BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 17 2009, pages 2208–2215
doi:10.1093/bioinformatics/btp365

Gene expression

Inferring progression models for CGH data

Jun Liu[1], Nirmalya Bandyopadhyay[1,*], Sanjay Ranka[1], M. Baudis[2] and Tamer Kahveci[1,*]
[1]Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA and [2]Institute for Molecular Biology, University of Zurich, Zurich, Switzerland

Received on March 11, 2009; revised on May 18, 2009; accepted on June 7, 2009
Advance Access publication June 15, 2009
Associate Editor: David Rocke

ABSTRACT

Motivation: One of the mutational processes that has been monitored genome-wide is the occurrence of regional DNA *copy number alterations (CNAs)*, which may lead to deletion or over-expression of tumor suppressors or oncogenes, respectively. Understanding the relationship between CNAs and different cancer types is a fundamental problem in cancer studies.

Results: This article develops an efficient method that can accurately model the progression of the cancer markers and reconstruct evolutionary relationship between multiple types of cancers using comparative genomic hybridization (CGH) data. Such modeling can lead to better understanding of the commonalities and differences between multiple cancer types and potential therapies. We have developed an automatic method to infer a graph model for the markers of multiple cancers from a large population of CGH data. Our method identifies highly related markers across different cancer types. It then builds a directed acyclic graph that shows the evolutionary history of these markers based on how common each marker is in different cancer types. We demonstrated the use of this model in determining the importance of markers in cancer evolution. We have also developed a new method to measure the evolutionary distance between different cancers based on their markers. This method employs the graph model we developed for the individual markers to measure the distance between pairs of cancers. We used this measure to create an evolutionary tree for multiple cancers. Our experiments on Progenetix database show that our markers are largely consistent to the reported hot-spot imbalances and most frequent imbalances. The results show that our distance measure can accurately reconstruct the evolutionary relationship between multiple cancer types.

Availability: All the code developed in this article are available at http://bioinformatics.cise.ufl.edu/phylogeny.html.

Contact: nirmalya@cise.ufl.edu; tamer@cise.ufl.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

This article develops a systematic way of understanding the progression of multiple types of cancers by analyzing aberrations in gene copy numbers. Alterations in the tumor genome affects the progression of tumors. It has been argued that the oncogenomic evolution leaves characteristic signatures of inheritance, thereby allowing to infer models of tumor progression by identification of these signatures in genome-wide mutational data (Bilke *et al.*, 2005). One of the mutational processes that can be monitored genome-wide is the occurrence of regional DNA *copy number alterations (CNAs)*, which may lead to deletion or over-expression of tumor suppressors or oncogenes, respectively.

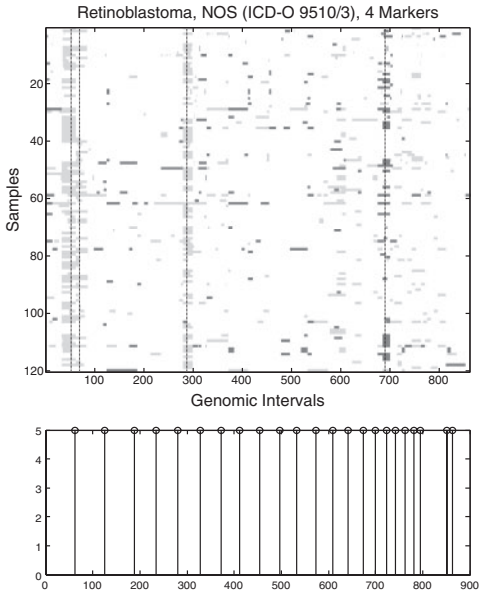

Fig. 1. A plot illustrating the copy number status of the CGH data for 121 cases belonging to the retinoblastoma cancer type. The *X*- and *Y*-axis denote the genomic intervals and samples, respectively. The gain and loss alterations are plotted in light gray and dark gray, respectively. The vertical lines show four cancer markers. The 25 irregularly spaced vertical lines at the bottom of the plot shows the starting/ending positions of the chromosomes, 1, 2, …, 22, *x* and *y* on the CGH data.

The distribution of CNAs in a given cancer type is not random, and alterations occur at recurrent locations. For a broad range of cancers or subtypes of the same clinico-pathological cancer entity, characteristic patterns of recurrent alterations have been observed (Forozan, 1997, Fig. 1). We call a CNA *recurrent* if it is found at the same location in sufficiently large percentage of the observed samples. Such recurrent alterations are also called *markers*. For example, Figure 1 highlights four markers. More than one set of alterations can trigger the same type of cancer. In other words, a cancer type can have several signatures. We call the disorder resulting from such alterations as *subtypes* of the same cancer.

An important method for genome-wide CNA screening is comparative genomic hybridization (CGH) (Kallioniemi *et al.*,

*To whom correspondence should be addressed.

1992). CGH is a molecular-cytogenetic analysis technique for detecting regions with genomic imbalances (gains or losses of DNA segments). Applying microarray technology to CGH allows the simultaneous, sequence-specific detection of the copy number state of thousands of individual DNA fragments (Pinkel and Albertson, 2005). Raw data from CGH experiments is expressed as the ratio of normalized fluorescence of tumor and reference DNA. Normalized CGH ratio data surpassing predefined thresholds are considered indicative for genomic gains or losses, respectively. For chromosomal CGH, several ratio measurements are used for the calculation of the regional copy number state (Jain *et al.*, 2001), while for array CGH various methods of averaging results from spatially related measurements are used (Hsu *et al.*, 2005). Chromosomal and array CGH data have proven an important resource for cancer cytogenetics (Desper *et al.*, 1999; Gray *et al.*, 1994; Hoglund *et al.*, 2005; Joos *et al.*, 2002; Mattfeldt *et al.*, 2001; Vandesompele *et al.*, 2005). For the communication of chromosomal CGH results (on which this article is based), a reverse *in situ* karyotype format (Mitelman, 1995) is used, describing imbalanced genomic regions with reference to their chromosomal location.

