JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 4, 2006

Pp. 853–865

New Probabilistic Network Models and Algorithms for Oncogenesis

MARCUS HJELM,[1] MATTIAS HÖGLUND,[2] and JENS LAGERGREN[1]

ABSTRACT

Chromosomal aberrations in solid tumors appear in complex patterns. It is important to understand how these patterns develop, the dynamics of the process, the temporal or even causal order between aberrations, and the involved pathways. Here we present network models for chromosomal aberrations and algorithms for training models based on observed data. Our models are generative probabilistic models that can be used to study dynamical aspects of chromosomal evolution in cancer cells. They are well suited for a graphical representation that conveys the pathways found in a dataset. By allowing only pairwise dependencies and partition aberrations into modules, in which all aberrations are restricted to have the same dependencies, we reduce the number of parameters so that datasets sizes relevant to cancer applications can be handled. We apply our framework to a dataset of colorectal cancer tumor karyotypes. The obtained model explains the data significantly better than a model where independence between the aberrations is assumed. In fact, the obtained model performs very well with respect to several measures of goodness of fit and is, with respect to repetition of the training, more or less unique.

Key words: cancer, chromosomal aberration, probabilistic model, learning algorithm, graphical representation.

1. INTRODUCTION

RAPID PROGRESS IN OUR UNDERSTANDING of the biology of human neoplasias has been achieved during the past decades. It is now well established that cancer arises through a multistep accumulation of somatic mutations (Fearon and Vogelstein, 1990; Hanahan and Weinberg, 2000), and that this process may proceed over many years. In most tumors, many genetic changes are microscopically visible as nonrandom and often disease-specific chromosomal abnormalities (Heim and Mitelman, 1995). Hence, the analysis of chromosomal changes in tumor cells have been rewarding both for the characterization of tumors and for the identification of genes involved in the tumorigenetic process. In particular, the chromosomal analysis of hematological malignancies has been important in that an increasing number of disease-specific balanced rearrangements, most often translocations, have been found (Mitelman *et al.*, 1997).

[1]SBC and Dept. of Numerical Analysis and Computer Science, KTH, Stockholm, SE-106 91, Sweden.
[2]Department of Clinical Genetics, Lund University Hospital, Lund, SE-221 85, Sweden.

However, many solid tumors exhibit a much more complex pattern of aberrations. Even though these chromosome changes invariably show a nonrandom distribution over the chromosome complement, tumor-specific aberrations are uncommon. Solid tumors also tend to contain a higher number of chromosomal changes than hematological malignancies and often exhibit extensive variability in the pattern of changes, even within the same histopathological entity. It has been suggested that this variability reflect a multistep pathogenetic process, resulting in both loss of tumor suppressor gene function and oncogene activation, seen as either loss or gain of chromosomal segments.

Cytogenetics and alternative approaches, such as comparative genomic hybridization (CGH), loss of heterozygosity (LOH), and array-based CGH, have been used to identify chromosomal segments altered during the development or progression of cancer. For the purpose of identifying global patterns of these changes, cytogenetic data has several advantages compared to other types of genomic information. For instance, genomic changes are described at a moderate level of resolution, and large datasets for individual tumor types are available (Mitelman *et al*., 2004).

The investigation of large sets of cytogenetically analyzed tumors have revealed several patterns of the evolution of the chromosomal changes of which the presence of specific developmental routes (chromosome aberrations pathways) as well as a temporal order of aberration acquisition are the most important (Höglund *et al*., 2005). Several types of models have been applied to describe tumor development in light of genetic changes. Fearon and Vogelstein (1990) used a linear model to describe the genetic changes leading to colorectal tumors. The linear model was developed further by Desper *et al*. (1999, 2000) to include tree-like models. These tree models have the property that any specific aberration a cannot happen unless an associated set of aberrations, appearing on the path from the root to a in the tree, have happened. Tree models, in general, assume that the development starts in a root and do not allow converging developmental pathways; that is, they cannot be used to model a situation where an aberration a and an aberration b lead to or, as is more likely, have a tendency to lead to an aberration c.

We develop a framework of algorithms with capacity to (1) derive a dynamic model explaining given data, which can be used to generate data and can be studied during the generation process, and (2) given a dynamic model, derive a graphical representation that communicates the dependencies in the data and the basic properties of the dynamic model. The type of model used is sufficiently powerful to capture converging as well as diverging pathways.

Starting from biologically reasonable assumptions concerning chromosome aberrations, we derive a Markov chain model. Markov chain models have been applied previously in the context of chromosomal aberrations by Simon *et al*. (2000), although they called their model the directed acyclic graph (DAG) model. Following Radmacher *et al*. (2001), we reduce the parameter complexity of the Markov chain model by introducing appropriate assumptions, basically, limiting dependencies to being pairwise. In addition, we allow a realistic representation of the event that the tumor is being "discovered" and thus that the tumorigenetic process is stopped.

Building on a technique to reduce the parameter complexity for Bayesian networks (Segal *et al*., 2003), we reduce the parameter complexity of our models by introducing modules of aberrations. That is, for each pair of modules, all pairs of aberrations of the respective modules have the same pairwise dependency. Furthermore, we add acyclic restrictions for cluster dependencies, thereby facilitating a good graphical representation of module models, where hopefully directed paths correspond to tumor progression pathways. We also give novel training procedures for our models. Finally, we apply our method to 461 cases of colorectal cancer with cytogenetically known chromosome aberrations (Höglund *et al*., 2002).

We use the standard machine learning approach of partitioning the data into a training set and a test set and evaluate the fitness of the derived models by how well they explain the test set. In contrast to Radmacher *et al*. who do not clearly explain their data significantly better than a model where independence is assumed, we do significantly better.

