THE UNIVERSITY *of* EDINBURGH

Computational Approaches to Discovering Differentiation Genes in the Peripheral Nervous System of *Drosophila melanogaster*

Giuseppe Gallone

Doctor of Philosophy
The University of Edinburgh
2012

Abstract

In the common fruit fly, *Drosophila melanogaster*, neural cell fate specification is triggered by a group of conserved transcriptional regulators known as proneural factors. Proneural factors induce neural fate in uncommitted neuroectodermal progenitor cells, in a process that culminates in sensory neuron differentiation. While the role of proneural factors in early fate specification has been described, less is known about the transition between neural specification and neural differentiation. The aim of this thesis is to use computational methods to improve the understanding of terminal neural differentiation in the Peripheral Nervous System (PNS) of *Drosophila*.

To provide an insight into how proneural factors coordinate the developmental programme leading to neural differentiation, expression profiling covering the first 3 hours of PNS development in *Drosophila* embryos had been previously carried out by Cachero et al. [2011]. The study revealed a time-course of gene expression changes from specification to differentiation and suggested a cascade model, whereby proneural factors regulate a group of intermediate transcriptional regulators which are in turn responsible for the activation of specific differentiation target genes.

In this thesis, I propose to select potentially important differentiation genes from the transcriptional data in Cachero et al. [2011] using a novel approach centred on protein interaction network-driven prioritisation. This is based on the insight that biological hypotheses supported by diverse data sources can represent stronger candidates for follow-up studies. Specifically, I propose the usage of protein interaction network data because of documented transcriptome-interactome correlations, which suggest that differentially expressed genes encode products that tend to belong to functionally related protein interaction clusters.

Experimental protein interaction data is, however, remarkably sparse. To increase the informative power of protein-level analyses, I develop a novel approach to augment publicly available protein interaction datasets using functional conservation between orthologous proteins across different genomes, to predict *interologs* (*inter*acting ortho*logs*). I implement this interolog retrieval methodology in a collection of open-source software modules called `Bio::Homology::InterologWalk`, the first generalised framework using web-services for "on-the-fly" interolog projection. `Bio::Homology::InterologWalk` works with homology data for any of the hundreds of genomes in Ensembl and Ensembgenomes Metazoa, and with experimental protein interaction data curated by EBI Intact. It generates putative protein interactions and optionally collates meta-data into a prioritisation index that can be used to help select interologs with high experimental support. The methodology proposed represents a significant advance over existing interolog data sources, which are restricted to specific biological domains with fixed underlying data sources often only accessible through basic web-interfaces.

Using `Bio::Homology::InterologWalk`, I build interolog models in *Drosophila* sensory

neurons and, guided by the transcriptome data, find evidence implicating a small set of genes in a conserved sensory neuronal specialisation dynamic, the assembly of the ciliary dendrite in mechanosensory neurons. Using network community-finding algorithms I obtain functionally enriched communities, which I analyse using an array of novel computational techniques. The ensuing datasets lead to the elucidation of a cluster of interacting proteins encoded by the target genes of one of the intermediate transcriptional regulators of neurogenesis and ciliogenesis, *fd3F*. These targets are validated *in vivo* and result in improved knowledge of the important target genes activated by the transcriptional cascade, suggesting a scenario for the mechanisms orchestrating the ordered assembly of the cilium during differentiation.

Acknowledgements

With thanks to Andy, Douglas and Ian, for being the best supervisors I could have asked for.

With thanks to all those who have made my Ph.D. years a wonderful experience, in particular my family, Eva, my colleagues, and my friends and flatmates in Edinburgh.

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.

(*Giuseppe Gallone*)

A Zio Giovanni, nostra fonte inesauribile di sapere.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The ultimate motivation for the work presented here is to gain better understanding on how nervous systems are formed.

Early during eukaryotic development, undifferentiated cells are assigned functional characterisation. They then take on specialised forms and functions. Specifically, complex nervous systems emerge when subsets of embryonic precursor cells evolve into neuronal and glial cell types at appropriate positions and in controlled numbers over time, in a process known as *neurogenesis*. Over the past decades, a significant body of research has tried to shed light on the genetics underlying cell determination and cell differentiation during neurogenesis. Studies conducted in the fruit fly, *Drosophila melanogaster*, have had a crucial role in improving our understanding of how cells commit to neuronal functional roles in higher eukaryotes. However, the dynamics leading committed neuronal precursors to take specialised forms are still an area of intense research.

In this thesis, I introduce a series of computational approaches aiming to improve the understanding of cell differentiation during neurogenesis, using as a model the Peripheral Nervous System (PNS) development in *Drosophila*.

The purpose of this introductory chapter is to link this aim with the content of this thesis, and to set out the approach I have taken. This is based on using computer science, graph theory and statistics to produce abstract descriptions of the *Drosophila* PNS able to inform experimental research.

1.1 Motivation

Drosophila has long been recognised as an important tool to elucidate the genetic bases of many classes of conserved biological dynamics. Its use as a model organism dates back to the pioneering work of Morgan [1917] and his students [Bridges and Morgan, 1923] who chose the fruit fly for explorations of inheritance and mutation, following on from the work of Mendel and Darwin. The reasons why *Drosophila* became a popular animal model for genetic studies were largely pragmatic. Its small size, short generation time and large brood size allow for

easy production and maintenance of large, comparatively inexpensive stocks in a laboratory environment. As a result of its early popularity between geneticists, the genetic toolkit for *Drosophila* has been refined for 100 years and is extremely powerful. Although simpler model organisms, such as yeast, have historically had advantages for studies addressing questions on cell autonomous functions, the fly has often complemented these studies, and has been used to model processes manifesting at the tissue-level or involving cell communication. Examples include metabolic, auditory, neurological, immune system and developmental processes.

Insight gained from *Drosophila* studies provides substantial help with the understanding of related mechanisms in higher eukaryotes. Specifically, the impact of *Drosophila* research on the understanding of human biology and human pathogenesis has been remarkable. A study by Reiter et al. [2001] represented one of the first attempts at compiling a map of human disease genes which could be studied using *Drosophila* as the model organism due to significant sequence similarity with fly sequences. The analysis found that 714 of 929 (77%) distinct human disease genes in the Online Mendelian Inheritance in Man (OMIM) dataset[1] have highly similar[2] cognates in *Drosophila*. Interestingly, almost one third of all human disease genes still have matches in the genome of *Drosophila* at extremely stringent E-value thresholds[3]. Subsequent research efforts (summarized, for instance, by Bier [2005]) estimated the number of disease genes having sufficiently conserved homologues in the fly at around 700, over a total of 2,309 human disease-gene entries.

Drosophila melanogaster is often used to address questions in human genetics when these have been difficult to resolve using mouse knock-out mutants or vertebrate cell culture systems. The reason why these systems may fail while a simple organism model is successful has to do with the greater genetic redundancy in vertebrates with respect to the fly. When a given biochemical function is redundantly encoded by two or more genes (as may be the case in organisms with large number of duplications and/or gene transfer events after speciation) mutations in one of these genes will have a reduced or, in general, less accountable effect on the functionality of the organism than expected from the genes' function [Kafri et al., 2006]. In such cases, *Drosophila* is often used for first approximation models and to design genetic schemes to answer the particular question at hand.

Over the course of the past few years, *Drosophila* models have been used to help elucidate a wide range of genetic-based processes in humans. The study of developmental processes has especially benefited from *Drosophila* research. Two classes of genes related to developmental processes can be studied in the fly: those that maintain human functionality through homology and those that work as part of some conserved pathway which has been co-opted for developmental purposes which are different in vertebrates and the fly. Examples of the first

[1] http://www.ncbi.nlm.nih.gov/omim
[2] blastP match with E-value $\leq 10^{-10}$.
[3] $E \leq 10^{-100}$.

class of genes include *PAX6* (*eyeless* in *Drosophila*), *SALL1* (homologous to *salm* and *salr* in *Drosophila*) and *TWIST1* (*twi* in *Drosophila*). Mutations in *PAX6* and *SALL1* cause defects in the eye and auditory systems, and mutations in *TWIST1* lead to malformations of mesodermal products in a large number of organisms. It is important to note however that the role of such homologues in their respective pathways is not always perfectly conserved between fly and vertebrates. For instance, *twi* has been known in the fly for its role as an activator of mesoderm genes [Kosman et al., 1991] such as the fibroblast growth factor receptor gene *heartless*. However, its mouse orthologue *TWIST1* has been found to act as a negative regulator of the *Fgfr2* gene, required for the formation of cranial sutures [Rice et al., 2000].

The second class of genes is involved in pathways which are reused for different capacities after speciation from the common ancestor. For instance, the *Notch* pathway (reviewed by Artavanis-Tsakonas et al. [1999]) is known to be involved in a large number of cell-fate decisions both in fly and in vertebrates, and some of these are clearly species-specific. In vertebrates, Notch signalling is essential for the segmentation of mesoderm, giving rise to skeletal elements. In *Drosophila*, it has a prominent function in limiting the width of wing veins.

In spite of species-specific pathway differences, crucial discoveries have been made by drawing inferences from one system and applying them to the other. Let us consider the Notch signalling pathway again. The pathway includes the *Delta* gene, which encodes a cell-surface ligand for the Notch receptor. Mutations of *Delta* were first identified in *Drosophila*, based on a thickened wing-vein phenotype [Bridges and Morgan, 1923]. Subsequent mouse studies showed that loss of function of the Delta-like 3 gene results in a family of related spinal malformations [Kusumi et al., 1998]. The work by Kusumi et al., together with other similar findings, then guided human genetic studies which revealed that the human *Delta* homologues *jagged 1* [Li et al., 1997] and *delta-like 3* [Bulman et al., 2000] play a role in spinal abnormalities associated with the Alagille syndrome and spondylocostal dysostosis.

Drosophila melanogaster has represented the starting point from which many of the mechanisms of neural development in vertebrates have been elucidated. Many of the genetic pathways that orchestrate basic neural developmental processes have remained largely intact through evolution [Bier, 2005]. Examples of conserved dynamics include the specification of segment identity along the anterior–posterior axis in embryos and the division of the ectoderm into neural versus non-neural domains along the dorsal–ventral axis. In this thesis, I will concentrate on the study of nervous system development using the *Peripheral Nervous System* (PNS) of *Drosophila* as the reference model.

1.2 *Drosophila* as a Model for Sensory Neuronal Development

Nervous system development in metazoans is triggered by the complex interplay between a number of transcription factors ad their DNA targets. The genes encoding proteins belonging to the bHLH (basic Helix-Loop-Helix) meta-family of transcription factors have been shown to play an essential role in early neurogenesis [Lee, 1997]. These bHLH proteins typically bind to a common DNA sequence (CANNTG) called the E-box sequence [Murre et al., 1989] through their basic region, while the two α helices are used for dimerisation.

Several sets of *Drosophila* genes have been found to share the bHLH sequence motif. The members of the *Achaete scute* bHLH complex have been described in the literature about three decades ago [Garcia-Bellido and Santamaria, 1978; Garcia-Bellido, 1979] and then again a decade later, when a number of important studies [Ghysen and Dambly-Chaudiere, 1988, 1989; Campuzano and Modolell, 1992] allowed the elucidation of the upstream controllers of the complex. Another key player in *Drosophila* PNS development is *atonal* [Jarman et al., 1993]. It has been shown that *atonal* is phylogenetically and structurally related (by means of its bHLH domain) to two additional genes, *cato* [Goulding et al., 2000a] and *amos* [Goulding et al., 2000b] whose isolation led to the definition of the *ato-like* family of bHLH genes.

bHLH transcription factors are necessary and sufficient to induce neural fate commitment in progenitor cells in metazoans [Bertrand et al., 2002]. Due to this property, they are commonly referred to as *proneural* factors. Proneural factors are expressed early in *Drosophila* quiescent ectodermal cells, when such cells have still both epidermal and neural potential. Proneural factors are not, however, ubiquitously expressed throughout the neuroectoderm. Rather, they appear in a spatially stereotyped pattern and are to be found in contiguous clusters of cells known as *proneural clusters*. Cells belonging to each proneural cluster show equivalent neural potential. A largely stochastic dynamic based on cell-cell competition and inhibitory feedback regulation, and reinforced by a mechanism mediated by Notch signalling and known as *lateral inhibition* [Pi and Chien, 2007] singles out, among the members of each cluster, one neural precursor that differentiates according to the assigned lineage (Figure 1.1-A).

While all proneural genes are responsible for providing neural competence to proneural clusters, different proneural genes endow cells with different types of competence. Members of the Achaete-Scute family are expressed in cells that will commit to an *external* sense organ (tactile/chemosensory bristles) lineage [Brunet and Ghysen, 1999] (Figure 1.1-D). *atonal* (*ato*), on the other hand, is expressed in ectodermal proneural clusters responsible for generating arthropod-specific internal sensory organs known as *chordotonal* organs (Figure 1.1-C), but also R8 photoreceptors [Jarman et al., 1994, 1995; Dokucu et al., 1996], olfactory sensilla and a subset of brain neurons [Powell and Jarman, 2008]. In this thesis, I will mostly be interested

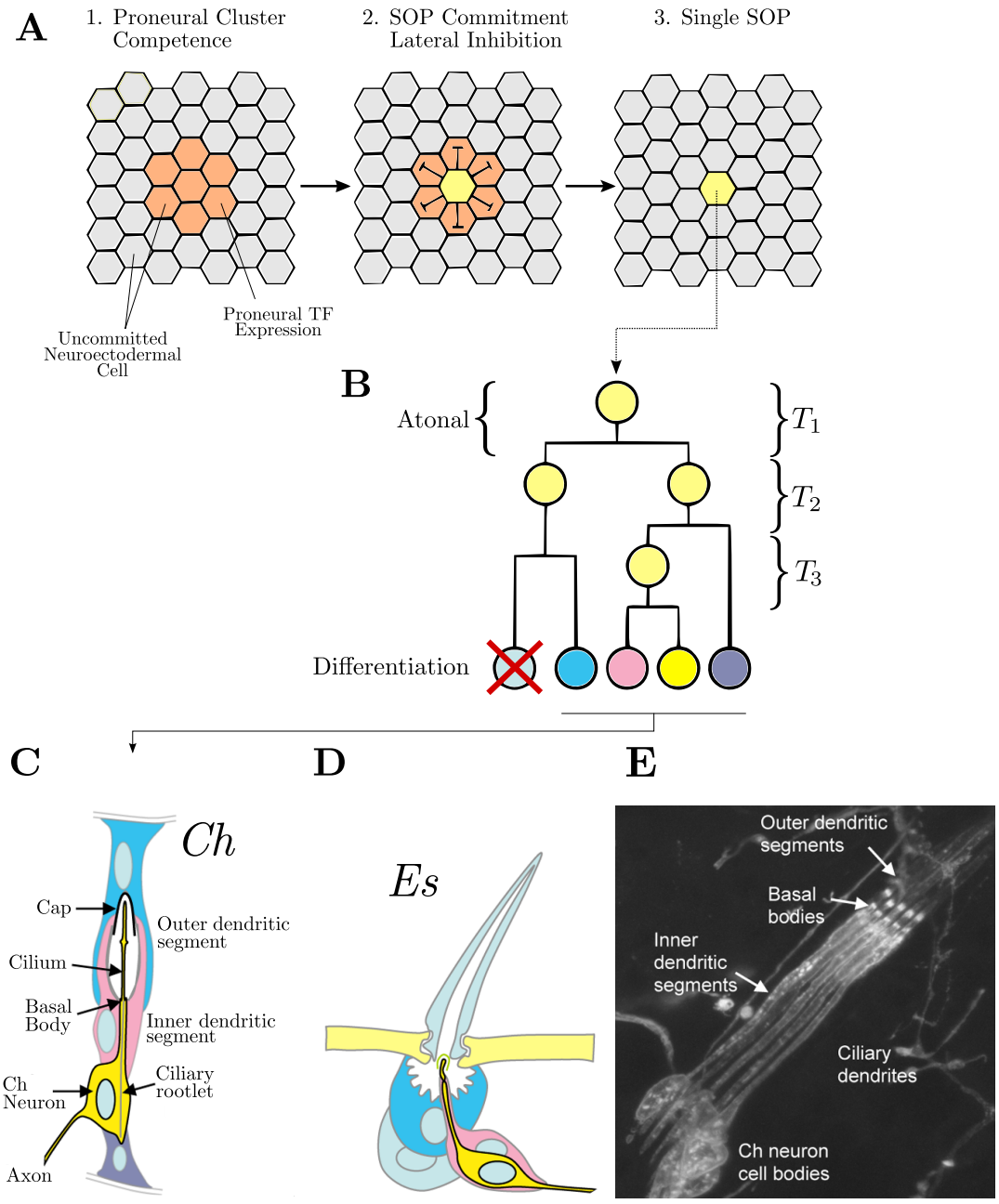

Figure 1.1: Sensory Neuron Specification and Differentiation. **A**: Specification of Sense Organ Precursor (SOP) from proneural cluster in the neuroectoderm. **B**: Schematic of cell lineage leading from an SOP to a Chordotonal organ. Ato is expressed at the SOP stage. The time points sampled for analysis are indicated approximately (T_1, T_2, T_3). **C-D**: Schematics of structural features of chordotonal (**C**) and external sensory (**D**) organs. **E**: Group of five Ch neurons in the larval lateral body wall, labelled with anti-HRP, which detects the cell body and inner dendritic segment (*Adapted from* Cachero et al. [2011]).

in *ato*-like proneural genes and related neuronal lineages[4].

In the case of *ato* neuron types, after commitment each SOP divides asymmetrically to give the 4-5 cells of an individual Ch organ (Figure 1.1-B,C). One of these cells differentiates to form a Ch neuron, while the remaining cells differentiate as support cells. A crucial point to note is that *ato* stops being expressed *before* the precursor candidate has started dividing [Chang et al., 2008]. Still, the division happens according to the plan specified by the proneural factor's blueprint. Therefore, a number of effectors of *ato* must be expressed downstream which will take the role of activating the relevant pathways needed to propagate the developmental plan as required. In other words, PNS neurogenesis is defined by a cascade of gene expression changes activated by proneural factors (Figure 1.1-B).

1.3 Linking Specification to Differentiation

In addition to their neural commitment role, proneural factors also influence the identity of the final neuron's subtype, indicating that proneural factors are at the root of neuronal diversity in the PNS [Bertrand et al., 2002]. While the dynamics of the interactions between *ato*-like factors and the notch pathway during the selection of the neural precursor are understood, less is known about how the high-level activity of proneural genes (Figure 1.1-A) leads to specific programs of neuronal differentiation (Figure 1.1-C).

Over the past decade, a number of studies have tried to shed light on the mechanisms determining gene activation downstream of proneural factors. Using the Gal4/UAS system[5] [Brand and Perrimon, 1993], Jarman and Ahmed [1998] and Brunet and Ghysen [1999] were able to prove that *ato* activity involves negative regulation of the neural selector gene *cut*, which encodes a homeodomain factor functioning as a critical bimodal switch between Ch and ES cell fates in the PNS: when no *ato* repression is in place, *cut* is responsible for external sense organ fate. Thanks to broadening knowledge regarding the functional specificity and the differences between bHLH proteins, zur Lage et al. [2003] proved that the gene *amos* acts as a repressor of bristle specification (promoted by *scute*), thus reinforcing a model where *ato-like* genes are responsible for negative control of external sensory organ fate and induction of alternative fates. Powell et al. [2004] investigated differences in proneural factor DNA binding site regions, defining an *atonal*-specific E-box binding consensus. Moving on to larger scale studies, Reeves and Posakony [2005] carried out a microarray analysis of wing-disc cells isolated using the technique of fluorescence-activated cell sorting (FACS). They obtained a list of 204 genes 2-fold enriched in *scute*-expressing proneural clusters with respect to the rest of the wing disc.

[4]The Ch and ES developmental programs are similar but terminal differentiation changes according to the specialised structures present in the two lineages.

[5]The Gal/UAS system allows expression of arbitrary transgenes only in the cells in which a particular enhancer is expressed. This allows a class of hypothesis testing to be done genetically in *Drosophila*, as opposed to mammals where similar tests would often require surgical intervention.

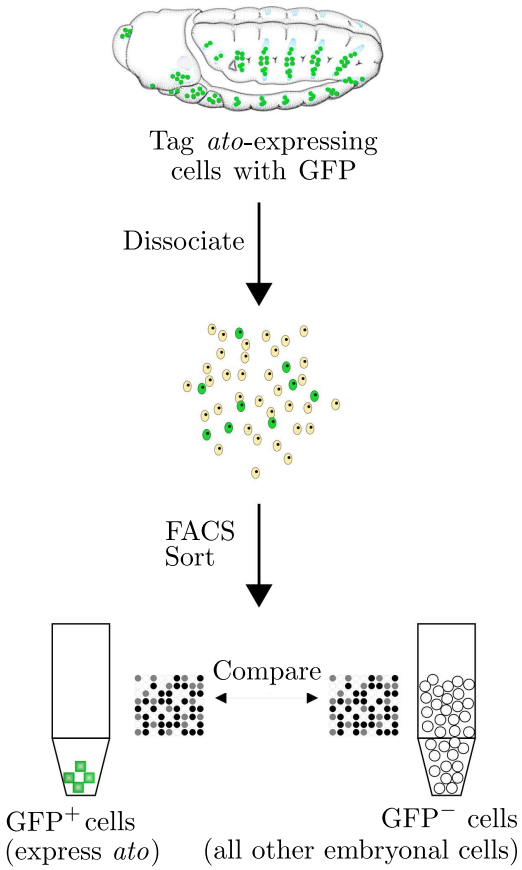

Figure 1.2: High resolution profiling of Ch cells carried out in Cachero et al. [2011]. Transgenic flies express GFP under the control of a proneural gene enhancer. Developmentally staged embryos are harvested and the cells dissociated. Subsequently, the cells are sorted by GFP fluorescence, their RNA is extracted and then hybridised to Affymetrix Dros2.0 microarray chips. Microarray processing and analysis was conducted using standard techniques recommended by Affymetrix. These experiments were performed for cells expressing *atonal*, *amos* and *cato* across developmental time points T_1, T_2 and T_3, and for an *ato*-defective mutant at T_1. (*Illustration courtesy Dr. Ian Simpson*)

They then used *in situ* hybridisation to restrict the list to a set of 27 genes some of which were shown to be direct proneural targets. Following a similar route, Ostrin et al. [2006] employed a combination of phylogenetic analysis, bioinformatics and transcription factor binding site analysis to identify targets of *Eyeless*, a transcription factor involved in retinal determination. Their analysis led to the identification of 20 putative targets, of which only one had been identified before. Another interesting route was explored by Aerts et al. [2010], who used transcriptome analysis of whole eye discs combined with computational analysis to discover direct *atonal* targets, suggesting that a major function of proneural factors might be to manipulate signalling pathways during neurogenesis.

Unfortunately, none of these studies is specifically concerned with the understanding of how neural development and subtype specification derive from *atonal* function. In order to understand the mechanisms leading from specification to neuronal differentiation, high resolution temporal profiling of embryonal Ch cells downstream of *atonal* function was carried out in the Jarman lab [Cachero et al., 2011]. *ato*-expressing cells were marked by GFP expression from an *ato*GFP reporter gene construct. This reporter gene is expressed in Ch SOP cells and their progeny[6]. *ato*GFP cells were isolated from cells in the rest of the embryo using fluorescence activated cell sorting (Figure 1.2). The data was collected from staged embryos at 3 time

[6] Although expression exists also in other *ato*-expressing cells.

points corresponding to the first 3 hours of neural development (Figure 1.1-B). Whole-genome affymetrix microarray chips were the chosen platform for data collection.

The results show a clear time course of expression changes which suggests an increase in the complexity of gene expression between specification and differentiation. At T_1, several known or suspected neural genes are part of the list of 141 genes which are 2-fold enriched in Ch cells. This includes many of the sense organ precursor genes found in previous studies. A large proportion of genes show an *intermediate Ch-enriched* pattern: a strong and early onset expression in the Ch lineage, but weak and later onset in the ES lineage. This suggests that subtype differences between the two main PNS lineages depend on a common differentiation program which is modulated in time and in expression levels. To prove that this modulation is ultimately attributable to differences in proneural gene function, Cachero et al. [2011] carry out a study of *ato*-expressing cells from *ato* mutants at T_1. Based on differences in enrichment between data for the wildtype and data from the mutant at T_1, the authors select a group of 11 genes which are good candidates for downstream targets. They then demonstrate that three of these, encoding the transcription factors Regulatory Factor X (Rfx), Cousin of atonal (Cato) and Fd3F, are directly regulated by Atonal, and are candidate intermediate regulatory factors linking proneural genes to differentiation.

Rfx [Emery et al., 1996] is a highly conserved transcriptional regulator necessary for ciliated sensory neuron differentiation [Dubruille et al., 2002]. Its sequence contains a single 76-residue DNA binding domain matching an X-box promoter sequence. This domain is very well conserved, showing about 40% identity between yeast, nematodes and mammals and close to 100% identity for the 9 aminoacid positions in direct contact with the X-box DNA sequence [Gajiwala et al., 2000]. The gene lists in Cachero et al. [2011] are significantly enriched for the presence of nearby X-box motifs, which indicates the presence of likely Rfx targets. The authors prove that *Rfx* is regulated through separable Ch and ES enhancers: the first binds Ato directly early during Ch development; the second is active only after *sc* expression is switched off, which indicates that Rfx is only an indirect *sc* target. This suggests that differences in Rfx regulation may be one way by which proneural factors regulate neuronal subtypes.

Like *Rfx*, *cato* has separable Ch and ES enhancers [zur Lage and Jarman, 2010]. An important basic-Helix-Loop-Helix (bHLH) transcription factor [Goulding et al., 2000a], *cato* is widely expressed in the developing PNS after neural precursor selection but before terminal differentiation [zur Lage and Jarman, 2010]. *cato* is a direct target of ato [zur Lage and Jarman, 2010], has been shown to be dynamically expressed during neurogenesis and is confined to the developing PNS.

atonal also regulates *fd3F*, a gene that encodes a novel forkhead family transcription factor that is exclusively expressed in differentiating chordotonal neurons from the precursor stage through to differentiation. In a more recent study [Newton et al., 2012] Fd3F was shown to

cooperate with Rfx to regulate specialised aspects of Ch neuronal physiology. In addition to regulating the intermediate transcriptional regulators *Rfx* and *fd3F*, Cachero et al. show that *ato* directly regulates a novel gene, *dilatory* (*dila*), a coiled-coil protein associated with ciliogenesis during neuronal differentiation. *dila* has been later implicated in the regulation of intraflagellar transport at the base of sensory cilia [Ma and Jarman, 2011].

1.3.1 A Specialisation of Mechanosensory Neurons – the Ciliary Dendrite

Cachero et al. report an increasing representation of genes known or suspected to be involved in the cell biological process of *ciliogenesis* along the three time points. Ciliogenesis is currently the focus of intensive research within the cell biology community (reviewed, for example, by Gerdes et al. [2009]). The *cilium* is a highly conserved cellular organelle important for sensory and motility functions in several species (Figure 1.3-A). In higher vertebrates, cilia are found on almost every cell. Examples of ciliated cells include photoreceptors, olfactory neurons, kidney, lung and embryonic cells.

The developmental processes leading to the formation of cilia are poorly understood. It is known that the dynamics of ciliary development involve the docking of the centrosome on the cell membrane where the latter becomes the basal body, from where the formation of the microtubule axoneme (forming the core of the cilium) begins. Axoneme extension and ciliary membrane expansion are coordinated and require the specialised transport protein complex known as Intraflagellar Transport (IFT) [Kozminski et al., 1993] (Figure 1.3-B). Cilia are crucial for signal transduction and the cell cycle, and disruptions to ciliogenesis can have devastating effects in humans, leading to a class of pathologies known as ciliopathies [Hildebrandt et al., 2011].

Drosophila PNS development is intrinsically related to ciliogenesis. In the fly, ciliogenesis is required to construct the sensory neuron dendrites [zur Lage et al., 2011]. The Ch neuron possesses a specialised ciliary subcellular structure that is required for processing mechanosensory signal transduction. ES cells also incorporate a modified cilium, which is anatomically and physiologically distinct from the one in Ch dendrites (Figure 1.3-C,D). Cachero et al. propose that functional and structural differences between sensory neurons are linked to differences in the regulation of ciliogenesis genes. This provides an opportunity to connect developmental proneural factors to a cell-biological pathway required for subtype differentiation. Subtype-specific variations in ciliogenesis must ultimately be regulated by the different neuronal specialisation activities of *atonal* and *scute*, and the group of intermediate transcription factors described earlier are ideally suited for a model linking proneural factors to Ch ciliary differentiation and ciliogenesis. The aforementioned ciliogenic factor Rfx, highly expressed in Ch neurons at all time points in Cachero et al. [2011], is a well known general activator of genes involved in sensory cilia formation. *fd3f* is not only a target of atonal, but has recently been

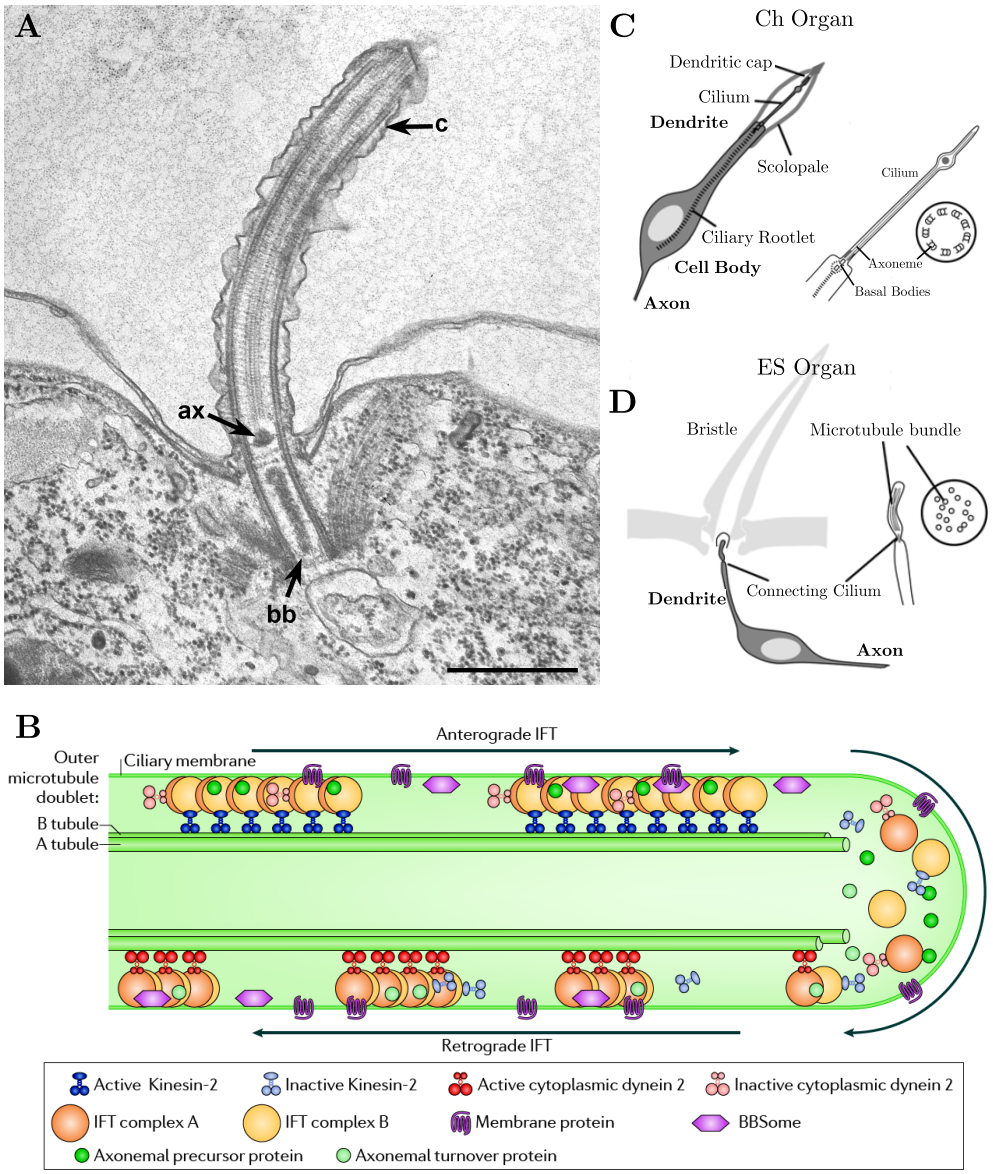

Figure 1.3: Cilia, ciliary dendrites and the conserved IFT pathway. **A**: Longitudinal section of a basal body-cilium complex. The cilium (*c*) is an organelle that arises from a basal body (*bb*) located in the cell's cortex, and is composed of 9 doublet fibers that are extensions of the tubulin triplets in the *bb*. Motile cilia have two singlet fibers in the centre of the cilium which are not present in the basal body. At least one of these singlets arises from the axosome (*ax*), at the proximal end of the cilium's core. (*bar* = 0.5μm, *EM image from* Allen [1967]) **B**: Intraflagellar Transport Machinery. The transport of ciliary proteins from the cytoplasm to the ciliary tip and back is mediated by IFT, a bidirectional movement of multiprotein complexes along the axoneme. Two IFT transport mechanisms have been described [Rosenbaum and Witman, 2002]: an anterograde one (from base to tip), mediated by the IFT-B complex and a retrograde one (from tip to base), mediated by the IFT-A complex. Movement of cargo proteins along microtubules is catalysed by kinesin (for IFT-B) and dynein (for IFT-A) motor proteins. (*Image from* Ishikawa and Marshall [2011]) **C-D**: Ciliary dendrites in Ch and ES sensory neurons. In addition to differences in their support cells, Ch and ES organs show distinctive differences in the dendrite, which is a modified cilium. In Ch neurons (**C**) the cilium at the dendrite tip is housed in a scolopale. In ES neurons (**D**) the cilium is reduced to a short segment containing a bundle of disorganized microtubules. (*Illustration from* zur Lage et al. [2011])

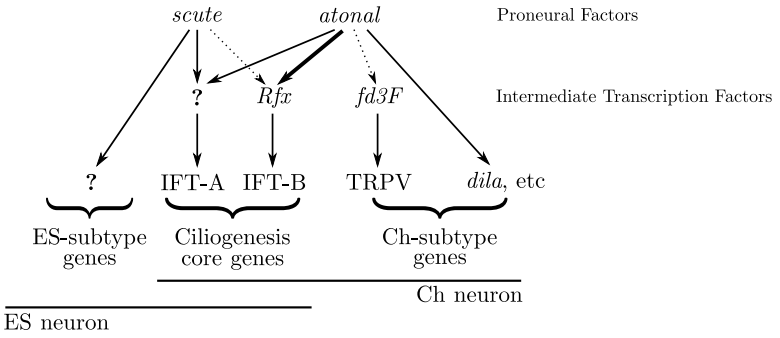

Figure 1.4: A model for the cascade leading from the proneural transcription factor *atonal* to neuronal subtype differentiation. Solid arrows indicate direct regulation, dashed arrows indicate indirect regulation. In this model, neuronal specificity is achieved from early Atonal specification in three ways. (1) *ato* regulates a Ch-specific transcription factor, *fd3F*, which regulates specialised aspects of sensory ciliary function (2) *ato* directly regulates a number of differentiation genes (e.g. *dila* is shown in the diagram). (3) *ato* regulates the ciliogenic regulator *Rfx*, which is expressed in both Ch and ES lineages. (source: [Cachero et al., 2011])

shown to be required for specialised aspects of sensory transduction and retrograde transport in Ch cilia [Newton et al., 2012]. The Forkhead gene family, of which *fd3F* is a diverged relative, has long been known for its role in ciliary modification in a number of species [Gerdes et al., 2009; Cruz et al., 2010].

Thus, the understanding of Ch neuron differentiation is linked to the understanding of ciliogenesis. The fruit fly represents an ideal model animal to study ciliogenesis, because it contains cilia on only two of its cell types, sperm cells (motile cilia) and sensory neurons (primary cilia). This restricted tissue distribution of cilia means that research focusing on *Drosophila* Ch transcriptome analysis is ideally suited to the detection of ciliogenesis genes and the elucidation of the process: mutations in ciliary genes result in flies that are phenotypically distinctive (they are uncoordinated and male sterile). Current work in the lab involves the elucidation of downstream direct targets of Rfx and Fd3F. The project is under way and some ciliary targets have been identified (for example, the Fd3F targets *nan* and *iav* [Newton et al., 2012], which will be discussed in Chapter 4) however the wealth of data available prompts for further enquiry in this direction.

1.4 Hypothesis and Project Goals

The data in Cachero et al. [2011] suggests a model which starts to bridge the gap between early specification in *ato*-expressing cells and Ch neuron differentiation, proposing that proneural factors control neuronal sub-type differences (Figure 1.4). This happens through two dynamics. In the first, proneural factors regulate both specific and common intermediate transcription factors, which in turn regulate genes directly involved in neuronal differentiation. In the second, proneural factors directly regulate differentiation genes. The results in Cachero et al.

also suggest that functional and structural differences between sensory neurons are linked to differences in the regulation of ciliogenesis genes. This provides an opportunity to connect developmental proneural factors to a cell-biological pathway required for subtype differentiation. The next step is to elucidate the nature and functionality of direct downstream targets of the intermediate transcription factors Rfx and Fd3f.

Finding these direct targets is not trivial. The wet-lab approach of using reporter gene analysis combined with mutagenesis of potential binding sites is the standard route, but is expensive and laborious. Computational approaches can however be used to prioritise hypotheses for experimental validation. Binding site analysis approaches have been attempted but these are error-prone due to problems such as high degeneracy of E-box sequences and high likelihood of non functional sites. Studies in this direction include the work by Rouault et al. [2010], who used a training set of enhancers from known genes expressed in sense organ precursors to identify candidate CIS-regulatory sequences and then use these to detect other SOP genes. Aerts et al. [2010] discovered over-represented sequence motifs associated with genes downstream of *atonal* in the eye disc, using an in-house computational tool, cisTargetX[7]. The approach led to the discovery of several new candidate *atonal* targets. The feasibility of using this approach with transcriptome data is being tested by other members of the lab.

The transcriptional analysis in Cachero et al. [2011] proposes hundreds of genes with a potential role in neuronal differentiation. Many of these are potentially direct targets of Rfx and Fd3f. However, only a selection of these genes can be followed up through experimental validation and integrated in a more detailed version of the regulatory network model proposed before.

In this thesis, I propose a computational approach to discover sensory neuron differentiation genes based on the idea that biological hypotheses supported by multiple heterogeneous data sources can represent stronger candidates for successive study. The approach I propose is based on a combined analysis of expression profiles and protein-protein interaction information. The literature describing methods combining multiple sources of data for gene functional discovery is vast (Sharan et al. [2007] provide a reasonably up-to-date summary). Here, I will mostly be interested in evidence showing that protein interaction clusters are often indicative of the existence of protein complexes and signal transduction pathways [Segal et al., 2003b] and that statistically significant transcriptome-interactome correlations have been described [Ge et al., 2001; Jansen et al., 2002; Hahn et al., 2005]. Genes with similar expression profiles are more likely to encode interacting proteins in both simple [Ge et al., 2001] and complex [Hahn et al., 2005] organisms. Conversely, pairs of genes linked by molecular interactors are more likely to have correlated expression profiles [Ideker et al., 2002].

Based on these and similar findings, the use of both information sources together for the

[7] med.kuleuven.be/cme-mg/lng/cisTargetX/

analysis of functional modules in protein interaction networks could help with the identification of smaller subsets of candidate genes. One candidate gene resulting from a combined proteomics/transcriptomics study could be considered *interesting* and deserving of a place in a candidate gene set if it satisfied both the following two conditions:

1. The transcriptomics data in Cachero et al. shows it is over-expressed or significantly enriched in Ch cells or else shows the gene has an interesting enrichment pattern over the time course;

2. Protein interaction data indicates that at least one protein product of the gene is part of a functional module in a protein interaction network; additionally, the other proteins in the functional module show evidence of involvement with sensory neuron differentiation.

Therefore, my main goal can be summarised as follows.

> *The transcriptional data in Cachero et al. lists many potential sensory neuron differentiation genes. One method to select well-supported candidates for experimental validation is to use a computational approach enabling the combination of evidence from informative heterogeneous datasets. Publicly available functional information at the protein level can be gathered to generate protein interaction networks. I wish to evaluate the possibility of isolating genes which are enriched in the transcriptional data and simultaneously have products which interact with other sensory neuron genes within functionally related protein network communities.*

The work in this thesis evaluates the potential for using protein interaction data, functional annotation data and transcriptomics data to a) support wet-lab predictions of transcription factor targets in sensory neurons, b) propose new targets to be followed up in the lab, c) isolate genes whose products interact in common neural differentiation sub-programs and d) help predict the function of these subprograms.

One of the themes in this study is the generation of protein interaction networks starting from genes showing enrichment in the PNS transcriptomics datasets discussed by Cachero et al. [2011]. On the one hand, we have transcript-level information providing a list of genes whose transcript concentration is significantly higher in cells expressing proneural factors known to be involved in late PNS differentiation (compared to the rest of the cells in the embryo). On the other, we can obtain proteomics data modelling how the final products of those transcripts are interacting with one another. While there are clearly a number of approximations involved[8], one can build a protein interaction dataset focusing on the products of these enriched genes and of their experimental and putative interactors. These additional molecules can carry valuable information. By definition, their transcript is not enriched in sensory neurons. However, the protein interaction data suggests they have a role in the PNS because they interact with

[8] e.g. no account for post-translational modifications.

proteins acting in a PNS-specific fashion. For example, they might be pan-neural genes with important roles both in the CNS and the PNS. A pure enrichment study cannot reveal these, but a combined proteomics-interactomics approach could implicate them in specific functional roles, suggesting hypotheses based on their interactions with enriched genes.

1.5 Organisation of the Thesis

The thesis is divided in three major parts arranged in three result chapters. The result chapters refer to relatively self contained groups of analyses. Each chapter, starting with a few introductory comments providing motivation and links to work done in the previous chapter, features a distinct methodology, result set and discussion. In each chapter, I discuss my findings while describing them, and provide final overall remarks in a concluding section.

In Chapter 2, I acknowledge the problem of the sparsity of available experimental protein interaction data. I propose a methodology to ameliorate the problem based on a software architecture (and related implementation) which builds putative *Drosophila* protein interactions based on the comparative biology concept of *interolog* (*int*eracting ortho*log*) mapping. I discuss implementation details, strengths and weaknesses of the approach, and present a collection of Perl modules, called `Bio::Homology::InterologWalk`. These allow one to retrieve, prioritise and visualize putative protein interactions through interolog mapping. The software package, released on the public domain under the GNU public license, is the first of its kind to use on-the-fly data retrieval from remote web services like EnsEMBL and IntAct, and works seamlessly with hundreds of genomes. It is currently being used by institutions as diverse as Rothamsted Research[9] and the National University of Taiwan[10] to build putative interactomes for poorly studied plant and animal pathogens. I run validation tests to demonstrate correct functionality of the tool and finally propose two analyses aimed at augmenting the full experimental interactomes of *Drosophila melanogaster* and of the malaria vector, *Plasmodium falciparum*.

In Chapters 3 and 4 I utilise data produced with the methodology presented in Chapter 2 and transcriptome data to address the biological questions sketched in the section above.

Specifically, in Chapter 3 I document a pilot study where I use information from protein interactions supported by transcriptional data to select novel *Drosophila* genes with a potential role in ciliogenesis. The approach is based on a putative protein network built from data compiled in the Drosophila Cilia and Basal Body (DCBB) dataset [Laurencon et al., 2007], a list of fly orthologues of genes with a tested role in ciliogenesis in several organisms. I identify network regions characterised by interesting connectivity and analyse related PNS transcript fold change information to build evidence implicating three candidate novel genes with ciliogenesis.

[9] www.rothamsted.ac.uk
[10] www.ntu.edu.tw/engv4

I conclude the chapter describing successful experimental validation of the candidates.

In Chapter 4 I rework the approach described in Chapter 3 to study genes active during *Drosophila* sensory neuron differentiation and known to have specific roles in mechanosensory cilia specialisation. I demonstrate the power of functionally enriched network communities to suggest functional hypotheses in a more specific biological domain than the one discussed in Chapter 3. I also propose an additional array of computational techniques to analyse protein interaction data which once again leads to promising experimental evidence.

1.5.1 Stylistic Conventions

In this thesis I have tried to keep to the following stylistic and grammatical conventions:

- Whenever possible I have tried to use the active voice in preference to the passive, at the risk of sounding more colloquial. This is in line with the style guide of the major peer reviewed journals in the field of Bioinformatics.
- I use the first person plural ("we", "our", "us") in two situations:
  1. When describing a point of view shared with my supervisors;
  2. When guiding the reader through the derivation of a mathematical expression.
- Throughout the thesis, I use `typewriter font` primarily to indicate dataset names. Specific naming conventions for dataset names will be introduced in each of the chapters. Additionally, I use typewriter for URLs, ontology concept names and to refer to the Perl software package presented in Chapter 2, `Bio::Homology::InterologWalk`.

CHAPTER 2

Augmenting Protein Interaction Datasets using Interologs

In the introductory chapter I discussed some of the important questions related to the study of sensory neurons in the fruit fly, and summarised the main results of a recent microarray-based time course investigation of the fly PNS transcriptome [Cachero et al., 2011]. Further to that, I wrote about a number of strategies which could aid with the reconstruction of the pathways underlying peripheral nervous system development in *Drosophila*. I proposed the usage of protein interaction networks based on the insight that statistically significant transcriptome-interactome correlations in many experimental contexts have been found [Ge et al., 2001; Jansen et al., 2002; Segal et al., 2003b] and pairs of genes linked by molecular interactors are more likely to have correlated expression profiles [Ideker et al., 2001]. In the last section of Chapter 1 I defined the scope and motivation for the thesis and postulated that the transcriptome data in Cachero et al. can be used as a reference set for a data mining experiment aiming to map literature-annotated protein interactions within maturing fly neurons.

During my survey of the field of interactomics, I found that the fidelity and completeness of protein interaction networks is limited by the relatively scarce amount of experimental interaction data available. Additionally, protein interaction data is restricted to just a few widely studied experimental organisms. Due to the sparse nature of current publicly available protein interaction information in *Drosophila melanogaster*, I concluded that a plain experimental protein interaction retrieval approach would risk to be insufficient. I conjectured that additional strategies to enrich available interaction datasets would be likely to lead to improved hypotheses regarding the mechanisms at the basis of fly neuronal development. In order to extend the utility of existing datasets, computational methods can be used that exploit functional conservation between orthologous proteins across *taxa* to predict putative interactions based on *homology mapping*. To date, most prediction efforts based on homology mapping have been restricted to specific biological domains with fixed underlying data sources and there are no software tools available that provide a generalised framework for customisable putative interaction prediction.

In this chapter, I introduce a methodology and a set of software tools to retrieve, prioritise and visualise putative protein-protein interactions using a novel orthology-mapping algorithm. The method I propose uses publicly available orthology and experimental interaction data to generate its predictions, and optionally collates meta-data into a prioritisation index that can be used to help in selecting predictions with high biological support for further analysis. I evaluate the methodology and its implementation on two sample datasets: the genomic interactomes of *Drosophila melanogaster* and of the malaria vector, *Plasmodium falciparum*. I discuss the resulting interaction networks and argue that the method proposes new biologically plausible interactome members and interactions that are candidates for future experimental investigation.

The proposed putative interaction prediction tool interfaces to up-to-date homology and interaction data sources to generate fresh predictions whenever required. This represents a significant advance on previous methods to perform similar predictions, as it allows the use of the latest orthology and protein interaction data. Additionally, the implementation works seamlessly with a very wide range of genomes: while its initial purpose was to build protein interaction datasets to aid with the understanding of fly PNS, we thought it would be a useful tool for several other usage scenarios (for example, the computational annotation of interactomes for recently sequenced obscure genomes). The software, methodology and part of the results presented in this chapter were published in BMC Bioinformatics [Gallone et al., 2011]. This chapter also includes new material, including data relative to the *Plasmodium falciparum* interactome and a related discussion.

Based on the results obtained in this chapter, and in spite of the limitations of the approach (discussed in some length in the conclusions to the chapter) I argue that the proposed methodology and data it produces can prove useful in exploring the interaction landscape within maturing neurons in the fly PNS. The next two chapters will concentrate on how this methodology represents a powerful tool for protein interaction discovery in sensory neurons.

2.1 Background

The study of large networks of protein interactions has attracted an increasing amount of interest over the past few years. Interactomics, as it has come to be known, is a discipline that focuses not only on the analysis, interpretation and visualisation of the interactions between biological molecules within cells, but also on the causes and effects of such interactions from a systemic point of view [Cusick et al., 2005]. This research field attempts to address some of the big questions raised by the abundant sequence data of the post-genomic era [Bray, 2003].

Protein interaction models of complete genomes are of great interest for a number of reasons. While reductionist, single-gene type experiments play a crucial role in understanding the fine details of discrete cellular constituents, shifting to a wider perspective and gaining insight

of the functional relationships that relate molecular components to one another may provide insights into the operation of the system as a whole. Indeed, most biological functions originate from the combined activity of many interacting molecules [Hartwell et al., 1999]. Defining the interactions among proteins is essential, because they play a role in virtually every biological process. The study of the structural properties of protein interaction networks can provide insights into the pathways and functional relationships between the molecules in the network. In *guilt by association* studies, for instance, the fact that a protein of unknown function participates in a highly connected complex of well studied, functionally related molecules might provide inferential clues on its own function [Oliver, 2000].

Several lab-based methods have been devised in order to investigate protein interactions. One key technological advance in this sense has been the introduction of high throughput techniques for experimental interaction detection. High throughput screen technology has allowed the characterisation of complete or quasi-complete interactomes, as opposed to just pairs of interactions. Two technologies, Yeast Two Hybrid (Y2H) [Bendixen et al., 1994] and Tandem Affinity Purification (TAP) [Puig et al., 2001] are crucially responsible for the rapid development of Interactomics. Y2H technology uses a mating assay where two proteins to be tested *in vitro* are expressed in yeast as fusion proteins. The first protein (*bait*) is fused to the DNA-binding domain (BD) of a transcription factor which binds a site upstream of a reporter gene. The second protein, (*prey*), is fused to a transcription activation domain (AD). If the bait and prey interact, the AD activates the reporter, leading to transcription and the formation of a colony on media. In TAP-based protein complex identification, a peptide called the *TAP tag* is fused to the C-terminus of a protein of interest. The resulting fusion protein goes through two cycles of affinity purification: first, it binds to beads coated with an Immunoglobulin G antibody. Then, the TAP tag is cleaved and a different part of the TAP tag binds to beads coated with Calmodulin. The protein complex resulting from the two purifications is examined for binding partners.

Y2H and TAP assays have unravelled interactions in a variety of unicellular and multicellular organisms, e.g. *Saccharomyces cerevisiae* [Yu et al., 2008a], *Campylobacter jejuni* [Parrish et al., 2007], *Plasmodium falciparum* [LaCount et al., 2005], *Caenorhabditis elegans* [Li et al., 2004b], *Drosophila melanogaster* [Giot et al., 2003], *Homo sapiens* [Stelzl et al., 2005]. Mass Spectrometry is also producing a large amount of protein interaction data [Figeys et al., 2001; Krogan et al., 2006; Ewing et al., 2007]. The data from these screens have proven extremely useful both for individual studies and for interactome modelling. The protein interaction maps generated for human pathogens, for instance, provide clues about proteins that might function together during pathogenesis and help identifying putative protein targets for drug development [Parrish et al., 2006b]. High throughput two-hybrid screens can sometimes focus on specific diseases or pathways and have proven very useful at identifying poorly characterised proteins

at the heart of conditions like inherited neurodegenerative disorders. For example, Lim et al. [2006] carried out a study of Purkinje cell degeneration proteins, linking many of the poorly characterized disease proteins to each other and to proteins with known functions, and providing new clues about the pathways involved in the ataxia diseases. The study found, amongst other results, that the majority of the ataxia-causing proteins interact either directly or indirectly, and that some of the physical interactors discovered are modifiers of ataxia phenotypes with important roles in neurodegeneration.

