EXPECTATION-MAXIMIZATION METHOD FOR RECONSTRUCTING TUMOR PHYLOGENIES FROM SINGLE-CELL DATA

G. Pennington
Computer Science Department, Carnegie Mellon University
Pittsburgh, PA 15213, USA

C. A. Smith and S. Shackney
Allegheny Singer Research Institute, Allegheny General Hospital
Pittsburgh, PA 15212, USA

R. Schwartz*
Department of Biological Sciences, Carnegie Mellon University
Pittsburgh, PA 15213 USA
Email: russells@andrew.cmu.edu

Recent studies of gene expression in cancerous tumors have revealed that cancers presenting indistinguishable symptoms in the clinic can represent substantially different entities at the molecular level. The ability to distinguish between these different cancers makes possible more accurate prognoses and more finely targeted therapeutics. Making full use of this knowledge, however, requires characterizing commonly occurring cancer sub-types and the specific molecular abnormalities that produce them. Computational approaches to this problem to date have been hindered by the fact that tumors are highly heterogeneous masses typically containing cells at multiple stages of progression from healthy to aggressively malignant. We present a computational approach for taking advantage of tumor heterogeneity when characterizing tumor progression pathways by inferring those pathways from single-cell assays. Our approach uses phylogenetic algorithms to infer likely evolutionary sequences producing cell populations in single tumors, which are in turn used to create a profile of commonly used pathways across the patient population. This approach is combined with expectation maximization to infer unknown parameters used in the phylogeny construction. We demonstrate the approach on a set of fluorescent in situ hybridization (FISH) data measuring cell-by-cell gene and chromosome copy numbers in a large sample of breast cancers. The results validate the proposed computational methods by showing consistency with several previous findings on these cancers. They also provide novel insights into the mechanisms of tumor progression in these patients.

1. INTRODUCTION

Computational studies have led to substantial revisions in thinking about how to treat and diagnose cancers. Although all cancers are characterized by a general pattern of uncontrolled cell growth, it has long been recognized that they represent many different diseases at the molecular level. Numerous different combinations of genetic abnormalities could potentially disrupt the controls on cell growth and produce essentially the same gross phenotypes. Classic chemotherapies for treating cancers thus typically target the phenotype of frequent cell division rather than any specific genetic state distinguishing cancerous from healthy cells, leading to treatments that are broadly but not consistently effective and that carry serious side-effects. The application of computational clustering methods to gene expression microarrays[5] has recently shown that most tumors can be grouped into one of a few common "cancer sub-types,"[6, 11, 14] each characterized by similar molecular abnormalities and potentially treatable by common "targeted therapeutics" addressing those specific abnormalities. Subtype identification has proven useful in predicting patient outcomes[22, 26, 25, 23] and in selecting appropriate treatment regimens.[3, 1] The most notable success of this new approach to targeted therapeutics is the drug trastuzumab (Herceptin), an antibody to the Her-2/neu gene that is specifically effective in a subset of breast cancers characterized by amplification of the Her-2/neu gene.[10]

The recognition of cancer sub-types was a signifi-

*Corresponding author

cant advance, but it is also a simplification. A cancer sub-type characterizes a general progression pathway or set of related pathways by which successively accumulating mutations transform once healthy cells into increasingly aggressive tumor cells. However, any given patient may have advanced to a greater or lesser degree along this pathway,[9, 19] and the degree of progression is itself a significant predictor of prognosis.[13] It is thus valuable to understand not just what changes distinguish advanced cancer cells on a particular pathway from those on a different pathway, but also the particular sequence of events by which those changes accumulate on any given pathway. Desper et al.[4] showed that it is possible to identify relationships among different tumors by constructing phylogenies, or evolutionary trees, using microarray gene expression data and a distance metric similar to those used in the prior clustering approaches. However, this approach oversimplifies in some ways because tumors are not homogeneous masses. As cells in a tissue progress along a given pathway through the accumulation of successive mutations, the earlier states do not die out, but rather leave remnant populations in the tumor. Figure 1 illustrates this process. The existence of multiple progression states within a single tumor can be expected to confound microarray-based approaches, which can only measure tissue-wide average expression levels. Cancer prognosis has indeed been shown to be affected by changes apparent in single cells, but not from such tissue-wide measurements.[20]

Our contributions: We present a new method that treats tumor heterogeneity as an asset rather than an obstacle to the inference of progression pathways by using single-cell measurements to infer progression pathways within and between patients. We develop an algorithm for inferring likely evolutionary trees across cells by combining phylogenetic methods with an expectation-maximization framework for learning model parameters. We then use trees inferred patient-by-patient to identify specific sequences of molecular changes that commonly underlie a particular tumor type. We apply our technique to a large set of single-cell fluorescent *in situ* hybridization (FISH) measurements from breast cancers in which copy numbers are assessed for the Her-2/neu oncogene, the p53 tumor suppressor gene, and chromosome 17, on which both genes are found.[7] The results validate our approach by recapitulating several previously observed features of the roles of these genes in breast cancers. They further provide new insights into nature of common progression pathways in these cancers with implications for the optimal diagnosis and treatment of cancer patients.

