the log-likelihood in the estimation method. Since all transition probabilities fall in the interval of (0, 1), a suitable prior for the transition probabilities can be chosen from a *beta* distribution. The consequence of this will be that genes in a region of over- or under-expression will tend to group together due to the high transition probability P_{11} or P_{44} and hence separate from singleton over or under-expressed genes.

Based on biological rationale we added a further constraint to the transition probability matrix, namely that $P_{23} = P_{33} = P_{53}$. The motivation for this is that the probability of transitioning into a normal state is the same, irrespective of whether the gene is in a singleton state or a normal state. Note that this constraint implies a non-Markov (independence) property of those three transition probabilities.

Penalized log-likelihood

Suppose x = $\{x_1, x_2, \ldots, x_n\}$ is the observed relative expression sequence for a chromosome of a tumor sample. The complete data is these observed expression values plus the unobserved hidden states. A HMM assumes independence between the observation given their hidden state, thus the penalized complete-data log-likelihood of the HMM for a single chromosome of a single sample can be written as

$$\log PL^C(\Psi) = \sum_{j=1}^{5} Z_j^1 \cdot \log p_j + \sum_{t=2}^{n} \sum_{j'=1}^{5} \sum_{j=1}^{5} Z_{j'j}^t \cdot \log P_{j'j} + \sum_{t=1}^{n} \sum_{j=1}^{5} Z_j^t \cdot \log f(x_t \mid \theta_j)$$
$$+ \log h_1(P_{11}) + \log h_2(P_{22}) + \log h_4(P_{44}) + \log h_5(P_{55}) \quad (2)$$

where $p_j = Pr(S_1 = j)$ is the initial probability, $\Psi = \{\{\theta_j\},\{p_j\},\{P_{ij}\}\}$ are the set of all parameters, and h_1, h_2, h_4, h_5 are the densities of the *beta* prior distributions for P_{11}, P_{22}, P_{44}, P_{55} respectively. Specifically,

$$h_1 \sim beta(\alpha_1,1), \quad h_2 \sim beta(1,\beta_2)$$
$$h_4 \sim beta(\alpha_4,1), \quad h_5 \sim beta(1,\beta_5).$$

We will select values of the tuning parameters $(\alpha_1,\alpha_4,\beta_2,\beta_5)$ to be greater than 1 to ensure that P_{11} and P_{44} are large, and P_{22} and P_{55} are small. The variables Z_j^t and $Z_{j'j}^t$ are indicator variables for the true hidden states defined as

$$Z_j^t = \begin{cases} 1, & \text{if } S_t = j \\ 0, & \text{otherwise} \end{cases}$$

$$Z_{j'j}^t = \begin{cases} 1, & \text{if } S_{t-1} = j' \text{ and } S_t = j \\ 0, & \text{otherwise} \end{cases}$$

The first 3 terms on the right hand side of (2) are the usual ones seen in the HMM literature. The whole penalized complete-data log-likelihood is then the sum over all chromosomes of all samples.

Estimation procedure

The parameters in (2) can be estimated by the method of maximum likelihood, which has been successfully used in a variety of mixture problems. For computation of the estimates it is natural to use the EM algorithm, Redner and Walker (1984). While the EM algorithm has good theoretical properties, it can be slow. In addition, in the E-step, the calculation of the expected value of Z_j^t, $Z_j^t = f_j^{(t)} b_j^{(t)} / \Sigma_{j=1}^5 f_j^{(n)}$, can be numerically unstable, when directly applying the forward-backward algorithm (Baum et al., 1970), as the forward and backward probabilities, $f_j^{(t)}$ and $b_j^{(t)}$ for gene t being in state j, become rapidly close to zero as n increases (Leroux, 1992). To address this issue, we use the method described by Leroux and Puterman (1992). We choose for each gene t a value of r for which $10^r \Sigma_j f_j^{(t)}$ lies between 0.1 and 1 and multiply $f_j^{(t)}$ by 10^{r} ($j = 1, \ldots, 5$). A similar procedure is applied to $b_j^{(t)}$. Then $f_j^{(t)}$ and the $b_j^{(t)}$ can be reconstructed for the purpose of computing the expected values of Z_j^t and $Z_{j'j}^t$.

In the M-step, there exist simple closed forms for finding the updated parameters in component densities of exponential family distributions and Markov chain transition probabilities for standard HMMs. However, as shown in (1), the condition of order reversibility for a Markov chain imposes 6 nonlinear constraints on the transition probabilities. This adds complexity on the estimation for the 14 free parameters in the transition probability matrix. To solve this problem, we employ sequential linear programming to find the solution in each M-step. We utilize the software, AMPL,[1] in conjunction with an R program for the forward-backward algorithm to obtain the parameter estimates and posterior probabilities of expression states for each gene.

