Published in final edited form as:

Biometrics. 2011 March ; 67(1): 164–173. doi:10.1111/j.1541-0420.2010.01417.x.

Learning oncogenic pathways from binary genomic instability data

Pei Wang, Dennis L. Chao, and Li Hsu*

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA

Summary

Genomic instability, the propensity of aberrations in chromosomes, plays a critical role in the development of many diseases. High throughput genotyping experiments have been performed to study genomic instability in diseases. The output of such experiments can be summarized as high dimensional binary vectors, where each binary variable records aberration status at one marker locus. It is of keen interest to understand how aberrations may interact with each other, as it provides insight into the process of the disease development. In this paper, we propose a novel method, LogitNet, to infer such interactions among these aberration events. The method is based on penalized logistic regression with an extension to account for spatial correlation in the genomic instability data. We conduct extensive simulation studies and show that the proposed method performs well in the situations considered. Finally, we illustrate the method using genomic instability data from breast cancer samples.

Keywords

Conditional Dependence; Graphical Model; Lasso; Loss-of-Heterozygosity; Regularized Logistic Regression

1. Introduction

Genomic instability refers to the propensity of aberrations in chromosomes such as mutations, deletions and amplifications in diseased tissues. It has been thought to play a critical role in the development of many diseases, for example, many types of cancers (Klein and Klein 1985). Unlike germline mutations, these somatic chromosomal abnormalities are not passed on from parents to children (Coop and Ellis 2009). Identifying which somatic aberrations contribute to disease risk, and how they may interact with each other during disease development is of keen interest. High throughput genotyping experiments have been performed to interrogate these aberrations in diseases, providing a vast amount of information on genomic instabilities on tens of thousands of marker loci simultaneously. These data can be organized as a $n \times p$ matrix where n is the number of samples, p is the number of marker loci, and the $(i, j)^{\text{th}}$ element of the matrix is the binary aberration status for the i th sample at the j th locus. Our goal is to infer the conditional dependencies between aberrations, which we refer to as oncogenic pathways, based on these binary genomic instability profiles.

Oncogenic pathways can be compactly represented by graphs, in which vertices represent aberrations and edges represent interactions between aberrations. Tools developed for

*Corresponding Author: 1100 Fairview Ave. N., M2-B500, Fred Hutchinson Cancer Research Center, Seattle, WA 98109.
lih@fhcrc.org.

graphical models (Lauritzen 1996) can therefore be employed to infer interactions among aberrations. Specifically, each vertex represents a binary random variable that codes aberration status at a locus, and an edge will be drawn between two vertices if the corresponding two random variables are conditionally dependent given all other random variables. Here, we want to point out that graphical models based on conditional dependencies provide information on “higher order” interactions compared to other methods (e.g., hierarchical clustering) which examine the marginal pairwise correlations. The latter does not tell, for example, whether a non-zero correlation is due to a direct interaction between two aberration events or due to an indirect interaction through a third intermediate aberration event.

There is a rich literature on fitting graphical models for a limited number of variables (see for example Edward 2000; Drton and Perlman 2004, and references therein). However, in genomic instability profiles, the number of genes p is typically much larger than the number of samples n . Under such high-dimension-low-sample-size scenarios, sparse regularization becomes indispensable for purposes of both model tractability and model interpretation. Some work has already been done to tackle this challenge for high dimensional continuous variables. For example, Meinshausen and Bühlmann (2006) proposed to perform neighborhood selection with lasso regression (Tibshirani 1996) for each node. Peng et al. (2009a) extended the approach by imposing the sparsity on the whole network instead of each neighborhood, and implemented a fast computing algorithm. Other approaches include penalized maximum likelihood (e.g., Yuan and Lin 2007 and Friedman et al. 2007b), various regularization methods (e.g., Li and Gui 2006; Schafer and Strimmer 2007), and Bayesian methods (e.g., Madigan et al. 1995).

In this paper, we consider binary variables and propose a novel method, LogitNet, for inferring edges, i.e., the conditional dependence between pairs of aberration events given all others. With proper assumptions on the topology of oncogenic pathways, we derive the joint probability distribution of the p binary variables, which naturally leads to a set of p logistic regression models with the combined $p \times p$ coefficient matrix being symmetric. We propose sparse logistic regression with a lasso penalty term and extend it to account for the spatial correlations within chromosomes. This extension together with the enforcement of symmetry of the coefficient matrix produces a group selection effect, which enables LogitNet to account for and also benefit from spatial correlation when inferring the edges.

LogitNet is related to the work by Ravikumar et al. (2009), which also utilized sparse logistic regression to construct a network based on high dimensional binary variables. The basic idea of Ravikumar et al. is the same as that of Meinshausen and Bühlmann's (2006) neighborhood selection approach, in which sparse logistic regression was performed for each binary variable given all others. A sparsity constraint was then imposed on each neighborhood and the sparse regression was performed for each binary variable separately. In this approach, the symmetry of conditional dependence obtained from regressing variable X_r on variable X_s and from regressing X_s on X_r is not guaranteed. As such, it can yield contradictory neighborhoods, which makes interpretation of the results difficult. It also loses power in detecting dependencies, especially when the sample size is small. The proposed LogitNet, on the other hand, makes use of the symmetry, which produces compatible logistic regression models for all variables and thus achieves a more coherent result with better efficiency than the Ravikumar et al. approach. We show by intensive simulation studies that LogitNet performs better in terms of false positive rate and false negative rate of edge detection.

The rest of the paper is organized as follows. In section 2, we will present the model, its implementation and the selection of the penalty parameter. Simulation studies of the

proposed method and the comparison with the Ravikumar et al. approach are described in Section 3. Real genomic instability data from breast cancer samples is used to illustrate the method in Section 4. We conclude the paper with a brief summary in Section 5.

