



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Reconstructing and Analysing Protein-Protein Interaction Networks of Synaptic Molecular Machines

Lysimachos Zografos



Doctor of Philosophy

Institute for Adaptive and Neural Computation

School of Informatics

University of Edinburgh

2012

Abstract

The postsynaptic density (PSD) is a complex, dynamic structure composed of ~2000 distinct proteins, found at the postsynaptic membrane. Interactions, of transient and non-transient nature, organise the PSD's constituent parts into a protein complex, which functions as an intricately regulated molecular machine, orchestrating the mediation and regulation of synaptic transmission and synaptic plasticity. Furthermore, many of the proteins found in this complex have been linked to synaptic and behavioural plasticity, basic cognition or disease. Although, through proteomics we have accumulated a lot of information on the constituent parts of this machine as well smaller sub-networks representing pathways, not a lot is known about the organisational principles of the PSD. In this project our aim is to develop a standardised approach to reconstructing protein interaction networks from PSD proteomics data, providing a descriptive integrative model. Using these models we also performed an analysis elucidating parts of these organisational principles. We applied this method on two murine postsynaptic density proteomics datasets and found a persistent modular architecture of biological significance. Furthermore, given the lack of substantial evidence on the composition and architecture of postsynaptic density interaction networks of other model organisms, we decided to perform an affinity purification of *Drosophila melanogaster* postsynaptic density proteins and perform a similar analysis. The resulting model corroborated theoretical predictions of a lower complexity but similar functionality and also showed a modular architecture. As a final analysis we compared the two models from a structural and evolutionary perspective attempting to elucidate the mechanisms of evolution of this molecular machine. The results of this analysis implied that a whole component rather than just individual proteins of the fly protein interaction network have been conserved, highlighting the importance of the aforementioned organisational principles.

Acknowledgements

This work was performed under the invaluable supervision and guidance of Prof. J Douglas Armstrong and Dr. Andrew Pocklington, without whom this project would have been impossible. Prof. Vincent Danos, Dr. Matthias Hennig, Bilal Malik and Dr. Joanna Rees contributed with their knowledge, technical guidance and insightful comments. I would also like to thank Prof. Seth Grant, Dr. Mike Croning and Dr. Esperanza Fernández of the Genes2Cognition consortium for a successful collaboration and their feedback with parts of this work. Additionally, a big thank you must go to all members of the Armstrong, Dr. Giusy Pennetta's and Prof. Andrew Jarman's labs: Dr. Giusy Pennetta, Prof. Andrew Jarman, Dr. Seymour Knowles-Barley, Dr. Joanna Young, Dr. Lynn Powell, Dr. Petra Zu Lage, Giuseppe Gallone, and Daniel Moor as well as all friends and colleagues who helped with the editing of this thesis. This project was funded through the EPSRC Doctorate Training Centres (DTC) programme (EP/D505984/1) and the research took place in the University of Edinburgh's Institute for Adaptive and Neural Computation Neuroinformatics DTC.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Lysimachos Zografos)

This work is dedicated to my family, friends, and partner. It was their love and support that made it possible.

“The model is not an oracle, it’s just an automation of our understanding”

J. Heath

Contents

1	Introduction and Background	1
1.1	The synapse and the postsynaptic density	1
1.2	The excitatory synapse	2
1.2.1	The architecture of an excitatory synapse	3
1.2.2	The molecular basis of cognition and disease	6
1.3	Systems neurobiology of the synapse	13
1.3.1	Neurobiology and the Omics	13
1.3.2	Proteomics of the synapse	14
1.3.3	Models of the PSD	18
1.3.4	Evolution of the PSD	26
1.3.5	Neuroproteomics informatics	31
1.4	Motivation, hypothesis and goals	32
1.4.1	Motivation	32
1.4.2	Hypothesis	33
1.4.3	Goals	34
2	Materials and Methods	35
2.1	<i>Drosophila</i> strains	35
2.1.1	Handling	35
2.1.2	Strains	36
2.2	Methods	39

2.2.1	Affinity purification of complexes	39
2.2.2	Mass spectrometry	43
3	Computational Methods	45
3.1	Background	45
3.2	Annotation methods	46
3.2.1	Data annotation	46
3.2.2	Interaction annotation	55
3.2.3	Mining data from the literature	59
3.3	Analysis methods	65
3.3.1	Biological network primer	65
3.3.2	Community structure in PPINs	68
3.3.3	Network topology features	74
3.3.4	Statistical significance	76
3.3.5	Visualisation	79
3.4	Comparative methods	81
3.4.1	Comparing models of PSDs	81
3.4.2	Implementation	82
3.5	Supplementary material	83
3.6	Concluding remarks	83
4	The PSD-95 associated proteins complex	89
4.1	Background	89
4.2	Genetics and proteomics	91
4.2.1	Construct design	91
4.2.2	<i>PSD – 95^{TAP}</i> mice phenotyping	91
4.2.3	Protocol	94
4.3	Results	94
4.3.1	The PSD-95 associated proteins complex	94

4.3.2	The PSD-95 associated proteins interaction network	98
4.4	Concluding remarks	105
5	The PSD interactome	111
5.1	Background	111
5.2	Integration and data mining	113
5.2.1	Dataset merging	113
5.2.2	Interaction mining	115
5.3	Results	116
5.3.1	The Union protein complex	116
5.3.2	The Union protein interaction network model	118
5.3.3	The PSD interactome and physiology	130
5.3.4	The PSD interactome and disease	130
5.3.5	Evolution of the PSD interactome	137
5.4	Concluding remarks	142
5.4.1	Issues	142
5.4.2	Biological significance of clustering	143
5.4.3	A core PSD protein interaction network dataset and model . .	145
5.4.4	Evolution of the PSD interactome	148
6	The fPSD interactome	151
6.1	Background	151
6.2	Results	153
6.2.1	Isolation of fPSD complexes	153
6.2.2	The fPSD proteomic catalogue	158
6.2.3	An integrated protein interaction network of the fPSD	166
6.2.4	Comparison of bait complexes	179
6.2.5	Mapping the fPSD to human disease	186
6.2.6	Evolution of the fPSD	187

6.3	Concluding remarks	189
7	Comparative analysis of PSD complexes and interaction networks	195
7.1	Background	195
7.2	Results	197
7.2.1	Homology	197
7.2.2	Families	198
7.2.3	Protein domains	201
7.2.4	GO annotation	202
7.2.5	Comparative interactomics	204
7.3	Concluding remarks	214
8	Discussion	219
8.1	General discussion	219
8.1.1	A broader view	219
8.1.2	Limitations	222
8.1.3	Critical examination	225
8.2	Future work	226
8.3	Conclusions	230
A	Supplemental Methods	235
A.1	fPSD complexes	235
A.1.1	Validation of affinity purifications	235
A.1.2	Mass spectrometry data filtering	235
B	Data tables	243
	Bibliography	255

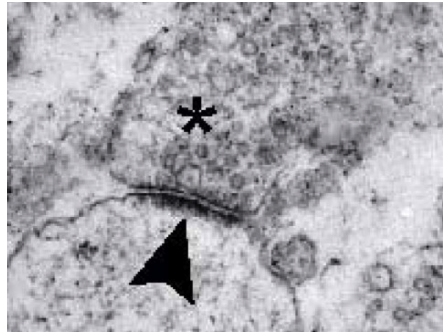
Chapter 1

Introduction and Background

1.1 The synapse and the postsynaptic density

Information in the nervous system is transmitted in patterns of action potentials - electrical pulses generated in neurons, and transmitted from one to another at specialised junctions known as synapses. At chemical synapses, the most abundant type in the nervous system, action potentials propagating through the presynaptic neuron are converted into release of a neurotransmitter, such as glutamate. This diffuses across the synaptic cleft and binds to receptors on the postsynaptic cell, resulting in transient local depolarization of the cell membrane. When the postsynaptic neuron becomes sufficiently depolarised, due to input from one or more synapses, a new action potential is generated. Synaptic input is also processed by the postsynaptic signalling machinery, which is closely linked to the intracellular side of the post-synaptic membrane in a structure known as the postsynaptic density (PSD) (Figure 1.1). The PSD is a complex, dynamic structure composed of ~2000 distinct proteins (Bai and Witzmann, 2007, Choudhary and Grant, 2004, Li et al., 2004, Collins et al., 2006, Emes et al., 2008, Li and Jimenez, 2008, Trinidad et al., 2008, Fernández et al., 2009, Croning et al., 2009), of which ~100 are thought to be present at an individual synapse (Sheng and Hoogenraad, 2007, Selimi et al., 2009). Physical interactions organise these pro-

Figure 1.1: The PSD, visible as an electron dense area in an electron microscopy scan. Synapses of mice neurons exhibit presynaptic vesicles (asterisks), a synaptic cleft and a distinct postsynaptic density (arrowheads). Figure from Heupel et al. (2008)



teins into signalling pathways that coordinate changes in synaptic strength in response to patterns of neuronal activity. These changes in synaptic strength can alter the flow of activity in neuronal networks and are widely thought to form the basis of behavioural learning and memory.

1.2 The excitatory synapse

The major classification of synapses reflects their function, as excitatory or inhibitory, depending on the type of primary neurotransmitters used. Glutamate is an example of excitatory neurotransmitter while GABA and serotonin are examples of inhibitory neurotransmitters. The studies performed in this work are centered around glutamate excitatory synapses.

Glutamate synapses have been implicated in various processes including neuronal development, neurotoxicity, and synaptic plasticity to name but a few. Both types of glutamate receptors, namely ionotropic (ion channel coupled) and metabotropic (second messenger coupled) receptors are differentially distributed on pre- and postsynaptic sites to contribute to action potential propagation and neuronal signal processing, functions that determine learning and memory formation (Bliss and Collingridge, 1993, Bear and Abraham, 1996, Riedel et al., 1996; 2003). Furthermore, glutamate receptors and the proteins they interact with have been implicated in various men-

tal diseases such as schizophrenia or various other forms of neurodegeneration such as Alzheimer's disease (Ellison, 1995, Pellicciari and Costantino, 1999, Millar et al., 2000, Jamain et al., 2003, Harrison and Weinberger, 2005, Redon et al., 2006, Jamain et al., 2008, Purcell et al., 2009, Pinto et al., 2010) (discussed in detail in subsection 1.2.2.3).

These glutamate receptor subtypes, as defined by the constituting subunits are involved in two types of transmission, fast and slow. Fast transmission is mediated by ionotropic receptor subtypes, namely N-Methyl-D-aspartic acid (NMDA), alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) and Kainate glutamate receptors. Ionotropic receptors are tetramer ion channels, composed of class-specific subunits that can form homo- or heteromers. Each subunit has an extracellular N- and intracellular C- terminus. Slow transmission is mediated by eight metabotropic glutamate receptor subtypes (mGluRs). Structurally, mGluRs have 7 transmembrane segments, but show sequence similarities with metabotropic GABA receptors (Kaupmann et al., 1997). For further details on the structure and function of glutamate receptor subtypes see Riedel et al., 2003. Although AMPAs are the basic synaptic transmission receptors, allowing the influx of cations once the agonist neurotransmitter has bound on the extracellular side, NMDA receptor function has been described as that of a coincidence detector (Bliss and Collingridge, 1993, Bourne and Nicoll, 1993, Stevens and Sullivan, 1998, Tang et al., 1999). More specifically NMDA receptors, sensing positive changes in the membrane potential caused by an initial depolarisation, are likely to expel the Mg^{2+} ion that blocks the channel from the outside, leading to further compartmentalised, thus input specific, Ca^{2+} influx, a property crucial to plasticity (Nicoll et al., 1988) as well as learning and memory processes (Tang et al., 1999, Tsien, 2000).

1.2.1 The architecture of an excitatory synapse

The function of molecular complexes on the presynaptic terminal is to mediate the release of glutamate from synaptic vesicles, in a four step process involving 1) vesicle

formation, 2) docking, 3) priming, and 4) fusion. Pre- and postsynaptic terminals are separated by the synaptic cleft, but held together by cell-adhesion molecules (CAMs) of the immunoglobulin, neuroligin, neuroligin, ephrin, ephrin receptor, synaptic cell-adhesion molecules (SynCAMs), and cadherin families (Dalva et al., 2007, Fogel et al., 2007; 2010, Rebsam and Mason, 2011), which provide this trans-synaptic linkage.

The PSD element of the synapse typically resides in mushroom like dendritic protrusions (dendritic spines) (Tada and Sheng, 2006, Sheng and Hoogenraad, 2007, Newpher and Ehlers, 2008). Dendritic spine morphology has been found to be associated with synaptic plasticity (Yuste and Bonhoeffer, 2001). An excellent recent review of the architecture of the excitatory synapse (Chua et al., 2010) classifies high-level PSD functions into three main purposes: 1) to cluster glutamate receptors and CAMs, 2) to recruit signalling proteins, and 3) to anchor these components to the microfilament cytoskeletal structures of the spine. The latter is achieved by an array of PSD proteins that create filamentous or lattice-like scaffolds and connect the receptor and signalling component to the cytoskeleton. A main example of this category of proteins are membrane-associated guanylate kinase (MAGUK) superfamily member families such as the protein products of the Dlg, Dlgap and Shank genes families. These proteins are organised in filaments, e.g. PSD-95 of the Dlg family (Chen et al., 2008) or lattice like structures, e.g. proteins of the Shank and Homer families (Hayashi et al., 2009), underlying the structural stability of the PSD. These proteins also serve by interacting directly or indirectly with receptors and signalling molecules. Shank proteins interact via Dlgap with PSD-95, which in turn interacts with NMDA receptors, K^+ channels, neuroligins and indirectly AMPA receptors via Stargazin interactions (Schnell et al., 2002, Kim and Sheng, 2004, Schoch and Gundelfinger, 2006, Newpher and Ehlers, 2008, Sturgill et al., 2009). The Shank family of proteins also binds this scaffold to the cytoskeleton via direct or indirect (e.g. with densin-180) interactions with actin binding proteins such as a-fodrin, Abp1, and α -actinin (Walikonis et al., 2001, Quitsch et al., 2005, Schoch and Gundelfinger, 2006, Kreienkamp, 2008).

The influx of Ca^{2+} after glutamate receptor activation activates calmodulin (CaM), which in turn activates a series of different signalling enzymes with downstream effect, acting as a signal integrator (Xia and Storm, 2005). These downstream proteins are recruited into PSD complexes by associating, directly or indirectly, with scaffolding proteins like PSD-95 and Shank and by extension coupling with the receptors and channels the scaffolding proteins interact with. Examples of indirectly associating proteins are adenylate cyclase isoforms and phosphodiesterase 1 (Pde1), all of which interact with PSD-95 interactor, Akap79 (Gorski et al., 2005, Efendiev et al., 2010), Ras GTPase interacting with PSD-95 via SynGAP (Kim et al., 1998) and Plc β interacting with Shank2 (Hwang et al., 2005). Furthermore, there are examples of direct PSD-95 associations such as the one between Src kinases family members (Lyn, Src, Yes, Fyn) (Tezuka et al., 1999, Kalia and Salter, 2003) and Rho-GEF family member karilin-7 (Penzes et al., 2001, Ma et al., 2008). It is worth noting that the signalling proteins not only interact with scaffolding molecules but in many cases modulate their interactions with other proteins, e.g. the CamK2 dependent phosphorylation of PSD-95, as reported by Gardoni et al. (2006). Also, there are direct interactions of signalling enzymes with receptor subunits, such as the interactions of NMDA receptor second subunit with PI-3 (Perkinton et al., 2002), CamK2 (Barria et al., 1997) and RasGRF1, which also connects it to the extracellular signal regulated kinase (Erk) / Mitogen activated kinase (Mapk) pathway (Krapivinsky et al., 2003).

One of the basic protein interaction domains found in scaffolding proteins discussed in this work is the PDZ domain (reviewed in Nourry et al., 2003, Kim and Sheng, 2004, Feng and Zhang, 2009). One could argue that PDZ mediated interactions are as much of a basic component for the synapse as the neurotransmitter receptors, which they also modulate (Chung et al., 2004, Iwamoto et al., 2004). PDZ domains can occur in one or multiple copies and are nearly always found in cytoplasmic proteins (Nourry et al., 2003). The MAGUK superfamily proteins contain PDZ domains (one or three), one SH3 domain, and a guanylate kinase domain (GuK), comprise a

characteristic PSD PDZ-containing group of proteins and have many representatives studied in Chapters 4, 5 and 6.

PSD-95, a protein central to the models of the aforementioned Chapters and representative of the MAGUKs, forms multimers in a head-to-head manner (Hsueh and Sheng, 1999), possibly allowing further clustering of its partners in large molecular complexes. There are many examples of such clustering involving key PSD proteins like K^+ channels (Kim et al., 2007) and NMDA receptors (Cho et al., 1992, Craven and Brecht, 1998). Except the organisation of such clusters PSD-95 also interacts with neuroligin, a postsynaptic membrane protein that interacts trans-synaptically with neuroligins (Irie et al., 1997), involving the PSD-95-based scaffold in synaptic adhesion as well. An illustration of the molecular organisation of the PSD around PDZ-containing proteins can be seen in Figure 1.2.

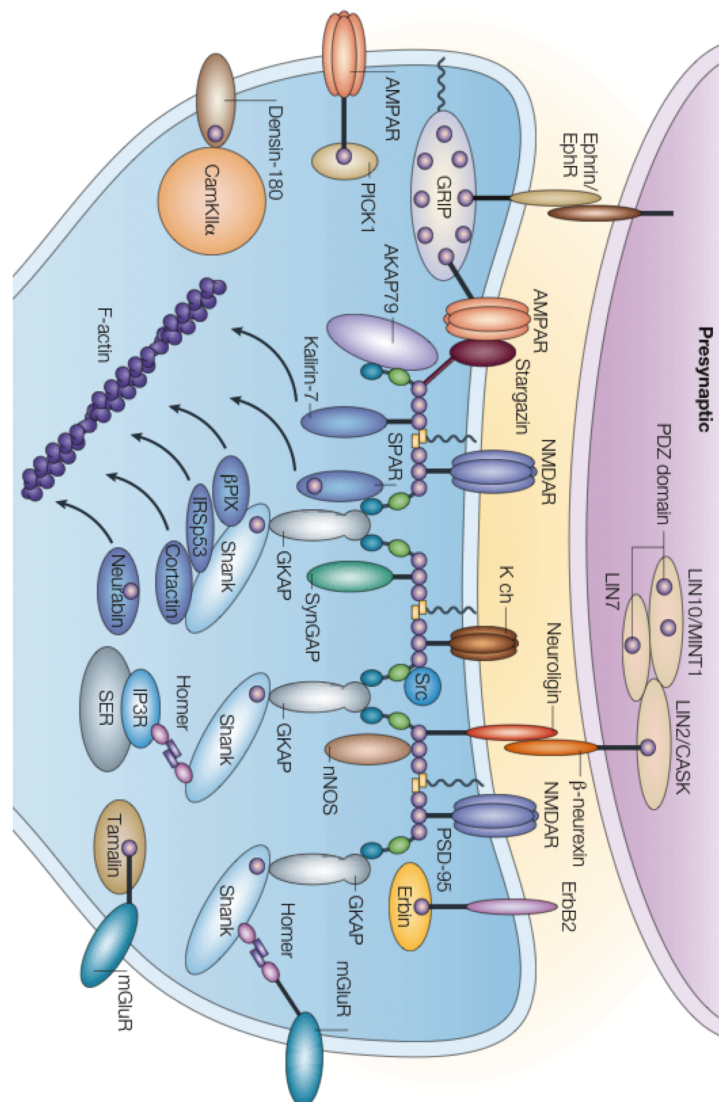
1.2.2 The molecular basis of cognition and disease

The protein interaction networks generated in this project are static models of mechanisms and pathways underlying molecular signalling, synaptic plasticity and disease. For that reason, the following paragraphs give an overview of synaptic plasticity and its associated signalling pathways as well as pathways involved in disease, in order to see these in a PSD context.

1.2.2.1 Synaptic plasticity

Synaptic plasticity is the modification (potentiation or depression) of synaptic transmission by experience, i.e. by other synaptic transmission. This modulated modification constitutes the basis of higher cognitive functions like learning and memory formation (Morris et al., 1990, Morris and Frey, 1997, Martin et al., 2000, Pastalkova et al., 2006, Whitlock et al., 2006). Although causal connections between plasticity and behaviour are hard to establish, some evidence, for example plasticity connection with fear conditioning (McKernan and Shinnick-Gallagher, 1997), show that it

Figure 1.2: Overview of the architecture of a synapse. The main PDZ-containing proteins of a glutamatergic synapse are shown, focusing on the postsynaptic density. PDZ domains are indicated by purple circles. The C-terminal cytoplasmic tails of membrane proteins are indicated by black lines. Specific protein–protein interactions are indicated by the overlap of proteins. Only a subset of known protein interactions is illustrated. Although not shown, Lin2, Lin7 and Lin10 are also present postsynaptically, and many of the proteins of the postsynaptic domain are also present in the presynaptic terminal. Green and blue ellipses in PSD-95 represent SH3 and GK domains, respectively. Figure from Kim and Sheng (2004).



is likely. Plasticity itself is a very wide term and it is well dissected into different types in an excellent review by Citri and Malenka (2008). There are two main types of plasticity, namely short term and long term. Short term plasticity, lasting between milliseconds to minutes has been observed in a wide range of invertebrates and mammals is generated by short or long trains of repetitive or tetanic stimulation (Zucker and Regehr, 2002). It can be either potentiating or depressing plasticity and can be attributed to mechanisms controlling the release of neurotransmitter, Ca^{2+} accumulation or postsynaptic mechanisms. Long term plasticity is expressed as long term potentiation (LTP) or depression (LTD) and has a longer temporal effect. For this reason it is an appealing mechanism to support Hebbian type learning (Lisman, 1989). The first evidence of long lasting activity dependent plasticity were given by Bliss et al. (Bliss and Gardner-Medwin, 1973), and further research gave rise to the, now well studied, CA1 hippocampal region model.

1.2.2.2 Synaptic plasticity in the context of PSD signalling

The main mechanistic distinction between LTP and LTD is based on their dependence upon the NMDA receptor (independent pathways include mGluR and endocannabinoid receptors). Many proteins have been implicated in the triggering, mediation/modulation and maintenance of LTP and LTD. Citri and Malenka cover the issue in their review, out of which we isolate and present highlights with specific PSD connections. In NMDA dependent LTP, a mechanism involved in long term-memory (Zola-Morgan and Squire, 1993, Martin et al., 2000), NMDA receptor activation is required (Malenka, 1991, Malenka and Nicoll, 1993) and the influx of Ca^{2+} beyond a threshold then acts as the initiator. This initiation is translated into LTP via downstream signalling pathways. A key component of these pathways is CamK2, which autophosphorylates after the triggering of LTP (Barria et al., 1997), but also phosphorylates AMPA receptors during LTP expression (Derkach et al., 1999; 2007). Also, activation of Pka kinase boosts the activity of CamK2, by inhibiting competing phos-

phatase activity (Blitzer et al., 1998). The Erk/Mapk pathway (Sweatt, 2004, Thomas and Huganir, 2004), Src kinase (Kalia et al., 2004) and Pkc (Pkm ζ isoform) (Hrabetova and Sacktor, 1996) play roles in stages of LTP induction and associated signalling. Beyond induction, expression and maintenance of LTP involves specific molecular mechanisms. Expression of LTP (in CA1) involves an increase of AMPA receptors clustered in the PSD (Bredt and Nicoll, 2003, Derkach et al., 2007). It is suggested that these receptors come from endosome recycling, a process mediated by Rab11a (Park et al., 2004), while MAGUKs like the Dlg are also good candidates for the process (Kim and Sheng, 2004, Montgomery et al., 2004), with PSD-95 being associated with surface expression of AMPAs (Ehrlich and Malinow, 2004). Maintenance of LTP also requires molecular signalling in order to establish protein synthesis (Zhou et al., 2006, Sutton and Schuman, 2006), depending on number of proteins including Pka, CamK4 and Erk-Mapk (Thomas and Huganir, 2004) as well as structural remodelling of the synapse (Lüscher et al., 2000).

In NMDA dependent LTD (Dudek and Bear, 1992) on the other hand, a modest Ca^{2+} influx can initiate this process (Mulkey and Malenka, 1992, Cummings et al., 1996), expressing the bidirectionality of plasticity. The signalling pathways that trigger LTD involve calcineurin, PP1 and inhibitor-1 (Lisman, 1989). The expression of NMDA dependent LTD, involves clathrin and dynamin mediated AMPA receptor endocytosis (Ashby et al., 2004, Blanpied et al., 2002). Other than the NMDA dependent LTP and LTD, signalling pathways have been involved in other types of LTP, LTD and/or plasticity like metaplasticity (“the plasticity of plasticity”) (Abraham and Bear, 1996, Abraham and Tate, 1997) and homeostatic plasticity (Turrigiano and Nelson, 2004). For example, p38, Erk, and Jnk have been shown to be involved in mGluR dependent LTD (Gallagher et al., 2004, Rush et al., 2002). Recent studies have also identified proteins of the cytoskeleton controlling spine morphology (Yoshihara et al., 2009, Tada and Sheng, 2006) and spine growth (Jaworski et al., 2009, Hoogenraad and Bradke, 2009), processes that have also been connected with plasticity pathways (Chen

et al., 2004, An et al., 2008, Lebeau et al., 2011).

It is speculated that since the triggering of plasticity is mediated by the integration of excitatory and inhibitory inputs to a neuron, there must be precise regulatory mechanisms to maintain the balance of excitatory and inhibitory transmission, the so called E/I balance (van Spronsen and Hoogenraad, 2010, Fritschy, 2008, Litvak et al., 2003, Gogolla et al., 2009). However, although knowledge has immensely progressed regarding molecular mechanisms of plasticity, there have been fewer attempts to put these within a larger PSD context.

1.2.2.3 The synapse and disease

Given the importance of synaptic signalling to normal brain function and development, it is natural to expect that mutations affecting synapse proteins may contribute to human psychiatric disorders. The Agency for Healthcare Research and Quality (<http://www.meps.ahrq.gov/mepsweb/>), cites a cost of \$57.5 billion in 2006 for mental health care in the U.S., equivalent to the cost of cancer care. Added to this cost is the cost of mental illness due to unemployment, expenses for social support and other indirect cost due to the individual's a chronic disability.

Indeed, functional genetic studies have shown that disruption of PSD proteins linked to glutamate receptor signalling alters cognitive function in rodents (Migaud et al., 1998, Husi et al., 2000, Grant, 2003), while drugs acting at synapses via antagonism of the glutamatergic NMDA receptors have long been known to result in a schizophrenia-like psychosis with cognitive disturbance. However, it is only comparatively recently that clear evidence has started to appear for a specifically synaptic involvement in complex psychiatric disorders such as autism (Jamain et al., 2003, Moessner et al., 2007, Berkel et al., 2010, Pinto et al., 2010, Hamdan et al., 2011) and schizophrenia (Kirov et al., 2009b). Additionally, a range of mental diseases or “abnormal” cognitive manifestations such as Fragile X syndrome (Pfeiffer and Huber, 2009, Dölen and Bear, 2008, Hagerman et al., 2005, Klemmer et al., 2011, Zalfa et al.,

2007), Parkinson's disease (Calabresi et al., 2006), compulsive behavior (Welch et al., 2007), and even addiction (Kauer and Malenka, 2007) have been shown to have synaptic or PSD related pathology. For example, of the 69 proteins linked to X-linked mental retardation 19 (28%) are postsynaptic proteins (Laumonnier et al., 2007). For some of the above diseases there are only gene association data, while for a very few cases, molecular mechanisms have been elucidated. A well studied example is the Fragile X mental related retardation, where it was recently shown that the affected FMRP protein directly interacts and affects the turnover of Dlg4 mRNA (Zalfa et al., 2007).

The earliest genetic studies, focusing on candidate genes, were based on small samples with only sufficient power to reliably detect disease-relevant mutations of relatively large effect. As a result, most reported genes failed to replicate in subsequent studies, and there was little consensus on which genes were the most strongly supported (for a review of schizophrenia studies see Harrison and Weinberger, 2005). To support the equivocal genetic data, comparisons were also made between gene expression and protein abundance between affected and unaffected individuals, some identifying differences in synaptic proteins. These studies were also of limited impact due to small sample sizes and problems in interpretation; in particular, it was unclear if the changes identified were primary causes of disease or secondary effects due to compensatory mechanisms or medication. When genome-wide association studies (GWAS) of common single nucleotide polymorphisms (SNPs) started to be performed, it became clear that conditions such as schizophrenia and bipolar disorder are highly polygenic, with potentially thousands of SNPs of small effect contributing to susceptibility (Purcell et al., 2009). The SNPs that have so far reached genome-wide levels of significance have not yet converged on a clear set of disease-relevant processes. Arguably the most productive area of research to date has been the study of rare structural variants, with early studies identifying a translocation of *Disc1* (Millar et al., 2000), *Pde4b* (Millar et al., 2005) and a micro-deletion causing Velocardiofacial syndrome as conferring increased risk of schizophrenia (for a recent review see Karayiorgou et al., 2010). There

have been cases where models have been proposed, involving whole pathways such as the Creb1 - Bdnf - Ntrk2 pathway in depression (Juhász et al., 2011). Genome-wide studies of copy number variants (CNVs), in which extended genomic sequences are duplicated or deleted, have discovered that large, rare CNVs contribute to both autism and schizophrenia (Redon et al., 2006, Walsh et al., 2008, Stone et al., 2008, Tam et al., 2009). Many CNVs disrupt multiple genes, making identification of the underlying risk factors difficult. Where it has been possible to link a CNV to disruption of a single gene, strong evidence for involvement of the trans-synaptic machinery has been found, with the identification of *Nrxn1* (Kim et al., 2008), *Nlgn3* (Jamain et al., 2003), *Nlgn4x* (Jamain et al., 2003), *Shank2* (Berkel et al., 2010, Pinto et al., 2010), *Shank3* (Moessner et al., 2007), *SynGAP1* (Hamdan et al., 2011, Pinto et al., 2010), *Dlgap2* (Pinto et al., 2010) and *Cntnap2* in autism, and *Nrxn1* (Kirov et al., 2009a) in schizophrenia. Interestingly, almost all of these genes regulate synapse structural organisation: the presynaptic neurexins (*Nrxn*) and their postsynaptic binding partners the neuroligins (*NLGN*) are also known to play a key role in synapse development and differentiation (Craig and Kang, 2007); while *Shank2*, *Shank3* and *Dlgap2* are PSD scaffolding molecules that organise postsynaptic signal transduction pathways, in various manners as discussed in earlier paragraphs.

In addition to being involved in mental diseases, PSD proteins are also involved in many types of neurodegeneration such as Huntington's disease (Harjes and Wanker, 2003, Goehler et al., 2004) and neurodegenerative dementias such as Alzheimer's disease. In the latter case, which has been well studied, various PSD proteins have been implicated including NMDA receptor subunits (Coleman and Yao, 2003, Shankar et al., 2007), AMPA receptor subunits (Armstrong et al., 1994), mGluRs (Allen et al., 1999), CamK2 (Gu et al., 2009), *Dlg4* (Gardoni, 2008), *Shank* and *SynGAP* (Gong et al., 2009), neuroligins (Zhong et al., 2008), integrin (Caltagarone et al., 2007) and cadherins (Serban et al., 2005), which have all been shown to be affected by the Alpha-beta amyloid toxicity.

Although there is no point listing cases of strong or weak genetic associations, what is striking from the waxing volume of data on synaptic proteins and disease is that, beyond the polygenic nature of many of those diseases, there is a clear overlap in genetic associations. Examples of genetic overlap between diseases include autism and fragile X syndrome (Budimirovic and Kaufmann, 2011), schizophrenia and bipolar disorder (Purcell et al., 2009) and autism and schizophrenia (Stone et al., 2008). This could lead to the speculation that while a gene might be involved in more than one diseases, it is the molecular context within which it acts that decides the manifesting phenotype. This molecular context is defined by the interactions of the gene products within pathways or molecular machines like the PSD.

1.3 Systems neurobiology of the synapse

1.3.1 Neurobiology and the Omics

The advance of molecular biology allowed neuroscience to move from studying neuronal circuits to studying molecules of interest. In a similar fashion the -omics era, with its continuous improvement of methods (e.g. nucleic acid arrays, mass spectrometry, next generation sequencing, etc) and associated analysis approaches has (again) revolutionised modern neurobiology, giving direction towards the combination of a wide range of data into system-wide models. This effect has also, partially at least, shifted research strategies from direct hypothesis testing to data driven approaches. In a recent review Geschwind and Konopka (2009) highlight a series of interface areas between neuroscience and systems biology. The first area is public data sharing and resource integration, including resources such as the Gene Expression Omnibus (GEO) database of microarray data (Barrett et al., 2009; 2011), the Allen Brain Atlas (Ng et al., 2009, Jones et al., 2009), GenePaint (Visel et al., 2004, Alvarez-Bolado and Eichele, 2006), GENSAT (Heintz, 2004), and BGEM (Magdaleno et al., 2006). All the latter resources allowed scientists to explore the transcriptome and proteome of neu-

rons and synapses. The second area of interface is genotype and phenotype integration using quantitative trait locus (QTL), CNV, or SNP analyses allowing to associate genes or genetic loci with a specific phenotype. The third area of interface is the accelerated discovery associated with next-generation sequencing, which allowed not only to re-sequence genes, but to quantify mRNA, RNA splicing, epigenetic phenomena, DNA binding, CNVs, insertions, and mutations. The final and most relevant are to this work is the move from lists, such as the ones provided by proteomics or microarray data, to networks. Biological networks allow the study of systems in an integrative manner and also, because of their nature, allow the utilisation of a series of analysis methods from network science. Moreover, this network perspective allowed vertical integration between different types of data including genomics, proteomics and phenotypes.

Since this work has a strong focus on the reconstruction of PSD protein interaction networks based on proteomics data, we will describe the basic aspects of this neurobiology discipline in the following paragraphs.

1.3.2 Proteomics of the synapse

The term proteome (Wasinger et al., 1995) was coined to describe the complete set of proteins expressed in a cell or organism, as an analogy to the term genome. Proteomics is a vast discipline in its own right. Proteomic methodology includes techniques such as protein electrophoresis and 2D electrophoresis (O'Farrell, 1975) as well technologies such as mass spectrometry (MS) and chip based methods (e.g. Schutkowski et al., 2005, Coba et al., 2008; 2009) to identify and catalogue proteomes of multi-protein complexes, organelles, cells, tissues, organs or whole organisms.

A typical approach for the identification of synaptic protein complexes requires the isolation of the complex from a, usually pre-fractionated, sample extracted using various types of detergents (e.g. Triton X-100, ComplexioLytes, and DOC). Pre-fractionation is achieved by extracting synaptic fractions from homogenised brain tissue through synaptosome isolation methods, such as the one described by Carlin et. al

(Carlin et al., 1980) and Wu et. al (Wu et al., 1986) or more recent optimisations such as the ones used in (Husi et al., 2000, Husi and Grant, 2001, Klemmer et al., 2009). The characterisation of synaptic complexes has generally been performed by MS applied to the protein sample, after it is separated by electrophoresis (e.g. SDS-PAGE or native PAGE): enzymes cut the proteins into fragments, which are then ionized, fired through an electromagnetic field, and their mass to charge ratio is then measured by a detector. The abundance of individual peptides is calculated from the resulting spectrum, and clusters of peptides corresponding to individual proteins (or sets of closely related proteins) identified. The ability to reliably detect a protein will depend on its abundance, the number of characteristic peptide fragments it is cleaved into, and how well these peptides ionise and 'fly' within the machine. With improving technology it has become possible to identify low abundance proteins, an inevitable side-effect of which is the increased identification of trace contaminants. This problem may be reduced by improvements in isolation techniques, more extensive validation of identified proteins and the removal of known common contaminants from results. The most prevalent MS method is liquid chromatography-mass spectrometry (LC-MS) (Yates et al., 1996), although matrix-assisted laser desorption/ionization reflector time-of-flight mass spectrometry (MALDI-TOF-MS) has also been used, primarily in early publications. For a review of the methods see (Domon and Aebersold, 2006).

The first cataloguing attempts of the proteome of mammalian synaptosomes revealed the presence of over 1000 proteins (Husi et al., 2000, Schimpf et al., 2005). However, in order to isolate specific components of the synaptic machinery affinity or immunoprecipitation methods can be used. In these a "bait" protein is immobilised on a resin via interaction with an antibody against an epitope or a genetically engineered tag; contaminants are removed with repeated washes; then the complex of "prey" proteins binding to the bait (both directly and through interactions with other proteins) are eluted and identified. A similar approach is the use of resin with a bound synthetic peptide acting as an artificial protein interaction domain. In analysing the composition

of such complexes, it must be remembered that affinity and immunoprecipitation based methods are susceptible to biases ranging from non-specificity of the affinity reagent to potential inability of a genetically tagged protein to be post-translationally modified in order to interact with some of its partners. Additional problems may arise if the immunoprecipitation epitope or affinity tag overlap interaction domains required by prey proteins or from the presence of promiscuous non-specific interactors. A common way to tackle these is by using multiple antibodies Schwenk et al. (2009). The transgenic Tandem Affinity Purification (TAP) method (Puig, 2001), which can tackle some of these issues by using two consecutive purifications with two different affinity tags, was used to generate some of the data in this project. The MS data in this case have to be compared to a negative control, where no affinity purification took place, revealing the true interactors of a bait protein.

Cataloguing the neural proteome in human brain regions is a project undertaken by the Human Brain Proteome Project (HBPP) (Hamacher et al., 2008). This is a complex task since different brain areas express different PSD proteins (Emes et al., 2008) and also synaptic plasticity has an effect on the composition of the synaptic proteome (McNair et al., 2006, Henninger et al., 2007, Piccoli et al., 2007). The aforementioned methods have been applied to the synaptic proteome both on the pre- and postsynaptic sides. Presynaptic proteome proteomics have yielded interesting models of the synaptic vesicle (Morciano et al., 2005, Takamori et al., 2006) and presynaptic protein complexes (Burré et al., 2006, Burré and Volkandt, 2007, Morciano et al., 2009) as well as common pre- and postsynaptic proteins (Collins et al., 2006, Phillips et al., 2005). Postsynaptic proteome studies have mostly been performed in mice, rats and humans. One of the earliest high throughput studies identified proteins in a series of 26 prominent multi-protein bands from synaptosome preparations using MALDI-TOF-MS (Walikonis et al., 2000). Other proteomic studies focusing on the PSD followed, including studies by Jordan et al. (2004), Peng et al. (2004), Yoshimura et al. (2004), Collins et al. (2006), and Klemmer et al. (2009). More recent additions to these lists

come from Hahn et al. (2009) and Bayés et al. (2010) who studied the human PSD, as well as Trinidad et al. (2008) and Coba et al. (2009), with the latter two focusing on phosphorylation and signaling associated proteins. With these constant additions the total number of proteins in the PSD has risen to more than 2000 (Bayés and Grant, 2009).

The high complexity of the PSD proteome is put into context when looked at from the perspective of protein complexes. A number of receptor protein complexes have been analysed including the AMPA (Husi et al., 2000), mGluR5 (Farr et al., 2004), serotonin (Bécamel et al., 2004), and nicotinic (Kabbani et al., 2007, Paulo et al., 2009) receptors as well as ion channels such as the K^+ channel Kir2.2 (Leonoudakis et al., 2004). The first studies to isolate protein complexes from within the PSD focused on the NMDA receptor (NMDAR), which is coupled to signalling pathways via MAGUK-family (and other) scaffold proteins and plays a major role in the induction of synaptic plasticity. Initially 100 proteins were identified in isolates using an antibody to the NR1 subunit of the receptor then 170 by peptide-affinity purification using a MAGUK-binding peptide from the C-terminus of the NR2B subunit (Husi et al., 2000, Husi and Grant, 2001, Farr et al., 2004, Collins et al., 2005). The combined set of 186 proteins, referred to as NRC/MASC (NMDA Receptor Complex/ MAGUK-Associated Signalling Complex) has been the subject of a number of subsequent analyses and will also be used and referenced throughout this project. More recently the transgenic TAP technique was applied by Fernández et al. (2009) to characterize 118 proteins in complexes containing PSD-95. The latter was part of this project and is discussed in Chapter 4.

Although isolation and identification methods have improved over the years, the latter studies revealed protein sets with an overlap in the area of 50%. An attempt to define a consensus PSD proteome or postsynaptic proteome (PSP), abbreviated cPSD, was made by Collins et al. (2006). Utilizing 1D gel electrophoresis of synaptosome protein extracts and LC-MS, 698 proteins were identified in the mouse postsynaptic

terminal, of which 620 had previously been found in PSD preparations. These were combined with data from other studies (Walikonis et al., 2000, Jordan et al., 2004, Peng et al., 2004, Yoshimura et al., 2004, Farr et al., 2004) to produce a list of 1126 postsynaptic proteins, of which 446 were found in two or more studies. The majority of PSD proteins had been identified only once (Collins et al., 2006) at the time, although, besides proteins of very central importance, such as receptors the situation remains similar and has to do with technical issues in the complex purification or MS analysis¹.

1.3.3 Models of the PSD

The existence of the synaptic proteome catalogues discussed in the previous subsection is what initiated the attempt to generate static models of PSD data, in order to understand not only the composition but also the organisation of the synaptic proteome. In the following paragraphs we will discuss important datasets, as well as the major modelling attempts of PSD proteomics data, which inspired this project.

1.3.3.1 The NRC/MASC model

The accumulation of PSD proteomic profiling data at the beginning of the last decade led to the first and seminal PSD interactome model, published by Pocklington et al. (2006). The authors reconstructed and analysed a model of the NMDA receptor and MAGUK associated molecules complexes (NRC/MASC) using rigorous annotation and curation of its constituent parts and the interactions between them. They also performed analysis of the network and annotations leading to the first model that described not only the architecture of the PSD proteome but its correlations with the molecular basis of plasticity and disease. The dataset was based on 186 proteins of the NMDA, AMPA and mGluR receptor complexes (Husi et al., 2000, Husi and Grant, 2001, Farr

¹Interesting extreme examples can be found within the AMPA receptor complexes where Cornichon, a very small AMPA receptor interactor that had not been discovered because of its small size (Schwenk et al., 2009) and Ckmp44, a plasticity associated protein, was not identified since it was, at the time a novel protein, not annotated in SwissProt (von Engelhardt et al., 2010).

et al., 2004, Collins et al., 2006), including receptors, key interacting components such as MAGUKs (e.g. protein of the DLG family), and signalling proteins. These proteins were annotated and classified in functional families. Protein interactions were retrieved and manually quality controlled, resulting in a protein interaction network of 105 proteins with 248 interactions, which segregated into 13 clusters or modules after computational analysis (Figure 1.3).

The results of the study showed that the NRC/MASC dataset had highly enriched protein domains (Table 1.1) with key synaptic signalling functionality including calcium binding, G-protein coupled signal transduction, phosphorylation, scaffolding and membrane localisation. Interestingly, 24% of the proteins in the NRC/MASC dataset were involved in plasticity, 29% in rodent behaviour, 23% in learning, and 29% in mental illness. Further statistical analysis of the data showed significant statistical correlations between functional families and specific annotations, e.g. the Glutamate receptors family showed correlations with synaptic plasticity, behaviour, cognitive disorders and schizophrenia and the Phosphatases family with synaptic plasticity. Also, interestingly there were significant overlaps between proteins involved in more than one annotations, for example synaptic plasticity showed overlap with disorders such as schizophrenia and bipolar disorder, revealing a possible overlap in the underlying biological mechanisms. Most importantly, statistical analysis showed correlations between modules and specific annotations, e.g. cluster 1 of the network appeared enriched in ionotropic glutamate receptors, as well as schizophrenia and synaptic plasticity correlations, while cluster 2 appeared enriched in metabotropic glutamate receptors and behaviour correlations. Another important finding from this analysis had to do with the position and the importance of the protein in the network and how that correlates with the effects of its mutation. Using all the data available the authors showed a correlation between the number of interactors a protein has and the effect its mutation has on measured LTP (Figure 1.4).

Overall the reconstruction and analysis of the NRC/MASC model yielded some

Figure 1.3: The NRC/MASC protein interaction network segregated in clusters (briefly summarised). Each of the resulting clusters showed enrichment in distinct groups key annotation terms of PSD functionality. Figure from Pocklington et al. (2006)

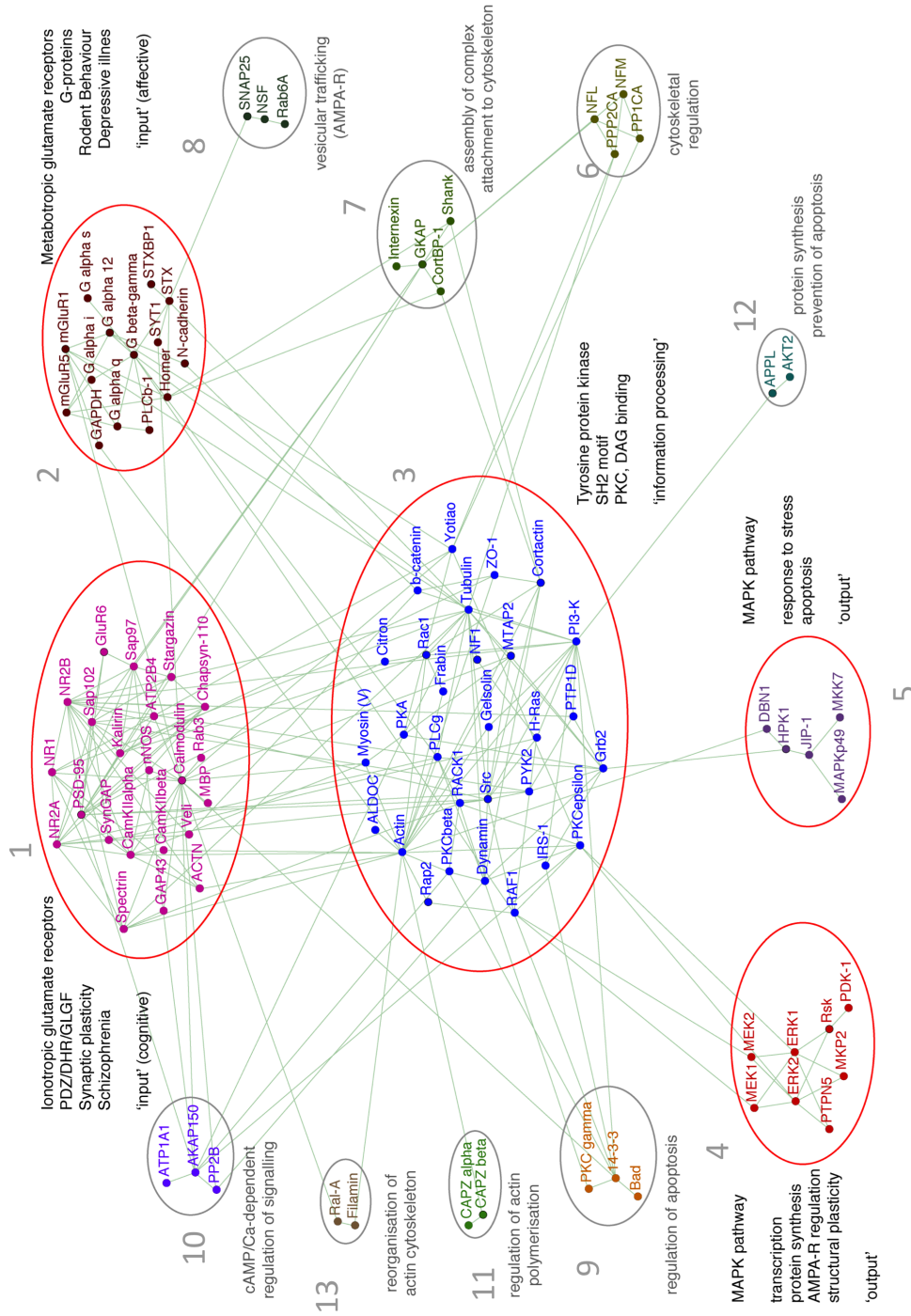
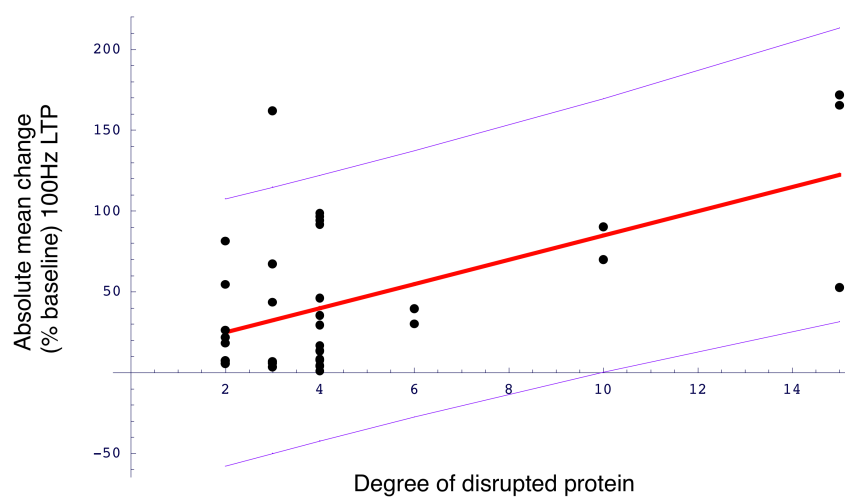


Table 1.1: Ten most common protein domains in NRC/MASC proteins. All domains in this list appear enriched compared to the genome and have key synaptic signalling functionalities, highlighting the involvement of the the complex in such processes.

Domain	n-fold enrichment compared to genome
IQ calmodulin-binding region	12.1
PDZ/DHR/GLGF	7.3
Serine/threonine-protein kinase domain	6
C2 calcium-dependent membrane targeting	5.9
Src homology-3 domain	5.3
Pleckstrin homology	4.7
Pleckstrin homology-type	4.5
Small GTP-binding protein	3.2
Protein kinase, catalytic domain	3.1
Calcium-binding EF-hand	2.9

Figure 1.4: An analysis performed on 36 proteins nodes in the NRC/MASC model (where data was available) shows that the effect of mutation of the node is significantly correlated to the node degree. The plot shows absolute change (% baseline) as a function of node degree. Data from personal communication with AJ Pocklington and Pocklington et al. (2006)

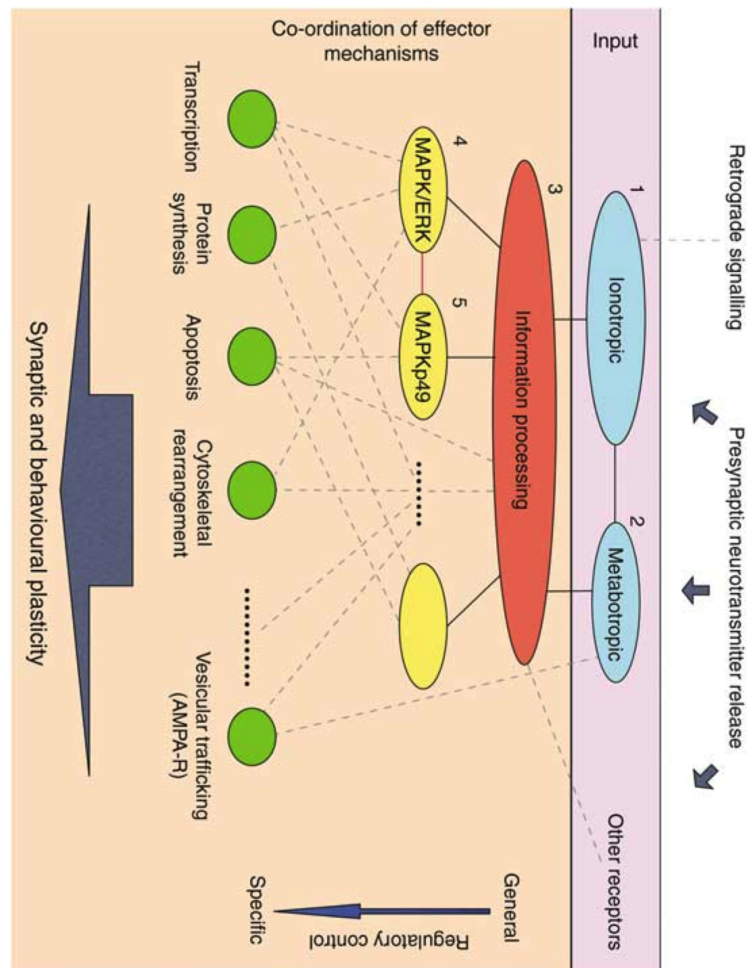


important findings regarding the modularity in the architecture of the PSD and how modules correlate with specific functionality, often in an overlapping manner. However, the most important conclusions were regarding the organisation of PSD protein complexes as a molecular computation machine. It was shown that proteins are organised in a modular protein interaction network, where each module shows specific correlations with molecular functions and biological processes. Analysis of this organisation showed that a higher level model of organisation (Figure 1.5) could be extracted from the protein interaction network. MASC proteins clustered into modules with specific functional roles. Primary signal reception modules (blue in Figure 1.5) are formed around ionotropic and metabotropic receptors. This 'input' layer directs information to a large 'information processing' layer (red in Figure 1.5) responsible for the integration and co-ordination of downstream processes. Other sources of input ('other receptors' in Figure 1.5) may feed into this module directly, or through smaller input/processing modules. Furthermore, modules appear not to have individual but multiple roles distributing information processing and regulation. The lower 'information processing layer consists of several intermediate modules (yellow in Figure 1.5), which regulate overlapping sets of pathways, while numerous small modules (green in Figure 1.5) are specific to individual effector responses, which constitute the 'output' layer. Interactions between components of these layers do not take place in with simple feed-forward mechanism, rather a dynamical balance between multiple functional processes.

1.3.3.2 The PSD phosphoproteome model

In light of the static models describing protein interaction networks of the PSD, namely the NRC/MASC model, as well as the models described in Chapters 4 and 5, the need to apply the same type of analysis on the dynamic aspect of these networks became apparent. Activation of NMDA receptors is known to modulate the activity of post-synaptic tyrosine, serine and threonine kinases (Kandel, 2001, Collins et al., 2005,

Figure 1.5: Modular structure and functional organisation of the NRC/MASC protein interaction network. Receptor complexes modules form the input layer (blue), which connects to signalling modules forming the information processing layer (red), which integrates the signals. The latter in turn connects to the output layer (green) which consists of modules directly regulating specific biological processes. Note that the connectivity is not strictly feed-forward manner, but there also are regulatory loops present. Figure from Pocklington et al. (2006)



Trinidad et al., 2006, Munton et al., 2007) and computational models have supported the advantage of the organisation of these phosphorylation interactions in signalling networks (Jordan et al., 2000, Ma'ayan et al., 2005). Coba et al. (2009) made the first attempt of modelling these synaptic phosphorylation networks, integrating data from previous knowledge on the PSD kinome with proteomic data on the change of the phosphorylation state of PSD proteins after LTD induction and receptor activation as well as phosphorylation peptide array data. Results showed that LTD induction elicited changes in a wide range of functional families of PSD proteins, including channels and receptors and scaffolding proteins accounting for 10% and 15% of proteins modulated respectively. Measuring the activation of 23 PSD kinases, 9, representing all four major kinase groups, showed NMDA dependent activity changes. Additionally, agonist activation of NMDA, dopamine and mGluR receptors, as well as Pka and Pkc kinases drove phosphorylation of NMDA and AMPA receptor subunits, revealing cross-talking regulatory pathways. Furthermore, using phosphoproteome (PPP) arrays, representing 300 phosphorylation sites of 92 PSD proteins, the authors probed 25 PSD kinases. The results revealed a complex kinase-substrate network (Figure 1.6), which was further investigated and found to be representing various phosphorylation mechanism motifs (single / multiple site, kinase divergence / convergence, pairing, and priming). Overall these results highlight the potential complexity within the already complex PSD molecular machine. Taking into account the dynamic nature of the phosphoproteome network the authors speculate that it could provide a framework for the transmission of information from a single neurotransmitter to a numerous output proteins in an orchestrated manner, also providing a basis of resilience to perturbations. Furthermore, with computational models indicating that two-state synapses perform poorly on memory storage compared to multi-state synapses (Fusi et al., 2005, Fusi and Abbott, 2007), this framework could provide a model for the latter type of synapse, via the wide range of possible states of kinases and substrates.

1.3.4 Evolution of the PSD

Most of the PSD studies in the literature have been performed using mammalian and most specifically mouse or rat samples. It is important, however, both for the basic understanding of the PSD's function, but also for future applications (e.g. identification of drug targets or disease susceptibility genes) to also look at the evolution of this molecular machine. This evolutionary study of the PSD extends both to organisms with "simpler" nervous systems, possessing PSD orthologs, as well as the human PSD, which is of major interest due to the cognitive and disease aspect of it. The following paragraphs summarise the state of the art regarding the evolution of the PSD and non-murine PSD models. Although our focus is on proteomics based datasets and models throughout this work, a lot if not most of the work on the study of the evolution of PSDs has been based on comparative genomics approaches, which with the exception of the human PSD, dominate this subsection.

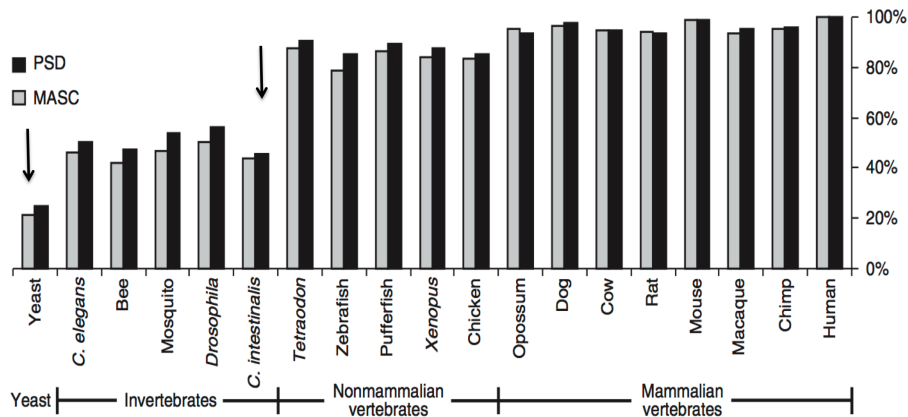
1.3.4.1 Evolution of PSD complexes

Sakarya et al. (2007) assembled phylogenies from 36 PSD gene families from the genomes of demosponge *A. queenslandica* (lacking neurons and nervous system), cnidarian *N. vectensis* (possesses a nerve net), reference invertebrate (*D. melanogaster*), and mammalian (*H. sapiens*) species. From comparison of the data the authors found that a surprising number of PSD proteins are present in the demosponge and cnidarian genomes. Furthermore, they found that domain organisation in dlg (representative of the Dlg family) was 100% conserved between these species and also that the conserved PSD genes were expressed in ciliated *A. queenslandica* cells. Other studies have confirmed that many families of synapse and cell signalling genes are present in the phylum Porifera (sponges), supporting the hypothesis that core synaptic signalling components were present at the base of animal kingdom (Yasuyama et al., 2002, Ruiz-Cañada et al., 2002, te Velthuis et al., 2007).

In an attempt to investigate the evolutionary origins and trajectory of the PSD and

its architecture a larger scale comparative genomics approach was applied by Emes et al. (2008). This approach was based on the proteomics data from the NRC/MASC and PSD datasets (Husi et al., 2000, Husi and Grant, 2001, Farr et al., 2004, Collins et al., 2006). One-to-one mappings of orthologues of genes encoding for proteins in the PSD were retrieved for 19 species. The species studied comprised of a wide range of animals with nervous systems of differing anatomical complexity, including invertebrates, non-mammalian vertebrates, mammals and also an out-group that does not possess a nervous system (*S. cerevisiae*). The authors observed that approximately 23% of all mammalian synapse proteins were detected in yeast (21.2% NRC/MASC, 25.0% PSD) and 45% were detected in invertebrates (46.2% NRC/MASC, 44.8% PSD) (Figure 1.7). Therefore, a substantial proportion of genes encoding MASC and PSD orthologues predated the origins of the nervous system, with apparent stepwise expansions following the divergence of metazoans from eukaryotes and vertebrates from invertebrates. Of these genes of course, the ones coding for membrane receptors, are absent from species lacking a nervous system. Further investigation suggested that most functional types of synaptic proteins were present in early metazoans and that the proto-synapse constructed from this core functionality has been elaborated on during the evolution of invertebrates and vertebrates. A more careful examination of individual functional families of PSD genes shows that this expansion appears to have primarily involved gene family expansion and diversification among upstream signalling and structural components (receptors, scaffolding proteins, and cytoskeletal, adhesion and signal transduction molecules). These gene family expansions, have possibly resulted from genome duplications and can be seen in Figure 1.7 as “jumps” in percentage of human PSD genes marked by the arrows. It must also be noted that in the light of the new human PSD proteomics data (see next subsection), a recent follow-up study (Emes and Grant, 2011) showed that there is conservation of human PSD genes, protein domains and even functional pathways that spans from prokaryotic organisms to human. A characteristic example is the chemotaxis system in *E. coli*, components

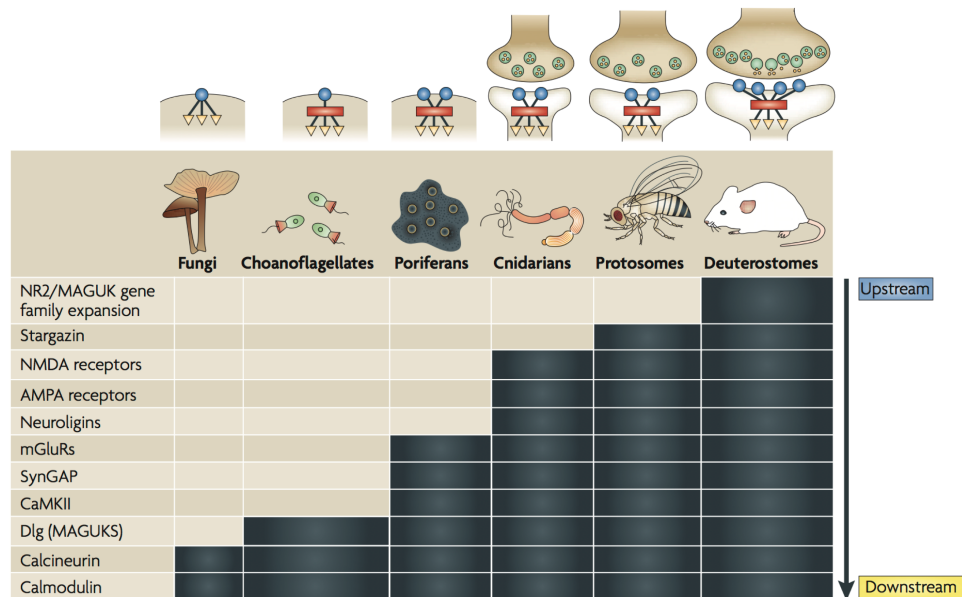
Figure 1.7: Comparison of PSD and MASC orthologues across species. The occurrences of human PSD and MASC orthologues found in each of the 19 species are shown as a percentage of those found in human. Black arrows mark major events of gene family expansions. Figure from Emes et al. (2008)



of which have homologs in the human PSD. Emes et al. obtained expression data for ~150 MASC genes in ~20 brain areas and showed that proteins in the upstream components ('input' and partially 'information processing' layers) showed greater variation in expression and were of more recent evolutionary origin.

Research on other non-mammalian PSD complexes has been done either in a gene or small pathway context, or in order to augment comparative genomics approaches. Among these, the case of model organism *D. melanogaster* (fruitfly) is interesting. Xia et al. (2005) showed that NMDA receptors mediate learning and memory in *D. melanogaster*. In a complementary study Leibl and Featherstone (Leibl and Featherstone, 2008) showed that the fruitfly genome has homologs for mammalian PSD receptors such as NMDA, AMPA and Kainate receptors as well as proteins associated with trafficking of these receptors and also demonstrated the effect of one of them (pod1) to glutamate receptor clusters. Interestingly, most synapse work in fruitfly has been done in a neuromuscular junction (NMJ) rather than a neuronal synapse context although the two are highly similar (Collins and DiAntonio, 2007). To our knowledge the only attempt to isolate and identify fruitfly PSD complexes was done by Emes et al. (Emes et al., 2008), where the authors used a synthetic hexapeptide NR2 C-terminus bait and isolated 220 fly MAGUK and MASC (fMASC) proteins. Although no net-

Figure 1.8: The emergence of titular MASC components across clades is illustrated. Proteins are ordered based on whether they are located 'upstream' or 'downstream' in synaptic signal transduction pathways of the NRC/MASC model. Non-coloured fields represent the absence of a given protein, whereas dark grey rectangles denote its presence. Diagrams of MASC structure are placed above each clade, along with an illustration of a representative model organism. Figure from Ryan et al. (2008)



work was reconstructed it was found that ~25% of the fMASC proteins were upstream components, in contrast with ~60% in the mouse case.

Ryan and Grant (2009) reviewed data from Emes et al. (2008), along with data on choanoflagellates, sponges and cnidarians. They give a descriptive review of the progression from organisms lacking a nervous system but having components to form a “protosynapse” to the complex synapses of mammals, via the simpler synapses of metazoans. Their review is summarised in Figure 1.8, where the authors show how NRC/MASC components started accumulating during evolution, allowing the formation of the deuterostome synapse. The high complexity of the deuterostome synapse according to the model is also a result of the gene family expansion of key PSD families like the DLGs and the second subunit of the NMDA receptor (NR2), which allowed more different interaction combinations. This is also supported by the evolutionary elaboration of the intracellular terminal of NR2B, which in contrast with the invertebrate gene that only interacts with PDZ domains, has more interaction interfaces (Ryan

et al., 2008). The authors also give some 'synapse first' speculations, hypothesising that current evidence shows that the presence of some synaptic components allowed further evolution in mechanisms of synaptogenesis.

1.3.4.2 The human PSD

The human PSD (hPSD) is of great interest due to the aforementioned involvement in cognition and disease. The first analysis of proteomics data from hPSD samples was not published until recently, partly because of issues with acquiring human brain tissue samples. Bayés et al. (2010) isolated hPSD complexes from fractionated human neocortical biopsies samples and using LC-MS identified 1461 proteins (748 detected in triplicate). This study focused more on disease and phenotype annotation, rather than network reconstruction and showed that 269 monogenic diseases result from mutations of 199 hPSD (14% of total) genes. Of these diseases 133 (~50%) are nervous system disorders, with ~80% being central nervous system disorders. The authors also showed that hPSD proteins and their mouse orthologs were enriched in 21 neural phenotype annotations from the Human Phenotype Ontology (HPO) (Robinson et al., 2008) and 77 neural phenotype annotations from the Mammalian Phenotype Ontology (MPO) (Smith et al., 2005), including cognitive and motor phenotypes. A reconstructed network model, although unpublished, showed overall high similarity to that of the mouse cPSD (L.N. van de Lagemaat, personal communication). The results of this analysis corroborates indications from the NRC/MASC mouse model, regarding the association of the hPSD in cognition and disease and also offers tangible evidence about the constituent parts of the human PSD and its similarity to the mouse PSD as speculated in earlier studies (see Emes et al., 2008 and Chapter 5).

Another important aspect regarding the evolution of the PSD under the light of the hPSD proteomics data was revealed after the analysis of dN/dS ratios (Hurst, 2002) (where dN represents amino-acid substitution frequency and dS the background rate of neutral DNA changes). Comparisons of human dN/dS ratios with mouse, chimp

and macaque showed that hPSD proteins were highly conserved compared to the rest of the genome, a phenomenon not restricted to the human lineage. It was also shown that hPSD proteins were more conserved than other brain expressed proteins, which are also known to evolve with a slower rate (Winter et al., 2004, Khaitovich et al., 2005, Wang et al., 2007). This agrees with previous reports showing that proteins with many interactions are more conserved (Fraser et al., 2003).

1.3.5 Neuroproteomics informatics

In order to make biological sense of the efforts to catalogue and model the PSD (or in general the synaptic proteome) it is of great importance to handle and organise the data (Kumar and Mann, 2009). Neuroproteomics (and proteomics in general) is notorious for the large volumes of data that it generates. Additionally, given the volume of data from annotations of genes and proteins as well as PPIN models of complexes, the need for neuroproteomics informatics is apparent. Although there is a large collection of freely available biological databases for functional annotation, known protein interactions and the associated phenotypes and physiology of genes, bespoke specialist neuroproteomics databases are now starting to appear. Prominent examples of these are SynDB (Zhang et al., 2007) (<http://syndb.cbi.pku.edu.cn/>) and Genes2Cognition (Croning et al., 2009) (G2Cdb, <http://www.genes2cognition.org/db/>) databases. SynDB is a database of genes and proteins with known (or predicted) synaptic function based on other annotations. G2Cdb is a data warehousing project, which compiles a catalogue of the mammalian synapse genes along with information on their synonyms, annotations, and associations with phenotypes of disease, behaviour, or physiology.

1.4 Motivation, hypothesis and goals

1.4.1 Motivation

The number of known synaptic proteins has increased 10-fold (Croning et al., 2009), since 2000, mostly as a result of improvements in the complex isolation, purification and characterisation methodology, and overall data acquisition process. However, even as recently as 2009, when Bayés and Grant reviewed the discipline of neuroproteomics, there was no standardised methodology for the analysis of the data. The effect of that can be seen in an otherwise excellent paper by Klemmer et al. (2009), where the authors deliver a high standard proteomic dataset but admit that the resulting reconstructed model (generated by expensive proprietary software) lacked many interactions and failed to integrate and represent the data. Due to that the analysis of the dataset was confined to the protein lists. From the previous paragraphs it must be now evident how reconstructions of models based on proteomics data can generate descriptive models. These models not only describe the data but also tease out hidden information that lies within the organisational principles in the PSD molecular machine. We find that the NRC/MASC model set a standard for the reconstruction and analysis of models from neuroproteomics data. This was achieved not only by successfully applying an approach to reconstruct the NRC/MASC complex but also because the resulting analysis showed a clear modular structure within it. The most interesting feature of this structure was that modules correlated with somewhat specific biological functions. We believe that being able to dissect and investigate that modular structure will lead to a better understanding of the function of the PSD molecular machine. Based on that we decided to create a standardized pipeline and apply it to different proteomics datasets in order to show that this modular structure and architecture is a prominent feature of PSD complexes and reflects a modularity in biological functions.

Furthermore, by applying this pipeline to proteomics data of the PSD, we noticed a missing link between work that has been done to characterise PSD protein in simple

organisms (e.g. cnidaria) and mammals. There was an impressive lack of information on the PSD complexes of the fruitfly. *D. melanogaster* is a model animal and because of that a vast repertoire of techniques and tools to genetically manipulate it is available. Also, given the evidence backed speculations that the fruitfly possesses similar but slightly “simpler” PSD complexes, we thought it would be interesting to attempt a first mapping of these using direct experimental methods rather than comparative genomics. This would allow not only to catalogue the fly PSD but also investigate whether the same modularity principles that govern complexes like the murine NRC/MASC apply there as well.

1.4.2 Hypothesis

The motivation to undertake this project was distilled into the following set of testable hypotheses:

1. PSD proteomics data, although substantial, is still sparse. These data can be augmented using new methods and affinity purification strategies isolating novel or known PSD proteins in the context of their protein complexes.
2. Protein interaction networks of the PSD have a modular architecture. This structure or modularity is persistent with respect to the addition of new data or the examination of different subsets of the PSD, as acquired by using different protein baits in affinity purification experiments. This modular architecture reflects, to some level, a functional significance.
3. The lower complexity in constituent parts of organisms with less intricate nervous systems and behavioural repertoires is reflected both in proteomics data and the reconstructed network models. However, although less complex from a constituent parts perspective we also hypothesise that there is evidence of the aforementioned modular architecture as well as a meaningful biological interpretation of it.

4. Reconstructed models of mouse and fruitfly PSDs are comparable, namely through their constituent parts and modular architecture. Furthermore, we expect to see some basic modules and their functionalities being conserved up to the level allowed by the presence of homolog gene products and their successful proteomic isolation.

1.4.3 Goals

In order to address these hypotheses we set out to achieve the following goals. Initially we compiled and put together the methods for the model reconstruction and analysis pipeline (Chapter 3). Followingly this pipeline was applied to a) a new mouse proteomics dataset (Chapter 4), addressing hypotheses (1) and (2), as well as b) a mouse proteomics dataset compiled from previous high quality datasets (Chapter 5), addressing hypothesis (2). Additionally, we set out to produce the first model of the fly PSD by isolating and identifying protein complexes and applying the same pipeline to this dataset (Chapter 6), addressing hypothesis (3). Finally, after obtaining these models, we discuss observations on their comparison in Chapter 7, addressing hypothesis (4).

Chapter 2

Materials and Methods

2.1 *Drosophila* strains

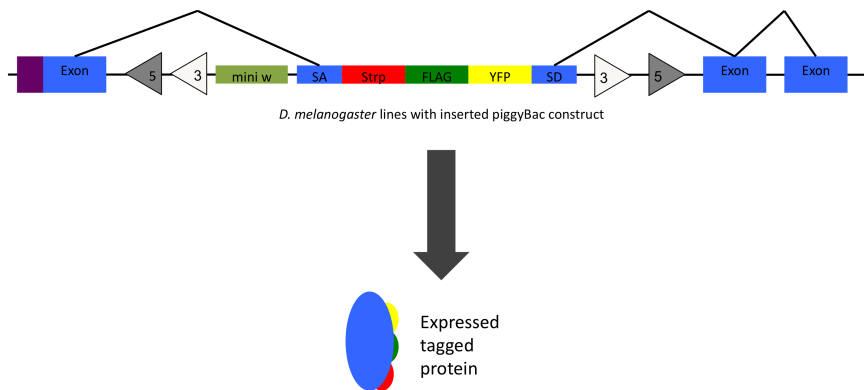
Materials and methods described here were used by the author to generate the fly PSD (fPSD) proteomics data discussed in Chapter 6. For the methods used for the generation of the mouse PSD data discussed in Chapters 4 and 5 refer to Fernández et al. (2009) and Husi et al. (2000), Husi and Grant (2001), Farr et al. (2004), Collins et al. (2005), Fernández et al. (2009) respectively. Generation of the two aforementioned datasets was *not* performed by the author.