Choice of tuning parameters

Because a *beta* distribution was used to define the penalty terms for P_{11}, P_{22}, P_{44}, P_{55}, the last four terms in equation (2) can be written as

$$(\alpha_1 - 1) \log P_{11} + (\alpha_4 - 1) \log P_{44} + (\beta_2 - 1) \log (1 - P_{22}) + (\beta_5 - 1) \log (1 - P_{55}) + \mathit{const}.$$

Thus, α_1,α_4,β_2,β_5 become the tuning parameters in (2). In standard penalized regression and classification problems tuning parameters are typically chosen by cross-validation. This is not possible with a HMM because the true state is not observed. To overcome this, we employ the idea of false positive rate (FPR) and false negative rate (FNR) to select the tuning parameters. We first define an event E_i as a sequence of genes with identical expression state and N_{Ei} the number of gene in the corresponding event. For example, for the sequence of expression states

331111333322333333444333335

there are eight events with $E_1 = 3$, $E_2 = 1$, $E_3 = 3$ etc. and, $N_E1 = 2$, $N_E2 = 4$, N_E3, etc. From the definition of the 5 expression states, the HMM should be able to classify a short sequence of under-expressed (or over-expressed) genes as an event of state 2 (or 5). Similarly, a long sequence of under-expressed (or over-expressed) genes should be classified as an event of state 1 (or 4). If we are willing to assume that any event E_i of an under-expressed gene with N_{Ei} being 1 should be in state 2, and any event E_j of an under-expressed gene with $N_Ej \geq 3$ should be in state 1, then the FPR and FNR for regions of under-expression can be defined as

$$FPR = \frac{\#\{E_i \mid N_{Ei} \geq 3,\ E_i = 2\}}{\#\{E_i \mid N_{Ej} \geq 3,\ E_i = 1 \text{ or } 2\}}$$

$$FNR = \frac{\#\{E_i \mid N_{Ei} = 1,\ E_i = 1\}}{\#\{E_i \mid N_{Ej} = 1,\ E_i = 1 \text{ or } 2\}}$$

The FPR and FNR for regions of over-expression can be defined in an analogous way. Therefore, the tuning parameters for the penalty terms can be adjusted in a way that the FPR and FNR for regions of under-expression or over-expression will be under some pre-selected thresholds. We use 0.05 as the threshold for both. However, we should note that this is not a unique way to select the tuning parameters because many different sets of tuning parameters may result in the same FPR and FNR.

RESULTS AND DISCUSSION

Analysis of lung cancer microarray data

We applied our method to a gene expression microarray experiment to search for regions of interest. The data were published by Beer et al. (2002). The data consist of gene expression values measured by Affymetrix HuGeneFL chips for 86 lung adenocarcinomas samples. In addition, there are data from 9 normal lung tissues taken from patients with lung cancer. The probeset expression summaries were calculated using the method described in Beer et al. (2002). On the Affymetrix HuGeneFL chip, there are 7,129 probesets with each probeset measuring the expression of one gene. However, there is redundancy in that there can be

[1] The reference is written by Fourer et al. (1993). The student edition can be downloaded for free from the web: www.ampl.com/cm/cs/what/ampl/DOWNLOADS.

more than one probeset for a given gene. To account for this we took the median value for any gene with multiple probesets, which resulted in 6,800 gene expression measurements. All the negative expression values were set at zero. We further added a constant 150 to all the expression values to avoid artificially large fold changes compared to the normal tissues for genes with small expression.

After this data preprocessing, the reference expression value for gene k on chromosome i is determined as the median of 9 normal lung samples, denoted as $N_i^{(k)}$. Then, the rescaled data for analysis are $X_{im}^{(k)} = \log_2(T_{im}^k / N_i^{(k)})$, where $T_{im}^{(k)}$ is the pre-processed expression for gene k on chromosome i and tumor sample m. For each gene, we have mapped its base pair location along the chromosome as described by Levin et al. (2001). Since some of the genes are not mapped through this program, the final data includes 5,707 genes for each tumor sample. We assume the density of the normal expression state is Gaussian. Since we have 9 normal tissue samples, we estimate the parameters $(\hat{\mu}_3, \hat{\sigma}^2{}_3)$ for this density using all the data from these normal tissues. The resulting parameter estimates are $(\hat{\mu}_3, \hat{\sigma}^2{}_3) = (0, 0.3^2)$. These estimates were then used in the EM algorithm as fixed. In addition, we assume truncated lognormal distributions with parameters (μ_j, σ_j), $j = 1, 2, 4, 5$ for the other four component densities $f(-x|\mu_1, \sigma_1)$, $f(-x|\mu_2, \sigma_2)$, $f(x|\mu_4, \sigma_4)$ and $f(x|\mu_5, \sigma_5)$. The truncated point is set at 0.4, so that there was no probability mass between -0.4 and 0.4 for genes in over- or under-expressed states. As described in the previous section, the prior we chose for the transition probabilities $\{P_{11}, P_{44}\}$ are $beta(\alpha_1, 1)$ and $beta(\alpha_4, 1)$ respectively, and the prior for $\{P_{22}, P_{55}\}$ are $beta(1, \beta_2)$ and $beta(1, \beta_5)$, respectively.

