A GRAPH REGULARIZATION BASED APPROACH FOR GENE SCORING AND GENE SET SELECTION AS APPLIED TO GLIOBLASTOMA MULTIFORME

STANFORD UNIVERSITY
Under the supervision of Prof. Daphne Koller

Zhenghao Chen
May 2012

(Daphne Koller) Principal Adviser

Abstract

One of the major challenges in cancer research is the problem of distinguishing driver mutations which drive neoplastic behavior and passenger mutations which have no functional purpose. The advent of high-throughput molecular assays have allowed tremendous amounts of data to be collected at fairly large sample sizes. However, it is difficult to approach the task of identifying a handful of critical genetic changes from the chaotic background alterations that is typically present in most tumor cells. This work presents a 2 stage gene scoring and gene set selection algorithm which is designed firstly to deal with missing information in a graceful manner and secondly to allow the user to leverage prior knowledge in the form of protein-interaction networks. The proposed approach is applied to the TCGA GBM dataset where it is shown that it robust to multiple forms of missing data and performs well at the task of gene prioritization and selection. Finally, the approach is used to analyze the Proneural patient cluster in GBM which shows poor response to aggressive treatment. A biologically plausible candidate pathway of 3 genes is found by our methods and is shown to be significantly associated with survival on an external validation dataset.

Acknowledgement

I would like to thank my advisor Prof. Daphne Koller for encouraging me and giving me the opportunity to do this thesis, it has truly been a learning experience for me.

I would also like to thank Alexis Battle, Andrew Beck, CS Foo and Sara Mostafavi for all their guidance, support, starter code and many many references.

Many thanks also to Pang Wei Koh and Hoon Cho for being my CS Honors buddies.

Thank you my parents and family for always being there for me.

Lastly, thanks Diane, I would not have finished this thesis without you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cancer is one of the major causes of death today, accounting for nearly one out of every four deaths in the United States. While great strides have been made in the development of more effective therapeutic options, prognosis remains bleak for certain forms of the disease.

Cancer is a disease of the genome and a better understanding of the genetic elements which drive neoplastic behavior is needed in order to devise more effective treatments. Rapid advancements in measurement technologies at the molecular scale have ushered in an age of high-throughput data gathering, the fundamental problem of identifying critical and meaningful mutations from tens to hundreds of thousands of candidate genes remains unsolved.

Large-scale data collection efforts such as The Cancer Genome Project (TCGA) have allowed researchers access to a wealth of data in multiple data modalities and projects such as the Human Protein Reference Database and Biogrid have also allowed hitherto unprecedented access to a significant fraction of the accumulated biological knowledge of our time.

In this work, I will present an approach towards scoring gene mutations and selecting groups of genes. This approach is heavily inspired by and adapts a class of random walk based graph regularization techniques which will be briefly presented in chapter 3.

The presented approach has the advantage of being robust to multiple forms of missing data and designed to leverage existing biological knowledge in the form of protein-interaction databases. This approach is applied to the TCGA GBM dataset and the HPRD protein-interaction network and some experimental results will be presented in chapter 4.

Chapter 2

Background and Related Work

2.1 Identifying Driver Pathways

An important problem in cancer research is the task of distinguishing driver mutations, which confer some form of selective advantage, from passenger mutations which do not contribute to neoplastic behavior (Greenman et al. (2007)). The identification of such driver mutations is a critical component in the development of effective therapeutic options. Examples of advancements in targeted treatments resulting from molecular characterization of cancers include the use of tyrosine kinase inhibitors such as geftinib and erlotinib for non-small cell lung cancer tumors harboring EGFR mutations (Kobayashi et al. (2005)) and the use of Herceptin in breast cancer (Baselga et al. (1998)).

A common class of approaches used to identify these driver mutations focuses on identifying recurrent mutations which occur in a significant subset of the patient population. Such approaches include copy-number based algorithms such as GISTIC which aim to identify regions significantly enriched in copy-number aberrations (CNAs) (Beroukhim et al. (2007)). However, copy-number based approaches suffer from the limitation that it is difficult to isolate a specific driver mutation from a given copy-number region which may contain many genes. Similar approaches which focus on identifying single genes which are mutated more often than expected by estimating the expected background mutation rate have also been proposed. (Ding et al. (2008); Sjöblom et al. (2006))

However, recent large-scale genomic studies of cancer have revealed a high degree of variability between individual tumors and efforts to identify the underlying drivers of oncogenesis have met with limited success because of this innate heterogenity of genetic alterations which is present even between morphologically similar tumors (Liang et al. (2005)).

While a diverse assortment of mutations affecting a large number of genes are known to occur in most cancer types, it is hypothesized that only a relatively small number of molecular pathways are perturbed in each case (Hahn et al. (2002); Vogelstein and Kinzler (2004)). The observed heterogeneity is thus due to the fact that a single pathway may be modified by mutations to any one of multiple genes along said pathway. There are therefore a potentially large combination of mutations which may result in carcinogenesis and driver mutations can be distributed over a relatively large number of genes making it difficult to identify driver mutations based solely on single gene frequency based approaches (Vandin et al. (2011)).

The need to identify driver pathways instead of driver mutations or driver genes motivates a number of approaches which first generate a list of candidate genes based on mutation data and subsequently identify known pathways or gene sets which contain a significant number of these candidate genes (Ding et al. (2008); Parsons et al. (2008); Sjöblom et al. (2006); McLendon et al. (2008)). However, such approaches have the disadvantage of having to rely on a fixed set of known pathways or gene groupings and may thus be unable to identify novel pathways which are not yet characterized.

Finally, a recent class of methods have been proposed which make use of regulatory or protein interaction networks in combination with mutation data to identify driver pathways such as (Vaske et al. (2010); Vandin et al. (2011)).

The work presented in this thesis falls under this category of network-based methods and a more detailed description of related work on such methods is presented in section 2.2.

2.2 Network Based Approaches

Gene and protein interactions have a natural representation as a graph structure in which nodes represent genes or gene products and edges represent interactions. In many cases, the term gene is also used to refer to its gene product and vice versa and we will refer to gene/protein-interaction networks in general as protein-interaction networks.

The use of protein-interaction networks has been used to incorporate existing information about relationships between genes/proteins in a variety of biological settings such as eQTL analysis (Suthram et al. (2008)), protein-disease associations (Navlakha and Kingsford (2010)) and protein function prediction (Ching et al. (2009); Li et al. (2010)).

In the context of cancer driver pathway discovery, network based approaches have been used to prioritize potential driver genes on the basis of various toplogical properties or novel graph-based statistics (Dezső et al. (2009); Jonsson and Bates (2006)). For example, Dezso et al. identified condition specific mutations using the topological features of genes within molecular networks while Taylor et al. identified robust gene biomarkers for breast cancer metastatsis using the loss of expression correlation with interaction neighbors as a gene statistic (Taylor et al. (2009)). Methods developed by Ideker et al. have also combined expression data with gene interaction networks in order to identify sub-networks within graphs which are differentially expressed in a group of interest and such methods have been applied to the problem of predicting breast cancer metastatis (Chuang et al. (2007)). Yet other approaches have tried to use various properties of genes in pathways such as the observation that mutations which affect the same pathway tend to be mutually exclusive to construct such pathways (Ciriello et al. (2012)).

