BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 10(10)

Differences in protein-protein association networks for lung adenocarcinoma: A retrospective study

Anisha Datta[1], Sinjini Sikdar[2] & Ryan Gill*[3]

[1]Louisville Collegiate School and Department of Mathematics, University of Louisville; [2]Department of Bioinformatics and Biostatistics, University of Louisville, [3]Department of Mathematics, University of Louisville, Louisville; Ryan Gill - Email: ryan.gill@louisville.edu; phone: 15028522729; *Corresponding author

Received October 04, 2014; Accepted October 05, 2014; Published October 30, 2014

Abstract:
Various methods to determine the connectivity scores between groups of proteins associated with lung adenocarcinoma are examined. Proteins act together to perform a wide range of functions within biological processes. Hence, identification of key proteins and their interactions within protein networks can provide invaluable information on disease mechanisms. Differential network analysis provides a means of identifying differences in the interactions among proteins between two networks. We use connectivity scores based on the method of partial least squares to quantify the strength of the interactions between each pair of proteins. These scores are then used to perform permutation-based statistical tests. This examines if there are significant differences between the network connectivity scores for individual proteins or classes of proteins. The expression data from a study on lung adenocarcinoma is used in this study. Connectivity scores are computed for a group of 109 subjects who were in the complete remission and as well as for a group of 51 subjects whose cancer had progressed. The distributions of the connectivity scores are similar for the two networks yet subtle but statistically significant differences have been identified and their impact discussed.

Keywords: protein-protein, networks, lung adenocarcinoma, expression data, protein-protein interaction, association networks, lung adenocarcinoma.

Background:
For some prevalent types of cancer such as lung adenocarcinoma where the effectiveness of standard chemotherapy is limited, alternative treatments based on targeting critical genes are desired **[1].** Consequently, it is important to examine the differences between the interactions of the proteins encoded by genes between patients whose cancer progress differently. Identifying such proteins or groups of proteins helps to identify potential targets for new treatments. In order to identify a protein or a class of proteins which is differentially expressed, we need a formal statistical framework. Previously, a framework for differential network analysis was developed **[2]** and applied to microarray data from a pair of networks. In this paper, methods from the above mentioned were adapted to analyze protein expression data and include tests for differential connectivity for individual proteins relative to all other proteins as well as tests for differential connectivity within a class of proteins.

Methodology:
Dataset used
Herein, the methods are applied to data from a study on lung adenocarcinoma. The data set is freely available in the International Cancer Genome Consortium (ICGC) data repository **[3].** It was also featured as one of the challenge data sets at the recent Critical Assessment of Massive Data Analysis (CAMDA) conference **[4].** We used version 14 of the data, which includes expression values of 174 protein antibodies. Some antibody IDs correspond to the same gene so the data set only includes protein expression values for 139 genes. There

are protein expression values for each of the 160 subjects, 109 of whom are in the complete remission group and 51 of whom are in the progression group.

Model

To quantify the strength of pairwise interactions between protein expression values within a group, we use connectivity scores based on the method of partial least squares. Specifically, the scores for each protein are computed by fitting a regression model with all of the other proteins as covariates using estimates obtained from partial least squares. For each network, this creates a $p \times p$ square matrix of coefficients for each pair of proteins where p is the number of proteins common to both networks. Finally, the matrix for the kth network is symmetrized to obtain the connectivity scores $s_{i,j;k}$ where i and j refer to the row and column numbers of the matrix of scores. See **[2]** and **[5]** for a complete description of the algorithm for obtaining association/interaction scores based on partial least squares and **[6]** for discussion of a freely available R package **dna** which provides a flexible implementation of the methods.

