Chromosome Abnormalities in Ovarian Adenocarcinoma III: Using Breakpoint Data to Infer and Test Mathematical Models for Oncogenesis

Richard Simon[1] *Richard Desper*[2] *Christos H. Papadimitriou*[3]
Amy Peng[4] *David S. Alberts*[5] *Raymond Taetle*[5] *Jeffrey M. Trent*[6]
Alejandro A. Schäffer[7,*]

[1] Biometric Research Branch, National Cancer Institute, NIH, Bethesda, MD USA
[2] Deutsches Krebsforschungzentrum, Abt. Theoretische Bioinformatik, Heidelberg, GERMANY
[3] Department of Elec. Eng. and Comp. Sci., U. C. Berkeley, Berkeley, CA USA
[4] The Emmes Corporation, Potomac, MD USA
[5] Departments of Medicine and Pathology, University of Arizona and Arizona Cancer Center, Tucson, AZ USA
[6] Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD USA
[7] Computational Biology Branch, National Center for Biotechnology Information, NIH, Bethesda, MD USA

Running Title: Models for Ovarian Adenocarcinoma Breakpoints
The work of Christos H. Papadimitriou was supported in part by NSF grant CCR-9626361. The collection of the ovarian adenocarcinoma data at the University of Arizona was supported by NIH grants CA41183, CA23074, and ES-95-32.
* **Correspondence to:** Alejandro A. Schäffer; NCBI/NIH; Bldg 38A, Room 8N805; 8600 Rockville Pike; Bethesda, MD 20894. USA.
e-mail: schaffer@helix.nih.gov
Phone: (301) 435-5884
FAX: (301) 480-9241
Keywords: Ovarian cancer, chromosome breakpoints, algorithms, phylogenetic trees, maximum branching, directed acyclic graphs, survival analysis.

Abstract

Cancer geneticists seek to identify genetic changes in tumor cells and to relate the genetic changes to tumor development. Because single changes can disrupt the cell cycle and promote other genetic changes, it is extremely hard to distinguish cause from effect. In this paper we illustrate how seven techniques from statistics, theoretical computer science, and phylogenetics can be used to infer and test possible models of tumor progression from single genome-wide descriptions of aberrations in a large sample of tumors. Specifically, we propose four tree models for tumor progression inferred from the large ovarian cancer data set described in the first two papers in this series. The models are derived from two different methods to select the non-random genetic aberrations and two different methods to infer the trees, given a set of events. Various aspects of the tree models are tested and extended by five methods: overall tests of independence, likelihood ratio tests, principal components analysis, directed acyclic graph modeling, and Bayesian survival analysis. All our methods lead to strikingly consistent conclusions about chromosomal breakpoints in ovarian adenocarcinoma, including: (1) the non-random breakpoints in ovarian adenocarcinoma do not occur independently; (2) breakpoints in regions 1p3 and 11p1 are important early events and distinguish a class of tumors associated with poor prognosis (3) breakpoints in 1p1, 3p1, 1q2 distinguish a class of ovarian tumors, and the breaks at 1p1 and 3p1 are associated with poor prognosis.

INTRODUCTION

We propose tree models to associate the occurrence of chromosomal breakpoints with tumor progression in ovarian adenocarcinoma. We use a variety of statistical techniques to test and confirm some of the predictions implicit in the tree models. We illustrate how mathematical modeling tools from statistics, theoretical computer science, and phylogenetics can be used to analyze a large data set of single samples from tumors of the same type. Identifying genetic changes in tumors and deciding which changes are significant markers of tumor progression is a multi-faceted complex problem at the core of cancer genetics. Because single genetic changes can disrupt mitosis and promote other genetic changes, it is extremely difficult to distinguish cause from effect.

The present study is based on a collection of 244 single samples of ovarian adenocarcinoma, focusing on 135 tumors that have at least 1 chromosomal breakpoint (Taetle et al., 1999a, 1999b) The two previous papers in this series gave a detailed catalog of the breakpoints, identified a set of 13 breakpoints that appear to be non-random and cover all but 14 of the cases, and showed that breakpoints on 1p1 and 3p1 are correlated with poor prognosis.

The thrust of this paper is that it is possible to do much more than look for *single* recurring abnormalities, when the sample size is moderately large. We suggest that it is possible to look at patterns of abnormalities occurring together in single tumors and to infer mathematically justified hypotheses for questions such as: (1) which events tend to occur early in tumor progression? (2) which events tend to occur together, possibly identifying a tumor subclass? (3) which pairs of events may have a cause-and-effect relationship, and in which direction?

The idea that multiple genetic alterations in one type of tumor can be related to progression is hardly new. One of the most celebrated achievements in cancer genetics is the work by Vogelstein's group on colorectal cancer including a model showing how 4 different genetic changes correspond to different stages in tumor progression (Fearon and Vogelstein 1990; Vogelstein et al., 1988). Unfortunately, there has been very limited success in finding such models for other types of tumors. We are not aware of any such progression model for ovarian cancer. The colorectal cancer model lists the stages and genetic changes in a single path going from normal cell to advanced tumor. It has been suggested (Kuukasjärvi et al., 1997) that such path models may not be appropriate for many types of cancer because they allow very limited opportunity to model heterogeneity.

We search for models in a much larger class that mathematicians and computer scientists call *trees*, that include the class of *paths* as a special case. Trees are much more appropriate than paths in modeling cancers of high heterogeneity. Tumors of a given histologic type are generally heterogeneous with regard to the breakpoints present, and they may also be heterogeneous with regard to the temporal relationship in which these breakpoints occur. We also test some of the predictions of the tree models using several statistical techniques of independent interest, and evaluate the relationships of breakpoints to survival.

MATERIALS AND METHODS

Patient and Data

Solid tumor samples for cytogenetic analyses were obtained from 244 patients with ovarian cancer at initial laparotomy as described (Taetle et al., 1999a). Collection and management of patient data, and Cox proportional hazards survival analysis are described in (Taetle et al, 1999a, 1999b). For the non-medical reader it may be useful to know that the data are presented as a table, where each row corresponds to one patient. For each patient, the data include: case number and other identifiers, date of biopsy, stage of cancer, site of biopsy sample, gender, overall cytogenetic profile (gross indication of whether cells look normal under the microscope), karyotype abnormality index (numerical score quantifying the ability to identify chromosomes in the cells), ploidy (chromosome count relative to the normal $2n$ for autosomes), and a list of cytogenetic abnormalities using a technical notation delineated and standardized in ISCN (1995).

Cytogenetic Analysis and Non-random Breakpoints

Cytogenetic analyses were performed as described (Thompson et al., 1994, Taetle et al., 1999a). Karyotypes were described according to the ISCN (1995). There were 135 samples with at least one breakpoint detected. The breakpoints were assigned to chromosomal regions and statistical analysis was carried out to detect non-random breakpoints (Taetle et al., 1999a). There were 21 recurrent breakpoints determined to be non-random, the smaller subset (Set I) of 13 breakpoints (1p1, 1q1, 1p2, 1q2, 1p3, 1q4, 3p1, 6q1, 7p1, 7p2, 11p1, 11q2, 12q2) represented in 121 of the 135 cases. This subset of 13 regions was used in subsequent analyses in (Taetle et al., 1999a, 1999b).

Set I was selected in (Taetle et al., 1999a) by a heuristic method aiming at maximizing coverage —the number of patients with at least one breakpoint in the set. There is a different, more established method of selecting non-random events (Brodeur et al. 1982). An implementation of this method is part of the software described herein. This method takes into account prior probabilities for each breakpoint to occur; we used the chromosome arm sizes in (Daniel, 1985) for the autosomes and (Daniel, personal communication) for the X-chromosome and 21q, to estimate these prior probabilities. The method of Brodeur et al. selected the following set (called Set II) of six breakpoints: 1p1, 1q1, 1p3, 7p2, 11p1, 11q2.