We ran the EM algorithm until the increment of the penalized log-likelihood is less than $10^{-4}\%$ of the penalized log-likelihood from the previous iteration. We tried many sets of tuning parameters $\{\alpha_1, \alpha_4, \beta_2, \beta_4\}$ to ensure that the false positive and false negative rates are less than 0.05. The expected value of Z^i_j s in the E-step from the last EM iteration is used as the posterior state probabilities to assign the expression state for each gene. A gene will be determined to be in state j if its corresponding posterior state probability is greater than 0.6. Thus, genes for which there was no clear indication of the state were not assigned to any particular state.

The estimated transition probability matrix is

$$\begin{bmatrix} 0.721 & 5\cdot 10^{-5} & 0.238 & 0.010 & 0.032 \\ 1\cdot 10^{-5} & 0.070 & 0.877 & 0.017 & 0.042 \\ 0.004 & 0.057 & 0.877 & 0.009 & 0.053 \\ 0.006 & 0.028 & 0.316 & 0.649 & 2\cdot 10^{-5} \\ 0.009 & 0.045 & 0.877 & 8\cdot 10^{-6} & 0.070 \end{bmatrix}$$

Note that the constraint on the transition probability matrix, has forced $P_{23} = P_{33} = P_{53}$. However, these three estimated transition probabilities were very similar when this constraint was removed. This transition probability matrix reveals several interesting phenomena. First, the transition probability of under-expression region to under-expression region (P_{11}) is larger than that of over-expression region to over-expression region (P_{44}). This may be due to the fact that the overall magnitude of the gene expression values relative to normal tissues in regions of under-expression is greater than that in regions of over-expression. This can be seen in Figure 1 where a larger proportion of genes in under-expression regions have values further away from zero than in over-expression regions. A possible explanation of this is that deletion of genes will result in disappearance of gene expression, resulting in a many fold change in gene expression relative to normal.

Second, P_{31} and P_{34} are close to 0. This suggests that there is very small number of regions of under- and over-expression in this data. And the number of regions of under-expression is even smaller than that of regions of over-expression ($P_{31} < P_{34}$). On the other hand, both P_{13} and P_{43} are the second largest transition probability in the first and fourth row of this matrix. This is because the majority of the genes are normally expressed. An interpretation of this is that genes in a region of under- or over-expression either continue in their current state or transit into a normal expression state. Furthermore, we calculated the limiting probabilities from the estimated transition probabilities. The resulting probabilities for the five different expression states are 1.78% for region of under-expression, 5.94% for singleton of under-expression, 84.6% for normal expression, 2.11% for region of over-expression, and 5.57% for singleton of over-

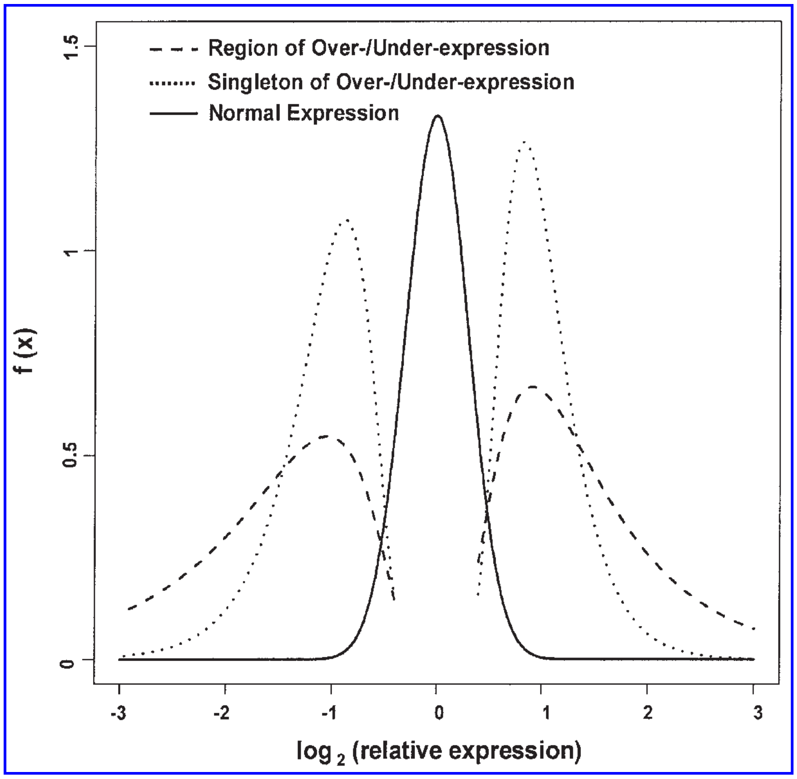

FIG. 1. Component densities of expression states estimated from hidden Markov model (HMM) for lung cancer data. The solid curve represents the density of relative expression values (in logarithm base 2 scale) in normal expression regions; dotted curves represent the densities of singleton of over- (right side) or under-expression (left side); dashed curves represent the density of region of over- (right side) or under-expression (left side). The relative expression is the ratio of the expression of a tumor over the median of nine normal expression values for any single gene.

expression. Note that the proportion of over-expressed genes (7.68%, state 4 and state 5) is close to the proportion of under-expressed genes (7.72%, state 1 and state 2).

