BAYESIAN STATISTICS 7, pp. 293–305
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2003

A Statistical Approach to Modeling Genomic Aberrations in Cancer Cells

M. A. NEWTON H. YANG
University of Wisconsin, Madison, USA
newton@stat.wisc.edu

P. GORMAN I. TOMLINSON R. ROYLANCE
Cancer Research, UK

SUMMARY

Whereas most cells in the body carry the normal complement of 23 chromosome pairs, the cells within a cancerous tumor very often present highly abnormal genomic structure. Deletions, amplifications, rearrangements and mutations are common at various scales and are highly variable amongst tumors, as indicated by molecular technologies which enable ever better measurement. It is an important statistical problem to separate sporadic abnormalities from those that may not be sporadic and that may have some biological significance. We discuss a modeling strategy for genomic aberration data which allows us to to infer combinations of aberrations that together increase the chance that a precancerous cell will have a descendant tumor lineage. The likelihood component involves a network of pathway structures. Markov chain Monte Carlo is used to sample from the space of these oncogenic networks. We illustrate the methodology with comparative genomic hybridizations from a recent breast cancer study, and we derive a likelihood formula for the larger class of tree-like networks.

Keywords: ONCOGENIC PATHWAYS; GENETIC NETWORKS; MODEL-BASED CLUSTERING; INSTABILITY AND SELECTION.

1. BACKGROUND

In much the same way the detective uses clues to infer what has happened at a crime scene, the cancer biologist attempts to identify critical genomic changes that have carried a normal cell to its highly aberrant state in a cancerous tumor. The extent and variety of genomic aberration in tumor cells are striking (e.g., see Knuutila, *et al.*, 1998, 1999, for surveys of cytogenetic aberrations), and support a statistical approach to data analysis. Lengauer, Kinzler, and Vogelstein (1998) comment that, "... all tumors are genetically unstable. Instability is the engine of both tumor progression and tumor heterogeneity, guaranteeing that no two tumors are exactly alike and that no single tumor is composed of genetically identical cells." Allowing that genetic instability somehow creates aberrations, it is equally important to know that a tumor cell lineage carrying some profile of aberrations is subjected to selective pressures which affect its fate (e.g., Tomlinson, *et al.*, 1996).

The concepts of instability and selection guide a general understanding of experimental results from cancer biology, but they also provide a framework for deriving probability models and statistical methodology to analyze measured aberrations (Newton, *et al*, 1998, 1999, 2000). Briefly, the probability distribution for measured aberrations is built in two steps: first, one postulates random genomic damage in a progenitor cell – a cell that could become ancestral to observable tumor cells; second, one allows that certain damage (such as deletion of a suppressor gene) is beneficial to the tumor and increases the probability of selection, that is, the probability that descendents of the progenitor cell will populate an observable tumor. Interestingly, Bayes rule is used to derive the sampling distribution of the data since we measure aberrations conditionally on selection having occurred. The resulting statistical methods have been used to characterize significant changes in bladder cancer (Yeager, *et al.* 1998), prostate cancer (Jarrard, *et al.* 1999), hepatocellular carcinoma (Teeguarden, *et al.* 2000), colon cancer (Shoemaker, *et al.* 1998), and breast cancer (Haag, *et al.* 1996). A limitation of the methodology used in those studies is that it did not account for statistical dependence among unlinked genomic aberrations. More specifically, it did not deal with the fact that the co-occurrence of several aberrations could increase the probability of selection. The term *oncogenic pathway* has been used to describe such a combination of aberrations that are relevant to oncogenesis, and this concept is the focus of our present effort.

The importance of computational schemes to identify oncogenic pathways is well recognized, and the available tree-based methods provide the first quantitative, multivariate approach (Desper, *et al.*, 1999, 2000; Jiang *et al.*, 2000). From the point of view of statistical inference, however, the available methods have several deficiencies. Notably, they rely on a high degree of initial data processing and they provide point estimates but no measures of confidence. In very recent work, it has been recognized that the instability-selection framework can incorporate oncogenic pathways (Newton, 2001). This finding provides a statistical basis to inference about oncogenic pathways. Here we review the *instability-selection-network* model from Newton (2001) and we illustrate the methodology on data from a recent breast cancer study. We also report a methodological extension of the work in Newton (2001) by solving the problem of how to calculate likelihoods for a class of overlapping-pathway models.

2. ILLUSTRATION

To be any more precise it is helpful to have an example. Figure 1 shows genomic aberrations taken on a set of 40 grade II invasive ductal breast cancers. The *profile* from each tumor is a vector $x = (x_1, \ldots, x_n)$ with binary elements $x_i \in \{0, 1\}$ that indicate whether or not aberration i occurs in that tumor. The number n of potential aberrations has to do with the resolution of the measurements, and turns out to be 82 in this study. This counts 41 possible amplifications and 41 possible deletions. Each of the 41 chromosome arms in the genome may exhibit a deletion of genomic material, an amplification of material (or both). (Recall that there are 23 chromosome pairs a normal cell. Except for several acrocentric chromosomes, there are two arms for each one – the short p arm and the long q arm, thus we measure 41 different arms.) The data in Figure 1 are obtained using comparative genomic hybridization (CGH), a general approach to assess DNA copy number variations in tumor cells. (See the appendix for further details.)