2. Methods

2.1 LogitNet Model and Likelihood Function

Consider a $p \times 1$ vector of binary variables $X^T = (X_1, \dots, X_p)$. Here the superscript T is a transpose. The pattern of conditional dependencies between these binary variables can be described by an undirected graph $\mathcal{G} = (V, E)$, where V is a set of p vertices $\{1, \dots, p\}$, representing binary variables $(X_1, \dots, X_p)$; and E is a set of pairs of vertices such that each pair are conditionally dependent given the rest of binary variables. We assume that the sizes of *maximum cliques* in the graph are all ≤ 2 . Under this assumption, the joint probability distribution $\Pr(X)$ can be represented as a product of functions of pairs of binary variables, i.e., cliques of size 2. Note that a *clique* is defined as a subset of vertices with each pair of vertices in the subset connected; and a *maximum clique* is a clique such that including any other vertex outside the clique will make the clique cease to be a clique. For example, in a set of four binary variables, if the edge set includes $(1,2)$, $(2,3)$, $(3,4)$ and $(1,4)$, under this assumption it can not include edge $(1,3)$ or $(2,4)$, as adding either edge will form maximum clique with size 3. However this assumption does not preclude the possibility of cycles. In the above example, we can see the edge set includes a cycle $1 - 2 - 3 - 4 - 1$.

Proposition 1—Let $V = \{1, \dots, p\}$; and $X_{-(r,s)}$ denote the vector of binary variables X excluding X_r and X_s for $r, s \in V$. Define the edge set

$$E = \{(r, s) | \Pr(X_r, X_s | X_{-(r,s)}) \neq \Pr(X_r | X_{-(r,s)}) \Pr(X_s | X_{-(r,s)})\}; r, s \in V, r < s\},$$

and $|E| = K$. If the sizes of maximum cliques in $\mathcal{G}$ are ≤ 2 , then there exist functions $\{h_k, k = 1, \dots, K\}$ such that

$$\Pr(X) = \prod_{k=1}^K h_k(X_{r_k}, X_{s_k}),$$

where $(r_k, s_k) \in E$ for $k = 1, \dots, K$.

The proof of Proposition 1 is largely based on the Hammersley and Clifford theorem (Lauritzen, 1996) and given in Supplementary Appendix A.

Assuming $\Pr(X)$ is strictly positive for all values of X , then the above probability distribution leads to the well known quadratic exponential model

$$\Pr(X=x) = \Delta^{-1} \exp(x^T \theta + z^T \kappa), \quad (1)$$

where $z^T = (x_1 x_2, x_1 x_3, \dots, x_{p-1} x_p)$, $\theta^T = (\theta_1, \dots, \theta_p)$, $\kappa^T = (\kappa_{12}, \kappa_{13}, \dots, \kappa_{(p-1)p})$, and Δ is a normalization constant such that $\Delta = \sum_{x_1=0}^1 \dots \sum_{x_p=0}^1 \exp(x^T \theta + z^T \kappa)$.

Under this model, the zero values in κ are equivalent to the conditional independence for the corresponding binary variables. The following proposition describes this result and the proof is given in Supplementary Appendix B.

Proposition 2—If the distribution on X is (1), then $X_r \perp\!\!\!\perp X_s \mid X_{-(r,s)}$ if and only if $\kappa_{rs} = 0$, for $1 \leq r < s \leq p$.

As the goal of graphical model selection is to infer the edge set E which represents the conditional dependence among all the variables, the result of Proposition 2 implies that we can infer the edge between a pair of events, say X_r and X_s , based on whether or not κ_{rs} is equal to 0. Interestingly, under model (1), κ can also be interpreted as a conditional odds ratio. This can be seen from

$$\frac{\Pr(x_s=1|x_1, \dots, x_{s-1}, x_{s+1}, \dots, x_p)}{\Pr(x_s=0|x_1, \dots, x_{s-1}, x_{s+1}, \dots, x_p)} = \exp(\kappa_{1s}x_1 + \dots + \kappa_{(s-1)s}x_{s-1} + \theta_s + \kappa_{s(s+1)}x_{s+1} + \dots + \kappa_{sp}x_p).$$

Taking the log transformation of the left hand side of this equation results in the familiar form of a logistic regression model, where the outcome is the j th binary variable and the predictors are all the other binary variables. Doing this for each of $x_1, x_2, \dots, x_p$, we obtain p logistic regressions models:

$$\begin{cases} \text{logit}\{\Pr(x_1=1|x_2, \dots, x_p)\} = \theta_1 + \kappa_{12}x_2 + \dots + \kappa_{1p}x_p, \\ \vdots \\ \text{logit}\{\Pr(x_p=1|x_1, \dots, x_{p-1})\} = \kappa_{1p}x_1 + \dots + \kappa_{(p-1)p}x_{p-1} + \theta_p. \end{cases} \quad (2)$$

The matrix of all of the regression coefficients from p logistic regression models can then be row combined, which we denote by $\mathcal{B}$. It is easy to see that the $\mathcal{B}$ matrix is symmetric, i.e., $\beta_{rs} = \beta_{sr} = \kappa_{rs}$, $i \neq j$ under model (2), where β_{rs} is the entry in the r th row and the s th column of $\mathcal{B}$. Vice versa, the symmetry of $\mathcal{B}$ ensures the compatibility of the p logistic conditional distributions (2), and the resulting joint distribution is the quadratic exponential model (1) (Joe and Liu, 1986). Thus, to infer the edge set E of the graphical model, i.e., nonzero off-diagonal entries in $\mathcal{B}$, we can resort to regression analysis by simultaneously fitting the p logistic regression models in (2) with symmetric $\mathcal{B}$.

