REVIEW

Reconstructing protein complexes: From proteomics to systems biology

J. Douglas Armstrong[1], *Andrew J. Pocklington*[1], *Mark A. Cumiskey*[2] *and Seth G. N. Grant*[2]

[1] School of Informatics, University of Edinburgh, Edinburgh, UK
[2] Wellcome Trust Sanger Institute, Hinxton, UK

Modern high throughput technologies in biological science often create lists of interesting molecules. The challenge is to reconstruct a descriptive model from these lists that reflects the underlying biological processes as accurately as possible. Once we have such a model or network, what can we learn from it? Specifically, given that we are interested in some biological process associated with the model, what new properties can we predict and subsequently test? Here, we describe, at an introductory level, a range of bioinformatics techniques that can be systematically applied to proteomic datasets. When combined, these methods give us a global overview of the network and the properties of the proteins and their interactions. These properties can then be used to predict functional pathways within the network and to examine substructure. To illustrate the application of these methods, we draw upon our own work concerning a complex of 186 proteins found in neuronal synapses in mammals. The techniques discussed are generally applicable and could be used to examine lists of proteins involved with the biological response to electric or magnetic fields.

Received: December 12, 2005
Revised: March 7, 2006
Accepted: March 8, 2006

Keywords:
Bio text mining / Membrane-associated guanylate kinase proteins (MAGUK)-associated signalling complex / Post-synaptic density / Schizophrenia / Systems biology

1 Membership of the network – proteins and nodes

Post-genomic biological methods, in particular high throughput proteomic methods can rapidly identify many tens to hundreds of molecules from biological specimens. The focus of this introductory review is to examine how we take such a list of proteins and test these data for functional relevance. Our own research has focussed on a set of proteins forming neurotransmitter receptor complexes in the post-synaptic densities of mammalian hippocampal neurons. The composition of this set was identified through immunoprecipitation with an N-methyl-D-asparate (NMDA) receptor subunit abundant in the post-synaptic density of mammalian neurons [1–4].

The NMDA receptor binds to membrane-associated guanylate kinase proteins (MAGUK) and forms of signalling complexes known as the NMDA receptor complex (NRC) or MAGUK-associated signalling complex (MASC) [1, 4]. NRC/MASC is located in the post-synaptic terminal of synapses and proteomic studies reveal it contains 186 proteins. These proteomic studies have been used as a starting point for bioinformatics studies, of which the approaches are described below, to construct a functional model of this complex in synaptic biology and disease.

No matter what technique is used to generate this list we need to recognise and acknowledge any inherent limitations [5] that may be present. Common limitations include sensi-

Correspondence: Dr. J. Douglas Armstrong, School of Informatics, University of Edinburgh, 5 Forrest Hill, Edinburgh, EH1 2QL, UK
E-mail: douglas.armstrong@ed.ac.uk
Fax: +44-7075-055-700

Abbreviation: MASC, membrane-associated guanylate kinase proteins (MAGUK)-associated signalling complex tivity – can the proteomic separation and identification techniques employed identify proteins present in very low amounts; cellular diversity – what is the likelihood that all the cells in the tissue sample contain a complex with the same protein complement; and contamination – what proteins might be present in the list as a result of methodological artefacts? However, proteomic methods are in a state of rapid development with improving sensitivity and accuracy. Further, the results of attempts to pull together multiple proteomic techniques for cross-validation of the protein component lists are encouraging [6, 7].

2 Functional annotation

Given such a list of proteins, one can actually learn a lot about the system before any attempt at mapping the individual proteins onto a functional or structural scaffold. One of the simplest initial analyses that gives a considerable amount of insight into the nature of a molecular complex and its possible functions is to look at the frequency of functional annotation within it. Here, a number of options exist. One of the richest sources is the Gene Ontology tool [8] that provides hierarchically structured information on molecular functions. However, in some situations, its use must be considered carefully as the depth and accuracy of annotation can vary widely with each molecule. Further, molecular function varies with cellular context and, for example, it has been reported that the naive use of Gene Ontology (GO; http://www.geneontology.org/) for classifying molecules in the nervous system can be misleading [9], as the annotations capture pleiotropic effects of protein functions in non-neuronal cells.