2.1.1 Handling

Flies were transferred from 18°C stock vials and reared on standard cornmeal food in bottles at 25°C in LMS cooled incubators. Food was ‘Dundee food’ (a standard yeast, cornmeal and agar medium) prepared in the Swann media kitchen, Kings Buildings at 18°C or room temperature. Precise numbers of flies were not maintained in vials but stocks were kept healthy by frequent changes into fresh food when required judging on the size of individuals and mobility of larvae. Stocks were tipped frequently to avoid mite pests and yeast paste (made from dried Allison’s yeast mixed with water) was added to vials with sick stocks when required. Small volumes (~0.5-3ml) of water

were occasionally added to vials or bottles when required, to prevent food from drying out.

Figure 2.1: Overview of the piggyBac construct used for the transformation of the CPT fly lines. The series of affinity tags (StrepII, FLAG and yellow fluorescent protein - YFP) flanked by splice donor and acceptor sites is visible in the middle of the construct.



2.1.2 Strains

2.1.2.1 Genetics

The *D. melanogaster* strains that were used are part of the Cambridge Protein Trap (CPT) line collection (Ryder et al., 2007). These strains were genetically modified with a combined piggyBac (Cary et al., 1989) and P-element transposon strategy, using an exon encoding a series of affinity tags (StrepII, FLAG and yellow fluorescent protein - YFP) flanked by splice donor and acceptor sites (Morin et al., 2001). This allowed the expression of a tagged protein product when inserted in correct orientation and reading frame. Figure 2.1 shows an illustration of the construct (top) and how it allows tagging of the expressed proteins.

This effort generated fly lines which were subsequently used to document and annotate the expression of *Drosophila* genes (Ryder et al., 2009) and also the in-vivo analysis of interactomes by parallel affinity capture (iPAC) (Rees et al., 2011).

2.1.2.2 Bait choice

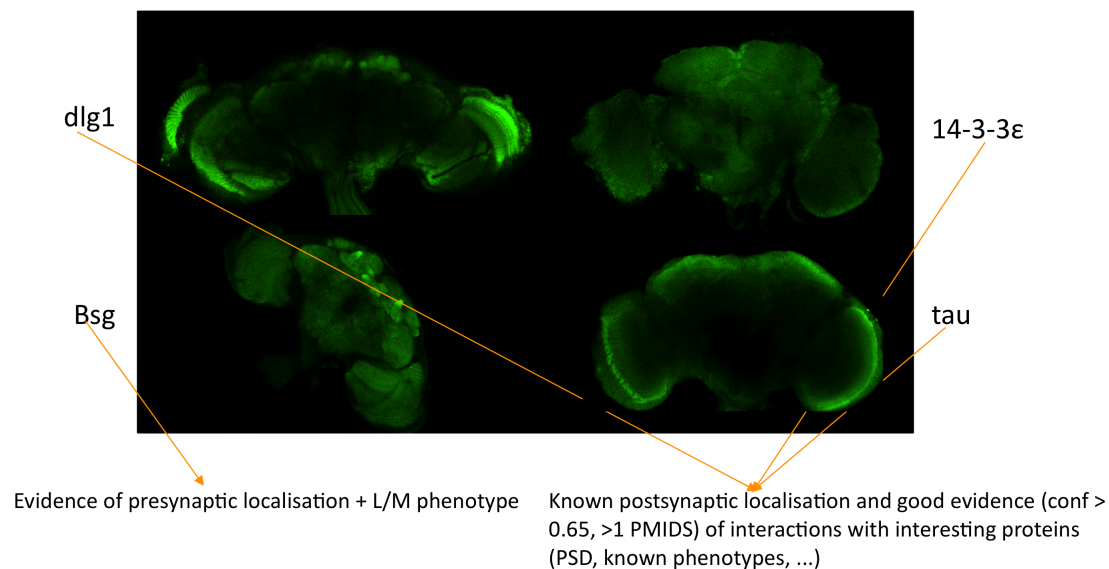
In order to isolate fPSD complexes we chose four different CTP strains, each expressing a tagged protein which we considered a candidate bait for affinity purification of the fPSD complexes. These proteins were *dlg1* (CPTI-0000207), *Bsg* (CPTI-0000062), *14-3-3 ϵ* (CPTI-0000034) and *tau* (CPTI-0000194). The entries in FlyProt database (Ryder et al., 2009) (<http://www.flyprot.org/>), which is an annotation portal for the CPT lines, shows that in all cases the insertions are included in all known splice variants.

One reason these bait proteins were chosen was their expression pattern in areas of the brain associated with learning and memory. We retrieved data from the Brain-Trap database (Knowles-Barley et al., 2010) (<http://fruitfly.inf.ed.ac.uk/braintrap/> - at the time only partial annotation was available), which integrates expression data for the subset of CTP lines that express the tagged protein in the brain. All of the above proteins are expressed in the brain (Figure 2.2). More specifically the data suggests that the above proteins are expressed in the ellipsoid body (*14-3-3 ϵ* , *Bsg*, *dlg1*), calyx (*dlg1*), stalk (*14-3-3 ϵ*), alpha-lobe (*14-3-3 ϵ*), and gamma-lobe (*14-3-3 ϵ* , *Bsg*). The bait proteins are also expressed in the whole (*14-3-3 ϵ* , *Bsg*, *dlg1*) of the mushroom body, the fan-shaped body (*14-3-3 ϵ* , *Bsg*, *dlg1*), the optic neuropil (*dlg1*, *tau*), the protocerebral bridge (*Bsg*, *dlg1*), the cerebral cortex (*14-3-3 ϵ* , *Bsg*, *tau*), the subesophageal ganglion (*dlg1*), the protocerebrum (*dlg1*, *tau*), the deutocerebrum (*dlg1*, *tau*), the nodulus (*dlg1*), lamina (*Bsg*, *tau*), antennal lobe (*14-3-3 ϵ* , *dlg1*), and optic (*Bsg*, *dlg1*) lobes. The aforementioned data are a good indicator of the involvement of these proteins in PSD processes, since the fly's mushroom bodies are associated with learning and olfactory memory (Strausfeld et al., 1998, Yu et al., 2006) while the fan-shaped body has been associated with visual memory (Liu et al., 2006a). Additionally, the ellipsoid body has been implicated in visual pattern memory via a protein kinase G pathway (Wang et al., 2008), NMDA receptor-dependent long-term memory consolidation via R2/R4m neurons (Wu et al., 2007) and spatial orientation memory via the R3/R4d

neurons (Neuser et al., 2008).

Another reason for choosing these baits was the central position of their mouse homologs in the corresponding PSD models. Dlg1 and 14-3-3 ϵ are known central component of the mammalian PSD. Additionally, tau, the homolog of the human tau protein (Mapt), is involved in pathways associated with Alzheimer's and other neurodegenerative diseases (Wittmann et al., 2001, Jackson et al., 2002, Scherzer-Attali et al., 2010, Ittner et al., 2010). Furthermore, Bsg is known to be required pre and post-synaptically in NMJs (Besse et al., 2007) and also its mouse homolog has been associated with abnormal behavioural responses (Igakura et al., 1996). Finally, we should note that an additional criterion to the bait choices along with the choice of specific CPT lines¹ was the quality of results obtained in preliminary experiments.

Figure 2.2: Expression of the 4 chosen bait proteins in the fly brain. Data from the Braintrap database Knowles-Barley et al. (2010). All proteins show localised expression in brain areas associated with *Drosophila* learning and memory. Also dlg1, 14-3-3 ϵ and tau have known postsynaptic localisation and known interactions with fPSD proteins, while Bsg has a learning and memory phenotype (see text for references)



¹There were more than one insertion lines per gene.

2.2 Methods

2.2.1 Affinity purification of complexes

2.2.1.1 Overview

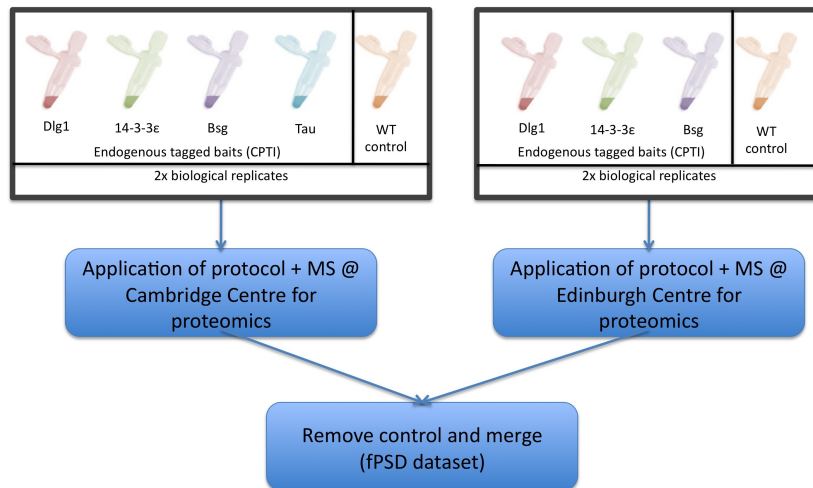
The proteomic isolation and purification of the fPSD complexes was an adaptation of the affinity purification protocol, described in Rees et al. (2011), optimised for fly head samples using the bait proteins discussed earlier. For each bait protein extract was mixed with resin slurry, which resulted in the immobilisation of the bait protein via resin - tag (StrepII) or antibody-tag (YFP) interactions. All experiments were performed in parallel with a wild type (WT) untagged control (*csw*– flies). Two separate sets of affinity purifications were performed and an overview of the workflow is shown in Figure 2.3. The first purification (Figure 2.3, left) involved four protein baits, using Strep-Tactin sepharose resin (IBA) and anti-YFP (Clontech) resins, against the Strep II and the YFP tag respectively, and was done in duplicate. Data from the YFP and Strep-Tactin purifications were merged after discussions with the proteomics facility. This affinity purification was performed in the Cambridge Centre from Proteomics and the resulting data will be hereon referred to as CCP data. The second purification (Figure 2.3, right) was performed using 3 out of 4 baits (tau was left out since it did not provide satisfying results in the first purification) using only a Strep-Tactin resin, also in duplicate. This affinity purification was performed in the Edinburgh Centre from Proteomics and the resulting data will be hereon referred to as ECP data. Note that in this case we chose a single-step purification protocol.

Although a tandem affinity purification ensures lower contamination rates we preferred a single step approach for two reasons. First, it would be simpler since the tagged constructs had not been extensively studied and we had no data on how good the exposure of a second tag would be. Second, based on in house data and discussions with the Genes2Cognition consortium, a two step purification would be suitable for a more detailed study of sub-complexes rather than an attempt to map and reconstruct as

much of the fPSD as possible.

Each affinity purification experiment in the workflow was experimentally validated for the presence and enrichment of the bait protein before the samples were sent to be analysed with LC-MS. This validation was done by western blotting and probing against the YFP tag. Results of this validation can be found in Appendix A.1.1.

Figure 2.3: Overview of the complex purification and identification experimental design.



2.2.1.2 Head sample preparation

The fPSD protein complexes were isolated using affinity purification from fly head protein extract. For each transgenic or wild type *Drosophila* line, independent samples were prepared for each replicate. An approximate number of adult (7-10 days old) flies was collected in a skirtless 50mL Falcon tube (Greiner), up to the 15mL line. The tube was flash frozen by immersing in liquid N_2 for at least 45s. The tube was then shaken on a bentchtop vortex for 10 seconds. This results in the heads, wings and legs being separated from the bodies. The content was poured through 2 tandem sieves (Fisher Scientific, top $710\mu m$ mesh, bottom $425\mu m$ mesh) immersed in liquid nitrogen. The sieves were then shaken sideways in order to separate coarser (bodies) from finer (heads) particles. The top sieve retained the bodies, while the bottom sieve retained the heads and let the legs, wings and other debris fall through. The heads

retained by the lower sieve were transferred and weighted into eppendorf tubes and then frozen at -80°C for the preparation of the protein extract ($\sim 80\text{mg}$ needed). We decided against using brain protein extract because of the time bottleneck introduced by the need for dissections in order to obtain enough sample material. Additionally, individual fly brains are much harder to handle, compared to the bulk preparations of heads.

2.2.1.3 Protein extraction protocol

For the protein sample preparation the tubes were transferred from -80°C on dry ice. 80mg of heads were weighted in an eppendorf tube. 1mL of lysis buffer (LB) was added to the heads and the mixture was homogenised on ice using a pestle and grinder (40 sec). The samples were left to stand on ice for 5min and then the pellet was removed by centrifugation (10.000 rpm for 10 mins) at 4°C . This results in a separation of the pellet and the protein extract. After optimisation, we found that a total volume of 1mL gave good separation of protein extract and pellet. The supernatant (S10) was removed avoiding the fatty layer between supernatant and pellet. During the optimisation of the protocol we noticed that the presence of this fatty layer in the affinity purification caused lower final yields. The concentration of the SN was calculated using a Nanodrop (Thermo Scientific). Typical concentrations varied between 30 and 60 mg/ml. All samples were normalised to 30 mg/ml. Buffer recipes can be found on Table 2.1.

This protocol was developed by Jo Rees (Cambridge Centre for Proteomics) and optimised by Jo Rees and the author. It has to be noted that initially we used the lysis buffer recipe used in Emes et al. (2008) (LB2) but decided against it since the recipe and protocol above gave slightly higher yields of protein.

2.2.1.4 Affinity purification protocol

For the StrepII tag purifications 50 μ l pre-washed Strep-Tactin sepharose resin was added to 1 ml (30 mg) protein extract S10 and incubated at 4°C for 2 hours on a rotary mixer. Non binding (NB) material was removed by centrifugation (3.000rpm for 2 min), while a small sample was kept for western blotting and the resin washed three times in ice cold LB, with a sample from the first wash (W) kept for western blotting. StrepII tagged bait protein, with any associating proteins, was eluted twice with 50 μ l of 10 mM desthiobiotin (Sigma) in LB for 30 min at 4°C on a rotary mixer. The two eluates (E1 and E2) were combined (E) and any residual resin was removed by centrifugation at 4.000 for 2 min. For the YFP purifications we followed an identical protocol although elution was performed using YFP high affinity peptide (Jo Rees, unpublished, sequence AcDFKEDGNILGHKLEYNYNSH/NH₂, 100 μ g/mL) in LB. Final eluates were reduced in volume to approximately 20 μ L in a SpeedVac (Savant). Buffer recipes can be found on Table 2.1.

2.2.1.5 Protein identification using immunoblotting

Initially enrichment of the protein extract in bait protein was verified via comparison with the amount of protein in the NB, W and S10 samples by immunoblotting. 2 μ L of E sample were with 18 μ L LB and 5 μ L 5 x sample buffer (5xSB). On all other samples (S10, NB, W) 20 μ L were mixed with 5 μ L of 5xSB. Precast 4-20% SDS-PAGE gels (PAGEgel) were loaded and run at 160V for ~1h. Transfer was done at 300A of current on nitrocellulose sheets (Amersham). Blots were blocked in 5% milk in wash buffer (WB) solution for 1h, probed with 1:1000 anti-JL8 (Clontech) in WB overnight, washed 3 times (15min) with WB and probed with secondary antibody (anti-mouse-HRP conjugate, Amersham). The gels were developed after an additional 3 15min washes with WB, using ECL Plus reagents (Amersham). The anti-JL8 antibody recognises the YFP tag expressed in the bait proteins. Buffer recipes can be found on Table 2.1.

Table 2.1: Recipes for buffers used in protein extraction (LB, LB2) and protein identification using immunoblotting (5xSB, WB, RB, TB)

Buffer	Recipe
LB	Final concentrations: Tris pH7.5 50mM, NaCl 125mM, MgCl 1.5mM, EDTA 1mM, Glycerol 5%, Nonidet P-40 0,4%, Tween 20 0.1%, PMSF 1mM. Make 10mL and add 1 tablet PhoStop (Roche), 1 tablet CompleteMini protease inhibitor (Roche) and DTT on tip of spatula.
LB2	Final concentrations: Tris pH 7.4 50 mM, Nonidet P-40 0.5 %, NaF 50 mM, $ZnCl_2$ 20 mM, o-vanadate 1 mM, PMSF 1mM, aprotinin 2 μ g/ml and leupeptin 2 μ g/ml.
5xSB	Final concentrations: Tris pH7.5 200 mM, Bromophenol blue 0.05%, Glycerol 20%, SDS 4% fc, B-mercaptoethanol 5%, PMSF in MeOH 2mM, EDTA 2 mM. Add fresh DTT on tip of spatula before using.
WB	PBS with 0.1% Tween-20
RB	RunBlue Fast Run Buffer (Expedeon): Tris-MOPS-SDS.
TB	3.03g Tris, 14.4g Glycine and 200ml Methanol. Make up to 1L and adjust pH to 8.3

2.2.2 Mass spectrometry

The ECP affinity purification data was processed in the Edinburgh Centre for Proteomics by Juri Rappsilber's group. An LTQ-Orbitrap mass spectrometer (Thermo Electron) was coupled online to an Agilent 1100 binary nanopump and an HTC PAL autosampler (CTC). To prepare an analytical column with a self-assembled particle frit (Ishihama et al., 2002), C18 material (ReproSil-Pur C18-AQ 3 μ m; Dr. Maisch, GmbH) was packed into a spray emitter (75- μ m ID, 8- μ m opening, 70-mm length; New Objective) using an air-pressure pump (Proxeon Biosystems). Mobile phase A consisted of water, 5% acetonitrile, and 0.5% acetic acid; mobile phase B, consisted of acetonitrile and 0.5% acetic acid. The peptides were loaded onto the column at a flow rate of 0.7 μ L/min and eluted at a flow rate of 0.3 μ L/min according to the gradient 0% to 20% buffer B in 75 min and then to 80% B in 13 min for two hours run. FTMS spectra were recorded at 30,000 resolution and the six most intense peaks of the MS scan were selected in the ion trap for MS2, (normal scan, wideband activation, filling

7.5×10^5 ions for MS scan, 1.5×10^4 ions for MS2, maximum fill time 150 msec, dynamic exclusion for 60s sec). Raw files were processed using DTA SuperCharge to obtain the peak list. Searches were conducted using MASCOT (MatrixScience, version 2.2) against a Drosophila database. The search parameters were: MS accuracy, 6 ppm; MS/MS accuracy, 0.6 Da; enzyme, trypsin; allowed number of missed cleavages, 2; fixed modification, carbamidometylation on Cysteine and variable modification, oxidation on Methionine.

The CCP affinity purification data was processed in the Cambridge Centre for Proteomics by Kathryn Lilley's group. 5 or 10 μ l peptide was loaded onto a precolumn (Presearch) then concentrated peptides were subsequently loaded onto a PepMap C18 reverse phase, 75 mm i.d., 15 cm analytical column (LC Packings) and eluted using an Eksigent nano LC system at a flow rate of 300 nl/min attached to a LTQ Orbitrap (Thermo Electron). Gradient was applied to resolve and elute the peptides into the LTQ ion trap. The two 30 min washes with 85% and 65% ACN were adopted to reduce carryover of residual abundant peptides, such as Actin, that bind non-specifically due to their sticky nature. The Orbitrap was operated in data dependant mode, MS then 2x MS/MS with data dependent settings set to excluded contaminant masses with a dynamic exclusion of 0.3 Da. m/z values were selected based on the protein abundance across multiple samples, including controls, from the same purification batch and from previous assays. Resulting fragment masses (MS/MS) were searched using the MASCOT (version 2.2) search engine against an in house database comprising FlyBase D. melanogaster genome (version 5.9) plus the FASTA sequence for YFP to confirm the presence of the tagged protein. The search parameters were: enzyme, trypsin; allowed number of missed cleavages, 2; fixed modification, carbamidometylation on Cysteine and variable modification, oxidation on Methionine. The decoy database option, comprising a scrambled D. melanogaster database in silico digested that generates a similar number of the same sized peptides, was checked to automatically calculate the protein false discovery rate (FDR).

Chapter 3

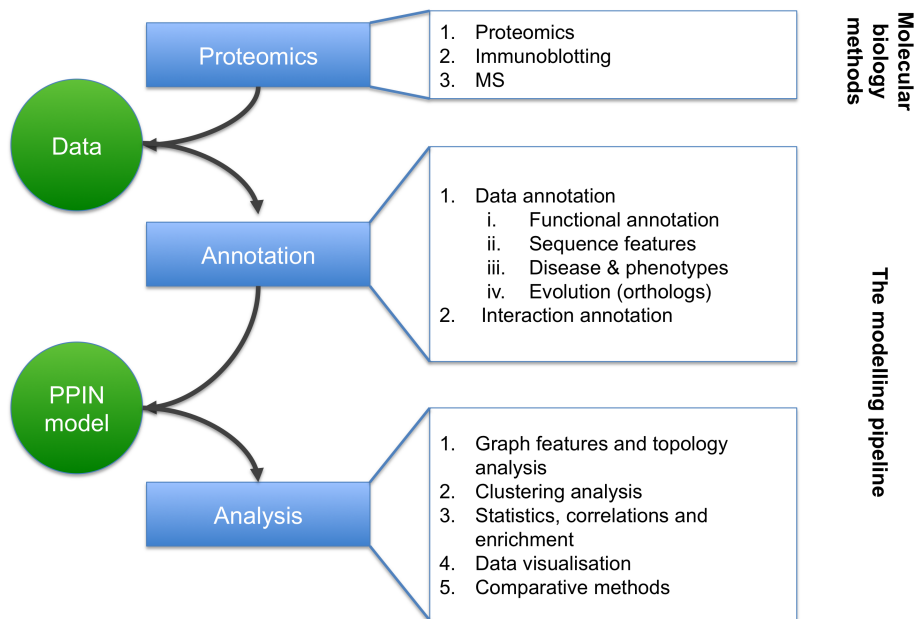
Computational Methods

3.1 Background

Models of protein-protein interaction networks (PPINs) are pivotal to systems biology approaches. Generation of valid PPIN models requires; a) proteomics data to provide a parts list and b) molecular interaction data to provide the connectivity of the model, with both datasets being of vital importance to the quality of the resulting model. One of the goals of this project was to design and implement a standardised workflow for the reconstruction of PSD PPIN models and their subsequent analysis in order to address and test the hypotheses. After background research and testing, we assembled a set of methods into a bioinformatics workflow, hereon referred to as the modelling pipeline, in order to reconstruct and analyse these models.

This modelling pipeline is in essence a series of annotation and computational analysis methods and software solutions that are applied in order to annotate the proteins and interactions of the model, reconstruct, visualise and finally analyse it (Figure 3.1). Although the contents of this chapter are in essence methods in the shape of algorithms, tools or data resources, a great deal of time was dedicated in compiling, reviewing, selecting and integrating the ones used in this work. For that reason we will not only mention these but give a very short background summary for each group.

Figure 3.1: Diagrammatic overview of the modelling pipeline.



3.2 Annotation methods

3.2.1 Data annotation

The first step when faced with a set of proteins is to collate information concerning their biological roles. Relevant features include known functional characteristics, phenotype and disease association, patterns of expression and evolution. Analysing these annotations can often yield insight into a complex even before constructing a network model.

3.2.1.1 Molecular function and pathways

Functional information can range from a broad classification of molecular function, to presence of functional domains and motifs, to involvement in biological processes and pathways. The most commonly used source of gene and protein annotations is the Gene Ontology database (Ashburner et al., 2000) (GO, <http://www.geneontology.org>), a hierarchical ontology of controlled terms organised into 3 independent domains: cel-

lular component (CC), biological process (BP), and molecular function (MF). With thousands of terms and up to 12 levels of specificity, GO can potentially provide quite detailed annotation. However, some terms are applied to only a handful of molecules, while others are so non-specific as to be virtually meaningless, and there can be extreme variability even between terms at the same level of the hierarchy. This is in part due to the perennial problem of literature bias - some genes have been subject to much greater research than others, which will be reflected in the reliability and depth of their annotation. Utilised in its entirety, GO imposes a large multiple testing burden on statistical analyses, and strategies for identifying a subset of meaningful, relatively independent terms must be devised (e.g. GO Slim subsets), in order to maximise the power of any test. More recent functional ontologies, such as PANTHER (Thomas et al., 2003, Mi et al., 2006) (<http://www.pantherdb.org/>), have tried to deal with some of the above issues. While PANTHER terms fully map to a subset of GO (and are backed up by the same types of evidence) they are much better balanced, being only 3 levels deep with more evenly sized gene sets at each level¹. Since GO is the standard in functional annotation but PANTHER is also superior in the aforementioned aspects, we have used both for the annotation of proteins in this project. More specifically PANTHER was used to classify proteins into functional families and subfamilies (examples in Table 3.1), while GO was used for annotation enrichment.

Protein domains - peptide sequences encoding structured, functional units - provide another important source of information (especially where gene-level data is poor). Sources of domain classifications include InterPro (Hunter et al., 2009) and PFAM (Finn et al., 2010). The UniProt database (Magrane and Consortium, 2011) annotates the amino acid sequence of each protein for known domains and other functional features such as binding and phosphorylation sites. Tools such as ELM (Gould et al., 2010) can also be used to identify short linear peptide motifs regulating subcellular tar-

¹While PANTHER provides accurate classifications for molecules of the nervous system, it does not come without flaws in that domain. Take the example of PANTHER's entry for *D. melanogaster*'s *dlg1* gene, which codes for one of the most central proteins in the fly PSD (and its homologs in the mouse and human PSDs). Its molecular function was "Unclassified" (as of 2010/02/15).

getting, physical interactions, phosphorylation and other processes. As these are short and often highly degenerate, they can easily occur by chance and many predicted sites will not be functional.

Table 3.1: Classification of the PSD genes and their products discussed in this thesis in functional families. This table shows examples of the types of families (and example subfamilies) of PSD proteins discussed throughout this work.

Family	Subfamilies (example)
Adaptor/ Regulatory	14-3-3, PDZ-domain containing scaffolders, non-PDZ-domain containing scaffolders
Cytoskeletal/ Structural/ Cell adhesion	Actin / Actin Related Proteins (ARP), Catenins, MAPs, Myelin, Other Cell Adhesion Molecules, Other Cytoskeletal Proteins, Other signalling molecules, Spectrin, Tubulin, actinin
Enzymes	ATP synthases, Other Enzymes
G-protein signaling	G-proteins, Modulators
Kinases	Ser/Thr Kinases, Tyr Kinase
Phosphatases	Protein Phosphatases
Receptors/ Channels/ Transporters	ATP synthases, Ca ²⁺ -ATPases, Glutamate Receptors, Inward rectifying K ⁺ channel, Na ⁺ /K ⁺ -ATPases, Other Channels and Receptors, Other signalling molecules, Transporters, Voltage-dependent anion channels, Voltage-gated K ⁺ -channel
Signalling molecules and Enzymes	Heat shock / Chaperones / Chaperonins, Mitochondrial Enzymes, NADH-Ubiquinone Oxidoreductase, Other Enzymes, Other signalling molecules, Translation
Transcription/ Translation	Ribosomal Proteins, Transcription, Transcription Elements
Unclassified	Other transmembrane, Uncharacterised / novel
Vesicular/ Trafficking/ Transport	Clathrin, Modulators, Motor Proteins, Other Enzymes, Other signalling molecules, Other transport, Synaptic vesicle

Information on involvement in biological processes and 'pathways' can be obtained from GO, PANTHER, KEGG (Kanehisa et al., 2010) and Reactome (Matthews et al., 2009). However, with the exception of well studied metabolic processes (and even here novel observations are still being made), most functional pathways are still poorly defined.

None of the annotation resources mentioned above captures the entire literature and the overlap between similar terms in different ontologies can be surprisingly low in some cases. Although all annotations can be improved by manual curation, this is a time-consuming process best reserved for highly focussed studies, such as refining the results of analyses based on one of the comprehensive ontologies. Evidence codes summarising the type of information linking an annotation to a gene (e.g. as supplied by GO) can be very useful for simple filtering without recourse to manual checking of references. Since each gene can have multiple functional annotations, these may capture pleiotropic effects in diverse cell-types that in some cases are misleading (Inlow and Restifo, 2004). On the other hand, this can be hard to disentangle from genes which truly have multiple functions - some of which were first observed in one cell-type, some in another - and whose disruption may have a more widespread effect on a complex than disruption of a highly specialised, single-function gene. The question of literature bias must always be kept in mind, especially when analysing the overlap between annotations. The fact that a gene has been extensively studied in one context (e.g. synaptic signalling) may make it more likely to have been studied in another (e.g. as a candidate gene for schizophrenia), making annotations based on these studies non-independent.

3.2.1.2 Diseases and phenotypes

Data covering involvement of genes in human Mendelian disorders is collated in the Online Mendelian Inheritance in Man (OMIM) database (McKusick and Amberger, 1994, McKusick, 2007) (<http://www.ncbi.nlm.nih.gov/omim/>). OMIM also covers complex disorders, but these are better dealt with using other resources. A number of recent resources, including the Genetic Association Database Becker et al. (2004), Alzgene (Bertram et al., 2007) (<http://www.szgene.org>), PDgene (Yu et al., 2008c) (<http://www.pdgene.org>), and SZgene (Allen et al., 2008) (<http://www.szgene.org>), collate the results of genetic association studies for a particular disorder, perform meta-

analyses and provide ranked lists of associated genes or loci, making them a useful gateway into the field for non-experts. While it is possible to use the top hits from these or other lists as a disease annotation, this does throw away a lot of information and can introduce the problem of literature bias. With many genome-wide studies of SNPs and CNVs now available through dbGAP (Wooten and Huggins, 2011) (<http://www.ncbi.nlm.nih.gov/gap>) it makes much more sense to use individual studies in their entirety, with access to multiple datasets, allowing replication of results.

Genetic and pharmacological manipulations of model organisms have uncovered developmental, physiological and behavioural roles for many genes. A substantial amount of this phenotypic data is available from organism-specific databases such as Mouse Genome Informatics (Blake et al., 2002, Eppig et al., 2005; 2007, Bult et al., 2008) (MGI, <http://www.informatics.jax.org/>), Rat Genome Database (Shimoyama et al., 2011) (RGD, <http://rgd.mcw.edu>), Flybase (Gelbart et al., 1997) (<http://www.flybase.org>) and Wormbase (Stein et al., 2001) (<http://www.wormbase.org>). MGI and RGD both use the Mammalian Phenotype ontology (Smith and Eppig, 2009), allowing them to be easily combined if necessary. As with virtually all resources, the databases listed above do not encapsulate the entire literature and can always be supplemented by text mining. As noted earlier, the representation of genes in the literature may be biased. Depending on the source, phenotype annotations can be based on a diverse array of evidence, and the ability to filter out certain types of studies is essential for ensuring data quality. We would strongly recommend separating single gene data from multi-gene manipulations, and suggest that the relevance of transgenic studies be carefully considered. Some additional resources have appeared in the literature, e.g. PhenomicDB (Kahraman et al., 2005, Groth et al., 2007) (<http://www.phenomicdb.de/>) that integrate data from multiple sources (including the databases listed above). These may also be of use, although care must still be taken to ensure data quality.

3.2.1.3 Gene and protein expression

Complexes are typically isolated from either whole brain preparations, or in some cases a particular anatomical region. Analysis of expression data can indicate the ways in which complex composition and function may vary across brain regions, cell-types and developmental stages. Expression can be measured in multiple ways. Western blots can detect the presence of a protein in a tissue, while immunohistochemistry allows its localisation to be determined as well. These methods are semi-quantitative at best, as is in-situ hybridisation which highlights mRNA localisation in a tissue. Microarrays and RNA sequencing both measure RNA abundance in a quantitative manner. In addition to producing more detailed information, exon arrays give more accurate gene-level expression measurements compared to older microarray chips, with the emerging RNAseq technology providing the cleanest data to date. When drawing upon multiple types of expression data it must be kept in mind that neurons can span multiple anatomical regions, with their cell-body (and most of the RNA) in one and axons and dendrites (and many proteins) extending into others. Useful resources include the MGI gene expression database (Ringwald et al., 1997; 2001, Smith et al., 2007b, Finger et al., 2011), the Brain Gene Expression Map (BGEM) (Magdaleno et al., 2006) and GENSAT (Heintz, 2004) (<http://www.gensat.org/index.html>) for mouse, and the Allen Brain Atlas (Jones et al., 2009) (<http://www.brain-map.org/>) for both mouse and human. Individual high quality datasets (Doyle et al., 2008a) can also be identified in the literature. Many of these can be downloaded from the large public repositories ArrayExpress (Parkinson et al., 2009) and the Gene Expression Omnibus (Barrett et al., 2009; 2011) (GEO, <http://www.ncbi.nlm.nih.gov/geo/>).

3.2.1.4 Evolution and orthology

Nervous systems and the behaviour repertoires they support vary tremendously in complexity. By identifying and comparing orthologous PSD genes across organisms we can start investigating the evolutionary mechanisms behind PSD complexes. This

may in turn shed light on the relative importance of particular complexes or classes of molecules in the behavioural complexity of different species. More pragmatically, identifying orthologous genes allows annotations to be transferred from one species to another e.g. when wanting to investigate the relevance to human disorders of complexes characterised in rodents. The Ensembl Compara database (Vilella et al., 2009) (<http://www.ensembl.org/>) concentrates upon vertebrates, but also contains a number of invertebrate and unicellular species commonly used as model organisms. The InParanoid database (Berglund et al., 2008, Ostlund et al., 2010) covers a more diverse range of organisms. Information retrieved from these databases will typically contain a large number of many-to-one and many-to-many mappings. These will need to be resolved, identifying the most closely related cross-species pair. An alternative approach to this is using MGI, which has pairwise one-to-one mappings between mouse, human and rat. A caveat of this type of precomputed annotation, when there is no manual sanity checking, is that if there is a misannotated ortholog, this mistake will “spread” since this misannotated ortholog is going to bring its homologs as orthologs of the query gene. The above has to be taken into account even when high quality resources, such as Compara, are used.

3.2.1.5 Implementation

Data annotation collection was implemented in PERL (<http://www.perl.org>). Implementation details are given on Tables 3.2 and 3.3. Note that also, since integrated disease associated SNP, GWAS and CNV resources were not available at the time we had to rely on database cross-references, text mining and manual curation for the annotation of these complexes.

3.2.1.6 Considerations

Various issues can arise during the annotation process. We have found that a typical problem in data annotation is keeping the protein and gene identifiers up-to-date.

Table 3.2: Data annotation implementation. All implementations in PERL using the CSV/XLS, OBO and XML parsers (CPAN)

Annotation	Source file	Comments
Database identifiers and gene/protein synonyms	UniPro XML and Ensembl Biomart query csv/xls (version 53)	UniPro has a synonyms field. Biomart should be queried for all database identifiers.
GO	Ensembl Biomart query exported csv/xls (version 53)	Ignore “Inferred from Electronic Annotation” (IEA). Parse the OBO file from GO and maintain the term hierarchy so over-counting is avoided.
PANTHER	PANTHER flat file (version 6)	The PANTHER GO-like ontology was eventually discontinued.

Table 3.3: Data annotation implementation. All implementations in PERL using the CSV/XLS parsers (CPAN).

Annotation	Source file	Comments
Functional families and subfamilies	PANTHER flat file	PANTHER family/subfamily classifications were mapped to the family/subfamily classifications and then were manually checked.
Protein domains	InterPro and PFAM flat files	InterPro and PFAM annotations were taken from the UniPro entry.
Evolution	Ensembl Biomart query exported csv/xls	Query Biomart for compara orthologs. Also include orthology type.
Disease & phenotypes	OMIM, G2C in-house association files	Most of the information gathered at this stage of annotations was also augmented with text mining data.

This type of analysis is done at a later stage, after the proteomics results have been obtained, and it is very common for some of the database IDs of the proteins or associated genes to have changed or become “stale” (outdated) resulting in a chain of misannotation events. A solution to this can be automated by running sanity check scripts on the dataset frequently and manually checking any inconsistencies. Also, choosing a database which uses identifiers that do not become “stale” often is suggested. We have found that MGI IDs rarely become “stale”.

Other issues that we have encountered involved the use of annotation resources

such as GO and PANTHER. The reader has to keep in mind that these annotations are based on various types of evidence which are denoted by an attached evidence code (<http://www.geneontology.org/GO.evidence.shtml>). Our practice, which we can suggest for similar tasks, is to choose a set of standards as to which evidence codes are deemed acceptable. These standards should be applied systematically throughout the annotation procedure. e.g. GO data with a general “Inferred from Electronic Annotation” (IEA) evidence code should usually be ignored or thoroughly checked before used. The same standards should be applied in cases of ambiguous annotation or similar issues that can only be resolved by a human annotator. Also, there are cases where an annotation will not fit the ontology format of GO or PANTHER since the user will want to have more control over parameters. An example of that would be differential expression of a gene in different cell types, where the user would want to control things like the threshold of the ratio of gene expression in different brain areas. These types of annotations have to be accommodated manually. Also, for compactness some annotation database files will only contain the lowest level terms for each gene, in which case it will be necessary to download the full ontology (OBO file) and assign all parent terms to avoid over-counting in correlation computations.

The process of annotation is lengthy, but rigorous systematic annotation can be assisted by partial automation. Ensembl offers the Biomart webservice which can be accessed programmatically (<http://www.ensembl.org/biomart/>) and automate the retrieval of such information. However we suggest, specially in cases of smaller datasets, that everything is manually checked. This manual curation has to be done using a set of rules, as to what is accepted and what not, and these rules have to be followed for all the annotation curation. As an extension to that reusing annotations from old data can save a lot of time and effort, so good archiving and regular updating of the central data repository is imperative.

As a final point, the reader should take into account that imbalances in literature affects most annotations. These imbalances range from bias towards the study of spe-

cific genes and their products, or specific contexts of the latter. Sources of imbalance, however, could be even more subtle and could e.g. have to do with biases in the experimental methods used to study specific gene products.

3.2.2 Interaction annotation

Once a list of constituent parts of a protein complex is obtained and annotated, the next step is gathering the connectivity information, in order to produce a PPIN model. Connectivity in the case of PPINs comes from binary interaction information. The following section describes the ways which interaction data may be obtained. There is a wide variety of data resources for protein-protein interactions ranging from single interaction studies to high throughput whole interactome studies. This subsection discusses methods for interaction annotation via direct data retrieval from these resources, while techniques for interaction annotation via text mining and manual literature curation will be discussed in section 3.2.3.

3.2.2.1 Experimental data resources

The constant improvement of protein complex affinity purification, mass spectrometry identification and other high throughput methods like Yeast Two-Hybrid (Y2H) screening (Young, 1998) and the mammalian protein interaction oriented LUMIER method (Barrios-Rodiles et al., 2005) have resulted in a great accumulation of protein-protein interaction data. Beyond the volume of research done on smaller complexes and interactions in a low throughput manner, as of today there are also a number of whole interactomes available, including organisms like yeast (Uetz et al., 2000, Schwikowski et al., 2000, Ho et al., 2002, Gavin et al., 2002; 2006, Krogan et al., 2006, Yu et al., 2008a), *C. elegans* (Walhout et al., 2000; 2002), *D. melanogaster* (Stuart et al., 2007), *H. pylori* (Rain et al., 2001) and human (Bouwmeester et al., 2004, Barrios-Rodiles et al., 2005, Rual et al., 2005, Ewing et al., 2007).

Several issues have been raised over the years with regard to Y2H screening, which

at the moment is the most high throughput of the experimental methods. The critique has focused on the high rate of false positives, analysed by Vidalain et al. (2004) as biological and technical false-positives and has caused dispute over the use of Y2H screening data in manually curated interaction models unless there is other supporting evidence. The first category includes interactions that occur in yeast cells, but do not occur in vivo in the organism of study, because there is no way to simulate differential gene expression and protein localisation. The only way to eliminate these is by obtaining this type of information regarding the studied proteins, which is not always available. The second category of technical false positives includes protein interactions that are identified in Y2H screens due to technical limitations of the system (e.g. auto-activation of reporter domains). Various approaches and frameworks have been proposed to minimise the false positives. These involve changes in the method itself (Vidalain et al., 2004), the use of combined results obtained by other methods (Mering et al., 2002), or the use of statistical methods in combination with functional annotation, in order to estimate the quality parameters of a Y2H screening experiment (Venkatesan et al., 2009). These high throughput based approaches are usually available as entries in protein interaction databases and are currently a major source for protein interaction data retrieval. However, in order to avoid the aforementioned caveats as much as possible, we manually curated the entries utilising specific standards (see subsection 3.2.3.7).

3.2.2.2 Databases

Technology allows us to take advantage of the accumulation of data coming from low and high throughput methods by organising it in databases. Because of the various different experimental approaches and data sources there is a lot of variety in the available data as well. This variety stems from the methods, different species, types of interactions (binding, phosphorylation etc), dataset quality, and confidence. Although several protein interaction databases are publicly available (see <http://ppi.fli->

leibniz.de/jcb_ppi_databases.html), we will focus on databases that include mammalian (or in the case of the fPSD complexes discussed in Chapter 6, also invertebrate) data. Databases have two major focuses, either being central protein complex repositories or curated databases of protein interactions focused on a specific set of organisms or type of interaction. An example of the former is IntAct (Hermjakob et al., 2004) (www.ebi.ac.uk/intact/), which is the one of the central repositories for protein interactions. It is managed by the European Bioinformatics Institute (EBI) and contains a mixture of literature curated entries and data submissions. The other category of databases include entries that usually come from manual or semi-automated curation of the literature or collections of high throughput interaction screening experiments. Databases that were used in the interaction annotation for this project are given in Table 3.4. UniHi is a very comprehensive database of the computational, predicted and experiment-based human protein interactions and has been extensively used in our workflows. It is based on merging different whole interactome maps from different data sources, including BioGrid, IntAct and DIP among others as well as Y2H screening data.

Table 3.4: Protein interaction databases used for the reconstruction of PSD PPIs.

Database	URL	Reference
BioGRID	http://www.thebiogrid.org/	Stark et al., 2006
DIP	http://dip.doe-mbi.ucla.edu/dip/	Salwinski et al., 2004
HOMOMINT	http://mint.bio.uniroma2.it/HomoMINT/	Persico et al., 2005
HPRD	http://www.hprd.org/	Peri et al., 2003
MINT	http://mint.bio.uniroma2.it/mint/	Ceol et al., 2010, Cesareni et al., 2008
UniHi	http://www.mdc-berlin.de/unihi	Chatr-aryamontri et al., 2007; 2008

Parsing data from these resources can (usually) be done automatically - when the data is available as for download or bulk searches. This is done by mapping gene IDs to the database's internal IDs and then retrieving all the relevant interactions. In some cases, where it is possible, setting confidence cut-offs is very useful since many

of the interactions have a low confidence score. Recently there was a very useful addition to the toolkit for protein interaction retrieval, namely the release of PSICQUIC (Aranda et al., 2011), a common interaction retrieval interface. Although currently not all databases are integrated in the system, when that becomes the case we speculate that PSICQUIC will become an indispensable tool.

3.2.2.3 Homology data

When two pairs of ortholog proteins from different species interact they are referred to as interolog pairs. Interolog prediction is a good way of inferring interactions but is also a thorny subject when only using the sequence homology as a criterion of similarity between interacting pairs. There are cases of very big length and sequence differences between orthologues in distant lineages, e.g. the NR2 subunit of the NMDA receptor in mice and flies (Ryan et al., 2008) and in these cases not all interolog interactions might take place. Interolog data should be used carefully and ideally filtered using a confidence score based on homology, correlation of gene expression, or functional annotations. An example of this approach has been implemented in DroID (Yu et al., 2008b), a database of interactions for *D. melanogaster*. Recently Gallone et al. (2011) developed a method for automated scoring of interolog-based protein interactions which, in the case of invertebrate complexes such as the fly PSD, will automate part of this process.

3.2.2.4 Implementation

For the part of the interaction data retrieved from databases, custom PERL scripts were written to parse flat files from databases in Table 3.4. Information retrieved included type of interaction, supporting pubmed ID and annotation type (if automatically or manually sourced from the paper). UniHi flat files were provided by Prof. Erich Wanker. The set of interactions used for the NRC/MASC model was also used as a source file. As mentioned in the next paragraphs, all evidence was manually checked.

3.2.2.5 Considerations

When using protein interaction databases it is important to consider the following factors. First of all a supporting pubmed ID must be accompanying every interaction. Second the protein interaction must be a direct interaction and not product of a spoke model expansion, which sometimes are included in the raw data file (e.g. IntAct). The most important factor is to consider which databases are manually curated and which include annotations from indirect protein interaction inference (e.g. co-citation of gene names in Pubmed abstracts). The latter type should be avoided.

3.2.3 Mining data from the literature

3.2.3.1 Text mining

While protein interaction and annotation databases like the ones mentioned in the previous paragraphs rapidly provide data for the models, their coverage is far from complete. Data is also buried within a corpus of hundreds of thousands of scientific papers in the existing literature. The volume of this corpus along with other issues (e.g. ambiguous terms and non-machine readable formats) makes the application of text-mining methods for information extraction imperative (Policies et al., 2008).

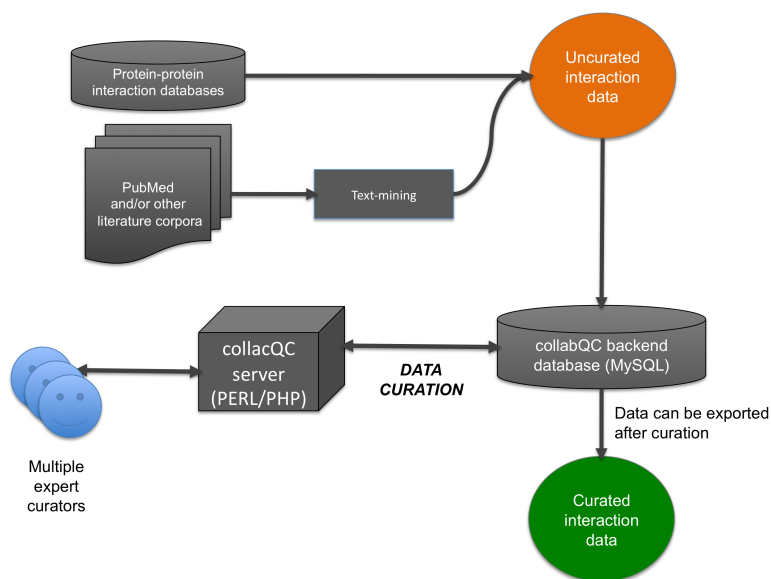
3.2.3.2 Corpus compilation and indexing

For biomedical text mining applications, local mirrors of PubMed, including all the meta-information are the standard type of corpus². The corpus has to be indexed in order to optimise the search of terms. This can be achieved with freely available software like Lucene (<http://lucene.apache.org/>). Additional application of natural language processing (NLP) tools can maximise the efficiency of text mining. These include pre-processing of the corpus with for the purposes of: tokenisation and sentence detection, part of speech (POS) tagging and abbreviation detection and named entity

²It has also been empirically found that figure legends provide an excellent additional corpus as well, although correctly extracting them from the papers still poses a challenge.

recognition (NER). In NER parts of the text referring to biological entities are tagged using classifiers trained for the biomedical domain, e.g. in Alex et al. (2007), identifying terms that represent protein names or interaction terms. Currently there are freely available tools like the ones offered by NaCTeM (<http://www.nactem.ac.uk>) which can be used for all the NLP steps. This collection of tools could perform most of the text tokenisation and POS tagging. The aforementioned, used in combination with tools like Biothesaurus (Liu et al., 2006b) or Biotagger-GM (Torii et al., 2009) for the NER can potentially provide higher efficiency.

Figure 3.2: Potential protein-protein interaction curation procedure with collabQC.



3.2.3.3 Queries and query expansion

Queries are made against the corpus, in order to retrieve abstracts containing relevant information. Obtaining the right keywords to retrieve relevant abstracts is another challenge. This is mainly because of the number of potential synonyms every gene or protein might have as well as their potential spelling variations. For some common terms or initials lists of variations can be manually compiled and combined with a list of synonyms for each protein and its associated gene name. Acquiring the gene and protein synonyms can be done by mining public database entries' "name", "synonyms"

and “gene name” fields or using some specialised service like BioMinT (Pillet et al., 2005) (<http://biomint.pharmadm.com/>). Query expansion is a key step in compiling the list of synonyms. The simplest form of query expansion would be to try all possible combinations of spelling taking things like special characters (e.g. spaces, hyphens) and roman numerals into account. Although pre-compiled thesauruses which can be very useful (Sasaki et al., 2010), there is no standalone solution for this problem. We have found that collecting all available synonyms for a gene and protein entity and applying rules like the one mentioned above will generate rich enough lists. An alternative to this is using the EFetch utility, as mentioned above, which uses PubMed’s built-in query expansion system although that was recently found to reduce precision in some cases (Schuemie et al., 2010). More recent innovations addressing this problem are based on the query itself, expanding it based on the biological context of the gene or protein, like e.g. QuExT (Matos et al., 2010).

3.2.3.4 Overview of the text mining process

Once the corpus is prepared and the list of queried terms is compiled it is easy to programmatically automate a process of submitting queries with the keywords of interest. If attempting to mine for protein annotations these queries will include all the synonyms of a specific gene or gene product (after query expansion) in all combinations with all the annotation terms of interest (e.g. [“gene name” OR “synonym”] AND “disease”). If mining for protein interactions, these queries will include all combinations of synonyms for a given pair of potentially interacting proteins. This will generate lists of results or “hits” which can then be prioritised and curated.

3.2.3.5 Implementation

Mining the literature for annotations of proteins Although annotation retrieval can be more difficult than protein interaction retrieval due to the more evasive nature of the types of annotations, annotating for specific features can be easier once

the keywords are defined. In the case of the PSD complexes annotation we initially used a combination of Lucene, Rainbow (<http://www.cs.cmu.edu/~mccallum/>) and Weka (Hall et al., 2009) for text classification. These annotations were performed for the NRC/MASC complex by Pocklington et al and were available as an in-house G2CDB dataset. At a later stage, in order to annotate novel mouse data and update the NRC/MASC annotations a custom script calling the PERL interface of the EFetch utility (<http://eutils.ncbi.nlm.nih.gov/entrez/eutils>) provided by PubMed was used. EFetch utilises PubMed's search function but can be used in programmatic workflows. The implementation included the gene name and synonyms along with the disease or phenotype term of interest. The abstracts retrieved from this process were manually checked.

Interaction retrieval In order to reconstruct mouse PSD complexes we used the TXM pipeline (Alex et al., 2008), a system initially developed for the BioCreativeII challenge (<http://www.biocreative.org/>). This pipeline queries an indexed corpus for co-citation of all possible combination of gene names in a complex (including synonyms). Further analysis of the co-citation hits utilises information within the structure of the text, in order to compute a confidence for each hit, based among others on the proximity of the references of the two potentially interacting entities in the text or the presence of interaction-associated terms.

3.2.3.6 Interaction curation using collabQC

While text-mining methods have become more accurate the results still contain false positive hits and have to be manually quality controlled. This manual quality control (curation) of each hit is done by human experts and is the single major bottleneck of the modelling pipeline.

Given the size of the text-mining task and volume of resulting data including hits and the data retrieved from databases, the procedure of quality control needs to be streamlined. After noticing that there is a lack of freely available software to assist with

that, we developed our own solution. CollabQC is a server-side application designed to organize and assist with the collaborative quality control of text-mining results for binary protein interactions, allowing multiple experts (curators) to annotate each potential interaction hit. Figure 3.2 illustrates an overview of how collabQC works as part of the modelling pipeline.

The design of collabQC was based on the need for a software solution that offers an intuitive, web-based, curation interface (Figure 3.3) and a database manipulation backend to store and manage the text-mining results. As a server-side application it is designed to provide an easy way to allow curators to annotate the text-mining results. The installation is simple and the software is optimized for large datasets. Most importantly, the curation procedure is done through an intuitive interface and that allows going through a significant volume of results in fewer man-hours.

Multiple experts can log into the server. Once they select one of the datasets they can view various grouped lists and start annotating hits. The grouping and sorting of the text-mining results is based on the protein names, PubMed IDs of the papers or likelihood of the hit (ranking), as provided by the text-mining methods. As such, these lists cover all the hits for a pair of interacting proteins or all hits found in a specific paper. The curator can go through and rapidly annotate the individual hits. Annotation of the individual hits is done on the curation page. This page is compact and informative. It includes the abstract of the paper in which the text-mining method found potential evidence of interaction and an annotation form. External protein interaction database entries, if found, are cross-linked from IntAct. The original query terms for the protein names and their synonyms are colour coded and highlighted in the text to aid the user to find the appropriate text and decide on the annotation. Next, the hit is annotated as a true positive or false positive. If the curator classifies the hit as a true positive, the type of interaction may be also annotated using a protein interaction term from the open biomedical ontology (OBO) (Smith et al., 2007a).

CollabQC requires apache server with PHP 5, MySQL 5 and PERL > 5. It will run

on standard Linux and Unix distributions. On the client side, the requirements are just a JavaScript enabled web browser. CollabQC is available under a GPL license from <http://fruitfly.inf.ed.ac.uk/~lzografos/cqc/>.

3.2.3.7 Considerations

Text-mining tools and methods keep getting more accurate, but the results always contain false positive hits and have to be manually quality-controlled. Furthermore, it is good practice to re-check the evidence supporting interactions retrieved from databases using the same criteria as in the text mining result check. This curation is a form of manual quality control, which is performed by reading the abstract of the paper linked as supporting evidence and verifying the interaction between the proteins along with the experimental evidence provided. This manual checking of individual papers, although in 90% of the cases just the abstract provides the related information, is the single major bottleneck of the modelling pipeline. Given the size of the text-mining task and volume of resulting hits, combined with the data retrieved from databases, the procedure of quality control needs to be streamlined.

When multiple curators collaborate, curation standards are imperative. These standards dictate if a potential physical interaction is accepted as a true positive or not and should be followed as an intact set of instructions throughout the data curation. The common curation standard used for the PSD complexes includes the following rules: 1) Clear mention of physical interaction in the abstract or full text (any evidence except “prior experimental knowledge”, unless backed up by experimental evidence). 2) Do not accept as true positive if the only supporting data only comes from co-localisation, protein complex pulldown, interolog pairs from distant lineages or Y2H with no other supporting evidence³. Regarding the last point we refer the reader to the relevant paragraphs of subsection 3.2.2.1. We advise that if Y2H data is used then it should be coming from datasets that have been thoroughly reviewed and shown to consider and

³At the time of this study the Y2H data quality dispute was at its peak.

tackle the caveat of false positives successfully.

3.3 Analysis methods

Up to this point the modules of the pipeline discussed have to do with annotating the nodes and edges of the biological network. More specifically in the PPIN case, annotating proteins and their interactions. The premise of network biology is that we can integrate the data in a network model (e.g. a PPIN) and abstract this network to a graph. The term graph is used as the mathematical equivalent of the network. Application of the modelling pipeline up to this point would result in a “coloured” graph. The term coloured in graph theory refers to specific attributes (colours) of the nodes and edges. Attributes can be of any type, i.e. gene name, family, subfamily, disease correlation or confidence for an interaction etc. Graphs are useful data abstractions because they can be manipulated by a wide variety of algorithms with which we can ask questions about the topology, architecture and most importantly latent structures or patterns in the network that we couldn't see otherwise.

3.3.1 Biological network primer

3.3.1.1 Networks and graph measures

Before we discuss any specific algorithms or workflows it is critical to mention some principles and central concepts of graph theory that are the basis of any analysis performed on PPINs. The most important is the definition of a graph. A graph is an abstract representation of a set of objects, called nodes or vertices, where some pairs of the objects are connected by links, called edges. These links may or may not have a specific direction, resulting in directed and undirected graphs respectively (there are “mixed” graphs as well).

The formal definition of a graph G is: $G = \{N, E\}$, where N is a set of k nodes, $N = \{n_1, \dots, n_k\}$ and E is a set of l edges, $E = \{e_1, \dots, e_l\}$. Each edge e is defined as

a relation of incidence that connects two nodes from N . The graphs resulting from the PPINs usually 1) finite: i.e. have a finite set of nodes and edges, 2) unweighted, i.e. no special value is associated with each edge, e.g. an association coefficient representing the strength of the interaction and 3) undirected, since binding is an bidirectional process. Note that (2) is not true if interaction confidence data is used and (3) is not true when modification (e.g. phosphorylation) interactions are included.

In order to use graphs for computations, they have to be represented in a manipulatable form. One such form is the adjacency matrix, A . A is a symmetric matrix defined as:

$$A_{ij} \begin{cases} = 1 & \text{if there is an edge between } n_i \text{ and } n_j \\ = 0 & \text{otherwise} \end{cases}$$

Figure 3.4 shows an illustration of an example where the adjacency matrix in panel A would result to the network in panel B.

There are a number of ways to describe and summarise a network, its topology and overall architecture. First we will look at the methods that describe the global topology of a graph as well as some basic properties of the nodes. Graph metrics of this type that are used for biological networks are:

- Degree: the degree of node i , k_i is the number of edges connected to it. (Figure 3.4C)
- Distance: the distance d_{ij} between nodes i and j is the shortest path (counted in edges) between them (Figure 3.4D).
- Diameter: the network's diameter is the maximum possible path length between any two nodes in the network (Figure 3.4E). Formally, $D = \max\{d_{ij} | i, j \in N\}$.
- Clustering coefficient (local): this measure is calculated for each node and shows the degree to which the neighbors of a particular node are connected to each

other. It is defined as $C_i = 2e_i / (k_i(k_i - 1))$, where e_i are the number of edges between the k_i nodes that connect to node i .

Distance in networks is measured by the path length, which tells us how many edges we need to cross in order to travel between any two nodes. As there are many alternative paths between two nodes, we choose the shortest path, i.e the path with the smallest number of links between the selected nodes. Formally, the shortest path problem is the problem of finding a path between two nodes such that the sum of the weights of its constituent edges is minimized. In unweighted graphs like most PPINs all edge weights are 1. There are different algorithmic approaches that can solve some of the categories. Some of these algorithms are: Dijkstra's algorithm (Dijkstra, 1959), the Bellman-Ford algorithm (Bellman, 1958) and the A* search algorithm (Hart et al., 1968) among others.

The concept of shortest paths allows us to introduce the notion of betweenness in graphs. Betweenness is a measurable property of nodes and edges. More specifically there are two types of betweenness in a graph. The first type is node betweenness, which for a node l is defined as:

$$b_l^n = \sum_{ij} p_{ij}(l) / p_{ij} \quad (3.1)$$

where $p_{ij}(l)$ is the number of shortest paths between nodes i and j that go through node l and p_{ij} is the total number of shortest paths between nodes i and j . The second type is edge betweenness, which for a edge k is defined as:

$$b_k^e = \sum_{ij} p_{ij}(k) / p_{ij} \quad (3.2)$$

where $p_{ij}(k)$ is the number of shortest paths between nodes i and j that include edge k and p_{ij} is the total number of shortest paths between nodes i and j .

3.3.1.2 Implementation

The computation of the above measures was implemented in Matlab (Mathworks Inc) with additional use of the MatlabBGL (<https://github.com/dgleich/matlab-bgl>) library,

a port of the Graph C++ library from Boost (<http://www.boost.org/>).

3.3.2 Community structure in PPINs

3.3.2.1 Community structure

Studies on all types of networks have shown that one of their universal properties is community structure (Ravasz et al., 2002, Guimerà and Amaral, 2005, Lagomarsino et al., 2007). Community structure is the segregation of nodes into groups, called clusters or communities. The characteristic feature of network clusters is that the nodes form densely connected groups or sub-graphs with sparse connections between them. An illustration of this concept can be seen in Figure 3.5. Notice how the three communities in the orange, yellow and red circles have more intra-connections rather than inter-connections between them.

3.3.2.2 Clustering algorithms

The identification of clusters or “clustering” within networks is a well studied general graph theory problem where numerous solutions have been proposed over the years. Clustering in biological networks is similar to the problem of graph partitioning in computer science and hierarchical clustering in the social sciences. In this section we will focus on clustering solution presented in the domain and context of systems biology. Due to the very high number of algorithms available we refer the reader to an excellent recent review by Wang et al. (Wang et al., 2010) and in this section we will present some of the widely used algorithms based on the classification used to present them in the aforementioned review. The algorithms can be split into two major categories: graph-based and combination-based. The first category encompasses algorithms that are based solely on the structure of the graph and act independently of the annotation of the nodes and the second category encompasses algorithms that use such information.

Graph-based algorithms These algorithms are based on local search around dense sub-graphs, hierarchical clustering or parameter optimisation.

Local search algorithms: In this category the clusters are defined as densely connected sub-graphs of the main network. The density of a sub-graph is defined as $d = 2n_{edges}/(n_{nodes} - 1)n_{nodes}$ (Spirin and Mirny, 2003) and reaches its maximum of 1 in a sub-graph where every two nodes are connected by an edge. In this case the sub-graph is called a clique (for more rigorous definitions and analysis see Erdős and Szckeres, 1987). Enumerating cliques in a graph is an NP-complete problem. However, protein networks make the enumeration less difficult because of their sparseness. Various solutions have been proposed implementing clique based methods including supermagnetic clustering (SPC) with Monte Carlo (MC) optimisation (Spirin and Mirny, 2003) and a quasi clique based method (Bu et al., 2003) among others. A widely used algorithm of this type is the molecular complex detection (MCODE) algorithm (Bader and Hogue, 2003). Although these algorithms tackle the issue of missing edges - or unknown interactions - in proteins interaction networks they suffer from issues related to the topology of the graphs as illustrated by Altaf-Ul-Amin et al. (2006).

Hierarchical clustering algorithms: Hierarchical clustering algorithms can be either agglomerative or divisive, depending on whether they add or removes edges to or from the network. One type of divisive algorithms are clustering coefficient based algorithms. Clustering coefficient represents a more local view of the network centered around a node and has been used as the basis for algorithms proposed by Radicchi et al. (2004) and Li et al. (2008). Newman and Girvan (2004) have proposed a betweenness based divisive hierarchical clustering algorithm. Along similar lines, authors have proposed other divisive algorithms as well with variations on the distance measures. These include HCS (Przulj et al., 2004), which is based on the minimum cut heuristic rule for grouping nodes, i.e. a configuration that separates two groups of nodes with the minimum number of edges between them (Hartuv and Shamir, 2000) and UVCLUSTER (Arnau et al., 2005) which is based on shortest paths instead of

edge betweenness.

Parameter optimisation algorithms: From a machine learning perspective parameter optimisation is based on the definition of a cost function which is then minimized searching through different clustering configurations. Markov Cluster Algorithm (MCL) (Enright et al., 2002) is a widely used parameter optimisation algorithm, which uses random walks on the network and then computes all the transition probabilities between nodes.

Combination based algorithms Unlike graph-based algorithms, combination based algorithms are not solely based on the graph that the PPIN represents but also the properties of its nodes. By taking the latter into account, these algorithms reduce the effects of false positive or false negative interactions. Properties can include genomic data (Jiang and Keating, 2005, Zhang et al., 2008), structural features of the proteins (Dittrich et al., 2008), gene co-expression data (Jansen et al., 2002, Hanisch et al., 2002, Ideker et al., 2002, Segal et al., 2003, Cho et al., 2006, Cline et al., 2007, Maraziotis et al., 2007, Lu et al., 2006, Ulitsky and Shamir, 2009, Jung et al., 2008) and ontology annotations (Lubovac et al., 2006, Ulitsky and Shamir, 2009) . These properties are integrated into frameworks like AVID (Jiang and Keating, 2005), PSIMAP (Park et al., 2005) and MATISSE (Ulitsky and Shamir, 2009), which usually utilise graph-based algorithms in light of the node properties data. Although these methods are evolving constantly the fact that all the availability of data of all properties examined is not always guaranteed and that has to be taken into account.

Within the class of combination based algorithms there is the variation of ensemble frameworks which use combinations of clustering methods and integrate their results into a common consensus. This type of approach was first proposed by Asur et al. (2007), followed by Greene et al. (2008) and Simpson et al. (2010). Although still in development as an approach ensemble clustering shows promise, if the choice of parameters, like which basic clustering methods to use and how to build the consensus,

is done with care.

3.3.2.3 Implementation

We have chosen the Newman and Girvan (Newman and Girvan, 2004) algorithm for our approach because it is a simple and elegant algorithm that allows not only to group proteins in clusters, but also to tease out hierarchical structures via the use of the dendrogram. Also, as a non-heuristic algorithm that runs within tractable time for the dataset sizes in hand and, when tested, it produced biologically meaningful results for our models. Making a choice for a clustering algorithm out of the variety of those available was based on mostly on empirical testing. We chose this approach because there was no common reference data set to compare them on (the number of clusters in PPINs is unknown and pathway based data sets are too small). After considering a number of algorithms available, we applied them on the datasets and manually compared the results. What we noticed was that some algorithms always returned smaller size clusters (e.g. MCL, also later verified by Wang et al., 2012) or left some nodes ungrouped (e.g. MCODE). We also noticed that the configurations obtained by the Newman and Girvan algorithm were biologically meaningful (i.e. recapitulated many pathways known from the literature). In addition to these evidence there were some practical issues that led us to choose the aforementioned algorithm for our approach, namely: 1) it uses more “global” view of the network architecture (Tuji et al., 2007), 2) the potential of recomputing Q after small permutations of the final configuration (this is very handy when one wants to move one or two nodes around in a given configuration).

Implementation was done in Matlab using MatlabBGL and using edge betweenness as a metric. Edges with high betweenness tend to be edges where the flow of information converges. This is because according to equation 3.2 on page 67, more shortest paths pass from these edges, thus they are parts of the path of least resistance. An interesting property of these paths is that they usually tend to connect the segre-

gated clusters of the network, as more thoroughly discussed in subsection 3.3.3. The steps of the Newman and Girvan algorithm are: 1) Compute betweenness score for all edges in the network. 2) Find the edge of highest betweenness and remove it. 3) Recompute betweenness score for all edges in the network. 4) Repeat from 2

Up to this point the application of the algorithm would eventually start breaking up the network in sub-networks. If one knew *a priori* how many communities were in a network, the algorithm would be stopped when reaching that number. However, this is not the case in most practical applications. For that reason, the authors also define a cost function or modularity quantity, Q . If one considers a particular division of a network into k communities, we can define a $k \times k$, e matrix whose elements e_{ij} are the fractions of all edges in the network that connect nodes in k_i with nodes in k_j , considering all edges in the original network including the ones removed so far. The trace of the aforementioned matrix, $Tr e = \sum_i e_{ii}$, would in practice give the fraction of edges in the network that connect nodes from the same community. However, the trace is not a good indicator because there are cases where $Tr e = 1$, without that being the best configuration, e.g. if all nodes were in the same community. For that reason the authors define the row or column sums $a_i = \sum_j e_{ij}$, which represent the fraction of edges that connect to nodes in community i . In a network where all edges connect nodes without regard to which community they belong to, that would mean $e_{ij} = a_i a_j$. Thus Q is defined as:

$$Q = \sum_i (e_{ij} - a_i^2) = Tr e - ||e||^2$$

where $||e||$ is the sum of elements of e .

In practice Q measures the fraction of edges between nodes of the same community over the edges between nodes of different communities. Q is monitored as the algorithm progresses and once all edges have been removed we can trace back to the configuration that resulted to the maximum value of Q , which also is the optimal community structure. By definition Q is found to be between 0 and 1, with low values reflecting configurations that are no better than random. Empirically, in biological net-

works the value of Q lies between 0.3 and 0.7. After trials we found that it is better to implement this algorithm allowing multiple random initialisations (5 to 10 is feasible) in order to find the maximum Q .

This algorithm was in Matlab using MatlabBGL. It should be noted that this algorithm has heavy demands on computational resources, running in $O(e^2n)$ time on an arbitrary network with e edges and n nodes, or $O(n^3)$ on a sparse network, where $n \sim e$. This restricts the algorithm to networks of a few thousand nodes. For that reason there has been a later modification by Clauset et al. (2004) based on more sophisticated data structures.

3.3.2.4 Considerations

It should be mentioned that, depending on their definition of a cluster, algorithms can either identify overlapping or non-overlapping clusters. Overlapping clusters can potentially represent multiple complexes that a protein can belong to as a result of transient associations or differential expression of the components. Although the Newman & Girvan algorithm does not identify overlapping clusters, we strongly suggest that this should be considered in the future, especially with more protein stoichiometry data becoming available (see section 8.2 of Chapter 8). It should be taken into account that more recent studies have shown that Q has some weaknesses and could be substituted by other measures, e.g. surprise, S (Aldecoa and Marín, 2011). Also, clustering, independently of which method is being used, has to be performed with two problems taken into account. These are 1) the presence of false positive and negative interaction data, and 2) the fact that we, in reality, do not know the number of clusters an algorithm should produce. Although these cannot be eradicated, the former can be partially minimized with careful data curation. Regarding the latter, it has to be taken into account that the results of clustering algorithms represent a mathematical computation or optimisation and do not necessarily accurately reflect biological reality, since an algorithm will always generate an output. As a final point we have to mention that because of

its implementation the Newman and Girvan algorithm the best local maximum after a series of randomised restarts ⁴

3.3.3 Network topology features

A widely reported feature of many networks, including biological, are their “scale-free architectures” (Barabási and Albert, 1999, Albert et al., 2000, Jeong et al., 2001, Barabási et al., 2003). The scale-free architecture of biological networks implies that the great majority of nodes only have a few edges connecting them to other nodes. On the contrary, there are only a few nodes in the network that have many edges connecting them to other nodes (Barabási and Albert, 1999). The scale-free property of the architecture can be formally described by the degree distribution of nodes which approximates a power law, $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability of a node having degree k . The “scale-free” term comes from the dependence of the $\frac{P(ak)}{P(k)}$ ratio only from a . This is illustrated in Figure 3.6. Power law curve fitting has been addressed many times in the literature with different methods, including least squares fitting as the most popular approach. However, the choice of method is crucial since substantial inaccuracies might arise (Clauset et al., 2007). There are also two important relevant points worth mentioning. The first is that the intuitive assumption that a scale-free network’s sub-networks are scale-free does not always hold (Stumpf et al., 2005). The second point applies to datasets that give rise to scale-free networks. In these cases the reader must be aware that the scale-free architecture might be an artifact caused by regularities and biases in the selection of the dataset (Han et al., 2005) and does not reflect any biological importance. Taking the above issues and caveats into account, we decided not to draw on power law fitting and interpretations in the analyses of the models described here.

However, an observable consequence of the scale-free structure in biological net-

⁴The maximum is inherently local since every time betweenness is computed, more than one edges might have the same, in those cases a random removal choice is made. For this reason we do multiple restarts of the algorithm and chose the best solution out of these.

works is that only a few nodes have many interactions and this can lead to robustness against random mutations (Albert et al., 2000). Additionally, the scale-free architecture, by definition, implies that there are nodes in the network that have more connections than others. When this is interpreted in the context of community structure of networks, the notion of hub nodes emerges. Hub nodes are nodes that interact with many partners. For that reason and in contrast with non-hub nodes, hub nodes are extremely sensitive to targeted mutation (Jeong et al., 2001). Hub nodes are points of convergence and in some cases connect different functional modules of the network that appear in the form of communities. There are two types of hubs in biological networks: party hubs, where most of the interactions are simultaneous, and date hubs, where different interactions take place at different times (Han et al., 2004).

Although the high degree of nodes in biological networks could imply importance of a specific node, betweenness is often used as a measure as well. By definition, nodes and edges of high betweenness accumulate the majority of shortest paths passing through them. Therefore, seen by a “path of least resistance” principle these nodes become the central points controlling the direction of information passing in the network. Newman and Girvan argue that high betweenness implies nodes or edges that connect modules in the network (Newman and Girvan, 2004) and thus promoting crosstalk. Additionally, it has been found that clustering on betweenness results in clusters with similar functional annotation (Dunn et al., 2005). However, although these claims might seem intuitive, there has been a lack of direct supporting evidence. Yu et al. (2007) reported evidence by bioinformatic analysis of yeast PPINs. In their work, the authors defined high betweenness nodes as “bottleneck” nodes and dissected the types of nodes to hub bottleneck, non-hub bottleneck, hub non-bottleneck, non-hub non-bottleneck. They showed evidence that non-hub bottleneck nodes tend to be essential when involved in non-transient interactions, are rarely parts of large complexes, and are joints for crosstalk.

Taking all this into account, along with some contradicting evidence in the litera-

ture, like the findings of Goh et al. (2003), who showed a correlation between betweenness and degree in social networks, the reader has to be cautious about the use of these measures for such predictions. Yu et al. (2007) also argue that degree might be a better predictor for PPINs specifically, but that is heavily affected by missing interactions. In any case, even partial corroboration of predictions by previous biological knowledge is advised.

3.3.4 Statistical significance

3.3.4.1 Annotation significance

Once the constituent protein parts of a complex have been annotated with specific attributes regarding, e.g. their functional classification or involvement in a certain phenotype, questions arise regarding the relation of these attributes. For instance we can ask questions such as: is a functional family A significantly associated with module K ?. Here the notion of significance represents a number of co-occurrences that is higher than expected at random. The simplest approach is to use Fisher's exact test (Fisher, 1922), either one- or two-sided depending on whether the test is specifically for enrichment, or both enrichment and depletion.

3.3.4.2 Multiple testing

Multiple testing is a general statistical concept of considering multiple statistical inferences simultaneously. In a more specific context, when dealing with annotation data, one can use multiple replicates of randomized data point sets in order to assess the null hypothesis, e.g. of an annotation having a high count due to chance. An example of that is that if an annotation appears k times in a protein list of N proteins, one can randomly sample multiple N sized samples the proteome in hand and see how many of the proteins possess that annotation.

Another important application of multiple testing, summarised by Noble (2009) is

correcting p-values obtained from tests like the one mentioned above - particularly in the cases of larger datasets. The most popular multiple testing p-value correction methods are Bonferoni and false discovery rate (FDR) estimation (Benjamini and Hochberg, 1995). Application of the Bonferoni method means that a p-value p is accepted if $p < a/n$, where a is the confidence threshold and n the number of separate tests. This approach can sometimes be too strict so the FDR estimation method or the Benjamini-Hochberg variation of the FDR procedure can be used as an alternative. In the former the FDR is computed using the empirical distribution of the null hypothesis while the latter uses the p-values (see also Benjamini and Yekutieli, 2001).

3.3.4.3 Implementation

A number of tools have been developed to perform gene set enrichment analyses including DAVID (Dennis et al., 2003, Hosack et al., 2003), FuncAssociate (Berriz et al., 2003), MAPPFinder (Doniger et al., 2003), GoMiner (Zeeberg et al., 2003), GoSurfer (Zhong et al., 2004), FatiGO (Al-Shahrour et al., 2004) and BINGO (Maere et al., 2005). These tools used similar variants of Fisher's exact test or the Hypergeometric test and the Z-statistic. The Hypergeometric test is identical to the corresponding one-tailed version of Fisher's exact test. Additionally, other alternatives for computing annotation significance are available including Barnard's test (Barnard, 1945), Chi-square tests (e.g. Vêncio and Shmulevich, 2007, Prifti et al., 2008) and Bayesian methods (e.g. Antonov et al., 2008).