Specifically, let $X_{n \times p}$ denote the data which consists of n samples each measured with p -variate binary events. We also define two other variables mainly for the ease of the presentation of the likelihood function: (1) Y is the same as X but with 0s replaced with -1 s; (2) X_r , $r = 1, \dots, p$ same as X but with r th column set to 1. We propose to maximize the summation of the log likelihood of the p logistic regressions in (2) as follows:

$$l(\mathcal{B}) = - \sum_{r=1}^p \sum_{i=1}^n \log \{1 + \exp(-X^r[i,] \mathcal{B}[r,]^T \cdot y_{ir})\}. \quad (3)$$

where $X^r[i,] = (x_{i1}^r, \dots, x_{ip}^r)$; and $\mathcal{B}[r,] = (\beta_{r1}, \dots, \beta_{rp})$. Note, here we have the constraints $\beta_{rs} = \beta_{sr}$ for $1 \leq r < s \leq p$; and β_{rr} now represents the intercept θ_r of the r th regression.

Recall that our interest is to infer oncogenic pathways based on genome instability profiles of tumor samples. Most often, we are dealing with hundreds or thousands of genes and only

tens or hundreds of samples. Thus, regularization on parameter estimation becomes indispensable as the number of variables is larger than the sample size, $p > n$. In the past decade, ℓ_1 norm based sparsity constraints such as lasso (Tibshirani 1996) have shown considerable success in handling high-dimension-low-sample-size problems when the true model is sparse relative to the dimensionality of the data. Since it is widely believed that genetic regulatory relationships are intrinsically sparse, we propose to use ℓ_1 norm penalty for inferring oncogenic pathways. The penalized loss function can be written as:

$$l_{\lambda}^{lasso}(\mathcal{B}) = -l(\mathcal{B}) + \lambda \sum_{r=1}^p \sum_{s=r+1}^p |\beta_{rs}|. \quad (4)$$

Note that ℓ_1 -norm penalty is imposed on all off-diagonal entries of $\mathcal{B}$ simultaneously to control the overall sparsity of the joint logistical regression model, i.e., only a limited number of β_{rs} , $r \neq s$ will be non-zero. We then estimate $\mathcal{B}$ by $\widehat{\mathcal{B}}(\lambda) := \arg \min_{\mathcal{B}} l_{\lambda}^{lasso}(\mathcal{B})$. In the rest of the paper, we refer to the model defined in (4) as LogitNet model, $\widehat{\mathcal{B}}(\lambda)$ as the LogitNet estimator and $\widehat{\beta}_{rs}(\lambda)$ as the rs th element of $\widehat{\mathcal{B}}(\lambda)$.

As described in the Introduction, the LogitNet model is closely related to the work by Ravikumar et al. (2009) which fits p lasso logistic regressions separately (hereafter referred to as SepLogit). The key difference between the two methods is that, LogitNet enforces symmetry when estimating the regression coefficients while SepLogit does not. Thus, for LogitNet, there are only about half of the parameters needed to be estimated as for SepLogit. As a result, LogitNet is more efficient and the resulting estimates are more interpretable than that of SepLogit. We want to point out that, in SepLogit, we used the same tuning parameter λ for all p lasso logistic regressions. Technically one can choose p different tuning parameters for the p regressions. However this strategy seriously overfits the data and additionally can cause substantial variability in model selection.

2.2 Model fitting

In this section, we describe an algorithm for obtaining the LogitNet estimator $\widehat{\mathcal{B}}(\lambda)$. The algorithm extends the gradient descent algorithm (Genkin et al. 2007) to enforce the symmetry of $\mathcal{B}$. Parameters are updated one at a time using a one-step Newton-Raphson algorithm, in the same spirit as the shooting algorithm (Fu 1998) and the coordinate descent algorithm (Friedman et al. 2007a) for solving the general linear lasso regressions.

More specifically, let $l(\beta_{rs})$ and $\dot{l}(\beta_{rs})$ be the first- and second- partial derivatives of log-likelihood $l(\mathcal{B})$ with respect to β_{rs} ,

$$\begin{aligned} \dot{l}(\beta_{rs}) &= \sum_{i=1}^n \frac{X^r[i,s]Y[i,r]}{1+\exp(R_r)} + \sum_{i=1}^n \frac{X^s[i,r]Y[i,s]}{1+\exp(R_s)}, \\ \ddot{l}(\beta_{rs}) &= \sum_{i=1}^n (X^r[i,s])^2 \frac{\exp(R_r)}{(1+\exp(R_r))^2} + \sum_{i=1}^n (X^s[i,r])^2 \frac{\exp(R_s)}{(1+\exp(R_s))^2}, \end{aligned}$$

where $R_r = X^r[i,s]\beta^T[r,Y[i,r]]$. Under the Newton-Raphson algorithm, the update for β_{rs} is $\Delta\beta_{rs} = -\dot{l}(\beta_{rs})/\ddot{l}(\beta_{rs})$. For the penalized likelihood (4), the update for $\widehat{\beta}_{rs}(\lambda)$ is

$$\Delta\beta_{rs}^{lasso} = -\frac{\dot{l}_{\lambda}^{lasso}(\beta_{rs})}{\ddot{l}_{\lambda}^{lasso}(\beta_{rs})} = \Delta\beta_{rs} - \frac{\lambda}{\ddot{l}(\beta_{rs})} \text{sgn}(\beta_{rs}), \quad (5)$$

where $\text{sgn}(\beta_{rs})$ is a sign function, which is 1 if β_{rs} is positive and -1 if β_{rs} is negative. The estimates are also thresholded such that if an update overshoots and crosses the zero, the update will be set to 0. If the current estimate is 0, the algorithm will try both directions by setting sgn to be 1 and -1 , respectively. By the convexity of (4), the update for both directions can not be simultaneously successful. If it fails on both directions, the estimate will be set to 0. The algorithm also takes other steps to make sure the estimates and the numerical procedure are stable, including limiting the update sizes and setting the upper bounds for $\hat{l}$ (Zhang and Oles 2001). See Supplemental Appendix C for more details.

To further improve the convergence speed of the algorithm, we utilize the Active-Shooting idea proposed by Peng et al. (2009a) and Friedman et al. (2009). Specifically, at each iteration, we define the set of currently non-zero coefficients as the current active set and conduct the following two steps: (1) update the coefficients in the active set until convergence is achieved; (2) conduct a full loop update on all the coefficients one by one. We then repeat (1) and (2) until convergence is achieved on all of the coefficients. Since the target model in our problem is usually very sparse, this algorithm achieves very fast convergence by focusing on the small subspace whose members are more likely to be in the model.

