

Statistical Analysis of Group Crosstalk in Networks

M.Sc. Thesis Report

Theodore J. McCormack

May 3, 2011



Stockholms Universitet
Stockholm Bioinformatics Center

Contents

1	Abstract	1
2	Introduction	2
2.1	Bioinformatics	2
2.2	Pathways	2
2.3	Interaction Networks and Pathways	4
2.4	Statistical Analysis	5
2.5	CrossTalkZ	6
3	Methods	7
3.1	Randomization Algorithms	7
3.1.1	Link Permutation (LP)	7
3.1.2	Node Permutaiton (NP)	8
3.1.3	Link Assignment (LA)	8
3.1.4	Link Assignment + Second-order (LA+S)	9
3.1.5	Fixing Degree Sequence Errors	9
3.2	Link Counting	10
3.3	Statistics	10
3.3.1	Z-score	10
3.3.2	P-value	11
3.3.3	False Discovery Rate	11
3.3.4	Reduced Chi-Squared	11
3.3.5	Analytical Estimation of Expected Links	11
4	Results	13
4.1	Estimation of True Positive Rate	13
4.2	Estimation of False Positive Rate	13
4.3	Comparison of Simulated and Analytical Expected Links	15
4.4	Degree of Topological Conservation	15
4.5	Null Model Quality	16
4.5.1	Distributions of z-score and p-value	17
4.5.2	Reduced Chi-Squared Validity	18
4.6	Software Performance	19
5	Discussion	21

Bibliography

23

1 Abstract

Analyzing groups of functionally coupled genes or proteins in the context of global interaction networks has become an important part of bioinformatic analysis. Typically, one wants to analyze the crosstalk, that is, the total connectivity between or within functional groups. However, this is only meaningful if statistical significance of the measured crosstalk is assessed. CrossTalkZ is a statistical method and software that can be used to assess the significance of crosstalk between pairs of gene or protein groups in large biological networks. It is shown that the standard z-score is generally an appropriate and unbiased statistic. Known biological pathways are used as a gold standard to evaluate the ability of four different null model generation methods to recover self-crosstalk. Two of these methods are previous art and two are novel. Based on null model quality and the true and false discovery rates, it is recommended that methods preserving second-order topological properties are best for crosstalk analysis.

2 Introduction

With the advent of high-throughput screening and sequencing techniques, large quantities of raw biological information is increasingly available. I use the term “raw” to elicit the notion that it is hard to interpret, digest or otherwise draw useful conclusions from it’s initial form. For instance, genetic sequencing results in a sequence of millions of nucleotide-bases (e.g. ...CAATCGGATCC...) that represent the genetic code of a given organism. To be able to answer biological questions given this type of information, it must be cataloged, transformed and interpreted in some meaningful way. Bioinformatics is a field that applies mathematics and computational methods in an attempt to do just this.

2.1 Bioinformatics

Powerful tools have been developed in the field of bioinformatics to assist in answering various biological questions. A typical workflow to understand more about an unknown protein sequence might go as follows. Using BLAST[2], which is a well known sequence alignment tool that can help identify or compare a DNA or protein sequence with previously cataloged sequences, one can identify known proteins that have similar sequences. Then, using UniProt[7], which is a database that organizes and annotates proteins, one can obtain functional information, catalytic activity, cofactors, related biological processes, published articles relating to similar proteins, and much more. One could go on with the analysis to predict the 3D structure via homology modeling, then find the native structure via molecular dynamics simulations, and analyze the kinetics of its active site(s) via quantum chemical calculations. This is only one small example of an analysis that can be performed using bioinformatic methods.

2.2 Pathways

One fascinating bioinformatic problem is understanding how genes, or genetic information, interact via their protein products to carry out functions that enable life. After completion of the Human Genome Project, it was discovered that humans have about 20,000 - 25,000 protein coding genes[5]. This may or may not seem like many, but only a fraction of them are expressed in any given cell. Which genes are expressed depends on many different variables such as: the cell tissue type (liver, skin,

brain, etc.), the temporal conditions (embryonic, infant, adult, etc.), environmental conditions (food availability, climate, etc.), to name a few. Due to these complicated conditions, biologists have spent a great deal of time and effort isolating and cataloging interactions between proteins and their reactants and products. Results of these efforts have been meticulously mapped into relational databases that represent various biological properties. One such database is the Kyoto Encyclopedia of Genes and Genomes (KEGG)[8]. It covers a wide range of biological information and relates proteins, enzymes, genes or whole genomes, diseases, drugs, and biochemicals. Some of these relationships can be referred to as biological pathways, where a pathway is a series of chemical interactions between biomolecules beginning with reactants and ending with products. Shown in Figure 2.1 is a representation of the chemical interactions and molecules involved in the KEGG photosynthetic pathway for adenosine triphosphate (or ATP, the “energy currency” of cells) synthesis.

