Cancer Phylogenetics from Single-Cell Assays

Gregory Pennington[*] **Stanley Shackney**[†]
Russell Schwartz[‡]

January 2006
CMU-CS-06-103

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[*]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[†]Allegheny Singer Research Institute, Allegheny General Hospital, Pittsburgh, PA 15212, USA
[‡]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh PA 15213, USA

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

Keywords: computational biology, cancer, FISH, phylogeny

Abstract

In the field of cancer biology, there is currently great interest in the development of "targeted therapeutics" that attack specific molecular abnormalities characterizing subsets of cancers. Computational methods have been essential in identifying subsets of tumors sharing a common molecular mechanism, making it possible to identify meaningful groupings for targeted therapy. To date, such approaches have been limited in their ability to infer the specific sequences of molecular changes, or progression pathways, by which a tumor forms and increases in aggressiveness. In the present work, we develop computational methods for inferring progression pathways from cell-by-cell assays. Our methods bypass important limitations of the current approaches by recognizing and taking advantage of tumor heterogeneity. We define a model for tumor progression and introduce a procedure for cancer phylogenetics based on the inference of likely progression pathways in individual patients. This procedure is formulated as a set of easily tractable graph problems. We demonstrate the methods on a set of fluorescence in situ hybridization (FISH) assays, which measure gene and chromosome gain and loss from a collection of fifty tumor samples. The results are consistent with prior knowledge about the role of the genes examined in cancer progression, and they suggest additional features of progression pathways involving the genes studied.

1 Introduction

Recent computational studies have profoundly transformed our understanding of cancer biology. It has long been understood that cancer is not a single disease, but rather a seemingly innumerable collection of possible ways different genetic lesions might uncouple cell growth from normal controls. Two cancers that appear identical to the clinician might in fact be driven by completely distinct causes at the molecular level and may therefore have very different optimal treatment regimens and patient outcomes. Computational methods have proven crucial in translating this basic insight into the practical discoveries. In particular, the application of clustering methods to gene expression microarray data [5, 22] has been able to identify sets of tumors exhibiting common underlying molecular aberrances. Such approaches have also been valuable in sorting clinically similar tumors into subsets representing common molecular causes [6, 14, 16], which in turn has shown practical value in predicting patient prognosis [23, 27, 26, 24] and in predicting the efficacy of particular treatments [3, 1]. Such successes have fueled hope that cancer treatment will be dramatically improved by moving from broadly useful chemotherapeutics to "targetted therapeutics" that are designed to address particular molecular lesions associated with a cancer subtype. The most notable example is the drug trastuzumab (Herceptin), an antibody to the Her-2/neu gene designed specifically to treat a subset of breast cancers in which that gene is known to be overexpressed [13].

Despite the utility of the notion of cancer subsets, it is a simplification. A cancer sub-type represents not a single static entity but rather a general pathway of progression by which cells accumulate molecular abnormalities [12, 20]. Any given patient may have progressed to varying degrees along this pathway. Furthermore, the degree of progression significantly influences prognosis even for patients proceeding along the same pathway [15]. If we wish to understand the molecular basis of cancer and maximize our ability to treat it, we need to know not only the general changes characterizing the pathway as a whole, but also the specific steps along any given pathway. A method was recently proposed by Desper et al. [4] to attempt to identify actual sequences of progression among tumors using phylogenetic methods. Working from microarray data and using a distance metric similar to those used in clustering approaches, they showed that tumors could be classified into meaningful evolutionary trees in which molecularly similar tumors group together and in which distance from a normal-like state apparently corresponds to degrees of progression.

That approach also has its limitations, primarily because of the limitations of the kind of data on which it is based. The method presumes that one can treat each tumor as a progression state and find the likely evolutionary tree among all of those states. However, a tumor is not homogeneous. Though cell proliferation rate generally increases with progression, cells in later states tend to augment, not replace, cells in earlier states [19]. Figure 1 illustrates this process. Individual cells in a tumor can thus be expected to span a range of progression states from fully healthy to advanced. This heterogeneity suggests that individual tumors should be treated as evolving populations each containing a partial record of the universe of cancer progression pathways. By finding likely evolutionary pathways within single tumors, we can more precisely infer the fine-level sequences of molecular events defining progression by any particular tumor type. Microarrays do not yield insight into this process, because they give tissue-wide average expression levels that obscure the specific fine-scale changes between individual progression states. Cancer prognosis

can be significantly influenced by changes assessable at the single-cell level but not apparent at the level of tissue-averaged assays [21].