Dataset. To test our model, we chose the 25 most frequent copy number changes in colorectal tumors reported in Höglund *et al*. (2002) among 569 karyotypes. Each of these karyotypes was assessed for presence or absence of the chosen aberrations. All karyotypes with at least 1 and at most 12 of the chosen aberrations were selected, resulting in a final dataset of 461 samples. The reason why karyotypes with more than 12 aberrations were excluded was partly that these were considered less informative and partly for efficiency considerations. For simplicity, we use the same notation for the aberrations as in Höglund *et al*. (2002), where more information about the aberrations can be found.

2. DEFINITIONS

The type of data used determines a set of possible aberrations; i.e., the aberrations may be chromosomal breakpoints, copy number changes for various segments, etc.

Let $[n] = \{1, \ldots, n\}$ represent the n aberrations chosen for the study. A sample $D = \{d_1, \ldots, d_k\}$ with respect to the study is then a subset of $[n]$. An ordering $d_{\sigma_1}, \ldots, d_{\sigma_k}$ of D represents the following scenario: first d_{σ_1} occurred, then d_{σ_2}, and so on. After the event d_{σ_k} occurred, the set D was reported by an aberration test. This event is called *stop* and denoted Ⓢ. Note that D alone does not give us any information which, among the $k!$ possible orders, is the true order.

3. A MARKOV CHAIN MODEL FOR ABERRATIONS

In this section, we derive a Markov chain model for accumulation of aberrations.

Let $\{X(t); t \geq 0\}$ be a random process, such that $X(t)$ represents the set of events that has occurred at time t. Initially, no events have occurred; i.e., $X(0) = \emptyset$. Since events can only be accumulated, $X(t_1) \subseteq X(t_2)$ if $t_1 < t_2$. Also, if Ⓢ occurs at time t_i, no additional events are allowed to occur and, hence, $X(t_j) = X(t_i)$ for all $t_j \geq t_i$.

Now, given that a set Q of events has occurred where Ⓢ $\notin Q$, all events in $Q^C = ([n] \setminus Q) \cup \{Ⓢ\}$ compete to become the next event to occur. The following assumptions are made.

(I) The time until event $x \in Q^C$ occurs in state Q, denoted T_x^Q, is exponentially distributed with intensity $\Lambda_x(Q)$.
(II) All T_x^Q for $x \in Q^C$ are independent.

Assumption (I) means that the process lacks memory (Norris, 1997); that is,

$$P(T_x^Q > s + t | T_x^Q > t) = P(T_x^Q > s).$$

This is a standard assumption for modeling failure of a system component.

Let T_x and T_y represent the time until event x and y occurs. Since x may affect the risk for y and vice versa, T_x and T_y are not guaranteed to be independent. However, being in a state Q, the number of events that have occurred stays constant, and thus assumption (II) seems reasonable.

Together, assumptions (I) and (II) give that $\{X(t); t \geq 0\}$ is a Markov chain process with the transition probability between state Q and $Q \cup \{x\}$ (Norris, 1997) given by

$$p_{Q,Q\cup\{x\}} = \frac{\Lambda_x(Q)}{\sum_{y \in Q^C} \Lambda_y(Q)}. \tag{1}$$

Furthermore, the probability for $d_{\sigma_1}, \ldots, d_{\sigma_k}$ occurring in this order is given by

$$p_{\emptyset,\{d_{\sigma_1}\}} \cdot \ldots \cdot p_{\{d_{\sigma_1},\ldots,d_{\sigma_{i-1}}\},\{d_{\sigma_1},\ldots,d_{\sigma_i}\}} \cdot \ldots \cdot p_{\{d_{\sigma_1},\ldots,d_{\sigma_{k-1}}\},D} \cdot p_{D,D\cup\{Ⓢ\}}. \tag{2}$$

Thus, each subset of $[n] \cup \{Ⓢ\}$ is a state in the Markov chain, and all states containing Ⓢ are absorbing states.

Figure 1 shows the Markov chain when $n = 2$. For example, in state $\{1\}$, 2 and Ⓢ compete to become the next event to occur, where $p_{\{1\}\{1,2\}} = \frac{\Lambda_2(\{1\})}{\Lambda_2(\{1\})+\Lambda_Ⓢ(\{1\})}$ and $p_{\{1\}\{1,Ⓢ\}} = \frac{\Lambda_Ⓢ(\{1\})}{\Lambda_Ⓢ(\{1\})+\Lambda_2(\{1\})}$.

4. THE NETWORK ABERRATION MODEL

Generally, a discrete Markov chain model requires one parameter for each pair of states. For the Markov chain model introduced in Section 3, this gives an exponential number of parameters in n. Instead, we will now present our network aberration model (NAM) that requires only a quadratic number of parameters.

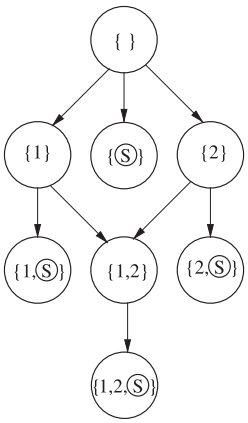


FIG. 1. A Markov chain with two aberrations.

4.1. Definition

A NAM M is a triple $M = (\lambda, \delta, \psi)$, where

- $\lambda = \{\lambda_1, \ldots, \lambda_n\}$ is a set of aberration intensities,
- $\delta = \{\delta_{ij} : 1 \leq i, j \leq n, i \neq j\}$ is a set of pairwise dependencies, and
- $\psi = \{\psi_0, \ldots, \psi_n\}$ is a set of stop intensities.