Existing works infer tumor progression models based on genetic events such as recurrent CNAs. Their models describe the evolutionary relationship between events and consequently expose the progression and development of tumors. One of the existing works by Bilke *et al.* (2005), focus on the progression of individual recurrent alterations. The time complexity of this approach grows exponentially with the number of cancer types. The Progenetix (Baudis and Cleary, 2001) database contains 20 different cancer types. The total number of cancer subtypes is even much larger than this. Liu *et al.* (2007) showed that on the average, each cancer can be explained with around six different marker sets. Thus, if we assume that each cancer is triggered by five different marker sets on the average, the number of cancer subtypes in nature will easily exceed 100. This makes the method by Bilke *et al*. impractical as its time complexity will exceed 2^{100} for this kind of dataset. A promising approach seems to consider the whole set of alterations of a cancer and infer a model based on the alteration patterns of different cancers. Its time complexity should also be polynomial of the different working parameters.

Such models effectively utilize the molecular characters of cancers and easily extend to large-scale analysis.

Contributions: in this article, our objective is to infer the progression model for multiple cancers (or cancer subtypes) based on the patterns of genetic alterations. (We will use term *cancer subtype* to denote both cancer subtype or stage of cancer.) We assume that similar evolutionary processes act on different cancers, so that closely related cancers exhibit similar alteration patterns. We identify the aberration patterns of a cancer based on the set of key recurrent CNAs in this cancer.

This article has two major technical contributions:

(1) We propose a computational method to infer a graph model for the markers of multiple cancers. We demonstrate the use of this model in determining the importance of markers in cancer evolution.

(2) We develop a new method to measure the evolutionary distance between different cancers based on their markers. We use existing distance matrix methods, such as Fitch–Margoliash, to infer progression models for multiple cancers.

Our experiments on a Progenetix dataset with 5918 CGH cases belonging to 23 clinico-pathological cancer categories (22 specific entities and one 'other') show that our markers are largely consistent to the reported hot-spot imbalances and most frequent imbalances. We also generate phylogenetic trees for 20 cancer entities and 58 cancer subtypes. The results show that cancers with the same histological compositions are well grouped together.

The rest of article is organized as follows. Section 2 briefly introduces the preliminary knowledge. Section 3 extends the Bilke *et al*.'s (2005) approach to infer a graph model for markers and discusses its use. Section 4 proposes the novel distance measure for multiple cancers based on a set of markers. Section 5 presents the experimental results and some observations. Section 6 concludes this article.

2 PRELIMINARIES

In this section, we briefly introduce some preliminary knowledge related to our proposed approach. Section 2.1 presents a brief discussion on the relationship between recurrent CNAs and cancer. In Section 2.2, we discuss the concept of markers, which define the key recurrent CNAs in a cancer. In Section 2.3, we discuss an approach proposed by Bilke *et al.* (2005), which we extend for inferring the progression of markers.

2.1 Markers and tumor development

Researchers have proposed a number of models to infer tumor progression based on genetic alterations, including recurrent CNAs. Vogelstein *et al.* (1988) inferred a chain model of four genetic events for the progression of colorectal cancer. The presence of all four events appears to be critical for colorectal cancer development. Desper and colleagues proposed a branching tree model (Desper *et al.*, 1999) and a distance-based tree model (Desper *et al.*, 2000) by assuming the recurrent CNAs as a set of genetic events that take place in some order. They inferred the models for renal carcinoma to demonstrate the progression of genetic events in that cancer type. Bilke *et al.* (2005) proposed a graph model based on the shared status of recurrent CNAs among different stages of cancer. They found that the pattern of recurrent CNAs in neuroblastoma (NB) is strongly stage dependent. Pennington *et al.* (2006) developed a mutation model for individual tumors and constructed an evolutionary tree for each tumor. They identified a consensus tree model based on the mutations shared by a substantial fraction of the tumor population (Pennington *et al.*, 2006). These and other studies were successful in setting in context prior knowledge about the role of individual cancer-related genes.

2.2 Marker detection

Due to the overlap between neighboring genomic intervals (Liu *et al.*, 2007), recurrent alteration intervals usually accumulate together and form a region of recurrent alterations, which we call *recurrent region*. Given a set of samples that belong to the same cancer, a marker is an independent key recurrent alteration representing a recurrent region. Previously, we proposed a dynamic programming algorithm to identify the best R markers for a set of CGH cases. We demonstrated that our markers capture the aberration patterns well and improve the clustering of CGH

cases (Liu *et al.*, 2007). In Figure 1, we plot the four markers identified in an example set of 121 CGH cases.

CGH data of an individual tumor can be considered as an ordered list of status values, where each value corresponds to a *genomic interval* (e.g. a single chromosomal band). The status can be expressed as a real number (positive, negative or zero for gain, loss or no aberration, respectively). We use this strategy and represent gain, loss and no change with +1, −1 and 0, respectively. Figure 1 plots CGH-derived copy number data from 121 cases of retinoblastoma.

The following notations are used for the rest of the article:

- *Genomic interval*: each chromosome in CGH data consists of an ordered list of intervals called genomic interval. The value of a genomic interval denotes the aberration type of that interval, which can be 0, +1 and −1 for gain, loss or no aberration, respectively.
- *Segment*: a segment is a contiguous array of intervals that have same aberrant status values for all the contained genomic intervals. Formally, $s_j[u,v]$ denotes a continuous run of intervals $\{x^j_u, x^j_{u+1}, \ldots, x^j_v\}$ in chromosome s_j that starts at the u-th interval and ends at the v-th where $x^j_u = x^j_{u+1} = \cdots = x^j_v \neq 0$, $x^j_{u-1} \neq x^j_u$, $x^j_{v+1} \neq x^j_v$.
- *Recurrent region*: a recurrent region is a segment that is present in sufficient number of samples of a cancer type.
- *Recurrent alteration*: a recurrent alteration is a single genomic interval present in an recurrent region.
- *Marker*: a marker is a recurrent alteration that is selected by a marker selection algorithm. If there is more than one recurrent alteration in proximity, only one of them is selected as a marker. Each marker m in a cancer is represented by an ordered pairs $<p, q>$, where p and q denote the position (genomic interval) and the aberration type, respectively. The aberration type of a marker is either gain or loss, denoted by 1 or −1, respectively.
- *Support*: Let S be a set of N CGH cases $\{s_1, s_2, \ldots, s_N\}$. Let x^j_d denotes the alteration value (i.e. 1, −1 or 0) for case j at the d-th genomic interval, $\forall d, 1 \leq d \leq D$, where D is the number of genomic intervals. Let $m = <p, q>$ be a marker. We denote the *independent support* of s_j to m as $\delta(s_j, m)$. Here, $\delta(s_j, m) = 1$ if and only if $x^j_p = q$. Otherwise, $\delta(s_j, m) = 0$. We define the total independent support value of marker m_t as the sum of its support from all the cases. Formally, $\text{Supt}(m) = \sum_{j=1}^{N} \delta(s_j, m)$. We will use term *support* to denote Supt(m) in this article.

