Community detection on a weighted graph generated through supervised classification using diverse data sources

Gavin Gray (s0805516)
Supervisors: Douglas Armstrong and Colin Mclean

28th April 2014

Abstract

The proposed project replicates previous work in the field of protein interaction prediction on a novel dataset. This is expected to be useful to the problem of community detection in Protein-Protein Interaction(PPI) networks with the aim of gaining insight on protein complexes and cellular function for disease-based research. The active-zone network investigated in this project comprises many proteins known to be vital for synaptic function. Protein interaction prediction is approached in a principled way through applying supervised classification algorithms. It is hoped that this will increase the likelihood of useful results.

Introduction

Disorders of the central nervous system, such as major depression or schizophrenia, affect 1 in 3 people in the developed world. At the synaptic level these diseases are likely related to proteins and their interactions found inside the synapse(Chua et al., 2010; SynSys Project, 2014). The diseases of the synapse are related to the proteins in the synapse(Chua et al., 2010), therefore understanding these proteins may help to treat disease(Li et al., 2010). This proposal focuses on proteins found in the pre-synapse using recent experimental data to identify a group of proteins that are likely to be relevant to disease.

Understanding the function and dynamics of these proteins can be approached through the interactions of these proteins(Chen et al., 2013). Protein-Protein Interaction (PPI) networks are graphical models of these interactions where each protein is represented by a node and each interaction by an edge. Groups of densely connected proteins, often called communities, in these networks are useful for inferring physical interactions between groups of proteins, co-complex relationships and functional associations(Qi, 2008).

Community detection in PPI networks is the practice of finding these communities, applied for example in Pocklington et al. (2006). Community structure is a characteristic of some graphs where the nodes can be grouped such that there are many connections within groups but few between groups(Newman, 2012). An important point is that there is no consensus on a definition for community structure; there are only example graphs that are agreed to exhibit this property.

High-throughput protein interaction experiments, such as Yeast Two Hybrid(Y2H) or Immuno-Precipitation, are often used to find these interactions(Wan Li et al., 2012). Unfortunately, these are prone to high error rates as a result of being sensitive to experimental conditions and the characteristics of the protein being studied(Qi, 2008). Indirect information, such as mRNA expression or genetic information about proteins, can be used to decrease these error rates(Qi et al., 2006). Identifying protein interactions from the experiment then becomes a prediction problem, specifically a classification problem which can be approached with many possible classification algorithms. One approach is to use a supervised classifier to estimate the probability that an interaction in the graph exists(Qi et al., 2006). This creates a weighted graph, as opposed to an unweighted graph where edges are binary. Community detection algorithms, such as a Modularity based shortest-path betweenness algorithm, which function on both weighted and unweighted graphs can then be used to search for communities(Newman and Girvan, 2004).

Localising protein interactions to the presynaptic bouton has been achieved through immunoprecipitation experiments. This type of experiment "pulls-out" protein complexes connected to chosen "bait" proteins, the additional proteins found being the "prey". Mass spectrometry techniques are then used to identify these "prey" proteins(Klemmer et al., 2009; Li et al., 2010; Wan Li et al., 2012). The experiments performed prior to this work have chosen bait proteins which cover the entire presynaptic region. A subset of these proteins which are based in the presynaptic active zone will be the focus of this proposal. Active zone proteins identified in Chua et al. (2010) are illustrated in figure 1.

Comparing the communities detected using both weighted and unweighted networks is the final stage of this project. The unweighted network will be generated by merging interactions obtained from multiple databases and experiment shown in figure 2. This will be discussed in more detail in the Methods section of this proposal.

Research Aims

The aim of this research is to improve the accuracy of current methods in analysing PPI networks with the purpose of co-complex detection. This will be achieved through the use of classifiers(Qi et al., 2006) that allow the creation of a weighted PPI network which incorporates information from databases and indirect information from, for example, mRNA expression. Combining these sources of information in a principled way using supervised classification should improve the accuracy of the information entering the community detection algorithm and improve the performance. Comparing this weighted approach to unweighted techniques is a key part of this project. Metrics such as the Normalised Mutual Information (NMI) and the results of a disease enrichment test will be used to compare the detected communities.