Examining the frequency of secondary sequence-level markers such as functional sites and structural motifs provides a simple and unbiased measure of enrichment for specific low-level molecular functions (*e.g.* kinase activity, calcium-binding domains). It is relatively simple to compare the frequency of any such tag to that expected within a random selection of proteins from the genome. There is a range of possible sources for these, but most of the commonly used annotations can be obtained directly through the InterPro database [10]. It is also useful to consider domains that are missing from the complex. For example, in our synapse proteome analysis (Table 1) we see strong enrichment for proteins containing domains linked to kinase activity and calcium binding, both prominent features of signalling pathways in the CNS [11]. Conversely, we see a paucity of domains more commonly associated with DNA-binding proteins, ribosomal subunits and proteolysis.

There are two possible reasons a specific molecular class (*e.g.* Interpro domain) to be significantly over represented in any biological sample. The first is that it reflects the true molecular composition and function of the sample. However, it may also represent a bias in the purification or detection method towards certain classes of molecule. Although such instances are often obvious, they are difficult to quantify to any degree of accuracy. For example, in our studies, we see enrichment for proteins containing PDZ domains and it is unclear whether this is a result of these domains being used as part of the purification process or a true enrichment within the complex of this class of molecule (in fact both are highly plausible in the complex we have studied).

Table 1. Protein features in MASC. Interpro domain frequencies in the MASC complex. Upper: the Interpro domains that are most highly enriched in the MASC complex compared to the entire mouse genome. Lower: domains most common within the mouse genome as a whole that are missing entirely from the MASC complex. Interpro ID are included for reference, see http://www.ebi.ac.uk/interpro/.

Domain	Accession	MASC	Genome
Enriched			
Protein kinase	IPR000719	11.8	3.75
Serine/threonine protein kinase	IPR002290	10.2	1.69
SH3	IPR001452	8.06	1.51
Pleckstrin-like	IPR001849	5.91	1.25
PDZ/DHR/GLGF	IPR001478	5.91	0.74
Small GTP-binding protein domain	IPR005225	5.38	1.49
Pleckstrin homology-type	IPR011993	4.84	1.08
Calcium-binding EF-hand	IPR002048	4.84	1.65
C2	IPR000008	4.84	0.82
IQ calmodulin-binding region	IPR000048	3.76	0.31
Missing			
Brix	IPR007109	0	10.5
Peptidase M13, neprilysin	IPR000718	0	7.85
FXYD	IPR000272	0	7.48
Malate dehydrogenase	IPR008267	0	5.65
20S proteasome, A and B subunits	IPR001353	0	4.91
Cystine knot	IPR006208	0	4.31
Protein tyrosine phosphatase, catalytic region	IPR003595	0	4.31
Zn-finger, C2H2 subtype	IPR007086	0	4.22
GABA A receptor, beta subunit	IPR002289	0	3.54
Uracil-DNA glycosylase	IPR002043	0	3.45

Perhaps the most difficult, yet in our case the most interesting type of analysis, is that of non-sequence-related biological annotation of molecules. General molecular ontologies such as the Gene Ontology (GO; http://www.geneontology.org/) do not capture the detailed domain knowledge (they were never designed to). In contrast, databases such as Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) catalogues and summarises the available evidence that relates genes to human diseases. Resources such as OMIM are not available for all areas of biology and when available are often incomplete in their coverage. For each of the 186 proteins in the MASC complex we examined on-line databases (such as OMIM) and also performed literature searches for evidence that a protein was involved in a range of biological processes including clinical disorders for human proteins and various aspects of rodent physiology and behaviour for the mouse orthologues.

3 Text mining for literature-based annotations

Querying the literature for functional information concerning the role (if any) of a small number of molecules in a single biological process is relatively painless. If we are lucky then even for a substantial list of proteins such information will be readily available from carefully curated public (*e.g.* OMIM) or commercial databases (*e.g.* TRANSPATH http://www.biobase.de). In most cases, such resources will not exist or be incomplete, requiring a substantial investment in manual searching of the scientific literature. This effort may be facilitated by text mining technologies. A complete review of text mining and the specific problems encountered in biological text mining is beyond the scope of this introduction but there are many dedicated reviews on these topics [12, 13].