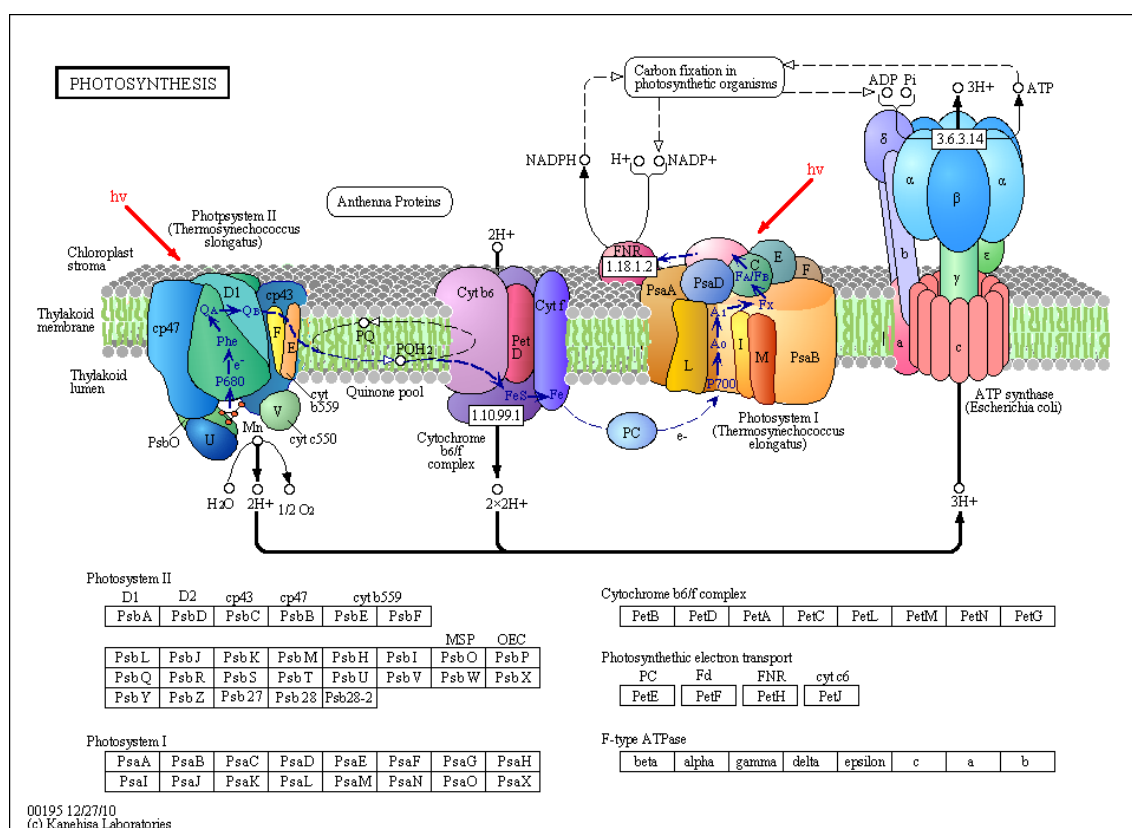


Figure 2.1: Photosynthetic pathway from KEGG.

Source: <http://www.genome.jp/kegg/pathway/map/map00195.html>
(accessed on May 18, 2011)

The bulky objects are different protein complexes shown embedded in a bilayer lipid membrane, with arrows representing the direction of reaction. The set of protein complexes, water, and two photons (represented in red $h\nu$) are the the major

reactants while the products are ATP and other molecules that become reactants of “downstream” pathways (e.g. carbon fixation, part of which is the production of sugars and starches). This modular deconstruction of biochemistry into pathways allows for easier interpretation of the cells complexity and allows one to formulate testable hypotheses more directly.

2.3 Interaction Networks and Pathways

Another informative way to view such complex interactions between genes is to create empirically evidenced gene or protein interaction networks. The FunCoup[1] networks generated at the Stockholm Bioinformatics Center are one set of global interaction networks for a variety of species that combine different types of evidence: protein-protein interactions, mRNA co-expression, sub-cellular co-localization, phylogenetic profile similarity, co-targeting by either miRNA or transcription factors, protein co-expression, and domain-domain interactions. These networks use Ensembl[6] gene id’s as nodes and are connected by links that represent some interaction between the two genes. Each link has an associated confidence score based on the listed evidences. An example of a protein network for yeast is shown in Figure 2.2.

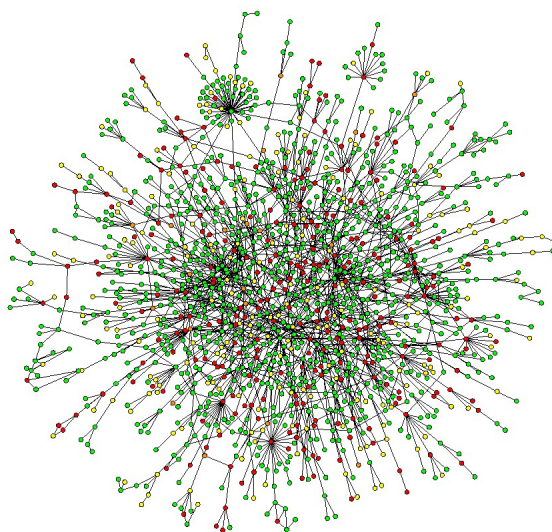


Figure 2.2: A yeast protein-protein interaction network

Source: <http://www.bordalierinstitute.com/target1.html>

(accessed on June 1, 2011)

Due to the quantity of data that goes into generating these networks, they can be very large. For example, the full (confidence cutoff = 0) human FunCoup (v1.1) network has 2,290,854 links between 17,150 genes. Raising this threshold to a confidence of 0.5 gives a network of 230,589 links between 10,885 genes. Therefore, it

becomes critical to be able to recognize and statistically evaluate patterns in these large networks to infer novel biological knowledge. Such patterns, or in network terminology, motifs, are likely to correspond to important processes. A common analysis is to identify clusters (e.g. sets of nodes with high connectivity to each other) of genes in such a network. Pathways composed of a group of genes are likely to be clustered in gene interaction networks. Additionally, these pathway clusters may be highly connected to other pathway clusters in the network, implying some important biological process between the two clusters. This inter-connectivity between pathways is termed pathway crosstalk (see Figure 2.3).

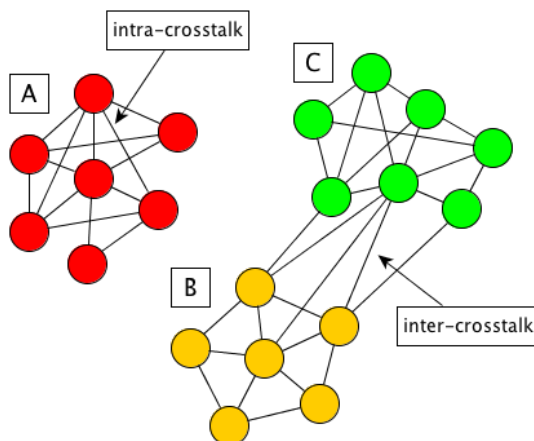


Figure 2.3: Crosstalk between and within pathways. Pathways A, B and C are densely self-connected (intra-crosstalk). Pathways B and C also have some connections with each other (inter-crosstalk).