If these results appear to be effective this could result in a publication. In addition, the program to generate weighted graphs could be incorporated into other work based on improving community detection algorithms. If it is effective, it would hopefully become a popular technique in the field and promote collaboration between researchers. If the clustering results are the same as that of the unweighted graph this project would show that information from existing databases is more useful than any extra features. This is unlikely, as existing work has already found that genetic features were more effective than the results of high-throughput experiments(Qi et al., 2006).

Douglas Armstrong will supervise this project as he has already completed research in the area(Armstrong and Sorokina, 2012). Colin McLean should also supervise as he has contributed to the formulation of this proposal.

Methods

There are two main techniques that are important to the work involved in this project. Firstly, community detection algorithms and their application to PPI graphs. Secondly, supervised classification used to combine data from varied sources. This section will focus on how these techniques will be used.

To build an unweighted PPI graph the method shown in figure 2 is used. Using the pull-down experiment a list of proteins is generated, which can be used as an input to the various databases available. For all the proteins of interest all possible interactions from the databases are then combined so that only the interactions between proteins in the list of interest are considered. This combines the known interactions from the illustrated sources for all of the proteins of interest, creating an unweighted protein interaction graph.

In figure 3 the proposed method for building a weighted interaction graph is shown. This process involves a larger number of data sources in the overall process, including those in the process shown in figure 2. There are two important steps in this process: feature extraction from the varied data sources and classification of the protein interactions.

The process of using the direct or indirect data to

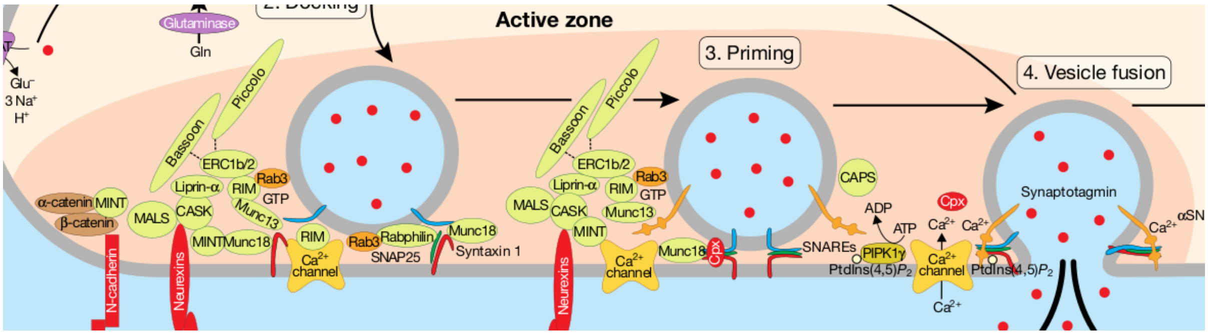
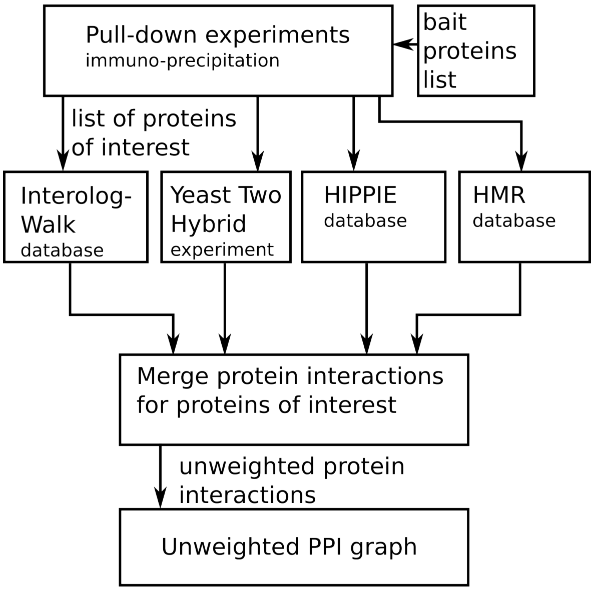
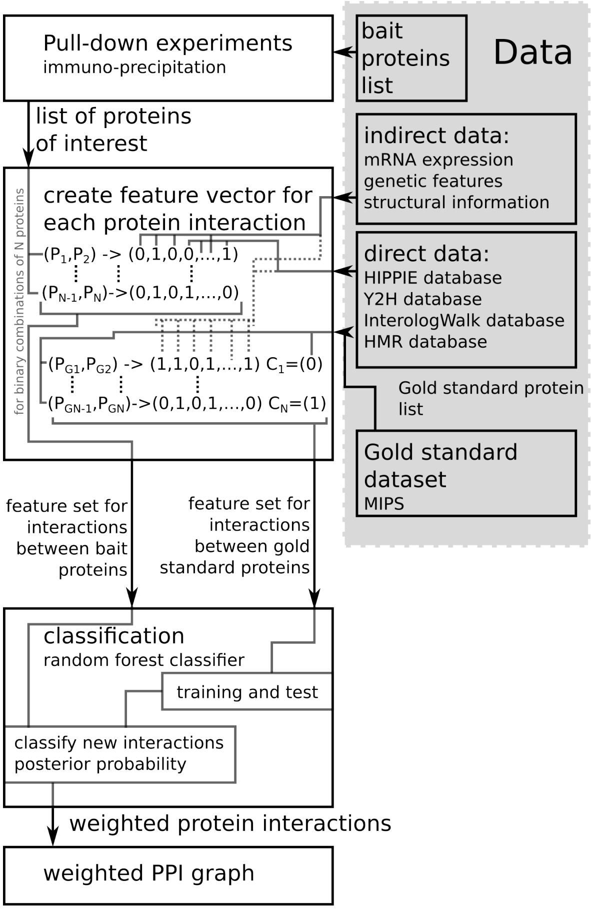