While lab assays and experimental quantification are the most sensible approach to protein interaction discovery there are a number of shortcomings that have led researchers to complement and corroborate *in vitro* results using other paradigms. For a large number of model organisms, experimental generation of interactions can still be extremely difficult or error prone. In the case of high throughput yeast two hybrid screens, establishing efficient strategies to mate large sets of BD (DNA Binding Domain) and AD (Activation Domain) yeast strains to sample all possible combinations of interactions can represent a challenge [Parrish et al., 2006b]. When testing for mammalian proteins, the reliability of the results can be compromised by the lack of certain post-translational modification mechanisms in yeast. For instance, lack of phosphorylation of a certain prey protein in the yeast host might lead to false negatives. When testing for bacterial interactions, Y2H might not report a true positive because the yeast might lack a chaperone required for proper protein folding and only produced by the bacterial host. The Y2H system relies on the proteins interacting in the nucleus, and it will not prevent two proteins, which reside in physically different cell locations in the original organism, to interact in the yeast system, thus reporting a false positive [Coates and Hall, 2003]. Furthermore, most high throughput experimental techniques are still very expensive. Genome screening projects continue to be carried out only around a limited number of popular model organisms, and amongst these coverage of the captured interactions is often biased to a particular domain and incomplete. Even when full genome sequencing efforts are undertaken on new species, very few of those will eventually become model organisms and attract enough attention to justify expensive experimental proteomics analyses.

In an attempt to address the relative paucity of data a number of computational protein interaction mapping techniques have been proposed. The underlying idea is that computational prediction of protein interactions can provide methods to highlight interesting proteins from large lists of potential candidates when little or no experimental evidence is available. In other terms, computational predictions can *prioritise* interaction candidates and produce hypotheses that can subsequently be tested *in vivo* [Valencia and Pazos, 2002; Berggård et al., 2007]. While the variety of computational interaction prediction approaches is large, here I focus on methods that share the basic idea that inter-species sequence conservation can inform protein function. By comparing interactomes of different organisms, observations on diverged and conserved

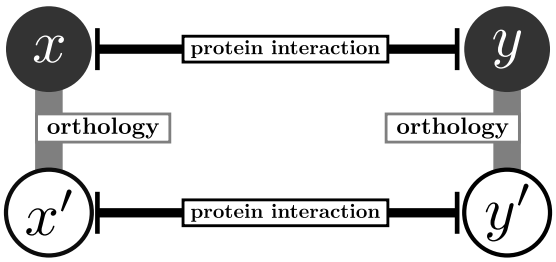

Figure 2.1: Given two proteins x and y that are known to interact, if the proteins x' (homologous to x) and y' (homologous to y) interact then the members of the pair $[I = (x,y), I' = (x',y')]$ are termed *interologs* (*inter*acting ortho*logs*).

pathways can be made on the basis of the amounts of variation and conservation reported [Wuchty et al., 2003], and the comparison of interaction maps from multiple organisms helps the prediction of additional interactions missing in one system but found in others.

This kind of insight has spurred the development of a field termed "comparative interactomics" [Cesareni et al., 2005]. Building on the interaction data available for some organisms, transfer of functional information to less annotated genomes has been attempted on the premise that sufficiently high sequence conservation is observed [Bork et al., 1998; Hegyi and Gerstein, 2001]. A number of studies have discussed the concept of cross-species orthology projection: If interacting proteins x and y in organism $\mathcal{A}$ have orthologues x' and y' in organism $\mathcal{B}$, under certain conditions the interaction will be conserved in organism $\mathcal{B}$, i.e. the x-y interaction can be mapped through the orthologies to obtain a putative x'-y' interaction. The pair of interactions (x,y) and (x',y') are named *interologs* [Walhout et al., 2000; Matthews et al., 2001] (Figure 2.1). Over the past few years, the potential of interolog mapping has been explored in a broad range of contexts. A class of studies addressed the lack of experimental protein interaction data in *Homo sapiens* [Huang et al., 2004; Lehner and Fraser, 2004; Brown and Jurisica, 2005; Persico et al., 2005; Kemmer et al., 2005; Gandhi et al., 2006; Huang et al., 2007]. The idea has been tested for similar reasons on a number of other organisms, including *Helicobacter pylori* [Wojcik et al., 2002], *Saccharomyces cerevisiae* [Sharan et al., 2005], *Plasmodium falciparum* [Wuchty and Ipsaro, 2007] and *Magnaporthe grisea* [He et al., 2008]. Yu et al. [2004], Michaut et al. [2008] and Wiles et al. [2010] have added elements to better quantify the predictions, using a variety of algorithms (Cluster of Orthologous Groups/String [von Mering et al., 2005] and Inparanoid [Remm et al., 2001] being two instances) and introducing scores to assess the reliability of the proposed transfer.

In spite of the fact that the data generated by these studies is frequently made available through web interfaces to databases (e.g. HomoMINT [Persico et al., 2005] and Ulysses [Kemmer et al., 2005]), most of these are *ad hoc* efforts. They consider a restricted set of organisms chosen to answer a specific set of questions, thus hindering the applicability of the ideas pro-

posed to different scenarios. Moreover, the publicly available datasets released are often static: data is either frozen at the moment of publication, or curated for a limited period of time, and subsequently abandoned. In some cases (e.g. InteroPORC [Michaut et al., 2008]) while the interface is maintained and developed, its data source stops being updated because the original project ceases to exist (Integr8 project [Kersey et al., 2005]). In other cases, data is still produced, but the generating algorithms are not state-of-the-art [Li et al., 2003]. In a field like comparative interactomics, where new orthology detection methods and new experimental interaction datasets are presented continually, static databases are destined to obsolescence.

More importantly, these kinds of approaches offer little flexibility when putative protein interactions are only an intermediate goal in a more complex discovery scheme. Pedamallu and Posfai [2010] have recently introduced *OpenPPI_predictor*, an open source program that implements a pipeline to obtain a putative network mapped from a reference genome. The tool also outputs results in a format compatible with the network manipulation program Cytoscape [Shannon et al., 2003]. While representing a step forward compared to the older generation, database-oriented approaches the tool is rather limited in its functionality: it can only be run on Unix, orthology data and interaction data must be manually obtained from reference databases, only one reference genome can be used at a time and there is no possibility of prioritising the putative interaction network obtained.

I decided to utilise interolog predictions to complement the available curated experimental protein interaction datasets in an effort to gain an insight into the dynamics happening during sensory organ development in *Drosophila*. Unfortunately, as evidenced in the previous two paragraphs, there is shortage of methodologies and implementations to build interolog datasets. No available tool would have been suitable for the task due to the limitations discussed. Therefore, I decided to design my own methodology, including a software architecture which would allow me to collect and collate data from remote repositories. This would then be used in conjunction with the transcriptomics data in Cachero et al. [2011] to predict interactions in developing fly neurons. Quite early on during the design of the methodology, however, I realised that a universal tool to build interolog predictions seamlessly across genomes would be of use in a wide number of scenarios, and would be a significant advance on the interolog databases described so far.

This chapter introduces therefore `Bio::Homology::InterologWalk`, a Perl module to retrieve, score and visualize putative protein interactions through interolog projection. `Bio::Homology::InterologWalk` is composed of a set of related libraries, freely available as a single source package on CPAN, the Comprehensive Perl Archive Network[1]. The tool is built on top of the Bioperl toolkit, the largest library of Perl modules to manage and manipulate life science information [Stajich et al., 2002], on the Ensembl and EnsemblGenomes Developer

[1] `www.cpan.org`

Application Programming Interface (API) and comparative genomic databases [Kersey et al., 2010; Flicek et al., 2010], and on the Proteomic Standard Initiative Common Query InterfaCe (PSICQUIC) web service [Aranda et al., 2011], an effort from the HUPO Proteomics Standard Initiative (HUPO-PSI) [Kerrien et al., 2007] to standardise the access to molecular interaction databases programmatically.

`Bio::Homology::InterologWalk` accepts as input a list of Ensembl gene accession numbers from any of the vertebrate or metazoan genomes within the Ensembl project, including the species in the Ensembl pan-taxonomic Compara database. The algorithm "walks" through the Ensembl and the PSICQUIC-enabled databases to collect, analyse and collate gene orthology data and protein interaction data, together with ancillary information. It then provides the option of filtering the putative interactions to retain those with strong experimental or phylogenetic support. Additionally, the module allows to query the experimental protein interaction database directly and collect all known interactions for the input gene list. This is useful to evaluate the significance of putative data in light of any existing experimental interaction data available for the domain. The software outputs plain text tab-separated files and can also output network representations of the protein interaction data and their attributes in a format compatible with the widely used biological network analysis tool Cytoscape [Shannon et al., 2003].

In the following sections I provide implementation details, a discussion of the main design decisions and a number of validation arguments for `Bio::Homology::InterologWalk`. Usage of the tool for the original purpose of investigating protein interactions in fly sensory neurons will be demonstrated in the next two chapters. Here, I will discuss results obtained using the method to investigate the potential of interolog projection using two generic sample datasets: the genome of the fruit fly, *Drosophila melanogaster* [Adams et al., 2000] and the genome of the protozoan parasite *Plasmodium falciparum* [Gardner et al., 2002]. In both test scenarios, the analysis generates a novel putative protein interaction network that increases the connectivity of the known interactomes and proposes new biologically-plausible interaction candidates, suggesting that the tool can be of great utility for protein discovery in fly sensory neurons.

2.2 Design and Implementation

2.2.1 Overview

A high-level schematic describing my implementation of the interolog walk principle is shown in Figure 2.2. The main purpose of `Bio::Homology::InterologWalk` is to obtain a list of putative protein interactions given a set of user-selected gene identifiers in a genome of interest. In order to be compatible with the module, the initial dataset must be a list of Ensembl IDs belonging to species in Ensembl Vertebrates, EnsemblGenomes Metazoa or Ensembl Pan-

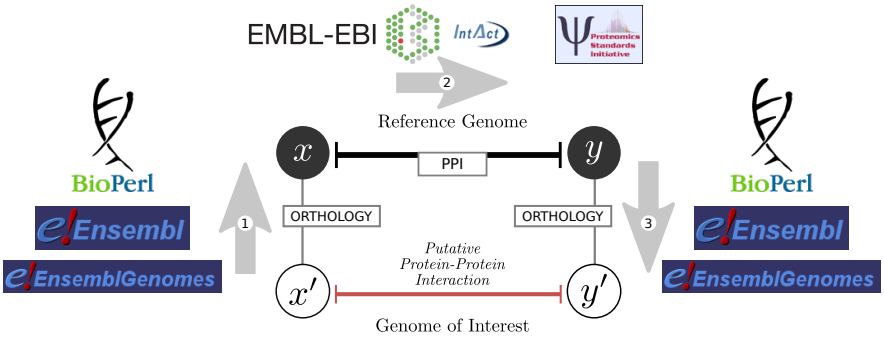

Figure 2.2: Schematic illustrating the principle behind interolog mapping as implemented in `Bio::Homology::InterologWalk`. Input data is a list of gene identifiers belonging to one genome of interest. In step **-1-**, orthologues of the genes of interest are obtained. These will belong to one or more *reference genomes*. In step **-2-**, the algorithm obtains any protein interactions for the list of orthologues obtained in step 1. In step **-3-**, the interactor list built in 2 is queried to find orthologues back in the original genome of interest. Across the three steps, options can be set to specify the stringency of the hits. The algorithm needs an internet connection and works by calling remote Web Services exposed by the data providers Ensembl and EBI IntAct. The algorithm also interprets, builds or retrieves supporting orthology and protein interaction metadata, and uses it to optionally output prioritisation metrics to help interpretation and selection of the results.

taxonomic Compara databases.

To carry out an interolog walk, `Bio::Homology::InterologWalk` will first query the gene identifiers chosen by the user against the Ensembl databases using the Ensembl Compara API [Vilella et al., 2009], retrieving a list of orthologous gene IDs. Next, the algorithm employs the Representational State Transfer (RESTful) interface [Fielding and Taylor, 2000] to interrogate a PSICQUIC-compliant protein interaction database with the list of orthologues returned by Ensembl, to retrieve the list of known interactions involving them. There are already several interaction databases implementing the PSICQUIC interface for programmatic data access [Prieto and De Las Rivas, 2006; Razick et al., 2008; Breitkreutz et al., 2008; Goll et al., 2008; Chautard et al., 2009; Jensen et al., 2009; Matthews et al., 2009; Ceol et al., 2010]. `Bio::Homology::InterologWalk` currently relies on the IntAct resource by the European Bioinformatics Institute (EBI) [Aranda et al., 2010].

Having obtained a list of interactors for the orthologues of the initial gene set, in the last step of the main data mining procedure `Bio::Homology::InterologWalk` will project the interactions retrieved (again, using the Ensembl Compara API) back to the original species of interest. The final output is a list of putative interactors for the initial gene set and several fields of supporting data for the forward orthology map, the protein interaction data collection, and the backward orthology map. The procedure is organised as a pipeline of related data-processing activities. The output of the basic pipeline can be further processed with the help of other methods in the module: it is possible to scan the results and compute counts, check for duplicate entries, isolate new gene IDs (i.e. not part of the original dataset) and save them in

another data file for further study.

An additional stand-alone functionality of the module is the *direct* interaction retrieval pipeline: it is possible to use `Bio::Homology::InterologWalk` to mine all the experimental protein interactions involving the initial gene list within the genome of interest[2]. This dataset is a "snapshot" of the current experimental interaction network for the input dataset. As such, it is useful both by itself (because it tells what is currently known in terms of experimental protein interactions for the initial genes) and as a term of comparison for the putative protein interactions (because it can be used to evaluate the amount of overlap between the known and putative networks, as well as the novelty of the putative data).

Once a putative protein interaction dataset has been obtained, it is possible to process it in a number of ways, including the computation of a prioritisation index and a conservation score, and the generation of Cytoscape-compatible network representations. One of the most important features of `Bio::Homology::InterologWalk` is that the retrieval of both orthology and interaction data happens on-the-fly. The user inputs a list of gene IDs plus a number of set-up parameters, and the data will be downloaded through web-service interfaces each time the program is run. To our knowledge, `Bio::Homology::InterologWalk` is the first project relying completely on web-services for homology and protein-protein interaction data retrieval. This means the user can repeat the data collection every time Ensembl and IntAct publish a new release of the data. Thus, an up-to-date dataset is easy to maintain for the final user. This represents a significant advance over existing interolog data projects, which are mostly static, pre-computed data repositories and in most cases do not undergo update cycles, thus becoming progressively less useful as new experimental protein interaction data and new orthology prediction methods lead to more refined data. Section 2.2.2 introduces the challenges related to programmatic access and integration of biological data from different providers.

2.2.2 Programmatic Access to Biological Data Resources

Biological data repositories have adopted a number of strategies to enable end users to search for information and manage search results efficiently. In most cases, the simplest form of data access is based on some form of web interface: a single identifier will be submitted through a textbox, using a limited number of pre-selection criteria. Usually, intermediate forms will then be presented to the user (mostly to allow for filtering of large result datasets), followed by the actual result sets. The National Center for Biotechnology Information[3] (NCBI), Ensembl[4] and most biological data repositories support this basic interaction paradigm (Figure 2.3-A). A slightly more advanced approach is based on customised query builders (Figure 2.3-B). Biomart [Kasprzyk, 2011] allows a biologist to construct customised queries of varying com-

[2] This implies no mapping to reference genomes using orthology.

[3] www.ncbi.nlm.nih.gov

[4] www.ensembl.org/index.html

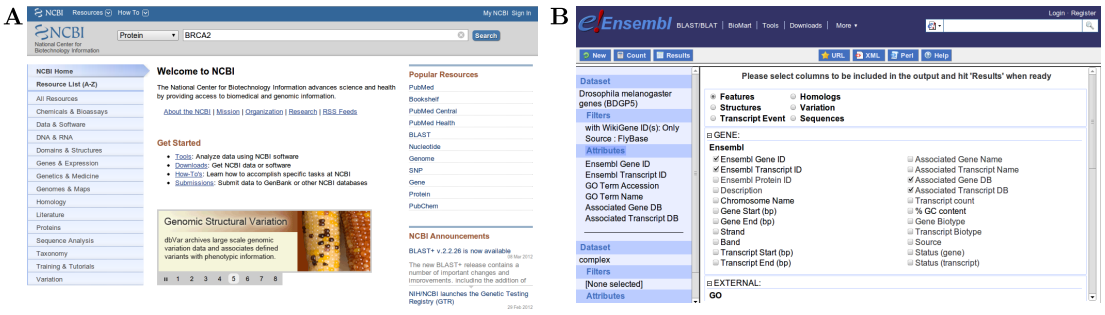

Figure 2.3: Approaches to user interaction to biological data. **A**: Text-box input web interface used by the NCBI. **B**: Biomart query builder, integrated in the Ensembl suite of web tools.

plexity through a set of web-forms and does not require any previous database experience. It is, effectively, a structured "abstraction layer" working between the user and the database data manipulation language, transparently turning user web-form input into queries. Compared to the simple data access paradigms described earlier, query builders allow a non-technical user to precisely describe the nature and characteristics of the information required. Since Biomart also allows to organize multiple distributed database systems into a virtual single integrated database, joining datasets, excluding entries on the basis of one or more properties and managing quite large collections of entries is simpler for the end-user.

Web-based access paradigms have another advantage: in order to offer the most up-to-date biological information, most large biological data providers follow a periodical release schedule, and their web interfaces typically reference the latest available data release[5]. This cyclic update process is completely transparent to the web-based user.

These two data-access paradigms are suited to quick comparison of data entries and are extremely helpful for non-technical users of the data. However, they are not designed for high-throughput, selective mass-retrieval of data entries and are close to useless when advanced real time analysis of the data is required. Pre-designed query builders are not ideal when huge amounts of data need multiple cycles of manipulation, selection, visualisation and evaluation before yielding useful content. When information is not explicit in the raw data and some form of post-processing needs to be performed to extract it, many biological data providers offer the possibility to download full dumps of the actual database tables to set-up off-line mirrors. This allows interested parties to work on large quantities of data avoiding the bottleneck of graphical interfaces.

Working on a local database mirror using an off-line copy of the data has several advantages. Because the data resides on local storage and there is no competition for remote resource access[6], queries and in general data manipulation will be much faster. Additionally, because

[5] Unless explicitly specified otherwise by the end user.

[6] or very low competition, in case a local client-server architecture has been set up.

no simplified web-interface exists between the developer and the data, the full array of relational database query language directives can be used to select very precise portions of the data and advanced analysis techniques can be used that are not available when exploring the data online. There are still cases however, when downloading a static dump of the data is still not the best option. Downloading the schemata and the data is certainly ideal in a "one-off" analysis scenario: the operations linked with downloading the data, learning the schema, setting up the environment occur one time only and what follows is some kind of discovery process linked to a hypothesis, validation, findings and so on. There are cases, however, when repeated observations of the data over time are required and a database dump will likely be out of synchronisation with the latest data version on the site. Additionally, since the database schema is often subject to refinement changes across releases, doing periodical data analysis on a local up-to-date dump might mean having to learn again the data architecture every time, with an ensuing metadata-maintenance overhead.

To address the drawbacks related to static database dumps some biological data providers have started to engineer frameworks for programmatic data access through API or Web Service Interfaces. In this context, an API serves as a middle-layer between a set of application programs or scripts and a biological database schema. It eliminates the necessity for a programmer to learn database design principles and schemata[7] and it does so by exposing a set of methods that programmers call in their scripts, allowing structured and repeatable access to large amounts of data. Similarly to the web-interface approach described above, API-access allows ready retrieval of the most recent dataset version. Unlike web-interface approaches, API-based access does not limit data manipulation to the expressive power of the web-interface: the power of the programming language employed by the API can be used to devise scripts of varying power and complexity.

A thorough definition of the concept of Web Service lies beyond the intended purpose of this dissertation. Here it shall suffice to say that a web-service interface is very similar to a *language-independent* API: a platform sitting between user-code and biological data that defines both a list of programming language-independent methods and a structured query language.

In the following section I shall define the scope and function of `Bio::Homology::InterologWalk`, a code library written in Perl that accesses biological database using both types of programmatic access described: an API-based one and a Web Service-based one. It does so to work simultaneously with data providers defining either only API-based access, or Web Service-based access. Specifically, `Bio::Homology::InterologWalk` uses the Ensembl API to obtain orthology and sequence data from the Ensembl databases and the PSICQUIC web-service interface to obtain interaction data from one of the PSICQUIC-compatible protein inter-

[7] Because any changes in the API-schema interface are the responsibility of the API developers.

action databases, EBI IntAct. The following sections will detail the rationale for these choices.

2.2.2.1 Orthology Predictions from Ensembl Compara

Orthologs are genes in different species that evolved from a common ancestral gene by speciation [Koonin, 2005]. Over the past few years, a growing body of literature has shown that orthology prediction is not a solved problem. In its simplest manifestation, given a gene from one genome, the gene from another genome with the highest sequence similarity is the orthologue[8]. This definition can in theory work when the two compared species are evolutionarily close, but at larger phylogenetic distances problems arise with this definition. If gene duplications occurred in each of the two lineages after the speciation event, the usage of a best-to-best sequence approach would miss most of the duplicated genes and fail to accurately describe the homology processes at play. Additionally, when the best hit is not highly statistically significant (as is the case many times with genes in very diverged species [Lonetto et al., 1992]) the risk of false positive orthologues is high, and conversely if the E-value cut-off is too stringent the risk of false negatives can be unacceptably high as well. Over the years, the large amount of new sequenced genomes and the need to compute orthology relationships between ever more distant species has made these drawbacks particularly important.

To deal with the problems inherent with (reciprocal best)-BLAST-hit approaches, and to attempt to distinguish a wider range of homology relationships, a number of alternative methods have been devised that are able to better appreciate the differences between speciation and duplication within and across genomes. One of the most effective has proven to be the phylogenetic tree-based method [Thornton and DeSalle, 2000; Storm and Sonnhammer, 2002], which discriminates speciation from duplication events by mapping species phylogenies onto phylogenetic gene trees [Goodman et al., 1979; Zmasek and Eddy, 2001].

`Bio::Homology::InterologWalk` uses the Ensembl Perl API[9] to access the comparative biology data provided by the Ensembl Project through Ensembl Compara. The orthology prediction method utilised by Ensembl is based on the second orthology method described above, species tree reconciliation [Vilella et al., 2009]. The choice fell on Ensembl Compara both for its advanced orthology labelling pipeline (interolog mapping crucially requires as correct as possible orthologue/paralogue labelling) and at the same time for its advanced API access to the data.

Going into some detail, Ensembl Compara implements a computational pipeline having at its core the TreeBeST algorithm[10]. TreeBeST is based on a modified version of the PhyML algorithm [Guindon and Gascuel, 2003]. The gene orthology and paralogy prediction pipeline

[8] (reciprocal best)-BLAST-hit approach

[9] `www.ensembl.org/info/data/api.html`

[10] `treesoft.sourceforge.net/treebest.shtml`

in Compara is based on the following 6 steps[11]:

1. load the longest translation of each gene from all species used in Ensembl
2. run WUBlastp+SmithWaterman of every gene against every other (both self and non-self species) in a genome-wise manner
3. Build a sparse graph of gene relations based on Blast scores and generate clusters using hcluster_sg1[12]
4. For each cluster, build a multiple alignment based on the protein sequences using a combination of multiple aligners and obtain a consensus using M-Coffee[Wallace et al., 2006]
5. For each aligned cluster, build a phylogenetic tree using the TreeBeST and the coding sequence back-translation of the protein multiple alignment from the original DNA sequences. A rooted tree with internal duplication tags is obtained at this stage, reconciling it with its species tree
6. From each gene tree, infer gene pairwise relations of orthology and paralogy types.

In step number 5 a total of five different trees are built — this includes a DNA-based tree, more accurate for closely related parts of the tree and a protein-tree, often more accurate for distant relationships. These are then fused into a consensus tree using TreeBeST merging algorithm.

Overall, Ensembl Compara identifies two main types of homology association between genes, *orthologues* and *paralogues*. Two genes in two different species are in an orthologous relationship if they derive from a single gene present in their last common ancestor [Sonnhammer and Koonin, 2002]. On the other hand, two genes are called paralogues if they stem from a single gene that underwent duplication within a genome. Given this basic distinction, a number of sub-categories are defined (Figure 2.4):

Co-orthologues multiple genes in an organism that are simultaneously orthologues of a gene in another organism are termed *co-orthologues* of such gene;

In-Paralogues and Out-Paralogues paralogy can be further distinguished, relative to a speciation event, into an ancestral form, called *out-paralogy*, and into a recent form, called *in-paralogy*. Two or more genes are called out-paralogues with respect to a speciation event if the duplication that created them predates the speciation. On the other hand, two or more genes are called in-paralogues relative to a speciation event if the duplication that created them follows the speciation event.

[11] From www.ensembl.org/info/docs/compara/homology_method.html
[12] treesoft.svn.sourceforge.net/viewvc/treesoft/branches/lh3/hcluster/

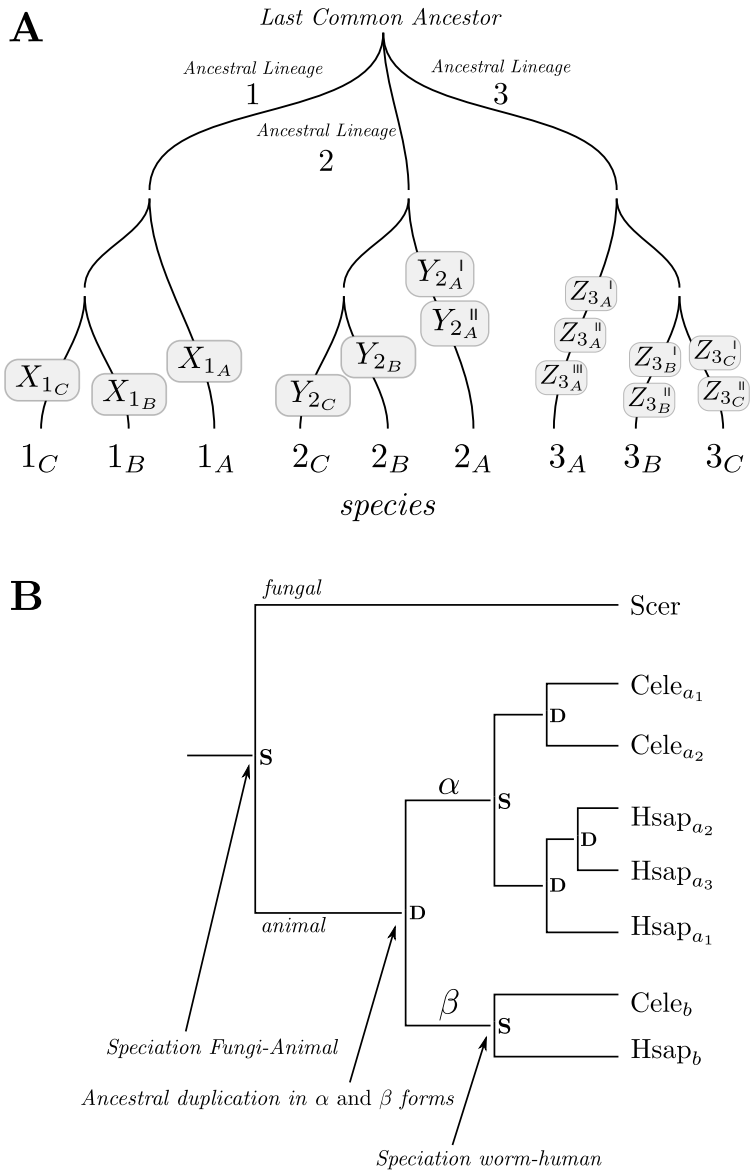

Figure 2.4: Basic orthology and paralogy definitions. **A**: 1:1, 1:many, many:many orthology examples. Each tree branching corresponds to a speciation event. From a common ancestor, three lineages, 1, 2, 3 are derived. Nine species appear as a result of speciation events in the three lineages. Gene X does not duplicate after the speciation events leading to 1_A, 1_B and 1_C — the three versions of this gene are 1:1 orthologues. By the same argument, genes Y_{2C} and Y^{1}_{2A} are 1:many orthologues, and genes Z^{1}_{3C} and Z^{11}_{3A} are many:many orthologues. **B**: In-paralogues, Out-paralogues and Co-orthologues. Each tree branching corresponds to S (speciation) or D (duplication). Let us consider an ancient gene inherited in *C. elegans*, *H. sapiens*, *S. cerevisiae*. Before the human/worm speciation, an ancestral gene duplication occurs. After the human-worm split, the α form is duplicated independently in the human and worm (unlike the β form). All human and worm genes are co-orthologues of the yeast one. All the genes in the Hsap_{a*} set are co-orthologues of all the genes in the Cele_{a*} set. Hsap_{a*} are in-paralogues of each other with respect to the human-worm speciation. Finally, Hsap_{a*} and Hsap_b are out-paralogues with respect to the human-worm speciation because the duplication that created them predates the speciation itself.

The concepts of in-paralogy and co-orthology are tightly related, and assume a particular significance within the context of any functional transfer study. Specifically, members of a co-orthology group are *in-paralogues* with respect to each other, and relative to the speciation event that created the orthology. Ensembl Compara will classify any gene α, for which at least one orthologue β is found, in one of the following categories:

ortholog one-to-one. α has not undergone duplication (better, according to the TreeBeST algorithm there are no duplicates after the speciation event) This corresponds to saying that α has no in-paralogues. Same for β.

ortholog one-to-many. α has not undergone duplication after the speciation event (α has no in-paralogues). However, β has undergone duplication after the speciation event and has a number of in-paralogues (i.e., β is part of co orthology group with respect to α).

ortholog many-to-one. α has undergone duplication after the speciation event, and has a number of in-paralogues. α is part of a co-orthology group with respect to β and only β.

ortholog many-to-many. Both α and β have undergone duplication after the speciation event. α, with all its in-paralogues, is an orthologue of β and of all its in-paralogues.

Within the context of this project we are particularly interested in *one-to-one* orthologues: these are orthology relationships where neither of the two members has (according to Compara) undergone duplication after the speciation event. It has been shown that gene duplication is related to neo-functionalisation and/or sub-functionalisation [Rastogi and Liberles, 2005]. The nature of orthologues and the classification introduced above represent a major decision step in an interolog-mapping algorithm. Multiplicity in homology relationships (1:many, many:many) and in-paralogy can represent a potential source of artefacts and noise in the final results. With 1:1 orthologues, functional conservation is more likely to be retained [Koonin, 2005; Hulsen et al., 2006], hence the need to discriminate between homology classes in the interolog-mapping implementation.

Based on the Ensembl annotation, `Bio::Homology::InterologWalk` can act in one of two ways:

- retain putative protein interactions in which *both* the orthology projections are of the one-to-one kind only, discarding all other classes;
- keep all putative protein interactions, regardless of class, and then optionally use prioritisation metrics to flag the predictions.

Ensembl also identifies cases tagged as *possible ortholog*: instances when the duplication vs. speciation nature of the phylogenetic tree could not be fully resolved by the Compara-TreeBeST algorithm, and partial evidence suggested the absence of a duplication event. As

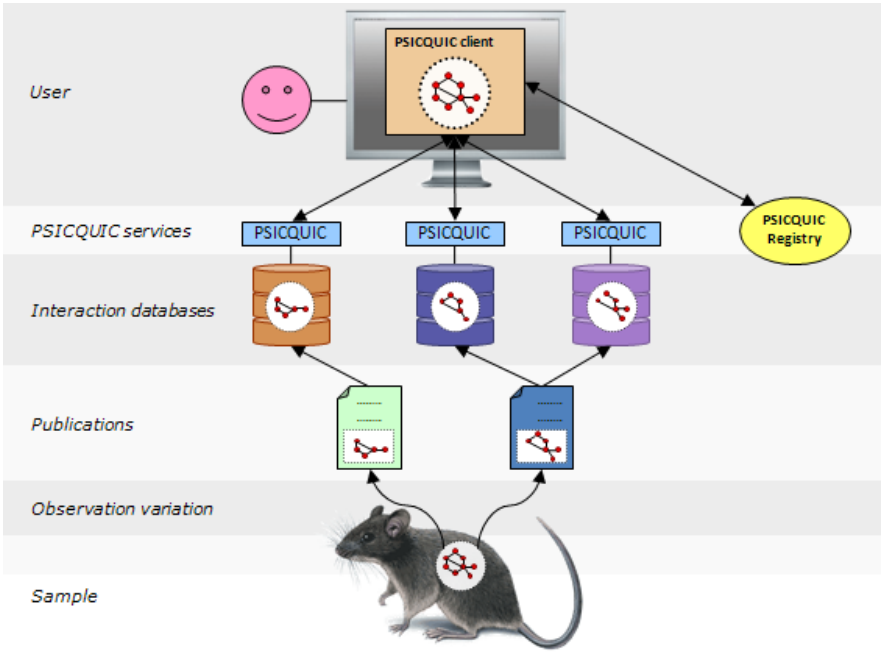

Figure 2.5: High level overview of the PSICQUIC web service architecture. PSICQUIC acts as a unifying interface to several biological data resources. It specifies a set of programming language-independent methods to query such data, as well as a query language called MIQL. [source: code.google.com/p/psicquic/]

reported by Ensembl, such cases might point to long distance relations that might be upgraded to *bona fide* orthologies in further versions of the gene tree pipeline.

2.2.2.2 HUPO PSICQUIC

The Proteomics Standard Initiative Common QUery InterfaCe [Aranda et al., 2011], also known as PSICQUIC, is an effort from the Human Proteome Organisation (HUPO) to standardise access to molecular interaction databases programmatically (Figure 2.5). The standard specifies the following:

1. A standard web service with a well defined list of methods, accessible using SOAP[13] or REST[14].

2. A common query language, MIQL[15], based on Lucene[16].

The standard provides unique designations for metadata terms specifying interaction detection method, taxon, interaction type, interactor type, database, interactor roles and more. Some of the protein interaction databases implementing the PSICQUIC specification are APID [Prieto

[13] code.google.com/p/psicquic/wiki/PsicquicSpec_1_0_Soap
[14] code.google.com/p/psicquic/wiki/PsicquicSpec_1_0_Rest
[15] code.google.com/p/psicquic/wiki/MiqlReference
[16] lucene.apache.org/core/

and De Las Rivas, 2006], iRefIndex [Razick et al., 2008], BioGrid [Breitkreutz et al., 2008], Reactome [Matthews et al., 2009], MPIDB[Goll et al., 2008], InnateDB [Lynn et al., 2008], MatrixDB [Chautard et al., 2009], STRING [Jensen et al., 2009], EBI IntAct [Aranda et al., 2010], MINT [Ceol et al., 2010] and ChEMBL[17]. In this project, I have selected EBI IntAct as the primary source of experimental protein interactions. The motivating factors for this choice will be detailed in the next section.

2.2.2.3 Protein Interaction Data from EBI IntAct

`Bio::Homology::InterologWalk` uses EBI IntAct [Aranda et al., 2010] as its source of experimental interactions. The IntAct platform is based on literature data and direct data deposition by expert curators, following a standardised mode, and a variety of data access methods is available. Furthermore, EBI IntAct is one of the oldest supporters and earliest adopters of the PSICQUIC interface and offers both SOAP- and REST-based data access. The data is organized according to the annotation rules defined in the HUPO-PSI Controlled Vocabulary [Hermjakob et al., 2004]. Additionally, IntAct PSICQUIC implementation is mature and matches the reference implementation very closely. As of June 2012, v. 1.2.0 of the IntAct database contains more than 285,000 curated binary protein interaction evidences[18].

I access IntAct data using its RESTful-based PSICQUIC implementation, and data is retrieved using the PSI-MI MITAB25 tab-delimited format [Kerrien et al., 2007]. There are several reasons why `Bio::Homology::InterologWalk` currently relies on EBI IntAct as its source of experimental interactions. A number of other PSICQUIC-enabled databases are mainly aggregators of interaction data found elsewhere (e.g. APID, MPIDB, MatrixDB, iRefIndex). Some are dedicated to collecting interactions belonging to specialised domains (e.g. InnateDB) while others contain data that is not suited for our purpose (e.g. ChEMBL). More importantly, some of the PSICQUIC-adopting data repositories were immediately discarded because they simply did not provide fully compliant and standardised data through the interface[19].

All protein interaction evidences in IntAct are binary entries: each data entry is a row identifying a binary interaction and its supporting evidence. However, some experimental methods,

[17] `www.ebi.ac.uk/chembl/`

[18] `www.ebi.ac.uk/Tools/webservices/psicquic/registry/registry?action=STATUS`

[19] As of December 2011, data in STRING and BioGRID, while technically compliant with the Mitab standard, is not equal to IntAct data in terms of MITAB column content. For instance, it is not possible to query these two data repositories using Ensembl IDs (the "properties" MITAB fields, which contain Ensembl IDs in the standard PSICQUIC Mitab, do not exist). Additionally, for MITAB field n. 12 (`Interaction Type`) String only offers the PSI-MI OBO ontology code for the interaction type (eg: `psi-mi:"MI:0190"`) unlike in the IntAct implementation, where the ontology concept name is present (eg: `psi-mi:"MI:0915"(physical association)`). These and other differences make the usage of some of the most advanced semantic-based features in `Bio::Homology::InterologWalk` inconvenient. However, Bruno Aranda, head of the PSICQUIC project, confirmed the existence of an ongoing effort to drive data providers to further standardise their interaction data within PSICQUIC (`code.google.com/p/psicquic/wiki/DataDistributionBestPractices`). This means all data fields should soon converge to a perfect standard, and automatic query of all dataset will be transparent for the algorithm presented here.

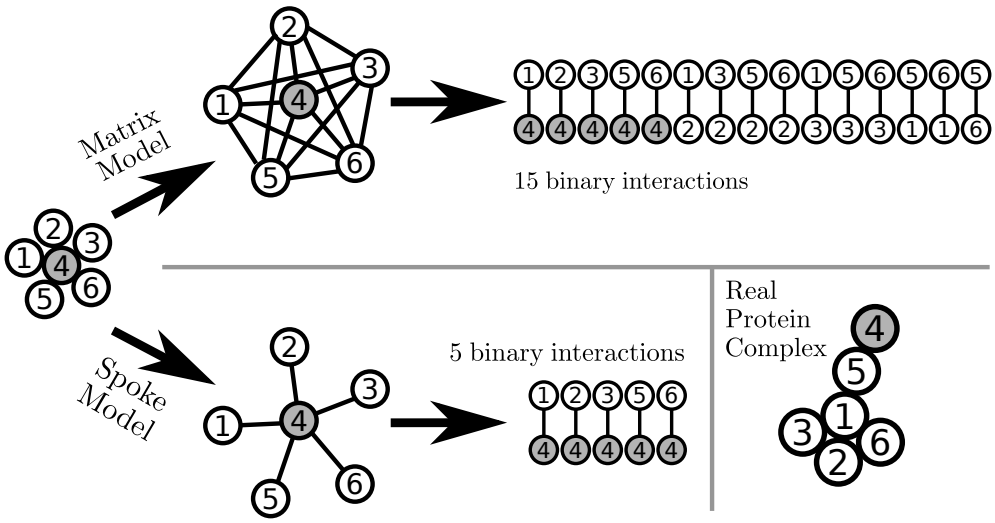

Figure 2.6: Computational expansion of protein complex interaction data. The real interaction pattern (bottom right) is unknown. Given experimental data on complex interaction (left, grey node is the bait, white nodes are preys), two choices can be made to obtain a set of binary interactions: In the *matrix* model, all possible interactions between the members of the complex are assumed. In the *spoke* model, only interactions between the bait and the rest of the proteins are assumed. EBI IntAct adopts the spoke complex expansion model. (Image adapted from Aranda et al. [2010])

such as Tandem Affinity Purification, generate molecular interaction predictions in the form of molecular complexes. When managing data relative to complexes of interacting proteins, protein interaction services make an assumption related to the decomposition of such complexes in binary associations. In particular, IntAct and other PSICQUIC-compliant databases utilise a computational complex expansion paradigm called *Spoke* (Figure 2.6).

In the Spoke model, the experimental results describing the purification of protein complexes are converted into pairwise interactions between bait and preys only. If the complex is composed of n proteins interacting in an unknown configuration, the Spoke model will generate $n-1$ binary associations. An alternative to the Spoke model would be a *fully connected matrix* model, which assumes all proteins in a complex to be connected to all others. Given an n-protein complex, the matrix model will generate $[n(n-1)]/2$ binary associations. Both methods have their shortcomings, and will report a certain number of false positives (fully connected matrix) or a smaller number of false positives but additionally false negatives (spoke expansion). One important piece of supplementary information provided by IntAct and processed by `Bio::Homology::InterologWalk` is the *complex expansion* flag. `Bio::Homology::InterologWalk` will deal with putative protein interactions derived from spoke-expanded complexes — and tagged accordingly by IntAct — in one of two ways: either by discarding them altogether, or by keeping them in the output dataset and penalizing them using a prioritisation index.

2.2.3 Implementation Details

2.2.3.1 Pipeline Schematics

Figure 2.7 shows the architecture of the data pipelines in `Bio::Homology::InterologWalk` in detail. Each dark block in Figure 2.7 represents a data collection function. In each block, the list of retrieved data fields is shown. Starting from the initial text file containing a plain list of Ensembl identifiers, two pipelines can be used to retrieve two different data sets. The experimental pipeline (Figure 2.7, *left*) is composed of two processing blocks. The first one is in charge of the actual connection to the remote experimental interaction dataset and with processing the MITAB data retrieved from remote. Once a complete local dataset is obtained, the second block post-processes it to calculate some useful statistics. As regards the putative pipeline (Figure 2.7, *right*), a number of additional processing blocks are introduced to deal with the orthology data retrieval from Ensembl. Up to three remote access steps are carried out for each input ID — two to Ensembl resources, and one to the IntAct resource. Once a "raw" local putative dataset is obtained, a further link can be added to the pipeline chain to refine the results through post-processing and ranking functions. A processing block common to both pipelines can then be added (2.7, *centre-bottom*) to obtain text file-based network representations of the output dataset.

2.2.3.2 Gene ID/Protein ID Conversion

`Bio::Homology::InterologWalk` retrieves Ensembl Compara orthology data using gene IDs. EBI IntAct, however, returns binary interaction information using UniprotKB protein identifiers. In order to return to Ensembl IDs for the backward part of the orthology retrieval, a conversion phase is required and Ensembl IDs must be obtained for each reference experimental interactor. Genes often produce several transcripts via alternative splicing, and these code for multiple protein isoforms. Currently, isoforms are mapped onto the parent gene. This is due to the absence of reliable methods to map isoforms between species, especially when their common ancestor is very distant. Additionally, there are currently no methods in the Ensembl API to obtain a gene object using a Uniprot ID complete with isoform information. Therefore, while the final data files stores Uniprot KB isoform information for the reference genome interaction, the projected interaction is at gene-level.

The module extracts Ensembl IDs from MITAB25 supplementary data fields provided by IntAct (when present). If such data is not available, a conversion algorithm is employed. The IDs in the IntAct data entry are generally preferable to avoid the computational burden associated with the ID look-up/conversion algorithm. However, I found that in some cases the gene IDs provided by IntAct pointed to obsolete or secondary IDs. As a consequence, `Bio::Homology::InterologWalk` gives the possibility of always double-checking gene ID consistency

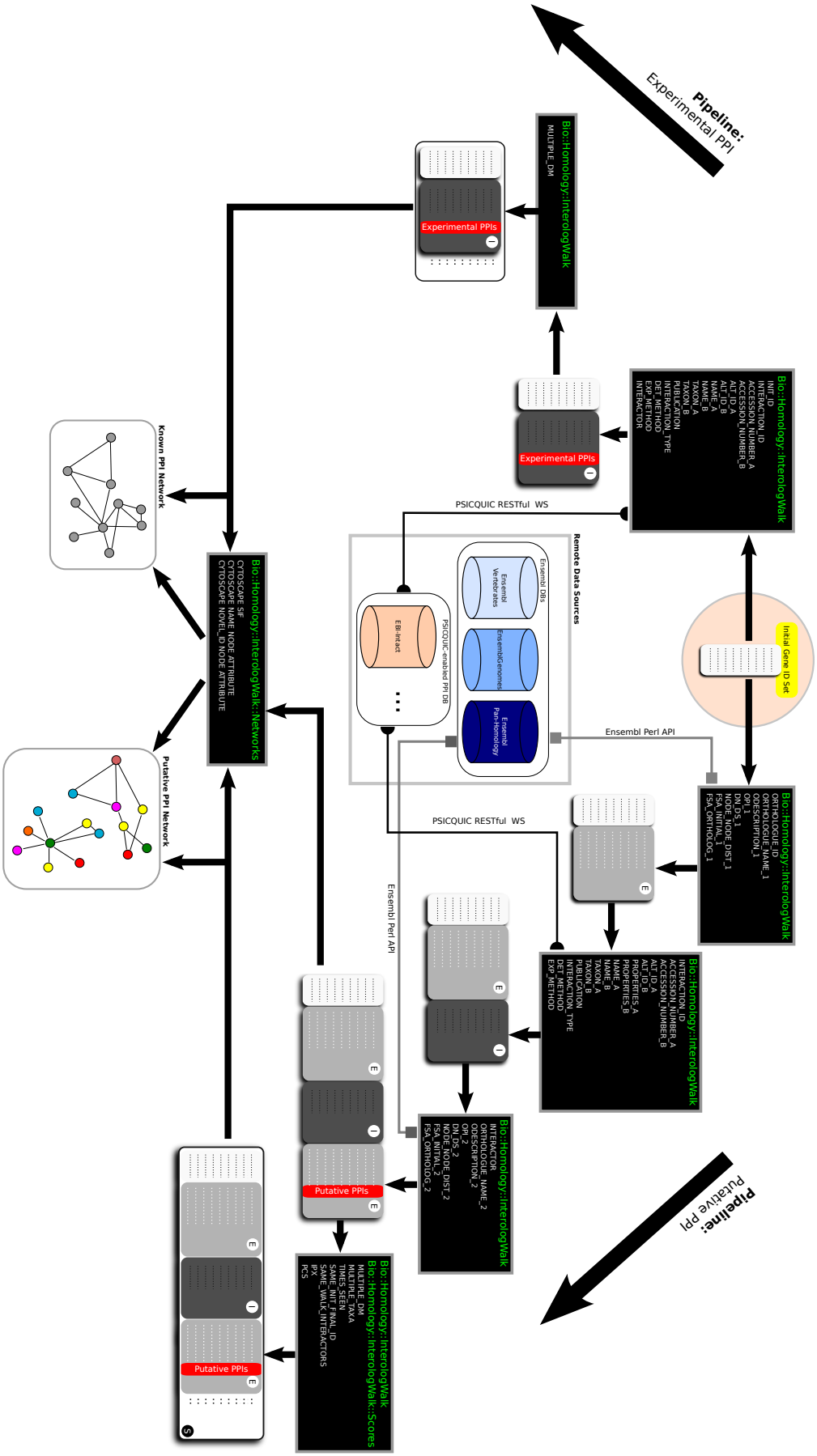

Figure 2.7: `Bio::Homology::InterologWalk` Pipelines.

against Ensembl. I relied on the conversion algorithm for the dataset analysis I discuss later in this dissertation.

2.2.3.3 Output Data Format

Details of the data format for the output produced by the module are in Table 2.1. `Bio::Homology::InterologWalk` stores its data in Tab Separated Value (TSV) text files. Each processing block in Figure 2.7 outputs temporary TSV files. Each block reads the output of the former block, and saves its output on disk for usage by the following block. All TSV data files are manipulated through a MySQL relational database Perl interface. This is based on the `Perl::DBI` module[20] and on its DBI driver for CSV files, `DBD::CSV`. The solution allows the usage of the full range of SQL data manipulation language capabilities on simple structured text files, which are small, readable by scripts and spreadsheet programs alike, easily distributed, easy to integrate in longer meta pipelines where interolog retrieval is only one of several steps in a longer data processing effort[21].

2.2.3.4 Using Semantic Similarity to Equalise PSI-MI Ontology Concepts

Both the putative and experimental pipelines collect, for each interaction, data about the interaction type and about the detection method used to detect it (Table 2.1). This information can be used by `Bio::Homology::InterologWalk` if desired to select, discard or penalise subsets of the results set, and it is used together with other supporting information as explained in the next two sections. Here, I would like to focus exclusively on how the information in the two fields `INTERACTION_TYPE` and `DET_METHOD` is manipulated in the present implementation. As described in Section 2.2.2.3, data distributed through any PSICQUIC-compliant implementation is organized according to the annotation rules defined in the HUPO-PSI ontology [Hermjakob et al., 2004]: `interaction type` and `interaction detection method` are two terms of this controlled vocabulary. Figure 2.8 shows the two sub-hierarchies branching down three levels from these two terms. The visualised data was extracted from the most recent version of the PSI-MI ontology[22]. Each of the two sub-trees is a hierarchy of concepts, with increasing level of semantic specialisation, from top to bottom. Figure 2.8-A shows how the generic "interaction detection method" is a generalisation of four more specific concepts. These are[23]:

- `Experimental interaction detection` Methods based on laboratory experiments to determine an interaction.

[20] `dbi.perl.org/`
[21] e.g. Taverna Workflows.
[22] `www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI`, v. 1.2, 02/2012.
[23] The definitions in the bullet list are taken from the actual HUPO PSI-MI ontology.

Stage	Field Designation	Description	Value type
FOP	INIT_ID*	Query Gene ID (in query genome)	Ensembl Gene ID
	ORTHOLOGUE_ID	Orthologue Gene ID (in ref genome)	Ensembl Gene ID
	ORTHOLOGUE_NAME_1	Orthologue Name	Ensembl Gene Name
	ODESCRIPTION_1	Co-orthologue configuration	[1:1|1:many|many:1|many:many]
	OPI_1	Overall Percentage Identity	Float
	DN_DS_1	dN/dS ratio	Float
	NODE_NODE_DIST_1	Node to Node Dist.	Float
	FSA_INITIAL_1	Distance from First Shared Ancestor (query gene)	Float
	FSA_ORTHOLOG_1	Distance from First Shared Ancestor (orth. gene)	Float
EPIR	INTERACTION_ID*	Protein Interaction ID	EBI IntAct ID
	ACCESSION_NUMBER_A*	Accession Number Prot. A	UNIPROT KB ID
	ACCESSION_NUMBER_B*	Accession Number Prot. B	UNIPROT KB ID
	ALT_ID_A*	Alternative ID(s) Prot. A	databaseName:aid
	ALT_ID_B*	Alternative ID(s) Prot. B	databaseName:aid
	PROPERTIES_A*	Properties Prot. A	databaseName:prop
	PROPERTIES_B*	Properties Prot. B	databaseName:prop
	NAME_A*	Protein name A	databaseName:n
	NAME_B*	Protein name B	databaseName:n
	TAXON_A*	NCBI Taxon ID A	NCBI ID
	TAXON_B*	NCBI Taxon ID B	NCBI ID
	PUBLICATION*	Publication ID	databaseName:id
	INTERACTION_TYPE*	Interaction Type	dataBaseName:id(interactionType)
	DET_METHOD*	Detection Method	databaseName:id(methodName)
	EXP_METHOD*	Experimental Method	[spoke|-]
BOP	INTERACTOR*	Interactor Gene ID	Ensembl Gene ID
	ORTHOLOGUE_NAME_2	Putative Interactor name	Ensembl Gene ID
	ODESCRIPTION_2	Co-orthologue configuration	[1:1|1:many|many:1|many:many]
	OPI_2	Overall Percentage Identity	Float
	DN_DS_2	dN/dS ratio	Float
	NODE_NODE_DIST_2	Node to Node Dist.	Float
	FSA_INITIAL_2	Distance from First Shared Ancestor (query gene)	Float
	FSA_ORTHOLOG_2	Distance from First Shared Ancestor (orth. gene)	Float
PP	MULTIPLE_DM*	Observed through multiple Det. Methods?	Int
	MULTIPLE_TAXA	Observed in multiple reference genomes?	Int
	TIMES_SEEN	Times putative PI seen?	Int
	SAME_INIT_FINAL_ID	Putative autointeraction?	Bool
	SAME_WALK_INTERACTORS	Mapped from autointeraction?	Bool
IPX	IPX	IPX	Float
PCS	PCS	PCS	Float

Table 2.1: `Bio::Homology::InterologWalk` output data fields for the putative pipeline and (* only) the experimental data pipeline. **FOP**: Forward Orthology Projection. **EPIR**: Experimental Protein Interaction Retrieval. **BOP**: Backward Orthology Projection. **PP**: Post Processing. **IPX**: Interaction Prioritisation Index. **PCS**: Protein Conservation Score.