For compatibility with both (Taetle et al., 1999a, 1999b) and also for completeness and comparison, we have performed our analysis on *both* sets.

Test of Independence

Several of the models described here are based on pairwise dependence between breakpoints in two bands. Although the methods are exploratory, we wished to utilize a rigorous statistical significance test of the (null) hypothesis of complete independence among breaks in different regions. We developed an exact randomization test of independence of regions containing breakpoints based on the smallest p-value for Fisher's exact test among all pairs of regions as described in the Appendix.

Optimal Trees

We present two types of oncogenetic tree models for ovarian cancer, and we call these *distance-based* and *branching* oncogenetic trees, for reasons explained below. A basic theory in cancer genetics is that tumors evolve clonally from a single cell. Our tree models incorporate this by using a special *vertex* called the *root* and denoted r. The root represents the normal cell that is about to suffer its first genetic alteration. We use the term *oncogenetic trees* to refer to trees that model tumor evolution based on genetic markers.

In *branching* trees, each vertex represents a breakpoint region and vertices are joined by directed edges. The probability that a tumor contains a breakpoint in a region represented by a vertex v depends only on the presence or absence of breakpoints represented by vertices upstream of v. We use the edge $i \rightarrow j$ to reflect the possibility that breakpoints represented by vertices i and j tend to occur together with i occurring before j. Events predicted to be early will be next to the root r. Events that tend to occur together will be clustered in subtrees. If the tree has an edge from event a to event b this represents a hypothetical cause-and-effect relationship. The term *branching* tree comes from the fact that our method to construct these trees uses a computer science technique called *maximum-weight branching*.

Our *distance-based* trees are similar to the distance-based trees used for phylogenetic analysis of species. Each breakpoint region is represented by a terminal vertex (a "leaf") of the tree. The unlabeled internal branch points represent hidden events that classify the breakpoints into different types. We define a distance function between all pairs of breakpoint regions based on the probability of co-occurrence of breakpoints in the regions. We then use an algorithm to identify the tree structure that fits the pairwise distances well, as described in the Appendix. Events predicted to be early will have a short distance to the root. Events that tend to occur together will have short distances to each other.

Some more details about each tree construction method can be found in the Appendix. The branching tree method is described fully in (Desper et al. 1999) and the distance-based method in (Desper et al. submitted). For each method we proved a convergence theorem that states that if the correct underlying model is a tree, and some other plausible assumptions hold, our methods will infer the correct tree structure given enough samples. Although 135 is a lot smaller than the number of samples needed for convergence to apply, we expect that many of the edges predicted by our models are correct.

Likelihood Ratios for Branching Trees

A branching tree is a graphical representation of the dependencies that may exist among the occurrence of breaks in different chromosomal regions. The natural null hypothesis that all events are independent corresponds to a *star* tree in which all vertices are directly connected to the root. We have developed (see Appendix) an expression for the ratio of the maximum likelihood of a branching tree to the maximum likelihood for the star tree. By forming ratios of such likelihoods relative to the star, we can compute the ratio of maximized likelihoods for any two branching tree models. This enables us to quantify the degree of support for specific features found in optimally fitting trees.

Principal Components Analysis

Principal components analysis is a classical statistical method for identifying patterns of variation in multivariate data (Morrison, 1967). For each tumor sample we represented the presence or absence of a breakpoint in each of the regions by a vector of ones and zeroes. The first principal component is the weighted average of the components of these vectors that accounts for the greatest variability across all of the tumor samples. The second principal component is the weighted average, which is linearly independent of the first principal component, and which accounts for the second greatest variability across all of the tumor samples, and so forth. The weights on the components indicate what patterns of breakpoints are most prominent in varying among tumors. We used S-Plus, version 4.0 (Becker et al., 1988) to perform a principal components analysis of the breakpoint regions in Set II and to make a picture of some of the results.

Each tumor can be represented in three dimensions by the values of the first three principal components. Such a three dimensional plot provides a graphical representation of groupings among tumor samples with regard to major breakpoint patterns. We provide such a plot for the breakpoint regions in Set II.

Directed Acyclic Graph Models

The natural generalization of tree models are directed acyclic graph (DAG) models. Each vertex in the DAG represents a subset of the breakpoint regions in Set II. Each vertex is also assigned a level. There are directed edges from the root to each of the six vertices on level one which represent the six possible breakpoint regions that may occur as the first genetic alteration in the tumor among the chromosomal regions in Set II. Associated with the edge from the root to the vertex on level one representing region i is a probability $p_i(1)$ that i occurs as the first genetic alteration. The vertices on level two represent pairs of regions. For each vertex i on level one and vertex $\{i, j\}$ on level two, there is an edge $i \rightarrow \{i, j\}$ representing that i occurs as the first genetic alteration and j as the second. The probability associated with this edge is the conditional probability that a breakpoint in region j occurs as the second genetic alteration given that a breakpoint in region i occurs first. The vertices on the subsequent levels are defined analogously. Edges join vertices of a level to vertices of the next level.

A count of the number of parameters, however, forces the realization that our data set is too small to reasonably fit a fully general DAG model or even to go many levels deep in the restricted form, as was also true for the model of Newton et al. (1994).

Since we are most interested in showing dependence among the breakpoint events, we begin by fitting the graph model based on the assumption of complete independence, and examine the adequacy of fit of the model to the data. If the fit is adequate, we go no further. If, however, the fit is inadequate, our modeling strategy is to introduce parameters expressing dependence between event pairs that best improve the fit. This modeling strategy is analogous to that used in step-wise regression.

The genetic independence model is quite tractable as it contains only $2n - 2$ independent probability parameters for n breakpoint regions. There are n single event probabilities, and for each level $2, 3, \ldots n - 1$, the probability that no further events occur. The model includes the 6 first event probabilities denoted $p_{1p1}, p_{1q1}, p_{1p3}, p_{7p2}, p_{11p1}, p_{11q2}$. We exclude data from tumors without any breakpoints in Set II. The probability that no further breakpoints in Set II occur after the j'th genetic alteration in that set is denoted by $p_0(j)$. More mathematical details can be found in the Appendix.

Survival Analysis

We evaluated the prognostic impact on survival of the breakpoints in Set I. We performed univariate log-rank tests of each breakpoint region and performed multivariate modeling using a Bayesian proportional hazards model (Faraggi and Simon, 1997). We used Bayesian proportional hazards (BPH) models including indicator variables for all breakpoint regions in Set I in addition to stage, grade and residual disease. In our models we grouped ovarian cancer stages I and II together and ovarian cancer stages III and IV together. We grouped well differentiated and moderately differentiated tumors together. We used 1 cm as our threshold for residual disease for grouping cases. The value of the regression coefficient corresponding to a specific region represents the prognostic effect of a breakpoint in that particular region with the breakpoint status of all other regions held fixed. The previous study (Taetle et al., 1999b) found that the presence of any breakpoint in Set I, as opposed to no breakpoint in Set I, was prognostic. The BPH model provides a method for evaluating the logical follow-up question: are specific regions in Set I prognostic more or less than the average effect? The previous study attempted to address this by using an indicator variable for whether or not a patient had any breakpoints in the set. That approach does not control for the number of breakpoints that a patient has, however. More details of how we set the parameters for BPH are in the Appendix.

Relationship to Clinical Variables

To evaluate whether presence of any of the 13 breakpoint regions is individually associated with any of the 3 clinical variables: stage, grade, and diameter of residual tumor, we evaluated 2×2 contingency tables and used the Chi-squared test with continuity correction as implemented in S-Plus version 4.0 (Becker et al., 1988).