Since pathways are helpful for understanding individual processes of the complex workings inside cells, it is important to know how the pathways interact (or crosstalk) for a more complete understanding. To draw sound conclusions about crosstalk, one must assess the significance of these interactions. Thus, a statistical analysis is needed to quantify the significance of pathway clustering and crosstalk in biological networks.

2.4 Statistical Analysis

For pathway crosstalk analysis it is important to determine how different the observed pattern of connectivity is from some random background. In statistics, the null model hypothesis can be stated simply as: the observed pattern is *equivalent* to (or belongs to the same equivalence class as) one generated from a random or control model. To reject the null hypothesis is to state that: the observed pattern is *different* from one generated from a random or control model. Then for crosstalk analysis, we can state the null hypothesis as: the number of links observed within

one or between two group(s) is *equivalent* to the number expected by chance from a random distribution. To reject this hypothesis and conclude the observed pattern is significant, one needs a set of networks that have equivalent nodes, but where the links between nodes have been randomized. Furthermore, the degree, or the number of links to/from each node, must be conserved. From this set of randomized networks the mean and standard deviation of the number of links within one or between two group(s) can be calculated. The mean represents the number of links expected by *chance* and can be compared to the number of links *observed* in the original interaction network. Given the distribution of the links expected by chance, one can calculate the following statistics: z-score, p-value, FDR adjusted p-value, and reduced chi-squared. See section 3.3 for more detail on the calculation and interpretation of these statistics.

2.5 CrossTalkZ

CrossTalkZ is a software package built for analyzing crosstalk within and between pathways in large biological networks. It implements four different null model generation algorithms that randomize a networks links. Two of these randomizing methods are previous art, while two are novel. A standard z-score, p-value, and FDR adjusted p-value statistics are calculated from the null model distributions and are reported along with a reduced chi-squared quality measure. CrossTalkZ was implemented in C++, is sufficiently fast, and has various command line options available to the user: network and pathway or group file specification, randomization and link counting method selection, number of randomization iterations, and link weight cutoff. The source is available for free at:
<http://sonnhammer.sbc.su.se/download/software/CrossTalkZ/>

3 Methods

As seen in section 2.4, to be able quantify statistical significance of crosstalk, a set of randomized networks must be generated. These random networks should have the same degree sequence (nodes have the same degree after randomization) as the original network. This is because a nodes degree tends to correspond with how many processes its related to. If degree sequence changes much after randomization, the network no longer represents the same number of processes. There are different ways to randomize networks and still preserve degree sequence. However, the four methods presented here all preserve degree sequence, so the “best” method for crosstalk analysis must then be determined. Lets start with an introduction to the different randomization algorithms in CrossTalkZ.

3.1 Randomization Algorithms

There are four different degree sequence preserving randomization methods implemented in the CrossTalkZ package. They are referred to as: Link Permutation (LP), Node Permutation (NP), Link Assignment (LA) and Link Assignment + Second-order (LA+S). Both of the permutation algorithms are prior art where as the assignment algorithms are novel.

3.1.1 Link Permutation (LP)

Maslov and Sneppen[10] proposed a method of swapping links between nodes. In this method, links A-B and C-D are swapped if A-D and C-B or A-C and B-D are new links in the network. Swapping links in such a way preserves the degree sequence and randomly rewires nodes in the network. In CrossTalkZ, this method is implemented in the following way:

1. Create a set linkSet that contains all links in the network. Create an empty set testedLinks.
2. Randomly choose two links from linkSet.
3. If two new links can be formed (links A-B and C-D can become either A-D and C-B OR A-C and B-D without duplicating links), swap the links appropriately and remove A-B and C-D from linkSet. Go to step 5.

4. Otherwise, add the link pair to testedLinks and go to step 5.
5. Continue to step 2 until the linkSet has less than 2 links OR the cardinality (size) of testedLinks equals the cardinality of the linkSet.

3.1.2 Node Permutaiton (NP)

The node permutation algorithm is a simple method by which nodes or the labels identifying nodes are swapped with other nodes. Since the links between nodes are unchanged, the network topology (specifically the degree sequence) is exactly same as before randomization. This method is implemented in CrossTalkZ as follows:

1. Bin all nodes based on their degree d into a set of bins established by the equivalence relation: $B[d] = \text{round}(\ln(d) + 1)$
2. For each node K in the network, pick a random node R from set of binned nodes having the same degree bin as K ($B[\text{deg}(R)] = B[\text{deg}(K)]$). Swap the nodes (or labels of) R and K .

A log scale is used so that scale free networks (such as observed in biological networks) will have approximately the same number of nodes in each bin.

3.1.3 Link Assignment (LA)

Link assignment is a novel algorithm for randomizing links in networks. It begins with a network of equivalent nodes as the original and then adds links between nodes chosen uniformly randomly until each node's original degree is recovered. It is implemented as follows:

1. Create a set of nodes paired with their degree in the original network, call it recordSet.
2. Randomly shuffle the ordering of recordSet and remove all links from the network.
3. For each record K in recordSet, generate a list of the indices of recordSet, call it recordIndexSet.
4. Until node K has the same degree as in the original network OR randSet is empty OR randIndexSet is empty do the following:
 - a) Pick a random index from recordIndexSet, call this node R
 - b) Test if a new link between node R and node K can be added: R is not equal to K , link $R - K$ is not already in the network, R AND K have not recovered their original degree yet.

- c) If the test passes, create the link between R and K and test if either of these nodes recovered their degree. If they have, remove them from recordSet, break from 4 loop and go to 3. Start 3 loop from the beginning of recordSet.
- d) If the test does not pass, erase the index corresponding to R from recordIndexSet and go to 4.