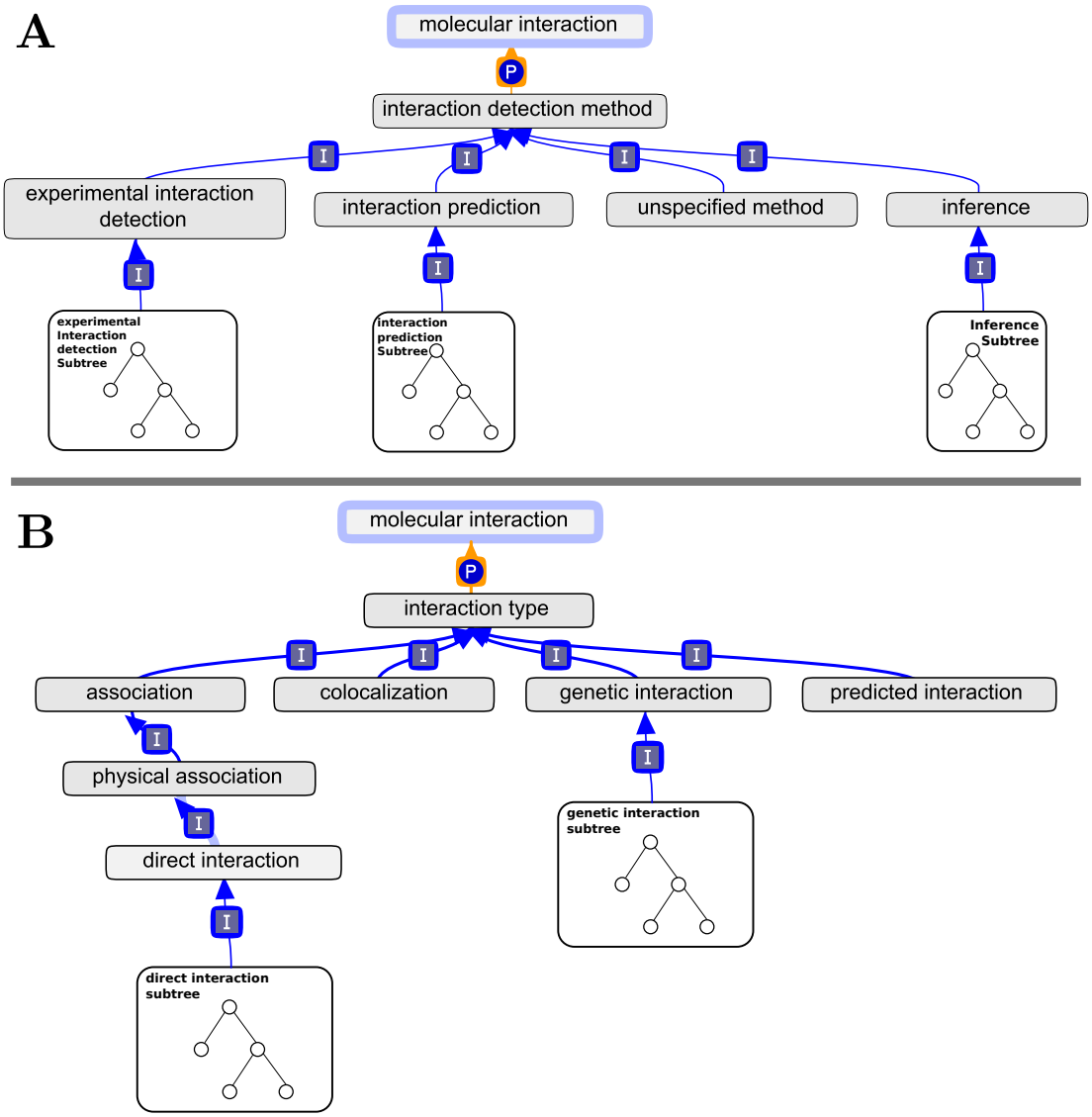

Figure 2.8: Two sub-hierarchies extracted from the HUPO PSI-MI ontology. **A**: `Interaction Detection Method` sub-tree. **B**: `Interaction Type` sub-tree. The PSICQUIC specification relies on this ontology to label all binary interaction occurrences in PSICQUIC-compliant databases. For each putative interaction retrieved, `Bio::Homology::InterologWalk` analyses the *interaction type* and *detection method* supplementary field values climbing up the ontology to find a matching concept and its parents up to the level visualised in figure. The sub-trees visualised here are used to inform the interaction filtering sub-system (Section 2.2.4, Page 41) and the IPX (Section 2.2.5, Page 42).

- `Interaction prediction` Computational methods to predict an interaction.
- `Unspecified method` Yet to be identified interaction detection method associated with interaction data imported from a third party database. This database may have potentially different standards of curation.
- `Inference` Evidence based on human assumption, either when the complete experimental support is not available or when the results are extended by homology to closely related orthologues sequences.

As regards Figure 2.8-B, the concept `interaction type` is further specialised into:

- `association` Molecules that are experimentally shown to be associated potentially by sharing just one interactor. Often associated molecules are co-purified by a pull-down or co-immunoprecipitation and share the same bait molecule.
- `colocalization` Coincident occurrence of molecules in a given sub-cellular fraction observed with a low resolution methodology from which a physical interaction among those molecules cannot be inferred.
- `genetic interaction` Two genes A and B "genetically interact" when the phenotype generated as the result of mutations in both genes (double mutant ab) is unexpectedly not just a combination of the phenotypes of the two single mutants a and b.
- `predicted interaction` Interaction has been predicted by either [*sic*] interolog mapping, by an algorithm or by a computational method.

Most of these concepts are further specialised within the ontology (this is indicated by a sub-tree image under a concept in Figure 2.8): for example, a specialisation of the generic experimental interaction detection can be `biochemical`, which can be further specialised in `affinity technology` then `solid phase assay` and then `bead aggregation assay`.

All the experimental interactions provided by a PSICQUIC service will be labelled by one detection method and one interaction type concept, which can belong to any level of the ontology. `Bio::Homology::InterologWalk` can analyse the two fields and climb the ontology up to the level shown in Figure 2.8 to infer the basic nature of the detection method and interaction type for the interaction examined. This is done to clearly classify an interaction as, for instance, experimental and physical/binary, as opposed to a less reliable computationally annotated prediction, or a colocalisation. The way this information is used by the algorithm is described in the next two sections.

2.2.4 Prioritisation of the Putative Interactions: Filtering

Depending on the size of the input dataset and on the amount of information available through homology mapping, `Bio::Homology::InterologWalk` can produce large numbers of putative interactions. In such cases it may be beneficial to filter and prioritise these in order to generate a smaller set of results for further study. As described earlier, the module is composed of a number of functions that can be executed in sequence to create pipelines for retrieving interologs. The output of this sequence of subroutines is a set of tab separated files containing one entry per line and closely resembling the MITAB tab delimited data exchange format from the HUPO PSI. Each row in the data files describes a binary putative interaction, plus 39 supplementary data fields (Table 2.1). The values of these supplementary data fields can be used to select subsets of the data based on specific requirements (such as thresholds on continuous variables or flag values on boolean variables). In the following subsection, I describe two possible prioritisation strategies that, I believe, may help finding interesting putative interactions in a wide number of cases. The following are mainly suggestions, and a specific strategy must be implemented on a case-by-case basis.

The following four metadata descriptors have been chosen to implement a filtering strategy in `Bio::Homology::InterologWalk`:

Spoke Interactions

The user can choose whether to return any "spoke" interactions when using interaction retrieval functions. As discussed in Section 2.2.2.3 (Page 33), Spoke interactions are binary interactions inferred from a complex of proteins that have been isolated together and as such the evidence for the interaction is indirect. Several of the most widely used interaction data repositories, including IntAct and BioGrid, explicitly draw the user's attention to the presence of spoke (or co-existence) interactions and provide the option of excluding them at an early stage. As a consequence, I have decided to extend this to putative interologs obtained from spoke interactions.

One-to-one Orthology

For each of the orthology mapping functions (forward orthology mapping and backward orthology mapping) the user can choose whether to restrict the mapping to explicit 1:1 relationships. This is likely to significantly reduce the number of orthologues retrieved as the evolutionary distance between mapped species increases. Restricting mappings to direct orthologues increases the likelihood that the mapped proteins retain some common functionality. Conversely considering *1-to-many* or *many-to-many* relationships that have arisen through duplication events risks connecting proteins and interactions whose functions have diverged [He and Zhang, 2005; Hittinger and Carroll, 2007]. Ensembl Compara explicitly draws the user's

attention to the nature of each orthology relationship it generates. I extend this to the putative interactions generated by the module.

Experimental Interactions

The user can specify whether to restrict the interactions retrieved to those that have been identified by experimental methods rather than by inference, prediction or an unspecified method (see Section 2.2.3.4 and Figure 2.8). This will discard computational prediction obtained with dubious or non trusted methods and make sure these will not inform the interolog predictions. It will also discard interactions obtained through other homology transfer algorithms.

Physical Interactions

The user can choose to retrieve only those interactions that test for physical association between proteins in the HUPO ontology, discarding generic unspecified associations, colocalisations, genetic interactions and predicted interactions (Figure 2.8). As is the case with the Experimental Interaction filter, usage of this filter will restrict the number of returned results, however it will make sure only experimental interactions annotated as representing physical protein interactions will be used to inform interolog predictions.

2.2.5 Prioritisation of the Putative Interactions: IPX

Filtering a result set based on the parameters described above can be sufficient to generate a small enough number of interaction hypotheses. However, in cases where the putative interaction data set is particularly large, it might be beneficial to further restrict the result set based on these and additional metadata features describing the orthology and interaction data. For this reason, I have created an *Interaction Prioritisation indeX* (IPX) and a *Protein interaction Conservation Score* (PCS). These can optionally be used in addition to filtering, or directly on the unfiltered data. The present section will describe the IPX, while the PCS will be introduced later (after a brief discussion of a number of relevant graph-theoretical concepts).

The IPX combines the contribution of several pieces of heterogeneous information collected during orthology projection and interaction retrieval. Figure 2.9 provides a general overview on which supplementary data fields in Table 2.1 are employed to build the IPX. When dealing with the Ensembl-based sub-components of the algorithm, I have used the Ensembl API and the BioPerl API to build the required supplementary metadata features in Table 2.1 if they were not already pre-calculated and stored in the database. In the case of IntAct, when these indicators were not presented explicitly, I have exploited some of the metadata structural properties (described in Section 2.2.3.4, Page 37) to derive them.

In the following paragraphs I shall describe an attempt at integrating the information of

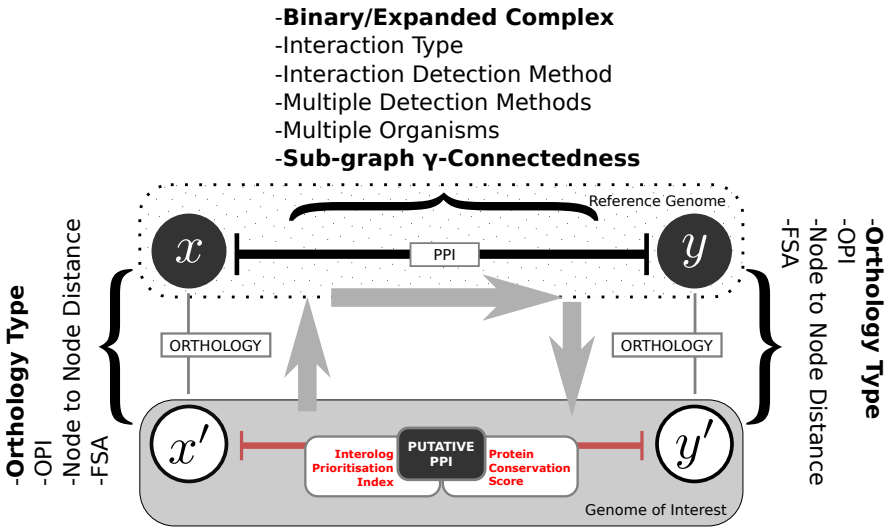

Figure 2.9: Schematic summarizing the features used to prioritise the putative protein interactions. For each putative interaction, a number of metadata fields are collected during the main steps of the algorithm. Two metrics can optionally be computed: an *Interolog Prioritisation Index* (IPX) and a *Protein interaction Conservation Score* (PCS).

these indicators together in such a way that each piece of evidence participates with a contribution of equal magnitude: due to the heterogeneity of the information collected, no conclusion must be drawn on the relative importance of the data fields. Therefore, an *agnostic* approach was chosen: each of the data fields should weight as much as any other.

In my implementation, the IPX is not intended to be a quantitative measure of interaction reliability, but rather an integration of biological information which can be used to single out highly supported evidences. The potential for integration of biological metadata to highlight interesting predictions has been explored before, for instance in the work of Huang et al. [2007] and Yu et al. [2004]. Yu et al. used sequence similarity between the orthologous proteins to build a joint similarity score, while Huang et al. proposed a scoring framework based on GO functional annotation, domain information, tissue specificity and sub-cellular localisation to rank interolog-based human putative protein interactions obtained from six eukaryotes. Another combination of biological features in meta-indicators was reported for example by Huang et al. [2007].

Following is a description of the prioritisation features I shall consider. Those related to the two orthology projections are:

Orthology Type

The kind of orthology relationship existing between an ID in the genome of interest and its orthologue in the reference genome. This feature indicates if there is a one-to-one mapping

of orthologues, or if in-paralogy events in one or both sides mean we are considering a one-to-many, many-to-one or many-to-many orthologous mapping (Figure 2.4, Page 30). As explained in the filtering section, we particularly value putative interactions where *both* orthology relationships are of the one-to-one kind. It has been shown [He and Zhang, 2005] that gene duplication is correlated with sub-functionalisation and neo-functionalisation. When the two orthologous pairs in the interolog walk are of the one-to-one kind I set a boolean variable, Θ, to a non-negative value in the score. I set $\Theta = 0$ otherwise.

OPI

OPI stands for Overall Percentage Identity. I utilise the method implemented in the BioPerl module `Bio::SimpleAlign`[24] to obtain overall percentage identity values for all putative interologs. BioPerl defines an OPI as the percentage identity of the identical columns between the orthology members' sequences. Compared to average sequence identity approaches, OPI is a more conservative choice[25]. Given N total samples, I define a *Joint OPI* as the geometric mean of the two OPIs (forward and backward orthology projection)

$$\mathbf{J}_{\mathrm{OPI}}^{(i)} = \sqrt{\mathrm{OPI}_1^{(i)} \times \mathrm{OPI}_2^{(i)}} \qquad \forall i \in 1,\ldots,N. \tag{2.1}$$

Node to Node Distance

A numerical indicator of the node-to-node distance in the consensus phylogenetic/species tree built by Ensembl Compara using Genetrees [Vilella et al., 2009] (Figure 2.10-A).

We consider

$$\mathbf{J}_{\mathrm{nnD}}^{(i)} = 1 - \frac{\max(\mathrm{nnD}_1^{(i)}, \mathrm{nnD}_2^{(i)})}{\mathrm{nnD}_{\max}} \qquad \forall i \in 1,\ldots,N, \tag{2.2}$$

where nnD_1 is the node-to-node distance between the two orthologues in the forward projection, nnD_2 is the node-to-node distance between the two orthologues in the backward orthology projection and I set

$$\mathrm{nnD}_{\max} = \max\left(\mathrm{nnD}_1^{(1)},\ldots,\mathrm{nnD}_1^{(N)},\mathrm{nnD}_2^{(1)},\ldots,\mathrm{nnD}_2^{(N)}\right). \tag{2.3}$$

FSA

A numerical indicator of the distance between the entry and the orthology pair's First Shared Ancestor in the consensus phylogenetic tree built by Compara/TreeBeST (Figure 2.10-A).

[24] doc.bioperl.org/releases/bioperl-1.2/Bio/SimpleAlign.html

[25] Because it only considers amino acids that are identical over all the members of the alignment, and then averages over the length of the multiple sequence alignment, including gaps.

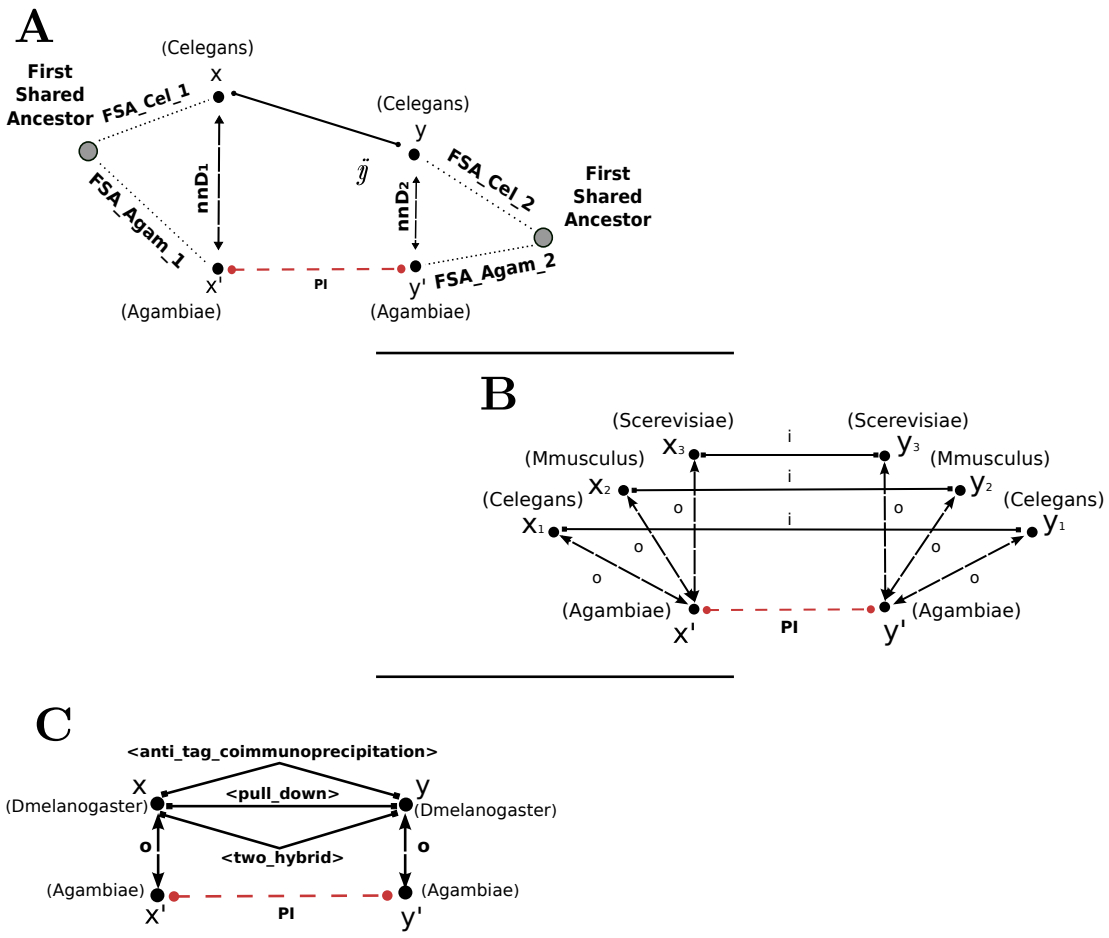

Figure 2.10: Some of the prioritisation features. **A**: Phylogenetic distances (according to TreeBeST). For each of the two orthologous pairs, a node-to-node distance (nnD_i) and two distances from the First Shared Ancestor (FSA_i^j) are computed. **B**: experimental interaction observed in multiple taxa — a component of the IPX is proportional to the number of reference genomes contributing to a putative interaction evidence. **C**: experimental interaction reconfirmed through multiple detection methods — a component of the IPX is proportional to the number of detection methods used to obtain experimental interaction evidence in the reference genome.

dN/dS Ratio

The ratio between the rate of non-synonymous and the rate of synonymous substitutions in the sequences. Also known as the K_a/K_s ratio, it can be interpreted as a measure of the evolutionary pressure acting on two orthologous genes [Kimura, 1991; Yang and Bielawski, 2000]. The following is usually held true:

1. $d_N/d_S \longrightarrow \infty$ strong positive selection
2. $d_N/d_S \longrightarrow 0$ strong stabilising selection
3. $d_N/d_S \approx 1$ some parts of the sequence under positive selection, some under stabilising selections so that overall effects cancel. In some cases, $d_N/d_S \approx 1$ is taken as an indicator for *neutral* selection.

Ensembl only calculates and provides dN/dS values for high coverage, closely related pairs of species: when the evolutionary distance between the two sequences is too large, the saturation of the dS values biases the estimated dN/dS ratio. Due to this, while the current release of `Bio::Homology::InterologWalk` collects dN/dS ratios whenever they are present, it ignores their contribution during calculation of the IPX. I decided to collect and keep dN/dS data, whenever available, due to its potential usefulness in scenarios where only recently diverged sequence data is considered (for example, an interolog walk run utilising *Drosophila melanogaster* as the reference genome and *Drosophila pseudoobscura* as the query genome).

As regards the interactions collected from the reference genome, I evaluate the following indicators:

Expanded Complex

Indicates whether the binary interaction has been extracted from a complex using the spoke expansion model. A boolean non negative term, Σ, is added to the score to reward each true binary interaction. $\Sigma = 0$ for spoke-expanded binary interactions.

Interaction Type & Interaction Detection Method

PSI-MI controlled vocabulary terms indicating, respectively, the type of interaction and the detection method used within the HUPO PSI-MI hierarchy. Terms contributing to the prioritisation index are shown in Table 2.2. If an interaction is annotated with a term that represents a specialisation of those in Table 2.2, `Bio::Homology::InterologWalk` climbs the hierarchy until one of terms in the table is reached. The protein interaction is labelled accordingly.

Interaction Type		Detection Method	
MI:0403	colocalization	MI:0045	experimental detection
MI:0208	genetic interaction	MI:0362	inference
MI:0914	association	MI:0063	interaction prediction
MI:0915	physical association	MI:0686	unspecified method
MI:0407	direct interaction		

Table 2.2: HUPO PSI-MI 2.5 Ontology Terms used to discriminate during the scoring phase of `Bio::Homology::InterologWalk` pipeline.

Protein interactions obtained with Multiple Methods & annotated in Multiple Organisms

This feature acknowledges experimental protein interactions reconfirmed through the usage of further detection methods and/or observed in multiple reference genomes (Figures 2.10-B and 2.10-C). The feature is based on evidence showing that experimental interactions reconfirmed through the usage of multiple detection methods and observed in multiple reference genomes potentially provide a stronger platform for functional transfer [Matthews et al., 2001; von Mering et al., 2002; Lehner and Fraser, 2004].

An Interaction Prioritisation IndeX

Overall, the **Interolog Prioritisation indeX** (IPX) is given by

$$\mathbf{IPX}^{(i)} = \omega_o \left[\mathbf{S}_{\mathrm{ORT}}^{(i)} + \Theta^{(i)}\right] + \omega_i \left[\mathbf{S}_{\mathrm{PPI}}^{(i)} + \Sigma^{(i)}\right] \qquad \forall i \tag{2.4}$$

where $\mathbf{S}_{\mathrm{ORT}}$ is the contribution to the IPX given by the orthology related parameters (equations 2.1 and 2.2)

$$\mathbf{S}_{\mathrm{ORT}}^{(i)} = \mathbf{J}_{\mathrm{OPI}}^{(i)} + \mathbf{J}_{\mathrm{nnD}}^{(i)} \qquad \forall i \tag{2.5}$$

and $\mathbf{S}_{\mathrm{PPI}}$ is the contribution to the IPX given by the normalised protein interaction related parameters

$$\mathbf{S}_{\mathrm{PPI}}^{(i)} = \frac{i^{(i)}}{\overline{I}_{\mathrm{dir}}} + \frac{d^{(i)}}{\overline{D}_{\mathrm{dir}}} + \frac{m_{\mathrm{dm}}^{(i)}}{\overline{M}_{\mathrm{dir}}} + \frac{m_{\mathrm{taxa}}^{(i)}}{\overline{M}_{\mathrm{taxa}}} \qquad \forall i. \tag{2.6}$$

Equation 2.6 agglomerates the terms relative to the protein interaction in the reference organism: i is a feature scoring the interaction type and d is a feature scoring the interaction detection method. m_{dm} acknowledges those experimental interactions present in the database more than once, with different detection methods (Figure 2.10-C). m_{taxa} is set to the number of reference genomes that possess an experimental interaction projecting back to the same putative interaction (Figure 2.10-B). The four features are normalised to make sure their values are comparable. The normalisation parameters in Equation 2.6 are obtained as follows:

- $\overline{I}_{\text{dir}}$, $\overline{D}_{\text{dir}}$, $\overline{M}_{\text{dir}}$ — mean values computed for the dataset of experimental interaction obtained from the initial gene list;

- $\overline{M}_{\text{taxa}}$ — this parameter cannot be obtained from the dataset of real interactions involving the starting gene set, where no projection to other organisms is involved and no statistics about taxa information is available. In order to normalise $m^{(i)}_{\text{taxa}}$ with a suitable value, `Bio::Homology::InterologWalk` randomly chooses N genomes from the Ensembl pool, and samples m random genes for each of them. For each of the N random gene sets the full interolog walk algorithm is run and putative protein interactions are retrieved. A mean $\overline{M}_{\text{taxa}}$ is computed for each so that

$$\overline{M}_{\text{taxa}} = \frac{\overline{M}^{r_1}_{\text{taxa}} + \cdots + \overline{M}^{r_i}_{\text{taxa}} + \ldots \overline{M}^{r_N}_{\text{taxa}}}{N} \tag{2.7}$$

  where we let $1 \leq m \leq 7$ (using at most 7 well represented taxa to draw random protein interactions from) and, given G genes in the initial query input file, $N = \min\{500, G\}$.

ω_i and ω_o are balancing weights for the two contributors in Equation 2.5 and 2.6. I set $\omega_i = \omega_o = 1$. Optimisation of these two weights based on training data will allow to reward either the interaction component or the orthology component of the score to optimise performance on a case-by-case basis. Lastly, Σ and Θ are boolean terms and I set $\Sigma = 0$ whenever the putative interaction has been inferred from a binary interaction derived from a spoke expanded complex ($\Sigma = n$, where $n > 0$ is an integer, otherwise), while $\Theta = n$ whenever the putative interaction has been inferred based exclusively on one-to-one orthology paths ($\Theta = 0$ otherwise).

Σ and Θ are boolean flags and unlike all other terms in Equation 2.4, they are not normalised. This is done to obtain a gross selection of putative interaction samples based on co-orthology/no co-orthology and spoke/no spoke information, prior to looking at other secondary metadata features. The value n was chosen to be the smallest integer bigger than the maximum spread of the distribution of the normalised IPX features. The IPX is composed of 6 features, $\mathbf{f} = [i, d, m_{dm}, m_{\text{taxa}}, \mathbf{J}_{\text{OPI}}, \mathbf{J}_{\text{nnD}}]$, where $0 \leq f_i \leq 1, \forall i \in 1, \ldots, 6$ and so $n = 7$.

Allowing Θ and Σ to be one order of magnitude bigger than other IPX features means the IPX distribution will take a roughly three-modal shape (Figure 2.11), depending on the combinatorial values of Σ and Θ, as follows:

1. $\Sigma = 0$, $\Theta = 0$ (*Low Tier*) — the experimental interaction is spoke-expanded and at least one of the two orthology projections is not one-to-one.

2. $(\Sigma = n, \Theta = 0) \vee (\Sigma = 0, \Theta = n)$ (*Mid Tier*) — either the experimental interaction is spoke expanded or at least one of the two orthology projections is not one-to-one

3. $\Sigma = n$, $\Theta = n$ (*High Tier*) — the experimental interaction is not expanded from a spoke-complex and the orthology projections are both one-to-one.

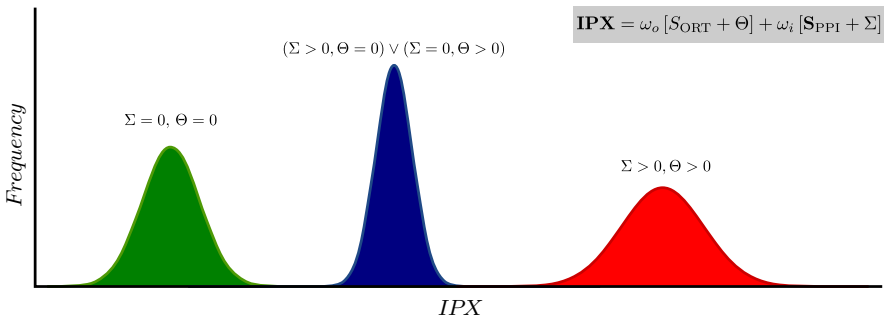

Figure 2.11: A graphical depiction of the typical tri-modal distribution obtained for the IPX values over Ensembl/IntAct interolog data. The three modes are due to the values taken by the boolean pre-selection variables, Θ (non-zero only for 1:1 interolog walks) and Σ (zero for experimental interactions obtained from spoke complex expansion). Unlike all other score features in Equation 2.4, Θ and Σ are not normalised. This is what generates the three modes. The green curve shows the distribution of the IPX for potentially unreliable interologs, while the red curve represents the distribution of the IPX for potentially more reliable interologs. Within each curve, better performing interologs are identified by better performance for the other scoring features in Equation 2.4.

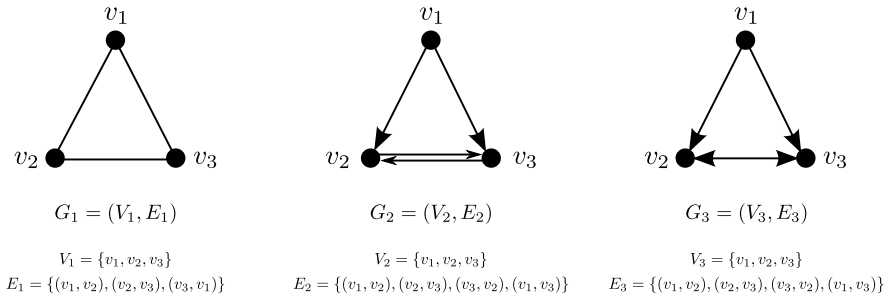

Figure 2.12: $G_1 = (V_1, E_1)$ is an undirected graph, while $G_2 = (V_2, E_2)$ is a directed graph. G_3 is an alternative visualisation for G_2.

Visual inspection of the modes in the IPX distribution can be used as strategy to filter out different sets of putative interactions, depending on the dataset considered and on the distribution of samples within the modes of the histogram. The choice of n provides good visual separation of the modes in the IPX distribution to facilitate inspection.

2.2.6 Graph Quasi-Completeness

A popular method to analyse the global properties of a protein interaction network is to model it as a *graph*. A graph is an abstract representation of a group of concepts or entities, usually called *nodes* (or *vertices*, or *points*) connected by links, called *edges*. Depending on the kind of network being modelled, either *directed* or *undirected* graph representations can be used. Figure 2.12 shows an example of directed versus undirected graph. In a directed graph, each edge has directionality attributes, meaning that the nodes it connects to are a source node and a target node (or both simultaneously). A directed graph is used when it is necessary to describe

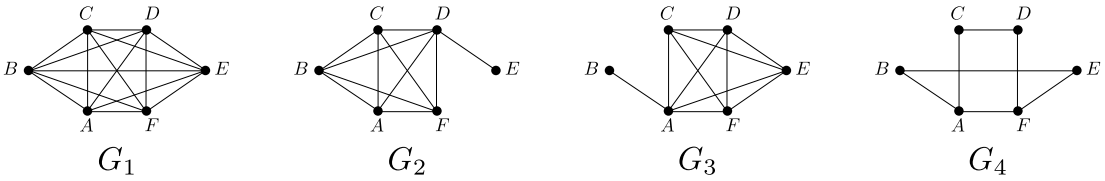

Figure 2.13: $\{A,C,D,F\}$ is a clique conserved in graphs G_1, G_2 and G_3.

some form of asymmetry or directional influence between the domain entities.

Undirected graphs on the other hand do not encode directionality information and are often used to model protein interaction networks. In this kind of abstraction, a graph node usually represents a protein[26] and an edge denotes the presence of an interaction between the nodes it connects to. Using graph models to describe protein interactions has allowed the application of the graph theory toolset to understanding protein networks. Many algorithms developed in graph theory to discover communities of nodes have found a natural application in the realm of protein interaction studies. The general context of a protein in its network, the number of its neighbours, the connectedness of its neighbours to one another, the information flow through selected nodes or edges, are examples of protein network signatures that have been quantified with the help of graph theory.

Of particular importance within the context of this study is the graph-theoretical definition of *clique*: let $G = (V,E)$ denote an undirected graph, defined by a set V of $|V|$ nodes and a set E of $|E|$ edges. A clique in G is a subset of the node set $\mathcal{C} \subseteq V$, such that for every two nodes in $\mathcal{C}$ there is a connecting edge. That is equivalent to saying that the sub-graph induced by $\mathcal{C}$ is complete (Figure 2.13). As discussed in Section 2.1, the study of densely connected sub-graphs in a network can lead to insights into the existence of interaction modules, or the functional relatedness between the proteins that compose a cluster. One way to model protein interaction modules is to adopt the clique perspective: several studies have investigated methods to scan protein interaction network to enumerate all cliques [Spirin and Mirny, 2003] or optimised algorithms to find cliques in protein networks [Ding et al., 2005].

Using cliques as a method to detect communities of related molecules in networks has drawbacks. A clique is a "perfect" abstract entity, a complete sub-graph. However, in most cases protein modules will not be composed by fully connected molecules. This can be either because not all proteins in the cluster interact with all other proteins in the same cluster, or because even if the unknown, true biological cluster is fully connected, data is not available for all interactions. What is needed is therefore a definition for an abstract entity able to describe protein modules that are *almost* fully connected, but not quite. In other words, it might be interesting to discover groups of proteins that are almost a clique, where each protein in the

[26] in some cases, a protein domain.

group is connected to at least a portion γ ($0 < \gamma < 1$) of the other proteins — with γ being a user-specified parameter. I follow Pei et al. [2005] in the notation and define the γ-*quasi-complete graph* and the γ-*quasi-clique* as follows

Definition 1 *A connected graph G is a γ-quasi-complete graph ($0 < \gamma \leq 1$) if every node in the graph has a degree of at least $\gamma \cdot (|V(G)| - 1)$.*

In a graph G, a subset of nodes $\mathcal{S} \subseteq V(G)$ is a γ-quasi-clique ($0 < \gamma \leq 1$) if $G(\mathcal{S})$ is a γ-quasi-complete graph, and no proper superset of $\mathcal{S}$ has this property.

If we let

$$\delta(G) = \min_{n \in V(G)} \deg(n) \tag{2.8}$$

it follows from definition 1 that if graph G is γ-quasi-complete the following inequality holds:

$$\gamma \leq \frac{\delta(G)}{(|V(G)| - 1)}. \tag{2.9}$$

Example 1 *(Quasi-complete graph) Consider graph G_3 in Figure 2.14-A. It is a 0.75-quasi-complete-graph, since $0.75 \leq 3/(5-1)$. Following Definition 1, it can also be noted that every node in G_3 has a degree of at least $0.75(5-1) = 3$. G_3 is closer to being a clique than G_1 and G_2, because in G_3 every node is connected at least to 75% of the remaining nodes.*

A clique is a special case of a γ-clique where $\gamma = 1$. In γ-quasi-complete graphs the parameter γ controls the density of the graph — larger values of γ correspond to denser graphs.

It is of particular interest to use the idea of γ-density in reverse: given a connected graph, we wish to find its maximum γ value. From Equations 2.8 and 2.9 it follows that

$$\gamma_{\max} = \max_{n \in V} \{\gamma(n)\} = \frac{\delta(G)}{|V(G)| - 1}. \tag{2.10}$$

The definition of γ density in Equation 2.10 differs from the classical concept of *graph D-density* for undirected networks. Given an undirected graph $G = (V, E)$ its D-density is defined by

$$D = \frac{|E(G)|}{\max\{|E(G)|\}}. \tag{2.11}$$

It can be proven that $\max\{|E(G)|\} = |V(G)|(|V(G)| - 1)/2$ and therefore

$$D = \frac{2|E(G)|}{|V(G)|(|V(G)| - 1)}. \tag{2.12}$$

Unlike γ-density, D-density in Equation 2.11 does not use information about graph node degree. It is simply a ratio between the number of observed nodes in the graph divided by the maximum number of possible nodes. This does not take into account the distribution of edges for each node. To illustrate this, let us consider a collection of m graphs

$$\mathcal{G} = \{G_1, \ldots, G_i, \ldots, G_m\} \tag{2.13}$$

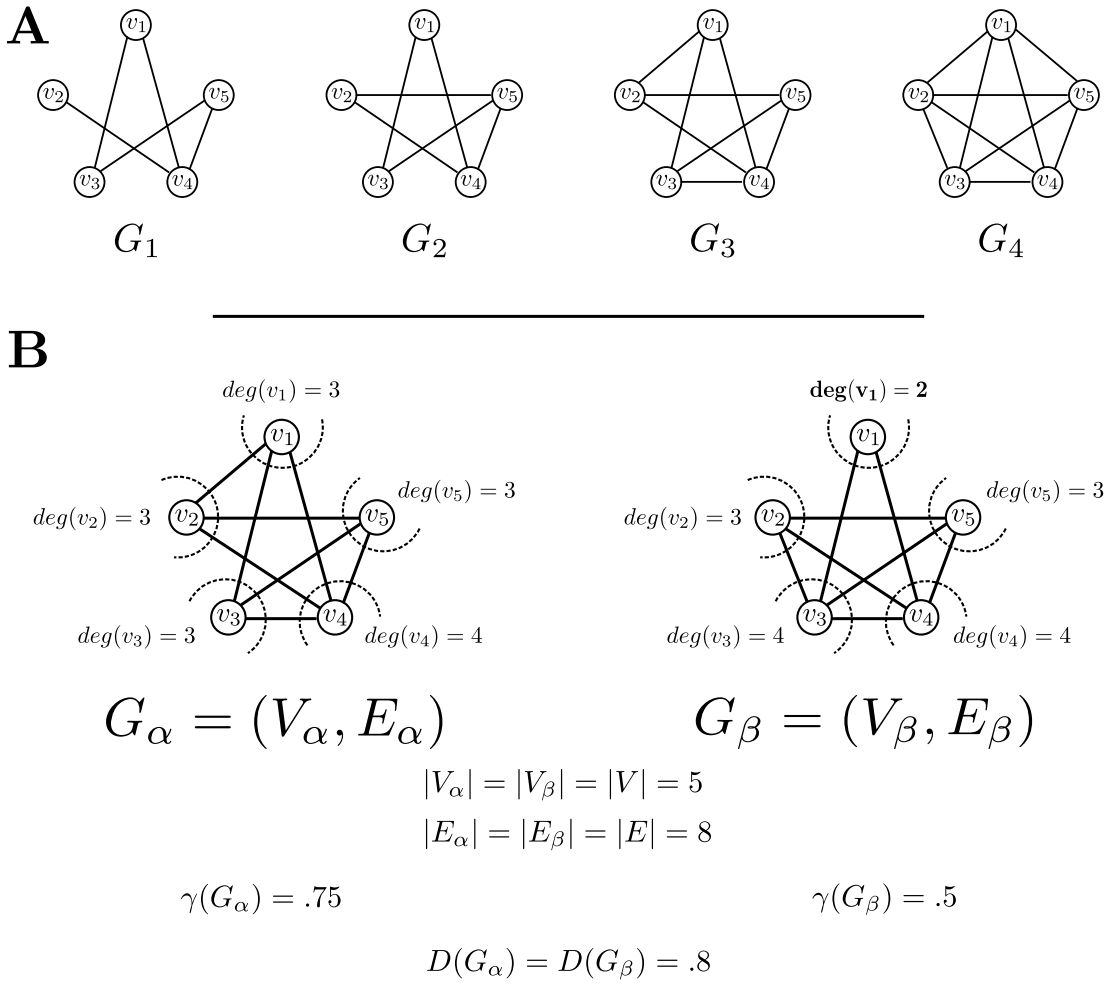

Figure 2.14: **A**: Four graphs with the same number of nodes $|V(G_1)| = |V(G_2)| = |V(G_3)| = |V(G_4)| = 5$ and increasing density from left to right. Since $\delta(G_1) = 1$, from Equation 2.9 it follows that Graph G_1 is a 0.25-quasi-complete graph. $\delta(G_2) = 2$, and G_2 is a 0.5-quasi-complete graph. $\delta(G_3) = 3$ and therefore $\gamma_{G_3} = 0.75$. Finally, G_4 is a 1-quasi-complete graph, i.e. G_4 is a complete graph. **B**: difference between γ-density and D-density. Graphs G_α and G_β have the same number of nodes and edges. However, they differ in the minimum node degree. In graph G_α all nodes have degree 3, apart from v_4 having degree 4. In graph G_β, v_1 stands out for having degree 2. The higher homogeneity in the density of G_α is picked up by the γ-density measure, which penalises G_β for its less uniform connectivity. However, the D-density, i.e. the ratio $|E|/\max|E|$, is equal for both graphs.

where

$$|E(G_1)| = \ldots = |E(G_i)| = \ldots = |E(G_m)| = E \tag{2.14}$$

and

$$|V(G_1)| = \ldots = |V(G_i)| = \ldots = |V(G_m)| = V. \tag{2.15}$$

The D-density for these graphs is identical

$$\begin{aligned} D(G_1) &= \frac{2|E(G_1)|}{|V(G_1)|(|V(G_1)|-1)} \\ &= \ldots \\ &= \frac{2|E(G_m)|}{|V(G_m)|(|V(G_m)|-1)} \\ &= D(G_m) \\ &= D_{\mathcal{G}}. \end{aligned} \tag{2.16}$$

However, the same cannot be said, in general, for their γ-density. Since the γ-density depends on the minimum graph node degree $\delta(G_i)$, from Equation 2.10 it follows that graphs with different minimum node degree will have different γ-densities, e.g.

$$\delta(G_j) < \delta(G_k) \Longrightarrow \gamma_{\max}(G_j) < \gamma_{\max}(G_k). \tag{2.17}$$

It follows from 2.16 and 2.17 that γ-density is able to discriminate finer differences between graphs. For instance, if a graph has several nodes with high degree and one "spurious" node with low degree, its γ-density will be lower than its corresponding D-density (Figure 2.14-B). The latter is blind to the topological arrangement of the edges within the network.

2.2.7 Prioritisation of the Putative Interactions: PCS

The Protein interaction Conservation Score (PCS) quantifies the potential for evolutionary conservation for the projected interaction by analysing the density of the sub-network in the neighbourhood of each experimental interaction used for the walk. It has been shown that the connectivity of well conserved proteins in interaction networks is negatively correlated with their rate of evolution [Fraser et al., 2002; Wuchty et al., 2003]. According to this theory, more connected proteins evolve at lower rate because they are subject to higher pressure to co-evolve with their interactors.

I use this insight to project the connectivity information from the reference genome (and the experimental interaction) to the query genome (and the putative interaction): a binary protein interaction part of a very well-connected sub-network in the reference genome is more likely to have retained its functional characterisation after the projection to the organism of interest. Figure 2.15-A,B provides a schematic illustration of the principle behind the PCS. One way of quantifying the connectivity of the experimental interaction sub-graph is to use the density

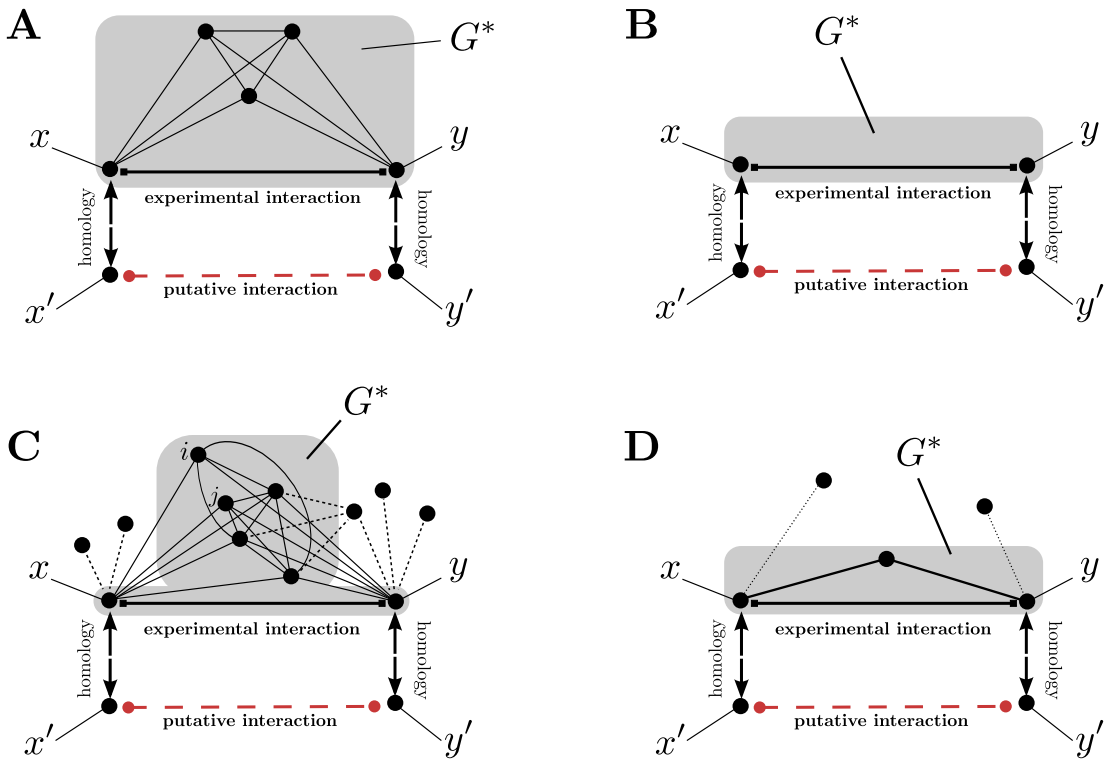

Figure 2.15: Analysing the neighbourhood connectivity for the experimental interaction (x, y) in the reference genome. **A-D**: hypothetical test cases - (x, y) is the experimental protein interaction, (x', y') is the putative interaction in the genome of interest and G^* is the graph composed by (x, y) and all their mutual interactors. Information about the density of G^* is processed to create a PCS value assigned to the putative interaction (x', y'). **A**: x and y are part of a complete graph G^* composed of 5 nodes and 10 edges. **B**: either no mutual interactors of (x, y) exist, or they do not exist in IntAct: in this case the PCS cannot give any information. **C**: proteins x and y share several documented experimental interactors. G^* is quasi-complete (the interaction (i, j) is missing). **D**: in this case, G^* is complete however it only has 3 nodes. In our biological application, case **C** is more useful than **D**, in spite of G^* in **D** being complete and G^* in **C** only quasi-complete. The PCS corrects the density indicator to reward larger quasi complete graphs and penalise small complete graphs.

indices discussed in Section 2.2.6. In my initial implementation, I have used the D-density (Equation 2.11). For each putative interaction, I compute $D(G^*)$, where G^* is the graph so defined:

- Its nodes are x and y, the two experimental interactors in the reference genome, and all their mutual experimental interactors;
- Its edges are all the edges between the nodes described above, minus any eventual self-interactions.

I realised however that using D naively would introduce a bias - very small complete networks would result in values $D(G^*) = 1$, while more biologically interesting, quasi-complete larger networks would receive values $D(G^*) < 1$ (Figure 2.15-C,D). I therefore adopted the method suggested by Huang et al. [2007], and defined the Protein interaction Conservation Score as

$$\mathbf{PCS} = D(G^*) \cdot |E(G^*)|, \tag{2.18}$$

Since the D-connectedness measure is biased towards maximally connected small sub-networks, it is relaxed by weighting it with the number of edges E.

There are many ways that an interolog could be prioritised. I aim for `Bio::Homology::InterologWalk` to be compatible with a diverse range of data and useful for many different kinds of users. Any prioritisation metric will be context-dependent and for this reason I offer a number of options to configure the process to suit the users' requirements and the coverage and quality of the data available to them. As such the generalised and customisable prioritisation scheme I provide here should provide the necessary flexibility to allow application across a broad range of biological domains.

2.3 Validation

2.3.1 Using the Algorithm to Retrieve Known Interactions

I tested the correct functioning of the `Bio::Homology::InterologWalk` package by assessing its ability to recover known interactions using the orthologue walking principle. The basic idea is to build a pair of complete experimental interactomes, using two generic genomes. A dataset of known interologs can then be constructed by selecting the subset of all binary protein interactions for the first genome made of interacting partners whose orthologues in the second genome are also linked by a binary protein interaction. In other words, we build a dataset containing all the experimental interologs between the first and the second genome. We can then pretend to *ignore* knowledge about the experimental interactions in one of the two species,

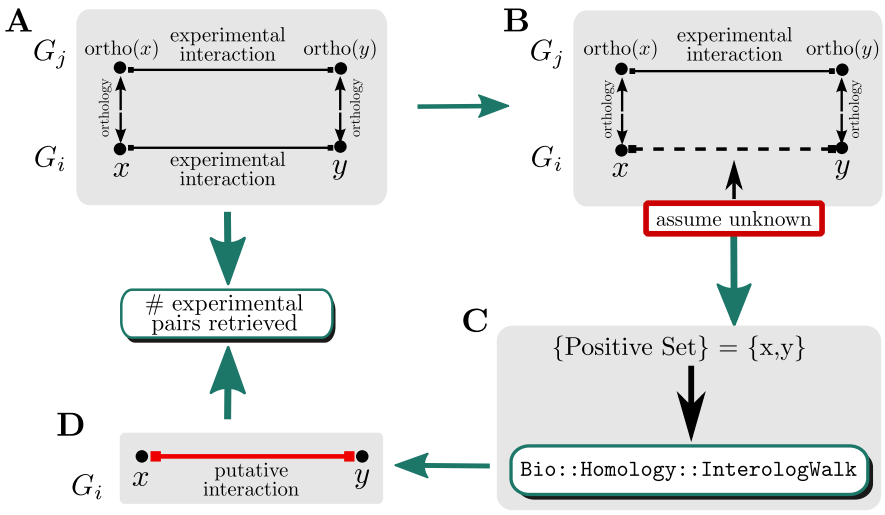

Figure 2.16: Schematic showing the rationale for the creation of the known positive sets $\mathcal{KP}_{G_iG_j}$ for validation. (**A**) Complete protein interaction datasets for two genomes G_i and G_j are retrieved. Only interactions conserved across the two species through orthology are retained. Interaction pairs in G_i satisfying this property constitute the known positive set $\mathcal{KP}_{G_iG_j}$. (**B**) Interaction information between the IDs in $\mathcal{KP}_{G_iG_j}$ is assumed unknown. (**C**) The gene IDs in $\mathcal{KP}_{G_iG_j}$ are the input for `Bio::Homology::InterologWalk`. (**D**) The putative interaction set obtained is compared with the experimental interaction known positive set.

and use `Bio::Homology::InterologWalk` to carry out an interolog walk. The expected result is the complete retrieval of all experimental interologs[27] (Figure 2.16).