One of the first things to see from the data in Figure 1 is the extent of marginal heterogeneity among aberrations. This variability is much larger than one expects by chance alone, for example, if these sample frequencies are identically distributed Binomial counts (calculation not shown). Extra copies of DNA on the long arm of chromosome one, i.e. $+1q$, is the most frequent aberration, occurring in 27 tumors. (On average, each aberration occurs in 5.5 tumors, and each tumor presents 11.2 aberrations.) Frequent aberrations are interesting for further study because they may be involved in the etiology of cancer. In considering oncogenic pathways, we are compelled to look beyond marginal frequencies towards sample correlations among

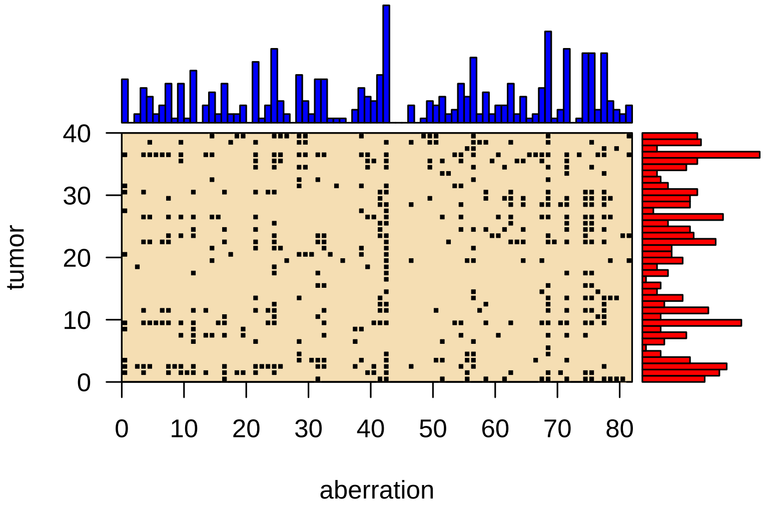

Figure 1. *Genomic aberrations (dark spots) in 40 grade II invasive ductal breast cancers (rows) for 82 potential aberrations (columns). The first 41 columns correspond to deletions on arms from* $1p$ *to* Xq*, and the next correspond to amplifications.*

aberrations. We performed a simple permutation test by shuffling within rows in Figure 1 and recomputing the sample covariance matrix among aberrations and we found that the extent of correlations is quite significant (calculation not shown).

One result of model-based computations described in the next section is a posterior probability that a given aberration resides on some oncogenic pathway, as opposed to being neutral. In the model, a neutral aberration is one for which its occurrence in a progenitor cell does not influence selection. We presume that if the aberration is not neutral then it is somehow relevant to oncogenesis and thus we should like to be able to compute the probability of this event. Figure 2 compares the posterior probability of being relevant to the marginal empirical frequency of occurrence. Some thirteen aberrations out of 82 are almost certainly relevant. Naturally, they tend to be the aberrations which occur most frequently, but the correspondence with empirical frequency is not perfect. For example, the deletion $-22q$ is probably relevant even though it occurs in only 8 of the 40 tumors; yet several other aberrations occur 10 times and they are probably neutral. Had we been processing only marginal frequencies then we would not expect this phenomenon; evidently, the way in which combinations of aberrations occur together is informative dependence.

Hierarchical clustering is a common tool in bioinformatics and provides a natural way to picture the dependencies between CGH aberrations. Figure 3 makes a comparison of two clusterings. Both use the default settings of the **hclust** function in **R** (Ihaka and Gentleman, 1996). In the first, we are clustering the raw profiles from Figure 1, using Euclidian distance between columns (aberrations) in Figure 1 and complete-linkage to build the tree. Perhaps not suprisingly, it is difficult to extract much information from the raw-data clustering. We find the second clustering rather more informative. Instead of clustering raw data, we are clustering certain posterior probabilities that come out of a Bayesian analysis of the data. More specifically, the distance between two aberrations is taken to be the posterior probability that they do not reside on a common oncogenic pathway, as calculated within the instability-selection-network model. A convenient feature of this clustering is that the many aberrations which are probably neutral (i.e., they are probably not on any oncogenic pathway), are compressed together on the right side of the plot. Dominant branches at height near 0 characterize collections of aberrations that probably constitute oncogenic pathways. This clustering provides one way to extract posterior summaries beyond point estimates of pathways. Furthermore, in our experience, the structure

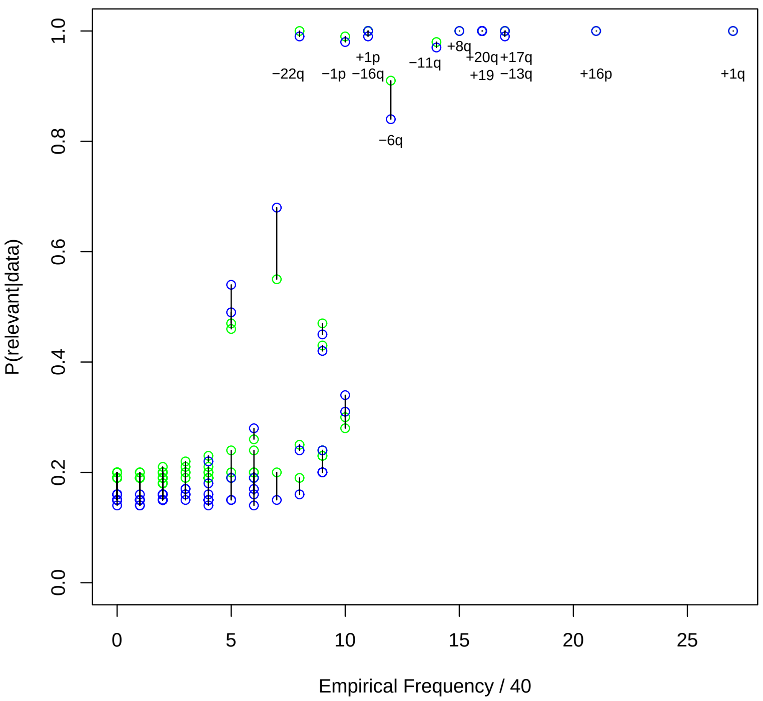

Figure 2. *Empirical aberration frequency and posterior probability that each aberration is relevant to oncogenesis. The posterior probabilities are computed twice, from independent MCMC runs, under the double Polya prior with $\tau = 5$. The level of Monte Carlo error is relatively low, as indicated by the fact that the two approximations (pairs of points connected by a line) are quite close.*