3.1.4 Link Assignment + Second-order (LA+S)

The algorithm for link assignment + second order is similar to the link assignment algorithm, but instead of choosing a random node from the whole network (represented by recordSet) for node K to connect to, it picks from a set of nodes that fall into the same log degree bin (as established in subsection 3.1.2 part 1) as the original neighbors of K . For example, say that node K was linked to a set of nodes that have the following degree sequence: $\{1, 30, 100, 400\}$. This can be termed the link degree sequence of node K . By choosing nodes from the same *binned* link degree sequence, the algorithm assures that node K has approximately the same link degree sequence before and after randomization.

3.1.5 Fixing Degree Sequence Errors

For the link assignment methods, a few nodes may not be able to preserve their degree due to conflicts in the test cases once the network becomes more densely connected. In that case the following procedure can be applied:

1. Iterate through the nodes in the randomized network and create a set of nodes that don't conserve their degree (errNodes). Sort these in increasing order of difference between the node's original and current randomized degree (deltaDeg).
2. For each of the nodes in errNodes, identify the first ($N1$) and second ($N2$) nodes in the list that have odd deltaDeg.
3. Iterate over links to find a link ($L1 - L2$) that passes the criterion:
 - a) if either node $L1$ OR $L2$ have an odd degree,
 - b) the nodes $L1$ AND $L2$ are not nodes $N1$ OR $N2$,
 - c) the network doesn't already have links $N1 - L1$ AND $N2 - L2$ OR links $N1 - L2$ AND $N2 - L1$.

Once link $L1 - L2$ satisfies the above conditions, delete link $L1 - L2$ and form either links $N1 - L1$ and $N2 - L2$ OR links $N1 - L2$ and $N2 - L1$ appropriately, break iteration over links and continue iteration over errNodes (go to step 2).

Note: After iterating through `errNodes` and fixing odd `deltaDeg` errors, all of the nodes in `errNodes` now have an even `deltaDeg`. The next part fixes the evens.

1. For each node K in `errNodes` calculate `deltaDeg`
2. For J from 0 to `deltaDeg/2` by 1 do the following:
 - a) Find a link ($L1 - L2$) that passes the test: $L1$ AND $L2$ are not K , AND the network doesn't already have links $K - L1$ AND $K - L2$.
 - b) When the test passes, break link $L1 - L2$ and create links $K - L1$ and $K - L2$.

3.2 Link Counting

For this network based crosstalk analysis it is important address how links between groups that have common members should be counted. For instance, if two groups were defined to be overlapping as $G1: \{A, B, C\}$ and $G2: \{A, X, Y\}$ where they share gene A , how should links $A-X$ or $A-C$ be counted? In `CrossTalkZ`, there are two different methods for counting links in this situation. The first and default method is not to count a link if *either* of its connected genes are in both groups. An alternate mode is not to count a link if *both* of its connected genes are in both groups. The reason for excluding these shared member links is that they enrich the number of links between groups and may wash out less strong yet important crosstalk features. In the above example, the number of links are potentially doubled compared to the default counting mode if links to A from B, C, X and Y were counted. This would make the existence or non-existence of the other four links: $B-X, B-Y, C-X,$ and $C-Y$ seem half as important statistically.

3.3 Statistics

The statistics that follow may refer to *expected* or *observed* links. It is implied that these are links within one or between two group(s) as dictated by either the *random* or *original* networks connectivity respectively.

3.3.1 Z-score

For a group or pair of groups, a standard score (z-score) can be calculated as follows: $Z = \frac{N_{obs} - N_{exp}}{\sigma_{exp}}$, where N_{obs} is the number of links observed in the interaction network, N_{exp} and σ_{exp} are the mean and standard deviation of the number of links expected by chance from the set of randomized networks. The z-score can be easily interpreted as the number of standard deviations away from the mean an observation point lies and is useful for quantifying the intra- or inter-crosstalk within one or between two group(s).

3.3.2 P-value

A z-score can be trivially transformed into a p-value via: $p(Z) = 1 - \frac{2}{\sqrt{\pi}} \int_0^{\frac{|Z|}{\sqrt{2}}} e^{-t^2} dt$, which in this case, is the probability of observing at least N_{obs} links given that the null hypothesis is true (that the observed value is no different than one expected by chance: $N_{obs} = N_{exp}$). It is generally held that for p-value < 0.05 one can reject the null hypothesis and conclude that the observed value is significantly different from that expected if the null hypothesis were true.

3.3.3 False Discovery Rate

To correct for false discoveries, or results that falsely reject the null hypothesis, the Benjamini-Hochberg[4] procedure can be applied to adjust the set of p-values. False discovery rate (FDR) adjusted p-values are more useful in multiple hypothesis testing because they account for the fact that a given distribution of p-values has some quantifiable number of false rejections.

3.3.4 Reduced Chi-Squared

Since the z-score is calculated under the assumption that the distribution of links expected by chance is a normal distribution, a test of this normality must also be available to ensure this requirement is met. The theoretical standard normal distribution has the cumulative density function for random variable X : $\Phi(X) = \frac{1}{2}[1 + \text{erf}(\frac{X-\mu}{\sigma\sqrt{2}})]$ where μ and σ are the mean and standard deviation of the normal distribution respectively, erf is the error function. Then for N iterations, the theoretical number of networks that have expected links between points a and b with $b > a$ is $T_{ab} = N \cdot [\Phi(b) - \Phi(a)]$, using $\mu = N_{exp}$ and $\sigma = \sigma_{exp}$. Let the function $C(x)$ be the cardinality of the subset of randomized networks that have x links within or between group(s). Then the number of random networks that have expected links in the region between a and b is simply $E_{ab} = \sum_{i=a}^b C(i)$. As the interval $[a, b)$ becomes infinitesimal, one expects $T_{ab} = E_{ab}$ if the expected distribution is exactly normal. However since a given number of links is integral this is never the case. Therefore a goodness-of-fit chi-squared statistic can be calculated as: $\chi^2 = \sum \left(\frac{E_i - T_i}{\sigma}\right)^2$ over an evenly spaced intervals i . The reduced chi-squared is $\tilde{\chi}^2 = \chi^2 / (N - 3 - 1)$ which normalizes for the number of data points (N) and number of constraints ($= 3 : \mu, \sigma$ and N). A reduced chi-squared ≤ 1 is considered a good fit.

