Diffusion Kernel to Identify Missing PPIs in Protein Network Biomarker

Dominic K. Bett, Ananda Mohan Mondal\*

Department of Mathematics and Computer Science

Claflin University, SC 29115, USA

\*[amondal@claflin.edu](mailto:amondal@claflin.edu)

Abstract

Little focus has been placed on neighborhood proteins in the protein-protein interaction (PPI) network that do not physically interact with each other but have a higher likelihood to interact than the actual PPIs in the network. Identifying these missing PPIs would complete the protein network biomarker representing a disease. In the present study, we check the capability of diffusion kernel in identifying these missing PPIs. A diffusion kernel is a computational framework which is based on a physical phenomenon of gas diffusion and a computer science concept of random walk on a graph or network. We seek to predict probable missing PPIs related to Allergy and Asthma using diffusion kernel, by employing a threshold on the kernel values. The completed protein network biomarker can be used as a better predictor for disease identification and classification. This would also help in better understanding of disease mechanism at protein network level.

Our results show that the network with high PPI score has better accuracy of predicting missing PPIs.

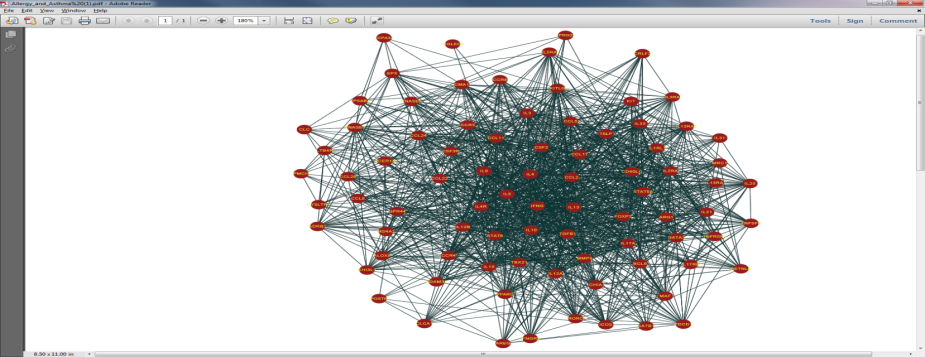
**Keywords:** Diffusion Kernel, Missing PPIs, Protein Network Biomarker, Biomarker, PPI Score.

# Introduction

Diffusion kernel is a computational framework that is based on the physical phenomenon of gas diffusion in a medium, which is also equivalent to the Computer Science concept of random walk on a graph [1]. Protein-protein interaction (PPI) network or protein network is a graph where each node represents a protein and a connection or an edge between two nodes represents the existence of an interaction between two proteins. A genome-wide PPI network comes with rich information about the signature of the disease process [2], protein functions [3], genetic interaction [4], protein subcellular localization [5-10], etc. The randomness of the flow of this information from one protein to another is hidden inside the complex structure of the PPI network, which makes it difficult to decipher this information. A diffusion kernel, since it is based on random walk on a graph, provides a suitable computational framework to extract meaningful biological information from the PPI network. Application of a diffusion kernel provides improved results for predicting protein functions [3], genetic interaction [4], and protein subcellular localization [8, 9]. These factors motivate us to use diffusion kernel in finding the missing PPIs in a protein network biomarker for a disease. A diffusion kernel generates edge weights (interpretable as similarity) between two proteins of all possible protein pairs, which is based on a global perspective of the network. If the weight of a non-existing edge is larger than that of the existing edges, then the non-existing edge can be thought of as a probable missing PPI.

# Data Preparation

Protein network biomarker for Allergy and Asthma is used for analysis in this study. The protein network biomarker is obtained by overlaying the single protein biomarkers, which are differentially expressed proteins, on top of genome-wide PPI network. 84 differentially expressed proteins for Allergy and Asthma are obtained from SABiosciences of Qiagen [11]. The genome-wide PPI network for human is obtained from STRING database [12], which comes with PPI score. The resulting network biomarker is composed of 84 proteins and 1,425 PPIs as shown in Figure 1.