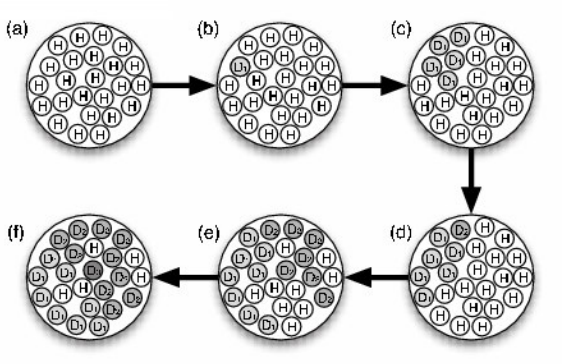

Fig. 1. Illustration of cancer progression resulting in tumor heterogeneity. **(a):** A healthy mass of cells labeled H. **(b):** A cell mutates into a diseased state D_1, which encourages proliferation and further progression. **(c):** The proliferating cell expands, leaving a heterogeneous population. **(d):** A D_1 cell reaches a further progression state D_2, increasing potential for proliferation. **(e):** Both populations continue to expand **(f):** The D_2 population becomes dominant, and an additional mutation results in a new disease state, D_3.

2. METHODS

Our method uses expectation maximization (EM) to learn several unknown parameters in a model of cell progression, applies an algorithm for the minimum cost arborescence problem to construct per-patient phylogenies consistent with the model, and then identifies commonly used pathways across patients. The remainder of this section defines the input data and phylogeny model and explains each step of the overall inference process. All algorithms described below were implemented using the functional programming language Objective Caml.

2.1. Input Data

Although our high-level approach is intended to apply to any form of cell-by-cell assay, we assume below an input format based on the FISH copy number data used in our validation experiments. These data count copy numbers of a single gene and a sin-

gle chromosome in individual cells. Each patient can thus be represented as an N by N two-dimensional array M, where N is some maximum observed count. For the present work, N is 10 and any counts above 10 are collectively grouped into a single row or column of M representing the count "greater than 10." Element m_{ij} of M is then the fraction of cells of a given sample that have i copies of the chromosome and j copies of the gene, which we call state (i, j). The FISH data is produced by manually counting fluorescent probes on labeled cell microscopy images, which can produce false counts if two probes are too close together or a single indistinct probe signal is incorrectly viewed as two distinct probes. We apply a preprocessing step to the input prior to our algorithm to reduce this noise. We assume that up to ten percent of cells from an observed state may have been misclassified and thus screen out from each patient's data any states for which the observed frequency is less than 10% of the sum of the frequencies of its neighbors.

2.2. Probability Model

We select phylogenies using a likelihood model that assumes cell states evolve from one another through four possible known molecular mechanisms for tumorigenesis: gene gain, gene loss, chromosome duplication, and chromosome loss. In gene gain, a cell gives rise to a new state with one extra copy of the gene. Gene loss produces a new state with one fewer gene copies. Chromosome duplication, modeling incomplete mitosis, doubles the complement of genes and chromosomes in a cell. Chromosome loss results in the loss of a single chromosome; as it is not clear how many gene copies might lie on the lost chromosome, we allow any number of gene copies to be lost simultaneously with the chromosome. Each of these operations is assumed to have some prior probability: p_{g+} for gene gain, p_{g-} for gene loss, p_{c+} for chromosome duplication, and p_{c-} for chromosome loss. We call the vector of these four prior probabilities θ. The prior probability of a full tree, $\Pr\{T|\theta\}$ is then defined to be the product of the prior probabilities of its edges.

Our model further defines the probability of the data given a tree, $\Pr\{M|T\}$, to be the product over all non-root nodes u of the frequency of u's parent node, where the root is always defined to be the (2,2) state. This model is meant to capture the intuition that a node is more likely to have descended from a well-populated state than from a sparsely populated state. Thus, we can define the full probability we seek to maximize, $\Pr\{M|T\}\Pr\{T|\theta\}$, to be $\prod_{e=(u,v)\in T} f_u p_e$, where f_u is the frequency of node u and p_e is the prior probability of the edge type of edge e given by θ. Figure 2 illustrates the definition, showing two possible trees for a given set of nodes and describing how the probability is derived for each tree. The goal of our computational methods is to find the θ maximizing $\Pr\{M|\theta\}$ over the full distribution of trees that might have produced M and, given this θ, find the tree T maximizing $\Pr\{M, T|\theta\} = \Pr\{M|T\}\Pr\{T|\theta\}$.