To identify known interologs for the validation analyses, I obtained the complete genomes for five well-annotated species (human, mouse, yeast, fly and worm) from Ensembl V. 61. Then, I extracted all the known experimental protein-protein associations for each of the five genomes G_i ($i = 1,\ldots,5$) from EBI IntAct.

Let us define I_{G_i} as the set of the N_{G_i} experimental protein-protein interaction pairs in G_i:

$$I_{G_i} = \left\{(x,y)^{(n)}\right\}_{n=1}^{N_{G_i}}. \tag{2.19}$$

Next, I chose five pairwise genome combinations G_iG_j: mouse-human, human-yeast, human-fly, fly-yeast and yeast-worm. For each G_iG_j, let us then define the *Known Positive Evidence* dataset $\mathcal{KP}$ as the following subset of I_{G_i}:

$$I_{G_i} \supset \mathcal{KP}_{G_iG_j} = \left\{(x,y) \in I_{G_i} : (\text{ortho}(x),\text{ortho}(y)) \in I_{G_j}\right\} \tag{2.20}$$

where ortho$(\cdot)$ is the orthology operator. $\mathcal{KP}_{G_iG_j}$ is the set of all binary interactions in G_i that match through orthology[28] in G_j (Figure 2.16-A).

The gene IDs in the five interaction sets in $\mathbf{KP} = \left[\left\{\mathcal{KP}_{G_iG_j}\right\}^{(k)}\right]_{k=1}^{5}$ were used as input for the module. To validate the ability of `Bio::Homology::InterologWalk` to recover known

[27] Plus, of course, a number of putative interologs with no countercheck in the available experimental interactomes.

[28] Or, equivalently, the set of all known interologs.

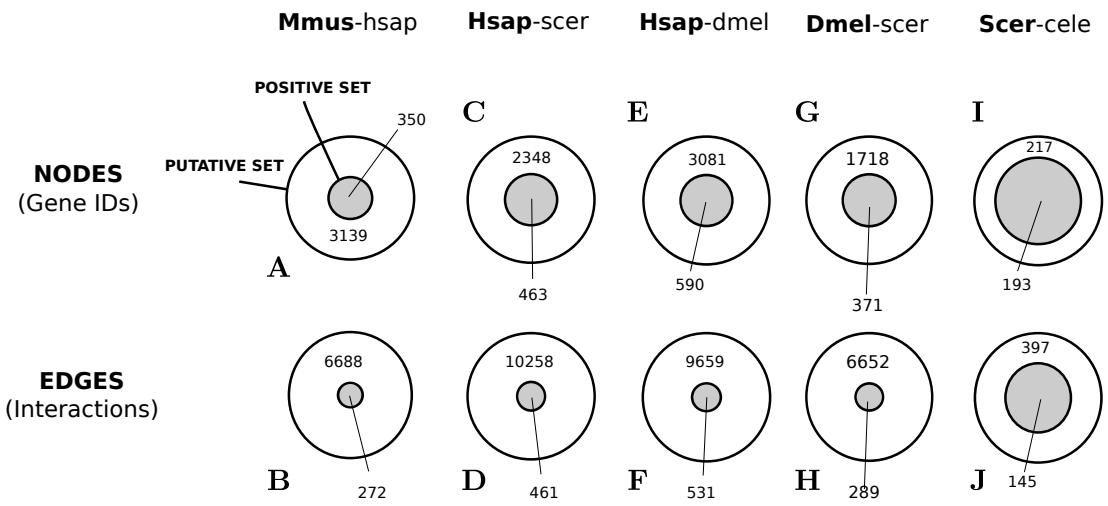

Figure 2.17: Venn diagrams showing, for five representative species-pair combinations, the overlap between known positive sets $\mathcal{KP}$ (grey circle) and `Bio::Homology::InterologWalk` predicted set (white circle). Mmus: mouse, Hsap: human, Scer: yeast, Dmel: fly, Cele: worm (bold typeface indicates query genome). In all observed cases, the algorithm completely rescues the known positive samples and, in addition, proposes new potential interactions and interaction candidates. The new predictions account for a minimum of 53% to a maximum of 90% of the total IDs produced and a minimum of 73% to a maximum of 96% of the total interactions. The results suggest that even in the case of well studied organisms — provided that the hypothesis of functional conservation between orthologues is correct — most physical protein associations are still unknown.

interologs (Figure 2.16-B,C,D), I compared the degree of overlap between predicted nodes (gene IDs) and edges (interactions) and known positive nodes and edges, for each of the five sets (Figure 2.17). For each Venn diagram, the grey set represents the known positive set $\mathcal{KP}$, while the white set corresponds to the algorithm's predictions. `Bio::Homology::InterologWalk` successfully retrieves 100% of the positive interactions in all cases considered, as expected. This confirms that the algorithm is correctly designed and the code does not have bugs causing it to miss data entries. Most importantly, this validation experiment confirms that the forward orthology retrieval module and the backward orthology retrieval module (Figure 2.7, Page 36) work in a perfectly symmetrical way — as loss of known interologs would have been caused by interactions mapped by the forward step, but not by the backwards step (or *vice-versa*). The results of this experiment lead me to conclude that, when using the code to retrieve putative interologs, no errors are committed due to problems with the data retrieval stages.

Interestingly, the known positive sets appear smaller than might be expected between closely related organisms like human and mouse. This might be due to a combination of factors such as (1) the parameters for orthology classification used by Ensembl are very stringent, (2) there are biases in experimental research across organisms (the bulk of experimental predictions in each of the two species might come from experiments in different cellular domain

and sub-systems) (3) experimental interaction data contains false positive interactions, which will not normally map through orthology. False positive interactions will only map through orthology if one of the following is true: a) the annotation/experimental error that generated the false positive interaction is found in both the query and the reference experimental interactome and the two false interactions are interologs; or b) a true positive interaction in one of the two interactomes happens to map through orthology to a false positive interaction occurred through an annotation/experimental method error in the other interactome.

It is also interesting to note that in the case of the yeast-worm pair (Figure 2.17-I,J) the number of novel IDs and novel interactions retrieved is one order of magnitude smaller than in the other four cases. This is consistent with the relatively limited amount of experimental interaction data available for *C. elegans* [Li et al., 2004b].

2.3.2 Assessing the IPX using ROC Analysis

There is no standardised, optimality-driven procedure to combine protein interaction information metadata into compound indicators for candidate prioritisation. Depending on the available metadata, customised heuristics have been proposed and evaluated on the basis of their actual utility in selecting interesting candidate interactions [Yu et al., 2004; Huang et al., 2007; Aranda et al., 2011]. As a consequence of the lack of an optimality-driven meta-analysis criterion, there is also no standardised procedure to validate such compound metadata indicators. However, it is standard practice in computer science to assess the behaviour of any indicator, predictor, or metric through quantitative analysis, to chart the stability of its response as a function of some variable.

In order to visualise and quantify the behaviour of the IPX as its discrimination threshold is varied, this section will employ techniques borrowed from the field of classification theory and detection theory. Specifically, I will be assessing the IPX through *sensitivity*-(1 - *specificity*) curves, also known as Receiver Operating Characteristics (ROC). ROC curves are graphical plots that illustrate the variation of the True Positive Rate (TPR) versus the False Positive Rate (FPR) for the predictions of a classifier, as a function of a predefined varying threshold. Let us consider, as an example of binary classifier, the spam filter commonly used by email services. The performance of a spam filter can be evaluated by comparing its TPR to its FPR (Figure 2.18). Its TPR would be the ratio of real spam emails correctly labelled as spam by the classifier, over the total number of spam emails tested, while its FPR is the ratio of emails incorrectly labelled as “spam” (when they are in fact regular emails) over the total number of non-spam emails available.

An ideal classifier would maximise its TPR (benefits) while minimizing its FPR (costs): however, in real world applications, a trade-off between the two is often sought, since an attempt to increase the TPR will result in an increase in the FPR (Figure 2.18-B,C). A ROC curve

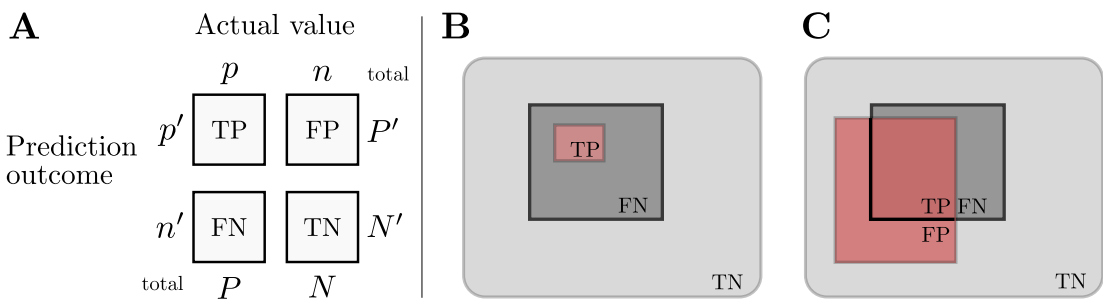

Figure 2.18: TPR/FPR-based classifier analysis. **A**: 2×2 confusion matrix. Suppose we have a dataset composed of P positives and N negatives — e.g. P spam emails, N non-spam emails. A classifier is designed to guess these two categories. Its predictions are P' and N'. Its accuracy can be evaluated on the basis of the similarity between P and P', and N and N'. For example two popular measures are the True Positive Rate $TPR = TP/TP+FN$ and the False Positive Rate $FPR = FP/FP+TN$. Ideally, $TPR = 1$ and $FPR = 0$. **B-C**: sample overlaps between real categories and predictions and TPR-FPR trade-off. Light grey square: all data samples. Dark grey square: spam emails. Red square: what the classifier predicts to be spam. In **B** the classifier's prediction only includes true positives, but not all of them. Therefore, while $\text{FPR}_B = 0$, $\text{TPR}_B << 1$. In **C**, the number of correct guesses is higher, however a number of negative samples have also been wrongly labelled. Therefore, while $\text{TPR}_C > \text{TPR}_B$, the improvement comes with a $\text{FPR}_C > 0$ trade-off. We use this set-up with the predictions in `Bio::Homology::InterologWalk`: however, in this case the exact P (true interologs) and N (false interologs) are unknown: the assessment can only provide a lower-bound estimate of accuracy.

shows these dynamics as a function of some varying parameter. A commonly used summary performance evaluator for a ROC curve is its Area Under the Curve (AUC): an AUC > 0.5 indicates that the classifier is predicting better than a random predictor, with an AUC $= 1$ representing perfect classification.

Going back to `Bio::Homology::InterologWalk` and the IPX, the idea is to use ROC curves to visualise how the relationship between TPR and FPR changes as a threshold over the IPX is increased or decreased. This can be done to quantify the benefit/cost trade-off when using the IPX. However, there are differences between a classification system (like the spam detector in the example above) and `Bio::Homology::InterologWalk`: first of all, the full positive P (interologs between species A and species B which really exist in nature) and negative N (interologs between species A and species B which do not exist) spaces are unknown, because the available interactomes are not complete. This entails that we do not have exact TP, FP, FN and TN datasets to gauge the tool's performance. As a consequence of the fact that the known TP set is only a subset of the real, unknown full TP set, and that the current known FP sets in fact contains undiscovered TP members, I expected this benefit/cost analysis to provide only an approximate estimate on the performance of the method[29]. Second point, `Bio::Homology::InterologWalk` is not strictly a classifier, but rather a discovery tool. Hence, the IPX is not designed to help classifying true positives while minimising false positives: rather, it is

[29] Because for every threshold value the real TPR would likely be bigger and the real FPR smaller.

designed to guide the *discovery* of new potential interactors, which in TPR/FPR terms means keeping the TPR as high as possible, while finding as many new putative interactors as possible. In the spam classifier example, we would design the system and its parameters to maximise the TPR and minimise the FPR as the threshold becomes more stringent. In `Bio::Homology::InterologWalk` (where the IPX is the thresholded parameter) as the threshold becomes more stringent, we would like to get a small number of very high scoring true positives, and a small number of very high scoring false positives — because new predictions are hiding in the high scoring false positive set.

Using the known true positive datasets in **KP** and using their IPX as the threshold parameter, I calculated ROC curves for each of the five species pairs (Figure 2.19). For each characteristic, the point at coordinate (1,1) corresponds to $\text{IPX}_{\text{thr}} = \min(\text{IPX})$, TPR $= 100\%$ and FPR $= 100\%$. The point at coordinate (0,0) corresponds to $\text{IPX}_{\text{thr}} = \max(\text{IPX})$, TPR $= 0\%$ and FPR $= 0\%$. Initially, $\text{IPX}_{\text{thr}} = \min(\text{IPX})$. Then, the score histogram is divided into 1000 segments and IPX_{thr} is incremented until $\text{IPX}_{\text{thr}} = \max(\text{IPX})$ is reached.

For all five datasets, the area under the curve AUC > 0.5, demonstrating that there is a positive relationship between known positives and the IPX. For all datasets, the decrease of TPR is slower than the decrease of FPR as $\text{IPX}_{\text{thr}} \rightarrow \max(\text{IPX})$. This means that, as the score threshold becomes more stringent, for all datasets the number of known positive samples lost stays smaller than the number of new predictions lost.

The correlation between TPR and the FPR is dataset-dependent: in the yeast-worm pair example, 98% of known positives are retrieved when the novel prediction retrieval rate (FPR) is down to about 76%. Conversely, in the human-yeast case, the TPR is down to about 92% for 98% FPR. The reason for these differences in accuracy lies in the completeness of the TP sets. As anticipated, for all five datasets, the real TP sets are unknown and the disparity between genome size and the number of known TP interologs means that they are likely to represent a small proportion of the real TP set. As a consequence, the AUC values are likely to underestimate the retrieval capability of the algorithm. This also suggests that the IPX may not be optimised. I anticipate that as improved coverage of protein interaction data becomes available it will be possible to optimise the IPX and improve these AUC values.

The reason why a number of known positives have a low index lies in the nature of the IPX and in the fact that ROC curves and TPR/FPR graphs are used to evaluate classifiers, not the IPX. The latter is designed to reward functionally conserved interologs obtained from binary experimental interactions: as stated in Section 2.2.5, the IPX penalises putative protein interactions obtained from orthology projections where co-orthologues exist or from binary interactions that have been artificially extracted from protein complexes. Some known TP interologs will fall into one or both of these two categories. In other words if, for a species pair, a large number of known TP interologs is constructed by matching spoke-expanded binary

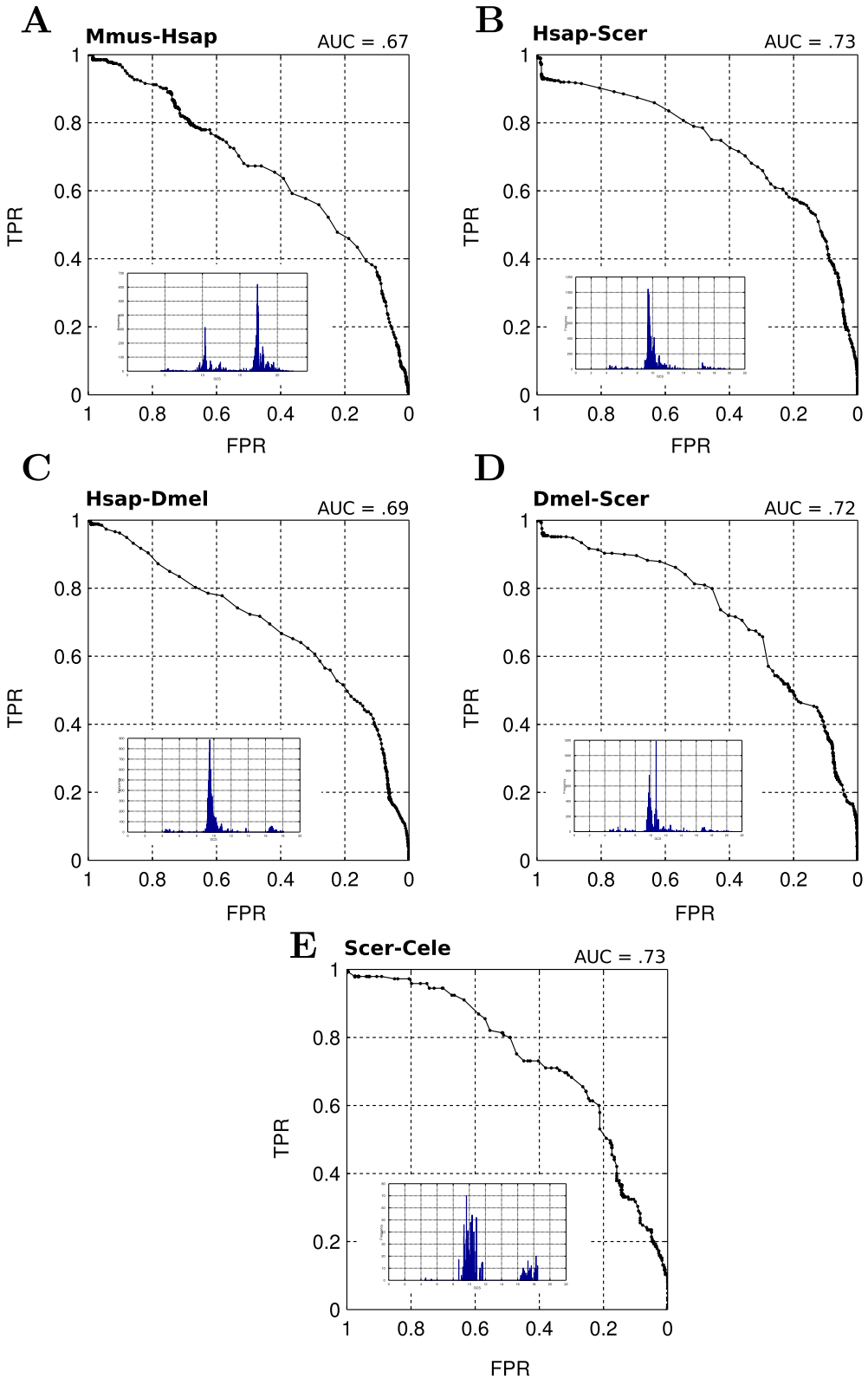

Figure 2.19: Mirrored ROC curves for the five genome pairs in the known positive sets in **KP**. *Inset*: IPX score distributions (reproduced for clarity in Figure A.1, Page 185).

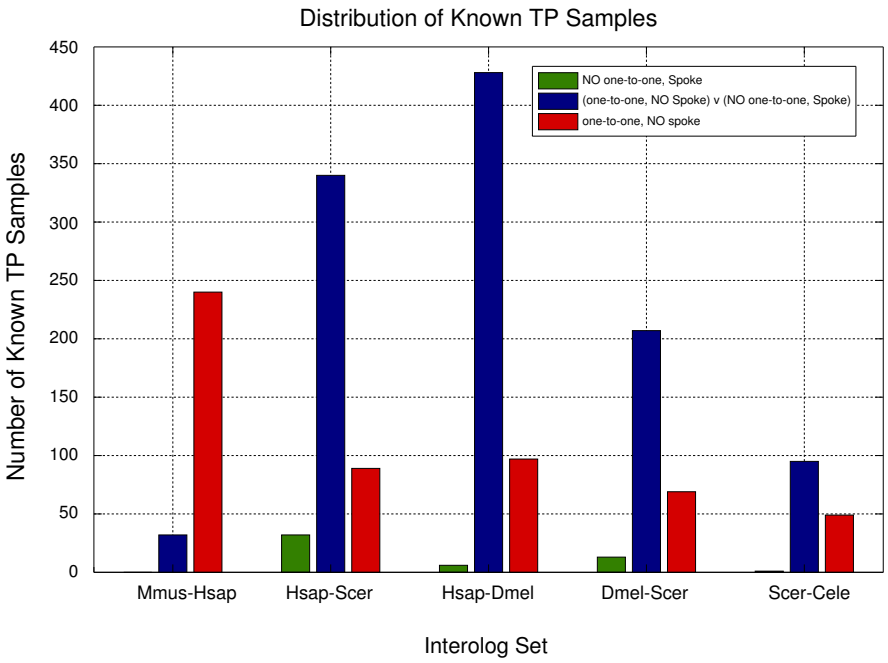

Figure 2.20: Distribution of known true positive samples in the IPX histograms. The chart shows, for each of the datasets in **KP**, the number of known true positive samples in the low (dark), average (medium) and high (bright) tiers of the IPX distribution (Figure 2.11, Page 49).

interactions and orthologues possessing in-paralogues, these samples will get a low IPX and will be eliminated early during the threshold sweep used to build the ROC curves. Distributions for the known TP samples within the IPX histograms are shown in Figure 2.20 (for reference, complete IPX distributions for all predictions are shown in Figure 2.19-inset and magnified in Figure A.1, Page 185). The chart in Figure 2.20 shows, for each dataset, how many true positive samples are in the low (dark), average (medium) and high (bright) tiers of the IPX distribution (Numerical quantities are in Table A.3, Page 186). As such, it quantifies what I discussed in the previous paragraph, and shows, for each sample species pair, how the known TP samples are distributed according to orthology type and expansion/no-expansion information. The largest amount of known TP samples falls in the second tier for all but the mouse-human dataset, which is the only one to have most of its positives in the high tier. The graph explains why some of the ROC curves in Figure 2.19 keep a higher TPR rate for longer as the threshold becomes more stringent and, conversely, why for some of them the TPR/FPR ratio seems to fall sharply early on: e.g. the *Hsap-Scer* and *Dmel-Scer* known TP sets have an initial sharp fall in their ROC, likely to be explained by a high amount of known TPs in the low tier, as shown in Figure 2.20.

As a side observation, I also looked at the relationship between the IPX threshold sweep and the loss of known TP data for the five sample datasets in more detail (Figure 2.21). The graph shows the decrease of the TPR (left) and of the FPR (right) as a function of the IPX threshold. When $\text{IPX}_{\text{thr}} = 0$, i.e. when no predictions are discarded, the TPR and FPR are at

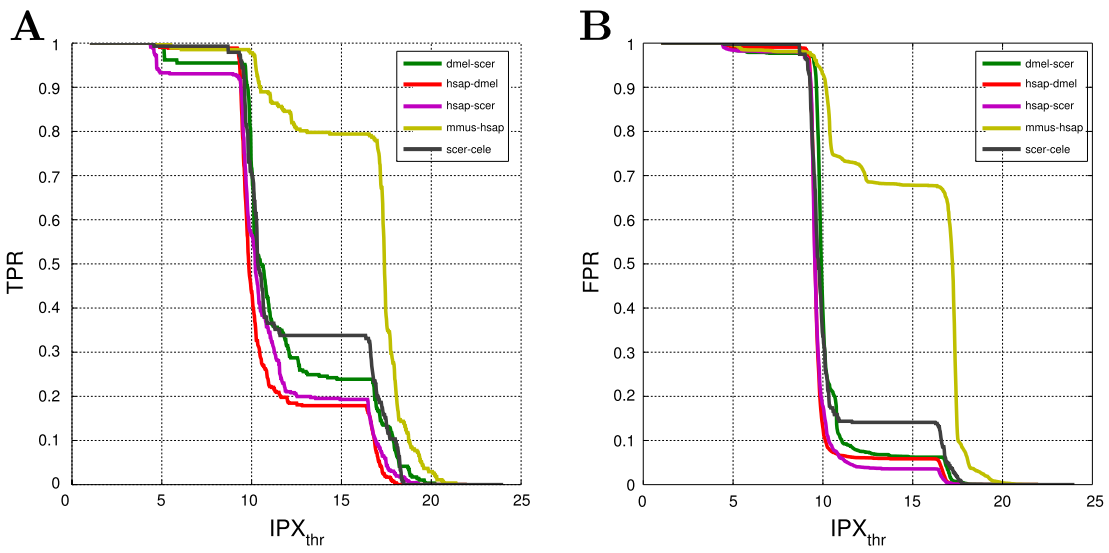

Figure 2.21: Relationship between TPR, FPR and IPX Threshold for the five putative protein interaction datasets obtained for the known TP datasets through `Bio::Homology::InterologWalk`.

their maximum value, meaning all known TP samples are retrieved by `Bio::Homology::InterologWalk`, but also all the FP samples. As we start sliding IPX_{thr}, the different datasets lose TP samples and FP samples at different rates, for the reasons detailed in the paragraph above. It is interesting to notice that for the mouse-human dataset (*mmus-hsap* in the legend for Figure 2.21) 80% of the known TP samples are retained at a $IPX_{thr} = 15$ threshold level. At the same threshold value, all of the other datasets retain a significantly lower number of known TP samples (Fisher/Chi-Square test, Table A.1, Page 186. Contingency tables used for the analysis are in Table A.2). This result may reflect the closer phylogenetic distance between mouse and human: as also shown in Figure 2.20, most of the known mouse-human interologs are in the high-tier of the IPX distribution: this indicates that the human-mouse homology maps in this area have been obtained through 1:1 orthologues, and no co-orthologues are known[30]. This is expected as the primate-rodent split has happened relatively recently (estimated between 65 and 85 million years ago, [Lee, 1999; Foote et al., 1999]) and less gene duplication will have occurred since divergence from their common ancestor in comparison to the other species pairs. This result suggests that the selection of putative interologs obtained through `Bio::Homology::InterologWalk` and the IPX is not in contradiction with sound phylogenetic evidence about the genome divergence of the species considered.

Overall, I have proven that `Bio::Homology::InterologWalk` is working correctly on the code, algorithm and implementation level, by showing that all known interologs are retrieved when labels are removed (Figure 2.17). Further to that, I have proven that according to a sensitivity/specificity curve analysis the selection of interologs returned by a sliding IPX threshold is

[30] According to Ensembl Compara V 61.

always favourable to the TPR. This suggests that the IPX, while by no means the best possible prioritisation index for the purpose (which cannot be quantified due to the lack of a confirmed positive/negative interolog set) is working correctly and is providing satisfactory classification performance. Such performance is likely to improve as more experimental interactions are mapped and added to the online database, thus refining the positive/negative boundary sets.

2.4 Results

In the previous section, I showed that `Bio::Homology::InterologWalk` correctly retrieves all known interologs for five popular interactome pairs, which proves that the methodology and the implementation work as expected. I also showed that the true positive rate is constantly higher than the false positive rate, for all putative interaction data in the five interactome pairs, for all IPX threshold values. This indicates that, regardless of what IPX threshold is employed to select a subset of the putative interactions, the number of true predictions is always higher than the number of false positives. Figures obtained for the Area under the ROC Curve (AUC) give values higher that 50% in all cases, and it is important to stress these are conservative estimates: some of the putative interactions falling in the false positive group are in fact unknown true positives — experimental interactions that exist in the interactome, but have not been verified experimentally and consequently have not yet been annotated in IntAct.

In the next step, I shall employ `Bio::Homology::InterologWalk` in sample biological scenarios to assess its utility in providing new interaction predictions. For this evaluation, I chose two experimental datasets: the full genomes of *Drosophila melanogaster* and of the protozoan parasite *Plasmodium falciparum 3D7*.

The bulk of available experimental information about the fruit fly's interactome comes from a few large scale high throughput Y2H studies [Giot et al., 2003; Stanyon et al., 2004; Formstecher et al., 2005]. DroID, the Drosophila Interactions Database[31], probably the most comprehensive aggregating resource for fly protein interactions, reports that the total number of fly interactions obtained through Y2H screens by Giot et al., Stanyon et al. and Formstecher et al. is in the region of 24000 [Murali et al., 2011]. Formstecher et al. [2005] start from 102 bait proteins, most of which are orthologous to human-cancer related and signalling proteins, and obtain about 2300 protein interactions, 710 of which are claimed to be high confidence. Formstecher et al. find very little overlap with the results by Giot et al. [2003], who describe 7048 interacting proteins and 20,405 interactions, and a high confidence subset of 4679 proteins and 4780 interactions. Stanyon et al. is a follow-up to the [Giot et al., 2003] study which aimed to cross-validate its results: it uses the same set of Drosophila open reading frames, but while the first uses a Gal4-based Y2H system, in Stanyon et al. [2004] a LexA-based

[31] www.droidb.org/

system is chosen. The study finds 1814 interactions among 488 proteins, with surprisingly little overlap with the data in Giot et al. [2003] (only 28 interactions in common between the two screens). The high number of new interactions discovered by each of these Y2H studies suggests that current fly experimental interaction mapping is far from complete, and many more new interactions remain to be found. Due to this reason, and also to provide interesting computational hypotheses to our lab (which, as stated in the Introduction, uses the fruit fly as a model system to understand organismal development) I decided to employ `Bio::Homology::InterologWalk` to attempt to augment the fly interactome with putative predictions inferred from orthology transfer.

Surprisingly little is known about protein interactions in malaria parasites. The full genome of the protozoan *Plasmodium falciparum* clone 3D7 has been published two years after the *Drosophila* one [Gardner et al., 2002]. The 23 megabase genome encodes about 5300 genes[32]. The only experimental study on the matter is the one by LaCount et al. [2005] where a high throughput Y2H screen was used to identify 2,846 unique pairwise interactions involving about 25% of the known proteome. One reason why the *Plasmodium*'s interactome has not been extensively studied experimentally has to do with the difficulties in expressing its proteins in heterologous systems (such as yeast): its genome has an overall (A + T) composition of about 80.6% [Gardner et al., 2002], which preclude the types of experimental validation that are available in model organisms. Other efforts in elucidating the parasite's interactome were purely computational [Date and Stoeckert, 2006; Wuchty et al., 2009]. In particular, Wuchty et al. [2009] derive a map of interactions utilising mainly interologs from four reference interactomes (fly, worm, yeast and *E.coli*). However, the interolog predictions are based on simple BLASTP sequence searches determined by the InParanoid script [Remm et al., 2001], which, unlike Ensembl and its TreeBeST-based homology prediction method, does not use any form of phylogenetic information support and gene-tree/species tree reconciliation [Page, 1994; Vilella et al., 2009] to inform its homology predictions. This can result in gross errors when dealing with orthology versus paralogy labelling of large gene families. Therefore the findings in Wuchty et al. are debatable in light of current state of the art orthology prediction methods and should ideally be reassessed. Additionally, we decided to utilise the *Plasmodium falciparum* genome as one of two sample datasets because we intended to explore the possibility of using `Bio::Homology::InterologWalk` to enrich the small experimental interactome of a non-model animal pathogen. As anticipated earlier in the chapter, one of the predicted usage scenarios for the algorithm involves the exploration of unknown interactomes for recently sequenced genomes of non-model organisms.

[32] Ensemblgenomes, as of release 14 (May 2012) lists about 5600 genes in total.

2.4.1 Overview and General Setup

I obtained the full list of *Drosophila* genes (termed `DS_DMEL` from now on) from Ensembl, Release 61, through its API. Ensembl fly gene information is a compendium of data from different sources, including BDGP[33], FlyBase[34] and the Drosophila Heterochromatin Genome Project[35]. The fly genomic sequence in Ensembl 61 was based on BDGP assembly release 5 (April 2006), and the annotations were imported from FlyBase release 5.25 (FB2010_02, dated 19 February 2010). The Ensembl database I used for the orthology projection was Ensembl Compara, containing data for a total of 53 genomes (as of release 61), including data for vertebrate genomes and for a selected set of popular non-vertebrate model organisms, including the fruit fly, *S. cerevisiae* and *C. elegans*.

As for the *P. falciparum* 3D7 parasite, I obtained the initial complete gene list (which I will refer to as `DS_PFAL` henceforth) from Ensemblgenomes, Release 7, which hosts annotation obtained from the two main *Plasmodium* resources, GeneDB[36] and PlasmoDB[37]. Since comparative genomics data for *P. falciparum* is not available from the Ensembl Vertebrate resource, the data source I chose this time was EnsemblGenomes *pan-homology*, a pan-taxonomic database that contains a representative sample of comparative data for a very large number of species (353 taxa as of release 7) encompassing metazoans, fungi, protists, plants and bacteria. By using two different homology databases (Ensembl Vertebrates and Ensemblgenomes Pan-homology) I also intended to demonstrate the seamless integration of `Bio::Homology::InterologWalk` with both the two homology resources it can interface with.

In both cases, I did not restrict the reference genomes to any specific species and used all the available taxa. I expected a proportion of the species in the sets to provide "dead-end" orthologues — orthologues in species for which no significant amount of experimental interaction discovery has been carried out. Nevertheless, given the relatively small size of the two initial datasets, from a computational point of view there was no significant advantage in querying only selected genomes in Ensembl versus querying all the available ones. Additionally, keeping data from all available reference genomes would allow the observation of potentially interesting patterns in the distributions of forward orthologues from fly/*Plasmodium* to the complete set of species.

I discarded all homology relationships belonging to the *paralog* class. This included cases identified by Ensembl as *possible ortholog* — instances when the duplication/speciation nature of the phylogenetic tree could not be fully resolved by the Compara–TreeBeST algorithm, but partial evidence suggested the absence of a duplication event. As reported by Ensembl,

[33] www.fruitfly.org
[34] flybase.org
[35] www.dhgp.org
[36] www.genedb.org
[37] www.plasmodb.org

such cases might point to long distance relations that might be upgraded to *bona fide* orthologies in further versions of the gene tree pipeline. There are several reasons why I decided not to consider paralogy relationships for these two pilot experiments. These will be thoroughly discussed in the final section of this chapter. Briefly, while it is possible to draw certain conclusions about the function of unknown genes by looking at their within-species duplicates, I initially developed `Bio::Homology::InterologWalk` to evaluate the possibility of making initial functional hypotheses in poorly studied interactomes based on what is known in more popular species, and the examples presented here follow this line of action. Additionally, the bulk of experimental interactions in the fly and *Plasmodium* comes from high throughput Y2H. I intended to check whether putative interactions obtained from low throughput model organisms interactions would tend to reconfirm some of the results observed in the large fly and *Plasmodium* Y2H studies. Lastly, it has been argued that protein duplication is responsible with sub-functionalisation and neo-functionalisation [He and Zhang, 2005; Hittinger and Carroll, 2007], although the debate is pretty much open, and an opposite line of thought regarding conservation of function between orthologues and paralogues exists [Mika and Rost, 2006].

As for homology relationships belonging to the *ortholog* class, all were accepted in the experiment. I decided to keep *one-to-many* and *many-to-many* relationships — in spite of their lower potential for reliable functional conservation — as a way to evaluate the prioritisation sub-module of the algorithm.

Regarding the protein interaction collection phase, I retrieved EBI IntAct interactions retaining only those that satisfied *both* the following criteria:

1. For the field `interaction detection method`, the interaction is tagged with a PSI-MI ontology code specifying an *experimental detection method* code or a specialisation thereof.

2. For the field `interaction type`, the interaction is tagged with a PSI-MI ontology code specifying a *physical association* code — or a specialisation thereof. This excluded relationships based on co-localization and genetic interaction evidence.

I did not discard interactions resulting from spoke-computational expansion of interacting complexes, provided they satisfied the previous two requirements. This was again to test the operation of the prioritisation routines in penalising spoke putative interactions.

As a parallel step in the experiment I processed both initial datasets, `DS_DMEL` and `DS_PFAL` through the direct interaction pipeline, to obtain all the experimental interactions available in IntAct. Again, only physical associations obtained through any of the experimental detection methods were considered, and spoke-expanded pairs were preserved.

Table 2.3 shows statistics for the resulting datasets using both `DS_DMEL` and `DS_PFAL` as the initial gene sets. I will henceforth adopt the following terminology:

	DS_DMEL Pipeline		DS_PFAL Pipeline	
	Putative	**Known**	**Putative**	**Known**
Datasets				
Gene IDs	**14869**	**14869**	**6213**	**6213**
Reference Genomes Used	52	1	283	1
Orthologues (Forward)	150968	NA	63337	NA
Interactions in Reference Genomes	37931	NA	49133	NA
Total Interactions	11316	51827	14594	5142
Unique PP Pairs	**4428**	**26622**	**4897**	**2629**
Surviving IDs (% **Gene IDs**)	2188 (14.7)	7779 (52.3)	1421 (22.9)	1250 (20.1)
Networks				
Nodes	**2188**	**7779**	**1421**	**1250**
Edges	**4428**	**26622**	**4897**	**2629**

Table 2.3: `Bio::Homology::InterologWalk` usage statistics using the *Drosophila melanogaster* interactome dataset (`DS_DMEL`) and the *Plasmodium falciparum* interactome dataset (`DS_PFAL`). Results obtained using the two available `Bio::Homology::InterologWalk` pipelines — putative and experimental — are shown. In the putative pipeline columns, the data shown are relative to interactions obtained through interolog mapping. In the experimental pipeline column, DS_DMEL and DS_PFAL has been queried against EBI IntAct to mine all the experimental molecular associations known in the literature. The field *Total Interactions* indicates the total number of final entries of the form $e = (\text{gene}_a, \text{gene}_b)$ obtained. Note that, in the putative pipeline, e can be observed several times through different walks.

1. `NET_<dataset>_known` — the network consisting of all the experimental physical associations between genes in `<dataset>` found in IntAct;
2. `NET_<dataset>_putative` — the network consisting of all the putative interactions between genes in `<dataset>` according to `Bio::Homology::InterologWalk`;
3. `NET_<dataset>_union` — the network obtained computing the union of (1) and (2) where:
   - each node is a node of `NET_<dataset>_known`, `NET_<dataset>_putative`, or both;
   - each edge is either an edge of `NET_<dataset>_known` or an edge of `NET_<dataset>_putative` (Note: duplicate edges were *not* collapsed into one).

In this terminology, the generic `<dataset>` is either `DS_DMEL` or `DS_PFAL`. I therefore obtained a total of six protein networks, three for each dataset (Table 2.4). Finally, I used the network-output sub-component of the `Bio::Homology::InterologWalk` package on each one of the six network dataset to create data representations compatible with the Cytoscape visualisation tool [Shannon et al., 2003].

Network	Nodes	Edges
NET_DS_DMEL_known	7779	26622
NET_DS_DMEL_putative	2188	4428
NET_DS_DMEL_union	8270	31050
NET_DS_PFAL_known	1250	2629
NET_DS_PFAL_putative	1421	4897
NET_DS_PFAL_union	2226	7527

Table 2.4: Six protein interaction networks for DS_DMEL and DS_PFAL.

GO-ID	term	p-val	corr p-val	#genes
6355	regulation of transcription, DNA-dependent	7.1132E-5	1.7647E-3	10
30260	entry into host cell	5.0058E-4	8.1501E-3	6
9408	response to heat	3.8178E-3	3.9781E-2	10
6511	ubiquitin-dependent protein catabolic process	3.5240E-4	7.3439E-3	13
8380	RNA Splicing	1.5125E-3	1.9219E-2	13
6413	translational initiation	8.3107E-4	1.2557E-2	8
280	nuclear division	2.0615E-3	2.4978E-2	5
Total Unique				65

Table 2.5: Summary of highest specificity functional information available for NET_DS_PFAL_known. Each of the 65 genes maps to only one of the GO terms in the table.

2.4.2 Selecting Sub-networks through GO Annotation

Due to the size and complexity of the genome-scale interaction networks in Table 2.4 I decided to restrict the analysis to subsets of the interacting nodes, the choice of which was informed by functional annotation provided by the Gene Ontology project [Ashburner et al., 2000]. I used BiNGO [Maere et al., 2005] to observe the available functional evidence for the genes in NET_DS_PFAL_known. BiNGO uses the hypergeometric test and False Discovery Rate Correction to determine which Gene Ontology categories are statistically over-represented in a set of genes or sub-network. I assessed over-representation of GO categories using the Benjamini & Hochberg False Discovery rate correction for multiple hypothesis testing, threshold p-value of 5%, the whole *P. falciparum* annotation[38] as a reference set, Biological Process as the namespace and the most recent[39] ontology from GO. Figure 2.22 and Table 2.5 show a summary of the enrichment data.

Figure 2.22 shows the biological processes observed in the *P. falciparum* interactome which are statistically enriched with reference to the complete genome[40]. The Gene Ontology sub-

[38] GOC Validation Date: 12/31/2010, Submission Date: 9/2/2008

[39] 18/02/2011

[40] In other words, for the terms in the coloured nodes in Figure 2.22, the probability of appearing with equal frequency in a random set of nodes of same size is < 0.05

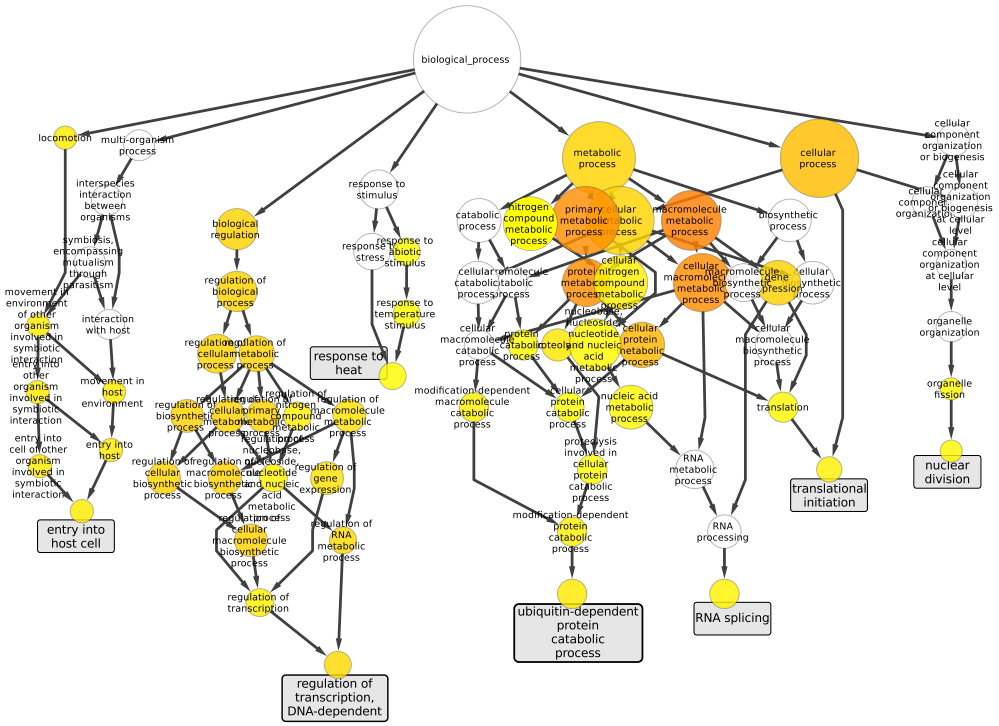

Figure 2.22: Summary of GO analysis results for `NET_DS_PFAL_known`. The diagram is a directed tree evidencing the go terms enriched in the network. Each coloured node is an enriched GO term, colour is proportional to P-value after FDR correction (darker colour corresponds to higher significance). White nodes do not pass 0.05 threshold and are included only to visualise connectivity. Concept specialisation increases from top to bottom: the most specialised enriched concepts (leaves of the tree) are highlighted by a grey description rectangle and described in Table 2.5.

ID	GO Evidence	Source
Plasmodium falciparum:		
MCM5	Inferred from Electronic Annotation	–
RPA1	Inferred from Direct Assay	[Voss et al., 2002]
PFD0950W	Inferred from Sequence or Structural Similarity	[Hall et al., 2002]
PFF1470C	Inferred from Sequence or Structural Similarity	[Hall et al., 2002]
Drosophila melanogaster:		
CG2714 (crm)	non-traceable author statement	[Harr, 2001]
CG7869 (SuUR)	inferred from expression pattern	[Makunin et al., 2002]
CG9241 (Mcm10)	inferred from mutant phenotype	[Christensen and Tye, 2003]
CG9633 (RpA-70)	inferred from direct assay	[Mitsis et al., 1993]
CG10262	inferred from sequence or structural similarity	(FlyBase, 1992-)
CG7413 (Rbf)	inferred from mutant phenotype	[Bosco et al., 2001]
CG10336	inferred from electronic annotation	[FlyBase Curators et al., 2004]
CG11301 (Mes4)	inferred by curator from GO:0003887	(FlyBase, 1992-)
CG4039 (Mcm6)	inferred from mutant phenotype	[Schwed et al., 2002]
CG4088 (lat)	non-traceable author statement	[Sokolowski, 2001]

Table 2.6: DNA Replication seed genes: provenance data

trees captured by the analysis give an idea on the current knowledge of over-represented functional roles for the genes in NET_DS_PFAL_known. Some of these genes (Table 2.5) are homologues of heat shock chaperones and play an essential role in folding of proteins participating in cell cycle regulation and signal transduction [Hall et al., 2002; Kumar et al., 2003]. They are likely to have been investigated due to the importance of the Hsp90 chaperone as a drug target — it has been shown that serum of mice and humans exposed to Plasmodium contains abundant amounts of antibody reactive to Hsp90, suggesting that the latter may have a major antigenic role in malaria [Goetz et al., 2003]. Six other genes, tagged with the term response to heat are of interest because of their potential role as vaccine candidates [Trucco et al., 2001]. Other enriched processes correspond to conserved basic cell-related dynamics.

Overall, it is clear that very little is known in terms of functional characterisation of most genes participating in the known *P. falciparum* interactome. I therefore decided to concentrate the analysis on a number of well conserved cell biology-related processes with low representation amongst the known protein interactors. I chose a well conserved process to maximise the possibility of meaningful interolog transfer, due to the large variety of diverged reference genomes being employed for the mapping. The choice fell on the DNA replication (GO:6260) process due to its very small representation in the interactome[41]. Only four genes are tagged with a DNA replication BP (Table 2.6). Of these, only one, RPA1, has been tested experimentally, while functional information about the other three is based on indirect evidence. As regards the fly data and its experimental interactome, NET_DS_DMEL_known, a

[41] Corrected $p = 7.6218E-1$ (BiNGO term enrichment analysis).

much wider variety of BP processes are enriched as visualised by BiNGO (data not shown). I decided to carry out a similar analysis for the fly interactome data: again nodes in the network were selected whenever Gene Ontology evidence for a role in DNA replication was found. The biological term `DNA replication` was again the choice due to its wide conservation across taxa. This time, a subset of 65 genes in the network shared DNA replication annotation. To keep visualisation manageable and comparable to the results obtained for *P. falciparum*, I further restricted this 65 gene set to a subset of 10 randomly selected fly DNA replication genes (Table 2.6).

Given the 2 sets of "seed" DNA replication genes for fly and *Plasmodium*, I then went back to the networks to examine the protein interaction neighbourhood of these functionally related genes. The purpose of this was to attempt to build interaction clusters from the seed genes in an attempt to discover novel DNA replication-related groups of genes. In other words, I wanted to evaluate if interactions of known DNA genes with unknown molecules could (via "guilt by association" inference) point towards interesting genes or clusters with a novel role in DNA replication. I therefore proceeded to retrieve all the nearest neighbours interactors for the two sets of seeds genes. These neighbours were extracted from `NET_DS_PFAL_known` in the case of the 4 *Plasmodium* genes and from `NET_DS_DMEL_known` for the fly seeds. For *Plasmodium*, I obtained a collection of 3 disconnected small network motifs (Figure 2.23-A). The fly seeds produced instead five disconnected sub-networks, the biggest of which features 4 DNA replication genes (Figure 2.23-B).

For both datasets, no direct interactions exist between any two GO annotated DNA replication genes. In both cases, seed genes are distributed in isolated complexes, the biggest of which features two seeds for the *Plasmodium* interactome, and four seeds for the *Drosophila* interactome.

To illustrate the utility of the interolog walk I performed the same procedure as above using this time `NET_DS_DMEL_union` and `NET_DS_PFAL_union`. This was done to observe the interaction context of the DNA replication seed genes using the union of the putative and experimental interactomes, for both the sample organisms. Again, I selected genes annotated with the `DNA replication` GO BP. In the *Plasmodium* case, this time I retrieved a set of 16 hits, a superset of the 4 found before — meaning that 12 additional DNA replication genes participate exclusively in putative protein interactions.

For the fly network example, I found that a set of 68 DNA replication genes participate in protein interactions, again meaning that 3 additional DNA replication genes are drawn in through the putative pipeline. As before, I selected the sub-networks of `NET_DS_PFAL_union` and `NET_DS_DMEL_union` composed of the seed genes and their nearest neighbours (again, for `NET_DS_DMEL_union` only the 10 randomly chosen seeds used earlier were selected). The ensuing *Drosophila* sub-network is shown in Figure 2.24, and discussed in Section 2.4.3. The

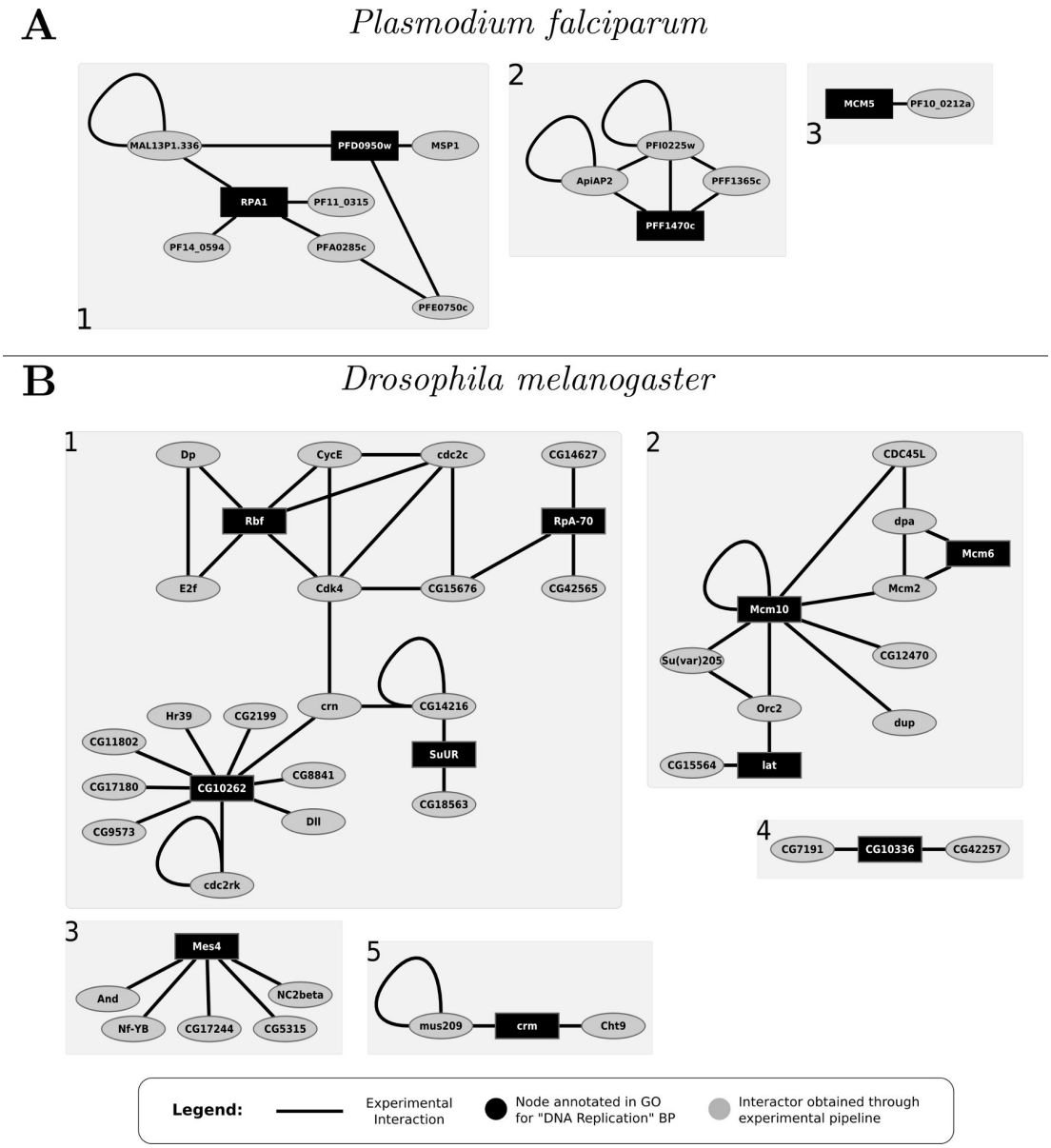

Figure 2.23: "DNA Replication" GO-annotated genes in experimental interactomes. **A**: DNA Replication seeded sub-networks in `NET_DS_PFAL_known`. **B**: DNA Replication-seeded sub-networks in `NET_DS_DMEL_known`. Data extracted as follows: 1. select all genes annotated with `DNA Replication` GO biological process (black nodes) 2. select all their nearest neighbours (grey nodes). Black connections are experimental protein interaction data from EBI IntAct. DNA replication GO-annotated genes never interact with each other in any case.

