Survival Outcome Prediction for Cancer Patients based on Gene Interaction Network Analysis and Expression Profile Classification

Final Project Report
Alexander Herrmann
Advised by Dr. Andrew Gentles

December 9, 2014

Special Notice: This project is undertaken for both CS 229 and CS 224W with instructor consent. Parts of the project used for CS 224W and CS 229, respectively, will be appropriately identified.

1 Introduction

The classification of gene expression profiles for diagnosis and survival outcome prediction for cancer patients is an important yet challenging task in systems biology. Conventional analysis of gene expression profiles generally succeeds in identifying distinct cancer cell types as well as healthy tissues. However, the reliable prediction of cancer subtypes and survival outcome parameters for patients suffering from the same type of cancer remains elusive because the expression patterns for the majority of genetic markers are not sufficiently distinct. In order to focus on the most informative marker subsets and allow subclassification of gene expression patterns, novel feature selection techniques for supervised machine learning methods must be developed. This project aims to address this problem by reconstructing a gene interaction network from steady-state gene expression profiles, characterizing certain subsets of genes according to their topological properties and verifying these subsets' utility as feature selection filters for a variety of classification algorithms.

2 Project Description and Related Work

For CS 229 the goal of the project was to subject a dataset of 13,334 genes and 1,106 lung adenocarcinoma patient samples to a number of supervised machine learning algorithms, e.g. logistic regression, softmax regression and Gaussian discriminant analysis (GDA) in order to classify gene expression profiles according to cancer subtype and survival outcome. The brute force approach was to apply each of the aforementioned learning algorithms to the entire dataset without feature selection. A more nuanced but nonetheless conventional approach was to adjust the number of features to address the tradeoff between bias and variance. Feature selection was performed by ranking genes according to the correlation between their expression values and survival outcome parameters, and choosing those with the highest correlation as machine learning features. For this purpose *Mutual Information (MI)* was used as a correlation measure. Both brute force and conventional feature selection were then tested and compared. The patient dataset was subdivided in a random but balanced fashion so as to allow training and testing on independent sets. After training each of the algorithms, test success rates for the different approaches were quantified for each algorithm and juxtaposed against each other.

For CS 224W, the goal was to provide a novel feature selection approach for machine learning by taking into account the topological properties of the gene interaction network whose states are described by the aforementioned gene expression profiles. One topological feature selection approach based on PageRank has been proposed in [1]. In this approach node j in the gene interaction network is ranked according to the following iteration recipe:

$$r_j^n = (1-d)c_j + d\sum_{i=1}^{N} \frac{A_{ij} r_i^{n-1}}{k_i}$$

where A is the adjacency matrix, k_i the degree of node i and c_j is the Pearson correlation value of gene j with the output measured in terms of survival time. The parameter d weighs how much influence is attributed to the Pearson correlation with the output and how much to the topology of the network via PageRank. Before a filter is applied, the optimal value of parameter d must be estimated. This is done by splitting the training set into two subsets for training and testing purposes. Several trials are performed for different values of d and the value with the best predictive result is chosen.

However, it is not obvious why a high PageRank score should be the determining criterion for feature selection as opposed to some other topological property with a different value range. It may therefore be worthwhile to choose a different approach and check other topological node properties for their utility as feature selection parameters. The first step in this project was to reconstruct the underlying gene interaction network with the network inference algorithm ARACNE [2]. This method uses steady-state gene expression data to predict direct interactions between gene pairs based on MI correlation and the data processing inequality (DPI). The adjacency matrix produced by ARACNE was then processed by SNAP to compute a set of $p = 18$ topological network properties for each node i. These values were used to embed each node as a coordinate point in a p-dimensional space in a manner similar to that in [3]. Subsequently, K-means clustering was applied to these embedded points to discover prominent gene clusters in p-space. Each cluster was then used as a gene pool for the MI feature selection filter to act upon. The performance of each pool was measured by applying the machine learning routines to the test set and comparing the resulting success rates with those of the brute force and the conventional feature selection approach.

3 Dataset

The dataset used in this project represents an agglomeration of clinical data from different patient groups with lung adenocarcinoma collected over many years. With 13,334 genes and 1,106 patients this data set is rather large, given that the usual number of patients used for machine learning algorithms in cancer biology is an order of magnitude lower. For machine learning purposes, the dataset was randomly split into a training and a test set of 563 and 543 patients, respectively, such that age, gender and other patient parameters remained balanced to prevent skewed results due to either of these factors.