An interesting result is found in Figure 1, which shows the five component densities. The densities suggest that the expression levels in regions of over-expression (under-expression) tend to be higher (lower) than those in singletons of over-expression (under-expression). Pollack et al. (2002) performed a parallel assessment of correlation between mRNA levels and DNA copy number changes using cDNA microarray and array CGH in a group of breast cancer lines and tumors. They observed that the overall patterns of gene amplification and elevated gene expression are quite concordant, and concluded that DNA copy number alterations have a pervasive global influence on gene expression. Thus, some of those regions we identified could be regions of gain or loss. Identification of the genes in those regions that are most highly expressed may lead to the discovery of oncogenes or genes important to the cancer process. Of the 86 × 5707 genes, the number of genes that are classified into state 1,2,3,4 and 5 are 2,419 (0.5%), 23,662 (4.8%), 422,218 (86.0%), 4,496 (0.9%), and 20,850 (3.5%), respectively. 17,159 (3.5%) genes are not classified into any of the five states because none of the five posterior state probabilities are greater than 0.6 for those genes.

Based on the definition of FPR and FNR in the previous section, the false positive and false negative rates for regions of under-expression (over-expression) are 5% (0.2%) and 0.5% (2.8%). The resulting number of regions and singletons of over- and under-expression is shown in Table 1. The definition of an event, E_i, in this table is given in the subsection, "Choice of tuning parameters." For example, 367 events of 2 genes in a region of under-expression mean that those regions all have only 2 genes in a row that are called in state of under-expression.

Table 1. Number of Events of Abnormal Expression Estimated from Hidden Mark or Model for Lung Cancer Data

Events size	*Region down*	*Singleton down*	*Region up*	*Singleton up*
1 gene	99 (0.5%)	20,194	534 (2.8%)	18,837
2 genes	367 (16%)	1968	671 (40%)	1005
3+ genes	455 (95%)	24	708 (99.8%)	1

Association of regional abnormality and clinical measures

We investigated the impact of regions of over- and/or under-expression on clinical measures, such as tumor stage and patient survival. Since we are more confident with the calls for those regions of over- or under-expression with at least three genes in a row, we calculated for each of the 86 tumor samples the number of regions of over- and under-expression with $NE_i \geq 3$ (the total number of events in the whole data are 708 and 455, respectively as seen in Table 1).

We first observed an interesting result that a strong correlation exists between the number of regions of over-expression and the number of regions of under-expression (Spearman correlation 0.67, $p < 0.0001$). We then investigated the association of region number with tumor stage and differentiation. The patients' tumors are either stage 1 (67 patients) or stage 3 (19 patients), and the differentiation states: are well (23 patients), moderate (42 patients), and poor (21 patients). The three box plots in Figure 2 show the distribution of the number