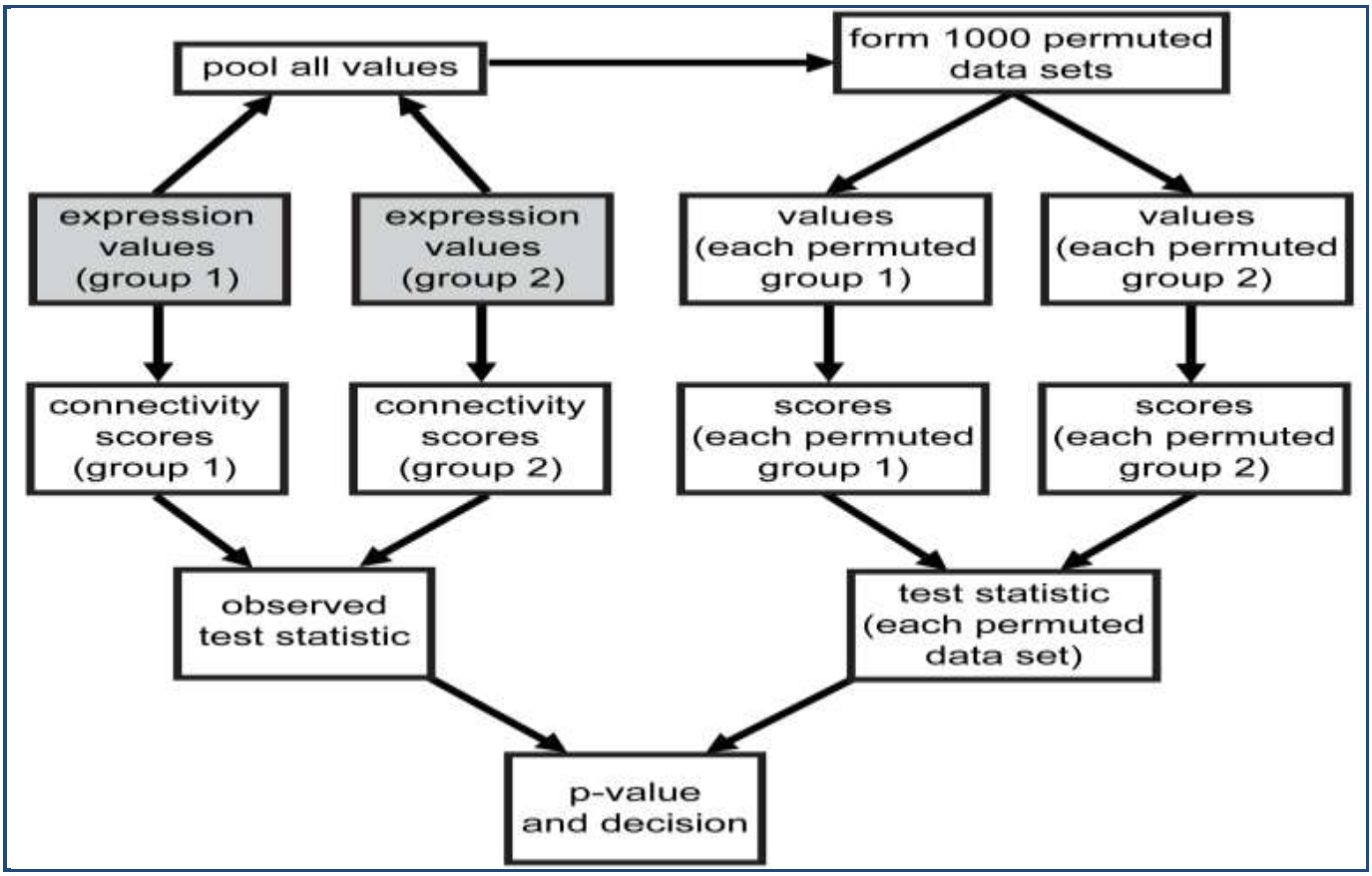

Figure 1: Flow diagram illustrating the test for differential connectivity for an individual protein, starting with the expression values from both groups. The values from both groups are pooled. Then the labels are randomly permuted 1000 times to form new pairs of groups for each data set. The connectivity scores are computed for each actual and permuted group. These connectivity scores are then used to calculate the test statistic for both the observed and permuted data sets. Finally, a p-value is determined by comparing the observed test statistic with the values of the test statistic based on the permuted data sets and is used to make a decision on whether there is a significant difference between the scores for the two proteins between the two groups.

Statistical tests

Next, formal statistical tests can be formulated based on these connectivity scores, similar to the framework proposed in **[2].** To test the differential connectivity of the scores corresponding to protein a, compared with all other proteins, we use the mean absolute difference statistic

$$d(a) = \frac{1}{p-1}\sum_{j=1}^{p}\left|s_{a,j;1} - s_{a,j;2}\right|.$$

This statistic measures the difference between the groups for all pairs of proteins involving the ath protein. The estimated p-value for this test statistic is computed by a permutation procedure. The observed protein expression values from both groups of proteins are first combined. Then 1000 new permuted data sets are constructed by randomly permuting the labels of the combined data set and splitting the combined data into two new groups. For each permuted data set, the test statistic is computed, and the distribution of test statistic values is compiled from the 1000 permuted data sets. If the null hypothesis that both networks are the same holds, then the random variable corresponding to the test statistic for the observed data has the same distribution as the permuted data sets and we do not expect a large value. On the other hand, large values of the observed test statistic give evidence against the null hypothesis in favor of the statement that the associations/interactions among the two networks differ for this protein. Hence, the estimated p-value is the proportion of