We note that in equation (5) the regularization shrinks the estimate towards zero by the amount determined by the penalty parameter λ and that each parameter is not penalized by the same amount: λ is weighted by the variance $\ddot{l}(\beta_{rs})^{-1}$ of $\hat{\beta}_{rs}$. In other words, parameter estimates that have larger variances will be penalized more than the ones that have smaller variances. It turns out that this type of penalization is very useful, as it would also offer us ways to account for the other features of the data. In the next section we propose a weight function to account for spatial correlations in genomic instability data.

2.3 Spatial correlation

Spatial correlation of aberrations is common in genomic instability data. When we perform the regression of X_r on all other binary variables, loci that are spatially closest to X_r are likely the strongest predictors in the model and will explain away most of the variation in X_r . The loci at the other locations of the same or other chromosomes, even if they are correlated with X_r , may be left out in the model. Obviously this result is not desirable because our objective is to identify the network among all of these loci (binary variables), in particular those that are not close spatially as we know them already.

One approach to accounting for this undesirable spatial effect is to downweight the effect of the neighboring loci of X_r when regressing X_r on the rest of the loci. Recall that in Section 2.2, we observed that the penalty term in (5) is inversely weighted by the variance of the parameter estimates. Following the same idea, we can achieve the downweighting of neighboring loci by letting the penalty term be proportional to the strength of their correlations with X_r . This way we can shrink the effects of the neighboring loci with strong spatial correlation more than those that have less or no spatial correlation. Specifically, the update for the parameter estimate β_{rs} in (5) can be written as

$$\Delta\beta_{rs}^{lasso} = \Delta\beta_{rs} - \lambda \frac{w_{rs}}{j(\beta_{rs})} \text{sgn}(\beta_{rs}),$$

where w_{rs} is the weight for the spatial correlation. Naturally the weight w_{rs} for X_r and X_s on different chromosomes is 1 and for X_r and X_s on the same chromosome should depend on the strength of the spatial correlation. As the spatial correlation varies across the genome, we propose the following adaptive estimator for w_{rs} :

1. Calculate the odds ratio α between every locus in the chromosome with the target locus by univariate logistic regression.
2. Plot the α 's by their genomic locations and smooth the profile by loess with a window size of 10 loci.
3. Set the smoothed curve $\tilde{\alpha}$ to 0 as soon as the curve starting from the target locus hits 0. Here "hits 0" is defined as $\alpha < \varepsilon$, where $\varepsilon = \text{median}_r |\alpha_r - \alpha_{r+1}|$.
4. Set the weight $w = \exp(\alpha)$.

It is worth noting that the above weighting scheme together with the enforcement of the symmetry of $\mathcal{B}$ in LogitNet encourages a group selection effect, *i.e.*, highly correlated predictors tend to be in or out of the model simultaneously. We illustrate this point with a simple example system of three variables X_1 , X_2 , and X_3 . Suppose that X_2 and X_3 are very close on the genome and highly correlated; and X_1 is associated with X_2 and X_3 but sits on a different chromosome. Under our proposal, the weight matrix w is 1 for all entries except $w_{23} = w_{32} = a$, which is a large value because of the strong spatial correlation between X_2 and X_3 . Then, for LogitNet, the joint logistic regression model

$$\text{logit}(X_1) \sim \beta_{11} + \beta_{12}X_2 + \beta_{13}X_3, \quad (6)$$

$$\text{logit}(X_2) \sim \beta_{12}X_1 + \beta_{22} + \beta_{23}X_3, \quad (7)$$

$$\text{logit}(X_3) \sim \beta_{13}X_1 + \beta_{23}X_2 + \beta_{33}, \quad (8)$$

is subject to the constraint $|\beta_{12}| + |\beta_{13}| + a|\beta_{23}| < s$. Because of the large value of a , β_{23} will likely be shrunk to zero, which ensures β_{12} and β_{13} to be nonzero in (7) and (8), respectively. With the symmetry constraint imposed on $\mathcal{B}$ matrix, we also enforce both β_{12} and β_{13} to be selected in (6). This grouping effect would not happen if we fit only the model (6) for which only one of β_{12} and β_{13} would likely be selected (Zou and Hastie 2005), nor would it happen if we didn't have a large value of a because β_{23} would have been the dominant coefficient in models (7) and (8). Indeed, the group selection effect of LogitNet is clearly observed in the simulation studies conducted in Section 3.

2.4 Penalty Parameter Selection

We consider two procedures for selecting the penalty parameter λ : cross validation (CV) and Bayesian Information Criterion (BIC).

Cross Validation—After we derive the weight matrix w based on the whole data set, we divide the data into V non-overlapping equal subsets. Treat the v^{th} subset $X^{(v)}$ as the v^{th} test set, and its complement X^{-v} as the v^{th} training set. For a given λ , we first obtain the LogitNet estimate $\hat{\beta}^v(\lambda)$ with the weight matrix w on the v^{th} training set X^{-v} . Since in our problem the true model is usually very sparse, the degree of regularization needed is often high. As a result, the value of $\hat{\beta}^v(\lambda)$ could be shrunk far from the true parameter values. Using such heavily shrunk estimates for choosing λ from cross validation often results in severe over-fitting (Efron et al. 2004). Thus, we re-estimate β using the selected model in the v^{th} training set without any shrinkage and use it in calculating the log-likelihood for the v^{th} test set. Denote the joint log likelihood of logistic regressions using the un-shrunk

estimates on the v^{th} test set as $l(\hat{\beta}_{uns}^{(v)}(\lambda)|X^{(v)})$. The optimal $\lambda_{cv} = \arg \max_{\lambda} \sum_{v=1}^V l(\hat{\beta}_{uns}^{(v)}(\lambda)|X^{(v)})$.