RESULTS

Overall Test of Independence

We performed the randomization test of independence of the 13 breakpoint regions in Set I as described in the first subsections of Materials and Methods and of the Appendix. The smallest significance level for Fisher's exact test among all pairs in the data is 0.000095 for the dependence between breakpoints in 1q4 and 6q1. The significance level for the randomization test was 0.012, indicating that the null hypothesis of independence should be rejected.

Tree Models

The best branching tree for Set I (the set of breakpoints selected in (Taetle et al., 1999a)) is shown in Figure 1. The best distance-based tree for this set of breakpoints is shown in Figure 2. The edge-lengths are meaningful in the distance-based tree, but not in the branching tree.

A basic motivation for constructing path or tree models is the assumption that there may be associations among the genetic alterations. If the alterations tend to occur independently, then we would expect the branching tree to look like a *star*, in which each event is adjacent to the root. In fact, our branching tree models for ovarian cancer do not look like stars, confirming the non-independence result based on 2×2 contingency tables.

The distance-based tree for Set I is in Figure 2 and is drawn to scale: the horizontal length of each edge on paper is directly proportional to the distance along the edge in the model. The vertical component of each edge is included merely to show the tree clearly. For this data set, the distance method was not conclusive about the entire tree structure. Other trees of significantly different topology were found which fit the data nearly as well as the tree shown above. However, all of the good solution trees shared a few common traits: clustering of the breakpoint at 7p2 with the one at 11q2, clustering of the breakpoint at 1q4 with those at 6q1 and 1p2, clustering of the breakpoints at 1p1,3p1 and 1q2, and a close relationship between the breakpoints at 7p1 1p3, and 11p1. The relatively rare breakpoint at 1q1 did not show a consistent placement. Also, the exact topological relationship between the clusters mentioned above, if there is one, cannot be conclusively determined from the data.

The two trees in Figure 1 and Figure 2 share the following characteristics. (1) The breakpoint at 1p3 is closest to the root in large subtrees containing many of the same other events. (2)

Breakpoints at 1p1 and 3p1 are clustered tightly together in one subtree, also containing 1q2. (3) Breakpoints at 6q1, 1q4, and 1p2 are tightly clustered together in a subtree, and seem to occur after breakpoints at 1p3 and 11p1. (4) The long edge leading to 1q1 in the branching suggest relative independence. In the branching, this property is reflected by the positioning of 1q1 as a leaf pendant from the root. (5) The breakpoints at 11q2 and 7p2 are close together in one subtree.

The most important inconsistency between the trees is in the placement of 12q2. When building the distance-based trees with PHYLIP (Felsenstein 1989), we found that using different program options varied the placement of 12q2 within the tree, so this discrepancy between the two trees is not too surprising.

The five striking similarities we noted suggest the following five hypotheses. (1) The breakpoint 1p3 is an important early event in ovarian adenocarcinoma, which confirms previous analysis of parts of this data set by different methods (Thompson et al., 1997). (2) the breakpoints in 1p1, 3p1, 1q2 distinguish a subtype of ovarian cancer and it would be consistent with (Taetle et al. 1999a) to hypothesize that these 3 events mark a subtype of ovarian adenocarcinoma that has an especially poor prognosis. (3) The breakpoints at 6q1, 1q4, and 1p2 distinguish another subtype of ovarian adenocarcinoma in which breakpoints at 1p3 and 11p1 tend to occur as earlier events. (4) The breakpoints at 11q2 and 7p2 distinguish another subtype of ovarian adenocarcinoma in which the breakpoint at 1p3 tends to occur as an early event. (5) There may be cause-and-effect relationships within the sets { 1p1,3p1,1q2}, {6q1,1q4,1p2}, {11q2, 7p2}.

We considered next the subset of 6 events in Set II. While the events in Set I covered 121 of the 135 samples with at least one breakpoint, the events of Set II covered only 99 of these 135 samples. This suggests perhaps less explanatory power in general, but with Set II we witnessed a greater degree of measurable correlation, which resulted in trees which fit the data more closely than did the trees generated from Set I.

The tree in Figure 4 was constructed in the same manner as the tree in Figure 2. However, in this case, the trees generated by the Neighbor and Fitch programs in PHYLIP (Felsenstein 1989) led to trees very closely related in topology (only two topological shifts across very short edges separated the output trees). It was possible to test all closely related topologies, of which the best one is shown in Figure 4.

Again, there are striking similarities to the trees resulting from the different methods. As with the branching in Figure 1, 1q1 is a leaf off the root, suggesting independence. The breakpoint at 1p3 is again at or close to the root of a large subtree, with the breakpoint at 11p1 following closely behind. Again, the breakpoints at 11q2 and 7p2 are clustering. Perhaps, most important, the six events are placed in exactly the same relative positions in the two branching trees, and also in exactly the same positions in the two distance-based trees.

We tested some of the predictions made by the tree models using other statistical techniques.

Likelihood Ratios for Branching Trees

As described in Materials and Methods and the Appendix, we derived a formula to compute the likelihood of a branching tree. We use this formula to compare the likelihoods of the predicted branching tree to various alternative branching trees.

Some of the features of the branching tree shown in Figure 1 were evaluated based on likelihood ratios. In all the cases we tested, the tree computed by branching has a higher likelihood than some plausible nearby alternative, even though the branching method is not a maximum likelihood method. Since the alternative trees we compare have similar edge sets, one would not expect the likelihood ratios to necessarily be very large. We computed the likelihood ratio for the tree in Figure 1 relative to a modified tree in which 1p3 remains the predecessor of 11p1 and 7p1 but the 1p3–11p1–7p1 branch is moved downstream to follow 1p2. In the modified tree, the 11q2–7p2 branch is attached to the 6q1 vertex. The rest of the tree in Figure 1 remains unchanged. This likelihood ratio provides a measure of the strength of evidence that 1p3 is an early event. The tree in Figure 1 is 4.1 times as likely as the modified tree.

To evaluate the observation that breakpoints at 1p1 and 3p1 are clustered together in a subtree containing 1q2, we computed the ratio of the likelihood of the tree in Figure 1 to that for a modified tree in which 1p1, 1q2, 3p1 and 12q2 are each attached to the root. The rest of the tree in Figure 1 remains unchanged. This likelihood ratio is greater than 148.

We evaluated the observation that breakpoints at 11q2 and 7p2 are close together in one subtree by computing the likelihood ratio for the tree in Figure 1 to one in which the vertex for 7p2 is attached to the root. The likelihood ratio is 4.1. Trees which did not cluster 6q1 and 1q4 as adjacent vertices were extremely unlikely relative to the tree of Figure 1.

The likelihood ratios reported above are based on using only those cases containing a breakpoint in at least one region in Set I. All of the likelihood ratios are much higher when based on all cases. This may, however, reflect a relatively non-specific additional dependence among breakpoint regions in patients who have at least one breakpoint. In all comparisons we tried, the predicted branching tree has higher likelihood. The method for choosing the best branching tree is not itself based on likelihood, so the likelihood test provides an independent confirmation that the predicted branching trees fit the data well.

Principal Components Analysis

We use a statistical method called *principal components analysis* to identify which breakpoint regions contribute most significantly to the variation between tumors. Figure 5 shows a pictorial representation of the 99 tumors that contained at least one breakpoint in Set II. The axes are the first three principal components (PC's). These three components account for 65% of the total variability among the tumors with regard to breakpoints in Set II. Table shows the weights of the breakpoint regions for the first three PC's. There is a strong correspondence between the PC's and the tree structures. The first PC separates the tumors with regard to whether regions 1p3 and 11p1 both contain breaks, whether only one of the regions contains a break or whether neither contains a break. The second PC distinguishes tumors with a break in the path 1p3–11q2–7p2 but not in 11p1, from those with a break in 11p1 but not 1p3. The third PC is less easily visualized with the view shown in the figure, but it distinguishes tumors with a break in 1p1 or 1p3 from those with breaks in 11p1, 11q2 or 7p2 without a break in 1p3.