3.3.5 Analytical Estimation of Expected Links

One can calculate the number links between two groups that are expected by chance if the links are chosen uniformly randomly. If there are N total links in the network

and non overlapping groups G1 and G2 have total number of links $N1$, $N2$ where $N1 = \sum_{i=1}^{|G1|} \text{deg}(G1_i)$ and $N2 = \sum_{i=1}^{|G2|} \text{deg}(G2_i)$ (the notation $G1_i$ implies the i th node of group G1). For overlapping sets, the nodes that are shared should not be counted (see section 3.2). A distribution for picking links at random with out replacement from a set of N possible links is best described by the hypergeometric distribution. Then the expected number of links between the two groups is $N_{exp} = \frac{N1 \cdot N2}{N}$. This can be used to compare with expected values obtained from null model distributions (see section 4.3).

4 Results

Several tests were designed to assess the true and false positive rate of crosstalk enrichment as well as the the null model quality for each of the four randomization methods. For all tests, the human FunCoup v1.1 network at confidence cutoff of 0.5 (resulting network has 230,589 links between 10,885 genes) was used along with 66 metabolic and 33 signaling KEGG pathways covering 2,004 unique genes. A crosstalk enrichment result is considered significant if it has p-value (or FDR adjusted p-value) < 0.05 and reduced chi-squared ≤ 1.0 (null model fits normal distribution well).

4.1 Estimation of True Positive Rate

To assess the true positive rate for the different randomization methods a “gold” standard is needed. For this, the KEGG pathways were used as they have been highly cited and are considered to be representative of “true” pathways. The first test was to simply quantify the “self” or intra-crosstalk for each pathway. At an FDR adjusted p-value cutoff < 0.05 all four methods found the 100% of the pathways to be significantly more connected than observed from a set of 150 randomized instances of the network. A second test was to split the pathways into random halves and determine the inter-crosstalk between these halves. This test is an indication of how well the split pathways can be rediscovered by CrossTalkZ. All four methods found at least 99.1% of the pathway halves inter-crosstalk significant.

4.2 Estimation of False Positive Rate

A false positive rate can be obtained by generating random pathways. These pathways have random gene members taken from the full network so that the following conditions are met:

1. No duplicate genes are allowed in a pathway.
2. No genes that were originally in the pathway are allowed.
3. The pathway must conserve its bin degree sequence, where bins are established as in subsection 3.1.2 part 1.

By generating sets of pathways this way we can estimate the null model for the hypothesis: the connectivity between pathways is the same as the connectivity expected by *chance*. This null hypothesis can be rejected then if there is significant intra- or inter-crosstalk enrichment between any given random pathway pair. Over the sets of random pathways, if the number of links expected is sufficient (see subsection 4.5.2), the estimated links for a pathway pair should then be normally distributed. Given that the z-score is only a shift and scale of the expected links, the z-scores should also be normally distributed. The quality of the random pathway null model can be assessed by how closely the distribution of z-scores is to a *standard* normal distribution. See subsection 4.5.1 for this assessment.

For the false positive test, 100 sets of random pathways were generated as described above. Then CrossTalkZ was used to estimate null model distributions and generate statistics. We found that the more conservative methods, LA+S and NP, gave the fewest fraction of false positives around 3.2% and 3.5% respectively. The less conservative LP and LA methods gave slightly higher false positive fraction of 4.7% and 3.6% respectively. Note that the generation of random pathways may result in pathways with “true” crosstalk enrichment, which amounted to approximately 1.5% of the test cases. This is the fraction of pathway pairs found to have significant crosstalk enrichment by all four methods. For false positive rate at varying p-value cutoff, see Figure 4.1.

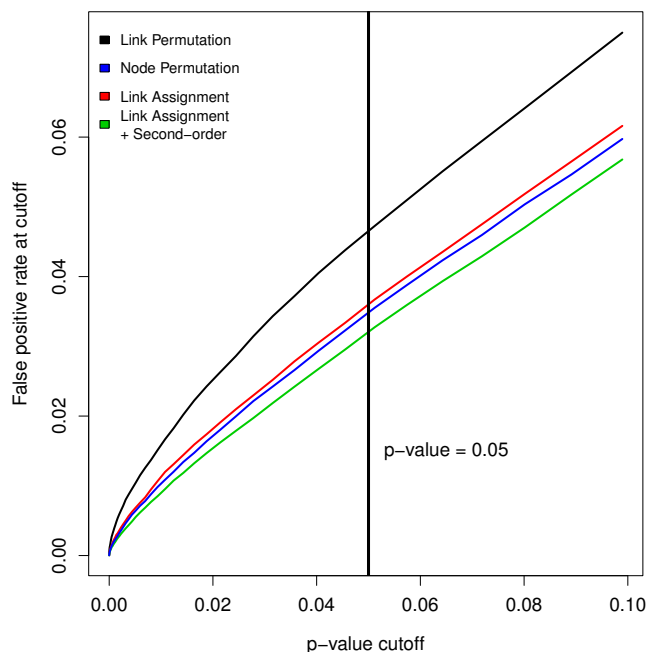


Figure 4.1: False positive rate at varying p-value cutoffs.

4.3 Comparison of Simulated and Analytical Expected Links

In subsection 3.3.5 it is described how the expected number of links drawn randomly between two groups can be calculated analytically from the hypergeometric distribution. Using this analytical estimate to compare with simulated estimates is a good test as to how close the randomization methods are to uniform random generation of links. Analytical estimates were calculated using the same system as all other tests (see introduction paragraph of chapter 4). Root mean squared deviations (RMSD) can then be calculated between the analytical and simulated number of links for each of the four methods (see Table 4.1). Here RMSD was calculated over all N pathway pairs as: $\text{RMSD} = \sqrt{\frac{1}{N} \sum_i (A_i - S_i)^2}$ where A_i and S_i are the number of links calculated analytically and from simulation, respectively, for pathway pair i . From this one can conclude that all four methods are approximately generating links as expected by a random selection process. Furthermore, the more conservative NP and LA+S tend to overestimate the links expected. LP and LA tend to underestimate the number of links expected, although to a lesser magnitude.