Figure 1: Proteins identified in the presynaptic active zone network(Chua et al., 2010).



Pull-down experiments
immuno-precipitation
bait proteins list
list of proteins of interest
Interolog-Walk database
Yeast Two Hybrid experiment
HIPPIE database
HMR database
Merge protein interactions for proteins of interest
unweighted protein interactions
Unweighted PPI graph

Figure 2: The method for building an unweighted PPI graph currently applied based on previous work in Pocklington et al. (2006). The databases and tools used are: InteroloWalk(Gallone et al., 2011), HIPPIE(Schaefer et al., 2012), HMR(combination of human, mouse and rat from BioGrid,CCSB,HPRD,Intact,MDC) and Yeast Two Hybrid (Edinburgh available Y2H data).



Data
bait proteins list
Pull-down experiments
immuno-precipitation
list of proteins of interest
indirect data:
mRNA expression
genetic features
structural information
direct data:
HIPPIE database
Y2H database
InterologWalk database
HMR database
Gold standard protein list
Gold standard dataset
MIPS
for binary combinations of N proteins
create feature vector for each protein interaction
(P1,P2) -> (0,1,0,0,...,1)
(PN-1,PN)->(0,1,0,1,...,0)
(PG1,PG2) -> (1,1,0,1,...,1) C1=(0)
(PGN-1,PGN)->(0,1,0,1,...,0) CN=(1)
feature set for interactions between bait proteins
feature set for interactions between gold standard proteins
classification
random forest classifier
training and test
classify new interactions
posterior probability
weighted protein interactions
weighted PPI graph

Figure 3: The proposed method for building weighted PPI networks is described. The relationship between the data sources and the final classification output is shown.

create feature vectors is known as feature extraction(Qi et al., 2006). Feature extraction and the design decisions it involves will be a major part of this project. The features used in this classification problem are derived either directly from high-throughput experimental databases or from indirect data sources. An example high-throughput data source which could be used to generate features would be the HIPPIE database(Schaefer et al., 2012), which itself combines data from a large number of sources. Indirect features could include:

- mRNA expression(Bader et al., 2004; Jansen et al., 2003; Lee et al., 2004; Zhang et al., 2004).
- Gene Ontology(GO)(Ashburner et al., 2000) derived features(Ben-Hur and Noble, 2005; Qi et al., 2006).
- Amino acid composition (AAC)(Roy et al., 2009).