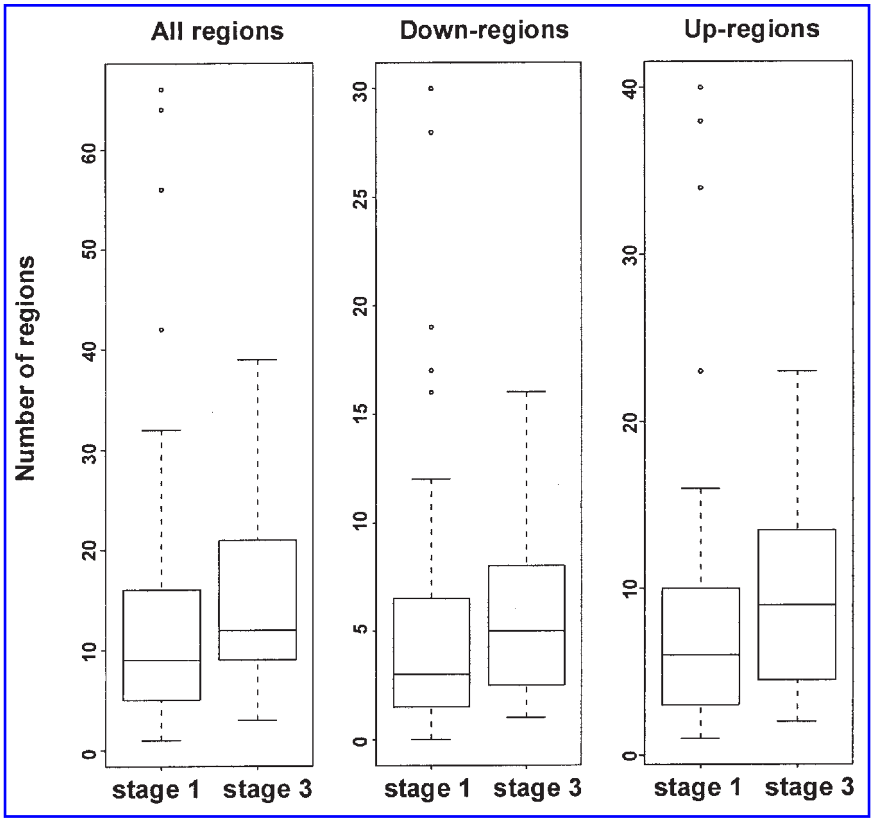

FIG. 2. The association of the number of regions of abnormal expression with stage in lung cancer data. Higher stage is associated with overall higher abnormal expression regions. The association is consistent for down- and up-regulated regions. However, four stage 1 tumors are found to have the highest numbers of abnormal expression regions in all three box plots.

of abnormal expression regions between stage 1 and stage 3 tumors. The number of all regions in the most left panel is the total number of regions of under- and over-expression. Although the difference of the number of abnormal expression regions between stage 1 and stage 3 tumors does not attain significant level (Wilcoxon rank sum test *p* value: 0.092, 0.067, and 0.203 for all regions, down regions, and up regions, respectively), stage 3 tumors tend to have more regions of over and/or under-expression than stage 1 tumors. The diluted significance level is largely due to the 4 stage 1 tumors that have most of abnormal expression regions. These four patients are all alive with survival times of 21.1, 87.7, 36, and 40 months (median survival time 29.5 months) with differentiation status being poor, poor, well, and moderate. With these 4 tumors removed, the tests show significant difference in the number of regions for stage (*p*-value: 0.028, 0.019, 0.078 for all regions, down regions, and up regions, respectively). An even stronger association can be seen in Figure 3 between the number of regions and the differentiation status. It is evident that the poorer the differentiation status of the tumor for each patient, the greater the number of regions of over- and/or under-expression. The Kruskal-Wallis rank sum tests result in *p*-values: 0.014, 0.017, and 0.013 for all regions, down-regions, and up-regions.

We finally examined the association of the number of regions with patient survival. We used Cox PH model (Cox, 1972) for survival analysis, treating the number of regions as a predictor. None of the *p*-values are significant for all 86 tumor samples. However, removing the four tumors results in *p*-values: 0.007, 0.003, and 0.049 for all regions, down-regions, and up-regions. Figure 4 is an example of this association in the all regions case. From the upper panel, we can see in general the greater number of regions of over- and under-expression for a patient the shorter the survival time. A Kaplan-Meier plot is shown in the lower panel, based on a dichotomization of the 82 patients based on a cut-off of 12 regions of over- and under-expression. It is evident that the group of patients with larger number of regions has a poor survival.

The process of carcinogenesis starts with the initiation of a normal cell which then progresses to a clonal