2.3 Tumor progression model

Bilke *et al.* (2005) proposed an approach of inferring a tumor progression model for NB with four different subtypes from CGH data. They describe the relationship between different subtypes based on the recurrent alterations shared by these subtypes. Their idea first identifies a set of recurrent alterations. Each recurrent alteration belongs to one of the following three categories: common (shared by all the subtypes), shared (shared by two or more subtypes) and unique (distinct to only one subtype). They propose a statistical model to identify recurrent alterations and compute the shared status of these alterations. Each shared status is a set of subtypes that contain this recurrent alteration.

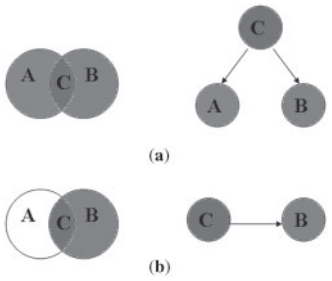

Fig. 2. Two examples of Venn diagram (left) of two sets and its corresponding graph model (right). The three sections in the Venn diagram are denoted as A, B and C, respectively. In (**a**) both the cancer types A and B have some common markers and each contains additional non-common markers. So they have inherited from an unobserved common subtype. In (**b**) the markers of A is a proper subset of that of B. So B is a derivative of type A.

The shared status of recurrent alterations can be described using a Venn diagram. For example, Figure 2 shows two Venn diagrams (left) of two sets, represented by two overlapping circles. Let S_1 and S_2 denote the left and right circle, respectively. There are three distinct areas (denoted as *sections*) marked by A, B and C in each Venn diagram. Each section represent a possible logical relationship between the two sets. For example, sections A and C represent $S_1 - S_2$ and $S_1 \cap S_2$, respectively. A section is called non-empty, if it contains some members. Each non-empty section is marked by a distinct color in Figure 2. The *component* of a non-empty section is defined to be the sets whose members are contained in this section. For example, the components of sections A and C are $\{S_1\}$ and $\{S_1, S_2\}$, respectively. In general, the number of distinct sections S in a Venn diagram of K sets can be as large as $S = 2^K - 1$, which is also the number of different shared status of a recurrent alteration between K cancer subtypes. Since each section can be empty or non-empty, there are totally 2^S distinct Venn diagrams for K sets.

The authors build a Venn diagram of four sets for the four different subtypes of NB. In the Venn diagram, each set corresponds to one of the four NB. The members of each set are the recurrent alterations that belong to that subtype of NB. The intersection of two main sets represents their shared recurrent alterations.

The authors proposed a graph model based on the structure of Venn diagram to infer the progression of four different subtypes of NB. The resulting graph is a directed acyclic graph with each vertex corresponding to a non-empty section in the Venn diagram. An edge connects from a vertex u to a vertex v if the recurrent alterations of u is a subset of that of v and there is no vertex w whose markers are a proper subset of that of v and a proper superset of that of u. The number of vertices in the resulting graph is bounded by $\min\{S, T\}$, where T is the number of recurrent alterations. For example, the graph models corresponding to cancer subtypes in Figure 2 is shown on the right of the figure. The authors demonstrate that, with the help of such a model, it is possible to identify tumor progression in CGH data. However, their approach has several limitations.

- First, their methods of calculating the shared status of each recurrent alteration is very computationally expensive. The time complexity is exponential to the number of cancers K.
- This method can model the progression of markers. It, however, cannot model the evolutionary relationship among different cancer types.

In addition to these limitations, Bilke *et al.* do not provide a systematic algorithm for mapping the Venn diagram to the graph model automatically. These limitations make it impractical to use their method for large-scale datasets composed of many cancers.

3 PROGRESSION MODEL FOR MARKERS

In this section, we extend Bilke's approach (Bilke *et al.*, 2005) to infer progression models for markers of multiple cancer types. Studies of the evolution of markers would be of obvious value to define gene loci relevant for the early diagnosis or treatment of cancer. It helps to answer questions about which markers tend to occur in many cancers, which markers are likely to occur together, etc. The main difference between our approach and the previous work is that we focus on markers instead of every recurrent alteration.

We compute the shared status of markers as follows. A marker identified in one cancer type represents a recurrent alteration region in this cancer type (Fig. 1). However, for any two or more cancers containing the same recurrent region, they may not have markers identified at the same position due to the noise in the aberration patterns. Therefore, markers in different cancers representing the same recurrent region should be considered shared by these cancers.

First, we define the *overlap coefficient* between a marker and its neighboring intervals. Let C denote a set of cases belonging to the same cancer. Let $m=<p,q>$ and $d, 1\leq d\leq D$ denote a marker in C and a genomic interval, respectively. For each case $s_j\in C$, we define $E_j(d,m)=1$ if there exists a segment $s_j[u,v]$ overlapping with both intervals d and p, i.e. $u\leq d, p\leq v$ and $x_u^j=q$, otherwise, $E_j(d,m)=0$. The function $E_j(d,m)$ indicates that the alterations at d and p belong to the same segment in s_j and can be caused by the same point-like genomic alteration. We compute the overlap coefficient between d and m as

$$\mathrm{OC}(d,m)=\frac{\sum_{j=1}^{|C|}E_j(d,m)}{\mathrm{Supt}(m)}$$

where $|C|$ denotes the size of C and $\mathrm{Supt}(m)$ denotes the support value of marker m in C. A large value of $\mathrm{Cor}(d,m)$ implies that intervals p and d belong to the same recurrent region that is represented by the marker m.