In closely related work, Vandin et al. proposed HotNet, a method of identifying potential driver pathways using a heat diffusion process (Vandin et al. (2011)). HotNet first builds an influence graph where edges between genes are weighted according to the amount of "heat" which diffuses from one gene to the other through a gene interaction network. Each edge in this influence graph is then weighted by the number of mutations observed affecting the end-points of the edge. Edges below a threshold are subsequently pruned from the influence graph in order to obtain a set of connceted components corresponding to potential driver pathways. Using HotNet, Vandin et al. were able to identify potential driver

pathways containing genes which were not significantly mutated on a individual gene level but which were significant when aggregated into pathways (Vandin et al. (2011)).

Chapter 3

Approach

3.1 Overview

The proposed gene scoring approach comprises of 2 main stages.

In the first stage, copy-number, mRNA expression and methylation data is combined to produce a putative mutation score for each patient-gene pair. Imputation of missing data and the combination of different data modalities into a single integrated score is performed using a Gaussian mixture model (fig. 3.1).

Once putative mutation scores have been obtained for a set of genes, nodes in a protein-interaction network which correspond to these genes are "labeled" with their respective mutation scores. An iterative "label spreading" random walk is used to produce a final significance score for each node in the protein-interaction network (fig. 3.1).

3.2 Data Imputation and Multiple Modality Integration

Integrated analysis of multiple molecular data modalities has been shown to be useful in discovering key genomic alterations in cancer (Masica and Karchin (2011); Parsons et al. (2008)) and co-ordinated changes in copy-number and gene expression or methylation levels and gene expression have been shown to be powerful indicators of important genomic alterations (Tayrac et al. (2009); Hirsch et al. (2003)). However, an issue often encountered when dealing with multi-modal datasets is the issue of missing data modalities. A common

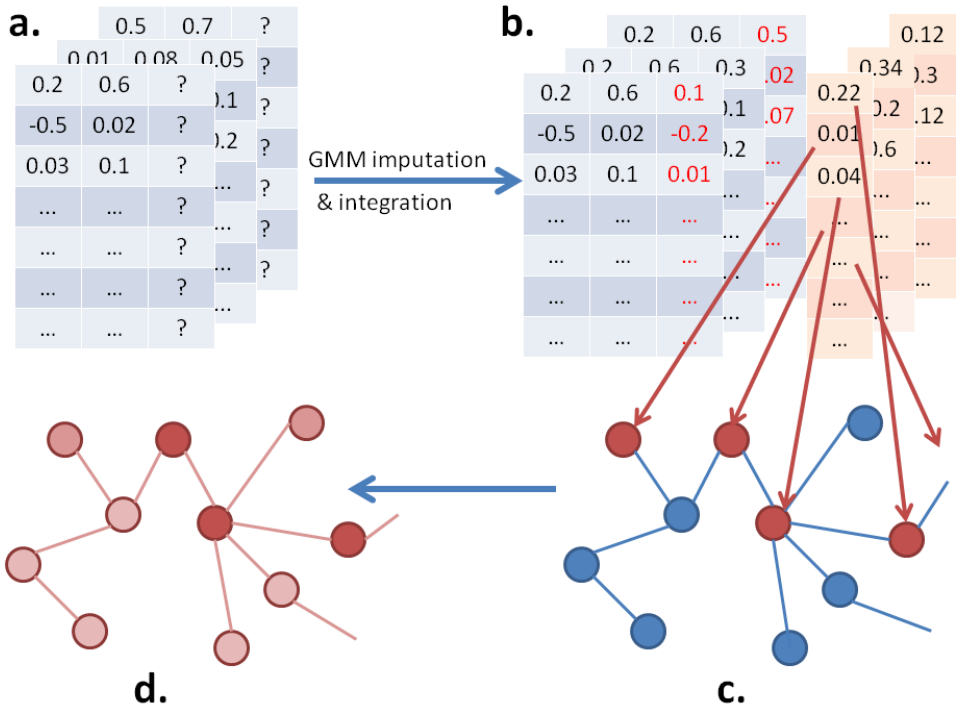

Figure 3.1: An overview of the proposed gene scoring method. a. Input consists of multiple observations (pages) of genes (rows) on different platforms (columns). Some columns may be entirely missing for certain observations. b. A Gaussian mixture model is fitted for each gene and used to impute the missing values (in red) and obtain an integrated mutation score for each gene (in orange). c. Nodes in a protein interaction network are associated with a mutation score (red) but some nodes may not be mapped with a mutation score (blue). d. A "label spreading" process is used to obtain a final significance score for all nodes in the interaction graph.

method used to deal with this problem is to discard observations which are incomplete, however, this greatly reduces the available sample size which may be undesirable. For example, in the TCGA glioblastoma multiforme dataset only roughly a fifth of available samples $(122/577)$ contain a full complement of molecular data while $(466/577)$ are missing one or less data modalities. This indicates that being able to deal with even a single missing data modality would allow studies to be conducted on significantly enlarged sample sizes.

Another problem related to working with multiple modalities is the problem of how to integrate data from different molecular assays and plaforms (Huang et al. (2011)), common methods employed include methods based on Singular Value Decomposition (SVD) (Berger et al. (2006)), Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) (Charlotte et al. (2010)).

In this work, we propose to deal with both problems together by using a Gaussian mixture model to model the vector of molecular assay measurements for each gene (e.g. CNA, mRNA expression). Each observation is assumed to be generated from a Gaussian distribution determined by a binary latent variable which represents whether the gene has undergone an amplification or deletion event. The mixing coefficients for each observation are later used to compute a combined mutation value for that gene observation.

Gaussian mixture models have previously been proposed for data imputation in expression microarrays wtih missing patient-gene pairs by modelling expression levels across genes as a mixture of Gaussians (Ouyang et al. (2004)). The same approach can be applied here to deal with missing data modalities by modelling the molecular profile of a single gene across multiple arrays with a mixture of Gaussians.

The motivation for using a Gaussian mixture model in the case of multi-modal data is three-fold:

1. Data imputation and data integration can be performed simultaneously.

2. Standard algorithms such as expectation-maximization (EM) can be used to learn model parameters and impute missing values.

3. Model fitting can be done in an unsupervised manner but prior assumptions can be incorporated in the form of the initial (soft) assignments of observations to Gaussian clusters.

In addition, restrictive parameterizations of the covariance structure can be used to enforce stronger prior beliefs regarding the underlying distribution (Ouyang et al. (2004)). In our case we choose to model the data with an unconstrained full covariance matrix which appears to yield reasonable results.

In order to encode the assumption that our Gaussian clusters correspond to amplification and deletion events, the initial mixing coefficients were computed based on a weighted sum of copy-number, mRNA expression and methylation status. Observations with high CNA, expression and low methylation values were assigned higher amplification cluster mixing coefficients while observations with low CNA, expression and high methylation levels were assigned higher deletion cluster mixing coefficients. This choice of initialization is also motivated by the intuition that co-ordinated changes in copy-number, expression and methylation status should receive a higher (more significant) mutation score (section 4.2).