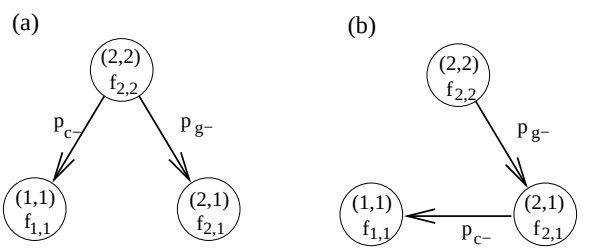

Fig. 2. Example illustrating the probability model for trees. Three cell states — (2,2), (2,1), and (1,1) with frequencies $f_{2,2}$, $f_{2,1}$, and $f_{1,1}$ — can be joined by two possible trees. (a) States (1,1) and (2,1) each descend directly from (2,2) by a chromosome loss and a gene loss respectively. The probability $\Pr\{M|T\}\Pr\{T|\theta\}$ has a contribution of $f_{2,2}p_{c-}$ from the chromosome loss and $f_{2,2}p_{g-}$ for the gene loss. (b) State (1,1) is descended from (2,1) by a chromosome loss and (2,1) from (2,2) by a gene loss. $\Pr\{M|T\}\Pr\{T|\theta\}$ has a contribution of $f_{2,2}p_{g-}$ from the gene loss as in (a) and a contribution of $f_{2,1}p_{c-}$ from the chromosome loss.

2.3. Optimal Tree Inference

Given the input M and a current set of parameters θ, we construct a directed graph $G = (V, E)$, where V is the set of observed states, and E is the set of all possible single mutation events that connect two states in V. If there exist states in G that are not reachable from the root, we add Steiner nodes to G using a heuristic method presented as Algorithm 2.1.

Once we have ensured that every node of G is reachable, we add a weight function $w(v, u) = f_v p_{(u,v)}$ to G and compute an optimal phylogenetic tree for the given patient using a classic algorithm for finding minimum weight arborescences (directed

minimum spanning trees) due to Chu and Liu.[2] Chu and Liu's algorithm is similar to Prim's greedy algorithm for undirected spanning trees[12] but with some additional complications to handle directed cycles. We specifically define the $(2, 2)$ state to be the root of the tree, which is part of the input to the arborescence algorithm. This algorithm is used after parameter inference to find the best-fit trees and is also used as a subroutine of the parameter inference method to initialize the Markov chain sampling for each patient on each EM round. A summary of our method for single-patient phylogeny inference is provided as Algorithm 2.2.

Algorithm 2.1 Heuristic algorithm for Steiner node inference

1: Given $G = (V, E)$. Let $G' = (V', E')$ be a directed graph containing all possible states and edges, and let $R \subseteq V$ be the set of vertices of G reachable from the root.
2: **while** $R \neq V$ **do**
3: Perform a breadth-first traversal of G' starting from R and stopping when we encounter a vertex $v \in V - R$. Let k be the distance from R to v (the length of the path found by BFS).
4: Consider all nodes in $V - R$ at distance k from R, and let v^* be the one from which we can reach the most nodes in $V - R$ (the largest island).
5: Solve the multiple source shortest path problem from R to v^* in G', where the weight of an edge $e \in E'$ is equal to $-\log p_e$ (minus the log of its probability).
6: Add the nodes and edges on the shortest path from V to v^* in G.
7: **end while**

2.4. Parameter Inference

We estimate the parameter set θ by EM. We treat the tree topology T as a set of latent variables corresponding to the presence or absence of each tree edge. In the expectation phase, we find the expectation of each of these latent variables by enumerating over possible trees T consistent with the output, weighted by the conditional probability $\Pr\{M|T\} \Pr\{T|\theta\} = \Pr\{M, T|\theta\}$. This expectation is evaluated by a Markov chain Monte Carlo method, in which states correspond to the possible trees and their stationary distributions are set to be proportional to $\Pr\{M|T\} \Pr\{T|\theta\}$. The frequency of occurrence of each possible tree edge in the Markov chain thus gives the expected value of the latent variable corresponding to that edge.

Algorithm 2.2 Procedure for tree inference from a matrix of cell counts S

1: Convert the FISH matrix for an individual patient into a graph G.
2: Add all edges to G allowed by the connectivity model of section 2.2, each weighted as minus the log of its probability.
3: Apply Algorithm 2.1 to add Steiner nodes until G is connected.
4: Find a minimum-cost arborescence on G by the method of Chu and Liu.[2]

In the maximization phase, these edge expectations are used to determine maximum-likelihood estimates of the parameters for the next EM round. This estimation is accomplished by counting the expected occurrences of each of the four edge types, summed over all potential tree edges of that type.

We initialize the method by assuming each of the four parameters is 0.25. We then construct an initial tree for each patient by running Algorithm 2.2 to provide a starting state for the Monte Carlo iteration. We then perform successive Monte Carlo steps as follows:

(1) Pick a node u from the tree uniformly at random from all nodes other than the root.
(2) For each possible parent v of u excluding current descendants of u, compute an edge weight $w(v, u) = f_v p_{(v,u)}$, where f_v is v's node frequency and $p_{(v,u)}$ is the prior probability of the edge type from v to u. Note that v might be u's current parent.
(3) Pick some v among all possible parents with probability $p_v = w(v, u) / \sum_x w(x, u)$.
(4) Delete the edge from u's current parent to u and replace it with an edge from v to u.