**Figure 1.** Protein network biomarker for Allergy and asthma using Cytoscape [13].

***Subnetworks***

For analysis, smaller networks or subnetworks are derived from the original network biomarker. The PPI scores, 150 through 999 representing the confidence of the interaction, are used to derive these subnetworks. The higher is the PPI score the higher is the confidence that two proteins are going to have an interaction. The PPIs were sorted in order of their scores and six sub-networks were generated from the original network for the purpose of this investigation. For example, the subnetwork N\_500 is composed of PPIs with scores equal to 500 or greater. Similarly, the subnetwork N\_950 is composed of PPIs with scores equal to 950 or greater. Table 1 shows the topology of these subnetworks along with the original network biomarker. As expected, networks with higher PPI scores have less number of proteins and PPIs as well as they are sparser as indicated by the average degree. The number of non-existing PPI (non-PPI) is calculated by subtracting the number existing PPIs from the maximum possible number of PPIs, (n\*(n-1)/2).

**Table 1.** Topology of original network biomarker and the derived subnetworks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Network** | **Proteins** | **PPI** | **Avg. Degree** | **Non-PPI** |
| **Original** | 84 | 1425 | 33.92 | 2061 |
| **N\_500** | 83 | 610 | 14.70 | 2793 |
| **N\_600** | 82 | 540 | 13.18 | 2781 |
| **N\_700** | 77 | 468 | 12.16 | 2458 |
| **N\_800** | 73 | 346 | 9.48 | 2282 |
| **N\_900** | 68 | 240 | 7.06 | 2038 |
| **N\_950** | 61 | 183 | 6.00 | 1647 |

# Methodology

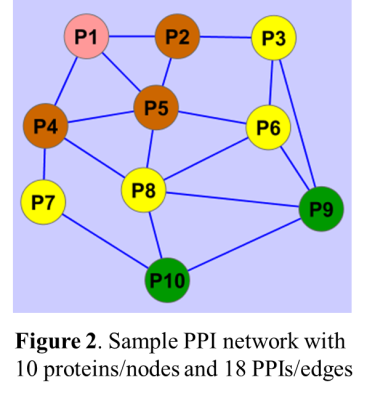
Here, we provide the definition of diffusion kernel on a PPI network and an example why diffusion kernel performs better than the direct use of a graph. Finally, a brief algorithm to identify probable missing PPIs is provided.

* 1. **Diffusion Kernel on a PPI Network**

The formal definition of diffusion kernel on a PPI network [1], equation 1, corresponds to a random walk with an infinite number of infinitesimally small steps.

In the formula, *I* is the identity matrix, *β* is the diffusion constant, *L* is a Laplacian matrix, a close relative to the adjacency matrix, is the number of interaction partners of protein *i*, and represents the matrix exponential of the Laplacian matrix *L*. In evaluating diffusion kernel for a PPI network, *β* is assumed to be 1 as used by others [3, 5-9].

* 1. **Example of Diffusion Kernel on a PPI Network**

Figure 2 shows an example of a PPI network composed of 10 proteins/nodes and 18 PPIs/edges. Table 2 shows the kernel values corresponding to this PPI network obtained using the equation of diffusion kernel, equation 1. A diffusion kernel generates edge weights, interpretable as similarity, between two proteins of all possible protein pairs as seen in Table 2, which is based on a global perspective of the network. For example, based on the direct use of a graph, proteins with the same shortest path distance will have the same similarity, while a diffusion kernel will produce a different similarity. This property makes the diffusion kernel perform better than the direct use of a graph. For example, from the protein P1, the green proteins (P9 and P10) at the shortest path distance of 3 (Figure 2) will have the same similarity value of 1/3 (inverse of distance) with the protein P1, but the diffusion kernel will produce different values of similarity (P9: 0.057and P10: 0.050), as seen in Table 2. Similarly, the diffusion kernel produces different values of similarity for the brown proteins at the shortest path distance of 1 as well as for the yellow proteins at the shortest path distance of 2.