The parameter λ_i represents the intensity for aberration i in the starting state, i.e., in $\emptyset$. Furthermore, the parameter δ_{ij} represents how the intensity for aberration j is changed when aberration i occurs and has the following interpretation: if $i \neq j \in Q^C \setminus \text{Ⓢ}$, then

$$\Lambda_j(\{Q, i\}) = \Lambda_j(Q) \cdot \delta_{ij}.$$

This assumption, namely that an aberration i always affects the intensity of an aberration j with the same factor, no matter when i occurs, gives the intensity for $j \in Q^C \setminus \{\text{Ⓢ}\}$ in general

$$\Lambda_j(Q) = \lambda_j \prod_{i \in Q} \delta_{ij}.$$

The NAM is based on the reasonable assumption that an aberration can increase or leave the probability of another aberration unchanged, i.e., that it cannot decrease the probability. So, $d_{ij} \geq 1$ for all $1 \leq i, j \leq n$ such that $i \neq j$. If $\delta_{ij} = \delta_{ji} = 1$, we say that aberration i and j are independent. A NAM where all pairs of aberrations are independent is said to be *independent*.

Concerning the intensity for Ⓢ, it seems reasonable that a late state and an early state differ w.r.t. the expected time until a tumor is discovered. However, all states of a NAM of equal size are assumed to have the same intensity for Ⓢ, where ψ_i denote the intensity for Ⓢ in states of size i; i.e.,

$$\Lambda_{\text{Ⓢ}}(Q) = \psi_{|Q|}.$$

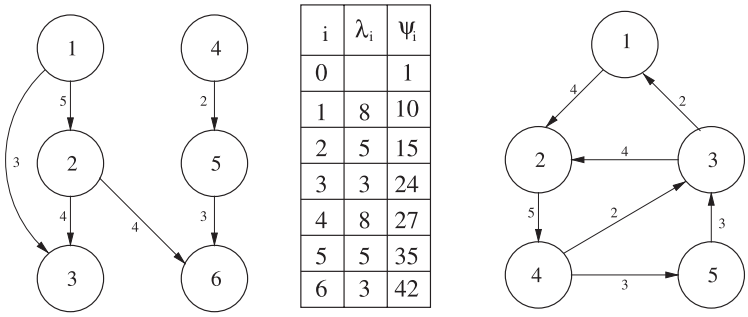
This completes the definition of all the intensities involved in (1).

4.2. Likelihood computation

The likelihood of a NAM M given a sample $D = \{d_1, \ldots, d_k\}$ is obtained by summing over all possible scenarios. That is,

$$L_D(M) = P(D|M) = \sum_{\sigma \in S_k} p_{\emptyset, \{d_{\sigma_1}\}} \cdot \ldots \cdot p_{\{d_{\sigma_1}, \ldots, d_{\sigma_{k-1}}\}, D} \cdot p_{D, D \cup \{\text{Ⓢ}\}}, \tag{3}$$

where S_k is the set of all permutations of $\{1, \ldots, k\}$. Computing (3) can be done in time $O(n2^k)$ using dynamic programming (Durbin *et al.*, 1998).



i	λ_i	ψ_i
0		1
1	8	10
2	5	15
3	3	24
4	8	27
5	5	35
6	3	42

FIG. 2. (**a**) A simple DG. (**b**) A table that defines λ and ψ. (**c**) A more complex DG.

4.3. Dependency graph

A dependency graph (DG) of a NAM is a graphical representation of the dependency parameters. In a DG, each node represents an aberration, and there exists a weighted directed edge from node i to node j labeled δ_{ij} if and only if $\delta_{ij} > 1$. A DG facilitates a visualization of the most common pathways in a disease, typically a cancer form, represented by a Markov chain.

Consider the DG in Fig. 2a. As can be seen, aberration 1 increases the intensity of aberrations 2 and 3, aberration 2 increases the intensity of aberrations 3 and 6, and so on. Hence, given a sample $\{1, 2, 3\}$, $1, 2, 3$ is a likely scenario; i.e., aberration 1 was first to occur, aberration 2 the second to occur, and so on. In particular, if λ and ψ take the values in Fig. 2b, the likelihood ratio $\frac{P(1,2,3|\lambda,\delta,\psi)}{P(3,2,1|\lambda,\delta,\psi)} \approx 24.6$.

One advantage of NAMs is that paths do not have to be followed strictly. Given a sample $\{1, 2, 3, 4\}$ and the same NAM as described above, the path $1 \to 2 \to 3$ affects the probability also of other orderings; for example, $1, 2, 4, 3$ is more likely than $3, 2, 4, 1$. Another advantage is that the probability for Ⓢ as the next event is allowed to vary between states of equal size. For example, in state $\{1, 2, 6\}$, aberration 3 is very likely to occur, and therefore $p_{\{1,2,6\},\{1,2,6,Ⓢ\}}$ is only 0.33. On the other hand, in state $\{4, 5, 6\}$, aberration 1, 2, and 3 have the same intensity as in $\emptyset$, and $p_{\{4,5,6\},\{4,5,6,Ⓢ\}} = 0.60$.

Now, consider the DG in Fig. 2c. Obviously, the most common paths in the Markov chain are all but easy to capture and, apparently, we need a way to limit the pairwise dependencies. We leave this issue for the moment and return to it in Section 8.

5. LEARNING A NAM

In this section, we describe a method to learn a NAM from a set $D = \{D_1, \ldots, D_m\}$ of samples. We will use the approach of maximum likelihood estimation, i.e., search for the assignment $\hat{M}$ that maximizes the likelihood, $L_D(M) = \prod_{i=1}^{m} L_D(M)$, or equivalently, the log likelihood, $\log L_D(M) = \sum_{i=1}^{m} \log L_D(M)$. Due to the complexity of the likelihood function, finding $\hat{M}$ is not an easy problem, and we consider heuristic methods. Most likely, one can devise an EM algorithm for learning a NAM; we have chosen not to do so, for two reason: our learning method is easy to implement, and it gives very good results. The algorithm can briefly be described as follows.

1. The parameters λ, δ, and ψ are initialized (initialization step).
2. $(\lambda^{opt}, \delta^{opt}, \psi^{opt}) \leftarrow (\lambda, \delta, \psi)$.

