Data-mining open source databases for drug repositioning using graph-based techniques

The analysis of 'Big Data' has great potential in drug discovery; however complications arise in integrating this data in a principled and coherent way. An important statistical tool to manage complexity is that of graph theory, which is as satisfying and attractive to the hard core data analyst as it is to the lay person. The statistician can marvel at the mathematics behind the theory while the lay-person can appreciate the highly visual and graphical information that shows links between objects and their relationships. In this paper we show how graphs and proteomic data are used to facilitate drug discovery.

By Dr Ken McGarry and Ukeme Daniel

Our knowledge of cellular functions and processes has been improved by the growth and complexity of biological data generated in recent years. There are now numerous resources available such as gene sequences, protein to protein interactions, chemical to protein interactions and several ontologies that structure and label known biological knowledge. This has led to an increased use of computational statistical techniques to pre-process, analyse and store this data. Other techniques such as the graphical representation of biological structures and relationships are important in order to display this complex information in a meaningful way. Furthermore, graph theory is a mathematical technique that can be applied to explain and predict the connectivity patterns between interacting proteins. Interestingly, it is not just proteomics and genomics that can benefit from this approach, many fields such as online social networks (eg data-mining Twitter or Facebook), collaborative networks between scientists, political science, marketing and economic systems have all utilised graph theory. In fact any system that contains interacting 'entities' that collaborate or connect with each other in a network are prime candidates for a statistical analysis using graph theory.

Graphs are generally with multi-edges (connections) and often with directed edges when cause and effect can be inferred, for instance in **Figure 1**, the graph can be described mathematically $G = (V, E)$ where V denotes a set of nodes (vertices) and E of pairs (u, v), $u, v \in V$ denotes a set of links called edges and if the pairs are unordered the graph is said to be undirected and adjacent nodes is termed the degree of a node and each edge that connects to v its adjacent node is incident to v. Based on the connectivity patterns of a network some statistical values can be computed that can explain the relationships between the nodes. Graph theory has provided us with several measures that can be applied to all scientific or social applications that describe the connectivity patterns such as hubs, locality, modularity, closeness, betweeness and centrality. In **Figure 2** we have displayed a fraction of a simple protein interaction network. Care needs to be taken when displaying network diagrams as it is all too easy to have a complete 'black out' in terms of a mass of nodes and incomprehensible connections.

In relation to drug discovery or drug repositioning we are interested in protein to protein and protein to chemical interactions. Proteins perform many functions within the cell and generally do not act alone but co-operate with other proteins with functionally similar roles in logical modules. Recent advances in high-throughput proteomic technologies have generated huge databases of known protein interactions. These interactions can be described by graphical networks, where the protein is the node or edge and the interaction between them is depicted by a link or vertice.

There are several different types of protein interaction databases such as Bimolecular Interaction Network Database (BIND), Database of interacting proteins (DIP), Search Tool for Interactions of Chemicals (STITCH) (which we used), The Human Protein Reference Database (HPRD), Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Protein-Protein Interaction Predictions (PIPS). These databases were brought into being by two different bodies known as the Proteomic Standard Initiative (PSI) and Human Proteome Organization (HUPO) and supported by Protein Interaction Data providers. In **Table 1** we highlight some of the biological databases that have seen much use in recent years.

The protein interaction databases such as the STRING collection presents data captured from several varieties of experiment's such as Y2H (yeast-2-hybrid) and even information extracted by automated text mining of scientific papers. They contain information indicating the source of information, the interaction between two proteins and a confidence score based on the accuracy and fidelity of the method, ie Y2H experiments typically have high scores while text mining usually has a much lower score.

DrugBank is a unique bioinformatics and chemoinformatics resource; it is an online database containing an extensive pharmaceutical, pharmacological and biochemical information about drugs, their targets and mechanism. The latest version of DrugBank is 4.0 and has been updated to contain more information on drug absorption, distribution, metabolism, toxicity, excretion, drug target and information on structure activity relationships.