Next, we define that a marker $m=<p,q>$ in cancer C_i is shared by C_j if and only if the following condition is reached: there is a marker $m'=<p',q'>$ in cancer C_j such that $q'=q$ and $\mathrm{OC}(p,m')>\epsilon$, where ϵ is a user-defined threshold. The larger the value of ϵ, the harder that a marker is shared among multiple cancers due to noise in the data. Intuitively, this definition indicates that a marker m_i in C_i is shared by another cancer C_j if there exists a marker m_j in C_j such that m_j is highly overlapped with m_i. To compute the shared status of a marker in C_i, we visit every cancer other than C_i. This makes the time complexity linear in the number of cancers K. We denote the shared status $\mathcal{S}(m)$ of a marker m as the set of cancers that share this marker, i.e. $\mathcal{S}(m)\in\mathcal{P}(\{C_1,\ldots,C_K\})$, where $\mathcal{P}$ denotes the power set operation.

We propose an algorithm that generates a progression model for K cancers based on markers. The progression model generated by our algorithm is a directed acyclic graph. Each node of this graph corresponds to a non-empty set of markers. The set of markers corresponding to different nodes of this graph do not intersect. Our algorithm consists of three steps:

- *First step:* we identify an optimal set of R markers for each cancer using our marker identification program (Liu *et al.*, 2007). These markers represent significant recurrent alterations specific to each cancer.
- *Second step:* for each marker in each cancer, we compute the shared status of this marker using the method we described above. If some markers in multiple disease entities are identical (both position and type), we think them as a single marker and compute its shared status once.
- *Third step:* the logical relationship between K cancers corresponds to a Venn diagram of K sets. There are totally $S=2^K-1$ distinct sections in this Venn diagram. Given a marker m with shared status $\mathcal{S}(m)$, the section corresponding to $\mathcal{S}(m)$ is non-empty. We mark all the non-empty sections in the Venn diagram based on the shared status of all markers. We then convert the Venn diagram to a graph model as follows. We create a vertex V for each non-empty section and associate it with the markers whose shared status corresponds to this section. We define the height of this vertex, denoted as $H(V)$, as the number of components in the corresponding section. We visit the vertices in the descending order of their heights. For each pair of vertices V_i and $V_j, H(V_i)<H(V_j)$, we create an edge from V_i to V_j if both of the following conditions hold:
  (1) The component set of the section corresponding to V_i is a true subset of that of V_j.
  (2) There is no other vertex V_k such that the component set of the section corresponding to V_k is a superset of that of V_i and a subset of that of V_j.

We analyze the time complexity of this algorithm as follows. The time complexity of the first step is $O(DNR)$ as analyzed in our previous work (Liu *et al.*, 2007), where D and N denote the number of genomic intervals and number of cases of all K cancers, respectively. The time complexity of the second step is $O(TNR)$, where T is the cardinality of set consisting of the union of all markers. In the third step, the number of vertices is bounded by $\min\{S,T\}$. Since $T\leq K\times R$, the time complexity of this step is $O(K^2R^2)$ in the worst case. Since we have $D\geq T$, the overall time complexity is $O(DNR)+O(K^2R^2)$. In general, we have $D\gg R, N\gg K^2$, the overall time complexity can be written as $O(DNR)$.

The graph created by our algorithm can be used to describe the hierarchical or evolutionary relationship between markers representing multiple stages between a single cancer type or among the markers of different cancer types. We term a node as a *root node* if it does not have any incoming edges. The nodes that are close to a root (there can be multiple roots) denote the aberrations that started in earlier stages. From this perspective, markers are not equally important. The markers that are parents of other markers in the hierarchical representation are common to multiple cancers. Thus, difference at parent marker positions should contribute more to the distance between different cancers than the child markers.

4 PROGRESSION MODEL FOR CANCERS

The aberration pattern defines the molecular characteristics of a cancer. We assume that cancers with similar aberration patterns

are close to each other in the evolutionary history. The proper identification of the similarities between cancers will expose the underlying mechanism of cancer development and benefit the diagnosis and treatment of cancers.

Phylogenetic tree is a simple and efficient model that infers evolutionary relationship among three or more cancers. A key challenge that needs to be addressed to employ existing distance matrix methods for tree construction is to find a biologically meaningful distance function between cancers. Next, we propose a novel measure for computing the distance between cancers based on their aberration patterns. Since markers are a set of recurrent alterations that define the aberration patterns of a cancer, our distance measure computes the distance between cancers based on their markers. Formally, let C_i and C_j denote two cancers. Let $M_i = \{m_{i,1}, \ldots, m_{i,R}\}$ and $M_j = \{m_{j,1}, \ldots, m_{j,R}\}$ denote the corresponding R markers identified in C_i and C_j, respectively, where $p_{i,1} < p_{i,2} < \cdots < p_{i,R}$ and $p_{j,1} < p_{j,2} < \cdots < p_{j,R}$. Please note that $p_{i,k}$ may not equal to $p_{j,k}$ for any $1 \leq k \leq R$. To compute the distance between C_i and C_j, we first align the markers in M_i to those in M_j. The goal of this alignment is to map M_i and M_j into two high-dimensional vectors $\hat{M}_i$ and $\hat{M}_j \in \mathbf{R}^g$, where $g \leq 2R$ is the number of dimensions of the new vectors, such that the new vectors contain consensus information about pattern of alterations in C_i and C_j.