Then, the algorithm iterates the following four steps until a local maximum is reached:

3. $\delta \leftarrow \delta^{opt}$.
4. One parameter in δ is changed (modification step).
5. Values of λ and ψ are found, by a heuristic method, such that $L_D(\lambda, \delta, \psi)$ is close to $\max_{\lambda,\psi} L_D(\lambda, \delta, \psi)$ (calibration step).
6. If $L_D(\lambda, \delta, \psi) > L_D(\lambda^{opt}, \delta^{opt}, \psi^{opt})$, then $(\lambda^{opt}, \delta^{opt}, \psi^{opt}) \leftarrow (\lambda, \delta, \psi)$.

The initialization, modification, and calibration steps are presented in more detail below.

5.1. Initialization step

First, δ is set so that all aberrations are independent. That is,

$$\delta_{ij} \leftarrow 1, \text{for all } i \text{ and } j \text{ such that } 1 \leq i, j \leq n, i \neq j.$$

Then, all intensity parameters are assigned the same value l, i.e.,

$$\lambda_i \leftarrow l, \text{ for all } i = 1 \ldots n, \text{ and}$$

$$\psi_i \leftarrow l, \text{ for all } i = n+1 \ldots 2n+1.$$

Finally, the initialization step ends with a call to the calibration method (see Section 5.3), which modifies λ and ψ.

5.2. Modification step

In the modification step, one pairwise dependency parameter, say δ_{jk}, is picked uniformly at random. The modification considered for δ_{jk} is with equal probability to subtract one from it or to add one to it, except if $\delta_{jk} = 1$, then $\delta_{jk} \leftarrow \delta_{jk} + 1$ with probability one. Thus, we restrict dependency parameters to be positive integers.

5.3. Calibration step

For a set $X = (X_1, \ldots, X_k)$ of samples, let

$$X_{[j]} = \begin{cases} \{X_i \in X : j \in X_i\} & \text{if } 1 \leq j \leq n \\ \{X_i \in X : |X_i| = j - (n+1)\} & \text{if } (n+1) \leq j \leq 2n+1. \end{cases}$$

For efficiency, data generation probability computations are avoided during calibration. Instead, how well a choice of intensities λ and ψ fit the data D, given that δ is known, is evaluated by generating a number of synthetic data sets $S^1, \ldots, S^r$ from (λ, δ, ψ), each of size m, and calculating the degree of over- and underrepresentation in $S = \{S^1, \ldots, S^r\}$ compared to D. We say that λ_j is underrepresented in S^i if $|S^i_{[j]}| < |D_{[j]}|$ and overrepresented if $|S^i_{[j]}| > |D_{[j]}|$. Similarly, ψ_j is underrepresented in S^i if $|S^i_{[j+(n+1)]}| < |D_{[j+(n+1)]}|$ and overrepresented if $|S^i_{[j+(n+1)]}| > |D_{[j+(n+1)]}|$. Now, let

$$I(S^i, j) = \begin{cases} 1 & \text{if } |S^i_{[j]}| > |D_{[j]}| \\ -1 & \text{if } |S^i_{[j]}| < |D_{[j]}| \\ 0 & \text{otherwise.} \end{cases}$$

We say that λ_j is the most underrepresented parameter in S if

$$\sum_{i=1}^{r} I(S^i, j) = \min_{k=1,\ldots,2n+1} \sum_{i=1}^{r} I(S^i, k)$$

and the most overrepresented if

$$\sum_{i=1}^{r} I(S^i, j) = \max_{k=1,\ldots,2n+1} \sum_{i=1}^{r} I(S^i, k).$$

Similarly, we say that ψ_j is the most underrepresented parameter in S if

$$\sum_{i=1}^{r} I(S^i, j + (n+1)) = \min_{k=1,\ldots,2n+1} \sum_{i=1}^{r} I(S^i, k)$$

and the most overrepresented if

$$\sum_{i=1}^{r} I(S^i, j + (n+1)) = \max_{k=1,\ldots,2n+1} \sum_{i=1}^{r} I(S^i, k).$$

The calibration algorithm iterates the following two steps:

1. A number of synthetic datasets $S^1, \ldots, S^r$ are generated from (λ, δ, ψ), each of size m.
2. If $\sum_{i=1}^{r} I(S^i, j) \approx 0$, for each $1 \leq j \leq 2n+1$, then $L_D(\lambda, \delta, \psi)$ is considered to be close to $\max_{\lambda,\psi} L_D(\lambda, \delta, \psi)$ and λ and ψ are accepted. If not, one intensity unit is passed from the most overrepresented parameter in $S = \{S^1, \ldots, S^r\}$ to the most underrepresented parameter in S.

Note that this approach allows intensity units to be passed between aberration intensity parameters and stop intensity parameters.

To increase the speed of the calibration, we introduce several calibration optimizations. For example, in an early calibration stage, we pass more than one intensity unit between the two parameters.

6. MODULE NAM

As stated in Section 2, a NAM requires $n^2 + n + 1$ parameters. In our ambition to further reduce the number of parameters, we will in this section define a module NAM (MNAM) inspired by the module network of Segal *et al.* (2003).

We start with some basic definitions. Two aberrations i and j are said to *share* dependencies if

1. $\delta_{xi} = \delta_{xj}$, for all $x \in [n]$ such that $x \neq i, j$,
2. $\delta_{ix} = \delta_{jx}$, for all $x \in [n]$ such that $x \neq i, j$, and
3. $\delta_{ij} = \delta_{ji}$.

Figure 3a shows a DG where aberrations 2, 3, and 4; and 5 and 6 share dependencies, respectively.

We define a module to be a nonempty set of aberrations, such that if two aberrations belong to the same module then they share dependencies. Figure 3b shows a partition of the aberrations in Fig. 3a into three modules. Since all aberrations within a module share dependencies, all directed edges leading from one module i to another module j have the same weight, something we call a module dependency and denote $\mathcal{D}_{ij}$. For example, the module dependency from module 2 to module 3 in Fig. 3b is 4, or shorter, $\mathcal{D}_{23} = 4$. In a MDG (Fig. 3c), nodes are modules and weighted directed edges between modules represent module dependencies. Furthermore, in a MDG, a weighed loop represents pairwise dependencies between aberrations within a single module.