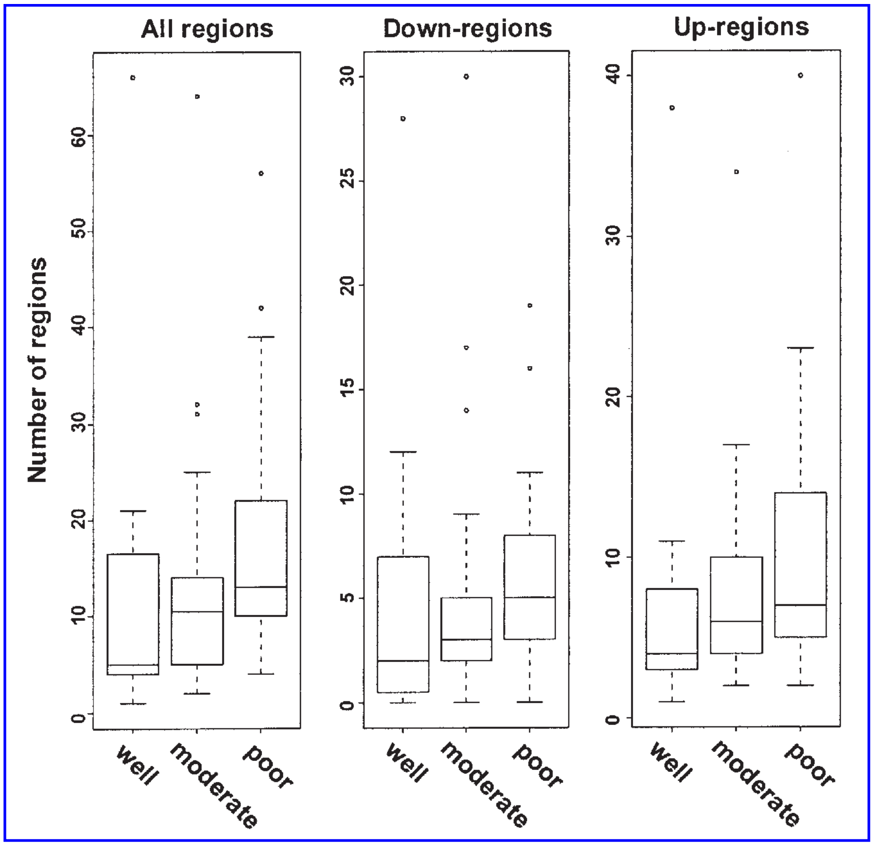

FIG. 3. The association of the number of regions of abnormal expression with differentiation status. More total abnormal, down- and up-regulated regions are found among the poorly differentiated tumors.

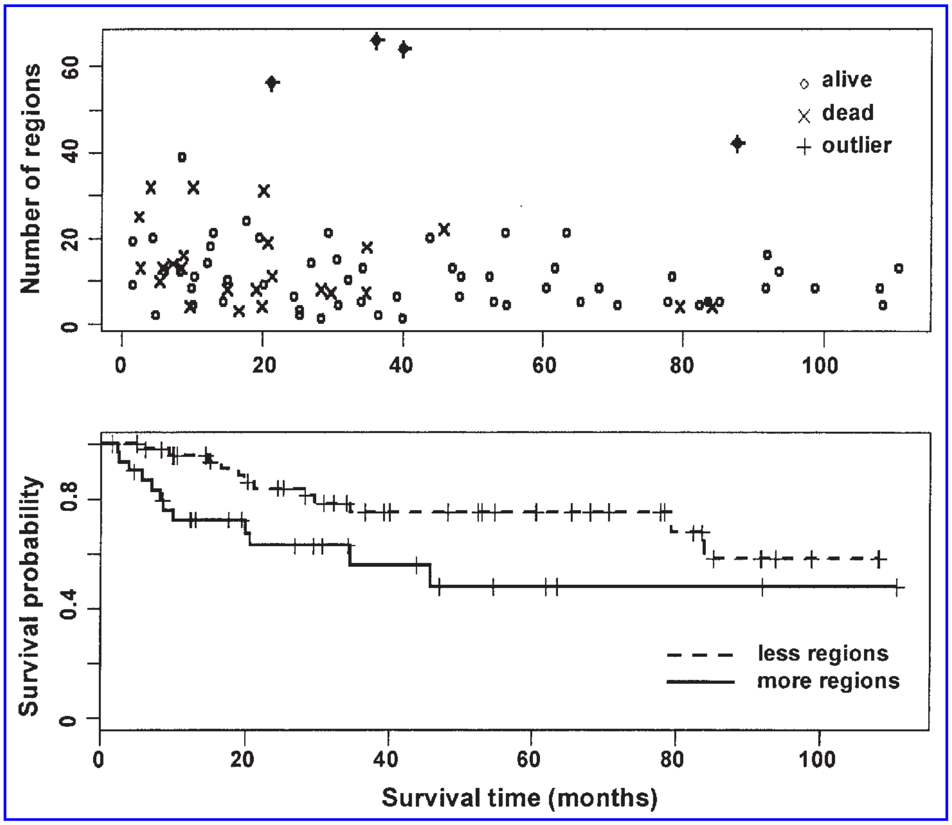

FIG. 4. The association of the number of regions of abnormal expression with patients' prognosis. Upper panel: The number of total abnormal expression regions (y-axis) is inversely associated with patients' survival time (x-axis). Lower panel: After removing four stage 1 tumors with the highest numbers of abnormal expression regions as observed in the upper panel, a significant difference of the survival distributions is observed between the two groups of patients with more (solid curve) or less (dashed curve) abnormal expression regions using 12 regions as the cutoff (p=0.03).

growth of malignant cells capable of invasion and metastasis. The mechanism driving this process is progressive DNA alterations, which occurs over time (Vogelstein and Kinzler, 2004). Under this model, it is the accumulation of DNA alterations in a cell which is involved in the progression from benign to the malignant state. Tumor stage and differentiation are well established surrogates of malignant potential. In this study, we provide a direct molecular measure of alterations from normal gene expression which reflects gross DNA damage (amplifications and deletions) as well as alterations in transcriptional control. As such, we expect these regional alterations in gene expression to reflect the accumulation of the DNA damage. Therefore, our findings that regional alterations in gene expression are correlated with stage and differentiation are consistent with the model of carcinogenesis described above.

Identification of regions of over- and underexpression

We further examine regions of over- or underexpression that occurs in many tumors and that may be of prognostic and therapeutic value. As the abnormal expression regions identified by our method may be caused by DNA copy number changes (amplifications or deletions), we will compare those regions to what have been reported in the literature such as Knuutila et al. (1998, 1999).