We say that a pair of markers $m_{i,k}$ and $m_{j,r}$ are *overlapping* if they satisfy either one of the following two conditions:

(1) Both markers appear at the same interval and have the same type, i.e. $p_{i,k} = p_{j,r}$ and $q_{i,k} = q_{j,r}$

(2) Both markers represent the same region of recurrent alterations, i.e. $\mathrm{OC}(p_{i,k}, m_{j,r}) > \epsilon$ and $\mathrm{OC}(p_{j,r}, m_{i,k}) > \epsilon$, where ϵ is a user-defined threshold for overlap constraint.

In Section 3, we argue that markers are not equally important in the progression of cancers. A marker that is common to many cancers usually represents a fundamental characteristic of cancers. Therefore, we assume that markers shared by many cancers are more important than those shared by a few cancers. The intuition behind this reasoning can be explained as follows. A marker that triggers most of the cancers has survived the evolution of cancer progression with high likelihood. The markers that are cancer specific have most likely appeared later in the evolutionary history and created the underlying cancer alteration pattern. As a result, the deviation in genomic alterations corresponding to older markers corresponds to larger distance between two cancer types as the age of the genomic alteration increases. We incorporate this idea into the mapping process. We assign weights to markers in each cancer. The weight of a marker is the number of cancers that share this marker. Let $W_i = \{w_{i,1}, \ldots, w_{i,R}\}$ and $W_j = \{w_{j,1}, \ldots, w_{j,R}\}$ be the vectors of weights for markers in M_i and M_j. Here, $w_{i,k}$ and $w_{j,k}$ denote the weights the k-th marker in M_i and M_j.

The mapping process works as follows. Each time we pick up a pair of markers from M_i and M_j. We add a pair of new dimensions to $\hat{M}_i$ and $\hat{M}_j$, respectively. The values of the added dimensions are determined by three attributes of markers: support, weight and type. Let $\Delta(m_{i,k}) = \mathrm{Supt}(m_{i,k}) \times w_{i,k} \times q_{i,k}$. If the two markers are overlapping, the values added into $\hat{M}_i$ and $\hat{M}_j$ are $\Delta(m_{i,k})$ and $\Delta_(m_{j,r})$, respectively. If two markers are not overlapping, we focus on the marker at a smaller genomic interval. Without loss of generality, we can assume $p_{i,k} < p_{j,r}$. There is no marker at interval $p_{i,k}$ in C_j. However, we need to compute the information of this interval across both cancers so that the difference of this interval can be taken into account. So we assume that there is a 'hypothetical' marker at $p_{i,k}$ in C_j. This marker is of the same type and weight as $m_{i,k}$. However, the support of this marker is computed based on the samples in C_j. Let $m' = <p', q'>$ in C_j denote this *hypothetical* marker. We have $p' = p_{i,k}, q' = q_{i,k}$ and $w' = w_{i,k}$. Please note that $\mathrm{Supt}(m')$ depends on the alteration pattern in C_j and may not equal to $\mathrm{Supt}(m_{i,k})$. We add the two values, $\Delta(m_{i,k})$ and $\Delta(m')$, into $\hat{M}_i$ and $\hat{M}_j$, respectively. Next, we choose another pair of markers and repeat the above procedure until all the markers have been processed.

The algorithm of the mapping process of two sets of markers is implemented as follows.

Inputs: $M_i = \{m_{i,1}, \ldots, m_{i,R}\}$ and $M_j = \{m_{j,1}, \ldots, m_{j,R}\}$ where $p_{i,1} < p_{i,2} < \cdots < p_{i,R}$ and $p_{j,1} < p_{j,2} < \cdots < p_{j,R}$. $W_i = \{w_{i,1}, \ldots, w_{i,R}\}$ and $W_j = \{w_{j,1}, \ldots, w_{j,R}\}$ are the vectors of weights for markers in M_i and M_j

1. **Initialize:** $\hat{M}_i = \hat{M}_j = []$; $k = r = 1$;
2. **while** $k \leq R$ and $r \leq R$
   (a) **if** $m_{i,k}$ and $m_{j,r}$ are overlapping
   $\hat{M}_i = [\hat{M}_i, \Delta(m_{i,k})]$; $\hat{M}_j = [\hat{M}_j, \Delta(m_{j,r})]$; $k = k+1$; $r = r+1$;
   (b) **else if** $p_{i,k} < p_{j,r}$
   Create a hypothetical marker m' same as $m_{i,k}$ in C_j; $\hat{M}_i = [\hat{M}_i, \Delta(m_{i,k})]$; $\hat{M}_j = [\hat{M}_j, \Delta(m')]$; $k = k+1$
   (c) **else if** $p_{i,k} > p_{j,r}$
   Create a hypothetical marker m' same as $m_{j,r}$ in C_i; $\hat{M}_i = [\hat{M}_i, \Delta(m')]$; $\hat{M}_j = [\hat{M}_j, \Delta(m_{j,r})]$; $r = r+1$
   (d) **else**
   $\hat{M}_i = [\hat{M}_i, \Delta(m_{i,k})]$; $\hat{M}_j = [\hat{M}_j, \Delta(m_{j,r})]$; $k = k+1$; $r = r+1$
3. **while** $k \leq R$
   Create a hypothetical marker m' same as $m_{i,k}$ in C_j; $\hat{M}_i = [\hat{M}_i, \Delta(m_{i,k})]$; $\hat{M}_j = [\hat{M}_j, \Delta(m')]$; $k = k+1$
4. **while** $r \leq R$
   Create a hypothetical marker m' same as $m_{j,r}$ in C_i; $\hat{M}_i = [\hat{M}_i, \Delta(m')]$; $\hat{M}_j = [\hat{M}_j, \Delta(m_{j,r})]$; $r = r+1$

Outputs: $\hat{M}_i, \hat{M}_j$

Once we have the aligned vectors $\hat{M}_i$ and $\hat{M}_j$, we use Extended Jaccard coefficient (Tan *et al.*, 2005) to compute the similarity between the two vectors. Extended Jaccard coefficient is widely used as a similarity measure in vector spaces. It retains the sparsity property of the cosine similarity while allowing discrimination of collinear vectors. For example, given two vectors $\hat{M}_i = [0.1, 0.3]$ and $\hat{M}_j = [0.2, 0.6]$, the cosine similarity does not discriminate the difference between them and the similarity value is computed as 1. However, in our case, $\hat{M}_i$ and $\hat{M}_j$ are different because they denote recurrent alterations in C_i and C_j with different frequencies. The Extended Jaccard coefficient is computed as follows.