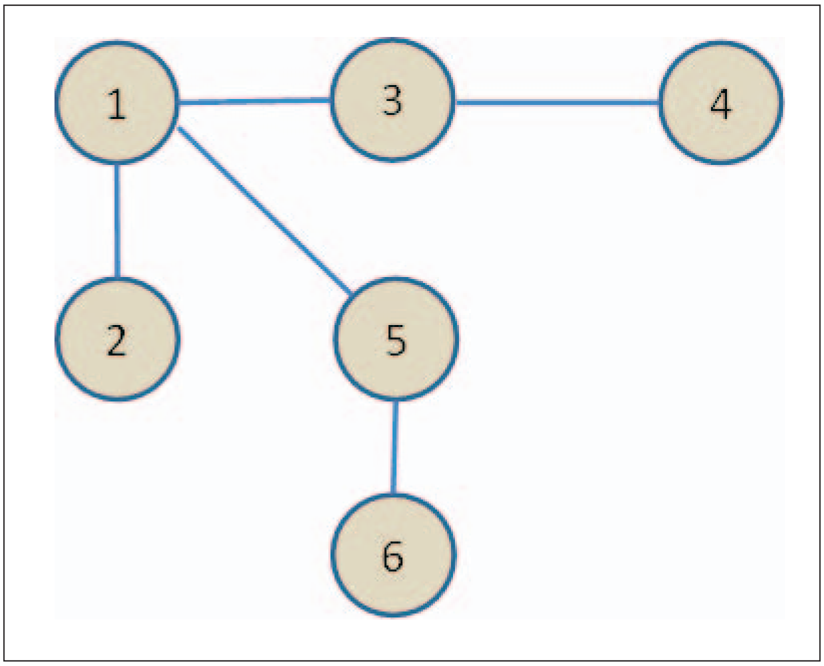


Figure 1: Simple undirected graph, consisting of six nodes and five connections (edges). Where the vertices or nodes are V = {1,2,3,4,5,6} and the edges are denoted by E = {{1,2},{1,3},{1,5},{3,4},{5,6}}

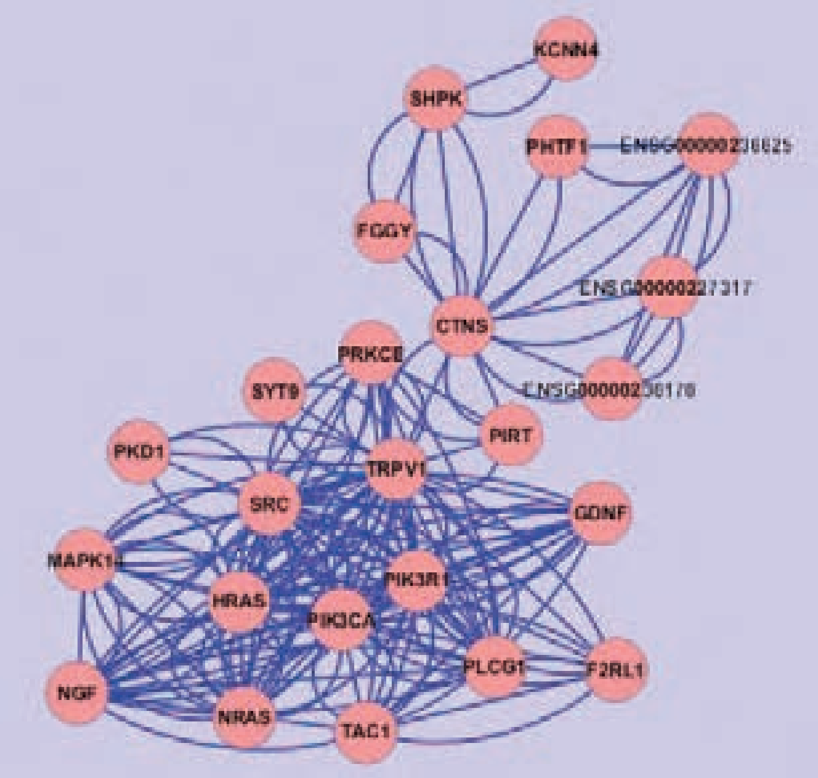


Figure 2: This shows a tiny fraction of the protein-to-protein interactions involved with our subset of lysosomal diseases. The screenshot is taken from the Cytoscape visualisation software which is a useful package for displaying network data. Here we use the basic settings, more elaborate set-ups can be used such as having different shapes and sizes of nodes along with different colours[4]