When analysing the PSD complexes discussed in this thesis we decided to re-implement the same in-house approach that we had experience with, as used by Pocklington et al. (2006), allowing us to easily incorporate it in our workflows (at the time no method offered satisfactory programmatic access).

Suppose now that for a set of N molecules, n_a and n_b possess the annotations a and b respectively. If these annotations were distributed randomly within the full set, the probability $h(n_{ab}, n_a, N, n_b)$ or $p(n_{ab})$, of a node possessing both annotations a and b ,

when the total number of nodes is N and n_a and n_b possess the annotations a and b respectively, is given by the following formula, according to Fisher's exact test.

$$p(n_{ab}) = h(n_{ab}, n_a, N, n_b) = \frac{n_a!(N - n_a)!n_b!(N - n_b)!}{[N!(n_a - n_{ab})!n_{ab}!(N - n_a - n_b + n_{ab})!(n_b - n_{ab})!]} \quad (3.3)$$

This probability distribution has one single maximum $p_{ml} = p(n_{ml})$, with n_{ml} the most likely overlap that would be occurring by chance ($n_{ml} + 1$ is equally likely depending on symmetry). If in our observed data μ_{ab} molecules possess both a and b annotations, we can evaluate the significance of deviation between μ_{ab} and n_{ml} by calculating the probability $P(\mu_{ab})$ of finding μ_{ab} molecules possessing both annotations, of finding an overlap as or less likely under the random distribution. In practice this means summing over all nodes n , where $p(n) \leq p(\mu_{ab})$, i.e.

$$P(\mu_{ab}) = \sum_n p(n) : p(n) \leq p(\mu_{ab}) \quad (3.4)$$

Using this definition in equation 3.4 above, for n_{ml} , $P(n_{ml}) = 1$, with both tails of the distribution contributing to P . This way P in 3.4 can be used to evaluate deviations in either direction of n_{ml} .

3.3.4.4 Considerations

We have to note that while we used our own implementation of Fisher's exact test, currently there are many tools that implement a range of methods, including the latter. Most of these tools are available online, with some offering an application programming interface (API) for programmatic access. Of these we highlight GSEA (Subramanian et al., 2005) and a later updated version of DAVID (Huang et al., 2009). Computations in the former are based on a variation of the Kolmogorov-Smirnov test while the latter uses a Fisher's exact test in combination with multiple testing p-value correction. A large and constantly updated list of these tools can be found on the GO website (<http://www.geneontology.org/GO.tools.shtml>).

When implementing these methods one has to take a few things into account. The

first is that sets of annotation variables are seldom independent and they range from mutually exclusive (e.g. the chromosomal location of the gene a protein is associated with) to redundant and overlapping. Another issue in the use of the method is our partial knowledge of the datasets. e.g. if a protein is not known to be associated with a disease mechanism, that could be either a true negative or the result of a bias of the experimental observations (false negative). For these reasons this method should be used carefully and its results should be trusted if appearing consistently and repeatedly.

3.3.5 Visualisation

3.3.5.1 Visualisation software

Since the models discussed in this work are descriptive and integrate annotations from different sources, this means that they carry a lot of information. Good and informative visualisation can convey abstract, complex information in intuitive ways. There are many software solutions for data visualisation. These vary from graph visualisation libraries like GraphViz (<http://www.graphviz.org/>), to more biological network oriented software. The former can be integrated in programmatic workflows, while the latter allow interaction with the model via a graphical user interface and a set of different functionalities and options.

Cytoscape (Shannon et al., 2003) was used for all the network visualisations, although alternatives such as BioLayout3D (Theocharidis et al., 2009) and BioLayout (Enright and Ouzounis, 2001) are available. In general, although tool choice is a matter of preference, performance issues might arise in bigger datasets. Cytoscape is an open source platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data. Features that make Cytoscape a useful tool are the support of many different formats and standards for input and output (e.g. plain text and XML based formats) as well as an API, which allows easy integration in all workflows, an

intuitive graphical interface, support of database web services for data import, efficient 2D visualisation with most visual parameters customisable according to attributes, and a variety of layout algorithms. Its plug-in architecture is probably the strongest feature of Cytoscape, because it allows the community to design and implement freely available add-ons.

3.3.5.2 Visualising networks

There can be many different approaches in visualising a PPIN. The typical approach is to draw the network in some informative manner, e.g. the nodes separated in communities or by subcellular location and highlight nodes with specific features, e.g. of a specific classification or associated with some disease. Sometimes, when the datasets are too big, it is more informative to visualise more collapsed versions of a network. Minimizing the clutter of visualisation by making the information more compact without reducing it can give a quick overview of a dataset, or even provide a collapsed representation that can tease information out of the model. That could be achieved in the form of a meta-network. Meta-networks are the same models visualised with the methods mentioned earlier, but in a collapsed form (Figure 3.7). An example of such collapsed form can be obtained by grouping the nodes based on a common property (e.g. family) and assigning a meta-edge if nodes with that property interact in the network. Information is visualised in order to make a structured collection of data shorter and concise. This higher-level view might allow observations that could not be made otherwise due to limitations in visualising large datasets.

An important note regarding visualisation of interaction networks in this thesis is our naming convention. Although displaying proteins, we name the nodes with the gene name. We took this decision since we do not take isoforms into account and also because the gene name has less synonyms and is not as easily confused.

3.4 Comparative methods

3.4.1 Comparing models of PSDs

It is crucial to compare models of PSD from different organisms or models of different PSD protein sub-complexes such as the ones acquired in this work. This comparison can answer questions on the evolution of the constituent parts of complexes, the evolution of their organisation, and how that reflects the evolution of the synapse. One category of these approaches is based on the inspection and comparison of the networks' constituent parts. Examples of this approach have been applied by comparing annotation distributions or enrichments which can give a quantitative result. Another category of approaches includes the comparison of the PPINs architectures and topologies (comparative interactomics). A simple, more qualitative approach of this category is using the metanetwork visualisation. A more detailed approach can be taken by using PPIN alignment algorithms such as GraphCrunch (Milenković et al., 2008, Kuchaiev et al., 2011), protein domain based alignment (Guo and Hartemink, 2009), neighbourhood topology (Singh et al., 2007), graph structure (Klau, 2009) or NetworkBLAST, which uses graph structure and sequence similarity (Kalaev et al., 2007). What we found using these algorithms comparing the mouse and fly PSD models (Chapter 7), is that all of them are geared towards global alignment based of very conserved network structures, something which was not true with our data (mostly due to the lack of protein interactions). NetworkBLAST however, which heavily draws on sequence similarity, was able to find a conserved component in the two networks. Although most of these comparative interactomics methods are still in development, we noticed a lack of methods that combine “blind”, structural only alignment (e.g. GraphCrunch) with annotations in the same spirit of the combination based clustering algorithms. An algorithm of this class is discussed in Chapter 8 as part of the future work.

3.4.2 Implementation

In order to compare the sets of proteins that the different PSD datasets are composed of, we decided to compare various aspects of their annotation as well as their organisation in PPINs. Qualitative comparison can be achieved by observing and comparing both the PPIN models or graphic illustrations of specific features. Other than comparing annotation enrichments or network measures, for the purpose of quantitative comparison, we compiled a series of methods in order to quantify differences or similarities in the annotations of the constituent parts of the datasets and models. Dataset comparison can, in practice, be reduced to comparisons between two sets of annotations (e.g. protein domains found in the fPSD dataset versus protein domains found in the mouse PSD dataset). In order to achieve the quantitative aspect we used the following measures.

3.4.2.1 Significance of difference in annotation enrichment

Cai et al. (2006) published a method that compares two genomes based on the differences in the count of GO terms in their annotations. We modified this method allowing us to quantify significance of the differences in the count of any annotations within two sets of genes. More specifically, for each annotation term we compute the enrichment to the background (genome where available), in both PSD datasets and using a chi-squared test followed by false discovery rate (FDR) correction, we compute if there is a significant difference in the abundances.

3.4.2.2 Semantic similarity of gene sets

Jain and Bader (2010) introduced the Topological Clustering Semantic Similarity (TCSS) algorithm, which scores the semantic similarity of the GO annotations of any two genes. This semantic similarity is based on the position of each gene's annotation terms on the directed acyclic graph (DAG) of GO. TCSS is based on finding subsets of the GO DAG that define similar concepts and achieves that by a topology based clus-

tering of the Gene Ontology. This approach by definition normalises the depth of the gene's annotations. After creating an "all versus all" TCSS reference comparison of gene products within fly and mouse PSD, we were able to compare any two gene subsets of the fPSD and Union datasets by computing the average of semantic similarities between all individual genes within the two subsets.

3.5 Supplementary material

Supplementary material, such as explorable versions of the figures, data tables, and a demo of collabqc are available at <http://fruitfly.inf.ed.ac.uk/~lzografos/thesis/>. A copy of this website is available as a DVD with this thesis.

3.6 Concluding remarks

This chapter presented an integrated modular modelling pipeline. Each module for annotation and analysis can be implemented with a programming language of choice and given raw data files or programmatic access to the databases as input. This is a semi-automated pipeline since there are steps that require manual quality control or curation, however, personal experience shows that, in many cases, this distinguishes good from bad models.

There are various commercial tools available that can perform similar tasks to the modelling pipeline such as Ingenuity's IPA (<http://www.ingenuity.com/>) and GeneGo's MetaCore (<http://www.genego.com/>), which incorporate custom databases of annotations and interactions. While these tools can be used to quickly construct annotated PPINs, care must be taken that the data is of the correct type (e.g. excluding genetic interactions when reconstructing a protein complex) and quality (e.g. is computational annotation acceptable?) for the purpose. They typically provide a fixed workflow which offers some alternative options and a good GUI, but are generally not very cus-

tomisable. Before spending large sums of money on such software, it is worth making sure that it is sufficiently flexible. Depending on the project needs this may include the ability to: check data provenance (e.g. via linked PubMed ids, virtually indispensable); filter the data based on your quality requirements, both manually and through simple rule-based filtering; incorporate qualitative/quantitative data of your own (e.g. task-specific annotations, expression data); combine annotations with each other or with quantitative data to generate new annotations (e.g. all channels and receptors with high expression in hippocampus); and give sufficient control over statistical testing (the ability to define an appropriate reference set is vital when performing enrichment analyses). It is also worth noting that the information incorporated in such tools can be biased towards particular areas of research, which may not overlap with the area of interest. We have found that while commercial software might be fine for a quick first-pass analysis (if one can afford the licenses), having control over all components of the workflow allows for efficient, agile and potentially more insightful research.

Finally, another important aspect of the modelling pipeline analysis component that has to be stressed again is that all the annotation data can suffer from the partial knowledge bias. Partial knowledge can affect clustering (false negative protein interactions) and statistical correlations (missing annotations). For that reason it is imperative to update the annotations and the analysis often in order to have an up-to-date reconstructed model.

Figure 3.4: Illustration of central graph concepts: the adjacency matrix (panel A) defines graph connectivity (panel B), degree of a node (panel C), distance (panel D) and network diameter (panel E).

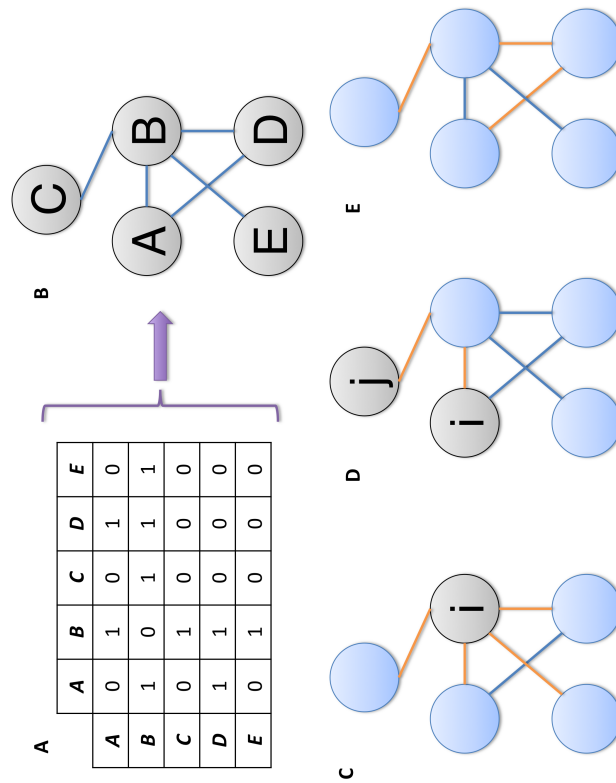


Figure 3.5: An illustration of community structure.

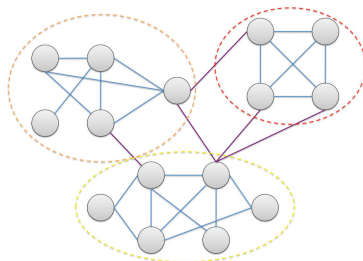


Figure 3.6: Degree distribution in a generated scale-free network. The network has 5000 nodes and was generated using the Barabási and Albert (1999) model of preferential attachment (2 nodes per step). Notice how the probability of a node connecting with many nodes decreases according to the $k^{-\gamma}$ power law.

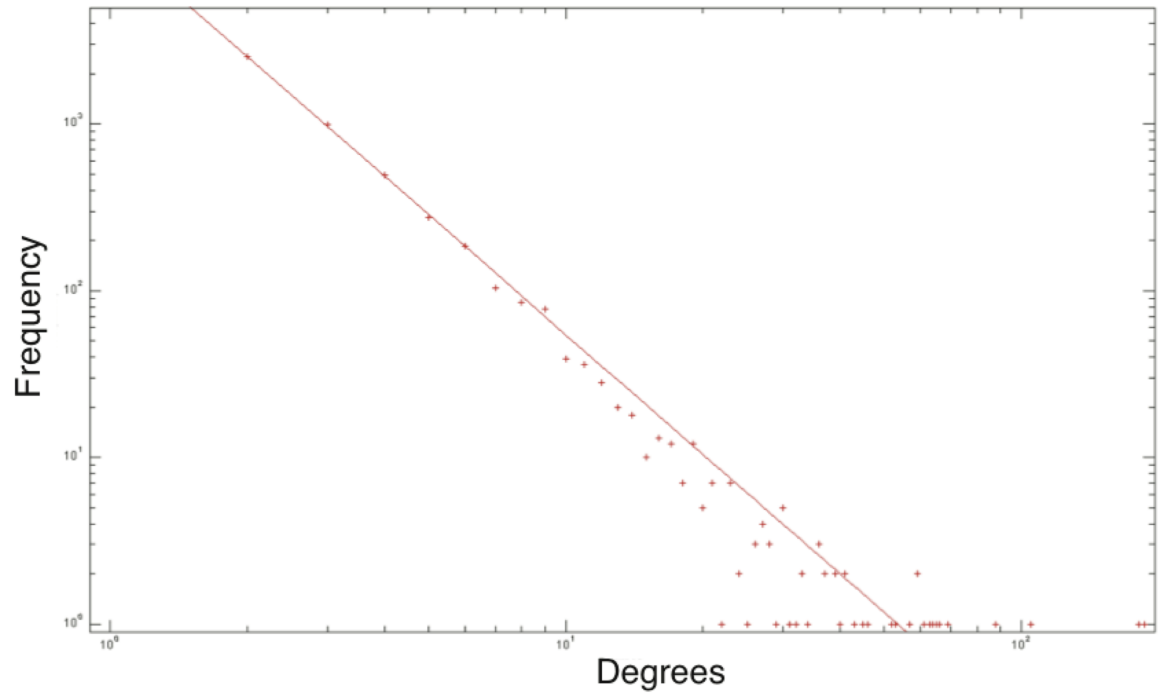
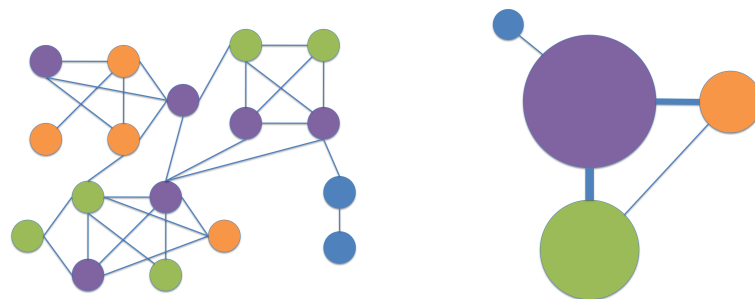


Figure 3.7: A general example of the network (left) to meta-network (right) transformation concept.



Chapter 4

The PSD-95 associated proteins complex

4.1 Background

In this chapter we will discuss the results of the targeted tandem affinity purification of PSD protein complexes, using the scaffolding protein PSD-95 as a bait, as described in (Fernández et al., 2009). The genetics and proteomics involved were performed at the Wellcome Trust Sanger Institute, Cambridge, UK by the co-authors and the annotation, model reconstruction and data analysis were performed by the author in the University of Edinburgh, UK and are the core of this chapter.

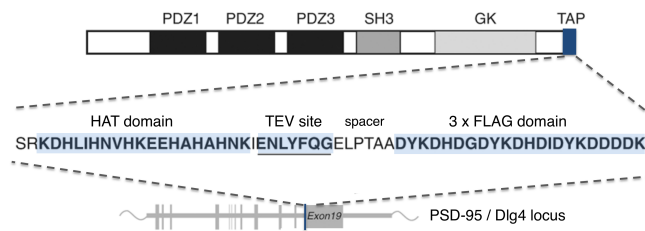
As previously discussed the PSD is a protein complex comprised of more than 2000 proteins. Although there have been numerous attempts to catalogue it the overlap between experiments is relatively low (see subsection 1.3.2). This is by itself reason enough to pursue new cataloguing endeavors, using different baits, however, one should also strive to improve the state-of-art methods. Proteomic analysis of complex sub-cellular structures such as the PSD could always be improved by introducing approaches that give datasets with less noise. This demand for new, more accurate, methods in combination with the need for more data was the premise behind the work

described in this chapter.

There is a series of reasons as to why PSD-95 was chosen as a bait for this purification. PSD-95 is encoded by the *Dlg4* gene and is one of the most abundant scaffolding proteins in the PSD of excitatory brain synapses. PSD-95 is a crucial factor to the PSD's organisation, via its protein-protein interactions (Nourry et al., 2003, Peng et al., 2004). PSD-95 is localized at the postsynaptic compartment, where it interacts with neurotransmitter receptors and ion channels (Dosemeci et al., 2007, Hunt et al., 1996, Husi et al., 2000, Kornau et al., 1995, Nehring et al., 2000) and signaling molecules like protein kinases Fyn (Tezuka et al., 1999), Cask (Chetkovich et al., 2002), and Prkca (Lim et al., 2002) to assemble signaling complexes. These complexes control neuronal plasticity (Migaud et al., 1998, Carlisle et al., 2008, Cuthbert et al., 2007), underlie learning and memory (Migaud et al., 1998) as well as drug addiction (Yao et al., 2004). Also, PSD-95 was expected to yield some new PSD proteins since it has different first degree interactors and different properties from previously used baits.

All previously PSD cataloging attempts (Husi et al., 2000, Farr et al., 2004, Husi and Grant, 2001, Sheng and Kim, 2002, Collins et al., 2006, Dosemeci et al., 2007, Klemmer et al., 2009, Paulo et al., 2009) used a single step protein complex purification step, an approach that is both limited to the specificity of the affinity reagent and potentially more prone to contaminants compared to an approach with more than one purification steps. With this being a very central problem - not only to PSD related proteomics - a solution was proposed and achieved in yeast with the fusion of a Tandem Affinity Purification (TAP) tag in the C or N terminals of a protein of interest. Application of this method allows a tandem isolation procedure, which overcomes many of the inherent specificity issues of other methods. Also, the endogenous gene integration of the TAP tag has a series of advantages over other methods that involve random insertions or over-expression (e.g. Brajenovic et al., 2004, Drakas et al., 2005, Angrand et al., 2006, Bürckstümmer et al., 2006), specially when the phenotype needs to be as close to the wild type as possible. This work shows the first example of

Figure 4.1: Domain structure of TAP modified PSD-95. PSD-95 domains, including three PDZ (PSD-95/discs large/zona occludens), a SH3 (Src homology 3), a GK (guanylate kinase) and C-terminal TAP-tag domain. Amino-acid sequence of the TAP tag comprising a histidine affinity tag (HAT)-domain, a TEV site and a 3XFLAG domain separated by a spacer. Figure from Fernández et al. (2009).



gene-targeted TAP tagging in mice that does not alter the expression or introduce any mutation but also demonstrates the advantages of two step purification in the analysis of multi-protein complexes.

4.2 Genetics and proteomics

4.2.1 Construct design

The 5kDa TAP tag used for the knock-in (Figure 4.1) consisted of a poly-histidine affinity tag (HAT) and a triple FLAG tag (Terpe, 2003) in tandem separated by a unique TEV-protease cleavage site. Note that the size of this tag is considerably smaller than the one originally used in yeast (20KDa) in (Rigaut et al., 1999). Although PSD-95 is known to have multiple isoforms of various lengths (130-767aa), they all have the C-terminal in common (Bence et al., 2005) and for that reason that was chosen as the insertion site. The recombination and integration method details are beyond the scope of this chapter and are described in Fernández et al. (2009).

4.2.2 $PSD - 95^{TAP}$ mice phenotyping

The next step after creating the $PSD - 95^{TAP}$ knock-in line was to ensure that it did not differ from the wild type mice in terms of protein expression, localisation and electrophysiology. Protein expression was tested on three different heterozygous $PSD -$

Figure 4.2: A) Immunoblot with PSD-95 antibody for immunoprecipitations. Three different heterozygous mice are shown ($PSD-95^{TAP/+}$, left panel). $PSD-95^{TAP/+}$ forebrain was also affinity purified with a FLAG antibody (right panel). B) Immunohistochemical staining of PSD-95 in sagittal hippocampus sections from $PSD-95^{TAP/TAP}$ and wild type mice showing CA1, CA3 and dentate gyrus (DG). Scale bar=1 mm. C) Long-term potentiation of fEPSPs induced by theta-burst stimulation in CA1 area of hippocampal slices is similar in $PSD-95^{TAP/TAP}$ (13 slices from 4 animals) and wild-type mice (15 slices from 4 animals). Figure from Fernández et al. (2009).

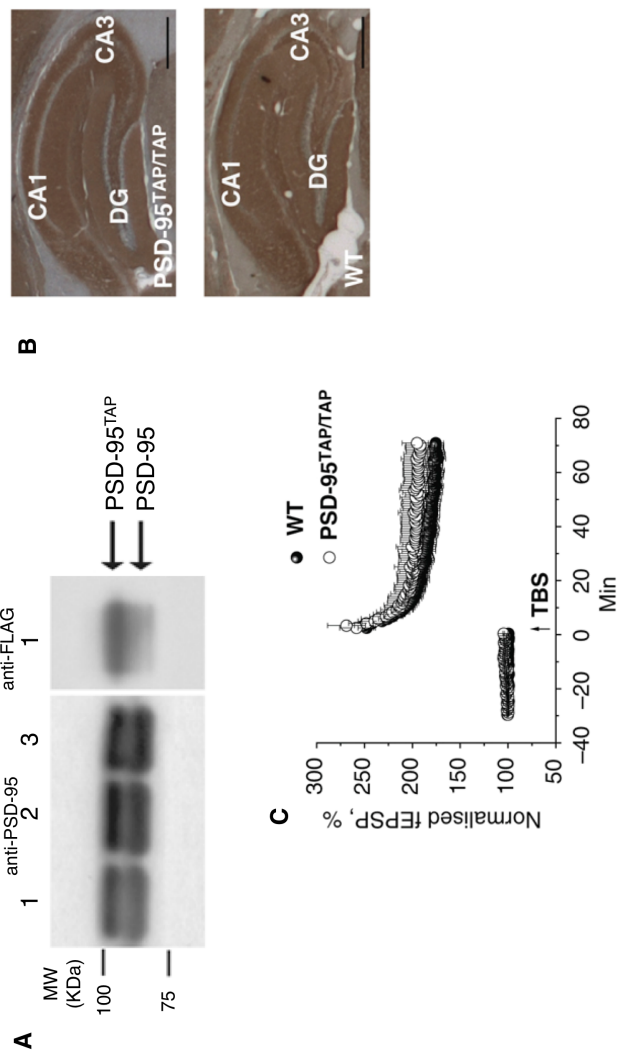
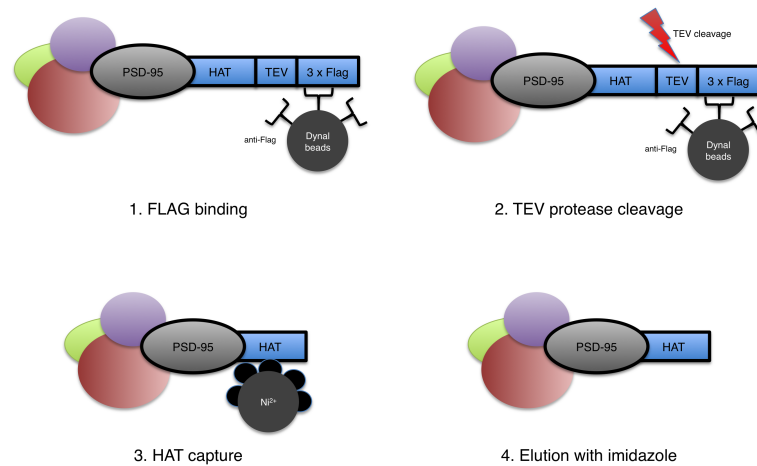
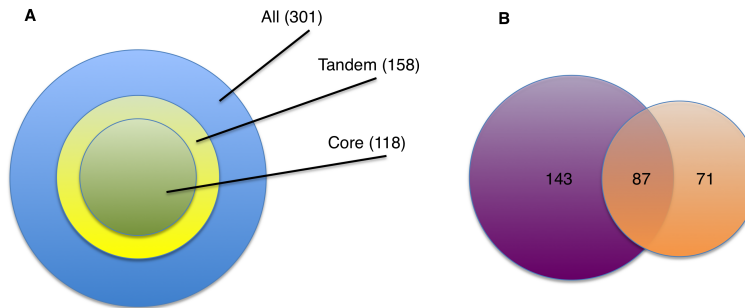


Figure 4.3: Overview of the TAP protocol. In the first step, the TAP-tagged PSD-95 was captured by FLAG antibody (1) and eluted by TEV cleavage (2). Cleaved TAP-tagged PSD-95 was then captured with Ni^{2+} -NTA-agarose beads (3) and eluted with 250 mM imidazole (4). Figure from Fernández et al. (2009)



$95^{TAP/+}$ mouse lines (Figure 4.2A, left panel) with anti-PSD-95 antibody. Two bands were observed, with the top band representing the TAP-tagged PSD-95, as also confirmed with anti-FLAG antibody (Figure 4.2A, right panel). Results were also confirmed with different protein extract concentrations and $PSD - 95^{TAP/TAP}$ samples (results not shown). Immunohistochemistry was carried out on sagittal brain sections, in order to examine localisation of the tagged protein. The expression of PSD-95 was similar in $PSD - 95^{TAP/TAP}$ and wild type animals (Figure 4.2B). Synaptic localisation was confirmed with various synaptic marker antibodies (anti-GluR1, anti-NR1, anti-MAP2B). Finally, synaptic physiology was examined in order to investigate if long and short term synaptic plasticity were affected as previously reported (Migaud et al., 1998, Komiyama et al., 2002, Béïque et al., 2006). Normalized fEPSPs after a theta-burst LTP inducing protocol on hippocampal slices were similar between $PSD - 95^{TAP/TAP}$ and wild type animals (Figure 4.2C). In conclusion, the $PSD - 95^{TAP}$ knock-in line did not show any signs of abnormal gene expression, localisation or functionality.

Figure 4.4: A) Schematic representation of the total number (301) of proteins identified in the combined single and tandem purifications. In four independent tandem purifications, a total of 158 proteins were identified and 118 appeared in at least three of four replicates (PSD-95 core complexes). B) Venn diagram with the number of proteins from either single or tandem purifications showing the common proteins (87) and proteins masked (71) in the single-step purification.



4.2.3 Protocol

The protocol followed for the isolation of PSD-95 associated protein complexes includes two purification steps (Figure 4.3). The TAP-tagged PSD-95 from *PSD – 95^{TAP/TAP}* mice is captured from brain extracts with an anti-FLAG antibody bound on Dynal beads. Isolated complexes were then eluted by cleavage using TEV protease, completing the first step of the purification. The second step of the purification is the recovery of the complexes using a Ni^{2+} -NTA-agarose column which binds the HAT part of the TAP tag. A detailed description of the protocol can be found in Fernández et al. (2009).

4.3 Results

4.3.1 The PSD-95 associated proteins complex

4.3.1.1 Characterisation of isolated complexes

Four replicate experiments were performed using *PSD – 95^{TAP/TAP}* mice brain extracts and a total of 301 proteins were identified by LC-MS/MS. These included proteins found in single step (step 1) and tandem (step 2) purifications. A total of 158 (52.5%) were found in the four independent replicate tandem purifications out of which a set

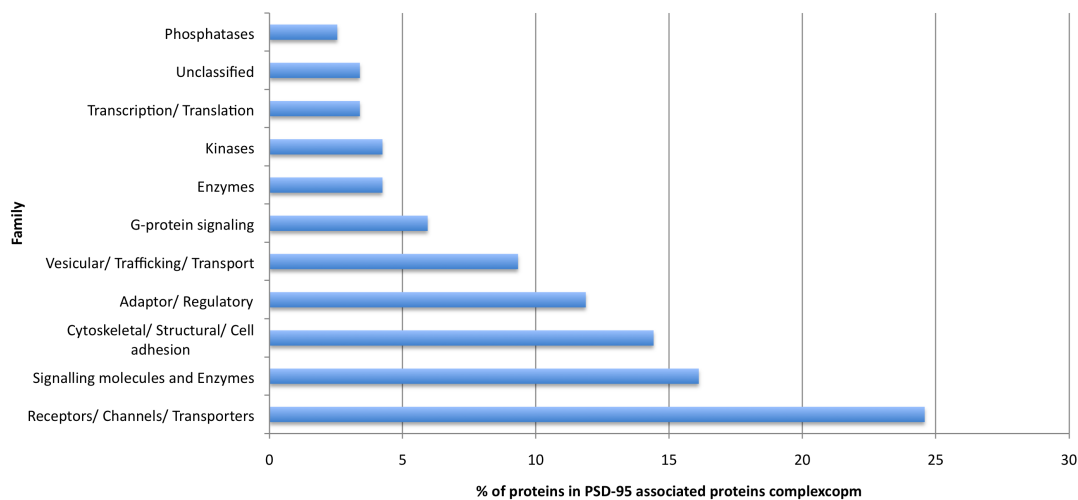
of 118 (39%) proteins (Tables 4.1 and 4.2) were found three or four times and are considered a 'core' complex (Figure 4.4A).

A significant advantage of the tandem method is that less abundant proteins, present in the single step, do not get masked in the LC-MS/MS process if they still appear after the second step of purification (Wang et al., 2006). In this case 71 (45%) of the 158 proteins were masked in the single step purification but appeared in the tandem (Figure 4.4B). To further verify the experimental results the presence of 13 known PSD-95 interactors was verified by immunoblotting.

4.3.1.2 Comparison with previous results

The PSD-95 associated proteins complex was compared with earlier studies of synapse proteomes. An earlier report (Dosemeci et al., 2007) using a single immunoprecipitation with a PSD-95 antibody identified 276 proteins from PSD fractions extracted in the absence of detergent. The comparison of this list with the PSD-95 core complexes of 118 proteins reported here shows 49 proteins in common. A peptide affinity method for binding PDZ domains of MAGUK proteins (Husi et al., 2000, Husi and Grant, 2001, Collins et al., 2005, Emes et al., 2008) was used in the same extraction conditions reported here and recovered 105 proteins (Collins et al., 2005). This peptide affinity method was not specific to PSD-95 as the peptides are known to bind PSD-93 and SAP102 (Lim et al., 2002, Chung et al., 2004). These 105 proteins and the proteins found by NMDA-receptor immunopurification were used to generate a list of 186 MASC proteins (Collins et al., 2006). Comparison of our 118 PSD-95 TAP list with the 186 proteins from the MASC complex shows 48 proteins in common (also discussed in Chapter 5). An important set of proteins that was recovered using the TAP method consisted of a) AMPA receptors and b) K^+ channels. It should be noted that this provides support to our first hypothesis, that new proteomics experiments using different baits or even technical approaches will augment our knowledge of the constituent parts of the PSD.

Figure 4.5: Distribution of families in the PSD-95 associated proteins complex.



4.3.1.3 Functional and disease annotation

Data collected included Gene Ontology (GO) and PANTHER terms associated with the proteins, protein domain data from InterPro, the corresponding homolog human genes and disease associations.

All proteins were grouped in ten groups (families) and their sub-groups (subfamilies) according to their classification in the PANTHER database. This classification is followed throughout this project for various other datasets. The families and subfamilies found in the PSD-95 associated proteins complex are shown in Tables 4.1 and 4.2, note that genes with * next to their name represent genes whose found peptides are common to other genes. Distribution of families in the 'core' dataset is illustrated in Figure 4.5. The most abundant families are Vesicular / Trafficking / Transport, Adaptor / Regulatory, Cytoskeletal / Structural / Cell adhesion, Signalling molecules and Enzymes and Receptors / Channels / Transporters, which account for more than 70% of the proteins in the core complex. This is indicative of the central position of PSD-95 in the organisation of this complex but also shows how the method is able to retrieve what is considered some of the most fundamental parts of the PSD like receptors, ion channels, scaffolding and signaling molecules.

Another point has to be made regarding the technical advantage of the tandem purification. As an initial comparison between the tandem and single step purification we looked at the distribution of families. The tandem set was enriched in Cytoskeletal/ Structural/ Cell adhesion, Receptors/ Channels/ Transporters and Adaptor/ Regulatory and depleted in Enzymes. This could reflect the fact that metabolic enzymes are more abundant and contaminate purifications (Chen and Gingras, 2007), resulting in the masking of other proteins. When we compared GO terms, the single step set showed significant over-representation of the terms “metabolism” ($p < 10^{-2}$). Also, we observed that the two-step procedure unmasked core interacting proteins that were not detected by mass spectrometry in the single-step purification: Ten known PSD-95 interactors, *Begain*, *Cit*, *Grik2*, *Grik5*, *Grin2c*, *Kcna4*, *Lrp1*, *Nlgn2*, *Nlgn3*, and *Shank1*, were present only after the tandem purification. Furthermore, we found 21 new proteins in the PSD-95 core complexes that were not reported in earlier PSD proteomic analysis (Collins et al., 2006, Dosemeci et al., 2007). These includes the adaptor proteins like *Dlgap3* and *Anks1*, receptors like *Gpr123* and *Grik5* and ion channels like *Kcnj10*, *Kcna1*, *Kcna3*, *Kcnab1* and *Kcna4*. This again suggests that the targeted TAP-tagging strategy produces greater depth and quality of interactors.

Looking at the protein domains that are over-represented in the PSD-95 associated proteins complex we can observe that the protein domains most commonly found in NRC/MASC and the PSD-95 associated protein complex (Table 4.3) were highly enriched (3-fold to 100-fold) when compared to their frequency in the genome as a whole. These top 10 most abundant domains represent key functionality associated with synaptic signaling: G-protein-coupled signal transduction (Extracellular ligand-binding receptor, Extracellular solute-binding protein, family 3), scaffolding (Src homology-3 domain, Variant SH3, PDZ/DHR/GLGF), channel polymerisation (BTB/POZ-like, BTB/POZ-fold), membrane localisation (Pleckstrin homology) and neurotransmitter related signaling (Ionotropic glutamate receptor, NMDA receptor, Glutamate receptor-related). These functional domain annotations clearly reflect spe-

Table 4.1: Families and subfamilies of the PSD-95 associated proteins complex. Continued in Table 4.2.

Family	Subfamilies	MGI Gene Symbol
Adaptor/ Regulatory	14-3-3, PDZ-domain containing scaffolders, non-PDZ-domain containing scaffolders	Ywhae, Anks1a, Anks1b, Baiap2, Begain, Dlgap1, Dlgap2, Dlgap3, Dlgap4, Dlg1, Dlg2, Dlg3, Dlg4, Slc9a3r1
Cytoskeletal/ Structural/ Cell adhesion	Actin / ARP, Catenins, MAPs, Myelin, Other Cell Adhesion Molecules, Other Cytoskeletal Proteins, Other signaling molecules, Spectrin, Tubulin, actinin	Arpc4, Plp1, Lgi1, Nrnx1, Ablim1, Adam22, Capza2, Cfl1, Dstn, Fscn1, Nefl, AI662250, Arc, Spnb2, Tubb6, Tuba1a*, Tubb2b*
Enzymes	ATP synthases, Other Enzymes	Atp5a1, Atp5b, Atp5c1, Atp5o, Gpx4
G-protein signaling	G-proteins, Modulators	Gnao1, Rac1, Sept11, Sept5, Abr, Kalrn, Syngap1
Kinases	Ser/Thr Kinases, Tyr Kinase	Camk2a, Camk2b, Mapk1, Pkg1, Pkm2
Phosphatases	Protein Phosphatases	Ppap2b, Pppp3ca, Ppp3cb

cialisation for the scaffolding associated with the Glutamate receptor activity. Compared to the NRC/MASC data, signaling related domains (e.g. Kinases) are absent from the top ten, but still very enriched (e.g. Protein kinase, catalytic domain: 1.38-fold, Guanylate kinase: 3-fold, Serine/threonine-protein kinase-like domain: 1.4-fold).

4.3.2 The PSD-95 associated proteins interaction network

4.3.2.1 Interaction mining

Interaction data for this reconstruction were collected through the interaction annotation part of the modelling pipeline with two main resources: protein interaction databases (as described in subsection 3.2.2.4 on page 58) and the TXM pipeline (as described in subsection 3.2.3.5 on page 62).

Hits from both resources were manually curated using collabQC. Note that, as

Table 4.2: Families and subfamilies of the PSD-95 associated proteins complex (continued from Table 4.1).

Family	Subfamilies	MGI Gene Symbol
Receptors/ Channels/ Transporters	ATP synthases, Ca ²⁺ -ATPases, Glutamate Receptors, Inward rectifying K ⁺ channel, NA ⁺ /K ⁺ -ATPases, Other Channels and Receptors, Other signaling molecules, Transporters, Voltage-dependent anion channels, Voltage-gated K ⁺ -channel	Atp6v0d1, Gria1, Gria2, Gria3, Gria4, Grik2, Grik5, Grin1, Grin2a, Grin2b, Grin2d, Kcnj10, Kcnj4, Atp1b1, Gpr123, Slc1a2, Slc4a4, Cacng2, Sfxn3, Slc25a4, Slc25a5, Vdac1, Vdac2, Kcna1, Kcna2, Kcna3, Kcna4, Kcnab1, Kcnab2
Signalling molecules and Enzymes	Heat shock / Chaperones / Chaperonins, Mitochondrial Enzymes, NADH-Ubiquinone Oxidoreductase, Other Enzymes, Other signaling molecules, Translation	Aco2, Msrb2, Sdha, Acat1, Acot7, Aldoc, Cnp, Gapdh, Cypin, Glul, Pdha1, Pdhb, Pgam5, Prdx1, Prdx2, Sucla2, Btbd11, Phb2, Pcbp1
Transcription/ Translation	Ribosomal Proteins, Transcription, Transcription Elements	Rps14, Rps3, Park7, Uba52*
Vesicular/ Trafficking/ Transport	Clathrin, Modulators, Motor Proteins, Other Enzymes, Other signaling molecules, Other transport, Synaptic vesicle	Cltc, Iqsec1, Iqsec2, Arf3, Cpne7, Nsf, Stx1b2, Stxbp1, Syt1, Vamp2*, Cpne4*

indicative numbers of the size of the task, that the high confidence output of TXM was 289 hits. Out of these 133 were curated as true positive binary interactions between 43 distinct pairs of proteins. Another 21 interactions were added from previous curations (NRC/MASC and in-house knowledge). The rest of the interactions were manually curated from UniHi database entries resulting in a total of 119 interactions between 50 of the 118 proteins of the core complex, excluding self-interactions.

4.3.2.2 Model reconstruction

The reconstructed protein interaction network included 50 out of 118 proteins, with 40 proteins forming a major connected component (MCC). The network is illustrated in Figure 4.6. Application of the Newman & Girvan algorithm (Newman and Gir-

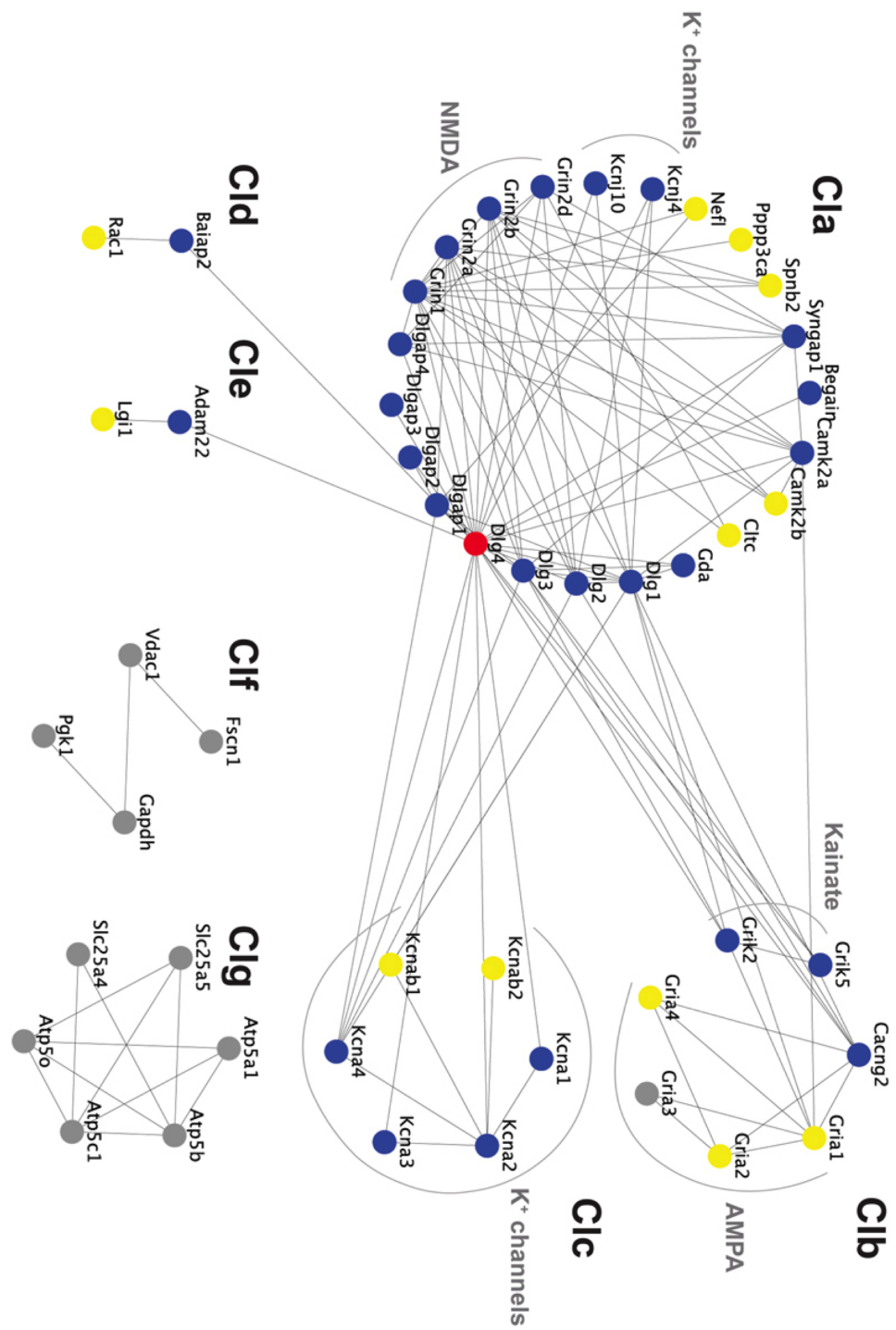
Table 4.3: Ten most common protein domains in the PSD-95 associated proteins complex. All domains have key synaptic signalling functionality.

Domain	n-fold enrichment compared to genome
Ionotropic glutamate receptor	101
NMDA receptor	90.9
Extracellular solute-binding protein, family 3	101
Extracellular ligand-binding receptor	22.03030303
Src homology-3 domain	4.954545455
Variant SH3	5.677083333
PDZ/DHR/GLGF	5.759493671
Pleckstrin homology	3.345588235
BTB/POZ-like	4.212962963
BTB/POZ fold	4.0625

van, 2004) to the MCC segregated the network into five clusters, hereon referred to as cluster a (Cla) to cluster e (Cle). In addition to the five MCC clusters, two further disconnected clusters ('Clf' and 'Clg') were found (see Table B.1, Appendix B). The modularity Q for this configuration was found to be 0.37. We found clustering configurations with higher Q values (up to 0.42); however, these configurations did not reflect the functional organization of the network as well as the one used¹. With the latter in mind, and also on the basis of the observation that 0.37 was over the average of examined configurations, we decided to use that. This segregation in clusters can be interpreted as supporting evidence for our hypothesis that some modular architecture is present in protein interaction networks of the PSD. Note that the protein, interactions, and annotations lists are available with the supplementary material DVD.

¹E.g. receptors of the same family did not cluster together. It has to be noted that solutions of clustering algorithms such as the one used might be local maxima, and not reflect the optimal solutions. Also, given the potential lack of interactors and interaction data we decided that it is reasonable to accept solutions with a lower Q .

Figure 4.6: Protein interaction network of PSD-95 interacting proteins. 50 proteins of the PSD-95 core complex were connected, with 119 interactions segregated into 5 clusters (C1a–C1e) forming the MCC and two separate small clusters C1f and C1g. PSD-95/Dlg4 is shown in red, primary interactors of PSD-95/Dlg4 are shown in blue and secondary interactors are shown in yellow. The glutamate receptors (NMDA, AMPA and kainate receptors) and potassium channels are bracketed. From Fernández et al. (2009). An interactive version of this figure with fully visible node labels is available in the additional material website (available in DVD format with this thesis)



4.3.2.3 Network topology analysis

It is interesting to note the location and proximity of the receptors and channels responsible for the postsynaptic depolarization and subsequent action potential generation. All NMDA, AMPA and kainate glutamate receptors were restricted to Cla and Clb and the voltage-dependent K⁺ channels were found in Cla and Clc (entirely comprised of K⁺ channels). These channels are known to couple to plasticity mechanisms (Chen and Sharp, 2004, Kim et al., 2007, Watanabe et al., 2002), and we noted that Cla contains important signaling enzymes involved in plasticity, including CaMK2 (Frankland et al., 2001) and SynGAP (Komiyama et al., 2002). It therefore seems that Cla, Clb and Clc are enriched with membrane proteins responsible for the electrical properties of the postsynaptic terminal as also verified in the statistical correlation of annotation analysis.

As PSD-95 was the bait for the biochemical isolation of the complexes, we examined the distribution of its primary interactors (proteins that directly bind PSD-95) and secondary interactors (proteins that do not bind PSD-95 directly, but bind one of its primary interactors) (Figure 4.6, top). Of the 39 MCC proteins (excluding PSD-95), 26 (67%) were primary interactors (blue symbols in Figure 4.6, top) and 12 (31%) were secondary interactors (yellow symbols in Figure 4.6, top) and only one protein, the AMPA receptor subunit Gria3, was a tertiary interactor. The majority of each cluster was comprised of primary interactors, more specifically Cla (74%), Clb (43%), Clc (67%), Cld (50%) and Cle (50%). To examine the centrality of each protein in the network the shortest path from each protein to every other protein was counted, and the average shortest path (ASP) calculated. For all proteins, the mean ASP was 2.25, indicating an intricate crosstalk. Ranking the ASP of each protein (not shown) showed PSD-95 had the lowest ASP (1.3), consistent with its central role in these networks.

4.3.2.4 Statistical correlation analysis

The distribution and enrichment of families across the clusters seems to be quite specific, (Table 4.4) with Clb, Clc and containing Receptors/ Channels/ Transporters only while Cla being significantly enriched in Receptors/ Channels/ Transporters and Adaptor/ Regulatory proteins, but containing some signaling proteins (Phosphatases, Kinases, G-protein signaling) as well. Also, Cle only contains Cytoskeletal/ Structural/ Cell adhesion proteins. Clg comprises of of four Enzymes and two Transporters.

When looking at cellular component (CC) GO term enrichments (Table B.2, Appendix B) Cla, Clb and Clc seem to be tightly or exclusively associated with the membrane. Cla seems to be predominately correlated with the postsynaptic membrane and the associated scaffolding with the exception of the soluble signaling molecules found there. Regarding clusters Clb and Clc all the lower level terms are membrane related, as expected since they both only contain receptors and channels with the exception of Cacng2. Finally, Clg is associated with the mitochondria. As expected Cla and Clb and Clg have the most diverse molecular function (MF) GO term enrichment with most of the enrichments being due to the channel or receptor activities and the binding interactions between the adaptor and scaffolding proteins with the channels, receptors or other molecules. The biological process (BP) GO term enrichments (Table B.4, Appendix B) show that Cla is associated with ion transport through its NMDA receptors and Potassium channels. Molecules like CaMK2a and CaMK2b indirectly associate it with Calcium transport. Via the signaling, scaffolding, and receptors proteins GO annotations, Cla is associated with the regulation of neuronal synaptic plasticity. Clb is dominated by AMPA receptor processes. Regarding Clc, its functionally is associated with Potassium transfer. Clg being mitochondrial, is associated with mitochondrial processes (e.g proton transport). Crosschecking the above using PANTHER's protein classification system, we found a similar distribution of functional groups of proteins (Table B.5, Appendix B).

PANTHER also offers pathway association information, shown in Table B.6, Ap-

Table 4.4: Significant cluster and family correlations in the PSD-95 associated proteins network model. P-values in parentheses.

Cluster	Families
Cla	Adaptor/ Regulatory ($< 10^{-6}$), Receptors/ Channels/ Transporters (0.02)
C1b	Receptors/ Channels/ Transporters ($< 10^{-6}$)
C1c	Receptors/ Channels/ Transporters ($< 10^{-6}$)
C1g	Enzymes ($< 10^{-5}$)

pendix B. Cla is associated with NMDA receptor and signaling but also a Huntington disease pathway. C1b is associated with the AMPA receptor signaling pathways. Also, C1d is correlated (note that there is a low count so it is not shown) with various pathways via Rac1's signaling activity.

It has to be noted that this distribution of functional families and their respective biological functions in clusters reflect some biological significance to the computed clustering configuration of the network. We have to keep in mind that clustering is computed with no knowledge of biological function. The fact that the latter is segregated in a biologically meaningful manner can be interpreted as evidence of this modular architecture having biological significance in how protein networks of molecular machines like the PSD function.

4.3.2.5 PSD-95 associated proteins complex and disease

Regarding disease annotation specifically data was collected for the core set of 118 PSD-95 associated proteins via OMIM, Genetic Association Database, data mining and manual curation. Out of the 118 proteins 49 (41.5%) were implicated in multiple (25 total) diseases like schizophrenia (28), mental retardation (6), bipolar disorder (13), Alzheimer's disease (6) and others (29). All disease associations are shown in Tables 4.5 and 4.6. For references look within Fernández et al. (2009) supplementary material. We next analyzed the pair-wise correlation between functional categories and disease type. The Receptors/ Channels/ Transporters family and Glutamate Receptors

subfamily were significantly correlated with schizophrenia ($p = 0.002$ and $p < 10e - 6$, respectively). Out of 28 schizophrenia-implicated proteins, 20 were mapped on the network model (orange in Figure 4.7). Of those 20 proteins, 70% fell into Cla, which was significantly enriched in schizophrenia-related proteins ($p = 0.0089$). All but one of the remaining schizophrenia-related proteins were found in cluster Clb. There was also a correlation between Kinases and depression, which corroborates previous evidence (Pocklington et al., 2006) but the count (2) made us considering it rather weak in this case. Mapping the primary interactors of these schizophrenia proteins recruited many other proteins found in the other modules of the network. More specifically 40 out of 50 proteins in the network were either associated with or first degree neighbours of genes associated with schizophrenia. This schizophrenia subnetwork (Figure 4.7) covers more than 92% of the MCC and 80% of the whole PSD-95 associated interactions network. Note that there are also 47 interactions (~39% of total number of interactions) between the schizophrenia susceptibility genes. All this implies that proteins and interactions in this core complex are very likely to be involved with schizophrenia.

4.4 Concluding remarks

The work described here contributed in multiple ways to different aspects of cataloguing and modelling the PSD. The first contribution has to do with the genetics and the proteomics and is that it is the first isolation of mouse proteins complexes using a knock-in TAP tag fused endogenous protein bait. This allows the gene's expression and the protein's localisation to be regulated by its endogenous regulatory elements. Also, it must be noted that the insertion of the TAP tagged gene did not seem to introduce any detectable mutation. From a proteomics aspect, the two step purification process yielded less contaminants, proving the TAP method as a good approach for isolating and purifying complexes with less noise generating and data obfuscating contaminants. One also has to consider how the TAP method in this case can overcome

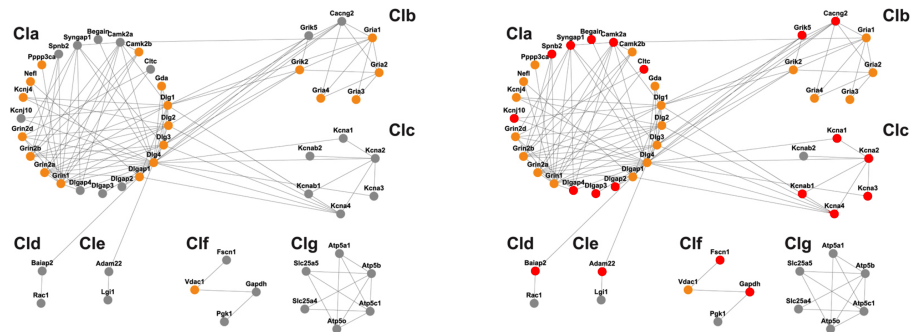
some of the common immunoprecipitation limitations such as when i) an antibody for the bait is not available or cross-reacts, ii) the antibody does not work due to binding of interactors, iii) the antibody affects the binding of interactors, and iv) the harsh elution conditions.

The other aspect of the contribution of this work is regarding the PSD. In this work we managed to co-precipitate NMDA, AMPA, Kainate receptor subtypes along with major K^+ channels, which shows that these proteins are found in native complexes as expected from major postsynaptic constituents responsible for synaptic transmission and shaping the postsynaptic electrophysiological response to presynaptic input.

From a protein interaction networks perspective an important observation is that the data described here connects PSD-95 with AMPA receptors, on which not a lot is known. PSD-95 has been shown to affect AMPA receptor-mediated excitatory synaptic transmission (Migaud et al., 1998, Béïque et al., 2006, Carlisle et al., 2008), via what is thought to be indirect interactions with stargazin, SAP-97, Adam22, Lgi1 and Nsf (Leonard et al., 1998, Osten et al., 1998, Fukata et al., 2006). Co-precipitation with Nsf reinforces the idea of PSD-95 involvement in synaptic vesicle trafficking and AMPA surface-expression modulation (Lüthi et al., 1999, Noel et al., 1999) since it implies some interaction between PSD-95 and Nsf. Other proteins involved in the trafficking and clustering of AMPA receptor are Arc/Arg3.1 (Chowdhury et al., 2006, Shepherd et al., 2006) and Rac1 (Wiens et al., 2005), and these were found within the complexes. The isolation of multiple AMPA-receptor modulators in the PSD-95 complexes underlines the importance of this complex in mediating synaptic plasticity.

From a mental disease point of view 19 genes involved in schizophrenia were significantly associated with the clusters Cla and Clb that contain all the glutamate receptors and MAGUK/Dlg proteins. There were 47 interactions between these nodes that create a schizophrenia associated subnetwork with nodes that have internal degrees (within this subnetwork) varying between one and ten edges. Mapping the primary interactors of these schizophrenia proteins recruited many other proteins found in the

Figure 4.7: Left: protein interaction network of PSD-95 interacting proteins. Schizophrenia susceptibility genes are shown in orange. Right: schizophrenia subnetwork in the PSD-95 associated proteins interaction network. Schizophrenia susceptibility genes are shown in orange and their primary interactors in red. From Fernández et al. (2009). An interactive version of this figure with fully visible node labels is available in the additional material website (available in DVD format with this thesis).



other modules of the network. This suggests that the overall network, its topology and interactions and its various clusters might play a role in schizophrenia, via pathway crosstalk and interplay, and not simply the glutamate receptors, as was generally considered in the 'glutamate hypothesis' of schizophrenia (Greene, 2001, Coyle, 2006, Lisman et al., 2008).

In its relation to our set of hypotheses this chapter presents evidence verifying that PSD proteomics data can be augmented by new methods and affinity purification strategies. More specifically, the co-precipitation of NMDA, AMPA and Kainate receptors along with major channels not only augmented the protein lists from the perspective of constituent parts of the PSD networks, but is also evidence of the interactions of the above proteins in functional complexes, as isolated from samples where nothing but native regulation of expression takes place. Additionally, the reconstruction PSD-95 associated proteins interaction network model shows that there is a clear modular architecture. Most importantly this modular architecture although similar to that of the NRC/MASC model was computed using data acquired from different experiments, giving further evidence supporting the presence of this modular architecture as a feature in different modules of the PSD networks. Finally, as the associations of functional annotation show, this modular architecture seems to have a biological sig-

Table 4.5: Diseases and associated genes in the PSD-95 associated proteins complex. Continued in Table 4.6.

Disease	Num of Genes	MGI Gene Symbols
Schizophrenia	28	Acot7, CamK2b, Cnp1, Dlg1, Dlg2, Dlg3, Dlg4, Dlgap1, Gda, Gnao1, Gria1, Gria2, Gria3, Gria4, Grik2, Grin1, Grin2a, Grin2b, Grin2d, Kcnj4, Mapk1, Nefl, Nrxn1, Nsf, Ppp3ca, S11a2, Stxbp1, Vdac1
Bipolar affective disorder	7	Atp5c1, Grin1, Grin2b, Msrb2, Slc25a4, Vdac1, Vdac2
Alzheimer's	6	Atp5a1, Gapdh, Gria1, Grin2a, Prdx1, Vdac1
Bipolar disorder	6	Cacng2, CamK2a, Dlg3, Dlg4, Nefl, Pgl1
Epilepsy	6	Adam22, Gria1, Gria2, Grin2b, Kcnj10, Lgi1
Depression	5	CamK2b, Dlg3, Mapk1, Pdha1, Plp1
Mental retardation	4	Capza2, Cltc, Grik2, Pgl1
ALS	2	Nefl, S11a2
Huntington disease	2	Grin2a, Grin2b
Seizure	2	Grin1, Kcnj10
X-Mental retardation	2	Dlg3, Gria3
Attention disorder	1	Grin1

nificance, reflecting the distribution of biological functions to different sub-networks of the bigger network. Regarding the latter two points, about the architecture of the complex, admittedly we can not infer conclusions for the whole network of PSD protein only by looking at a subset of the it. However, the evidence presented in this chapter offer the first evidence towards that direction regarding the molecular machine of the PSD.

Table 4.6: Diseases and associated genes in the PSD-95 associated proteins complex (continued from Table 4.5).

Disease	Num of Genes	MGI Gene Symbols
Autism	1	Nrxn1
CMT1	1	Nefl
CMT2	1	Nefl
Demyelinating disease	1	Plp1
Episodic ataxia type 1	1	Kcna1
Miller-Dieker lissencephaly	1	Ywhae
Multiple sclerosis	1	Plp1
Neurodegeneration	1	Atp1b1
Ophtalmoplegia	1	Slc25a4
Parkinson's	2	Pgk1, Prdx2
Pelizaeus-Merzbacher disease	1	Plp1
Rett syndrome	1	Atp1b1
Spastic paraplegia	1	Plp1

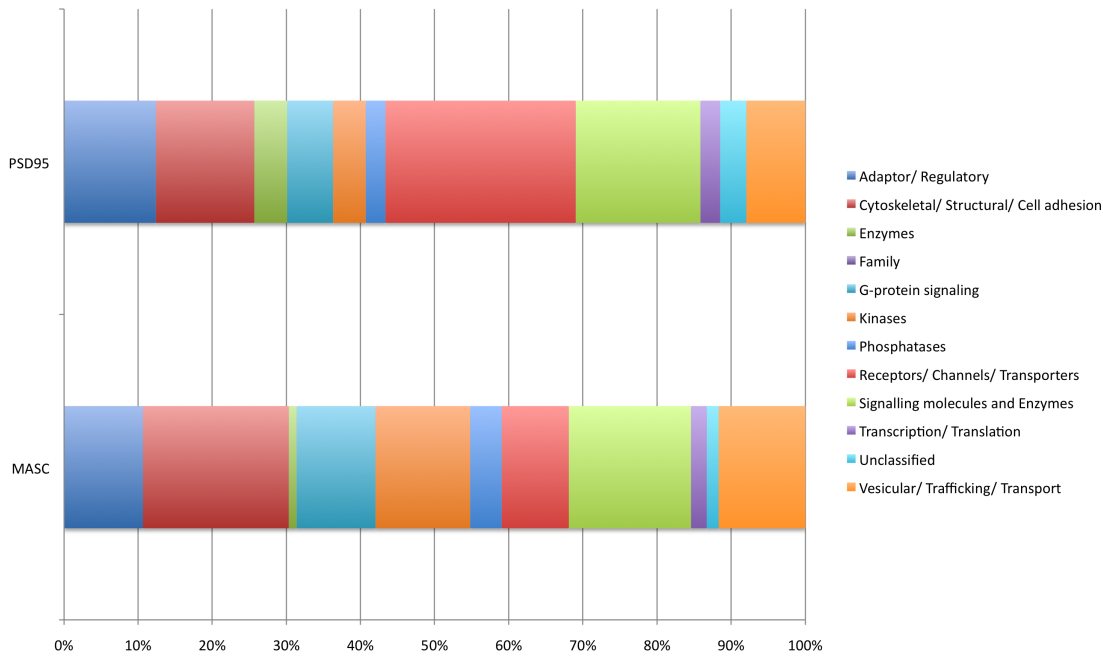
Chapter 5

The PSD interactome

5.1 Background

Two major murine PSD datasets have been reconstructed into models, the NRC/MASC and the PSD-95 associated proteins complex. Network models for these complexes were reconstructed separately by Pocklington et al. (2006) and in Chapter 4 respectively. The contribution of individual complexes towards different aspects of a wider model of the mouse PSD can become evident after examining the datasets' compositions. One can look at the data contribution of the individual models with various approaches, one of which is to see the contribution of specific protein families (Figure 5.1). One instantly observable difference is in the Receptors / Channels / Transporters family, which occupies ~15% more of the PSD-95 associated proteins complex compared to the NRC/MASC complex. If we focus on the distribution of subfamilies of this family (Figure 5.2) we can see how the PSD-95 associated proteins complex contributes to a more cross-membrane “lateral” perspective of the PSD by contributing insight on the composition of the electrical component (Voltage-gated K^+ channels, Inward rectifying K^+ channels) but also with glutamate receptors (AMPA family) and members of the Other Channels and Receptors subfamilies. The NRC/MASC on the other hand contributes, with Glutamate Receptors, G-proteins, Kinases and Phos-

Figure 5.1: Protein family distribution in the NRC/MASC and PSD-95 associated proteins complexes.

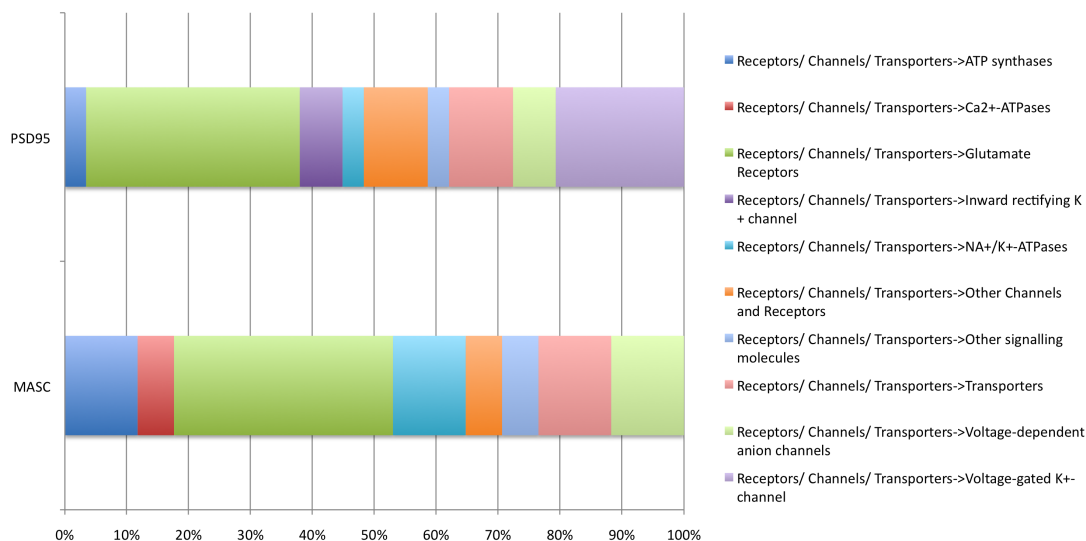


phatases as well as with different MAGUKS of the Adaptor / Regulatory family.

When looking at the protein lists of the two complexes there is a 48 protein overlap, which is also reflected on the protein interaction network models of the two complexes (Figure 5.3). A total of 16 proteins were common to the aforementioned network models ($p < 10^{-7}$) with overlap centered (10/16 proteins, $p < 10^{-3}$) on Cla and NRC/MASC cluster 1. These clusters are part of what is described as the 'input layer' of these molecular machines, since they contain many of the proteins responsible for receiving the extracellular signal and relaying it further downstream.

Given the aforementioned overlap in both the context of the protein lists and the protein interaction network models and also the different contributions of NRC/MASC and the PSD-95 associated proteins complex we decided to merge the two complexes into one larger model. Application of the modelling pipeline would lead to a manually curated and annotated model of all mouse PSD proteomics data produced by the Genes2Cognition consortium at that time. Using high quality proteomics data from one source ensures that the core model has high confidence and reproducibility. Such a "gold standard" model of considerable size, will be useful for integrating existing

Figure 5.2: Subfamily distribution for the Receptors / Channels / Transporters family in the NRC/MASC and PSD-95 associated proteins complexes.



data, adding data that is underway, but also for a comparative analysis of the murine synaptic molecular machine with that of other organisms, like *D. melanogaster* or human models.

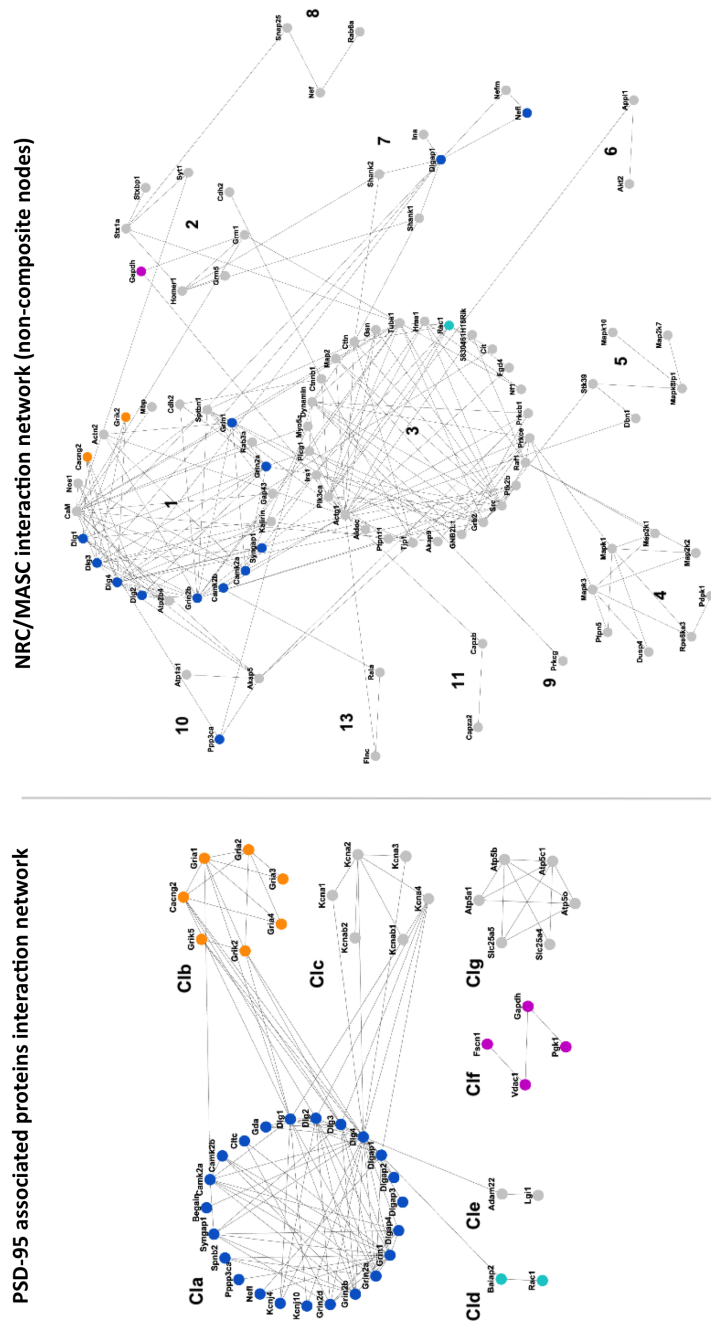
The overlap between the aforementioned two PSD protein interaction models also provided evidence for the presence of modular architecture with biological significance in protein interaction networks like the PSD. More specifically, we saw how the modular architecture was not only present in two independently derived datasets, but also how most of the overlap present in the network was found to be clustered together from two independently computed clustering configurations, showing a persistence of this modular architecture.

5.2 Integration and data mining

5.2.1 Dataset merging

The first step towards this model was merging the NRC/MASC and PSD-95 associated proteins complex protein lists. We merged using the 'core' PSD-95 associated

Figure 5.3: Overlap of the PSD-95 associated proteins interaction network and NRC/MASC interaction network (composite nodes removed) . Node colours reflect clusters of the PSD-95 associated proteins interaction network (left), some of the proteins in these highlighted clusters also belong to the overlap with the NRC/MASC interaction network (right). An interactive version of this figure with fully visible node labels is available in the additional material website (available in DVD format with this thesis).



proteins dataset, after removing protein entries where there was ambiguity for the corresponding gene (113 proteins) and the NRC/MASC list after removing composite entries (188 proteins - two of the proteins of overlap were within the composite entries part of NRC/MASC¹). As mentioned earlier there was an overlap of 48 proteins, so the total number of proteins in the complex was 253. GO annotations for all entries were also merged and updated. The dataset resulting from the union of NRC/MASC and PSD-95 associated proteins complex dataset will be hereon referred to as Union.

5.2.2 Interaction mining

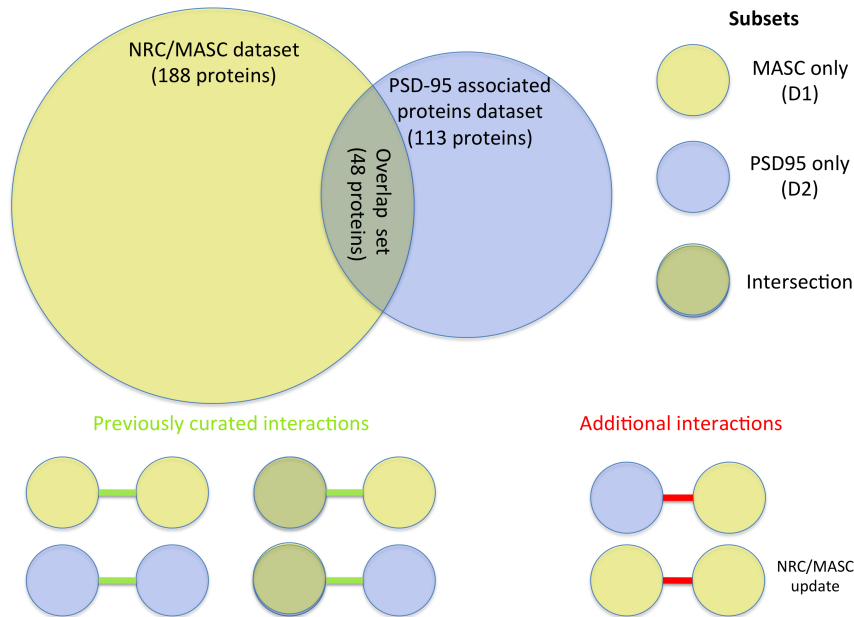
Interactions have been previously curated both for the NRC/MASC and the PSD-95 associated proteins complex individually. For that reason the subsets we had to mine interactions for would be proteins belonging to the differences of the two sets. Also, since the PSD-95 associated proteins model curation had been performed 2 months previous to the reconstruction of this model, we decided that there is no need for an update. On the other hand the NRC/MASC interaction curation had been done almost 3 years before so we decided to update it. An illustration of the above concept is shown in Figure 5.4.

Due to the time bottleneck associated with running a dedicated text-mining service we decided against it and mined two other main data resources for interactions. One was protein interaction databases and the other was the FIND protein interaction dataset. The FIND dataset (Ewa Stocka, unpublished) was provided by the Dietrich Rebholz's group at the EBI. It was compiled, with tools developed by the group, by mining PubMed abstracts based on co-citation of protein and gene names. There were 1052 potential interactions hits from FIND and 248 from protein interaction databases. Out of these, interactions between 505 pairs of proteins were curated as true positive.²

¹Composite entries contained proteins not identified by MS.

²The rest were curated out of the dataset as false positive hits, something very common to an error prone approach such as co-citation.

Figure 5.4: Overlap and difference between the NRC/MASC and PSD-95 associated proteins datasets (top). Additional interactions (bottom). Interactions within the NRC/MASC dataset as well as interactions between the differences of the PSD-95 associated proteins and NRC/MASC datasets had to be updated and curated respectively.