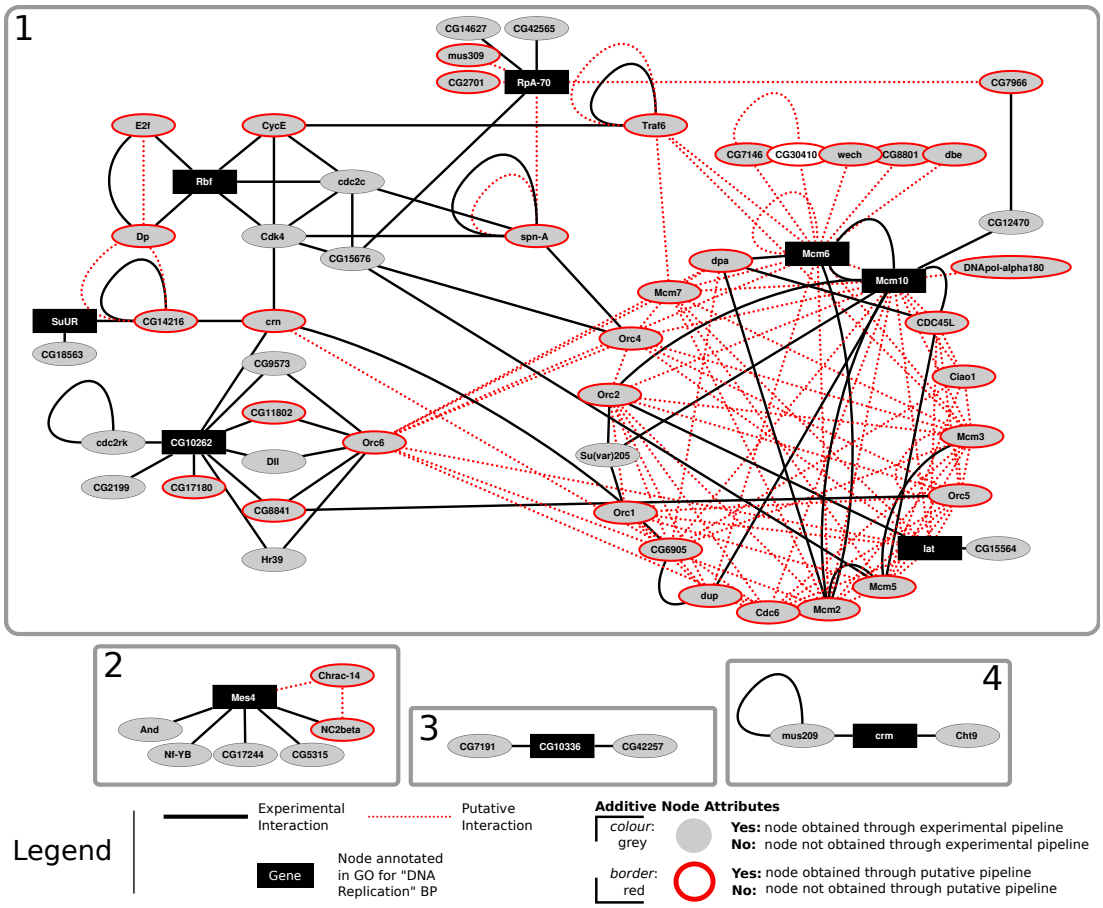

Figure 2.24: "DNA Replication" GO-annotated genes in experimental+putative *D. melanogaster* interactome. Data extracted from `NET_DS_DMEL_union` as follows: 1. select all genes annotated with `DNA Replication` GO biological process (68 genes, black nodes) 2. select 10 random genes out of those 68 (the same used for the experimental-only network) 3. select all their nearest neighbours. Solid connections (black) are experimental protein interactions obtained from EBI IntAct and originally in `NET_DS_DMEL_known`. Dotted connections (red) are putative predictions originally in `NET_DS_DMEL_putative`. Node colour code explained in legend.

Plasmodium network, shown in Figure A.2, Page 187 was too large to deal with effectively and was further processed as discussed in section 2.4.4, Page 76.

2.4.3 A *D. melanogaster* Putative DNA Replication Network

The sub-network shown in Figure 2.24 (and designated `NET_DMEL_DNArep` from now on) is a DNA replication-centred summary of direct experimental protein interaction evidence, as well as putative protein interaction evidences projected from a set of reference genomes. It is composed of 68 nodes and 165 edges and has greatly increased connectivity compared to its counterpart composed of fly experimental data (Figure 2.23-B). Indeed, the main connected component in `NET_DMEL_DNArep` now comprises 55 genes and 153 interactions, and wires together 7 of the 10 core DNA replication genes. The introduction of putative interaction data is in our opinion interesting for two reasons:

1. *Extending the existing interactome* — the genes represented by a white circle enclosed in red are purely putative interactors: they have no interacting partners known in the experimental data[42] and are part of `NET_DMEL_DNArep` purely due to a projection through orthology. In this sense, such purely putative nodes represent hypotheses that *extend* the known interactome, and represent first-approximation functional associations for genes with no previous interactome or pathway role.

2. *Reshaping the existing interactome* — those genes represented by a grey circle enclosed in red in Figure 2.24 are simultaneously experimental and putative fly interactors: they already have a role in the experimental data[43], however they are also captured by the interolog walk algorithm: this means that either they hold a very conserved role across a large number of taxa, and the comparative data is reconfirming what already known in the fly, or a new pathway/interactome role is being suggested based on what happens in one or more of the reference genomes used for the transfer. In this sense, the information brought in by these genes is reshaping the experimental interactome — because although the gene was there, the topology is changing to accommodate new connectivity related to it.

A number of observations can be made about the data visualised in the putative clusters (Figure 2.24). In sub-cluster 1 the tightly interconnected element on the right part of the image putatively links most of the proteins in the mini-chromosome maintenance complex (MCM). MCM is a component of the pre-replication complex, which is in turn a component of the licensing factor, and is a hexamer of sequence-related polypeptides (mcm2-7) that form a characteristic ring structure. Most of the complex subunits are observed in the putative cluster. MCM is

[42] I.e. they do not exist in `NET_DS_DMEL_known`
[43] I.e. they exist in `NET_DS_DMEL_known`

known to have a role in both the initiation and the elongation phases of eukaryotic DNA replication and specifically in the formation and elongation of the replication fork [Fletcher et al., 2003]. *Cell division cycle 6* (CDC6), present in the sub-cluster purely via the putative pipeline, is an essential regulator of DNA replication in eukaryotic cells. Its best-characterized function is the assembly of the pre-replication complex (pre-RC) during the G(1) phase of the cell cycle. Additionally, CDC6 has also been linked to the activation and maintenance of the mechanisms that coordinate S phase and mitosis, and recent studies have unveiled a related proto-oncogenic activity [Borlado and Méndez, 2008].

Another important DNA replication complex observed in the cluster is the *Origin Recognition Complex* (ORC) a heterohexameric complex that specifically binds to origins of DNA in an ATP-dependant manner [Matsuda et al., 2007]. One of the sub-units, ORC6, is a "hub", in protein network terms — an element at the interface between tightly connected sub-networks. ORC6 appears to connect the sub-clusters described above with another interesting group of genes (Figure 2.24, sub-cluster 1, *left*). The group includes a DNA clamp protein, CG10262 (*Proliferating cell nuclear antigen*) an auxiliary of DNA polymerase delta involved in the control of the polymerase's processibility during elongation of the leading strand.

Additionally, interactions between SuUR, *Suppressor of Under-Replication*, Rbf (*Retinoblastoma-family protein*) and the RpA-70 (*Replication Protein A 70*) suggest that the cluster has extracted a number of important control proteins and complexes with key importance in DNA replication. Additionally, we observe genes with no DNA replication annotation whatsoever: this is the case for CG8841, CG11802, crn, Dll, which according to current annotation are not related with cell cycle processes. While some of these links may undoubtedly be artefactual, others might indicate untested roles in DNA replication and repair. The fact that for some of them membership to the cluster through experimental interaction is reconfirmed through putative interactions constitutes additional reason to consider them interesting hypothesis in a potential experimental study.

To summarise, the data shows that the algorithm returns putative interactors which create tightly interconnected clusters composed of members operating in clearly related biological processes: some of the linking proteins appear to be poorly known and annotated, which means that the algorithm might be creating functional association hypotheses worthy of lab testing.

2.4.4 A *P. Falciparum* Putative DNA Replication Network

The sub-network obtained for *P. falciparum* (designated `NET_PFAL_DNArep` from now on) comprises 146 nodes and 514 edges (Figure A.2, Page 187). It relates DNA replication genes to one another in a vast complex, wiring together 15 of the 16 seed molecules. Given the relatively high number of interactions and participating genes in the sub-network, I decided to carry out a refinement of the interaction candidates obtained, using the prioritisation metrics in

`Bio::Homology::InterologWalk`. As described earlier, `Bio::Homology::InterologWalk` can optionally output a summary index called the IPX (Section 2.2.5, Page 42) for each of the putative interactions produced. This numerical estimate can be employed to select a high support "backbone" network, based on strong biological evidence, by pruning nodes connected through low support putative protein interactions. Figure 2.25-A shows the IPX distribution for `NET_DS_PFAL_putative`. As expected from the description in Section 2.2.5, Page 42, the distribution is roughly tri-modal and the IPX values are divided in three major groups: Figure 2.25(A-I): the experimental interaction is spoke-expanded and at least one of the two orthology projections is *not* one-to-one. Figure 2.25(A-II): either the experimental interaction is spoke-expanded or at least one of the two orthology projections is *not* one-to-one. Figure 2.25(A-III): the experimental interaction is not expanded from a spoke-complex and the orthology projections are *both* one-to-one.

In order to visualise the composition of putative protein interactions in `NET_PFAL_DNArep`, I set two IPX thresholds, one cutting off the data distributed around the left-most mode, and another discarding the data distributed around the first two modes. An analysis of the modes in Figure 2.25-A yields

$$\mathrm{IPX}_{\mathrm{thr}_1} = 9, \qquad \mathrm{IPX}_{\mathrm{thr}_2} = 16. \tag{2.21}$$

I then mapped IPX values to edge thickness in `NET_PFAL_DNArep`, obtaining the graphs in Figure 2.25-C,D. Figure 2.25-D shows the putative protein network which, according to the algorithm, retains the highest support putative interactions. From 2.25-D, removing all genes not directly interacting with any `DNA replication` genes and all experimental interactions derived from spoke-expanded complexes, I obtained a core *P. falciparum* DNA replication model, `NET_PFAL_DNArep_HQ`, schematised in Figure 2.26. As with the fly example, the introduction of putative protein interactors for *Plasmodium* has organised the existing GO annotated DNA replication genes in interesting clusters. Additionally, due to the incompleteness of the GO annotation, many genes with clear roles in DNA replication (or orthologues of genes with known roles in DNA replication) having no such biological process labelling in GO[44] are part of `NET_PFAL_DNArep_HQ`. Lastly, a number of new genes lacking functional annotation join the clusters, and due to their high support putative connectivity with dense clusters of DNA replication genes represent interesting candidates for experimental analysis.

The largest connected component in `NET_PFAL_DNArep_HQ` features 7 seed genes organised in three high connectivity areas and is shown in Figure 2.26-1. One of these clusters (Figure 2.26-1, blue shading, *right*) is composed almost exclusively of putative interactions, and four of the genes (white node, red border) are in the overall interactome (`NET_DS_PFAL_union`) purely through interolog transfer. The remaining seven have been observed both in `NET_DS_PFAL_known` and in `NET_DS_PFAL_putative`, but they are connected to the genes in Figure 2.26

[44] Meaning they were not picked when selecting the 16 DNA replication seeds

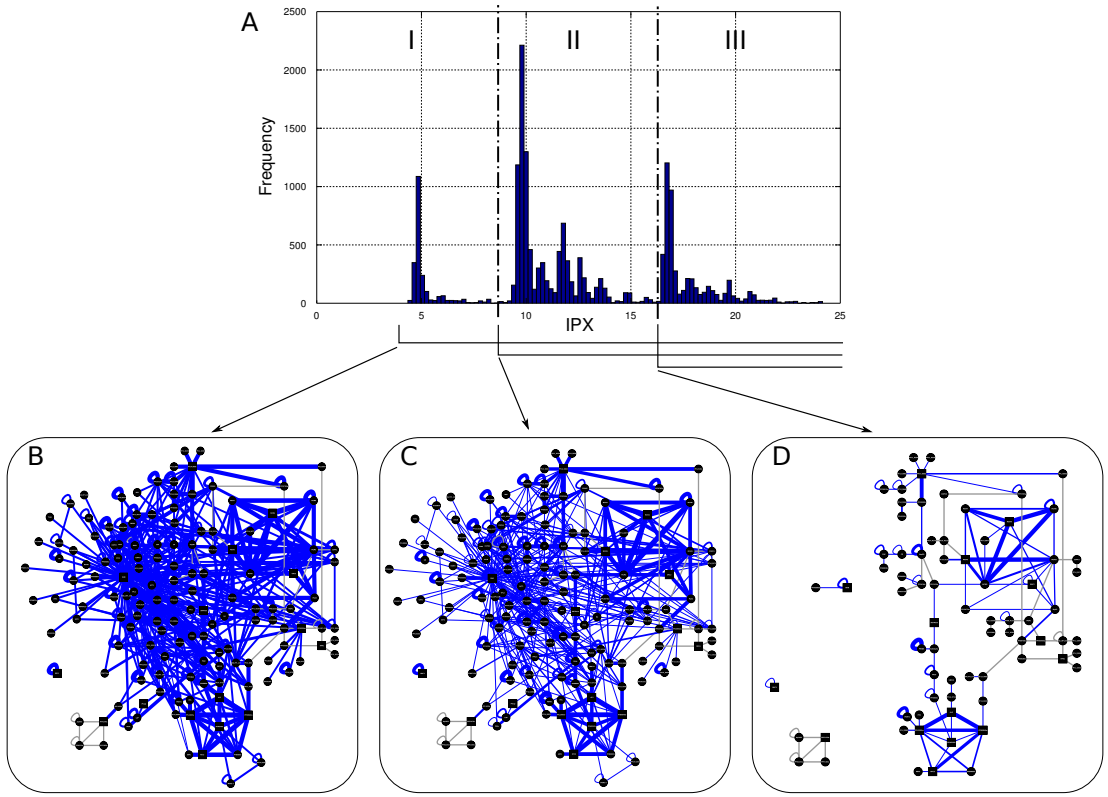

Figure 2.25: Effect of varying IPX cut-off levels on sample putative protein interaction network. **A**: Interolog Prioritisation Index (IPX) distribution for `NET_DS_PFAL_putative`. The three quadrants (AI, AII, AIII) highlight three modes in the distribution. **I**: putative interactions labelled with an IPX in this quadrant are projections of proteins belonging to low scoring spoke-expanded complexes. Moreover, at least one member of the two orthologous pairs (forward and backward) has in-paralogues (i.e., at least one of the two orthologous pairs is a 1:many or many:many orthology). **II**: putative interactions labelled with a score in this quadrant are projections of proteins in complexes or have been obtained through not-optimal orthologues (again, 1:many, many:many). **III**: members of this high-scoring sector are projections of binary experimental interactions mapped strictly through 1:1 orthology relationships. **B**: `NET_PFAL_DNArep`. **C**: subset of `NET_PFAL_DNArep` where the IPX cut-off is after the 1st quadrant **D**: subset of `NET_PFAL_DNArep` where the IPX cut-off is after the first two quadrants. **B-D** Thickness of blue edges is proportional to its IPX.

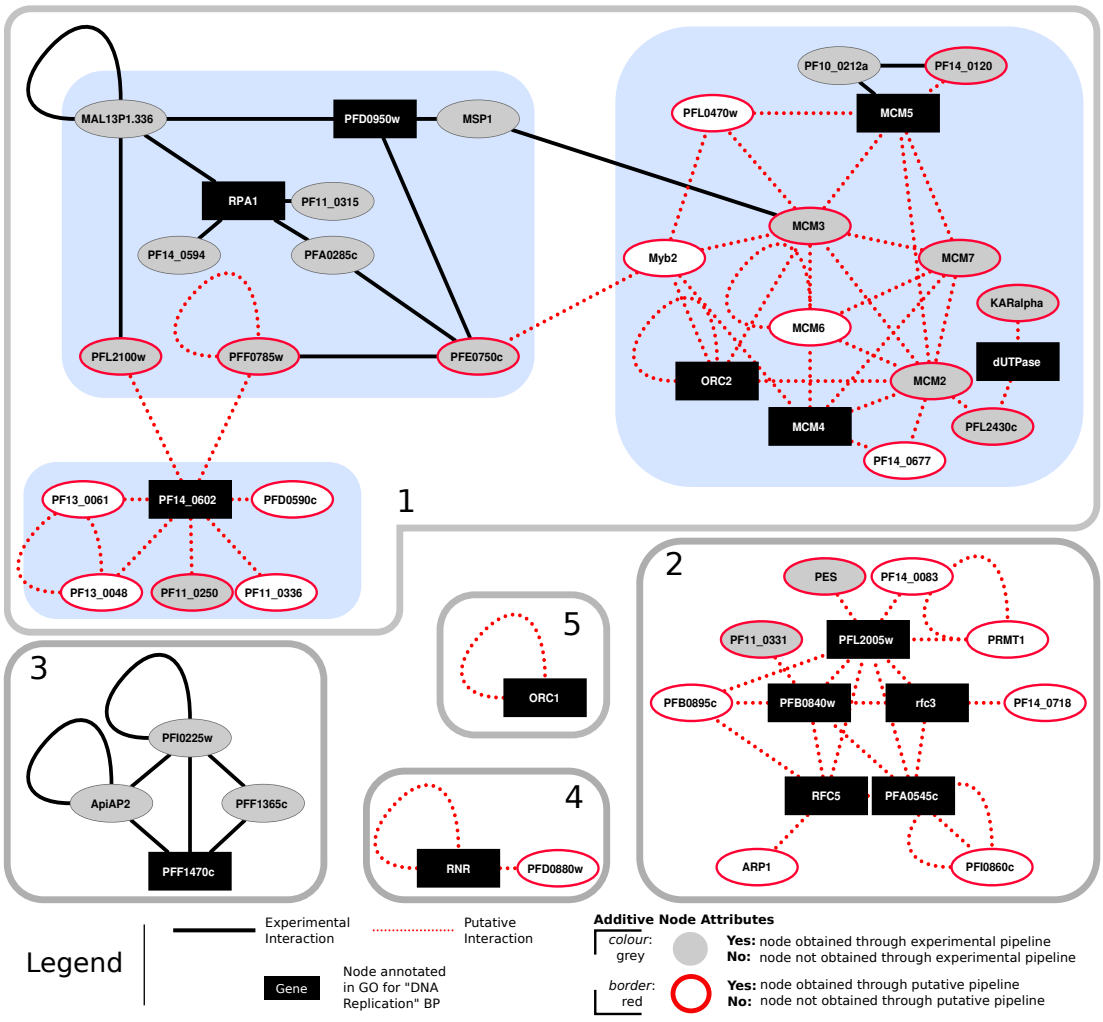

Figure 2.26: High confidence `DNA Replication` GO-annotated genes in experimental and putative *P. falciparum* interactome. Data extracted from `NET_DS_PFAL_union` as follows: 1. select all genes annotated with `DNA replication` GO biological process 2. select all their nearest neighbours 3. use IPX threshold ($\mathrm{IPX}_{\mathrm{thr}_2} = 16$), prune all non-direct interactors of DNA replication genes, remove experimental interactions obtained from spoke-expanded complexes. Solid connections (black) are EBI IntAct experimental interactions, dotted connections (red) are putative predictions originally in `NET_DS_PFAL_putative`. Nodes are described in key.

exclusively through putative interactions[45]. From a functional perspective, this sub-cluster is of particular interest because it almost completely describes the Mini Chromosome Maintenance complex which, as discussed earlier, is a well conserved hexamer of polypeptides organised in a ring structure, with a crucial role in both the initiation and the elongation phases of eukaryotic DNA replication. The loading of the MCM complex to the origin of replication is orchestrated by Cdc6/Cdc18 and Cdt1 equivalents (the network suggests that the transcription factor myb2, orthologue of CDC5, could have such a role in *Plasmodium*) and happens after the binding of the Origin Recognition Complex (ORC) to the replication origin. As described in Section 2.4.3, the ORC is a multi-subunit DNA binding complex: one of the units, ORC2, is present in the sub-cluster and correctly represented as an interactor of 2 MCM subunits. ORC1 (the only sub-unit in the complex having ATPase activity in yeast [Klemm et al., 1997]) also appears in the network, though in a separate sub-cluster. This was probably disconnected from the remaining sub-unit during the IPX thresholding, due probably to the presence of in-paralogues.

Moving on to the second sub-cluster in Figure 2.26-1 (blue shading, *top left*) we notice that this appears to be mainly composed of molecules with functions in nucleotide/RNA binding proteins (PFA0285c, PFE0750c), translational initiation (MAL13P1.336) and ATP binding (PF11_0315), interacting with the Replication Protein A1 (RPA1), the *Plasmodium* homologue [Voss et al., 2002] of a well conserved eukaryote protein, known to bind and stabilise single-stranded DNA intermediates preventing complementary DNA from reannealing [Wold, 1997]. The existing functional annotation linked to the proteins in this cluster seem to suggest a generic role in cell cycle control and cell growth. PF14_0602 is the parasite's homologue of the DNA Polymerase Alpha sub-unit B (POLA2) [Collins et al., 1993] which, together with the second sub-unit, POLA1, is responsible for coupling the polymerase alpha/primase complex to the cellular replication machinery during the early stages of replication. The *Plasmodium* orthologue of POLA1, PFD0590c, is retrieved by a putative interaction and appears in the cluster. Lastly, the connected cluster in Figure 2.26-2 is a complete reconstruction of the Replication Factor C complex in *Plasmodium* based exclusively on putative data retrieved by the algorithm: RFC is a heteropentamer with sub-units Rfc1 (PFB0895c), Rfc2 (PFB0840w), Rfc3 (rfc3), Rfc4 (PFL2005w), and Rfc5 initially found in yeast [Cullmann et al., 1995].

The results once again show that the methodology employed has allowed to relate highly conserved proteins in meaningful complexes, has reintroduced in the DNA replication model DNA replication-related proteins that had been missed by existing GO annotation and at the same time has proposed evidence regarding genes lacking any form of functional assignment, thus producing a list of highly supported genes which represent good candidates for further testing.

[45] Meaning they have other interactors somewhere else in `NET_DS_PFAL_known`

2.5 Conclusions and Further Work

In this chapter I presented a methodology to retrieve, prioritise and visualise putative protein interactions using interolog mapping. I implemented the method in `Bio::Homology::InterologWalk`, a collection of programming scripts in Perl. Unlike previous efforts, this Perl library (a) automatically connects to orthology and protein interaction data web-services to generate up-to-date predictions "on the fly" (b) outputs its predictions in the form of simple text files, allowing to use its methods, or the data it produces, within the context of pipeline-based workflows of wider scope (c) optionally flags the predictions on the basis of related biological metadata through a prioritisation index, allowing the selection of a subset of candidates with high biological support, for *in vivo* validation.

I formally validated the accuracy of the tool, the correctness of the implementation, and presented a ROC curve-based analysis to assess the association between the IPX and known true positive interologs across several inter-species reference sets. I tested the potential of the method to retrieve putative interactions on the genomes of two eukaryotes, *Drosophila melanogaster* and *Plasmodium falciparum*, obtaining large putative interactomes for both. I looked more closely at a number of subsets of these interactomes, based on annotated evidence for functional roles linked to the DNA replication gene ontology biological process. In parallel, I utilised IPX thresholding to create a core network from a large *Plasmodium* DNA replication-related network, evidencing a smaller sub-network for which there is strong biological and experimental support. The usage of `Bio::Homology::InterologWalk`, in combination with these analyses and selection techniques, allowed the identification of several novel interactions that interconnected known domain-related genes in biologically meaningful clusters, as well as a number of novel nodes with no previous known link to DNA replication.

I made the implementation freely available for non-commercial purposes by uploading it on the Comprehensive Perl Archive Network[46]. `Bio::Homology::InterologWalk` is modifiable under the GNU GPL license to allow whoever is interested to make corrections, enhancements, improvements, as well as customisations to adapt the software to specific projects. The package includes full documentation and example scripts to simplify usage.

A few points about the algorithm and methodology need further discussion. The interaction prioritisation index (IPX) is designed to encapsulate biologically relevant principles that relate directly to the assessments currently made manually by many researchers using interaction data. I would like to stress, however, that the IPX measure for an interaction is not fully explored here and that a full validation is not possible due to the current poor coverage of protein interaction data across species. In our experience however, the IPX has proven to be a useful summary of biological metadata for protein interactions. As such, it must be intended

[46] search.cpan.org/~ggallone/Bio-Homology-InterologWalk

as a pragmatic aid to candidate prioritisation, with room for improvement and refinement.

There are other aspects of the implementation showing room for improvement. As anticipated in Section 2.2.3, the interologs produced by `Bio::Homology::InterologWalk` are based on interactions at the gene level: the tool suggests pairs of interactors based on gene identifiers, and is not currently able to provide putative information at the transcript level: isoform information from the reference interaction, if present, is not used during the backward orthology projection, due to Ensembl being unable to return gene information starting from isoform-level UniprotKB information. The complete UniprotKB IDs for each experimental interaction used by `Bio::Homology::InterologWalk` to produce interologs are, however, always available in the output datasets for manual inspection. As a consequence of this observation, the predictions produced by the tool are more akin to gene interaction networks, and would require additional manual curation if evidence at the transcript level was needed.

The API used by the module for the orthology manipulations is based on the Perl programming language. This is a general purpose, dynamic type, interpreted language not originally designed for complex applications. As a consequence of Perl being a dynamic type language, the Perl memory garbage collector is unable to handle circular references: this can create memory leaks in complex, layered software like a large API and, in general, performance issues which can at times impact on the operation of `Bio::Homology::InterologWalk` (depending on the input size). It would be interesting in the future to adapt the code to API written in more optimised languages, or to redesign the code to use REST/SOAP fully throughout the data collection process (should one day Ensembl enable REST/SOAP programmatic data access, of course).

The potential of the Protein Conservation Score has not been investigated using the sample datasets discussed. It would be interesting to evaluate the top scoring interologs for *Drosophila* and *P. falciparum*, and the topological properties of the clusters showing the highest PCS values and verify the agreement or disagreement between PCS cluster and IPX clusters. Additionally, the algorithm used to select the best quasi-clique containing the reference interaction is not optimal: I relied on a rather crude heuristic to allow reasonably acceptable performance. This heuristic is based on a hard limit — an upper boundary on the maximum number of nodes for which the quasi-clique optimisation is attempted: if two experimental interactors are part of a complex of $n > n_{\text{thrs}}$ nodes (with n_{thrs} a hard limit set within the code), quasi-clique optimisation is aborted, and the density value for the initial complex is returned. A number of efficient, stochastic clique finding algorithms have been proposed, for example the Reactive Local Search algorithm (RLS) [Battiti and Protasi, 2001] and the Dynamic Local Search for Maximum Clique (DLS-MC) [Pullan and Hoos, 2006]. The DLS-MC algorithm is based on the idea of assigning penalties to nodes that are selected to be part of a clique. The Reactive Local Search algorithm operates by maintaining a current clique and modifying it with two

basic moves: node addition and node removal. After a node is removed or added, it will not be reconsidered for addition or removal before S steps have been performed. Recently, an attempt to adapt these algorithms for the γ-quasi clique case has been proposed [Brunato et al., 2008]. It would be interesting to evaluate the possibility of optimising the PCS algorithm in light of such recent research.

The decision to test the methodology proposed on the *Plasmodium falciparum* genome and the focus on its DNA replication protein interaction sub networks needs further discussion. The parasite is the cause for one of the most life-threatening diseases — deaths linked to Malaria are estimated at 2 millions per year [Voss et al., 2002]. DNA-related processes in *P. falciparum* are of special interest for the scientific community, due to the complexity of the parasite life cycle (involving several cycles of invasion, growth and schizogony) and to its ability to effectively adapt its DNA metabolism to thwart the immune system of its host. It is hypothesized that particularly specialised DNA replication pathways are responsible for the success of the parasite and for its ability to counteract immune response and antimalarial drugs [White and Kilbey, 1996]. This and the unusually high AT rich genome might indicate peculiarities in the parasite's replication machinery: it is, therefore, extremely important to evidence differences in the DNA replication mechanism between *P. falciparum* and other eukaryotes to gain insights which might eventually lead to the identification of new potential drug targets for malaria therapy.

However, the utility of comparative data to help elucidating the interactome of *P. falciparum* through functional inference from model interactomes might be questionable. According to a study published a few years ago [Suthram et al., 2005] the protein interaction network of *P. falciparum* diverges from the ones in eukaryote model organisms: the study finds very little conservation between complexes in *Plasmodium* and complexes in yeast, fly, worm and *Helicobacter pylori*. However, the study uses rather old orthology prediction methods (Pathblast[47], Blast Best Hits, E-value thresholds) which, for reasons explained earlier, yield non-optimal predictions of homologues between diverged *taxa* where several duplication events have occurred. Additionally, only one source of Y2H protein-protein interaction is used [LaCount et al., 2005], and it would be interesting to know if the same conclusions would have been found today with improved orthology prediction methods, larger, low throughput *Plasmodium* datasets available, and the usage of more genomes. A later study [Wuchty et al., 2009] again concludes that the gene/protein sequences of *Plasmodium* feature peculiarities which hamper the detection of orthologues in other organisms, however as before Wuchty et al. [2009] utilise best-hit homologue detection methods and a similarly limited choice of reference genomes, which lead us to argue that further research might be needed to provide additional experimental evidence to corroborate this "peculiarity" hypothesis. Interestingly, one of the

[47] www.pathblast.org/

few conserved clusters discussed by Suthram et al. [2005], describing the MCM complex and a group of heat shock proteins, resembles quite closely one of those resulting from the study carried out in Section 2.4.4, and this provides independent support for the results obtained with `Bio::Homology::InterologWalk`.

The usage of paralogy rather than (or in addition to) homology information for information transfer with `Bio::Homology::InterologWalk` has not been discussed or tested. While it is technically straightforward to utilise the methodology and algorithm to map protein interactions through paralogy, in both of the examples presented paralogues have been discarded immediately. This has been done mainly for two reasons.

Firstly, one of the initial design ideas driving the development of the tool was to provide a way to annotate very poor or non existent interactomes through information transfer from well studied model interactomes. Even when considering the case of organisms with relatively rich experimental interactomes (e.g. *Drosophila*) I hypothesized that, due to various biases in experimental research[48] a new perspective on the interactome could be gained by projecting interactions from different species.

The second reason why paralogues have not been used has to do with the acknowledgement of the so called "standard model" of evolutionary genomics. It is widely assumed that orthologues are more likely to retain the ancestral gene function; also, evidence points toward higher conservation of structural parameters like intron position [Henricson et al., 2010], protein structure [Peterson et al., 2009] and domain architecture [Forslund et al., 2011] between orthologues. Paralogues, on the other hand, appear to provide the raw material from which functional diversity evolves: they have been linked to neo-functionalisation by some [Ohno, 1970], while others [Lynch and Conery, 2000] have argued that gene duplications and paralogy involve primarily non-adaptive substitutions leading to either non-functionalisation of one duplicate, or to sub-functionalisation, but not neo-functionalisation. In both cases, the result is that paralogues have been deemed less reliable for functional transfer [Tatusov et al., 1997].

It must be said, however, that large experimental studies corroborating or refusing this model are scarce [Studer and Robinson-Rechavi, 2009]. Given the recent availability of genome-wide reliable orthology predictions and comprehensive Gene Ontology-based functional annotations for some genomes, this "standard model" has come under scrutiny. One paper has recently discussed this "orthology conjecture" [Nehrt et al., 2011] with surprising results: paralogues appear to be more functionally similar than orthologues. Nehrt et al. [2011] devise a large scale test utilising homology data from human and mouse and functional similarity measures based on human/mouse Gene Ontology annotation. The most surprising finding shows an absence of correlation between functional similarity and sequence identity in human-mouse

[48] Many times, scientists study one specific organism (rather than any other) to answer particular questions; many times, scientists studying one organism rather than the other share a particular *forma mentis* which sets their results apart from those they might have had studying a different organism, and so on.

orthologues sequences; on the other hand, functional similarity between paralogues is positively correlated with sequence identity. The paper has several serious flaws, however one of the most relevant is probably the rejection of the "orthologue conjecture" almost entirely on the basis of computational analysis of existing GO annotations, with no in-depth analysis of specific examples. Another point that raises some concern is the sole usage of mouse-human examples to draw rather universal conclusions. Overall, what Nehrt et al. can infer from their study is not that the standard model is wrong, but rather that the gene ontology data for human and mouse is biased: a bias in annotations arises because research programs in human and mouse models tend to discover aspects of orthologous gene function that are not completely independent. However, Nehrt et al. fail to put in place a series of anti-bias measures and an extensive experimental validation in support of their claim. The paper was harshly criticised for its shortcomings on post-publication review forums[49] and a rebuttal came from a publication signed by the Gene Ontology consortium [Thomas et al., 2012], which mainly focused on the wrong usage Nehrt et al. [2011] had done of Gene Ontology information. Subsequently, another study defined and addressed the shortcomings of the work by Nehrt et al. [2011] once again re-establishing the "standard model" by large scale data analysis of data from 13 genomes [Altenhoff et al., 2012]. Altenhoff et al. [2012] find, amongst other results, that after controlling for bias GO molecular function appears to be "strongly conserved between even distant homologues, which supports the received wisdom of predicting this type of annotation on the basis of conserved protein domains".

[49] f1000.com/12462957?key=5g7rjmt7xzv2y32

CHAPTER 3

A Protein Interaction Network for *Drosophila* Ciliogenesis

In the introductory chapter I have briefly summarised wet-lab work done in the Jarman laboratory, aimed at understanding some of the dynamics behind peripheral nervous system development in the fly. In Chapter 2 I have described a computational methodology and proposed an implementation aimed at gathering large amounts of putative protein interaction data which can be used to provide computational hypotheses to guide a prioritisation study of PNS genes in *Drosophila*. In this and in the next chapter I will describe a number of options to support protein interaction datasets produced with the software in Chapter 2 using insights from the transcriptome data described in the introductory chapter.

Specifically, here I will document a pilot study where I used information from protein interactions supported by transcriptional data to select novel *Drosophila* genes with a potential role in ciliogenesis. The protein interaction data represents the main data source: the approach is based on a putative protein network built from data compiled in the Drosophila Cilia and Basal Body (DCBB) dataset [Laurencon et al., 2007], a list of fly orthologues of genes with a tested role in ciliogenesis in several organisms. The supporting source of information is represented by the PNS transcriptome time series data described in the introductory chapter [Cachero et al., 2011]. Starting from a network of putative fly interactions for the genes in the DCBB, I identified network regions characterised by interesting connectivity and employed the Interolog Prioritisation Index (described in Chapter 2), together with PNS transcript fold change information from Cachero et al. [2011] to build evidence implicating three candidate novel genes, CG17599, CG30441 and CG31320 with ciliogenesis. I conclude the chapter describing experimental validation of these candidates conducted by members of the lab, which followed the computation prediction and confirmed that the genes are required for sensory neuron function in *Drosophila*, providing experimental evidence linking the genes to fly ciliogenesis.

3.1 Ciliogenesis and the DCBB dataset

Cilia are eukaryotic microtubule-based cell surface protrusions which perform a wide range of motility-related and sensory tasks. In spite of evident functional and structural diversity, the core pathways defining the process of cilia development, or ciliogenesis, are well conserved [Ishikawa and Marshall, 2011]. It is hypothesized that cell-type specific gene expression programs are required to adapt and modify this common set of assembly programmes to generate cilia diversity [Silverman and Leroux, 2009], however little is known about the details of these programmes and about the transcription factors controlling the underlying regulatory events. As discussed in the introductory chapter, *Drosophila melanogaster* represents an ideal model for the study of ciliary development, as it only contains cilia on two of its cell types, sperm and the sensory neurons (as opposed to higher eukaryotes, where cilia are nearly ubiquitous [Pazour and Witman, 2003]). This greatly facilitates *in vivo* analysis [Lee et al., 2008]. Also, as anticipated, in fly Ch- and ES- sensory neurons dendrite endings are modified cilia responsible for sensory signal reception and transduction [Dubruille et al., 2002]. It follows from this that our study of PNS development and gene regulation in sensory neurons is ideally suited to investigate the process of ciliogenesis. Indeed, a number of ciliogenesis-specific sub-systems have been elucidated thanks to evident sensory defects after mutations in fly [Han et al., 2003; Lee et al., 2008].

In *Drosophila*, a general activator of genes involved in sensory cilia formation is *Rfx*, introduced in Chapter 1. The majority of *Rfx* targets feature an X-box regulatory sequence about 150 to 50 nucleotides upstream of the translation start site [Swoboda et al., 2000]. Recently, Laurencon et al. [2007] showed that *Rfx* target genes are largely conserved between *C. elegans* and *Drosophila*. Starting from a subset of known ciliogenesis genes in worm and fly (known to be regulated by *Rfx* in fly and by its orthologue DAF-19 in worm) Laurencon et al. used divergent *Drosophila* species to determine a consensus *X*-box binding sequence. Using the sequence, they scanned the fly genome to build a candidate *Rfx* target list, which was then refined to varying degrees based on consensus sequences of increasing stringency. A final validation on the most stringent 83 target set shared between worm and fly led to the characterisation of 16 genes under Rfx control. Of these, 11 had not been described as Rfx targets before. Additionally, while 9 of these 11 genes showed evidence for an involvement in ciliogenesis in the literature, the remaining 2 had never been related to ciliogenesis in any organism. Reporter analysis was then utilised to prove that 3 of the 11 novel Rfx targets encode proteins specifically localized in the ciliated endings of *Drosophila* sensory neurons. Interestingly, Laurencon et al. also showed that all of the *Drosophila* orthologues of genes that had, until then, been implicated in human Bardet-Biedl syndrome (BBS, a human ciliopathy) are under the control of Rfx. Additionally, the only BBS protein with no worm orthologue, BBS4, is 17-fold

down-regulated in *Rfx*-mutant flies. These findings drew an interesting link between *Rfx* and the Bardet-Biedl syndrome which provided insights for vertebrate ciliopathy research studies.

One of the datasets presented by Laurencon et al. acquires particular relevance in the rest of this chapter. In order to test if their X-box gene lists were enriched for ciliogenesis genes, Laurencon et al. collected all genes showing evidence for ciliogenesis involvement in a number of heterogeneous studies conducted on several distinct species. These included human [Ostrowski et al., 2002; Andersen et al., 2003; Gherman et al., 2006], the single-cell green flagellated alga *Chlamydomonas reinhardtii* [Pazour et al., 2005; Stolc et al., 2005], a flagellated protozoan, *Trypanosoma brucei* [Broadhead et al., 2006] and *C. elegans* [Efimenko et al., 2005; Blacque et al., 2005]. The dataset also included genes implicated in ciliogenesis by two studies via comparative analyses of ciliated versus non-ciliated genomes. The first of these studies, by Avidor-Reiss et al. [2004], uses a mixture of five ciliated (human, *C. elegans*, *P. falciparum*, *C. reinhardtii* and *T. brucei*) and three non-ciliated organisms (*A. thaliana*, *S. cerevisiae* and *D. discoideum*); the second, by Li et al. [2004a], uses *Arabidopsis* against human and *Chlamydomonas*. The blast hit-based fly orthologue list (815 genes) of these genes is termed by the authors the 'Drosophila Cilia and Basal Body' (DCBB) knowledge base. Based on this DCBB list, the team was able to show that their X-box candidate genes are statistically enriched for ciliogenesis genes which led them to argue that the X-box conservation is a good marker for ciliogenesis association.

I decided to use the DCBB list as a starting set for a protein interaction retrieval experiment based on `Bio::Homology::InterologWalk` and the IPX. This will be described in the next section and will constitute the bulk of this chapter. There are multiple reasons why I selected the DCBB knowledge base for a protein interaction retrieval experiment to find potential novel ciliogenesis genes. The DCBB represented the most comprehensive fly ciliogenesis compendium available at the time of my study of this subject. Additionally, I hypothesized that the heterogeneity of the underlying data (and the large number of genomes involved) would make it less prone to bias than a dataset of comparable scope obtained from a single study using a single reference genome. Having said that, the DCBB was also chosen for its obvious flaws. Firstly, the evidence associating some of the genes in the reference genomes to ciliary function is, in some cases, purely computational — when available, experimental validation was generally obtained only for a selection of the proposed candidates. Secondly, the techniques used to obtain the fly interologs from the reference genomes are not optimal, given the large divergence between some of the species employed in the study. Finally, rather arbitrary X-box distance assumptions were made and site-hopping between divergent species was not taken into account. By retrieving binary fly protein interactions between DCBB members I hypothesized I could corroborate available functional association with protein interaction-based, *guilt by association* evidence. Additionally, by using the robust orthology retrieval paradigms

in `Bio::Homology::InterologWalk` I expected to reconfirm a subset of the fly ciliogenesis homologues obtained by Laurencon et al.. Overall, I decided to use the interolog walk methodology and `Bio::Homology::InterologWalk` to systematically expand what had been obtained using customised annotation methods by Laurencon et al.

3.2 A DCBB putative protein interaction network

I obtained the list of 815 DCBB gene IDs from the original supplementary data published in Laurencon et al. [2007]. I then double-checked each gene ID against Ensembl for consistency, and obtained current Flybase IDs for each identifier. These updated IDs represented the input dataset for `Bio::Homology::InterologWalk`. The Ensembl database I chose for the orthology data collection was Ensembl Vertebrate. While its main focus is the annotation of higher eukaryotes, as of release 59[1] Ensembl Vertebrates still included genome and homology/variation data for the main well-studied invertebrate model organisms, including *S. cerevisiae*, *C. elegans* and the fly.

As for the `Bio::Homology::InterologWalk` run set-up, all homology relationships belonging to the *paralog* class were discarded, for the reasons explained in Chapter 2. This included cases identified by Ensembl as *possible ortholog*, instances where the duplication vs. speciation nature of the event cannot be reliably resolved by TreeBeST: these are often cases pointing to long-distance relations which tend to be upgraded to *bona-fide* orthologues in successive version of the Compara pipeline. Regarding the interaction collection phase, I used the filtering options in `Bio::Homology::InterologWalk` (Chapter 2, Section 2.2.4, Page 41) to query EBI IntAct retaining only interactions satisfying both the following criteria:

1. For the PSI-MI field `detection method`, the interaction is tagged with a code specifying an experimental detection method or a specialisation of it. This excluded entries based on other computational prediction methods and entries lacking complete annotation.

2. For the PSI-MI field `interaction type`, the interaction is tagged with an ontology code specifying a physical association or specialisation thereof. This excluded entries based on co-localisation or genetic interaction evidence.

I decided against discarding binary evidences obtained from spoke-expanded complexes and marked accordingly in IntAct. The main reason for this is that ciliogenesis is heavily reliant on the activity of protein complexes and the products of many ciliogenesis genes are known to act in complexes (e.g. Intraflagellar Transport (IFT), a bidirectional motility along axonemal microtubules that is essential for the formation and maintenance of cilia, is led by two groups

[1] V.59, 5 August 2010.

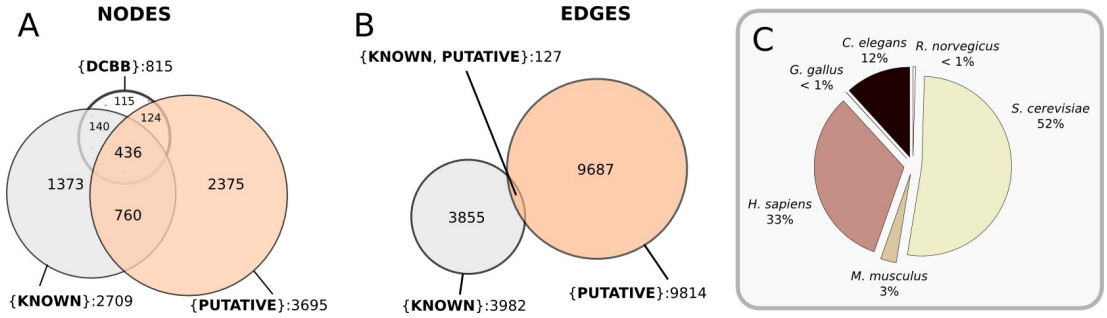

Figure 3.1: **A**: Quantification of overlap between three sets: (1) DCBB: initial gene ID set. (2) PUTATIVE: all unique gene IDs participating in the putative protein interaction network (3) KNOWN all unique gene ids participating in the known protein network **B**: Overlap between the edges of the direct protein network and those of the putative network. **C**: Reference genome break-up for the putative interaction data.

	B::H::I Pipeline	
	Putative	**Known**
Datasets		
Gene IDs	**815**	**815**
Reference Genomes used	51	1
Orthologues	57321	NA
Interactions in Reference Genomes	34931	NA
Failed Backward Orthologies	377	NA
Total Interactions	15262	4392
From Spoke-expansion	911(5.9%)	10 (.2%)
Unique PP Pairs	**9814**	**3982**
Surviving IDs (% **Gene IDs**)	560 (68.7)	576 (70.7)
Networks		
Nodes	**3695**	**2709**
Edges	**9814**	**3982**
Novel Nodes (% **Nodes**)	3135 (84.8)	2133 (78.7)

Table 3.1: `Bio::Homology::InterologWalk` usage statistics using the *Drosophila* Cilia and Basal Body (DCBB) gene list as the input dataset. Homology data is from Ensembl 59 (5 August 2010), protein interaction data from EBI IntAct (September 2010, v. 1.1.6). Results obtained using the two available `Bio::Homology::InterologWalk` pipelines — putative and experimental — are shown. In the putative pipeline, the data shown are relative to interactions obtained through interolog mapping. In the experimental pipeline, the DCBB dataset has been queried against EBI IntAct to mine all the experimental molecular associations known in the literature.

of proteins that can be biochemically fractionated as complexes [Rosenbaum and Witman, 2002]). By discarding information from isolated complexes (obtained for example, through Tandem Affinity Purification) I hypothesized I would miss important insight into ciliogenesis processes discovered in some of the reference genomes. The obvious trade-off coming with this choice (the addition of pairwise interactions for which there is no direct evidence) was managed through penalisation of the spoke-evidence through IPX prioritisation (as done in Chapter 2).

I carried out two parallel runs of `Bio::Homology::InterologWalk`: one processed the DCBB through the putative pipeline, while the other processed the same genes through the direct pipeline, to obtain all fly experimental interactors for the dataset. Table 3.1 shows statistics relative to the resulting putative and direct datasets, and Figure 3.1-A,B presents an overview of set overlaps between the obtained direct and putative interaction networks.

`Bio::Homology::InterologWalk` identifies a total of 9814 putative protein interactions,

98.7% of which are unseen in the known interaction space for the dataset. These putative interactions involve a total of 3695 molecules. More than one third (36%) of these interactors are part of initial DCBB set or of the list of experimental interactors of the DCBB set (Figure 3.1-A). This suggests that more than one third of the molecules in the new putative network are known to the domain. While the algorithm is proposing new relationships between molecules absent in the DCBB set, it is also retrieving new relationships between DCBB genes and their experimental interactors. It is also worth noting that a portion of the original gene set accounting for 29% of the 815 genes had no known experimental interaction annotation whatsoever. Of these genes showing no interaction, 52% are found to be involved in putative associations based on the orthologue projection data.

3.3 DCBB Network Topology Analysis

In Chapter 2 I validated the methodology used in the implementation of `Bio::Homology::InterologWalk`, showing that the implementation is working correctly and that the prioritisation index shows good true positive classification performance, and is thus acting better than a random predictor. In Section 3.2 of this chapter I have generated two protein interaction networks, one based on experimental fly data and one based on interologs. My hypothesis is that, by computing the union of the two networks, an additional level of detail on the process of ciliogenesis can be gained, because the putative network is bringing in meaningful biological information, derived from reference interactomes, to the incomplete fly DCBB interaction data.

This hypothesis cannot be fully proven[2]. However, we can gather evidence to reject it: we can observe some mathematical properties of the putative network and compare them to the same properties observed for the experimental network. By doing this, we can evaluate the possibility that the putative data is grossly artefactual, i.e. composed mainly of noise aggregated through homology projection, in which case the putative data would actually be contributing damage, rather then benefit, to the understanding of this system.

3.3.1 Preliminary Network Topology Observations

In order to better appreciate the effect of adding putative protein interactions to an experimental dataset, I observed the amount of similarity between the experimental and putative networks to evaluate whether the putative network would show some of the topological signatures that have been found to be typical of many non-random networks [Maslov and Sneppen, 2002; Ravasz et al., 2002; Barabasi and Oltvai, 2004]. Again, I intended this as an "open world" test: a necessary but not sufficient condition to evaluate the biological utility of putative interolog data in this context. By open world test I shall here intend the following:

[2]Because a full experimental validation of all the putative candidates is not within the purpose of this thesis.

1. Supposing the putative data shows typical biological network signatures, nothing can be inferred, because a full experimental validation would still be needed to support claims of biological meaningfulness;

2. Supposing the putative data clearly does not show biological network signatures but instead random-noise characteristics, then it can be concluded that an extra degree of caution should be observed in making inferences out of this putative data, because the data is not in full agreement with expected topological patterns. As such, it might be as useful as randomly assembled data as far as the understanding of the system is concerned.

In the following discussion I shall adopt the following terminology to describe the networks being investigated:

1. `NET_DCBB_known` (2709 nodes, 3982 edges) — the network consisting of all the experimental physical associations involving genes in DCBB, according to EBI IntAct;

2. `NET_DCBB_putative` (3695 nodes, 9814 edges) — the network consisting of all the putative interactions involving genes in the DCBB list according to `Bio::Homology::InterologWalk`;

3. `NET_DCBB_union` (5208 nodes, 13796 edges) — the network obtained computing the union of (1) and (2) where:

   - each node is a node of `NET_DCBB_known`, `NET_DCBB_putative` or both;
   - each edge is an edge of `NET_DCBB_known` or an edge of `NET_DCBB_putative`, or both[3]

Initially, I used Network Analyzer [Assenov et al., 2008] to plot, for each of the three networks, a total of four topological indices: Degree Distribution, Average Neighbourhood connectivity, Betweenness Centrality and Average Clustering Coefficient. Plots of the resulting distributions are presented in Figure 3.2.

