International Journal of Biotechnology and Biomedical Sciences
p-ISSN 2454-4582, e-ISSN 2454-7808, Volume 2, Issue 3; July-December, 2016 pp. 234-237

http://www.krishisanskriti.org

Development of a Tool that Create Complete Protein Circle for a Given Input Protein

Prateek Sukumar[1], Disha Mandal[2] and Yasha Hasija[*3]

[1,2,3]Delhi Technological University
E-mail: [1]prateek.dtu@hotmail.com, [2]dmdishamandal@gmail.com, [3]yashahasija@gmail.com

Abstract—*There is complex molecular network at play imperative to complete any biological process. To understand the working of proteins and how they function, it is quintessential to know and map protein-protein interaction networks. With increased availability of protein-protein interaction data, mainly due to the development of genome-wide experimental methods such as protein chips, two-hybrid test and mass spectrometric analysis, various computational methods and tools are being developed to analyze interaction data in order to reveal the mechanism of the cause of the disease, as the production of research data without any ground results is not beneficial. As one protein affects the working of another protein, any disturbances in these networks are usually linked to the disease and, therefore, protein networks are increasingly serving as tools to unravel the molecular basis of disease. However, most of these tools and database of protein interaction networks is complex to use. Some tools require interaction data to make their visualization e.g. Cytoscape, while some tools provide only the static graphic tool showing limited interactions. In the present work, we intend to present a novel tool that outputs the complete protein interaction of the given input protein by using interaction data from STRING database (www.string-db.org). The tool is currently available in-house.*

1. INTRODUCTION

The US National Human Genome Research Institute (NHGRI) described challenges for the community concerning the genomics and proteomics, which is the conversion of knowledge of genomics into the methodology to understand working of disease, development of drugs, diagnostics, and clinical therapy. The direction is fundamentally "genomics to health" and not just production of research data without any ground results and now the time has come to use this knowledge for welfare of the society [1].

A major challenge faced by scientists is to decipher the molecular details that cause a disease. Only knowing the genetic basis of the disease is not enough, as understanding the molecular mechanism underlying the disease is necessary. For diseases that are oligogenic, synergistic contribution of genes from several loci could explain disruptions in their products, in particular when these proteins are directly or indirectly interacting. Various models like the dosage [2] and the Poisson [3] model are there that explain the molecular mechanisms of the disruption. The dosage model explains disruptions of two proteins within a complex. It states the relationship between a mutation and the phenotype. It says that mutation in one of the protein of a protein complex only weaken the interaction and thus does not cause changes in the phenotype. It is only when the proteins of the complex are mutated, then the interaction is lost and phenotype is affected. The Poisson model on the other hand says that the mutation in one protein can potentially disrupt the complex though but there are still enough other interacting proteins or unchanged complexes that can preserve the function.

The molecular models described earlier could be also used to explain indirect interactions between proteins (i.e. proteins that do not physically interact but participate in the same functional pathway). The increasing knowledge about protein networks can be used towards identifying new genes and genetic mechanisms behind diseases. For instance, if the gene products (proteins) have any functional interaction, one could trace these proteins back to their respective genes and identify the genes responsible for the disease. Identifying genes associated with complex diseases from all possible candidates generated from genome-wide genetic linkage studies would involve searching through hundreds of genes.

The energy of bioinformatics development has moved away from understanding networks encoded by model species towards understanding the networks underlying human disease [4]. Protein networks are increasingly serving as tools to unravel the molecular basis of disease [5]. Genome-wide association studies (GWAS) have increased in number in recent years, but the variants that have been found have generally explained only a tiny proportion of the estimated genetic contribution to phenotypic variation, suggesting that the current approach lacks the necessary power to detect the bulk of risk variants [6].Network analysis is now conveniently used to analyze the GWAS. Apart from the lack of statistical power, there are also other possible reasons for this: (1)If the variant is rare, the coverage of true genetic variants(SNPs) may still be low (2) phenotypic heterogeneity is more pervasive than suspected earlier; and (3) environmental heterogeneity among cohorts combined for

GWAS may hide the true signals.[7] Methods to arrive at high-precision predictions that are translatable to effective steps in disease prevention, diagnosis and prognosis should be the goal of PPI studies. Then, generated leads should be tested experimentally to determine their relevance [8].