**Table 2.** Kernel values corresponding to PPI network of Figure 2

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Protein | P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | SUM |
| P1 | 0.182 | 0.148 | 0.079 | 0.125 | 0.130 | 0.077 | 0.068 | 0.082 | 0.057 | 0.050 | 1.000 |
| P2 | 0.148 | 0.179 | 0.128 | 0.091 | 0.122 | 0.094 | 0.041 | 0.077 | 0.076 | 0.045 | 1.000 |
| P3 | 0.079 | 0.128 | 0.176 | 0.061 | 0.093 | 0.131 | 0.039 | 0.092 | 0.127 | 0.073 | 1.000 |
| P4 | 0.125 | 0.091 | 0.061 | 0.137 | 0.110 | 0.078 | 0.133 | 0.099 | 0.072 | 0.093 | 1.000 |
| P5 | 0.130 | 0.122 | 0.093 | 0.110 | 0.121 | 0.101 | 0.070 | 0.099 | 0.083 | 0.070 | 1.000 |
| P6 | 0.077 | 0.094 | 0.131 | 0.078 | 0.101 | 0.137 | 0.056 | 0.111 | 0.126 | 0.090 | 1.000 |
| P7 | 0.068 | 0.041 | 0.039 | 0.133 | 0.070 | 0.056 | 0.263 | 0.094 | 0.077 | 0.161 | 1.000 |
| P8 | 0.082 | 0.077 | 0.092 | 0.099 | 0.099 | 0.111 | 0.094 | 0.117 | 0.114 | 0.115 | 1.000 |
| P9 | 0.057 | 0.076 | 0.127 | 0.072 | 0.083 | 0.126 | 0.077 | 0.114 | 0.144 | 0.123 | 1.000 |
| P10 | 0.050 | 0.045 | 0.073 | 0.093 | 0.070 | 0.090 | 0.161 | 0.115 | 0.123 | 0.180 | 1.000 |
| SUM | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |  |

* 1. **Algorithm to Identify Missing PPIs**

Diffusion kernel generates kernel values or similarity weight between two proteins in the PPI network for all possible PPIs (both existing and non-existing), Table 2. Usually, for smaller network, kernel values for the non-existing PPIs (non-PPIs) are less than that of the existing PPIs as is the case for PPI network in Figure 2. The same is not true for a large network. Following definitions are introduced in order to develop the algorithm for identifying the missing PPIs.

***Probable Missing PPIs:*** In case of a large network, there might be some non-PPIs that will have larger kernel values greater than that of the existing PPIs, which are considered as the probable missing PPIs.

***Missing PPIs:*** These are the PPIs that are subset of probable missing PPIs having kernel values equal to or greater than a threshold kernel value.

***Threshold Kernel Value:*** Three different kernel values can be used as the threshold: a) Minimum kernel value of existing PPIs, b) Median kernel value of existing PPIs, and c) Average kernel value of existing PPIs.

**Algorithm:**

***Step-1:*** Find Laplacian matrix, *L*, given the PPI network.

***Step-2:*** Find matrix exponential, , which produces the kernel values.

***Step-3:*** Select a threshold type. Find the threshold value from the matrix elements found in step-2.

***Step-4:*** Find the missing PPIs using the threshold chosen.

***Step-5:*** Put the missing PPIs back in the network.

***Step-6:*** Repeat steps 1 through 5 until there is no missing PPI.

# Results and discussion

We applied diffusion kernel [1] to identify the possible missing PPIs in a given protein network biomarker for Allergy and Asthma. For analysis, six subnetworks, namely, N\_500, N\_600, …, N\_950, are derived from the original network biomarker based on PPI score as described in the data preparation section.

* 1. **Laplacian Matrix and Kernel Value Matrix**

Table 3 shows the Laplacian matrix for the original protein network biomarker for Allergy and Asthma. The column and row headings represent the names of proteins. Laplacian matrix is a symmetric matrix consisting of elements either 0 (represents non-PPI) or 1 (represents existing PPI) except for the diagonal. The diagonals are the number of interactions for the corresponding proteins with other proteins preceded by a negative (-) sign. So, it is clear from this table that the protein ADAM33 interacts with 29 different proteins.