The Gaussian mixture model is fit using the expectation-maximization (EM) algorithm (Aitkin and Rubin (1985)), derivation of the updates provided below can be found within the literature (Delalleau (2012)) and have been omitted for brevity.

Formally, let $x^i = \left(x_1^i, x_2^i, \ldots, x_m^i\right)$ represent the i-th observation of a gene and x_j^i represents the j-th component of the gene profile. Furthermore, let x_o^i denote the vector consisting of the observed components of x^i and x_m^i denote the vector of missing components of x^i.

The EM updates for the Gaussian mixture model with missing data are:

E-Step:

$$\hat{p}_{ij} = \mathcal{N}\left(x_o^i, \mu_o^j, \Sigma_{oo}^j\right)$$

$$p_{ij} = \frac{\hat{p}_{ij}}{\sum_{k=1}^{h} \hat{p}_{ik}}$$

Where p_{ij} is the mixing coefficient for the j-th Gaussian on the i-th observation and μ_o^j and Σ_{oo}^j denote the current estimates of the mean and covariance matrix of the j-th Gaussian for the observed components of x^i.

M-Step: Impute the missing values x_m^i with the conditional expectation of the missing components given the observed components weighted by the mixing coefficients p_{ik}.

$$x_m^i = \sum_{k=1}^{h} p_{ik} \left(\mu_k + \Sigma_{mo}^k \left(\Sigma_{oo}^k \right)^{-1} \left(x_o^i - \mu_o^k \right) \right)$$

Re-estimate the mean of each Gaussian in the mixture

$$\mu^j = \frac{\sum_{i=1}^n p_{ij} x^i}{\sum_{i=1}^n p_{ij}}$$

where the imputed value of x_j^i is used if it was originally missing.

The estimate for the covariance is

$$\Sigma^j = \frac{\sum_{i=1}^n p_{ij} \left(x^i - \mu^j \right) \left(x^i - \mu^j \right)^T}{\sum_i^n p_{ij}} + \tilde{\Sigma}^j$$

where $\tilde{\Sigma}_j$ is the covariance of the imputed values computed by initializing $\tilde{\Sigma}^j = 0$ and then updating for all x^i,

$$\left(\tilde{\Sigma}^j \right)_{mm} += \left(\frac{p_{ij}}{\sum_{l=1}^n p_{lj}} \Sigma_{mm}^j - \Sigma_{mo}^j \left(\Sigma_{oo}^j \right)^{-1} \Sigma_{om}^j \right)$$

where $\Sigma_{mo}^j = \left(\Sigma_{om}^j \right)^T$ denotes the sub-matrix of Σ^j taking the rows corresponding to the missing components of x^i and columns corresponding to the observed components of x^i.

In practice it was found that regularizing the covariance matrix by adding a small constant to the diagonal of the covariance matrix helped ensure faster convergence.

After data imputation is carried out, a combined mutation score for each gene observation is computed using the mixture coefficients learnt for each observation. Let $p_{A_i}^{(j)}$ and $p_{D_i}^{(j)}$ denote the mixing coefficients for the amplification and deletion Gaussian components for the i-th observation of the j-th gene.

The mean mutation score $z_j \in (0,1)$ of gene j is given by $z_j = \frac{1}{n}\sum_{i=1}^{n} 2 \cdot \left[\max\left(p_{A_i}^{(j)}, p_{D_i}^{(j)}\right) - 0.5\right]$, this score captures the notion of the average level of confidence that some significant modification has occurred to gene j in each of its n observations. We note that the score is rectified to be positive regardless of whether it was an amplification or deletion event.

3.3 Label Spreading and Graph Regularization

The "label spreading" process is part of a class of random walk methods known as label propagation algorithms. Label propagation is so named because it operates by taking as input a graph wherein a subset of nodes are assigned labels and these labels are "propagated" to the rest of the graph by simulating a random walk starting from the labeled nodes (Delalleau (2012)).

This class of label propagation algorithms was first presented in the context of semi-supervised learning (Zhu et al. (2003); Szummer and Jaakkola (2002)) in which nodes represent data examples and the edges are specified by a suitable metric measuring the similarity of two examples. For example, a common example of a similarity graph is the k-nearest neighbor graph where an example is connected to the k other examples which are closest to it by some distance metric with an edge of weight 1.

In the current context of gene scoring, the use of label propagation algorithms has been motivated by the assumption that genes which are "near" each other in the protein-interaction graph are also more likely to participate in common pathways. The presence of many significantly mutated genes in the neighborhood of an unknown gene may thus indicate that this unknown gene is part of a driver pathway and hence a potential oncogene.

Let W denote the adjacency matrix representing a molecular interaction network and $V = \{v_1, \ldots, v_G\}$ denote the set of genes present in the network. Let S denote the set of genes i which are mapped to an associated mutation score z_i (for example the mean mutation score defined in section 3.2).

The "label spreading" process is defined as (Delalleau (2012)):

Algorithm 3.1 The "label spreading" process

$Y_i^{(0)} = \begin{cases} z_i & \text{if } v_i \in S \\ 0 & \text{otherwise} \end{cases}$
$D_{ii} := \sum_j W_{ij}$
$\mathscr{L} := D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$
Repeat until convergence:
$Y^{(t+1)} := (1-\alpha)\mathscr{L}Y^{(t)} + \alpha Y^{(0)}$

The claim that this "label spreading" process converges for values of $\alpha \in [0,1)$ proceeds from the fact that, by construction, $\mathscr{L}$ has eigenvalues between -1 and 1 (Delalleau (2012)).

It follows from above that at convergence,

$$Y^{(\infty)} = \alpha (I - (1-\alpha)\mathscr{L})^{-1} Y^{(0)}$$

This "label spreading" random walk process is closely related to the minimization of a least-squares objective with a regularization term based on the graph structure specified by W (Delalleau (2012)).

$$C(Y) = \sum_{v_i \in S} (Y_i - z_i)^2 + \frac{\mu}{2} \sum_{j,l} W_{j,l} \left(\frac{Y_j}{\sqrt{D_{jj}}} - \frac{Y_l}{\sqrt{D_{ll}}} \right)^2 \tag{3.1}$$

the relationship between the two can be shown by minimizing $C(Y)$ by setting its derivative with respect to Y to 0. The minimizing assignment $Y^\star$ takes on the expression

$$Y^\star = ((1+\mu) I - \mu \mathscr{L})^{-1} Y^{(0)}$$

which is equivalent (modulo a factor of $\frac{1}{\alpha^2}$) to $Y^{(\infty)}$ if we let $\mu = \frac{1-\alpha}{\alpha}$.

Intuitively, we can think of minimizing the cost function $C(Y)$ as attempting to keep the predicted values of Y_i close to their associated mutation scores (for $v_i \in S$) while also trying to minimize the difference between the predicted values of neighboring nodes in the graph.