In order to further control the false positive findings due to stochastic variation, we employ the cv.vote procedure proposed by Peng et al. (2009b). The idea is to derive the “consensus” result of the models estimated from each training set, as variables that are consistently selected by different training sets should be more likely to appear in the true model than the ones that are selected by one or few training sets. Specifically, in the following simulation and real application sections, we report edges that have been detected in more than $V/2$ training sets as our final result.

BIC—We can also use BIC to select λ :

$$\lambda_{BIC} = \arg \min_{\lambda} \left\{ -2l(\hat{\beta}_{uns}(\lambda)|X) + \log(n) \sum_{r < s} I(\hat{\beta}_{uns,rs}^{(v)}(\lambda) \neq 0) \right\} \quad (9)$$

where $\sum_{r < s} I(\hat{\beta}_{uns,rs}^{(v)}(\lambda) \neq 0)$ gives the dimension of the parameter space of the selected model. Here again, *un-shrunk estimates* $\hat{\beta}_{uns}^{(v)}(\lambda)$ is used to calculate the log likelihood.

3. Simulation Studies

In this section, we investigate the performance of the LogitNet method and compare it with SepLogit which fits p separate lasso logistic regressions with the same penalty parameter value (Ravikumar et al., 2009). We use R package *glmnet* to compute the SepLogit solution and the same weight matrix as described in Section 2.3 to account for the spatial correlation. In addition, since the SepLogit method does not ensure the symmetry of the estimated β matrix, there will be cases that $\beta_{rs} = 0$ but $\beta_{sr} \neq 0$ or vice versa. In these cases we interpret the result using the “OR” rule: X_r and X_s are deemed to be conditionally dependent if either β_{rs} or β_{sr} is not 0. We have also used the “AND” rule, *i.e.* X_r and X_s are deemed to be conditionally dependent if both β_{rs} and β_{sr} are not 0. The “AND” rule always yields very high false negative rate. Due to space limitations, we omit the results for the “AND” rule.

We generated background aberration events with spatial correlation using a homogenous Bernoulli Markov model. It is part of the instability-selection model (Newton et al. 1998), which hypothesizes that the genetic structure of a progenitor cell is subject to chromosomal instability that causes random aberrations. The Markov model has two parameters: δ and ν , where δ is the marginal (stationary) probability at a marker locus and ν measures the strength of the dependence between the aberrations. So δ plays the role of a background or sporadic aberration when ν affects the overall rate of change in the stochastic process. Under this model, the infinitesimal rate of change from no aberration to aberration is $\nu\delta$, and from aberration to no aberration is $\nu(1 - \delta)$. We then super-imposed the aberrations at disease

loci, which were generated according to a pre-determined oncogenic pathway, on the background aberration events. The algorithm for generating an aberration indicator vector $X^T = (X_1, \dots, X_p)$ is given below:

1. Specify the topology of an oncogenic pathway for the disease loci and the transitional probabilities among the aberrations on the pathway. The K disease loci are indexed by $\{s_1, \dots, s_K\}$, where $s_i \in \{1, \dots, p\}$ for $i = 1, \dots, K$.
2. Generate background aberrations denoted by a $p \times 1$ vector Z according to the homogenous Bernoulli Markov process with preselected values of $\delta = 0.05$ and $v = 15$.
3. Generate aberration events at disease loci following the oncogenic pathway specified in Step 1. This is denoted by a $p \times 1$ vector U , where indices $\{s_1, \dots, s_K\}$ are disease loci. If disease locus s_i has an aberration ($U_{s_i} = 1$), we also assign aberrations to its neighboring loci $U_t = 1$, for $t \in [s_i - a_i, s_i + b_i]$, where a_i and b_i are independently sampled from Uniform[0, 30]. The rest of the elements in U are 0.
4. Combine the aberration events at disease loci and the background aberrations by assigning $X_i = 1$ if $U_i + Z_i > 0$ and 0 if $U_i = Z_i = 0$, for $i = 1, \dots, p$.

We set $n = 200$ and $p = 600$ to mimic the dimension of the real data set used in Section 4, so $V = \{1, \dots, 600\}$. We assume the 600 marker loci fall into 6 different chromosomes with 100 loci on each chromosome. We consider two different oncogenic pathway models: a chain shape and a tree shape (see Figure 1). Each model contains 6 aberration events: $M = \{A, B, C, D, E, F\}$. Without loss of generality, we assume these 6 aberrations are located in the middle of each chromosome, so the indices of A–F are $s_A = 50, s_B = 150, \dots, s_F = 550$, respectively. For any $u \in M$, $X_{s_u} = 1$ means aberration u occurs in the sample.

We evaluate the performance of the methods by two metrics: the false positive rate (FPR) and the false negative rate (FNR) of edge detection. Denote the true edge set $E = \{(u, v) | X_u \text{ and } X_v \text{ are conditionally dependent, } u \in V, v \in V\}$. We define a non-zero $\hat{\beta}_{rs}$ a false detection if its genome location indices (r, s) is far from the indices of any true edge:

$$|r - I_u| + |s - I_v| > 30, \quad \forall (u, v) \in E.$$

A cutoff value of 30 is used here because in the simulation setup (see Step 3) we set the maximum aberration size around the disease locus to be 30. Similarly, we define a conditionally dependent pair $(u, v) \in E$ is missed, if there is no non-zero β falling in the grey diamond. We then calculate FPR as the number of false detections divided by the total number of non-zero $\hat{\beta}_{rs}$, $r < s$; and calculate FNR as the number of missed $(u, v) \in E$ divided by the size of E .