The prominence of 1p3 and 11p1 in the first two principal components supports our other analyses indicating that 1p3 and 11p1 are important early events. Group A in the figure consists of those tumors containing breaks in both 1p3 and 11p1. Tumors in Group C do not contain breaks in either of those regions. Group B(i) consists of tumors with a break in 1p3 but not in 11p1 and Group B(ii) consists of those tumors with a break in 11p1 but not in 1p3.

Directed Acyclic Graph Modeling.

The complete level directed acyclic graph (DAG) model based on independence of breakpoints in Set II was fit to the tumor data. The principal aim is to test the prediction of both tree models that 1p3 is an early event and affects the probability of occurrence of subsequent events. We used Set II since it has only 6 events, leading to 6 vertices on level 1 and 15 vertices on level 2, so that the number of probability parameters is not too large.

Using the assumption of independence of breakpoints, the DAG model did not provide a good fit to the observed number of tumors with breaks in 1p3 and 11p1, the most common events. There were 12 tumors with breaks in those two regions and no other regions in Set II but only 7.5 were expected using the maximum likelihood estimators for the independence model. The one-tailed statistical significance level for this underestimate is 0.05. Although this is not definitive evidence of a lack of fit, it is consistent with other evidence that breakpoint regions 1p3 and 11p1 are positively associated. The sum of squares of discrepancies between number of breakpoints observed and number expected over all sets of breakpoints in Set II was 92.98 for the independence model.

We introduced two new parameters to attempt to improve the fit. One parameter represents the probability that a break at 11p1 is the second event given that the first event was a break at 1p3 and that a second event occurred. The other extra parameter represents the probability that 1p3 was the second event given that the first event was a break at 11p1 and that a second event occurred. Using these two extra parameters improves the fit substantially; the total sum of squares of discrepancies is reduced to 49.24. The parameter estimates for the resulting model are shown in Table .

The model parameters indicate that 1p3 is the most likely first event, followed by 11p1 and 11q2. Among the events in Set II, 1p3 is the first breakpoint to occur 26% of the time. Breaks in regions 1p3 and 11p1 differ from those in the other regions in that they occur predominantly as either first or second events. Using the parameters in Table 2, we can predict that of the 52 tumors containing a breakpoint in 1p3, the break occurs as the first event 53% of the time and as the second event 41% of the time. For the 40 tumors containing a break in 11p1, the model predicts that the 11p1 break occurred as a first event 49% of the time and as a second event 43% of the time. Given that a break in 1p3 occurred as the first event, the model predicts a 79% chance that 11p1 will be the second event among events in Set II, if there is a second event. If a break occurs in 11p1 as the first event in Set II, the model predicts a 90% chance that 1p3 will be the second event, if there is a second event. Hence, this model supports the view that 1p3 and 11p1 are relatively early linked events. This model is quite consistent with the placement of 1p3 and 11p1 near the root in all 4 trees.

By using precise estimates of event probabilities, the DAG model gives more quantitative evidence of the early role of 1p3 and 11p1 breaks. In contrast, the tree models yield predictions about all the events, but with less detail.

Survival Analysis

It was reported in (Taetle et al, 1999b) that a breakpoint in regions 1p1 or 3p1 is associated with shortened survival based on univariate analysis of the 13 regions in Set I. A number of other regions were also prognostically significant at the nominal .05 level. We found that the number of breakpoint regions in Set I was also highly prognostic. We fit a Bayesian proportional hazards (BPH) model including all 13 breakpoint regions. The regression coefficient associated with a region in this model represents the logarithm of the hazard of death for a patient with a tumor containing a breakpoint in the specified region relative to a tumor for a patient without the specified breakpoint but with all other breakpoints the same.

For the BPH model, the posterior probability that a breakpoint in 3p1 is associated with poor prognosis is 0.985. This posterior probability is 0.88 for 1p1. The posterior probability that regions 1p3 and 11p1 are associated with poor prognosis is 0.94 for each region. We also fit a BPH model containing the 13 regions as well as stage, grade and residual disease. The posterior probability that a region was associated with poor prognosis within stage, grade and residual disease groups was 0.97 for 3p1, 0.96 for 1p1, 0.74 for 1p3, and 0.93 for 11p1.

These analyses support the previous observations that breaks in 1p1 and 3p1 are associated with poor prognosis and indicate that these observations were not the result of multiple comparisons, and that the effects are not the result of associations with the prognostic effect of number of regions in Set I containing breaks. The BPH analyses also provide evidence that breaks in 1p1 and 3p1 are associated with poor prognosis independently of stage, grade, and residual disease. These analyses suggest that breakpoints in regions 1p3 and 11p1 are also associated with poor prognosis in a specific way and not just by association with number of regions containing breaks. The prognostic effect of 1p3 is reduced, however, after adjustment for stage, grade and residual disease.

Relationship to Clinical Variables

We evaluated whether presence of any of the 13 breakpoint regions is individually associated with any of the 3 clinical variables: stage, grade, and diameter of residual tumor. For each variable we divided the possible values into two categories as shown in Table . Only 2 of the 39 Chi-squared tests (1q2 and 6q1 associated with smaller residual disease) showed significance at the nominal 0.05 level, and none of the results is significant after correcting for multiple tests. Our interpretation, consistent with the theme of this entire paper, is that the patterns of genetic changes in ovarian adenocarcinoma are too heterogeneous and complex to yield clear results by basic univariate analysis with contingency tables.

DISCUSSION

We presented many mathematical analyses of the data in (Taetle et al., 1999a, 1999b) on chromosomal breakpoints in a set of 244 adenocarcinomas of which 135 have at least one breakpoint. As discussed in those two papers, the analysis of single breakpoints is reasonably consistent with earlier smaller studies (Pejovic et al., 1992; Jenkins et al., 1993). Our goal in this paper has been to present more global analyses that consider the co-occurrences of multiple breakpoints. Except for analysis of translocations that involve 2 breakpoints at a time, the previous papers did not attempt to model more than 1 breakpoint at a time. Some of the tumors we analyzed do have genetic changes other than breakpoints, such as chromosomal imbalances. However, the authors of (Taetle 1999a, 1999b) recognized that none of the non-breakpoint events occurred in more than a few tumors, so a statistically meaningful analysis including the non-breakpoint events seems infeasible. For consistency with (Taetle 1999a, 1999b) and so as to highlight our new methods we did our analysis on the same breakpoint data used in those two papers.

Our analyses worked with two different sets of breakpoint regions that could be considered as non-random. The two sets reflect different approaches to identifying which events should be of interest. The method of Brodeur et al. (1982) unconditionally selects those events that occur more often than would be expected by chance. The method of Taetle et al. (1999a) selects events conditionally based on which events have been selected already. Thus the latter method will be more inclined to select at least one event from most tumors, and it may tend to select events that occur in a smaller subset of the tumors. We found that the method of Brodeur led to distance-based trees that fit the data more closely.

We sought to introduce to the cancer genetics community techniques from theoretical computer science, phylogeny, and statistics that are relevant to modeling various types of cancer. We have developed freely available software that can be used to produce either branching trees or distance-based trees. The software is available by sending email to **r.desper@dfkz-heidelberg.de** or **schaffer@helix.nih.gov**. The software was designed to analyze chromosomal gain and loss aberrations detected by comparative genome hybridization (Kallioniemi et al., 1992). However, the notion of what constitutes an "aberration" is encoded in a sufficiently abstract manner that the software is readily amenable to analyzing chromosomal breakpoint aberrations instead.