Repeatedly applying this move creates a Markov model we call H.

Note that this move set cannot produce a cycle

in the graph because, when selecting a new parent v of u, the move set specifically prohibits the selection of any v that is a current descendant of u. This guarantees v is not reachable by any directed path from u. The newly added edge (v, u) thus cannot create any directed cycle.

We can further show that H samples among all possible trees according to their relative probabilities as defined in our probability model. This result is established by the following theorem:

Theorem 2.1. *For any two trees S and T and any input data set M, the ratio of the stationary probabilities $\frac{\pi_S}{\pi_T}$ in M will be equal to $\frac{\Pr\{M|S\}\Pr\{S|\theta\}}{\Pr\{M|T\}\Pr\{T|\theta\}}$.*

Proof. Each non-root node has exactly one parent, so a tree T is completely defined by its list of parent assignments $p_T : V \to V$, where $p_T(v)$ is the parent of V in T. H is ergodic, since we can reach any tree T from any tree S by reassigning parents in S to match those in T in breadth-first order of the nodes in T. Any cycle from tree T returning to itself will contain some (possible empty) sequence of changes to the parent of V: $p_T(u), u_1, u_2, \ldots, u_k, p_T(u)$. These changes will contribute a factor of $w(u_1, v)w(u_2, v)\cdots w(u_k, v)w(p_T(v), v)/W^{k+1}$ for some W to the probability of traversing the cycle. The probability of traversing the cycle in the opposite direction is $w(u_k, v)w(u_{k-1}, v)\cdots w(u_1, v)w(p_T(v), v)/W^{k+1}$, i.e. the same value. Counting contributions for all $v \in V$, this establishes by the Kolmogorov criterion that H converges on a unique stationary distribution obeying detailed balance. It then suffices to show that for any two neighboring trees T_1 and T_2 that the ratio of their transition probabilities $\frac{p_{T_1\to T_2}}{p_{T_2\to T_1}}$ is equal to $\frac{\Pr\{M|T_2\}\Pr\{T_2|\theta\}}{\Pr\{M|T_1\}\Pr\{T_1|\theta\}}$. The fact that they are neighbors means that they differ by a single parent assignment, u_1 versus u_2, of a node v.

$$\begin{aligned}\frac{p_{T_1\to T_2}}{p_{T_2\to T_1}} &= \frac{w(u_2, v)/W}{w(u_1, v)/W}\\ &= \frac{w(u_2, v)}{w(u_1, v)}\\ &= \frac{\Pr\{M|T_2\}\Pr\{T_2|\theta\}}{\Pr\{M|T_1\}\Pr\{T_1|\theta\}}\end{aligned}$$

□

In order to establish that the Markov chain is adequately sampling states, we also need to show it is rapidly mixing. If we define $P_S(t, T)$ to be the probability of encountering tree T at step t from starting tree S then we can do this by showing there is some t_0 polynomial in n for which we have a small *variation distance* between $P_S(t_0, T)$ and the stationary distribution Π, where we follow Jerrum and Sinclair[8] in defining variation distance as $\Delta_t = \frac{1}{2}\sum_T |P_S(t_0, T) - \pi_T|$. We establish this by the following theorem:

Theorem 2.2. *The Markov chain H initialized with some state S reaches variation distance ϵ from Π in time $O(n\phi(\ln \epsilon^{-1} + \ln \pi_S^{-1}))$ where n is the number of distinct cell states and ϕ is the maximum ratio of any two probabilities from $\theta = \{p_{c-}, p_{c+}, p_{g-}, p_{g+}\}$.*

Proof. We can prove rapid mixing using the canonical path method,[21] in which we define a path $\gamma_{U,V}$ between any two states U and V. Space does not permit a detailed tutorial on the method, so we provide only the details specific to our problem here and refer the interested reader to Jerrum and Sinclair[8] for an excellent tutorial on the method. We establish a canonical path between any tree S and tree T in which we convert parents of nodes in S to their parents in T according to the breadth-first order of those nodes in T. Suppose we examine a step on the canonical path from S to T transitioning from some S^* to T^*, in which we change the parent of some node v from u_S to u_T. Then the other canonical paths using that transition will be those between any S' and T' for which S' and T' have the same parents as T for nodes before v in breadth-first order and the same parents as S for nodes after v, $p_{S'}(v) = p_S(v)$, and $p_{T'}(v) = p_T(v)$. The canonical path method depends on bounding a quantity called the *edge loading*, defined for a transition $e = (S^*, T^*)$ as $(\pi_{T^*}p_{T^*,S^*})^{-1}\sum_{S',T' s.t. e\ni\gamma_{S^*,T^*}} \pi_{S'}\pi_{T'}|\gamma_{S',T'}|$. $\sum_{S',T'} \pi_{S'}\pi_{T'} \leq \pi_T p_{S^*,T^*}$ and $|\gamma_{S',T'}| \leq n$, so edge loading for H is bounded by $(\pi_{T^*}p_{T^*,S^*})^{-1}(n\pi_{T^*}p_{S^*,T^*})$, which is itself bounded by $n\phi$. This establishes the mixing time bound of $n\phi(\ln \epsilon^{-1} + \ln \pi_S^{-1})$. □