**Table 3.** Laplacian matrix corresponding to protein network biomarker for Allergy and Asthma

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **ADAM33** | **ADRB2** | **ALOX5** | **…** | **TNFSF4** | **TPSAB1** | **TSLP** |
| **ADAM33** | -29 | 1 | 0 | … | 0 | 0 | 0 |
| **ADRB2** | 1 | -27 | 0 | … | 0 | 0 | 1 |
| **ALOX5** | 0 | 0 | -22 | … | 0 | 0 | 0 |
| **…** | … | … | … | … | … | … | … |
| **TNFSF4** | 0 | 0 | 0 | … | -25 | 0 | 1 |
| **TPSAB1** | 0 | 0 | 0 | … | 0 | -12 | 0 |
| **TSLP** | 0 | 1 | 0 | … | 1 | 0 | -40 |

Table 4 shows the kernel value matrix corresponding to the Laplacian matrix of subnetwork N\_700. The sums of the values both in rows and columns are 1, not shown in the table due to space constraint.

**Table 4.** Kernel value matrix for subnetwork N\_700

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **ADAM33** | **ALOX5** | **AREG** | **…** | **TNFSF4** | **TSLP** |
| **ADAM33** | 0.1718 | 0.0083 | 0.0042 | … | 0.0097 | 0.0111 |
| **ALOX5** | 0.0083 | 0.0482 | 0.0056 | … | 0.0107 | 0.012 |
| **AREG** | 0.0042 | 0.0056 | 0.4038 | … | 0.007 | 0.0078 |
| **…** | … | … | … | … | … | … |
| **TNFSF4** | 0.0097 | 0.0107 | 0.007 | … | 0.0274 | 0.0154 |
| **TSLP** | 0.0111 | 0.012 | 0.0078 | … | 0.0154 | 0.0147 |

* 1. **Missing PPIs using Single-Pass Algorithm**

Table 5 shows the average kernel values for the existing PPIs, which were used as the thresholds, and the number of missing PPIs discovered using these thresholds.

**Table 5.** Average kernel values (thresholds) and number of discovered missing PPIs in first iteration.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Network** | **PPI** | **Proteins** | **Non-PPI** | **Avg. Kernel value** | **Missing PPIs** |
| **Original** | 1425 | 84 | 2061 | 0.011926 | 1252 |
| **N\_500** | 610 | 83 | 2793 | 0.013495 | 18 |
| **N\_600** | 540 | 82 | 2781 | 0.014591 | 15 |
| **N\_700** | 468 | 77 | 2458 | 0.017008 | 5 |
| **N\_800** | 346 | 73 | 2282 | 0.022212 | 11 |
| **N\_900** | 240 | 68 | 2038 | 0.028909 | 15 |
| **N\_950** | 183 | 61 | 1647 | 0.033594 | 10 |

It is clear that the average kernel values (thresholds) increases from 0.011926 to 0.033594 with the decrease of size of the network. The reason is that the sum of the kernel values for each row and column remain the same as 1.

* 1. **Missing PPIs using Multi-Pass Algorithm**

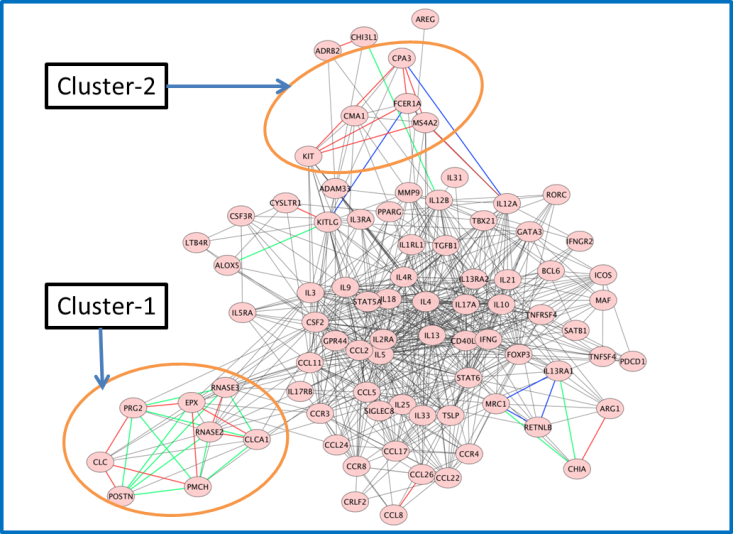
In multi-pass algorithm, the missing PPIs, discovered or identified in an iteration, are put back with the existing PPI for next iteration. So, the existing PPIs for 2nd iteration is the sum of existing PPIs of 1st iteration and the missing PPIs found in the 1st iteration. Table 6 summarizes the results of multi-pass algorithm by showing the numbers of existing PPI and discovered missing PPI in each of the iteration. For space constraint, results for the subnetwork N\_800, N\_900, and N\_950 are not shown. It is clear from this table that the algorithm runs at most 5 iterations. It is also clear that the most of the missing PPIs are discovered in the first iteration and fewer PPIs are found in subsequent iterations.