Simulation I — Chain Model

For the chain model, aberrations A–F occur sequentially on one oncogenic pathway. The aberration frequencies and transitional probabilities along the oncogenic pathway are illustrated in Figure 1(a). The true conditionally dependent pairs in this model are

$$E = \{(s_A, s_B), (s_B, s_C), (s_C, s_D), (s_D, s_E), (s_E, s_F)\}.$$

Based on this chain model, we generated 50 independent data sets. The heatmap of one example data matrix 200×600 is shown in Supplemental Figure S-1. We then apply

LogitNet and SepLogit to each simulated data set for a series of different values of λ . Figure 2 shows the FPR and FNR of the two methods as a function of λ . For both methods, FPR decreases with λ while FNR increases with λ . Comparing the two methods, LogitNet clearly outperforms SepLogit in terms of FPR and FNR. For LogitNet, the average optimal sum of the error rates (FPR+FNR) across the 50 independent data sets is 0.014 (s.d.=0.029; FPR=0.014, FNR=0); while for SepLogit it is 0.211 (s.d.=0.144; FPR=0.116, FNR=0.028). Here the optimal sum of the error rates for each data set is defined as the smallest value for the sum of FPR+FNR that could be achieved and is obtained by examining a series of models corresponding to different values of λ for this data set. In Figure 2, it is the nadir of the graph for the sum of the FPR+FNR error rates in the bottom panel for each data set. Specifically, taking the data set shown in the Supplemental Figure S-1 as an example, the optimal sum of the error rates achieved by LogitNet on this data set is 0 (FPR= 0, FNR= 0), while the optimal sum of the error rates achieved by SepLogit is 0.549 (FPR= 0.549, FNR= 0). The corresponding two coefficient matrices $\hat{\beta}$ are illustrated in Figure 3. As one can see, there is a large degree of asymmetry in the result of SepLogit: 378 out of the 398 non-zero $\hat{\beta}_{rs}$ have inconsistent transpose elements, $\hat{\beta}_{sr} = 0$. On the contrary, by enforcing symmetry our proposed approach LogitNet has correctly identified all five true conditionally dependent pairs in the chain model. Moreover, the non-zero $\hat{\beta}_{rs}$'s plotted by red dots tend to be clustered within the grey diamonds. This shows that LogitNet indeed encourages group selection for highly correlated predictors, and thus is able to make good use of the spatial correlation in the data when inferring the edges.

We evaluated the two penalty parameter selection approaches: CV and BIC, for LogitNet. Table 1 summarizes selected λ values, FPR and FNR for CV and BIC. We can see that on average the selected λ values are close to the lowest point of the “U” shape curves of the sum of FPR+FNR in the bottom panel of Figure 2, suggesting the tuning strategies perform reasonably well. Comparing CV with BIC, the CV criterion tends to select smaller λ values which means larger models and thus more FPR and fewer FNR than the BIC. The average sum of the error rates for CV is 0.079, which is slightly smaller than that for BIC, 0.084.

We also investigated the impact of sample size on the performance of LogitNet and SepLogit by conducting a simulation experiment using $n = 50, 100, 150, 200$. For each n , we generated 10 independent data sets based on the Chain Model, and applied both methods to each simulated data set for a series of λ values. The average optimal sum of the error rates (FPR+FNR) for each n and each method are shown in Figure 4, and the corresponding FPR and FNR values are listed in Supplementary Table S-1. From the figure and the supplementary Table S-1 we can see that for both methods the sum of the error rates decreases as the sample size increases and LogitNet has much lower sum of the error rates than SepLogit for all n s. Interestingly, the relative improvements from increasing sample size are nearly the same for both methods.

Simulation II — Tree Model

For the tree model, we used the empirical mutagenic tree derived in Beerenwinkel et al. (2004) for an HIV data set. The details of the model are illustrated in Figure 1(b). The true conditionally dependent pairs in this model are

$$E = \{(s_A, s_B), (s_B, s_E), (s_A, s_C), (s_C, s_F), (s_A, s_D)\}.$$

The results of LogitNet and SepLogit for these data sets are summarized in Supplementary Figure S-2. Again, LogitNet outperforms SepLogit in terms of FPR and FNR. The average optimal sum of the error rates (FPR and the FNR) achieved by LogitNet across the 50

independent data sets is 0.163 (s.d.=0.106); while for SepLogit it is 0.331 (s.d.=0.160), twice as large as LogitNet. We also evaluated CV and BIC for LogitNet. The results are summarized in Table 1. Both CV and BIC give much higher FNRs under the tree model than under the chain model. This is not surprising as some transition probabilities between aberration events along the pathway are smaller in the tree model than in the chain model. As in the chain model, we also observe that BIC gives smaller FPR and higher FNR than CV, suggesting CV tends to select larger models and thus yields less false negatives but with more false positives in detecting edges.

4. Application to a Breast Cancer Data Set

In this section, we illustrate our method using genomic instability data from a breast cancer study (Loo et al., 2008). In this study, the GeneChip Mapping 10K Assay (Affymetrix, Santa Clara, CA) which contains 9670 SNPs with annotated genome locations was used to genotype 166 breast tumors and their constitutional normals. The genomic instability is measured by loss of heterozygosity (LOH), a common genomic alteration in breast tumors. An LOH event at a marker locus for a tumor is a binary variable, which is 1 if the locus is heterozygous in the constitutive normal DNA but homozygous in the tumor, and 0 if the locus is heterozygous in both the normal and tumor DNA. Even though 90% of the SNPs have call rates over 95%, 75% of these SNPs are homozygous in the constitutive normal samples, hence their LOH status can not be determined. To overcome this high rate of non-informativeness at individual locus, we binned the SNPs by cytogenetic bands (cytoband) and treated each cytoband as a genomic unit in our analysis. We then re-defined LOH for a cytoband as follows: the LOH status is 1 if at least 2 informative SNPs in the cytoband show LOH and 0 otherwise. In total, there were 765 cytobands covered by these SNPs. Cytobands which had noninformative rates above 20% or showed LOH in < 5 samples were removed. The the average LOH rate of the remaining 601 cytobands was 12.3%.

After binning the SNPs by cytobands, about 7.5% of LOH status still couldn't be determined. We used the multiple imputation approach to impute the unknown LOH status based on the conditional probability of LOH at the target cytoband given the available LOH status at adjacent cytobands. If the LOH status for both adjacent cytobands were also unknown, we imputed the genotype using only the marginal probability of the target cytoband. See Supplemental Appendix D for details of the multiple imputation algorithm.