Method	RMSD (# links)	μ (# links)	σ (# links)
LP	0.793	0.228	0.759
NP	-1.193	-0.413	1.120
LA	0.382	0.035	0.380
LA+S	-2.064	-0.493	2.004

Table 4.1: RMSD, mean, and standard deviation for 4950 intra- and inter-pathway pair estimations of expected links. The sign on RMSD is taken from the mean deviation. Note that the mean and standard deviation (μ and σ) are for *differences* between analytical and simulated expected links.

4.4 Degree of Topological Conservation

Functional crosstalk between biological processes depends upon physically interacting “neighbor” molecules. In this sense, neighbor is used to describe the fact that sets of biomolecules are more or less likely to directly interact with other specific sets of molecules. In this way, one can describe degree of biomolecular interaction by how many “neighbors” are between two particular biomolecules. In interaction networks this “neighbor”-ness manifests itself in secondary topological features. Some of these features can be described assortativity [11], the s-metric [9], link degree sequence as in subsection 3.1.4, and overall percent identity of links in common between the original and randomized networks.

Assortativity is the degree - degree correlation of connected nodes, it is positive if

nodes of similar degree tend to be connected and negative if nodes of more differing degree tend to connect. For instance, if hubs (high degree nodes) tend to connect to other hubs instead of low degree nodes, the network is said to be assortative. If they on the other hand connect to low degree nodes more, the network is said to be disassortative. An assortative network is likely to be more robust to hub removal. In biological terms, if an “hub” gene loses function, other processes won’t be as affected if the interaction network is assortative. Indeed, the human FunCoup v1.1 network has an assortativity of 0.20. Table 4.2 shows the assortativity for each method. Note that the more conservative methods LA+S and NP preserve the assortativity better than LA and LP.

The s-metric is the sum of the product of node degrees connected by a link for all links in a network. By calculating the ratio of the random network s-metric divided by the original network s-metric, one can determine how well the link degree sequences are preserved. Again, the s-metric ratio (see Table 4.2) is preserved with the more conservative methods LA+S and NP.

Percent identity is also given in Table 4.2 for the four methods. This indicates the fraction of links that are identical in the randomized and original network. Note that LP has the lowest percent identity, while LA+S has the highest. Percent identity can be considered a good measure of how conservative each method is.

Method	Percent identity	Random assortativity	S-metric ratio
LP	7.38 \pm 0.04	-0.09 \pm 0.00	0.80 \pm 0.00
NP	12.12 \pm 0.11	0.20 \pm 0.00	1.00 \pm 0.00
LA	10.47 \pm 0.21	-0.15 \pm 0.02	0.76 \pm 0.01
LA+S	14.79 \pm 0.07	0.14 \pm 0.01	0.96 \pm 0.00

Table 4.2: Topological properties of randomized networks from all four methods compared with the original network. The average percent identity, or the fraction of links the randomized network had in common with the original network (link A-B present in both). The assortativity of the original network was 0.20. The s-metric ratio is calculated by dividing the random network s-metric by the original network s-metric. All means and standard deviations were generated from 50 iterations for each method using the human FunCoup v1.1 network at confidence cutoff = 0.5.

4.5 Null Model Quality

For crosstalk analysis as described in the methods it is important to have a null model that is representative of a random sample space. It should also be unbiased, meaning that it should not tend to favor any one pattern over another. This section describes different ways to quantify the bias and quality of null model distributions.

4.5.1 Distributions of z-score and p-value

When generating random pathways for the false positive test in section 4.2 one expects that the distribution of z-scores for intra- and inter-crosstalk to be standard normal. Figure 4.2 shows the random pathway z-score and p-value distributions for each of the four randomization methods. Also indicated are the z-score distribution skewness, mean, standard deviation, and significant p-value excess. A perfectly standard normal distribution will have skewness = 0, mean = 0 and standard deviation = 1. The p-value excess is expressed as a ratio of the number of results in bin $p\text{-value} < 0.05$ to the average number of results in all other bins. Ideally, this ratio would be unity. Any deviation from unity is indicative of how biased the z-score distribution is and therefore a measure of the null model distribution quality. Based on these measures, one would conclude that LA+S gives the least bias null model distribution, while LP has the most bias.