test statistic values among the permuted data sets which are at least as large as the observed test statistic, and the hypothesis that the networks are the same is rejected if the p-value is sufficiently small. A flowchart summarizing the procedure for this significance test is given in **Figure 1**. A more detailed mathematical description of a similar permutation test is given in **[2].**

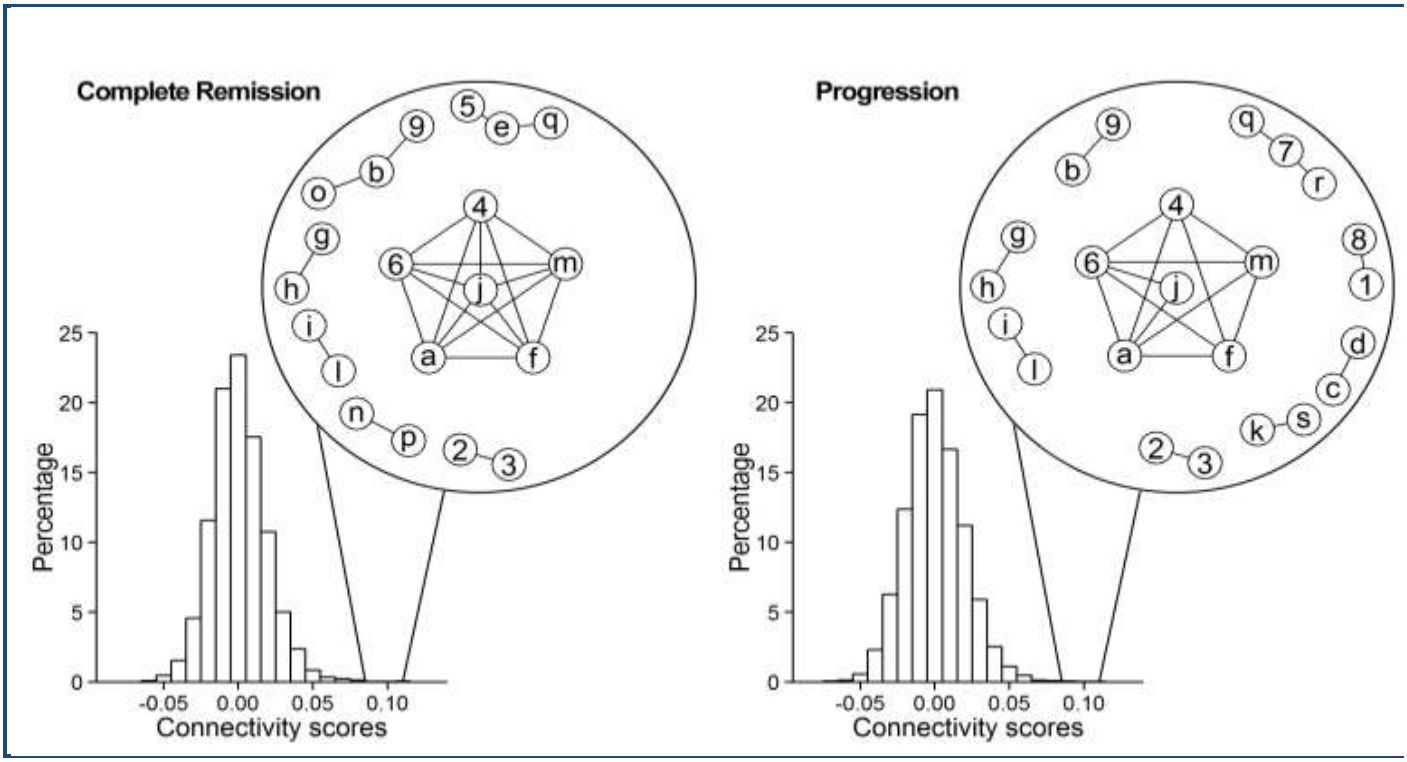

Figure 2: Histograms for the distribution of connectivity scores for the complete remission and progression networks. The scores were computed for each pair of proteins using expression values for 174 proteins from a group of 109 subjects with lung adenocarcinoma who went into complete remission and from a group of 51 subjects with lung adenocarcinoma whose cancer progressed. For each network, the connections involving scores greater than 0.085 are illustrated in a graph to the right of each corresponding histogram. The edges represent pairs of proteins with connectivity scores which exceed 0.085. The proteins in the graph (with labels for vertices in parentheses) are alpha-Catenin(1), ACC_pS79(2), ACC1(3), c-Met(4), Caspase-3(5), Caspase-8(6), CD20(7), E-Cadherin(8), EGFR_pY1068(9),ERCC1(a),HER2_pY1248(b),MAPK_pT202_Y204(c), MEK1_pS217_S221(d), p27_pT157(e), PARP(f), PKC-alpha(g),PKC-alpha_pS657(h),Rab25(i),Rb(j), 6_pS235_S236(k), SETD2(l), Snail(m), Src(n), Src_pY416(o), TIGAR(p), XBP1(q), YB-1(r), and YB-1_pS102(s).

Alternately, to test the differential connectivity of the scores within a particular subset A of proteins, we use a similar statistic. Without loss of generality, suppose that A is the first L proteins. Then, to test the differential connectivity of the proteins within the subset of proteins in A, we use the test statistic

$$\Delta(A) = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \left| s_{i,j;1} - s_{i,j;2} \right|.$$

The procedure for obtaining the p-value and performing a permutation test based on this statistic is analogous to the procedure described in the previous paragraph for testing the differential connectivity of an individual protein.

Results & Discussion:

The connectivity scores were computed for each group of proteins. A histogram illustrating the distributions of the connectivity scores for the proteins in each network is shown In **Figure 2**. Overall, the distributions for the two networks appear to be very similar visually, which is to be expected since all of the patients were lung adenocarcinoma patients. However, differences were identified when the networks were tested formally with the statistical framework, which shows how essential these formal tests are when trying to detect important differences in association. More notably, the pairs of proteins corresponding to the largest connectivity scores are very similar for both networks, as shown in **Figure 2**. The vertices in **Figure 