Rewriting the objective $C(Y)$

$$\begin{aligned} C(Y) &= \sum_{v_i \in S} (Y_i - z_i)^2 + \frac{\mu}{2} \sum_{j,l} W_{j,l} \left(\frac{Y_j}{\sqrt{D_{jj}}} - \frac{Y_l}{\sqrt{D_{ll}}} \right)^2 \\ &= \sum_{v_i \in S} (Y_i - z_i)^2 + \sum_{v_i \notin S} Y_i^2 + \mu \left(D^{-\frac{1}{2}} Y \right)^T \mathscr{L} \left(D^{-\frac{1}{2}} Y \right) \end{aligned}$$

we observe that other than the term $\mu \left(D^{-\frac{1}{2}} Y \right)^T \mathscr{L} \left(D^{-\frac{1}{2}} Y \right)$ which serves as a smoothness penalty over the labelings Y on the graph W, $C(Y)$ contains a separate regularization term $\sum_{v_i \notin S} Y_i^2$ which discourages unobserved nodes from taking values far from 0 (Delalleau (2012)). This corresponds to our prior that most gene nodes in the network do not harbor potential driver mutations and hence should be assigned a low final significance score.

3.4 Candidate Gene Subset Selection

This recasting of the "label spreading" process as an iterative approximation for the minimization of (3.1) provides a useful interpretation of the role that the free parameter α plays in the "label spreading" process. Smaller values of α correspond to enforcing stricter penalties for differences in values between adjacent nodes, at the expense of allowing Y to stray far from the original observed labels. On the other hand, larger values of α allow more leeway for sudden changes in label values but penalize deviations from the initial node labels more heavily in comparison.

This motivates an extension to the "label spreading" process which can be used to select a subset of genes within a network based on the scores assigned to them by the "label spreading" process.

Algorithm 3.2 Subset selection using the "label spreading" process

$$Y_i^{(0)} = \begin{cases} z_i & \text{if } v_i \in S \\ 0 & \text{otherwise} \end{cases}$$
$\alpha := 0.99$
$S_+ := V$
While $\alpha > 0.1$ and $S_+ \neq \emptyset$
  $D_{ii} := \sum_j W_{ij}$
  $\mathscr{L} := D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$
  Repeat until convergence:
    $Y^{(t+1)} := (1-\alpha)\mathscr{L}Y^{(t)} + \alpha Y^{(0)}$
  $S_+ := S_+ - \left\{ v_i : Y_i^{(t)} \leq Q\left(Y^{(t)}, \gamma\right) \right\}$
  $W := W_{S_+}$
  $Y^{(0)} := Y^{(0)}_{S_+}$
  $Y^{(t)} := Y^{(t)}_{S_+}$
  $\alpha := \beta \cdot \alpha$

Pseudo-code for iterative subset selection using the "label spreading" process is shown in algorithm 3.2 where $\beta \in (0,1]$, $\gamma \in (0,1]$ are free parameters, $Q(Y^{(t)}, \gamma)$ denotes the γ-th quantile of $Y^{(t)}$ and $W_{S_+}/Y^{(0)}_{S_+}/Y^{(t)}_{S_+}$ denote the rows and columns (where applicable) of $W/Y^{(0)}/Y^{(t)}$ corresponding to the genes in S_+.

The subset selection process maintains an active set S_+ of candidate genes which is iteratively refined by removing the lowest scoring fraction of genes after the "label spreading" algorithm has run to convergence. The fraction of genes removed at each iteration is determined by the free parameter γ. Each time the active set S_+ is pruned, the adjacency matrix W, the initial label values $Y^{(0)}$ and the current label vector $Y^{(t)}$ are updated to remove the discarded genes and a new graph laplacian $\mathscr{L}$ is computed and the "label spreading" process is repeated on the reduced network graph.

At the end of each iteration, the parameter α is shrunk by a constant factor β and the process iterates until α goes below a certain minimum threshold (set here to be 0.1).

Since a constant fraction γ of the active set is removed at each iteration and the number of iterations depends on the shrinking coefficient β, the resulting size of S_+ is determined jointly by these free parameters. In practice we find that the subset selection algorithm is

robust to a fairly wide range of values for γ near 0 and we set it to 0.1 for convenience leaving β the sole free parameter we use to determine the desired size of the candidate set.

The subset selection algorithm is designed to start off with a large value of α (weak regularization) and shrink it down as the active set of genes are reduced (stronger regularization). This is done in order to avoid assigning overly high scores to genes based solely on the graph topology in the beginning when the graph is still relatively densely connected with potentially irrelevant associations but still allow the graph regularization to effectively "propagate" the labels once the graph has been suitably restricted.

The subset selection algorithm is thus fairly sensitive to the choice of value for β. If β is set too conservatively (near 1), many genes which would be selected indirectly based on the mutation status of their interaction partners would be pruned before α was reduced sufficiently for their values to grow above the threshold. If β is set too aggressively (too near 0), a large number of false positives may result due to the insensitivity of the selection algorithm to the initial assignment of mutation scores.

A useful heuristic is to set the value for β to be the largest value possible such that the top n genes are declared significant (where significant is defined as being included in S_+ after subset selection has terminated) no more than $\frac{0.05}{n}$ of the time under a suitable null hypothesis, this heuristic was used to set the value of β used in the results of this work. The null distribution considered is generated on a per gene basis for gene i by permuting the labeled values of $Y^{(0)}$ and setting $Y_i^{(0)} = 0$.

Chapter 4

Results and Discussion

4.1 Dataset

The methods developed in chapter 3 are applied to the The Cancer Genome Atlas (TCGA) glioblastoma multiforme (GBM) dataset. The TCGA GBM dataset comprises of a total of 577 samples with associated surival and molecular assay data. Of these samples, 111 samples which did not have at least 2 out of 3 of the following molecular measurements were removed:

- Integrated mRNA expression values from the Affymetrix U133, Affymetrix Exon, and Agilent platforms
- Copy-number abberation data from the Affymetrix SNP6 array platform
- Methylation values from the Infinium HumanMethylation27 platform

This resulted in a final sample size of 466 samples of which 196 samples were complete and 270 contained one missing data modality.

Protein-protein interaction data from the Human Protein Reference Database (HPRD) was used to construct a protein-interaction network for the following experiments (Prasad et al. (2009)). The set of TCGA genes was filtered down to a set of 9104 genes which were common to both the TCGA and the HPRD datasets.

mRNA expression and copy-number abberation values were log-transformed prior to all experiments while all methylation values used are in the form of beta values.

4.2 Gaussian Mixture Data Integration Captures Co-ordinated Amplifications/Deletions

After initial preprocessing, a Gaussian mixture model was used to impute missing data and obtain an integrated putative mutation score as described in section 3.2. In order to validate that the putative mutation scores were providing a good metric of how often a gene was significantly mutated, the top 20 genes by mutation score were extracted and analyzed (fig. 4.1).

Within the top 20 genes, 8 genes *EGFR*, *METTL*1, *TSFM*, *CTDSP*2, *CDK*4, *SLC*25*A*3, *CHIC*2 were also in the top 20 genes when ranked by linear correlation between copy-number and expression values (Team). Furthermore, all but 5, namely, *SLC*25*A*3, *KRT*8, *KISS*1, *LGALS*3 and *KRT*6*A* were amplified in more than 10% of TCGA GBM cases.