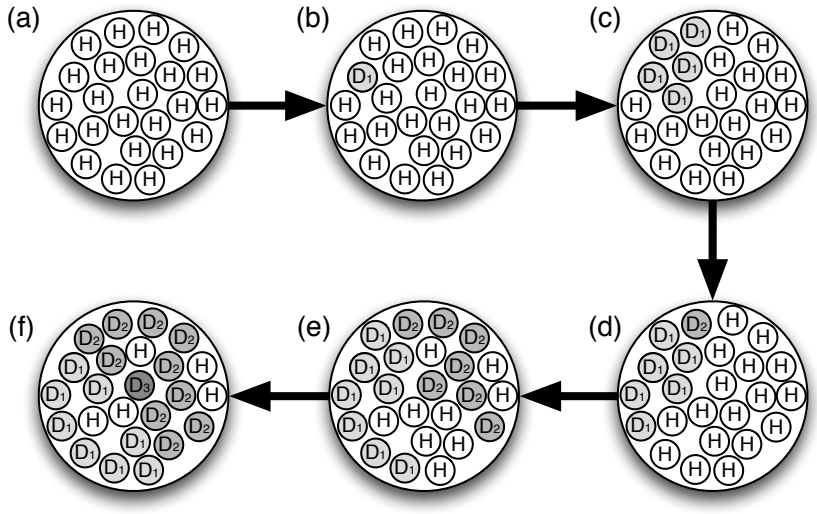


Figure 1: Illustration of cancer progression resulting in tumor heterogeneity. **(a):** A healthy mass of cells labeled H. **(b):** A cell mutates into a diseased state D_1, which encourages proliferation and further progression. **(c):** The proliferating cell expands, leaving a heterogeneous population. **(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.

Fortunately, it is possible to collect single-cell data on tumor samples, which has the potential to resolve these issues. This type of data includes any kind of measurement that assays individual cells rather than a tissue-wide average. One such technique is fluorescent *in situ* hybridization (FISH), in which fluorescent probes are hybridized to cellular DNA and microscopy is used to count the probes that anneal. FISH can be used to test for loss and gain of individual genes, genetic regions, or entire chromosomes—all frequent events in cancer progression. Single-cell methods are of great value because if we can identify cell populations in different progression states within a single tumor and determine how they relate to one another, then we can pinpoint the specific sequence of changes that led to cancer progression in that individual. Furthermore, if we can do this for many individuals, we can identify the common pathways across the human population.

In the present work, we pursue the problem of performing phylogenetic analysis on single-cell cytometric data from collections of patients to infer commonly used progression steps across patient populations. In the remainder of this paper, we formalize a model for inference on this data, present a method for inferring common progression pathways, and apply the method to a real data set. We focus here on application to FISH data, demonstrating the methods on a set of such data gathered for a prior study [7]. However, we believe the methods developed will generalize to other kinds of single-cell tumor assays.

2 Methods

The intuition behind our approach derives from non-computational analyses of similar single-cell data [18, 20]. We wish to find common progression pathways in the population by exploiting heterogeneity in individual tumors. For the present study, we apply a two-stage approach. First, we develop an evolutionary model for individual patients and construct an evolutionary tree for each individual, using variants of standard methods for phylogeny inference. Then, we identify pathways shared by a substantial fraction of the full patient population. The method, at a high level, is illustrated in figure 2.