2** represent proteins and edges are shown for a pair of proteins if the connectivity score for the pair exceeds a specified threshold; two proteins are said to be in the same module if there is a path connecting them. There are 23 pairs of proteins with connectivity scores exceeding 0.085 in the complete remission network and 21 pairs (9 modules) exceeding 0.085 in the progression network. The largest module in each network includes 6 proteins (ERCC1, PARP, Snail, c-Met, Caspase-8, Rb); the difference is that in the complete remission network, every pair of proteins is connected while in the progression network, three of the connections are lost for the Rb protein. There are several smaller modules which have similar connections based on this threshold. The connections for the pairs PKC-alpha and PKC-alpha_pS657, Rab25 and SETD2, and ACC1 and ACC_pS79 based on this threshold are identical for the two networks.

Also, the proteins HER2_pY1248 are connected in both networks, but HER2_pY1248 is also connected to Src_pY416 in the complete remission network.

Although the overall distributions of connectivity scores are similar and many of the pairwise interactions between proteins with the top connectivity scores are similar, there are some differences between the networks which are statistically significant. The individual proteins with the smallest p-values based on the test for differential connectivity are listed in **Table 1** (**see supplementary material)** along with the corresponding genes which encode the protein, the observed values of the MDA test statistic, and the estimated p-values based on 1000 permutations. Several of these proteins have been connected with lung cancer in previous studies. Specifically for lung adenocarcinoma, it was recently demonstrated in **[7]** that expression of the CDH2 gene might be increased by the inhibition of the microRNA miR-218 through ADAM9 (a protein not included in our data set) which increases the protein expression of N-Cadherin and consequently results in metastasis of the cancer. Also, an inverse relationship was found between the IRS-1 gene and the presence of neutrophil elastase which affects growth in tumor cells in subjects with human lung adenocarcinoma **[8].** Furthermore, *in vivo* and *in vitro* experiments in **[9]** demonstrate functional roles for a Yap/Taz pathways in the progression and metastasis of lung adenocarcinoma tumors. Another recent study by **[10]** a pathway involving PTEN for both small cell lung cancer and lung adenocarcinoma. The CASP9 gene which is connected with regulation of apoptosis and **[11]** polymorphisms of this gene were found to be associated with the risk of lung cancer. The BCL2L1 gene inhibits activation of caspases preventing cell death **[12].** Additionally, recent work involving anticancer drug development for lung cancer based on PARP-1 inhibitors is discussed in **[13].**