We then generated 10 imputed data sets. We applied LogitNet on each of them taking into account the spatial correlation. It is worth noting that LOHs here are meant to detect somatic mutations in tumors and thus the spatial correlation among LOHs has a different pattern (for example, larger blocks) compared to the spatial correlation (linkage disequilibrium) for the inherited genetic variation, though in both cases the same SNP chips have been used. We used 10-fold CV for penalty parameter selection. To stabilize the penalty parameter

selection, we averaged the CV scores $\sum_{v=1}^V l(\hat{\mathcal{B}}_{uns}^{(v)}(\lambda) | X^{(v)})$ over 10 imputed datasets for a series of λ values and selected the value that gives the smallest average CV score. We then applied this λ value to each imputed data set and obtained the final model. The total number of edges inferred on each imputed data set is summarized in Supplementary Table S-2. We can see the number of edges selected is consistent across 10 imputed data sets. We then examined the consensus edges across different imputation data sets. Three edges are inferred in at least 4 imputed datasets. These are (11q22.3, 13q33.1), (11q24.2, 13q21.33), and (11q25, 13q14.11) and they occur in 6, 10, and 4 out of 10 imputed data sets, respectively. Particularly, the interaction between 11q24.2 and 13q21.33 has been consistently detected in all of the 10 imputation data sets. The LOH frequencies at these two cytobands are shown in Supplementary Table S-3. Cytoband 11q24.2 harbors the CHEK1 gene, which is an

important gene in the maintenance of genome integrity and a potential tumor suppressor. DACH1 is located on cytoband 13q21.33 and has a role in the inhibition of tumor progression and metastasis in several types of cancer (e.g., Wu et al., 2009). Both CHEK1 and DACH1 inhibit cell cycle progression through mechanisms involving the cell cycle inhibitor, CDKN1A. See Supplemental Figure S-3 for the pathway showing the interaction between CHEK1 on 11q24.2 and DACH1 on 13q21.33.

5. Final Remarks

In this paper, we propose the LogitNet method for learning oncogenic pathways using high dimensional binary data. When the size of maximum cliques ≤ 2 , the dependence parameters in the joint probability distribution of binary variables can be estimated by fitting a set of logistic regression models with a symmetric coefficient matrix. For high-dimension-low-sample-size data, this result is especially appealing as we can use sparse regression techniques to regularize the parameter estimation. We implemented a fast algorithm for obtaining the LogitNet estimator. For example, in our simulation study where there are 600 variables and 200 samples, it took less than 30 seconds (on average) to fit the LogitNet model on a server with two Dual/Core CPU 2.5 GHz and 6 GB RAM. With extensive simulation studies, we demonstrate that this method achieves good power in edge detection, and also performs favorably compared to an existing method by Ravikumar et al. (2009).

The likelihood function (3) is a summation of log-likelihood functions corresponding to p logistic regression models, and is not an exact likelihood function. A recent work by Hofling and Tibshirani (2009) showed that approximate pseudo-likelihood estimators were much faster to obtain than exact likelihood estimators and only had slightly worse accuracy. Our basic likelihood function (3) is exactly the same as the approximate pseudo-likelihood in the Hofling and Tibshirani paper. Even though we further developed the likelihood function to accommodate specific features of genomic instability data, we would still expect the same observation from Hofling and Tibshirani hold in our situation.

In LogitNet, the weighting scheme together with the enforcement of symmetry encourage a group selection effect, *i.e.*, highly spatially correlated variables tend to be in and out of the model simultaneously. It is conceivable that this group selection effect may be further enhanced by replacing the lasso penalty with the elastic net penalty proposed by Zou and

Hastie (2005) as $\lambda_1 \sum_{1 \leq r < s \leq p} |\beta_{rs}| + \lambda_2 \sum_{1 \leq r < s \leq p} \beta_{rs}^2$. The square ℓ_2 norm penalty may facilitate group selection within each regularized logistic regression. More investigation along these lines is warranted.

In our example using real data, the genomic instability data consist of binary variables indicating the LOH status along the genome. There are also other types of genomic instability data which may have multiple categories. For example, DNA copy number alteration at a locus (or a genomic region) is usually summarized into one of the following three status: loss, no change, and gain. In these cases, it would be more appropriate to treat variables as polytomous than binary and use polytomous logistic regression models. We want to point out that regardless of the data type, the data pre-processing is an important step in the data analysis, particularly for data mining such as our network analysis. Data preparation and filtering steps can take considerable amount of processing time. However, analyzing data that has not been carefully curated not only makes the discovery of networks more difficult but also can produce misleading results.

R package LogitNet is available to public from CRAN (<http://www.r-project.org/>).

5.1 Supplementary Materials

Web Appendices referenced in Sections 2, 3, and 4 are available under the Paper Information

link at the Biometrics website <http://www.biometrics.tibs.org>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to Drs. Peggy Porter and Lenora Loo for providing the genomic instability data set to us, which has motivated this methods development work. The authors are in part supported by grants from the National Institute of Health: R01GM082802 (Wang), R01AG14358 (Chao and Hsu), and P01CA53996 (Hsu).

