Protein-Protein Interaction Networks: Predicting Protein Functionality

By Brian Rappaport, ECE '18

Introduction

Proteins are present in all forms of life, from bacteria up on to humans, and a human cell can contain more than a billion of them. Proteins are responsible for all work within the cell, from breaking down glucose and water to provide energy to managing DNA and RNA transcription to create templates for new proteins. They are made up of chains of amino acids (basic biological molecules); since there are 20 amino acids and proteins can be hundreds of acids long, there are a huge number of possible proteins, even though many combinations are impossible.

Although individual proteins are important and worthy of study, it is in the sequence of proteins working together that true complexity achieved, so protein-protein interactions and their networks are studied in many different areas of cell biology. In our work this semester, we used the interaction networks for a basic task: clustering different proteins by their function within the cell, from respiration to transportation.

Proteins

Structure

Protein structure is described at four distinct levels. At the first level, called the primary structure, this refers to the chain of amino acids itself. The next level, the secondary structure, is made up of common structures formed by the strand in the course of folding, such as α-helixes or β-strands. Tertiary structure describes the arrangement of the protein itself as a whole, and quaternary structure determines how several proteins interact with each other. The fourth level is mainly what interests us in this project, since it determines which proteins will bind together and which will not, which should imply a similar function.

Function

The sheer number of ways proteins can be put together and interact with each other means the study of the structure is a daunting task, so for practical results proteins are often studied by their functions instead.

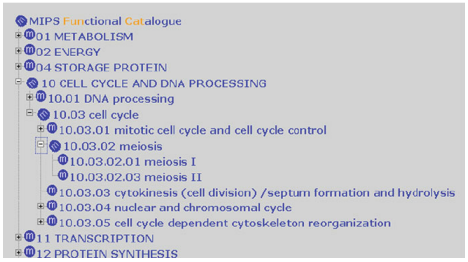

Figure 1: Section of MIPS Functional Catalogue annotations. This example shows labels four levels deep, to the different types of meiosis.

The large number of possible uses has been carefully studied and many classifications are available in different databases; for example, the MIPS (Munich Information center for Protein Sequencing) Functional Catalogue provides annotations at up to six levels of specificity, with more than 1300 different annotations. These databases are used to verify predicted results.

Interactions

Proteins can function paired in many different ways; for instance, a protein can carry another protein through a membrane, or multiple enzymes (proteins which act as biological catalysts) may work together to produce macromolecules for use in cellular respiration or many other domains. Quaternary protein structures often take the form of *protein complexes*, which are groups of associated chains of the polypeptides (amino acids) making up a larger structure with a particular utility. These can be assembled by protein *subunits*, and these subunits are commonly used to determine interactions: if a protein subunit can assemble two proteins together into a complex, there exists an interaction between them. Recent research has shown that essential genes (those required for an organism to stay alive) are disproportionally found in protein complexes; these are also often strongly correlated in interaction mechanisms.

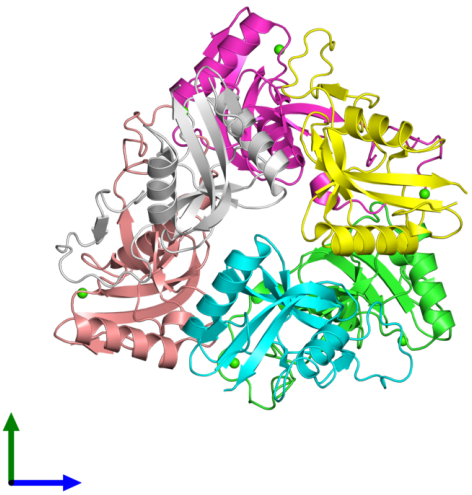

Figure 2: An example of a protein complex found in snake venom. Note that 6 different peptide chains are involved in a single complex.