The data is comprised of two parts. The first part is a table of gene expression values, in which rows correspond to genes and columns to patients. Sporadically occurring missing values were imputed with the function knnimpute in MATLAB using the 15 nearest neighboring columns as measured by Euclidean distance, a standard procedure in bioinformatics. This adjusted dataset is the starting point for the ARACNE algorithm from which the adjacency matrix is inferred. For performance reasons, ARACNE was executed with a set of one hundred patients, the minimum admissible number for the algorithm to produce a useful reconstruction of the gene interaction network.

The second part of the dataset consists of a file containing patient IDs as rows and survival parameters and other clinical information as columns. There are two kinds of survival parameters considered in this project. The first measures the patient status as to whether the patient is dead or alive after 5 years. The second measures the cancer subtype, a multivalued measure assuming the values $\{I, II, III, IV\}$ representing increasing severity and aggressiveness of the cancer.

4 Network Inference Algorithm and Other Methods

4.1 ARACNE

ARACNE's reconstruction of the gene interaction network is not based on time series data of gene expression profiles. Instead, stationary gene expression profiles are subjected to correlation analysis in order to infer the probability of direct links between gene pairs. ARACNE is based on the following definition of irreducible direct interactions between gene pairs. If $\{g_i\}$ represents the stationary expression of genes for $i = 1, \ldots, n$, the authors in [2] define the joint probability distribution for the stationary gene expression as

$$P(\{g_i\}) = \frac{1}{Z} e^{-\Sigma_i^N \phi_i(g_i) - \Sigma_{i,j}^N \phi_{ij}(g_i,g_j) - \Sigma_{i,j,k}^N \phi_{ijk}(g_i,g_j,g_k) - \ldots}$$

where Z is the normalization constant. Truncating the Hamiltonian $H = -\Sigma_i^N \phi_i(g_i) - \Sigma_{i,j}^N \phi_{ij}(g_i, g_j) - \Sigma_{i,j,k}^N \phi_{ijk}(g_i, g_j, g_k) - ...$ after the second order term, ARACNE declares a gene pair $\{g_i, g_j\}$ non-interacting if the second-order potential term $\phi_{ij} = 0$. Higher order potentials ϕ are intentionally neglected because the occurrence of pure higher order interactions, i.e. interactions that produce a higher-order term without generating a second-order term, is a rare event. ARACNE proceeds by examining all possible pairwise interactions and then deleting all those which do not pass through its filters. The are two filters:

(1) *Mutual information (MI) filter*: MI is an information theoretic relatedness measure employed to remove all links $\{i, j\}$, for which the mutual information estimate $I(g_i, g_j) < I_0$, i.e. for which the calculated mutual information approximation falls below a certain threshold I_0 determined by the user either directly or by setting a p-value. It should be noted that, in contrast to exactly computed MI, its estimate is usually nonzero but small for independent variables, which explains why a finite-size threshold is needed to filter out weakly correlated interaction pairs. MI is well-defined for continuous variables so that there is no need to discretize the gene expression data before applying it. For continuous random variables X and Y we have:

$$I(X,Y) = \int_Y \int_X p(x,y) \cdot log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$$

However, MI cannot be computed exactly from a finite dataset of gene-patient-samples and thus has to be estimated by conditioning it on the given data. Given a set of N genes and M patients the values x_i and y_i of genes x and y sampled from patients $i \in \{1, \ldots, M\}$ serve to approximate the mutual information between genes x and y as follows:

$$I(x,y) = \frac{1}{M}\sum_i log\left(\frac{f(x_i, y_i)}{f(x_i)f(y_i)}\right) \text{ with } f(x,y) = \frac{1}{M}\sum_i \frac{G\left(\frac{|(x,y)-(x_i,y_i)|}{h}\right)}{h^2}$$

Here, $f(x)$ and $f(y)$ are the marginals of $f(x, y)$. G is the bivariate normal distribution and h is the kernel width, a scaling parameter used for fitting the normal distribution to the given data. The kernel width must be estimated by the algorithm during preprocessing, which constitutes the approximation step. This parameter is precomputed for the entire ensemble in the first version of ARACNE before any MI is calculated. A newer and faster version of ARACNE, which is also implemented in the software package Workbench 2.5.1 used for this project, uses an estimator of conditional mutual information based on adaptive partitioning to circumvent the computationally intensive calculation of the kernel width. This latter option will be used for the project as it delivers faster performance without sacrificing precision.

(2) *DPI filter*: Steady-state network inference algorithms are notoriously susceptible to inferring false positives, i.e. non-existent links between pairs of nodes, since strong correlations between two genes' activity patterns can result from indirect long-range interactions in the network. In order to remove inferred but non-existent direct links, ARACNE examines all triangles $\{g_i, g_j, g_k\}$ and, using the data processing inequality (DPI) $I(g_1, g_3) \leq min\{I(g_1, g_2), I(g_2, g_3)\}$, marks the weakest link in each triad for deletion, subsequently disposing of all marked links and with them of all three-gene-loops in the network. This amounts to deleting most of the interactions for which $\phi_{ij} = 0$. As the calculated MI is an estimate, ARACNE provides an additional parameter called DPI tolerance which specifies the sampling error. The higher the tolerance the fewer edges are deleted. A DPI tolerance of 0.0 was set for the network reconstruction in this project.