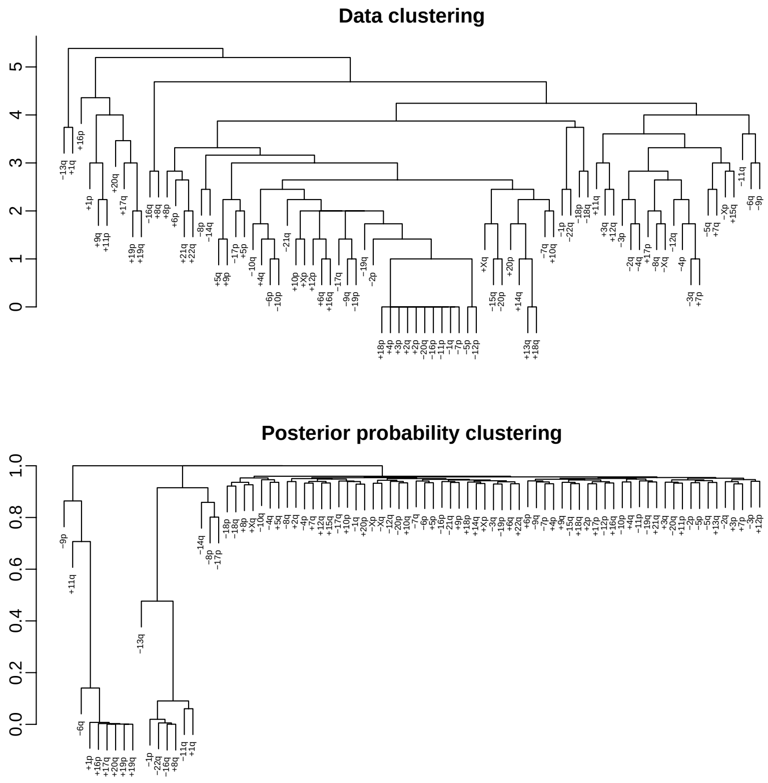

Figure 3. *Dependence among aberrations via hierarchical clustering.*

exposed in this clustering is rather insensitive to the prior distribution.

The summary calculations reviewed above are based on a specific probability model for measured aberrations, a prior distribution over the space of oncogenic pathways, and MCMC methods for posterior analysis. We consider these elements next.

3. A NETWORK MODEL

3.1. *Non-overlapping pathways*

Whether or not an aberration is neutral is unknown *a priori*. We represent these parameters with a vector $a = (a_1, \ldots, a_n)$ of binary indicators. The pathway structure amongst relevant (i.e., non-neutral) aberrations is represented in Newton (2001) as a set partition using a label vector $c = (c_1, \ldots, c_n)$. The ith label c_i has no meaning in isolation; but for relevant aberrations i and j, $c_i = c_j$ means that i and j reside on the same oncogenic pathway. An oncogenic pathway is a collection of relevant aberrations, and is also unknown. The assumption of non-overlapping pathways is not realistic but it leads to a feasible likelihood calculation.

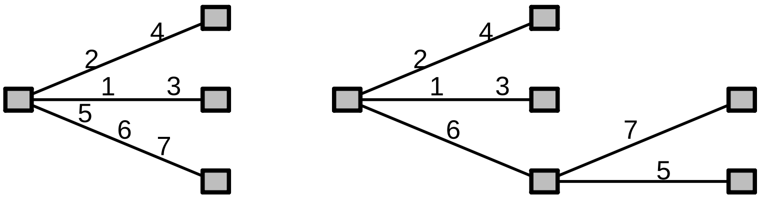

Figure 4. *Two hypothetical networks. The one on the left has three non-overlapping paths:* $\{2,4\}$, $\{1,3\}$ *and* $\{5,6,7\}$. *The one on the right has four paths:* $\{2,4\}$, $\{1,3\}$, $\{6,7\}$ *and* $\{5,6\}$. *The number of leaves on each tree corresponds to the number of paths. The node on the left in each case is the root.*

The pathway labels c_i for relevant aberrations ($a_i = 1$) describe some collection of pathways. For example, in Figure 4, the diagram on the left considers a hypothetical cancer comprised of three oncogenic pathways and seven relevant aberrations. In addition to measurable (i.e., overt) damage indicators x_i, the model envisions latent binary variables y_i that stand for covert aberrations which may be generated by genetic instability but which we are unable to measure. Genetic instability in the progenitor cell creates either overt damage $x_i = 1$ or covert damage $y_i = 1$ randomly: i.e. $x_i \sim_{iid}$Bernoulli(α) and $y_i \sim_{iid}$Bernoulli(β), for unknown rates α and β. We say that instability has opened a pathway if all the aberrations on that pathway have occurred either overtly or covertly. Thus, the probability that pathway e, say, is open is, $p_e = \theta^{n_e}$, where $\theta = 1-(1-\alpha)(1-\beta)$ is the chance that a given aberration occurs somehow, and n_e is the size of the oncogenic pathway. The progenitor cell, having incurred damage x and y, is selected by oncogenesis if there is at least one open pathway. In other words, the probability of selection is

$$\begin{aligned} P(\text{SEL}) &= 1 - P(\text{all pathways closed}) \\ &= 1 - \prod_e (1 - p_e). \end{aligned}$$