5.3 Results

5.3.1 The Union protein complex

When examining the twenty most common abundant protein domains (Table 5.1) in the dataset the functionality represented includes G-protein-coupled signal transduction (Extracellular solute-binding protein, family 3), scaffolding (Src homology-3 domain, Variant SH3, PDZ/DHR/GLGF), mitochondrial domains (Mitochondrial Rho-like), membrane localisation (Pleckstrin homology) and neurotransmitter related signaling (Ionotropic glutamate receptor, NMDA receptor, glutamate receptor-related), and phosphosignalling (Serine/threonine-protein kinase domain, Protein kinase, ATP binding site, Tyrosine-protein kinase, catalytic domain). This shows a clear overlap with the abundant domains in the respective datasets, revealing, as expected PSD functionalities.

Table 5.1: Twenty most common protein domains in the Union complex. All domains have key synaptic signalling functionalities.

Domain	n-fold enrichment compared to genome
C2 calcium-dependent membrane targeting	5.83
Protein kinase, catalytic domain	3.48
Serine-threonine/tyrosine-protein kinase	4.05
Ionotropic glutamate receptor	43.08
Src homology-3 domain	5.82
PDZ/DHR/GLGF	6.04
NMDA receptor	40.81
Extracellular solute-binding protein, family 3	45.61
Ras GTPase	5.43
Pleckstrin homology domain	3.83
Serine/threonine-protein kinase domain	3.46
Ras small GTPase, Rab type	4.76
Small GTP-binding protein domain	4.67
C2 calcium/lipid-binding domain, CaLB	5.14
Protein kinase-like domain	3.37
Variant SH3	6.56
Mitochondrial Rho-like	4.85
Ras	5.14
Serine/threonine-protein kinase-like domain	3.56
Tyrosine-protein kinase, catalytic domain	3.55

5.3.2 The Union protein interaction network model

5.3.2.1 Model reconstruction

The resulting protein nodes and interaction edges generated a protein interaction network with a major connected component (MCC) of 164 nodes and 458 interactions. The protein interaction network was partitioned in 15 clusters using the Newman and Girvan (2004) algorithm (modularity, $Q = 0.52$), as illustrated in Figure 5.5. This figure also illustrates the distribution of families and specific key protein groups such as glutamate receptors, ion channels, scaffolding and phosphosignalling related proteins. The families and proteins in each cluster are shown in Tables B.7, B.8 and B.9, found in Appendix B. Note that the protein, interactions, and annotations lists are available with the supplementary material DVD.

As previously discussed there was a 48 protein overlap between the initial NRC/MASC and PSD-95 associated proteins datasets. The merging of the datasets and addition of new interactions allowed more of that overlap to enter the network model. More specifically 33 (c.f., 16 before interaction curation) out of the 48 proteins are now in the model. The distribution of PSD-95 associated proteins complex-only, NRC/MASC-only and overlap proteins varied from cluster to cluster (Figure 5.6). There are cases like cluster A where contribution of the original datasets was equal or cases like clusters B and D where one complex contributed many more proteins. There are also the case of cluster I, which originates only from NRC/MASC

5.3.2.2 Network topology analysis

As mentioned before the network segregates in 15 clusters. Significant correlations of GO annotations, which include the domains of molecular function (MF), biological process (BP) and cellular component (CC) are shown in tables B.10, B.11 and B.12 respectively, found in Appendix B. Again, as hypothesised earlier, the network appears to have a modular architecture parts of which (clusters) have correlations with specific

Figure 5.5: The Union protein interaction network. 164 proteins connected with 455 interactions segregated into 15 clusters forming the MCC. Nodes are coloured by family. K^+ channels are parallelograms, glutamate receptors V-shaped, Scaffolding proteins are triangular and Kinases / Phosphatases square. An interactive version of this figure with fully visible node labels is available in the additional material website (available in DVD format with this thesis).

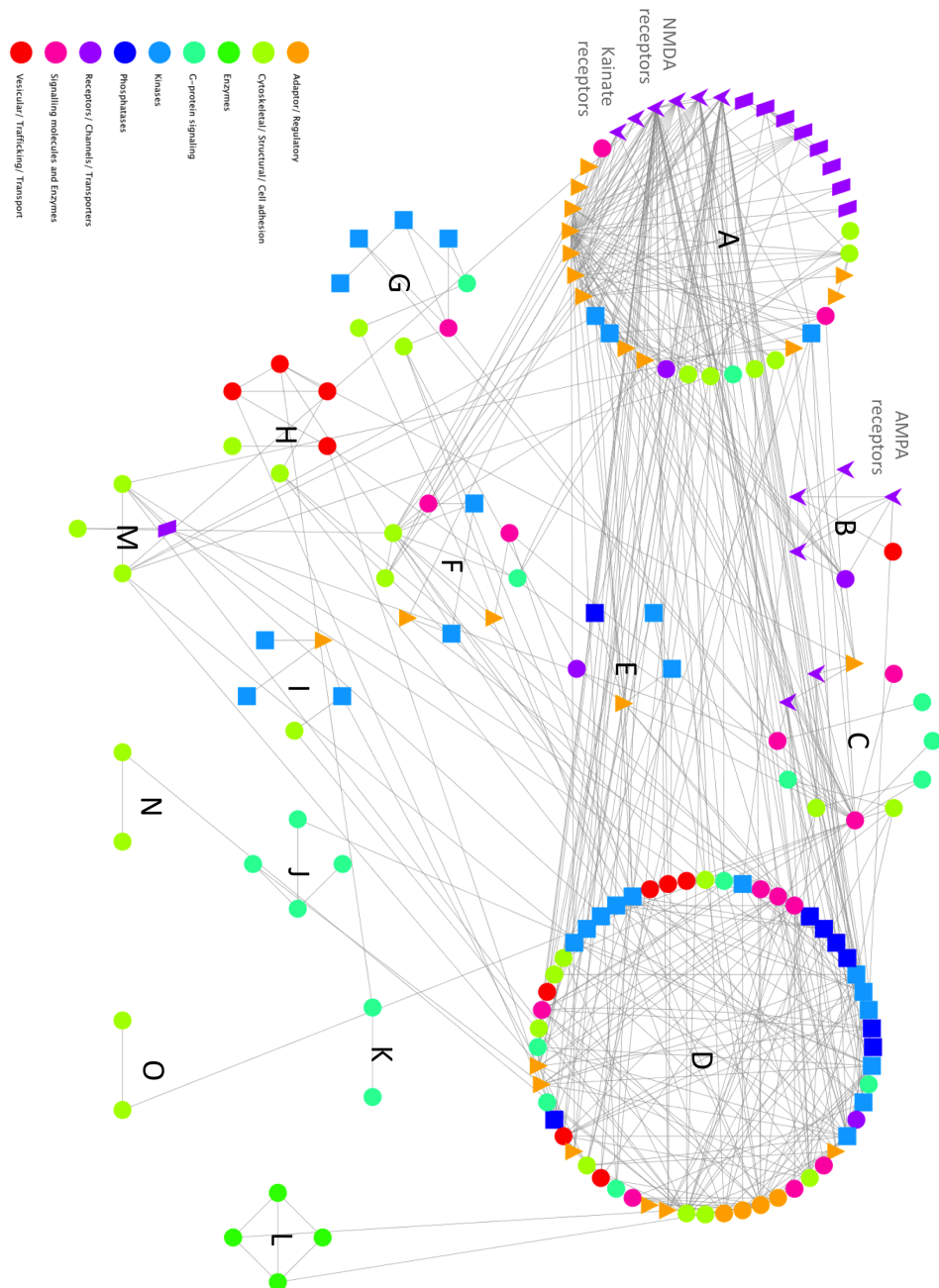
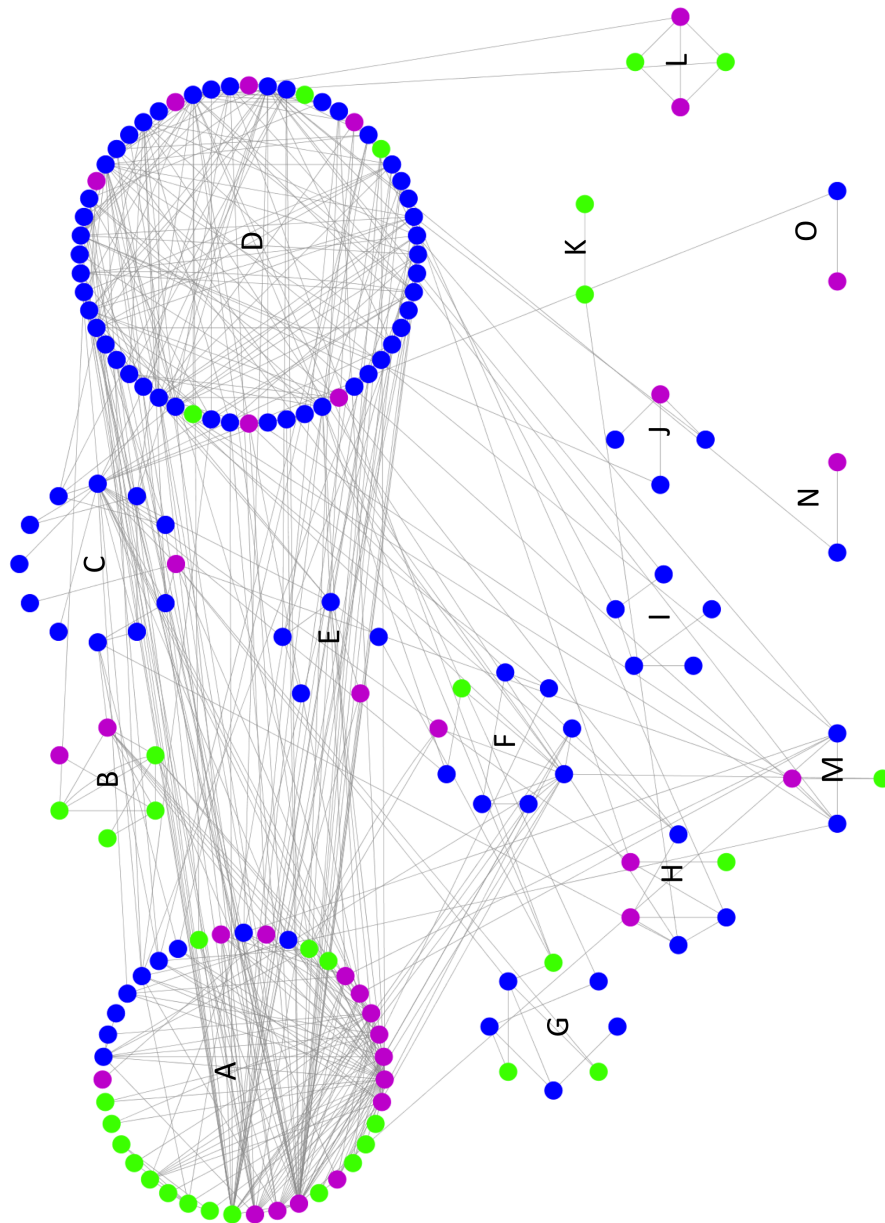


Figure 5.6: Node contribution in the Union protein interaction network. Nodes in the PSD-95 associated proteins complex are green, nodes from the NRC/MASC are blue and overlap nodes are purple.



biological functions. From the above the following observations can be made about these clusters:

- A: Consists of 39 proteins. Includes all NMDA and Kainate glutamate receptors. It also includes all but one of the K^+ channels in the network model. The Receptors/ Channels/ Transporters family correlates significantly with the cluster ($p = 2.31E - 005$) and takes up ~35% of the nodes within it. Another ~30% is taken up by proteins of the Adaptor/ Regulatory family, namely MAGUKs of the PDZ and non-PDZ domain containing scaffolders subfamilies, which also correlates with the cluster ($p = 0.01$). The former subfamily includes all the proteins of the Dlg gene family (Dlg1-4). This cluster also contains cytoskeletal proteins of the PSD's structural lattice like Shank1. According to the GO annotations, molecular function (Table B.10) seems to be dominated by receptor and ion channel function related annotation terms. Similarly, the biological processes (Table B.11) are clearly NMDA receptor, ion transport and synaptic transmission associated. Also, subcellular localisation (Table B.12) of this cluster seems to be highly post-synaptic membrane associated or bound.
- B: Consists of 6 proteins and is correlated with the Receptors/ Channels/ Transporters family ($p < 10^{-6}$). This cluster contains all the AMPA receptors (Gria1-4) in the dataset. It also contains Cacng2 and Nsf, which have been associated with AMPA receptor trafficking (Hashimoto et al., 1999, Chen et al., 2000 and Nishimune et al., 1998 respectively). Nsf has also been shown to influence AMPA receptor subunit (Gria2) via endocytosis (Braithwaite et al., 2002). GO annotations show that the cluster's molecular function, cellular component and the biological processes it is involved in are dominated by the AMPA receptors terms.
- C: Consists of 12 proteins. This cluster includes the metabotropic glutamate receptors in the model (Grm1 and Grm5). The cluster also includes members

of and is correlated with the G-protein signalling family ($p = 0.03$), but also contains signalling enzymes and cytoskeletal proteins. Again, this cluster is enriched in membrane related GO CC terms and its dominant GO BP is signal transduction. Homer1 is a member of this cluster. Homer1 is an PDZ-domain containing scaffolder that interacts with Shank1 (found in cluster A) to create a dynamic polymeric network scaffolding lattice structure for the PSD (Hayashi et al., 2009). Also Homer1 - mGluR interactions are responsible for the clustering of mGluRs (Kammermeier, 2006).

- D: Consists of 56 proteins and is the larger cluster in the network. This cluster is of a mixed nature, but ~46% of it consists of the signalling related protein families, Kinases (~21%), Phosphatases (~12.5%) and Signalling Molecules and Enzymes (12.5%). This cluster contains all the members of the phosphodependent chaperones 14-3-3 subfamily (Ywhae, Ywhag, Ywhah, Ywhaz) of the Adaptor/Regulatory family of proteins. The 14-3-3 proteins interact densely with the Kinases and Signalling Molecules and Enzymes members but not that much with the Phosphatases, which in turn have a similar number of interactions with the Kinases and Signalling Molecules and Enzymes members. This cluster contains 7 out of the 8 Phosphatases family members in the network model and highly correlates with the family ($p < 10^{-6}$), while the Kinases family members seem to be more distributed in the network. However, cluster D encapsulates the well-studied Erk/Mapk signalling pathway cluster. From a GO term enrichment perspective the cluster appears to have its molecular functions highly correlated with kinase related activity and a mixed cytoplasmic, cytoskeletal and membrane GO CC annotation. The results are similar in the GO BP annotation domain, where phosphorylation related terms prevail, along with a cytoskeletal regulation (regulation of cell shape) via the GO annotations of Ppp2ca and Nefl.
- E and F: Consist of 5 and 9 proteins respectively and are of a mixed functional

nature. There are no dominant correlated GO terms or families associated with these clusters, except for the case of GO CC for cluster F, which is correlated with the cytoplasm and the nucleus. Both clusters contain a mixture of Adaptor/Regulatory proteins as well as members of the Signalling Molecules and Enzymes and / or Kinases families. Cluster E also contains Na/K channel *Atp1a1*, which is responsible for the maintenance of the postsynaptic resting potential (Dobretsov and Stimers, 1997).

- G: Consists of 8 proteins and is also highly correlated with the Kinases family ($p = 0.02$). The cluster is highly correlated with the transferase activity in MF terms and the glycolysis process in GO BP terms. Its also correlated with the cytosol and the nucleus cellular components.
- H: Consists of 6 proteins and is highly correlated with the Vesicular/ Trafficking/ Transport protein family ($p < 10^{-6}$). The rest of the proteins of this family are found in cluster D. The majority of the cluster's interactions are with clusters A, C and D and the cluster is highly correlated with the neurotransmitter secretion GO BP term.
- I: Consists of 5 proteins and is highly correlated with the cytoplasm annotation in the GO CC domain of GO. Again, this seems to be a cluster centered around a scaffolding protein (mitogen-activated protein kinase 8 interacting protein 1 - *Mapk8ip1*) that is associated with three kinases (*Map2k7*, *Stk39*, *Mapk10*).
- J and K: Consist of 4 and 2 proteins respectively and both exclusively contain members of the G-protein signalling family. Cluster K is too small to show any significant correlations, but cluster J appears, as expected, correlated with GO terms related with G-protein signalling as the cluster itself is correlated with the protein family ($p < 10^{-6}$). The clusters appear disconnected on a first degree neighbour level from the other G-protein containing cluster C, with the metabotropic glutamate receptors. Cluster J connects to cluster D and cluster K

to cluster H, via an interaction with Stx1a. Also, note all proteins in cluster J are involved in the Muscarinic acetylcholine receptor 2 and 4 signaling pathway (Table B.14, Appendix B).

- **L:** Consists of 4 proteins. All proteins in the Cluster are of the Enzymes family ($p < 10^{-7}$), and more specifically ATP proton transporters. The cluster is highly correlated with the associated GO MFs and BPs and is also correlated with mitochondrion associated GO CC terms.
- **M, N and O:** Consist of 4, 2 and 2 proteins respectively. All proteins but one belong to the Cytoskeletal/ Structural/ Cell adhesion family. Cluster M includes Vdac1, an outer mitochondrial membrane protein, while the rest of the proteins are of cytoskeletal nature.

Pocklington et al. divided the NRC/MASC complex into three layers, or components, by grouping the clusters it was segregated in. These layers were input (or upstream component), information processing (or midstream component) and output (or downstream component), based on their overall biological function. We looked at how this division mapped on the clustering of the Union network, given the addition of new nodes, we expected some of the original mapping to have changed. Clusters A, B and C, mapped almost completely to the input layer with the exception of four output component proteins that were also clustered in the aforementioned clusters due to interactions with their members. An example of such protein is Shank1, which in the Union network has 4 additional interactions with cluster A members. Cluster H is also comprised of 50% input layer proteins, with the exception of Snap25, Nefm and Nrnx1, which also give it a more structural character. Clusters J and K - comprised of G proteins - can also be considered part of the input layer, along with cluster O, which is part of the input layer's cytoskeletal support. Given the above, the Union network's input layer comprises of clusters A, B, C, J, K and O. Cluster H is likely to belong to this layer. The information processing layer maps almost exclusively to cluster D, with

the exception of three proteins (2 in F and 1 in M). Cluster G, which also contains 4 Ser/Thr kinases can be considered part of the information processing layer. Therefore, in the Union network the information processing layer corresponds to cluster D and G. When the NRC/MASC network's output layer was mapped on to the Union network we noticed that part of it mapped to clusters I (exclusively output layers proteins) and D, with the latter cluster appearing to be of a mixed information processing and output nature. The rest of the clusters E, L, M and N can be considered part of the molecular machine that is further downstream of the output and are responsible for individual effector responses (e.g. cytoskeleton rearrangement, mitochondrial function, vesicular trafficking).

5.3.2.3 Architecture

To our experience the most informative metrics are node degree and betweenness as well as the average shortest path (ASP) to node. Since the latter two tend to follow similar trends we present the twenty nodes from the network with the highest betweenness in descending order in Table B.15 of Appendix B. As previously discussed in 3.3, nodes of high betweenness are central points in the network where the information flow converges and they tend to connect modules of the network. Also, the ASP of a node can be a measure of its importance in pathway crosstalk.

When looking at the top-20 nodes of highest betweenness we see various types of proteins. Two PDZ domain and two non-PDZ domain containing scaffolders are in this set of twenty proteins, with a total of 5 proteins belonging to the Adaptor/Regulatory family being in this set. The PDZ containing domain proteins are both members of the Dlg gene family. As mentioned in Chapter 4, Dlg4 is central to the organisation of the scaffolding lattice of the PSD and is at the top of the list, with a betweenness of 0.19. Scaffolding proteins in general are expected to hold central positions in a network since they connect the receptors with the downstream signalling. Another prominent group in this set are the Grin1, Grin2b and Grin2d NMDA receptor subunits. These

receptors all have a high degree in the network and are expected to hold positions of high betweenness since they interact with many adaptor, scaffolding and signalling proteins.

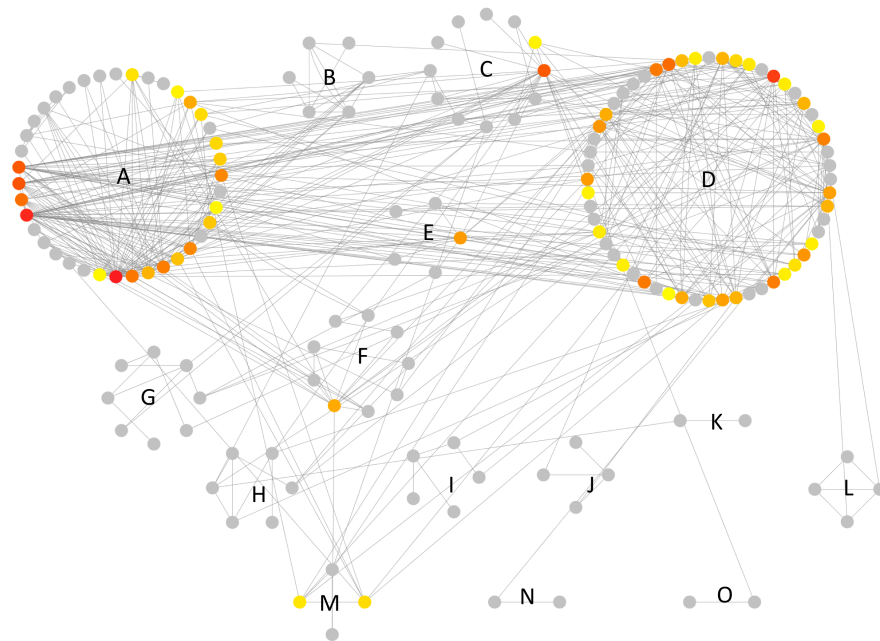
Looking at the family distribution in this set of 20 nodes, we see that 6 (30%) are of the Kinases family, with a total of 9 (45%) being comprised of proteins with signalling functionality (Kinases and Signalling molecules and Enzymes families). This could be interpreted on the basis that these types of proteins are points of information flow convergence from the receptors towards more downstream signalling (Salter and Kalia, 2004)

From an ASP perspective, all but three nodes in the top 20 list have an ASP < 3, with two exceptions, with the network's average ASP being 3.35. Looking at the general distribution of nodes with a low ASP (< 3) in Figure 5.7, we can see that they appear gathered in clusters A and D as would be expected. Low ASP is indicative of the importance of a node to the information flow in the network, since it usually indicates high crosstalk between modules. An example of such crosstalk is between clusters A and D, via tyrosine kinases Ptk2b in A and Src in D. Again, tyrosine kinases in this case are a point of convergence for multiple signalling pathways regulating NMDA receptor activity found in cluster A (Salter and Kalia, 2004).

5.3.2.4 Module and family interactions

An interesting aspect of a network model is that we can observe interactions between modules of the network and interpret them from a functional perspective using annotation information. In the following paragraph we will look at the interactions between the clusters of the network which represent functional modules. In the Union network there are 3 receptor clusters (A, B, C) and a major cluster of signalling nature (D). However, we notice signalling related proteins (Kinases and Signalling molecules and Enzymes families) spread in other clusters as well (E, F, G). From a functional module connectivity point of view it would be interesting to see how the input component

Figure 5.7: Average shortest paths to nodes in the Union interaction network. Nodes with ASP > 3 are grey. Nodes with ASP < 3 vary from yellow (lower) to red (higher).



of the network (receptors) connects to the more downstream signalling component. The NMDA receptor cluster, A, seems to be well connected (66 edges) to D, with its first degree neighbours covering most of the latter cluster. Similarly, cluster C (metabotropic glutamate receptors) is connected via 12 edges via proteins from various families including Adaptor/Regulatory, Kinases and Signalling molecules and Enzymes. Cluster B (AMPA receptors) seems to be more cut off, connecting to D via and interaction between *Gria4* and *Prkcc*. Except direct connections there seem to be some indirect connections of cluster A to cluster D via smaller clusters that are connected to both with equal or comparably similar numbers of edges. There is also the case of cluster G that connects the metabotropic glutamate receptor cluster C to the signalling cluster D. Clusters C and G are connected via an interaction between *Gapdh* and *Tpi1*. *Tpi1* interacts via another protein in cluster G (*Cfl1*) with clusters D and E. Cluster G also interacts with cluster D via interactions with cytoskeleton regulating protein *Cse1l*. Another cluster that connects to many other is M. M contains cytoskele-

tal proteins and a voltage-dependent anion channel (Vdac1) and connects to clusters A, C, D and F. Interactions are achieved mostly via Vdac1, actinin Actn4, and gelsolin (Gsn).

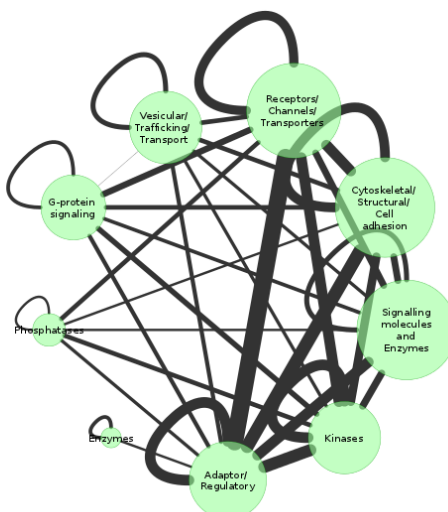
Another perspective of observations is that of interactions between protein families (Figure 5.8). If the diagonal of the illustration is ignored (as it represents interactions within the same family), one can observe that members of the Kinases family interact with the Adaptor/ Regulatory (31 edges), Receptors/ Channels/ Transporters (25 edges), Cytoskeletal/ Structural/ Cell adhesion (16 edges) and the Signalling molecules and Enzymes (13 edges) families. The Adaptor/ Regulatory interacts a lot with Receptors/ Channels/ Transporters (52 edges), Cytoskeletal/ Structural/ Cell adhesion (35 edges) and Signalling molecules and Enzymes (26 edges) families. These numbers sketch the foundation of the PSD's organisation, i.e. the cytoskeletal proteins interacting with the adaptor proteins, creating a lattice and scaffold. This scaffold becomes a substratum not only for the receptors and ion channels, but for part of the signalling component as well. Finally, interactions of the receptors and channels with signalling and phosphosignalling molecules of the PSD allow the propagation of downstream signals to the cell, regulating aspects of synaptic transmission such as plasticity.

The aforementioned notion of basic architecture could also be illustrated using a metanetwork. In Figure 5.9 we can observe how the dominant (thicker) metaedges connect proteins of the Adaptor/ Regulatory with the structural proteins of the Cytoskeletal/ Structural/ Cell adhesion family and the receptors and ion channels of the Receptors/ Channels/ Transporters family. The Adaptor/ Regulatory family also connects with proteins of a signalling nature (Kinases and Signalling molecules and Enzymes families) - providing a hub for the downstream propagation of the signal. In addition to these interactions, there are also direct interactions of the signalling and structural proteins with the receptors and channels adding to the complexity of the network.

Figure 5.8: Table of the protein family to protein family connectivity in the Union network (normalised). Cells are coloured according on a scale of white (smallest) to red (highest) . Note how prominent the central interactions between Adaptor and Cytoskeletal, Kinases and Receptors are.

	Adaptor/ Regulatory	Cytoskeletal/ Structural/ Cell adhesion	Enzymes	G-protein signaling	Kinases	Phosphatases	Receptors/ Channels/ Transporters	Signalling molecules and Enzymes	Vesicular/ Trafficking/ Transport
Adaptor/ Regulatory	0.054585153	0.076419214	0.004366812	0.013100437	0.06768559	0.008733624	0.113537118	0.054585153	0.013100437
Cytoskeletal/ Structural/ Cell adhesion	0	0.045851528	0	0.015283843	0.034934498	0.004366812	0.054585153	0.026200873	0.015283843
Enzymes	0.004366812	0	0.010917031	0	0	0	0	0	0
G-protein signaling	0	0	0	0.010917031	0.019650655	0	0.024017467	0.010917031	0.002183406
Kinases	0	0	0	0	0.054585153	0.015283843	0.054585153	0.028384279	0.008733624
Phosphatases	0	0	0	0	0	0.004366812	0.015283843	0.004366812	0
Receptors/ Channels/ Transporters	0	0	0	0	0	0	0.041484716	0.026200873	0.015283843
Signalling molecules and Enzymes	0	0	0	0	0	0	0	0.017467249	0.008733624
Vesicular/ Trafficking/ Transport	0	0	0	0	0	0	0	0	0.017467249

Figure 5.9: Metanetwork of nodes in the union network. Each metanode represents a family and groups all nodes of it. Metaedges represent normalised ratios of edges between proteins of the same family in the original network model. The weight of the metaedges is proportional to the edge count in the original network.



5.3.3 The PSD interactome and physiology

We annotated the proteins in the Union dataset using previous in-house curation and G2Cdb data (for references see within Pocklington et al., 2006, Fernández et al., 2009). Figure 5.10 shows the Union protein interaction network with the nodes involved in synaptic plasticity, behaviour or both highlighted. Most of the nodes in the model are correlated with both plasticity and behaviour. Most nodes involved in either one or both are spread across clusters A - E, with the exception of synaptic vesicle protein Syt1 in cluster H and Vdac1 in cluster M. As expected these clusters are the ones associated with receptors, signalling and scaffolding of the PSD.

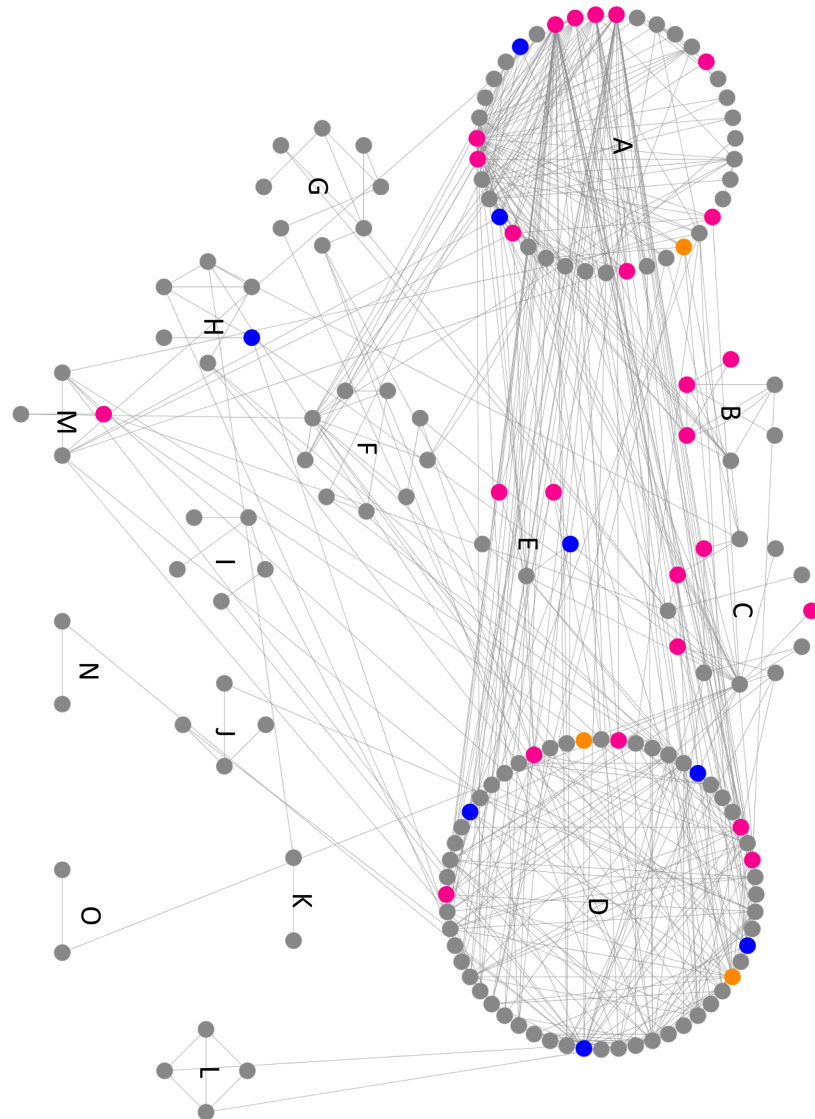
The Receptors / Channels / Transporters and Signalling molecules and Enzymes families correlated with behaviour ($p < 10^{-6}$ and $p = 0.05$ respectively). The Receptors / Channels / Transporters family also correlated with plasticity ($p = 0$). The Glutamate receptors subfamily of the Receptors / Channels / Transporters family correlates both with plasticity and behaviour ($p < 10^{-6}$ in both cases) and also the Tyr Kinases subfamily of the Kinases family correlates with behaviour ($p = 0.01$). Cluster E was found to correlate with behaviour ($p = 0.05$). This is interesting because as previously discussed, cluster E is one of the signalling associated clusters that connect with A and D via equal number of interactions - and could function as an information relay in their communication. A position like that would be likely to affect the flow of information from input (receptors) to the more downstream components of signalling. Among the top twenty nodes of highest betweenness in Table B.15 on page 253, found in Appendix B, 9 have plasticity and 9 have behavioural phenotype associations.

5.3.4 The PSD interactome and disease

5.3.4.1 Disease annotation and correlation

We collected disease associations for all proteins in the Union dataset from the previously curated NRC/MASC and PSD-95 associated proteins datasets (for references see

Figure 5.10: Proteins associated with behaviour (blue), plasticity (orange) or both (pink) in the Union network. An interactive version of this figure with fully visible node labels is available in the additional material website (available in DVD format with this thesis).



within Pocklington et al., 2006, Fernández et al., 2009). An additional semi-automatic curation was done in case any new data had appeared. The disease associations of proteins in the Union dataset are shown in tables 5.2 and 5.3. The dataset associated with a total of 23 diseases. Out of these we will focus on schizophrenia, bipolar and bipolar affective disorder (grouped as bipolar disorder), mental retardation and depression since all of the aforementioned diseases correlate with more than 10 proteins in the dataset. A total of 59 nodes in the network are associated with one or more of these diseases. From a functional correlation point of view, Receptors/ Channels/ Transporters correlate with schizophrenia ($p < 10^{-4}$) and Kinases with depression ($p = 0.02$) - as previously shown for the PSD-95 associated proteins dataset. Another correlation that appeared by merging NRC/MASC to the aforementioned dataset is a correlation of the Receptors/ Channels/ Transporters family with bipolar disorder ($p = 0.01$) - corroborating previous correlations between proteins involved in schizophrenia and bipolar disorder (Pocklington et al., 2006). A more fine grain examination of the functional family annotation in correlation with disease reveals that in the Receptors/ Channels/ Transporters family it is the Glutamate Receptors subfamily that is correlated with schizophrenia ($p = 3.93E - 7$) and the Voltage-dependent anion channels subfamily that is correlated with bipolar and bipolar affective disorder ($p = 0.01$).

5.3.4.2 Disease in the network

It is interesting to look at how disease associations correlate with modular structure. Besides the correlations of specific modules with disease it is also interesting to examine the distribution of disease nodes and their primary interactors across the network. This way we can isolate a disease associated sub-network and assert its spread within the model.

Schizophrenia correlates with all clusters A ($p = 0.03$) and B ($p = 0$) where NMDA and AMPA glutamate receptors appear. It also correlates with clusters D ($p = 0.02$) and H ($p = 0.02$). Other than the NMDA receptors in cluster A, cluster D contains

Table 5.2: Diseases and associated genes in the Union complex. Continued in Table 5.3.

Disease	Num of Genes	MGI Gene Symbols
Schizophrenia	43	Acot7, CamK2b, Cnp, Dlg1, Dlg2, Dlg3, Dlg4, Dlgap1, Dpysl2, Dusp4, Flna, Gap43, Gda, Gnao1, Gnas, Gria1, Gria2, Gria3, Gria4, Grik2, Grin1, Grin2a, Grin2b, Grin2d, Grm5, Hspa1b, Kcnj4, L1cam, Mapk1, Mapk3, Nefl, Nos1, Nrxn1, Nsf, Pik3ca, Pla2g4a, Ppp3ca, Slc1a2, Snap25, Stx1a, Stxbp1, Vdac1, Ywhah
Bipolar or Bipolar Affective Disorder	19	Atp1a3, Atp5c1, Cacng2, CamK2a, Dlg3, Dlg4, Gnas, Gnb211, Grin1, Grin2b, Gsk3b, Msrb2, Nefl, Pgk1, Pik3ca, Slc25a4, Snap25, Vdac1, Vdac2
MentalRetardation	26	Actg1, Calb2, Capza2, Cltc, Dlg3, Dpysl2, Gnas, Gnb211, Gria3, Grik2, L1cam, Ldhb, Msrb2, Nefl, Nefm, Nf1, Nrxn1, Pfk1, Pgk1, Pik3ca, Pklr, Ptpn11, Rab3a, Rac1, Rap2a, Rps6ka3, Tpi1
Depression	12	CamK2b, Dlg3, Dusp4, Gnas, Hspa1b, Mapk1, Mapk3, Pdha1, Pla2g4a, Plcb1, Plp1, Ptk2b
Epilepsy	6	Adam22, Gria1, Gria2, Grin2b, Kcnj10, Lgi1
ALS	2	Nefl, Sl1a2

Table 5.3: Diseases and associated genes in the Union (continued from Table 5.2).

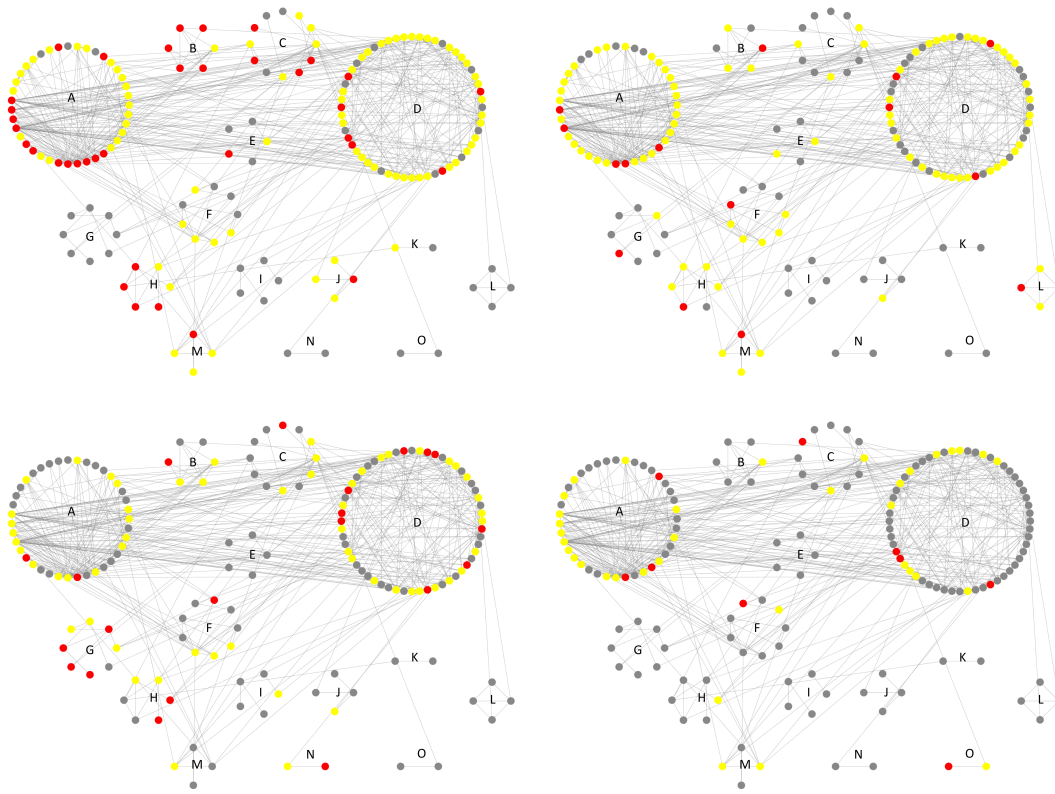
Disease	Num of Genes	MGI Gene Symbols
Huntington disease	2	Grin2a, Grin2b
Seizure	2	Grin1, Kcnj10
X-Mental retardation	2	Dlg3, Gria3
Attention disorder	1	Grin1
Autism	1	Nrxn1
CMT1	1	Nefl
CMT2	1	Nefl
Demyelinating disease	1	Plp1
Episodic ataxia type 1	1	Kcna1
Miller-Dieker lissencephaly	1	Ywhae
Multiple sclerosis	1	Plp1
Neurodegeneration	1	Atp1b1
Opthalmoplegia	1	Slc25a4
Parkinson's	2	Pgk1, Prdx2
Pelizaeus-Merzbacher disease	1	Plp1
Rett syndrome	1	Atp1b1
Spastic paraplegia	1	Plp1

some key signalling molecules involved in the disease like Mapk1, Mapk3 and Nefl. Cluster H, as mentioned earlier, contains 4 out of 6 synaptic vesicle proteins in the network (not including Clathrin or motor proteins). All 4 synaptic vesicle proteins in cluster H are associated with schizophrenia. The total of 36 nodes associated with schizophrenia seem to be in 8 out of 15 clusters in the network (Figure 5.11 top left). The primary interactors to these nodes seem to be distributed in 10 out of 15 clusters, covering >70% of the network, mostly due to the fact that some of the associated genes are of very high degree (e.g. Grin1, Grin2b, Grin2d, Dlg3 and Dlg1). Also, notice how cluster F, with no directly associated proteins, consists of >50% of their primary interactors. Another strong impression is made by how much clusters A - D are affected.

Bipolar disorder does not correlate with any of the clusters in particular. The total of 15 genes associated with it are distributed over 8 out of 15 clusters in the network (Figure 5.11 top right), with their primary interactors in 11/15 clusters covering ~46% of the network. With the exception of cluster G, K and L the clusters affected are the same with schizophrenia, with the main NMDA and AMPA receptors and signalling clusters A and D affected in similar extent with schizophrenia.

Mental retardation (Figure 5.11 bottom left) correlates with cluster G ($p = 0.02$). The disease correlates with 3 out of 4 Ser/Thr Kinases found in that cluster (Pfk1, Pklr, Pgk1) and signalling associated enzyme Tpi1. There is a total of 21 proteins associated with mental retardation, interacting with 56 proteins on a first degree level. What is also noticeable, is that clusters A and D are not as affected as they are in the schizophrenia and bipolar disorder cases. Another relevant observation is that when comparing mental retardation to bipolar disorder, we see that even if more proteins are actually directly associated in the former the number of first degree interactors is lower. That is a result of the type of nodes involved. In the case of mental retardation all associated nodes with the exception of Actg1, Dlg3 and Pik3ca have a degree smaller than 10 nodes.

Figure 5.11: The Union protein interaction with protein nodes associated with disease highlighted in red and primary interactors in yellow. Top left: schizophrenia, top right: bipolar disorder, bottom left: mental retardation, bottom right: depression. An interactive version of this figure with fully visible node labels is available in the additional material website (available in DVD format with this thesis).



Depression (Figure 5.11 bottom right) is also not correlated with any particular clusters. There are a total of 9 proteins associated with depression interacting with another 39 nodes on a first degree level. Four of the associated proteins belong to the Kinases family. Again, in this case, clusters A and D are less affected.

It is also interesting to note that among the top twenty nodes of highest betweenness in Table B.15 on page 253, 9 are associated with schizophrenia, 7 with bipolar disorder, 2 with mental retardation and 1 with depression, while 9 of the nodes have no disease association (with two of them associated with plasticity and behaviour).

5.3.5 Evolution of the PSD interactome

Following up on analysis originally performed by Emes et al. (2008) we decided to examine the Union complex from an evolutionary perspective in order to elucidate open questions about the evolution of synaptic molecular machines. Based on this approach and the additional network model, which at the time the original study was performed was unavailable, we performed a comparative genomics analysis of the Union network and also superimposed the results on the protein interaction network in order to explore possible links between evolutionary origin and network architecture.

For each protein in the Union dataset, we retrieved the corresponding gene's orthologues across 18 species (the genome of *A. melifera* was unavailable via Compara). This, besides retrieving the database accession numbers for all orthologs of a gene also allowed us to define another attribute. This attribute, hereon referred to as the “boundary of appearance in evolution” or simply “boundary”, represents the first appearance of a gene in evolution. Due to the nature of our data we divided on three lineage boundaries: Eukaryotic, Metazoan and Chordate.

The distribution of orthologs across the 18 species, visualised as a percentage of the total human genes (Figure 5.12), corroborated the results by Emes et al. (2008). More specifically, and from a perspective of appearance in evolution, in the Union dataset ~21% of the proteins were present in Eukaryotes, ~41% in Metazoans and ~39% were in Chordates. Rises in Union complexity, by the addition of more genes are visible as “jumps” at the Eukaryotic-Metazoan and Metazoan-Chordate boundaries (arrows in 5.12), are speculated to be coinciding with genome duplications.

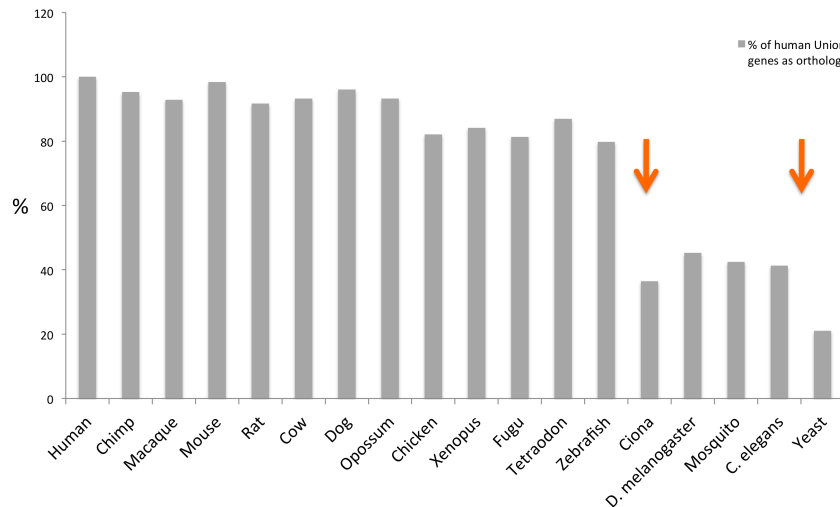
The rises in complexity are a result of a specific elaboration mechanism based on expansion by duplication and diversification of existing gene families (e.g. receptors, scaffolders, etc) rather than the addition of new genes (Emes et al., 2008). This conclusion is drawn from observations like the following. Although representatives of all families were present in Eukaryotes (Figure 5.13, top), specific subfamilies only appeared in the complex after the Metazoan boundary, once a nervous system ap-

peared. Examples of these subfamilies are the PDZ-domain containing scaffolders of the Adaptor/ Regulatory family the Glutamate Receptors, Inward rectifying K^+ channel, Voltage-dependent anion channels and Voltage-gated K^+ channel subfamilies of the Receptors/ Channels/ Transporters family (Figure 5.13, bottom). It is also interesting to notice how these key components got even more elaborate in variety after the Chordate boundary. For example, ~33% of Glutamate receptors and 64% of the non-PDZ domain containing scaffolders subfamilies variety was added after the Chordate boundary. From an enrichment perspective, the Enzymes, Signalling molecules and Enzymes, Transcription/ Translation families are significantly enriched in proteins appearing at the Eukaryotic boundary ($p = 0$, $p = 0$ and $p = 0.01$ respectively). Also, the Adaptor/ Regulatory, Signalling molecules and Enzymes and Cytoskeletal/ Structural/ Cell adhesion families are significantly enriched in proteins appearing at the Chordate boundary ($p = 0.04$, $p = 0.04$ and $p = 0.02$ respectively).

All this reflects the existence of a generic signalling toolkit along with the basic cytoskeletal and adaptor scaffolding in Eukaryotes (some of the PSD components are also parts of basic cellular toolkits like the ribosome or mitochondria). This signalling toolkit got more elaborate in Metazoans after the formation of the first synapse (protosynapse). Chordates possess a molecular machine of even greater complexity, with a great variety of key components like glutamate receptors, PDZ-domain and non-PDZ domain containing scaffolders and signalling molecules.

The novelty in this approach was attempting to use the Union protein interaction network model as a scaffold for these evolution annotations. We annotated the network with the boundary of evolutionary appearance for all proteins (Figure 5.14). As expected, cluster A appears almost exclusively of Metazoan or Chordate origin with the exception of Ca^{+2} Atpase Atp2b4. Similarly, cluster B contains only one protein of Eukaryotic appearance, synaptic vesicle protein Nsf, which as previously mentioned is associated with trafficking of other proteins in the cluster. Clusters C and D appear to have a more mixed composition since they're associated with signalling, which could

Figure 5.12: Distribution of Union gene orthologues across 18 species. Orthologs are plotted as a percentage of total human Union genes as per (Emes et al., 2008), the results of which are recapitulated as indicated by the jumps (arrows) in complexity, potentially coinciding with genome duplication events.



be part of the protosynaptic toolkit. Also, clusters G and L are correlated with the proteins of Eukaryotic appearance ($p = 0.01$ and $p < 10^{-4}$ respectively). As mentioned earlier, cluster L is associated with the mitochondrion, an organelle present in all Eukaryotes examined.

The next step in the analysis was to use the interaction network and the boundary annotations in order to draw some conclusions on how the increase of complexity over evolutionary time, by the addition of new nodes, affected the connectivity of the network. Two metrics that can describe this are the node degree and the ASP to node. As mentioned earlier, the first shows how connected a node is and the second how central a node is in the crosstalk of pathways, by being in a short distance from other nodes. The cumulative distribution functions of the node degree and ASP are illustrated in Figure 5.15. We noticed that the degree cumulative distributions are similar (Kolmogorov-Smirnov test $p < 0.05$) and independent of the evolutionary origin of the proteins. However, the ASP to node distributions are similar for proteins of Metazoan or Chordate origin and have values lower than those for proteins of Eukaryotic origin (Kolmogorov-Smirnov test $p < 0.05$ in both cases). These results can be interpreted in the following way: while the degree of a protein in the network is independent of its

Figure 5.13: Top: Distribution of boundaries of appearance in evolution (Eukaryotic, Metazoan or Chordate) for families of proteins in the Union dataset. Bottom: Distribution of boundaries of appearance in evolution (Eukaryotic, Metazoan or Chordate) for subfamilies of the Adaptor/ Regulatory and Receptors/ Channels/ Transporters families of proteins in the Union dataset.

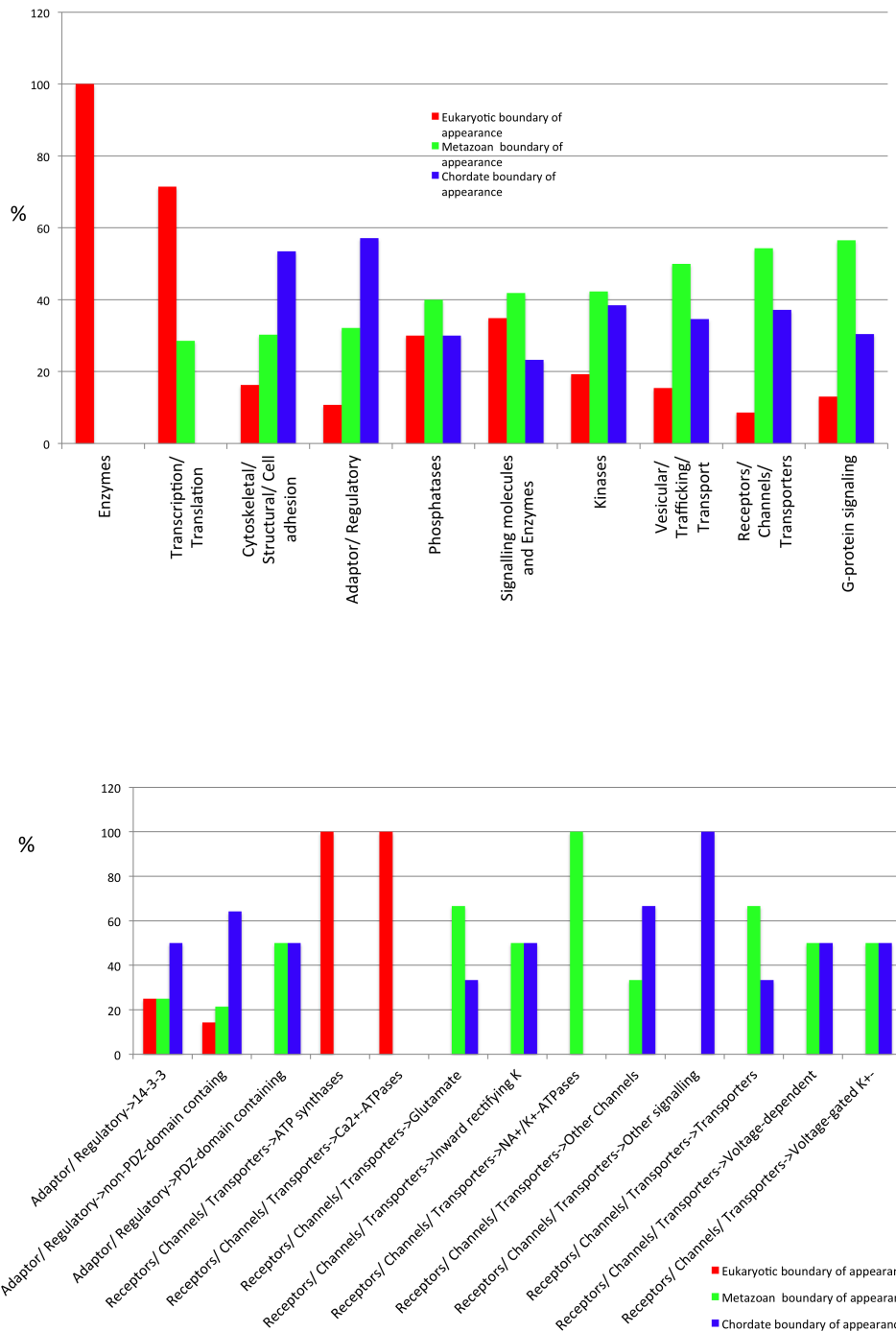
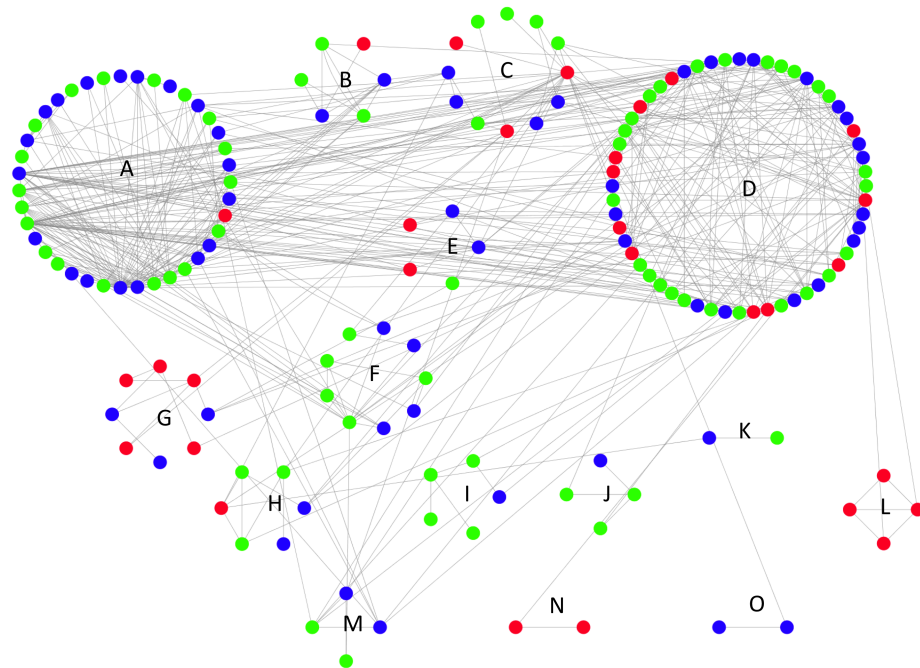


Figure 5.14: The Union protein interaction network with nodes annotated according to their boundary of appearance in evolution: Eukaryotic (red), Metazoan (green) and Chordate (blue) .



evolutionary origin, the average distance separating proteins is not. The average distance separating proteins of Eukaryotic origin from other proteins in the network tends to be larger and also proteins of the Metazoan and Chordate origin classes tend to be closer to each other and be separated with shorter paths from the rest of the proteins in the network. From an architecture point of view this means that proteins of the “basic” cellular (e.g. mitochondrial) and signalling (e.g. kinases) toolkit, that were available before the early protosynapse are present but not as central in pathway crosstalk. Proteins of the Metazoan and Chordate synapse, which were mostly added by gene family duplication and diversification tend to be closer connected to each other from an ASP to node perspective. This is probably a result of proteome evolution, where a node gets duplicated but still retains some of the original’s interactions, thus keeping homolog nodes closer in the network.

From a more physiological perspective of the 36 genes associated with plasticity or behaviour in the complex 15 are of Chordate, 16 are of Metazoan and 5 of Eukaryotic

origin. Additionally, out of the genes associated with one or more of the four mental diseases we focus on, 21 are of Chordate, 25 of Metazoan and 13 of Eukaryotic origin. Again, due to partial knowledge we cannot test these numbers but it seems to be the case that out of the proteins which have a physiological effect mostly are of Metazoan or Chordate origin. This can be partially because of the wider effect of mutations on proteins of Eukaryotic origin since they would be expected to be of a more “housekeeping” nature.

5.4 Concluding remarks

The Union model is the biggest manually curated and annotated ‘gold standard’ model of the PSD’s protein interaction network coming from mouse affinity purified complexes data. This model described the core of the mouse PSD (mPSD) protein interaction network since it contains key proteins of the receptor, scaffolding, electrical and signalling components.

5.4.1 Issues

One issue that has to be kept in mind when discussing annotations is partial knowledge. All data on protein interactions, functional, behavioural or disease annotation are likely to be incomplete. This is due to inherent experimental method drawbacks and research bias towards specific genes (interactions, annotation, association with behaviour and disease). In other words, protein interaction detection methods could have some false negative results by missing some protein interactions. There are also false positive interactions, most of which we assume we have resolved by combining multiple protein interaction resources and manual curation. As for behaviour and disease annotations, the main causes of false negatives are inherent research strategy bias towards “known suspects” or lack of solid experimental evidence (rather than weak literature associations). Nevertheless, given that careful manual curation is likely to

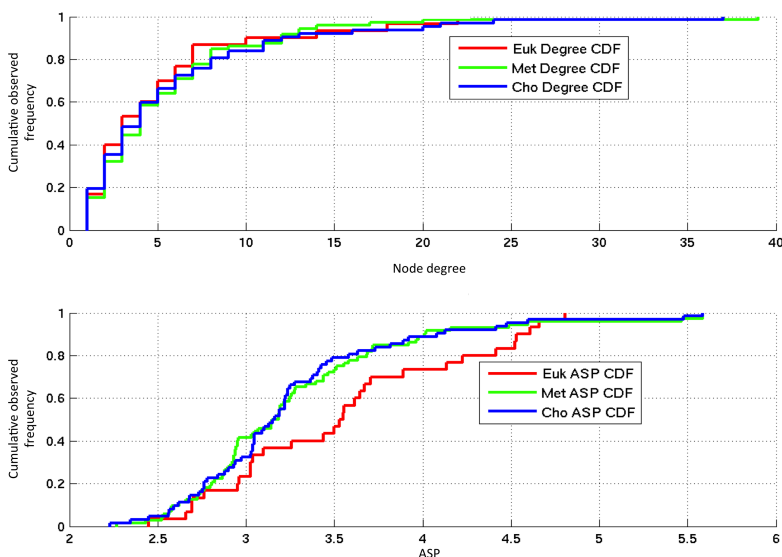
remove some or most false positives, working with partial knowledge still provides information. Some correlations might be obfuscated due to low number or missing instances of a specific annotation, but the strongest are still visible and might give an overall approximation. Specifically regarding missing protein interactions, we can see from the clustering analysis that the grouping of the proteins fits to previous biological knowledge and research findings (clusters of receptors and signalling etc), so our feeling is that the overall structure of the network is well represented in this model.

Another observation that has to be kept in mind is regarding the data used to reconstruct this complex. In both the PSD-95 associated proteins complex and the NRC/MASC cases, we have addressed the complexes as if they were identical at all synapses, which is known not to be the case in all parts of the central nervous system (Porter et al., 2005). Studies of PSD proteins using microarray data and protein localisation show a high degree of co-expression for most proteins in the Union complex in forebrain structures, including hippocampus, cortex, striatum and amygdala (Zapala et al., 2005). If one takes into account the number of proteins and interactions leading to the resulting complexity it seems likely that any given synapse will contain a distribution of complexes of varying composition.

5.4.2 Biological significance of clustering

Computation of the clustering configuration shows how a modular architecture is present and has persistence between the different overlapping datasets used to form the Union dataset. This modular architecture appears to have biological significance since specific correlations of annotations appear in some clusters. It has to be taken into account that these enrichments have been computed based on the dataset annotation background (rather than the whole genome). Union is a dataset that already has a lot of annotation bias towards specific PSD related annotations, because of the proteomics isolation methods (proteins PSD protein baits). The appearance of even more specific significant annotation enrichments of terms within clusters shows that the chosen con-

Figure 5.15: Cumulative distributions of node degree (top) and average shortest path (ASP) (bottom) for the Union protein interaction network. Degree and ASP were computed on the whole Union network.



figuration is better than random. To our knowledge, the scientific community considers this sufficient, since it appears to be the common approach in the biological network analysis literature (not only limited to the PSD).

Ideally we would need to quantify how much better than random this configuration is. This is hard because of two reasons there is no known “perfect” solution we could compare it to. Nevertheless a potential approach could be multiple testing based. This means generating random clustering configurations (of the same cluster number and size, otherwise the problem becomes computationally intractable) and testing them for some metric such as semantic similarity of annotations using TCSS (Jain and Bader, 2010). An approach like this is limited in the sense that the number of randomised configurations to be checked are limited (nodes within a cluster must still interact, we can not just chose random nodes in the network). Alternatively there have been recent approaches that can measure the “goodness” of a configuration based on architecture and enrichment of KEGG pathway annotation terms (Martha et al., 2011), as well as other more data dependent methods (Tang et al., 2011).

Additional evidence for the significance of the clustering can also be given by com-

paring the average semantic similarity of individual cluster GO annotations with the corresponding average in the network. In all domains the majority of clusters appears to be higher than the background of the proteins within the network (Figure 5.16).

5.4.3 A core PSD protein interaction network dataset and model

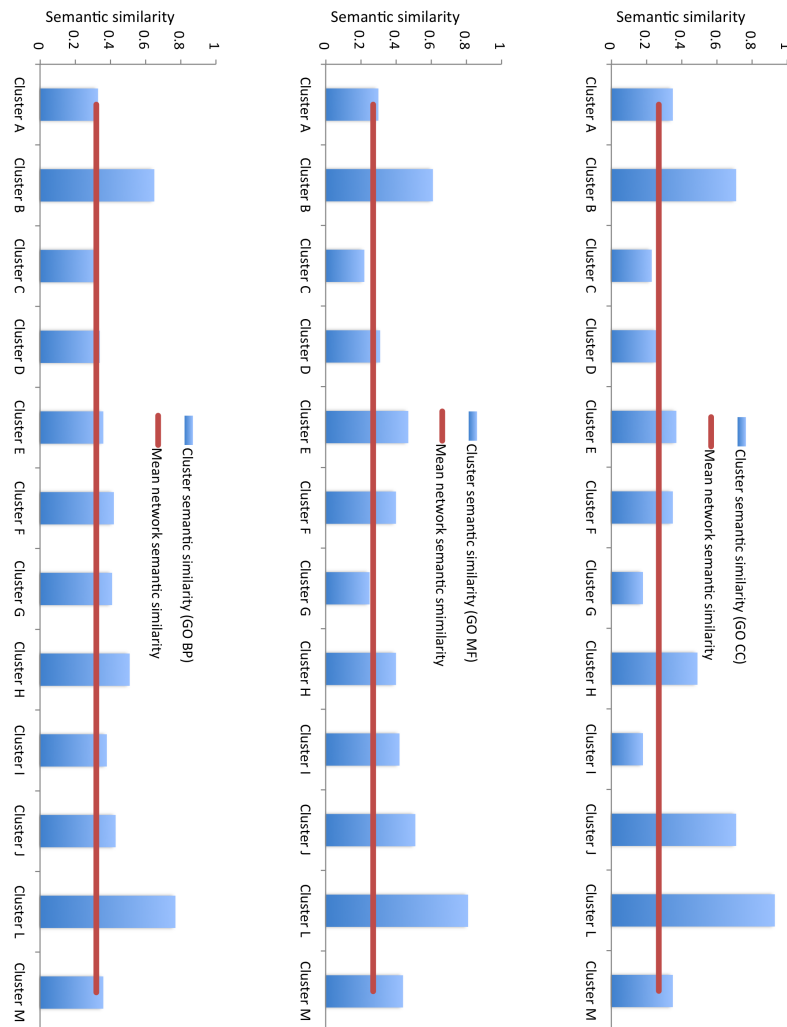
Merging of the two different datasets contributed various different or overlapping components to the resulting Union dataset. This highlights how we can start obtaining a more complete PSD model by augmenting the datasets with new proteomics data and then merging them or adding them to existing models, as hypothesised earlier. Observations on the Union dataset and model also highlight the basic principles of the PSD's structure, organisation, and architecture. The most common protein domains and GO MF annotations highlight the main functionality in the complex. Similarly, the classification of proteins into families and subfamilies outlines the basic types of proteins involved in the complex. Combining this with the protein interaction network we can get a basic grasp of how these protein classes interact and form the basic foundation for PSD's functions as a molecular machine. Also as mentioned earlier we saw that the modularity in the organisation of this molecular machine is biologically significant and persistent. Furthermore, proteins and their modules in the model correlate to synaptic plasticity, behaviour and disease, highlighting the involvement of the PSD in such processes

Most genes associated with synaptic plasticity and/or behaviour are distributed around clusters A-E, with the glutamate receptors associated with both and Tyr Kinases with behaviour. It is also interesting to note that the representatives of the latter subfamily found in this complex, appeared after the Chordate boundary. When we examined the 4 more prevalent diseases in the network, we identified schizophrenia as the disease most closely associated with the network, with the directly associated genes and their primary interactors covering ~70% of the network and significantly correlating with clusters A, B, D and H. Another cognitive disorder, mental retardation, covers

a similar percentage (~71%) of the network when the primary interactors of the associated genes are taken into account. Mental retardation correlates only with cluster H and has a smaller occupancy of clusters A and D in terms of associated proteins and primary interactors. Bipolar disorder and depression were not found to correlate with specific clusters. Although depression had less spread over the network, the spread of primary interactors of bipolar disorder seems to cover a portion of the network comparable to that of schizophrenia's because it is correlated with fewer nodes which, however, have a higher degree. Regarding depression specifically, its first degree association with association with the NMDA receptors in cluster A might corroborate evidence of depressive symptoms being improved by altering the actions of glutamate using sub-anesthetic doses of antagonist Ketamine (Zarate et al., 2006). Proteins associated with schizophrenia significantly overlap with depression (8 proteins, $p < 10^{-4}$), bipolar and bipolar affective disorder (9 proteins, $p < 10^{-4}$) and mental retardation (9 proteins, $p = 0.02$). This overlap is reflected in an overlap in the molecular pathways involved in these diseases. Also, the same proteins significantly overlap with proteins associated with synaptic plasticity (17 proteins, $p < 10^{-6}$) and behaviour (18 proteins, $p < 10^{-7}$). Additionally, proteins associated with mental retardation have a lower but significant overlap (9 proteins, $p < 10^{-8}$) with proteins associated with behaviour. The latter is possibly evidence of the true cognitive nature of the two diseases from a molecular perspective. Possibly the most important conclusion that can be drawn from disease associations is that a given mental disease can associate with a given - finite - number of proteins, but its manifestation mechanisms are more complicated since it is part of a complex network of interacting agents. What also stems from placing disease associated genes and their protein products in the context of a protein interaction network is that we are not looking at independent pathways - but at overlapping, guilt-by-association subnetworks.

It is also vital to clarify that the model of the core PSD protein interaction network presented in this chapter is only a snapshot of the molecular machine in action. If one

Figure 5.16: Plots of semantic similarity within the clusters of the Union network (bars) compared to the average network background value (red line). Most clusters appear to have a higher internal semantic similarity compared to the average corresponding value in the network in all three GO domains (CC: top, MF: middle, BP: bottom). Note that there were not enough annotations to compute semantic similarity for clusters K, N and O. .

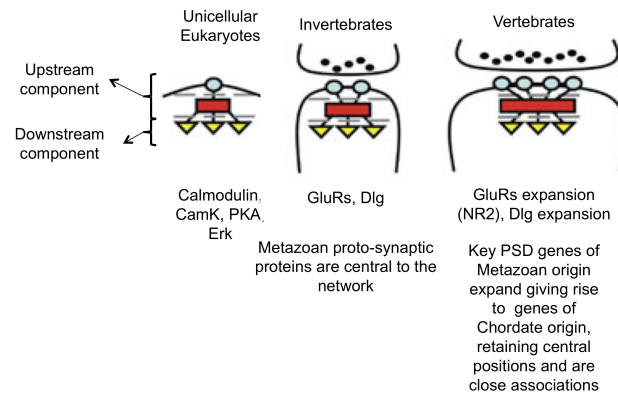


factors in dynamics such as the variations in protein abundance due to different protein turnover rates or differential gene expression and change of binding capacities due to modifications like phosphorylation, there is an explosion in the emergent complexity of the resulting network.

5.4.4 Evolution of the PSD interactome

The findings regarding evolution of the PSD corroborate results by Emes et al. (2008). In that paper the authors propose a potential mechanism of evolution (Figure 5.17) where a good proportion of the PSD's protein components predated synapses and were utilised in order to form a synaptic proteome once it first appeared (Cnidarian phylum synapse). During the course of evolution the PSD's upstream (receptors, adhesion proteins) and midstream (signalling and scaffolding proteins) components expanded while the downstream component, mostly comprised of Eukaryotic downstream signalling proteins and basic cellular toolkits (e.g. protein synthesis), remained similar in size. Also, according to the results of the Union protein interaction network analysis, proteins of Metazoan and Chordate origin tend to possess similarly more central positions in the network, providing crosstalk paths between modules. Proteins of Eukaryotic origins, which are mostly found in more conservatively expanded downstream clusters, on the other hand, occupied less central positions and provide paths for the molecular machine to communicate with more basic cellular mechanisms. The increased complexity of postsynaptic signalling complexes in vertebrates was possibly the result of an expansion of the upstream component, including the receptors, which provided a wider range of ligands to be added as extracellular signals. According to the model proposed by Ryan and Grant (2009) the complexes themselves show differences in their organisation. More specifically invertebrates only have one NMDA receptor 2nd subunit and a single Dlg adaptor protein, with one resulting complex. Vertebrates on the other hand have 4 of each, allowing 16 different combinations. The authors also argue that the different plasticity and distinct cognitive learning task phenotypes

Figure 5.17: Illustration of the evolutionary expansion of the PSD's molecular machine. Genes of interest to learning and plasticity are listed where they first arise. The schematic representations of signaling complexes use three interlinked shapes (blue circles represent upstream receptor/adhesion proteins, the red box indicates signaling proteins and yellow triangles are downstream proteins). The amount of blue circles and the size of the red box increased, illustrating the relative expansion of the up- and midstream components. Note that the expansion of mammalian brain size occurred after the expansion of synaptic proteome complexity. Adapted from Emes et al. (2008).



of Dlg2, Dlg3, Dlg4 mutant mice (Cuthbert et al., 2007, Migaud et al., 1998) support the above. It has been shown that in tissue specific networks, proteins that appeared first in evolution have an overall higher degree distribution (Bossi and Lehner, 2009). In the PSD's case, proteins of Eukaryotic that appeared first were not specific to that network, since it was before the existence of a synapse, thus do not fall into the tissue specific category. Proteins of Metazoan origin on the other hand, appeared after the synapse, but share similar degree distributions with proteins of Chordate origins. This is because of their expansion via duplication and diversification mechanisms.

Chapter 6

The fPSD interactome

6.1 Background

As seen in the previous chapters the majority of PSD proteomics data comes from mammalian samples, with the exception of representatives of the Cnidarian phylum (te Velthuis et al., 2007, Ryan and Grant, 2009), which possess the earliest form of nervous system and are the point of the emergence of postsynaptic ionotropic glutamate receptors (Nakazawa et al., 2004, Kessels and Malinow, 2009). Research has also been done in other species, including model organisms like *D. melanogaster* (Lee et al., 2008, Emes et al., 2008) or *Aplysia sp* (Belvin and Yin, 1997, Zhu et al., 2007), however, in these cases the methodology used was genomics rather than proteomics based, or focused in very small sub complexes, rather than attempting to reconstruct a wider descriptive model and elucidate its organisational principles.

Genomics and bioinformatics approaches can predict the potential composition of PSD complexes of organisms, although these approaches cannot offer lineage specific information about the architecture and organisation of the complexes. Knowledge about the composition and organisation of PSDs of other model organisms would allow us to draw conclusions about its molecular and architectural evolution, in parallel with the evolution of the synapse and the nervous system. This knowledge would also

allow us to use a model organism with a smaller but present behavioural repertoire as a model. Acquiring the information on proteomic composition, protein interactions and the PSD's organisation would also be indispensable tools for using the organism as a platform for understanding the basic molecular basis of cognition. *D. melanogaster* is an ideal organism for the above. Its short generation time, in combination with the low maintenance cost and the wide range of genetics and molecular biology tools available, would result in a versatile and fast neurobiology research platform. For the aforementioned reasons we decided it would be interesting to perform a purification, identification as well as the first reconstruction and analysis of fly PSD (fPSD) protein complexes.

Systematic attempts to elucidate the proteomic constituent parts of the fPSD have never been made before, with the exception of one published experiment by (Emes et al., 2008). A significant volume of data is available on the fruitfly's neuromuscular junction (NMJ) (Schuster, 2006a, Budnik et al., 2006, Ruiz-Cañada and Budnik, 2006a, Peron et al., 2009) and many of the genes found to be involved in the fPSD in have been characterised within the context of the NMJ. The NMJ is glutamatergic making it similar in composition and function to mammalian CNS synapses (Collins and DiAntonio, 2007). Also, the NMJ has been used as a model system for research on many fPSD genes as well as synaptic function and development since it is easier to isolate and study (e.g, Keshishian et al., 1996, Personius and Balice-Gordon, 2002, Hebbar et al., 2006, Schuster, 2006b, Ruiz-Cañada and Budnik, 2006b). With the major neurotransmitters, as with all insects, being acetylcholine (ACh), γ -Aminobutyric acid (GABA) and glutamate (Brotz et al., 2001, Kolodziejczyk et al., 2008), one would expect to find proteins of the associated pathways within the fPSD. If neurotransmitters like glutamate, that are central to mammalian brain function are equally important in insect brain function, remains to be seen.