Several genes in the top 20 genes such as *EGFR*, *CDK*4, *PDGFRA* and *MDM*2 are prominent oncogenes/tumor suppressor genes with known associations with GBM (Verhaak et al. (2010)). For example, Verhaak et al. characterized 4 sub-types of GBM based on *EGFR* and *PDGFRA* mutations (Verhaak et al. (2010)). *CDK*4 is a known oncogene commonly observed to be altered in GBM and *CDK*4 inhibitors have been recently shown to be effective in suppressing GBM tumor growth in mice (Michaud et al. (2010)). Similarly, *MDM*2 has been shown to be universally overexpressed in GBM tumors (regardless of *EGFR*/*TP*53 status) (HALATSCH et al. (2006)) .

The aforementioned results suggest that the putative mutation scores produced by Gaussian mixture models are correctly able to assign high scores to genes which are commonly mutated in a co-ordinated manner across many samples. However, a more detailed analysis of the top 20 genes shows 11 out of 20 show a highly correlated pattern of concordant amplification (fig. 4.1). Of these 11 genes, all 11 are located on the chromosome $12q13-15$, a region which is recurrently amplified in more than 15% of GBM tumors (Fischer et al.

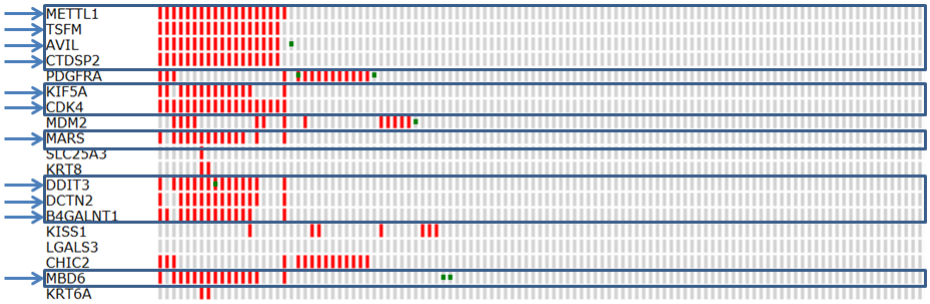

Figure 4.1: OncoPrint visualization (cBio Cancer Genomics Portal) of the top 20 genes by putative mutation score (excluding EGFR for visibility issues). Each column corresponds to a tumor sample, boxes shaded in red are significant amplification events (determined by GISTIC) while green boxes denote somatic mutation events. Blue arrows and boxes indicate genes on chromosome $12q13-15$.

(2008)) while all 11 of these genes are consistently amplified in a significant proportion of GBM samples, it is difficult to infer based on molecular data alone which of these 11 genes are the true target of $12q13-15$ amplification and hence probable driver genes or promising drug targets. For example, as mentioned above, $CDK4$ has been confirmed as a well known oncogene and the $CDK4/6$ inhibitor PD0332991 is currently undergoing clinical trials as a potential treatment for GBM (Wiedemeyer et al. (2010)). In contrast, MBD6 appears to be a typical passenger mutation with no known clinical association with GBM or other cancers.

On a similar note, both $CHIC2$ and $PDGFRA$ show highly co-ordinated amplification in a large subset of GBM tumors (fig. 4.1). However, while $PDGFRA$ has been identified as an important therapeutic gene of interest targeted by drugs such as imatinib mesylate (Gleevec), the main clinical importance of CHIC2 stems from the fact that deletion of the $CHIC2$ locus is an observable indication of $PDGFRA$ fusion with $FIP1L1$ (Pardanani et al. (2003)).

4.3 Graph Regularized Gene Rankings are Significantly Enriched for Important GBM genes

As noted above, while the putative mutation scores returned by our Gaussian mixture models are a good measure of the frequency of amplification and deletion events affecting a gene, it is difficult to distinguish important driver mutations and drug targets from passenger mutations based on the mutation status of a gene alone.

Furthermore, critical oncogenes or tumor suppressor genes may not always be clearly differentially expressed or amplified/deleted in a large enough subset of tumors to be identified directly from mutation data. For example, $RB1$ and $TP53$, two important tumor suppressor genes implicated in GBM (Goldhoff et al. (2012); Nakamura et al. (2001)) rank 208 and 2939 respectively out of 6165 genes based on their putative mutation score.

We therefore apply the subset selection algorithm described in sections 3.3 and 3.4 to select a set of potential driver genes and ranked them based on their final significance scores. The top 25 genes by final significance score are shown in table.

In contrast to the earlier list, almost all genes on this list are known to have direct associations with GBM, in order to give a rough gauge of the relevance of this list, a quick literature search was carried out using the GBM research resource site GBM base (GBM-Base). For each gene, the number of GBM related academic publications found on GBM base is reported and only 3 out of 25 of the listed genes do not have a GBM related publication associated with them on GBM Base.

As mentioned above, $TP53$ and $RB1$ are notable tumor suppressor genes implicated in a wide variety of cancers. The SRC proto-oncogene is a potential target gene in GBM treatment and *SRC* inhibitors such as dasatinib are currently being tested in GBM in phase II trials (de Groot and Milano (2009)).

$AKT1$has also been shown to mediate malignant tumor progression in GBM and is known to be frequently hyper-activated in GBM tumors, like *SRC* it is a promising drug target especially since it is known to be implicated in several other cancer types (Altomare and Testa (2005)).

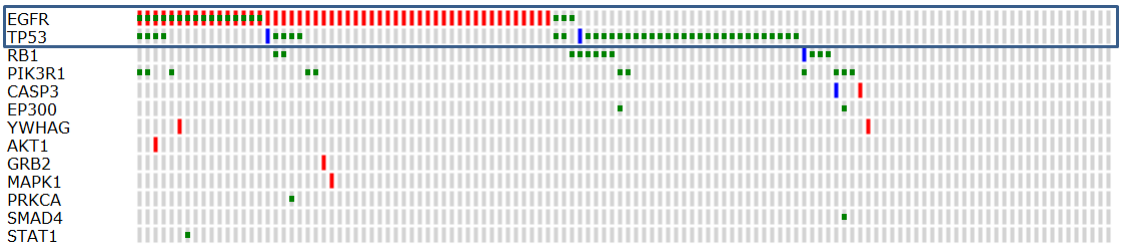

Figure 4.2: OncoPrint visualization (cBio Cancer Genomics Portal) of the subset of the top 25 genes (with at least 1 somatic mutation). Each column corresponds to a tumor sample, boxes shaded in red are significant amplification events and blue boxes are deletion events (both determined by GISTIC). Green boxes denote somatic mutation events. Blue box indicates EGFR and TP53, note the mutual exclusivity of EGFR amplification and TP53 mutation.

*MAPK*1is a key part of the *MAPK* signalling pathway and it has been suggested that GBM cell growth could be suppressed by inhibiting the *MAPK* pathway thereby down-regulating *CDK*4 (Loilome et al. (2009)).