It is widely believed that it takes an accumulation of genetic changes to cause solid tumor development. Therefore, it is important to identify patterns of genetic changes that occur together, and genetic changes that occur early in tumor formation. The breakpoints analyzed here represent large genetic alterations that may affect expression of many genes. Our tree models of these breakpoints suggest areas of the genome in which genes which cooperate to induce ovarian cancers. This information on regions and relationships of breakpoint events can be used in conjunction with results from techniques such as microarrays (DeRisi et al. 1996) or SAGE (Zhang et al. 1997) that evaluate gene expression levels of genes across the genome. Most of the cancer genetics literature has investigated markers by looking at single recurring events. One notable exception is the work of Vogelstein's group on modeling colorectal cancer (Vogelstein et al., 1988; Fearon and Vogelstein 1990).

For any pair of genetic events, the path model represents the idea that one event is temporally prior to or after the other event in all tumors. The branching tree model is more general in that it provides for two events to be on different paths, and thus one may occur without the other and without a temporal orientation. However, in the branching tree there is only a single path from the root r to any other event vertex. The branching tree model may still be too restrictive to model the true sequences of genetic changes because it is possible that the paths of changes should converge, so that a later change such as the breakpoint in 3p1 might be modeled as occurring after several different sets of earlier changes.

The natural generalization of the branching tree model is the directed acyclic graph (DAG) model that allows the paths of changes to converge. In a DAG, edges are still directed away from the root, so there cannot be any feedback, but there may be multiple paths from the root to any vertex. A principal difficulty with DAG models is that DAGs may have many more edges than trees, and consequently there may be many more parameters to estimate. One must be careful not to fall into the trap of overfitting the data.

In this paper we demonstrated seven distinct multivariate techniques applied to a single ovarian cancer data set: 1) tests of independence, 2) branchings in graphs, 3) phylogenetic tree construction, 4) likelihood analysis of tree models, 5) principal components analysis, 6) directed acyclic graph models, and 7) Bayesian survival analysis. No one of these techniques is a magic bullet, but together they permit the formulation and validation of many hypotheses. In the canonical uses of phylogenetic tree construction in biology, the leaves represent living species and the internal nodes represent hypothetical common ancestors. These properties do not hold in our application of phylogenetic tree construction to cancer data; the tree does not provide a direct trace of the evolutionary history of a tumor as in the DAG models. Nevertheless, the oncogenic tree models represent one of *many* different applications of fitting a distance-based tree to a matrix of distances (Barthélemy and Guénoche, 1991).

It would be nice to reduce the number of techniques needed for convenience, but multiple techniques are needed since we are trying to answer multiple questions: (1) are the events dependent? (2) which events tend to occur early? (3) which events tend to occur together? (4) which events are associated with poor survival? We recommend to start with tests of independence to directly address the first question. The two tree building methods are overlapping attempts to suggest hypotheses for the second and third questions. We prefer to use both methods, so as to boost confidence when they make consistent predictions, and raise immediate doubts when they make inconsistent predictions. The tree models have the limitation that they do not give a very useful measure of how much better the best tree is than the next best; this is provided by the likelihood analysis, The partial directed acyclic graph (DAG) models provide a more direct and detailed representation of the sequence of genetic events than the tree models The DAG models have far too many parameters in general, and so we were able to use them effectively only to confirm and strengthen the branching tree predictions about 1p3 and 11p1, but not the other events. In general, the restricted DAG models serve to test predictions about 1 or 2 events much more than to make predictions about the entire data set. The principal components analysis is less essential but provides useful confirmatory evidence. Survival analysis is a different problem requiring one or more different methods. The trees and principal components analysis can suggest hypotheses to test with survival analysis, so as to diminish the amount of multiple testing.

There are several important statistical problems in biomedicine such as phylogenetic tree construction, sequence homology searching, genetic linkage analysis, and protein structure prediction for which no single technique is clearly best and using multiple techniques to confirm any questionable conclusions is highly recommended. Therefore, it should not be surprising that we are proposing combined use of multiple techniques for evaluating models of oncogenesis.

In the ovarian adenocarcinoma data set, the seven methods lead to strikingly consistent conclusions, including: (1) The non-random breakpoints in ovarian adenocarcinoma do not occur independently. (2) Breakpoints in 1p3 and 11p1 are important linked early events and distinguish a class of tumors associated with poor prognosis. (3) The breakpoint in 1p3 appears to occur before 11p1 frequently. (4) The events 1p1, 3p1, 1q2 distinguish a class of ovarian tumors, and the events 1p1 and 3p1 are associated with poor prognosis.

APPENDIX

Test of Independence

Several of the models developed here are based on pairwise dependence between breakpoints in two bands. Although the methods developed here are exploratory, we wished to utilize a rigorous statistical significance test of the (null) hypothesis of complete independence among breaks in different regions. A test of the independence hypothesis is complicated for two reasons. Firstly, there are many types of multi-way interactions possible, and one cannot test for all of them and retain substantial statistical power. Secondly, many tests of independence are based on large sample distributions and the resulting significance levels may not be accurate for the data at hand.

We developed a randomization test of independence which is sensitive to pairwise dependencies. The test statistic is the smallest p-value for Fisher's exact test among all 2×2 contingency tables determined by pairs of breakpoints. We simulated the distribution of this test statistic under the null hypothesis of independence in the following way. We estimated the marginal frequency of a break in each region. We then simulated a breakpoint profile for each patient using these marginal frequencies. We held fixed the number of breakpoints observed in the actual data for each patient. Having simulated a breakpoint profile for each patient under this independence model, we computed the test statistic. We repeated this simulation 10,000 times to estimate the null distribution of the test statistic. The significance level for the test was the proportion of the null distribution with a value of the test statistic less than or equal to that calculated from the actual data. Because the modeling was based on tumors with at least one breakpoint in the set of interest, we performed the randomization test in the same way.

Optimal Trees

Our methods of inferring tree models use basic probabilities of event (co)occurrences that can be directly estimated from the breakpoint data. To relate the estimated probabilities of events to a tree, we consider a probabilistic model that encompasses a large space of all possible trees. From this large space our method selects one *optimal* tree, the one that fits best the breakpoint data. We have proved elsewhere (Desper et al., 1999; Desper et al., submitted) that each of our methods for choosing a tree converges in a formal sense to the correct tree, given enough samples. However, the theoretical guarantees should be viewed cautiously. The convergence results in (Desper et al., 1999; Desper et al., submitted) are asymptotic and some cautions about applying the asymptotic results to finite data sets are discussed in (Desper et al. 1999).

A *rooted tree*, or simply *tree*, on V is a triple $T = (V, E, r)$. The elements of V are called *vertices* and they are events (breakpoints) or the special vertex r. The elements of E are called *edges* and they are ordered pairs of (vertex) events. The item r is a special vertex called the *root*. A tree satisfies the conditions that:

1. For each vertex $v \in V$ there is at most one edge $u \to v \in E$ with v as its second component,

2. There is no edge $u \to r$ entering r,

3. There is no *cycle*, that is, no sequence of edges in E of the form $v_0 \to v_1 \to v_2 \cdots, v_{k-1} \to v_k \to v_0$,

4. For each $v \in V$, there is a sequence of edges $\mathcal{P}_v = r \to v_1 \to v_2 \cdots \to v_k$, such that $v_k = v$, and $v_{i-1}, v_i \in V$ for all i.

The vertices that have no outgoing edge are called *leaves*.

In our branching model, each vertex represents a breakpoint region. In our distance-based model, each leaf represents a breakpoint region. In our branching tree, the probability that a tumor contains a breakpoint in a region represented by a vertex v depends only upon the presence or absence of breakpoints represented by vertices upstream of v. Consequently we think of the upstream events as occurring earlier than the downstream events. The probability of obtaining a breakpoint in a region represented by a leaf vertex directly connected to the root r is independent of the occurrence of breakpoints in other regions.