Table 1: Biodatabases used for constructing protein networks

DATABASE NAME	ABBREVIATION	DATABASE	STATISTICS	WEBSITE
Search Tool For Interaction Of Chemicals	STITCH	PPI & protein-chemical interaction	1,133 organisms, 2.6 million proteins	http://stitch.embl.de/
Biological General Repository Interaction Datasets	BioGRID	PPI and others	49 organisms, 500, 239 non-redundant interactions, 725,045 raw interactions	http://thebiogrid.org/
MalaCards – The Human Disease Compendium	MalaCards	HDN	19,486 diseases from 63 sources, 82,947 interactions, >9 species of organism	http://www.malacards.org/
Database of Interacting Proteins	DIP	PPI	665 organism, 26,743 proteins, 77,514 interactions, 6861 articles	http://dip.doe-mbi.ucla.edu/dip/News.cgi
Gene Ontology	GO		36 organisms	http://www.geneontology.org/GO.contents.doc.shtml
BioCyc database collection	BioCyc	Ecocyc & Metacyc databases	Ecocyc obtains information from 17,000 publications, 1 organism only (E. coli), Matacyc 19,000 articles sited, 1,200 metabolic pathways, 1,500 organisms	http://biocyc.org/
PDZ domain protein-protein interaction database	PDZBase	PPI	339 interactions, > 200 articles	http://abc.med.cornell.edu/pdzbase
HIV-1, Human Protein Interaction Database at NCBI	HIV-1 at NCBI	HIV and PPI	1 organism (Homo sapiens only)	http://www.ncbi.nlm.nih.gov/projects/RefSeq/HIVInteractions/
Kyoto Encyclopedia of Genes and Genomes	KEGG	PPI and others	3,086 organisms, 231,041 articles, 299,364 pathways	http://www.genome.jp/kegg/
DrugBank	DrugBank	Protein-chemical database	4,282 non-redundant protein, 1,549 FDA approved drugs, 155 FDA- approved biotech, 89 nutraceuticals and > 6,000 experimental drugs	http://www.drugbank.ca/
Molecular Interaction Database	MINT	PPI only	35,553 proteins, 241,458 interactions, 5,554 pmids	http://mint.bio.uniroma2.it/mint/Welcome.do

DrugBank has become the referential source of drug data for some other databases, for example KEGG and UniProt.

Our own interests involve drug repositioning for orphan diseases and the development of pro-drugs, our recent work and the work of colleagues has concentrated on Lysosomal disorders and in particular Cystinosis. Any disease that affects a small fraction of the population is termed a rare or an orphan disease. Some of these diseases are genetics and therefore are seen throughout the whole life of an affected individual, some of these diseases appear early in life and approximately 30% of children with these rare diseases die before the age of five years. There is no cut-off number to classify a disease as rare, for instance, according to the United States Rare Disease Act of 2002, any condition that affects less than 200,000 individuals is a rare disease, while in Japan, a rare disease is any disease or condition that affects less than 50,000