Many other genes on the list are also known oncogenes which have been implicated in other forms of cancer such as *JUN* (sarcoma), *CREBBP* (leukemia, ALL/AML), *EP*300 (colorectal, breast, pancreatic) and *SMAD*4 (colorectal, pancreatic) (Trust).

While we have seen that many of the top genes which were selected by the subset selection algorithm are important gene targets in GBM, we note that the converse is also true, while our subset selection algorithm correctly selected *CDK*4 (218th) and *PDGRFA* (442nd) as being significant genes other genes with similar mutation scores such as *METTL*1, *MBD*6, *TSFM*, *AVIL*, *MARS*, *B*4*GALNT* and *CHIC*2 were not selected by our semi-supervised subset selection algorithm. The 3 remaining genes *DDIT*3 (702nd), *DCTN*2 (1399th) and *KIF*5*A* (2304th) were also ranked significantly lower than *CDK*4 or *PDGRFA* when ranked by the final significance score assigned by our semi-supervised "label spreading" process.

Upon closer examination of the distribution of somatic mutations among the top 25 genes (fig.4.2), it is clear that the failure to incorporate somatic mutations as one of our data modalities when calculating the putative mutation score resulted in $TP53$ receiving a low score since TP53 has a low frequency of amplification/deletion mutations even though it has the highest rates of somatic mutations among all genes in GBM. However, we note that incorporating prior biological knowledge about gene and protein interactions through our graph regularization framework allowed us to identify $TP53$ as a significant gene of interest (in fact the most significant gene of interest according to our subset selection algorithm) even without incorporating somatic mutation data into our framework.

4.4 Graph Regularized Gene Subset Selection is Robust to Missing Gene Observations

The main advantage of our semi-supervised "label spreading" approach over similar existing graph diffusion based approaches such as HotNet is the ability to transductively generalize to genes for which mutation data is not yet available (Vandin et al. (2011)).

In order to evaluate how robust our subset selection algorithm is to missing genes, we took the top 25 genes returned by our subset selection approach and simulated a scenario in which the mutation data for each of those genes was not observed. Effectively, this means that we zeroed out the putative mutation score of the "removed" gene before running our subset selection algorithm on this modified vector of mutation scores to see if we would still be able to recover the "removed" gene in our final gene set.

Recall that the free parameter β for our subset selection algorithm is chosen to ensure that for any gene i, under the null distribution generated by censoring gene i and randomly permuting the other mutation values, we would declare gene i significant less than $\frac{0.05}{25}$ of the time.

The results of this validation experiment are shown in table 4.1 where the red genes denote genes which we were **not** able to recover after censoring their mutation data. Overall, the subset selection algorithm was fairly robust to missing gene observations, it recovered 9 out of the top 10 genes and 15 out of the full list of 25.

Rank	Gene Name	GBM publications found on GBM Base	Rank	Gene Name	GBM publications found on GBM Base
1	*TP53*	161	13	*ESR1*	0
2	*EP300*	0	14	*STAT3*	19
3	*CTNNB1*	13	15	*JUN*	6
4	*RELA*	9	16	*SMAD2*	2
5	*CREBBP*	2	17	*YWHAG*	1
6	*GRB2*	3	18	*FYN*	3
7	*CSNK2A1*	2	19	*CASP3*	8
8	*SRC*	12	20	*PIK3R1*	6
9	*MAPK1*	19	21	*AKT1*	43
10	*SMAD3*	1	22	*RB1*	13
11	*PRKCA*	16	23	*SMAD4*	1
12	*EGFR*	96	24	*AR*	7
			25	*STAT1*	0

Table 4.1: Top 25 genes by final significance score returned by subset selection algorithm. Almost all genes have been studied in relation to GBM in peer reviwed publications. Genes in red are genes which were not recovered after mutation data was censored.

While this result seems to indicate that the subset selection algorithm is only able to recover genes which would be near the top of the list if they were observed, we note that the particular choice of β used in this case was unnecessarily conservative. In actual applications, β could be set to a lower value to apply a stronger prior especially if one suspects that there are a number of significant genes which are unobserved.

4.5 Graph Regularized Gene Subset Selection Identifies $MUS7-GALNT14-MUS13$ as Potentially Clinically Relevant Pathway

While much effort has been made in identifying the key driver mutations which drive GBM tumorigenesis in general, many major advances in the molecular characterization of GBM

have focused on the inherent heterogenity of the disease and developing a better understanding of the different subtypes of GBM (Verhaak et al. (2010); Mischel et al. (2003); Choe et al. (2002)).

Recent work by Verhaak and colleagues have identified 4 clinically distinct subtypes of GBM characterized by genetic alterations to PDGFRA, IDH1, EGFR and NF1 (Verhaak et al. (2010)). Further efforts to build on this result have resulted in the further subcharacterization of the proneural subtype which is characterized by poor response to aggressive/high intensity treatment. For example, Noushmehr et al identified a glioma-CpG island methylator phenotype which was used to further segregate the proneural subtype into G-CIMP positive and G-CIMP negative tumors (Noushmehr et al. (2010)).

In order to identify molecular characteristics of the proneural subtype as compared to non-proneural tumors, we considered the set of genes which were determined to be highly significant by our subset selection algorithm for the proneural group but not significant at all for the non-proneural group.

To do this, we compute 2 separate score statistics at the end of the data imputation and integration stage of our gene selection method,

$$z_j^P = \frac{1}{|P|}\sum_{i\in P} 2\cdot\left[\max\left(p_{A_i}^{(j)}, p_{D_i}^{(j)}\right) - 0.5\right]$$

$$z_j^{NP} = \frac{1}{n-|P|}\sum_{i\notin P} 2\cdot\left[\max\left(p_{A_i}^{(j)}, p_{D_i}^{(j)}\right) - 0.5\right]$$

where z_j^P denotes the mutation score of gene j for the subgroup of proneural tumors and z_j^{NP} is the corresponding score for the rest of the tumors. Graph regularized gene subset selection is then run separately on both sets of mutation scores to get a prioritized gene set for both the proneural and non-proneural tumors. The 200 most significant genes for the proneural subgroup are then searched for genes which were not determined to be significant in the case of the non-proneural subgroup.

Our search yielded just 11 such candidate genes of which the gene *GALNT*14 (5th) was the most significant gene in the proneural set of genes which was not found to be significant in the non-proneural group. Attention was further focused on *GALNT*14 since

it had another 2 known interaction partners *MUC*13 and *MUC*7 which were also included in our set of 11 candidates.

*GALNT*14 codes for a class of enzymes known as O-glycosyltransferases which glycosylate a large range of peptides including mucins, a class of glycoproteins. The overexpression of mucin coding genes is typical in various types of solid tumors such as *MUC*1 overexpression in breast and ovarian cancer and *MUC*13 overexpression in colorectal cancer (Zaretsky et al. (2006); Van Elssen et al. (2010); Walsh et al. (2007); Shimamura et al. (2005)) and it is hypothesized that abberations in glycosylation changes are responsible for tumor resistance to certain anti-tumor agents (Varki (1999); Park et al. (2010)). Similar O-glycosyltransferases have been proposed as targets for cancer therapies for example *GALNT*6 knockdown has been shown to inhibit breast cancer tumor growth. Recent studies have also shown that glycosylation inhibitors can enhance the effects of 5-FU effect (Shah (2009)).