Our method of selecting a branching tree starts with a *weighted complete graph* with vertex set V and two directed edges, one in each direction, between each pair of vertices. If there are k vertices in V, then the weighted complete graph contains $k(k-1)$ edges.

To obtain a tree model we assign to each of the $k(k-1)$ edges a weight based on the occurrences of the two events at its ends as explained in (Desper et al. 1999) Then we select a subset of $k-1$ edges such that:

1. there is one root to the tree, so that all edges are directed away from the root
2. the root of the tree is r
3. there is exactly one path from the root to each vertex
4. the sum of the edge weights is maximized

The number of possible trees satisfying properties 1,2,3 grows exponentially as a function of k, and is far too large to search exhaustively to find the maximum-weight tree. Nevertheless, Edmonds (1967) discovered a method that finds the tree satisfying properties 1,3,4 and takes time that grows as polynomial in k. Others (Chu and Liu 1965; Karp 1971; Tarjan 1977) improved the method, so that the time grows proportional to k^2, a bound which is completely practical for the number of events considered in this application. A directed tree selected from an edge-weighted graph to satisfy properties 1,3,4 is called a *maximum-weight branching* in the computer science literature, so we call oncogenetic trees derived by this method *branching trees*.

Phylogenetic Methods.

Our distance based trees are derived from known solutions to a basic problem in phylogeny called the *numerical taxonomy problem* (Swofford and Olsen, 1990; Barthélemy and Guénoche, 1991) that is similar to the problem of inferring oncogenetic trees. In the numerical taxonomy problem, we start with a set S of living species and a square, symmetric non-negative matrix D, such that D_{ij} is the *distance* between species i and j. The desired output is a phylogenetic tree T in which:

1. each species is represented by a leaf
2. there are distances on the edges
3. the tree distance T_{ij} is the sum of the distances on the edges between i and j.
4. the tree distances are a good approximation to distances specified in the input matrix D.

We use (Desper et al. submitted) the function:

$$D_{ij} = -2 \log \hat{p}_{ij} + \log p_i + \log p_j$$

for the distance between the vertex representing the breakpoint regions i and j. The distances along sequences of edges can be added. We can use the tree fitting algorithm as summarized below (Desper et al. submitted).

1. The Fitch (Fitch and Margoliash, 1967) and Neighbor (Saitou and Nei, 1987) programs from the PHYLIP (version 3.5c) package (Felsenstein, 1989) were used on the distance matrix to obtain initial topologies.

2. A linear program was used upon these topologies to find optimal edge-lengths. Often, this step will return some zero-length edges. If this happens, nearest-neighbor interchanges (NNIs) would be tested across each non-pendant zero-length edge, with edge-lengths re-optimized. With this data set, no zero-length edges were found by the linear program. A NNI test across the shortest edges indicated solution trees were local optima.

Likelihood Ratios for Branching Trees

A branching tree model assumes that the probability of an event represented by a given vertex depends only on the presence or absence of events represented upstream of that vertex. We have assumed here that this probability depends only on the presence or absence of the event represented by the immediate predecessor vertex. This assumption is necessary for likelihood calculations for datasets of limited size. The likelihood function of the data is thus the product over the vertices of the probability of the observed number of breakpoints in the region represented by the vertex conditional on the number of breakpoints in the region represented by the predecessor vertex.

Let n_1^i and n_0^i respectively denote the number of tumors with and without breaks in the region represented by vertex i. Let $n_1^{p(i)}$ and $n_0^{p(i)}$ respectively denote the number of tumors with and without breaks in the region represented by parent $p(i)$ of vertex i in the branching tree. Then the likelihood L of a branching tree can be written mathematically as:

$$L = \prod_i Pr[n_1^i, n_0^i \mid n_1^{p(i)}, n_0^{p(i)}],$$

where the product is over all of the vertices other than the root.

For the important special case of the star tree, where all breakpoints are independent, the likelihood is similar but the conditional probabilities are replaced by unconditional probabilities. We maximize the likelihood for the star model by using the observed marginal frequency of occurrence of a breakpoint in each region. We maximize the likelihood for other models by using these same frequencies for regions represented by vertices directly attached to the root, but by using the observed conditional probabilities for other vertices.

Let n_{01}^i, n_{00}^i, n_{10}^i, and n_{11}^i denote respectively: the number of tumors with a breakpoint in the region represented by i but not in the region represented by $p(i)$, the number of tumors without breakpoints in either region, the number of tumors without a breakpoint in the region represented by vertex i but with a breakpoint in the predecessor region and the number of tumors with breakpoints in both regions. Let n denote the total number of tumors with breakpoints in the set being considered. It can be shown that the ratio of maximized likelihoods for a general branching tree to the maximum likelihood for the star model can be written as

$$LR = \prod_i \frac{[(n_{01}^i)^{n_{01}^i}(n_{00}^i)^{n_{00}^i}(n_{11}^i)^{n_{11}^i}(n_{10}^i)^{n_{10}^i} n^n]}{[(n_0^{p(i)})^{n_0^{p(i)}}(n_1^{p(i)})^{n_1^{p(i)}}(n_0^i)^{n_0^i}(n_1^i)^{n_1^i}]}$$

This product is over all vertices i not adjacent to the root. From this expression for the ratio of maximized likelihoods for any branching tree model relative to the star model, we can compute the ratio of maximized likelihoods for any two branching tree models.

Directed Acyclic Graph Models

The natural generalization of tree models are directed acyclic graph (DAG) models. A directed acyclic graph is a triple $D = (V, E, r)$ satisfying conditions 2,3, and 4 of the definition of a rooted tree. All of the edges in a DAG are still directed away from the root. The DAG, however, provides for multiple paths from the root r to other vertices, i.e. the path specified in condition 4 need not be unique. We use a type of DAG, which we call a *complete level DAG*, to represent the 6 events of Set II. Each vertex v in the DAG represents a *subset* $s(v)$ of Set II. Since Set II has 6 events, there are $2^6 = 64$ vertices in the full DAG. Each vertex v is assigned a level from $0, 1, \ldots, 6$ according to the number of events in $s(v)$. Like the trees described previously, the DAG has a root vertex r, which represents the empty set, and is the only vertex on level 0.

There are 6 vertices on level 1 representing the 6 possible breakpoint regions that may occur as first events among Set II . There is an edge $r \to \{i\}$ from the root to each vertex on level 1, representing the single event i. Associated with this is edge is probability $p_i(1)$ that i occurs as the first genetic alteration.

There are 15 vertices in level 2 corresponding to the 15 pairs of events in Set II. For each vertex $\{i\}$ on level 1, and each vertex $\{i,j\}$, $i \neq j$, on level 2, there is an edge $\{i\} \to \{i,j\}$ representing the possibility that i occurs first and j occurs second. Associated with this edge is $p_{i,j}(2)$, which is the probability that event j occurs second in a tumor that already has a breakpoint at i. The sets are unordered, but the probability subscripts are ordered. For example, there will be an edge $\{7p2\} \to \{7p2, 11p1\}$ with probability $p_{7p2,11p1}(2)$ and an edge $\{11p1\} \to \{7p2, 11p1\}$ with probability $p_{11p1,7p2}(2)$. The first edge represents tumors where a breakpoint in 7p2 occurs first, and a breakpoint in 11p1 occurs second. The second edge represents tumors where a breakpoint in 11p1 occurs first, and a breakpoint in 7p2 occurs second.

It is, of course, possible that there is no second event. We denote the probability of having the process terminate with only a breakpoint in region i by $p_{i,0}(2)$. We use $p_0(2)$ to denote the probability that a tumor has no second event. The vertices at later levels and the associated edges can be defined analogously.

Since we are most interested in showing dependence among the breakpoint events, we begin by fitting the graph model based on the assumption of complete independence, and examine the adequacy of fit of the model to the data. If the fit is adequate, we go no further. If, however, the fit is inadequate, our modeling strategy is to introduce parameters expressing dependence between event pairs that best improve the fit. This modeling strategy is analogous to that used in step-wise regression.