On the surface, neither manual searching nor text mining approaches initially appear to be a significant problem. However, consider the case in detail: Proteins generally have more than one specific name to which they have been referred to in the literature (*i.e.* synonyms) and these lists of alternative names can be extracted from the gene and protein sequence annotation databases with relative ease. While the main protein names in the curated sequence databases are often unique (at least within the species) it is often the case that the other synonyms can hit multiple proteins entities within the same species, proteins in other species and non-protein entities entirely. The number of synonyms attached to a protein varies widely but can easily be more than ten. The BioMinT service (http://biomint.oefai.at/) provides a system to query synonyms across genes and proteins in 14 species [14]. A further complexity is often added by papers that deviate from using any community agreed nomenclatures such as EC numbers and introduce more subtle variations in spelling and formatting of names raising the number of potential synonyms to search several fold (known as query expansion).

For each search performed on a protein name or synonym, we need to add to the query a selection of keywords associated with the process we want to annotate. Finally, we also want to add in specific conditions to help rule out false positive hits. To improve speed and accuracy, abstracts are usually classified and indexed before searching. For example, all the abstracts including keywords such as *phosphorylation*, *phosphorylated* and *phosphorylating* are merged into a single classification group (*phosphorylation*). This group can then used to select evidence for phosphorylation of any specific protein on a pair-wise basis. Inferring biological information from the literature directly often results in false positive associations. Recent improvements in text mining technologies to include machine learning are making automated approaches even more accurate. However, to ensure data quality it is important that manual curation forms an integral part of the process.

For our own studies, we developed an in-house solution to these issues by linking together several publicly available packages. These were Lucene [15] for indexing and searching, Rainbow (McCallum A. K. 1996 http://www.cs.cmu.edu/~mccallum/bow) and weka [16] for text classification.

4 Statistical analysis of annotations

Having generated multiple sets of annotations covering sequence-based properties, phylogeny and various higher-level functions/phenotypes, a potentially rich source of information becomes available through their comparative analysis. The significance of any overlap between a pair of annotations may be evaluated by calculating its probability under a random distribution.

Suppose that out of a set of N molecules, n_a and n_b possess annotations a and b, respectively. If these annotations are distributed randomly within the full set, the probability $p(n_{ab})$ of n_{ab} possessing both is given by the function:

$$p(n_{ab}) = n_a!(N - n_a)!n_b!(N - n_b)!/[N!(n_a - n_{ab})!\, n_{ab}!(N - n_a - n_b + n_{ab})!(n_b - n_{ab})!] \quad (1)$$

This probability distribution has a single maximum $p_{ml} = p(n_{ml})$, with n_{ml} the most likely overlap occurring by chance (depending on symmetry, n_{ml} and $n_{ml} + 1$ may be equally likely). Given that μ_{ab} proteins possess both annotations, we can evaluate the significance of any deviation of μ_{ab} from n_{ml} by calculating the probability $P(\mu_{ab})$ of finding an overlap as or less likely under a random distribution [*i.e.* sum over all n for which p(n) is less than or equal to $p(\mu_{ab})$]:

$$P(\mu_{ab}) = \Sigma_n\, p(n) : p(n) \leq p(\mu_{ab}) \quad (2)$$

Under this definition, $P(n_{ml}) = 1$. Note that both tails of the distribution contribute to P, which may be used to evaluate deviations from n_{ml} in either direction. While it is possible to adjust P to account for the number of comparisons made, this does not necessarily improve the test. Sets of annotations are seldom independent, ranging from mutually exclusive terms (*e.g.* chromosomal location) to semi-redundant functional classifications. This is especially the case when investigating function at multiple levels, hierarchical ontologies (*e.g.* GO terms) providing a simple example. Such dependencies make it difficult to estimate the contribution of false positives to the number of $P(\mu_{ab})$ below a given significance threshold. On a more fundamental level, why should the significance of an overlap between protein function and psychiatric disorder depend on whether or not chromosomal location was investigated? A major contributing factor to these problems is the incompleteness and potentially uneven nature of much of the data being analysed, which means that many of the most interesting results may be of borderline significance. With these considerations in mind, it is perhaps better to use $P(\mu_{ab})$ and search for consistent patterns within the results, acknowledging the potential drawbacks.