6.2 Results

The methods used for the proteomic isolation and mass spectrometry identification of the fPSD complexes were discussed in Chapter 2. In summary the two resulting raw datasets, namely ECP and CCP, were a result of affinity purifications using four chosen protein baits believed to be involved in fly PSD complexes (dlg1, tau, 14-3-3ε, Bsg).

6.2.1 Isolation of fPSD complexes

6.2.1.1 Data filtering and stringency

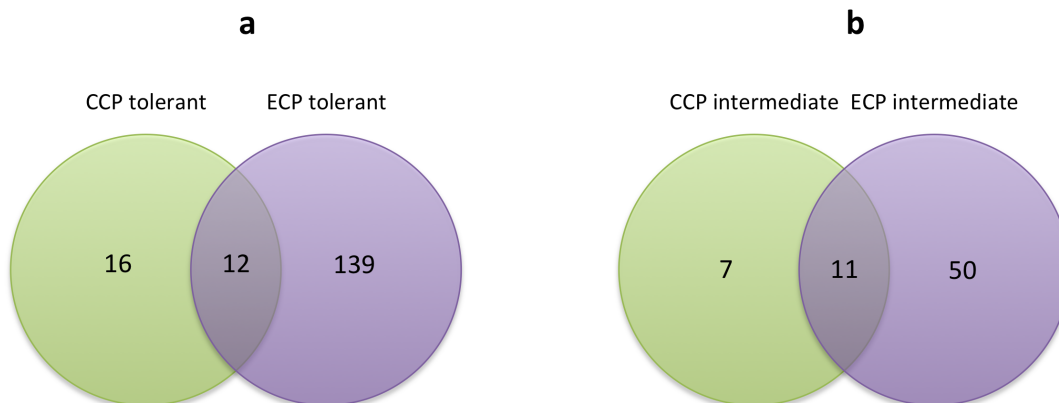
The chosen criteria cut-offs for the raw MS data are discussed in detail in Appendix A, subsection A.1.2. Three levels of stringency were chosen for the ECP and CCP raw datasets, generating three distinct labelled datasets according to stringency level: *tolerant*, *intermediate* and *strict*.

The first level, generating the filtered datasets labelled as *tolerant* (ECP *tolerant*, CCP *tolerant*), uses criteria values slightly above what each mass spectrometry facility uses as standard, reflecting a higher standard baseline. The second level, generating datasets labelled as *intermediate* (ECP *intermediate*, CCP *intermediate*), uses the same cut-offs but with an additional filter that allowed in only previously known proteins¹. Finally, the third level, generating datasets labelled as *strict* (ECP *strict*, CCP *strict*), uses the same cut-offs but only allows in lists appearing in more than one affinity purification replicates. The filtered datasets of corresponding stringency levels from the ECP and CCP raw datasets were then integrated to generate the final protein lists.

Before examining the aforementioned lists we will discuss observations on the sizes and overlaps between prays of different bait proteins. The numbers of proteins identified by each different affinity purification after data filtering as well as the number of total distinct proteins are shown in Table 6.1. It is interesting to look at the overlap between the two datasets prior to integration. When examining the total of distinct

¹Found in an in-house (Bilal Malik, unpublished data) and data published by Emes et al. (2008)

Figure 6.1: Venn diagrams of the overlap between the CCP and ECP datasets. a) *tolerant* label and b) *intermediate* label.



prey proteins of the CCP and ECP datasets (Figure 6.1) we can notice that the contribution of proteins by the CCP is 16 or ~9% of the total dataset in the *tolerant* dataset. That number drops to 7, which is the same percentage of the total size, when looking at previously known proteins only (*intermediate* dataset). This highlights the smaller contribution of the CCP dataset, which is nevertheless, considered reliable. Additionally, we can notice that the overlap of the CCP and ECP datasets has been isolated in previous purification almost in its entirety. The difference between the contributed data between CCP and ECP is also discussed in Appendix A, subsection A.1.2.4 and is mainly attributed to differences in the processing workflows of the facilities.

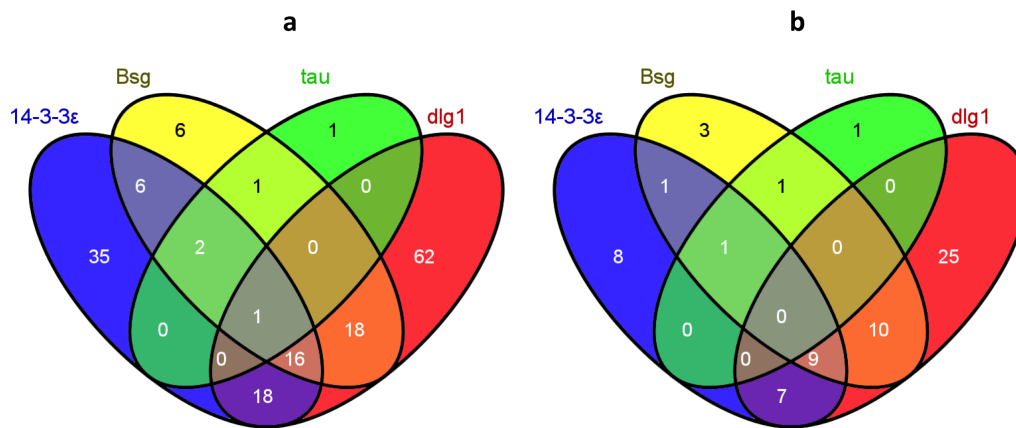
6.2.1.2 Data integration

For the purpose of integration we merged datasets of the same stringency level (e.g. CCP *tolerant* and ECP *tolerant*), resulting in the generation of two merged datasets or masterlists (*tolerant* and *intermediate*). The *strict* datasets were too small to provide strong statistical associations and were not used for the subsequent analysis. The small size of the datasets is due to low reproducibility we found in the data, an issue discussed in Appendix A, subsection A.1.2.4. The *tolerant* merged dataset contains 166 proteins out of which 66 (~40%) were known from previous affinity purifications and constitute

Table 6.1: Number of proteins in the CCP and ECP datasets prior to integration.

Dataset (<i>label</i>)	CCP:		CCP:		ECP:		ECP:	
	<i>tolerant</i>	<i>intermediate</i>	<i>intermediate</i>	<i>strict</i>	<i>tolerant</i>	<i>intermediate</i>	<i>strict</i>	
Proteins identified in the 14-3-3 ϵ purification	12	7	7	2	65	24	2	
Proteins identified in the Bsg purification	13	7	7	5	36	18	3	
Proteins identified in the tau purification	4	4	4	4	(N/A)	(N/A)	(N/A)	
Proteins identified in the dlg1 purification	12	8	8	12	110	49	3	
Total (distinct):	28	18	18	20	151	61	6	

Figure 6.2: Prey overlaps between affinity purifications using different baits (merged datasets). a) *tolerant* label and b) *intermediate* label



the *intermediate* merged dataset. The *tolerant* merged dataset will be used for further analysis and the *intermediate* merged dataset will serve as a quality control since it comprises of proteins previously known to be part of fPSD complexes. Note that the data from tau purifications were not used, except in the network models, where we found interactions with other proteins.

Before proceeding with the annotation and analysis of the proteomics lists, we examined overlaps of preys between affinity purification with different baits (Figure 6.2). What can be observed is that there is a good overlap of 14-3-3ε and Bsg preys with dlg1 preys. More specifically both 14-3-3ε and Bsg have 18 protein preys in common with dlg1 in the *tolerant* merged dataset (~10% of the dataset in each case). There is also an overlap of 16 proteins that are common preys to 14-3-3ε, Bsg and dlg1 in the *tolerant* merged dataset (~9% of the dataset in each case). These numbers go down when examining the *intermediate* dataset but their relative proportion to the dataset size slightly rises to the 11%-15% area. It is interesting to look at the composition of this common protein core (Table 6.2). Although we noticed potential common abundant contaminants like tubulin, enzymes and proteins of the mitochondria, we also notice kinases, vesicular proteins, a G-protein.

Table 6.2: Composition of the common protein core found in the prey overlap of 14-3-3 ϵ , Bsg and dlg1.

Gene symbol	Family	Subfamily
alphaTub84B	Cytoskeletal/ Structural/ Cell adhesion	Tubulin
CG8193	Enzymes	Other Enzymes
Cat	Enzymes	Other Enzymes
Gbeta76C	G-protein signaling	G-proteins
PyK	Kinases	Other Enzymes
Argk	Kinases	Other Kinases
Sod2	Signalling molecules and Enzymes	Mitochondrial Enzymes
Sccs-fp	Signalling molecules and Enzymes	Mitochondrial Enzymes
kdn	Signalling molecules and Enzymes	Other Enzymes
ade5	Signalling molecules and Enzymes	Other Enzymes
Ald	Signalling molecules and Enzymes	Other Enzymes
norpA	Signalling molecules and Enzymes	Phosphodiesterases
Ef2b	Transcription/ Translation	Transcription Elements
Ef1alpha48D	Transcription/ Translation	Transcription Elements
Chc	Vesicular/ Trafficking/ Transport	Clathrin
Mhc	Vesicular/ Trafficking/ Transport	Motor Proteins

Table 6.3: Selected GO terms that were found enriched compared to the genome in the *tolerant* merged dataset. These selected term reflect neuron and neuron associated biological processes, molecular functions and cellular components. Continues in Table 6.4.

Term (count)	Enrichment (p-value)
synaptic transmission (7)	9.12 (0.03)
behavior (21)	3.8 (0)
locomotory behavior (11)	5.5 (0.02)
regulation of G-protein coupled receptor protein signaling pathway (6)	28.84 (<0.000001)
synaptic vesicle coating (4)	32.36 (0.02)
synaptic vesicle transport (6)	10.23 (0.05)
neurotransmitter secretion (10)	8.71 (<0.000001)
neurotransmitter transport (10)	6.92 (0.01)
regulation of neurotransmitter levels (11)	9.12 (<0.000001)
response to external stimulus (17)	8.13 (<0.000001)

6.2.2 The fPSD proteomic catalogue

The genes associated with the proteins in the obtained merged dataset lists were annotated using Gene Ontology (GO) (retrieved from FlyBase, v10.08) and protein domain information from InterPro. The genes were also grouped in functional families and subfamilies using the same methodology as in Chapters 4 and 5, so the results could be comparable.

6.2.2.1 Gene ontology

In order to get an overview of the merged dataset we initially examined all GO annotation terms (with more than 5 occurrences in the dataset, exceptions were made for interesting terms with lower counts). This can give a more general picture of the function, localisation and processes the complexes are involved in and also evaluate

what types of proteins this specific affinity purification protocol is able to isolate and how these proteins might be connected to fPSD function. The frequently occurring and enriched GO annotations of the *tolerant* merged dataset, presented separated by GO domain are:

- **Cellular component (CC)** (Figure 6.3). All terms show an n-fold enrichment in the range of ~4-30, compared to the whole genome. The only exception is the nucleus term which appears depleted compared to the whole genome. Within the annotation terms we also find synapse (3) and postsynaptic membrane (1). The above terms can outline cellular components and organelles that the proteins appear in. The majority of proteins found in the lipid particle (45), microtubule associated complex (38) and cytoplasm (33). The presence of the plasma membrane, synaptic vesicle, and lipid particle shows that the current form of the affinity purification protocol is able to isolate membrane associated proteins to some extent. However, examination of the list shows a lack of receptors that we would expect to be there (e.g. NMDA receptor as a prey). This is more extensively discussed in the end of this subsection. The presence of nuclear proteins could be a result of contamination because of the use of whole heads rather than synaptosome preparations.
- **Molecular function (MF)** (Figure 6.4). All terms show an n-fold enrichment in the range ~3-37, compared to the whole genome. Terms like calcium ion binding and phosphorylative mechanism are common in the mPSD and its sub-complexes. Additionally, protein binding seems to be a prominent function, as expected by a set of proteins that function in complexes. Due to potential contaminations associated with proteins from neuron nuclei or cytoplasm, as seen from the GO CC annotations, we have to keep in mind that not all terms might be equally relevant.
- **Biological process (BP)** (Figure 6.5). All terms show an n-fold enrichment in

the range ~7-38. It also worth mentioning, that between annotations with lower counts we find more neuron related process terms (olfactory behavior, synaptic vesicle coating, synaptic vesicle endocytosis, axonogenesis, nervous system development, peripheral nervous system development, central nervous system development, axon midline choice point recognition, synaptic growth at neuromuscular junction) and terms previously associated with the mPSD and the Union protein complex (synaptic vesicle priming, G-protein coupled receptor protein signaling pathway, synapse assembly). Again, there is a mixture of annotations describing biological processes that are either common or clearly associated with neuronal function. Example of the latter type are synapse associated terms (neurotransmitter secretion, synaptic transmission, axon guidance). Terms like organisation of cytoskeleton (cytoskeleton organization) and protein transport and folding (protein folding, intracellular protein transport) could also be associated with PSD functions. The rest of the functions could also reflect processes that contaminants are involved in.

We also examined significant enrichments compared to the whole genome ($p < 0.05$). Among the enriched terms we also found terms implying that many of these proteins are organised in complexes (protein complex, cellular protein complex assembly, enriched ~4 and ~9 fold respectively, $p < 10^{-6}$), associated with protein folding and targeting (protein folding, establishment of localization in cell, enriched ~7 and ~4 fold, $p = 0.04$ and $p < 10^{-6}$ respectively) and cation transfer activity (inorganic cation transmembrane transporter activity, enriched ~9 fold, $p < 10^{-6}$). All these processes, although having general applications in all cells are also related with synaptic function. Between the ~200 terms, we isolated terms that are associated with neurons or neuronal processes (Tables 6.3 and 6.4). Within these ms we notice terms associated with synaptic transmission, behaviour, G-protein signalling, synaptic vesicles, neurotransmitter secretion and regulation, response to various types of stimuli, and a set of terms associated with light perception. Terms belonging to the latter set, associated

Table 6.4: Selected GO terms that were found enriched compared to the genome in the *tolerant* merged dataset. These selected term reflect neuron and neuron associated biological processes, molecular functions and cellular components. Continued from Table 6.3.

Term (count)	Enrichment (p-value)
response to light stimulus (12)	12.02 (<0.000001)
response to stimulus (43)	3.72 (<0.000001)
response to chemical stimulus (15)	3.63 (0.04)
detection of abiotic stimulus (8)	14.79 (<0.000001)
detection of external stimulus (8)	13.18 (<0.000001)
detection of light stimulus (8)	16.6 (<0.000001)
detection of light stimulus involved in visual perception (8)	16.98 (<0.000001)
detection of stimulus (8)	9.12 (0.01)
detection of stimulus involved in sensory perception (8)	11.48 (<0.000001)

with the neurons innervating ommatidia and photoreceptor cells, occur often in this set and are possibly a result of *dlg1* abundance in these.

6.2.2.2 Protein families

In order to obtain a concise representation of the constituent parts of the isolated complexes, we examined the distribution of different protein families in the *tolerant* and *intermediate* merged datasets (Figure 6.6). Keeping in mind that the *intermediate* dataset represents the previously known portion of the *tolerant* dataset, we can observe the contribution of novel fPSD associated proteins by this dataset. More specifically we can see that some of the biggest contributions are classified under the Signalling Molecules and Enzymes family. Proteins in this functional category have a function

Figure 6.3: Histogram of GO Cellular Component annotation counts of the fPSD merged tolerant dataset.

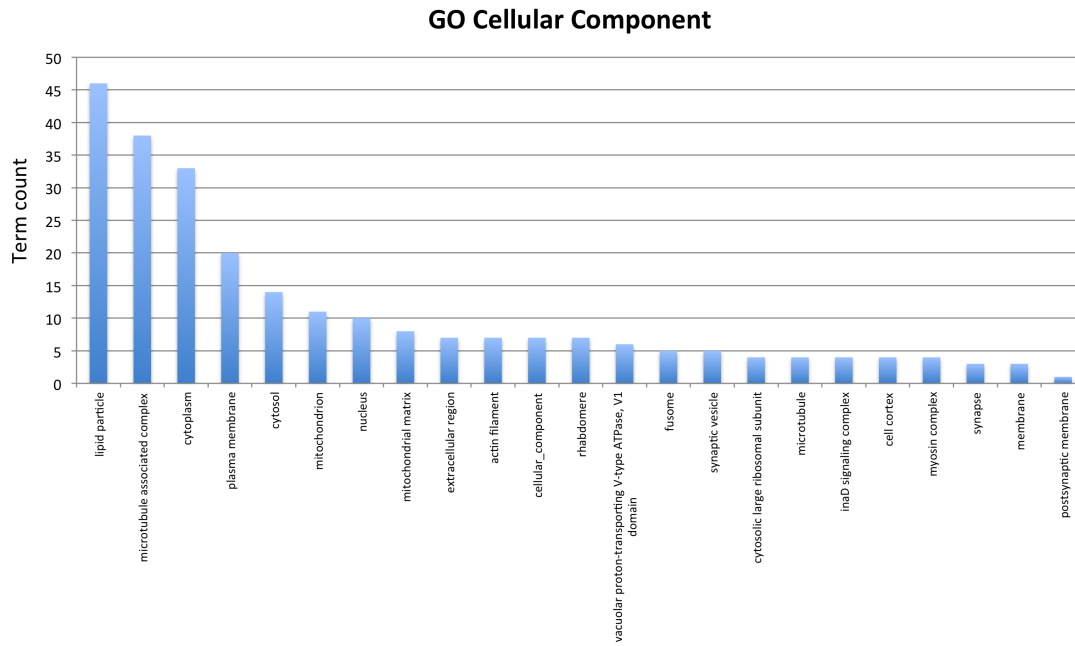


Figure 6.4: Histogram of GO Molecular Function annotation counts of the fPSD merged tolerant dataset.

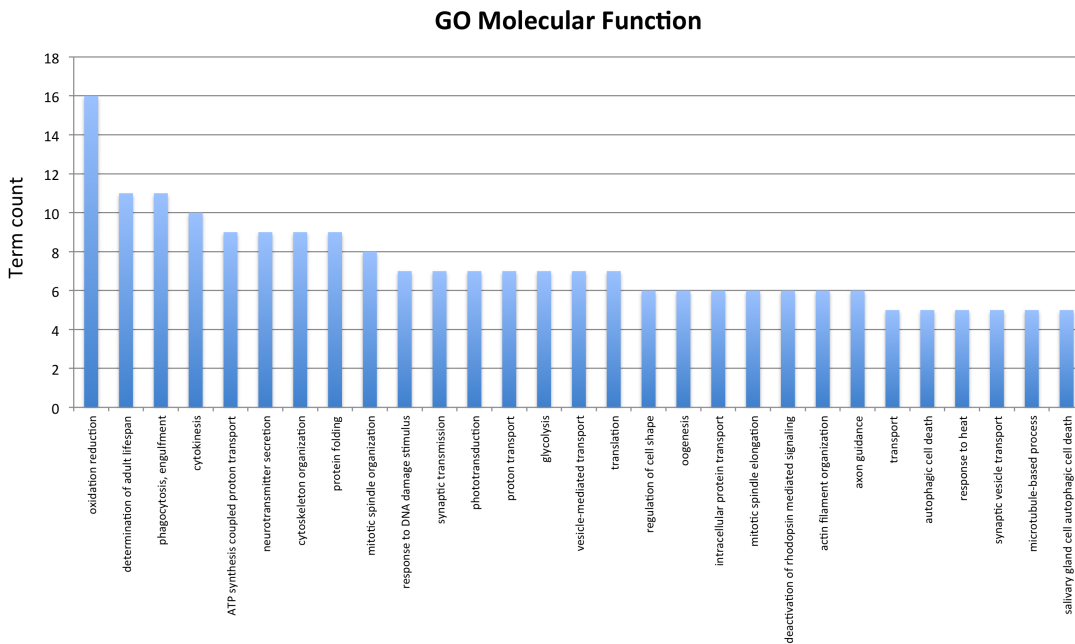
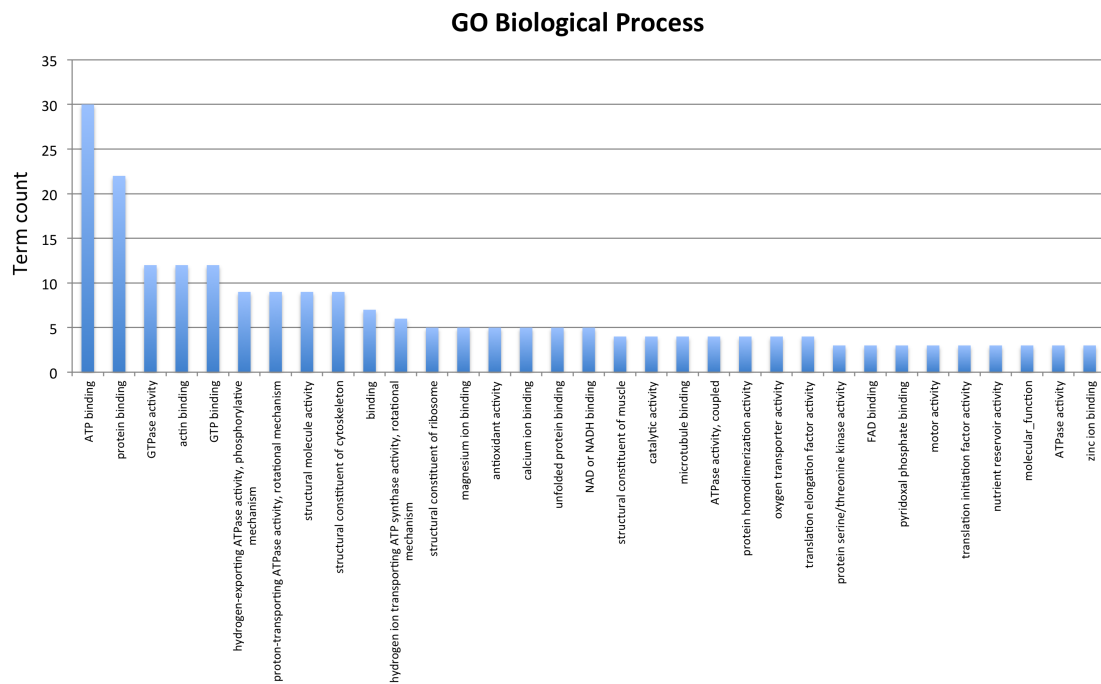
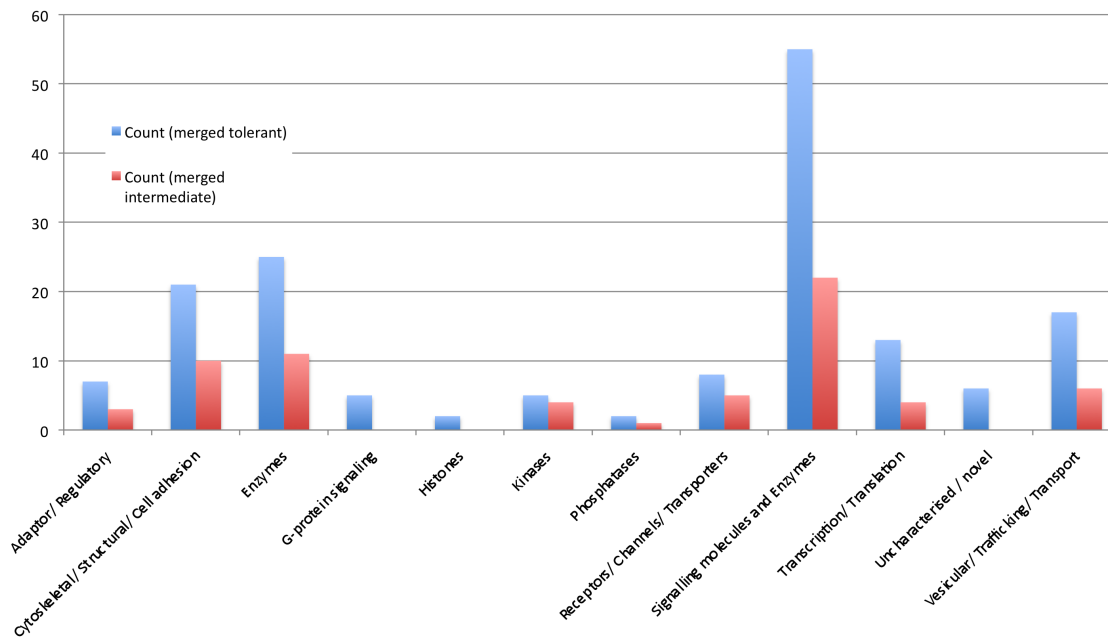


Figure 6.5: Histogram of GO Biological Process annotation counts of the fPSD merged tolerant dataset.



that is either signalling associated (e.g. proteases) or are involved in pathways with signalling associated products. Although the count of this family is dominated by the latter category which is more general, we noticed contributions of proteins belonging to the Heat shock / Chaperones / Chaperonins (Hsp60, jdp, Pdi, Hsp83) and Proteases (DppIII, CG12163, CG3107, Rpn2, Pros28.1, Cp1) subfamilies that had not been previously associated with the complexes. Although there are contributions in all families, sizable contributions are seen in the case of the Cytoskeletal/ Structural/ Cell adhesion, Enzymes, Transcription/ Translation and Vesicular/ Trafficking/ Transport families. All these families have representatives in the mPSD, but in the case of the fPSD, isolating novel proteins from these families can give us a wider picture of their role in the function of the molecular machine. There are cases where the contribution is not as high in count, but for example in the cases of Kinases, Phosphatases and G-protein signalling families, which have a central role in the signalling component of PSDs (Ballyk and Goh, 1993, Miller et al., 1995, Dosemeci and Reese, 1995, Shen and Johnson, 1997, Lüscher et al., 1997, Castro et al., 2003, Coba et al., 2008; 2009), even

Figure 6.6: Distribution of protein families in the *tolerant* (blue) and *intermediate* (red) datasets.

a small contribution could be important in drafting the components of these pathways in the PPIN analysis to follow.

6.2.2.3 Protein domains

Another aspect of functional annotation is that of the protein domains present within a complex. We examined the frequency of occurrence and enrichment of domains compared to the whole genome. Domains with high occurrence frequency (>5) and high enrichment (>10-fold) included mitochondrial (ATPase, F1/V1/A1 complex) and cytoskeletal (Actin/actin-like, Actin/actin-like conserved site and Actin, conserved site) domains. Both these structures have been known to be parts of the PSD (MacAskill et al., 2010, Hotulainen and Hoogenraad, 2010). Interestingly, among domains with lower counts we found many domains that were highly enriched in the Union complex (see subsection 5.3.1, Chapter 5). These included domains that are involved in G-protein signalling (Small GTP-binding protein, Ras small GTPase, Rab type), Calmodulin and Calcium binding (IQ calmodulin-binding region, C2 calcium/lipid-binding

domain, CaLB, C2 calcium-dependent membrane targeting, EF-Hand 1, calcium-binding site), phosphorylation (Protein kinase-like domain, Serine-threonine/tyrosine-protein kinase, Protein kinase, ATP binding site, Serine/threonine-protein kinase, active site, Guanylate kinase), scaffolding (PDZ / DHR / GLGF, Src homology-3 domain), cytoskeleton (Pleckstrin homology), and mitochondrion (MIRO-like).

6.2.2.4 Discussion of annotations

Overall the annotation of this complex contains many GO terms and protein domains that would be expected in a PSD complex, as highlighted in the paragraphs above. However, the two main differences with the mPSD datasets discussed in previous chapters are, a) the lower count and enrichment factor of these annotations when compared to similar annotations from the mPSD datasets and b) the presence of a wider variety of annotations and protein domains. Although hard to quantify, we believe that this variation is due to a mixture of the following reasons. First, the constituent parts of the fPSD are partially different to these of the mPSD. Also, mouse nervous system proteins are expected to be better annotated in GO compared to those of *D. melanogaster*. Finally, the fly head protein extract used was not as enriched in fPSD proteins as expected. Nevertheless, given the circumstances domains and annotations representative of the mPSD were found, although their lower counts and enrichment could also reflect a masking of the fPSD proteins by contaminants or an effect of lack of annotations.

Another issue observed with this dataset is the lack of neurotransmitter receptor proteins. While we were able to isolate part of the scaffolding, signalling, G-protein signalling protein synthesis and turnover components we were not able to isolate the receptor component. Although the complex shows a number of membrane related GO CC annotations, almost all were after manual inspection, found to be associated with proteins found near rather than spanning the cell membrane. Although the protocol included steps that assist the enrichment in membrane proteins, some PSD receptors like the NMDA receptor subunits are notoriously hard to isolate. Proposed solutions

to this are discussed at the end of the chapter.

6.2.3 An integrated protein interaction network of the fPSD

6.2.3.1 Integration with previous data

In order to produce the network model reconstructions presented in the rest of this chapter we decided to merge the dataset at hand with data from previous affinity purifications. The premise behind this merge was to include a body of previously known data based on affinity purification approaches, in order to provide the basis for reconstructing a network model that includes all the known data. More specifically we use the following data:

- Nmdar2 C-terminal affinity purification data (Emes et al., 2008). This dataset was acquired with an affinity purification using a synthetic C-terminal peptide from the Nmdar2 subunit as a bait. The use of this dataset also allows us to include Nmdar2 as an element in the reconstructed networks.
- Bsg affinity purification data (Bilal Malik, unpublished data). This dataset is the result of a successful purification using the method described in this chapter.

We added the 219 proteins from the Nmdar2 C-terminal affinity purification data and the 92 proteins from the Bsg affinity purification data to the *tolerant* merged dataset (166 proteins). The resulting dataset contained 402 proteins. 192 of these proteins were unique to the Nmdar2 C-terminal affinity purification data only, which was expected since the method was slightly different and performed using a different bait.

6.2.3.2 Interaction mining

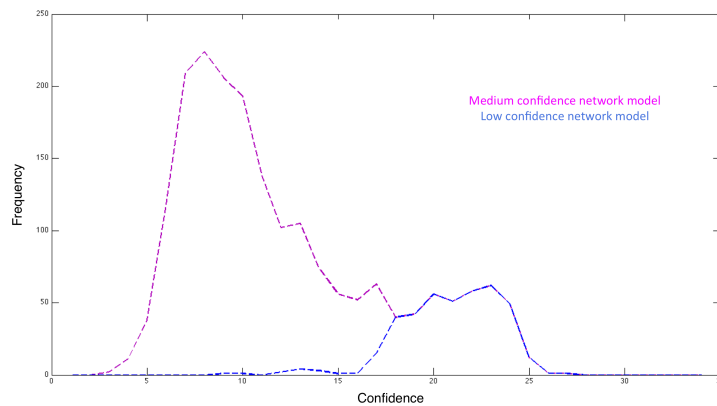
The current availability of protein interaction data for *Drosophila* is rather low. Additionally, only a few (Giot et al., 2003, Stanyon et al., 2004, Formstecher et al., 2005) Y2H datasets are available. Also, it has been shown that the majority of interaction

data come only from a few studies and have not been otherwise verified (Razick et al., 2008). For the reconstruction of the fPSD network model we decided to use DRoID (Yu et al., 2008b) as a protein interaction resource. DRoID integrates data from various resources including Y2H datasets, the literature, other databases and interactions of homolog proteins from other organisms. Additionally, the data integration process behind DRoID allows the computation of a logistic regression based interaction confidence score (Giot et al., 2003), which is independent of the original data source and type of experiment.

When considering interaction resources we also tested alternative resources. An example is the emerging STRING database (Mering et al., 2007, Jensen et al., 2009, Szklarczyk et al., 2011). STRING contains more than 12 million protein and genetic interactions, most of which are not curated. The evidence used to compile such a dataset range from the acceptable previous knowledge, homology and direct experimental data approaches to predictions based on empirically weak features such as co-citations in PubMed or association in other databases. Weak features such as these can lead to false positives, e.g. we noticed many cases where a Reactome entry was cited as evidence, but in the original entry the protein complex was annotated having an unknown topology.

When testing various confidence score cut-off values we encountered a problem: while the core of the resulting networks had many protein interactions with a high score (> 0.5), most of the protein interactions connecting baits with preys had confidence scores between 0.2 and 0.4 (Figure 6.7). For that reason we decided to use a specific cut-off for the core of the network, but not filter interactions between baits and prey proteins in the same affinity purification. The choice of a cut-off score was 0.5 (medium confidence model) was based on empirical observations. There were also interactions without a confidence score and in these cases we performed a manual check before including these in the model. A confidence score cut-off above 0.5 also allowed for a substantially sized major connected component in the interaction network

Figure 6.7: Distribution of protein interaction confidences in the low and medium confidence network models. A confidence score cut-off above 0.5 (medium confidence model) also allowed for a substantially sized major connected component in the interaction network and ensured that almost all proteins that had a nervous system associated functional annotation remained within the connected portion of the model



and ensured that almost all proteins that had a nervous system associated functional annotation remained within the connected portion of the model.

6.2.3.3 The fPSD protein interaction network

The protein interaction network included 182 nodes and 516 edges. Many of these interactions were between pairs of proteins that did not interact with the main component of the network. The network's major connected component (MCC) consists of 105 nodes and 389 edges which segregates into 9 communities using the Newman and Girvan algorithm ($Q = 0.51$). Further analysis showed that segregation into 10 clusters was more biologically meaningful (a sub-cluster of nuclear proteins found in the first cluster was assigned to a separate cluster). Additionally, to those, there are 5 more disjoint communities of substantial size (≥ 5) including a total of 149 nodes. For the rest of this analysis we will focus on the 10 clusters of the MCC and the 5 disjoint clusters, shown on Figure 6.8. Note that the protein, interactions, and annotations lists are available with the supplementary material DVD.

6.2.3.4 Clusters and statistical correlations

As mentioned earlier the network segregates in a total of 15 clusters, with 10 of them being connected in the MCC. These clusters are

- **Cl1:** This cluster was initially found to consist of 12 nodes and contains bait (in previous experiments) protein Nmdar2. After close inspection we noticed that 4 of these nodes formed a tightly interconnected sub-cluster which interacted via the Nmdar2 subunit with the rest of Cl1. Inspection of the annotations for these proteins shows that they are all nuclear proteins - possibly contaminants with respect to the fPSD. We manually removed these to an independent cluster (Cl10), with no significant change to Q and proceeded with the annotation analysis. The new resulting cluster consists of 8 proteins and is composed of 37% - and significantly correlated with - the Kinases family ($p = 0.03$) and kinase GO annotations. 50% of the cluster are direct interactors of Nmdar2 and the cluster also contains the PDZ-domain containing scaffolding protein Patj. No specific synapse associated GO BP associated terms were found enriched compared to the whole network - although there were enrichments in other types of processes like establishment of tissue polarity or digestive tract development, intracellular signalling and light perception. This could be also attributed to the effect of pleiotropy on nervous system annotations (see section 6.3), as well as biased, missing or wrong annotations in GO.
- **Cl2:** Consists of 8 proteins and includes bait protein dlg1. It is dominated by (~ 62%) and significantly correlated with the Cytoskeletal/ Structural/ Cell adhesion family ($p = 0.03$). The cluster has a clear localisation at the postsynaptic terminal and its membrane element, showing enrichments in associated GO CC terms. From a GO MF perspective, binding is prevalent and significantly correlated ($p < 0.05$). The GO BP terms in this case include enrichments in terms like synaptic transmission, transmission of nerve impulse, generation of neurons

and regulation of synapse structure and activity (all with $p < 0.05$). The cluster contains bait protein dlgl1 which is among other functions is associated with synapse and receptor complex assembly (Budnik et al., 1996, Chen and Featherstone, 2005).

Clusters C11 and C12 contain the more well studied fPSD proteins, we can observe how a complex between veli, dlgl1 and Nmdar2 appears in the model. The interaction between veli and dlgl1 is known to be cell type specific and important for the localisation of veli (Bachmann et al., 2004). It has also been observed that the aforementioned interaction controls proper structural organisation of NMJs (Bachmann et al., 2010). Cluster C12, contains dlgl1, which also recruits scrib via a binary interaction requiring gukh (Mathew et al., 2002) which however, is not present in these purifications. The mouse homologs of scrib, Scribble1 has been found to be essential to neuronal plasticity in other studies (Moreau et al., 2010). Mutations to another direct interactor of dlgl1, futsch, has been found to cause neurodegeneration-like symptoms to *Drosophila* via caused defects to cytoskeleton organisation and axon transport mechanisms. The results are similar to the over-expression of fly or bovine tau, suggesting a certain degree of functional redundancy of microtubule-associated proteins (da Cruz et al., 2005). Regarding β -Spectrin, also belonging to C12, it is known to be associated with the localisation of various PSD membrane receptors in mice (Bloch and Morrow, 1989, Daniels, 1990, Wechsler and Teichberg, 1998). Interestingly, mutations causing disruptions to synaptic transmission and plasticity mutations have been found in *Drosophila*, but their causes seem to be of a presynaptic nature (Featherstone et al., 2001).

- C13: Consists of 7 proteins, includes bait protein 14-3-3 ϵ , and is dominated and significantly correlated with the Phosphatases ($p < 10^{-4}$) family and that reflects on the GO terms enrichments found for that cluster that are mostly dephosphorylation and phosphatase complex related (all with $p < 0.05$).

- Cl4: Consists of 5 proteins and is of a mixed family nature. It shows no particular correlation with specific GO terms.

Clusters Cl3 and Cl4 appear to be organised around 14-3-3 ϵ and 14-3-3 ζ respectively. While the mouse homolog of 14-3-3 ϵ has a central position in the network, in the fPSD case it seems to directly interact with 3 proteins. Nevertheless, this could be a result of the lack of interaction data. 14-3-3 ϵ is known to be an activator of the Ras pathway (Chang and Rubin, 1997) which is known to possess a role in synaptic transmission and plasticity (Brambilla et al., 1997). On the other hand 14-3-3 ζ (leo), mutations of which have also been associated with memory consolidation defects in *Drosophila* (Skoulakis and Grammenoudi, 2006), associates with 6 proteins on a first degree. One of these proteins is drk, also associated with the Ras pathway (Olivier et al., 1993), has been associated with olfactory learning and memory (Moressis et al., 2009). Another protein associated with 14-3-3 ζ is Clh (clathrin), which among other functions mediates changes in receptor number and distribution via endocytosis mechanisms, affecting LTD (Wang and Linden, 2000).

- Cl5: This is the second biggest cluster in the network consisting of 21 proteins and contains the bait protein tau. It is dominated by (66%) and is significantly correlated ($p < 10^9$) to the Cytoskeletal/ Structural/ Cell adhesion family of proteins. All the annotations its is significantly correlated with, are cytoskeleton related. This seems to be the cytoskeletal component of the fPSD. The cluster also contains two heat shock proteins, Hsp83 and Hsc70-4 both known to have a synaptic function but most likely of a presynaptic nature (Bronk et al., 2001, Neal et al., 2006). Bait protein tau is also in this cluster, via its only interaction, with abundant cytoskeletal protein Act42A.
- Cl6: Consists of 8 proteins and is significantly correlated with the Vesicular/ Trafficking/ Transport family ($p = 0.02$) as well as the associated GO CC terms. Cluster Cl6 contains rl (rolled), an extracellular signal-regulated (Erk) kinase,

down-regulation of which is associated with the correct structural formation of synapses (Wairkar et al., 2009). This protein associates, among others, with ras which mediates the synthesis of Guanine nucleotides, which in turn are key players in mediating growth-cone signaling during neural development. Also, Cl6 interacts with the MCC only through an interaction with Cl10, so although synaptic it might not be necessarily considered postsynaptic.

- Cl7: Consists of 3 proteins and is a small tubulin associated cluster.
- Cl8: Consists 6 proteins and is associated with metabolism related GO terms. However, this cluster includes tpi, a gene which has been show to be associated with locomotion with a mechanism independent of general bioenergetic impairment (Celotto et al., 2006).
- Cl9 is the biggest cluster in the network (35 proteins) and consists of a set of highly interconnected ribosomal proteins as well as the parts of the ubiquitination mechanism. The cluster is associated with protein synthesis and turn-over. GO BP terms that appear enriched include translation and ubiquitin-dependent protein catabolism (both with $p < 0.0005$). Most of its cross-cluster interactions are via ubiquitin Ubi-p63E to clusters Cl5, Cl7 and Cl8.
- Cl10 is the cluster resulting from the manual separation of Cl1 and consists of nuclear proteins that interact with the MCC only via the Nmdar2 subunit with a very low confidence interaction. Thus, this cluster can be safely considered a result of contamination.

The rest of the clusters are disjoint from the MCC.

- Cl11 consists of 17 proteins and is protease and proteasome associated. GO CC terms do not offer more information about its localisation. It has however, been shown that postsynaptic impairment of the Ubiquitin-Proteasome system demonstrated that postsynaptic proteasome function limits neurotransmission

strength (Haas et al., 2007). We have to note that the only Ubiquitin associated protein in the network is found in cluster C19 (Ubi-p63E) and is not connected with C111 with any interactions - although that could be a result of missing interactions in the databases.

- C112 consists of 11 proteins of mitochondrial nature, verified by relevant term enrichments in all domains of GO. Mutations to ADP/ATP translocase *sesB* has been shown to affect the NMJ, causing activity-dependent neurotransmission lesions (Trotta et al., 2004).
- C113 consists of 6 proteins, mostly of vesicular nature and also includes and PDZ-domain containing scaffolding protein (*skf*). It is associated with clathrin vesicles and by examination of its GO BP terms, it is likely to be of a presynaptic nature. Nevertheless, it is interesting that all the vesicular proteins have been isolated from purifications using Nmdar2 C-terminal peptides and *dlg1* baits which are known to be postsynaptic. The interacting pair of *comt* (*comatose* or *nsf1*) and *nsf2* are found in this cluster. These two highly related isoforms of NEM-sensitive fusion protein (*Nsf*) are required for many intracellular membrane trafficking steps, including presynaptic neurotransmitter vesicle priming (Kawasaki and Ordway, 2009). Again, supporting the potential presynaptic relevance of this cluster, is the presence of *lap*, a key factor in receptor-mediated endocytosis (Zhang et al., 1998).
- C114 and C115 consist of 5 proteins each associated with vacuoles and the Golgi apparatus respectively

The previously discussed distribution of protein families in the network can be seen in Figure 6.9. In link to the second part of hypothesis 3, we notice that regardless of potential contaminants and drawbacks of the protein extraction methods regarding receptor proteins a modular architecture is present. While we speculate that lack of both constituent parts and interaction data does not give a “final” cluster configuration,

the modular architecture is analogous to the one of mPSD models. An example of this architecture can be seen in C11 and C12 organised around Nmdar2 and dlg1 respectively. Similarly, clusters C13 and C113 organise around 14-3-3 ϵ and skf respectively. Again, GO annotation correlations and previous knowledge verify some biological significance to the modularity.

6.2.3.5 Cluster interactions and architecture

With the exception of C19, inter-cluster connections seem to be rather sparse - 24 edges between the 10 MCC clusters with the average confidence being in the 0.5 area. Clusters C11 to C15 are interconnected with C15 having most connections due to its size. C16, C17 and C18 connect to the MCC via C19. It is possible that these interactions are not fPSD specific - although 11 out of the total of 17 proteins have been purified with more than one baits. This implies that they are either common contaminants of the preparation or that the interactions connecting them to the fPSD are not yet known or below the threshold. Similarly, cluster C110 is connected to the MCC via an interaction between the Nmdar2 subunit and CG30122. The interaction is of particularly low confidence, reinforcing the notion that this interaction as well as the whole of C110 might be out of the fPSD context.

Cluster C19 has a big impact when calculating any network measure due to its very high rate of intra-connections, which account to more than 60% of the edges in the MCC with 73% of these interactions having intermediate (0.5) to high (0.7) confidence. For this reason members of that cluster appear to have relatively low average shortest path (ASP) length leading to them. Setting these nodes aside, other nodes with low ASP lengths to them include, C15 members Act42A, Hsc70-4, Actn, Act87E and Act5C ranging between 3.1 and 3.5 edges. These protein are in majority cytoskeletal with the exception of Hsc70-4. Dlg1 is found further down the ranked list with an ASP of 3.8 edges, considerably higher than the Dlg family values in the mPSD complexes. Similarly, 14-3-3 ϵ and Nmdar2 have high ASPs of 4.25 and 4.6 respectively. This

can be interpreted as a result of the combination of highly dense interactions in C19 in combination with a potential lack of interaction information between members of the network belonging to other clusters, leading to longer shortest paths between key members of the complex.

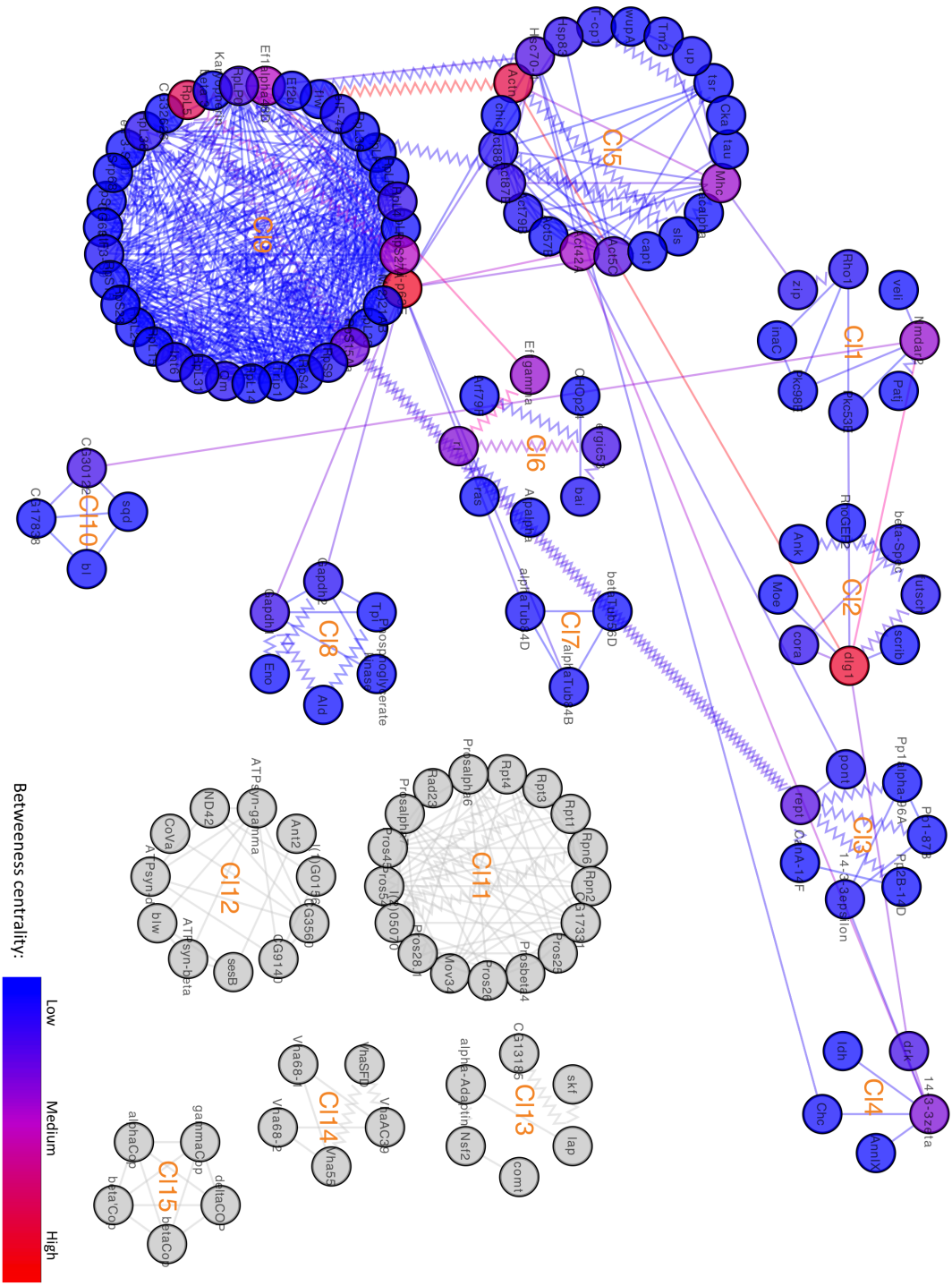
We also examined the network with respect to betweenness centrality (Figure 6.10). This revealed a different set of results where members of the Adaptor/ Regulatory (Ubi-p63E, dlg1), Cytoskeletal/ Structural/ Cell adhesion (Act42A), Kinases (rl) and Receptors/ Channels/ Transporters (Nmdar2) are found on the top of the ranking. So while the role of proteins known to be central in other PSD networks is somewhat obfuscated when using the ASP measure, use of the centrality measure manages to highlight this role. Again, this interpretation shows a similarity to the organisation of mPSD networks where NMDA receptor subunits and proteins of the Dlg family appear central to the network. We also have to take into consideration that due to the quality of the interaction data, there could potentially be missing or false positive interactions in the dataset altering the results.

6.2.3.6 Protein expression

We retrieved a set of expression data from the BrainTrap database. This set could not under any circumstances be considered complete, however, we considered interesting to examine, using the data available, the distribution of expression of proteins in the brain. We also considered FlyAtlas (Chintapalli et al., 2007) as an alternative but decided against it since it does not provide with enough granularity on brain regions.

After retrieving a set of expression data from the BrainTrap database, we found no particular statistical correlations between brain regions and clusters of the fPSD network. This was expected since the sample in hand represents a brain average. We also examined the gene lists of the dataset used for the reconstruction of the fPSD network model as well as the list of 182 genes that made it in the connected model for expression annotation in BrainTrap. Out of the 402 proteins in the reconstruction dataset, 39

Figure 6.10: The fPSD protein interaction network annotated with network measures. More specifically the colour gradients of the edges and nodes are based on a their edge and node betweenness respectively(zig-zag shaped edges have an unknown confidence).



had annotated entries in BrainTrap. Out of these 39 genes, 13 were expressed in the whole brain, 15 showed expression in the cerebral cortex and 35 (89%) were expressed in some component of the central complex. Similarly, out of the 182 genes encoding for proteins in the network, 21 had annotated entries in BrainTrap. Out of these 21 entries, 19 included some central complex component. As mentioned previously components of the central complex such as the mushroom, ellipsoid and fan-shaped body are involved in various form of learning and memory (see 2.1.2.2). Although BrainTrap does not offer full or unbiased coverage of the fPSD or genome, these evidence verify that at least part of the dataset is expressed in areas of brain associated with learning and memory.

6.2.4 Comparison of bait complexes

Although the focus of this work is to reconstruct and analyse a protein interaction network model for the fPSD, it is also interesting to examine sub-complexes of the individual baits used for the affinity purifications with the corresponding preys. For that reason, in the following paragraphs we will present and discuss the position of bait proteins 14-3-3 ϵ , dlg1, Nmdar2 and Bsg from a different perspective.

For reference and comparison purposes we also reconstructed a network using a 0 confidence cut-off and the *tolerant* merged dataset (low confidence network model). It is evident that the low confidence network model contains many more interactions, with the majority being below a 0.4 confidence (Figure 6.7), allowing more proteins from the affinity purification to be included with the trade-off of containing non-specific and/or false positive interactions, or non-fPSD context specific interactions of contaminants. For these reasons, the low confidence model will only be used as a reference when examining the bait-prey complexes rather than a model of the fPSD network.

6.2.4.1 Bait-prey complexes of 14-3-3 ϵ

14-3-3 ϵ purifies a total of 71 preys in the *tolerant* merged dataset. Out of these 71 proteins 30 (~42%) and 56 (~78.8%) appear in the medium (Figure 6.11, top) and low confidence network models respectively. However, in both cases, the only bait-prey interaction found in both cases is between Pp1-87B and 14-3-3 ϵ . Also, within the set of 26 proteins that are not included in the medium confidence network model, we see enrichment in GO BP terms associated with light perception (response to abiotic stimulus, $p < 10^{-4}$) and photoreceptor associated phospholipase C signalling (activation of phospholipase C activity, $p < 10^{-4}$). This could be evidence of contamination via non specific interactions due to the affluence of ommatidia in the prepared head sample. Looking at the immediate neighbourhood (1st-3rd degree interactors) of 14-3-3 ϵ in the medium confidence network (Figure 6.11, bottom), we observe that only 3 of the 31 prey proteins are included. Adding one more degree of neighbours (up to 4th) includes 12 (~38% of the preys appearing in the network) preys spanning clusters C13, C14, C15, C17 and C19. Some of the preys appear to be in a 8th degree neighbourhood (e.g. Arf79F) while others are disconnected from the MCC. It is interesting that a 14-3-3 ϵ bait two of its interactors (drk and AnnIX), but not its homolog and interactor 14-3-3 ζ .

6.2.4.2 Bait-prey complexes of dlg1

Dlg1 purifies a total of 113 proteins in the *tolerant* merged dataset. Out of these 50 (~44%) and 99 (88.7%) proteins appear in the medium (Figure 6.12, top) and low confidence network models respectively. In the medium confidence network dlg1 appears connected with Moe, but when examining the low confidence network we can see that this pair of interacting proteins connects to another 30 proteins via an interaction with 14-3-3 ζ . Part of this tightly connected component of the complex is also visible in the medium confidence network, although the majority of interactions gets filtered out in that model. The set of 49 proteins that are not included in the medium confidence network model, are involved in biological processes that are associated with metabolism

(e.g. glycogen metabolic process, $p < 10^{-3}$) and metabolic response to inorganic substances (e.g. response to reactive oxygen species, $p < 10^{-3}$). Again, there are enriched sets of annotations for photoreceptor associated signaling (detection of light stimulus) involved in visual perception and deactivation of rhodopsin mediated signaling, both $p < 10^{-3}$, which could be attributed to contaminations due to non-specific interactions or non-fPSD context specific proteins due to the abundance of ommatidia in the sample. Looking at the immediate neighbourhood (1st-3rd degree interactors) of *dlg1* in the medium confidence networks (Figure 6.12, bottom), we can see that 10 of the prey proteins are included. This number rises to 19 (16.8% of the preys appearing in the network) on the by adding one more degree of neighbours (4th degree neighbours) with preys spanning clusters C11-C15, C19 and C110.

6.2.4.3 Bait-prey complexes of *Nmdar2*

The *Nmdar2* C-terminal peptide purifies 217 proteins in the *tolerant* merged dataset. Out of these 100 (~45%) and 160 (~73%) appear in the medium (Figure 6.13, top) and low confidence network models respectively. In the medium confidence network *Nmdar2* is connected with 14 prey proteins, spanning clusters C11, C12, C15 and C110, in a tightly connected component. Other sets of preys also appear in sets of appear in tightly connected components varying in size from 23 to 4 prey proteins. These components appear interconnected into a larger component in the low confidence network via a set of low confidence interactions. The set of 60 proteins that are not included in the medium confidence network model, are involved in biological processes that are associated with organic acid metabolism (e.g. pyruvate metabolism, $p < 10^{-3}$) and fatty acid transfer (e.g. long chain fatty acid transfer, $p < 10^{-3}$). The immediate neighbourhood (1st - 3rd degree interactors) of *Nmdar2* in the medium confidence network (Figure 6.13, bottom) includes 15 preys, while adding one more degree of neighbours includes 24 preys (24% of the preys appearing in the network) spanning clusters C11, C12, C14, C15, C19 and C110.

which 37 out of 57 appear to be in a tightly connected component. Bsg is the also the only case where the difference set of the of 48 proteins that are included in the low, but not the medium confidence network model, have an fPSD context annotations. More specifically the set showed an enrichment for the term neurotransmitter receptor metabolic process ($p < 10^{-3}$), which is associated with Gad1 and Gs2. Gad1 is associated with GABA synthesis (Küppers et al., 2003) and Gs2 with glutamate catabolism (Featherstone et al., 2002).

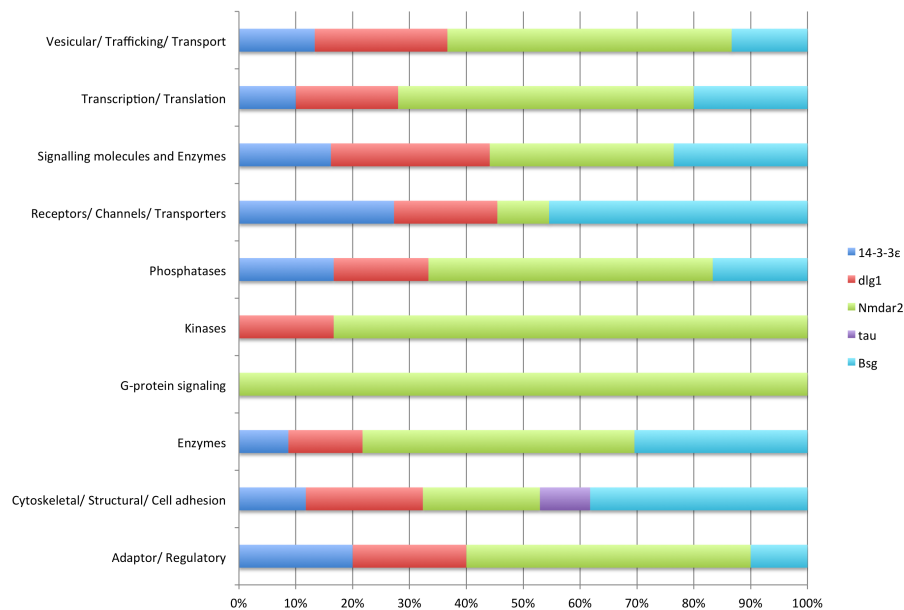
6.2.4.6 Proteins family distribution

From a protein family distribution (Figure 6.15) point of view 14-3-3 ϵ and Bsg complexes include 8 out of 10 families of preys, with Bsg's complexes including most of the Receptors/ Channels/ Transporters family and a substantial proportion of the total proteins. Nmdar2 complexes include members of all families. These complexes also include all G-protein signaling associated proteins as well as 5 out of 6 kinases, 5 out of 10 Adaptor/ Regulatory and 3 out of 6 phosphatases in the model. Interestingly, members of the Enzymes, Singalling molecules and Enzymes families, which are likely to include contaminants, were purified from all baits (except tau).

6.2.5 Mapping the fPSD to human disease

Using Biomart (Ensembl Genes version 64) and the Compara service we mapped all genes in the *tolerant* merged dataset to human orthologs and also retrieved the human orthologs disease associations from OMIM. The merged *tolerant* dataset showed associations with 141 diseases via human orthologs (28 diseases when only using one to one ortholog mappings). Proteins in the merged *tolerant* network showed associations with 44 diseases (2 diseases when only using the one to one ortholog mappings). Interestingly, in the case of the *tolerant* merged dataset scu, Nmdar2, Gdi and Sc2 were linked to forms of X-linked and autosomal mental retardation, with scu linked via a one to one orthology. Also, Idgf2, Idgf4, Chit and slgA were linked to schizophrenia,

Figure 6.15: Protein family composition of individual bait complexes.



out of which *slgA* with a one to one orthology. Additionally, *Mdh* and *Nmdar2* were also linked with forms of epilepsy. Further to these, other interesting peripheral nervous system diseases had links with proteins in the dataset such as distal hereditary motor neuropathy (via *Hsp83*) and agenesis of the corpus callosum with peripheral neuropathy (via *CG5594*). The evidence, although not as impressive as in the case of the mPSD, still highlights the relevance of the fPSD dataset with regards to its human homolog set.

6.2.6 Evolution of the fPSD

We have already presented data on the evolution of the mPSD (section 5.3.5), examining PSD evolution from the viewpoint of the mouse and human data. Here we describe a different perspective of the evolution of the PSD, from the viewpoint of the fPSD data. We obtained only one to one human, mouse and yeast orthologs of genes in the *tolerant* merged dataset via *Compara* (version 59). Results of the homolog retrieval

are shown in Table 6.5. Focusing on one to one homologs means excluding interesting proteins that have undergone gene family expansion (e.g. *dlg1* - the mapping of these is one to many by definition). These homologies, however, are examined in the next chapter.

A similar number of one to one homologs of the fly genes were retrieved from mouse and human, while fewer were found in yeast in accordance with previous findings. We looked up genes found in these homolog lists in the updated Emes et al. dataset in Chapter 5, which contains homologs of PSD genes across species including human, mouse and yeast. These queries showed that 30 of the one to one human homologs, 36 of the mouse one to one homologs and 23 of the yeast one to one homologs were in that list, linking these directly to PSD complexes. These numbers are quite satisfactory given the fact that we have knowingly excluded many PSD proteins which have undergone family expansion.

A breakdown of the results is shown in Figure 6.16. We can observe how most of the aforementioned hits map to the general (consensus) human PSD, rather than its core sub-complexes like NRC/MASC or the PSD-95 associated proteins complex (Figure 6.16A). Most of these hits belong predominantly to the Signalling molecules and Enzymes family and to the Transcription/ Translation and Vesicular/ Trafficking/ Transport families (Figure 6.16B). Interestingly, the common hits between the one to one homolog lists include 12 proteins (Figure 6.16C), out of these 7 are in the Signalling molecules and Enzymes, 2 in the Transcription/ Translation, 2 in the Vesicular/ Trafficking/ Transport and 1 in the Receptors/ Channels/ Transporters families. This can lead to the conclusion that the part of the fPSD that has not undergone gene duplication belongs mostly to the Vesicular/ Trafficking/ Transport and Signalling molecules and Enzymes families and maps to the “general” part of the consensus PSD, rather than the components more closely associated with the membrane.

Furthermore, when we compared the found fly homologs of the human PSD in the updated Emes et al. dataset with the *tolerant* merged dataset we found that out of the

Table 6.5: Homolog retrieval results for genes in the *tolerant* merged dataset.

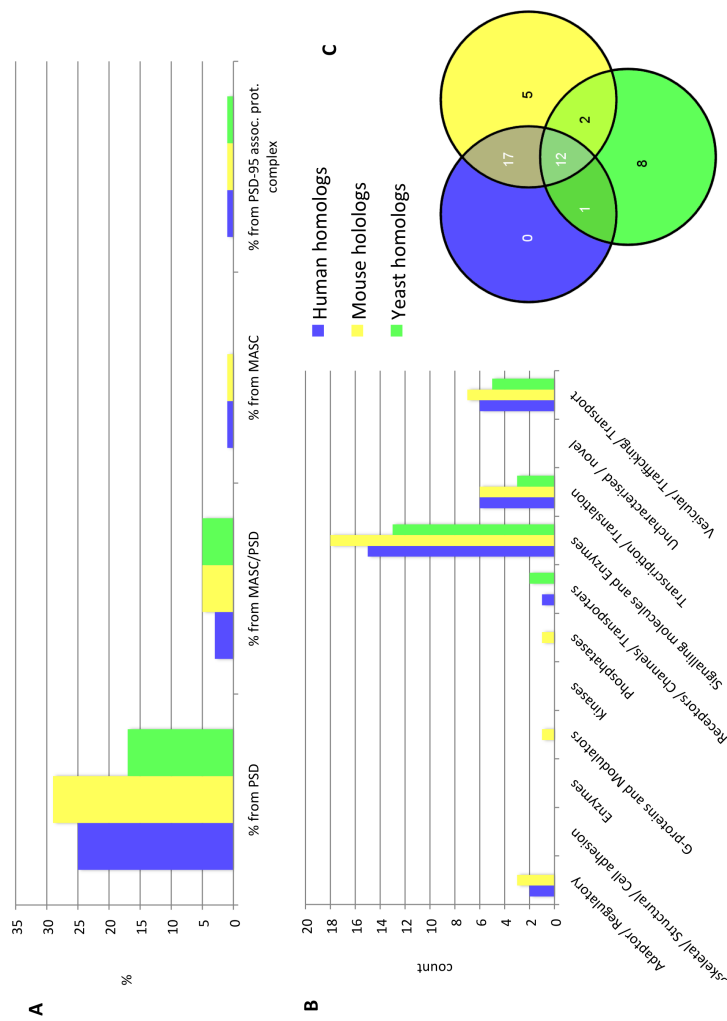
Total genes in fPSD model	Human homologs (one to one)	Mouse homologs (one to one)	Yeast homologs (one to one)
402	129 (~32%)	124 (~31%)	89 (22%)

402 genes in the dataset 96 (23.8%) had been predicted according to that comparative genomics approach. This at least partially corroborates this 23% of the *tolerant* merged dataset as likely by prediction.

6.3 Concluding remarks

Overall this chapter presented the first draft protein interaction network model of the fPSD based of proteomics data as well as an analysis of its architecture. However, affected by the shortcomings of the specific method, contaminants, and lack of quality interaction or annotation data, the model captures a number of the molecules involved in the fPSD and associated pathways. This, along with annotation similarities establishes a basis of a similar functionality to the mPSD. The fPSD was also hypothesised to have a lower complexity. We believe that this is reflected by the lower counts and enrichments of domains and annotations. Although hard to distinguish at this point, the latter could be due to the presence of contaminants or misannotations that lower the enrichment of the aforementioned PSD functionality. However, some additional support for the lower enrichments reflecting lower complexity comes from looking at isolated modules of the network (e.g. C11 and C12) where we see a very similar structure to mPSD clusters, with the only difference being the lack of family expansion. This results in e.g. specific domains and their annotations having a comparatively lower count. Also, when looking at the evolution of the proteins found here we find almost 1/4th being in list generated by comparative genomics approaches based on mPSD. Given that the aforementioned examination also excludes key proteins such as the NMDA receptor and *dlg1*, we consider this as giving some additional confidence

Figure 6.16: Summary of fPSD one to one homolog query in the human PSD associated genes homologs list across 18 species. Homologs for the *tolerant* merged fPSD dataset were retrieved from human, mouse and yeast. A) Homolog one to one hits from the list separated by which sub-complex they belong to in the human PSD, showing how PSD complexes have higher representation compared to NRC/MASC or PSD-95 associated protein complexes. C) Family distribution of hits, showing that while most families are represented the dataset is biased towards specific families (e.g. Signalling molecules and Enzymes) D) Venn diagram showing the overlaps between hits from different sets of fPSD one to one homologs showing a core of 12 proteins conserved from yeast to humans.



in the results.

It is evident from the results that the parts list and the network model captured part of the protein synthesis and turnover, signalling - including kinases, phosphatases, G-proteins and other enzymes, cytoskeletal and vesicular components of the fPSD. It is worth mentioning that the majority of the isolated proteins are soluble and in some cases the annotations did not provide any evidence of their relative localisation with regards to the postsynaptic membrane. This highlights one issue of the applied method, the fact that, although steps were taken in order to purify parts of the membrane associated proteins, it is not optimised to isolate them. Further improvements to the proteomics protocol including a more detailed study of the fatty layers of the protein extract or the application of a detailed study of separation might give better insights regarding the membrane associated proteins found in the fPSD. Use of different detergents would be the first suggested strategy. It is known that detergents such as DOC and ComplexioLytes can be successful in solubilising receptors and ion channels (Li et al., 2010). This would resolve issues regarding another key family in PSD complexes, that of Channels and Receptors. A key example of the previous is the NMDA receptor, which in spite of having been shown to mediate memory consolidation in *Drosophila* (Xia et al., 2005) was not isolated in these purifications as it was not in the only other previous proteomic characterisation (Emes et al., 2008), where it was - strictly speaking - present only as a peptide bait.

To our opinion the drawbacks of this particular model are the presence of contaminants and the issues we found with GO annotations. It is hard to truly dissect the effect of these two to our final conclusions. Regarding the contaminants in the model we can say that while there are some more obvious cases (e.g. C110), most cases are ambiguous. For example, one cannot exclude the protein synthesis machinery from the PSD, but such close association as the one found here could be suspicious. Similarly, the high presence of Signalling molecules and Enzymes family, could indicate contaminations - since these soluble proteins are abundant and known to contaminate (Chen

and Gingras, 2007). On the other hand the potential of some compensation signalling mechanisms (in lack for such a wide repertoire of kinases) could not be excluded. Regarding GO annotation there are several issues such as protein function pleiotropy with respect to annotations in the nervous system. GO annotations are known to capture pleiotropic effects of protein functions in a non-neuronal context which can result to misleading and skewed annotations (Inlow and Restifo, 2004). Also, since the fly synapse has not been extensively studied from the PSD perspective it might be the case that some of molecules have not been annotated yet, or have been annotated from other perspectives of function.

Contamination issues could be addressed by changes in the experimental method. Fly heads, due to their size are very hard to handle in order to remove a sufficient number of brains, quickly enough so that the protein complexes remain intact and do not degrade. A potential solution would be to approach this problem after the collection of the heads but before the extraction of the protein. More specifically this can be achieved with the use of a synaptosome preparation with gradient centrifugation. This could possibly reduce potential contaminants and also give the proteomic analysis a subcellular location context that is closer to the PSD also addressing the issue of transmembrane or membrane associated proteins. The successful isolation of a synaptosome preparation has been attempted few times in insects with unclear results. When designing the experiment we were not convinced of the efficiency of the method - since it has not become standardised although it appeared in the literature around the 1980s. More specifically synaptosomes have been isolated from various mammalian nervous tissues. However, the conditions established for mammalian systems are not necessarily suitable for insect systems (Breer and Jeserich, 1980). There have been few attempts to actually isolate synaptosomes from other insects (Breer and Jeserich, 1980, Breer and Knipper, 1984, Torrence-Campbell et al., 1991) including two in *Drosophila* (Kelly, 1981, Ramarao et al., 1987). The latter work by Ramarao et al. proposes a method using ficoll floatation technique, which is an adaptation of the method of Breer

and Jeserich. This method circumvents the use of liquid nitrogen - which can damage the fine structures of the synaptosome but is also necessary in our high-throughput head collection protocol, in order to avoid protein degradation and also be able to collect enough tissue for a proteomic pull down. Although not impossible, developing a method that balances avoiding protein degradation with the retrieval of intact synaptosomes, would be very time-consuming for this project. Also most of this, rather limited, knowledge is based on NMJ synapses rather than neuron synapses. Another consideration that has also been reported, is that synaptosome preparations provide limited amount of material for affinity purification, a scenario which is not ideal for the general fPSD mapping we were trying to achieve. In conclusion, we believe that while using a synaptosome preparation would be a valid strategy, its optimisation for an initial mapping of the fPSD would be exceeding what is necessary, however, future endeavors would probably need to take this approach into account.

Chapter 7

Comparative analysis of PSD complexes and interaction networks

7.1 Background

After obtaining datasets and reconstructing protein interaction network models describing PSD complexes of two different organisms, the next step was to compare them in a qualitative and quantitative manner. Such comparison could provide insight on the differences and similarities between the datasets and models from two perspectives of interest. The first perspective will allow us to highlight differences in the proteomics methods and data availability for the model reconstruction. This takes into account that the fPSD complexes were isolated using a less optimised approach and also that the protein interaction models were reconstructed with data of lesser quality and coverage. The second, and perhaps most relevant, perspective is that of the comparison with respect to the constituent parts and organisation of these two PSD molecular machines. As hypothesised in the introduction, the fPSD model shows evidence of lower complexity with regards to its constituent parts. However, as shown in the previous chapter, the fPSD model has a modular architecture comparable to this of the mPSD, and using this modular architecture we will also present evidence on how some module

Table 7.1: Homologies between genes across the Union and fPSD network models. Abbreviations of orthology types are o2m: one to many; o2o: one to one; m2m: many to many. Continued in Table 7.2

Union gene symbol	fPSD gene symbol	Homology type (Compara)	Union gene symbol	fPSD gene symbol	Homology type (Compara)
Atp1a1	Atpalpha	<i>o2m</i>	Gapdh	Gapdh1	<i>o2m</i>
Atp5a1	blw	<i>o2o</i>	Gapdh	Gapdh2	<i>o2m</i>
Atp5b	ATPsyn-beta	<i>o2o</i>	Grb2	drk	<i>o2o</i> (<i>apparent</i>)
Atp5c1	ATPsyn-gamma	<i>o2o</i>	Grin2a	Nmdar2	<i>o2m</i>
Cfl1	tsr	<i>o2m</i>	Grin2b	Nmdar2	<i>o2m</i>
Cltc	Chc	<i>o2o</i>	Grin2d	Nmdar2	<i>o2m</i>
Dlg1	dlg1	<i>o2m</i>	Lin7a	veli	<i>o2m</i>
Dlg2	dlg1	<i>o2m</i>	Mapk1	rl	<i>o2m</i>
Dlg3	dlg1	<i>o2m</i>	Mapk3	rl	<i>o2m</i>
Dlg4	dlg1	<i>o2m</i>			

or modules have been conserved in evolution.

In order to compare the mouse and fly PSD complexes we chose the two datasets and their resulting PPIN models respectively. More specifically, for the mouse PSD we chose the Union dataset (253 proteins) and protein interaction network model (164 proteins and 458 interactions), as described in Chapter 5. For the fly PSD we chose the fPSD dataset and model as described in Chapter 6. Since there were different versions of the latter dataset and model we chose the largest, namely the merged version used to reconstruct the model, described section 6.2.3.1 on page 166, hereon referred to as fPSD dataset (402 proteins) and protein interaction network model (149 proteins and 495 interactions).

7.2 Results

The following subsections describe the comparisons of certain annotation aspects (e.g. proteins domains, families, GO annotation), along with a comparative interactomics approach, where we attempt to highlight conserved components between the two networks.

7.2.1 Homology

We retrieved ortholog data from Ensembl's compara service (version 50) for the Union and fPSD datasets. We found 70 orthology relations of different types (many to many - 24, one to many - 37, one to one - 9), between 64 (25.3%) genes of the Union and 52 (12.9%) genes of the fPSD dataset. The one to many category includes the genes that have undergone gene family expansion such as *dlg1*, *Nmdar2*, *spectrin* (beta-Spec) and *Atpalpha*. When examining the subsets of genes that have proteins in the network models these numbers changed to 32 (19.5% of proteins in the network) and 27 (18.1% of proteins in the network) genes respectively, with a total of 37 orthology relations between the network models. These ortholog genes encode proteins existing in both datasets and include the representatives from most functional families¹. A closer look at the subfamilies reveals that these are key protein types with known PSD functionality such as PDZ-domain containing scaffolders and non-PDZ-domain containing scaffolders, Ser/Thr Kinases, Protein Phosphatases, G-protein signaling proteins, Glutamate Receptors, Voltage-dependent anion channels and Motor Proteins.