Figure 3.2-A (black circles) shows the degree distribution for `NET_DCBB_known` on a log-log scale. Visual inspection of the distribution reveals some heterogeneity, with a shape that might indicate *heavy-tail* characteristics[4]. This appears to be the case for `NET_DCBB_union` (Figure 3.2-A, red squares) as well. Much has been written about what can or cannot be inferred from the observation of heterogeneous node degree distributions in biological and social

[3] multiple edges were *not* collapsed into one at this stage of the analysis, to visualize overlap of experimental/putative interactions. Multiple edges were collapsed during actual network analysis, for those topological measures requiring edges of multiplicity 1.

[4] The distribution of a random variable X is said to have a heavy right tail if $\lim_{x\to\infty} e^{\lambda x} \Pr[X > x] = \infty, \forall \lambda > 0$, i.e. the distribution has heavier right tail than the exponential distribution.

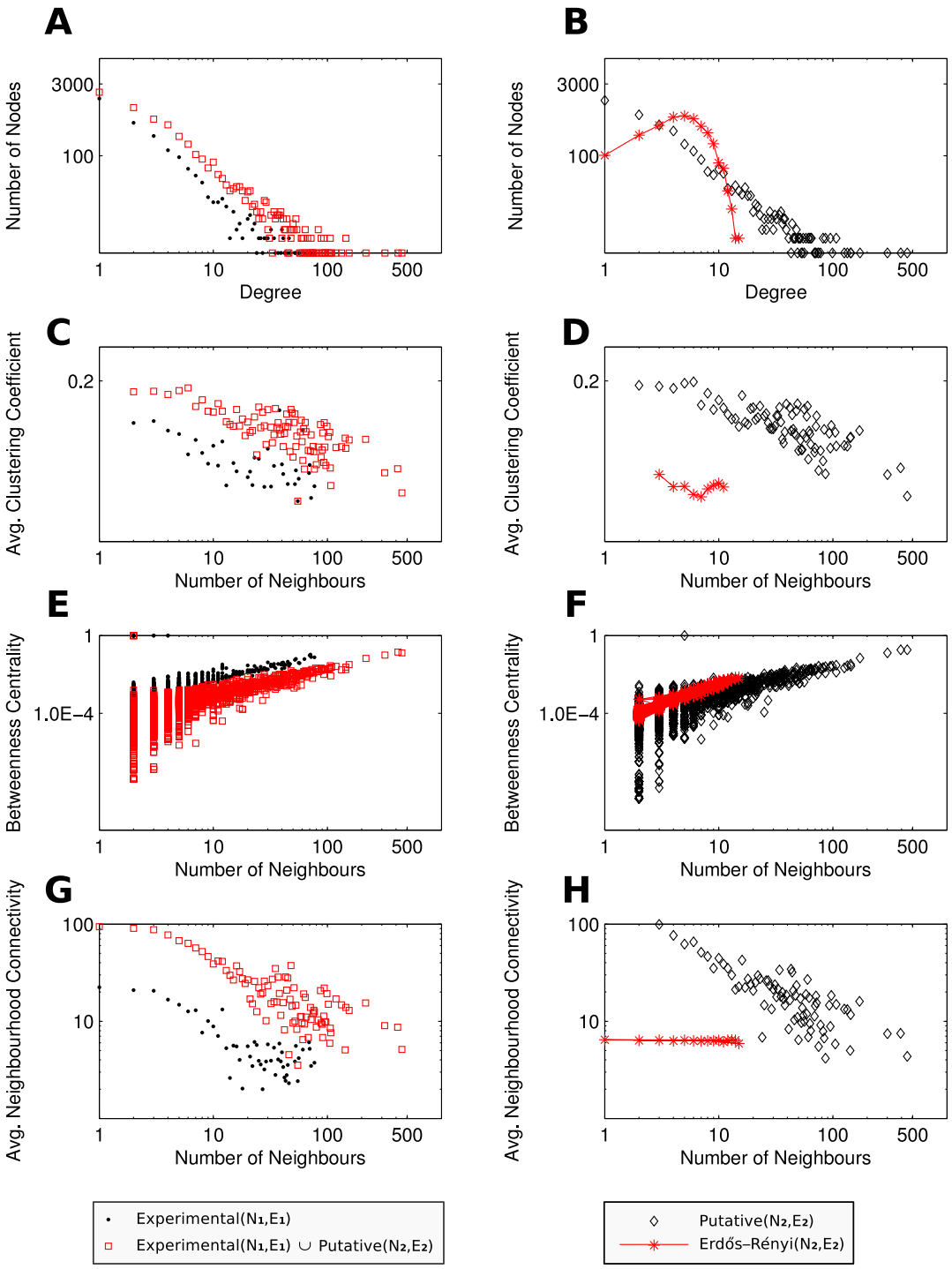

Figure 3.2: Complex Network Parameters for the DCBB protein networks. (**A**, **C**, **E**, **G**): graphs comparing, for each parameter, `NET_DCBB_known` (black dots) to `NET_DCBB_union` (red squares). (**B**, **D**, **F**, **H**): graphs comparing, for each parameter, `NET_DCBB_putative` (black diamonds) to `NET_DCBB_putative_rand` (red stars). **A** and **B**: node degree distribution. **C** and **D**: average clustering coefficient. **E** and **F**: betweenness centrality. **G** and **H**: average neighbourhood connectivity.

networks. A particular class of heavy-tail distribution, the power law, has attracted a large amount of interest. A power law is a probability distribution of the kind

$$p(x) \propto x^{-\alpha} \tag{3.1}$$

where α is a constant parameter known as the *scaling parameter*, and usually $2 < \alpha < 3$. Network studies from several disciplines (for instance Albert et al. [1999]) have reported results claimed to fit power laws. However, it is very difficult to be certain that a sample of data is drawn from a power law distribution and most of the power law claims observed in the literature have been proven to be inexact [Clauset et al., 2009]. In order to evaluate the possibility that the degree distributions shown in Figure 3.2-A might be consistent with a model like the one in Equation 3.1, a statistical analysis will be carried out later in this section. Plain observation of Figure 3.2-A can merely denote similar tail irregularities for both `NET_DCBB_known` and `NET_DCBB_union`, suggesting that the putative network does not grossly upset the degree distribution of the known network.

The Clustering Coefficient represents a measure of the degree of interconnectivity in the neighbourhood of a node. Evidence suggests that in most real world networks nodes tend to create tightly knit groups characterised by a relatively high density of connection [Holland and Leinhardt, 1971]. The Clustering Coefficient distribution (Figure 3.2-C,D) summarises the relationship between node connectivity and average neighbourhood density in the network. The negative correlation between the variables in `NET_DCBB_known` (Figure 3.2-C, black dots) has been observed in many biological networks before [Ravasz et al., 2002] and has been interpreted as an indication of hierarchical network organisation with embedded modularity.

Figure 3.2-E,F show data relative to the Betweenness Centrality [Sabidussi, 1966] of the network nodes. In general, given a graph $G = (N, E)$ with N nodes and E edges, we can define a path from $s \in N$ to $t \in N$ as a sequence of nodes and edges beginning with s and ending with t. If we define $\sigma_{st} = \sigma_{ts}$ as the number of *shortest paths* from $s \in N$ to $t \in N$, the betweenness centrality of n is defined as

$$C_B(n) = \sum_{s \neq n \neq t \in N} \frac{\sigma_{st}(n)}{\sigma_{st}}, \tag{3.2}$$

and the fraction of shortest paths from s to t that pass through n, with $0 \leq C_B(n) \leq 1$. Therefore, $C_B(n)$ is a measure of the amount of importance that node n exerts over the interactions of other nodes in the network and, as such, it is higher when n is crucial to the communication between dense communities of nodes, while it is smaller when n lies inside a sub-network and does not contribute to the flow of information through the shortest paths between all the couples of other nodes in the network. The overall distribution of betweenness centrality values plotted against node connectivities (Figure 3.2-E) appears, on visual inspection, to be roughly conserved when putative protein interaction data is added.

The Average Neighbourhood Connectivity graph (Figure 3.2-G,H) is a measure describing the relationship between the connectivity of a node and the average connectivity of its nearest neighbours. A decreasing monotone network connectivity graph like the one in Figure 3.2-G suggests that the most connected nodes in the network (sometimes called *hubs*) tend to have neighbours characterised by low connectivity. The systematic suppression of direct links between hubs in biological networks has been interpreted as a robustness measure [Maslov and Sneppen, 2002] that limits the propagation of deleterious information between network areas.

To visualise how these parameter distributions would be distributed when a completely random network is analysed, I generated a random network having the same number of nodes and edges in `NET_DCBB_putative`, using the Erdős–Rényi model [Erdos and Renyi, 1960], which chooses a graph $G(N,E)$ uniformly at random from the collection of all graphs which have N nodes and E edges[5]. I again analysed the ensuing network, `NET_DCBB_putative_rand`, using Network Analyzer, and again obtained four parameter distributions (Figures 3.2-B,D,F,H, red stars). I superimposed these distributions onto those obtained for `NET_DCBB_putative`, to visualise if a random network of the same size of the putative network would produce results still interpretable in terms of distinctive biological network signatures.

The results provide anecdotal evidence suggesting that the distributions for the random network show different patterns, when compared to those relative to the biologically-sourced network. It is interesting to look at the plot showing the betweenness centrality values for `NET_DCBB_putative` against `NET_DCBB_putative_rand` (Figure 3.2-F). The latter shows less variability in the C_B values, especially for nodes with few neighbours. This might indicate that while in the putative networks, due to the presence of community structure, there is an amount of variability in the importance of low connectivity nodes (with some being extremely "peripheral" and negligible, and others being relatively important) in the random network the absence of sub-networks or communities homogenises the betweenness of most of the nodes to very similar values.

3.3.2 Analysing Node Degree Distribution Data

Let us now go back to the degree distributions observed in Figure 3.2-A,B. Based on simple visual inspection, not much can be stated about the nature of the heterogeneity observed in the three degree distributions. Therefore, I decided to carry out a number of statistical analyses to support or refute the hypothesis that these observations might be drawn from power law distributions.

In early network studies, whether a given distribution could be appropriately described by a power law was indeed largely determined by visual inspection [Albert et al., 1999; Liljeros

[5]For example, in the $G(3,2)$ model, each of the three possible graphs on three vertices and two edges are included with probability 1/3.

et al., 2001]. Through careful application of statistical principles, it has later been possible to refute several of these power law signatures [Clauset et al., 2009]. Specifically, it has been shown that not all protein interaction network degree distributions show power law characteristics [Tanaka et al., 2005] and it has also been argued that the network properties of incomplete network dataset are rarely informative about the signatures of the complete (and often unknown) networks they were extracted from [Stumpf et al., 2005]. These arguments will be matter of further discussion for the last part of the chapter.

Here, I will run a series of tests which, while unable to give a definitive answer regarding the possibility that the degree distributions for the `NET_DCBB_*` networks agree with the power law hypothesis, will provide sound arguments to refute this hypothesis. Additionally, these tests will also provide arguments to verify if the distributions in object are still related *independently* from any power law considerations. This would prove that the heterogeneities they share are similar. The approach I will adopt is described by Clauset et al. [2009]. It combines maximum-likelihood based methods with goodness-of-fit tests based on the Kolmogorov-Smirnov statistics and likelihood ratios to compare power law fits to other heavy tail distributions.

Firstly, I computed power laws fitting the degree distributions for `NET_DCBB_known`, `NET_DCBB_putative` and `NET_DCBB_union`. I did not use linear regression to obtain the fits: the usage of linear least-squares regression to fit models to log-log plots like those in Figure 3.2-A has been shown to be problematic and to lead to systematic errors in the estimation of the parameters[6]. Here, I estimated the scaling parameter α and the lower bound[7] $x_{\min}$ of power law behaviour using the maximum likelihood method, which is proven to be an unbiased estimator in the asymptotic limit of large[8] sample size $n \to \infty$ [Barndorff-Nielsen and Cox, 1994]. In order to select the best $x_{\min}$, I used a goodness-of-fit test based on the Kolmogorov-Smirnov statistics. This provides a measure D of the distance between the probability distribution of the actual data and the probability distribution of the best-fit power law. I estimated α via ML and calculated D for each possible choice of $x_{\min}$. The final $x_{\min}$ value is the one that gives the minimum value D over all values of $x_{\min}$. Figure 3.3 shows the distributions of the three datasets together with their power law fits using the estimated parameters. Each plot shows the complementary Cumulative Distribution Function (CDF) for its corresponding dataset instead of the density function PDF, because the visual form for the CDF better evidences any fluctuations in the tail of the distributions due to finite sample size. The plots show good agreement for low degree nodes. However agreement is increasingly lost for the higher degree regions of the plots. The fits deviate more, in the high degree region, for `NET_DCBB_putative` and `NET_DCBB_union`, due probably to a lack of proportionality between the increase in the number of

[6] A typical example is the incorrect estimation of the α parameter during a power law fit of the yeast interactome degree distribution in Yu et al. [2008a].

[7] It is normally the case that data following a power law do so only for values of x above some lower bound $x_{\min}$.

[8] For finite data sets of the size we deal with, biases can be ignored because they are much smaller than the statistical error of the estimator - while biases decay as $O(n^{-1})$, the statistical errors decay as $O(n^{-\frac{1}{2}})$.

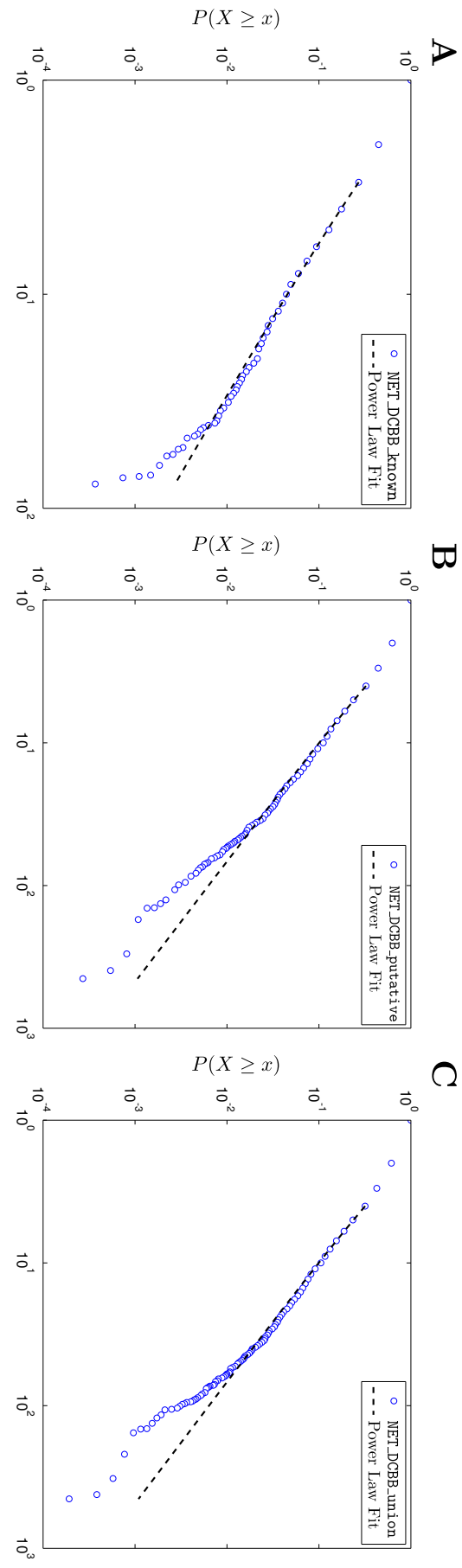

Figure 3.3: Power law fits for `NET_DCBB_known` (**A**), `NET_DCBB_putative` (**B**) and `NET_DCBB_union` (**C**) using Maximum Likelihood estimates for the α scaling and minimisation of the K-S distance for $x_{\min}$. Circles represent the CDF $P(x)$, lines represents the power law fits.

Data Set	Maximum Likelihood				Support for
	α	$x_{\min}$	n_{tail}	$p(\pm 0.03)$	Power Law
NET_DCBB_known	2.35 ± 0.07	3 ± 0.55	187.82	**0.53**	OK
NET_DCBB_putative	2.18 ± 0.09	4 ± 2.57	293.76	0	none
NET_DCBB_union	2.17 ± 0.03	4 ± 0.19	101.47	0.01	none

Table 3.2: Summary results for the power law parameter estimation and K-S goodness of fit test. Significance is for $p > 0.1$. S.e. for α (slope of power law fit), $x_{\min}$ (lower cut-off at which power law no longer applies) and n_{tail} (number of observations in power law region) are provided.

very high degree nodes versus low degree nodes.

Next, I estimated uncertainties for the computed α and $x_{\min}$ parameters. I followed Clauset et al. [2009] and implemented a non-parametric bootstrap method: given n measurements, I obtained a synthetic dataset with a similar distribution to the original by drawing a new sequence of n points uniformly at random from the original data. I repeated the process 1000 times and retrieved α and $x_{\min}$ from each randomised dataset. I took the standard deviations of these sets as my uncertainty estimates.

The power law fits obtained so far respond to optimality criteria, however they do not indicate whether a power law is a plausible model for the degree distributions in NET_DCBB_known, NET_DCBB_putative and NET_DCBB_union: a power law can always be fitted, regardless of the true nature of the data generating process behind the observations. Therefore, I decided to test the power law hypothesis quantitatively. The approach relied once more on a goodness-of-fit test based on the Kolmogorv-Smirnov distance. I sampled several synthetic data sets from a true power law distribution, measured the deviation these show from the power law form, and compared the distribution of these deviations with the deviation observed from the node degree data in the DCBB networks. The null hypothesis is that the *empirical* distance is much larger than the *synthetic* distance - in other words, the p-value is defined as the proportion of synthetic distances over the total which are larger than the empirical distance.

Therefore, a small p indicates that only a small number of synthetic distances are larger than the empirical distance, and in this case the power law is not a plausible model for the data. A large p indicates that, given a specified confidence interval, a power law might be a plausible model for the empirical data. For this test, power law is ruled out if $p \leq 0.1$. Summary data showing results for the three experiments described (power law fit parameters, parameter uncertainties and power law/data GOF test) are shown in Table 3.2. The results show that we can readily rule out power law behaviour for NET_DCBB_putative and NET_DCBB_union. In both cases, agreement with a power law generating process is strong in the low to average degree area of the plots, and the obtained parameters show agreement with NET_DCBB_known

and a similar data generating process. However, the agreement is lost for the higher degree regions. One reason for this is bias in the protein interaction subsets [Sprinzak et al., 2003] or the fact that these DCBB sub-networks are samples of the complete (unknown) *Drosophila* interactome [Stumpf et al., 2005] or most likely a combination of both reasons.

Necessary and not sufficient evidence for power law behaviour has ruled out power law characteristics for `NET_DCBB_putative` and `NET_DCBB_union`. A power law is still a plausible explanation for `NET_DCBB_known`. However, a large p-value does not prove that the power law is *the* correct data-generating process: there might be other distributions matching the data as good as, or better than, a power law[9]. Therefore I proceeded to compare the power law hypothesis for the degree distribution in `NET_DCBB_known` with alternative hypotheses: by combining p-value calculations for the power law and several other plausible competing data-generating distributions, we can obtain further information regarding the likelihood that a power law is the data generating distribution: if the p-value for the power law is high, while the p-values for the competing distributions are all low, then the case in favour of the power law is strengthened. Several established, statistically principled approaches for model comparison exist: the cross-validation approach [Stone, 1974], the fully Bayesian approaches [Kass and Raftery, 1995] or the minimum description length approach [Grünwald, 2007] are all valid alternatives. Here, I will rely on the likelihood ratio test described in Clauset et al. [2009], and based on a method by Vuong [1989], because a working implementation in R and C is freely available[10]. The likelihood ratio test computes the likelihood of the data under two alternative distributions, with the distribution having the higher likelihood (in absolute value) representing the better fit. I carried out likelihood tests for `NET_DCBB_union` and `NET_DCBB_putative`, too, to evaluate the possibility that a closer fit for both datasets could be obtained through one of the alternative distributions tested.

Results for the likelihood tests are presented in Table 3.3. For `NET_DCBB_known`, the results show that one of the alternative data-generating models carries some weight: a power law with an exponential cut-off[11] is a slightly better predictor for the data in `NET_DCBB_known` than a pure power law. The data also shows that we can rule out, for `NET_DCBB_known`, the possibility that the data comes from a Poisson, an Exponential, or a Weibull distribution. Results for `NET_DCBB_putative` and `NET_DCBB_known` are equally interesting. The table reproduces, for all three networks, the p-values shown in Table 3.2. These show that, according to the GOF tests, there is no support for power law for the putative and union degree distributions. Earlier, I suggested this loss of support could be due to the irregular shape of the tail of the distributions, possibly because of non-proportional increase of node degree with respect to

[9] A straight CDF on a log-log axes is a necessary but not sufficient condition for power law behaviour. There are many kinds of data that look straight on log-log axes but are not power law distributed.

[10] `tuvalu.santafe.edu/~aaronc/powerlaws/Rcode_README.txt`

[11] This is simply a power law multiplied by an exponential, i.e. a model of the form $p(x) \propto x^{\alpha} e^{\beta x}$.

		Poisson		Log-normal		Exponential		Weibull		Yule		PL + cut-off		Support for
Data Set	p	LR	p	LR	p	LR	p	LR	p	LR	p	LR	p	**power law**
NET_DCBB_known	**0.53**	7.38	**0.00**	-1.13	0.26	7.14	**0.00**	76.7	**0.00**	-0.33	0.74	-2.59	**0.02**	with cut-off
NET_DCBB_putative	0.00	4.97	**0.00**	-1.66	**0.09**	5.83	**0.00**	87.56	**0.00**	-0.59	0.56	-5.54	**0.001**	other
NET_DCBB_union	0.01	6.47	**0.00**	-1.80	**0.07**	7.82	**0.00**	103.01	**0.00**	-0.13	0.89	-7.37	**0.00**	other
NET_DCBB_putative_rand	0.00	-16.23	0.00	—		—		—		—		—		none

Table 3.3: Likelihood Ratio Tests for Model Comparison. For each data set, the table shows a p-value for the power law fit and likelihood ratios for the alternatives. Additional p-values refer to the significance of the LR tests [Vuong, 1989]. For each alternative, if the likelihood ratio is significantly different from zero (corresponding to $p < 0.1$), then the sign indicates whether or not the alternative is favoured over the power law model: positive values of the LR indicate that the power law is favoured over the corresponding competing model. Statistically significant p-values are denoted in bold. The Erdős–Rényi model NET_DCBB_putative_rand (which is, by definition, Poisson distributed) is included for reference. The column "support for power law" lists judgement about the statistical support for the power law hypothesis. "with cut-off": the power law with exponential cut off is a better model for the data than the power law. "other": one or more of the alternative generating distributions fit the data better than a power law. For both NET_DCBB_union and NET_DCBB_putative, there is mild support for a log-normal and higher support for a power law with exponential cut off. (Yule: Yule-Simon distribution)

NET_DCBB_known. Table 3.3 shows that, for both networks, two alternative models are more likely that a power law, the log-normal one and the power law with exponential cut-off. The latter hypothesis carries the highest weight, and shows that, while for a number of reasons the new putative information brought in through interologs seems to break a potential power law behaviour, all three networks are plausibly generated by a more specialised model, the power law with exponential cut-off, and this seems to account for the slightly thinner-than-power law tail in the distributions (Figure 3.3). Of course, there are infinite data-generating probabilities which I have not considered in this analysis and might better model these datasets, but the point remains: the putative data does not significantly upset the experimental protein network degree distribution.

Table 3.3 provides further insight which confirms some of the hypotheses discussed earlier in the section. The likelihood ratio analysis rules out the possibility that NET_DCBB_putative and NET_DCBB_union could have been generated by a Poisson model[12]. I also performed a likelihood ratio test of degree distribution data in NET_DCBB_putative_rand (the Erdős–Rényi model having the same number of nodes and edges in NET_DCBB_putative, introduced before). This is included as a control in Table 3.3. The power law/Poisson test agrees with the theory: Erdős–Rényi node degrees are Poisson-distributed by definition [Erdos and Renyi, 1960]. The fact that for NET_DCBB_putative and NET_DCBB_union a Poisson distribution is ruled out proves that the nodes in these network are organised in a non-random fashion.

3.3.3 Global Topological Parameters

Some of the results presented confirm that the data in NET_DCBB_putative follows the signatures observed in NET_DCBB_known more closely than does a random network of the same size. However, comparison with a completely random network is not entirely a fair test, as it might be argued that any random sample of protein interactions might show distributions of network properties compatible with *bona fide* protein interaction networks such as NET_DCBB_known. While, as stated before, this computational analysis cannot provide a definitive answer to the biological optimality of the putative predictions, the ROC-based validation provided in Chapter 2 partly addresses this showing that Bio::Homology::InterologWalk consistently retrieves more true positives than false positives, for any threshold set on the IPX. On top of that, an additional network experiment will now be introduced to investigate at least a partial answer to this issue.

Specifically, here I address the following question: are the topological signatures observed in NET_DCBB_putative significantly different from those that would be observed in any other "biological-looking" network of the same size? A test can be devised to statistically evaluate any differences between NET_DCBB_putative versus a large population of random networks

[12] And by an Exponential or a Weibull distribution.

of the same size which share *some, but not all* the topological signatures in `NET_DCBB_putative`. In order to create a random network able to resemble a typical biological network in some parameters, one can perform a *degree distribution-conserving* randomisation of `NET_DCBB_putative`. A degree distribution conserving network randomisation has equal number of nodes and edges as the original, and is obtained by shuffling the edges between the nodes, in such a way that the distribution of node degrees is unchanged. This is normally used as a random control in network analysis, and it outputs a more constrained random network for a set of N nodes and E edges compared to the Erdős–Rényi model. Because the degree distribution is unchanged, a degree-conserving randomisation can be considered closer to an originating biological network than a bare Erdős–Rényi random model. Hence, any differences between a test network and its degree-conserving randomisations are more relevant than the differences between the same test network and a random model of the same size.

To design this randomisation experiment, I devised a mixed Perl/C framework (which will be used and described thoroughly in Chapter 4) based on the igraph C network manipulation library [Csardi and Nepusz, 2006]. For this experiment, I chose a total of four global network parameters: network diameter, average path length (also called characteristic path length, or average geodesic length), global clustering coefficient (also known as global transitivity) and the average of the local clustering coefficients (local transitivities) for all nodes. I selected these indices because they provide a single value summary of interesting network properties, unlike the complex parameters shown in Figure 3.2, where the information is encoded by distribution signatures and is difficult to summarise and compare within a randomisation experiment. The only surviving complex parameter is the local clustering coefficient (Figure 3.2-C,D) which is now summarised in an average over all N nodes

$$\overline{C_L} = \frac{1}{N}\sum_{i=1}^{N} C_L(i) \tag{3.3}$$

as proposed by Watts and Strogatz [1998]. For the test, I stripped `NET_DCBB_putative` of all, if any, edges with multiplicity > 1, and computed 1000 randomisations. I then obtained the four topological indices for `NET_DCBB_putative` and for the 1000 sample networks. Results are shown in Table 3.4. To evaluate how the values for the four parameters would compare to the values obtained for the observed network, I carried out a Z-test and derived p-values. Full results are available in the Appendix, Table A.4, Page 192, and a graph summarising the result is in Figure 3.4. The results show that the difference between the actual network and its degree-conserving randomisations is significant ($p < 0.001$) for both the characteristic path length and the connectivity coefficients. The findings related to the connectivity coefficients are compatible with several reports describing higher clustering coefficients observed in real networks when compared to their randomisations [Ravasz et al., 2002] and can be interpreted as a signature of modular structure in biological networks: based on these results, `NET_DC-`

Network	D	CPL	C_G	$\overline{C_L}$
NET_DCBB_putative *	9	4.009462	0.015709	0.083513
DCR(NET_DCBB_putative *)-(μ,σ)	9.142,0.676974	3.705467,0.010824	0.013380, 0.000582	0.044560, 0.002735
Erdős–Rényi(NET_DCBB_putative *)	11	5.077331	0.001551	0.001336
NET_DCBB_known *	12	5.203123	0.004962	0.009684
Erdős–Rényi(NET_DCBB_known *)	18	7.334419	0.000780	0.000571

Table 3.4: Results for the NET_DCBB_putative randomisation experiments. The asterisk indicates that edges with multiplicity > 1, if any, have been collapsed onto one edge. For the 1000 degree conserving randomisations (DCR), mean and standard deviation are shown. Parameters values for NET_DCBB_known and for one Erdős–Rényi random network having the same number of nodes and edges as NET_DCBB_known are shown for reference.

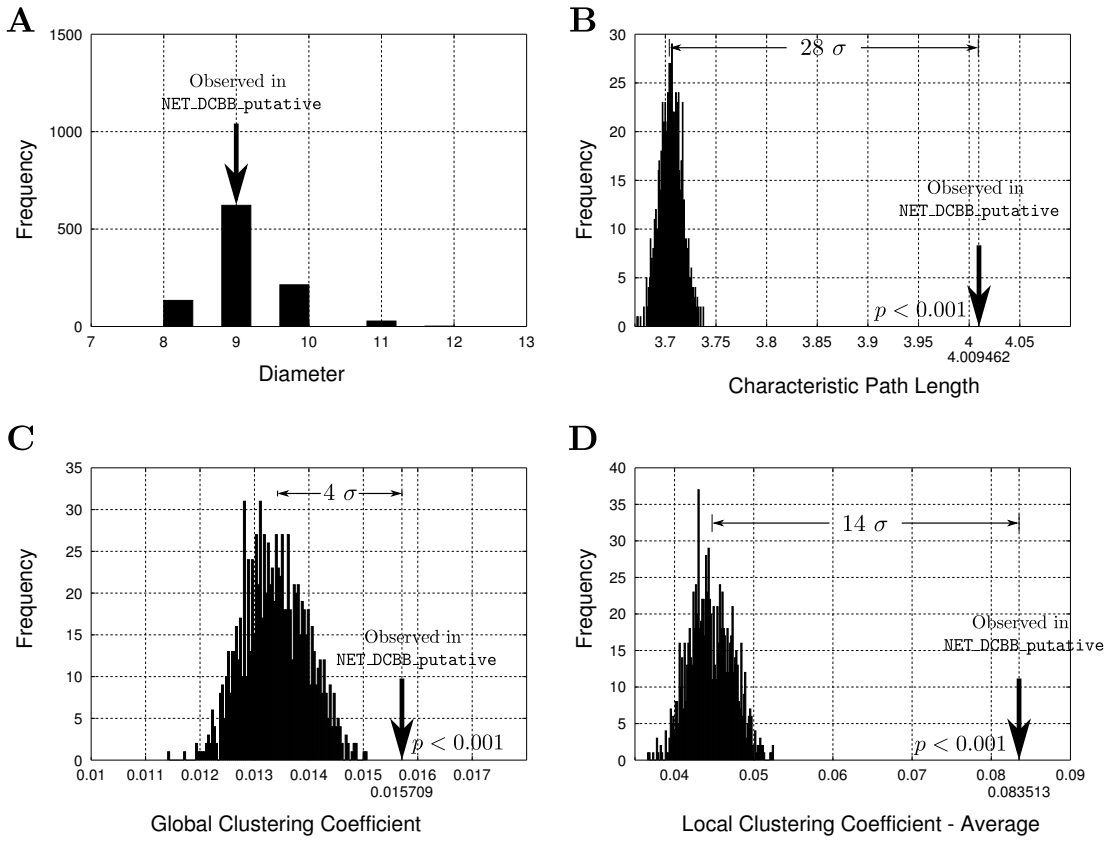

Figure 3.4: Randomisation test results for NET_DCBB_putative. All distributions represent data obtained from 1000 degree-distribution conserving randomisations of NET_DCBB_putative. **A**: Network Diameter. **B**: Average Path Length. **C**: Global Connectivity Coefficient. **D**: Local Connectivity Coefficient. The corresponding values observed in the actual NET_DCBB_putative are indicated by a down-pointing arrow. The null hypothesis 'observed value is a sample from the distribution' has been rejected at the 0.01% significance level in **B**, **C**, **D**.

BB_putative shows significantly higher modularity than any other network with the same number of nodes, edges, and biologically compatible degree distribution. Interestingly, Table 3.4 also shows that NET_DCBB_known features very low values for both the global and average local clustering coefficient: one reason for this is probably the incompleteness of NET_DCBB_known compared to NET_DCBB_putative: the latter is assembled from protein interaction data from six reference organisms (Figure 3.1-C). If less data is available, fewer nodes will be surrounded by well connected cliques and the average local coefficient will be smaller. For the same reasons, less complete triplets of nodes will occur and the global connectivity coefficient will be smaller.

The characteristic path length is significantly higher than the one observed in the degree-conserving randomisations. This data required more careful interpretation in light of the results described in the literature. Watts and Strogatz [1998] defined the concept of "small-world" network based on two properties:

$$C >> C_{\text{random}} \qquad \text{CPL} \gtrsim \text{CPL}_{\text{random}}, \tag{3.4}$$

where C is the clustering coefficient and CPL is the characteristic path length. According to this definition, a small-world network is characterised by 1) a large clustering coefficient C, indicating that each node is linked to a relatively well connected set of neighbouring nodes and 2) a relatively small characteristic path length CPL[13] indicating the presence of several short-cuts between far away communities in the network. While NET_DCBB_putative satisfies property 1, it does not seem to completely agree with property 2, because its characteristic path length CPL is significantly larger (28 standard deviations from the mean, Table A.4) than in the randomly rewired samples. This indicates that while NET_DCBB_putative has denser than random local communities, it has overall less long-distance node relationships compared to any of its randomised copies.

Therefore, based only the definition by Watts and Strogatz, one cannot conclude that NET_DCBB_putative has small-world network signatures. The reasons why NET_DCBB_putative has a lower than random number of far reaching edges and simultaneously high local cliqueness could be several. The most likely has to do with the nature of this interolog network: it is a putative network assembled using fly homologues of experimental interactions obtained from six unrelated reference interactomes. We could conjecture that, as experimental work is carried out for different reasons and in different sub-systems in different species, NET_DCBB_putative is a 'patchwork' of information that, while biologically correct at the local level (i.e. within communities of nodes projected from the same species) it lacks bridging edges *across* far away communities (i.e. few experimental nodes tying communities mapped from different reference interactomes exist).

[13] Comparable in size to that of a random network of similar dimensions.

However, it is also important to consider that Watts and Strogatz did not actually look at protein interaction data in their famous publication. Three empirical examples were provided: a collaboration graph of actors in feature films, an electrical power grid, and the neural network of *C. elegans*. The only biological network shown to feature the two famous signatures is the worm neural network, yet somehow the authors' small world hypothesis has been ubiquitously summoned to label compatible behaviour in a very diverse range of biological networks. Based on this observation, a more interesting explanation can be provided. This would support the initial hypothesis that `NET_DCBB_putative` features all the traits of an experimental protein interaction network and is based on new data and new evidence proposed by Zhang and Zhang [2009] and Xu et al. [2011]. Surprisingly, both teams specifically investigated experimental up-to-date interactomes for several reference genomes, and compared the obtained characteristic path lengths to those obtained from degree conserving randomisations. Xu et al. [2011] found that, for all species apart from *P. falciparum*[14] the characteristic path length is significantly larger than expected in the randomizations. Additionally, Zhang and Zhang [2009] reported that even a modest increase in characteristic path length can significantly favour modularity of the network, and argue that in this sense longer path lengths could represent a biological advantage in a protein interaction network scenario.

In summary, these results suggest that, based on theoretical network signatures, the original hypothesis holds true: the network `NET_DCBB_putative`, obtained through interolog mapping with `Bio::Homology::InterologWalk`, not only is more similar to the experimental `NET_DCBB_known` than a random Erdős–Rényi model of the same size; it is also more similar to it than any degree conserving randomisation. This represented topology-based evidence that interolog data is not random noise and prompted us to use interolog-based hypotheses to study sensory neurons in *Drosophila*.

3.4 A Ciliogenesis Protein Interaction Sub-network

Having discussed some of the theoretical and topological properties of `NET_DCBB_known` and `NET_DCBB_putative`, I next proceeded to look more closely at some of the actual information and protein interaction hypotheses found in these networks. This was done to evaluate the possibility of obtaining functional insight into the roles of novel proteins which might be implicated in ciliogenesis, and to see if the DCBB members would show any form of cluster organisations in the networks (suggesting or supporting hypotheses of their collaboration in some biological processes) or rather if they would appear in scattered, disconnected components. The latter possibility would render an interolog-based protein interaction analysis less informative.

[14] Incidentally, some of the problems with the data available for this interactome have been discussed in Chapter 2.

In order to clearly visualise the predictions produced by `Bio::Homology::InterologWalk`, I processed `NET_DCBB_known` and `NET_DCBB_union` using Cytoscape [Shannon et al., 2003]. Due to the size and complexity of the interaction networks considered (Table 3.1, Page 91) I decided to restrict the analysis to subsets of the interacting nodes. Similar to the approach I used in Chapter 2, the selection of nodes to analyse was informed by functional annotation provided by the Gene Ontology project [Ashburner et al., 2000]. Specifically, I decided to restrict the analysis to the subset of the 815 DCBB seed genes in `NET_DCBB_known` annotated with the biological processes `cilium assembly` and `cilium morphogenesis` in the Gene Ontology. This was done to obtain a seed gene set which would be smaller than the full 815 genes for clarity purposes, and which at the same time would include genes validated by the largest functional annotation project available. I downloaded the *Drosophila* annotation (25/09/2010) from Flybase[15] and the most recent gene ontology from the GO website[16]. I obtained 17 hits[17]. Provenance information for these 17 genes is provided in the top part of Table 3.5. I then retrieved all the nearest neighbours of these seed genes found in `NET_DCBB_known`. This resulted in a collection of 12 disconnected small sub-networks (Figure 3.5). Figure 3.5 shows that, according to current protein interaction data, 17 GO-annotated ciliogenesis genes (black nodes) possess physical interactors (grey nodes). The 17 genes do not interact with each other, and are distributed in isolated complexes, the biggest of which features 4 ciliogenesis genes.

In order to verify if the data added to `NET_DCBB_known` through the computational pipeline introduced in Chapter 2 was able to produce interesting protein interaction clusters, I repeated the same procedure, this time using `NET_DCBB_union`. Again, I selected genes annotated with the `cilium assembly` and `cilium morphogenesis` GO biological process. This yielded a set of 23 hits, a superset of the 17 found before — meaning that 6 additional ciliogenesis genes participate exclusively in putative interactions (Table 3.5, bottom). These 23 hits (again, a subset of the 815 genes in the DCBB dataset) shall be known henceforth as the *seed* ciliogenesis gene set. As before, I selected the sub-network of `NET_DCBB_union` composed by the 23 seed genes and their nearest neighbours. The ensuing sub-network, `NET_cilium`, composed of 193 nodes and 224 edges, is shown in Figure 3.6.

While Figure 3.5 shows a collection of isolated network motifs where only few of the ciliogenesis genes are connected together through experimental protein interactions, Figure 3.6 relates ciliogenesis genes to each other in a large connected complex. The main connected component in `NET_cilium` comprises 165 genes and 202 interactions, and ties together 16 (about 70%) of the 23 seed genes. The existence of large clusters connecting several DCBB genes suggests that including putative interactions can allow guilt by association-based func-

[15] `ftp.flybase.net`

[16] `www.geneontology.org`

[17] These were, as expected, a subset of the 815 DCBB genes, meaning that no genes annotated in GO for `cilium` were missing from the list in Laurencon et al.

ID	GO Evidence	Source
	NET_DCBB_known and NET_DCBB_union:	
CG1399	inferred from sequence model	[Avidor-Reiss et al., 2004]
CG17599	inferred from sequence model	[Avidor-Reiss et al., 2004]
Oseg1	inferred from sequence model	[Avidor-Reiss et al., 2004]
Oseg4	inferred from sequence model	[Avidor-Reiss et al., 2004]
CG15161	inferred from sequence model	[Avidor-Reiss et al., 2004]
asl	inferred from mutant phenotype	[Blachon et al., 2008]
CG3259	inferred from expression pattern	[Avidor-Reiss et al., 2004]
CG11048 (Efhc1.2)	inferred from sequence model	[Avidor-Reiss et al., 2004]
CG14870	inferred from expression pattern	[Avidor-Reiss et al., 2004]
CG14367	inferred from sequence model	[Avidor-Reiss et al., 2004]
Klp64D	inferred from mutant phenotype	[Sarpal et al., 2003; Jana et al., 2011]
CG1126	inferred from expression pattern	[Avidor-Reiss et al., 2004]
Sas-4	inferred from mutant phenotype	[Basto et al., 2006]
CG8853 (Hippi)	inferred from sequence model	[Avidor-Reiss et al., 2004]
nompB	inferred from sequence model	[Avidor-Reiss et al., 2004]
osm-6	inferred from sequence model	[Avidor-Reiss et al., 2004]
Kap3	inferred from mutant phenotype	[Sarpal et al., 2003; Jana et al., 2011]
	Only NET_DCBB_union:	
BBS4	inferred from expression pattern	[Avidor-Reiss et al., 2004]
BBS8	inferred from expression pattern	[Avidor-Reiss et al., 2004]
CG30441	inferred from sequence model	[Avidor-Reiss et al., 2004]
CG7735	inferred from expression pattern	[Avidor-Reiss et al., 2004]
dnd	inferred from sequence model	[Avidor-Reiss et al., 2004]
rempA	inferred from expression pattern	[Avidor-Reiss et al., 2004]

Table 3.5: Provenance data for the genes annotated in GO for Cilium Assembly and/or Cilium Morphogenesis BP terms which appear in NET_DCBB_known and NET_DCBB_union.

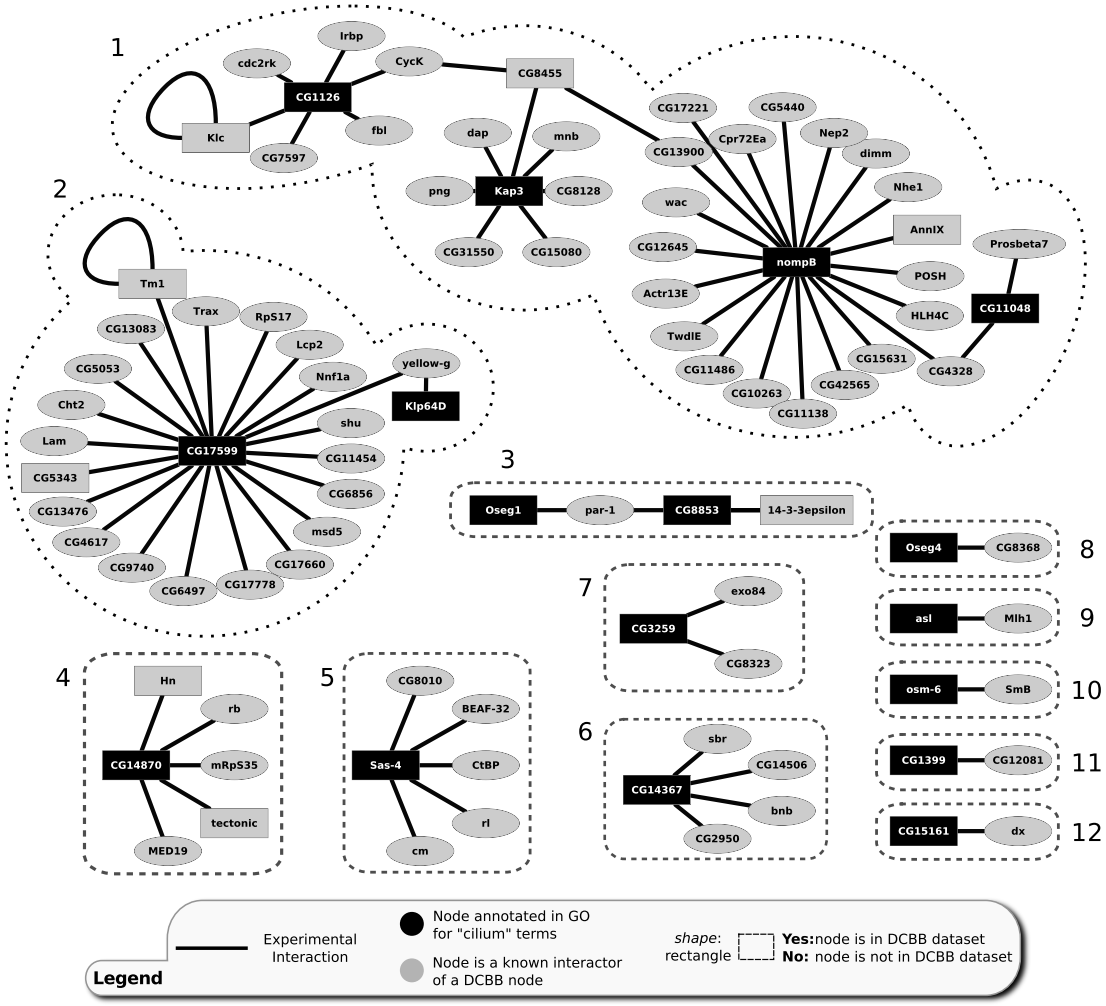

Figure 3.5: Data extracted from `NET_DCBB_known` as follows: a) select all genes annotated with GO biological processes `cilium assembly` and `cilium morphogenesis` (17 genes, black nodes) b) select all their nearest neighbours (78 genes, grey nodes). Black connections are experimental protein interaction data from EBI IntAct. Ciliogenesis GO-annotated genes never interact with each other. 12 disconnected components are observed, the biggest of which (1) connects 4 seed genes.

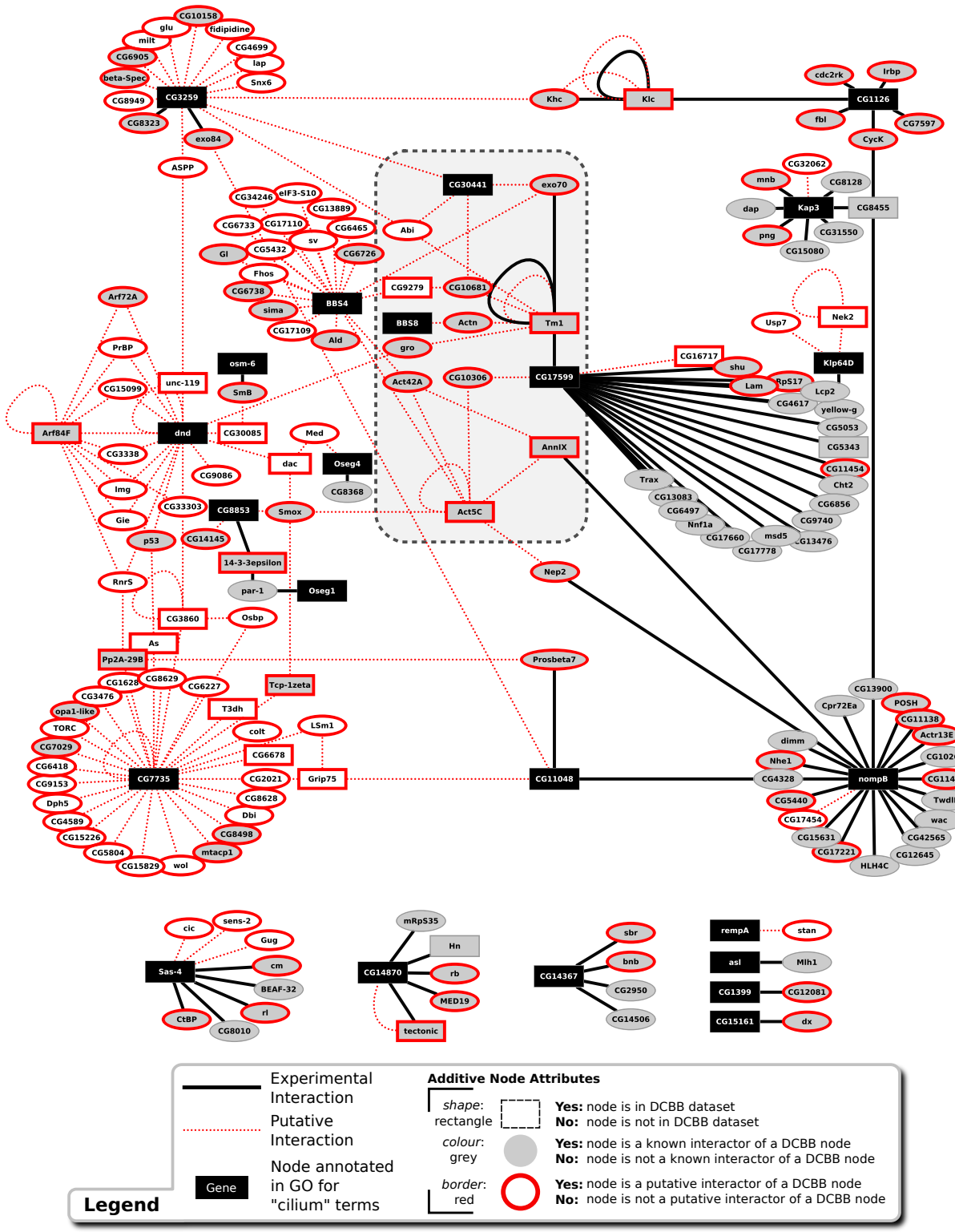

Figure 3.6: `NET_cilium`. Data extracted from `NET_DCBB_union` as follows: a) select all genes annotated with GO biological processes `cilium assembly` and `cilium morphogenesis` (23 genes, black nodes) b) select all their nearest neighbours (170 genes). Solid connections are experimental protein interactions (from EBI IntAct) belonging to `NET_DCBB_known`, while red dotted lines are putative predictions from `NET_DCBB_putative`. Nodes are described in key. The network is roughly divided into an experimental region (right), representing experimental interaction evidence on the domain, connecting to a putative region (left), representing putative interaction evidence. The shaded area (centre) highlights a highly connected complex of genes that can be thought of as communication hub between the two sections of the network.

tional inferences which would not be possible lacking large connected components based on ciliogenesis genes. Thus, including putative interactions appears to be useful to define a ciliogenesis network.

A broad analysis of `NET_cilium` reveals that several new genes, not formerly known to be related to ciliogenesis, are retrieved through putative protein interactions. Additionally, some of the DCBB ciliogenesis genes initially left out (due to absent GO annotation, which excludes their presence in the 23 genes seed set) are now part of the network (Figure 3.6, *white rectangles*). Finally, the putative interactions wire most of the seed ciliogenesis genes in a large connected sub-network. Genes that were known to be involved in ciliogenesis now interact with genes for which no evidence for ciliogenesis involvement existed — meaning new potential candidates for ciliogenesis activity are drawn in to build a more complete picture of the domain. Moreover, relationships between pairs of genes where each member was known to be involved in ciliogenesis, but for which there was no experimental interaction evidence indicate that prior weak evidence[18] is being reconfirmed through this independent method of inquiry. The fact that DCBB genes lacking GO annotations appear in the network by purely topological arguments led us to argue that the method is reconfirming functional labelling obtained with one-off methods by Laurencon et al., and is also adding new hypotheses to test.

A number of experimental interactions are reconfirmed via interologs: for instance, *Kinesin heavy chain* and *Kinesin light chain*, *tectonic* and *CG14870* and the auto-interaction of *Tropomyosin 1*. About 28% of the nodes in `NET_cilium` are both experimental and putative interactors of `DS_DCBB` genes. This indicates nodes whose affiliation to the network has been obtained experimentally and reconfirmed via interologs.