5 Protein interactions – where to get them?

Reconstructing a model of a protein complex requires the information to assemble the individual components from the list obtained in the initial studies. For the static models we discuss here, we use basic graph theory to represent the complex as a network in which proteins are represented by nodes and interactions by edges [17].

For small networks of proteins, one could always screen each protein against each other using a biochemical interaction assay such as yeast-2-hybrid [18]. For many species, high throughput studies have been published using yeast-2-hybrid assays and the results indexed on databases for rapid retrieval. These include yeast, *C. elegans* and *Drosophila*. These high throughput techniques are starting to mature and with the latest methods, recently applied to the human proteome [19], there is now a greater degree of confidence in the resulting data.

The data from most of these high throughput studies, in addition to many other smaller focussed studies have been collated into a variety of public databases. The first step for collecting these data for most people will be one of the major on-line protein interaction databases such as BIND (http://www.bind.ca/Action), IntAct (http://www.ebi.ac.uk/intact), MINT (http://mint.bio.uniroma2.it/mint/), and DIP (http://dip.doe-mbi.ucla.edu/).

The key to accessing the information in these databases is a suitable list of protein ID (each database tends to use different ID). Tools now exist on the web that are good for converting between the various ID but in many cases it will need to be done by hand. Ariadne Genomics provides an on-line service that allows users to paste in a list of ID in one format and convert to any one of a range of others (http://www.ariadnegenomics.com/services/idmap.html).

For more detailed information on a protein-by-protein basis, the EnsMart tool at EBI provides a unified query interface to a number of databases and their ID (http://www.ensembl.org/Multi/martview).

In many cases, the protein interaction information from closely related species (interlogs) may be of interest. This can either be directly included or used to add extra confidence to interactions with less reliable sources. There are several databases that list cross-species orthologue maps, normally based at the gene level. The ones we commonly use are the Ensemble genome database [20] at http://www.ensembl.org/index.html and InParanoid [21], which also has software for building custom maps available for download (http://inparanoid.cgb.ki.se/index.html).

An important source for known protein-protein interactions is the scientific literature. Searching on combinations of synonyms and keywords can be used to rapidly extract abstracts from public literature databases such as PubMed (discussed above). A number of databases now exist that collate information extracted from abstracts using text-mining techniques alone or through a combination of text mining and subsequence manual curation. For all of these a careful check of the provenance of the data is advised: where did the information come from; what queries were used to define the proteins and the interaction. Do these queries make sense in terms of the accuracy of the protein nomenclature (*i.e.* does the database refer to the molecule you think it does), the species and any other constraints you may have on the 'quality' of the interaction (*e.g.* database entries may refer to indirect associations between proteins in large complexes using text that is very similar to that describing direct biochemical binding interactions).

We have used two literature-mined databases. The first of these, iHOP is freely available on-line [22] at http://www.ihop-net.org. Given a synonym, it will look up alternative names and provide a list of abstracts that might refer to potential interactions with other molecules in any species. It provides a brief summary of the provenance of each entry, highlighting the fragment of the abstract used to classify its inclusion in the report for that molecule and a link to PubMed or another data source for further details. This feature allows the rapid screening out of obvious false leads but then requires systematic curation of the remaining results by the end user. We have also looked at a commercial database NetPro (from www.molecularconnections.com) that has similar features but has additionally been subject to manual curation by human experts. Each NetPro entry has associated unique ID tags and some detail on the type of interaction and the type of experimental evidence available to support the interaction.

In our own studies, we have augmented all of the above methods by implementing our own text mining tools to check for further interactions missed by the on-line tools and databases. An expert then assessed evidence gathered from all sources for inclusion in the model, often referring at this stage to the full text of scientific publications. Finally, a second independent expert assessed each interaction. The research group then discussed any discrepancies between two rounds of assessment (around 1% of interactions).