Intuitively, a MNAM is simply a NAM where pairwise dependencies are given by a module DG (MDG) instead of by a DG. Formally, a MNAM M is a quadruple $M = (\lambda, \mathcal{C}, \mathcal{D}, \psi)$, were $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$ represents the set of modules and $\mathcal{D} = \{\mathcal{D}_{ij} : 1 \leq i, j \leq N\}$ the set of module dependencies. Since

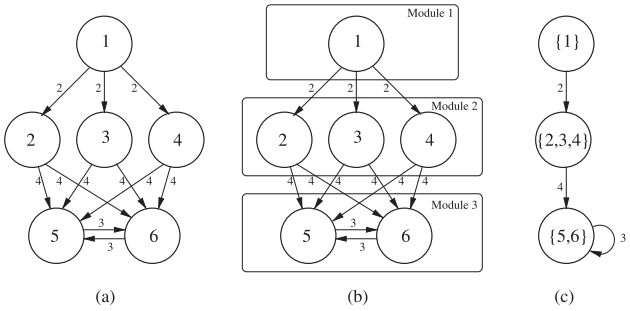


FIG. 3. (**a**) A DG. (**b**) The same DG partitioned into three modules. (**c**) A MDG representation of the DG.

$\mathcal{C}$ and $\mathcal{D}$ induce δ in a natural way, a MNAM induces a NAM. Obviously, in order to benefit from a MNAM, many aberrations must, to a first approximation, share dependencies. This is, however, not an unrealistic assumption (Höglund *et al.*, 2001). If $\mathcal{C}$ is known, the total number of parameters for a MNAM is $N^2 + 2n + 1$ where, hopefully, $N \ll n$. Of course, $\mathcal{C}$ is unknown, and we will deal with this problem in the next section.

7. LEARNING A MNAM

In this section, we describe a method to learn a MNAM from a set $D = \{D_1, \ldots, D_m\}$ of samples. Note that if $\mathcal{C}$ is known, we can use the NAM learning algorithm described in Section 5 to learn a MNAM. The only modification needed is that instead of changing one parameter in δ explicitly in each modification step, we change one parameter in $\mathcal{D}$ and, hence, one or more parameters in δ implicitly. This modified version of the NAM learning method is applied to various module configurations in a search for a good configuration. The algorithm can briefly be described as follows.

1. The parameters λ, $\mathcal{C}$, $\mathcal{D}$, and ψ are initialized (initialization step).
2. $(\lambda^{opt}, \mathcal{C}^{opt}, \mathcal{D}^{opt}, \psi^{opt}) \leftarrow (\lambda, \mathcal{C}, \mathcal{D}, \psi)$.

Then, the algorithm iterates the following four steps until a local maximum is reached:

3. $\mathcal{C} \leftarrow \mathcal{C}^{opt}$.
4. The multiset $\mathcal{C}$ is modified (modification step).
5. The values of λ, $\mathcal{D}$, and ψ are determined by the modified NAM learning algorithm.
6. If $\text{score}(\lambda, \mathcal{C}, \mathcal{D}, \psi) > \text{score}(\lambda^{opt}, \mathcal{C}^{opt}, \mathcal{D}^{opt}, \psi^{opt})$, then $(\lambda^{opt}, \mathcal{C}^{opt}, \mathcal{D}^{opt}, \psi^{opt}) \leftarrow (\lambda, \mathcal{C}, \mathcal{D}, \psi)$ (greedy step).

The initialization, modification, and greedy steps are presented in more detail below.

7.1. Initialization step

First, all aberrations starts from a single module, i.e.,

$$\mathcal{C}_1 \leftarrow [n].$$

Second, a call to the modified NAM learning algorithm assigns λ, $\mathcal{D}$, and ϕ.

7.2. Modification step

In the modification step, $\mathcal{C}$ is modified by one of the following four operations: *Split, Merge, Move*, or *Swap*, chosen at random. The *Split* operation splits one randomly chosen module into two, such that each aberration in the original module has the same probability to appear in any of the two new modules. For large modules, however, the *Split* operation is extended such that a split can result in more than two new modules. The *Merge* operation merges two randomly chosen modules into one, whereas the *Move* operation moves one randomly chosen aberration from one module to another. Finally, the *Swap* operation changes the module membership of two randomly chosen aberrations.

7.3. Greedy step

When choosing between $(\lambda, \mathcal{C}, \mathcal{D}, \psi)$ and $(\lambda^{opt}, \mathcal{C}^{opt}, \mathcal{D}^{opt}, \psi^{opt})$, we have to consider that they can be of different complexity, i.e., that $|\mathcal{C}| \neq |\mathcal{C}^{opt}|$. A model with more parameters can more easily fit the data, so evidently, merely comparing the likelihoods is not a good idea. In particular, if all modules in $\mathcal{C}$ are a subset of modules in $\mathcal{C}^{opt}$, then independent of the data

$$\max_{\lambda, \mathcal{D}, \psi} L(\lambda, \mathcal{C}^{opt}, \mathcal{D}, \psi) \geq \max_{\lambda, \mathcal{D}, \psi} L(\lambda, \mathcal{C}, \mathcal{D}, \psi).$$

In order to penalize more complex models, we define a penalizing function $H(\mathcal{D})$ as

$$H(\mathcal{D}) = \epsilon \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} (\mathcal{D}_{ij} - 1),$$

where ϵ is a penalizing parameter. The score for a model M is then defined as

$$\text{score}(M) = \log L_D(M) - H(\mathcal{D}).$$

The major advantages of $H(\mathcal{D})$ are that it is simple to use, it does not penalize complex models when the dependency parameters are independent, and it prevents overfitting since a dependency must be strong to be invoked.