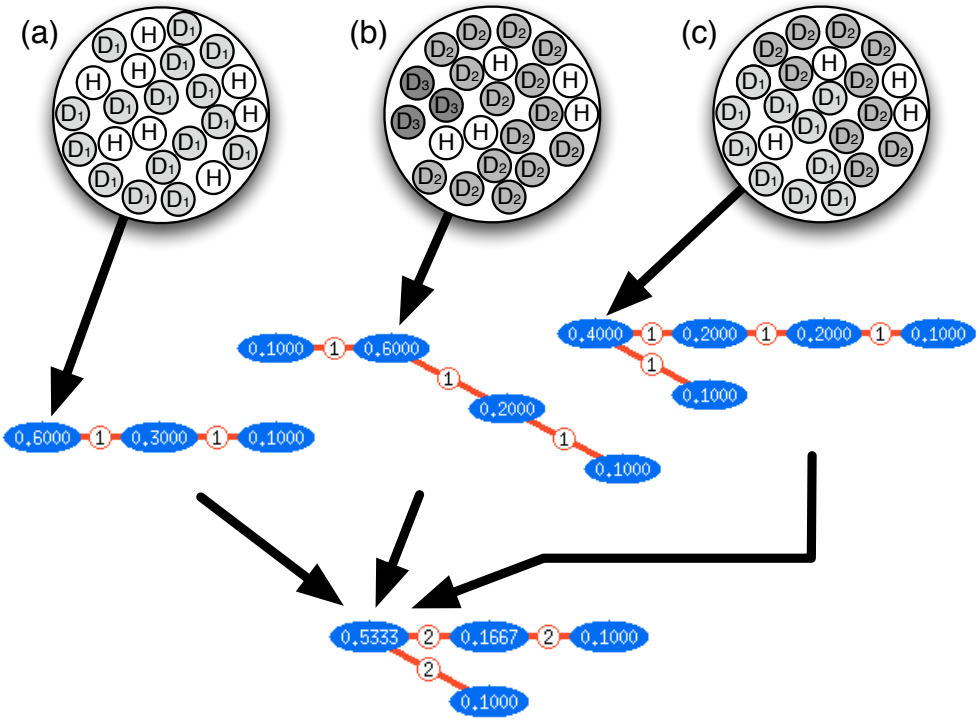


Figure 2: High-level overview of the proposed cancer phylogenetics approach, illustrating the method with three hypothetical tumor samples: (a) sample containing disease state D_1, (b) sample containing disease state D_2 and a small population of D_3, and (c) sample containing large populations of D_1 and D_2. First, likely phylogenies are referred for individual tumors based on cell-by-cell measurements and a model of allowed progression steps. Second, a consensus graph is formed by finding paths occurring in a significant subset of the individual phylogenies (two in this illustration). The arrows point to the root (normal state) of the phylogenies.

2.1 Input Data

While the high-level approach is intended to be generic with regard to the data type, for the present work we assume a particular data format available for this study. Input is presumed to be FISH

copy number data for a set of cells, with each FISH assay counting copies of a targeted gene and the centromere of its chromosome. We can therefore represent the input 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 the sample that have i copies of the chromosome and j copies of the gene. We refer to these cells as the population of cells at state (i, j).

2.2 Mutation Model

For the purpose of phylogeny construction, we must establish a model for mutation events. Aberrant cell populations result from a series of progressive deviations from the normal diploid state, $(2, 2)$, and intermediate populations are often still identifiable in FISH data. For example, in a matrix with a population at $(2, 4)$ corresponding to diploid chromosomes but two extra copies of the gene, it is likely that there were at least two mutation events corresponding to the addition of surfeit chromosomes, genes, or chromosome/gene pairs. We would therefore expect to see some other population that is closer to the normal state (at $(2, 3)$ for example). By linking "neighboring" states (a term we will define presently) into a tree rooted at the normal state, we create a hypothetical phylogeny for tumor cell populations.

To formalize the problem of finding the optimal phylogeny, we translate the input matrix into graph, G, in which we create a node (i, j) for each element m_{ij} of M for which $m_{ij} > 0$. We then presume a limited number of allowed "local moves" from any given tumor state. We assume in the model that it is possible for a single mutation event to cause gain or loss of a single gene copy, or the simultaneous gain or loss of the gene and its chromosome. Additionally, if a cell is known to contain a chromosome that is missing an allele (if $i > j$), we assume chromosomal gain or loss may occur without a simultaneous change in allelic count. However, a single mutation cannot result in simultaneous chromosomal loss and allelic gain or chromosomal gain and allelic loss. We represent the potential mutations affecting state (i, j) with the connectivity matrix in Figure 3. These local connectivity constraints define most of the edge set of G.

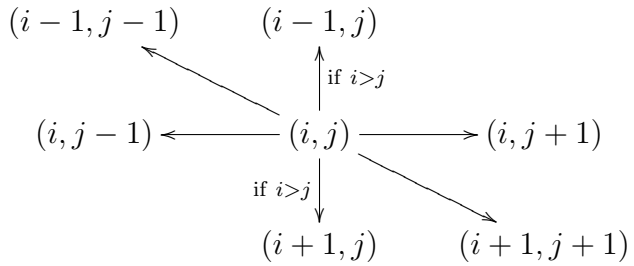


Figure 3: Connectivity of a population of cells in a FISH matrix.