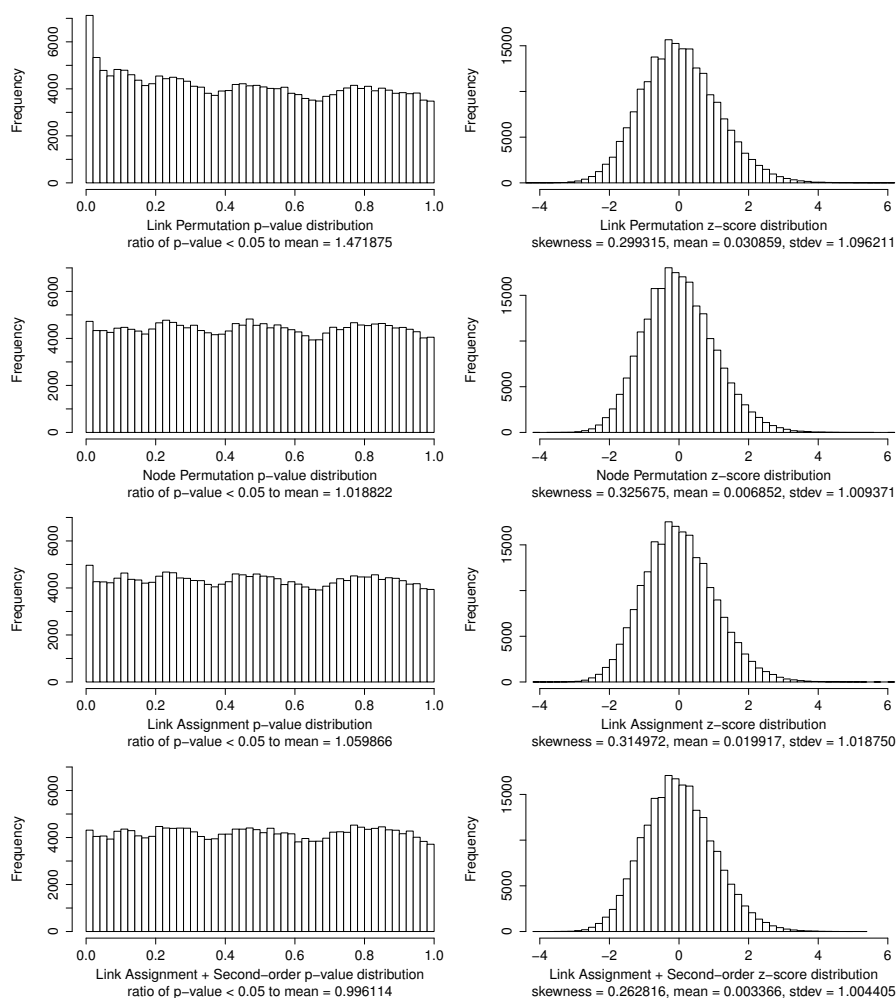


Figure 4.2: Distributions of z-score and p-value for the random pathways along with quality metrics.

4.5.2 Reduced Chi-Squared Validity

A reduced chi-squared statistic (see subsection 3.3.4) is provided for each intra- or inter-group crosstalk analysis. It is a measure of how well the distribution of expected links fits a normal distribution. A reduced chi-squared ≤ 1 is considered to be a good fit. Generally, if the number of expected links is small, the distribution will poorly fit a normal distribution because there are no “negative” links possible. This results in a half or partial normal distribution and is invalid for calculating a proper z-score. Conversely, if the number of expected links is large, the distribution of expected links generally fits a normal distribution. For example, in the random pathway test (see section 4.2) when averaged over all four methods, only $4.7 \pm 1.5\%$ of the cases had expected number of links below 5 and fit a normal distribution (reduced chi-squared ≤ 1). On the other hand, for the cases with expected links above 5, $97.6 \pm 0.8\%$ had reduced chi-squared ≤ 1 . In Table 4.3 are plots of the reduced chi-squared vs expected links from the random pathway data for all four methods. One should note, however, that reduced chi-squared = 1 can potentially be an overly stringent criterion if the expected number of links is small, leading to false negatives.

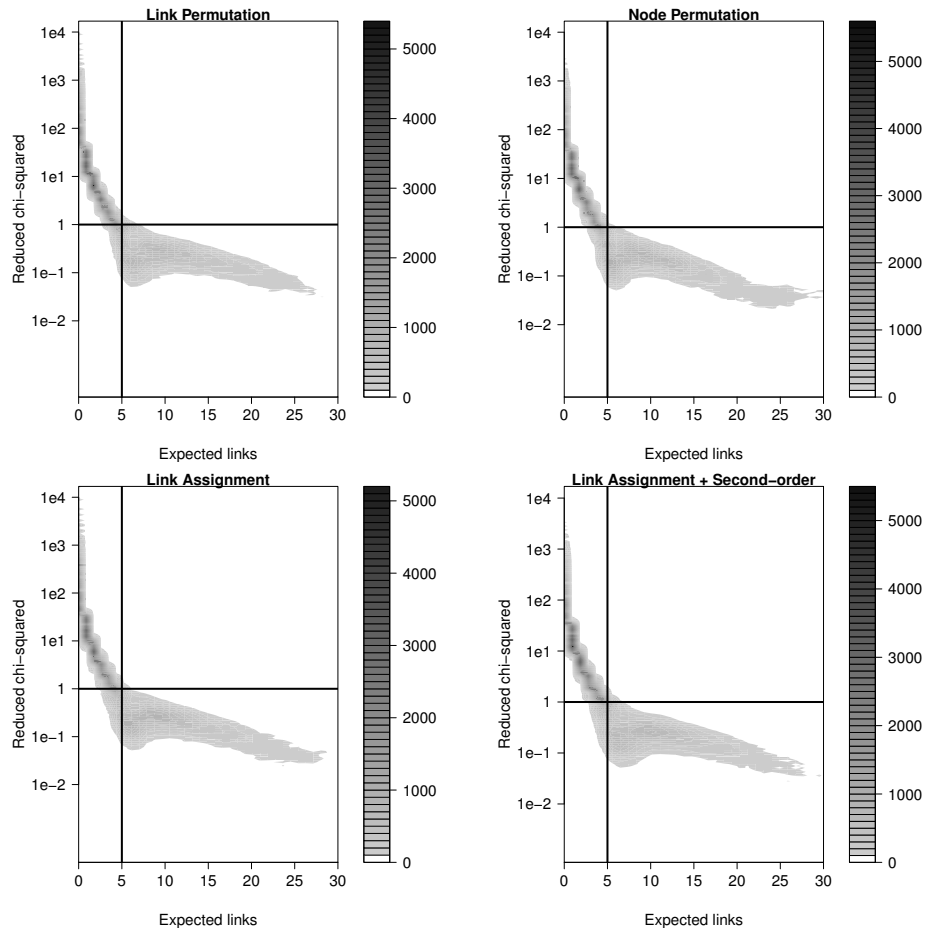


Table 4.3: Reduced chi-squared vs the number of expected links for all four methods. The horizontal line represents the threshold used for normality (chi-squared = 1), and the vertical line is at expected number of links = 5. The number of points is indicated by the density, lighter colors are lower density.