There are many other features for consideration during the project based on what is available, these are described in: Qi et al. (2006);Bader et al. (2004);Jansen et al. (2003);Lee et al. (2004);Zhang et al. (2004);Roy et al. (2009);Mering et al. (2005);Rhodes et al. (2005);Ben-Hur and Noble (2005).

The classification problem itself can be solved with a variety of algorithms. All of these are supervised classification algorithms. These are learning algorithms that must be trained with a data set that contains correct classifications for the existence or non-existence of interactions(Qi et al., 2006). This data set is known as the gold-standard data set and for the task of co-complex detection the recommended source for this task is the Munich Information Center for Protein Sequences (MIPS) complex catalogue[Mewes et al. (2004);qi_learning_2008].

The chosen classifier for this problem is the Random Forest classifier. It was found in Qi et al. (2006) to be the most effective of the classification algorithms tested. Random Forest classifiers deal with the dependencies in the data more readily than other methods by averaging the predictions made by multiple decision trees.

Once trained these algorithms can evaluate the probability that a new feature vector is or is not a real interaction based on the previously observed gold standard dataset. This probability is used as the weight in the PPI graph.

Community detection on both the weighted and unweighted graphs will be approached with the same algorithm. The chosen algorithm is a geodesic betweenness algorithm which has been optimised and applied at Edinburgh(Achar, 2011). This algorithm operates by grading edges in the network to determine which edges to remove to partition the graph into individual communities(Newman and Girvan, 2004). The geodesic betweenness score is found through measuring the signals through each edge considering a single signal sent between each node in the network along geodesic paths.

A possible extension to this project would be to approach the entire complex detection task as a semi-supervised learning problem. Something similar to this has been attempted in Shi et al. (2011) using a neural network model and features derived from the graph structure as opposed to the features used in this project, which are directly related to the data. Semi-supervised data would indicate that a deep learning architecture could be successful, drawing on the research in the field of object recognition(Bengio, 2009). The training set to use an algorithm like this would have to be extended to include positive and negative examples of proteins such as that used in Qi et al. (2008);Shi et al. (2011).

Shown in figure 4 is an overview of the project as a whole. The final stages are comparison of the results with two metrics: NMI and Disease enrichment. Normalised Mutual Information measures the consistency between two different community sets. This measure depends on the number of nodes in each community for each different version of the graph - specifically, these are used to calculate the entropy of each community structure and the mutual information between the two. The mutual information is then normalised using the entropies of both graphs.

Disease enrichment comparison links the proteins in a cluster to associated diseases to calculate a p-value indicating homogeneity within the cluster. This metric has direct relevance to disease research so differences between the two approaches here are important to the aims of this project.