While other mutation events are possible — for example, duplication of a chromosome containing multiple gene copies, thus adding multiple genes in a single step — we assume a conservative

mutation model as we lack an empirical basis for weighting probabilities of relatively rare single moves. Since we will focus on the consensus of many trees, we expect the omission of very unlikely transitions to have no ultimate effect on our analysis.

We have to modify the graph to deal with two special cases. First, it is possible G may be disconnected, due either to an inaccurate assumption in the model or failure to detect some true intermediate state. In order to connect the graph, we define a set of Steiner nodes and edges connecting each disconnected component of the graph, using the shortest possible Steiner path between any two islands, using the maximum frequency of the end points to break ties in the choice of path. Second, we presume the normal diploid state is the root of all inferred phylogenies. In the rare case that no cells are observed in the normal state, we create a $(2, 2)$ node with zero weight.

We next assign edge weights to the graph to represent likely probabilities of different transitions. Intuitively, a well populated state is more likely than a lightly populated state to have be an ancestor of another state that is near them both. A formulation that captures that intuition would be to make G a directed graph, where the weight of an edge from state (i, j) to (k, l) is $-m_{ij}$. While we could solve for this metric as a directed minimum spanning tree problem, it is algorithmically convenient to use an equivalent undirected formulation in which we define G' to be an undirected graph containing the same vertices as G, where the weight of an edge from (i, j) to (k, l) is defined to be $-(m_{ij} + m_{kl})$.

Theorem 2.1 *t is a minimum spanning tree for G' iff t is a minimum spanning tree for G.*

Proof: Let T be the set of all minimum spanning tree for G and let T' be the set of all minimum spanning tree for G'. Then, $(V, E) \in T$ is a spanning tree that minimizes

$$\sum_{(v_1,v_2)\in E} -p(v_1)$$

over all possible spanning trees in G. Similarly, $(V, E') \in T'$ is a spanning tree that minimizes

$$\sum_{(v_1,v_2)\in E} -p(v_1) - p(v_2) = \left(\sum_{(v_1,v_2)\in E} -p(v_1)\right) - \left(\sum_{(v_1,v_2)\in E} p(v_2)\right)$$

over all possible spanning trees in G'. Because every node in a tree has in-degree 1, $\sum_{(v_1,v_2)\in E} p(v_2)$ is a constant, and so $T = T'$. □

We complete our model by specifying the metric for optimization. We assume each node has a unique ancestor, and therefore we seek a tree G for each patient. Given our definition of edge weight, an optimal tree T is simply a minimum spanning tree on G. Effectively, this means that the optimal tree is one such that the sum of the weights of the end-points of all edges is maximized. Because the root node is known to us, directionality in the final graph is implicit.

2.3 Algorithms

Our computational task consists of two steps: finding the optimal patient phylogenies given the preceding model and finding frequently used pathways given the phylogenies. Both problems are fairly straightforward computationally. High level pseudocode for the full procedure is presented as Algorithm 1.

Given the problem definition, single-patient phylogeny construction can be trivially solved by any classic minimum spanning tree (MST) algorithm. For this purpose, we apply Kruskal's algorithm [9]. For a group of n samples, application of the MST method to each patient sample will result in a forest F of n phylogenies.

We next seek to infer frequent paths across all phylogenies, which we propose represent likely progression states. We define a frequent path to be any path starting from the $(2, 2)$ state that occurs in a fraction at least f of the single-tumor phylogenies, where f is a user-specified parameter.

We can find common paths by iterating over all phylogenies and, for each phylogeny, over all non-root nodes by depth first search. This procedure identifies candidate paths. For each path, we can then count how many phylogenies contain all edges in the path. If the count exceeds $f \cdot n$, then we record the path. Finally, we can take the union of all such frequent paths, creating a consensus graph C of frequent progression pathways. In practice, we use a faster but more involved method.