4.6 Software Performance

We tested the performance of each of the four methods on random scale free networks that were obtained using the Barabási–Albert[3] scale free model. Table 4.4 shows the performance of the methods for 150 randomizations when either a) keeping the number of nodes constant and varying the number of links or b) keeping the number of links constant and varying the number of nodes. The node permutation method is the fastest in all test cases and its speed depends only on the number of nodes in the network. Conversely, the link permutation method depends only on the number of links, but is orders of magnitude slower on the test set. The link assignment methods have a more complex performance curves, but still randomize a network of 10^4 nodes and 10^6 links in approximately 10 seconds. All benchmarks were conducted on a 2

GHz processor with 4GB of memory.

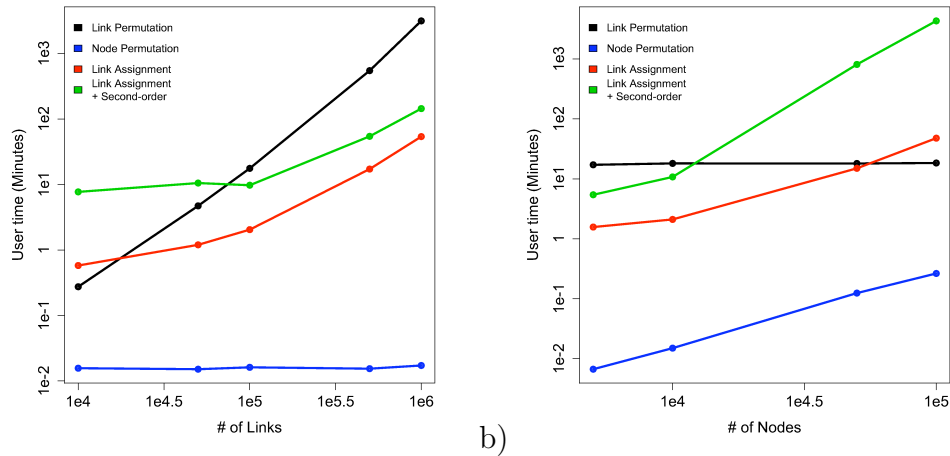


Table 4.4: Software performance for 150 randomizations for each of the four methods. a) Computational performance with increasing number of links and constant number of nodes. The node permutation method is independent of the number of links and only depends on the number of nodes. b) Computational performance with increasing number of nodes and constant number of links. The link permutation method is independent of the number of nodes and only depends on the number of links.

5 Discussion

CrossTalkZ implements four network randomization algorithms to estimate null model distributions for statistical assessment of group crosstalk enrichment in large networks. The algorithms conserve the scale-free topology (degree sequence) of the original network and differ mostly by the extent to which they conserve second-order topological properties. Link permutation attempts to swap *all* links in the original network and therefore generates random networks that have the least links in common with the original network. A result of which is that it potentially underestimates the number of links between groups expected by chance, giving overall higher fractions of significant crosstalk enrichment. Link assignment provides random networks that have links between nodes drawn uniformly from the whole network, a result of which is disassortative mixing (high degree nodes prefer low degree nodes). The more conservative methods LA+S and NP have restricted sets of links or nodes, respectively, to choose from when randomizing and therefore potentially overestimate the number of links between groups expected by chance. Therefore, these methods result in lower fractions of significant crosstalk enrichment.

Which randomization method to use depends on the question to be addressed. If second-order properties are important, such as in crosstalk analysis, methods that preserve second-order features should be used (LA+S or NP). However, if only the degree sequence need be preserved and the process to be modeled is a random selection of links, the LA method is ideal. Finally, if the model process should be as different as possible from the observed network, LP would be best applied.

There are several directions for applications or improvements to methods that can be done at this point. Starting with improvements, the dynamic binning only uses a natural log scale. Different functions may be more applicable to networks that are not necessarily scale-free. Also, the current binning method does not account for number of members in each bin, which could skew results if the cardinality between bins is sufficiently different. These issues affect LA+S and NP methods only.

Second, the z-score approach taken here is generally not valid in the limit of low expected links (see subsection 4.5.2). This can be overcome by a permutation test. A permutation test would use several orders of magnitude more random network instances in order to obtain p-values (no z-score available). Then the p-value is simply ratio of the number of instances that have at least of many links as observed originally ($N_{obs} \leq N_{exp}$) to the total number of random instances.

Some future applications of CrossTalkZ may include analysis or verification of pathway and interaction network databases, disease gene interaction networks, functional

robustness to removal of specific genes or proteins, and general network motif functional significance. Note however that the application of CrossTalkZ is not limited only to gene or protein networks and sets, but any network type where one wants to understand undirected relational interaction within or between sets of nodes. For example, entire ecological systems can be represented by a network where the nodes are species and links between nodes are some trophic or symbiotic relationship. Assessing the significance of relationships between groups of species could help to, for instance, direct protection efforts or predict effects of species invasion. However, CrossTalkZ is best applied in areas where confidence of network links are less than certain.

Bibliography

- [1] Andrey Alexeyenko and Erik L L Sonnhammer. Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Research*, 19(6):1107–1116, 2009.
- [2] S F Altschul, W Gish, Webb Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
- [3] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):11, 1999.
- [4] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, 57(1):289–300, 1995.
- [5] Human Genome Sequencing ConsortiumInternational. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, October 2004.
- [6] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas Kähler, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Bert Overduin, Bethan Pritchard, Harpreet Singh Riat, Daniel Rios, Graham R S Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Giulietta Spudich, Y Amy Tang, Stephen Trevanion, Jana Vandrovcova, Albert J Vilella, Simon White, Steven P Wilder, Amonida Zadissa, Jorge Zamora, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, Jan Vogel, and Stephen M J Searle. Ensembl 2011. *Nucleic Acids Research*, 39(suppl 1):D800–D806, 2011.
- [7] Ursula Hinz. From protein sequences to 3D-structures and beyond: the example of the UniProt Knowledgebase. *Cellular and molecular life sciences CMLS*, 67(7):1049–1064, 2010.
- [8] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.
- [9] Lun Li, David Alderson, Reiko Tanaka, John C Doyle, and Walter Willinger. Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version). *Power*, 2(4):44, 2005.

- [10] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science (New York, N.Y.)*, 296(5569):910–3, 2002.
- [11] M E J Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):5, 2002.