$$\mathrm{EJ}(\hat{M}_i, \hat{M}_j) = \frac{\hat{M}_i \cdot \hat{M}_j}{\|\hat{M}_i\|^2 + \|\hat{M}_j\|^2 - \hat{M}_i \cdot \hat{M}_j}$$

The Extended Jaccard coefficient of any two vectors takes value within the range of [0, 1]. It is easy to convert Extended Jaccard coefficient to distance by subtracting it from one, i.e. $D(C_i, C_j) = 1 - \mathrm{EJ}(\hat{M}_i, \hat{M}_j)$. We compute the distance $D(C_i, C_j)$ for any $1 \leq i, j \leq R$. As a result, we construct the distance matrix for K cancers. We apply existing distance matrix method, such as unweighted pair group method with arithmetic mean (UPGMA), to construct the phylogenetic tree.

5 EXPERIMENTAL RESULTS

Dataset: with 15 127 cases from 571 publications as of December 2007, Progenetix is the largest resource for published chromosomal CGH data (Baudis and Cleary, 2001) (http://www.progenetix.net/). For the purpose of this article, we use a dataset with 5918 clearly malignant epithelial neoplasias (ICD-O-3 xxxx/2 and xxxx/3), a descriptive overview of which had been published previously (Baudis, 2007). From the biomedical perspective, this dataset could be divided into 22 clinico-pathological disease categories. Additional entities consisting of <40 cases each were summarily moved to an 'other' category.

As result of the Progenetix database format transformation, for each case the genomic imbalance status for 862 ordered intervals had been extracted from the karyotype annotation. This information represents the whole genome copy number status information, in the maximum resolution feasible for cytogenetic methods. The value of each interval is 1, −1 or 0, indicating the gain, loss and no change status, respectively. We have used the cases corresponding to 20 different carincomas in this dataset. The number of cases for these carincomas vary from 42 to 640. The details of the dataset is shown in Table 1 of the Supplementary Material. For simplicity, we use the following abbreviations for some of these carincomas. CRC: colorectal adenocarcinoma; HCC: hepatocellular adenocarcinoma; HNSCC: head-neck squamous cell carcinoma; NSCLC: non-small cell lung carcinoma; ES: esophagus carcinoma; RCC: renal carcinoma; NPC: nasopharynx carcinoma; NE: neuroendocrine carincoma and carcinoid; MEL: melanocytic; PAC: pancreas adenocarcinoma; and SCLC: small cell lung carcinoma.

System specifications: we developed our code using MATLAB and ran our experiment on a system with dual 2.59 GHz AMD Opteron Processors, 8 GB of RAM and a Linux operating system.

5.1 Results for marker models

In this experiment, we infer a progression model for markers. We perform each step one by one and discuss the results of each step as follows. In the first step, we identify an optimal set of 20 markers for each cancer. We use the marker selection algorithm from Liu *et al.* (2007). The number of markers is decided based on observation by Baudis (2007) that most of the cancer subtypes can be effectively represented by around 20 markers. Please note that we exclude 100 (peri) centromeric intervals because (i) they mostly consist of repetitive sequence (ALU repeats, etc.) without encoding genes; and (ii) they have technical or interpretation difficulties. The markers are identified from the remaining 762 intervals.

An existing work by Baudis (2007) has identified the imbalance hot spots in clinico-pathological entities in the same dataset, using an 'average profile'-based approach. We compared our markers with the reported imbalance hot spots for validation test. Due to the limitation of space, here we only present the comparison results for HNSCC disease category.

- Imbalance hot spots identified by Baudis (2007):
  gains: 3q26 (59.2%), 8q24 (40.8%), 11q13 (31.9%, many specific high-level), 5p (26.5%), Xq, 1q, 7q(21), 12p, 17.
  losses: 3p (30.1%), 18q(22) (22.4%), 9p (22.4%), 11q24 (19.2%), 4, 5q, 8p, 13.
- Markers identified by our method:
  gains: 3q26.2 (57.2%), 8q24.3 (41%), 11q13.4 (31.9%), 5p14.3 (26.5%), Xq28 (23%), 7q21.3 (20.9%), 12p13.1 (17.7%), 17q25.3 (17.7%), 20q12 (17.7%), 19p13.11 (16.8%), 1q31.3 (16.2%), 18p11.23 (15.9%).
  losses: 3p26.3 (30.7%), 18q23 (22.7%), 9p23 (22.4%), 11q25 (19.2%), 4p14 (18%), 5q21.3 (15.3%), 8p23.3 (16.2%), 13q21.33 (16.5%) .

In the above results, markers or hot spots are listed with detailed locus and frequency information. Gains and losses are evaluated separately. The hot spots or markers are sorted in descending frequency of occurrence. We identify markers as individual intervals, while Baudis identified the regional hot spots from summary data. Our results are highly compatible to reported results if we consider a marker as a representative of a region. We successfully identify all the hot spots identified by Baudis. We also identify additional hot spots (e.g. 18q23) that has significant support.

In the second step, for each disease entities, we compute the shared status of each marker identified in this cancer using the method we described in Section 3. We set the threshold ϵ to 0.8. We tried with different values for ϵ and used 0.8 as it was giving best results. However, we believe that our method is not too sensitive in terms of those parameters, if we select those parameters from the near neighborhoods of the given values. To compare with the reported most frequent imbalances over all cancers, we analyze the markers that are in the same regions. The comparisons of imbalance with top frequencies are shown as follows.