2. REVIEW OF LITERATURE

In vivo, as observed protein doesn't perform as single isolates to carry their task. [9]. It can be revealed through the analysis of protein that is annotated that proteins work by forming an interaction network to carry out cellular processes [10]. We can even define a function of an unknown protein by finding out its interacting partner whose function is already known. By mapping protein- protein interactions, apart from deciphering the functions of the unknown protein, more important is the formation of functional pathways of cellular processing and to know the molecular mechanism behind each cellular function. To understand the working of proteins and how they function, it is quintessential to know and map protein-protein interaction networks. Annotating, characterizing and analyzing the protein-protein interaction networks of the given cellular proteome is now the way to understand the biochemistry of the cell.

There are many ways to show the results of interaction of two or more proteins that interact together to carry out a functional task. Phizicky and Fields show the measurable effects of protein interactions [11]. Protein interactions can:

- Cause subtle changes on substrate binding or allosteric binding effects resulting in the altering of the enzyme kinetics;
- Work together to provide substrate channeling;
- For small effector molecules they provide binding sites;
- Kill or inactivate a protein;
- Specificity for binding of a protein with its substrate can be changed by interactions of multiple proteins;.

Contrary to as earlier suspected, the protein-protein interactions are much more wide spread. They have the power for large amount of degree regulation. To completely be aware of their role and potential in a cell, one need to first identify the interactions, see to what is the degree of these interactions and what is the result of a particular interaction.

From some last years the databases of protein interaction networks are growing at an exponential pace thus providing wet lab biologist references to work upon[12-14], and also these repositories provide data for computer scientist to look for patterns or write algorithms to decipher the structure of protein networks[15]. With increased capacity and development of genome-wide experimental methods like the protein chips, two-hybrid test and mass spectrometric analysis, the number of reported interactions has increased exponentially. [16]. On one side, due to such increase in availability of protein interaction data has caused development of disease classifiers and other tools(Table 1) that exploit the inherent global properties of protein interaction network data and on the other side it has led to the challenge of providing better visualization and analysis tools that can utilize to the maximum the information stored in these networks.

It is now being question and debated the method of collection of the interaction data by these databases. It is being watched that the literature source through which data is taken must be reliable [17]. In this debate, a public repository is to organize experiments supporting PPIs and collect data and organize into comprehensive sets of accurately annotated data [18].

Table 1 List of various protein interaction network analysis Tools

Tool	URL	Features
BioLayout Express 3D	http://www.biolayout.org/	Analyse microarray data
Cytoscape	http://www.cytoscape.org/	Contains many apps and different ways of visualization
Large Graphic Layout(LGL)	http://sourceforge.net/projects/lgl	Is used to represent large graph
Osprey	http://biodata.mshri.on.ca/osprey/servlet/Index	Have different layouts and network and connectivity filters
Pajek	http://vlado.fmf.uni-lj.si/pub/networks/pajek/	Is exclusively used to visualize large data networks
Visant	http://visant.bu.edu	Gene Ontologies can be analysed using this tool

Other than this the user must know the different types of interaction data repositories available, the difference between them and which databases are most annotated and comprehensive.

For instance, the KEGG pathway has such way of representation that depicts the direction and properties of the links, but on the other hand PPI network does not provide such directional information. The bimolecular elements (i.e., the nodes) in both networks are generally similar, and the information that can be deduced from them is complementary, each single view being enriched by the other. One thing that KEGG will miss is the interaction between two different pathways while PPI data can miss some molecules of the pathway. In conclusion, the use of PPI data combined with related pathways allows for a useful and detailed exploration of protein networks. This approach may bring about better comprehension of the complex functional roles that the proteins play by physically interacting in living systems.