`NET_cilium`, obtained by a purely structural extension of the GO annotated ciliogenesis genes, wires most of these together in an interacting complex. This is because several of the new genes brought in by the putative pipeline (a) interact with more than one seed ciliogenesis gene and (b) interact with each other. An illustrative example of this is observed in the sub-network around *BBS4*, *BBS8*, *CG17599*, *CG3259* and *CG30441*. These known ciliogenesis genes have not been shown to interact mutually before (according to current EBI IntAct annotation). However, `Bio::Homology::InterologWalk` extracts, for each of them, sets of putative interactions that overlap very well with one another, thus building a completely novel sub-complex of participating units worth investigating further. A number of cases can be observed. Firstly, *Abelson interacting protein* (*Abi*) appears in this network purely through computational arguments. Secondly, *CG10681*, α-*actinin* and *groucho* have been brought in `NET_cilium` by putative interactions with the seed DCBB genes — although they also participate in experimental interactions within `NET_DCBB_union`. Lastly, *CG9279* and *Tropomyosin*

[18] Most of the 23 seed genes had only been associated to ciliogenesis through computational comparison in [Avidor-Reiss et al., 2004] (Table 3.5).

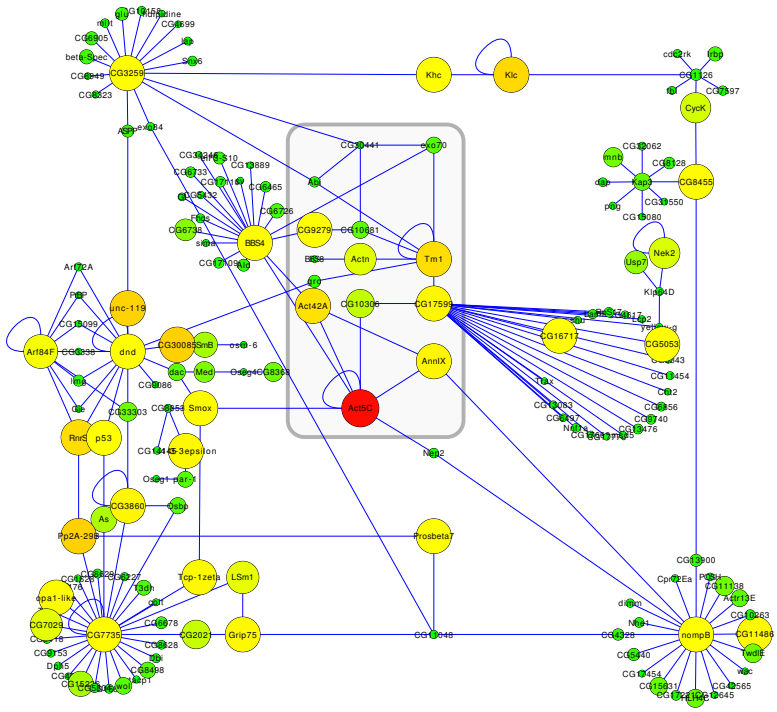

Figure 3.7: Betweenness Centrality-mapped main connected component in NET_cilium. Alternative visualisation for the main sub-network in Figure 3.6. For each node, size and colour have been mapped to the node's betweenness centrality in the original network, NET_DCBB_union. Bigger node diameter and darker colour correspond to higher betweenness centrality values.

1, two members of DCBB, are captured in putative interactions by the module, but are not part of the GO cilium-related seed set. *Tropomyosin 1* is, moreover, a hub for the experimentally obtained portion of NET_cilium (Figure 3.6, *solid connections*).

A number of ciliogenesis-annotated genes connected together by putative genes intuitively seem to act as information flow gateways between more peripheral clusters of genes (Figure 3.6, *inset*). In order to look at the meaningfulness of the increased connectivity introduced by the putative elements, and to have a topology-based measure of the relative importance of the nodes in NET_cilium from an information-flow perspective, I went back to NET_DCBB_union, collapsed all edges with multiplicity > 1 into one and computed the betweenness centrality of all nodes. Then, I selected the main connected component of NET_cilium (165 genes) and modified the visualisation in Figure 3.6 by mapping, for all nodes, their size and colour to their betweenness centrality index (Figure 3.7). Data for the 15 nodes with the highest betweenness centrality values in the main connected component of NET_cilium are shown in Table 3.6. Interestingly, of the 15 highest centrality nodes in NET_cilium 5 belong to the large connecting area evidenced earlier (Figure 3.6 and 3.7, *inset*). This suggests that while many known ciliogenesis genes are, as expected, of importance in terms of communication flow within the network, some putative genes retrieved by the algorithm appear — in terms of this

Gene	B_C	C	D	N_C
Act5C*	0.0662	0.0127	219	15.4305
CG30085	0.0139	0.0053	83	7.2651
Pp2A-29B	0.0138	0.0182	85	17.6988
unc-119	0.0132	0.0078	72	8.2778
Klc	0.0115	6.734E-4	62	3.5454
Tm1*	0.0103	0.0204	55	22.3921
RnrS	0.0103	0.0327	38	36.2631
Act42A*	0.0093	0.0924	35	70.2000
AnnIX*	0.0040	0.0461	26	29.4615
CG3860	0.0037	0.0353	37	13.4286
CG17599*	0.0037	0.0	23	5.1739
14-3-3epsilon	0.0032	0.0085	28	20.3333
Grip75	0.0032	0.0	24	9.1667
CG11486	0.0031	0.0444	11	57.7000
Tcp-1zeta	0.0031	0.0417	16	45.1875

Table 3.6: Node to C_B map for the 15 nodes with the highest C_B values in the main connected component of `NET_cilium`. Nodes marked by an * are situated in the central network component observed in Figures 3.6 and 3.7. Nodes in bold face are part of the GO annotated 23 genes seed list.

centrality analysis — as important as the seed DCCB genes, and many are connected with one another in a tight group of high centrality nodes situated in the central areal of `NET_cilium`.

One last observation pertains to the 23 seed DCBB genes, which have varying amounts of evidence implicating them in ciliogenesis-related processes within the Gene Ontology (Table 3.5). Most of them are functionally associated with cilium processes only as a result of comparative studies[19]. `Bio::Homology::InterologWalk` reinforces the available evidence by placing the 23 genes in a tightly-connected complex of functionally related proteins.

3.4.1 Pruning the Network via IPX thresholding

`Bio::Homology::InterologWalk` returned 3695 genes in total for DCBB — representing roughly 24% of the *D. melanogaster* genome (14869 genes, source: Ensembl, V. 59). This indicates that the settings used were not specific enough. The following step in the analysis was to attempt to refine the candidate list within the dataset through IPX thresholding, in order to find highly supported 'backbone' sub-networks of `NET_cilium` by pruning nodes connected through putative interactions appearing with poor biological support.

Figure 3.8-A shows the IPX distribution for `NET_DCBB_putative`. As expected from the methodology discussion in Chapter 2, in both cases a tri-modal distribution appears because the binary/spoke index and the orthology class index for both the forward and backward steps (summarised by the reward/penalisation terms Σ and Θ, equation 2.4, Page 2.4) are not nor-

[19] Like the aforementioned one by Avidor-Reiss et al. [2004], a genomics screen comparing the genomes of ciliated and non-ciliated organisms for functional inference.

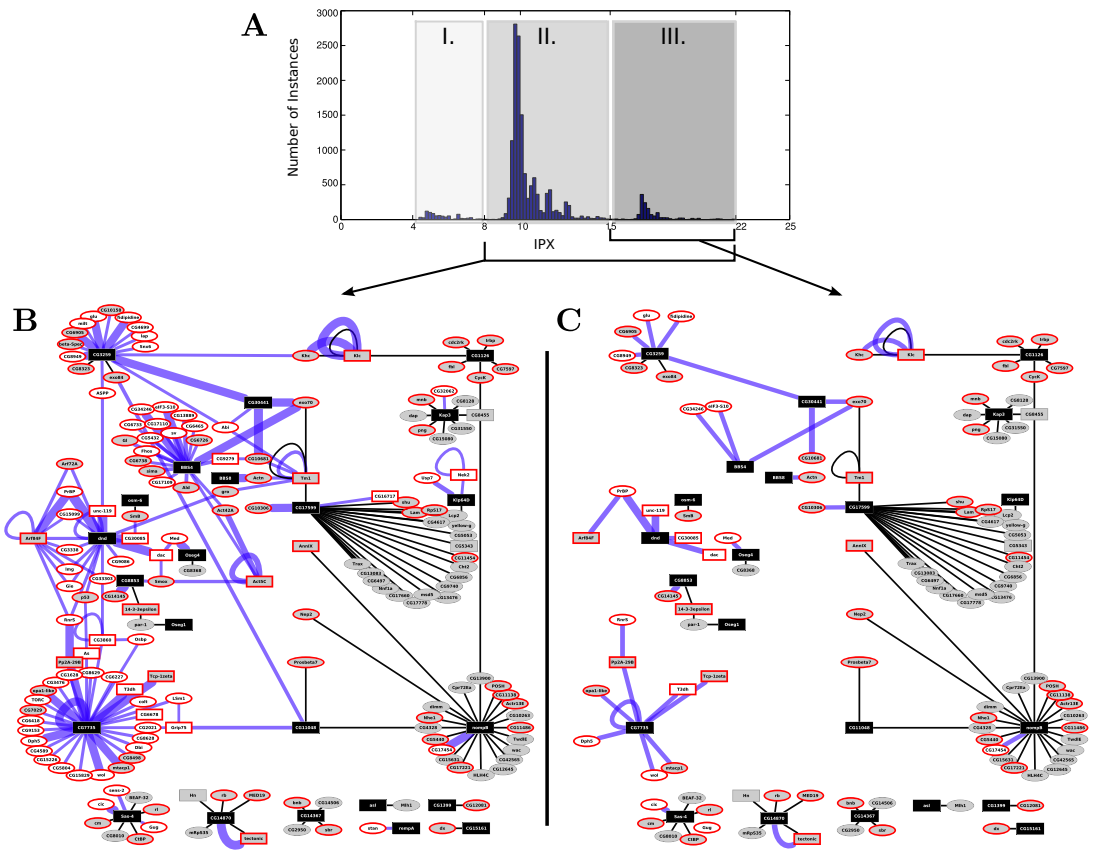

Figure 3.8: Effect of varying IPX cut-off levels on putative protein interaction network. **A**: Global confidence score distribution for `NET_DCBB_putative`. The three quadrants highlight three modes in the distribution. **I**: putative interactions labelled with a score in this quadrant are projections of proteins belonging to low scoring spoke-expanded complexes. Moreover, at least one member of the two orthologous pairs (forward and backward) has in-paralogues (i.e., at least one of the two orthologous pairs is a 1:many or many:many orthology). **II**: putative interactions labelled with a score in this quadrant are projections of proteins in complexes or have been obtained through not optimal orthologues (again, 1:many, many:many). **III**: members of this high-scoring sector are projections of binary experimental interactions mapped strictly through 1:1 orthology relationships. **B**: network in Figure 3.6 when the score cut-off is after the 1st quadrant **C**: network in Figure 3.6 when the score cut-off is after the first two quadrants.

malised. Depending on the combinations of values for these two parameters, the IPX values in Figure 3.8-A are divided in three groups: (I) the experimental interaction is spoke-expanded and at least one of the two orthology projections is *not* 1:1. (II) either the experimental interaction is spoke-expanded or at least one of the two orthology projections is *not* 1:1 (III) the experimental interaction is not expanded from a spoke complex and the orthology projections are *both* 1:1.

In order to look at the composition of putative interactions in `NET_cilium`, I set two IPX thresholds. The first, IPX_{thr_1}, discards the data distributed around the left-most mode; the second, IPX_{thr_2}, discards the data distributed around the first two modes. Setting $IPX_{thr_1} = 8$, $IPX_{thr_2} = 15$ and mapping IPX values to edge thickness in `NET_cilium`, I obtain the graphs in Figures 3.8-B and 3.8-C. Figure 3.8-B shows that the connectedness of `NET_cilium` is roughly preserved when the lowest-support set of putative interactions is filtered out of the network. Figure 3.8-C shows the putative interactions characterised by the highest biological support according to the IPX. While the connectedness of the main component of the network is broken at this threshold level, we notice that the network motif discussed earlier (Figure 3.6, *inset*) survives around a subset of essential genes.

The interactions remaining after applying the stringent IPX_{thr_2} cut-off have the highest biological support and, I argue, are good candidates for validation experiments.

3.4.2 Using the network to prime the role of CG17599, CG30441 and CG31320 in Ciliogenesis

Figure 3.8-C shows a sub-network of `NET_cilium` containing either experimental protein interactions between the genes shown (black solid edges) or putative interactions obtained exclusively from fully 1:1 orthology projections of experimental binary interactions in the reference organisms (purple edges). Although the connected component of `NET_cilium` observed in Figure 3.6 has largely disappeared, most of the genes in the central hub highlighted in the shaded area of Figure 3.6 have survived the strictest threshold cut.

I used this central hub to guide a gene prediction experiment. In order to obtain a very small putative network involving high confidence interactors, which might include potentially interesting candidate ciliogenesis genes, I went back to the original network, `NET_DCBB_union`. Again, I selected the 23 seed genes annotated for `cilium` in Gene Ontology. This time, a number of additional genes lacking cilia-related GO annotation were also considered for membership to the seed list. These were mainly genes showing interesting expression enrichment profiles in *ato*-expressing cells, according to the transcriptional data in [Cachero et al., 2011]. One of these genes, CG31320, stood out for its participation to two putative interactions in `NET_DCBB_union`, both of them with members of the high confidence sub-network in Figure 3.8-C. I manually added this gene to the 23-seed list. I then selected their first neighbours from

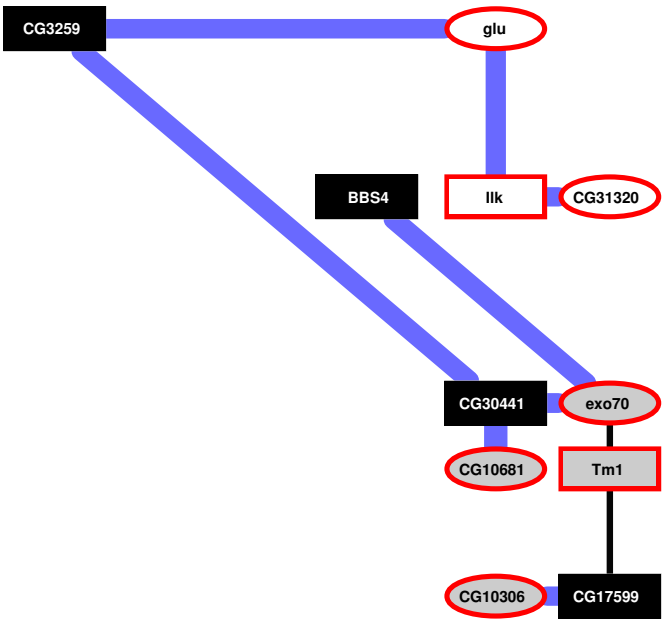

Figure 3.9: Thresholded connected component. This has been extended from 3.8-C by adding CG31320 and its first neighbours from NET_DCBB_union. Black edges are experimental interactions, while blue edges are putative interactions. For the blue edges, thickness is proportional to the corresponding IPX value. All putative interactions in this sub-network pass the strict threshold $IPX_{thr_2} > 15$ described in subsection 3.4.1. Node colour-code legend is as in Figure 3.6.

NET_DCBB_union and employed a stringent threshold cut-off to prune less supported putative interactions. Figure 3.9 reproduces the resulting connected component. I decided to study this sub-network in some detail.

Provenance data for the interactions in Figure 3.9 are provided in Table 3.7 (relative to the experimental interactions) and Table 3.8 (relative to the putative interactions). A number of functional annotation cues link some of the genes in the sub-network to ciliary development processes. BBS4 is the only fly homologue of the human *Bardet-Biedl syndrome 4 protein*, one of the members of the BBSome complex [Nachury et al., 2007]. The BBSome complex is known to have a role in ciliogenesis: it associates with the ciliary membrane and binds to RAB3IP/Rabin8, the guanosyl exchange factor (GEF) for Rab8. Rab8-GTPase localizes to the cilium and promotes docking and fusion of carrier vesicles to the base of the ciliary membrane. Defects in BBS4 are the cause of Bardet-Biedl syndrome type 4 [Kim et al., 2004]. Evidence for CG3259 association with ciliary assembly comes from one of the screen performed by Avidor-Reiss et al. [2004], who detected CG3259 in their list of novel ciliary compartment genes. No Expressed Sequence Tags were found for the gene, however an X-Box binding sequence is present in the upstream sequence of the gene in both *Drosophila* and *C. elegans*. Additionally, the gene was biologically validated and found to be selectively expressed in ciliated sensory neurons. Ilk (Integrin linked kinase) is the homologue of a human intracellular serine/threonin protein kinase that mediates the integrin signalling in diverse types of cells [Wu and Dedhar, 2001; Hannigan et al., 2005]. Dysregulation of its expression has been implicated in the pathogenesis of a wide variety of chronic kidney diseases, including nephrotic syndrome and diabetic and obstructive nephropathy [de Paulo Castro Teixeira et al., 2005].

Somehow less relevant information is available for Tm1 (Tropomyosin-1) and for exo70

Interaction	EBI_ID	Det.Method	Publication
Tm1–exo70	EBI-507515	two hybrid fragment pooling approach	[Formstecher et al., 2005]
CG17599–Tm1	EBI-251087	Y2H	[Giot et al., 2003]

Table 3.7: Provenance data for the experimental protein interactions in Figure 3.9. For all entries, the annotated interaction type corresponds to the PSI-MI code MI:0915 (physical association).

Interaction	Ref.Interaction	Ref.Species	Det.Method	Publication	IPX
CG3259–CG30441	TRAF3IP1–IFT20	Hsap	Y2H	[Camargo et al., 2006]	16.83
CG3259–glu	TRAF3IP1–SMC4	Hsap	Y2H	[Camargo et al., 2006]	16.83
glu–Ilk	SMC4–ILK	Hsap	anti bait coip	[Ewing et al., 2007]	17.04
Ilk–CG31320	ILK–HEATR2	Hsap	anti bait coip	[Ewing et al., 2007]	16.78
CG30441–exo70	IFT20–EXOC7	Hsap	two hybrid pooling approach	[Rual et al., 2005]	16.91
CG30441–CG10681	IFT20–KXD1(C19orf50)	Hsap	two hybrid pooling approach	[Rual et al., 2005]	17.87
BBS4–exo70	BBS4–EXOC7	Hsap	Y2H	[Oeffner et al., 2008]	16.9
CG17599–CG10306	dyf-3 eif-3.K	Cele	two hybrid pooling approach	[Li et al., 2004b]	16.77

Table 3.8: Provenance data for the putative protein interactions in Figure 3.9. The fields *Ref.Interaction*, *Ref.Species*, *Det.Method* and *Publication* are relative to the experimental interaction in the reference genome used by `Bio::Homology::InterologWalk` to infer the putative interaction in *Interaction*. For all entries, the annotated interaction type corresponds to the PSI-MI code MI:0915 (physical association).

(Exocyst complex component 7). Tropomyosin, in association with the troponin complex, plays a central role in the calcium dependent regulation of muscle contraction, while exo70 is part of the exocyst complex, an octameric protein complex involved in vesicle trafficking implicated in a number of cell processes, including exocytosis and also cell migration and growth. CG10681 is a homologue of the human KXD (KxDL motif-containing protein 1) and of C13F10.2 in worm. The latter is required during embryo development [Sonnichsen et al., 2005]. For these genes however, there are no previous indications of roles in ciliogenesis.

CG17599 and CG30441 are interesting from several points of view. CG17599 is the fly homologue of the human gene CLUAP1 (Clusterin-associated protein 1) and of dyf-3 in *C. elegans*. The latter has been associated with sensory cilium formation and specifically intraflagellar transport [Murayama et al., 2005]. CG30441 is an IFT20 (Intraflagellar transport protein 20) homologue. This is a component of the IFT complex B, known to be involved in ciliary process assembly (Figure 1.3-B, Page 10). Specifically, the protein is believed to play a role in the trafficking of ciliary membrane proteins from the Golgi complex to the cilium [Follit et al., 2006]. It is also known that IFT20 acts as an adapter between the IFT complex B and Kinesin II through its interactions with IFT57 and KIF3B via coiled-coil domains [Baker et al., 2003] and is thus indispensable for anterograde IFT and cilium assembly. Very little is known about the last gene, CG31320 — a homologue of the *HEAT repeat containing 2* protein. HEAT domain containing proteins are hypothesized to function as scaffolds for protein-protein interaction

Flybase ID	Affymetrix ID	Ensembl Hits
FBgn0050441	1630985_at	CG10395 (7), CG30441 (14)
FBgn0050441	1629101_a_at	CG10395 (14), CG30441 (8)

Table 3.9: CG30441: affymetrix probe ambiguity

surfaces.

For two of these three genes, gene expression data was available from the transcriptome profiling of *ato*-expressing sensory neurons by Cachero et al. [2011]. I also accessed an additional dataset of gene expression in cells positive for expression of another proneural factor, *cousin of Atonal* (cato) [Cachero *et al.*, unpublished data]. This dataset is closely related to the *ato* one, however *cato* is a much more specific marker: the *cato*-GFP protein is exclusively expressed in the neuronal lineage (as opposed to *ato*-GFP, which is expressed both in the neuronal lineage and in support cells). Additionally, the onset of *cato* expression happens later in development compared to *ato*, allowing the isolation of a cleaner list of candidates to understand processes related to neuronal differentiation. Transcriptional data in *cato*-expressing cells will represent a central point in Chapter 4, therefore a longer discussion of these concepts will be presented there.

Figure 3.10 shows, for both CG17599 and CG31320, the fold change of transcript quantity relative to non *ato*-expressing cells (3.10-A) and non *cato*-expressing cells (3.10-B). For the *ato* experiments, data was available for developmental time points T_1, T_2 and T_3 (roughly corresponding to the first 3 hours of neural development) while due to the later onset of *cato* activity the *cato* experiments are offset by one hour and correspond to developmental time points T_2, T_3 and T_4. In *ato*-expressing cells, CG17599 and CG31320 show no appreciable difference in expression from the baseline during the first two time points. However, they then come on around the third hour after formation of the chordotonal neuron precursor cells, corresponding with the onset time for genes related with differentiation. Relative transcript quantities for *cato*-expressing cells show the genes already active in Cato cells at T_2, with a 7-fold increase in expression with respect to the baseline around the 4th hour after formation of the Ch neuron precursor cells. As regards CG30441, the available expression data cannot be considered informative. This is due to unavailability of affymetrix probes matching exclusive CG30441 oligonucleotides. There is no 1:1 mapping between the Flybase ID for CG30441 (FBgn0050441) and any of the probe IDs to which expression data is associated: the two available probes are detailed in Table 3.9. The probes both map to CG30441 as well as another protein coding gene, CG10395. It cannot be excluded that for both probes the signal obtained is a combination of the two mRNA signals. Expression data for CG30441 (Figure 3.10-D and 3.10-E) cannot be reliably utilized to support or negate the hypothesis of increased transcript

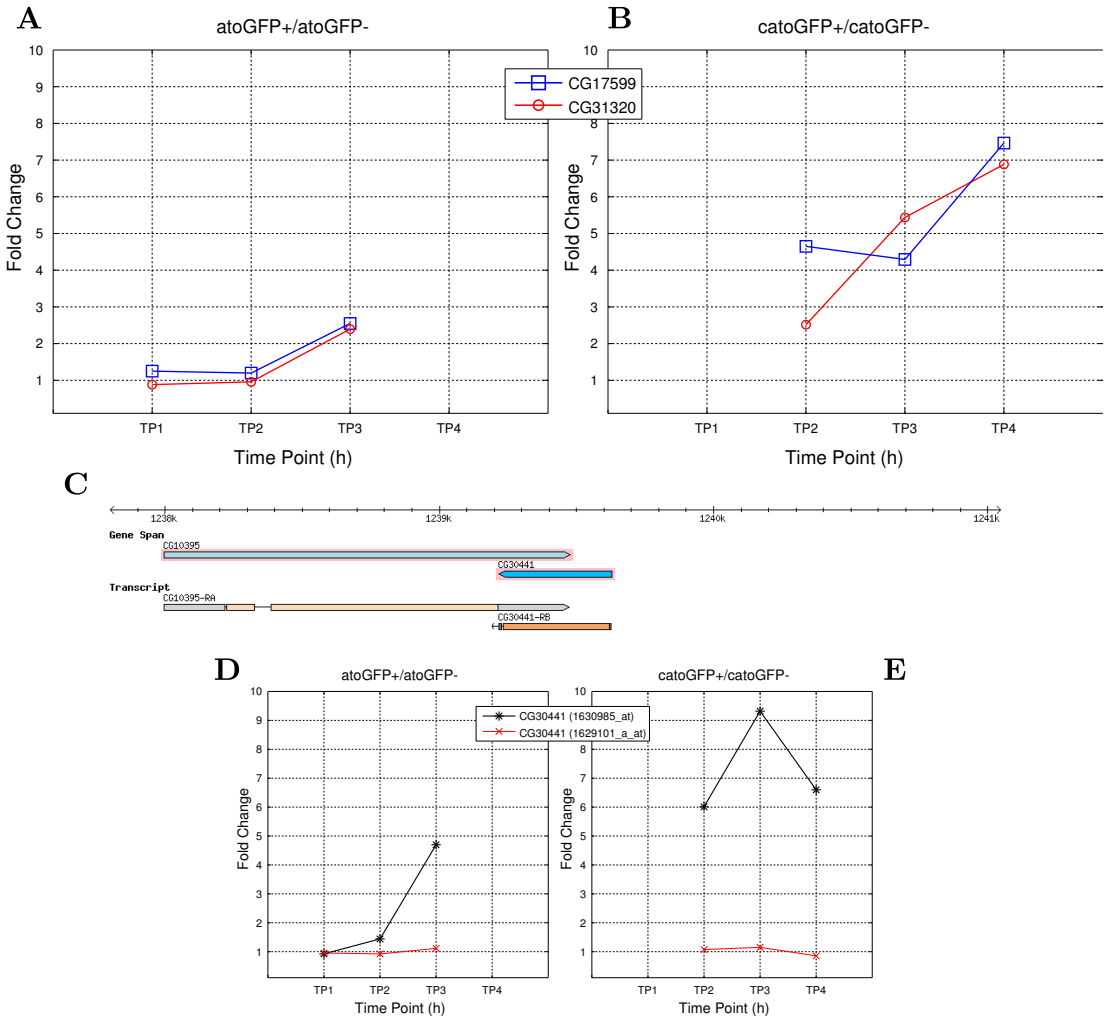

Figure 3.10: Temporal expression data in early *Drosophila* embryos for the genes CG31320, CG17599 and CG30441. **A**: Fold change of transcript quantity in cells positive for expression of the proneural factor Atonal (atoGFP+) versus cells negative for expression of Atonal (atoGFP-) [Cachero et al., 2011]. **B**: Fold change of transcript quantity in cells positive for expression of the proneural factor *Cousin of Atonal* (catoGFP+) versus cells negative for expression of Cato (catoGFP-). **C**: Genomic sequence and transcript overlap for genes CG30441 and CG10395. **D,E**: Fold change of transcript quantity for gene CG30441. Data shown is relative to both oligos mapping to the gene, `1630985_at` and `1629101_a_at`. For either or both oligos, the expression signal is potentially a mixture of the two mRNAs and cannot be considered reliable.

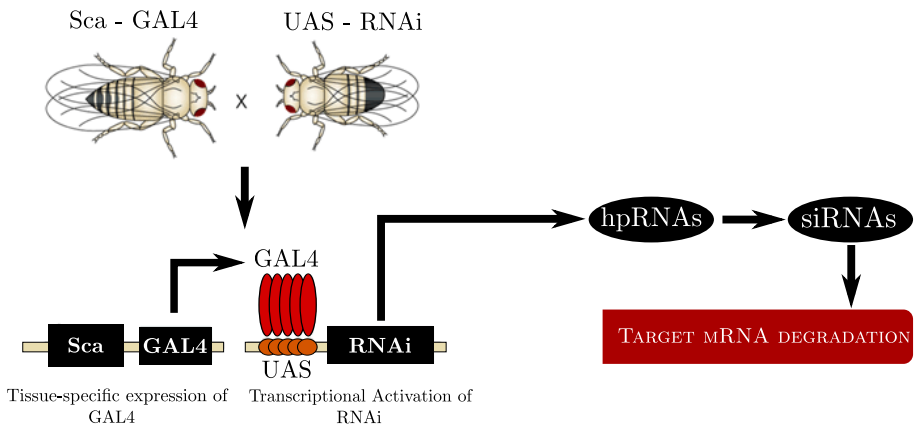

Figure 3.11: Schematics of the Sca-GAL4/UAS system utilised. To investigate the functional significance of the CG17599, CG31320 and CG30441 expression, flies with an RNAi gene for the three genes downstream of a yeast upstream activating sequence (UAS) were crossed to flies expressing the GAL4 protein under the control of a neural-specific gene promoter, *scabrous* (expressed in the embryonic developmental stages 6 to 16). *scabrous* drives neuronal-specific expression of GAL4, which in turn acts on UAS and drives expression of the RNAi genes. The output of the system is hairpin RNAs (hpRNAs), which are processed by the cellular machinery into small interfering RNAs (siRNAs). The latter bind to their target mRNA leading to their endonucleolytic degradation, causing gene silencing. (Image adapted from Muqit and Feany [2002])

abundance towards the late stages of neural development.

3.4.3 Experimental Validation

The connected sub-network described and the amount of previous evidence from independent studies prompted selection of CG30441, CG17599 and CG31320 for elucidation of functional significance in a series of wet-lab experiments. This was the work of Girish Mali in the Jarman lab; in this section, I will give a brief account of his main findings.

The study aimed, firstly, at investigating the localisation of the expression of the three candidates, to verify whether they are in the sensory neurons as expected from the transcriptome profiling data. RNA *in situ* hybridisation using riboprobes for CG31320 and CG30441 labelled with Digoxygenin and LacZ antibody staining for CG17599 was performed. For both CG31320 and CG17599, chordotonal neuron-specific expression was observed. For CG30441, no suitable pattern was detected (data not shown), and further attempts to make a suitable riboprobe using another cloning-based protocol also failed. Next, the GAL4/UAS system [Brand and Perrimon, 1993] combined with RNAi knock-down was used to create three separate mutant fly lines, each showing reduced expression for one of the three genes. Figure 3.11 provides a brief high-level explanation of the GAL4/UAS system employed. Having obtained suitable RNAi knock-downs of CG31320, CG17599 and CG30441 (Figure 3.12 show the effect mRNA reduction for CG31320), a series of locomotion assays were performed (Figure 3.13) to test

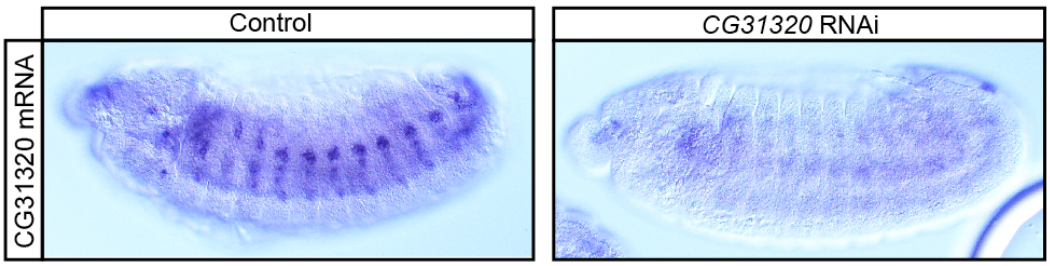

Figure 3.12: RNA *in-situ* hybridisation for CG31320, showing strong mRNA reduction in the knock down.

if the mutant flies would show significant lack of coordination and/or locomotion defects. Indeed, in all three cases reduced scores in the climbing assay suggested impaired function of Ch neurons, suggesting that the genes are required for some aspect of Ch neuron structural or functional task. The most abnormal behavioural phenotype was observed in CG30441 lines, where almost no flies managed to cross the 50ml threshold within 20 seconds ($p = 4.1E-10$). Other behavioural characteristics were also observed, including inability to fly, high propensity to fall whilst climbing, increase in time spent being stationary.

To see if their requirement in Ch neurons might be related to the construction of the ciliary dendrite, a further series of experiments was performed. These evaluated the impact of the expression knock-down on ciliary morphology. Immunohistochemistry was performed on the RNAi lines and on a control line to assess the structure of the chordotonal organs and of the modified primary cilia within (Figure 3.14-A,B). Cilia in embryos from the CG31320 and CG17599 RNAi lines showed marked outer dendritic segment defects compared to control cilia (figure 3.14-C,D). Lack of ciliary dilation and loss of thickness were determined. CG30441 mutant line embryos showed dramatic alterations in ciliary morphology (Figure 3.14-E). A further set of immunohistochemistry experiments were performed to investigate whether changes in ciliary morphology at the embryo stage also manifested as alterations in morphology at the larval stage. Other experiments in course of completion include the experimental validation of some of the protein interactions included in the cluster under study. Overall, these observations were not conclusive but suggest an abnormal phenotype. Follow up work will include the utilisation of transmission electron microscopy to look at the cilia ultrastructure.

As mentioned, CG30441 encodes an IFT20 homologue which has a role in anterograde IFT and cilium assembly. CG30441 knock-down results in shortened cilia and phenotypic defects in adult flies, which would suggest conservation of IFT20 function for CG30441. CG17599 knock-down also resulted in shortened cilia without dilation, suggesting an involvement in IFT. The current hypothesis postulates that the gene is a novel IFT-B complex protein in *Drosophila*. Lastly, CG31320 appears not to be directly involved in cilium formation, functioning instead

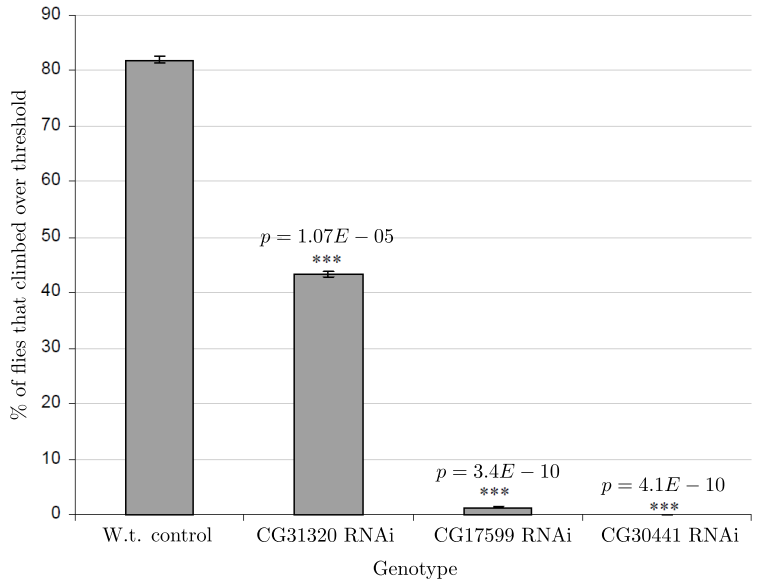

Figure 3.13: Adult fly locomotor assay. Fly mutants obtained through the expression knock-down experiment were tested for lack of coordination and for potential locomotion defects. Flies were introduced at the bottom of a vertical tube and the time taken to climb past a 50ml threshold was measured. An appropriate heterozygous control was also tested (*first bar, left*). On the y axis, the percentage of total flies which crossed the 50ml threshold within 20 seconds is indicated. 60 flies were tested in total, with 3 independent batches of 20 flies, with 5 replicates within each batch. Significance is indicated with * (2-tail student's t-test.)

as a scaffold protein (a role it would keep from its HEATR2 homologue). One hypothesis is that CG31320 has a role in the stabilisation of the IFT-B complex, indirectly through secondary interactions with CG30441. While not definitive, these results, together with existing comparative biology-based functional annotation, clearly suggest a role for each of the three proteins. While this is not completely surprising for CG30441 and CG17599, it is the first demonstration that the *Drosophila* orthologues for these two genes are indeed likely to be IFT genes. As shown in Table 3.3 (Page 101), the only evidence linking the two candidates to ciliogenesis comes from inference hypotheses in Avidor-Reiss et al. [2004]. As regards CG31320, the results are completely novel and suggest the discovery of a new type of cilia gene.

3.5 Conclusions

In this chapter, I used a number of computational methods to gain functional insight into lists of poorly known genes, implicating a small number of them with the cell-biological process of ciliogenesis. I relied on two convergent approaches: on the one side, the methodology and the software presented in Chapter 2 were used to deploy a number of protein interaction networks based on a list of fly genes with evidence for involvement in ciliogenesis, known as the Drosophila Cilia and Basal Body (DCBB) dataset [Laurencon et al., 2007]. This produced a

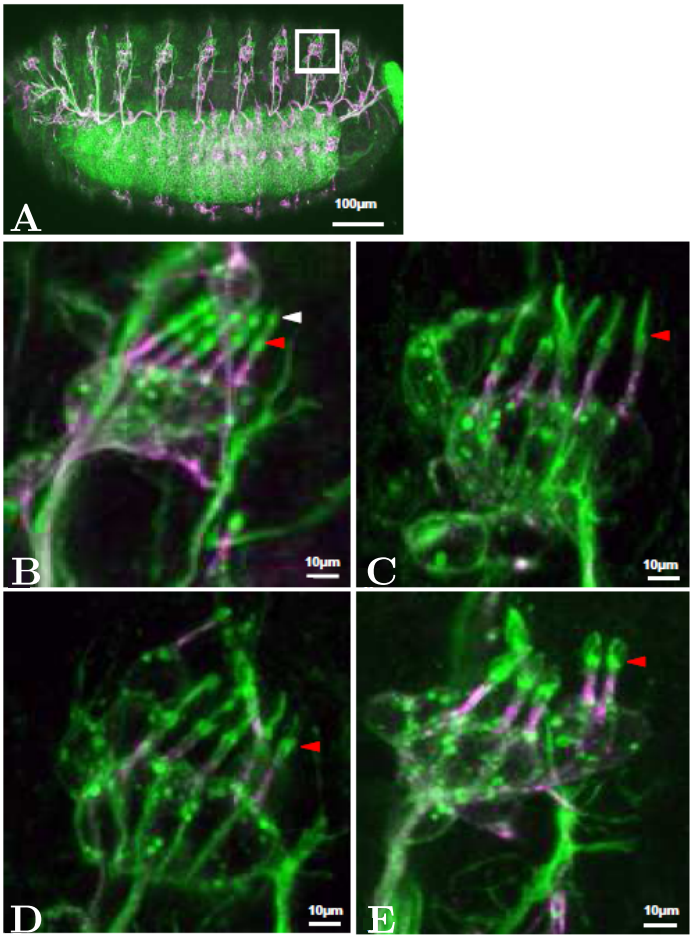

Figure 3.14: Ciliary morphology in control and RNAi embryos. RNAi knock-down of CG31320, CG17599 and CG30441 alters ciliary morphology in late stage embryos. **A**: patterning of PNS in control embryos. **B**: zoomed view of abdominal hemisegments in **A** (area surrounded by white square). **C,D,E**: abdominal hemisegments in CG31320, CG17599 and CG30441 RNAi embryos respectively. White arrow marks ciliary dilation. Red arrow denotes basal-body lumina. All images are at similar late developmental stage. Anterior is left and dorsal is up.

number of protein function hypotheses through guilt by association. On the other side, PNS transcriptional data generated in the lab [Cachero et al., 2011] was used to support the selection of some of these protein interaction-based hypotheses by highlighting, amongst the genes proposed by the protein network approach, a few candidates which showed interesting enrichment patterns in cells expressing proneural genes during sensory neuron differentiation. This is based on the idea that transcript-level information and protein-level information can inform each other [Ge et al., 2001; Segal et al., 2003a]. If a gene of unknown function has protein products participating in clusters of known ciliogenesis proteins, and additionally at least one of its transcripts is shown to be enriched in embryos during ciliary differentiation, then the two informations can be combined together to prioritise this gene and carry out additional evaluation to implicate it with ciliogenesis.

Specifically, starting from a network of fly interologs for the genes in Laurencon et al., I first carried out a number of statistical analyses to verify that the putative network was — from a structural point of view — more similar to the experimental protein interaction than to random networks. This was done to support the hypothesis that the putative data is not introducing random noise (instead of potential new hypotheses) into the system. I then identified network regions characterised by interesting connectivity and employed the Interolog Prioritisation Index (described in Chapter 2) together with the PNS transcriptional data in Cachero et al. [2011] to build evidence implicating two poorly characterised genes, CG17599 and CG30441, and one gene of unknown function, CG31320, with ciliogenesis. I concluded the chapter describing experimental validation of these candidates done by other members of the lab which followed the computational predictions. It is important to remark that, as I used an initial dataset with known functional implication to ciliogenesis to guide a discovery workflow leading to more ciliogenesis genes, the example presented in this chapter is a conservative proof of principle for the methodology introduced in Chapter 2 — which will in fact represent a useful tool for gene prioritisation in a much wider number of scenarios.

The site of synthesis of ciliary proteins in the cell body is far from the site of assembly of the cilia, therefore transport of polypeptides towards the distal tip of the flagellum must happen through the combined activity of a number of transport proteins. Two classes of molecules have an important role in this process. Firstly, IntraFlagellar Transport (IFT) proteins, which are organised in two complexes (IFT-A and IFT-B) and carry materials for the assembly and maintenance of the ciliary axoneme and membrane. Secondly, unidirectional molecular motors (dyneins and kinesins) which move IFT particles and their load from the cytoplasm to the cilia (kinesins), where the IFT proteins release their cargoes, and back to the cytoplasm (dyneins), where they recycle themselves. Defects in IFT lead to defects in the assembly of cilia, which cause a range of diseases including polycystic kidney disease (PKD) and retinal degeneration [Pazour and Rosenbaum, 2002]. One of the fly proteins considered here, CG30441, encodes the

orthologue of a human IFT-B protein, IFT20 [Yin et al., 2003]. CG30441 knock-down results in shortened cilia and phenotypic defects in adult flies, which would suggest conservation of IFT20 function for CG30441.

The knock down of CG17599, another protein evidenced by the computational screen presented here, also resulted in shortened cilia without dilations, loss of thickness and rigidity and in some cases reduction in the length of the cilium. Evidence for implication of the worm orthologue of CG17599 (dyf-3) in cilium formation [Murayama et al., 2005] also converges to suggest a role for CG17599 gene in ciliary transport, with the current hypothesis postulating the gene as a novel IFT-B complex protein in *Drosophila*.

Lastly, one gene (CG31320) was evidenced by the computational analysis because of a series of converging bits of evidence. While it is not a direct interactor of the seed set of DCBB genes labelled for `cilium` biological processes in Gene Ontology, it is a direct interactor of one of the DCBB genes not yet labelled in GO (at the time of writing). Additionally, it participates to a tightly connected group of genes in `NET_cilium`. This connected group of genes is notable because of some interesting topological properties relating its nodes with respect to the full network, and because it maintains its connectedness when the strictest IPX threshold is applied — suggesting high biological support for all the putative interactions contained in it. Another hint is that CG31320 transcript is enriched in the time course data in Cachero et al. [2011], suggesting that the gene is significantly more expressed in embryonic cells expressing two crucial proneural genes, *ato* (active early in PNS development) and *cato* (active later, during PNS and ciliary differentiation and expressed exclusively in neuronal lineages). Experimental validation proved that CG31320 has Ch-specific expression. Cilia in CG31320 RNAi mutant line embryos show marked outer dendritic segment defects, with alterations in ciliary morphology similar to those observed in CG17599 lines. However, the range of locomotion defects and uncoordinated phenotypes for CG31320 is milder compared to those observed in the other two gene lines, suggesting that CG31320 might not to be directly involved in cilium formation, but rather be functioning as a scaffold protein (not unlike its mammalian HEATR2 orthologue). While not definitive, these results suggest a role for each of the three proteins. Speculative, comparative biology-based evidence for CG30441 and CG17599 existed [Avidor-Reiss et al., 2004], however this computational study led to the first demonstration that the Drosophila orthologues for these two genes are indeed likely to be IFT genes. As regards CG31320, the results are completely novel and suggest the discovery of a new type of cilia gene.

Here, I showed that interolog data produced through the methodology introduced in Chapter 2 can help understanding poorly characterised biological processes by providing small set of hypotheses that can guide wet-lab research. In Section 3.3 I observed some of this putative protein interaction data from a network theoretical perspective. A number of statistical techniques allows quantifiable claims around the structural properties of biological networks

[Clauset et al., 2009], a practice that helps avoiding most of the pitfalls typical of some early studies of biological network structure based on speculation and anecdotal evaluation of visual evidence [Albert et al., 1999]. However, even when sound statistical analysis is employed, inferring biological conclusions from protein network structure is not a universally accepted principle.

Hakes et al. [2008] discuss the issue of the incompleteness and, most importantly, of the biased nature of protein interaction networks. First of all, experimental interaction networks are sparse. In spite of this, important topological inferences are routinely made through them and extended to full, unknown interactomes. For instance, network topology and the relationships between node entities have driven protein essentiality studies [Jeong et al., 2001; Han et al., 2004; Pržulj et al., 2004] and analysis of network structure has been used to propose evolutionary mechanisms for the appearance of cellular complexity [Berg et al., 2004; Ispolatov et al., 2005]. Currently available protein interactions are a sample of complete interactomes, and the sample is not random, but affected by multiple biases: issues with sampling biases are still poorly understood. Even when a high-confidence, validated subset of protein interactions is used, the ensuing protein interaction network is not necessarily representative of the network as a whole: most networks analysed today offer only partial insights into the true complete networks [Stumpf and Wiuf, 2005]. Stumpf and Wiuf, in particular, have studied the relationship between networks samples and full networks with respect to the transfer of topological parameters, finding that, in most cases, two distinct flavours of network samples exist: firstly, those that consist of all the nodes (and the edges between these) in a region of the full network, which, although not representative of the network as a whole, may offer valuable insight into the nature of some biological process within a certain neighbourhood. The second kind of sample network is obtained when each node of the global network is included in the sub-network with probability p, and only the edges between pairs of nodes which are included in the sub-network are studied. This type of sub-network is more commonly found in biological network studies, and, according to Stumpf and Wiuf, is more problematic: the degree distributions of the global network and sampled sub-networks will be qualitatively different [Stumpf et al., 2005]. For instance, it has been shown that some scale-free architectures might be an artefact caused by regularities and biases in the selection of the datasets and, according to some studies, may not reflect any biological importance [Han et al., 2005].

While not all criticism is fair, it is evident that currently available protein interaction data has flaws. Additionally, many early interactomics studies based on network structure analysis rested on arguably shaky ground, which produced sometimes inaccurate results and controversial claims for ubiquitous biological signatures. However, if a lesson can be learnt from these insights, is that extra care must be used when manipulating protein data to make biological inferences. In this chapter I have discussed a protein interaction dataset and a series of prin-

cipled analyses which resulted in biologically interesting experimental results. I would argue that, when curated through some form of manual or automatic technique, filtered to retain only samples showing high experimental support and used in conjunction with other well supported data of different nature to build small prioritisation datasets, protein interaction data can lead to interesting biological insights.

CHAPTER 4

Network Communities and Cilia Specialisation Genes

In this chapter, I extend the workflow introduced in Chapter 3 and adopt a similar approach to study genes active during *Drosophila* sensory neuron differentiation and suspected to have specific roles in mechanosensory cilia specialisation. This chapter builds on Chapter 3 because it is based on a similar principle: to use a combined proteomics-transcriptomics approach for highlighting a manageable number of candidate *Drosophila* PNS development genes and use the results to inform wet lab research. However, while in Chapter 3 I discussed options for computational analysis of generic ciliogenesis genes, universally required for cilium formation, here the focus is on a more specific group of ciliogenesis genes, some of which are active only in the developing chordotonal neurons, and are known to be regulated by the Forkhead Transcription Factor Fd3F — recently associated with the control of sensory cilia specialisation [Newton et al., 2012]. In addition to discussing a more specific biological domain, this chapter proposes an additional array of computational techniques to analyse interolog data and once again leads to promising experimental evidence.

4.1 Mechanosensory Cilia Specialisation and Fd3F

As anticipated in the introductory chapter, the Ch-neuron specific gene forkhead factor Fd3F directly regulates genes related to specialised aspects of Ch cilium differentiation and function. Phenotypic analysis has shown that mutants which do not express *fd3F* present cilia that lack their motile apparatus and normal specialised sub-compartments [Newton et al., 2012]. Fd3F cooperates closely with the ciliogenic factor Rfx and acts as a cell-type-specific modulator of Rfx target gene specificity (Figure 4.1). Here, we are particularly interested in the regulatory dynamics happening downstream of Fd3F. A number of Fd3F target genes have already been identified which are implicated in cilia motor function and in delineating the ciliary compartments.

Two functionally and spatially distinct zones constitute chordotonal cilia: a *distal* zone

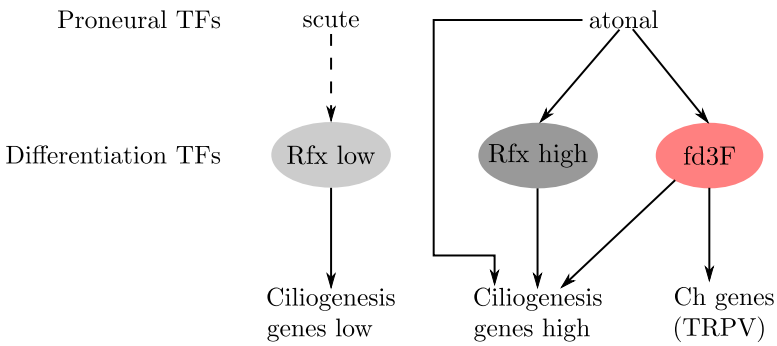

Figure 4.1: Rfx and Fd3f linking specification to differentiation in PNS development. Scute is not a direct regulator of Rfx, hence the dashed arrow. "Low" and "High" refer to transient and persistent expression in ES and Ch cells. (*Image adapted from* zur Lage et al. [2011]).

(sensory and non motile) and a *proximal* zone, which is motile, and contains, amongst other molecules, axonemal dyneins and Transient Receptor Potential (TRP) ion channels.

Axonemal dyneins are molecular motors: proteins that work by undergoing a number of shape changes and converting chemical energy from adenosine triphosphate (ATP) into mechanical energy. Three families of molecular motors are known, dyneins, myosins and kinesins. Both dyneins and kinesins transport various cellular cargoes along microtubules in a unidirectional fashion: towards the minus-end of the microtubule in the case of dyneins, and towards the plus-end of the microtubule in the case of the kinesins. Specifically, *axonemal* dyneins, also called ciliary or flagellar dyneins, are found in the axoneme, the microtubule-based inner core structure of cilia and flagella. Typically, the axoneme of motile cilia consists of nine doublet microtubules arranged cylindrically around a pair of singlet microtubules. Axonemal dyneins are distributed along each doublet as inner and outer rows of arms (Figure 4.2-A). Fd3f has been shown to regulate genes required for several aspects of dynein arm formation [Newton et al., 2012].