- Most frequent imbalances reported by Baudis (2007):
  +8q: ubiquitously high (exception NE and thyroid)
- Markers identified by our method and their shared status:
  +8q23.1, +8q23.2, +8q23.3: 19 cancers (exception thyroid)
  +8q24.13, +8q24.23, +8q24.3: 18 cancers (exception NE and thyroid)
- Most frequent imbalances reported by Baudis (2007):
  −13q: occurring in most carcinoma types (exception cholangio and SQS)
- Markers identified by our method and their shared status:
  −13q21.1, −13q21.2, −13q21.33: 18 cancers (exception CRC, gastric, cholangio and SQS)
  −13q22.3: 15 cancers (exception SCLC, CRC, prostate, thyroid, gastric, cholangio and SQS)

The results show that our approach discovers the most frequent markers in a consistent way to Baudis' work. Please note that markers are individual intervals instead of chromosomal regions. Additionally to the markers reported by Baudis *et al*. as top-scorers in the different entities, our method detected other regions, for example +17q and +7p which both are shared by more than 12 cancers types.

Table 1. Quality of the phylogenetic trees according to three different criteria: NMI, entropy and parsimony. Large NMI value and small entropy and parsimony values are desirable. The numbers given in bold show the best result obtained among all trees in each category.

Tree construction method	Markers	Quality of the trees		
		NMI	Entropy	Parsimony
Fitch–Margoliash	Weighted	0.68	**0.69**	**8**
	Unweighted	0.62	0.82	9
Neighbor joining	Weighted	0.67	0.81	9
	Unweighted	**0.69**	0.74	9
UPGMA	Weighted	0.67	0.80	10
	Unweighted	0.60	0.89	10

In the third step, we build a graph model based on the shared status of markers. The model contains 119 vertices and 385 edges, which makes it hard to display in this article format. However, we have uploaded the graph in a tabular format in a separate file as a Supplementary Material. In that file, each entry corresponds to a vertex. Each vertex is a set of markers that are shared by some cancer subtypes. The model conveys useful information about the importance of markers. We use this information in our next experiments in Section 5.2.

5.2 Results for phylogenetic models

In this experiment, we infer progression models for cancer subtypes using the distance-based approach described in Section 4. We compute the distance matrix of 20 cancers listed in Table 1 of the Supplementary Materials based on the markers reported in Section 5.1. We test three different tree construction method, namely Fitch–Margoliash, neighbor joining and UPGMA in PHYLIP package (Felsenstein, 1989).

5.2.1 Quantitative evaluation Our first experiment measures the effect of computing the distance between cancers based on the importance of markers on phylogenetic tree construction. To do this, we construct phylogenetic tree with and without assigning weights to the markers. We then label each cancer type with the histology group it belongs to. We quantitatively evaluate the goodness of each tree using three different measures:

- NMI: this metric measures how well a given set of clusters separate labeled data. It takes values in [0, 1] interval, where 1 shows perfect separation. We measure the NMI at each internal node by considering the nodes in its left and right subtree as two clusters. We report the average NMI of all internal nodes.
- ENTROPY: Shannon's entropy of an internal node of a tree measures the uniformity of the labels of all the nodes of the subtree rooted at it. It takes values in [0, 1] interval, where 0 shows that all the nodes have the same label.
- PARSIMONY: this value shows the minimum number of unit mutations needed explain a given phylogenetic tree. The unit mutation in this tree changes one histology label to another.

Table 1 shows the average quality of all the trees we tested. Several observations follow from these results. First, Fitch–Margoliash produces the best tree in terms of entropy and parsimony. It is the second to the neighbor joining in terms of NMI, but the difference is little between the two. Particularly, we observe significant improvement in terms of the parsimony measure. The definition of parsimony measure implies that the probability of having a tree is exponential in the number of mutations needed for that tree. Assume that, on the average, the probability of mutating the genes to transform the cancer in one histology to another is p. Then the likelihood of the UPGMA tree is p^2 times that of the Fitch–Margoliash tree as UPGMA requires two more mutations. Our final important observation from this table is that weighing the markers often improves the quality of the trees. The exception was the neighbor joining algorithm, where the quality drop was not big.

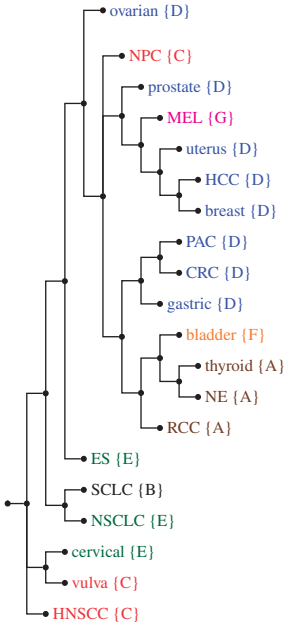

Fig. 3. Phylogenetic tree of 20 cancer entities. The labels/colors indicate the following histologies. A: endocrine and clear, B: small cell neuroendo, C: squamous, D: adenocarcinomas, E: mixed squamous/adeno, F: transitional, and G: melanoma.

5.2.2 Qualitative evaluation From Table 1, we conclude that the Fitch–Margoliash method with weighted markers produces the best tree. We thus use this method in the rest of this section.

Figure 3 shows the phylogenetic tree constructed on all cases of all cancer types. The leaf nodes of the trees correspond to cancers (e.g. clinico-pathological cancer entities). We mark these cancers using different colors as well as capitalized letters based on the histological composition of majority of cases in this cancer. Each color corresponds to a capitalized letter. Different colors (letters) encode different histological compositions of cancers. The internal nodes represent hypothetical cancers. Since these intermediate cancers may contain daughter branches from completely different histological cancer, they have to be viewed as common biological feature sets rather than truly occurring clinico-pathological cancer entities.

The phylogenetic tree in Figure 3 organizes cancer types with same histological composition closely in the same subtree for many of the cancer types. This correlation is in concordance with the

view that cancer clones may arise from tissue-specific cancer stem cells (Reya *et al.*, 2001), with a similar regulatory program targeted by genomic aberrations in related tissues.