(The product form is valid only when two different pathways do not share a common aberration, but see the next section for an extension of this.) We can by a similar argument calculate

$$P(\text{SEL}|x) = 1 - \prod_e [1 - p_e(x)].$$

where $p_e(x) = P(\text{pathway } e \text{ is open}|x)$. If all the aberrations on e have occurred overtly, then $p_e(x) = 1$. Covert aberrations may be necessary to open the pathway, and thus, more generally,

$p_e(x) = \beta^{t_e}$ where $t_e = \sum_{i \in e}(1 - x_i)$. Combining the instability and selection terms, the likelihood contribution from a tumor presenting data x is:

$$P(x|\mathrm{SEL}) = \alpha^{\sum_i x_i}(1-\alpha)^{\sum_i (1-x_i)} \left\{ \frac{1 - \prod_e [1 - p_e(x)]}{1 - \prod_e (1 - p_e)} \right\} \tag{1}$$

Newton (2001) first derived this model, showed some of its sampling properties, and presented Bayesian model fitting methods. It is interesting that the act of selection creates heterogeneity in marginal aberration rates, positive dependence between aberrations on the same pathway, negative dependence between aberrations on different pathways, and independence of all neutral aberrations.

Bayesian analysis was proposed in Newton (2001) to infer the unknown oncogenic pathways using the likelihood (1) and a prior distribution over rates α and β and, more importantly, over the network of oncogenic pathways. The number and composition of these pathways is unknown. The prior implemented by the software developed in Newton (2001) (and used in the example from Section 2) uses a Bernoulli-Polya prior for the vector a and then a second Polya-type prior for the pathways. With suggestive notation, the label subvector $c[a]$ indicates the pathway structure amongst relevant aberrations. The prior distribution is:

$$\begin{aligned} \pi(c[a], a) &= \pi(c[a]|a)\ \pi(a) \\ &= \left[\frac{\tau^K \Gamma(\tau)\, [\prod_e \Gamma(n_e)]}{\Gamma(\tau + m)} \right] \left[\frac{\Gamma(m+1)\Gamma(n-m+1)}{\Gamma(n+2)} \right] \end{aligned}$$

where K is the number of pathways, n_e is the length of pathway e, $m = \sum_e n_e$ is the total number of relevant aberrations, $n - m$ is the number of neutral aberrations, and $\Gamma()$ is the Gamma function. Among other things, the prior induces a uniform prior distribution over m. A single hyperparameter $\tau > 0$ affects the extent of clustering.

Markov chain Monte Carlo is the obvious approach to enable posterior analysis. The algorithm in Newton (2001) involves a collection of move types which alter the activation vector a, the cluster vector c and the unknown rates α and β. It is interesting to note that some simple move types are not very effective. For example, in changing a neutral aberration to a relevant aberration it seems reasonable to attempt to add a new pathway carrying that single aberration. Such a local change in network space may correspond to a very large change in likelihood (1), since short pathways are expected to present very particular statistical properties. On the other hand, two networks that differ by one long pathway may have very similar likelihood.

Figure 5 shows trace plots of log-posterior and network size m for two independent MCMC runs using the 40 breast cancer profiles. We show results for the particular hyperparameter $\tau = 5$, and we note that similar results obtained using $\tau = 1$ and $\tau = 10$. The mixing is reasonable but not ideal; it is encouraging that posterior summaries vary very little across replicate runs. The posterior mean for α is 0.094 and β is 0.085. The prior was uniform over the triangle $\alpha < \beta$.

A point estimate of the network is the one which achieves the highest posterior probability. For these data, we obtain two pathways of equal length:

e_1	+1q	+8q	-16q	-13q	-22q	-1p	-11q
e_2	+19q	+20q	+16p	+17q	+19p	+1p	-6q

(The MAP estimate was similar for the three hyperparameter values considered.)

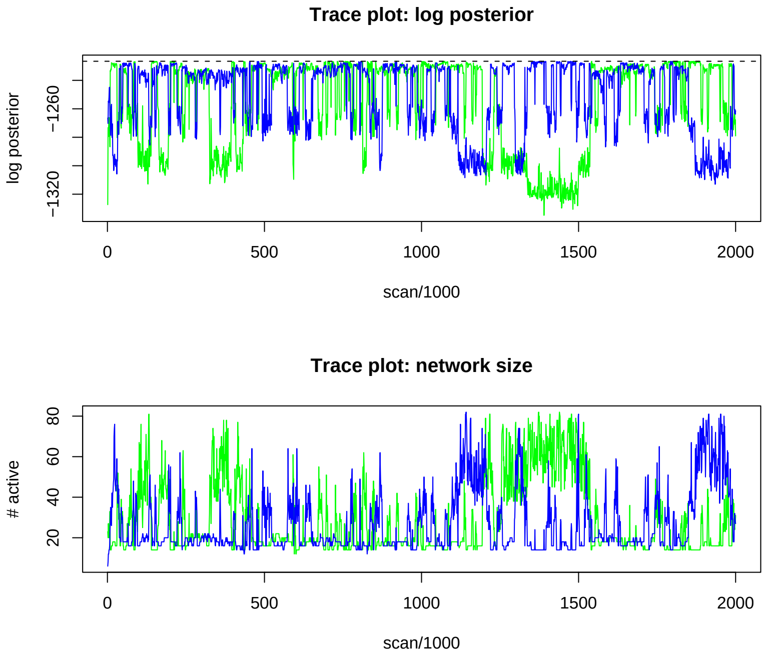

Figure 5. *Trace plots from MCMC: log posterior of each network (top) and number of relevant aberrations (bottom). Each chain started at a random network, proceeded for* 2×10^6 *scans, and was subsampled every 1000 scans.*