This led us to explore the possibility that *GALNT*14, *MUC*7 and *MUC*13 could be a possible pathway driving resistance of the proneural group to high intensity treatment (defined here as concurrent radiation and chemotherapy).

To test this hypothesis, we performed a multivariate Cox regression on the putative mutation scores of *MUC*13, *MUC*7 and *GALNT*14 against survival data on the TCGA GBM dataset. It was found that *MUC*13, *MUC*7 and *GALNT*14 were significantly associated with survival (p-value: 0.029).

As a form of external validation, we obtained a separate unseen dataset from Horvath et al consisting of gene expression levels and survival data for 65 GBM patients from UCLA. Although *MUC*7 expression was not available in this separate validation dataset however *MUC*13 and *GALNT*14 gene expression levels were also found to be significantly associated with survival (p-value: 0.024). Data and Statistical R Code: Analysis of Oncogenic Signaling Networks in Glioblastoma: Identification of ASPM as a Novel Target

While these results are far from definitive, such results seem to suggest that *GALNT*14 and its glycosylation targets *MUC*7 and *MUC*13 represent potential targets for further study into the variability in response to cancer treatment present in GBM.

Chapter 5

Conclusion

5.1 Conclusion

This work presents a 2 stage gene scoring and gene set selection algorithm which is able to deal with missing data at different scales from single entries to modalities and entire genes in a graceful manner. The approach presented in this work utilizes ideas from graph regularization methods developed in the context of semi-supervised learning to allow users to leverage prior biological knowledge in the form of protein interaction networks and thereby generalize even to genes which are unobserved in the input data.

This approach was applied to the TCGA GBM dataset where it was able to identify $TP53$ as an important GBM gene even though most mutations which occur on $TP53$ are somatic mutations which was not provided to the algorithm.

The approach was also able to identify important GBM associated genes from a large amplified chromosomal region at $12q13$ and correctly determine the significance of most passenger mutations.

Simulations of situations with missing data also showed that this approach is fairly robust to the censoring of all observations across all platforms of important genes and is able to recover more than half of the top 25 genes in the GBM dataset.

Finally, the approach was used to attempt to identify molecular signatures of the proneural GBM subtype discovered by Verhaak et al (Verhaak et al. (2010)). The proposed approach was able to produce a short candidate list of 11 genes in which 3 genes $GALNT14$,

*MUC*7 and *MUC*13 were known to be interaction partners. A literature search revealed that those genes were often implicated in determining drug response in other cancer types and Cox regression both on the TCGA GBM dataset and an external validation dataset indicated significant gene associations with survival.

Bibliography

M. Aitkin and D. Rubin. Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 67–75, 1985.

D. Altomare and J. Testa. Perturbations of the akt signaling pathway in human cancer. *Oncogene*, 24(50):7455–7464, 2005.

J. Baselga, L. Norton, J. Albanell, Y. Kim, and J. Mendelsohn. Recombinant humanized anti-her2 antibody enhances the antitumor activity of paclitaxel and doxorubicin against her2 overexpressing human breast cancer xenografts. *Cancer Research*, 58(13):2825, 1998.

J. Berger, S. Hautaniemi, S. Mitra, and J. Astola. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 3(1):2, 2006.

R. Beroukhim, G. Getz, L. Nghiemphu, J. Barretina, T. Hsueh, D. Linhart, I. Vivanco, J. Lee, J. Huang, S. Alexander, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences*, 104(50):20007, 2007.

T. cBio Cancer Genomics Portal. URL `http://www.cbioportal.org/`.

S. Charlotte, L. Henrik, F. Thoas, and F. Magnus. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. 2010.

W. Ching, L. Li, Y. Chan, and H. Mamitsuka. A study of network-based kernel methods on protein-protein interaction for protein functions prediction. In *International Symposium on Optimization and Systems Biology*, volume 11, pages 25–32. APORC., 2009.

G. Choe, J. Park, L. Jouben-Steele, T. Kremen, L. Liau, H. Vinters, T. Cloughesy, and P. Mischel. Active matrix metalloproteinase 9 expression is associated with primary glioblastoma subtype. *Clinical cancer research*, 8(9):2894–2901, 2002.

H. Chuang, E. Lee, Y. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(1), 2007.

G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, 22(2):398–406, 2012.

J. de Groot and V. Milano. Improving the prognosis for patients with glioblastoma: the rationale for targeting src. *Journal of neuro-oncology*, 95(2):151–163, 2009.

O. Delalleau. Apprentissage machine efficace: théorie et pratique. 2012.

Z. Dezső, Y. Nikolsky, T. Nikolskaya, J. Miller, D. Cherba, C. Webb, and A. Bugrim. Identifying disease-specific genes based on their topological significance in protein networks. *BMC systems biology*, 3(1):36, 2009.

L. Ding, G. Getz, D. Wheeler, E. Mardis, M. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. Muzny, M. Morgan, et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069–1075, 2008.

U. Fischer, A. Keller, P. Leidinger, S. Deutscher, S. Heisel, S. Urbschat, H. Lenhof, and E. Meese. A different view on dna amplifications indicates frequent, highly complex, and stable amplicons on 12q13-21 in glioma. *Molecular Cancer Research*, 6(4):576, 2008.

GBMBase. URL `http://beta.gbmbase.org`.

P. Goldhoff, J. Clarke, I. Smirnov, M. Berger, M. Prados, C. James, A. Perry, and J. Phillips. Clinical stratification of glioblastoma based on alterations in retinoblastoma tumor suppressor protein (rb1) and association with the proneural subtype. *Journal of Neuropathology & Experimental Neurology*, 71(1):83, 2012.

C. Greenman, P. Stephens, R. Smith, G. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, 2007.

W. Hahn, R. Weinberg, et al. Modelling the molecular circuitry of cancer. *Nature Reviews Cancer*, 2(5):331–341, 2002.

M. HALATSCH, U. SCHMIDT, A. UNTERBERG, and V. VOUGIOUKAS. Uniform mdm2 overexpression in a panel of glioblastoma multiforme cell lines with divergent egfr and p53 expression status. *Anticancer research*, 26(6B):4191–4194, 2006.

F. Hirsch, M. Varella-Garcia, P. Bunn Jr, M. Di Maria, R. Veve, R. Bremnes, A. Barón, C. Zeng, and W. Franklin. Epidermal growth factor receptor in non–small-cell lung carcinomas: Correlation between gene copy number and protein expression and impact on prognosis. *Journal of clinical oncology*, 21(20):3798–3807, 2003.

N. Huang, P. Shah, and C. Li. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in Bioinformatics*, 2011.

P. Jonsson and P. Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297, 2006.

S. Kobayashi, T. Boggon, T. Dayaram, P. Jänne, O. Kocher, M. Meyerson, B. Johnson, M. Eck, D. Tenen, and B. Halmos. Egfr mutation and resistance of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 352(8):786–792, 2005.