To strengthen the claim that cancers that are in proximity in the phylogenetic tree are closely related we refer to available literature. First, we focus on cancers with same histologies. According to Lee *et al.* (2005), PIK3CA gene, which is an oncogene, is frequently mutated in breast carcinomas and hepatocellular carcinomas. Katoh *et al.* (1996) suggest a similarity of gastric and colorectal adenocarcinoma in terms of GSTM1 and GSTT1 genetic polymorphism. An obvious question is whether similar evidences exist for the cancer types that belong to different histologies, but are located closely in the phylogenetic tree. Indeed, Kurzrock *et al.* (1995) show that abnormalities in the PRAD1 (CYCLIN D1/BCL-1) oncogene are frequent in cervical and vulvar squamous cell carcinoma cell lines.

5.2.3 Running time results We executed our code on a Linux machine with Intel Xeon 2.7 GHz processor and 5 GB RAM. The first step that uses Rsim for clustering ran for almost 6 h. The second step that generates the progression model of markers completed in 15 min. The next step that saves the distance matrix for all cancer entities required 30 min. The final step that generates the phylogeny model for the cancer types required a few seconds. The entire program completed in 7 h.

Due to the space limitations, we report further experimental results that analyze a subtree of the phylogenetic tree in Figure 3 in Supplementary Material.

6 CONCLUSIONS

We have developed an automatic method to infer a graph model for the markers of multiple cancers. We demonstrated the use of this model in determining the importance of markers in cancer evolution. We also developed a new method to measure the evolutionary distance between different cancers based on their markers. We used this measure to create an evolutionary tree for multiple cancers.

With the application of our modeling approach to a set of more than 4600 epithelial neoplasias (carcinomas) with genomic imbalances, we can draw some preliminary conclusions:

(1) Marker determination and marker-dependent subset generation are powerful tools for structuring large CGH datasets.

(2) Phylogenetic modeling of 58 cancer subtypes with unique genomic marker sets shows a high concordance between branch association and histological subtype

(3) Cancer subtypes with a high level of genomic instability have overall similar imbalance patterns, which may reflect their origin from earlier, less-determined progenitor cells and/or tissue-independent mechanisms responsible for high-order genomic instability.

The important oncogenomic result of our work is the description of a closer relation between some tumor subsets/entities, which is related to rough histopathological grouping (e.g. adenocarcinomas versus squamous cell). This goes beyond the single gene aberrations described before, and supports statements made by us based on frequency-based clustering (Baudis, 2007).

While our approach as described here used rough histological group classification as a reference, a refined dataset combined with different reference qualities (e.g. clinical parameters) should provide a significant contribution to the overall perception of genomic instability in cancer development.

ACKNOWLEDGMENTS

This work was supported partially by NSF under grants CCF-0829867, DBI-0606607 and IIS-0845439, and UF Research Initiatives grant (00072365).

Conflict of Interest: none declared.

REFERENCES

Baudis,M. (2007) Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer*, **7**, 226.

Baudis,M. and Cleary,M.L. (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, **17**, 1228–1229.

Bilke,S. *et al.* (2005) Inferring a tumor progression model for neuroblastoma from genomic data. *J. Clin. Oncol.*, **23**, 7322–7331.

Desper,R. *et al.* (2000) Distance-based reconstruction of tree models for oncogenesis. *J. Comput. Biol.*, **7**, 789–803.

Desper,R. *et al.* (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comput. Biol.*, **6**, 37–52.

Forozan,F. *et al.* (1997) Genome screening by comparative genomic hybridization. *Trends Genet.*, **13**, 405–409.

Felsenstein,J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

Gray,J.W. *et al.* (1994) Molecular cytogenetics of human breast cancer. *Cold Spring Harb. Symp. Quant. Biol.*, **59**, 645–652.

Hoglund,M. *et al.* (2005) Statistical behavior of complex cancer karyotypes. *Genes Chromosomes Cancer*, **42**, 327–341.

Hsu,L. *et al.* (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.

Jain,A.N. *et al.* (2001) Quantitative analysis of chromosomal CGH in human breast tumors associates copy number abnormalities with p53 status and patient survival. *Proc. Natl Acad. Sci. USA*, **98**, 7952—7957.

Joos,S. *et al.* (2002) Classical hodgkin lymphoma is characterized by recurrent copy number gains of the short arm of chromosome 2. *Blood*, **99**, 1381–1387.

Kallioniemi,A. *et al.* (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Katoh,T. *et al.* (1996) Glutathione S-transferase M1 (GSTM1) and T1 (GSTT1) genetic polymorphism and susceptibility to gastric and colorectal adenocarcinoma. *Carcinogenesis*, **17**, 1855–1859.

Kurzrock,R. *et al.* (1995) Abnormalities in the PRAD1 (CYCLIN D1/BCL-1) oncogene are frequent in cervical and vulvar squamous cell carcinoma cell lines. *Cancer*, **75**, 584–590.

Lee,J.W. *et al.* (2005) PIK3CA gene is frequently mutated in breast carcinomas and hepatocellular carcinomas. *Oncogene*, **24**, 1477–1480.

Liu,J. *et al.* (2007) Markers improve clustering of CGH data. *Bioinformatics*, **23**, 450–457.

Mattfeldt,T. *et al.* (2001) Cluster analysis of comparative genomic hybridization cgh data using self-organizing maps: application to prostate carcinomas. *Anal. Cell. Pathol.*, **23**, 29–37.

Mitelman,F. (ed.) (1995) *International System for Cytogenetic Nomenclature*. Karger, Basel.

Pennington,G. *et al.* (2006) Cancer phylogenetics from single-cell assays. *Technical Report CMU-CS-06-103*, School of Computer Science, Carnegie Mellon University, Pittsburgh.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37** (Suppl.), S11–S17.

Reya,T. *et al.* (2001) Stem cells, cancer, and cancer stem cells. *Nature*, **414**, 105–111.

Tan,P.-N. *et al.* (2005) *Introduction to Data Mining*, 1st edn. Addison Wesley, Boston, MA.

Vandesompele,J. *et al.* (2005) Unequivocal delineation of clinicogenetic subgroups and development of a new model for improved outcome prediction in neuroblastoma. *J. Clin. Oncol.*, **23**, 2280–2299.

Vogelstein,B. *et al.* (1988) Genetic alterations during colorectal-tumor development. *N. Engl. J. Med.*, **319**, 525–532.