Algorithm 1 High-level procedure for forming consensus trees

given a collection of samples C
for all $S \in C$ **do**
    convert the FISH matrix for S into a graph G
    add weighted edges to G according to the connectivity model
    join islands in G by the minimum length, maximally popular path
    find a minimum spanning tree which contains all the nodes of G
let P be the set of all paths that occur in at least a fraction f of the individual MSTs
the consensus graph C is the union of all the paths in P

3 Results and Discussion

We applied our methods to a data set that was gathered for a prior study on amplification patterns in human breast cancers [7] and used FISH to obtain cell-by-cell counts of gene and chromosome centromere copy numbers [18, 7]. Our dataset consists of a subset of tumor samples from 50 patients, each of which was examined by two FISH assays. One assay measured copy numbers of the Her-2/neu gene and chromosome 17 (on which Her-2/neu is found). The second examined copy numbers of c-myc and chromosome 8 (on which c-myc is found). These genes are of interest because both are known to be cancer-associated. Her-2/neu amplification promotes cell proliferation and is associated with a class of breast cancers [17, 25], while c-myc amplification is associated with aneuploidy, particularly when in combination with the loss of the p53 tumor suppressor gene [10, 2, 11, 29]. The data set contained an average of 98 single-cell measurements

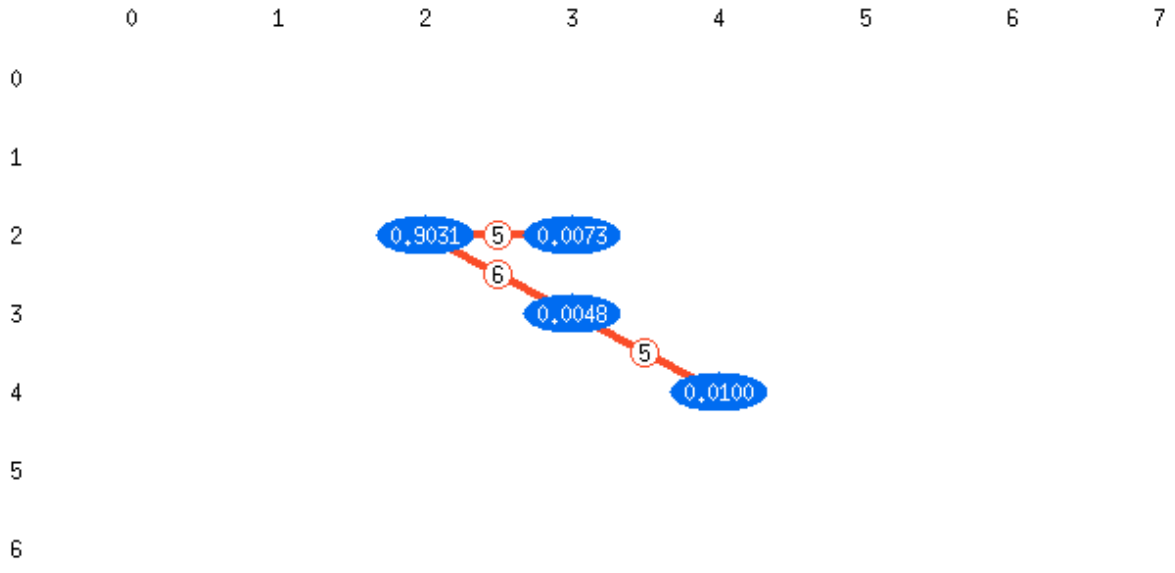
per patient for Her-2/neu and chromosome 17 and an average of 52 single-cell measurements per patient for c-myc and chromosome 8.

We first inferred phylogenies for the c-myc data. For this data set, 86% of patients show strictly normal profiles (two copies of both gene and chromosome). This result is consistent with prior indications that amplification of c-myc is a relatively late event in progression of a subset of cancers [19]. In our dataset, four of the five patients that amplify c-myc also amplify Her-2/neu, likewise consistent with the indications that c-myc is a late event in particularly aggressive Her-2/neu amplifying tumors. Figure 4(a) shows the inferred progression pathways common to at least 10% of patients. The figure is quite simple, showing only two short chains of progression. One corresponds to a single amplification of c-myc alone, without amplification of the chromosome. The other corresponds to amplification of both the gene and chromosome, with triploid and tetraploid states observed in a significant fraction of the patients. This result again fits prior expectations, as c-myc amplification is known to be associated with stable tetraploidy in cancerous cells [29]. Each of the two pathways corresponds to approximately 10% of the patients, again consistent with c-myc amplification being part of either a rare pathway or a late stage in the pathway.