As found the tools and database of protein interaction network are complex to use. Some tools require interaction data to make their visualization like Cytoscape. Some tools provide

only the static graphic tool only showing limited interactions. To use such tools user have to first go to interaction databases and then collect dataset from them and then input these dataset in shortest path finding tools. Hence we tried to make things simpler for the user by making such a tool where he can just enter a protein name and get static graphical view of interaction network with proteins interacting with it.

3. METHODOLOGY

The procedure can be broadly divided into following subparts –

1. Use of Online resource of String Database to download Protein Interaction Network Database.
2. Use of Awk and Shell scripts to manipulate the downloaded data.
3. Use of Uniprot-Id mapping service
4. Writing of C++ to develop the tool(Fig. 1).
5. Working with CGI(Common Gateway Interface) to combine HTML code and C++ code to form the GUI.
6. Connecting Cytoscape with the GUI.

```
string readuni()
{
   ifstream ip("uniprot.csv");
   if(!ip.is_open()) std::cout << "ERROR: File Open" << '\n';

   while(ip.good())
   {
   getline(ip, uniprot_id,',');
   getline(ip,stringdb_id,'\n');

   name_uni[uniprot_id] = stringdb_id;
   }
   cout<<"Enter the Uniprot ID of protein: ";
   getline(cin,uniprot_id);
   cout<<"\nThe StringDB ID of protein is: "<<name_uni[uniprot_id]<<endl;
   ip.close();
   return name_uni[uniprot_id];
};

void readconfi()
{
 //Reading Confident Human Interactions file

   ifstream ix("confident.csv");
   if(!ix.is_open()) std::cout << "ERROR: File Open" << '\n';

   while(ix.good())
   {
   getline(ix, prot1,',');
   getline(ix, prot2,',');
   getline(ix, score,'\n');
   name_confi.insert(makepair(prot1,prot2));
   }

   cout<<"\nEnter the StringDB ID of protein for which you want path for: ";
   getline(cin,prot1);

   for(it = name_confi(); it != name_confi.end(); it++)
      cout<<it->first<<" "<<it->second<<endl;
   ix.close();
}
```

Fig. 1: Source Code.

First of all, protein network data for *Homo Sapiens* with Taxonomy ID 9606 was downloaded from STRING database. Awk and shell scripts was used to extract interactions data of *Homo Sapiens* from the database whose confidence is greater than 0.7. Then, the String IDs were converted into Uniprot IDs so that the user of the tool can enter the uniprot id of his protein name. Then, C++ code was developed for making the protein circle for input of any uniprot id of a protein.

4. CONCLUSION

The result is a protein interaction network circle, whose every protein has at least one interaction connection with any other protein of that circle. The present tool, ProteinCircle, provides a visual protein-protein interaction network to show the potential of a protein to influence the other protein in its circle. The proteins of one circle can be influenced by each other by means of passing the message from one protein to another and so on. But this chain of passing the message will break as soon as a protein lying outside the circle (which is essentially the protein which has no interaction with any protein in that circle) comes.

REFERENCES

[1]Collins, FS, Green, ED, Guttmacher, AE, Guyer, MS., *A vision for the future ofgenomics research,* Nature, 2003, 422(6934):835.

[2]Fuller MT, *Interacting genes identify interacting proteins involved in microtubule function in Drosophila.* Cell Motil Cytoskeleton, 1983, 114:128–35.

[3]Stearns, T, Botstein D, *Unlinked noncomplementation:isolation of new conditional-lethal mutations in each of the tubulin genes of Saccharomyces cerevisiae,* Genetics, 1988, 119:249–60.

[4]Kann, M.G, *Protein interactions and disease: Computational approaches to uncover the etiology of diseases*, Brief, Bioinform, 2007, 8:333–346.

[5]Ideker, T, Sharan, R, *Protein networks in disease* ,Genome Res,2008, 18: 644-652.

[6]McCarthy, MI, Hirschhorn, JN, *Genome-wide association studies:Potential next steps on a genetic journey*, Hum Mol Genet 17, 2008, R156–R165.