**Table 6.** Existing PPI and missing PPI discovered in multi-pass algorithm

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Original | | N\_500 | | N\_600 | | N\_700 | |
| Iteration | e-PPI | m-PPI | e-PPI | m-PPI | e-PPI | m-PPI | e-PPI | m-PPI |
| 1 | 1425 | 1252 | 610 | 18 | 540 | 15 | 468 | 5 |
| 2 | 2677 | 179 | 628 | 15 | 555 | 8 | 473 | 2 |
| 3 | 2856 | 403 | 643 | 5 | 563 | 2 | 475 | 2 |
| 4 | 3259 | 118 | 648 | 0 | 565 | 7 | 477 | 7 |
| 5 | 3377 | 0 | 648 | 0 | 572 | 0 | 484 | 0 |

***e-PPI: existing PPI; m-PPI: missing PPI***

Figure 3 shows the subnetwork N\_500 with the discovered missing PPIs. 18 PPIs discovered in iteration-1 are shown in red, 15 PPIs discovered in iteration-2 are shown in green, and 5 PPIs discovered in iteration-3 are shown in blue. It is noticeable that missing PPIs form two clusters: cluster-1 with 8 proteins and cluster-2 with 5 proteins. These clusters might be related to some functions or pathways for which further analysis is required.

****

**Figure 3.** Network N\_500 with discovered missing PPIs. Original PPIs are in black. Discovered PPIs - Iteration-1: 18 PPIs in red; Iteration-2: 15 PPIs in green, and Iteration-3: 5 PPIs in blue.

* 1. **Capability of Diffusion Kernel in Identifying Missing PPIs**

A simple formula is used to describe the capability of diffusion kernel in identifying the missing PPIs.

Total number of discovered PPIs means the sum of PPIs discovered in each of the iterations for a subnetwork. Actual missing PPIs are the subset of discovered PPIs that belong to the original network biomarker. Table 7 summarizes the accuracy of diffusion kernel for different subnetworks derived based on PPI scores. Based on multi-pass algorithm, it is clear that the subnetworks with higher PPI scores (>=800) produce better accuracy (>= 50%) compare to subnetworks with low PPI scores.

**Table 7.** Accuracy of diffusion kernel in identifying missing PPIs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Network | Protein | PPI | Discovered PPIs | |  |
|  |  |  | Total | Actual | Accuracy |
| N\_500 | 83 | 610 | 38 | 14 | 37% |
| N\_600 | 82 | 540 | 32 | 10 | 31% |
| N\_700 | 77 | 468 | 16 | 5 | 31% |
| N\_800 | 73 | 346 | 14 | 7 | 50% |
| N\_900 | 68 | 240 | 17 | 13 | 76% |
| N\_950 | 61 | 183 | 11 | 8 | 73% |

Further investigation is required to check whether all of the actual PPIs are discovered in the 1st iteration or not. This study can further be improved by using a better method to evaluate the capability of diffusion kernel in identifying the missing PPIs. For example, derive a subnetwork of fixed size by randomly selecting the PPIs from the original network biomarker. Then find the missing PPIs and accuracy. This experiment will be carried out 5 times by selecting the same number of PPIs randomly. The average of these five accuracies will be considered as the accuracy at the size of the subnetwork. The same approach will be used for other sizes of subnetworks.