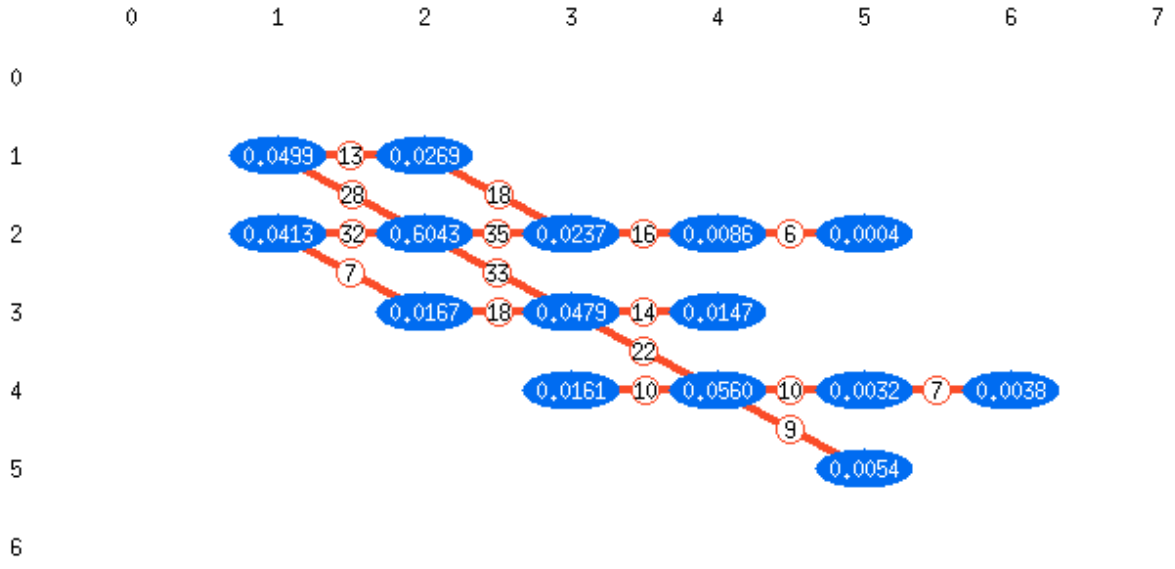
We next inferred phylogenies for the Her-2/neu data. Figure 4(b) shows all pathways inferred for at least 10% of the patients. The graph is far more complicated than the one inferred for c-myc. Several features are worth noting. The two largest components of the graph correspond to Her-2/neu amplifying pathways. One pathway, observed in 70% of patients, proceeds to the right from the root node, corresponding to amplification of the gene independent of the chromosome. A second, observed in 66% of patients, proceeds diagonally down from the root node, corresponding to amplification of the gene and chromosome simultaneously. This latter pathway shows some additional variability around the major diagonal axis, with subsets of patients exhibiting gene gain or loss. The frequency of the pathways indicates that a large fraction of patients exhibit both of them in independent sets of cells. It has been previously observed that Her-2/neu gene amplification can occur in diploid cells or coincident with polyploidy [7], consistent with the observation of these two dominant pathways. We also observe that individual cells can shift amplification mechanisms, as at least 42% of chromosome-amplifying lines contain populations that appear to switch from chromosome amplification to gene amplification at some point in the progression. We do not, however, observe a significant fraction of gene-amplifying cells switching to chromosome amplification late in the tree.

We can further observe prominent shorter pathways in the Her-2/neu consensus graph corresponding to gene or chromosome loss. Large fractions of cells exhibit Her-2/neu loss and simultaneous loss of chromosome 17 and Her-2/neu. These pathways are not explained by the known activity of Her-2/neu in promoting certain cancers by amplification. However, we do not observe similar loss states in the c-myc data, and therefore believe that their observance at high frequency is reliable.

We can partially attribute these extra states to the coincidental fact that chromosome 17 also contains p53, a critical tumor suppressor gene whose loss is implicated in many human cancers. FISH data for p53 and chromosome 17 is available from the same study for a subset of 13 of the 50 cancer patients, allowing us to validate this hypothesis. Of the 13 patients for which we have both Her-2/neu and p53 data, 6 exhibit Her-2/neu loss, and all of them also exhibit p53 loss. This



(a) Consensus tree for c-myc and chromosome 8.