6 Network analysis

Having gone to the effort of constructing a new model for a protein complex, the first challenge is to visualise it in some way. There are a number of graph visualisation tools that have 2-D and 3-D layout algorithms directly embedded within them. The two that we use most often are Pajek [23], an advanced package unfortunately limited to the Windows operating system and BioLayout (*e.g.* Fig. 1) [24], a JAVA-based freeware package that is platform independent. Most of the public databases have applet-based tools that have basic layout and presentation functionality but are usually augmented with links from the images back to the database, allowing the network to be explored. There is a clear need from the community for a package that has the layout and analysis functionality of Pajek, that is platform independent and can be linked to databases to support point and click queries of the underlying information (for example, a click on an edge or interaction could pull up the supporting evidence for the interaction). A number of projects are underway in this area, with Cytoscape [25] currently the most advanced.

Before proceeding with the analysis of an interaction network, it is worth evaluating the extent to which it captures the various functional processes being investigated. It is highly unlikely that interaction data will be available for all

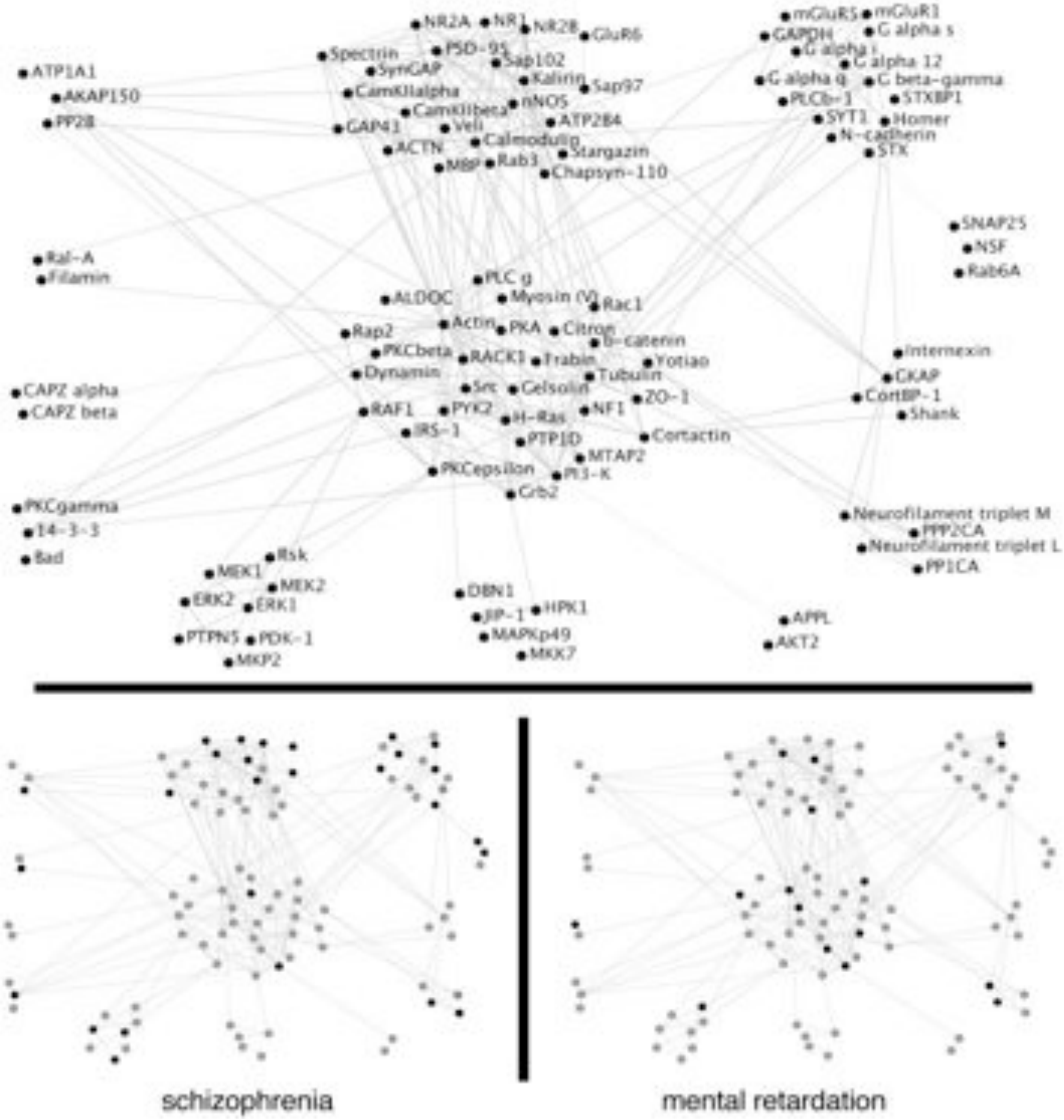

Figure 1. The connected proteins of the MASC complex. Schematic diagrams of the MASC complex based on protein-protein interactions. Upper: The MASC protein interaction network with common protein names as described in [11]. The lower panels show the distribution of functional annotation. Lower left: the dark grey proteins are those involved in schizophrenia and statistical analysis suggests they are more tightly associated with the two clusters are the top of the network. Lower right: proteins linked to an involvement in mental retardation proteins are evenly distributed across the network.

molecules of interest and such data as there is may be subject to bias. In addition, methods of analysis tend to focus on the properties of single connected sets of molecules, of which there may be several. The composition of a connected network component may be analysed using the statistical approach described earlier. Treating membership of the component as a new annotation, this can be used to evaluate the significance of any overlap between the component and other functional/phenotypic annotations. A component does not have to be an unbiased representation of the full set of molecules to be of interest – in our investigation of synaptic protein complexes, we found a single, large component enriched for glutamate/calcium signal transduction and containing the majority of all proteins with known electrophysiological, behavioural and disease phenotypes. The component thus captured key aspects of network functionality, primarily those associated with signal reception and integration [11].