The model of event independence includes the 6 first event probabilities $p_i(1)$ described above. We denote these $p_{1p1}, p_{1q1}, p_{1p3}, p_{7p2}, p_{11p1}, p_{11q2}$ and suppress the parenthesis. We exclude data from tumors without any breakpoints in Set II. For the independence model we assume that:

$$p_{i,j}(2) = \frac{(1 - p_0(2))p_j}{1 - p_i},$$

for all events pairs i, j. This formula encodes the assumption that the probability of having the second event be j conditional on the first event having been i equals the product of two probabilities. The first term in the product is the probability that there is a second breakpoint, which is $(1 - p_0(2))$, and is assumed to be independent the first event to occur. The second term in the product is the conditional probability that the second breakpoint is in region j given that the first event was i and that a second event occurs. The second term is computed using the same event probabilities as were used for first events adjusted for the fact that a break has already occurred in region i. This independence model assumes that the first stage events do not influence which second stage events are favored. Levels 3,4,5,6 can be modeled the similarly.

The genetic independence model is quite tractable as it contains only $2n - 2$ independent probability parameters for n breakpoint regions. There are $n - 1$ parameters for the n single event probabilities, and one parameter for each level $2, 3, \ldots n - 1$, representing the probability that no further events occur.

We used the nlminb function of the S-Plus (Becker et al., 1988) statistical package to find the parameter values that maximize the likelihood function. The likelihood function can be written:

$$\prod_v [\sum_{\mu \in P(v)} Pr(\mu)]^{m(s(v))}$$

The product is over all the vertices of the graph. For each vertex v, the sum is over the paths from the root terminating in vertex v. The elements of the summation are the product of the edge probabilities along each path μ leading from the root and terminating in the vertex v. The exponent $m(s(v))$ denotes the number of tumors having all the breakpoints in the set $s(v)$, represented by vertex v, and without any other breakpoints in Set II.

Survival Analysis

The BPH model (Faraggi and Simon 1997) provides a method for controlling the "multiple comparison problem" in survival analysis. The BPH model is based on the assumption that each regression coefficient is *a priori* most likely to be 0, and that small effects are more likely than large effects. The prior distribution used for the regression coefficients is a Gaussian distribution with mean 0 and variance σ^2. The value of the variance determines the strength of the multiple comparison adjustment. The Bayesian analysis shrinks each regression coefficient towards zero, the degree of shrinkage depending on the amount of data available for estimating the value of the regression coefficient. If very large values of σ^2 are used, then BPH becomes equivalent to the non-Bayesian proportional hazards model. We used the value $\sigma^2 = 0.5$ as suggested by Faraggi and Simon (1997) for models containing the breakpoint bands in Set I and $\sigma^2 = 2$ for stage, grade and residual disease.

ACKNOWLEDGEMENTS

Special thanks to Martin Farach-Colton for his assistance as Richard Desper's Ph.D. advisor at Rutgers University. Thanks to Olli-P. Kallioniemi for posing the problem inferring tree models of oncogenesis and for many helpful discussions. Thanks to Leslie Biesecker for referring us to the tables in (Daniel 1995) and for putting us in contact with Dr. Daniel to fill in the X-chromosome band sizes.

REFERENCES

Barthélemy J-P, Guénoche A. 1991. Trees and Proximity Representations. New York:Wiley.

Becker RA, Chambers JM, Wilks AR. 1988. The New S Language. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Brodeur GM, Tsiatis AA, Williams DL, Luthardt FW, Green AA. 1982. Statistical analysis of cytogenetic abnormalities in human cancer cells. Cancer Genet Cytogenet 7:137–152.

Chu YJ Liu TH. 1965. On the shortest arborescence of a directed graph. Sci Sinica 14:1396–1400.

Daniel A. 1985. The size of prometaphase chromosome segments. Clin Genet 28:216–224.

DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nature Genet 14:457–460

Desper R, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH, Schäffer AA. 1999. Inferring tree models for oncogenesis from comparative genomic hybridization data. J Comp Biol 6:37–51.

Edmonds J. 1967. Optimum branchings. J Res Nat Bur Stand 71B:233–240.

Faraggi D, Simon R. 1997. Large sample Bayesian inference on the parameters of the proportional hazard models. Stat Med 16:2573–85.

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

Felsenstein J. 1989. PHYLIP – Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166. See also http://evolution.genetics.washington.edu/phylip.html.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. Science 155:279–284.

ISCN. (1995) International System for Human Cytogenetic Nomenclature. Mitelman F, ed. Basel:S. Karger.

Jenkins RB, Bartelt D Jr, Stalboerger P, Persons D, Dahl RJ, Podratz K, Keeney G, Hartmann L. 1993. Cytogenetic studies of epithelial ovarian carcinoma. Cancer Genet Cytogenet 71:76–86.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258:818–821.

Karp RM. 1971. A simple derivation of Edmonds' algorithm for optimum branching. Networks 1:265–272.

Kuukasjärvi T, Karhu R, Tanner M, Kähkönen M, Schäffer A, Nupponnen N, Pennanen S, Kallioniemi A, Kallioniemi OP, Isola J. 1997. Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. Cancer Res 57:1597–1604.

Morrison DF. 1967. Multivariate Statistical Methods. New York, NY: McGraw–Hill.

Newton MA, Wu SQ, Reznikoff CA. 1994. Assessing the significance of chromosome-loss data: where are suppressor genes for bladder cancer. Stat Med 13:839–858.

Pejovic T, Heim S, Mandahl N, Baldetorp B, Elmfors B, Flodérus UM, Furgyik S, Helm G, Himmelmann A, Willén H, Mitelman F. 1992. Chromosome aberrations in 35 primary ovarian carcinomas. Genes Chromosomes Cancer 4:58–68.

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–424.

Swofford DL, Olsen GJ. 1990. Phylogeny reconstruction. In Hillis, D. M., and Moritz, C. eds. Molecular Systematics, 411–501. Sunderland, MA: Sinauer Associates Inc.

Taetle R, Aickin M, Yang JM, Panda L, Emerson J, Roe D, Adair L, Thompson F, Liu Y, Wisner L, Davis JR, Trent J, Alberts DS. 1999a. Chromosome abnormalities in ovarian adenocarcinoma: I. Non-random chromosome abnormalities from 244 cases. Genes Chromosomes Cancer 25:290-300.

Taetle R, Aickin M, Panda L, Emerson J, Roe D, Thompson F, Davis J, Trent J, Alberts D. 1999b. Chromosome abnormalities in ovarian adenocarcinoma: II. Prognostic impact of nonrandom chromosome abnormalities in 244 cases. Genes Chromosomes Cancer 25:46–52.

Tarjan RE. 1977. Finding optimum branchings. Networks 7:25–35.

Thompson FH, Liu Y, Emerson J, Weinstein R, Makar R, Trent JM, Taetle R, Alberts DS. 1994. Simple numeric abnormalities as primary karyotype changes in ovarian carcinoma. Genes Chromosomes Cancer 10:262–266.

Thompson FH, Taetle R, Trent JM, Liu Y, Massey-Brown K, Scott KM, Weinstein RS, Emerson JC, Alberts DS, Nelson MA. 1997. Band 1p36 abnormalities and t(1;17) in ovarian carcinoma. Cancer Genet Cytogenet 96:106–110.

Vogelstein B, Fearon E, Hamilton S, Kern S, Preisinger A, Leppert M, Nakamura Y, White R, Smits A, Bos J. 1988. Genetic alterations during colorectal-tumor development. N Engl J Med 319:525–532.