(b) Consensus tree for Her-2/neu and chromosome 17.

Figure 4: Commonly occurring progression pathways detected in the breast cancer data set. Each graph consists of the union of all pathways from the root state (2, 2) found in at least 5 (10%) of the patients. Increasing gene copy numbers proceed to the right and increasing chromosome copy numbers proceed down. Labels on nodes show the fraction of all cell measurements corresponding to the given state and labels on edges the number of patients whose inferred phylogenies possessed the given edge.

supports our hypothesis that the chromosome loss observed in the Her-2/neu data is in fact part of progression pathways characterized by early p53 loss, with simultaneous Her-2/neu loss being incidental. Overall, 11 patients (85%) either exhibit loss in both Her-2/neu and p53 datasets or loss in neither.

This hypothesis does not explain why we would observe many cases of loss of the Her-2/neu gene without chromosome 17; Her-2/neu and p53 sit on opposite arms of the chromosome, so it is unlikely both would be lost without losing the chromosome centromere. The Her-2/neu gene loss pathway may in fact reflect a more complicated series of steps than the single mutation event inferred by the model, or it may be an incidental result of some other mutation process. In particular, we can propose that this state may reflect a pathway involving allelic loss of the tumor suppressor BRCA1, another breast-cancer associated gene which happens to co-occur on the same arm of chromosome 17 as Her-2/neu. While BRCA1 is predominantly associated with hereditary breast cancer, somatic loss or mutation of the gene has been found to be associated with sporadic cases of breast cancer [28, 8].

4 Conclusions

We have defined a model and methods for computationally inferring cancer progression from heterogeneous tumor samples and demonstrated them on cell-by-cell gene and chromosome copy number data from breast cancer tumors. The results are consistent with prior knowledge about the role of the genes studied in cancer progression and suggest several features of their progression pathways beyond that derived from the prior work. This work shows how an important problem in human biology can be addressed by adapting methods from a well-studied field of computational biology.

While the work presented is a first step, much more remains to be done. Several aspects of the model might be improved. Better models of experimental noise may lead to more effective ways to infer long paths without overwhelming them with erroneous states. More sophisticated computational tools from the field of phylogenetics may prove useful in improving the quality of the inferences, particular with regard to inferring missing (Steiner) states. It is also possible that better use of multiple data sets could be made by converting the two-stage approach adopted here into a single optimization and by developing methods to join trees across multiple assays. Other kinds of cytometric assay, particularly single-cell expression measurements, may also prove more broadly informative than gene and chromosome copy numbers. Similar methods for phylogenetics on cytometric assays may also prove useful in other applications, such as examining expression or epigenetic changes in organismal development or in cell signaling networks of mature tissues.

References

[1] M. Ayers, W. F. Symmans, J. Stec, A. I. Damokosh, E. Clark, K. Hess, M. Lecocke, J. Metivier, D. Booser, N. Ibrahim, V. Valero, M. Royce, B. Arun, G. Whitman, J. Ross, N. Sniege, G. N. Hotoagyi, and L. Pusztai. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. Journal of Clinical Oncology, 22(12):1- 10, 2004.

[2] O. Chernova, M. Hernov, Y. Ishizaka, M. Agarwal, and G. Stark. MYC abrogates p53-mediated cell cycle arrest in N-(phosphonacetyl)-L-aspartate-treated cells, permitting CAD gene amplification. Molecular Cell Biology. 18:536-545, 1998.

[3] H. E. Cunliffe, M. Ringnér, S. Bilke, R. L. Walker, J. M. Cheung, Y. Chen, and P. S. Meltzer. The gene expression response of breast cancer to growth regulators: patterns and correlation with tumor expression profiles. Cancer Research, 63:7158-7166, 2003.

[4] R. Desper, J. Khan, and A.A. Schaffer. 2004. Tumor classification using phylogenetic methods on expression data. J Theor Biol 228: 477-496.

[5] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. PNAS USA, 95:14863-14868, 1998.

[6] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeej, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531-537, 1999.

[7] L.E. Janocko, K.A. Brown, C.A. Smith, L.P. Gu, A.A. Pollice, S.G. Singh, T. Julian, N. Wolmark, L. Sweeney, J.F. Silverman, and S.E. Shackney. Distinctive patterns of Her-2/neu, c-myc, and cyclin D1 gene amplification by fluorescence in situ hybridization in primary human breast cancers. Cyometry, 46(3):136-149, 2001.