Testing for Protein Interactions

Several tests are commonly employed to identify protein-protein iteractions. One of the most frequently used is a test known as two-hybrid screening. In this test, two proteins that might interact are each bound together, which form a transcription factor (a protein which will write a particular type of DNA) only if the two proteins interact. Then the resulting cells are tested for the presence of that type of DNA. If the two proteins interacted, the transcription factor is created, and the DNA will be produced, but if it is not present, then the two proteins must not have interacted. Another test, tandem affinity purification, involves modifying the cells to give a "tag" to a particular protein, letting the protein bind to anything, then separating out everything not bound to the protein and seeing which proteins are left in the system, all of which must have been bound to the original protein.

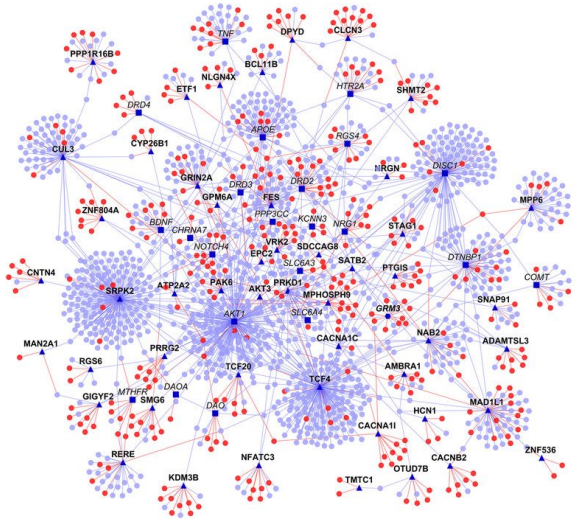

Figure 3: Protein-protein interaction network, representing proteins found in patients with schizophrenia.

Both of these methods have obvious disadvantages: they require lab work and a lot of time to perform the experiments correctly. Each of these tests also only test a single protein or a single interaction. Given the astronomical number of possible interactions, or even the much-less-but-still-large number of actual proteins in a human cell (20,000 - 100,000), these tests are infeasible on the scale required. Other work, including our work, uses the (relatively) small number of known interactions to predict others, using various algebraic and machine learning methods. In this way, interactions can be predicted much more fully and comprehensively than would be possible in labwork.

Challenges

Many problems in biology are hard because there is no known ground truth to go off of, and clustering protein-protein interaction networks is such a problem. Several databases exist containing labels, but all are incomplete and some contradict each other. Moreover, the majority of testing has been done on simple organisms, primarily yeast, but many connections that we hope to detect are more complex than perhaps yeast would exemplify at all. In clustering, we did not take structure into account, but rather treated the interactions as black boxes: in reality, there is much more that we could get from knowing the structure in addition, at the cost of increased complexity. Finally, just because two proteins interact does not necessarily mean they are complementary in function.

Conclusion

Protein-protein interaction networks are an important factor in cell biology. A full map of these interactions would allow biologists to have a much better understanding of many diseases and how to cure them, along with many other benefits. It is our hope that our work will help to facilitate the greater understanding of these influential networks.

References

[1] M. Cao, "New Graph Metrics Improve Network-based Protein Function Prediction." pp. 1–137, 2016.

[2] Ganapathiraju, Madhavi K. et al. "Schizophrenia Interactome with 504 Novel Protein–protein Interactions." NPJ Schizophrenia 2 (2016): 16012. PMC. Web. 2017.

[3] Braun, P. and Gingras, A.-C. "History of protein–protein interactions: From egg-white to complex networks." Proteomics, 12: 14781498. 2012.

[4] Ruepp, A. et al. "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." Nucleic Acids Res 32, 5539-5545. 2004.

[5] Mizuno H, Fujimoto Z, Koizumi M, Kano H, Atoda H, et al. "Structure of coagulation factors IX/X-binding protein, a heterodimer of C-type lectin domains." Nat. Struct. Biol. pp. 438–41. 1997.

[6] Hart, G Traver, Insuk Lee, and Edward R Marcotte. "A High-Accuracy Consensus Map of Yeast Protein Complexes Reveals Modular Nature of Gene Essentiality." BMC Bioinformatics 8 (2007): 236. PMC. Web. 17 Apr. 2018.