8. LEARNING AN ACYCLIC MNAM

Our final step is a heuristic approach to learn a MNAM such that each cycle in the MDG is a loop. Given such a graph, which we here call acyclic, we get a clear overview of the different tumor progression pathways. First, let us focus on the problem for a DG. In that case, one known approach is to allow only dependencies between two aberrations going from the more frequent aberration to the less frequent (Bulashevska *et al.*, 2004). However, this assumption is not appropriate for converging pathways ($a \rightarrow c \leftarrow b$), where c is very likely to be more frequent than a and b. In order to distinguish early aberrations from late, we will instead define a measure called *average number of aberrations per sample*, denoted AVA. A similar measure has been used by Desper *et al.* (1999). For an aberration j and a set $X = (X_1, \ldots, X_k)$ of samples, $\text{AVA}(j, X)$ is defined as the average size of the sets in X where j is present. More formally,

$$\text{AVA}(j, X) = \frac{\sum_{X_i \in X_{[j]}} |X_i|}{|X_{[j]}|}.$$

The idea is then to allow a dependency from u to v if and only if

$$\text{AVA}(u, X) \leq \text{AVA}(v, X).$$

For a module $\mathcal{C}_i$, we simply define $\text{AVA}(\mathcal{C}_i, X)$ as

$$\text{AVA}(\mathcal{C}_i, X) = \sum_{j \in \mathcal{C}_i} \frac{|X_{[j]}|}{\sum_{k \in \mathcal{C}_i} |X_{[k]}|} \text{AVA}(j, X).$$

In the same manner, we allow a dependency between module $\mathcal{C}_u$ and $\mathcal{C}_v$ if and only if $\text{AVA}(\mathcal{C}_u, X) \leq \text{AVA}(\mathcal{C}_v, X)$. Note that these dependency restrictions are active during the whole learning process; i.e., the MDG is not pruned in the end. This is very important to point out since pruning in the end, as in Bulashevska *et al.* (2004), can totally change the behavior of the model.

9. PENALTY PARAMETER SELECTION

Before applying the MNAM learning, we have to specify the penalty parameter ϵ. In order to find the best value of ϵ for the colorectal cancer (CC) data, we used a 5-fold cross-validation approach which now will be described.

The CC data was randomly partitioned into five equally sized partitions. For each value of ϵ considered, we (1) trained five models, each on a different set of four of the five partitions, (2) calculated the log-likelihood for each of the five models with respect to its held out partition, and (3) summarized the log-likelihoods obtained in step 2.

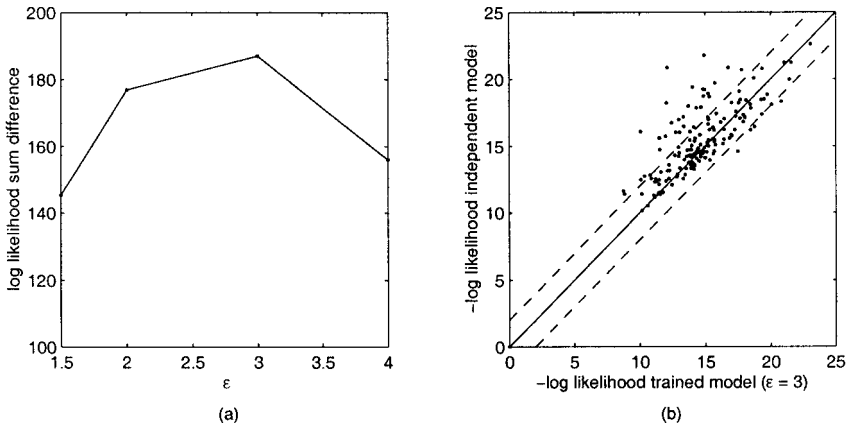


FIG. 4. (**a**) Cross validation log-likelihood sum difference between the trained model and the independent model for various values of ϵ. (**b**) Cross validation log-likelihood comparison between the trained model and the independent model for individual samples with more than four aberrations ($\epsilon = 3$).

Steps 2 and 3 were also done for five independent models, i.e., five models, each initialized on a different set of four of the five partitions. We then calculated the difference between the log-likelihood sum of the trained models for ϵ = 1.5, 2, 3, and 4 and the log-likelihood sum of the independent models (Fig. 4a). The trained models performs much better on unseen data than the independent models, and the best trained models are obtained for $\epsilon = 3$. As expected, a too low value of ϵ gives overtrained models, and a too high value of ϵ gives models too close to the independent model.

To be more specific, we compared the log-likelihood for each sample with more than four aberrations obtained in step 2 for the trained models with $\epsilon = 3$ and the independent models (Fig. 4b). As can be seen, many samples are much more likely to be generated by a trained model compared to an independent model, but very few vice versa.

10. VALIDATION

In this section, we derive and apply a goodness of fit measure (Section 10.1) and investigate the uniqueness of the optimal solution found by the learning algorithm (Section 10.2).

10.1. Goodness of fit

To quantify a model's goodness of fit, we want to measure how closely the model reproduces the empirical distribution. Standard measures such as the cosine distance (for example, used by Beerenwinkel *et al.* [2004]) are not well suited for our data, because of the huge sample space. It is easy to obtain an ML estimation of the independent model, which will have the best possible fit to the data if the number of occurrences of each aberration is measured. Since our main interest is the dependencies between aberrations, we introduce a measure called *co-occurrence distance*, measuring how frequently two aberrations occur together.