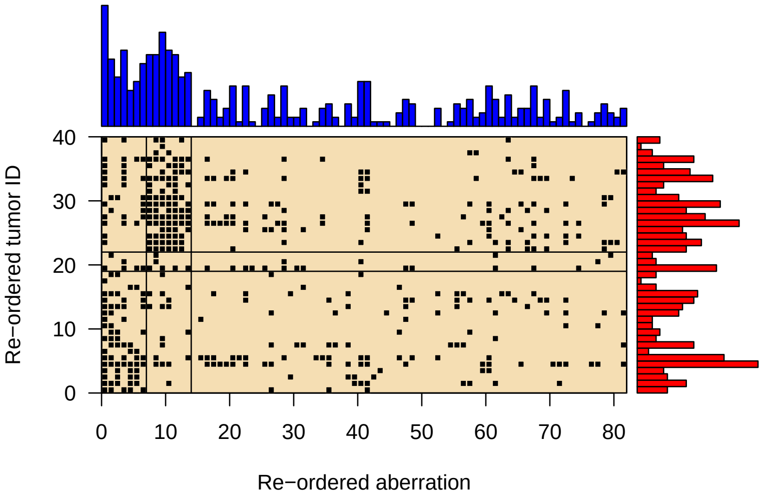

Figure 6. *Genomic aberrations like in Figure 1, but with rows and columns rearranged. The left block of columns corresponds to the first estimated pathway, the second block to the second estimated pathway, and the remainder are probably neutral. The rows (tumors) are organized by chances of following each pathway. The lower block of 19 tumors probably followed the first pathway* e_1*, the upper block of 18 tumors probably followed* e_2*, and it's a toss up for three of the tumors.*

Having an estimated network allows us to perform a model-based clustering of the tumors. The notion is that each tumor has traversed (at least) one of the pathways. Conditional on the data x, we can compute the probability that each of the estimated pathways was open for that tumor. Figure 6 shows a summary of this calculation. It is analogous to Figure 1, except that aberrations are rearranged according to the estimated pathways and tumors are reordered according to the pathway they probably followed. These pathway predictions might provide a useful biomarker to relate to other properties of the tumor or clinical outcomes.

3.2. *Tree-like pathways*

A limitation of the methods developed in Newton (2001) and reviewed here is that pathways cannot overlap. This simplifying assumption makes it feasible to calculate the likelihood (1) of a network, but the general understanding from cancer biology compels us to do better. Rather than allow all possible pathway arrangements, it is useful to restrict attention to tree-like networks, as Desper and colleagues have done (Desper *et al.*, 1999, 2000; Jiang, *et al.*, 2000). The question is can we develop instability-selection calculations for pathways arranged in a tree? Take the right panel in Figure 4 as an example. The tree is comprised of edges, as before, that are disjoint collections of relevant aberrations. No longer is there a 1-1 correspondence between edges and oncogenic pathways. A pathway is a series of edges moving from the root to a leaf node. To be more precise, each edge e has a parent edge $\text{PA}(e)$, allowing that the root can be a parent edge also (e.g., the parent of the edge containing aberration 6 is the root.) Some edges are not parents of any edge, and these are the leaves. What makes the network tree-like is the absence of loops in this graph. Note that the number of oncogenic pathways equals the number of leaves.

To calculate a likelihood, note that, as before, each edge e has a probability p_e of being open, and a conditional probability $p_e(x)$ of being open in light of overt damage x. Each edge e also represents the ancestral edge of a branch of the tree. For example (using the aberration label/s to label the edge), in Figure 4, e_6 begins a branch containing e_6, e_7 and e_5. Edges like $e_{1,3}$ constitute a single branch, as does any edge which is a leaf. Noting this special tree structure, we can define two other probabilities on each edge:

$$\tilde{p}_e = P(\text{branch starting at } e \text{ is open})$$

and

$$\tilde{p}_e(x) = P(\text{branch starting at } e \text{ is open}|x)$$

For leaf edges, $\tilde{p}_e = p_e$ and $\tilde{p}_e(x) = p_e(x)$. For non-leaf edges, the event that the branch starting at e is open is the event that some descendant pathway in that branch is open. We immediately obtain the recursive equations:

$$\tilde{p}_e = p_e \left[1 - \prod_{h:\text{PA}(h)=e} (1 - \tilde{p}_h) \right]$$

and

$$\tilde{p}_e(x) = p_e(x) \left[1 - \prod_{h:\text{PA}(h)=e} [1 - \tilde{p}_h(x)] \right]$$

For example, $\tilde{p}_6 = p_6[1-(1-p_5)(1-p_7)]$. Given any tree-like network we can compute $\tilde{p}_e$ and $\tilde{p}_e(x)$ for all edges by moving recursively from the leaves towards the root. The chance of selection and the likelihood involve $\tilde{p}_e$ and $\tilde{p}_e(x)$ for edges emanating from the root. Extending (1), we obtain

$$P(x|\text{SEL}) = \alpha^{\sum_i x_i}(1-\alpha)^{\sum_i (1-x_i)} \left\{ \frac{1 - \prod_{e:[\text{PA}(e)=\text{root}]}[1 - \tilde{p}_e(x)]}{1 - \prod_{e:[\text{PA}(e)=\text{root}]}(1 - \tilde{p}_e)} \right\} \tag{2}$$

This formula allows us to extend the domain of instability-selection modeling to more realistic oncogenic pathway structures. It remains to develop a useful prior distribution over tree-like networks and to implement a posterior sampling algorithm.