Table 2: The available treatments for lysosomal storage diseases

DISEASE	GENERIC NAME	TRADE NAME	MANUFACTURER	INDICATION
Cystinosis	Cysteamine	cystaran™	Sigma-Tau Pharmaceuticals Inc	Cystaran™ is an ophthalmic solution: a hydrochloric, cysteine-depleting agent for treating patients with cysteine crystal accumulation
	Cysteamine	cystagon® and procysbi	Mylan Pharmaceuticals Inc and Raptor Pharmaceuticals Corporation respectively	Cystagon® and Procysbi are bitartrate salts which have been used for the oral treatment of nephropathic cystinurea and cystinosis. For treatment of alpha-galactosidase A deficiency known as Fabry disease
Fabry disease	Agalsidase-beta	Fabrazyme®	Genzyme Corporation	Agalsidase-beta is used for the treatment of alpha-galactosidase A deficiency also known as Fabry disease. http://www.genzyme.com/
Gauchers disease	Imiglucerase	Cerezyme®	Genzyme Corporation	Cerezyme® is an enzyme replacement therapy for adults and pediatrics diagnosed with type I Gauchers disease. http://www.genzyme.com/
	Velaglucerase alpha	VPRIV®	Shire Human Genetic Therapies	Velaglucerase alpha is an enzyme replacement therapy: a hydrolytic lysosomal glucocerebroside-specific enzyme for treating pediatric and adult patients with type I Gaucher disease
	Taliglucerase	Elelyso™	Pfizer	For treating adults with type I Gaucher disease
Glucagon storage disease	Alglucosidase alpha	Lumizyme®	Genzyme Corporation	Lumizyme is a lysosomal glycogen-specific enzyme used in treating patients 8 years and older with non-infantile onset of Pompe disease without evidence of cardiac hypertrophy. http://www.genzyme.com/
	Alglucosidase alpha	Myozyme®	Genzyme Corporation	Used in treating patients with infantile onset of Pompe disease to improve ventilator-free survival. http://www.genzyme.com/
Mucopolysacch aridosis	laronidase	Aldurazyme	Genzyme Corporation	Used in treating patients with Scheie and Hurler forms of mucopolysaccharidosis I with moderate to severe symptoms. http://www.genzyme.com/
	Idusulfase	Elaprase	Shire Human Genetic Therapies	For the treatment of Hunter syndrome or mucopolysaccharidosis ii in patients aged 5 and older
	Galsulfase	Naglazyme™	BioMarin Pharmaceutical Inc	For treating children and adults with Mucopolysaccharidosis vi

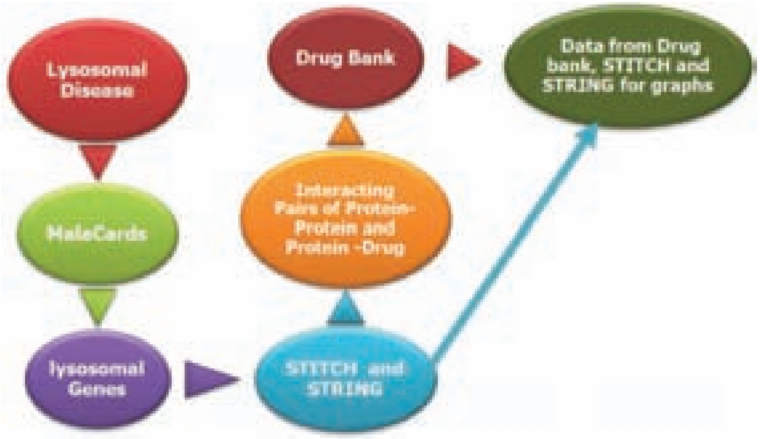

Lysosomal Disease
Drug Bank
Data from Drug bank, STITCH and STRING for graphs
MalaCards
Interacting Pairs of Protein-Protein and Protein -Drug
lysosomal Genes
STITCH and STRING

Figure 3
Methodological model for identifying potential therapeutic candidate treatment for Lysosomal diseases

patients and the European Commission defines rare disease as life-threatening conditions that occur in 1:2,000 people. However, this definition varies from country to country.

There are approximately 6,000-7,000 known rare diseases and about 25 million Americans or approximately 10% of the US population suffer from them. Most of these rare diseases are as a result of an alteration in a gene that substantially affects life expectancy, is severely disabling, impairs mental and physical abilities, reduces patient quality of life and their potential for education and earning capabilities.

In the past the process of repositioning was always by serendipity, for example thalidomide was initially prescribed to pregnant women as a sedative but was withdrawn from the market in 1961 due to its teratogenic effect. However, it later received approval from FDA for the treatment of glioblastoma and renal cell carcinoma and a complication in leprosy. In addition, metformin is the most commonly prescribed drug for treating individuals with type II diabetes mellitus. Recent studies have shown that metformin has anticancer properties most significantly for colon cancer, hepatocellular carcinoma and pancreatic cancer.

One of the main components of eukaryotic cells are the lysosomes. These cellular machines have a

MAXimize reproducibility. Minimize costs.

pipetmaX™

Advance your science and increase the pace of your experiments by allowing PIPETMAX to do the routine pipetting tasks. Focus on what is really important to your research: scientific analysis and publications.