To ensure adequate mixing, we apply the Monte Carlo move $100n^3$ times per patient counting edge types every $10n^2$ moves. The fraction of edges assigned to each type provides a maximum likelihood

estimate of that edge type's probability for the next EM round. We repeat the above steps until all parameters converge with an error of less than one percent. We perform two versions of this inference: a global inference, in which we establish the four edge type probabilities for the whole population by pooling edge counts across all patients on each EM round, and a per-patient inference, in which we establish distinct parameters for each patient by performing the complete EM algorithm on each patient individually.

2.5. Identifying a Global Consensus Network

A final stage of analysis is performed with the EM-inferred parameters to find a best-fit tree for each patient and a global consensus network for the entire population. We first fit a phylogeny to each patient using Algorithm 2.1. We then find a global consensus network by identifying all pathways used in at least some fraction t of all patients. Given the per-patient trees $T_1, \ldots, T_n$, we can identify consensus pathways by searching depth-first through each tree individually and then, for each node, counting how many other trees have the same node and exhibit the same pathway from that node to the root. Those pathways occurring in a t fraction of trees are added to the global consensus network. For the present study, t=5%. Note that this consensus network need not itself be a tree, since a node may be reachable by more than one common pathway in different individual trees.

3. RESULTS AND ANALYSIS

3.1. Data

We applied our phylogeny inference methods to two data sets collected for a previous study on human breast cancer progression[7] using a protocol for FISH-based analysis of gene and chromosome copy numbers.[17] One data set consists of Her-2/neu gene copy numbers and chromosome 17 copy numbers assayed in cells from 118 individuals, with an average of 63 cell assays per patient. The second consists of p53 and chromosome 17 copy numbers assayed in 113 individuals with an average of 68 assays per patient. These two data sets were chosen for the present study in part because of the importance of both genes in breast cancer progression. Her-2/neu amplification promotes cell proliferation and is associated with a class of breast cancers[15, 24]. p53 is a crucial tumor suppressor gene whose loss or inactivation is implicated in approximately half of all human cancers.[18] Furthermore, the fact that both genes occupy chromosome 17 provides some means for validation of the method, as inferred patterns of chromosome gain and loss should be the same in both datasets.

3.2. Global Consensus Trees

We first performed a single consensus inference, fitting one set of prior probabilities to each of the full data sets. Table 1 shows the inferred probabilities from the two data sets. Both show similar frequencies of chromosome duplication and loss, with slightly higher rates of loss than duplication. p53 and Her-2 show very different patterns of gene gain and loss, though. Her-2 shows a notable preference for gene gain over loss, consistent with the fact that Her-2 amplification characterizes a subset of breast cancers.[15, 24] p53, on the other hand, shows a slight excess of gene loss over gain, consistent with the fact that p53 is implicated in cancers through loss of function, rather than amplification.

Table 1. Global consensus probabilities inferred for chromosome duplication (p_{c+}), chromosome loss (p_{c-}), gene gain (p_{g+}), and gene loss (p_{g-}).

	p_{c+}	p_{c-}	p_{g+}	p_{g-}
Her-2/neu	0.268	0.282	0.319	0.131
p53	0.274	0.290	0.198	0.238

Figure 3(a) shows a consensus phylogenetic network for Her-2/neu and chromosome 17. Two dominant edges project from the root, one corresponding to chromosome duplication and the other to gene gain, with lesser amounts of chromosome and gene loss. The two dominant edges lead to two prominent pathways in the graph. One pathway exhibits successive gene gains without changes in chromosome copy number while the other shows a pattern of alternating chromosome duplication and loss. There is support in the literature for both of these pathways. A large fraction of breast cancers exhibit diploidy with substantial amplification of Her-2/neu,[10, 24] consis-

tent with the pure gene gain pathway. The alternating pattern of chromosome duplication and loss has also previously been predicted based on mathematical models and is supported by evidence from several classes of solid tumors.[16] We further note, however, that these two pathways are not rigidly separated, but rather exhibit some ability to interconvert. Gene gain or loss events occasionally branch off of the chromosome gain/loss pathway and chromosome abnormalities occasionally appear off of the gene amplification pathway. This observation is, to our knowledge, novel. Examination of individual phylogenies (data not shown) suggests that individual patients may follow one or the other of these two dominant pathways exclusively or may combine the two.