APPENDIX

The data was obtained by using comparative genomic hybridization, a method of screening the entire genome for gains and losses of genetic material in a single experiment. Differentially labelled test or tumor DNA (green) and reference or normal DNA (red) are co-hybridized to normal metaphase spreads. Differences in the copy number between test and reference DNA are seen as differences in the ratio of green to red fluorescence intensity on the metaphase chromosomes. Images of the metaphases are captured using an epifluorescence microscope, equipped with a cooled CCD (charge-coupled device) camera. Quantification of the fluorescence ratios is performed using a digital image analysis system. For each tumor, 5-10 metaphases are analysed and an average fluorescence ratio for each chromosome obtained. Regions of chromosomal gain are seen as an increased fluorescence ratio, while regions of loss are seen as a decrease in the fluorescence ratio. Conventionally, gains and losses are considered significant when the ratio is $> 1.15 : 1$ and $< 0.85 : 1$ respectively.

REFERENCES

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

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

Gray, J. W. and Collins, C. (2000). Genome changes and gene expression in human solid tumors. *Carcinogenesis* **21**, 443–452.

Haag, J. D., L.-C. Hsu, Newton, M. A. and Gould, M.N. (1996). Allelic Imbalance in Mammary Carcinomas Induced by Either 7,12-Dimethylbenz[a]anthracene or Ionizing Radiation in Rats Carrying Genes Conferring Differential Susceptibilities to Mammary Carcinogenesis. *Mol. Carcinog.* **17**, 134–143.

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics, *J. Comp. Graph. Statist.* **5**, 299–314. (See `www.r-project.org`.)

Jarrard, D. F., Sarkar, S., Shi, Y., Yeager, T. R., Magrane, G., Kinoshita, H., Nassif, N., Meisner, L., Newton, M. A., Waldman, F. M. and Reznikoff, C. A. (1999). p16/pRb Pathway Alterations Are Required for Bypassing Senescence in Human Prostate Epithelial Cells. *Cancer Research* **59**, 2957–2964.

Jiang, F., Desper, R., Papadimitriou, C. H., Scäffer, A. A., Kallioniemi, O.-P., Richter, J., Schraml, P., Sauter, G., Mihatsch, M. J. and Moch, H. (2000). Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data. *Cancer Research* **60**, 6503–6509.

Knuutila S, Björkqvist A.-M., Autio K., Tarkkanen, M., Wolf, M., Monni, O., Szymanska, J., Larramendy, M. L., Tapper, J., Pere, H., El-Rifai, W., Hemmer, S., Wasenius, V.-M., Vidgren, V. and Zhu, Y. (1998). DNA copy number amplifications in human neoplasms. Review of comparative genomic hybridization studies. *Am. J. Pathol.* **152**, 1107–1123.

Knuutila, S., Aalto, Y., Autio, K., Björkqvist, A.-M., El-Rifai, W., Hemmer, S., Huhta, T., Kettunen, E., Kiuru-Kuhlefelt, S., Larramendy, M. L., Lushnikova, T., Monni, O., Pere, H., Tapper, J., Tarkkanen, M., Varis, A., Wasenius, V.-M., Wolf, M. and Zhu, Y. (1999). DNA copy number losses in human neoplasms. Review. *Am. J. Pathol. Online* **155**, 683–694.

Lengauer, C., Kinzler, K. W. and Vogelstein, B. (1998). Genetic instabilities in human cancers *Nature* **396**, 643–649.

Newton, M. A. (2001). A statistical method to discover significant combinations of genetic alterations associated with cancer using comparative genomic hybridization profiles. *Tech. Rep.* **128**, Dept. Biostatistics and Madical Informatics, UW Madison, USA.

Newton, M. A., Gould, M. N. Reznikoff, C. A. and Haag, J. D. (1998). On the statistical analysis of allelic-loss data. *Statist. Med.* **17**, 1425-1445.

Newton, M. A., Yeager, T. and Reznikoff, C. A. (1999). A statistical analysis of cancer genome variation. *Statistics in Genetics, IMA Volumes in Mathematics and its Applications*, M. E. Halloran and S. Geisser (eds.), **112**, 223–236, New York: Springer.

Newton, M. A. and Lee, Y. J. (2000). Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics* **56**, 1088-1097.

Shoemaker, A. R., Moser, C. A., Midgley, L., Clipson, M. A., Newton, W. F. Dove (1998). A resistant genetic background leading to incomplete penetrance of intestinal neoplasia and reduced loss of heterozygosity in Apc$^{\mathrm{Min}/+}$ mice. *Proc. Natl. Acad. Sci. USA* **95**, 10826–10831.

Teeguarden, J. G., Newton, M. A., Dragan, Y. P. and Pitot, H. C. (2000). Genome-wide loss of heterozygosity analysis of chemically induced rat hepatocellular carcinomas reveals elevated frequency of allelic imbalances on chromosomes 1, 6, 8, 11, 15, 17, and 20. *Molecular Carcinogenesis* **28**, 51–61.

Yeager, T. R., DeVries, S., Jarrard, D. F., Kao, C., Nakada, S. Y., Moon, T. D., Bruskewitz, R., Stadler, W. M., Meisner, L. F., Gilchrist, K. W., Newton, M. A., Waldman, F. M. and Reznikoff, C. A. (1998). Overcoming cellular senescence in human cancer pathogenesis. *Genes and Development* **12**, 163–174.

DISCUSSION

SCOTT C. SCHMIDLER *(Duke University, USA)*

I would like to begin by congratulating the authors on a very interesting paper, which describes an elegant stochastic model for genomic aberration data. This paper builds on a body of previous work by the authors, and as such represents a well-developed and relatively mature statistical methodology. Thus my comments will focus as much on the scientific questions under study as on the statistical methods, with comments on the latter restricted primarily to the methodological extensions introduced in this paper for incorporation of dependency structure between aberrations.