# Conclusion and Future Work

Diffusion kernel, a computational framework based on gas diffusion which is also equivalent to random walk on a graph, is used to identify the missing PPIs in a protein network biomarker. The protein network biomarker for Allergy and Asthma is used for analysis. Six subnetworks were derived from the original network biomarker and tested for finding the missing PPIs. Missing PPIs are identified as the non-existing PPIs for a subnetwork those have the higher kernel values than a threshold.

In the present work, average kernel value for the existing PPIs was used as the threshold. The capability of identifying missing PPIs by the diffusion kernel depends on the PPI score of the derived subnetworks. The higher is the PPI score of a subnetwork the higher is the accuracy of identifying the missing PPIs.

In order to understand better the capability of diffusion kernel in identifying the missing PPIs, following are the future work. ***Future work-1:*** Use the multi-pass algorithm to find the missing PPIs using the minimum kernel value for the existing PPIs as the threshold. ***Future work-2:*** Use the multi-pass algorithm to find the missing PPIs using the median kernel value for the existing PPIs as the threshold. ***Future work-3:*** Functional analysis of cluster proteins formed by the discovered missing PPIs. ***Future work-4:*** Further investigation is required to check whether all of the actual PPIs are discovered in the 1st iteration or not. ***Future work-5:*** Use a better method to evaluate the capability of diffusion kernel in identifying the missing PPIs such as by fixing a size of subnetwork. The results with the future work will be published in an extended version of this paper.

# Acknowledgement

This work is partially supported by NSF HBCU-UP grant, Award number: 1332516 and Center for Excellence in Teaching of Claflin University.

Bibliography

[1] R. Kondor and J.-P. Vert. *Diffusion Kernels*. Available: <http://www.its.caltech.edu/~risi/papers/KondorVert04.pdf>

[2] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Res,* vol. 18, pp. 644-52, Apr 2008.

[3] H. Lee, Z. Tu, M. Deng, F. Sun, and T. Chen, "Diffusion kernel-based logistic regression models for protein function prediction," *OMICS,* vol. 10, pp. 40-55, Spring 2006.

[4] Y. Qi, Y. Suhail, Y.-y. Lin, J. D. Boeke, and J. S. Bader, "Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions," *Genome research,* vol. 18, pp. 1991-2004, 2008.

[5] A. M. Mondal and J. Hu, "NetLoc: Network Based Protein Localization Prediction Using Protein-Protein Interaction and Co-expression Networks," in *IEEE International Conference on Bioinformatics & Biomedicine (BIBM2010)*, Hong Kong, 2010, pp. 142-148.

[6] A. M. Mondal, J. Lin, and J. Hu, "**Network Based Subcellular Localization Prediction for Multi-Label Proteins**," in *BIBM-International Workshop on Biomolecular Network Analysis (IWBNA)*, 2011.

[7] A. M. Mondal and J. Hu, "Protein Localization by Integrating Multiple Protein Correlation Networks," in *The 2012 International Conference on Bioinformatics & Computational Biology, BIOCOMP'12*, Las Vegas, USA, 2012, pp. 82 – 88.

[8] A. M. Mondal and J. Hu, "Scored Protein-Protein Interaction to Predict Subcellular Localizations for Yeast Using Diffusion Kernel," in *Pattern Recognition and Machine Intelligence*, ed: Springer, 2013, pp. 647-655.

[9] A. Mondal and J. Hu, "Network based prediction of protein localisation using diffusion Kernel," *International Journal of Data Mining and Bioinformatics,* vol. 9, pp. 386-400, 2014.

[10] J.-R. Lin, A. M. Mondal, R. Liu, and J. Hu, "Minimalist ensemble algorithms for genome-wide protein localization prediction," *BMC bioinformatics,* vol. 13, p. 157, 2012.

[11] (May 16, 2013). *Biomarkers*. Available: <http://www.sabiosciences.com/Biomarker.php>

[12] C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini*, et al.*, "STRING: known and predicted protein-protein associations, integrated and transferred across organisms," *Nucleic Acids Res,* vol. 33, pp. D433-7, Jan 1 2005.

[13] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage*, et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research,* vol. 13, pp. 2498-2504, 2003.