It should be noted that different proteins are differentially expressed when the groups are analyzed marginally by traditional statistical tests. For instance, the two sample t-test identifies ER-alpha (p-value = .02), Cyclin_B1 (p-value = .02), YB-1 (p-value = .02), GATA3 (p-value = .04), and XBP1 (p-value=.04) as significantly different at level .05. Thus, differential network analysis identifies important differences in protein expression values not found by analyzing each protein alone. Also, there are 7 genes (EIF4EBP1, EGFR, SRC, PRKCA, GSK3A|GSK3B, CDKN1B, and AKT1|AKT2|AKT3) in the data set which encode 3 or more proteins. The test for differential connectivity within each of these classes of proteins is performed for each gene, but only one gene exhibited significant differences between connectivity scores; the test statistic for the group of protein antibodies Akt_pT308, Akt_pS473, and Akt encoded by gene AKT1|AKT2|AKT3 is $\Delta = 0.0298$ with corresponding p-value 0.022. The role that AKT plays in lung adenocarcinoma is discussed in **[14].** On the other hand, for example, the test was not rejected (p-value = 0.114) for the protein antibodies encoded by the three protein antibodies (p27, p27_pT157, p27_pT198) encoded by CDKN1B.

Conclusions:

A method has been presented for testing whether proteins and groups of proteins interact differently with other proteins in two groups. The method was applied to protein expression data on lung adenocarcinoma. Analysis shows that the two networks are similar in appearance. However, the method was successful at identifying proteins with statistically significant differences in the connectivity scores for the complete remission and progression groups. A group of proteins encoded by the AKT1|AKT2|AKT3 gene was found to be significantly different between the two classes. These were known to be associated with cancer. Thus, we describe a method for analyzing expression levels to help identify processes, which differ between cancer patients with different progressions. This type of analysis finds application in gene discovery for cancer treatment.

References:

[1] Greulich H, *Genes Cancer*. 2010 **1**: 1200 [PMID: 21779443]
[2] Gill R *et al. BMC Bioinformatics*. 2010 **11**: 95 [PMID: 20170493]
[3] https://dcc.icgc.org/download/release_14/LUAD-US
[4] Camda2014.bioinf.jku.at/doku.php/contest_dataset
[5] Pihur V *et al. Bioinformatics* 2008 **24**: 561 [PMID: 18204062]
[6] Gill R *et al. Bioinformation* 2014 **10**: 233 [PMID: 24966526]
[7] Sher YP *et al. PLoS One.* 2014 **9**: e94065 [PMID: 24705471]
[8] Metz HE & Houghton AM, *Clin Cancer Res.* 2011 **17**: 206 [PMID: 20966354]
[9] Lau AN *et al. EMBO J.* 2014 **33**: 468 [PMID: 24497554]
[10] Cui M *et al. Mol Cancer Res*. 2014 **12**: 654 [PMID: 24482365]
[11] Park JY *et al. Hum Mol Genet*. 2006 **15**: 1963 [PMID: 16687442]
[12] http://www.uniprot.org/uniprot/Q07817
[13] Lee YR *et al. PLoS ONE*. 2013 **8**: e56284 [PMID: 23451039]
[14] Siegelin MD & Borczuk AC, *Lab Invest*. 2014 **94**: 129 [PMID: 24378644]

Edited by P Kangueane

Citation: **Datta *et al.*** Bioinformation 10(10): 647-651 (2014)

BIOINFORMATION

open access

Supplementary material:

Table 1: Tests for differential connectivity of individual protein antibodies between the complete remission and progression network with p-values less than 0.050.

Antibody ID	Gene	*d*	p-value
ACC1	ACACA	0.0199	0.001
ACC_pS79	ACACA | ACACB	0.0210	0.001
Caspase-9	CASP9	0.0184	0.005
N-Cadherin	CDH2	0.0121	0.006
Alpha-Catenin	CTNNA1	0.0163	0.013
IRS1	IRS1	0.0178	0.019
Smad3	SMAD3	0.0200	0.027
TAZ	WWTR1	0.0171	0.031
Bcl	BCL2L1	0.0142	0.031
PARP	PARP1	0.0125	0.037
Snail	SNAI2	0.0122	0.044
PTEN	PTEN	0.0206	0.046