Accomplish more by automating your sample prep protocols for:

- qPCR
- PCR
- NGS library prep
- ELISA
- most tip-based protocols

GILSON®

instruments • accessories • software • solutions ▶ Meet MAX at **www.pipetmax.com**
Visit www.gilson.com to download Speaking Volumes, our interactive publication for your tablet!

semipermeable membrane that contains multiple transporters and about 50 different enzymes that degrade extracellular material. The problems occur with the intralysosomal accumulation of sphingolipids, mucopolysaccharides and oligosaccharides leading to deficiency in lysosomal proteins. Deficiency occurs as a result of interruption in the metabolic pathways that recycle the degrading product of the macromolecules; the direct consequence of the genetic effect is as a result of progressive accumulation of these primary and secondary storage products and the formation of histological storage lesions.

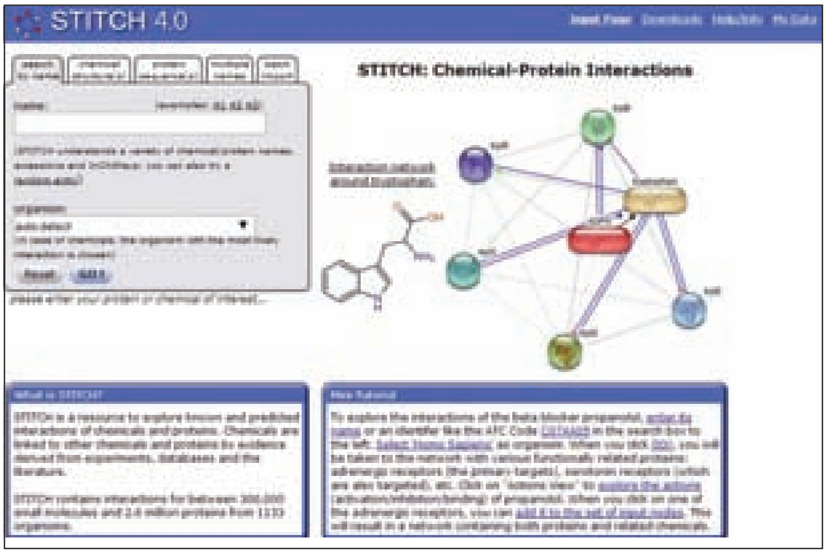
Cystinosis is of particular interest to us and occurs as a result of deficiency in lysosomal transport mechanism for cysteine and mutations in CTNS (Cysteine Transport Nephrotic Syndrome). CTNS is a gene on chromosome 17 that encodes the cysteine transport protein and breaks down the mechanism that removes excess cysteine leading to accumulation of cysteine which eventually forms crystals within the body cells. This prevents proper functioning of these cells leading to initial kidney problems before it progresses to other parts of the body including the liver, eyes, thyroid gland and an impaired growth. It is a rare inherited disease characterised by a wide spread deposition of amino acids and may be seen in every 200,000 live births in the developed countries and sufferers rarely survive into adulthood[1].

Methods

The Lysosomal diseases and their top affiliating gene were obtained from the MalaCard online system. MalaCard incorporates data from a wide variety of sources (approximately 90) such as HGNC (The Human Gene Nomenclature Committee), Ensemble and NCBI (National Center for Biotechnology Information) databases. It contains information on human proteins, genes, disease-related information, clinical and functional information[2]. The top affiliating gene for each of the lysosomal diseases identified by the MalaCard human malady compendium were entered into the STITCH database.

We accessed version 4.0 of the STITCH protein database through its portal site http://stitch.embl.de/. Each top affiliating gene for the selected lysosomal diseases (identified by the MalaCard human malady compendium) was entered into the STITCH database and the resultant network was saved as an Excel file. The Excel file was loaded into our statistical software (the R programming environment) package for deeper analysis.

There are many software tools available for data analysis, but since an element of programming is required we decided to use the R programming environment for the majority of our analysis work. The R language is available free from the comprehensive R archive network http://www.r-project.org/. We recommend that any user of R should also use the RStudio



STITCH: Chemical-Protein Interactions

name: examples: #1 #2 #3

(STITCH understands a variety of chemical/protein names, accessions and InChIKeys; you can also try a random entity)

organism: auto-detect

(in case of chemicals, the organism with the most interactions is chosen)

Reset GO!

please enter your protein or chemical of interest...

Interaction network around tryptophan

What is STITCH?