Figure 3(b) shows the consensus phylogenetic network inferred for p53. Like the Her-2/neu network, the p53 one shows one prominent pathway exhibiting alternating chromosome duplication and loss. It is to be expected that the same chromosome patterns would be observed, as p53 and Her-2/neu are found on the same chromosome, and this finding thus validates the data and the analysis methods. Gene gain and loss is much less prominent for p53 than it was for Her-2/neu, however. Some gene gain and loss does occur, but it is comparatively rare and does not produce any long chains of successive amplifications, as is seen with Her-2/neu. Patterns of p53 gain and loss are likely to be difficult to interpret directly from copy number data, as they may involve partial inactivation of the gene rather than total loss.[18] p53 is, however, a tumor suppressor, so we would not expect to see a prominent p53 amplification pathway in cancers.

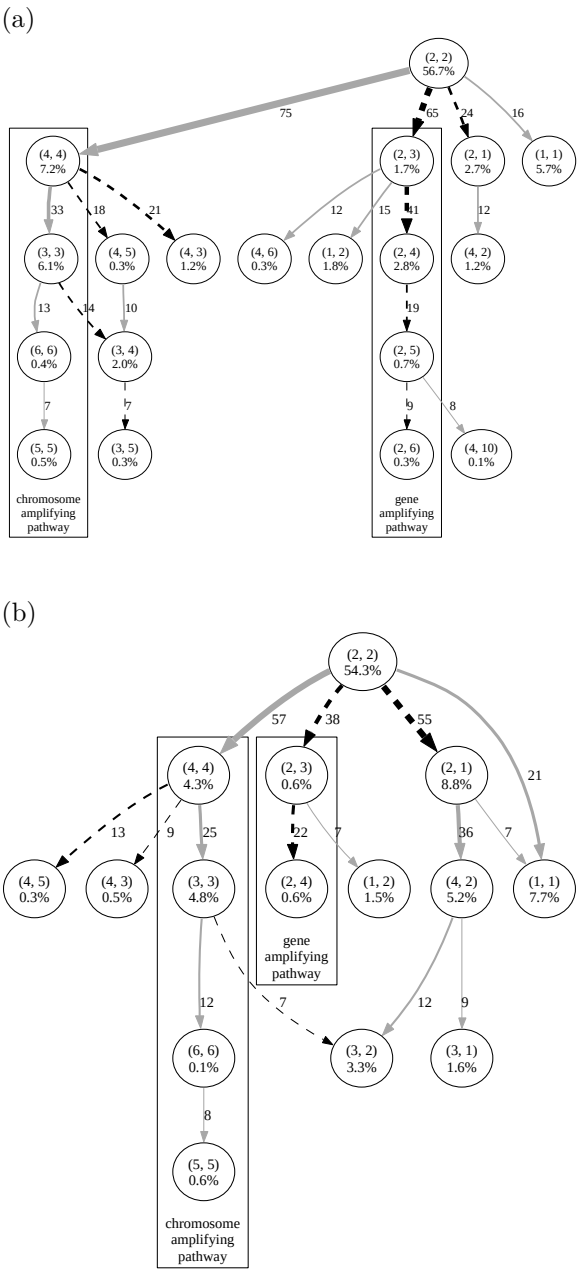

Fig. 3. Consensus networks inferred from pathways found in at least 5% of patients. Nodes represent cell states with chromosome and gene counts in parentheses and frequency of the state as a percentage of observed cells. Black dashed edges denote gene events and gray solid edges chromosome events. Edge label and thickness indicates the number of patients exhibiting the given edge. (a) Her-2/neu and chromosome 17 network. (b) p53 and chromosome 17 network.