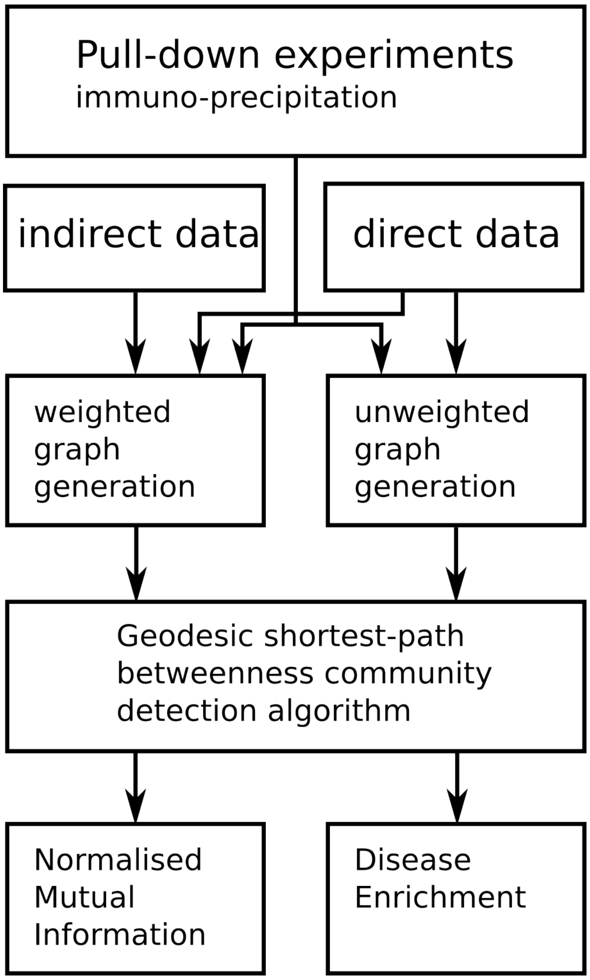

Pull-down experiments
immuno-precipitation
indirect data
direct data
weighted graph generation
unweighted graph generation
Geodesic shortest-path betweenness community detection algorithm
Normalised Mutual Information
Disease Enrichment

Figure 4: An overview of the project as a whole, illustrating the components of the project.

The flow chart of the project in figure 4 illustrates how the components of the project fit together. Figures 2 and 3 describe the unweighted and weighted graph generation sections, respectively. The techniques in these flow charts have been discussed to show that they can be applied to this problem readily.

Schedule

This project is proposed as a 10 week summer project. The steps of the project are illustrated in figure 4 and are sequentially as follows:

1. Feature extraction.
    1. Including choosing, gathering and processing biological data.
    2. Choosing feature encoding.
2. Training and test performance on Gold dataset.
    1. Replication of results in Qi et al. (2006) for verification.
3. Test algorithm on active zone proteins.
    1. Check initial results against unweighted results.
4. Community detection on both weighted and unweighted networks.
5. Comparison of performance using NMI and disease enrichment.

The first step of this project is likely to be the most complicated. Dealing with large amounts of biological data and processing this into useful features is a difficult task. The remaining tasks should not take as much time:

- Classification algorithms are well documented(Murphy, 2012).
- Community detection algorithms are well studied at Edinburgh(Achar, 2011; Pocklington et al., 2006).
- NMI is not complicated to implement.
- Disease enrichment is already implemented.

Taking these observations into account a rough schedule can be constructed:

- Data gathering and feature extraction - 3 weeks
- Classification, community detection and comparison - 2 weeks
- Writeup - 2 weeks
- Overflow - 1 week

Budget

The supervisors recommend a large monitor for working with the large data sets involved in this project. A 28 inch monitor at this point costs £235 and would be sufficient for this project. No other hardware or software is required so the total budget for this project is £235.

Conclusion

Based on the results of previous work combining data to improve protein interaction prediction this project has the potential to improve the accuracy of the PPI graph(Qi and Noble, 2011). It is hoped that this will improve the performance of the community detection algorithm in it's task of protein complex detection. This should provide additional information on the functioning of the synapse and provide useful information for treating diseases affecting the synapse.

References

Achar, V., 2011. Optimising community structure discovery in large protein interaction networks (MSc thesis). University of Edinburgh.

Armstrong, J.D., Sorokina, O., 2012. Evolution of the cognitive proteome: from static to dynamic network models, in: Advances in Systems Biology. Springer, pp. 119–134.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. Nature Genetics 25, 25–29. doi:10.1038/75556

Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J., 2004. Gaining confidence in high-throughput protein interaction networks. Nature Biotechnology 22, 78–85. doi:10.1038/nbt924

Ben-Hur, A., Noble, W.S., 2005. Kernel methods for predicting protein–protein interactions. Bioinformatics 21, i38–i46. doi:10.1093/bioinformatics/bti1016

Bengio, Y., 2009. Learning deep architectures for AI. Foundations and trends® in Machine Learning 2, 1–127.