ARACNE's complexity is $O(N^3 + N^2M^2)$ with N and M being the number of nodes and patient samples, respectively. The N^3 cube term results from the DPI filter whereas the N^2M^2 term is due to the MI filter.

4.2 Machine Learning Algorithms

In order to embed the nodes in p-space, 18 topological parameters were chosen and computed for each node i in the network. These parameters were selected as follows:

(1) degree
(2) average degree of neighbors
(3) clustering coefficient
(4) PageRank score
(5) HITS Hub score/HITS Authority score
(6) farness centrality
(7) betweenness centrality
(8) average betweenness centrality of neighbors
(9-18) the i-th component of the eigenvector associated with the 2nd (smallest nonzero) through the 11-th eigenvalue of the Laplacian matrix

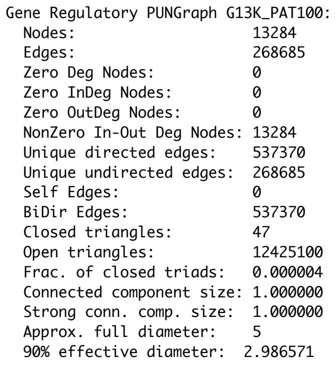
After computing each of the 18 values for each of the nodes in the ARACNE network, they were normalized to have mean zero and standard deviation one, so as to reduce clustering distortion due to scaling. As already stated, the unsupervised learning method for the grouping of network nodes according to their topological properties in p-space was K-means clustering.

The supervised machine learning algorithms logistic regression and GDA were used for survival outcome prediction. Softmax regression was used for cancer stage prediction. The optimization of parameters of the maximum log-likelihood function for logistic regression and softmax regression was performed with stochastic gradient ascent. The optimization of parameters for GDA was carried out by finding their maximum likelihood values analytically. The MI estimator function used for the MI feature selection filter was adapted from a code repository.

5 Results and Findings

The network inference was completed for a subset of 100 patients using ARACNE2 in Workbench 2.5.1, an improved and faster version of the original ARACNE algorithm. The corresponding adjacency matrix was exported as an NNF file and loaded into SNAP after appropriate conversion. The undirected network constructed by ARACNE had one giant connected component consisting of 13284 nodes and 268685 edges with a minimal fraction of closed triangles resulting from application of the data processing inequality (see below). Nodes not connected to the giant component were ignored. General Information about the network is listed in Figure 1. The degree distribution is plotted in Figure 2.

The K-means clustering algorithm discovered 5 prominent clusters in p-space, with 5703, 2075, 641, 124 and 4741 nodes. Each of these clusters was then subjected to MI correlation ranking. For each cluster and for each node an MI estimate measured the correlation of the gene/feature values with the cancer survival and cancer subtype output values, thus ranking each node within the cluster in descending order. For each cluster, the 10, 15 and 20 nodes



```
Gene Regulatory PUNGraph G13K_PAT100:
  Nodes:                     13284
  Edges:                     268685
  Zero Deg Nodes:            0
  Zero InDeg Nodes:          0
  Zero OutDeg Nodes:         0
  NonZero In-Out Deg Nodes: 13284
  Unique directed edges:     537370
  Unique undirected edges:  268685
  Self Edges:                0
  BiDir Edges:               537370
  Closed triangles:          47
  Open triangles:            12425100
  Frac. of closed triads:    0.000004
  Connected component size: 1.000000
  Strong conn. comp. size:  1.000000
  Approx. full diameter:     5
  90% effective diameter:  2.986571
```