We determine that a gene is within a region of over- or under-expression if the posterior state probability is greater than 0.6. Since regions may contain various numbers of genes, we only select those that have at least three genes to avoid false positive calls for the small regions (one or two genes). However, each tumor is likely to have different genes that show abnormal expression levels in a specific chromosomal region. Therefore, we count the number of tumor samples in which a given gene is in a region of over- or underexpression.

We observed that some chromosomes have more abnormal expression regions than others. Regions of over-expression in chromosomes 1, 6, 7, 11, 12, 17, 19, 22, and X are more frequent. Many studies using CGH have reported amplifications in non-small-cell lung cancer (NSCLC) in these chromosomes as well. For example, 12p amplicon harbors the KRAS2 gene that has been detected in two lung adenocarcinomas (Bjorkqvist et al., 1998). In addition, amplifications of 6p12-pter and 17q24-qter are frequent in NSCLC (Knuutila et al., 1998). For these three regions, we observed in our data five genes (D6S51E, D6S52E, CSNK2B, CLIC1, and VARS2) in the 6p21.33-p22.1 region with frequent counts in four to seven tumor samples, three genes (TNFRSF1A, SCNN1A, and LTBR) in 12p13.2-p13.33 with frequent counts in six to seven tumor samples, and three genes (GALK1, ITGB4, and LLGL2) in 17q24.3-q25.3 with frequent counts in five tumor samples. We also report here the four most frequent regions of over-expression in our data: (1) 7p22.3 (JTV1, RAC1, KDELR2) with frequency of 10–15; (2) 21q22.3 (TFF1, TFF2, TFF3) with frequency of 10; (3) 11q23.3-q25 (SC5DL, unknown gene, SORL1) with frequency of 9; and (4) 11q12.1-q13.5 (EMS1, PPFI1A, FADD) with frequency of 3–8. We note that some of the survival genes reported in Beer et al. (2002) are in regions of over-expression and include VEGF, ERBB2, WNT10B, and IGFBP3, although the frequency is rare (1 or 2 only).

Regions of under-expression are much less frequent in our data as compared to the detection of regions of deletions in NSCLC (Knuutila et al., 1999). For example, we didn’t find any region of under-expression at either region 3p14.2 or 9p21. We observed in our data two chromosomes, 16 and 19, that have the most regions of under-expression. We also observed 3 regions in chromosomes 6, 10, and 17. These regions are (1) 6p21.33 (HLA-DRA, HLA-DRB1, HLA-DRB5) with a frequency of 21; (2) 10q22.1-q22.2 (SFTPA1, SFTPA2, SFTPD) with a frequency of 42; and (3) 17q11.2-q21.1 (SCYA14, SCYA5, SCYA4) with a frequency of 15. Although DNA deletion and amplification could account for expression changes in the regions identified above, the possibility of other mechanisms that may result in the expression abnormality should not be excluded. For example, DNA methylation, a type of chemical modification of DNA that can be inherited without changing the DNA sequence, can lead to gene silencing when combined with histone acetylation (Baylin et al., 2001). In addition, up-regulation of gene expression in a chromosomal region may be caused by the opening of the chromatin of an entire neighborhood as a result of activation of a target gene within the neighborhood (Spellman and Rubin, 2002). Overall, our findings suggest the possibility that there are regional changes associated with genomic structure changes; these merit further investigation for potential clinical utility.

CONCLUSION

In this paper, we have integrated gene expression data with genomic location information and have applied a HMM to discover chromosomal regions of abnormal expression. Several regions identified in this paper are consistent with known regions of amplification reported in the literature. This integrated analysis provides biological interpretation regarding the expression changes in certain lung tumors. Recent development of high throughput SNP arrays has allowed researchers to directly identify changes in the DNA level reflective of LOH, deletion, and amplification. Computational methods (Lin et al., 2004; Huang et al., 2004) have been developed to identify these abnormal DNA events. As cancers are complex biological systems, no single type of global profiling approach can entirely elucidate tumor behavior. Therefore, combining mRNA and DNA high throughput data in conjunction with sophisticated statistical algorithms may be needed to identify important molecular causes of carcinogenesis and characterize the fundamental processes of tumor growth.

ACKNOWLEDGMENTS

We acknowledge the contributions of Dr. Albert M. Levin towards the ideas presented in the manuscript and thank him for providing genome map location data. We also acknowledge the helpful comments of the reviewers of this manuscript. This work was supported by National Cancer Institute grant: U19 CA-85953.

REFERENCES

BAYLIN, S.B., ESTELLER, M., ROUNTREE, M.R., et al. (2001). Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. Hum Mol Genet **10,** 687–692.

BAUM, L.E., PETRIE, T., SOULES, G., et al. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. Ann Math Stat **41,** 164–171.

BEER, D.G., KARDIA, S.L.R., HUANG C.-C., et al. (2002). Gene expression profiles predict survival of patients with lung adenocarcinoma. Nat Med **8,** 816–824.