References

- Beerenwinkel N, Rahnenfhrer J, Däumer M, Hoffmann D, Kaiser R, Selbig J, Lengauer T. Learning Multiple Evolutionary Pathways from Cross-Sectional Data. *Journal of Computational Biology*. 2005; 12:584–598. [PubMed: 16108705]
- Coop, A.; Ellis, MJ. The nature and development of cancer. In: Warrell, DA.; Cox, TM.; Firth, JD., editors. *Oxford Textbook of Medicine*. Oxford University Press; London: 2009.
- Drton M, Perlman MD. Model selection for Gaussian concentration graphs. *Biometrika*. 2004; 91:591–602.
- Edward, D. *Introduction to Graphical Modelling*. 2. New York: Springer; 2000.
- Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. *Annals of Statistics*. 2004; 32:407–499.
- Friedman J, Hastie T, Hofling H, Tibshirani R. Pathwise Coordinate Optimization. *Annals of Applied Statistics*. 2007a; 1:302–332.
- Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2007b; 9:432–441. [PubMed: 18079126]
- Friedman, J.; Hastie, T.; Tibshirani, R. *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Technical report. 2009. <http://www-stat.stanford.edu/jhf/ftp/glmnet.pdf>
- Fu W. Penalized Regressions: the Bridge vs the Lasso. *Journal of Computational and Graphical Statistics*. 1998; 7:397–416.
- Genkin A, Lewis DD, Madigan D. Large-scale Bayesian logistic regression for text categorization. *Technometrics*. 2007; 49:291–304.
- Hofling H, Tibshirani R. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*. 2009; 10:883–906.
- Joe H, Liu Y. A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regression. *Statistics & Probability Letters*. 1996; 31:113–120.
- Klein G, Klein E. Evolution of tumors and the impact of molecular oncology. *Nature*. 1985; 315:190–195. [PubMed: 2987696]
- Lauritzen, SL. *Graphical Models*. Clarendon Press; Oxford, United Kingdom: 1996.
- Li H, Gui J. Gradient Directed Regularization for Sparse Gaussian Concentration Graphs, with Applications to Inference of Genetic Networks. *Biostatistics*. 2006; 7:302–317.
- Loo L, Ton C, Wang YW, Grove DI, Bouzek H, Vartanian N, Lin MG, Yuan X, Lawton TL, Daling JR, Malone KE, Li CI, Hsu L, Porter P. Differential patterns of allelic loss in estrogen receptor-positive infiltrating lobular and ductal breast cancer. *Genes, Chromosomes and Cancer*. 2008; 47:1049–66. [PubMed: 18720524]
- Madigan D, York J. Bayesian graphical models for discrete data. *International Statistical Review*. 1995; 63:215–232.

- Meinshausen N, Buhlmann P. High dimensional graphs and variable selection with the Lasso. *Annals of Statistics*. 2006; 34:1436–1462.
- Newton MA, Gould MN, Reznikoff CA, Haag JD. On the statistical analysis of allelic-loss data. *Statistics in Medicine*. 1998; 17:1425–1445. [PubMed: 9695190]
- Peng J, Wang P, Zhou N, Zhu J. Partial Correlation Estimation by Joint Sparse Regression Model. *Journal of the American Statistical Association*. 2009a; 104:735–746. [PubMed: 19881892]
- Peng, J.; Zhu, J.; Bergamaschi, A.; Han, W.; Noh, DY.; Pollack, JR.; Wang, P. Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer. Technique Report. 2009b. <http://arxiv.org/abs/0812.3671>
- Ravikumar P, Wainwright M, Lafferty J. High-dimensional Ising model selection using l_1 -regularized logistic regression. *Annals of Statistics*. 2009 to appear.
- Schafer J, Strimmer K. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*. 2005; 4 Article 32.
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*. 1996; 58:267–88.
- Wu K, Katiyar S, Witkiewicz A, Li A, McCue P, Song L, Tian L, Jin M, Pestell RG. The Cell Fate Determination Factor Dachshund Inhibits Androgen Receptor Signaling and Prostate Cancer Cellular Growth. *Cancer Research*. 2009; 69:3347–3355. [PubMed: 19351840]
- Yuan M, Lin Y. Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*. 2007; 94:19–35.
- Zhang T, Oles F. Text categorization based on regularized linear classifiers. *Information Retrieval*. 2001; 4:5–31.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*. 2005; 67:301–320.

(a) The Chain Model

(b) The Tree Model

Figure 1.

(a) A chain shape oncogenic pathway; (b) A tree shape oncogenic pathway. Numbers in the yellow boxes are the marginal frequencies of the aberration in the disease population. The numbers on/below the arrows are the conditional probabilities between mutations along the oncogenic pathway. For example, consider the arrow from Mut A to Mut B in panel (a), the " $P = 0.6$ " above the arrow means: $\text{Prob}(\text{Mutation B happens} \mid \text{Mutation A happens}) = 0.6$; while the " $P' = 0.05$ " below the arrow means: $\text{Prob}(\text{Mutation B happens} \mid \text{Mutation A does not happen}) = 0.05$. The remark " $P' = 0.05$ for all arrows" in panel (b) suggests: $\text{Prob}(\text{The right side mutation happens} \mid \text{the left side mutation does not happen}) = 0.05$.

Figure 2. Results of LogitNet and SepLogit for the chain model in Figure 1(a). Each grey line represents one of the 50 independent data sets. The solid line is the mean curve with the two dashed lines represent mean $\pm$ one s.d.

(a) Estimated $\mathcal{B}$ by LogitNet at the λ giving the smallest sum of the error rates (FPR+FNR).

(b) Estimated $\mathcal{B}$ by SepLogit at the λ giving the smallest sum of the error rates (FPR+FNR).

Figure 3.

Results of LogitNet and SepLogit on the example data set shown in Supplemental Figure S-1. Each red dot represents a non-zero β_{rs} . Points in the grey diamond are deemed as correct detections. The dashed blue lines indicate the locations of aberration A–F.

Figure 4. Average optimal *sum of the error rates* (FPR + FNR) versus sample size for LogitNet and SeptLogit.

Table 1

Summary of FPR and FNR for LogitNet for using CV and BIC to choose optimal λ . Each entry is the mean (S.D.) over 50 independent data sets

	Chain Model			Tree Model		
	λ	FPR	FNR	λ	FPR	FNR
CV	12.5(1.23)	0.08(0.05)	0(0)	14.4(2.72)	0.06(0.06)	0.28(0.17)
BIC	20.5(2.50)	0.03(0.04)	0.06(0.10)	25.9(5.94)	0.02(0.04)	0.44(0.20)