The Data: Modern biologists often wish to ask a wide variety of questions of the data, and the model-based approach and simulation methods described in the paper are particularly attractive in such settings. However it is also worth asking what are likely to be the primary goals of the type of analysis described in the paper. Although this is not stated explicitly, one presumes that one such goal may be to identify aberrations implicated in tumorigenesis for more detailed study; another perhaps to find diagnostic markers for tumor identification. In both cases I wonder whether the data described are adequate to answer such questions in the absence of a control population. Figure 2 of the paper shows that high empirical frequency generally leads to high posterior probability of relevance, but perhaps some aberrations are more frequent in non-tumor cells as well. Do the authors have other data or domain knowledge which applies here?

A second question involves the interpretation of the “oncogenic pathways” under study. I find it somewhat difficult to ascribe a meaning to these pathways in terms of molecular events. This makes validation and interpretation of the inferred pathways unclear. Put another way, what does their inference tell us about the underlying biological system?

A minor point is that given the large amount of work invested here in modeling and computation, the reduction of the data to a set of directly observable $x_i \in \{0,1\}$ seems undesirable. Might more quantitative information be obtained simply by treating the x_i's as missing data and writing a conditional density $\prod_k P(f_k \,|\, x_i)$ for the fluorescence ratio observations themselves?

The Model: A few relatively minor questions regarding the model itself: Does x_i independent of x_j make sense when x_i denotes deletion of a site and x_j amplification of the same site? Are the inferences sensitive to such assumptions?

In reality, $\alpha_i \not\equiv \alpha$ and similarly for β_i. Moreover, these differences are likely to be of real interest. Is it possible with the currently available data to make inference about differences in these parameters?

The Bernoulli-Polya prior chosen has uniform marginal distribution on m. In previous work, the first author have showed that the resulting MAP inference can miss larger networks, suggesting that perhaps this prior over-penalizes large networks. In a related problem involving priors on set partitions, Schmidler (2002) found that such combinatorial priors which provide adequate marginal inferences often provide poor MAP estimates. This may be checked by performing the MCMC using a uniform prior $P(c\,|\,a) \propto const$. This can be expected to give upwardly biased marginal probabilities of aberrations sharing a pathway, but the resulting MAP networks may be informative.

The Results: A particularly attractive aspect of this paper is the elegant solution it provides to both pathway identification and tumor clustering using a common statistical model. The visualization of marginal posterior probabilities of dependence via hierarchical clustering is quite informative, and likely to have applications in other areas.

The posterior mean values obtained for α and β are .094 and .085, respectively. I find this slightly troubling in light of the prior used requiring $\beta < \alpha$. It seems to suggest that the "covert" aberrations in the model are playing nearly as big a role in inference as the aberrations actually observed. Perhaps the β's provide too much flexibility? Placing additional restrictions on the β's or directly on $\sum_i y_i$ may be desirable.

Finally, an important question not addressed in the paper is the validation of the inferred pathways. The MAP networks are interesting, but there will always be a MAP network - how can we verify whether it is correct? This relates back to the question of interpretation of oncogenic pathways - is there a measurable quantity that can provide ground truth?

Model extensions: The paper discusses but does not implement relaxing the assumption of non-overlapping pathways via a tree structure. Note that there is no inherent ordering among aberrations in a "pathway", so many network structures may also be rearranged to form trees as well. It should be noted that the recursive calculation given applies equally well to general network structures described in the paper, and the authors have indicated that they are aware of this as well.

The word "pathway" is a bit of a misnomer here, due to the lack of any real ordering among aberrations sharing a pathway. We may instead view pathways as simply a subset of aberrations all of which must occur, and therefore we may define a pathway as a new random variable formed by a logical AND, *e.g.*, $e = a_i \wedge a_j \wedge \ldots \wedge a_k$. This suggests more general logic statements, *i.e.*, use of logical OR and NOT. The trees described introduce a restricted form of OR, and networks extend this further. However use of negation may be valuable as well. Here I envision certain aberrations which may contribute to tumorigenesis, but which are lethal when occurring simultaneously. Such cases would best be modeled by statements of the form $a_i \oplus a_j = (a_i \vee a_j) \wedge \neg(a_i \wedge a_j)$.

General remarks: The approach developed here appears to have direct applications to related problems in genomics and bioinformatics, such as the discovery of genetic regulatory networks and metabolic pathways. Have the authors considered applying this approach to these other important and open problems? It is also worth noting the relation to the latent factor models of West *et al.* (2001) where the a_i's are not restricted to binary values. Another question is whether it is possible leverage the large amounts of genomic sequence data recently deposited into public databases, perhaps in helping specify informative priors for aberrations occurring in the region of known genes.

Finally, I would like to congratulate the authors again on an elegant and practical model-based solution to an important applied problem in molecular biology. Too often problems in bioinformatics have been dominated by heuristic algorithmic formulation, ignoring the underlying uncertainty and implicit statistical assumptions. In recent years, a number of such problems have been treated from a formal (often Bayesian) statistical modeling perspective, including global and local sequence alignment (Lawrence *et al.* (1993) and Krogh *et al.* (1994)), protein structure prediction (Schmidler *et al.* (2000)), molecular structure analysis (Wu *et al.* (1998), Schmidler (2002)), and an enormous literature on analysis of gene expression experiments (*e.g.*, Kerr *et al.* (2000), Li and Wong (2001)). The authors have been among the pioneering statisticians in addressing important problems in molecular biology and genetics, and it is a great pleasure to be able to add oncogenic pathway discovery to this list of problems with proper statistical solutions.