STITCH is a resource to explore known and predicted interactions of chemicals and proteins. Chemicals are linked to other chemicals and proteins by evidence derived from experiments, databases and the literature.

STITCH contains interactions for between 300,000 small molecules and 2.6 million proteins from 1133 organisms.

Mini Tutorial

To explore the interactions of the beta blocker propranolol, enter its name or an identifier like the ATC Code C07AA05 in the search box to the left. Select 'Homo Sapiens' as organism. When you click GO!, you will be taken to the network with various functionally related proteins: adrenergic receptors (the primary targets), serotonin receptors (which are also targeted), etc. Click on "Actions view" to explore the actions (activation/inhibition/binding) of propranolol. When you click on one of the adrenergic receptors, you can add it to the set of input nodes. This will result in a network containing both proteins and related chemicals.

Figure 4: The STITCH database web front-end

References

1 Anderson, R, Cairns, D, Coulthard, M, Terry, F (2012). Cystinosis and its treatment. The Pharmaceutical Journal, 269, 615-616.

Continued on page 70

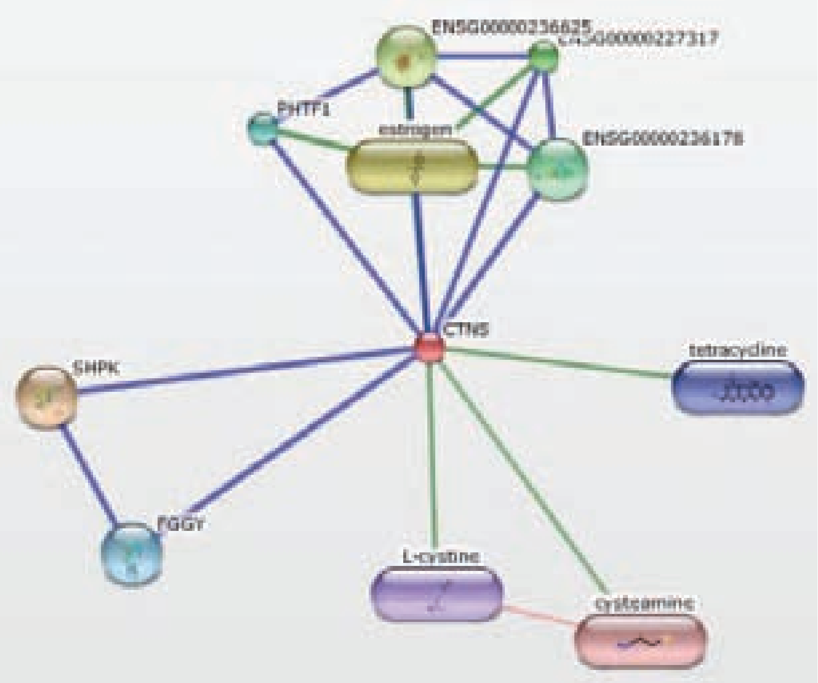


Figure 5: Top affiliating genes for cystinosis (CTNS) and interacting chemicals (oblongs) and proteins (circles), the colour and width of the connections indicate stronger associations represented by thicker lines. Protein-protein interactions are shown in blue, chemical-protein interaction in green and interactions between chemicals in red. STITCH also assigns a confidence score for each of these interacting pairs of proteins and chemicals

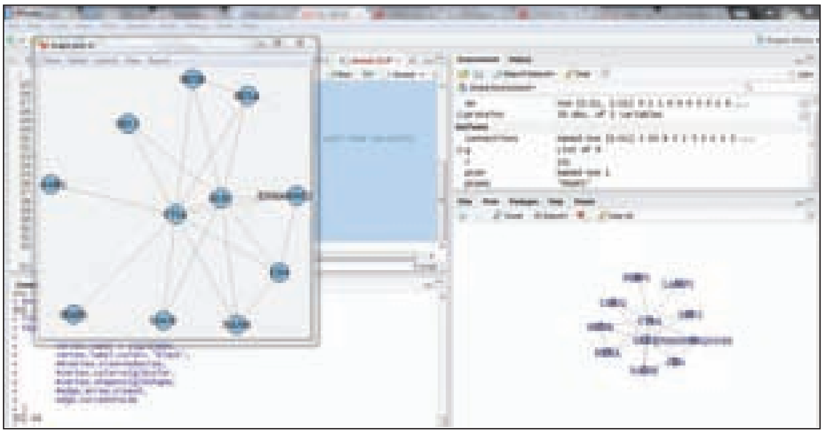


Figure 6: The RStudio development environment. The top left panel displays a drag and drop diagram of the protein connectivity patterns, the bottom left panel shows the R program code that created it

Continued from page 69

2 McGarry, K (2013). Discovery of functional protein groups by clustering community links and integration of ontological knowledge. Expert Systems with Applications, 40, 5101-5112.