[7]Pedroso, I , Breen, G, *Gene Set Analysis and Network Analysis for Genome-Wide Association Studies*, Cold Spring Harbour Protocols, 2011.

[8]Rao, A Bulusu, GK, Srinivasan, R, Joseph T,*Protein-Protein Interactions and Disease,* Chapter 8 , Protein Interactions, 2012.

[9]Yanagida, M Chromatogr, B, Functional proteomics; current achievements, 2012, J. 771:89–106.

[10] Mering, C., Krause, R., Snel, B.; Cornell, M. Oliver, S.G., *Comparative assessment of large-scale data sets of protein-protein interactions*, Nature, 2002, 417:399–403.

[11]Phizicky, E.M. Fields, S, *Protein-protein interactions: methods for detection and analysis*, Microbiol. Rev., 2011, 59:94–123.

[12]Jain, E Bairoch, A Duvaud, S Phan, I Redaschi N *et al., Infrastructure for the life sciences: design and implementation of the UniProt website,* BMC Bioinformatics, 2009, 10: 136.

[13]Apweiler, R; Martin, MJ; O'Donovan, C; Magrane, M; Alam-Faruque, Y; *et al., The Universal Protein Resource (UniProt) in 2010*. Nucleic Acids Res, 2010, 38: D142–D148.

[14]Cusick, ME; Klitgord, N; Vidal, M; Hill, DE, *Interactome: gateway into systems biology*, Hum Mol Genet, 2005, 14 Spec No. 2: R171–R181.

[15]Blow, N, S*ystems biology: Untangling the protein web*, Nature,2009, 460: 415–418.

[16]Mackay, JP; Sunde, M; Lowry, JA; Crossley, M, Matthews JM, *Protein interactions: is seeing believing?* Trends Biochem Sci, 2007, 32: 530–531.

[17]Cusick, ME; Yu, H; Smolyar, A; Venkatesan, K; Carvunis, AR*; et al, Literature-curated protein interaction datasets.* Nat Methods, 2009, 6: 39–46.

[18]Salwinski, L; Licata, L; Winter, A; Thorneycroft, D; Khadake, J; *et al. Recurated protein interaction datasets,* Nat Methods, 2009, 6: 860–861.

[19]Fields, S; Song O. *A novel genetic system to detect protein-protein interactions.* Nature, 1989, 340:245–6.

[20]Giot, L.; Bader, JS; Brouwer,C.; *et al., A protein interaction map of* Drosophila *melanogaster*, Science, 2003, 302:1727–36.

[21]Ito,T; Tashiro, K; Muta, S; *et al., Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins,* Proc Natl Acad Sci U S A, 2000, 97:1143–7.

[22]Li, S; Armstrong, CM; Bertin, N; *et al. A map of the interactome network of the metazoan C. elegans,* Science, 2004, 303:540–3.

[23]Rual, J F; Venkatesan, K; Hao, T; *et al., Towards a proteome-scale map of the human protein-protein interaction network,* Nature, 2005, 437:1173–8.

[24]Stelzl, U; Worm U; Lalowski, M; *et al, A human protein-protein interaction network: A resource for annotating the proteome,* Cell, 2005, 122:957–68.

[25]Uetz, P; Giot, L; Cagney, G; *et al, A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae,* Nature, 2005, 403:623–7.

[26]Gavin, AC; Bosche, M; Krause, R; *et al., Functional organization of the yeast proteome by systematic analysis of protein complexes*, Nature, 2002, 415:141–7.

[27]Ho, Y; Gruhler, A; Heilbut, A; *et al, Systematic identification of protein complexes in* Saccharomyces cerevisiae *by mass spectrometry*, Nature, 2002, 415:180–3.

[28]C; Huynen, M; Jaeggi, D; Schmidt, S; Bork P, Snel B, *STRING: a database of predicted functional associations between proteins,* , 2003, 1;31(1):258-61.

[29]Stutz, M, *Get Started with GAWK:AWK language fundamentals.developerWorks*, IBM, 2006.