First, let the co-occurrence of aberration i and j in a set $X = (X_1, \ldots, X_k)$ of samples, denoted $\mathrm{CO}_X(i, j)$, be defined as

$$\mathrm{CO}_X(i, j) = |X_{[i]} \cap X_{[j]}|.$$

The co-occurrence distance between X and Y is then defined as the sum of squares of co-occurrence differences between X and Y; i.e.,

$$\mathrm{dist}(X, Y) = \sum_{1 \le i < j \le n} (\mathrm{CO}_X(i, j) - \mathrm{CO}_Y(i, j))^2.$$

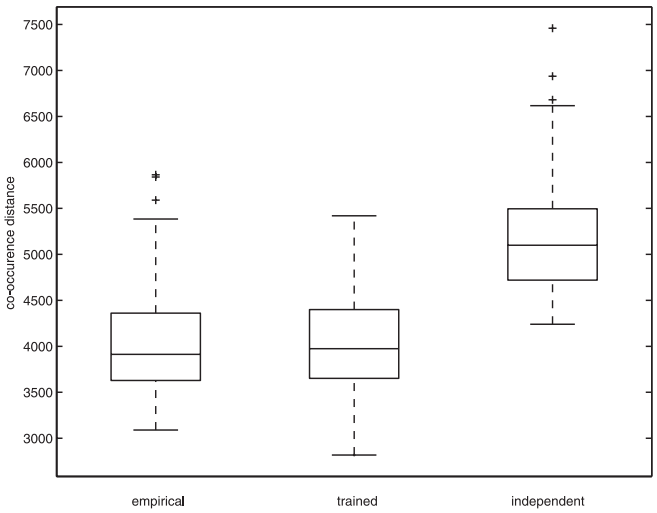


FIG. 5. Cross validation co-occurrence distance between the test data and the empirical, trained, and the independent distribution, respectively.

We do not normalize this measure, since more frequent aberrations should have a larger influence than less frequent aberrations. Using cross-validation in the same way as described above, we sum for each partition into test and training data, the co-occurrence distance between the test data and data, of equal size, generated from

1. the empirical distribution induced by the training data,
2. the trained model, and
3. the independent model.

Figure 5 shows a box-plot of 100 5-fold cross-validation runs. While the trained model performs nearly as well as the empirical training distribution, the independent model shows poor model fit.

10.2. Uniqueness of obtained solution

When the MNAM learning algorithm is applied to the same set of data several times, it should preferably converge to the same, or at least approximately the same, MDG.

In order to investigate the uniqueness of the obtained solution for the CC data, we trained five models on the whole dataset. The MDG for the best model, i.e., the one with the highest likelihood, is presented in Fig. 6. As desired, all five MDGs were nearly identical. First, all five models predicted seven modules, and second, at most two *move* operations (see Section 7.2) were needed to go from the MDG in Fig. 6 to any of the other four MDGs. Aberrations predicted to belong to different modules were -10, -14p, and -15, i.e., very late events in the MDG.

11. DISCUSSION

We give a generative probabilistic model, based on realistic assumptions, for chromosomal aberrations. The model can be used to generate synthetic datasets, where data points are generated aberration by aberration. This facilitates studies of process dynamics and evaluations based on comparisons of datasets, such as the co-occurrence tests we perform.

Tree models, such as those described in Desper *et al.* (1999, 2000), have several disadvantages. First, the distance-based tree models are nonprobabilistic. Second, there is no reason to believe that the distances obtained from chromosomal aberration data are additive (i.e., can be represented by a tree) or close to

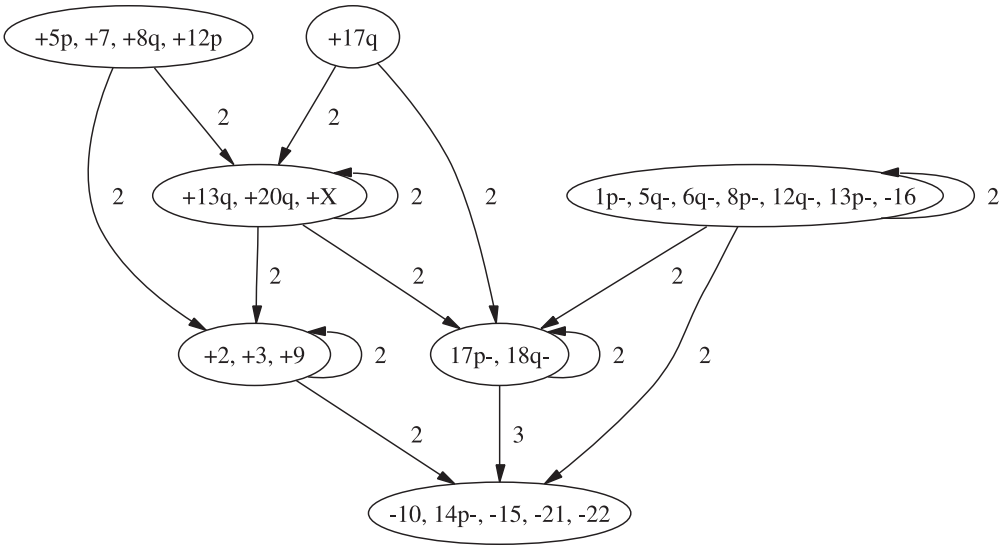


FIG. 6. The MDG with the highest likelihood from the CC data.

additive. Third, there is no natural interpretation of proximity between two aberrations in the tree. Finally, a tree cannot simultaneously represent diverging and converging pathways. The latter is a severe problem, since such pathways are common as well as important (Höglund *et al*., 2005).

Recently, Bulashevska *et al*. proposed Bayesian networks for analysis of oncogenesis. The major disadvantage of Bayesian networks is that an aberration can depend only on a limited number of aberrations or the parental conditional probability function will be overfitted. For example, in Bulashevska *et al*. (2004), one aberration depends on six aberrations, which means that 64 conditional probabilities have to be estimated based on merely 123 samples. In our framework, any subset of the aberrations is a state, so the probability of an event is dependent not only on a small subset of the previously occurred events, but on all aberrations that have occurred. This reflects the fact that tumor progression is a historical process. Also, we make sure not to overfit our model by not introducing more pairwise dependencies than the data allow for. When, for instance, the MDG in Fig. 6 is compared with an alternative more detailed description, the possibility that the more detailed one is overfitted must be taken into account.