Continued on page 71

development and working environment from https://www.rstudio.com/. The RStudio is designed to assist users to be more productive with their programming/analysis since it is a well-developed environment consisting of a set of integrated tools.

Each of the following measures can be defined mathematically, and by applying them to the network generated by STITCH we can determine the relevance and importance of each protein[3].

- Hubs. Non-essential disease genes which represent the majority of genes tend not to be hubs and segregate at the functional periphery of the interactome.
- Local. Proteins involved in the same disease have an increased tendency to interact with each other.
- Modularity. Many cellular components tend to be clustered together and form networks.
- Parsimony. Various pathways often coincide with the shortest molecular paths between known disease-associated pathways.
- Shared components. Genes that share cellular components show phenotypes and co-morbidity.

Thus a protein with high closeness, compared to the average closeness of the network, will be easily central to the regulation of other proteins but with some proteins not influenced by its activity. However, it is also useful to analyse proteins with low closeness, in contrast to the average closeness of the network, as these proteins, although less relevant for that specific network, are possibly behaving as intersecting boundaries with other protein networks (crosstalk between functional modules).

Since our aim was to identify potential therapeutic candidates for the treatment of Lysosomal diseases, we further analysed our data by matching disease gene symbols CTNS, CTSA and other MalaCards proteins, and hub proteins for Lysosomal diseases with drug banks to identify potential drug candidates. Drugs were connected to these proteins if the protein is a known target of the drug. In many cases some of these proteins did not match any drug in the drug bank, such as CTSA, but we did find it was affiliated to tetracycline from the STITCH database[5].

The analysis revealed some interesting new potential therapeutic candidates for the treatment of these diseases. However, this is only a bioinformatics approach based on topological analysis of the graphs, conclusive proof can only be provided by a clinical trial. Our study suggested Oseltamivir and Cysteamine as potential therapy for primary hyperoxaluria. Another candidate was Velaglucerase alpha as potential therapy for metabolic syndrome x. Furthermore, argifin, agadin and allosamizoline were identified as potential drug treatments for Gaucher, Fabry and Krabbe disease. Finally, tetracycline and lactacystine appeared potential therapy for cystinosis and galactosialidosis respectively.