Figure 1: General Information for ARACNE Gene Interaction Network

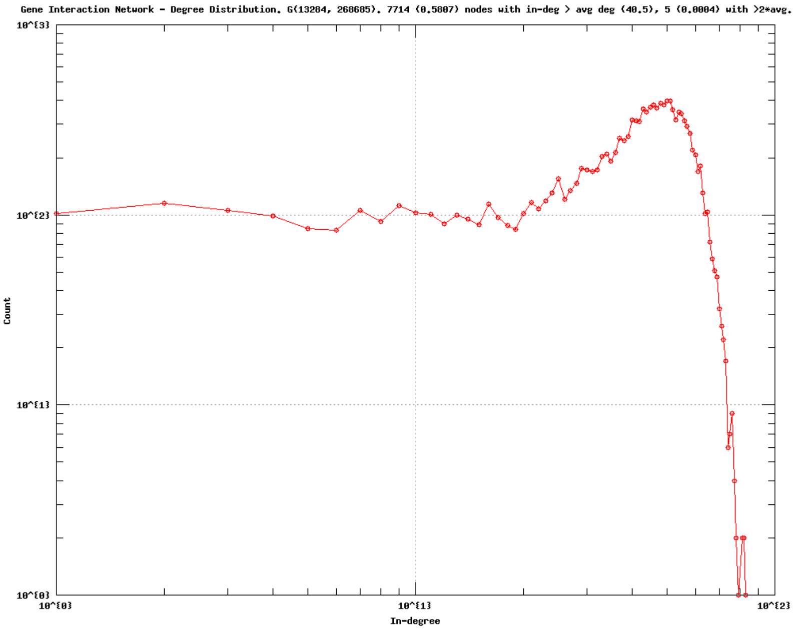
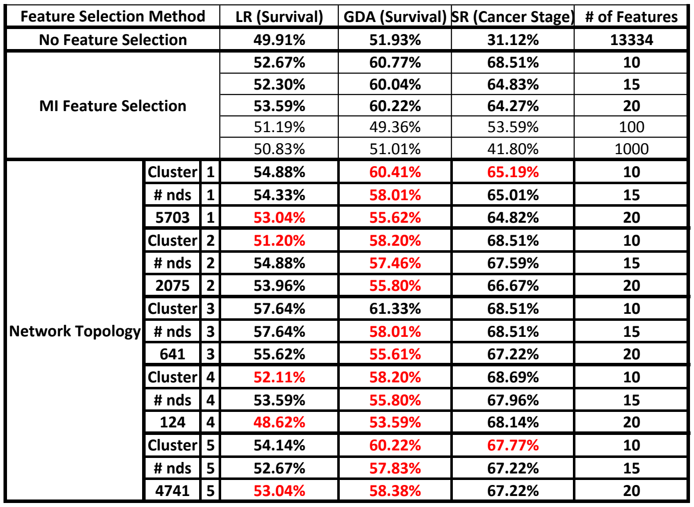


Figure 2: Degree Distribution



Feature Selection Method			LR (Survival)	GDA (Survival)	SR (Cancer Stage)	# of Features
No Feature Selection			49.91%	51.93%	31.12%	13334
MI Feature Selection			52.67%	60.77%	68.51%	10
			52.30%	60.04%	64.83%	15
			53.59%	60.22%	64.27%	20
			51.19%	49.36%	53.59%	100
			50.83%	51.01%	41.80%	1000
Network Topology	Cluster	1	54.88%	60.41%	65.19%	10
	# nds	1	54.33%	58.01%	65.01%	15
	5703	1	53.04%	55.62%	64.82%	20
	Cluster	2	51.20%	58.20%	68.51%	10
	# nds	2	54.88%	57.46%	67.59%	15
	2075	2	53.96%	55.80%	66.67%	20
	Cluster	3	57.64%	61.33%	68.51%	10
	# nds	3	57.64%	58.01%	68.51%	15
	641	3	55.62%	55.61%	67.22%	20
	Cluster	4	52.11%	58.20%	68.69%	10
	# nds	4	53.59%	55.80%	67.96%	15
	124	4	48.62%	53.59%	68.14%	20
	Cluster	5	54.14%	60.22%	67.77%	10
	# nds	5	52.67%	57.83%	67.22%	15
	4741	5	53.04%	58.38%	67.22%	20

Figure 3: Machine Learning Success Rates

with the highest MI value were used as features for machine learning. After training and testing for each of these instances the following results for the success rates of output value prediction were obtained (see Figure 3).

While none of the clusters appears to yield a significant improvement in success rates over conventional MI feature selection, especially for GDA, Cluster 3 with 641 nodes seems to be the most promising candidate for the search of predictive features/genes as it raises the success rates of both logistic and softmax regression. For future work, it may be more instructive to try an iterative procedure and create a subgraph of the cluster yielding the largest improvement at each step, subsequently subjecting this subgraph to the same clustering routine and discovering subclusters which may yield a further improvement still. Subdividing clusters in this manner may be a method to identify predictive features and examine what topological patterns they have in common.

References

[1] Winter C, Kristiansen G, Kersting S, Roy J, Aust D, et al. Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes. PLoS Comput Biol 8(5) 2012: e1002511. doi:10.1371/journal.pcbi.1002511.

[2] Margolin A, et al.: ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 2006, 7(Suppl 1):S7 doi:10.1186/1471-2105-7-S1-S7.

[3] Nishikawa T, Motter AE: Discovering Network Structure Beyond Communities. Sci. Rep. 1, 151; DOI:10.1038/srep00151 (2011)