Chen, B., Fan, W., Liu, J., Wu, F.-X., 2013. Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. Briefings in Bioinformatics bbt039. doi:10.1093/bib/bbt039

Chua, J.J.E., Kindler, S., Boyken, J., Jahn, R., 2010. The architecture of an excitatory synapse. Journal of Cell Science 123, 819–823. doi:10.1242/jcs.052696

Gallone, G., Simpson, T.I., Armstrong, J.D., Jarman, A.P., 2011. Bio::Homology::InterologWalk - a perl module to build putative protein-protein interaction networks through interolog mapping. BMC Bioinformatics 12, 289. doi:10.1186/1471-2105-12-289

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M., 2003. A bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302, 449–453. doi:10.1126/science.1087361

Klemmer, P., Smit, A.B., Li, K.W., 2009. Proteomics analysis of immuno-precipitated synaptic protein complexes. Journal of Proteomics 72, 82–90. doi:10.1016/j.jprot.2008.10.005

Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M., 2004. A probabilistic functional network of yeast genes. Science 306, 1555–1558. doi:10.1126/science.1099511

Li, K.W., Klemmer, P., Smit, A.B., 2010. Interaction proteomics of synapse protein complexes. Analytical and Bioanalytical Chemistry 397, 3195–3202. doi:10.1007/s00216-010-3658-z

Mering, C. von, Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., Bork, P., 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Research 33, D433–D437. doi:10.1093/nar/gki005

Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Güldener, U., Mannhaupt, G., Münsterkötter, M., Pagel, P., Strack, N., Stümpflen, V., Warfsmann, J., Ruepp, A., 2004. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Research 32, D41–D44. doi:10.1093/nar/gkh092

Murphy, K.P., 2012. Machine learning: A probabilistic perspective, ed. The MIT Press, London.

Newman, M.E.J., 2012. Communities, modules and large-scale structure in networks. Nature Physics 8, 25–31. doi:10.1038/nphys2162

Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical Review E 69, 026113. doi:10.1103/PhysRevE.69.026113

Pocklington, A.J., Cumiskey, M., Armstrong, J.D., Grant, S.G.N., 2006. The proteomes of neurotransmitter receptor complexes form modular networks with distributed functionality underlying plasticity and behaviour. Molecular Systems Biology 2, n/a–n/a. doi:10.1038/msb4100041

Qi, Y., 2008. Learning of protein interaction networks (PhD thesis). Universitat Pompeu Fabra, Spain.

Qi, Y., Balem, F., Faloutsos, C., Klein-Seetharaman, J., Bar-Joseph, Z., 2008. Protein complex identification by supervised graph local clustering. Bioinformatics 24, i250–i268. doi:10.1093/bioinformatics/btn164

Qi, Y., Bar-Joseph, Z., Klein-Seetharaman, J., 2006. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins: Structure, Function, and Bioinformatics 63, 490–500. doi:10.1002/prot.20865

Qi, Y., Noble, W., 2011. Protein interaction networks: Protein domain interaction and protein function prediction, in: Lu, H.H.-S., Schölkopf, B., Zhao, H. (Eds.), Handbook of Computational Statistics: Statistical Bioinformatics. Springer.

Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., Chinnaiyan, A.M., 2005. Probabilistic model of the human protein-protein interaction network. Nature Biotechnology 23, 951–959. doi:10.1038/nbt1103

Roy, S., Martinez, D., Platero, H., Lane, T., Werner-Washburne, M., 2009. Exploiting amino acid composition for predicting protein-protein interactions. PLoS ONE 4, e7813. doi:10.1371/journal.pone.0007813

Schaefer, M.H., Fontaine, J.-F., Vinayagam, A., Porras, P., Wanker, E.E., Andrade-Navarro, M.A., 2012. HIPPIE: integrating protein interaction networks with experiment based quality scores. PLoS ONE 7. doi:10.1371/journal.pone.0031826

Shi, L., Lei, X., Zhang, A., 2011. Protein complex detection with semi-supervised learning in protein interaction networks. Proteome Science 9, S5. doi:10.1186/1477-5956-9-S1-S5

SynSys Project, 2014. synsys - a european expertise network on building the synapse.

Wan Li, K., Chen, N., Klemmer, P., Koopmans, F., Karupothula, R., Smit, A.B., 2012. Identifying true protein complex constituents in interaction proteomics: The example of the DMXL2 protein complex. PROTEOMICS 12, 2428–2432. doi:10.1002/pmic.201100675

Zhang, L.V., Wong, S.L., King, O.D., Roth, F.P., 2004. Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics 5, 38. doi:10.1186/1471-2105-5-38