When looking at the subset of conserved constituent parts within the protein interaction network models (Tables 7.1 and 7.2), we can notice a trend in key protein components being conserved. An exception to this is the G-protein signalling family which has 5 representatives in the fPSD dataset conserved constituent parts core but none in the network's conserved constituent parts core. Results are similar, but not

¹with the exception of Histones, Extracellular Proteins and Uncharacterised/ novel

as prominent for the Kinases and Receptors/ Channels/ Transporters families. Manual inspection shows that the majority of the above can be attributed to lack of interactions or missing interaction partners in the fPSD network model.

The above is evidence of conservation at least at the level of constituent parts, showing that although the mPSD contains expanded gene families (e.g. DLGs) a substantial number of these appear in the fPSD dataset and network, revealing a potential of a molecular machine of similar composition but lower complexity. How the latter is actually organised in conserved modules is approached in subsection 7.2.5.2, where network connectivity is taken into account.

7.2.2 Families

From a functional annotation perspective, we can compare the distribution of families found in the datasets and the network models. Their ratios to the total number of constituent parts are in Figure 7.1, where significant (χ^2 test p-value < 0.05) differences are highlighted. There are significant composition ratio differences within most families, with the Union dataset having a higher ratio of Adaptor/ Regulatory, Cytoskeletal/ Structural/ Cell Adhesion, G-protein signaling and Kinases. fPSD ratios are higher for the Enzymes, Signalling molecules and Enzymes and Transcription/ Translation families. This propagates in the ratios within the network models, with the exception of the Enzymes family, where the difference is not significant anymore. This could potentially highlight an issue with a number of contaminants which belong to the latter family. After all it has been shown that metabolic enzymes are more abundant and often contaminate purifications (Chen and Gingras, 2007). It is also worth noticing how the Phosphatases and Vesicular/ Trafficking/ Transport families' ratios are not significantly different between the models, possibly reflecting good affinity purification of these families, but also suggesting a similar extent of involvement in the two molecular machines.

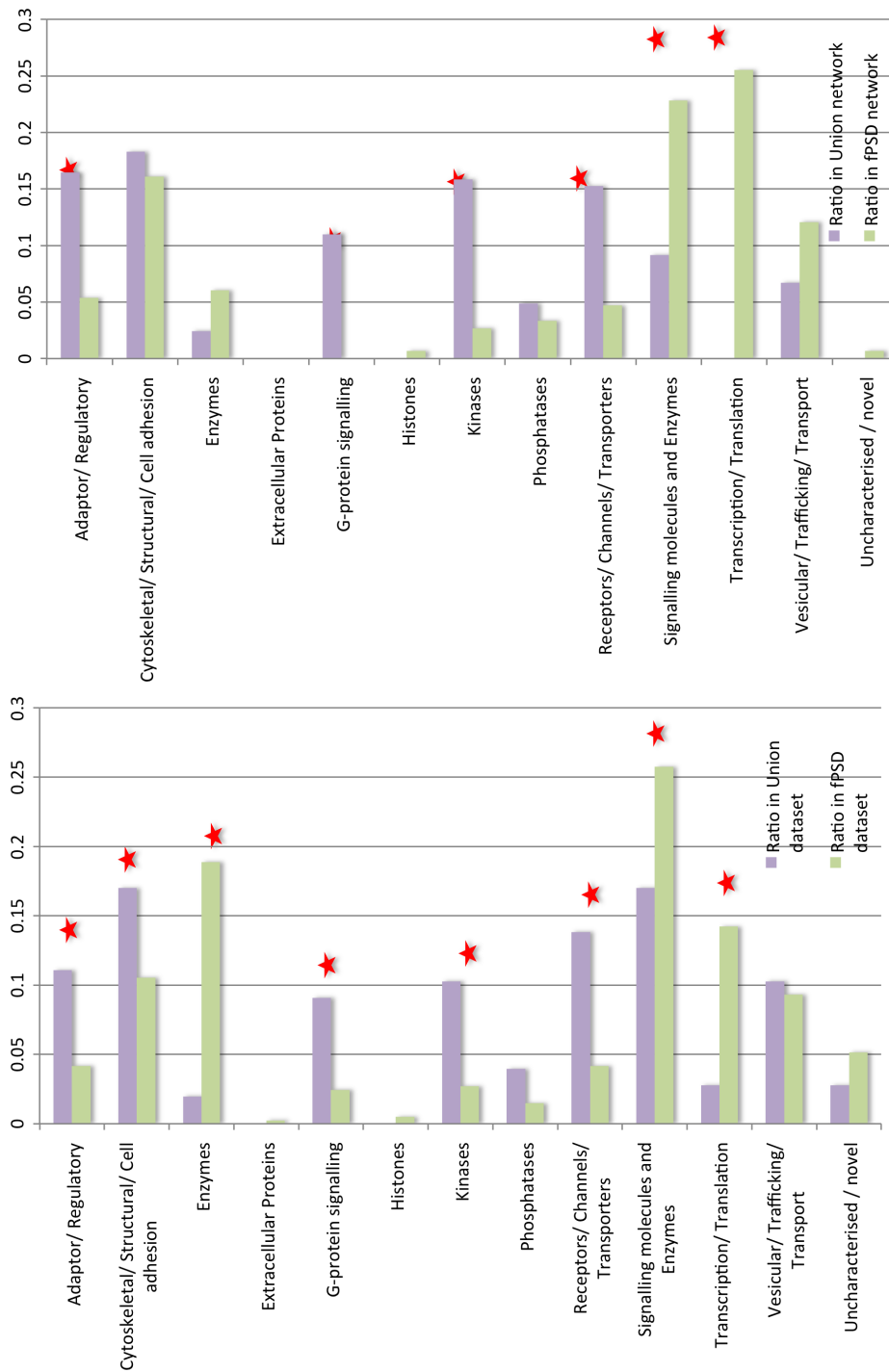
Another observation is how the fPSD network contains more than twice the ratio of

Table 7.2: Homologies between genes across the Union and fPSD network models. Abbreviations of orthology types are o2m: one to many; o2o: one to one; m2m: many to many. Continued from Table 7.1

Union gene symbol	fPSD gene symbol	Homology type (Compara)	Union gene symbol	fPSD gene symbol	Homology type (Compara)
Mpp3	skf	<i>o2m</i>	Prkcc	Pkc53E	<i>o2m</i>
Mtap2	tau	<i>o2m</i>	Prkcc	inaC	<i>o2m</i>
Myh10	zip	<i>o2m</i>	Prkce	Pkc98E	<i>o2m</i>
Myh9	zip	<i>o2m</i>	Slc25a4	Ant2	<i>o2m</i>
Nsf	comt	<i>o2m</i>	Slc25a4	sesB	<i>o2m</i>
Nsf	Nsf2	<i>o2m</i>	Spnb2	beta-Spec	<i>o2m</i>
Pgk1	Phosphoglycerate kinase	<i>o2m</i>	Spnb3	beta-Spec	<i>o2m</i>
Prkcb	inaC	<i>o2m</i>	Tpi1	Tpi	<i>o2o</i>
Prkcb	Pkc53E	<i>o2m</i>	Ywhae	14-3-3epsilon	<i>o2o</i>

Signalling molecules and Enzymes family compared to the Union network. This can be interpreted, from a functional perspective, as a compensation for the lower ratio in other signalling related families (e.g. Kinases and G-protein signalling), however, the factor of contamination is still present in this case. Another big difference both in the dataset and network contexts is the ratio of the Transcription/ Translation family, which is much higher in the fPSD case. The majority of this family represents translation associated proteins (e.g. ribosomal proteins) and although many mass spectrometry oriented groups consider them contaminants (Peng et al., 2004), they have been found in the PSD numerous times (e.g. Steward and Falk, 1991, Krichevsky and Kosik, 2001). Furthermore, local transcription is considered part of the mechanism for plastic phenomena (Gardiol et al., 1999).

Figure 7.1: Family ratios in the Union and fPSD datasets (top panel) and networks (bottom panel). Families indicated with a red star have significant (χ^2 test p-value < 0.05) difference in their ratio within the datasets or networks.



7.2.3 Protein domains

We then examined if there are significant (χ^2 test p-value < 0.05) differences in the counts² of specific domains within the datasets (Figure 7.2) and networks (Figure 7.3). With the exception of one, all domains with significant differences have a higher count in the Union dataset. The NAD(P)-binding domain, a domain found in enzymes, with some neuronal function (e.g. Nitric oxide synthase) has a comparatively high count in the fPSD data because of its presence on many enzymes with metabolic functionality (Enzymes family) or signalling functionality (Signalling molecules and Enzymes family).

Most of the domains with significant differences have some direct association with the signalling or receptor associated processes of the PSD (see Table 5.1, Chapter 5). As also reflected by the differences in the ratio of the Receptors/ Channels/ Transporters family, the complete lack of ion channels is evident in the fPSD dataset by the lack of the associated domains (e.g. Ionotropic glutamate receptor, Potassium channel domains, BTB/POZ). A significant difference in the counts of phosphorylation associated domains (e.g. Serine-threonine/tyrosine-protein kinase, Protein kinase, catalytic domain, etc) and SNARE protein domains (t-SNARE) can also be noted. In all the above cases the Union dataset has more proteins with such domains, which in many cases are absent from the fPSD dataset (e.g. Potassium channels and glutamate receptors).

When looking at the network data the observations are similar, although none of the domains with significant differences is completely absent from the fPSD. We can also notice how there are some domains (ATPase, AAA-type, conserved site, Actin, conserved site, Actin/actin-like conserved site and Proteasome, subunit alpha/beta) are absent from the Union network. If we assume that the presence of proteins with such domains in the network reflects the existence of the associated interactions, then we can

²We consider that counts are a better measure for protein domains since percentages in the dataset are not as clearly defined (it is possible that not all domains are annotated)

credit these differences to biases in the protein complex isolation methods, although, as mentioned in Chapter 6 the proteasome is a known component of the PSD (Haas et al., 2007).

Overall we can say that the Union dataset seems to have a higher count of all PSD signalling and neurotransmitter receptor associated domains. This could be partially attributed to: a) the purification methods used, which did not isolate enough trans-membrane proteins in the fPSD case, b) the expansion of gene families carrying specific domains, e.g. the Dlg family (see Emes et al., 2008 and subsection 5.3.5), and c) differences in the complexity of the two PSD complexes. While the first explanation is a technical issue, the latter two are indicative, as previously hypothesised, of less complex fPSD, which maintains the same protein functions, but with less diversity.

7.2.4 GO annotation

We applied the modified version of the method by Cai et al. (2006), in order to examine significant (FDR corrected p -value < 0.05) differences in the GO annotations of the Union and fPSD data. More specifically we applied the method to both the datasets and the subset of the datasets, which constitute the protein interaction networks.

The results are presented in Treemap illustrations³, generated using Revigo (Rivals et al., 2007) (<http://revigo.irb.hr>). This specific visualisation is for 2 levels of GO annotation (the starting level is decided by the Revigo algorithm). Revigo also clusters terms (squares) based not only on if they have a common ancestor but also based on semantic similarity of the terms.

From a GO BP annotation perspective, of the terms which have a significant comparative enrichment in the fPSD (Figure 7.4) dataset, we see the presence of proteasome components again. Among the other terms we see either fruitfly specific (e.g.

³Treemaps display hierarchical (tree-structured) data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub-branches. These treemaps retain the parent-child information inherent from the GO structure. More specifically terms appear in nested squares according to the count of their or their child terms occurrences.

Figure 7.2: Significant differences in normalised domain counts in the Union and fPSD datasets.

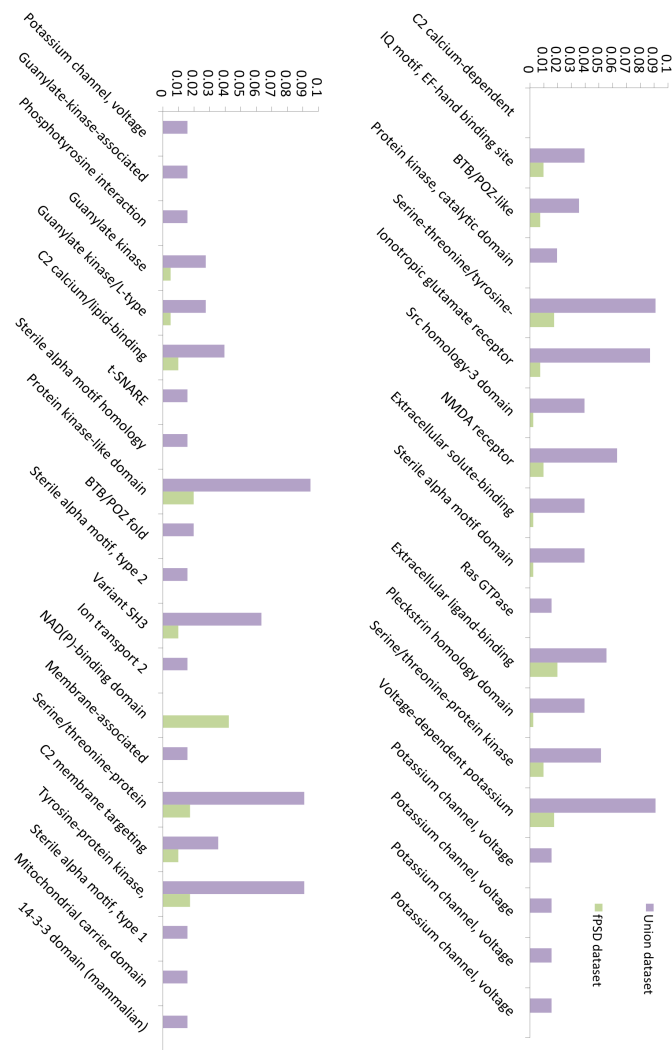
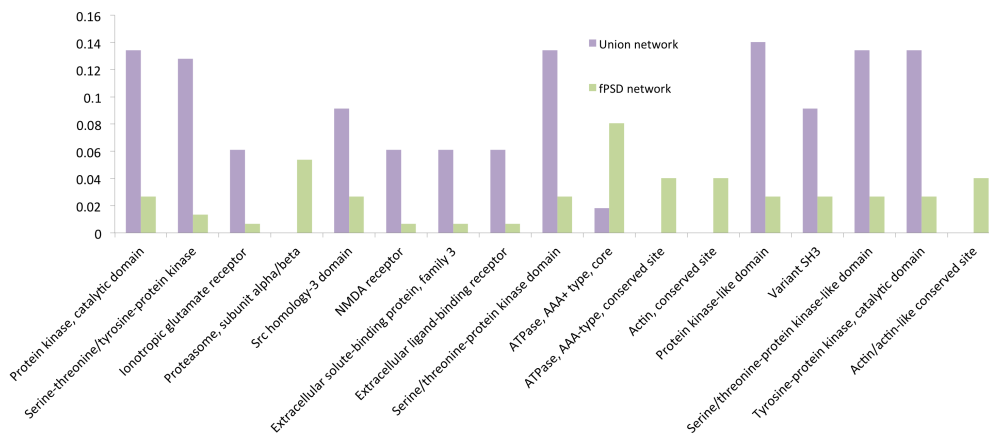


Figure 7.3: Significant differences in normalised domain counts in the Union and fPSD networks.



instar larval or pupal development), transcriptional mechanism associated, and a set of terms associated with phagocytosis, probably due to contamination. The terms that have a significant comparative enrichment in the Union dataset (Figure 7.5) on the other hand, encompass a majority of PSD and PSD signalling associated terms, showing how the Union dataset is comparatively more enriched in such processes. From a GO MF annotation perspective both terms with a significant comparative enrichment in the fPSD (Figure 7.6) and the in the Union (Figure 7.7) datasets show functions that are reflected at GO BP level (e.g. peptidase activity for the proteasome or kinase activity for protein phosphorylation). The results are similar to the ones of the GO BP comparative enrichment, showing the Union dataset significantly more enriched for PSD associated GO MFs than its fPSD counterpart. Finally, from a GO CC annotation perspective, while there are no significant comparative enrichments for the fPSD dataset, the Union dataset (Figure 7.8) shows comparative enrichment for many known PSD associated subcellular locations (e.g. postsynaptic membrane) as well as for membrane bound or transmembrane proteins. Overall the comparison suggests that key annotation terms appear more enriched in the Union rather than the fPSD dataset, where in most cases they are present but in low counts. This can be attributed to reasons such as contamination of the fPSD dataset or overall better quality of the Union dataset. However, we have to note that this is one of the cases where GO annotation quality plays a significant role. We have found a number of cases where the annotation of fly orthologs is poorer compared to that of the mouse and we believe that this plays a role in the results above.

7.2.5 Comparative interactomics

7.2.5.1 Comparison of basic network architecture

When examining the path length distribution (Figure 7.9, top), we notice how the fPSD has a wider distribution (with lower frequencies), meaning that there are some longer

Figure 7.4: Terms with comparative enrichment in the fPSD dataset (GO BP).

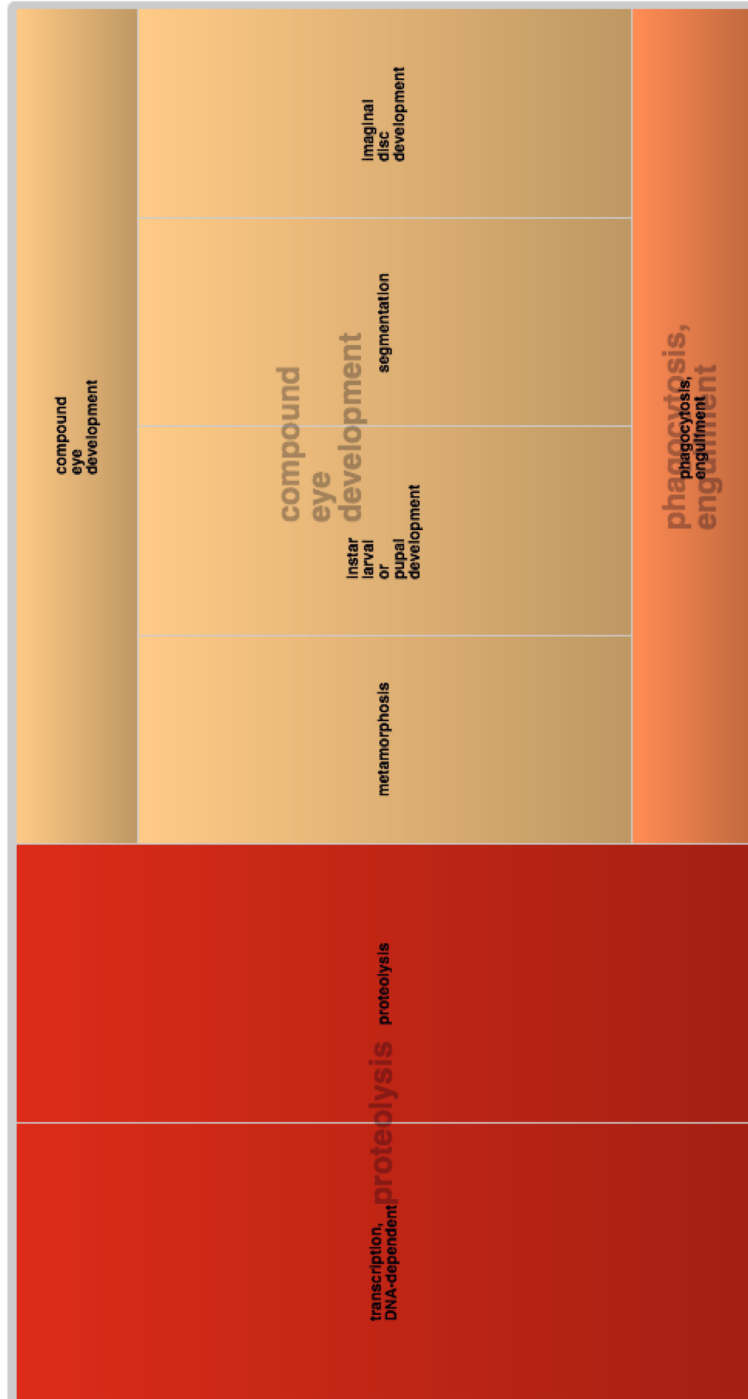


Figure 7.5: Terms with comparative enrichment in the Union dataset (GO BP).

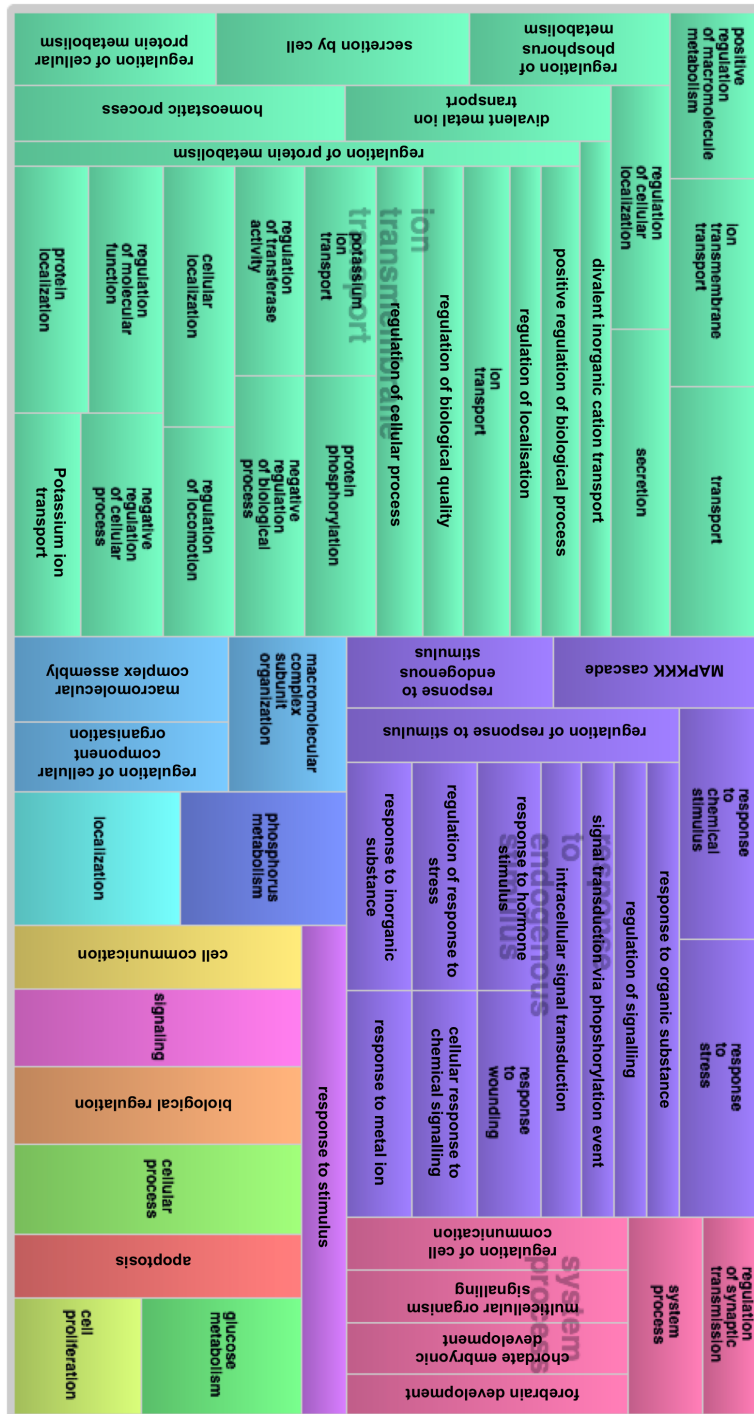


Figure 7.6: Terms with comparative enrichment in the fPSD dataset (GO MF).

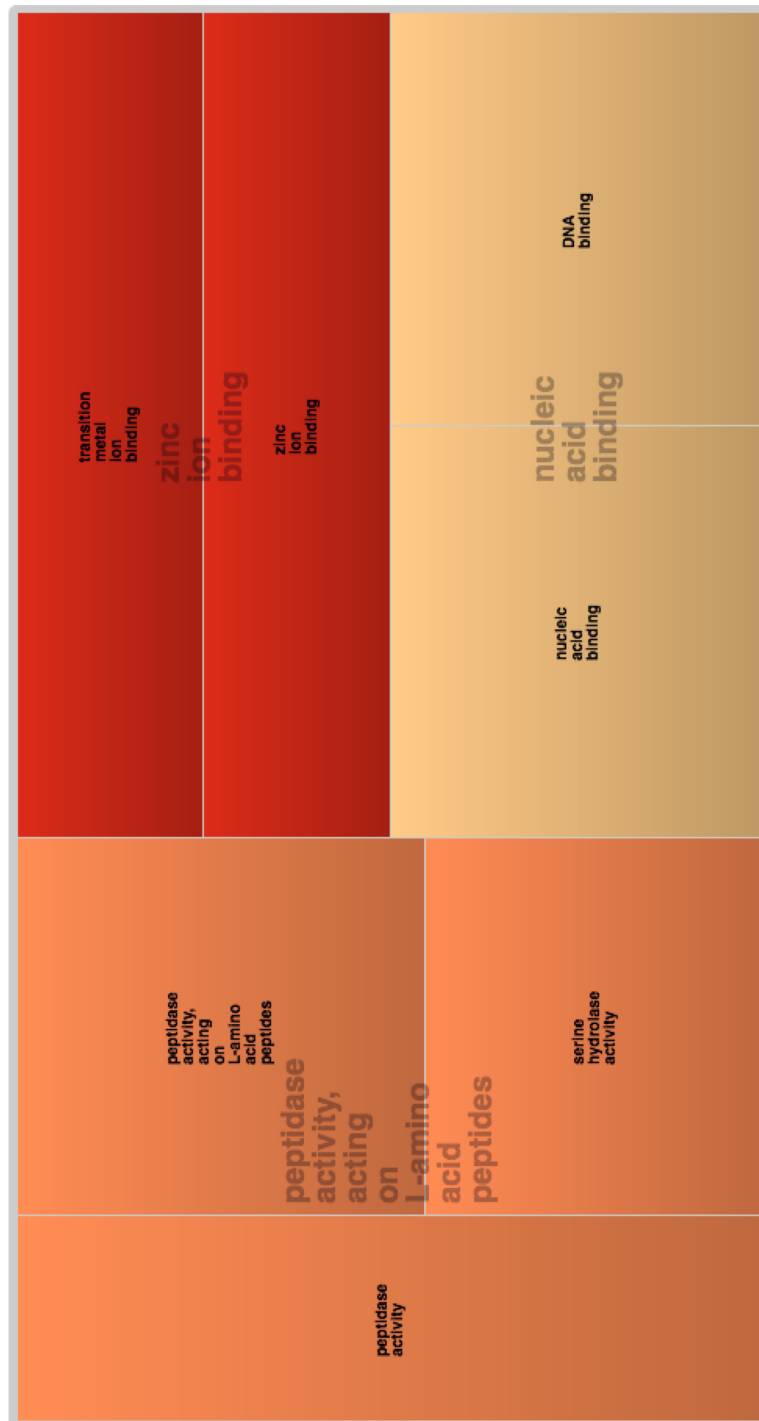


Figure 7.7: Terms with comparative enrichment in the Union dataset (GO MF).

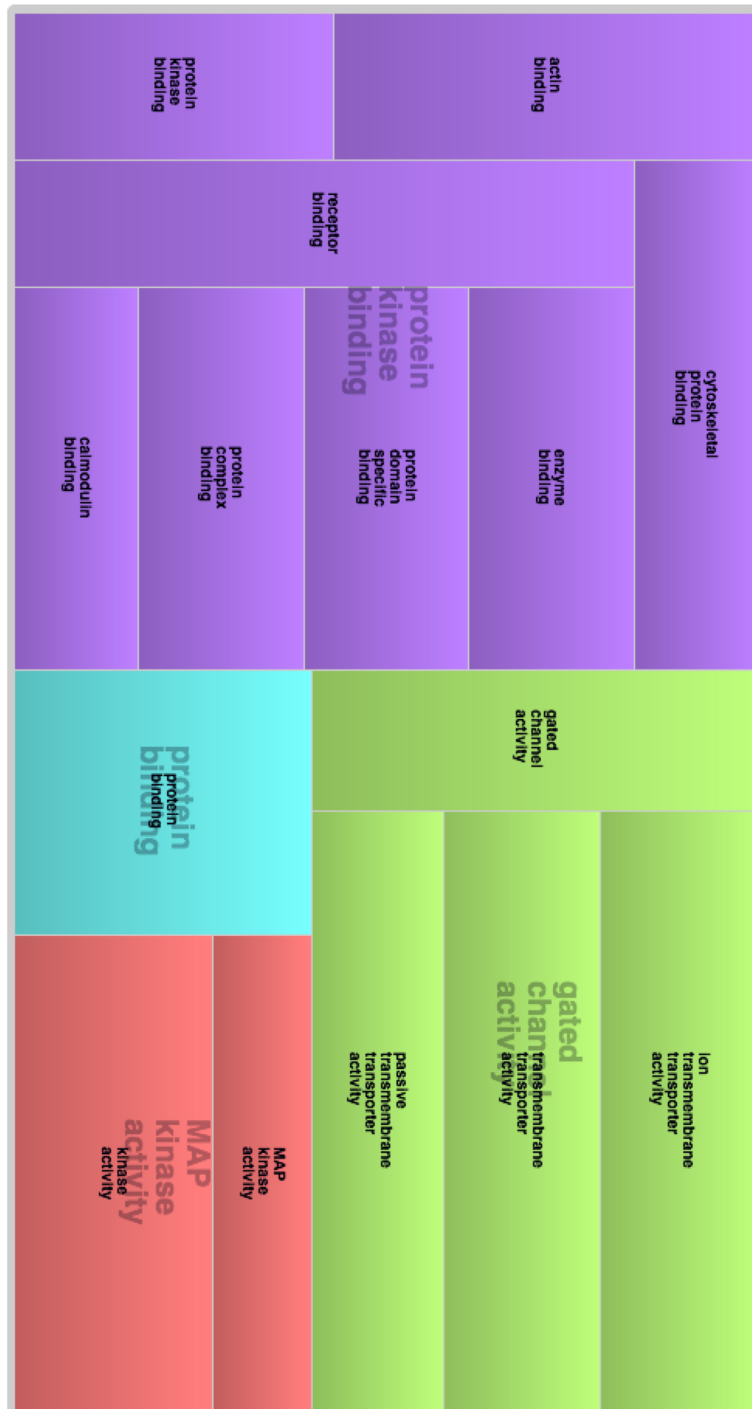
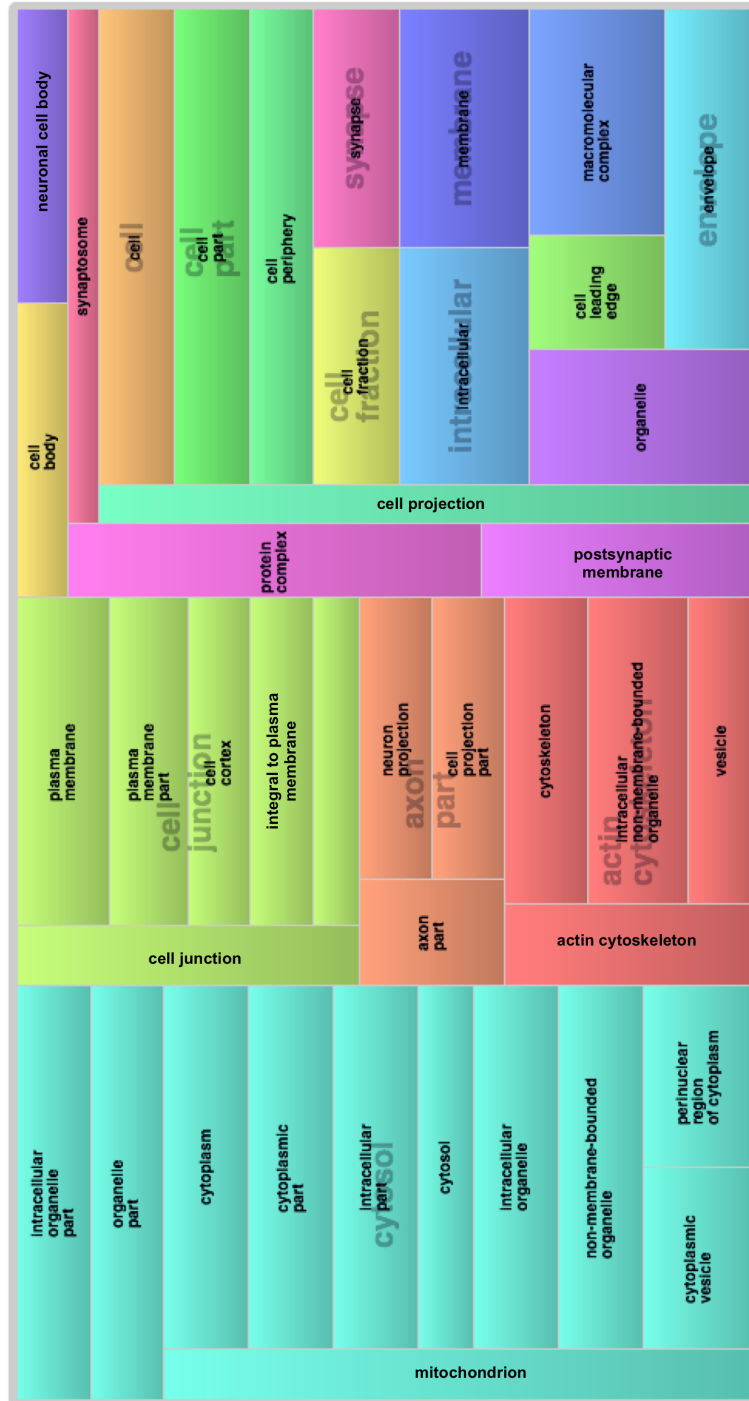


Figure 7.8: Terms with comparative enrichment in the Union dataset (GO CC).



paths in the network. We believe however, that this is because of biases towards specific parts of the fPSD network (e.g. C19), which have more interactions present compared to nodes in the rest of the network, as a result of a literature bias towards the more well studied proteins (e.g. translation associated protein of C19). This becomes evident when analysing the network, while ignoring C19, where the path length distributions become more similar (although the number of interactions is reduced). The data also shows that the path distributions are significantly different (Figure 7.9, bottom).

The average shortest path (ASP) length metric that represents cross-talk between modules as well as betweenness centrality. The fPSD protein interaction network appears to have an average ASP length of 3.56 ± 1.48 while the Union protein interaction network has an average ASP length of 3.35 ± 0.67 . However, although the average ASP lengths are similar, the fPSD protein interaction network's ASPs are affected by the highly dense interactions in C19 (or possibly contaminants), in combination with a potential lack of interaction information between members of the network (subsection 6.2.3.5). For these reasons, while the Union network (subsection 5.3.2.3) shows Adaptor/Regulatory, Kinases and Receptors/Channels/Transporters in the top rankings for shortest individual ASP length, this is not the case for the fPSD network. There, the same families are lower in the ranking, having however, still considerably low ASP lengths. For this reason we focused on the conserved homology core of constituent parts found in the networks (see 7.2.1). Within this subset of nodes we notice that in most cases we see that the average ASP length value of functional families tested (Adaptor/Regulatory, Kinases and Receptors/Channels/Transporters) is higher than the average of the network. However, when we examined the betweenness centrality of this subset we noticed something different. More specifically the average betweenness centrality of the Cytoskeletal/Structural/Cell adhesion, Kinases and Receptors/Channels/Transporters appears higher than average, in contrast with that of the Union network. Although this is not necessarily indicative of a more centralised connectivity we believe it still shows the relatively central position of the latter families in the

fPSD network model, regardless of the effect of interaction data quality and availability on the network.

We also compared metanetworks of the Union and fPSD protein interaction models. The metanetworks collapsed on family and are shown in Figure 7.10. An immediate observation is the difference between the variety of inter-family interactions. This of course is attributed partially to the differences in numbers, however, it could be a case of missing interaction data. This visualisation is also useful for summarising observations from the previous sections, highlighting differences in the counts of different families.

7.2.5.2 Identification of conserved components

With both the Union and fPSD network models available it would be interesting to examine the two in a comparative manner. A comparative interactomics approach augments the orthology data, by giving a context of the conserved network components which contain them.

We applied the NetworkBLAST algorithm (default parameter values) to the two networks after running an all against all BLAST (default parameter values, BLOSUM62 substitution matrix) for the proteins in the Union and fPSD network models. The e-values resulting from the BLAST run were used as a similarity measure for NetworkBLAST, along with the protein interaction data. The results of the algorithm are illustrated in Figure 7.11. The figure illustrates the two conserved components of the fSPD and Union networks members, connected by homology (zig-zag edges in Figure 7.11). Proteins within one conserved component interact with protein interactions, most of which are partially conserved between the two components. Although there are no long pathways conserved, what we see preserved is the typical architecture that emerges from all PSD related data. That architecture includes cytoskeletal proteins (Actn / ACTN2, ACTN4) interacting with scaffolding proteins (dlg1 / DLG1, DLG2, DLG3, DLG4) which in turn interact with receptors (Nmdar2 / GRIN1, GRIN2A,

Figure 7.9: Path length frequency (top) and cumulative frequency (bottom) comparison between the Union and fPSD protein interaction network models. A Kolmogorov-Smirnov test showed significant difference ($p = 10^{-5}$).

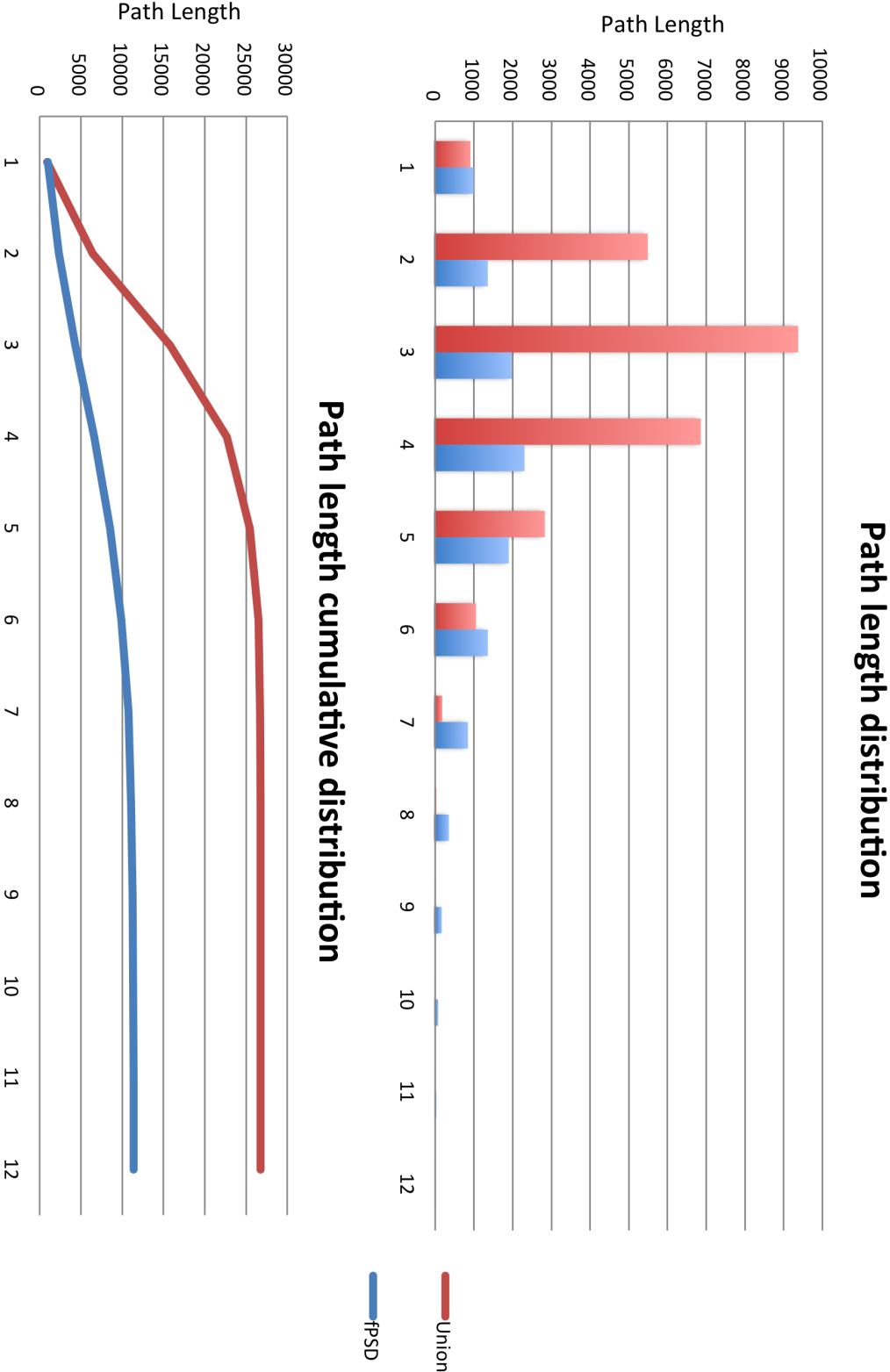
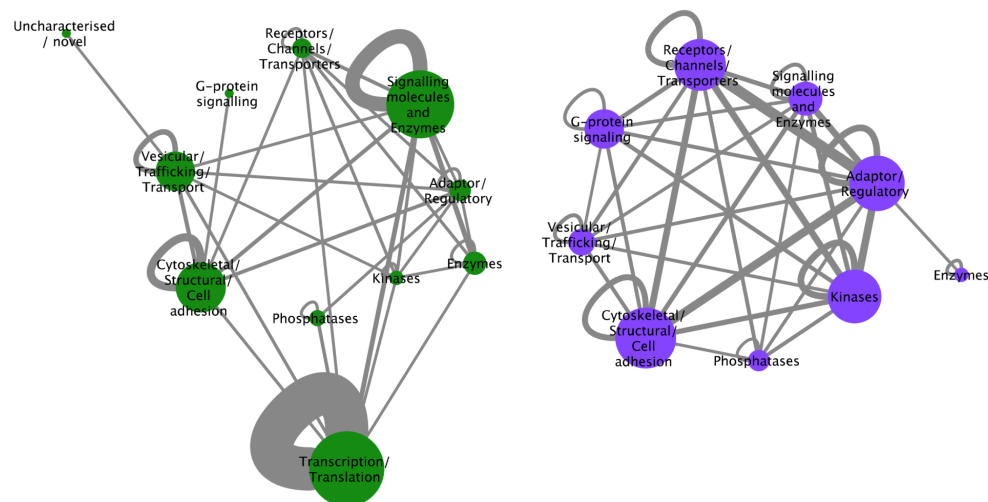


Figure 7.10: Family metanetworks of the fPSD (left) and Union (right) network models. Node size is proportional to family count and edge size proportional to family-family interaction counts.



GRIN2B, GRIK2). The receptors in turn interact with cytoskeletal proteins (e.g. *veli / LIN7A*) that assist with differential modulation (*veli / LIN7A* - see Iwamoto et al., 2004) or kinases as downstream signalling molecules that modulate signal transduction (*Pkc53E / PRKCC*, *Pkc98E / PRKCE*, *inaC / PRKCB* - see Lan et al., 2001).

What this result shows is the existence of a core component which is present in both networks and has evolved after the gene family expansion. This component is only a small fraction of both networks and this is not because of the lack of homologies, but possibly due to the confidence of the corresponding interactions, either due to the lack of interaction data or to their absence. Fox et al. (2009) recently showed that hub proteins tend to have more conserved interactions and this one of this cases. Also, recently Zinman et al. (2011) showed that interactions which are parts of functional modules are conserved at much higher rates than interactions which are not. In this case the conserved component of the network is speculated to be a functional module which had its interactions conserved. It has also been shown that protein interactions overlap at a low rate with conservation of binding partners in whole proteomes (Gandhi et al., 2006). In this case due to the potential partial nature of the datasets we can not test such findings. Finally, Figure 7.11 also recapitulates characteristic examples of the

evolution via gene family expansion mechanism proposed by Emes et al. (2008).

7.2.5.3 Semantic similarity of conserved components

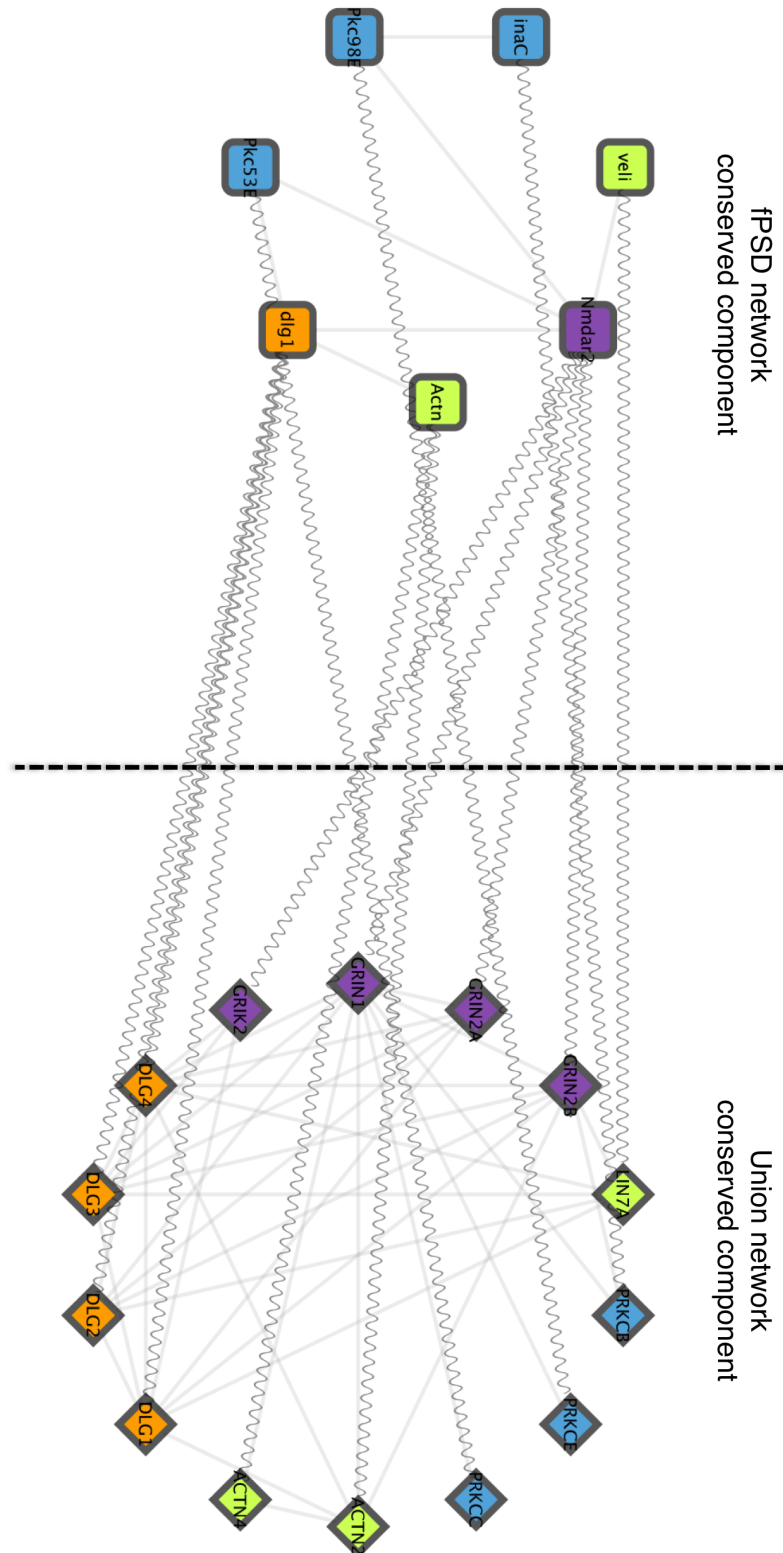
Using an adapted version of TCSS (Jain and Bader, 2010), we attempted to compare various subsets of the two networks regarding the semantic similarity of their annotations. Since semantic similarity is measured between 0 and 1, we needed to set a standard for high semantic similarity. For that reason we compared the semantic similarity between all pairs of Union and fPSD dataset homologs. The average semantic similarity was 0.49 ± 0.28 for GO CC, 0.51 ± 0.21 for GO BP and 0.69 ± 0.25 for GO MF.

The conserved network components of fPSD and Union showed averaged semantic similarities of 0.31, 0.33 and 0.24 for GO CC, GO BP and GO MF respectively (c.f. values for whole networks: 0.24, 0.24 and 0.17), with all values below the cutoffs. However, when examining homolog pairs, the semantic similarity results improve (Table 7.3). Interestingly, adding the first degree neighbours to each component results in a small reduction of semantic similarity (0.26, 0.28 and 0.23 for GO CC, GO BP and GO MF respectively), which still remains higher than for the whole networks. This can be considered as additional evidence that the conserved component (along with the respective immediate neighbours in the network models) show overall similarity in their functional and localisation annotations.

7.3 Concluding remarks

This chapter discusses a direct comparison between the representative datasets and protein interaction network models of the mPSD and fPSD. The comparison is done in from two distinct perspectives. We compared the data bearing in mind that it was the first time that fPSD complexes were isolated and characterised using a less optimised approach. In the beginning of this project we hypothesised that PSD complexes of the fly and mouse would be similar enough to be comparable both from a constituent

Figure 7.11: Conserved components of the fPSD (left) and Union (right) networks, as predicted by the NetworkBlast algorithm. Node colour represents functional families. Blue: Kinases, Orange: Adaptor/Regulatory, Purple: Receptors/Channels/Transporters, Lime: Cytoskeletal/Structural/Cell Adhesion. An interactive version of this figure with fully visible node labels is available in the additional material website (available in DVD format with this thesis).



parts and a network perspective. The results showed that while both datasets contained annotations relevant to PSD functionality, as it is at least understood from the mPSD data, the fPSD data contained these annotations with lower counts and enrichments. This reflects how the basic mPSD functionality is also present in the fPSD, although GO annotations for the fly data are more rare, since it has not been studied as extensively. Furthermore, as hypothesised, comparison between the datasets showed a conserved core of constituent parts. A small part of this core, was also reflected in the network models as conserved sub-networks of nodes and their interactions, reflecting again basic mPSD functionality present in the fPSD in the form of interacting proteins. This provided evidence not only for the conservation between the mPSD and fPSD complexes but also suggests that the mechanism of that conservation might be based on conservation of particular modules of the networks. As it has been mentioned before, both by Emes et al. (2008) and in Chapter 6, gene family expansion is one of the proposed mechanisms of evolution and this mechanism would be compatible with the results shown here, where the conserved component evolved by expanding with ortholog gene products, which maintain similar patterns of interactions.

Table 7.3: Semantic similarities between ortholog pairs within the conserved components of fPSD and Union.

Union genes	fPSD genes	Semantic similarity for GO CC	Semantic similarity for GO BP	Semantic similarity for GO MF
ACTN2, ACTN4	Actn	0.44	0.45	0.79
DLG1, DLG2, DLG3, DLG4	dlg1	0.55	0.38	0.32
LIN7A	veli	0.6	0.39	0
GRIN1, GRIN2A, GRIN2B, GRIK2	Nmdar2	0.67	0.34	0.8
PRKCC	Pkc53E	0	0.33	0.77
PRKCE	Pkc98E	0	0.58	0.77
PRKCB	inaC	0.42	0.33	0.77

Chapter 8

Discussion

8.1 General discussion

8.1.1 A broader view

This project describes our effort to catalogue and analyse the complex postsynaptic protein machinery by the reconstruction and analysis of PSD protein interaction network models based on proteomics data. Although the datasets and networks presented and discussed in Chapters 4 and 5 are models of the murine PSD, the effort of Chapter 6 is to model a non-mammalian PSD complex, by using *Drosophila* as a model organism. Furthermore, Chapter 7 compares the mouse and fly models, highlighting some emerging similarities in the data, which are also reflected in the network models. Besides the contribution of each chapter in methods, data and models there is a significant wider perspective towards which this work has contributed.

Perhaps the most exciting part of postgenomic-era biology was moving away from the single gene or protein studies towards the study of whole sets of molecules, working together via their interactions. It is the combination of constituent parts and interactions that gives these sets the emergent properties of a molecular machine. Studying something as evasive and intangible as emergent properties is hard. However, the mediators of emergent properties are the organisational principles of these molecular

machines and in practice it was these principles that we tried to elucidate with our approach.

Chapters 4, 5 and 6 model the basic organisation of PSD networks and assess if there are correlations of specific functions with modules of these networks. That allows us not only to functionally characterise protein complexes but also to tease out some of their basic organisational principles. Note that the modelling pipeline, although applied to the PSD molecular machine example, is not limited to that.

In Chapter 4, the initial attempt was to model a subset of proteins organised around PSD-95, which was used as a bait for the tandem affinity purification. This dataset not only augmented PSD data, but moreover the use of a different bait assisted with revealing slightly different subsets of the PSD (a more “lateral” membrane view). Further to the model revealing the organisation of these proteins around PSD-95 in a tightly connected protein interaction network of modular architecture, it also showed how very basic components of the toolkit of an excitatory synapse are closely interconnected (i.e. the simultaneous isolation of NMDA and AMPA glutamate receptors along with K^+ channels). Additionally, the association of some of these proteins with various types of mental disease not only highlights general observations regarding the importance of this central core of the PSD molecular machine in disease mechanisms, but also shows that these mechanisms could make a whole group of proteins more susceptible candidates (e.g. glutamate receptors and MAGUK/Dlg proteins of Cla and Cib).

An extension to the analysis of the PSD-95 associated proteins complex data was to combine it with other recent proteomics data (Collins et al., 2006, Husi and Grant, 2001, Husi et al., 2000), previously modelled by Pocklington et al. (2006), in order to create a more inclusive model of the PSD, was presented in Chapter 5. The resulting model did not only corroborate previous results but allowed the statistical and topological description of a protein interaction network with all the key components of the excitatory synapse present (including components not modelled previously, e.g. AMPA receptors and K^+ channels). Subsequent analysis also showed that the model

possesses a modular architecture. This, in combination with Chapter 4 and the results of Pocklington et al. (2006), shows that this modular architecture is persistent, i.e. present when modelling sub-networks of the PSD reconstructed from different datasets. The modular architecture is also biologically relevant, as highlighted by the correlation of molecular function and phenotype (plasticity, behaviour or disease) annotations to specific modules of the network. In the same manner that a car engine has parts that work together to perform specific sub-tasks (e.g. fuel injection, combustion, cooling etc), a molecular machine has functional modules performing sub-tasks. Crosstalk between these modules is vital not only in order to successfully fulfill the task in hand, but also to regulate it. Examples of this regulation can be found with the phosphorylation of the NMDA receptor (Coba et al., 2009; 2008). Furthermore, in the same chapter the dataset is analysed from the viewpoint of evolution, not only of its constituent parts, but also of their organisation in protein complexes.

After noticing how this modular architecture is persistent and has biological importance, we furthermore theorised that the PSDs of organisms with less intricate nervous systems and behavioural repertoires will also have some evidence of it. Chapter 6 presents our findings in a study of the fPSD. In this part of the work, we acquired the first catalogue and model of the fPSD based on affinity purification of selected in-vivo protein baits. Although the protein complex purification methods had their shortcomings and the dataset was not as optimised and noise-free as the mouse counterparts, a model was successfully reconstructed. Analysis of the annotations in this model showed presence of basic mouse PSD functionality. Although the fPSD appears to have a lower complexity in constituent parts, as previously predicted in the literature, the resulting network model appears to be organised in the same modular manner.

Application of the modelling pipeline in the aforementioned chapters allowed only for species specific conclusions to be drawn. The natural extension of this was the direct comparison of the PSD complexes. In Chapter 7 we compared the annotations found in the two datasets along with the network models and their architectural fea-

tures. These comparisons, beyond homologies in constituent parts and similarities in annotations, corroborated the comparative genomics prediction that the fPSD would be less complex in constituent parts. Most importantly however, by using the network models we showed how parts of the fPSD that have been conserved both structurally and architecturally, and revealed a conserved core component of the network. An important aspect of this evolutionary conservation is that it shows how a specific type of interaction configuration of receptors with cytoskeletal, adaptor and signalling proteins proved to be so useful as a general model of function and regulation that practically remained the same between species with very different cognitive repertoires.

8.1.2 Limitations

It is very important to consider that the results and contributions summarised above are the outcome of a process, based on a series of assumptions which have to be acknowledged. Furthermore, there are potential limitations of the methods used, which also have to be recognized. All of the former can at be least partially addressed with some future work. These assumptions and limitations can be grouped in the following categories:

Data assumptions and limitations: This category includes assumptions on the proteomics data. Such assumptions are that we are first of all modelling an “average” synapse, i.e. not a specific population of cells but tissue preparation from a specific brain region or whole brains. Differential expression of genes in different neuronal cell types has been shown several times and with different approaches including mRNA assays (Magdaleno et al., 2006, Doyle et al., 2008b) and protein assays (Emes et al., 2008). Also, electron microscopy studies of single synapses in rodents reveal that this gene expression diversity (for Grin2 subunits and MAGUK proteins) distinguishes individual synapses (Sans et al., 2000, Petralia et al., 2005). Another assumption is that of the presence of noise in data. Although this issue stretches more into MS data analysis, we have to bear in mind that contaminants are always present

in the dataset and that even the reconstruction of a protein interaction network does not guarantee their removal since some proteins might interact with members of the complex, but outside its PSD context. The reverse is also the case sometimes, when common contaminants, such as those found in “beadomes” (Rees et al., 2011) or that are common in cell lysates, like cytoskeletal proteins, might indeed have functionally important interactions within the PSD complexes.

There are also some less specific assumptions, applying to the fPSD. More specifically it is widely accepted that glutamate is a principal excitatory neurotransmitter in the mammalian CNS. Although the role of glutamate in fly synapses was established years ago (Jan and Jan, 1976, Chase and Kankel, 1987), the literature is mostly focused in the NMJ context (e.g. see Schuster et al., 1991). There is also evidence supporting the presence of glutamate receptors in the fly CNS (Ultsch et al., 1992, Parmentier et al., 1996, Völkner et al., 2000). Glutamate has been shown to have both excitatory and inhibitory actions in animals like molluscs (VYu et al., 1991) and *Aplysia sp* (Ke-hoe, 1994). What the previous research has not extensively covered is if glutamatergic transmission in *Drosophila* is as central as it is in mammalian CNSs. In turn the fPSD model reconstructed here appears somewhat “simpler” since the gene duplication events that made it evolve to the current complexity (Ryan and Grant, 2009) took place at the Deuterostome Bilaterian boundary. However, we must note that this apparent simplicity is based on a glutamate-centric model of PSDs and might or might not reflect the full fPSD functionality since it is likely that other neurotransmitter receptors and their associated protein complexes might account for additional functionality.

Another issue with PSD data in general is the identification of many presynaptic proteins in all datasets. Some of these might be normally expressed in the postsynaptic cell, while others might be affinity purified due to trans-synaptic complex formation or just because of interaction promiscuity. Identifying these categories can be challenging but until these annotations are available we also assume that the PSD models examined are not strictly “post”-synaptic.

There are also some technical issues to be acknowledged, as discussed in Chapter 6. These issues are centered around the protein complex isolation and purification method used to acquire the fly PSD data. More specifically, the method was based on whole brain extract rather than synaptosome extract. This was unavoidable since we were unable to find a standard synaptosome preparation method for insect neurons and were limited in time available for developing one. This resulted to a variety of contaminants in the dataset, some of which could be attributed to this fact (e.g. proteins abundant in the ommatidia). Additionally, with transmembrane proteins of the PSD like the NMDA receptor, being hard to isolate, again due to the time-frame of this project, we decided to trade-off a full optimisation towards these proteins with a more general protocol that would give a first catalogue of the fly PSD, rather than to strictly focus on MASC-type complexes.

Modelling and model interpretation assumptions and limitations: This category includes assumptions relevant to the reconstruction of the models using protein interaction data as well as the analysis and interpretation of these models using annotations and clustering methods. When it comes to protein interactions, admitting partial knowledge along with the presence of both false positive and negative data is common, is essential. Similarly crucial is a critical evaluation of their accuracy, biases, overlaps and complementarities (Mering et al., 2002). Similarly, gene and gene product annotation and its use in statistical inferences has to be based on the assumption that the dataset in hand is well annotated. This, however, does not always hold and along with issues such as the skewed annotations generated due to pleiotropic effects of protein functions in non-neuronal contexts (Inlow and Restifo, 2004), can cause issues. Another related issue is that fly synaptic proteins are not as well studied, thus annotated, as mouse synaptic proteins. The latter becomes even more evident when looking for the very scarce behavioural phenotype associations of fly proteins.

Another limitation of the modelling approach is that when using the Newman and Girvan algorithm we are unable to identify overlaps between clusters. In the light of

new data, this approach might be more useful in order not only to identify modules but also multiple simultaneous sub-complexes.

8.1.3 Critical examination

Given the contributions, the “bigger picture” and the acknowledged limitations presented in the previous paragraphs we can examine how far the initial proposed set of hypotheses was tested. From a proteomics data perspective (hypothesis 1), it is evident that new affinity purifications not only augment PSD in the form of lists of constituent parts. Namely, beyond the new proteins that are isolated and the higher confidence in the presence of certain proteins that tend to re-appear often, the use of different baits can better isolate different sub-complexes of the PSD. From a network organisation perspective (hypothesis 2) we were able to provide evidence of a persistent modular architecture with biological interpretation or significance, although the confidence varied from case to case due to data limitations (e.g. noise or different contamination levels in datasets, unavailable or false annotations and availability of protein interactions). The biological significance of this modular architecture, although impossible to precisely quantify, can be supported by the correlation of specific annotations with specific modules. The latter is also supported by a common concept in the literature, where proteins that interact are expected to have common or semantically more similar annotations (example references Mahdavi and Lin, 2007, Lord et al., 2003). Regarding the fPSD (hypothesis 3), we were able to obtain a list of the constituent parts based on affinity purifications and reconstruct the first model using the modelling pipeline. Although the method suffered from technical drawbacks, we were able to produce an initial map of the fPSD, showing how it has similar functionality and modular architecture to that of the mPSD. Given these drawbacks and the additional issues with the lack of protein interaction data and annotations, it can be said that this was the least supported of our hypotheses. Finally, regarding the comparison of PSD complexes (hypothesis 4), we presented a set of approaches and adapted method to quantify sim-

ilarities and differences and showed how components of the network were conserved. Again, in this case our approach was not as fruitful as expected since it suffered from the issues of the fPSD model. However, we believe that given the quality of the model, due to contamination and missing interaction, as well as annotation data, the quality of the resulting evidence is satisfactory.

8.2 Future work

Some of the drawbacks and limitations in this work are technically more addressable than others, since there are clear steps to be taken in order to improve them. For example the isolation and purification of protein complexes of the fPSD, can be further improved in two steps. The first step would be the optimisation of the purification protocols towards transmembrane proteins. This could be achieved by testing different lysis buffer compositions with varying concentrations of different detergents (Triton X-100, DOC, ComplexioLytes, digitonin). By targeting more towards proteins of the membrane fraction, we could potentially isolate more receptors and channels relevant to the fPSD. The second step would include the use of different proteins as baits. Potential bait candidates could include synthetic peptides of interaction domains of other insect neurotransmitter receptors, potentially also central to the fPSD. Other bait candidates could include prey proteins that were found to be central components in this work and also known interactors of current baits. Since a first draft of, the fly PSD was obtained here, it would also be interesting to attempt two step purifications using TAP, in order to reduce noise from contaminants. Another alternative to that would be to use parallel affinity capture of complexes utilising two tags of the CPTI lines of interest. Additionally, and in retrospect we think that it would have been an interesting endeavor to attempt purification using the same bait but different protocols, one more optimised towards proteins localised in or near the cell membrane and one for soluble proteins. This way one can generate affinity purifications under conditions suitable for a wider

range of proteins. A similar approach to this would be optimising a more inclusive set of bait-prey interactions by trying antibodies that recognise different epitopes of the bait proteins.

Other directions of the future work are relevant to models of PSD complexes in general, and would probably be initially applicable to the more well studied and annotated mouse PSD. The constituent parts of the static models described in this work did not include two important annotation components, namely of differential expression by brain region and of protein abundance. Although this information was partially available at the time, the coverage of the datasets was small. More specifically regarding differential expression in brain areas, a proposed approach would be measuring transcription data from BACTRAP (Doyle et al., 2008b) and array (e.g. Magdaleno et al., 2006) approaches. Although some comparative analysis suggests correlation between transcriptional and translational expression for the majority of genes (Mijalski et al., 2005), the subject of general transcriptome - proteome correlation is still debatable. Efforts to analyze noisy data with simple correlation metrics have resulted in weak positive correlations (Lian et al., 2001, Griffin et al., 2002, Roch et al., 2004), while on the other hand an analysis with more robust statistic framework have yielded stronger correlations (Gygi et al., 1999). Being able to identify differences in the composition of the PSD will allow us to better understand not only its evolution as neuron cell types diversified, but also if this differential composition has functional significance. Regarding protein abundance there have been various quantitative solutions proposed ranging from relative abundance within the dataset, to quantitative methods involving metabolic (e.g. SILAC - Ong et al., 2002), chemical labeling (e.g. iTRAQ - Ross et al., 2004), or label free methods such as spectral counting (Liu et al., 2004, Vogel and Marcotte, 2008). For a review of methods see Ong and Mann, 2005, Bantscheff et al., 2007. Abundance information is not only important in static models such as the ones described here, but it is also the first step towards more detailed models that take features like complex stoichiometry into account.

Incorporating the aforementioned parameters to PSD models would mean richer descriptive models which would still, nevertheless, be static models describing what is fundamentally an adaptive and highly dynamic structure. For that reason another major direction would be that of dynamic models. Recently Sorokina et al. (2011) presented a solution towards quantitative modelling of the PSD, using a core of molecules from the PSD model described in Chapter 5, illustrating the possibility of extending a qualitative protein-protein interaction map into a quantitative executable model and capturing the dynamic complexity expected to be found in the PSD. Dynamic “executable” models will require even more parameters (e.g. binding site affinities, quantitative proteomics etc) and so extending from reduced pathway models of a few molecules to the scale of models described here (>100s of molecules) will take a significant investment in both modelling and biochemical characterisation. However, this problem could be partially addressed by computational inference.

Part of this dynamic complexity is also the presence of one protein in multiple simultaneous complexes, the identification of which could show us a different type of modularity, this of protein organising or taking part in multiple modules. Identification of overlapping complexes could be achieved by using appropriate clustering algorithms that allow that, for example MCODE(Bader and Hogue, 2003), OMIM(Wang et al., 2012), or CPM(Bu et al., 2003). However, identifying these overlapping modules makes the concept of modularity multidimensional, which will result to computation of statistical correlations being far more complicated. Also, due to technical restrictions and lack reference data availability there have been few attempts to compare the performance of these algorithms. In the comparisons we are aware of, MCL shows good performance and robustness to noise (Brohee and van Helden, 2006, Vlasblom and Wodak, 2009).

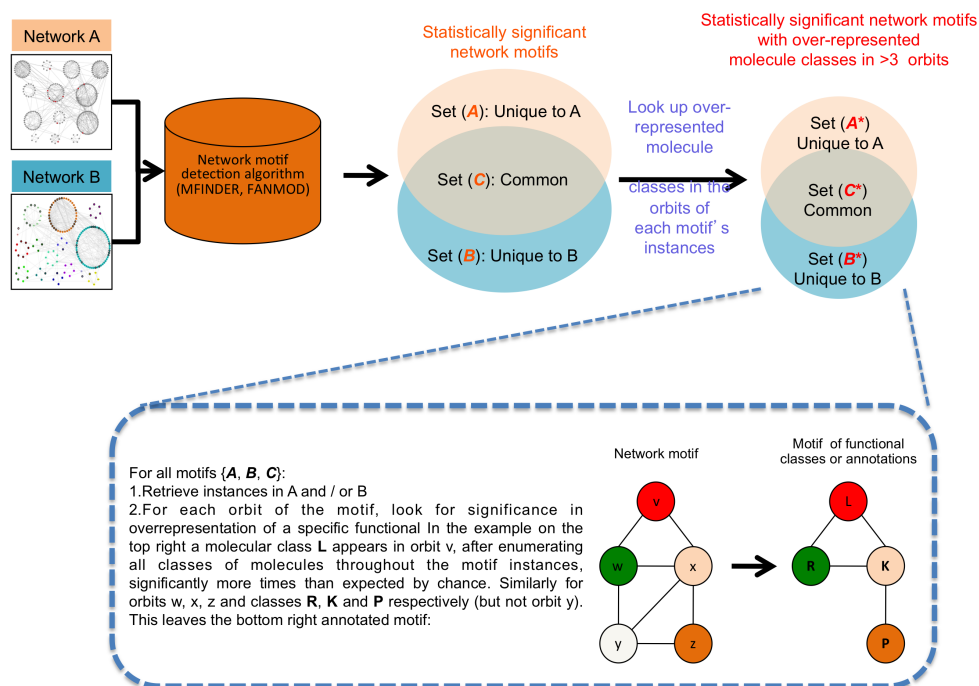
Yet another direction of dynamic models will be the challenge of the PSD phosphoproteome. This issue was addressed by Coba et al. (Coba et al., 2008; 2009) who collected phosphorylation data on a set of PSD kinases and their known substrates,

after range of NMDA receptor stimulation protocols. While the range of the stimulations protocols had sufficient coverage, the range of kinases and substrates did not provide full PSD coverage. Nevertheless, the resulting model was one of impressive complexity, possibly explaining some of the combinatorial computational power of synapses.

The other aspect of future work would be that of completing the comparative PSD model analysis started here after acquiring more data to complete the fly PSD model. Besides the comparison of parts lists and distribution of functionalities or annotations, comparative interactomics approaches could be used in order to compare how these parts lists are organised in networks. There are various solutions proposed recently in order to compare network models using computational methods, including Network-BLAST (Sharan et al., 2005), also used here, graphlet degree distribution (Przulj, 2007, Kuchaiev and Przulj, 2011) and other graph based methods, e.g. (Klau, 2009). These methods allow the identification of common conserved cores or topological features of the networks based on protein homology and/or interaction motifs, usually heavily depending on the former, with the exception of graphlets. If the case is that there is more overlap to be found between fly and mouse PSDs, these methods will indicate how this overlap constitutes a conserved signalling network, extending work in Chapter 7. However, these methods rely heavily on either sequence and/or network architecture similarity¹. The author has developed a comparative method (summarised in Figure 8.1), which is based on network motifs (Milo et al., 2002, Shen-Orr et al., 2002, Alon, 2007) common in the two networks under comparison. Network motifs are statistically significant recurring pattern of connections between nodes in a network. These motifs, if present, represent similar organisation patterns between the two networks. Motifs also have a local organisation (a specific protein in every orbit, i.e. node position with respect to symmetries). In the proposed approach common motifs are detected and tested for statistical correlations of specific orbits with a functional

¹Both found to be problematic at times.

Figure 8.1: A proposed workflow for finding conserved motifs of functional family interactions.



family. In turn, these correlations in combination with the motifs represent conserved interaction topologies of specific functions. The method is still under development and the author was not able to present a detailed application to these datasets due to time limitations.

Furthermore, another interesting aspect is the evolution of this network which could be modelled with application of the Pastor-Satorras et al. (2003) model. This model is based on gene duplication events and the respective protein binary interaction duplication, taking place with a probability a . This modelling approach is very appealing in the PSD case since gene family expansion via gene duplication is very frequent in central proteins (e.g. DLG family, 2nd unit of the NMDA receptor, etc).

8.3 Conclusions

Systems Biology (and Systems Neuroscience) analysis of the brain aims to provide a framework upon which we can understand the brain at all levels of its complexity from bio-molecular events at synapses through complex networks of neurons, brain

regions and systems and ultimately to individual and social behaviour. Here we have focussed on strategies and methods that help us capture, explore and understand molecular complexes identified from primary biochemical analysis of neural tissues. Protein interaction network models provide a powerful scaffold for knowledge integration and hypothesis generation. By combining data annotation, analysis and protein interaction network reconstruction we model and investigate proteomics datasets. Utilising the resulting models as integrative descriptive tools offers an overview of the major constituent parts of molecular machines and also gives insight on how these parts are combined to give rise to the properties of a complex system such as the receptor signalling complexes embedded in synapse proteomes.

Closer examination of these models can also be used to explore the validity of disease hypotheses. In the case of the PSD for example, using models like the ones described here one can see that the primary interactors of genes associated with disease, e.g. schizophrenia, are spread throughout many modules within the networks. Also, GWAS of common SNPs showed that conditions such as schizophrenia and bipolar disorder are highly polygenic, with potentially thousands of SNPs of small effect contributing to susceptibility (Purcell et al., 2009). However, the SNPs that have so far reached genome-wide levels of significance have not yet converged on a clear set of disease-relevant processes. This suggests that the overall network and its various clusters might play a role in these diseases, and in schizophrenia for example while enriched, the glutamate receptors may not be the entire story, as per the “glutamate hypothesis” of schizophrenia (Greene, 2001, Coyle, 2006, Lisman et al., 2008). Similarly, we can also start to query common mechanisms that might be shared across multiple diseases. For example, the reconstructed PSD models uncover 43 proteins linked by various lines of evidence to schizophrenia, of which 20 are also implicated in other diseases (bipolar disorder, depression mental retardation). Also, genome-wide studies of copy number variants (CNVs), in which extended genomic sequences are duplicated or deleted, have discovered that large, rare CNVs contribute to both autism

and schizophrenia (Redon et al., 2006, Walsh et al., 2008, Stone et al., 2008). This is supporting evidence of a rising trend of the belief that diseases like schizophrenia and bipolar disorder, which have been described in psychiatry as a spectrum of disorders with heterogeneous presentation, have genetic mechanisms behind them can lead to a sub-classification of these diseases into subtypes. But it is not only the genetics but also the organisation of affected proteins in functional complexes. Recent data (Frank et al., 2011) have shown also shown schizophrenia and bipolar disorder diseased individuals possessed an increased load of deleteriousness from multiple concurrent rare and common coding variants. This, by first observation, raises the question of the missing heritability (Maher, 2008, Gunter, 2009). However, analysis of these models along with other analysis of GWAS data (Maher, 2008, Gershon et al., 2011) suggests that it is the multiple rare variants combined via an epistatic interplay can be causing the disease or varying the symptoms. So in the schizophrenia and bipolar disorder cases the interplay of compound genetic coding variants, distributed among glutamate receptors and their interacting proteins, could contribute to the pathogenesis and phenotype of the diseases. As confidence in the data underpinning these models increases with time, these methods will start to deliver on their potential. However, as mentioned throughout, one must always bear in mind the significant limitations of these models when making any decisions based upon them.