REPLY TO THE DISCUSSION

The discussant has identified critical issues surrounding the proposed methodology and he has been generous in his overall assessment. We appreciate the informed comments.

To clarify our goals, the primary goal is to characterize potentially important combinations of genomic aberrations so as to guide future cancer studies. Such studies might derive useful diagnostic markers for clinical work or they could involve focused searches for the genes and molecules involved in the cancer process. Our approach is statistical, of course, and it is meant to provide a quantitative summary of the information collected in a given CGH study. One hopes that such an analysis can complement related investigations of tumor biology. Our calculations do not provide direct insight into molecular events, but that would be a lot to ask from the CGH data alone. Regarding controls, it is important to recognize that the genomic damage measured by CGH does not occur in normal tissue. That is, except for a very low rate of measurement error, Figure 1 would be all white (no aberrations) if non-tumor DNA was under investigation. In machine-learning terms, we have an unsupervised learning problem.

Fairly, the discussant criticizes our data reduction scheme. The binary data retain most of the signal, we suspect, but certainly one could attempt to model the whole intensity profile. Each binary arm-level indicator is an inference about whether or not some aberration has occurred in a large genomic region, so it is aggregating many intensity measurements together. This issue requires some serious thought for high-resolution array-based CGH measurements that are now becoming available. A brute-force model-based approach might be difficult to implement; it seems quite practical to segment the intensity profiles somehow for our downstream network calculations. Precisely how to do this is an open problem.

The discussant suggests that the marginal rates of occurrence of overt and covert aberrations ought to depend on the genomic location i. This may be so but it raises some interesting questions about the whole enterprise to look for cancer genes in aberration hot spots. The issue is investigated in Newton *et al.* 1998. Our model entails prior homogeneity (of aberration rates) that becomes heterogeneity in tumors because of selection. If genomic instability targets a location i, rather than being random (α_i is larger than other α_j, rather than complete equality), then aberration i may occur frequently in tumors, whether or not it is relevant to oncogenesis. Our premise, essentially, is that instability is less directed than that, implying that aberration hot spots in observed cells correspond somehow to relevant genes. One could allow independent but heterogeneous genomic instability. Then it would be the second order statistical dependencies only, rather than the marginal frequencies that would deliver inferences about relevant aberrations. There seem to be clear benefits to the present model; for one thing, the estimated network provides a simple reorganization of the aberrations according to both marginal frequency and joint co-occurrence.

We are intrigued by discussant's related work suggesting that a flat prior over networks may be preferable when computing the MAP estimate, even though the informed prior produces reasonable marginal inferences. One might expect the opposite if shrinkage results from continuous parameter spaces provide any guidance.

Our prior constraint $\alpha > \beta$ was used to enforce some regularity since too many covert aberrations will quench the effect of selection. That the posterior pushes β near α suggests to us that the allowable pathways have not captured the dependence patterns effectively. This is why we have been so keen to develop tools for overlapping pathway models. As we have reported, we can compute the likelihood using a recursive algorithm, but we continue to build MCMC methods for this general case.

The potential for the instability-selection-network methodology to characterize statistical patterns in many forms of biological data has not gone unnoticed by us; indeed, we are excited about possible developments beyond chromosome-based CGH. The term "pathway" may be imprecise, as the discussant notes, since we use it to represent an unordered collection of markers; the term "ensemble" may be more appropriate (Newton, 2002). Whatever we call it, the framework is compelling. Data measuring aberrant features in the current state of a living system exhibit patterns; these patterns arose somehow prior to observation, and, insofar as they were favorable, they became observable. In allowing overlapping pathways, the class of models based on logical AND seems quite rich; we have not considered the more general class but the discussant's proposal ought to be followed up. With ever richer models the issue of identifiability seems to be important, in spite of the fact that a Bayesian analysis of some data with some model need not regard the issue. From the perspective of model development, we think it is useful to know what can be recovered in principle from a large data set.

Finally, we comment on "ground truth" and how to verify that an estimated network is correct in some sense. The best solution we can suggest is via prediction. Each CGH profile is a high-dimensional marker of tumor type; the modeling effort provides a clustering of tumors according to the probable pathways that each tumor has experienced (Figure 6). If the approach has any merit, it may be that the probable-pathway marker is an efficient low-dimensional predictor of other attributes, such as clinical outcome. Currently we are investigating the relationship between pathway predictions and clinical data.

ADDITIONAL REFERENCES IN THE DISCUSSION

Kerr, M. and Churchill, G. (2000). Analysis of variance for gene expression microarray data. *J. Comp. Biol.* **7**, 819–837.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994). Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214.

Li, C. and Wong, W.H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* **98**, 31–36.

Newton, M.A. (2002). Discovering combinations of genomic aberrations associated with cancer. *J. Amer. Statist. Assoc.* **97** (to appear).

Schmidler, S. C. (2002). *Statistical Models and Monte Carlo Methods for Protein Structure Prediction*. Ph.D. Thesis, Stanford University, USA.

Schmidler, S. C. (2002). Statistical shape theory and protein structure analysis. *Tech. Rep.*, Duke University, USA.

Schmidler, S. C., Liu, J. S. and Brutlag, D. L. (2000). Bayesian segmentation of protein secondary structure. *J. Comp. Biol.* **7**, 233–248.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Marks J.R. and Nevins J.R. (2001). Predicting the clinical status of human breast cancer using gene expression profiles. *Proc. Natl. Acad. Sci.* **98**, 11462–11467.

Wu, T.D., Schmidler, S.C., Hastie, T. and Brutlag, D.L. (1998). Regression analysis of multiple protein structures. *J. Comp. Biol.* **5**, 597–607.