BISOGNIN, A., BORTOLUZZI, S., and DANIELI, G.A. (2004). Detection of chromosomal regions showing differential gene expression in human skeletal muscle and in alveolar rhabdomyosarcoma. BMC Bioinform **5,** 68.

BJORKQVIST, A.M., HUSGAFVEL-PURSIAINEN, K., ANTTILA, S., et al. (1998). DNA gains in 3q occur frequently in squamous cell carcinoma of the lung, but not in adenocarcinoma. Genes Chromosomes Cancer **22,** 79–82.

CARON, H., VAN SCHAIK, B., VAN DER MEE, M., et al. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domain. Science **291,** 1289–1292.

COHEN, B.A., MITRA, R.D., HUGHES, J.D., et al. (2000). A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet **26,** 183–186.

COX, D.R. (1972). Regression models and life tables J R Soc Stat B **34,** 187–220.

CUI, Y., ZHOU, M., and WONG, W.H. (2004). Integrated analysis of microarray data and gene function information. OMICS **8,** 106–117.

FOURER, R, GAY, D.M., KERNIGHAN, B.W. (1993). *AMPL: a modeling language for mathematical programming*. (Scientific Press, New York).

FUJII, T., DRACHEVA, T., PLAYER, A., et al. (2002). A preliminary transcriptome map of non-small cell lung cancer. Cancer Res **62,** 3340–3346.

GAO, F., FOAT, B.C., and BUSSEMAKER, H.J. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinform **5,** 31.

HUANG, J., WEI, W., ZHANG, J., et al. (2004). Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics **1,** 287–299.

HUSING, J., ZESCHNIGK, M., BOES, T., et al. (2003). Combining DNA expression with positional information to detect functional silencing of chromosomal regions. Bioinformatics **19,** 2335–2342.

KOSKI, T. (2002). *Hidden Markov models for bioinformatics* (Springer/Kluwer Academic Publishers, New York).

KNUUTILA S., BJORKQVIST, A.M., AUTIO, K., et al. (1998). DNA copy number amplification in human neoplasms: review of comparative genomic hybridization studies. Am J Pathol **152,** 1107–1123.

KNUUTILA, S., AALTO, Y., AUTIO, K., et al. (1999). DNA copy number losses in human neoplasms. Am J Pathol **155,** 683–694.

LEE, T.I., RINALDI, H.J., ROBERT, F., et al. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science **298,** 799–804.

LEROUX, B.G. (1992). Maximum-likelihood estimation for hidden Markov models. Stochastic Proc Appl **40,** 127–143.

LEROUX, B.G., and PUTERMAN, M.L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. Biometrics **48,** 545–558.

LEVIN, A.M., LEVIN, A., and KARDIA, S.L.R. (2001). A physical transcriptome map for chromosome level analysis of gene expression data. Presented at the American Society of Human Genetics 51st Annual Meeting, San Diego.

LEVIN, A.M., GHOSH, D., CHO, K.R., et al. (2005). A model-based scan statistic for identifying extreme chromosomal regions of gene expression in human tumors. Bioinformatics **21,** 2867–2874.

LIN, M., WEI, L.J., SELLERS, W.R., et al. (2004). dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics **20,** 1233–1240.

MCLACHLAN, G., and PEEL, D. (2000). *Finite mixture model* (Wiley, New York).

NARIAI, N., KIM, S., IMOTO, S., et al. (2004). Using protein-protein interactions for re-fining gene networls estimated from microarray data by Bayesian networks. Proc Pac Symp Biocomput **9,** 336–347.

POLLACK, J.R., PEROU, C.M., ALIZADEH, A.A., et al. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet **23,** 41–46.

POLLACK, J.R., SRLIE, T., PEROU, C.M., et al. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci USA **99,** 12963–12968.

RABINER, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE **77,** 257–286.

REDNER, R.A. and WALKER, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. Soc Indust Appl Math Rev **26,** 195–239.

ROSS, S.M. (1993). *Introduction to Probability Models*, 5th ed. (Academic Press, San Diego).

SPELLMAN, P.T., and RUBIN, G.M. (2002). Evidence for large domains of similarly expressed genes in the *Drosophila* genome. J Biol **1,** 5.

TAMADA, Y., KIM, S., BANNAI, H., et al. (2003). Estimating gene networks from gene expression data by combining Bayesian netowrk model with promoter element detection. Bioinformatics **Suppl. 2,** 227–236.

VOGELSTEIN, B., and KINZLER, K.W. (2004). Cancer genes and the pathways they control. Nat Med **10,** 789–799.

ZHOU, X., KAO, M.J., and WONG, W.H. (2002). Transitive functional annotation by shortest path analysis of gene expression data. Proc Natl Acad Sci USA **99,** 12783–12788.

Address reprint requests to:
Dr. Chiang-Ching Huang
Department of Preventive Medicine
Northwestern University
680 N. Lake Shore Dr., Ste. 1102
Chicago, IL 60611-4402

E-mail: huangcc@northwestern.edu