3.3. Heterogeneity Between Patients

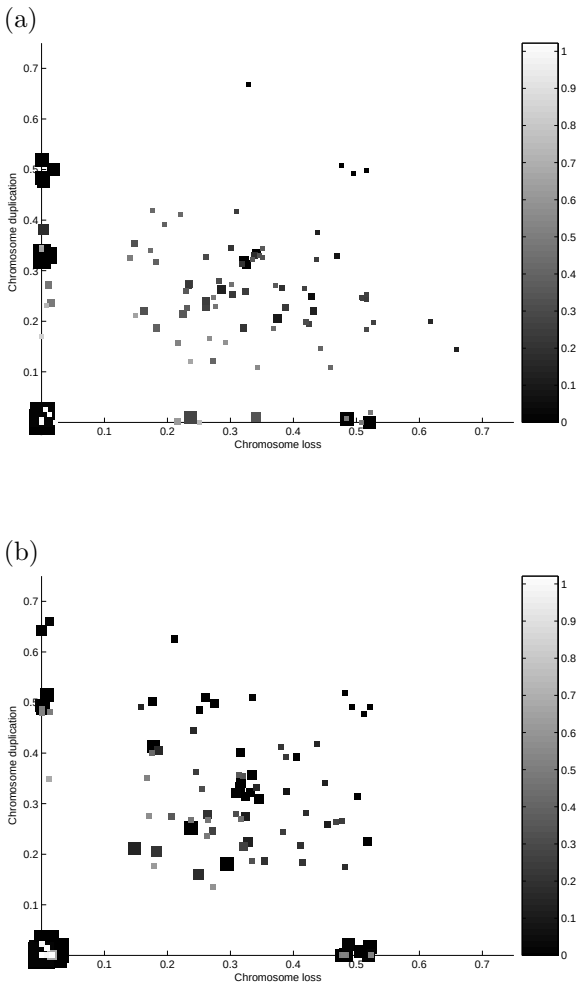

Fig. 4. Visual representation of the space of inferred prior probability parameters from per-patient data. Each image shows data points for the four inferred probability parameters on individual patients. p_{c-} (chromosome loss) and p_{c+} (chromosome duplication) determine as the x and y positions of the points. p_{g-} (gene loss) determines point size, with point size proportional to $1 + 10p_{g-}$. p_{g+} determines the color of the point, ranging from black for $p_{g+} = 0$ to white for $p_{g+} = 1$. Point positions are perturbed by a random factor up to 0.025 in x and y dimensions in order to make points with the same positions visible as distinct entities. (a) Parameters for Her-2/neu. (b) Parameters for p53.

While the global analysis gives us a reasonable best estimate of the overall frequencies of each of the possible genetic abnormalities, it is also useful to assess differences between patients. Figure 4 provides a graphical display of edge-type distributions derived by performing the EM inference one patient at a time instead of globally.

Figure 4(a) shows parameters for Her-2/neu and chromosome 17. A substantial fraction of points cluster on the axes and especially at the origin, corresponding to tumors that exhibit no chromosome loss, no chromosome duplication, or no loss and no duplication; these tumors cover a spectrum of gene gain and gene loss probabilities. Many points exhibit no gene loss (appearing as small squares in the figure) and these are scattered throughout the graph. A relatively small number of points exhibit almost exclusively chromosome events. A substantial fraction of all points lie in the middle of the plot, exhibiting some balance of all four event types. These observations are consistent with what was seen in the consensus phylogenies, suggesting that a large fraction of patients use both the gene and chromosome amplifying pathways, with other groups exhibiting exclusively one pattern or the other.

Figure 4(b) shows parameters for p53 and chromosome 17. The plot is superficially similar to that of Her-2/neu but with some notable differences. First, pure gene gain or gene loss in the absence of the other is comparatively rare for p53. Of those points on the axes or origin, though, a comparatively greater portion of them show up as having high gene loss and low gene gain (large black squares) as opposed to high gene gain and low gene loss (small white squares). This again appears consistent with the fact that p53 amplification is not associated with breast cancer, while Her-2/neu amplification is.

4. DISCUSSION

We have developed a novel computational method using phylogeny reconstruction algorithms to infer tumor progression pathways from cell-by-cell assays. The method allows us to produce likely progression trees for individual patients and to identify common progression pathways across distinct patients. Application to a set of FISH data on two known cancer-related genes gathered from breast cancer tumors validates the method by recapitulating several previously identified properties of these genes and their role in breast cancer. It further provides novel insights into the progression mechanisms acting in these tumors.

This work may have several important conse-

quences for cancer biology in general and in the specific types studied here. Her-2/neu amplifying tumors show two dominant pathways, chromosome amplifying and gene amplifying, which is consistent with prior knowledge. Our study also reveals, though, that these pathways can work in concert in individual patients. Approaches to cancer sub-type identification based on clustering of tissue-averaged measurements would not generally be able to recognize that these hybrid tumors are in fact using combinations of two fundamental pathways and may require therapeutics directed at both. This problem may be particularly significant for the classification of Her-2/neu tumors because current clinical standards for detecting Her-2/neu amplification and prescribing anti-Her-2/neu therapy use a protocol tuned for diploid cells and normalized by chromosome counts;[27] the protocol would be expected to be less sensitive to Her-2/neu amplification in aneuploid cells and thus potentially to fail to recommend anti-Her-2/neu therapy to patients whose tumors are genuinely Her-2/neu amplifying but are also aneuploid. We can anticipate that similar issues will arise with other tumor types as more targeted therapies become available. Accurate inference of progression pathways within tumors is thus likely to be a key step in developing a more rational approach to the targeted treatment of cancers.

There are several future directions to be explored in this work. One current limitation is that it looks at only a small number of measurements per cell simultaneously (one gene and one chromosome in the present work). The nature of the assay precludes much improvement in the experimental data, but computational inferences could in principle correlate states across different sets of copy data. For example, one might infer which Her-2/chromosome 17 states and which p53/chromosome 17 states overlap to produce likely Her-2/p53/chromosome 17 phylogenies. There are also other kinds of single-cell cytometry data to which this method could be applied, such as single-cell protein expression data. Finally, there are many avenues for advancement in developing more realistic models of the tumorigenesis process and more sophisticated phylogeny algorithms for the core inference and sampling steps, for example to deal more robustly with the inference of Steiner nodes.

Acknowledgments

R.S. and G.P. were supported by a grant from the Berkman Faculty Development Fund at Carnegie Mellon University.