Over the past 5 years, there has been an intense interest in the analysis and modelling of biological networks and their mathematical properties. The renewed interest in this field was largely a result of work on metabolic network architecture [27], which led to the hypothesis that most, if not all, biological networks follow a similar set of organisational rules (evident in the widely observed power-law connectivity distribution) and can be considered 'scale-free'. There is a great deal of debate and on-going research into the degree to which different networks are truly scale-free, modular, hierarchical, bow-tie or some combination of these various network architectures. The common feature that has emerged is that connectivity in biological networks almost without exception follows an approximate power-law distribution where the probability of finding a node with *n* connections varies with *n* to the power *k*, where *k* is a constant for that network. Put simply, nodes or proteins with a small number of connections are extremely common whereas nodes or proteins with large numbers of connections occur but are rare. Another common feature is a 'small-world' property that can result from a power-law distribution of connections. In small-world networks, there is generally a relatively short path (via one of the highly connected hub nodes/proteins) between any pair of molecules in the complex. In signalling pathways these features have two main effects, they result in a generally robust network architecture and at the same time introduce cross talk between the pathways embedded in the complex. This makes understanding or predicting the effect of a molecular or genetic disruption really quite difficult.

7 Functional prediction based on architecture

How does the architecture fit with the functional annotation? The first such analysis was performed in yeast [17] where a correlation between vertex degree (number of interactions per protein) and the viability of single gene knockouts was observed. The so-called hub proteins (the rare, highly connected proteins) were statistically more likely to be lethal when mutated than the more common non-hub proteins (with just a small number of connections). Several studies since have taken this forward using more complex measures such as the characteristic path length (sometimes called diameter). The characteristic path length is a measure of how closely interconnected a network is, being the average of the shortest path length between all pairs of nodes. Removing a highly connected node tends to increase network diameter, while removing a node with few connections tends to have minimal effects. However, if these connections are important in the architecture the diameter can be affected more severely.

Assuming the complex has a uniform functional distribution (*i.e.* that the entire network is equally likely to be involved in the process of interest) then these measurements tend to correlate well with functional annotation. In most cases, we are limited to qualitative data where we know a set of molecules that are involved in a process or disease and we can correlate the likelihood of annotation with a network parameter such as vertex degree or network diameter. There are several drawbacks in these approaches. First, the functional annotation is often very incomplete and negative examples are often missing (*i.e.* that protein X is not involved). Secondly, where functional annotation and network connectivity are literature based there is an obvious potential for bias – proteins that more people work on are more likely to have more connections and more likely to have been linked to the process or disease than those that few people are interested in.