Another problem with Bayesian networks occurs when there are several starting points. In reality, the starting points are typically negatively correlated, since after one has been acquired the progression has started and it is likely that tumor discovery occurs before the other starting points occur. However, in a Bayesian network, only one starting point in any pair of starting points can depend directly on the other. In our framework, such negative correlations are accomplished by an increased intensity for a subset of other aberrations.

The model we derive of karyotypic evolution in colorectal cancer fits well with previous results obtained by methods based on correlation. Höglund *et al*. (2002) presented a model of the karyotypic evolution obtained through principal component analysis. This analysis revealed two major pathways of which one was dominated by gains, preferentially +7, +8q, +17q, +12p, +5p, +9, +13q, and +20 and a second by losses, particularly by 1p−, 5q−, 8p−, and 6q−. Both of these pathways are reproduced in the MDG with the highest likelihood (Fig. 6). Interestingly, 17p− and 18q−, which together form a central late node in the MDG in Fig. 6, were important changes in both pathways. Furthermore, previous cytogenetic and molecular studies have pointed to the importance of +7, +13, and +20, and 1p− as early changes in colorectal tumors (Bomme *et al*., 1996, 2001; Bardi *et al*., 1993).

Again, in contrast to the model derived by Radmacher *et al*., which does not explain their data significantly better than a model where independence is assumed, the model we derive does significantly better. This can possibly be explained by a combination of (i) the novel training procedure, (ii) the further reduction of parameters, and (iii) the different modeling of tumor discovery. Another possibility is that we consider copy number aberrations in colorectal cancer while Radmacher *et al*. consider breakpoints in melanoma.

REFERENCES

Bardi, G., Pandis, N.C.F., Kronborg, O., Bomme, L., and Heim. S. 1993. Deletion of 1p36 as a primary chromosomal aberration in intestinal tumorigenesis. *Cancer Res.* 8, 1895–1898.

Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D, Kaiser, R., Selbig, J., and Lengauer, T. 2004. Learning multiple evolutionary pathways from cross-sectional data. *Proc. 8th Ann. Int. Conf. on Research in Computational Biology*, 36–44.

Bomme, L., Bardi, G., Pandis, N., Fenger, C., Kronborg, O., and Heim, S. 1996. Chromosome abnormalities in colorectal adenomas: Two cytogenetic subgroups characterized by deletion of 1p and numerical aberrations. *Human Pathol.* 11, 1192–1197.

Bomme, L. Lothe, R., Bardi, G., Fenger, C., Kronborg, O., and Heim, S. 2001. Assessments of clonal composition of colorectal adenomas by fish analysis of chromosomes 1, 7, 13 and 20. *Int. J. Cancer* 6, 816–823.

Bulashevska, S., Szakacs, O., Brors, B., Eils, R., and Kovacs, G. 2004. Pathways of urothelial cancer progression suggested by Bayesian network analysis of allelotyping data. *Int. J. Cancer* 110, 850–856.

Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H., and Schaffer, A.A. 1999. Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comp. Biol.* 6, 37–51.

Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H., and Schaffer, A.A. 2000. Distance-based reconstruction of tree models for oncogenesis. *J. Comp. Biol.* 7, 789–803.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis*, Cambridge University Press, London.

Fearon, E.R., and Vogelstein, B. 1990. A genetic model for colorectal tumorigenesis. *Cell* 61, 759–767.

Hanahan, D., and Weinberg, R.A. 2000. The hallmarks of cancer. *Cell* 100, 57–70.

Heim, S., and Mitelman, F. 1995. *Cancer Cytogenetics*, Wiley–Liss, New York.

Höglund, M., Frigyesi, A., Säll, T., Gisselsson, D., and Mitelman, F. 2005. Statistical behavior of complex cancer karyotypes. *Gene Chromosome Cancer* 42, 327–341.

Höglund, M., Gisselsson, D., Hansen, G.B., Säll, T., Mitelman, F., and Nilbert, M. 2002. Dissecting karyotypic patterns in colorectal tumors: Two distinct but overlapping pathways in the adenoma-carcinoma transition. *Cancer Res.* 62, 5939–5946.

Höglund, M., Gisselsson, D., Mandahl, N., Johansson, B., Mertens, F., Mitelman, F., and Säll, T. 2001. Multivariate anayses of genomic imbalances in solid tumors reveal distinct and converging pathways of karyotypic evolution. *Gene Chromosome Cancer* 31, 156–171.

Mitelman, F., Johansson, B., and Mertens, F. 2004. Mitelman database of chromosome aberrations in cancer. *http://cgap.nci.nih.gov/Chromosomes/Mitelman*.

Mitelman, F., Mertens, F., and Johansson, B. 1997. A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.* 15(Spec. No.), 417–474.

Norris, J.R. 1997. *Markov Chains*, Cambridge University Press, London.

Radmacher, M.D., Simon, R., Desper, R., Taetle, R., Schaffer, A.A., and Nelson, M.A. 2001. Graph models of oncogenesis with an application to melanoma. *J. Theor. Biol.* 212, 535–548.

Segal, E., Pe'er, D., Regev, A., Koller, D., and Friedman, N. 2003. Learning module networks. *Proc. 19th Conf. on Uncertainty in Artificial Intelligence*, 525–534.

Simon, R., Desper, R., Papadimitriou, C.H., Peng, A., Alberts, D.S., Taetle, R., Trent, J.M., and Schaffer, A.A. 2000. Crhomosome abnormalities in ovarian adenocarcinoma: III. Using breakpoint data to infer and test mathematical models of oncogenesis. *Gene Chromosome Cancer* 28, 106–120.

Address correspondence to:
Jens Lagergren
Dept. of Numerical Analysis and Computer Science
KTH, Stockholm
SE-106 91, Sweden

E-mail: jensl@nada.kth.se