References

1. Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, Hess K, Lecocke M, Metivier J, Booser D, Ibrahim N, Valero V, Royce M, Arun B, Whitman G, Ross J, Sniege N, Hotoagyi GN, Pusztai L. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. *J Clin Oncol* 2004; **22(12)**: 1–10.
2. Chu Y, Liu T. On the shortest arborescence of a directed graph. *Sci Sinica* 1965; **14**: 1386–1400.
3. Cunliffe HE, Ringnér M, Bilke S, Walker RL, Cheung JM, Chen Y, and Meltzer PS. The gene expression response of breast cancer to growth regulators: patterns and correlation with tumor expression profiles. *Cancer Res* 2003; **63**: 7158–7166.
4. Desper R, Khan J, Schaffer AA. Tumor classification using phylogenetic methods on expression data. *J Theor Biol* 2004; **228**: 477–496.
5. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; **95**: 14863–14868.
6. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeej M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**: 531–537.
7. Janocko LE, Brown KA, Smith CA, Gu LP, Pollice AA, Singh SG, Julian T, Wolmark N, Sweeney L, Silverman JF, ShackneySE. Distinctive patterns of Her-2/neu, c-myc, and cyclin D1 gene amplification by fluorescence in situ hybridization in primary human breast cancers. *Cytometry* 2001; **46(3)**: 136–149.
8. Jerrum M, Sinclair A. The Markov chain Monte Carlo method: An approach to approximate counting and integration. In Hochbaum DS (ed.), *Approximation Algorithms for NP-Hard Problems* PWS Publishing, Boston. 1996: 482–520.
9. Nowell PC. The clonal evolution of tumor cell populations. *Science* 1976; **194**: 23–28.
10. Pegram MD, Konecny G, and Slamon DJ. The molecular and cellular biology of HER2/neu gene amplification/overexpression and the clinical development of herceptin (trastuzumab) therapy for breast cancer. *Cancer Treat Res* 2000; **103:** 57–75.
11. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale A-L, Brown PO, and Botstein D. Molecular portraits of human breast tumors. *Nature* 2000; **406:** 747–752.
12. Prim RC. Shortest connection networks and some gener-

alizations. *Bell System Technical Journal* 1957; **36**: 1389–1401.

13. Ried T, Heselmeyer-Haddad K, Blegen H, Schrock E, and Auer G. Genomic changes defining the genesis, progression, and malignancy potential of solid human tumors: a phenotype/genotype correlation. *Genes Chromosomes Cancer* 1999; **25(3)**: 195–204.
14. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, van de Rijn M, Waltham M, Pergamenschikov A, Lee JCF, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, and Brown PO. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 2000; **24**: 227–235.
15. Salmon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 1987; **235**: 177–182.
16. Shackney SE, Smith CA, Miller BW, Burholt DR, Murtha K, Giles HR, Ketterer DM, Pollice AA. Model fo the genetic evolution of human solid tumors. *Cancer Res* 1989; **49**: 3344–3354.
17. Shackney SE, Singh SG, Yakulis R, Smith CA, Pollice AA, Petruolo S, Waggoner A, Hartsock RJ. Aneuploidy in breast cancer: a fluorescence in situ hybridization study. *Cytometry* 1995; **22(4)**: 282–291.
18. Shackney SE, Shankey TV. Common patterns of genetic evolution in human solid tumors. *Cytometry* 1997; **29**: 1–27.
19. Shackney SE, Silverman JF. Molecular evolutionary patterns in breast cancer. *Adv Anat Pathology* 2003; **10(5)**: 278–290.
20. Shackney SE, Smith CA, Pollice A, Brown K, Day R, Julian T, Silverman JF. Intracellular patterns of Her-2/neu, ras, and ploidy abnormalities in primary human breast cancers predict postoperative clinical disease-free survival. *Clin Cancer Res* 2004; **10**: 3042–3052.
21. Sinclair A. Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combin Probab Comput* 1992; **1**: 351–370.
22. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale A-L. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001; **98(19)**: 10869–10874.
23. Valk PJM, Verhaak RGW, Beijen MA, Erpelinck CAJ, van Waalwijk can Doorn-Khosrovani SB, Boer JM, Beverloo HB, Moorhose MJ, van der Spek PJ, Löwenberg B, Delwel R. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med* 2004; **350(16)**: 1617–1628.
24. van de Vijver M, van de Bersselaar R, Devilee P, Cornelisse C, Peterse J, Nusse R. Amplification of the neu (c-erbB-2) oncogene in human mammary tumors is relatively frequent and is often accompanied by amplification of the linked c-erbA oncogene. *Mol Cell Biol* 1987; **7(5)**: 2019–2023.
25. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002; **347**: 1999–2009.
26. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**: 484–485.
27. Winston JS, Ramanaryanan T, Levine E. Her-2/neu evaluation in breast cancer: are we there yet? *Am J Clin Pathol* 2004; **121**: S33-S49.