Human Protein Meta-Interaction Database (HPMD) Potentiates Integration for Meta-Analysis

Gil Alterovitz[1], Dmitriy Patek[2], Isaac S. Kohane[2,3], and Marco F. Ramoni[2,3]
[1] Health Science and Technology/Electrical Engineering & Computer Science,
Massachusetts Institute of Technology, Cambridge, 02139, USA.
[2] Children's Hospital Informatics Program, Harvard Medical School, Boston, 02115, USA.
[3] Harvard Partners Center for Genetics and Genomics, Boston, 02115, USA.

***Abstract*— Difficulties in integration of biological databases has been a long standing issue [1]. Problems in combining databases include different identification schemes, redundancies, and varying levels of information cataloged within each database. This work seeks to address these concerns and design a corresponding implementation. Analysis via Monte-Carlo samples the shortest distances between pairs of proteins.**

The result of this work is the creation of the largest human protein-protein interaction database, containing over 103,000 interactions from over half a dozen databases/sources. For pairs of proteins within the same component (graph), the Monte Carlo simulation suggested the minimum separation distance was 4.3 interactions.

An implemented version that can dynamically integrate standard Proteomics Standards Initiative- Molecular Interaction (PSI-MI)-formatted databases (e.g. DIP, IntAct, MINT, cPath, HPRD) is available via web site request: http://chocolate.chip.org/∼protcoop/bap_req.html

I. Introduction

As the human genome project draws to a close, recent work has decreased the estimate of the number of genes to between 20-25 thousand, not far from the number of genes in a simple worm (i.e. C. elegans) [2]. Thus, the complexity of humans must be derived from other sources such as the interactions of the genes' products, or proteins [3]. A number of different databases and standards have emerged for cataloging protein-protein interactions [4]. Each contains different types of data sources. For example, some databases exclusively store interactions inferred from orthologs that have been shown to interact- while others contain direct human protein interactions. Even when directly measured, there can be a differences in quality and quantity based on whether the experiment was done in automated manner versus manually.

What is needed is a standard. One relatively new standard is the PSI-MI format [5]. While that initiative seeks to standardize the structure of databases, the actual content is left rather ambiguous. Plus, data in these fields can vary somewhat across databases. So, for those databases that actually support PSI-MI format, even the actual proteins can be referred to by different identifiers ranging from Uniprot, NCBI GI numbers, Ensembl, and International Protein Index (IPI) references. Also, virtually no database contains all of the PSI-MI format fields. Both inter- and intra-database redundancies of cataloged interactions is a common problem. What is needed is a way to establish a common language between these databases so they can be integrated (and without redundancies)- not just a common structured format as exists with PSI-MI.

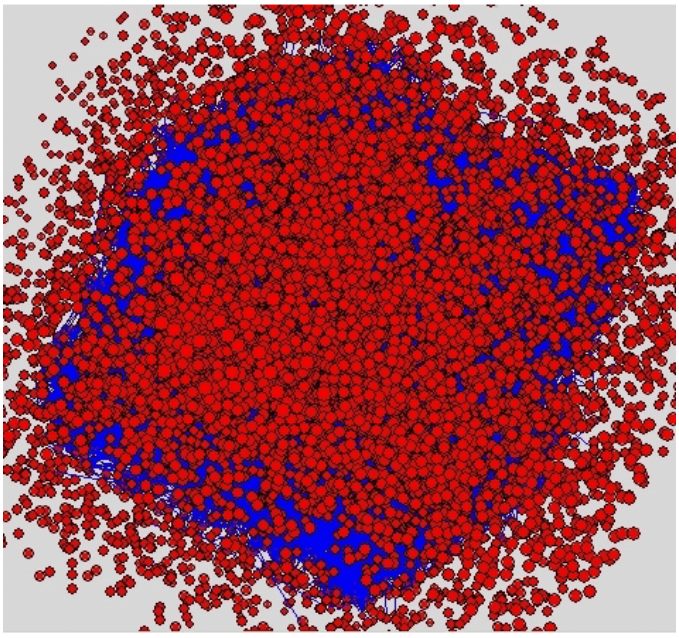

Fig. 1. 3-D visualization of the Human Protein Meta-Interaction Database

II. Methods

In this paper, an automated protocol was designed and implemented in Matlab. Using this, seven protein interaction databases were integrated into the Human Protein Meta-Interaction Database (HPMD). These include: DIP [6], IntAct [7], MINT [8], BIND [9], cPath, HPRD [10], and the Sanger Institute Interaction Map [11].