For our work on the synaptic protein complex, this was a concern, as both network information and functional annotation were literature based. However, quantitative information was available in the LTP literature where single gene knockouts were assessed quantitatively for physiological function. Here again, we found significant correlation between network properties and the magnitude of the phenotype.

8 Network substructure

The elegant work done on the yeast proteome worked so well because a global function (viability) was used as a measure [17]. Within such a large complex, sub-structure does exist with different regions having distinct functions. In the context of molecular interaction networks, sub-structure refers to the existence of molecular clusters characterised by a high density of intra-cluster interactions and much sparser connectivity with the rest of the network. A large amount of research has gone into methods and analysis of clustered networks, particularly in metabolic networks where processes are directional. Clustering of proteomic networks can be done using a range of methods each with advantages and disadvantages. As all methods are heuristics for identifying the clustering, which best reflects interaction data, it is important to evaluate any results. The modularity score [26] provides an objective measure of how well a given clustering reflects network structure and may be used to compare alternatives generated by the same or different algorithms.

An exhaustive review of clustering methods is beyond the scope of this article, as numerous methods exist and new ones are constantly being developed. The performance of each algorithm typically depends on the type of data being analysed. Two reviews [27] (Berkhin, P., *Technical report, Accrue Software, San Jose, California*, 2002. http://citeseer.comp.nus.edu.sg/berkhin02survey.html) discuss and compare a large number of clustering algorithms, including those most commonly used. In addition, we would note the recent use of an information-based clustering algorithm [28] in the study of modularity in genetic networks [29]. In our own work, a divisive clustering algorithm [24] has proven useful. Simply, this algorithm searches for the single edge/interaction within the network that occurs most frequently on all possible paths between vertices. This edge is then removed from the network and the process restarted. As the network fragments into distinct clusters, the protein-to-cluster assignments are recorded and the modularity score used to identify which level of the resulting hierarchy best reflects network structure (see Fig. 1).

Once we have a clustered network we can look at the distribution of annotations over clusters using the statistical methods described earlier. In the MASC network, for example, we observe a cluster that is significantly enriched for association with schizophrenia, ionotropic glutamate receptors, synaptic plasticity and PDZ domain containing proteins when compared to the rest of the network. On the other hand, proteins associated with mental retardation are evenly distributed throughout the complex (Fig. 1). These observations suggest that the complex has general functional properties in neuronal function where perturbation results in generalised mental retardation (in a manner similar to the yeast proteome and viability) but that sub-structures are associated with more specific functionality (*e.g.* schizophrenia).

9 Concluding remarks

We have attempted to provide an introductory walkthrough of the methods and resources available for reconstructing a protein complex model given a list of molecules hot off a mass-spec machine. Limitations with methods and sources of error must always be considered carefully as it is all too easy to construct a model on extremely shaky foundations. The tools and methods described here are developing extremely rapidly and we expect that the databases in particular will increase in coverage and accuracy at an extremely fast rate.

At present, expertly curated datasets provide the highest quality information for model construction. However, curation is very expensive and the literature is biased towards a small subset of molecules. High throughput technologies are becoming more reliable and this will enable more complete models of protein complexes to be constructed. The first studies using such static models of protein complexes are predicting new molecules to be involved in various functional processes. A lot of further research is required to test these predictions and to refine and develop methods of analysis.

We have discussed the available methods in the context of our own studies that focus on the proteins in the mammalian post-synaptic density. However, these methods are equally applicable to other biological fields where samples containing protein complexes are being characterised.

Douglas Armstrong studied genetics with Kim Kaiser in Glasgow before moving to Rice where he worked with Kathleen Beckingham on the genetic analysis of gravitaxic behaviour. In both positions, JDA developed a suite of bioinformatics tools for structure/function analysis in the nervous system. He started his own group in 2001 at Edinburgh in the School of Informatics and Centre for Integrative Physiology. His group works in the area of Systems Biology of Cognition with both informatics and wet-lab research programmes.