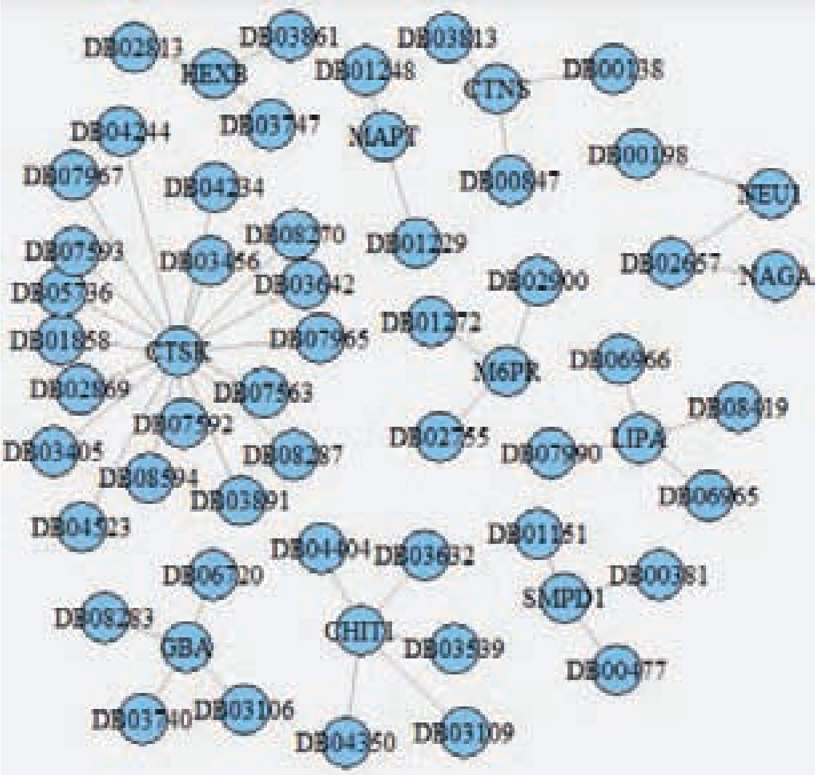


Figure 7: Bipartite graph of top affiliating genes of each lysosomal disease and their drug target. The DB prefixes refer to the DrugBank candidates identified as interacting with the key protein partners associated with the diseases

Table 3: Interacting drugs identified by DrugBank for a subset of the lysosomal diseases

DISEASE	PROTEIN	DRUGBANK IDENTIFIER	NAME OF DRUG/CHEMICAL
fucosidosis	NAGA	DB02657	Glucosamine 6-Phosphate
cystinosis	CTNS	DB00847	Cysteamine
cystinosis	CTNS	DB03813	2-Decenoyl N-Acetyl Cysteamine
cystinosis	CTNS	DB00138	L-Cystine
Primary hyperoxaluria	NEU1	DB00198	Oseltamivir
Primary hyperoxaluria	NEU1	DB02657	Glucosamine 6-Phosphate
Metabolic syndromex	GBA	DB03106	Myo-Inositol
Metabolic syndromex	GBA	DB03740	2-(Acetylamino)-2-Deoxy-a-D-Glucopyranose
Metabolic syndromex	GBA	DB06720	Velaglucerasealfa
Metabolic syndromex	GBA	DB08283	(2R,3R,4R,5S)-2-(HYDROXYMETHYL)-1-NONYLPIPERIDINE-3,4,5-TRIOL
Fabry disease	CHIT1	DB03109	N-Acetyl-D-Allosamine
Fabry disease	CHIT1	DB03539	2-(Acetylamino)-2-Deoxy-6-O-Methyl-Alpha-D-Allopyranose
Fabry disease	CHIT1	DB03632	Argifin
Fabry disease	CHIT1	DB04350	Argadin
Fabry disease	CHIT1	DB04404	Allosamizoline

The future

Recent advances in genomic and proteomic techniques have revolutionised our understanding of our genes, their function and how they co-operate with other genes in biological processes. One of the main discoveries is that several genes tend to operate together in modules that perform specific functions and that genes can belong to several modules. This information has now been captured and stored in databases that can be mined for interesting trends and associations. It is quite possible that hidden in these databases are existing drugs that may become future cures that will save the cost of discovery and development of a new treatment. The first stage will be to run a bioinformatics analysis on the proteomic/DrugBank databases which it is hoped will shorten the pathway to drug approval and maximise the potential of existing treatments. DDW

Dr Ken McGarry is a senior lecturer in statistics for health sciences at the University of Sunderland. He teaches statistics and research methods to UG and PG students and also advises staff and PhD students on statistical issues related to their research. This includes a wide range of subjects from biostatistics to economic data. He has 20 years' experience of data analysis. His own interests include the investigation of genomic and proteomic data for constructing bioinformatics models of protein interactions using a variety of computational techniques and has published more than 60 research papers.

Ukeme Daniel is a researcher at the University of Sunderland. Her interests lie in drug discovery and development and in particular drug repositioning. Previously she studied at the University of Uyo, Nigeria where she obtained her B.Pharm Degree and is a registered pharmacist in Nigeria.

Continued from page 70

3 Barabasi, A, Gulbahce, N, Losalzo, J (2011). Network Medicine: a network based approach to human disease. Nature Review Genetics, 12, 56-68.

4 Cline, M et al (2007). Integration of biological networks and gene expression data using Cytoscape, Nature Protocols 2, 2366 – 2382, doi:10.1038/nprot.2007.324.

5 Kuhn, M, Szklarczyk, D, Frankild, S, Blicher, T, Merging, C, Jensen, L, Bork, P (2013). STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Research, 42, D401-D407.