L. Li, W. Ching, Y. Chan, and H. Mamitsuka. On network-based kernel methods for protein-protein interactions with applications in protein functions prediction. *Journal of Systems Science and Complexity*, 23(5):917–930, 2010.

Y. Liang, M. Diehn, N. Watson, A. Bollen, K. Aldape, M. Nicholas, K. Lamborn, M. Berger, D. Botstein, P. Brown, et al. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proceedings of the National Academy of Sciences of the United States of America*, 102(16):5814, 2005.

W. Loilome, A. Joshi, C. Ap Rhys, S. Piccirillo, V. Angelo, G. Gallia, and G. Riggins. Glioblastoma cell growth is suppressed by disruption of fibroblast growth factor pathway signaling. *Journal of neuro-oncology*, 94(3):359–366, 2009.

D. Masica and R. Karchin. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Research*, 71(13): 4550, 2011.

R. McLendon, A. Friedman, D. Bigner, E. Van Meir, D. Brat, G. Mastrogianakis, J. Olson, T. Mikkelsen, N. Lehman, K. Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

K. Michaud, D. Solomon, E. Oermann, J. Kim, W. Zhong, M. Prados, T. Ozawa, C. James, and T. Waldman. Pharmacologic inhibition of cyclin-dependent kinases 4 and 6 arrests the growth of glioblastoma multiforme intracranial xenografts. *Cancer research*, 70(8): 3228, 2010.

P. Mischel, R. Shai, T. Shi, S. Horvath, K. Lu, G. Choe, D. Seligson, T. Kremen, A. Palotie, L. Liau, et al. Identification of molecular subtypes of glioblastoma by gene expression profiling. *Oncogene*, 22(15):2361–2373, 2003.

M. Nakamura, Y. Yonekawa, P. Kleihues, and H. Ohgaki. Promoter hypermethylation of the rb1 gene in glioblastomas. *Laboratory investigation*, 81(1):77–82, 2001.

S. Navlakha and C. Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063, 2010.

H. Noushmehr, D. Weisenberger, K. Diefes, H. Phillips, K. Pujara, B. Berman, F. Pan, C. Pelloski, E. Sulman, K. Bhat, et al. Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell*, 17(5):510–522, 2010.

M. Ouyang, W. Welsh, and P. Georgopoulos. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, 20(6):917–923, 2004.

A. Pardanani, R. Ketterling, S. Brockman, H. Flynn, S. Paternoster, B. Shearer, T. Reeder, C. Li, N. Cross, J. Cools, et al. Chic2 deletion, a surrogate for fip1l1-pdgfra fusion, occurs in systemic mastocytosis associated with eosinophilia and predicts response to imatinib mesylate therapy. *BLOOD-NEW YORK-*, 102(9):3093–3096, 2003.

J. Park, T. Nishidate, K. Kijima, T. Ohashi, K. Takegawa, T. Fujikane, K. Hirata, Y. Nakamura, and T. Katagiri. Critical roles of mucin 1 glycosylation by transactivated polypeptide n-acetylgalactosaminyltransferase 6 in mammary carcinogenesis. *Cancer research*, 70(7):2759, 2010.

D. Parsons, S. Jones, X. Zhang, J. Lin, R. Leary, P. Angenendt, P. Mankoo, H. Carter, I. Siu, G. Gallia, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science's STKE*, 321(5897):1807, 2008.

T. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, et al. Human protein reference database 2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772, 2009.

A. Shah. Evaluation of mucin glycosylation as a barrier to drug uptake: a quantitative approach. 2009.

T. Shimamura, H. Ito, J. Shibahara, A. Watanabe, Y. Hippo, H. Taniguchi, Y. Chen, T. Kashima, T. Ohtomo, F. Tanioka, et al. Overexpression of muc13 is associated with intestinal-type gastric cancer. *Cancer science*, 96(5):265–273, 2005.

T. Sjöblom, S. Jones, L. Wood, D. Parsons, J. Lin, T. Barber, D. Mandelker, R. Leary, J. Ptak, N. Silliman, et al. The consensus coding sequences of human breast and colorectal cancers. *science*, 314(5797):268–274, 2006.

S. Suthram, A. Beyer, R. Karp, Y. Eldar, and T. Ideker. eqed: an efficient method for interpreting eqtl associations using protein networks. *Molecular systems biology*, 4(1), 2008.

M. Szummer and T. Jaakkola. Information regularization with partially labeled data. *Advances in Neural Information processing systems*, 15:1025–1032, 2002.

I. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. Wrana. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology*, 27(2):199–204, 2009.

M. Tayrac, A. Etcheverry, M. Aubry, S. Saïkali, A. Hamlat, V. Quillien, A. Treut, M. Galibert, and J. Mosser. Integrative genome-wide analysis reveals a robust genomic glioblastoma signature associated with copy number driving changes in gene expression. *Genes, Chromosomes and Cancer*, 48(1):55–68, 2009.

T. T. G. Team. Glioblastoma multiforme: Correlations between copy number and mrna expression. URL `http://gdac.broadinstitute.org/runs/analyses__2012_03_21/reports/cancer/GBM/index.html`.

T. W. Trust. URL `http://www.sanger.ac.uk/genetics/CGP/Census/`.

C. Van Elssen, P. Frings, F. Bot, K. Van de Vijver, M. Huls, B. Meek, P. Hupperets, W. Germeraad, and G. Bos. Expression of aberrantly glycosylated mucin-1 in ovarian cancer. *Histopathology*, 57(4):597–606, 2010.

F. Vandin, E. Upfal, and B. Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522, 2011.

A. Varki. *Essentials of glycobiology*. Cold Spring Harbor Laboratory Pr, 1999.

C. Vaske, S. Benz, J. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

R. Verhaak, K. Hoadley, E. Purdom, V. Wang, Y. Qi, M. Wilkerson, C. Miller, L. Ding, T. Golub, J. Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.

B. Vogelstein and K. Kinzler. Cancer genes and the pathways they control. *Nature medicine*, 10(8):789–799, 2004.

M. Walsh, J. Young, B. Leggett, S. Williams, J. Jass, and M. McGuckin. The muc13 cell surface mucin is highly expressed by human colorectal carcinomas. *Human pathology*, 38(6):883–892, 2007.

W. Wiedemeyer, I. Dunn, S. Quayle, J. Zhang, M. Chheda, G. Dunn, L. Zhuang, J. Rosenbluh, S. Chen, Y. Xiao, et al. Pattern of retinoblastoma pathway inactivation dictates response to cdk4/6 inhibition in gbm. *Proceedings of the National Academy of Sciences*, 107(25):11501, 2010.

J. Zaretsky, I. Barnea, Y. Aylon, M. Gorivodsky, D. Wreschner, and I. Keydar. Muc1 gene overexpressed in breast cancer: structure and transcriptional activity of the muc1 promoter and role of estrogen receptor alpha (erα) in regulation of the muc1 gene expression. *Molecular cancer*, 5(1):57, 2006.

X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 20, page 912, 2003.