In conclusion, after considering potential flaws of the approach described in this work, namely the limitations of the methods for obtaining proteomics and interaction data as well as the static nature of it. Nevertheless, these static models managed to efficiently describe a complex molecular machine such as the PSD and give insight to some of the underlying principles of its organisation. We are acutely aware that in the longer term we need to look to more dynamic modelling approaches that cover the computational complexity of synaptic molecular machines, however, we also see these approaches stemming and evolving out of methods and models similar to the ones described (and referenced) in this work and feel that this is part of our main

contribution.

Appendix A

Supplemental Methods

A.1 fPSD complexes

A.1.1 Validation of affinity purifications















Figure A.1 shows the western blot for the ECP and CCP datasets. Note that the migration lengths of the bands correspond to the molecular weight of known isoforms with the added weight of the tag. Blots were probed with the JL-8 antibody (Clonetechn) against the YFP tag. The CCP blot was slightly under-exposed, but the bands are still visible. The bands were verified based on molecular weight (MW). More specifically $MW_{target} = MW_{isoform} + MW_{tag}$, where $MW_{tag} \approx 30KDa$. Also, in the second replicate of *dlg1* in the CCP data the band was very faint. Given that the film was underexposed, we decided to proceed with the mass spectrometry analysis of the sample.

A.1.2 Mass spectrometry data filtering

A.1.2.1 Filtering criteria

Since we had to deal with two substantially different raw data analysis workflows and sets of exported results from the two different facilities, the following paragraphs will explain the data filtering and integration process for the CCP and ECP datasets.

Figure A.1: Experimental validation of affinity purification samples. Each well was loaded with 18 μ l of sample with loading dye. Note that the molecular weights of the bands are similar to weights of known isoforms plus ~30KDa, which is the weight of the expressed tag construct. Red arrows show the weight of the closest molecular weight marker.

Bait	MW including CPTI insert (Kda)	ECP data		CCP data	
		Replicate 1	Replicate 2	Replicate 1 (Strep, YFP)	Replicate 2 (Strep, YFP)
dlg1 (CPTI-0000207)	~130				
Bsg (CPTI-0000062)	~100				
14-3-3 ϵ (CPTI-0000034)	~60				
tau (CPTI-0000194)	~70	N/A	N/A		

Due to the intricacies of the different mass spectrometry workflows the datasets had to be filtered using different combinations of the following criteria:

1. Peptide count (*pepcount*): the number of distinct peptides that belong to a specific protein sequence.
2. % sequence coverage (*%cov*): per cent coverage of a protein sequence by the distinct peptides found
3. MASCOT score (*M_score*): MASCOT score is computed by the Mascot software and is based on false discovery from a “negative control” database. Although the specifics are beyond the scope of this chapter, the definition is $M_score = -10 * \log_{10}(P)$, where P is the false positive match probability. To illustrate this we can use the following example: in a database of $5 * 10^5$ entries, a 0.001 chance of getting a false positive match is a probability of $P = 1 / (10^3 * 5 * 10^5)$, which is equivalent to a Mascot score $M_score = 87$.
4. Peptide enrichment ratio (R_{pep}): given the data filtering strategy of choice, if a protein is found both in the control and a affinity purification sample, it can either considered contamination of the affinity purification sample and discarded or not. In the second case we can apply a dynamic filtering approach where the peptide enrichment ratio is the criterion. If for every protein X found both in control and an affinity purification sample we define:

$$R_{pep} = \frac{\frac{n_{sample}(X)}{N_{sample}}}{\frac{n_{control}(X)}{N_{control}}}$$

Where $n(X)$ is the peptide count of X and N is the total peptide count in the sample and control respectively. This way we can control for enrichment in the relative proportion of protein specific peptides in a pool of peptides. Note that in case protein X is absent from the control $R_{pep} = +\infty$. Note that in order to use this model of dynamic cut-offs it makes more sense to adjust the peptide count a

low value (1 peptide with a MASCOT score above 60 is considered acceptable by the Edinburgh facility).

5. Number of replicates a protein appears in. Since each affinity purification was performed in duplicate, one approach is to allow in the dataset proteins appearing in one or more replicates another more stringent approach is to allow in proteins appearing in both replicates. In our case this caused an issue for one case (dlg1 in the CCP data) where one of the replicates failed to purify any baits probably due to unsuccessful purification.

On suggestion from each of the two facilities, we decided to use the criteria they used in-house for the filtering of the datasets generated in each facility. For the CCP data criteria 1,2 and 5 were used and for ECP data criteria 1, 3, 4 and 5.

A.1.2.2 ECP data filtering

The minimum number of peptides (criterion 1) was set to 2, with 1 being the default value the Edinburgh facility uses. This in combination with a MASCOT score cut-off of 80 (60 being the facility's default) guarantees confidence in the identification. After experimenting with the R_{pep} ratio we observed that any value above 4 gives similar number of proteins in the resulting dataset so we chose a cut-off of 6. We had to manually add the 14-3-3 ϵ and Bsg baits in their respective affinity purification. Although the presence of all baits was verified with western blotting, 14-3-3 ϵ was under the 2 peptide cut-off and Bsg peptides were probably masked. Regarding Bsg specifically - it is not the first time that the mass spectrometer of the Edinburgh facility has been unable to identify its presence in samples where it has been otherwise verified. The above are also summarised in table A.1.

Table A.1: Cut-off criteria for the ECP dataset filtering.

Dataset label	Number of Peptides	Number of replicates	MASCOT score	R _{pep}	Keep only proteins known from previous pulldowns
<i>tolerant</i>	2	≥ 1	80	6	no
<i>intermediate</i>	2	≥ 1	80	6	yes
<i>strict</i>	2	> 1	80	6	no

Table A.2: Cut-off criteria for the CCP dataset filtering.

Dataset label	Number of Peptides	Number of replicates	Coverage	Keep only proteins known from previous pulldowns
<i>tolerant</i>	3	≥ 1	15%	no
<i>intermediate</i>	3	≥ 1	15%	yes
<i>strict</i>	3	> 1	15%	no

A.1.2.3 CCP data filtering

The minimum number of peptides (criterion 1) was set to 3, with 2 being the default value the Cambridge facility uses. The minimum coverage (criterion 2) was set to 15% of the protein sequence, with 10%-12% being the standard values the facility uses. The minimum number of replicates a protein should appear in was set to ≥ 1 (one or more) for the *tolerant* and *intermediate* labelled datasets. For the *strict* labeled dataset we chose exactly the same parameters but only allowed in proteins appearing in one or more replicate. Although all baits were identified by mass spectrometry, we had to manually add the tau and 14-3-3 ϵ baits in the *intermediate* labelled dataset because they were filtered out as previously unknown (these proteins have not been isolated as preys in our in-house small scale experiments). The above are also summarised in table A.2. Note: in the case of dlg1 in the CCP data, one replicate returned a very short list which we considered as failed and for this reason the dlg1 data were excepted from the number of replicates criterion.

A.1.2.4 Issues with the data

Although affinity purifications experiments were always performed by the author using the same types of reagents, with all different baits and replicates performed in parallel with one exception (see below), in an attempt to minimize the effect of systematic error on the purification level, we noticed two main issues with the data after the mass spectrometry results were returned. These were:

- **Difference in numbers of identified proteins (for the same bait):** It has been noted before that when using these two mass spectrometry facilities, the Edinburgh Centre for Proteomics always returns larger lists of identified proteins. Although this is an empirical observation it has occurred many times and we believe it has to do with the workflow of processing raw data in the Edinburgh and Cambridge facilities. One of the major effects of this is the difference in dataset sizes and subsequently the contribution of each affinity purification identification in the final dataset. Nevertheless, we manually checked the overlap between the datasets and found their overlaps statistically significant ($p < 0.05$) using a Fischer's exact test and given the dataset sizes.
- **Low reproducibility between replicates:** an overall low reproducibility between replicates within the same experiment was also noticed. In the case of the CCP datasets this was observed with the 14-3-3 ϵ and Bsg purifications. In the ECP datasets it was the case with all affinity purifications. A possible explanation for that is that one of the two replicate samples was destroyed by the facility and had to be repeated as an independent purification.

In order to control for this issue we compared the *tolerant* dataset with a list of proteins known from previous independent proteomics experiments. These "previously known" proteins come from an in-house collection of unpublished data from smaller scale experiments (using a Nmdar2 N-terminal bait and a Bsg affinity purification). We filtered out all proteins from the tolerant datasets if

they were not in this list and generated a new dataset. Using these versions of the *tolerant* datasets, labelled *intermediate*, we observed that while there is a considerably low reproducibility between replicates the both the CCP and ECP datasets isolate previously known proteins. More specifically ~50% and ~40% of the proteins identified in the CCP and ECP datasets respectively were in the “previously known” list. Taking into account that the only common previous bait was Bsg these numbers can be considered satisfactory.

Appendix B

Data tables

Table B.1: List of proteins per cluster in the PSD-95 associated proteins network.

Cluster	Proteins
Cla	Begain, Camk2a, Camk2b, Cltc, Cypin, Dlg1, Dlg2, Dlg3, Dlg4, Dlgap1, Dlgap2, Dlgap3, Dlgap4, Grin1, Grin2a, Grin2b, Grin2d, Kcnj10, Kcnj4, Nefl, Pppp3ca, Spnb2, Syngap1
Clb	Cacng2, Gria1, Gria2, Gria3, Gria4, Grik2, Grik5
Clc	Kcna1, Kcna2, Kcna3, Kcna4, Kcnab1, Kcnab2
Cld	Baiap2, Rac1
Cle	Adam22, Lgi1
Clf	Fscn1, Gapdh, Pkg1, Vdac1
Clg	Atp5a1, Atp5b, Atp5c1, Atp5o, Slc25a4, Slc25a5

Table B.2: Significant cluster and “cellular component” gene ontology (GO) terms correlations in the PSD-95 associated proteins network.. P-values in parentheses.

Cluster	Cellular Component (GO)
Cla	N-methyl-D-aspartate selective glutamate receptor complex (0.04), integral to membrane (0.01), synapse (0.04)
Clb	alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid selective glutamate receptor complex (0.01), cell junction (0), dendrite (0), integral to membrane (0), membrane (0.01), membrane fraction (0.01), neuronal cell body (0), outer membrane-bounded periplasmic space (8.53E-05), perikaryon (0), plasma membrane (0.04), postsynaptic density (0.05), postsynaptic membrane (0), presynaptic membrane (0.03), synapse (0.01), terminal button (9.91E-06)
Clc	integral to membrane (0), juxtaparanode region of axon (0), voltage-gated potassium channel complex (6.51E-05)
Clg	mitochondrial inner membrane (4.41E-07), mitochondrial proton-transporting ATP synthase complex (0), mitochondrion (1.32E-05), proton-transporting ATP synthase complex, catalytic core F(1) (6.51E-05)

Table B.3: Significant cluster and “molecular function” gene ontology (GO) terms correlations in the PSD-95 associated proteins network.. P-values in parentheses.

Cluster	Molecular Function (GO)
Cl _a	N-methyl-D-aspartate selective glutamate receptor activity (0.04), calmodulin binding (0.04), cation channel activity (0.04), protein domain specific binding (0.01)
Cl _b	PDZ domain binding (0.02), extracellular-glutamate-gated ion channel activity (8.53E-05), ion channel activity (0), ionotropic glutamate receptor activity (8.53E-05), receptor activity (0), transporter activity (0)
Cl _c	ion channel activity (0.02), voltage-gated ion channel activity (5.29E-06), voltage-gated potassium channel activity (6.29E-08)
Cl _g	ATPase activity (0), hydrogen ion transporting ATP synthase activity, rotational mechanism (6.51E-05), proton-transporting ATPase activity, rotational mechanism (0)

Table B.4: Significant cluster and “biological process” gene ontology (GO) terms correlations in the PSD-95 associated proteins network.. P-values in parentheses.

Cluster	Biological Process (GO)
Cl _a	calcium ion transport (0.04), cell-cell signaling (0.04), ion transport (0), regulation of neuronal synaptic plasticity (0.04), startle response (0.04)
Cl _b	ion transport (0), regulation of membrane potential (0.03), transport (0)
Cl _c	ion transport (0), potassium ion transport (1.76E-06), transmembrane transport (5.29E-06)
Cl _g	ATP synthesis coupled proton transport (6.51E-05), proton transport (6.51E-05)

Table B.5: Significant cluster and PANTHER protein classes correlations in the PSD-95 associated proteins network.. P-values in parentheses.

Cluster	Protein Class
Cl _a	transmembrane receptor regulatory/adaptor protein (0)
Cl _b	ionotropic glutamate receptor (8.53E-05), ligand-gated ion channel (0)
Cl _c	potassium channel (1.76E-06), voltage-gated ion channel (1.32E-05)
Cl _g	ATP synthase (6.51E-05), hydrolase (0), ligand-gated ion channel (0.04)

Table B.6: Significant cluster and PANTHER pathways correlations in the PSD-95 associated proteins network. P-values in parentheses.

Cluster	Pathways
Cl _a	Huntington disease->N-methyl-d-aspartate receptor (0.04), Ionotropic glutamate receptor pathway->N-methyl-D-aspartate Receptor (0.04), Metabotropic glutamate receptor group I pathway->N-methyl-D-aspartate Receptor (0.04), Metabotropic glutamate receptor group III pathway->N-methyl-D-aspartate Receptor (0.04), Muscarinic acetylcholine receptor 1 and 3 signaling pathway->N-methyl-D-aspartate Receptor (0.04)
Cl _b	Ionotropic glutamate receptor pathway->AMPA Receptor (0), Ionotropic glutamate receptor pathway->AMPA/Kainate Receptor (4.41E-07), Ionotropic glutamate receptor pathway->Glutamate receptor, ionotropic, AMPA 1 (0), Ionotropic glutamate receptor pathway->Glutamate receptor, ionotropic, AMPA 2 (0), Ionotropic glutamate receptor pathway->Glutamate receptor, ionotropic, AMPA 3 (0), Ionotropic glutamate receptor pathway->Glutamate receptor, ionotropic, AMPA 4 (0), Metabotropic glutamate receptor group III pathway->AMPA/Kainate Receptor (4.41E-07)

Table B.7: List of proteins per cluster in the Union network. Continues in Table B.8.

Cluster	Number of proteins	Families in cluster	MGI gene symbol
A	39	Adaptor/ Regulatory, Cytoskeletal/ Structural/ Cell adhesion, G-protein signaling, Kinases, Receptors/ Channels/ Transporters, Signalling molecules and Enzymes	Actn2, Adam22, Atp2b4, Baiap2, Begain, Camk2a, Camk2b, Dlg1, Dlg2, Dlg3, Dlg4, Dlgap1, Dlgap2, Dlgap3, Gda, Grik2, Grik5, Grin, Grin2a, Grin2b, Grin2d, Kcna1, Kcna2, Kcna3, Kcna4, Kcnab1, Kcnab2, Kcnj10, Kcnj4, Lgi1, Lin7a, Mpp2, Mpp3, Nos1, Ptk2b, Shank1, Shank2, Spnb2, Syngap1
B	6	Receptors/ Channels/ Transporters, Vesicular/ Trafficking/ Transport	Cacng2, Gria1, Gria2, Gria3, Gria4, Nsf
C	12	Adaptor/ Regulatory, Cytoskeletal/ Structural/ Cell adhesion, G-protein signaling, Receptors/ Channels/ Transporters, Signalling molecules and Enzymes	Calm2, Flna, Gap43, Gapdh, Grm1, Grm5, Homer1, Pla2g4a, Rab2a, Rab3a, Rala, Spnb3

Table B.8: List of proteins per cluster in the Union network (continued from Table B.7). Continues in Table B.9.

Cluster	Number of proteins	Families in cluster	MGI gene symbol
D	56	Adaptor/ Regulatory, Cytoskeletal/ Structural/ Cell adhesion, G-protein signaling, Kinases, Phosphatases, Receptors/ Channels/ Transporters, Signalling molecules and Enzymes, Vesicular/ Trafficking/ Transport	Ablim1, Actg1, Akap9, Anks1, Bad, Cit, Cltc, Ctnn, Dlgap4, Dnm1, Dusp4, Fgd4, Gnb2l1, Grb2, Hras1, Ina, Irs1, Klc2, Lmnb1, Map2k1, Map2k2, Map2k3, Mapk1, Mapk3, Mtap2, Myh10, Myh9, Myo5a, Nefl, Nf1, Pdpk1, Pgam5, Pik3ca, Plcg1, Ppp1cc, Ppp2ca, Ppp2r1a, Ppp5c, Prkcb, Prkcc, Prkce, Ptpn11, Ptpn5, Raf1, Rap2a, Rps6ka3, Slc25a4, Src, Tjp1, Trp53bp1, Tuba1b, Vegfa, Ywhae, Ywhag, Ywhah, Ywhaz
E	5	Adaptor/ Regulatory, Kinases, Phosphatases, Receptors/ Channels/ Transporters	Akap5, Atp1a1, Ppp3ca, Prkacb, Prkar2b
F	9	Adaptor/ Regulatory, Cytoskeletal/ Structural/ Cell adhesion, G-protein signaling, Kinases, Signalling molecules and Enzymes	Akt2, App11, Cdh2, Ctnnb1, Fus, Gsk3b, Plcb1, Rac1, Slc9a3r1

Table B.9: List of proteins per cluster in the Union network (continued from Table B.8).

Cluster	Number of proteins	Families in cluster	MGI gene symbol
G	8	Cytoskeletal/ Structural/ Cell adhesion, G-protein signaling, Kinases, Signalling molecules and Enzymes	Cfl1, Cse11, Pfk1, Pkg1, Pklr, Pkm2, Ran, Tpi1
H	6	Cytoskeletal/ Structural/ Cell adhesion, Vesicular/ Trafficking/ Transport	Nefm, Nrnx1, Snap25, Stx1a, Stxbp1, Syt1
I	5	Adaptor/ Regulatory, Cytoskeletal/ Structural/ Cell adhesion, Kinases	Dbn1, Map2k7, Mapk10, Mapk8ip1, Stk39
J	4	G-protein signaling	Gnao1, Gnb1, Gnb2, Gnb4
K	2	G-protein signaling	Sept11, Sept5
L	4	Enzymes	Atp5a1, Atp5b, Atp5c1, Atp5o
M	4	Cytoskeletal/ Structural/ Cell adhesion, Receptors/ Channels/ Transporters	Actn4, Fscn1, Gsn, Vdac1
N	2	Cytoskeletal/ Structural/ Cell adhesion	Capza2, Capzb
O	2	Cytoskeletal/ Structural/ Cell adhesion	Mbp, Plp1

Table B.10: Significant cluster and molecular function gene ontology (GO) terms correlations in the Union network. P-values in parentheses.

Cluster	Molecular Function (GO)
A	ATP binding (0.05), N-methyl-D-aspartate selective glutamate receptor activity (0), PDZ domain binding (0), cation channel activity (0), extracellular-glutamate-gated ion channel activity (0.01), ion channel activity (0), ionotropic glutamate receptor activity (0.01), nucleotide binding (0), receptor activity (0.05), transporter activity (0.04), voltage-gated ion channel activity (4.18E-005), voltage-gated potassium channel activity (0)
B	extracellular-glutamate-gated ion channel activity (0), ion channel activity (0), ionotropic glutamate receptor activity (0), receptor activity (0), transporter activity (0)
D	ATP binding (0.01), insulin receptor binding (0.01), motor activity (0.01), phosphotyrosine binding (0.01), protein domain specific binding (0.01), protein kinase activity (0.01), protein serine/threonine kinase activity (0.01), protein tyrosine kinase activity (0.05)
G	transferase activity (0.02)
J	signal transducer activity (7.23E-006)
L	hydrogen ion transporting ATP synthase activity (3.44E-008)

Table B.11: Significant cluster and biological process gene ontology (GO) terms correlations in the Union network. P-values in parentheses.

Cluster	Biological Process (GO)
A	calcium ion transport (0.01), ion transport (0), potassium ion transport (0), regulation of excitatory postsynaptic membrane potential (0), regulation of long-term neuronal synaptic plasticity (0.02), regulation of membrane potential (0), regulation of neuronal synaptic plasticity (0.01), startle response (0), synaptic transmission (0), synaptic transmission (0.05), transmembrane transport (0), transport (0.04)
B	ion transport (0), transport (0)
C	signal transduction (0.04)
D	MAPKKK cascade (0.02), intracellular signaling cascade (0), protein amino acid dephosphorylation (0.01), protein amino acid phosphorylation (0.03), regulation of cell shape (0.01)
G	glycolysis (3.56E-007)
H	neurotransmitter secretion (2.56E-006)
J	G-protein coupled receptor protein signaling pathway (5.16E-007), signal transduction (0)
L	ATP synthesis coupled proton transport (3.44E-008), ion transport (0), proton transport (3.44E-008)

Table B.12: Significant cluster and cellular component gene ontology (GO) terms correlations in the Union network. P-values in parentheses.

Cluster	Cellular Component (GO)
A	N-methyl-D-aspartate selective glutamate receptor complex (0), cell junction (2.31E-005), cytoplasm (0.01), integral to membrane (0), integral to plasma membrane (0), membrane (0.01), outer membrane-bounded periplasmic space (0.01), postsynaptic density (0), postsynaptic membrane (2.02E-006), presynaptic membrane (0.04), synapse (7.90E-006), synaptosome (0.01), voltage-gated potassium channel complex (0.01)
B	cell junction (0.01), integral to membrane (0), membrane (0.01), outer membrane-bounded periplasmic space (0), postsynaptic density (0), postsynaptic membrane (0), synapse (0.01)
D	Golgi apparatus (0.02), cell cortex (0.05), cytoplasm (0.03), membrane (0.03), microtubule (0), mitochondrial outer membrane (0.05), plasma membrane (0.03)
F	cytosol (0), lamellipodium (0), nucleus (0.04)
G	cytosol (0.04), nucleus (0.02)
I	cytoplasm (0.03)
L	mitochondrial inner membrane (1.20E-006), mitochondrion (0), proton-transporting ATP synthase complex (3.44E-008)

Table B.13: Significant cluster and family correlations. P-values in parentheses.

Cluster	Families
A	Adaptor/ Regulatory (0.01), Receptors/ Channels/ Transporters (2.31E-005)
B	Receptors/ Channels/ Transporters (0)
C	G-protein signaling (0.03)
D	Phosphatases (0)
G	Kinases (0.02)
H	Vesicular/ Trafficking/ Transport (0)
J	G-protein signaling (0)
L	Enzymes (3.44E-008)

Table B.14: Significant cluster and PANTHER pathways correlations. P-values in parentheses.

Cluster	Pathways
A	Huntington disease->N-methyl-d-aspartate receptor (0), Iontropic glutamate receptor pathway->N-methyl-D-aspartate Receptor (0), Metabotropic glutamate receptor group I pathway->N-methyl-D-aspartate Receptor (0), Metabotropic glutamate receptor group III pathway->N-methyl-D-aspartate Receptor (0), Muscarinic acetylcholine receptor 1 and 3 signaling pathway->N-methyl-D-aspartate Receptor (0)
B	Iontropic glutamate receptor pathway->AMPA Receptor (5.16E-007), Iontropic glutamate receptor pathway->AMPA/Kainate Receptor (7.59E-006), Iontropic glutamate receptor pathway->glutamate receptor, ionotropic, AMPA 1 (5.16E-007), Iontropic glutamate receptor pathway->glutamate receptor, ionotropic, AMPA 2 (5.16E-007), Iontropic glutamate receptor pathway->glutamate receptor, ionotropic, AMPA 3 (5.16E-007), Iontropic glutamate receptor pathway->glutamate receptor, ionotropic, AMPA 4 (5.16E-007), Metabotropic glutamate receptor group III pathway->AMPA/Kainate Receptor (7.59E-006)
D	EGF receptor signaling pathway->14-3-3 (0.01), FGF signaling pathway->14-3-3 (0.01), PI3 kinase pathway->14-3-3 (0.01), Parkinson disease->14-3-3 (0.01), p53 pathway->14-3-3 (0.01), p53 pathway->14-3-3 sigma (0.01)
J	Metabotropic glutamate receptor group II pathway->G-protein (3.44E-008), Muscarinic acetylcholine receptor 2 and 4 signaling pathway->Gi (3.44E-008)

Table B.15: Top twenty nodes of highest betweenness in the Union network.

Mgi gene symbol	Betweenness	Average Shortest Path	Cluster
Dlg4	0.19	2.25	A
Grin1	0.17	2.28	A
Calm2	0.15	2.47	C
Ywhag	0.14	2.61	D
Src	0.08	2.36	D
Actg1	0.08	2.71	D
Grb2	0.06	2.71	D
Gapdh	0.06	3.04	C
Grin2d	0.05	2.47	A
Prkce	0.05	2.54	D
Dlg1	0.05	2.6	A
Raf1	0.05	2.77	D
Prkcb	0.04	2.58	D
Ctnnb1	0.04	2.74	F
Gnb2l1	0.04	2.78	D
Tuba1b	0.04	2.78	D
Mapk3	0.04	2.95	D
Stk39	0.04	3.66	I
Tpi1	0.04	3.67	G
Grin2b	0.03	2.45	A

Bibliography

- W. C. Abraham and M. F. Bear. Metaplasticity: the plasticity of synaptic plasticity. *Trends in Neurosciences*, 19(4):126–30, Apr 1996.
- W. C. Abraham and W. P. Tate. Metaplasticity: a new vista across the field of synaptic plasticity. *Prog Neurobiol*, 52(4):303–23, Jul 1997.
- F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–80, Mar 2004.
- R. Albert, H. Jeong, and A. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.
- R. Aldecoa and I. Marín. Deciphering network community structure by surprise. *PLoS ONE*, 6(9):e24195, Jan 2011.
- B. Alex, B. Haddow, and C. Grover. Recognising nested named entities in biomedical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 65–72, 2007.
- B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, R. Tobin, and X. Wang. Automating curation using a natural language processing pipeline. *Genome Biol*, 9 Suppl 2:S10, Jan 2008.
- J. W. Allen, B. A. Eldadah, and A. I. Faden. Beta-amyloid-induced apoptosis of cerebellar granule cells and cortical neurons: exacerbation by selective inhibition of group i metabotropic glutamate receptors. *Neuropharmacology*, 38(8):1243–52, Aug 1999.
- N. C. Allen, S. Bagade, M. B. McQueen, J. P. A. Ioannidis, F. K. Kavvoura, M. J. Khoury, R. E. Tanzi, and L. Bertram. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the szgene database. *Nat Genet*, 40(7):827–34, Jul 2008.
- U. Alon. Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6):450–61, Jun 2007.
- M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, and S. Kanaya. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC bioinformatics*, 7(1):207, 2006.

- G. Alvarez-Bolado and G. Eichele. Analysing the developing brain transcriptome with the genepaint platform. *J Physiol (Lond)*, 575(Pt 2):347–52, Sep 2006.
- J. J. An, K. Gharami, G.-Y. Liao, N. H. Woo, A. G. Lau, F. Vanevski, E. R. Torre, K. R. Jones, Y. Feng, B. Lu, and B. Xu. Distinct role of long 3' utr bdnf mrna in spine morphology and synaptic plasticity in hippocampal neurons. *Cell*, 134(1):175–87, Jul 2008.
- P.-O. Angrand, I. Segura, P. Völkel, S. Ghidelli, R. Terry, M. Brajenovic, K. Vintersten, R. Klein, G. Superti-Furga, G. Drewes, B. Kuster, T. Bouwmeester, and A. Acker-Palmer. Transgenic mouse proteomics identifies new 14-3-3-associated proteins involved in cytoskeletal rearrangements and cell signaling. *Mol Cell Proteomics*, 5(12):2211–27, Dec 2006.
- A. Antonov, T. Schmidt, Y. Wang, and H. Mewes. Profcom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Research*, 36(suppl 2):W347, 2008.
- B. Aranda, H. Blankenburg, S. Kerrien, F. S. L. Brinkman, A. Ceol, E. Chautard, J. M. Dana, J. D. L. Rivas, M. Dumousseau, E. Galeota, A. Gaulton, J. Goll, R. E. W. Hancock, R. Isserlin, R. C. Jimenez, J. Kerssemakers, J. Khadake, D. J. Lynn, M. Michaut, G. O'Kelly, K. Ono, S. Orchard, C. Prieto, S. Razick, O. Rigina, L. Salwinski, M. Simonovic, S. Velankar, A. Winter, G. Wu, G. D. Bader, G. Cesareni, I. M. Donaldson, D. Eisenberg, G. J. Kleywegt, J. Overington, S. Ricard-Blum, M. Tyers, M. Albrecht, and H. Hermjakob. Psicquic and psiscore: accessing and scoring molecular interactions. *Nat Methods*, 8(7):528–9, Jul 2011.
- D. M. Armstrong, M. D. Ikonovic, R. Sheffield, and R. J. Wenthold. Ampa-selective glutamate receptor subtype immunoreactivity in the entorhinal cortex of non-demented elderly and patients with alzheimer's disease. *Brain Res*, 639(2):207–16, Mar 1994.
- V. Arnau, S. Mars, and I. Marín. Iterative cluster analysis of protein interaction data. *Bioinformatics*, 21(3):364, 2005.
- M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, and J. Eppig. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25, 2000.
- M. C. Ashby, S. A. D. L. Rue, G. S. Ralph, J. Uney, G. L. Collingridge, and J. M. Henley. Removal of ampa receptors (ampars) from synapses is preceded by transient endocytosis of extrasynaptic ampars. *J Neurosci*, 24(22):5172–6, Jun 2004.
- S. Asur, D. Ucar, and S. Parthasarathy. An ensemble framework for clustering protein–protein interaction networks. *Bioinformatics*, 23(13):i29, 2007.
- A. Bachmann, M. Timmer, J. Sierralta, G. Pietrini, E. D. Gundelfinger, E. Knust, and U. Thomas. Cell type-specific recruitment of drosophila lin-7 to distinct maguk-based protein complexes defines novel roles for sdt and dlg-s97. *J Cell Sci*, 117(Pt 10):1899–909, Apr 2004.

- A. Bachmann, O. Kobler, R. J. Kittel, C. Wichmann, J. Sierralta, S. J. Sigrist, E. D. Gundelfinger, E. Knust, and U. Thomas. A perisynaptic ménage à trois between dlg, dlin-7, and metro controls proper organization of drosophila synaptic junctions. *J Neurosci*, 30(17):5811–24, Apr 2010.
- G. Bader and C. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):2, 2003.
- F. Bai and F. Witzmann. Synaptosome proteomics. *Subcellular Proteomics*, pages 77–98, 2007.
- B. A. Ballyk and J. W. Goh. A postsynaptic g-protein in hippocampal long-term potentiation. *Brain Res*, 611(1):81–6, May 1993.
- M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem*, 389(4):1017–31, Oct 2007.
- A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.
- A. Barabási, Z. Dezso, E. Ravasz, S. Yook, and Z. Oltvai. Scale-free and hierarchical structures in complex networks. *AIP Conference Proceedings*, 661:1, 2003.
- G. Barnard. A new test for 2 x 2 tables. *Nature*, Jan 1945.
- T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muerter, and R. Edgar. Ncbi geo: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37(Database issue):D885–90, Jan 2009.
- T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muerter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. Ncbi geo: archive for functional genomics data sets—10 years on. *Nucleic Acids Res*, 39(Database issue):D1005–10, Jan 2011.
- A. Barria, D. Muller, V. Derkach, L. C. Griffith, and T. R. Soderling. Regulatory phosphorylation of ampa-type glutamate receptors by cam-kii during long-term potentiation. *Science*, 276(5321):2042–5, Jun 1997.
- M. Barrios-Rodiles, K. R. Brown, B. Ozdamar, R. Bose, Z. Liu, R. S. Donovan, F. Shinjo, Y. Liu, J. Dembowy, I. W. Taylor, V. Luga, N. Przulj, M. Robinson, H. Suzuki, Y. Hayashizaki, I. Jurisica, and J. L. Wrana. High-throughput mapping of a dynamic signaling network in mammalian cells. *Science*, 307(5715):1621–5, Mar 2005.
- A. Bayés and S. Grant. Neuroproteomics: understanding the molecular organization and complexity of the brain. *Nat Rev Neurosci*, 10(9):635–646, 2009.

- A. Bayés, L. van de Lagemaat, M. Collins, M. Croning, I. Whittle, J. Choudhary, and S. Grant. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat Neurosci*, 14(1):19–21, 2010.
- M. F. Bear and W. C. Abraham. Long-term depression in hippocampus. *Annu Rev Neurosci*, 19:437–62, Jan 1996.
- C. Bécamel, S. Gavarini, B. Chanrion, G. Alonso, N. Galéotti, A. Dumuis, J. Bockaert, and P. Marin. The serotonin 5-ht_{2a} and 5-ht_{2c} receptors interact with specific sets of pdz proteins. *J Biol Chem*, 279(19):20257–66, May 2004.
- K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang. The genetic association database. *Nat Genet*, 36(5):431–2, May 2004.
- J.-C. Béïque, D.-T. Lin, M.-G. Kang, H. Aizawa, K. Takamiya, and R. L. Huganir. Synapse-specific regulation of ampa receptor function by psd-95. *Proc Natl Acad Sci USA*, 103(51):19535–40, Dec 2006.
- R. Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16(1), 1958.
- M. P. Belvin and J. C. Yin. Drosophila learning and memory: recent progress and new approaches. *Bioessays*, 19(12):1083–9, Dec 1997.
- M. Bence, M. I. Arbuckle, K. S. Dickson, and S. G. N. Grant. Analyses of murine postsynaptic density-95 identify novel isoforms and potential translational control elements. *Brain Res Mol Brain Res*, 133(1):143–52, Jan 2005.
- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.
- Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
- A.-C. Berglund, E. Sjölund, G. Ostlund, and E. L. L. Sonnhammer. Inparanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Research*, 36(Database issue):D263–6, Jan 2008.
- S. Berkel, C. R. Marshall, B. Weiss, J. Howe, R. Roeth, U. Moog, V. Endris, W. Roberts, P. Szatmari, D. Pinto, M. Bonin, A. Riess, H. Engels, R. Sprengel, S. W. Scherer, and G. A. Rappold. Mutations in the shank2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nat Genet*, 42(6):489–91, Jun 2010.
- G. F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–4, Dec 2003.
- L. Bertram, M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi. Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. *Nat Genet*, 39(1):17–23, Jan 2007.

- F. Besse, S. Mertel, R. J. Kittel, C. Wichmann, T. M. Rasse, S. J. Sigrist, and A. Ephrussi. The ig cell adhesion molecule basigin controls compartmentalization and vesicle release at drosophila melanogaster synapses. *J Cell Biol*, 177(5):843–55, Jun 2007.
- J. A. Blake, J. E. Richardson, C. J. Bult, J. A. Kadin, J. T. Eppig, and M. G. D. Group. The mouse genome database (mgd): the model organism database for the laboratory mouse. *Nucleic Acids Research*, 30(1):113–5, Jan 2002.
- T. A. Blanpied, D. B. Scott, and M. D. Ehlers. Dynamics and regulation of clathrin coats at specialized endocytic zones of dendrites and spines. *Neuron*, 36(3):435–49, Oct 2002.
- T. V. Bliss and G. L. Collingridge. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361(6407):31–9, Jan 1993.
- T. V. Bliss and A. R. Gardner-Medwin. Long-lasting potentiation of synaptic transmission in the dentate area of the unanaesthetized rabbit following stimulation of the perforant path. *J Physiol (Lond)*, 232(2):357–74, Jul 1973.
- R. D. Blitzer, J. H. Connor, G. P. Brown, T. Wong, S. Shenolikar, R. Iyengar, and E. M. Landau. Gating of camkii by camp-regulated protein phosphatase activity during ltp. *Science*, 280(5371):1940–2, Jun 1998.
- R. J. Bloch and J. S. Morrow. An unusual beta-spectrin associated with clustered acetylcholine receptors. *J Cell Biol*, 108(2):481–93, Feb 1989.
- A. Bossi and B. Lehner. Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, Jan 2009.
- H. R. Bourne and R. Nicoll. Molecular machines integrate coincident synaptic signals. *Cell*, 72 Suppl:65–75, Jan 1993.
- T. Bouwmeester, A. Bauch, H. Ruffner, P. Angrand, G. Bergamini, K. Croughton, C. Cruciat, D. Eberhard, J. Gagneur, and S. Ghidelli. A physical and functional map of the human tnf-a / nf-kb signal transduction pathway. *Nature cell biology*, 6(2):97–105, 2004.
- S. P. Braithwaite, H. Xia, and R. C. Malenka. Differential roles for nsf and grip/abp in ampa receptor cycling. *Proc Natl Acad Sci USA*, 99(10):7096–101, May 2002.
- M. Brajenovic, G. Joberty, B. Küster, T. Bouwmeester, and G. Drewes. Comprehensive proteomic analysis of human par protein complexes reveals an interconnected protein network. *J Biol Chem*, 279(13):12804–11, Mar 2004.
- R. Brambilla, N. Gnesutta, L. Minichiello, G. White, A. J. Roylance, C. E. Herron, M. Ramsey, D. P. Wolfer, V. Cestari, C. Rossi-Arnaud, S. G. Grant, P. F. Chapman, H. P. Lipp, E. Sturani, and R. Klein. A role for the ras signalling pathway in synaptic transmission and long-term memory. *Nature*, 390(6657):281–6, Nov 1997.

- D. S. Bredt and R. A. Nicoll. Ampa receptor trafficking at excitatory synapses. *Neuron*, 40(2):361–79, Oct 2003.
- H. Breer and G. Jeserich. A microscale floatation technique for the isolation of synaptosomes from nervous tissue of locusta migratoria. *Insect Biochemistry*, Jan 1980.
- R. Breer and M. Knipper. Characterization of acetylcholine release from insect synaptosomes. *Insect Biochemistry*, Jan 1984.
- S. Brohee and J. van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, 7(1):488, 2006.
- P. Bronk, J. J. Wenniger, K. Dawson-Scully, X. Guo, S. Hong, H. L. Atwood, and K. E. Zinsmaier. Drosophila hsc70-4 is critical for neurotransmitter exocytosis in vivo. *Neuron*, 30(2):475–88, May 2001.
- T. M. Brotz, E. D. Gundelfinger, and A. Borst. Cholinergic and gabaergic pathways in fly motion vision. *BMC Neurosci*, 2:1, Jan 2001.
- D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, and N. Zhang. Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Research*, 31(9):2443, 2003.
- D. B. Budimirovic and W. E. Kaufmann. What can we learn about autism from studying fragile x syndrome? *Developmental neuroscience*, Sep 2011.
- V. Budnik, Y. H. Koh, B. Guan, B. Hartmann, C. Hough, D. Woods, and M. Gorczyca. Regulation of synapse structure and function by the drosophila tumor suppressor gene dlg. *Neuron*, 17(4):627–40, Oct 1996.
- V. Budnik, M. Gorczyca, and A. Prokop. Selected methods for the anatomical study of drosophila embryonic and larval neuromuscular junctions. *Int Rev Neurobiol*, 75: 323–65, Jan 2006.
- C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, J. A. Blake, and M. G. D. Group. The mouse genome database (mgd): mouse biology and model systems. *Nucleic Acids Research*, 36(Database issue):D724–8, Jan 2008.
- T. Bürckstümmer, K. L. Bennett, A. Preradovic, G. Schütze, O. Hantschel, G. Superti-Furga, and A. Bauch. An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat Methods*, 3(12):1013–9, Dec 2006.
- J. Burré and W. Volkandt. The synaptic vesicle proteome. *J Neurochem*, 101(6): 1448–62, Jun 2007.
- J. Burré, T. Beckhaus, H. Schägger, C. Corvey, S. Hofmann, M. Karas, H. Zimmermann, and W. Volkandt. Analysis of the synaptic vesicle proteome using three gel-based protein separation techniques. *Proteomics*, 6(23):6250–62, Dec 2006.
- Z. Cai, X. Mao, S. Li, and L. Wei. Genome comparison using gene ontology(go) with statistical testing. *BMC bioinformatics*, 7(1):374, 2006.

- P. Calabresi, B. Picconi, L. Parnetti, and M. D. Filippo. A convergent model for cognitive dysfunctions in parkinson's disease: the critical dopamine-acetylcholine synaptic balance. *Lancet Neurol*, 5(11):974–83, Nov 2006.
- J. Caltagarone, Z. Jing, and R. Bowser. Focal adhesions regulate abeta signaling and cell death in alzheimer's disease. *Biochim Biophys Acta*, 1772(4):438–45, Apr 2007.
- R. K. Carlin, D. J. Grab, R. S. Cohen, and P. Siekevitz. Isolation and characterization of postsynaptic densities from various brain regions: enrichment of different types of postsynaptic densities. *J Cell Biol*, 86(3):831–45, Sep 1980.
- H. J. Carlisle, A. E. Fink, S. G. N. Grant, and T. J. O'Dell. Opposing effects of psd-93 and psd-95 on long-term potentiation and spike timing-dependent plasticity. *The Journal of Physiology*, 586(Pt 24):5885–900, Dec 2008.
- L. C. Cary, M. Goebel, B. G. Corsaro, H. G. Wang, E. Rosen, and M. J. Fraser. Transposon mutagenesis of baculoviruses: analysis of trichoplusia ni transposon ifp2 insertions within the fp-locus of nuclear polyhedrosis viruses. *Virology*, 172(1):156–69, Sep 1989.
- M. E. Castro, A. Diaz, E. del Olmo, and A. Pazos. Chronic fluoxetine induces opposite changes in g protein coupling at pre and postsynaptic 5-ht1a receptors in rat brain. *Neuropharmacology*, 44(1):93–101, Jan 2003.
- A. M. Celotto, A. C. Frank, J. L. Seigle, and M. J. Palladino. Drosophila model of human inherited triosephosphate isomerase deficiency glycolytic enzymopathy. *Genetics*, 174(3):1237–46, Nov 2006.
- A. Ceol, A. C. Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. Mint, the molecular interaction database: 2009 update. *Nucleic Acids Res*, 38(Database issue):D532–9, Jan 2010.
- G. Cesareni, A. Chatr-aryamontri, L. Licata, and A. Ceol. Searching the mint database for protein interaction information. *Curr Protoc Bioinformatics*, Chapter 8:Unit 8.5, Jun 2008.
- H. C. Chang and G. M. Rubin. 14-3-3 epsilon positively regulates ras-mediated signaling in drosophila. *Genes Dev*, 11(9):1132–9, May 1997.
- B. A. Chase and D. R. Kankel. A genetic analysis of glutamatergic function in drosophila. *J Neurobiol*, 18(1):15–41, Jan 1987.
- A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni. Mint: the molecular interaction database. *Nucleic Acids Res*, 35(Database issue):D572–4, Jan 2007.
- A. Chatr-aryamontri, A. Zanzoni, A. Ceol, and G. Cesareni. Searching the protein interaction space through the mint database. *Methods Mol Biol*, 484:305–17, Jan 2008.

- G. I. Chen and A.-C. Gingras. Affinity-purification mass spectrometry (ap-ms) of serine/threonine phosphatases. *Methods*, 42(3):298–305, Jul 2007.
- H. Chen and B. M. Sharp. 1471-2105-5-147. *BMC Bioinformatics*, 5(1):147, Jan 2004.
- K. Chen and D. E. Featherstone. Discs-large (dlg) is clustered by presynaptic innervation and regulates postsynaptic glutamate receptor subunit composition in drosophila. *BMC Biol*, 3:1, Jan 2005.
- L. Chen, D. M. Chetkovich, R. S. Petralia, N. T. Sweeney, Y. Kawasaki, R. J. Wenthold, D. S. Brecht, and R. A. Nicoll. Stargazin regulates synaptic targeting of ampa receptors by two distinct mechanisms. *Nature*, 408(6815):936–43, Jan 2000.
- X. Chen, C. Winters, R. Azzam, X. Li, J. A. Galbraith, R. D. Leapman, and T. S. Reese. Organization of the core structure of the postsynaptic density. *Proceedings of the National Academy of Sciences of the United States of America*, 105(11):4453–8, Mar 2008.
- Y. Chen, J. Bourne, V. A. Pieribone, and R. M. Fitzsimonds. The role of actin in the regulation of dendritic spine morphology and bidirectional synaptic plasticity. *Neuroreport*, 15(5):829–32, Apr 2004.
- D. M. Chetkovich, R. C. Bunn, S.-H. Kuo, Y. Kawasaki, M. Kohwi, and D. S. Brecht. Postsynaptic targeting of alternative postsynaptic density-95 isoforms by distinct mechanisms. *J Neurosci*, 22(15):6415–25, Aug 2002.
- V. R. Chintapalli, J. Wang, and J. A. T. Dow. Using flyatlas to identify better drosophila melanogaster models of human disease. *Nat Genet*, 39(6):715–20, Jun 2007.
- K. O. Cho, C. A. Hunt, and M. B. Kennedy. The rat brain postsynaptic density fraction contains a homolog of the drosophila discs-large tumor suppressor protein. *Neuron*, 9(5):929–42, Nov 1992.
- Y.-R. Cho, W. Hwang, and A. Zhang;. Efficient modularization of weighted protein interaction networks using k-hop graph reduction. *Bioinformatics and BioEngineering, 2006. BIBE 2006. Sixth IEEE Symposium on*, pages 289 – 298, 2006.
- J. Choudhary and S. G. N. Grant. Proteomics in postgenomic neuroscience: the end of the beginning. *Nat Neurosci*, 7(5):440–5, May 2004.
- S. Chowdhury, J. D. Shepherd, H. Okuno, G. Lyford, R. S. Petralia, N. Plath, D. Kuhl, R. L. Huganir, and P. F. Worley. Arc/arg3.1 interacts with the endocytic machinery to regulate ampa receptor trafficking. *Neuron*, 52(3):445–59, Nov 2006.
- J. J. E. Chua, S. Kindler, J. Boyken, and R. Jahn. The architecture of an excitatory synapse. *J Cell Sci*, 123(Pt 6):819–23, Mar 2010.
- H. J. Chung, Y. H. Huang, L.-F. Lau, and R. L. Huganir. Regulation of the nmda receptor complex and trafficking by activity-dependent phosphorylation of the nr2b subunit pdz ligand. *J Neurosci*, 24(45):10248–59, Nov 2004.

- A. Citri and R. C. Malenka. Synaptic plasticity: Multiple forms, functions, and mechanisms. *Neuropsychopharmacology*, 33(1):18–41, Jan 2008.
- A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.
- A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data. *Arxiv preprint arXiv:0706.1062*, 2007.
- M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P.-L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader. Integration of biological networks and gene expression data using cytoscape. *Nature protocols*, 2(10):2366–82, Jan 2007.
- M. P. Coba, L. M. Valor, M. V. Kopanitsa, N. O. Afinowi, and S. G. N. Grant. Kinase networks integrate profiles of n-methyl-d-aspartate receptor-mediated gene expression in hippocampus. *Journal of Biological Chemistry*, 283(49):34101–34107, Aug 2008.
- M. P. Coba, A. J. Pocklington, M. O. Collins, M. V. Kopanitsa, R. T. Uren, S. Swamy, M. D. R. Croning, J. S. Choudhary, and S. G. N. Grant. Neurotransmitters drive combinatorial multistate postsynaptic density networks. *Sci Signal*, 2(68):ra19, Jan 2009.
- P. D. Coleman and P. J. Yao. Synaptic slaughter in alzheimer’s disease. *Neurobiol Aging*, 24(8):1023–7, Dec 2003.
- C. A. Collins and A. DiAntonio. Synaptic development: insights from drosophila. *Curr Opin Neurobiol*, 17(1):35–42, Feb 2007.
- M. O. Collins, L. Yu, M. P. Coba, H. Husi, I. Campuzano, W. P. Blackstock, J. S. Choudhary, and S. G. N. Grant. Proteomic analysis of in vivo phosphorylated synaptic proteins. *J Biol Chem*, 280(7):5972–82, Feb 2005.
- M. O. Collins, H. Husi, L. Yu, J. M. Brandon, C. N. G. Anderson, W. P. Blackstock, J. S. Choudhary, and S. G. N. Grant. Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *J Neurochem*, 97 Suppl 1:16–23, Apr 2006.
- J. T. Coyle. Glutamate and schizophrenia: beyond the dopamine hypothesis. *Cell Mol Neurobiol*, 26(4-6):365–84, Jan 2006.
- A. M. Craig and Y. Kang. Neurexin-neuroigin signaling in synapse development. *Curr Opin Neurobiol*, 17(1):43–52, Feb 2007.
- S. E. Craven and D. S. Bredt. Pdz proteins organize synaptic signaling pathways. *Cell*, 93(4):495–8, May 1998.

- M. D. R. Croning, M. C. Marshall, P. McLaren, J. D. Armstrong, and S. G. N. Grant. G2cdb: the genes to cognition database. *Nucleic Acids Research*, 37(Database issue):D846–51, Jan 2009.
- J. A. Cummings, R. M. Mulkey, R. A. Nicoll, and R. C. Malenka. Ca²⁺ signaling requirements for long-term depression in the hippocampus. *Neuron*, 16(4):825–33, Apr 1996.
- P. C. Cuthbert, L. E. Stanford, M. P. Coba, J. A. Ainge, A. E. Fink, P. Opazo, J. Y. Delgado, N. H. Komiyama, T. J. O'Dell, and S. G. N. Grant. Synapse-associated protein 102/dlgh3 couples the nmda receptor to specific plasticity pathways and learning strategies. *J Neurosci*, 27(10):2673–82, Mar 2007.
- A. B. da Cruz, M. Schwärzel, S. Schulze, M. Niyiyati, M. Heisenberg, and D. Kretschmar. Disruption of the map1b-related protein futsch leads to changes in the neuronal cytoskeleton, axonal transport defects, and progressive neurodegeneration in drosophila. *Mol Biol Cell*, 16(5):2433–42, May 2005.
- M. B. Dalva, A. C. McClelland, and M. S. Kayser. Cell adhesion molecules: signalling functions at the synapse. *Nat Rev Neurosci*, 8(3):206–20, Mar 2007.
- M. P. Daniels. Localization of actin, beta-spectrin, 43 x 10(3) mr and 58 x 10(3) mr proteins to receptor-enriched domains of newly formed acetylcholine receptor aggregates in isolated myotube membranes. *J Cell Sci*, 97 (Pt 4):615–26, Dec 1990.
- G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, Jan 2003.
- V. Derkach, A. Barria, and T. R. Soderling. Ca²⁺/calmodulin-kinase ii enhances channel conductance of alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionate type glutamate receptors. *Proc Natl Acad Sci USA*, 96(6):3269–74, Mar 1999.
- V. A. Derkach, M. C. Oh, E. S. Guire, and T. R. Soderling. Regulatory mechanisms of ampa receptors in synaptic plasticity. *Nat Rev Neurosci*, 8(2):101–13, Feb 2007.
- E. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.
- M. Dittrich, G. Klau, A. Rosenwald, T. Dandekar, and T. Muller. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–31, 2008.
- M. Dobretsov and J. R. Stimers. Na⁺/k pump current in guinea pig cardiac myocytes and the effect of na leak. *J Cardiovasc Electrophysiol*, 8(7):758–67, Jul 1997.
- G. Dölen and M. F. Bear. Role for metabotropic glutamate receptor 5 (mglur5) in the pathogenesis of fragile x syndrome. *J Physiol (Lond)*, 586(6):1503–8, Mar 2008.

- B. Domon and R. Aebersold. Mass spectrometry and protein analysis. *Science*, 312(5771):212–7, Apr 2006.
- S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, and B. R. Conklin. Mappfinder: using gene ontology and genmapp to create a global gene-expression profile from microarray data. *Genome Biol*, 4(1):R7, Jan 2003.
- A. Dosemeci and T. S. Reese. Effect of calpain on the composition and structure of postsynaptic densities. *Synapse*, 20(1):91–7, May 1995.
- A. Dosemeci, A. J. Makusky, E. Jankowska-Stephens, X. Yang, D. J. Slotta, and S. P. Markey. Composition of the synaptic psd-95 complex. *Mol Cell Proteomics*, 6(10):1749–60, Oct 2007.
- J. P. Doyle, J. D. Dougherty, M. Heiman, E. F. Schmidt, T. R. Stevens, G. Ma, S. Bupp, P. Shrestha, R. D. Shah, M. L. Doughty, S. Gong, P. Greengard, and N. Heintz. Application of a translational profiling approach for the comparative analysis of cns cell types. *Cell*, 135(4):749–62, Nov 2008a.
- J. P. Doyle, J. D. Dougherty, M. Heiman, E. F. Schmidt, T. R. Stevens, G. Ma, S. Bupp, P. Shrestha, R. D. Shah, M. L. Doughty, S. Gong, P. Greengard, and N. Heintz. Application of a translational profiling approach for the comparative analysis of cns cell types. *Cell*, 135(4):749–762, Nov 2008b.
- R. Drakas, M. Prisco, and R. Baserga. A modified tandem affinity purification tag technique for the purification of protein complexes in mammalian cells. *Proteomics*, 5(1):132–7, Jan 2005.
- S. M. Dudek and M. F. Bear. Homosynaptic long-term depression in area ca1 of hippocampus and effects of n-methyl-d-aspartate receptor blockade. *Proc Natl Acad Sci USA*, 89(10):4363–7, May 1992.
- R. Dunn, F. Dudbridge, and C. Sanderson. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC bioinformatics*, Jan 2005.
- R. Efendiev, B. K. Samelson, B. T. Nguyen, P. V. Phatarpekar, F. Baameur, J. D. Scott, and C. W. Dessauer. Akap79 interacts with multiple adenylyl cyclase (ac) isoforms and scaffolds ac5 and -6 to alpha-amino-3-hydroxyl-5-methyl-4-isoxazole-propionate (ampa) receptors. *J Biol Chem*, 285(19):14450–8, May 2010.
- I. Ehrlich and R. Malinow. Postsynaptic density 95 controls ampa receptor incorporation during long-term potentiation and experience-driven synaptic plasticity. *J Neurosci*, 24(4):916–27, Jan 2004.
- G. Ellison. The n-methyl-d-aspartate antagonists phencyclidine, ketamine and dizocilpine as both behavioral and anatomical models of the dementias. *Brain Res Brain Res Rev*, 20(2):250–67, Feb 1995.

- R. D. Emes and S. G. N. Grant. The human postsynaptic density shares conserved elements with proteomes of unicellular eukaryotes and prokaryotes. *Front Neurosci*, 5:44, Jan 2011.
- R. D. Emes, A. J. Pocklington, C. N. G. Anderson, A. Bayes, M. O. Collins, C. A. Vickers, M. D. R. Croning, B. R. Malik, J. S. Choudhary, J. D. Armstrong, and S. G. N. Grant. Evolutionary expansion and anatomical specialization of synapse proteome complexity. *Nat Neurosci*, 11(7):799–806, Jul 2008.
- A. Enright and C. Ouzounis. Biolayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, 17(9):853, 2001.
- A. Enright, S. V. Dongen, and C. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575, 2002.
- J. T. Eppig, C. J. Bult, J. A. Kadin, J. E. Richardson, J. A. Blake, A. Anagnostopoulos, R. M. Baldarelli, M. Baya, J. S. Beal, S. M. Bello, W. J. Boddy, D. W. Bradt, D. L. Burkart, N. E. Butler, J. Campbell, M. A. Cassell, L. E. Corbani, S. L. Cousins, D. J. Dahmen, H. Dene, A. D. Diehl, H. J. Drabkin, K. S. Frazer, P. Frost, L. H. Glass, C. W. Goldsmith, P. L. Grant, M. Lennon-Pierce, J. Lewis, I. Lu, L. J. Maltais, M. McAndrews-Hill, L. McClellan, D. B. Miers, L. A. Miller, L. Ni, J. E. Ormsby, D. Qi, T. B. K. Reddy, D. J. Reed, B. Richards-Smith, D. R. Shaw, R. Sinclair, C. L. Smith, P. Szauter, M. B. Walker, D. O. Walton, L. L. Washburn, I. T. Witham, Y. Zhu, and M. G. D. Group. The mouse genome database (mgd): from genes to mice—a community resource for mouse biology. *Nucleic Acids Research*, 33(Database issue):D471–5, Jan 2005.
- J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, and M. G. D. Group. The mouse genome database (mgd): new features facilitating a model system. *Nucleic Acids Research*, 35(Database issue):D630–7, Jan 2007.
- P. Erdős and G. Szekeres. A combinatorial problem in geometry. *Classic Papers in Combinatorics*, pages 49–56, 1987.
- R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore, S. Zhang, O. Ornatsky, Y. V. Bukhman, M. Ethier, Y. Sheng, J. Vasilescu, M. Abu-Farha, J.-P. Lambert, H. S. Duewel, I. I. Stewart, B. Kuehl, K. Hogue, K. Colwill, K. Gladwish, B. Muskat, R. Kinach, S.-L. Adams, M. F. Moran, G. B. Morin, T. Topaloglou, and D. Figeys. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology*, 3:89, Jan 2007.
- C. D. Farr, P. R. Gafken, A. D. Norbeck, C. E. Doneanu, M. D. Stapels, D. F. Barofsky, M. Minami, and J. A. Saugstad. Proteomic analysis of native metabotropic glutamate receptor 5 protein complexes reveals novel molecular constituents. *J Neurochem*, 91(2):438–50, Oct 2004.
- D. E. Featherstone, W. S. Davis, R. R. Dubreuil, and K. Broadie. *Drosophila* alpha- and beta-spectrin mutations disrupt presynaptic neurotransmitter release. *J Neurosci*, 21(12):4215–24, Jun 2001.

- D. E. Featherstone, E. Rushton, and K. Broadie. Developmental regulation of glutamate receptor field size by nonvesicular glutamate release. *Nat Neurosci*, 5(2): 141–6, Feb 2002.
- W. Feng and M. Zhang. Organization and dynamics of pdz-domain-related supramodules in the postsynaptic density. *Nat Rev Neurosci*, 10(2):87–99, Feb 2009.
- E. Fernández, M. Collins, R. Uren, M. Kopanitsa, N. Komiyama, M. Croning, L. Zografos, J. Armstrong, J. Choudhary, and S. Grant. Targeted tandem affinity purification of psd-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Molecular Systems Biology*, 5(1), 2009.
- J. H. Finger, C. M. Smith, T. F. Hayamizu, I. J. McCright, J. T. Eppig, J. A. Kadin, J. E. Richardson, and M. Ringwald. The mouse gene expression database (gxd): 2011 update. *Nucleic Acids Research*, 39(Database issue):D835–41, Jan 2011.
- R. D. Finn, J. Mistry, J. Tate, P. Coghill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman. The pfam protein families database. *Nucleic Acids Research*, 38(Database issue):D211–22, Jan 2010.
- R. Fisher. On the interpretation of chi squared from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.
- A. I. Fogel, M. R. Akins, A. J. Krupp, M. Stagi, V. Stein, and T. Biederer. Syncams organize synapses through heterophilic adhesion. *J Neurosci*, 27(46):12516–30, Nov 2007.
- A. I. Fogel, Y. Li, J. Giza, Q. Wang, T. T. Lam, Y. Modis, and T. Biederer. N-glycosylation at the syncam (synaptic cell adhesion molecule) immunoglobulin interface modulates synaptic adhesion. *J Biol Chem*, 285(45):34864–74, Nov 2010.
- E. Formstecher, S. Aresta, V. Collura, A. Hamburger, A. Meil, A. Trehin, C. Reverdy, V. Betin, S. Maire, C. Brun, B. Jacq, M. Arpin, Y. Bellaiche, S. Bellusci, P. Benaroch, M. Bornens, R. Chanet, P. Chavrier, O. Delattre, V. Doye, R. Fehon, G. Faye, T. Galli, J.-A. Girault, B. Goud, J. de Gunzburg, L. Johannes, M.-P. Junier, V. Mirouse, A. Mukherjee, D. Papadopoulo, F. Perez, A. Plessis, C. Rossé, S. Saule, D. Stoppa-Lyonnet, A. Vincent, M. White, P. Legrain, J. Wojcik, J. Camonis, and L. Daviet. Protein interaction mapping: a drosophila case study. *Genome Research*, 15(3):376–84, Mar 2005.
- A. Fox, D. Taylor, and D. K. Slonim. High throughput interaction data reveals degree conservation of hub proteins. *Pac Symp Biocomput*, pages 391–402, Jan 2009.
- R. A. W. Frank, A. F. McRae, A. J. Pocklington, L. N. van de Lagemaat, P. Navarro, M. D. R. Croning, N. H. Komiyama, S. J. Bradley, R. A. J. Challiss, J. D. Armstrong, R. D. Finn, M. P. Malloy, A. W. Maclean, S. E. Harris, J. M. Starr, S. S. Bhaskar, E. K. Howard, S. E. Hunt, A. J. Coffey, V. Ranganath, P. Deloukas, J. Rogers, W. J. Muir, I. J. Deary, D. H. Blackwood, P. M. Visscher, and S. G. N. Grant. Clustered

- coding variants in the glutamate receptor complexes of individuals with schizophrenia and bipolar disorder. *PLoS ONE*, 6(4):e19011, Jan 2011.
- P. W. Frankland, C. O'Brien, M. Ohno, A. Kirkwood, and A. J. Silva. Alpha-camkii-dependent plasticity in the cortex is required for permanent memory. *Nature*, 411(6835):309–13, May 2001.
- H. B. Fraser, D. P. Wall, and A. E. Hirsh. A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol Biol*, 3:11, May 2003.
- J.-M. Fritschy. Epilepsy, e/i balance and gaba(a) receptor plasticity. *Front Mol Neurosci*, 1:5, Jan 2008.
- Y. Fukata, H. Adesnik, T. Iwanaga, D. S. Bredt, R. A. Nicoll, and M. Fukata. Epilepsy-related ligand/receptor complex Igi1 and adam22 regulate synaptic transmission. *Science*, 313(5794):1792–5, Sep 2006.
- S. Fusi and L. F. Abbott. Limits on the memory storage capacity of bounded synapses. *Nat Neurosci*, 10(4):485–93, Apr 2007.
- S. Fusi, P. J. Drew, and L. F. Abbott. Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611, Feb 2005.
- S. M. Gallagher, C. A. Daly, M. F. Bear, and K. M. Huber. Extracellular signal-regulated protein kinase activation is required for metabotropic glutamate receptor-dependent long-term depression in hippocampal area ca1. *J Neurosci*, 24(20):4859–64, May 2004.
- G. Gallone, T. I. Simpson, J. D. Armstrong, and A. P. Jarman. Bio::homology::interologwalk—a perl module to build putative protein-protein interaction networks through interolog mapping. *BMC Bioinformatics*, 12:289, Jan 2011.
- T. K. B. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, S. S. Mohan, S. Sharma, S. Pinkert, S. Nagaraju, B. Periaswamy, G. Mishra, K. Nandakumar, B. Shen, N. Deshpande, R. Nayak, M. Sarker, J. D. Boeke, G. Parmigiani, J. Schultz, J. S. Bader, and A. Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–93, Mar 2006.
- A. Gardiol, C. Racca, and A. Triller. Dendritic and postsynaptic protein synthetic machinery. *J Neurosci*, 19(1):168–79, Jan 1999.
- F. Gardoni. Maguk proteins: new targets for pharmacological intervention in the glutamatergic synapse. *Eur J Pharmacol*, 585(1):147–52, May 2008.
- F. Gardoni, F. Polli, F. Cattabeni, and M. D. Luca. Calcium-calmodulin-dependent protein kinase ii phosphorylation modulates psd-95 binding to nmda receptors. *Eur J Neurosci*, 24(10):2694–704, Nov 2006.

- A. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick, A. Michon, and C. Cruciat. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.
- A. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. Jensen, S. Bastuck, and B. Dümpelfeld. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.
- W. Gelbart, M. Crosby, B. Matthews, W. Rindone, J. Chillemi, S. Twombly, D. Emmer, M. Ashburner, R. Drysdale, and E. Whitfield. Flybase: a drosophila database. the flybase consortium. *Nucleic Acids Research*, 25(1):63, 1997.
- E. S. Gershon, N. Alliey-Rodriguez, and C. Liu. After gwas: searching for genetic risk for schizophrenia and bipolar disorder. *Am J Psychiatry*, 168(3):253–6, Mar 2011.
- D. Geschwind and G. Konopka. Neuroscience in the era of functional genomics and systems biology. *Nature*, 461(7266):908–915, 2009.
- L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of drosophila melanogaster. *Science*, 302(5651):1727–36, Dec 2003.
- H. Goehler, M. Lalowski, U. Stelzl, S. Waelter, M. Stroedicke, U. Worm, A. Droege, K. S. Lindenberg, M. Knoblich, C. Haenig, M. Herbst, J. Suopanki, E. Scherzinger, C. Abraham, B. Bauer, R. Hasenbank, A. Fritzsche, A. H. Ludewig, K. Büsow, K. Buessow, S. H. Coleman, C.-A. Gutekunst, B. G. Landwehrmeyer, H. Lehrach, and E. E. Wanker. A protein interaction network links git1, an enhancer of huntingtin aggregation, to huntington’s disease. *Mol Cell*, 15(6):853–65, Sep 2004.
- N. Gogolla, J. J. Leblanc, K. B. Quast, T. Südhof, M. Fagiolini, and T. K. Hensch. Common circuit defect of excitatory-inhibitory balance in mouse models of autism. *J Neurodev Disord*, 1(2):172–181, Jun 2009.
- K. Goh, E. Oh, B. Kahng, and D. Kim. Betweenness centrality correlation in social networks. *Physical Review E*, 67(1):017101, 2003.
- Y. Gong, C. F. Lippa, J. Zhu, Q. Lin, and A. L. Rosso. Disruption of glutamate receptors at shank-postsynaptic platform in alzheimer’s disease. *Brain Res*, 1292:191–8, Oct 2009.
- J. A. Gorski, L. L. Gomez, J. D. Scott, and M. L. Dell’Acqua. Association of an a-kinase-anchoring protein signaling scaffold with cadherin adhesion molecules in neurons and epithelial cells. *Mol Biol Cell*, 16(8):3574–90, Aug 2005.

- C. M. Gould, F. Diella, A. Via, P. Puntervoll, C. Gemünd, S. Chabanis-Davidson, S. Michael, A. Sayadi, J. C. Bryne, C. Chica, M. Seiler, N. E. Davey, N. Haslam, R. J. Weatheritt, A. Budd, T. Hughes, J. Pas, L. Rychlewski, G. Travé, R. Aasland, M. Helmer-Citterich, R. Linding, and T. J. Gibson. Elm: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Research*, 38(Database issue):D167–80, Jan 2010.
- S. Grant. Synapse signalling complexes and networks: machines underlying cognition. *Bioessays*, 25(12):1229–1235, 2003.
- D. Greene, G. Cagney, N. Krogan, and P. Cunningham. Ensemble non-negative matrix factorization methods for clustering protein–protein interactions. *Bioinformatics*, 24(15):1722, 2008.
- R. Greene. Circuit analysis of nmdar hypofunction in the hippocampus, in vitro, and psychosis of schizophrenia. *Hippocampus*, 11(5):569–577, 2001.
- T. J. Griffin, S. P. Gygi, T. Ideker, B. Rist, J. Eng, L. Hood, and R. Aebersold. Complementary profiling of gene expression at the transcriptome and proteome levels in *saccharomyces cerevisiae*. *Molecular & cellular proteomics : MCP*, 1(4):323–33, Apr 2002.
- P. Groth, N. Pavlova, I. Kalev, S. Tonov, G. Georgiev, H.-D. Pohlenz, and B. Weiss. Phenomicdb: a new cross-species genotype/phenotype resource. *Nucleic Acids Res*, 35(Database issue):D696–9, Jan 2007.
- Z. Gu, W. Liu, and Z. Yan. Beta-amyloid impairs ampa receptor trafficking and function by reducing ca²⁺/calmodulin-dependent protein kinase ii synaptic distribution. *J Biol Chem*, 284(16):10639–49, Apr 2009.
- R. Guimerà and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, Feb 2005.
- C. Gunter. Schizophrenia: missing heritability found? *Nature Reviews Neuroscience*, Jan 2009.
- X. Guo and A. J. Hartemink. Domain-oriented edge-based alignment of protein interaction networks. *Bioinformatics*, 25(12):i240–1246, Jun 2009.
- S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mrna abundance in yeast. *Mol Cell Biol*, 19(3):1720–30, Mar 1999.
- K. F. Haas, S. L. H. Miller, D. B. Friedman, and K. Brodie. The ubiquitin-proteasome system postsynaptically regulates glutamatergic synaptic function. *Mol Cell Neurosci*, 35(1):64–75, May 2007.
- R. J. Hagerman, M. Y. Ono, and P. J. Hagerman. Recent advances in fragile x: a model for autism and neurodegeneration. *Curr Opin Psychiatry*, 18(5):490–6, Sep 2005.

- C.-G. Hahn, A. Banerjee, M. L. Macdonald, D.-S. Cho, J. Kamins, Z. Nie, K. E. Borgmann-Winter, T. Grosser, A. Pizarro, E. Ciccimaro, S. E. Arnold, H.-Y. Wang, I. A. Blair, and D. Fox. The post-synaptic density of human postmortem brain tissues: An experimental study paradigm for neuropsychiatric illnesses. *PLoS ONE*, 4(4):e5251, Apr 2009.
- M. Hall, E. Frank, G. Holmes, and B. Pfahringer. The weka data mining software: An update. *ACM SIGKDD Explorations*, Jan 2009.
- M. Hamacher, T. Hardt, A. van Hall, C. Stephan, K. Marcus, and H. E. Meyer. Inside smp proteomics: six years german human brain proteome project (hbpp) - a summary. *Proteomics*, 8(6):1118–28, Mar 2008.
- F. F. Hamdan, H. Daoud, A. Piton, J. Gauthier, S. Dobrzeniecka, M.-O. Krebs, R. Joobert, J.-C. Lacaille, A. Nadeau, J. M. Milunsky, Z. Wang, L. Carmant, L. Mottron, M. H. Beauchamp, G. A. Rouleau, and J. L. Michaud. De novo syngap1 mutations in nonsyndromic intellectual disability and autism. *Biological Psychiatry*, Jan 2011.
- J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, Jul 2004.
- J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol*, 23(7): 839–44, Jul 2005.
- D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18 Suppl 1:S145–54, Jan 2002.
- P. Harjes and E. E. Wanker. The hunt for huntingtin function: interaction partners tell many different stories. *Trends in Biochemical Sciences*, 28(8):425–33, Aug 2003.
- P. J. Harrison and D. R. Weinberger. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Mol Psychiatry*, 10(1):40–68; image 5, Jan 2005.
- P. Hart, N. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics SSC4*, 4(2), 1968.
- E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information processing letters*, Jan 2000.
- K. Hashimoto, M. Fukaya, X. Qiao, K. Sakimura, M. Watanabe, and M. Kano. Impairment of ampa receptor function in cerebellar granule cells of ataxic mutant mouse stargazer. *J Neurosci*, 19(14):6027–36, Jul 1999.

- M. K. Hayashi, C. Tang, C. Verpelli, R. Narayanan, M. H. Stearns, R.-M. Xu, H. Li, C. Sala, and Y. Hayashi. The postsynaptic density proteins homer and shank form a polymeric network structure. *Cell*, 137(1):159–171, Mar 2009.
- S. Hebbar, R. E. Hall, S. A. Demski, A. Subramanian, and J. J. Fernandes. The adult abdominal neuromuscular junction of drosophila: a model for synaptic plasticity. *J Neurobiol*, 66(10):1140–55, Sep 2006.
- N. Heintz. Gene expression nervous system atlas (gensat). *Nat Neurosci*, 7(5):483, May 2004.
- N. Henninger, R. E. Feldmann, C. D. Fütterer, C. Schrempp, M. H. Maurer, K. F. Waschke, W. Kuschinsky, and S. Schwab. Spatial learning induces predominant downregulation of cytosolic proteins in the rat hippocampus. *Genes Brain Behav*, 6(2):128–40, Mar 2007.
- H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. Intact: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue):D452–5, Jan 2004.
- K. Heupel, V. Sargsyan, J. J. Plomp, M. Rickmann, F. Varoqueaux, W. Zhang, and K. Krieglstein. Loss of transforming growth factor-beta 2 leads to impairment of central synapse function. *Neural Dev*, 3:25, Jan 2008.
- Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180–3, Jan 2002.
- C. C. Hoogenraad and F. Bradke. Control of neuronal polarity and plasticity—a renaissance for microtubules? *Trends Cell Biol*, 19(12):669–76, Dec 2009.
- D. A. Hosack, G. Dennis, B. T. Sherman, H. C. Lane, and R. A. Lempicki. Identifying biological themes within lists of genes with ease. *Genome Biol*, 4(10):R70, Jan 2003.
- P. Hotulainen and C. C. Hoogenraad. Actin in dendritic spines: connecting dynamics to function. *J Cell Biol*, 189(4):619–29, May 2010.
- S. Hrabetova and T. C. Sacktor. Bidirectional regulation of protein kinase m zeta in the maintenance of long-term potentiation and long-term depression. *J Neurosci*, 16(17):5324–33, Sep 1996.

- Y. P. Hsueh and M. Sheng. Requirement of n-terminal cysteines of psd-95 for psd-95 multimerization and ternary complex formation, but not for binding to potassium channel kv1.4. *J Biol Chem*, 274(1):532–6, Jan 1999.
- D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, Jan 2009.
- C. Hunt, L. Schenker, and M. Kennedy. Psd-95 is associated with the postsynaptic density and not with the presynaptic membrane at forebrain synapses. *The Journal of neuroscience*, 16(4):1380, 1996.
- S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. A. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. Interpro: the integrative protein signature database. *Nucleic Acids Research*, 37(Database issue):D211–5, Jan 2009.
- L. D. Hurst. The ka/ks ratio: diagnosing the form of sequence evolution. *Trends Genet*, 18(9):486, Sep 2002.
- H. Husi and S. G. Grant. Isolation of 2000-kda complexes of n-methyl-d-aspartate receptor and postsynaptic density 95 from mouse brain. *J Neurochem*, 77(1):281–91, Apr 2001.
- H. Husi, M. A. Ward, J. S. Choudhary, W. P. Blackstock, and S. G. Grant. Proteomic analysis of nmda receptor-adhesion protein signaling complexes. *Nat Neurosci*, 3(7):661–9, Jul 2000.
- J.-I. Hwang, H. S. Kim, J. R. Lee, E. Kim, S. H. Ryu, and P.-G. Suh. The interaction of phospholipase c-beta3 with shank2 regulates mglur-mediated calcium signal. *J Biol Chem*, 280(13):12467–73, Apr 2005.
- T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1: S233–40, Jan 2002.
- T. Igakura, K. Kadomatsu, O. Taguchi, H. Muramatsu, T. Kaname, T. Miyauchi, K. Yamamura, K. Arimura, and T. Muramatsu. Roles of basigin, a member of the immunoglobulin superfamily, in behavior as to an irritating odor, lymphocyte response, and blood-brain barrier. *Biochem Biophys Res Commun*, 224(1):33–6, Jul 1996.
- J. K. Inlow and L. L. Restifo. Molecular and comparative genetics of mental retardation. *Genetics*, 166(2):835–81, Feb 2004.
- M. Irie, Y. Hata, M. Takeuchi, K. Ichtchenko, A. Toyoda, K. Hirao, Y. Takai, T. W. Rosahl, and T. C. Südhof. Binding of neuroligins to psd-95. *Science*, 277(5331):1511–5, Sep 1997.