Zhang L, Zhou W, Velculescu V, Kern S, Hruban R, Hamilton S, Vogelstein B, Kinzler K. 1997. Gene expression profiles in normal and cancer cells. Science 276:1268–1272.

Table 1. Coefficients in the first 3 principal components for each of the 6 elements of Set II.

Region	PC-1	PC-2	PC-3
1p1	-0.21	0.20	-0.61
1p3	0.66	-0.46	-0.50
1q1	-0.26	0.30	-0.03
7p2	0.03	-0.28	0.34
11p1	0.64	0.48	0.40
11q2	-0.21	-0.59	0.33

Table 2. Parameters and estimated values for the two-level DAG model, using events in Set II.

Parameter	Maximum Likelihood Estimate
p_{1p1}	0.14
p_{1p3}	0.26
p_{1q1}	0.10
p_{7p2}	0.11
p_{11p1}	0.23
p_{11q2}	0.16
$p_0(1)$	0.43
$p_0(2)$	0.63
$p_0(3)$	0.86
$p_0(4)$	1.00
$p_{\mathrm{1p3,11p1}}$	0.79
$p_{\mathrm{11p1,1p3}}$	0.90

Table 3. Number (percent) of tumors containing a break in chromosomal bands in Set I classified by clinical stage, clinical grade, and diameter of residual disease. * indicates nominal $p < 0.05$ by contingency Chi-square test with continuity correction. 1 tumor missing the stage information and 12 tumors missing the residual disease information were excluded in analyzing those variables.

Breakpoint	Stage		Grade		Residual Disease	
Region	1-2	3-4	mod–well	poor	<1cm	$\geq$ 1cm
1p1	3(11%)	20(9%)	7(8%)	17(11%)	14(10%)	9(9%)
1q1	1(4%)	16(7%)	10(12%)	7(4%)	11(8%)	5(5%)
1p2	1(4%)	12(6%)	2(2%)	11(7%)	5(4%)	8(8%)
1q2	0(0%)	20(9%)	4(5%)	16(10%)	*6(4%)	13(13%)
1p3	4(15%)	47(22%)	15(17%)	37(23%)	24(18%)	25(26%)
1q4	3(11%)	15(7%)	5(6%)	13(8%)	9(7%)	8(8%)
3p1	0(0%)	14(6%)	5(6%)	10(6%)	6(4%)	8(8%)
6q1	2(7%)	20(9%)	4(5%)	18(11%)	*7(5%)	14(14%)
7p1	2(7%)	16(7%)	4(5%)	14(9%)	12(9%)	6(6%)
7p2	2(7%)	17(8%)	4(5%)	15(9%)	9(7%)	10(10%)
11p1	3(11%)	36(17%)	14(16%)	26(16%)	21(16%)	16(16%)
11q2	2(7%)	24(11%)	9(10%)	17(11%)	14(10%)	11(11%)
12q2	4(15%)	15(7%)	4(5%)	15(9%)	9(7%)	10(10%)
Cases	27	216	86	158	134	98

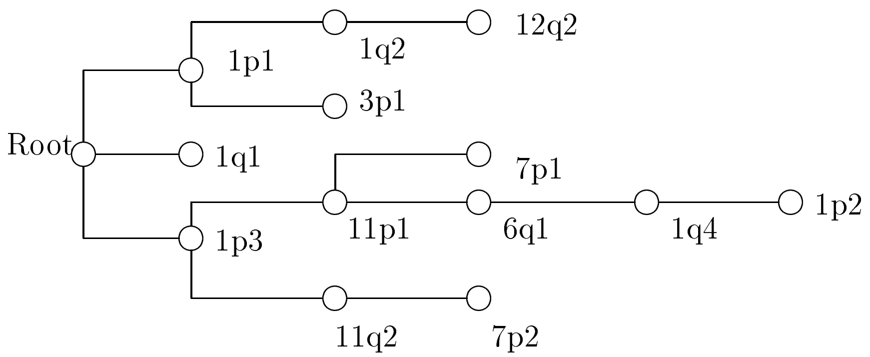

Figure 1: Branching tree for Set I

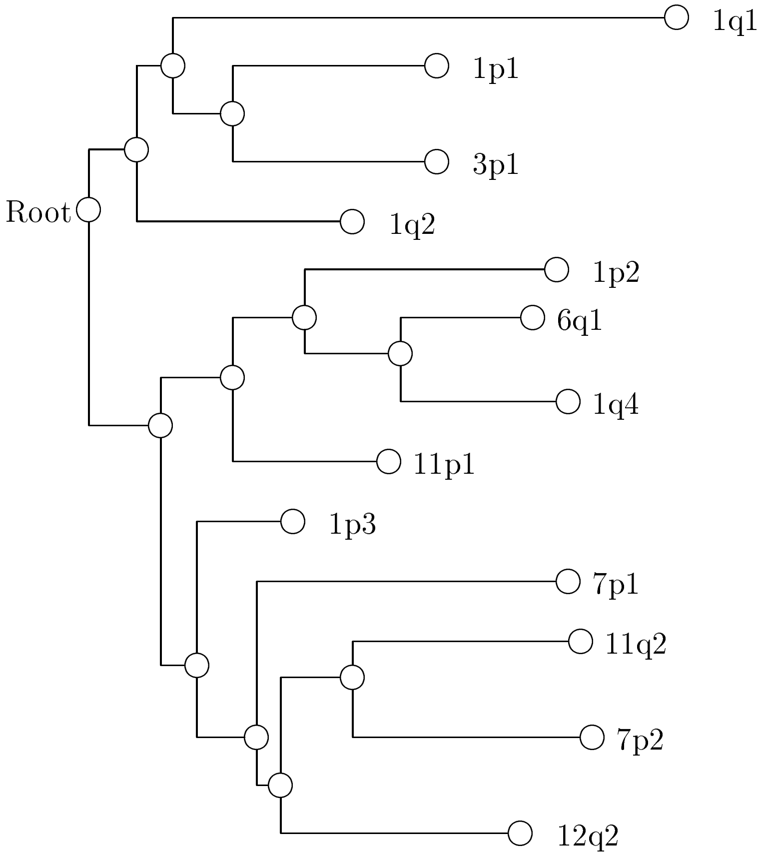

Figure 2: Distance-based tree for Set I

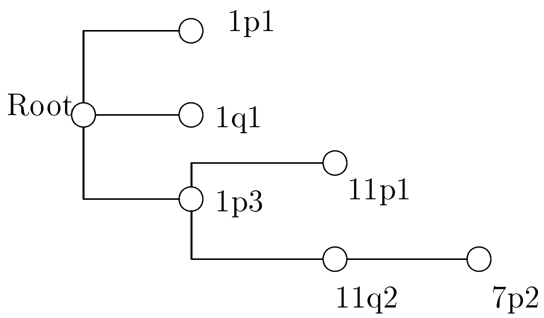

Figure 3: Branching tree for Set II

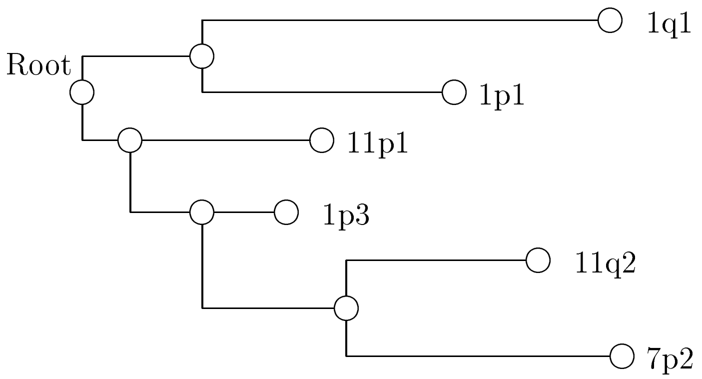

Figure 4: Distance-based tree for Set II