First, the XML/flat files of databases were parsed. Then, the different protein identification numbers were converted to NCBI Entrez Protein GI numbers. This was done by sequentially querying SeqHound [12] via remote Java Application Protocol Interface and AliasServer [13] through Simple

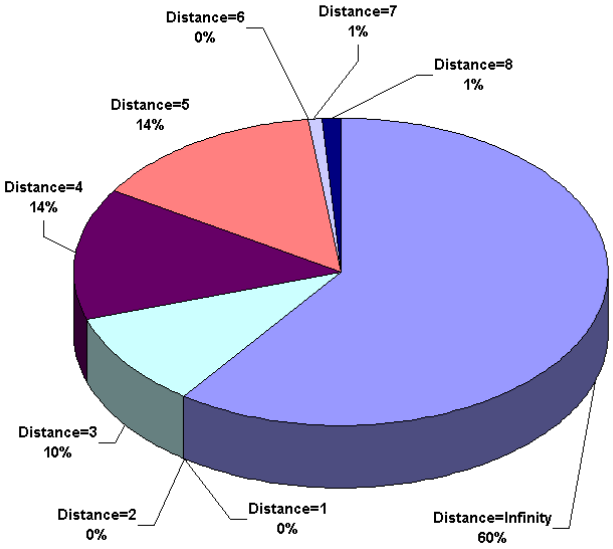

Fig. 2. Pairwise Dijkstra shortest distances

Object Access Protocol (SOAP). Also, the IPI cross-reference indexes, Ensembl cross-reference indexes, and Entrez Protein database were queried to match the disparate ID's with appropriate NCBI GI numbers. Next, SeqHound was used to find redundant GI numbers. The best annotated version of the protein (from a group of database entries referring to the same protein sequence) was then used. With a common identifier, the databases could then be merged- with duplicates with removed from the new collection.

The meta-database was then saved into several formats for graph analysis and visualization including Pajek, dot, and GraphML. Once the protein interactions were in the form of a graph, a Monte-Carlo simulation was done to analyze the network connectivity. Here, a pair of proteins is selected at random from HPMD. The pairwise Dijkstra shortest distances were then calculated between these proteins.

III. Results

The Human Protein Meta-Interaction Database integrates seven major protein-protein interaction databases for a total of 103,657 interactions (edges), 16,683 proteins (vertices), and 5,733 components (separate graphs)- making it by far the largest human protein-protein interaction database.

A 3-D version Fruchterman-Reingold Force-directed Placement algorithm [14] was used to plot HPMD within the Pajek environment [15]. This visualization is shown in Figure 1. The outer 'cortex' potion of the cube contains the proteins vertices, while the inner 'medulla' contains interaction edges.

Figure 2 shows the results of the Monte-Carlo simulation. Each slice is proportional to the percent (see labels) that a randomly chosen protein pair was connected by a minimum of **n** edges. If the proteins were in unconnected graphs, then the distance was classified as infinity. For protein pairs within the same component (graph), the average pairwise Dijkstra shortest distance was just over 4 (i.e. 4.3 edges).

IV. Discussion and Conclusion

This paper approaches the first step in database integration via establishing a common language for some crucial database fields. More work can be done on formalizing these notions and developing standards that allow translation and migration between databases- rather than centralized integration as was done here.

In order to fully leverage the knowledge contained in this database, information from proteins and their interactions will need to be examined within the context of existing gene expression and regulation networks. Additional work in this area could potentially integrate such knowledge by connecting protein interaction network information with DNA-binding proteins and other elements that bridge genomics and proteomics.

V. Acknowledgement

This work was supported in part by a fellowship from the Whitaker Foundation.

References

[1] L. D. Stein, "Integrating biological databases," *Nature Reviews Genetics*, vol. 4, pp. 337-45, 2003.
[2] I. Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931-945, 2004.
[3] G. Alterovitz, E. Afkhami, and M. Ramoni, "Robotics, Automation, and Statistical Learning for Proteomics," in *Focus on Robotics and Intelligent Systems Research*, vol. 1, F. Columbus, Ed. New York: Nova Science Publishers, Inc., 2005 (In press).
[4] I. Xenarios and D. Eisenberg, "Protein interaction databases," *Current Opinion in Biotechnology*, vol. 12, pp. 334-9, 2001.
[5] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, et al., "The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data," *Nature Biotechnology*, vol. 22, pp. 177-83, 2004.
[6] I. Xenarios, L. Salwinski, X. J. Duan, et al., "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Res*, vol. 30, pp. 303-5, 2002.
[7] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, et al., "IntAct: an open source molecular interaction database," *Nucleic Acids Res*, vol. 32 Database issue, pp. D452-5, 2004.
[8] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, et al., "MINT: a Molecular INTeraction database," *FEBS Letters*, vol. 513, pp. 135-140, 2002. [9] G. D. Bader,
[9] D. Betel, and C. W. Hogue, "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Res*, vol. 31, pp. 248-50, 2003.
[10] S. Peri, J. D. Navarro, R. Amanchy, et al., "Development of human protein reference database as an initial platform for approaching systems biology in humans," Genome Res, vol. 13, pp. 2363-71, 2003.
[11] B. Lehner and A. G. Fraser, "A first-draft human protein-interaction map," *Genome Biology*, vol. 5, pp. R63, 2004.
[12] K. Michalickova, G. D. Bader, M. Dumontier, et al., "SeqHound: biological sequence and structure database as a platform for bioinformatics research," *BMC Bioinformatics*, vol. 3, pp. 32, 2002.
[13] F. Iragne, A. Barre, N. Goffard, et al., "AliasServer: a web server to handle multiple aliases used to refer to proteins," *Bioinformatics*, vol. 20, pp. 2331-2, 2004.
[14] T. M. J. Fruchterman and E. M. Reingold, "Graph Drawing by Force-directed Placement," *Software: Practice and Experience*, vol. 21, pp. 1129 - 1164, 1991.
[15] V. Batagelj and A. Mrvar, "Pajek - Program for Large Network Analysis," *Connections*, vol. 21, pp. 47-57, 1998.