- Y. Ishihama, J. Rappsilber, J. S. Andersen, and M. Mann. Microcolumns with self-assembled particle frits for proteomics. *J Chromatogr A*, 979(1-2):233–9, Dec 2002.
- L. M. Ittner, Y. D. Ke, F. Delerue, M. Bi, A. Gladbach, J. van Eersel, H. Wölfing, B. C. Chieng, M. J. Christie, I. A. Napier, A. Eckert, M. Staufienbiel, E. Hardeman, and J. Götz. Dendritic function of tau mediates amyloid- β toxicity in alzheimer's disease mouse models. *Cell*, 142(3):387–397, Jun 2010.
- T. Iwamoto, Y. Yamada, K. Hori, Y. Watanabe, K. Sobue, and M. Inui. Differential modulation of nr1-nr2a and nr1-nr2b subtypes of nmda receptor by pdz domain-containing proteins. *J Neurochem*, 89(1):100–8, Apr 2004.
- G. R. Jackson, M. Wiedau-Pazos, T.-K. Sang, N. Wagle, C. A. Brown, S. Massachi, and D. H. Geschwind. Human wild-type tau interacts with wingless pathway components and produces neurofibrillary pathology in drosophila. *Neuron*, 34(4):509–19, May 2002.
- S. Jain and G. D. Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC bioinformatics*, 11:562, Jan 2010.
- S. Jamain, H. Quach, C. Betancur, M. Råstam, C. Colineaux, I. C. Gillberg, H. Soderstrom, B. Giros, M. Leboyer, C. Gillberg, T. Bourgeron, and P. A. R. I. S. Study. Mutations of the x-linked genes encoding neuroligins nlg3 and nlg4 are associated with autism. *Nat Genet*, 34(1):27–9, May 2003.
- S. Jamain, K. Radyushkin, K. Hammerschmidt, S. Granon, S. Boretius, F. Varoqueaux, N. Ramanantsoa, J. Gallego, A. Ronnenberg, D. Winter, J. Frahm, J. Fischer, T. Bourgeron, H. Ehrenreich, and N. Brose. Reduced social interaction and ultrasonic communication in a mouse model of monogenic heritable autism. *Proc Natl Acad Sci USA*, 105(5):1710–5, Feb 2008.
- L. Jan and Y. Jan. L-glutamate as an excitatory transmitter at the drosophila larval neuromuscular junction. *The Journal of Physiology*, 262(1):215, 1976.
- R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12(1):37, 2002.
- J. Jaworski, L. C. Kapitein, S. M. Gouveia, B. R. Dortland, P. S. Wulf, I. Grigoriev, P. Camera, S. A. Spangler, P. D. Stefano, J. Demmers, H. Krugers, P. Defilippi, A. Akhmanova, and C. C. Hoogenraad. Dynamic microtubules regulate dendritic spine morphology and synaptic plasticity. *Neuron*, 61(1):85–100, Jan 2009.
- L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. V. Mering. String 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue):D412–6, Jan 2009.
- H. Jeong, S. Mason, A. Barabasi, and Z. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

- T. Jiang and A. Keating. Avid: an integrative framework for discovering functional relationships among proteins. *BMC bioinformatics*, 6(1):136, 2005.
- A. R. Jones, C. C. Overly, and S. M. Sunkin. The allen brain atlas: 5 years and beyond. *Nat Rev Neurosci*, 10(11):821–8, Nov 2009.
- B. A. Jordan, B. D. Fernholz, M. Boussac, C. Xu, G. Grigorean, E. B. Ziff, and T. A. Neubert. Identification and verification of novel rodent postsynaptic density proteins. *Mol Cell Proteomics*, 3(9):857–71, Sep 2004.
- J. D. Jordan, E. M. Landau, and R. Iyengar. Signaling networks: the origins of cellular multitasking. *Cell*, 103(2):193–200, Oct 2000.
- G. Juhasz, J. S. Dunham, S. McKie, E. Thomas, D. Downey, D. Chase, K. Lloyd-Williams, Z. G. Toth, H. Platt, K. Mekli, A. Payton, R. Elliott, S. R. Williams, I. M. Anderson, and J. F. W. Deakin. The creb1-bdnf-ntkr2 pathway in depression: multiple gene-cognition-environment interactions. *Biological Psychiatry*, 69(8):762–71, Apr 2011.
- S. Jung, W. Jang, H. Hur, B. Hyun, and D. Han. Protein complex prediction based on mutually exclusive interactions in protein interaction network. *Genome Informatics*, 21:77–88, 2008.
- N. Kabbani, M. P. Woll, R. Levenson, J. M. Lindstrom, and J.-P. Changeux. Intracellular complexes of the beta2 subunit of the nicotinic acetylcholine receptor in brain identified by proteomics. *Proceedings of the National Academy of Sciences of the United States of America*, 104(51):20570–5, Dec 2007.
- A. Kahraman, A. Avramov, L. G. Nashev, D. Popov, R. Ternes, H.-D. Pohlenz, and B. Weiss. Phenomicdb: a multi-species genotype/phenotype database for comparative phenomics. *Bioinformatics*, 21(3):418–20, Feb 2005.
- M. Kalaev, M. Smoot, T. Ideker, and R. Sharan. Networkblast: comparative analysis of protein networks. *Bioinformatics*, 24(4):594–596, Dec 2007.
- L. V. Kalia and M. W. Salter. Interactions between src family protein tyrosine kinases and psd-95. *Neuropharmacology*, 45(6):720–8, Nov 2003.
- L. V. Kalia, J. R. Gingrich, and M. W. Salter. Src in synaptic transmission and plasticity. *Oncogene*, 23(48):8007–16, Oct 2004.
- P. J. Kammermeier. Surface clustering of metabotropic glutamate receptor 1 induced by long homer proteins. *BMC Neurosci*, 7:1, Jan 2006.
- E. R. Kandel. The molecular biology of memory storage: a dialogue between genes and synapses. *Science*, 294(5544):1030–8, Nov 2001.
- M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(Database issue):D355–60, Jan 2010.

- M. Karayiorgou, T. J. Simon, and J. A. Gogos. 22q11.2 microdeletions: linking dna structural variation to brain dysfunction and schizophrenia. *Nat Rev Neurosci*, 11 (6):402–16, Jun 2010.
- J. A. Kauer and R. C. Malenka. Synaptic plasticity and addiction. *Nat Rev Neurosci*, 8(11):844–58, Nov 2007.
- K. Kaupmann, K. Hugel, J. Heid, P. J. Flor, S. Bischoff, S. J. Mickel, G. McMaster, C. Angst, H. Bittiger, W. Froestl, and B. Bettler. Expression cloning of gaba(b) receptors uncovers similarity to metabotropic glutamate receptors. *Nature*, 386(6622): 239–46, Mar 1997.
- F. Kawasaki and R. W. Ordway. Molecular mechanisms determining conserved properties of short-term synaptic depression revealed in nsf and snap-25 conditional mutants. *Proc Natl Acad Sci USA*, 106(34):14658–63, Aug 2009.
- J. Kehoe. Glutamate activates a k⁺ conductance increase in aplysia neurons that appears to be independent of g proteins. *Neuron*, 13(3):691–702, Sep 1994.
- L. Kelly. The regulation of protein phosphorylation in synaptosomal fractions from drosophila heads: the role of cyclic adenosine monophosphate and calcium/calmodulin. *Comparative Biochemistry and Physiology Part B: . . .*, Jan 1981.
- H. Keshishian, K. Broadie, A. Chiba, and M. Bate. The drosophila neuromuscular junction: a model system for studying synaptic development and function. *Annu Rev Neurosci*, 19:545–75, Jan 1996.
- H. W. Kessels and R. Malinow. Synaptic ampa receptor plasticity and behavior. *Neuron*, 61(3):340–50, Feb 2009.
- P. Khaitovich, I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Pääbo. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, 309(5742):1850–4, Sep 2005.
- E. Kim and M. Sheng. Pdz domain proteins of synapses. *Nat Rev Neurosci*, 5(10): 771–81, Oct 2004.
- H. Kim, S. Kishikawa, A. Higgins, I. Seong, D. Donovan, Y. Shen, E. Lally, L. Weiss, J. Najm, and K. Kutsche. Disruption of neurexin 1 associated with autism spectrum disorder. *The American Journal of Human Genetics*, 82(1):199–207, 2008.
- J. Kim, S.-C. Jung, A. M. Clemens, R. S. Petralia, and D. A. Hoffman. Regulation of dendritic excitability by activity-dependent trafficking of the a-type k⁺ channel subunit kv4.2 in hippocampal neurons. *Neuron*, 54(6):933–47, Jun 2007.
- J. H. Kim, D. Liao, L. F. Lau, and R. L. Huganir. Syngap: a synaptic rasgap that associates with the psd-95/sap90 protein family. *Neuron*, 20(4):683–91, Apr 1998.
- G. Kirov, D. Rujescu, A. Ingason, D. A. Collier, M. C. O'Donovan, and M. J. Owen. Neurexin 1 (nrxn1) deletions in schizophrenia. *Schizophr Bull*, 35(5):851–4, Sep 2009a.

- G. Kirov, I. Zaharieva, L. Georgieva, V. Moskvina, I. Nikolov, S. Cichon, A. Hillmer, D. Toncheva, M. J. Owen, and M. C. O'Donovan. A genome-wide association study in 574 schizophrenia trios using dna pooling. *Mol Psychiatry*, 14(8):796–803, Aug 2009b.
- G. W. Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 10 Suppl 1:S59, Jan 2009.
- P. Klemmer, A. B. Smit, and K. W. Li. Proteomics analysis of immuno-precipitated synaptic protein complexes. *Journal of proteomics*, 72(1):82–90, Feb 2009.
- P. Klemmer, R. M. Meredith, C. D. Holmgren, O. I. Klychnikov, J. Stahl-Zeng, M. Loos, R. C. V. D. Schors, J. Wortel, S. Spijker, D. C. Rotaru, H. D. Mansvelder, A. B. Smit, and K. W. Li. Proteomics, ultrastructure and physiology of hippocampal synapses in a fragile x syndrome mouse model reveals pre-synaptic phenotype. *J Biol Chem*, May 2011.
- S. Knowles-Barley, M. Longair, and J. D. Armstrong. Braintrap: a database of 3d protein expression patterns in the drosophila brain. *Database (Oxford)*, 2010:baq005, Jan 2010.
- A. Kolodziejczyk, X. Sun, I. A. Meinertzhagen, and D. R. Nässel. Glutamate, gaba and acetylcholine signaling components in the lamina of the drosophila visual system. *PLoS ONE*, 3(5):e2110, Jan 2008.
- N. H. Komiyama, A. M. Watabe, H. J. Carlisle, K. Porter, P. Charlesworth, J. Monti, D. J. C. Strathdee, C. M. O'Carroll, S. J. Martin, R. G. M. Morris, T. J. O'Dell, and S. G. N. Grant. Syngap regulates erk/mapk signaling, synaptic plasticity, and learning in the complex with postsynaptic density 95 and nmda receptor. *Journal of Neuroscience*, 22(22):9721–32, Nov 2002.
- H. C. Kornau, L. T. Schenker, M. B. Kennedy, and P. H. Seeburg. Domain interaction between nmda receptor subunits and the postsynaptic density protein psd-95. *Science*, 269(5231):1737–40, Sep 1995.
- G. Krapivinsky, L. Krapivinsky, Y. Manasian, A. Ivanov, R. Tyzio, C. Pellegrino, Y. Ben-Ari, D. E. Clapham, and I. Medina. The nmda receptor is coupled to the erk pathway by a direct interaction between nr2b and rasgrf1. *Neuron*, 40(4):775–84, Nov 2003.
- H.-J. Kreienkamp. Scaffolding proteins at the postsynaptic density: shank as the architectural framework. *Handb Exp Pharmacol*, (186):365–80, Jan 2008.
- A. M. Krichevsky and K. S. Kosik. Neuronal rna granules: a link between rna localization and stimulation-dependent translation. *Neuron*, 32(4):683–96, Nov 2001.
- N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete,

- J. Vlasblom, S. Wu, C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. S. Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili, and J. F. Greenblatt. Global landscape of protein complexes in the yeast *saccharomyces cerevisiae*. *Nature*, 440(7084):637–43, Mar 2006.
- O. Kuchaiev and N. Przulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics (Oxford, England)*, Mar 2011.
- O. Kuchaiev, A. Stevanović, W. Hayes, and N. Pržulj. Graphcrunch 2: Software tool for network modeling, alignment and clustering. *BMC Bioinformatics*, 12(1):24, Jan 2011.
- C. Kumar and M. Mann. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Lett*, 583(11):1703–12, Jun 2009.
- B. Küppers, N. Sánchez-Soriano, J. Letzkus, G. M. Technau, and A. Prokop. In developing drosophila neurones the production of gamma-amino butyric acid is tightly regulated downstream of glutamate decarboxylase translation and can be influenced by calcium. *J Neurochem*, 84(5):939–51, Mar 2003.
- M. C. Lagomarsino, P. Jona, B. Bassetti, and H. Isambert. Hierarchy and feedback in the evolution of the escherichia coli transcription network. *Proc Natl Acad Sci USA*, 104(13):5516–20, Mar 2007.
- J. Y. Lan, V. A. Skeberdis, T. Jover, S. Y. Grooms, Y. Lin, R. C. Araneda, X. Zheng, M. V. Bennett, and R. S. Zukin. Protein kinase c modulates nmda receptor trafficking and gating. *Nat Neurosci*, 4(4):382–90, Apr 2001.
- F. Laumonier, P. C. Cuthbert, and S. G. N. Grant. The role of neuronal complexes in human x-linked brain diseases. *Am J Hum Genet*, 80(2):205–20, Feb 2007.
- G. Lebeau, L. DesGroseillers, W. Sossin, and J.-C. Lacaille. mrna binding protein staufen 1-dependent regulation of pyramidal cell spine morphology via nmda receptor-mediated synaptic plasticity. *Mol Brain*, 4:22, Jan 2011.
- J. Lee, A. Ueda, and C.-F. Wu. Pre- and post-synaptic mechanisms of synaptic strength homeostasis revealed by slowpoke and shaker k⁺ channel mutations in drosophila. *Neuroscience*, 154(4):1283–96, Jul 2008.
- A. S. Leonard, M. A. Davare, M. C. Horne, C. C. Garner, and J. W. Hell. Sap97 is associated with the alpha-amino-3-hydroxy-5-methylisoxazole-4-propionic acid receptor glur1 subunit. *J Biol Chem*, 273(31):19518–24, Jul 1998.
- D. Leonoudakis, L. R. Conti, C. M. Radeke, L. M. M. McGuire, and C. A. Vandenberg. A multiprotein trafficking complex composed of sap97, cask, veli, and mint1 is associated with inward rectifier kir2 potassium channels. *J Biol Chem*, 279(18):19051–63, Apr 2004.

- K. W. Li and C. R. Jimenez. Synapse proteomics: current status and quantitative applications. *Expert review of proteomics*, 5(2):353–60, Apr 2008.
- K. W. Li, M. P. Hornshaw, R. C. V. D. Schors, R. Watson, S. Tate, B. Casetta, C. R. Jimenez, Y. Gouwenberg, E. D. Gundelfinger, K.-H. Smalla, and A. B. Smit. Proteomics analysis of rat brain postsynaptic density. implications of the diverse protein functional groups for the integration of synaptic physiology. *J Biol Chem*, 279(2): 987–1002, Jan 2004.
- K. W. Li, P. Klemmer, and A. B. Smit. Interaction proteomics of synapse protein complexes. *Anal Bioanal Chem*, 397(8):3195–3202, Aug 2010.
- M. Li, J. Wang, and J. Chen;. A fast agglomerate algorithm for mining functional modules in protein interaction networks. *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, 1:3 – 7, 2008.
- Z. Lian, L. Wang, S. Yamaga, W. Bonds, Y. Beazer-Barclay, Y. Kluger, M. Gerstein, P. E. Newburger, N. Berliner, and S. M. Weissman. Genomic and proteomic analysis of the myeloid differentiation program. *Blood*, 98(3):513–24, Aug 2001.
- F. L. W. Liebl and D. E. Featherstone. Identification and investigation of drosophila postsynaptic density homologs. *Bioinform Biol Insights*, 2:375–387, Nov 2008.
- I. A. Lim, D. D. Hall, and J. W. Hell. Selectivity and promiscuity of the first and second pdz domains of psd-95 and synapse-associated protein 102. *J Biol Chem*, 277(24): 21697–711, Jun 2002.
- J. Lisman. A mechanism for the hebb and the anti-hebb processes underlying learning and memory. *Proc Natl Acad Sci USA*, 86(23):9574–8, Dec 1989.
- J. E. Lisman, J. T. Coyle, R. W. Green, D. C. Javitt, F. M. Benes, S. Heckers, and A. A. Grace. Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. *Trends in Neurosciences*, 31(5):234–42, May 2008.
- V. Litvak, H. Sompolinsky, I. Segev, and M. Abeles. On the transmission of rate code in long feedforward networks with excitatory-inhibitory balance. *J Neurosci*, 23(7): 3006–15, Apr 2003.
- G. Liu, H. Seiler, A. Wen, T. Zars, K. Ito, R. Wolf, M. Heisenberg, and L. Liu. Distinct memory traces for two visual features in the drosophila brain. *Nature*, 439(7076): 551–6, Feb 2006a.
- H. Liu, R. G. Sadygov, and J. R. Yates. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem*, 76(14):4193–201, Jul 2004.
- H. Liu, Z.-Z. Hu, J. Zhang, and C. Wu. Biothesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics (Oxford, England)*, 22(1):103–5, Jan 2006b.

- P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83, Jul 2003.
- H. Lu, B. Shi, G. Wu, Y. Zhang, X. Zhu, Z. Zhang, C. Liu, Y. Zhao, T. Wu, and J. Wang. Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochemical and Biophysical Research Communications*, 345(1):302–309, Jun 2006.
- Z. Lubovac, J. Gamalielsson, and B. Olsson. Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins*, 64(4):948–959, Jun 2006.
- C. Lüscher, L. Y. Jan, M. Stoffel, R. C. Malenka, and R. A. Nicoll. G protein-coupled inwardly rectifying k⁺ channels (girsks) mediate postsynaptic but not presynaptic transmitter actions in hippocampal neurons. *Neuron*, 19(3):687–95, Sep 1997.
- C. Lüscher, R. A. Nicoll, R. C. Malenka, and D. Muller. Synaptic plasticity and dynamic modulation of the postsynaptic membrane. *Nat Neurosci*, 3(6):545–50, Jun 2000.
- A. Lüthi, R. Chittajallu, F. Duprat, M. J. Palmer, T. A. Benke, F. L. Kidd, J. M. Henley, J. T. Isaac, and G. L. Collingridge. Hippocampal ltd expression involves a pool of ampars regulated by the nsf-glu2 interaction. *Neuron*, 24(2):389–99, Oct 1999.
- X.-M. Ma, D. D. Kiraly, E. D. Gaier, Y. Wang, E.-J. Kim, E. S. Levine, B. A. Eipper, and R. E. Mains. Kalirin-7 is required for synaptic structure and function. *J Neurosci*, 28(47):12368–82, Nov 2008.
- A. Ma’ayan, S. L. Jenkins, S. Neves, A. Hasseldine, E. Grace, B. Dubin-Thaler, N. J. Eungdamrong, G. Weng, P. T. Ram, J. J. Rice, A. Kershenbaum, G. A. Stolovitzky, R. D. Blitzer, and R. Iyengar. Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science*, 309(5737):1078–83, Aug 2005.
- A. F. MacAskill, T. A. Atkin, and J. T. Kittler. Mitochondrial trafficking and the provision of energy and calcium buffering at excitatory synapses. *Eur J Neurosci*, 32(2):231–40, Jul 2010.
- S. Maere, K. Heymans, and M. Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–9, Aug 2005.
- S. Magdaleno, P. Jensen, C. L. Brumwell, A. Seal, K. Lehman, A. Asbury, T. Cheung, T. Cornelius, D. M. Batten, C. Eden, S. M. Norland, D. S. Rice, N. Dosooye, S. Shakya, P. Mehta, and T. Curran. Bgem: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system. *PLoS Biol*, 4(4):e86, Apr 2006.
- M. Magrane and U. Consortium. Uniprot knowledgebase: a hub of integrated protein data. *Database (Oxford)*, 2011:bar009, Jan 2011.

- M. A. Mahdavi and Y.-H. Lin. False positive reduction in protein-protein interaction predictions using gene ontology annotations. *BMC Bioinformatics*, 8:262, Jan 2007.
- B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21, Nov 2008.
- R. C. Malenka. Postsynaptic factors control the duration of synaptic enhancement in area ca1 of the hippocampus. *Neuron*, 6(1):53–60, Jan 1991.
- R. C. Malenka and R. A. Nicoll. Nmda-receptor-dependent synaptic plasticity: multiple forms and mechanisms. *Trends Neurosci*, 16(12):521–7, Dec 1993.
- I. A. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos. Growing functional modules from a seed protein via integration of protein interaction and gene expression data. *BMC Bioinformatics*, 8:408, Jan 2007.
- V. Martha, Z. Liu, L. Guo, Z. Su, Y. Ye, H. Fang, D. Ding, W. Tong, and X. Xu. Constructing a robust protein-protein interaction network by integrating multiple public databases. *BMC Bioinformatics*, 12(Suppl 10):S7, 2011.
- S. J. Martin, P. D. Grimwood, and R. G. Morris. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annu Rev Neurosci*, 23:649–711, Jan 2000.
- D. Mathew, L. S. Gramates, M. Packard, U. Thomas, D. Bilder, N. Perrimon, M. Gorczyca, and V. Budnik. Recruitment of scribble to the synaptic scaffolding complex requires guk-holder, a novel dlg binding protein. *Curr Biol*, 12(7):531–9, Apr 2002.
- S. Matos, J. P. Arrais, J. Maia-Rodrigues, and J. L. Oliveira. Concept-based query expansion for retrieving gene related publications from medline. *BMC Bioinformatics*, 11:212, Jan 2010.
- L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, 37(Database issue):D619–22, Jan 2009.
- M. G. McKernan and P. Shinnick-Gallagher. Fear conditioning induces a lasting potentiation of synaptic currents in vitro. *Nature*, 390(6660):607–11, Dec 1997.
- V. McKusick. Mendelian inheritance in man and its online version, omim. *Am J Hum Genet*, 80(4):588, 2007.
- V. McKusick and J. Amberger. The morbid anatomy of the human genome: chromosomal location of mutations causing disease. *J Med Genet*, 31(4):265–279, 1994.
- K. McNair, C. H. Davies, and S. R. Cobb. Plasticity-related regulation of the hippocampal proteome. *Eur J Neurosci*, 23(2):575–80, Jan 2006.
- C. V. Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.

- C. V. Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krüger, B. Snel, and P. Bork. String 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research*, 35(Database issue):D358–62, Jan 2007.
- H. Mi, N. Guo, A. Kejariwal, and P. Thomas. Panther version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Research*, 35(suppl 1):D247, 2006.
- M. Migaud, P. Charlesworth, M. Dempster, L. C. Webster, A. M. Watabe, M. Makhinson, Y. He, M. F. Ramsay, R. G. Morris, J. H. Morrison, T. J. O'Dell, and S. G. Grant. Enhanced long-term potentiation and impaired learning in mice with mutant postsynaptic density-95 protein. *Nature*, 396(6710):433–9, Dec 1998.
- T. Mijalski, A. Harder, T. Halder, M. Kersten, M. Horsch, T. M. Strom, H. V. Liebscher, F. Lottspeich, M. H. de Angelis, and J. Beckers. Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues. *Proc Natl Acad Sci USA*, 102(24):8621–6, Jun 2005.
- T. Milenković, J. Lai, and N. Pržulj. Graphcrunch: A tool for large network analyses. *BMC Bioinformatics*, 9(1):70, Jan 2008.
- J. K. Millar, J. C. Wilson-Annan, S. Anderson, S. Christie, M. S. Taylor, C. A. Semple, R. S. Devon, D. M. S. Clair, W. J. Muir, D. H. Blackwood, and D. J. Porteous. Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum Mol Genet*, 9(9):1415–23, May 2000.
- J. K. Millar, B. S. Pickard, S. Mackie, R. James, S. Christie, S. R. Buchanan, M. P. Malloy, J. E. Chubb, E. Huston, G. S. Baillie, P. A. Thomson, E. V. Hill, N. J. Brandon, J.-C. Rain, L. M. Camargo, P. J. Whiting, M. D. Houslay, D. H. R. Blackwood, W. J. Muir, and D. J. Porteous. Disc1 and pde4b are interacting genetic factors in schizophrenia that regulate camp signaling. *Science*, 310(5751):1187–91, Nov 2005.
- L. D. Miller, J. J. Petroszino, and J. A. Connor. G protein-coupled receptors mediate a fast excitatory postsynaptic current in ca3 pyramidal neurons in hippocampal slices. *J Neurosci*, 15(12):8320–30, Dec 1995.
- R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–7, Oct 2002.
- R. Moessner, C. R. Marshall, J. S. Sutcliffe, J. Skaug, D. Pinto, J. Vincent, L. Zwaigenbaum, B. Fernandez, W. Roberts, P. Szatmari, and S. W. Scherer. Contribution of shank3 mutations to autism spectrum disorder. *Am J Hum Genet*, 81(6):1289–97, Dec 2007.
- J. M. Montgomery, P. L. Zamorano, and C. C. Garner. Maguks in synapse assembly and function: an emerging view. *Cell Mol Life Sci*, 61(7-8):911–29, Apr 2004.

- M. Morciano, J. Burré, C. Corvey, M. Karas, H. Zimmermann, and W. Volkandt. Immunoisolation of two synaptic vesicle pools from synaptosomes: a proteomics analysis. *J Neurochem*, 95(6):1732–45, Dec 2005.
- M. Morciano, T. Beckhaus, M. Karas, H. Zimmermann, and W. Volkandt. The proteome of the presynaptic active zone: from docked synaptic vesicles to adhesion molecules and maxi-channels. *J Neurochem*, 108(3):662–75, Feb 2009.
- M. M. Moreau, N. Piguel, T. Papouin, M. Koehl, C. M. Durand, M. E. Rubio, F. Loll, E. M. Richard, C. Mazzocco, C. Racca, S. H. R. Oliet, D. N. Abrous, M. Montcouquiol, and N. Sans. The planar polarity protein scribble1 is essential for neuronal plasticity and brain function. *J Neurosci*, 30(29):9738–52, Jul 2010.
- A. Moressis, A. R. Friedrich, E. Pavlopoulos, R. L. Davis, and E. M. C. Skoulakis. A dual role for the adaptor protein drk in drosophila olfactory learning and memory. *J Neurosci*, 29(8):2611–25, Feb 2009.
- X. Morin, R. Daneman, M. Zavortink, and W. Chia. A protein trap strategy to detect gfp-tagged proteins expressed from their endogenous loci in drosophila. *Proc Natl Acad Sci USA*, 98(26):15050–5, Dec 2001.
- R. G. Morris and U. Frey. Hippocampal synaptic plasticity: role in spatial learning or the automatic recording of attended experience? *Philos Trans R Soc Lond, B, Biol Sci*, 352(1360):1489–503, Oct 1997.
- R. G. Morris, S. Davis, and S. P. Butcher. Hippocampal synaptic plasticity and nmda receptors: a role in information storage? *Philos Trans R Soc Lond, B, Biol Sci*, 329(1253):187–204, Aug 1990.
- R. M. Mulkey and R. C. Malenka. Mechanisms underlying induction of homosynaptic long-term depression in area ca1 of the hippocampus. *Neuron*, 9(5):967–75, Nov 1992.
- R. P. Munton, R. Tweedie-Cullen, M. Livingstone-Zatchej, F. Weinandy, M. Waidelich, D. Longo, P. Gehrig, F. Potthast, D. Rutishauser, B. Gerrits, C. Panse, R. Schlappbach, and I. M. Mansuy. Qualitative and quantitative analyses of protein phosphorylation in naive and stimulated mouse synaptosomal preparations. *Mol Cell Proteomics*, 6(2):283–93, Feb 2007.
- K. Nakazawa, T. J. McHugh, M. A. Wilson, and S. Tonegawa. Nmda receptors, place cells and hippocampal spatial memory. *Nat Rev Neurosci*, 5(5):361–72, May 2004.
- S. J. Neal, S. Karunanithi, A. Best, A. K.-C. So, R. M. Tanguay, H. L. Atwood, and J. T. Westwood. Thermoprotection of synaptic transmission in a drosophila heat shock factor mutant is accompanied by increased expression of hsp83 and dnaj-1. *Physiol Genomics*, 25(3):493–501, May 2006.
- R. Nehring, E. Wischmeyer, F. Döring, R. Veh, M. Sheng, and A. Karschin. Neuronal inwardly rectifying k⁺ channels differentially couple to pdz proteins of the psd-95/sap90 family. *The Journal of Neuroscience*, 20(1):156, 2000.

- K. Neuser, T. Triphan, M. Mronz, B. Poeck, and R. Strauss. Analysis of a spatial orientation memory in drosophila. *Nature*, 453(7199):1244–1247, Jun 2008.
- M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):1–15, Feb 2004.
- T. M. Newpher and M. D. Ehlers. Glutamate receptor dynamics in dendritic microdomains. *Neuron*, 58(4):472–97, May 2008.
- L. Ng, A. Bernard, C. Lau, C. C. Overly, H.-W. Dong, C. Kuan, S. Pathak, S. M. Sunkin, C. Dang, J. W. Bohland, H. Bokil, P. P. Mitra, L. Puelles, J. Hohmann, D. J. Anderson, E. S. Lein, A. R. Jones, and M. Hawrylycz. An anatomic gene expression atlas of the adult mouse brain. *Nat Neurosci*, 12(3):356–62, Mar 2009.
- R. A. Nicoll, J. A. Kauer, and R. C. Malenka. The current excitement in long-term potentiation. *Neuron*, 1(2):97–103, Apr 1988.
- A. Nishimune, J. T. Isaac, E. Molnar, J. Noel, S. R. Nash, M. Tagaya, G. L. Collingridge, S. Nakanishi, and J. M. Henley. Nsf binding to glur2 regulates synaptic transmission. *Neuron*, 21(1):87–97, Jul 1998.
- W. S. Noble. How does multiple testing correction work? *Nature Biotechnology*, 27(12):1135–1137, Dec 2009.
- J. Noel, G. S. Ralph, L. Pickard, J. Williams, E. Molnar, J. B. Uney, G. L. Collingridge, and J. M. Henley. Surface expression of ampa receptors in hippocampal neurons is regulated by an nsf-dependent mechanism. *Neuron*, 23(2):365–76, Jun 1999.
- C. Nourry, S. G. N. Grant, and J.-P. Borg. Pd domain proteins: plug and play! *Sci Signal*, 2003(179):RE7, Apr 2003.
- P. H. O’Farrell. High resolution two-dimensional electrophoresis of proteins. *J Biol Chem*, 250(10):4007–21, May 1975.
- J. P. Olivier, T. Raabe, M. Henkemeyer, B. Dickson, G. Mbamalu, B. Margolis, J. Schlessinger, E. Hafen, and T. Pawson. A drosophila sh2-sh3 adaptor protein implicated in coupling the sevenless tyrosine kinase to an activator of ras guanine nucleotide exchange, sos. *Cell*, 73(1):179–91, Apr 1993.
- S.-E. Ong and M. Mann. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol*, 1(5):252–62, Oct 2005.
- S.-E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann. Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics : MCP*, 1(5):376–86, May 2002.
- P. Osten, S. Srivastava, G. J. Inman, F. S. Vilim, L. Khatri, L. M. Lee, B. A. States, S. Einheber, T. A. Milner, P. I. Hanson, and E. B. Ziff. The ampa receptor glur2 c terminus can mediate a reversible, atp-dependent interaction with nsf and alpha- and beta-snaps. *Neuron*, 21(1):99–110, Jul 1998.

- G. Ostlund, T. Schmitt, K. Forslund, T. Köstler, D. N. Messina, S. Roopra, O. Frings, and E. L. L. Sonnhammer. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38(Database issue):D196–203, Jan 2010.
- D. Park, S. Lee, D. Bolser, M. Schroeder, M. Lappe, D. Oh, and J. Bhak. Comparative interactomics analysis of protein family interaction networks using psimap (protein structural interactome map). *Bioinformatics*, 21(15):3234, 2005.
- M. Park, E. C. Penick, J. G. Edwards, J. A. Kauer, and M. D. Ehlers. Recycling endosomes supply ampa receptors for ltp. *Science*, 305(5692):1972–5, Sep 2004.
- H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwani, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. Arrayexpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37(Database issue):D868–72, Jan 2009.
- M. L. Parmentier, J. P. Pin, J. Bockaert, and Y. Grau. Cloning and functional expression of a drosophila metabotropic glutamate receptor expressed in the embryonic CNS. *J Neurosci*, 16(21):6687–94, Nov 1996.
- E. Pastalkova, P. Serrano, D. Pinkhasova, E. Wallace, A. A. Fenton, and T. C. Sacktor. Storage of spatial information by the maintenance mechanism of ltp. *Science*, 313(5790):1141–4, Aug 2006.
- R. Pastor-Satorras, E. Smith, and R. Sole. Evolving protein interaction networks through gene duplication. *Journal of theoretical biology*, 222(2):199–210, 2003.
- J. Paulo, W. Brucker, and E. Hawrot. Proteomic analysis of an 7 nicotinic acetylcholine receptor interactome. *J Proteome Res*, Jan 2009.
- R. Pellicciari and G. Costantino. Metabotropic g-protein-coupled glutamate receptors as therapeutic targets. *Curr Opin Chem Biol*, 3(4):433–40, Aug 1999.
- J. Peng, M. J. Kim, D. Cheng, D. M. Duong, S. P. Gygi, and M. Sheng. Semiquantitative proteomic analysis of rat forebrain postsynaptic density fractions by mass spectrometry. *J Biol Chem*, 279(20):21003–11, May 2004.
- P. Penzes, R. C. Johnson, R. Sattler, X. Zhang, R. L. Huganir, V. Kambampati, R. E. Mains, and B. A. Eipper. The neuronal rho-gef kalirin-7 interacts with pdz domain-containing proteins and regulates dendritic morphogenesis. *Neuron*, 29(1):229–42, Jan 2001.
- S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjana, B. Muthusamy, T. K. B. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. N. Shivashankar, B. P. Rashmi, M. A. Ramya,

- Z. Zhao, K. N. Chandrika, N. Padma, H. C. Harsha, A. J. Yatish, M. P. Kavitha, M. Menezes, D. R. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. K. Anand, V. Madavan, A. Joseph, G. W. Wong, W. P. Schiemann, S. N. Constantinescu, L. Huang, R. Khosravi-Far, H. Steen, M. Tewari, S. Ghaffari, G. C. Blobe, C. V. Dang, J. G. N. Garcia, J. Pevsner, O. N. Jensen, P. Roepstorff, K. S. Deshpande, A. M. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–71, Oct 2003.
- M. S. Perkinson, J. K. Ip, G. L. Wood, A. J. Crossthwaite, and R. J. Williams. Phosphatidylinositol 3-kinase is a central mediator of nmda receptor signalling to map kinase (erk1/2), akt/pkb and creb in striatal neurones. *J Neurochem*, 80(2):239–54, Jan 2002.
- S. Peron, M. A. Zordan, A. Magnabosco, C. Reggiani, and A. Megighian. From action potential to contraction: neural control and excitation-contraction coupling in larval muscles of drosophila. *Comp Biochem Physiol, Part A Mol Integr Physiol*, 154(2):173–83, Oct 2009.
- M. Persico, A. Ceol, C. Gavrilu, R. Hoffmann, A. Florio, and G. Cesareni. Homomint: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC bioinformatics*, 6 Suppl 4:S21, Dec 2005.
- K. E. Personius and R. J. Balice-Gordon. Activity-dependent synaptic plasticity: insights from neuromuscular junctions. *Neuroscientist*, 8(5):414–22, Oct 2002.
- R. S. Petralia, N. Sans, Y.-X. Wang, and R. J. Wenthold. Ontogeny of postsynaptic density proteins at glutamatergic synapses. *Mol Cell Neurosci*, 29(3):436–52, Jul 2005.
- B. E. Pfeiffer and K. M. Huber. The state of synapses in fragile x syndrome. *Neuroscientist*, 15(5):549–67, Oct 2009.
- G. R. Phillips, L. Florens, H. Tanaka, Z. Z. Khaing, L. Fidler, J. R. Yates, and D. R. Colman. Proteomic comparison of two fractions derived from the transsynaptic scaffold. *J Neurosci Res*, 81(6):762–75, Sep 2005.
- G. Piccoli, C. Verpelli, N. Tonna, S. Romorini, M. Alessio, A. C. Nairn, A. Bachi, and C. Sala. Proteomic analysis of activity-dependent synaptic plasticity in hippocampal neurons. *J Proteome Res*, 6(8):3203–15, Aug 2007.
- V. Pillet, M. Zehnder, A. K. Seewald, A.-L. Veuthey, and J. Petrak. Gpsdb: a new database for synonyms expansion of gene and protein names. *Bioinformatics*, 21(8):1743–4, Apr 2005.
- D. Pinto, A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, J. Almeida, E. Bacchelli, G. D. Bader, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bölte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, S. E. Bryson, A. R. Carson, G. Casallo, J. Casey,

- B. H. Y. Chung, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, C. Cytrynbaum, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, A. Green, J. Green, S. J. Guter, H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Iglizzi, C. Kim, S. M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. L. Couteur, B. L. Leventhal, A. C. Lionel, X.-Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles, M. Pilorge, J. Piven, C. P. Ponting, D. J. Posey, A. Poustka, F. Poustka, A. Prasad, J. Ragoussis, K. Renshaw, J. Rickaby, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, L. J. Bierut, J. P. Rice, J. Salt, K. Sansom, D. Sato, R. Segurado, A. F. Sequeira, L. Senman, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, O. Stein, N. Sykes, V. Stoppioni, C. Strawbridge, R. Tancredi, K. Tansey, B. Thiruvahindrapduram, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. V. Engeland, J. B. Vincent, F. Volkmar, S. Wallace, K. Wang, Z. Wang, T. H. Wassink, C. Webber, R. Weksberg, K. Wing, K. Wittemeyer, S. Wood, J. Wu, B. L. Yaspan, D. Zurawiecki, L. Zwaigenbaum, J. D. Buxbaum, R. M. Cantor, E. H. Cook, H. Coon, M. L. Cuccaro, B. Devlin, S. Ennis, L. Gallagher, D. H. Geschwind, M. Gill, J. L. Haines, J. Hallmayer, J. Miller, A. P. Monaco, J. I. N. Jr, A. D. Paterson, M. A. Pericak-Vance, G. D. Schellenberg, P. Szatmari, A. M. Vicente, V. J. Vieland, E. M. Wijsman, S. W. Scherer, J. S. Sutcliffe, and C. Betancur. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, Jun 2010.
- A. J. Pocklington, M. Cumiskey, J. D. Armstrong, and S. G. N. Grant. The proteomes of neurotransmitter receptor complexes form modular networks with distributed functionality underlying plasticity and behaviour. *Mol Syst Biol*, 2:1–14, Jan 2006.
- F. Policies, D. Policy, and I. Subscribers. Seeking a new biology through text mining. *Cell*, Jan 2008.
- K. Porter, N. H. Komiyama, T. Vitalis, P. C. Kind, and S. G. N. Grant. Differential expression of two nmda receptor interacting proteins, psd-95 and syngap during mouse development. *Eur J Neurosci*, 21(2):351–62, Jan 2005.
- E. Prifti, J. Zucker, K. Clement, and C. Henegar. Funnet: an integrative tool for exploring transcriptional interactions. *Bioinformatics*, 24(22):2636, 2008.
- N. Przulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–83, Jan 2007.
- N. Przulj, D. A. Wigle, and I. Jurisica. Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–8, Feb 2004.

- O. Puig. The tandem affinity purification (tap) method: A general procedure of protein complex purification. *Methods*, 24(3):218–229, Jul 2001.
- S. Purcell, N. Wray, J. Stone, P. Visscher, M. O'Donovan, P. Sullivan, P. Sklar, D. Ruderfer, A. McQuillin, and D. Morris. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.
- A. Quitsch, K. Berhörster, C. W. Liew, D. Richter, and H.-J. Kreienkamp. Postsynaptic shank antagonizes dendrite branching induced by the leucine-rich repeat protein densin-180. *J Neurosci*, 25(2):479–87, Jan 2005.
- F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proc Natl Acad Sci USA*, 101(9):2658–63, Mar 2004.
- J. C. Rain, L. Selig, H. D. Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schächter, Y. Chemama, A. Labigne, and P. Legrain. The protein-protein interaction map of helicobacter pylori. *Nature*, 409(6817):211–5, Jan 2001.
- C. Ramarao, S. Acharya, K. Krishnan, and U. Kenkare. High affinity uptake of l-glutamate and gamma-aminobutyric acid in drosophila melanogaster. *Journal of Biosciences*, 11(1):119–135, 1987.
- E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–5, Aug 2002.
- S. Razick, G. Magklaras, and I. M. Donaldson. irefindex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1):405, Jan 2008.
- A. Rebsam and C. A. Mason. Cadherins as matchmakers. *Neuron*, 71(4):566–8, Aug 2011.
- R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–54, Nov 2006.
- J. S. Rees, N. Lowe, I. M. Armean, J. Roote, G. Johnson, E. Drummond, H. Spriggs, E. Ryder, S. Russell, D. S. Johnston, and K. S. Lilley. In vivo analysis of proteomes and interactomes using parallel affinity capture (ipac) coupled to mass spectrometry. *Molecular & cellular proteomics : MCP*, Mar 2011.
- G. Riedel, W. Wetzell, and K. G. Reymann. Comparing the role of metabotropic glutamate receptors in long-term potentiation and in learning and memory. *Prog Neuropsychopharmacol Biol Psychiatry*, 20(5):761–89, Jul 1996.

- G. Riedel, B. Platt, and J. Micheau. Glutamate receptor function in learning and memory. *Behavioural Brain Research*, 140(1-2):1–47, 2003.
- G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–2, Oct 1999.
- M. Ringwald, G. Davis, A. Smith, L. Trepanier, D. Begley, J. Richardson, and J. Eppig. The mouse gene expression database gxd. *Semin Cell Dev Biol*, 8(5):489–97, Oct 1997.
- M. Ringwald, J. T. Eppig, D. A. Begley, J. P. Corradi, I. J. McCright, T. F. Hayamizu, D. P. Hill, J. A. Kadin, and J. E. Richardson. The mouse gene expression database (gxd). *Nucleic Acids Research*, 29(1):98–101, Jan 2001.
- I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics (Oxford, England)*, 23(4):401–7, Feb 2007.
- P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 83(5):610–5, Nov 2008.
- K. G. L. Roch, J. R. Johnson, L. Florens, Y. Zhou, A. Santrosyan, M. Grainger, S. F. Yan, K. C. Williamson, A. A. Holder, D. J. Carucci, J. R. Yates, and E. A. Winzeler. Global analysis of transcript and protein levels across the plasmodium falciparum life cycle. *Genome Res*, 14(11):2308–18, Nov 2004.
- P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlett-Jones, F. He, A. Jacobson, and D. J. Pappin. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics : MCP*, 3(12):1154–69, Dec 2004.
- J. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. Berriz, F. Gibbons, M. Dreze, and N. Ayivi-Guedehoussou. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.
- C. Ruiz-Cañada and V. Budnik. Synaptic cytoskeleton at the neuromuscular junction. *Int Rev Neurobiol*, 75:217–36, Jan 2006a.
- C. Ruiz-Cañada and V. Budnik. Introduction on the use of the drosophila embryonic/larval neuromuscular junction as a model system to study synapse development and function, and a brief summary of pathfinding and target recognition. *Int Rev Neurobiol*, 75:1–31, Jan 2006b.
- C. Ruiz-Cañada, Y. H. Koh, V. Budnik, and F. J. Tejedor. Dlg differentially localizes shaker k⁺-channels in the central nervous system and retina of drosophila. *J Neurochem*, 82(6):1490–501, Sep 2002.

- A. M. Rush, J. Wu, M. J. Rowan, and R. Anwyl. Group i metabotropic glutamate receptor (mglur)-dependent long-term depression mediated via p38 mitogen-activated protein kinase is inhibited by previous high-frequency stimulation and activation of mglurs and protein kinase c in the rat dentate gyrus in vitro. *J Neurosci*, 22(14):6121–8, Jul 2002.
- T. J. Ryan and S. G. N. Grant. The origin and evolution of synapses. *Nat Rev Neurosci*, 10(10):701–12, Oct 2009.
- T. J. Ryan, R. D. Emes, S. G. Grant, and N. H. Komiyama. Evolution of nmda receptor cytoplasmic interaction domains: implications for organisation of synaptic signalling complexes. *BMC Neurosci*, 9(1):6, Jan 2008.
- E. Ryder, J. Rees, and D. Johnston. Genome-wide mapping and characterisation of protein expression and interaction in drosophila melanogaster, using a hybrid piggybac/p-element yfp gene trap system with tandem affinity tags. *Poster*, pages 1–1, May 2007.
- E. Ryder, H. Spriggs, E. Drummond, D. S. Johnston, and S. Russell. The flannotator—a gene and protein expression annotation tool for drosophila melanogaster. *Bioinformatics*, 25(4):548–549, Jan 2009.
- O. Sakarya, K. A. Armstrong, M. Adamska, M. Adamski, I.-F. Wang, B. Tidor, B. M. Degnan, T. H. Oakley, and K. S. Kosik. A post-synaptic scaffold at the origin of the animal kingdom. *PLoS ONE*, 2(6):e506, Jan 2007.
- M. W. Salter and L. V. Kalia. Src kinases: a hub for nmda receptor regulation. *Nat Rev Neurosci*, 5(4):317–28, Apr 2004.
- L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):D449–51, Jan 2004.
- N. Sans, R. S. Petralia, Y. X. Wang, J. Blahos, J. W. Hell, and R. J. Wenthold. A developmental change in nmda receptor-associated proteins at hippocampal synapses. *J Neurosci*, 20(3):1260–71, Feb 2000.
- Y. Sasaki, J. McNaught, and S. Ananiadou. The value of an in-domain lexicon in genomics qa. *J Bioinform Comput Biol*, 8(1):147–61, Feb 2010.
- R. Scherzer-Attali, R. Pellarin, M. Convertino, A. Frydman-Marom, N. Egoz-Matia, S. Peled, M. Levy-Sakin, D. E. Shalev, A. Caffisch, E. Gazit, and D. Segal. Complete phenotypic recovery of an alzheimer’s disease model by a quinone-tryptophan hybrid aggregation inhibitor. *PLoS ONE*, 5(6):e11101, Jan 2010.
- E. Schnell, M. Sizemore, S. Karimzadegan, L. Chen, D. S. Bredt, and R. A. Nicoll. Direct interactions between psd-95 and stargazin control synaptic ampa receptor number. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21):13902–7, Oct 2002.

- S. Schoch and E. D. Gundelfinger. Molecular organization of the presynaptic active zone. *Cell Tissue Res*, 326(2):379–91, Nov 2006.
- S. P. Schrimpf, V. Meskenaite, E. Brunner, D. Rutishauser, P. Walther, J. Eng, R. Aebersold, and P. Sonderegger. Proteomic analysis of synaptosomes using isotope-coded affinity tags and mass spectrometry. *Proteomics*, 5(10):2531–41, Jul 2005.
- M. J. Schuemie, N. Kang, M. L. Hekkelman, and J. A. Kors. Genee: gene and protein query expansion with disambiguation. *Bioinformatics*, 26(1):147–8, Jan 2010.
- C. M. Schuster. Experience-dependent potentiation of larval neuromuscular synapses. *Int Rev Neurobiol*, 75:307–22, Jan 2006a.
- C. M. Schuster. Glutamatergic synapses of drosophila neuromuscular junctions: a high-resolution model for the analysis of experience-dependent potentiation. *Cell Tissue Res*, 326(2):287–99, Nov 2006b.
- C. M. Schuster, A. Ultsch, P. Schloss, J. A. Cox, B. Schmitt, and H. Betz. Molecular cloning of an invertebrate glutamate receptor subunit expressed in drosophila muscle. *Science*, 254(5028):112–4, Oct 1991.
- M. Schutkowski, U. Reineke, and U. Reimer. Peptide arrays for kinase profiling. *Chembiochem*, 6(3):513–21, Mar 2005.
- J. Schwenk, N. Harmel, G. Zolles, W. Bildl, A. Kulik, B. Heimrich, O. Chisaka, P. Jonas, U. Schulte, B. Fakler, and N. Klöcker. Functional proteomics identify cornichon proteins as auxiliary subunits of ampa receptors. *Science*, 323(5919):1313–9, Mar 2009.
- B. Schwikowski, P. Uetz, and S. Fields. A network of protein–protein interactions in yeast. *Nat Biotechnol*, 18(12):1257–1261, 2000.
- E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19 Suppl 1:i264–71, Jan 2003.
- F. Selimi, I. M. Cristea, E. Heller, B. T. Chait, and N. Heintz. Proteomic studies of a single cns synapse type: the parallel fiber/purkinje cell synapse. *PLoS Biol*, 7(4):e83, Apr 2009.
- G. Serban, Z. Kouchi, L. Baki, A. Georgakopoulos, C. M. Litterst, J. Shioi, and N. K. Robakis. Cadherins mediate both the association between ps1 and beta-catenin and the effects of ps1 on beta-catenin stability. *J Biol Chem*, 280(43):36007–12, Oct 2005.
- G. M. Shankar, B. L. Bloodgood, M. Townsend, D. M. Walsh, D. J. Selkoe, and B. L. Sabatini. Natural oligomers of the alzheimer amyloid-beta protein induce reversible synapse loss by modulating an nmda-type glutamate receptor-dependent signaling pathway. *J Neurosci*, 27(11):2866–75, Mar 2007.

- P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, Nov 2003.
- R. Sharan, S. Suthram, R. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences*, 102(6):1974–1979, 2005.
- K. Z. Shen and S. W. Johnson. A slow excitatory postsynaptic current mediated by g-protein-coupled metabotropic glutamate receptors in rat ventral tegmental dopamine neurons. *Eur J Neurosci*, 9(1):48–54, Jan 1997.
- S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1):64–8, May 2002.
- M. Sheng and C. C. Hoogenraad. The postsynaptic architecture of excitatory synapses: a more quantitative view. *Annu Rev Biochem*, 76:823–47, Jan 2007.
- M. Sheng and M. J. Kim. Postsynaptic signaling and plasticity mechanisms. *Science*, 298(5594):776–80, Oct 2002.
- J. D. Shepherd, G. Rumbaugh, J. Wu, S. Chowdhury, N. Plath, D. Kuhl, R. L. Huganir, and P. F. Worley. Arc/arg3.1 mediates homeostatic synaptic scaling of ampa receptors. *Neuron*, 52(3):475–84, Nov 2006.
- M. Shimoyama, J. R. Smith, T. Hayman, S. Laulederkind, T. Lowry, R. Nigam, V. Petri, S.-J. Wang, M. Dwinell, H. Jacob, and R. Team. Rgd: a comparative genomics platform. *Human Genomics*, 5(2):124–9, Jan 2011.
- T. I. Simpson, J. D. Armstrong, and A. P. Jarman. Merged consensus clustering to assess and improve class discovery with microarray data. *BMC Bioinformatics*, 11(1):590, Dec 2010.
- R. Singh, J. Xu, and B. Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. *Research in computational molecular biology*, pages 16–31, 2007.
- E. M. C. Skoulakis and S. Grammenoudi. Dunces and da vincis: the genetics of learning and memory in drosophila. *Cell Mol Life Sci*, 63(9):975–88, May 2006.
- B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, O. Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–5, Nov 2007a.
- C. L. Smith and J. T. Eppig. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip Rev Syst Biol Med*, 1(3):390–9, Jan 2009.

- C. L. Smith, C.-A. W. Goldsmith, and J. T. Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*, 6(1):R7, Jan 2005.
- C. M. Smith, J. H. Finger, T. F. Hayamizu, I. J. McCright, J. T. Eppig, J. A. Kadin, J. E. Richardson, and M. Ringwald. The mouse gene expression database (gxd): 2007 update. *Nucleic Acids Research*, 35(Database issue):D618–23, Jan 2007b.
- O. Sorokina, A. Sorokin, and J. D. Armstrong. Towards a quantitative model of the post-synaptic proteome. *Mol Biosyst*, Aug 2011.
- V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12123, Oct 2003.
- C. A. Stanyon, G. Liu, B. A. Mangiola, N. Patel, L. Giot, B. Kuang, H. Zhang, J. Zhong, and R. L. Finley. A drosophila protein-interaction map centered on cell-cycle regulators. *Genome Biol*, 5(12):R96, Jan 2004.
- C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–9, Jan 2006.
- L. Stein, P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spieth. Wormbase: network access to the genome and biology of *caenorhabditis elegans*. *Nucleic Acids Research*, 29(1):82–6, Jan 2001.
- C. F. Stevens and J. Sullivan. Synaptic plasticity. *Curr Biol*, 8(5):R151–3, Feb 1998.
- O. Steward and P. M. Falk. Selective localization of polyribosomes beneath developing synapses: a quantitative analysis of the relationships between polyribosomes and developing synapses in the hippocampus and dentate gyrus. *J Comp Neurol*, 314(3):545–57, Dec 1991.
- J. Stone, M. O'Donovan, H. Gurling, G. Kirov, D. Blackwood, A. Corvin, N. Craddock, M. Gill, C. Hultman, and P. Lichtenstein. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210):237–241, 2008.
- N. J. Strausfeld, L. Hansen, Y. Li, R. S. Gomez, and K. Ito. Evolution, discovery, and interpretations of arthropod mushroom bodies. *Learn Mem*, 5(1-2):11–37, Jan 1998.
- L. M. Stuart, J. Boulais, G. M. Charriere, E. J. Hennessy, S. Brunet, I. Jutras, G. Goyette, C. Rondeau, S. Letarte, H. Huang, P. Ye, F. Morales, C. Kocks, J. S. Bader, M. Desjardins, and R. A. B. Ezekowitz. A systems biology analysis of the drosophila phagosome. *Nature*, 445(7123):95–101, Jan 2007.
- M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci USA*, 102(12):4221–4, Mar 2005.

- J. F. Sturgill, P. Steiner, B. L. Czervionke, and B. L. Sabatini. Distinct domains within psd-95 mediate synaptic incorporation, stabilization, and activity-dependent trafficking. *J Neurosci*, 29(41):12845–54, Oct 2009.
- A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102(43):15545–50, Oct 2005.
- M. A. Sutton and E. M. Schuman. Dendritic protein synthesis, synaptic plasticity, and memory. *Cell*, 127(1):49–58, Oct 2006.
- J. D. Sweatt. Mitogen-activated protein kinases in synaptic plasticity and memory. *Curr Opin Neurobiol*, 14(3):311–7, Jun 2004.
- D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguéz, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. V. Mering. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, 39(Database issue):D561–8, Jan 2011.
- T. Tada and M. Sheng. Molecular mechanisms of dendritic spine morphogenesis. *Curr Opin Neurobiol*, 16(1):95–101, Feb 2006.
- S. Takamori, M. Holt, K. Stenius, E. A. Lemke, M. Grønberg, D. Riedel, H. Urlaub, S. Schenck, B. Brügger, P. Ringler, S. A. Müller, B. Rammner, F. Gräter, J. S. Hub, B. L. D. Groot, G. Mieskes, Y. Moriyama, J. Klingauf, H. Grubmüller, J. Heuser, F. Wieland, and R. Jahn. Molecular anatomy of a trafficking organelle. *Cell*, 127(4):831–46, Nov 2006.
- G. W. C. Tam, R. Redon, N. P. Carter, and S. G. N. Grant. The role of dna copy number variation in schizophrenia. *BPS*, pages 1–8, Oct 2009.
- X. Tang, J. Wang, B. Liu, M. Li, G. Chen, and Y. Pan. A comparison of the functional modules identified from time course and static ppi network data. *BMC Bioinformatics*, 12(1):339, 2011.
- Y. P. Tang, E. Shimizu, G. R. Dube, C. Rampon, G. A. Kerchner, M. Zhuo, G. Liu, and J. Z. Tsien. Genetic enhancement of learning and memory in mice. *Nature*, 401(6748):63–9, Sep 1999.
- A. J. W. te Velthuis, J. F. Admiraal, and C. P. Bagowski. Molecular evolution of the maguk family in metazoan genomes. *BMC Evol Biol*, 7:129, Jan 2007.
- K. Terpe. Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol*, 60(5):523–33, Jan 2003.
- T. Tezuka, H. Umemori, T. Akiyama, S. Nakanishi, and T. Yamamoto. Psd-95 promotes fyn-mediated tyrosine phosphorylation of the n-methyl-d-aspartate receptor subunit nr2a. *Proc Natl Acad Sci USA*, 96(2):435–40, Jan 1999.

- A. Theocharidis, S. van Dongen, A. J. Enright, and T. C. Freeman. Network visualization and analysis of gene expression data using biolayout express(3d). *Nat Protoc*, 4(10):1535–50, Jan 2009.
- G. M. Thomas and R. L. Huganir. Mapk cascade signalling and synaptic plasticity. *Nat Rev Neurosci*, 5(3):173–83, Mar 2004.
- P. Thomas, M. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129, 2003.
- M. Torii, Z. Hu, C. H. Wu, and H. Liu. Biotagger-gm: a gene/protein name recognition system. *J Am Med Inform Assoc*, 16(2):247–55, Jan 2009.
- C. Torrence-Campbell, H. Gates, and U. Effiong. . . . An improved invertebrate synaptosomal preparation with cholinergic properties. *Journal of neuroscience . . .*, Jan 1991.
- J. Trinidad, A. Thalhammer, C. Specht, A. Lynn, P. Baker, R. Schoepfer, and A. Burlingame. Quantitative analysis of synaptic phosphorylation and protein expression. *Molecular & Cellular Proteomics*, 7(4):684, 2008.
- J. C. Trinidad, C. G. Specht, A. Thalhammer, R. Schoepfer, and A. L. Burlingame. Comprehensive identification of phosphorylation sites in postsynaptic density preparations. *Mol Cell Proteomics*, 5(5):914–22, May 2006.
- N. Trotta, C. K. Rodesch, T. Fergestad, and K. Broadie. Cellular bases of activity-dependent paralysis in drosophila stress-sensitive mutants. *J Neurobiol*, 60(3):328–47, Sep 2004.
- J. Z. Tsien. Linking hebb’s coincidence-detection to memory formation. *Curr Opin Neurobiol*, 10(2):266–73, Apr 2000.
- H. Tuji, M. Altaf-UI-Amin, M. Arita, and H. Nishio. . . . Comparison of protein complexes predicted from ppi networks by dplus and newman clustering algorithms. *Information and Media . . .*, Jan 2007.
- G. G. Turrigiano and S. B. Nelson. Homeostatic plasticity in the developing nervous system. *Nat Rev Neurosci*, 5(2):97–107, Feb 2004.
- P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshon, V. Narayan, M. Srinivasan, and P. Pochart. A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.
- I. Ulitsky and R. Shamir. Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, 25(9):1158–64, May 2009.
- A. Ultsch, C. M. Schuster, B. Laube, P. Schloss, B. Schmitt, and H. Betz. Glutamate receptors of drosophila melanogaster: cloning of a kainate-selective subunit expressed in the central nervous system. *Proc Natl Acad Sci USA*, 89(21):10484–8, Nov 1992.

- M. van Spronsen and C. C. Hoogenraad. Synapse pathology in psychiatric and neurologic disease. *Curr Neurol Neurosci Rep*, 10(3):207–14, May 2010.
- R. Vêncio and I. Shmulevich. Probcld: enrichment analysis accounting for categorization uncertainty. *BMC bioinformatics*, 8(1):383, 2007.
- K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A.-S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A.-L. Barabási, and M. Vidal. An empirical framework for binary interactome mapping. *Nat Methods*, 6(1):83–90, Jan 2009.
- P.-O. Vidalain, M. Boxem, H. Ge, S. Li, and M. Vidal. Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods*, 32(4):363–70, Apr 2004.
- A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, 19(2):327–35, Feb 2009.
- A. Visel, C. Thaller, and G. Eichele. Genepaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic Acids Res*, 32(Database issue):D552–6, Jan 2004.
- J. Vlasblom and S. J. Wodak. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*, 10:99, Jan 2009.
- C. Vogel and E. M. Marcotte. Calculating absolute and relative protein abundance from mass spectrometry-based protein expression data. *Nature protocols*, 3(9):1444–51, Jan 2008.
- M. Völkner, B. Lenz-Böhme, H. Betz, and B. Schmitt. Novel cns glutamate receptor subunit genes of drosophila melanogaster. *J Neurochem*, 75(5):1791–9, Nov 2000.
- J. von Engelhardt, V. Mack, R. Sprengel, N. Kavenstock, K. W. Li, Y. Stern-Bach, A. B. Smit, P. H. Seeburg, and H. Monyer. Ckamp44: a brain-specific protein attenuating short-term synaptic plasticity in the dentate gyrus. *Science*, 327(5972):1518–22, Mar 2010.
- B. VYu, S. A. Gapon, and L. G. Magazanik. Different types of glutamate receptors in isolated and identified neurones of the mollusc planorbium corneum. *J Physiol (Lond)*, 439:15–35, Aug 1991.
- Y. P. Wairkar, H. Toda, H. Mochizuki, K. Furukubo-Tokunaga, T. Tomoda, and A. DiAntonio. Unc-51 controls active zone density and protein composition by downregulating erk signaling. *J Neurosci*, 29(2):517–28, Jan 2009.
- A. J. Walhout, R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg, and M. Vidal. Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287(5450):116–22, Jan 2000.

- A. J. M. Walhout, J. Reboul, O. Shtanko, N. Bertin, P. Vaglio, H. Ge, H. Lee, L. Doucette-Stamm, K. C. Gunsalus, A. J. Schetter, D. G. Morton, K. J. Kemphues, V. Reinke, S. K. Kim, F. Piano, and M. Vidal. Integrating interactome, phenome, and transcriptome mapping data for the *c. elegans* germline. *Curr Biol*, 12(22):1952–8, Nov 2002.
- R. S. Walikonis, O. N. Jensen, M. Mann, D. W. Provance, J. A. Mercer, and M. B. Kennedy. Identification of proteins in the postsynaptic density fraction by mass spectrometry. *J Neurosci*, 20(11):4069–80, Jun 2000.
- R. S. Walikonis, A. Oguni, E. M. Khorosheva, C. J. Jeng, F. J. Asuncion, and M. B. Kennedy. Densin-180 forms a ternary complex with the (alpha)-subunit of ca^{2+} /calmodulin-dependent protein kinase ii and (alpha)-actinin. *J Neurosci*, 21(2):423–33, Jan 2001.
- T. Walsh, J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, S. M. Stray, C. F. Rippey, P. Roccanova, V. Makarov, B. Lakshmi, R. L. Findling, L. Sikiach, T. Stromberg, B. Merriman, N. Gogtay, P. Butler, K. Eckstrand, L. Noory, P. Gochman, R. Long, Z. Chen, S. Davis, C. Baker, E. E. Eichler, P. S. Meltzer, S. F. Nelson, A. B. Singleton, M. K. Lee, J. L. Rapoport, M.-C. King, and J. Sebat. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, 320(5875):539–43, Apr 2008.
- H.-Y. Wang, H.-C. Chien, N. Osada, K. Hashimoto, S. Sugano, T. Gojobori, C.-K. Chou, S.-F. Tsai, C.-I. Wu, and C.-K. J. Shen. Rate of evolution in brain-expressed genes in humans and other primates. *PLoS Biol*, 5(2):e13, Feb 2007.
- J. Wang, S. Rao, J. Chu, X. Shen, D. N. Levasseur, T. W. Theunissen, and S. H. Orkin. A protein interaction network for pluripotency of embryonic stem cells. *Nature*, 444(7117):364–8, Nov 2006.
- J. Wang, M. Li, Y. Deng, and Y. Pan. Recent advances in clustering methods for protein interaction networks. *BMC Genomics*, 11(Suppl 3):S10, Dec 2010.
- X. Wang, L. Li, and Y. Cheng. An overlapping module identification method in protein-protein interaction networks. *BMC Bioinformatics*, 13 Suppl 7:S4, Jan 2012.
- Y. T. Wang and D. J. Linden. Expression of cerebellar long-term depression requires postsynaptic clathrin-mediated endocytosis. *Neuron*, 25(3):635–47, Mar 2000.
- Z. Wang, Y. Pan, W. Li, H. Jiang, L. Chatzimanolis, J. Chang, Z. Gong, and L. Liu. Visual pattern memory requires foraging function in the central complex of drosophila. *Learn Mem*, 15(3):133–142, Feb 2008.
- V. C. Wasinger, S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams, and I. Humphery-Smith. Progress with gene-product mapping of the mollicutes: *Mycoplasma genitalium*. *ELECTROPHORESIS*, 16(7):1090–4, Jul 1995.

- S. Watanabe, D. A. Hoffman, M. Migliore, and D. Johnston. Dendritic k⁺ channels contribute to spike-timing dependent long-term potentiation in hippocampal pyramidal neurons. *Proc Natl Acad Sci USA*, 99(12):8366–71, Jun 2002.
- A. Wechsler and V. I. Teichberg. Brain spectrin binding to the nmda receptor is regulated by phosphorylation, calcium and calmodulin. *EMBO J*, 17(14):3931–9, Jul 1998.
- J. M. Welch, J. Lu, R. M. Rodriguiz, N. C. Trotta, J. Peca, J.-D. Ding, C. Feliciano, M. Chen, J. P. Adams, J. Luo, S. M. Dudek, R. J. Weinberg, N. Calakos, W. C. Wetsel, and G. Feng. Cortico-striatal synaptic defects and ocd-like behaviours in sapap3-mutant mice. *Nature*, 448(7156):894–900, Aug 2007.
- J. R. Whitlock, A. J. Heynen, M. G. Shuler, and M. F. Bear. Learning induces long-term potentiation in the hippocampus. *Science*, 313(5790):1093–7, Aug 2006.
- K. M. Wiens, H. Lin, and D. Liao. Rac1 induces the clustering of ampa receptors during spinogenesis. *J Neurosci*, 25(46):10627–36, Nov 2005.
- E. E. Winter, L. Goodstadt, and C. P. Ponting. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res*, 14(1):54–61, Jan 2004.
- C. W. Wittmann, M. F. Wszolek, J. M. Shulman, P. M. Salvaterra, J. Lewis, M. Hutton, and M. B. Feany. Tauopathy in drosophila: neurodegeneration without neurofibrillary tangles. *Science*, 293(5530):711–4, Jul 2001.
- E. C. Wooten and G. S. Huggins. Mind the dbgap: The application of data mining to identify biological mechanisms. *Mol Interv*, 11(2):95–102, Apr 2011.
- C.-L. Wu, S. Xia, T.-F. Fu, H. Wang, Y.-H. Chen, D. Leong, A.-S. Chiang, and T. Tully. Specific requirement of nmda receptors for long-term memory consolidation in drosophila ellipsoid body. *Nat Neurosci*, 10(12):1578–1586, Dec 2007.
- K. Wu, R. Carlin, and P. Siekevitz. Binding of l-[3h]glutamate to fresh or frozen synaptic membrane and postsynaptic density fractions isolated from cerebral cortex and cerebellum of fresh or frozen canine brain. *J Neurochem*, 46(3):831–41, Mar 1986.
- S. Xia, T. Miyashita, T.-F. Fu, W.-Y. Lin, C.-L. Wu, L. Pyzocha, I.-R. Lin, M. Saitoe, T. Tully, and A.-S. Chiang. Nmda receptors mediate olfactory learning and memory in drosophila. *Curr Biol*, 15(7):603–15, Apr 2005.
- Z. Xia and D. R. Storm. The role of calmodulin as a signal integrator for synaptic plasticity. *Nat Rev Neurosci*, 6(4):267–76, Apr 2005.
- W.-D. Yao, R. R. Gainetdinov, M. I. Arbuckle, T. D. Sotnikova, M. Cyr, J.-M. Beaulieu, G. E. Torres, S. G. N. Grant, and M. G. Caron. Identification of psd-95 as a regulator of dopamine-mediated synaptic and behavioral plasticity. *Neuron*, 41(4):625–38, Feb 2004.

- K. Yasuyama, I. A. Meinertzhagen, and F.-W. Schürmann. Synaptic organization of the mushroom body calyx in *Drosophila melanogaster*. *J Comp Neurol*, 445(3):211–26, Apr 2002.
- J. R. Yates, A. L. McCormack, A. J. Link, D. Schieltz, J. Eng, and L. Hays. Future prospects for the analysis of complex biological systems using micro-column liquid chromatography-electrospray tandem mass spectrometry. *Analyst*, 121(7):65R–76R, Jul 1996.
- Y. Yoshihara, M. D. Roo, and D. Muller. Dendritic spine formation and stabilization. *Curr Opin Neurobiol*, 19(2):146–53, Apr 2009.
- Y. Yoshimura, Y. Yamauchi, T. Shinkawa, M. Taoka, H. Donai, N. Takahashi, T. Isobe, and T. Yamauchi. Molecular constituents of the postsynaptic density fraction revealed by proteomic analysis using multidimensional liquid chromatography-tandem mass spectrometry. *J Neurochem*, 88(3):759–68, Feb 2004.
- K. H. Young. Yeast two-hybrid: so many interactions, (in) so little time. *Biol Reprod*, 58(2):302–11, Feb 1998.
- D. Yu, D.-B. G. Akalal, and R. L. Davis. *Drosophila* alpha/beta mushroom body neurons form a branch-specific, long-term cellular memory trace after spaced olfactory conditioning. *Neuron*, 52(5):845–55, Dec 2006.
- H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein. The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4):e59, Jan 2007.
- H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrikapa, C. Fan, A.-S. D. Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, Oct 2008a.
- J. Yu, S. Pacifico, G. Liu, and R. L. Finley. Droid: the *Drosophila* interactions database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, 9:461, Jan 2008b.
- W. Yu, A. Wulf, T. Liu, M. J. Khoury, and M. Gwinn. Gene prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC bioinformatics*, 9:528, Jan 2008c.
- R. Yuste and T. Bonhoeffer. Morphological changes in dendritic spines associated with long-term synaptic plasticity. *Annu Rev Neurosci*, 24:1071–89, Jan 2001.
- F. Zalfa, B. Eleuteri, K. S. Dickson, V. Mercaldo, S. D. Rubeis, A. di Penta, E. Tabolacci, P. Chiurazzi, G. Neri, S. G. N. Grant, and C. Bagni. A new function for the fragile x mental retardation protein in regulation of psd-95 mRNA stability. *Nat Neurosci*, 10(5):578–87, May 2007.

- M. A. Zapala, I. Hovatta, J. A. Ellison, L. Wodicka, J. A. D. Rio, R. Tennant, W. Tynan, R. S. Broide, R. Helton, B. S. Stoveken, C. Winrow, D. J. Lockhart, J. F. Reilly, W. G. Young, F. E. Bloom, D. J. Lockhart, and C. Barlow. Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci USA*, 102(29): 10357–62, Jul 2005.
- C. A. Zarate, J. B. Singh, P. J. Carlson, N. E. Brutsche, R. Ameli, D. A. Luckenbaugh, D. S. Charney, and H. K. Manji. A randomized trial of an n-methyl-d-aspartate antagonist in treatment-resistant major depression. *Arch Gen Psychiatry*, 63(8): 856–64, Aug 2006.
- B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28, Jan 2003.
- B. Zhang, Y. H. Koh, R. B. Beckstead, V. Budnik, B. Ganetzky, and H. J. Bellen. Synaptic vesicle size and number are regulated by a clathrin adaptor protein required for endocytosis. *Neuron*, 21(6):1465–75, Dec 1998.
- B. Zhang, B.-H. Park, T. Karpinets, and N. F. Samatova. From pull-down data to protein interaction networks and complexes with biological relevance. *Bioinformatics (Oxford, England)*, 24(7):979–86, Apr 2008.
- W. Zhang, Y. Zhang, H. Zheng, C. Zhang, W. Xiong, J. G. Olyarchuk, M. Walker, W. Xu, M. Zhao, S. Zhao, Z. Zhou, and L. Wei. Synldb: a synapse protein database based on synapse ontology. *Nucleic Acids Research*, 35(Database issue):D737–41, Jan 2007.
- N. Zhong, K. Scearce-Levie, G. Ramaswamy, and K. H. Weisgraber. Apolipoprotein e4 domain interaction: synaptic and cognitive deficits in mice. *Alzheimers Dement*, 4(3):179–92, May 2008.
- S. Zhong, K.-F. Storch, O. Lipan, M.-C. J. Kao, C. J. Weitz, and W. H. Wong. Go-surfer: a graphical interactive tool for comparative analysis of large gene sets in gene ontology space. *Appl Bioinformatics*, 3(4):261–4, Jan 2004.
- Y. Zhou, H. Wu, S. Li, Q. Chen, X.-W. Cheng, J. Zheng, H. Takemori, and Z.-Q. Xiong. Requirement of torc1 for late-phase long-term potentiation in the hippocampus. *PLoS ONE*, 1:e16, Jan 2006.
- X. Zhu, M. Gerstein, and M. Snyder. Getting connected: analysis and principles of biological networks. *Genes Dev*, 21(9):1010–24, May 2007.
- G. E. Zinman, S. Zhong, and Z. Bar-Joseph. Biological interaction networks are conserved at the module level. *BMC Systems Biology*, 5(1):134, Jan 2011.
- S. Zola-Morgan and L. R. Squire. Neuroanatomy of memory. *Annu Rev Neurosci*, 16: 547–63, Jan 1993.

R. S. Zucker and W. G. Regehr. Short-term synaptic plasticity. *Annu Rev Physiol*, 64: 355–405, Jan 2002.