JDA, MAC and SGNG were supported by the Wellcome Trust (UK) Genes to Cognition programme. AJP was supported by the Medical Research Council (UK) through a Special Research Training Fellowship in Bioinformatics.

10 References

[1] Husi, H., Ward, M. A., Choudhary, J. S., Blackstock, W. P., Grant, S. G., *Nat. Neurosci.* 2000, *3*, 661–669.

[2] Husi, H., Grant, S. G., *J. Neurochem.* 2001, *77*, 281–291.

[3] Farr, C. D., Gafken, P. R., Norbeck, A. D., Doneanu, C. E. *et al.*, *J. Neurochem.* 2004, *91*, 438–450.

[4] Collins, M. O., Husi, H., Yu, L., Brandon, J. M. *et al.*, *J. Neurochem.* 2006, *97*, Suppl 1: 16–23.

[5] Yates, J. R. 3rd., Gilchrist, A., Howell, K. E., Bergeron, J. J., *Nat. Rev. Mol. Cell. Biol.* 2005, *6*, 702–714.

[6] Gavin, A-C., Bösche, M., Krause, R., Grandi, P. *et al.*, *Nature* 2002, *415*, 141–147.

[7] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D. *et al.*, *Nature* 2002, *415*, 180–183.

[8] Gene Ontology Consortium, *Nat. Genet.* 2000, *25*, 25–29.

[9] Inlow, J. K., Restifo, L. L., *Genetics* 2004, *166*, 835–881.

[10] Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A. *et al.*, *Nucleic Acids Res.* 2005, *33*, D201–205.

[11] Pocklington, A. J., Cumiskey, M. A., Armstrong, J. D., Grant, S. G. N., *Mol. Systems Biol.* 2006, *2*, msb4100041-E1-E14.

[12] Hearst, M., *Proc. ACL* 1999, 3–10.

[13] Hirschman, L,. Park, J. C., Tsujii, J., Wong, L., Wu, C. H., *Bioinformatics* 2002, *18*, 1553–1561.

[14] Pillet, V., Zehnder, M., Seewald, A. K., Veuthey, A. L., Petrak, J., *Bioinformatics* 2005, *21*, 1743–1744.

[15] Hatcher, E., Gospodnetic, O., *Lucene In Action Manning* 2004.

[16] Witten, H., Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Edn., Morgan Kaufmann, San Francisco 2005.

[17] Jeong, H., Mason, S. P., Barabasi, A. L., Oltvai, Z. N., *Nature* 2001, *411*, 41–42.

[18] Vidal, M., *FEBS Lett.* 2005, *579*, 1834–1838.

[19] Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T. *et al.*, *Nature* 2005 *437*, 1173–1178.

[20] Hubbard, T., Andrews, D., Caccamo, M., Cameron, G. *et al.*, *Nucleic Acids Res.* 2005, *33*, D447–D453.

[21] Remm, M., Storm, C. E. V., Sonnhammer, E. L. L., *JMB* 2001, *314*, 1041–1052.

[22] Hoffmann, R., Valencia, A., *Nat. Genet.* 2004, *36*, 664.

[23] Batagelj, V., Mrvar, A., *Graph Drawing Software*. Springer, Berlin. 2003, pp. 77–103.

[24] Goldovsky, L., Cases, I., Enright, A. J., Ouzounis, C. A., *Appl. Bioinformatics* 2005, *4*, 71–74.

[25] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S. *et al.*, *Genome Res.* 2003, *13*, 2498–2504.

[26] Newman, M. E. J., Girvan, M., *Phys. Rev. E.* 2004, *69*, 026113.

[27] Shamir, R., Sharan, R., in: Jiang, T., Smith, T., Xu, Y., Zhang, M. Q., (Eds.), *Current Topics in Computational Biology*. MIT press, Cambridge, MA, USA 2001.

[28] Slonim, N., Atwal, G. S., Tkacik, G., Bialek, W., *Pro.c Natl. Acad. Sci. USA* 2005, *102*, 18297–18302.

[29] Slonim, N., Elemento, O., Tavazoie, S., *Mol. Systems Biol.* 2006, 2, doi:10.1038/msb4100047.