[8] M. Janatova, M. Zikan, P. Dundr, B. Matous, and P. Pohlreich. Novel somatic mutations in the BRCA1 gene in sporadic breast tumors. Human Mutation. 25(3):319, 2005.

[9] J. B. Kruskal. On the shortest spanning subtree of the graph and the traveling salesman problem. Proceedings of the American Mathematical Society, 7:48-57, 1956.

[10] S. Mai, M. Fluri, D. Siwarski, and C. Huppi. Genomic instability in MycER-activated Rat1A-MycER cells. Chromosome Research, 4:365-371, 1996.

[11] S. McCormack, Z. Weaver, S. Deming, et al. Myc/p53 interactions in transgenic mouse mammary development, tumorigenesis, and chromosome instability. Oncogene, 16:2755-2766, 1998.

[12] P. C. Nowell. The clonal evolution of tumor cell populations. Science, 194:23-28, 1976.

[13] M.D. Pegram, G. Konecny, and D.J. Slamon. 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., 103:57-75, 2000.

[14] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, Øystein Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A.-L. Børresen-Dale, P. O. Brown, and D. Botstein. Molecular portraits of human breast tumors. Nature, 406:747-752, 2000.

[15] T. Ried, K. Heselmeyer-Haddad, H. Blegen, E. Schrock, and G. Auer. Genomic chnages defining the genesis, progression, and malignancy potential of solid human tumors: a phenotype/genotype correlation. Genes Chromosomes Cancer, 25(3):195-204, 1999.

[16] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. van de Rijn, M. Waltham, A¿ Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics, 24:227-235, 2000.

[17] D. J. Salmon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, W. L. McGuire. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 235: 177-182, 1987.

[18] S.E. Shackney, S.G. Singh, R. Yakulis, C.A. Smith, A.A. Pollice, S. Petruolo, A. Waggoner, R.J. Hartsock. Aneuploidy in breast cancer: a fluorescence in situ hybridization study. Cytometry. 22(4):282-291, 1995.

[19] S.E. Shackney and T.V. Shankey. Common patterns of genetic evolution in human solid tumors. Cytometry. 29:1-27, 1997.

[20] S. E. Shackney and J. F. Silverman. Molecular evolutionary patterns in breast cancer. Advances in Anatomic Pathology, 10(5):278-290, 2003.

[21] S. E. Shackney, C. A. Smith, A. Pollice, K. Brown, R. Day, T. Julian, and J. F. Silverman. Intracellular patterns of Her-2/neu, ras, and ploidy abnormalities in primary human breast cancers predict postoperative clinical disease-free survival. Clinical Cancer Research. 10: 3042-3052, 2004.

[22] D. Slonim. From patterns to pathways: gene expression data analysis comes of age. Nature Genetics, 32:502-508, 2002.

[23] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A.-L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. PNAS USA, 98(19):10869-10874, 2001.

[24] P. J. M. Valk, R. G. W. Verhaak, M. A. Beijen, C. A. J. Erpelinck, S. B. van Waalwijk can Doorn-Khosrovani, J. M. Boer, H. B. Beverloo, M. J. Moorhose, P. J. van der Spek, B. Löwenberg, and R. Delwel. Prognostically useful gene-expression profiles in acute myeloid leukemia, The New England Journal of Medicine, 350(16):1617-1628, 2004.

[25] M. van de Vijver, R. van de Bersselaar, P. Devilee, C. Cornelisse, J. Peterse, and R. Nusse. 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. Molecular and Cellular Biology, 7(5): 2019-2023, 1987.

[26] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine, 347: 1999-2009, 2002.

[27] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. 415:484-485, 2002.

[28] Q. Yang, G. Yoshimura, M. Nakamura, Y. Nakamura, T. Suzama, T. Umemura, I. Mori, T. Sakurai, and K. Kakudo. BRCA1 in non-inherited breast carcinomas. Oncology Reports. 9(6):1329-1333, 2002.

[29] X. Yin, L. Grove, N. Datta, M. Long, and E. Prochownik. c-myc overexpression and p53 loss cooperate to promote genomic instability. Oncogene, 18:1177-1184, 1999.