As regards TRP ion channels, these are a group of cellular sensors involved in a wide variety of cellular processes and were initially discovered in a *trp* mutant strain of *Drosophila* [Minke, 2010]. TRP channels are usually tetramers formed by subunits with six transmembrane domains showing high calcium cation permeability. The molecular architecture of TRP channels is reminiscent of voltage-gated channels and comprises six putative transmembrane segments (S1-S6), intracellular N- and C- termini and a pore-forming loop between S5 and S6 [Gaudet, 2008]. The physiological functions of these channels range from sensory tasks such as temperature sensation and taste transduction, to motile functions, such as muscle contraction and vaso-motor control [Gees et al., 2010]. The TRP family is a well conserved set of about 30 channels divided, from a sequence homology perspective (Figure 4.2-B), in seven subfamilies: TRPC ('canonical'), TRPV ('vanilloid'), TRPM ('Melastatin'), TRPP ('polycystin'), TRPA ('Ankyrin'), TRPML ('Mucolipin') and TRPN (Nomp-C homologues) [Pedersen et al.,

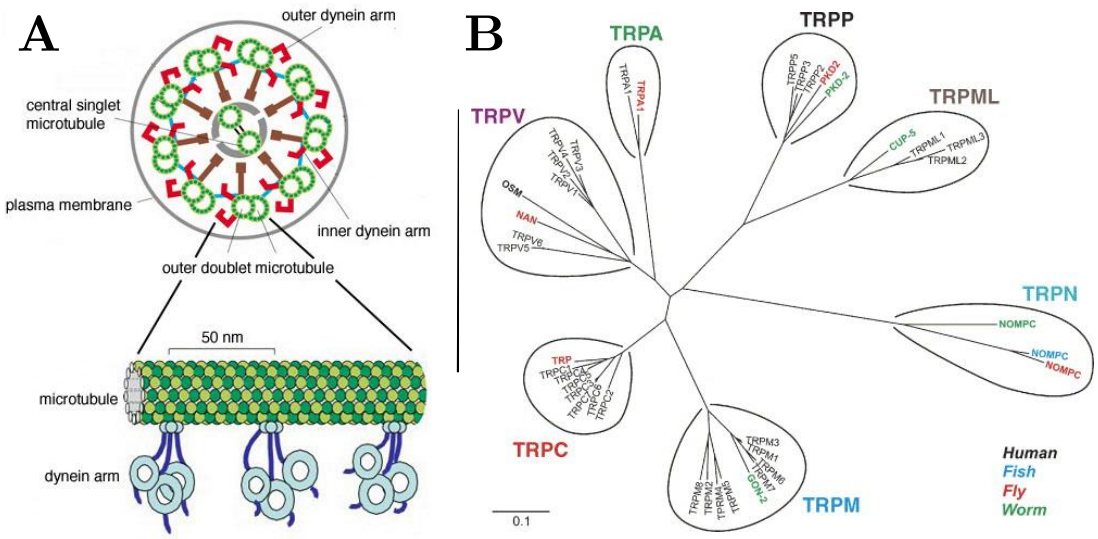

Figure 4.2: **A**: Schematics showing section of motile cilia with the characteristic '9+2' structure, composed by two central single microtubules encircled by nine outer double microtubules. The outer and inner dynein arms slide along each other double microtubule (*image credit*: worms.zoology.wisc.edu/). **B**: Phylogenetic tree of the Transient Receptor Potential Ion Channel super-family (*image credit*: [Pedersen et al., 2005]).

2005]. In particular, the TRPV *vanilloid* subgroup contains six mammalian members, all of which have been associated to Ca^{2+} entry channel function, gated by a variety of physical and chemical stimuli. The group includes the fly proteins *nanchung*[1], and *inactive* (iav), expressed in vivo exclusively in chordotonal organs and known to be related to sensory perception [Kim et al., 2003; Gong et al., 2004]. These two fly homologues of TRPV4 have additionally been related to mechanosensory transduction mechanisms in fly chordotonal cilia [Gopfert et al., 2006], and Cachero et al. [2011] showed that both are regulated by Fd3F, explaining why *fd3F* mutant flies are deaf/uncoordinated. Further work in Newton et al. [2012] provided evidence showing that Fd3F directly regulates the two genes[2].

Overall, Newton et al. [2012] show that Fd3F is required for the expression of several chordotonal-specific axonemal dyneins and Transient Receptor Potential Vanilloid (TRPV) ion channels which are required for sensory transduction. Fd3F also regulates retrograde transport genes, which are required to differentiate the distinct motile and sensory ciliary zones. It follows from this that Fd3F, which was already known as an intermediate factor regulated by *ato*, is in turn an important activator of genes for specialised aspects of chordotonal cilium differentiation and function. This means that the elucidation of the regulatory cascade going from neuronal specification to Ch-neuron differentiation must pass through *fd3F*. In this chapter, I evaluate the possibility of using again protein networks and network theory analysis to a) aid

[1] Korean for deafness.

[2] *fd3F* regulates *nan* and *iav* jointly with *Rfx*. This demonstrated that Rfx, a pan-ciliary transcription factor required for both Ch and ES neurons, also contributes to the regulation of Ch-specific *fd3F* targets.

in understanding known Fd3F targets by virtue of their predicted protein interactions, b) aid with the prediction of further Fd3F targets and c) hypothesize the roles that such potential new targets may play in the Fd3F sub-program.

4.1.1 Cato-GFP Data and Known Fd3F Targets

This section briefly describes the data I used to build the protein networks analysed and discussed in the rest of the chapter. The underlying rationale is to assemble a protein interaction network and once again augment it using interologs produced with `Bio::Homology::InterologWalk`. Unlike the examples in Chapter 2 though, here I do not start from the full *Drosophila* interactome, because I wish to obtain a much closer view on a small subset of potentially related proteins. A more fitting candidate list from a size point of view would be the DCBB data set used in Chapter 3. However, here I aim to find an even smaller initial gene set with direct *Drosophila* evidence of relatedness to ciliary specialisation processes.

One way to assemble a restricted list of related fly genes to feed to `Bio::Homology::InterologWalk` was to choose genes based on transcript enrichment in late PNS development in cells expressing a proneural factor[3], possibly one known to be active after SOP commitment and around the onset of mechanosensory cilia specialisation. By selecting genes enriched in Ch cells expressing a late onset PNS proneural factor I anticipated I could build a protein interaction dataset involving the products of these enriched genes plus a number of additional experimental and putative interactors. These interactors, while not enriched in the same cells at the same developmental time-points, can produce interesting hypotheses because, arguably, many important PNS development molecules will be up- or down-regulated in an *absolute*, rather than *relative* sense. This is based on the insight that proteins with a role in Ch neurons *and* and a role outside Ch neurons will not be detected by a transcript enrichment study. However, they might be detected through their protein interactions with genes showing exclusive increase of activity in proneural factor-expressing cells.

As a first step in this direction, I initially examined the transcriptional time course data for *ato*-expressing cells introduced by Cachero et al. [2011]. The transcriptional data collected by Cachero et al. analysed embryonal cells at three time points, corresponding to the first 3 *h* of neural development. The expression profiling was designed to reveal dynamics of differentially expressed genes in cells expressing the proneural gene *ato* compared to cells that did not express it — with T_1 representing the point of maximal *ato* expression, and T_2 and T_3 reflecting subsequent post-*ato* development as the precursors divide leading up to differentiation. One option was to create a starting gene set based on genes which are at least 2-fold enriched in *ato*-expressing cells, time-wise as close to differentiation as possible, i.e. during T_3. However

[3] Meaning genes producing significantly more transcript in cells expressing a transcription factor with respect to cells that do not express said transcription factor.

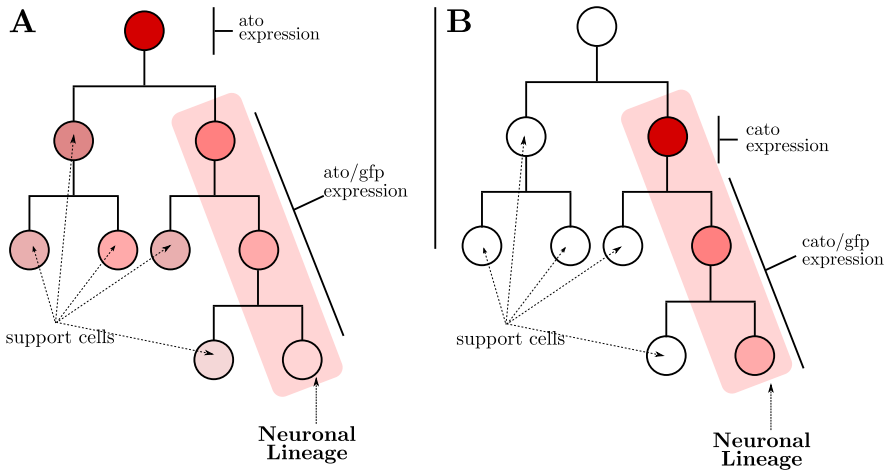

Figure 4.3: Specificity of *ato*-GFP vs *cato*-GFP expression in SOP lineages. **A**: the proneural gene *atonal* is a high-level regulator and the *ato*-GFP marker gene is expressed both in the neuronal lineage and in the lineages producing supporting cells. **B**: the gene *cousin of atonal* is a target of atonal, and codes for a bHLH transcription factor expressed in the developing PNS after neural precursor selection but before terminal differentiation. The *cato*-GFP reporter line is very much concentrated in the sensory part of the sense organ precursor lineage, thus making *cato* GFP-tagged cells more specific markers than *ato* GFP ones in differentiation studies.

there are two problems with this choice: firstly, *ato* is expressed only transiently during sense organ precursor formation: it has been shown to be a high-level regulator, which initiates a regulatory cascade eventually leading to differentiation genes. Most importantly, *ato* is not exclusively expressed in the sensory neuron part of the sense organ precursor lineage: *ato*-GFP marked expression is present in support cells as well (Figure 4.3-A).

A better option was represented by transcriptional data of genes enriched in cells expressing *cousin of atonal* (cato) [Cachero *et al.*, unpublished data], another bHLH transcription factor [Goulding et al., 2000a] widely expressed in the developing PNS after neural precursor selection but before terminal differentiation [zur Lage and Jarman, 2010]. *cato* is a direct target of *ato* [zur Lage and Jarman, 2010], has been shown to be dynamically expressed during neurogenesis and is confined to the developing PNS: differently from *ato*, *cato* is expressed almost exclusively in the sensory neuron part of the SOP lineage (Figure 4.3-B). The modality of expression is very different from that of high-level proneural genes [Goulding et al., 2000a] in that it is initiated in sense organ precursors after their formation. Its expression continues in the division products of the precursors, after high-level proneural genes such as *ato* and *achaete-scute* are switched off. The higher specificity of *cato*-GFP expression, together with the availability of enrichment data for *cato*-expressing cells in T_4 (one hour later than the last *ato* time point) prompted me to utilise genes enriched in *cato*-GFP-expressing cells at time point T_4 as a starting gene set for a protein interaction analysis based on the principles outlined above.

Such an analysis is likely to highlight a number of potentially interesting candidates but is

also likely to extract a high amount of noise and unrelated cellular housekeeping genes. One thing that can be done to further restrict the results produced by a protein interaction analysis like the one described above is to, once again, seed it[4] using genes sharing some form of functional role. In the past two chapters I used GO annotation to achieve this.

For this analysis, I decided to use first hand information resulting from recent work in the lab, some of which is published in Newton et al. [2012]. Thus, the set of seed genes in this analysis is the list of 26 known and potential Fd3f targets shown in Table 4.1. Some of the genes in the table are orthologues of genes known to be regulated by the vertebrate gene *foxj1*. The gene codes for a protein that, in mammals, has been implicated in activating a set of genes essential for motile cilia formation and function [Yu et al., 2008b]. Fd3F targets having human orthologues regulated by Foxj1 are mostly functionally implicated in axonemal dyneins motor activity. This suggests as the ancestral function of *fd3F* the regulation of the motility apparatus in chordotonal cilia [Newton et al., 2012]. Several of the known target genes have a conserved X-box + Forkhead domain binding motif combination, usually within 100bp of the transcriptional start site. As the X-box binds Rfx and the Forkhead domain binds Fd3f, this suggests that Rfx and fd3F may cooperatively regulate this group of genes.

By highlighting genes showing evidence for Fd3F regulation and mining their closest interactors in a protein network of highly enriched late specialisation genes, I hypothesised I could obtain smaller, more manageable interaction datasets and analyse these to look for regions containing potential new Fd3F targets. A group of approaches to doing this will be described in the following sections.

4.2 A Cato-T4 Protein Interaction Network

The transcriptional data for *cato* T_4 considered here was obtained by lab members in parallel with the *ato* data described by Cachero et al. [2011], using the technique described in the introductory chapter.

For this analysis, I downloaded the data from a repository maintained by Dr. Ian Simpson[5]. The data had been pre-processed and quality-checked as detailed in Cachero et al. [2011] (*Materials and Methods*). Whenever I could not obtain a 1:1 match between affymetrix probeset ID and Flybase ID, I discarded the corresponding expression data point. I then filtered the remaining non-ambiguous data points by fold change and by false discovery rate: I chose FC ≥ 2 and FDR ≤ 0.01, for consistency with the analysis of significant enrichment of *ato* data in Cachero et al. [2011]. The resulting 285-gene list ranked by decreasing fold change is shown in Table A.6, Page 196.

I double checked each gene ID against Ensembl for consistency and obtained up-to-date

[4] See Chapter 2, Section 2.4.2, Page 69, and Chapter 3, Section 3.4, Page 106.
[5] neuroregulatorygenomics.org/

Gene	Orthologue	Function	X+F Motif	*Foxj1* target
tektin-A	TEKT4	Axoneme stability	Y	M,X
CG8800	DNAL1	Motility - Axonemal dynein	Y	M,X
CG9313	DNAI1	Motility - Axonemal dynein	Y	M,X,Z
CG34192	DYNLRB1	Motility - Axonemal dynein	Y	M,X
CG13930	WDR78	Motility - Axonemal dynein	Y	M,Z
CG6971	DNALI1	Motility - Axonemal dynein	Y	M,X
Dhc16F	DNAH6	Motility - Axonemal dynein, inner arm	Y	M
Dhc62B	DNAH3	Motility - Axonemal dynein, inner arm	Y	M
Dhc93AB	DNAH9	Motility - Axonemal dynein, outer arm	Y	M,X,Z
CG10064	WDR16	Motility related	Y	X
CG14905	ODA1 (*C reinhardtii*)	Axonemal dynein assembly	N	
tilB	LRRC46	Axonemal dynein assembly	Y	
btv	DYNC2H1	Retrograde transport - dynein 2	Y	
CG3769	DYNC2LI1	Retrograde transport - dynein 2	Y	
Oseg1	IFT122	Retrograde transport - IFT-A	Y	
rempA (Oseg3)	IFT140	Retrograde transport - IFT-A	Y	
Oseg4	WDR35, ifta-1 (*c elegans*)	Retrograde transport - IFT-A	Y	
Oseg6	WDR19	Retrograde transport - IFT-A	Y	
CG5780	IFT43	Retrograde transport - IFT-A	Y	
nan	TRPV[1-6]	TRPV ion channel, active amplification control	Y	
iav	TRPV[1-6]	TRPV ion channel, active amplification control	Y	
CG6980	TTC12	Unknown, TPR motifs	Y	
CG10339	-	Unknown	Y	
CG31320	HEATR2	Unknown	Y	
CG11253	ZMYND10	Unknown	Y	
CG16984	ENKUR	Unknown - TRP-C Interacting?	-	

Table 4.1: Map of 26 known and potential Fd3F targets. All genes are at least 2-fold enriched in the transcriptome of Ch- ato-expressing cells, according to the time course microarray data in Cachero et al. [2011]. Additionally, there are genes from a screen performed by Newton et al. [2012] showing mRNA expression of axonemal dynein genes in wild-type and *fd3F* mutant embryos, retrograde transport candidates [Ishikawa and Marshall, 2011] and axonemal motility candidates [Wickstead and Gull, 2007]. The column *Foxj1 target* indicates whether the fly gene has homologues which are known Foxj1 targets in any of mouse (M) [Jacquet et al., 2009], xenopus (X) [Stubbs et al., 2008], zebrafish (Z) [Yu et al., 2008b].

	B::H::I Pipeline	
	Putative	**Known**
Datasets		
Gene IDs	**285**	**285**
Reference Genomes used	375	1
Unique PP Pairs	**948**	**975**
Surviving IDs (% **Gene IDs**)	113 (39.6)	153 (53.7)
Networks		
Nodes	**789**	**887**
Edges	**948**	**975**
Novel Nodes (% **Nodes**)	676 (85.7)	734 (82.7)

Table 4.2: `Bio::Homology::InterologWalk` output using the Cato T_4 enriched list as the input dataset.

Flybase IDs for each identifier where necessary. These IDs represented the input dataset for `Bio::Homology::InterologWalk`. For the orthology data collection, I chose the EnsemblGenomes Pan-homology database[6], V. 63. As regards the `Bio::Homology::InterologWalk` set-up, I discarded all homology relationships belonging to the `paralog` class, for the reasons explained in Chapter 2. For the interaction collection phase the set-up was identical to the one introduced for the DCBB analysis in Chapter 3 (section 3.2, page 90).

Two parallel runs of `Bio::Homology::InterologWalk` were completed: one processed the Cato T_4 dataset through the putative pipeline, while the other processed the same genes through the direct pipeline to obtain all fly experimental interactors for the dataset. Table 4.2 shows statistics relative to the resulting putative and direct datasets. Once again, I imported the two networks in Table 4.2 in Cytoscape [Shannon et al., 2003] and merged them to obtain their union as described in Chapter 3. The resulting union network, designated `NET_CATOT4_union` (1607 nodes, 2072 edges) features a large connected component of 1493 nodes and 1999 edges and was again processed to retrieve a smaller sub network based on the seed genes set. Of the 26 Fd3F targets in Table 4.1, I found 19 in `NET_CATOT4_union`. I extracted these 19 seeds together with their nearest neighbours.

The resulting sub-network, `NET_CATOT4_NN`, (118 nodes, 117 edges, Figure A.3, Page 188) is a collection of 11 clusters, the largest of which connects via protein interactions 5 of the 18 Fd3F targets. One cluster in the network (Figure A.3-2, *shaded area*) links one of the genes studied through the analysis done in Chapter 3, CG31320, to 3 other interesting Fd3F targets, one of them (CG11253) having very little to no experimental functional evidence [FlyBase Curators et al., 2004].

In order to better evaluate the interaction context of this and other poorly annotated Fd3F targets I decided to analyse `NET_CATOT4_union` adopting an alternative approach, and em-

[6] 30 June 2011, 375 genomes.

ployed community-finding algorithms to evidence densely connected areas in the network.

4.3 Network Community Detection

One way to discover interesting relationships between molecules in a protein network is to use algorithms to decompose the network in communities of nodes. These are sometimes known as *groups*, *modules*, or *clusters*[7].

In general, cluster analysis is the mathematical study of methods for recognizing natural groups within a class of entities [van Dongen, 2000]. In the context of network theory, cluster analysis can be described as the problem of finding a sensible decomposition of a graph into sub-graphs according to some metric, in such a way that the nodes in each sub-graph have more to do with each other than with outsiders. The nodes in each sub-graph are then said to be in the same community.

The study of the algorithms and of the metrics needed to obtain network communities is a field that has always been driven by demand from various disciplines engaged in exploratory data analysis [Everitt et al., 2009]. Systems of interest to the scientific community that can be modelled by networks and studied from a cluster perspective include social groups [Wasserman and Faust, 1994], the Internet [Faloutsos et al., 1999], food webs [Dunne et al., 2002], biochemical networks [Kauffman, 1969] and protein networks [Spirin and Mirny, 2003].

Several approaches have been proposed to study the problem of finding communities within networks. Early propositions include the Kernighan-Lin algorithm [Kernighan and Lin, 1970] and spectral methods [Fiedler, 1973; Pothen et al., 1990], however the performance of these methods has been shown to be inherently topology-dependent [Newman, 2004b]. A number of algorithms have been proposed over the years to obtain increased robustness and speed: Fortunato [2010] provides an in-depth survey of options and recent advances in network clustering algorithms. In this study, I explored two classes of popular community-finding algorithms. My choice was dictated by availability of publications, support, source code, implementations and quality of results in comparative studies (for instance, Brohee and van Helden [2006] and Vlasblom and Wodak [2009]). The two approaches will now be described further.

The first one is based on divisive algorithms employing edge betweenness as a metric to identify the boundaries of communities. The algorithm works by progressively removing the edges having the highest betweenness centralities, producing a dendrogram with node-community mappings. The idea was proposed by Girvan and Newman [2002] and successfully applied to a vast number of phenomena ranging from gene networks [Wilkinson and Huberman, 2004] to jazz collaboration networks [Gleiser and Danon, 2003]. The original algorithm by Girvan and Newman is unfortunately very computationally demanding, running in $O(m^2n)$

[7] The concepts of community and cluster, while not equivalent in some disciplines (e.g. physics), will be considered equivalent in the context of this thesis.

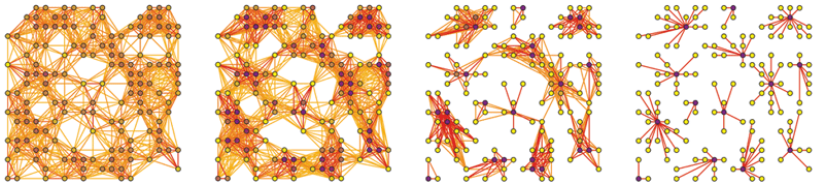

Figure 4.4: MCL is a fast and scalable unsupervised clustering algorithm based on simulations of stochastic flow in graphs (Image courtesy micans.org/mcl/).

time on a network with n nodes and m edges or $O(n^3)$ time on a sparse graph (most real-world networks of interest fall into the latter category), which limits practical usability to networks of a few hundreds of nodes at most. In a later publication, one of the two authors then addressed this problem and described a much faster algorithm [Newman, 2004a], based on the greedy optimisation of a measure called the *modularity* of the network. The algorithm produces communities which are very similar to the ones in Girvan and Newman [2002], but runs in $O((m+n)n)$, or $O(n^2)$ for sparse networks, and thus brings within reach the study of communities in networks of hundreds of thousands of nodes. This was then further improved with the help of more sophisticated data structures in Clauset et al. [2004], who suggested an algorithm able to produce communities identical to those in Newman [2004a] in less time[8]. Additional improvements on this family of algorithms came from the work of Wakita and Tsurumi [2007], who introduced an additional set of tweaks based on heuristics optimising the balance of communities being merged by the algorithm in Clauset et al. [2004], which pushed performances further.

Another interesting group of graph clustering algorithms is based on the idea of stochastic simulations of random walks. A graph showing some form of community structure will have many links within a community, and fewer links between the communities: a random walk starting from a generic node and then randomly visiting connected nodes is more likely to stay within a cluster than hopping between clusters. By randomly traversing the graph, an algorithm based on random walks will find out where the flow tends to gather and will classify areas of intense flow as clusters. This idea is at the core of the Markov CLustering Algorithm (MCL) [van Dongen, 2000] a fast algorithm based on flow simulation (Figure 4.5). MCL has been applied in a variety of contexts, including EST analysis [Dunn et al., 2008], protein networks community-finding [Enright et al., 2002], phylogenomic analysis [Robbertse et al., 2006]. It works by building the graph's transition matrix and applying an iterative algorithm having at its core two steps: in the initial *expansion* phase, the algorithm takes the e-th power of the matrix, where e is the expansion parameter. In the *inflation* phase, each non-zero value in the matrix is raised to a power (the inflation parameter I, user-selectable) followed by a diagonal

[8] Their implementation scales almost linearly with the number of network nodes for networks showing some form of hierarchical structure.

Parameter	Value
Clusters	73
Average size	22.014
Maximum size	130
Minimum size	2
Modularity	0.787

Table 4.3: Results for fast greedy Newman Girvan on `NET_CATOT4_union`.

scaling of the result. During each iteration, all values below a certain threshold are dropped from the matrix after normalisation. The expansion phase can be thought of as a "spreading" out of the flow which becomes more homogeneous, while the inflation phase corresponds to a "contraction" of the flow: in other terms, the expansion operator is responsible for allowing flow to connect different regions of the graph, while the inflation operator is responsible for both strengthening and weakening of current. The process converges towards a partition of the graph, with a set of high-flow regions separated by boundaries with no flow. The value of the inflation parameter I controls the extent of the effect of this operator on the matrix and, ultimately, influences the granularity of the output clusters.

4.4 Communities in Cato-T4 Network

For this study I mostly relied on the community-finding algorithms based on the method proposed in Girvan and Newman [2002] and on its refinements, although I also performed separate tests using MCL, to evaluate the difference in the output clusters produced. In order to obtain communities for `NET_CATOT4_union` I initially used the Cytoscape plug-in clusterMaker v. 1.9[9] which implements both families of algorithms: for the Newman-Girvan family, clusterMaker incorporates Glay [Su et al., 2010] a plug-in based on the work by Wakita and Tsurumi [2007]. A visualisation of the resulting communities using the `NET_CATOT4_union` input network and the Girvan-Newman fast greedy algorithm is shown in Figure 4.5. An alternative visualisation (showing node provenance information with respect to the original Cato T_4 seed list, the putative pipeline, and the experimental pipeline) is provided in Figure A.4, Page 189. Output statistics are in Table 4.3.

In parallel with this analysis, I processed `NET_CATOT4_union` using MCL. All user-selectable parameters were set to default, apart from the inflation parameter I, were I set $I = 1.7$ instead of the default $I = 2.0$. This was done in order to slightly increase the granularity of the resulting clusters as the default value produced hundreds of small motifs of no practical interest. Results for the MCL clustering are in Figure A.5, Page 190.

I found the Girvan-Newman based algorithm to give more topologically articulate commu-

[9] www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.html

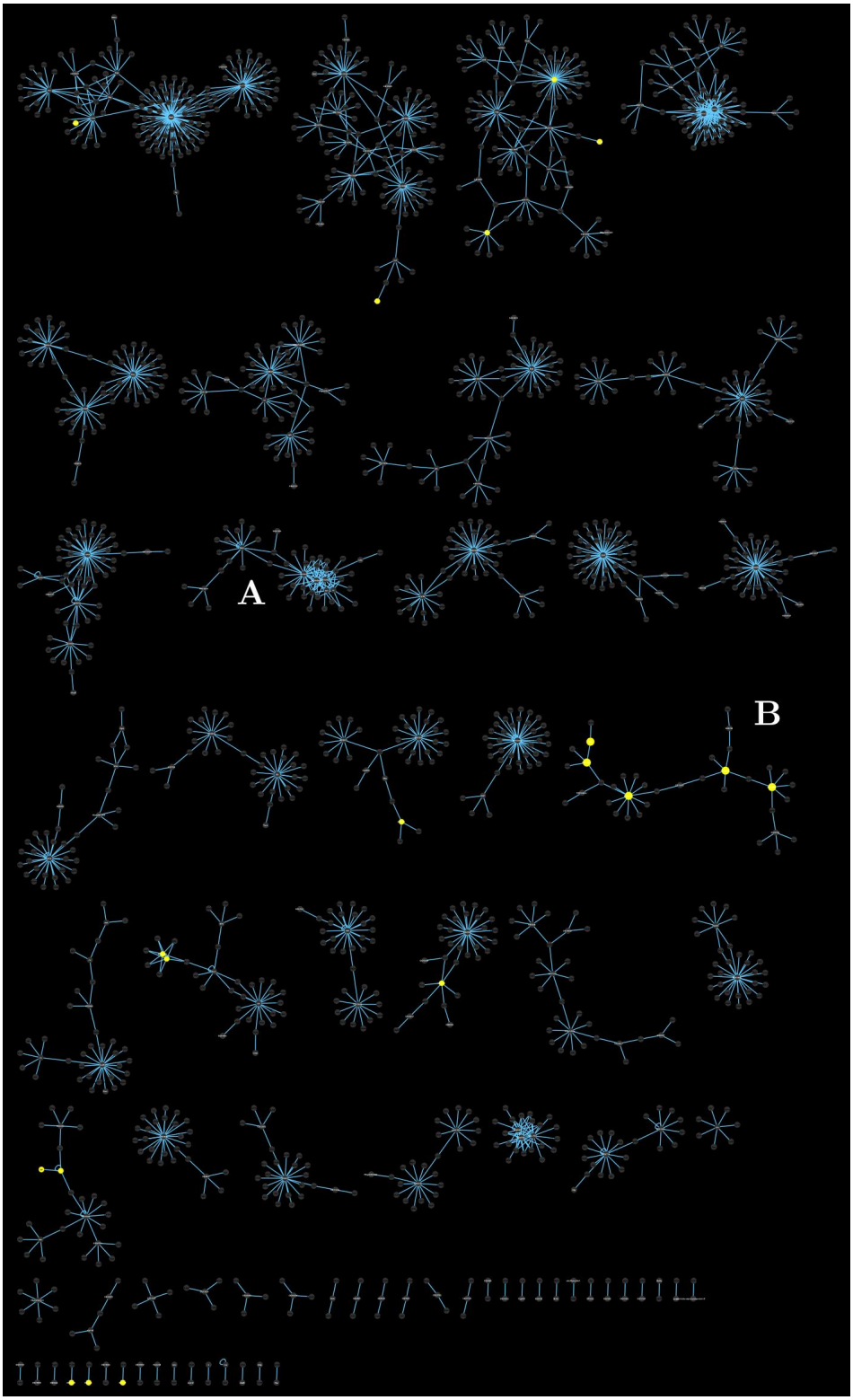

Figure 4.5: `NET_CATOT4_union` — Communities found using the optimised greedy Newman-Girvan algorithm implemented in Glay [Su et al., 2010]. Yellow nodes are Fd3F target genes (Table 4.1, Page 135). Two communities are labelled. **A**: (53 nodes, 92 edges) proneural and HES TFs. **B**: (38 nodes, 37 edges) 5 of the 19 total Fd3F targets.

nities compared to those retrieved by MCL. The latter partitioned the network in simple "tumbleweed" sub-networks — spoke-shaped collections of nodes each basically containing one hub and its neighbours. For this reason I shall be primarily following up the results obtained through the Girvan-Newman method. MCL has been shown to work better with weighted networks (for instance, BlastP E-values are often used as weights in those studies employing MCL for sequence homology clustering). In my tests, `NET_CATOT4_union` was unweighted, which might be one of the reasons for the spoke-like nature of the clusters retrieved: as there is no difference in weight between the edges attached to hub nodes and the edges attached to transition-zone nodes, the expansion-inflation steps have probably failed to break the spoke topologies surrounding hubs while all connectivity around transition-zone nodes has been disrupted. It would be interesting to repeat the MCL experiments using a weighted version of `NET_CATOT4_union` obtained using the IPX or PCS (Chapter 2). Another reason why MCL only retrieved clusters composed of hubs and their direct neighbours might be the lack of dense local structure in `NET_CATOT4_union`: MCL has been described to work better with larger graphs showing dense local structure [Pereira-Leal et al., 2004].

The NG community partition identifies communities which approximate functionally related groups of genes in Ch neurons. In some cases, the predictions are remarkably precise. Figure 4.5-A is a cluster of 53 genes including several basic Helix-Loop-Helix proneural factors (ato, cato, achaete, scute, asense, daughterless), bHLH transcriptional co-repressors from the *Enhancer of split* family (E(spl), HLHM3, HLHm5, HLHm7, HLHmbeta, HLHmdelta, HLHmgamma), other bHLH transcriptional corepressors implicated in neurogenesis, segmentation and sex determination [Paroush et al., 1994] (groucho, hairy) and Zinc finger domain proteins (eagle). Overall, this cluster describes interactions between the products of genes implicated in neural fate commitment within proneural clusters (proneural factors) and co-repressors implicated in transcriptional inhibition, maintaining progenitor cells in undifferentiated state through lateral inhibition mediated by the Notch pathway [Kageyama et al., 2007]. This cluster assignment suggests that the algorithm can capture communities composed by genes functionally related according to the literature.

Figure 4.5 also shows how the 19 interacting Fd3F targets described in the previous section are distributed in terms of community membership. One such community of Fd3F targets will be discussed in the next section.

4.4.1 A Community of Fd3F Targets

A gene cluster particularly stood out for the high number of Fd3F targets involved in interactions (Figure 4.5-B). The community (38 nodes, 37 edges) models protein interactions between some of the members of the list in Table 4.1, namely those with the smallest amount of available functional evidence, but with plenty of circumstantial evidence for a role in cilia. *Community*

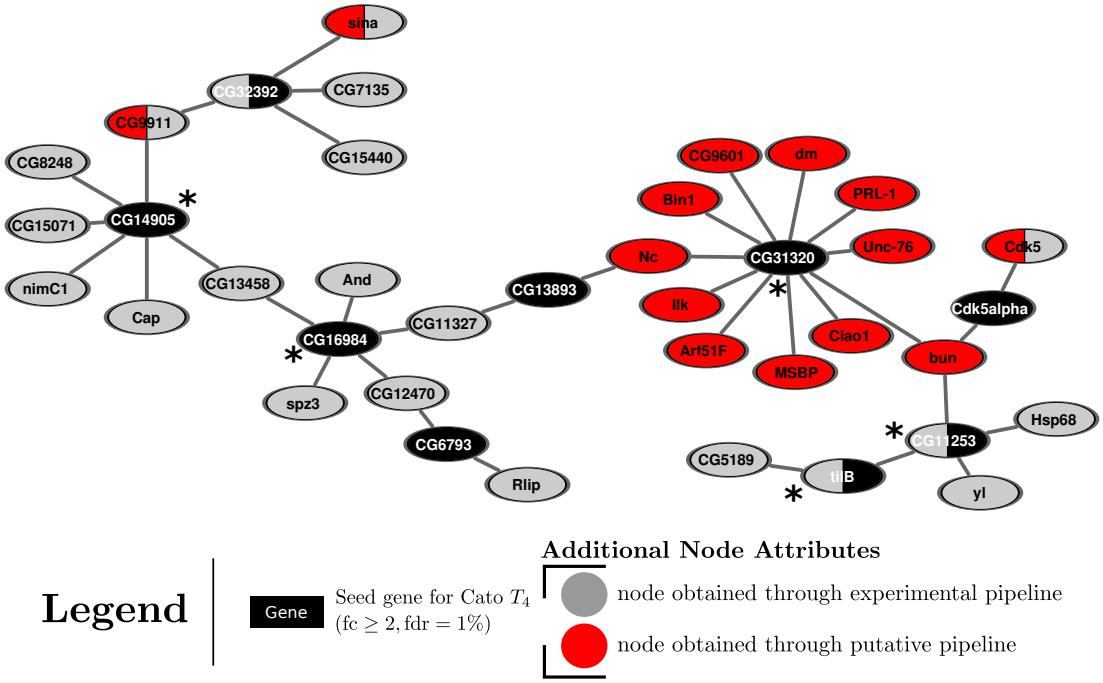

Figure 4.6: Close-up view on community B from Figure 4.5, Page 140. The community features 5 genes (nodes marked by an asterisk) over a total of 19 interacting genes with evidence of a role in cilia differentiation (Table 4.1, Page 135).

B is reproduced for clarity in Figure 4.6. This community is clearly interesting from a biological perspective and might again indicate that the community-finding algorithm is successful at grouping together related sets of genes. However, this hypothesis must be tested from a statistical point of view. The next section proposes a number of approaches to assessing the statistical significance of Community B and of the high number of Fd3F targets concentrated in it.

4.4.1.1 Significance Analysis of Community B

There is no definitive standard in the graph theory literature on how network community significance should be evaluated, although recently a number of studies in the physics community [Lancichinetti et al., 2010; Mirshahvalad et al., 2012] have started to address the need for standardised testing. The popular graph manipulation API igraph [Csardi and Nepusz, 2006] implements no community detection significance methods at the time of writing. Therefore I evaluated the significance of community B using a combination of approaches.

As a first test, I calculated the internal and external degree for each vertex in the community (i.e. the number of incident edges leading inside and outside the community, respectively), and ran a Mann-Whitney U-test. The null hypothesis H_0 is that the distribution of internal and external degrees is the same. If the null hypothesis is correct, the community is not a real

one because it is just as dense inside then outside. The test will reject it if the community is a significant one. The result ($p = 2.02E-08$) is highly significant at the 1% level and H_0 is rejected. From this test it can be concluded that the partition of communities determined by the fast greedy Newman-Girvan approach is sound in the case of community B: proteins in the community are topologically more related to themselves than to proteins outside the community.

For my second test, I evaluated a group of more specialised measures for significance detection [Lancichinetti et al., 2010]. These approaches are based on comparing a community whose significance needs to be evaluated with the *best* expected result for a general null random model. This is motivated by the insight that community detection algorithms will in general produce the best possible partitioning of a graph even if the graph is random. Community significance is obtained as the extreme probability [Janke et al., 2003] of finding a cluster equal or better than the one built in a set of random equivalent graphs [Lancichinetti et al., 2010]. I assessed two metrics proposed by Lancichinetti et al. [2010] on Community B. First, I tested the more stringent $\mathcal{C}$-score. This uses statistics relative to the *worst* node in the community to assess statistical significance. The worst node w in community B is defined as the one with the lowest internal degree[10] k_w^{int}. Lancichinetti et al. assume that in a random network there is no drastic variation between k_w^{int} and the internal degree k^{int} of the best nodes outside the group. Formally, the $\mathcal{C}$-score is defined as the probability that, given an optimised community in an equivalent random graph partitioning, k_w^{int} is higher or equal than the one observed in the community of interest. The measure ultimately represents the probability of occurrence of a group with the same properties[11] in a null model where links are randomly placed. Since relying on the worst node only for significance assessment can be too stringent a criterion for many applications, Lancichinetti et al. also propose a second score, the $\mathcal{B}$-score, which uses an algorithm to build a "border" $\mathcal{B}$, i.e. a set including the worst nodes in the community[12]. The $\mathcal{B}$-score is the probability that the sum of the scores of the worst t nodes of an optimised community in several random networks is smaller than the one given for Community B in Figure 4.5. The quantity is equivalent to the $\mathcal{C}$-score for $t = 1$ (only the worst node is considered). The $\mathcal{C}$-score still makes sense when computational constraints are of importance: computational complexity grows linearly for the $\mathcal{C}$-score and quadratically for the $\mathcal{B}$-score, meaning the more stringent test is preferable for very large networks. As a last note, the algorithm is able to find cores in the input communities: these are called $\mathcal{C}-q$ or $\mathcal{B}-q$ cores. Based on a user-selectable significance level, for each community the algorithm returns its largest core having a $\mathcal{C}$-score

[10] The number of edges towards other nodes within the community.

[11] I.e. same number of nodes, nodes with the same degree sequence and same internal connections.

[12] Here, the worst nodes in the community are defined iteratively as the vertices with the highest value for $r_i = \sum_{q=k_i^{\text{int}}}^{k_i} f(q)$, where $f(\cdot)$ is given by the hypergeometric distribution over k_i^{int} given in eq. 1 of Lancichinetti et al. [2010].

(respectively, $\mathcal{B}$-score) smaller than q (if any). This is useful to verify if some of the input communities which do not show significance under this framework do instead contain smaller significant cores.

I downloaded the C++ code for the statistical tests in Lancichinetti et al. from the corresponding author's website[13]. I set the q cores threshold to 0.05 as suggested in the paper. The results (C = 0.70, B = 0.57) indicate no significance for community B. In order to evaluate if any of the other communities would pass the significance test, I then repeated the analysis for all the clusters in Figure 4.5. Global results are in Table A.5, Page 193. Communities having a $\mathcal{B} \leq 0.05$ and considered significant based on the thresholds in Lancichinetti et al. [2010] are indicated by an asterisk close to the corresponding cluster in Figure A.6, Page 191. Additionally the only community with $\mathcal{C} \leq 0.05$ is evidenced by a grey rectangle in the same figure.

The results obtained for Community B suggest that we cannot rule out the possibility that a cluster having the same properties of Community B[14] can be found in a random network. As regards the rest of the partition, a total of ten communities satisfy the criterion for significance of the more relaxed B-score (Figure A.6, significant communities smaller than 5 nodes not shown) while only one spoke-like group of 8 genes satisfies both scores at the 0.05 level (Figure A.6, shaded cluster).

The lack of significance for community B is quite interesting as it does not seem to support the other evidence for significance obtained with the basic analysis discussed before. Taken together, the results seem to suggest that community B, while definitely a community because the internal nodes are more connected to themselves that with the outside (Mann-Whitney U-test), it is not however a special community, in the sense that a topologically optimal cluster with similar characteristics could have been found in degree-conserving randomisations of the original network. Interestingly though, all clusters having low B-score in Figure A.6 seem to have similar topological traits, namely at least one hub of very large degree and dense radial neighbour organisation (a "spoke-like" motif not unlike those observed after performing MCL-based clustering of `NET_CATOT4_union`, Figure A.5). Most of the clusters identified as significant through these scores do no seem to be describing clearly related biological processes or well-defined molecular complexes. It should be noted that the approach by Lancichinetti et al. is quite recent and very little biological validation data is discussed in the related paper. Therefore, I cannot exclude the possibility of systematic bias in the predictions of the software and the algorithm. Some of the assumptions in Lancichinetti et al. might not hold in particular domains — for example, the justification given to the null model employed is somewhat lacking and alternative null models are not described. An additional explanation for the result is

[13] `sites.google.com/site/andrealancichinetti/software`

[14] By "properties" Lancichinetti et al. mean 'same number of nodes, nodes with the same degree and same internal connections'.

that NET_CATOT4_union is only a subset of the full interactome. Missing protein and missing interactions could mean loss of local structure potentially resulting in biologically interesting communities becoming not significant in topology-based studies. A final possibility is that the greedy modularity optimisation algorithm used to obtain the clusters [Wakita and Tsurumi, 2007] uses optimisations resulting in slightly different communities than those modelled in Lancichinetti et al. This would imply that the clusters tested could be sub-optimal, resulting in absence of significance in some cases: the high number of statistically significant $\mathcal{B}$ cores (25 clusters at the $q = 0.05$ significance, Table A.5, Page 193) indicates the presence of significant optimal cores within many of the identified clusters and seem to support this hypothesis.

The fact that so many Fd3F targets are connected in one community is quite remarkable and represents inconclusive evidence that target genes of a transcription factor might encode proteins that functionally interact. In the two tests discussed above I evaluated over-representation of interactions within a cluster, compared to the rest of the network. I showed that community B is richer in interactions within itself, compared with the outside world. In a last test, I looked instead at over-representation of interacting Fd3F targets in this community. I set out to evaluate how likely it is to observe 5 interesting genes (in this case, Fd3f target genes) in a cluster of the size of community B given the total number of genes in the network and the total number of interesting genes. The probability of obtaining k successes in n draws from a population of size N containing m total successes without replacement was assessed through a hypergeometric test. The null hypothesis H_0 is that the observed number of Fd3F-regulated genes (successes) in the observed number of draws (genes in the community) from a population (network) containing m total Fd3F-regulated genes can be obtained by chance. The result ($p = 4.91E-05$) is significant at the 1% level and H_0 is rejected. From this test it can be concluded that, given the size of NET_CATOT4_union at 1% significance level a community of the size of community B with such a high number of Fd3f targets compared to the total is very unlikely to be obtained by chance alone.

The overall interesting results obtained from these tests prompted us to continue work on this community. I next looked at the functional evidence and literature annotation available for the genes in the community.

4.4.1.2 Functional Annotation Survey

Provenance information for the protein interactions captured in community B is in Table 4.4 (for the experimental interactions) and Table 4.5 (for the putative interactions). As regards the implicated interactors, I discussed evidence for a role in ciliogenesis for CG31320 in Chapter 3, where I also provided an overview of experimental work aiming to validate the computational predictions obtained. Like most of the genes in Table 4.1, CG31320 has a fairly conserved binding motif combination (X+F) within 100 bp of the transcriptional start site, and is reg-

Interaction	Publication	Detection Method	Interaction Type
CG14905-CG9911	[Giot et al., 2003]	Y2H	PA
CG14905-Cap	[Giot et al., 2003]	Y2H	PA
CG14905-CG8248	[Giot et al., 2003]	Y2H	PA
CG14905-nimC1	[Giot et al., 2003]	Y2H	PA
CG14905-CG13458	[Giot et al., 2003]	Y2H	PA
CG14905-CG15071	[Giot et al., 2003]	Y2H	PA
CG16984-CG13458	[Giot et al., 2003]	Y2H	PA
CG16984-spz3	[Giot et al., 2003]	Y2H	PA
CG16984-And	[Giot et al., 2003]	Y2H	PA
CG16984-CG11327	[Giot et al., 2003]	Y2H	PA
CG16984-CG12470	[Giot et al., 2003]	Y2H	PA
CG11253-yl	[Giot et al., 2003]	Y2H	PA
CG11253-Hsp68	[Giot et al., 2003]	Y2H	PA
CG11253-tilB	[Giot et al., 2003]	Y2H	PA
tilB-CG5189	[Giot et al., 2003]	Y2H	PA
CG6793-CG12470	[Giot et al., 2003]	Y2H	PA
CG6793-Rlip	[Giot et al., 2003]	Y2H	PA
CG32392-CG7135	[Giot et al., 2003]	Y2H	PA
CG32393-CG15440	[Giot et al., 2003]	Y2H	PA
CG32392-sina	[Giot et al., 2003]	Y2H	PA
CG32392-CG9911	[Giot et al., 2003]	Y2H	PA
CG13893-CG11327	[Giot et al., 2003]	Y2H	PA

Table 4.4: Evidence for the experimental interactions in community B (Figure 4.6, Page 142). PA is `physical association`.

Interaction	Source Taxon	Publication	Detection Method
CG31320-Unc-76	Hsap	[Wanker *et al.*, Cell 2005]	two hybrid pooling
CG31320-Ciao1	Hsap	[Figeys *et al.*, Mol Syst Biol 2007]	anti-bait coip
CG31320-dm	Hsap	[Figeys *et al.*, Mol Syst Biol 2007]	anti-bait coip
CG31320-PRL-1	Hsap	[Figeys *et al.*, Mol Syst Biol 2007]	anti-bait coip
CG31320-MSBP	Hsap	[Figeys *et al.*, Mol Syst Biol 2007]	anti-bait coip
CG31320-Bin1	Hsap	[Figeys *et al.*, Mol Syst Biol 2007]	anti-bait coip
CG31320-CG9601	Hsap	[Figeys *et al.*, Mol Syst Biol 2007]	anti-bait coip
CG31320-Nc	Hsap	[Figeys *et al.*, Mol Syst Biol 2007]	anti-bait coip
CG31320-Arf51F	Hsap	[Figeys *et al.*, Mol Syst Biol 2007]	anti-bait coip
CG31320-bun	Hsap	[Figeys *et al.*, Mol Syst Biol 2007]	anti-bait coip
CG11253-bun	Hsap	[Vidal *et al.*, Nature 2005]	pull down
CG11253-bun	Hsap	[Vidal *et al.*, Nature 2005]	two hybrid pooling
Cdk5alpha-Cdk5 (**S**)	Hsap	[Hisanaga *et al.* Biochem Biophys RC 2003]	protein kinase assay
Cdk5alpha-Cdk5	Hsap	[Musacchio *et al.* J Med Chem 2005]	x-ray crystallography
Cdk5alpha-bun	Hsap	[Vidal *et al.*, Nature 2005]	two hybrid pooling

Table 4.5: Evidence for the putative interactions in community B (Figure 4.6, Page 142). Hsap is *Homo sapiens*. **S**: spoke inference.

ulated cooperatively by Rfx and Fd3F [Newton et al., 2012]. Its human homologue, *HEAT repeat-containing protein 2* (HEAT2) contains 10 HEAT repeats (tandemly repeated, 37-47 amino acid long modules occurring in a number of cytoplasmic proteins) and had so far been implicated in intracellular transport processes, although for some HEAT proteins there is prior evidence of a role in Huntington's disease [Andrade and Bork, 1995]. Recently, clinical evidence of a family with a mutation in HEAT2 has been found [Jarman et al., manuscript in preparation]. The family has Primary Ciliary Dyskinesia (PCD), an autosomal recessive genetic disorder affecting motile cilia [Chodhari et al., 2004]. Around 90% of individuals with PCD have ultrastructural defects affecting proteins in the outer and inner dynein arms [Zariwala et al., 2007]. This phenotype adds further evidence to the evidence built in our lab for the role of CG31320 in ciliogenesis: Fd3F regulates cilia motor genes and genes implicated in delineating the compartments, and CG31320 is one of them.

The gene *touch insensitive larva B* (*tilB*) has been linked to cilia construction and maintenance [Kavlie et al., 2010] and its homologues have been implicated in a number of ciliary-related processes. Its human and mouse homologue *Leucine-rich repeat-containing protein 6* (LRRC6) has a testis-specific expression pattern [Xue and Goldberg, 2000], while a study in *Trypanosoma brucei* [Morgan et al., 2005] found that the gene homologue in this organism, TbLRTP, localises in the basal body and is implicated in basal body duplication. The most interesting information comes from a zebrafish study [Serluca et al., 2009], showing that the protein plays a crucial role in regulating cilia motility: it is required for assembly of the axonemal dynein motors in the cytoplasm before their transport into the cilia. Additionally, polycystic kidney disease onset has been observed in mutants [Kishimoto et al., 2008].

The other three seed genes in community B have varying degrees of implication in cilia specialisation. Prior to the series of experimental evaluations currently being carried out in the lab, what was known about CG16984 came exclusively from functional transfer: CG16984 is the fly homologue of enkurin (ENKUR), a mouse TRP-C channel protein which seems to be required in sperm motile cilia [Sutton et al., 2004]. The transcriptome analysis done in the lab [Cachero et al., 2011] shows 13-fold over-expression of the gene in cato-expressing embryonal cells during the 4th time point (Table A.6, Page 196). Its expression pattern was checked by *in situ* hybridisation, showing that it is expressed exclusively in Chordotonal neurons. This suggests it might be associated with motile cilia, and therefore be an Fd3F target gene. CG16984 transcript has then been shown to disappear in *fd3F* mutant embryos [Jarman et al, manuscript in preparation]. No phenotypic data for CG16984 exists at the time of writing — due to the absence of a suitable RNAi line, however this is matter of current research in the lab. CG14905 is the fly homologue of the *Outer dynein arm docking complex protein* (ODA1) in *Chlamydomonas reinhardtii*, the single celled biflagellated green alga where most initial discoveries on the mechanism of assembly of cilia have been done [Takada et al., 2002].

As was the case for CG31320 (Chapter 2), one of these hypotheses, CG16984, has been experimentally identified as an Fd3F target based on this protein interaction analysis. Another one, CG14905, was being studied due to promising evidence making it a good potential target, and this analysis provided further elements to substantiate the decision of pursuing this gene further.

Information regarding the linking genes in the community is sparse. For CG13458 only electronic-derived annotation is available, while CG13893 is the human homologue of *SEC14-like protein 2*, required for transport of secretory proteins from the Golgi complex. *Nedd2-like caspase* (Nc), a putative interactor of CG31320 (from a human projection, Table 4.5) is ubiquitously expressed in embryos during early development and is dramatically enriched in the salivary glands and midgut of late third-instar larvae [Dorstyn et al., 1999] before tissue apoptosis. It has been shown to play a role in sperm differentiation in *Drosophila* [Arama et al., 2003] while there is evidence for involvement in epidermal differentiation of one of its mouse orthologues [Kuechle et al., 2001]. The role of *bunched* (also known as *shortsighted*) in the differentiation of the eye imaginal disc is well-established [Treisman et al., 1995] and recent evidence paints a more global picture where the gene codes for a factor required for determining proper dorsal cell fates leading to the formation of the dorsal appendages [Carreira et al., 2011].

Overall, available functional annotation suggest a degree of homogeneity in the functional roles of the proteins in the community. While this observation had initially not been supported statistically (I will present a statistical functional annotation analysis in Section 4.4.3) the agreement between the protein interaction evidence, the transcriptional data in Cachero et al. and ongoing work in the lab motivated further work on some of the members of the community. Importantly, the fact that tilB appears to be required for assembly of the axonemal dynein motors in the cytoplasm before their transport into the cilia allows to implicate the other members of community B in the same biological process as a testable hypothesis. Due to its association with tilB in the community, CG11253 has been considered a very good candidate to be involved in axonemal dynein assembly and is being followed up in experimental work in the lab.

4.4.1.3 Experimental Validation of CG11253

One of the five Fd3F-related genes in community B, CG11253, is the object of an analysis currently being carried out in the lab by Daniel Moore. Some of his findings will be summarised in the next paragraphs and in Figure 4.7.

The gene was part of an initial selection of candidates prioritised through transcriptome analysis [Cachero et al., 2011]. The list was further restricted to genes featuring relevant gene ontology annotation, or having orthologues featuring relevant annotation. From an experimen-

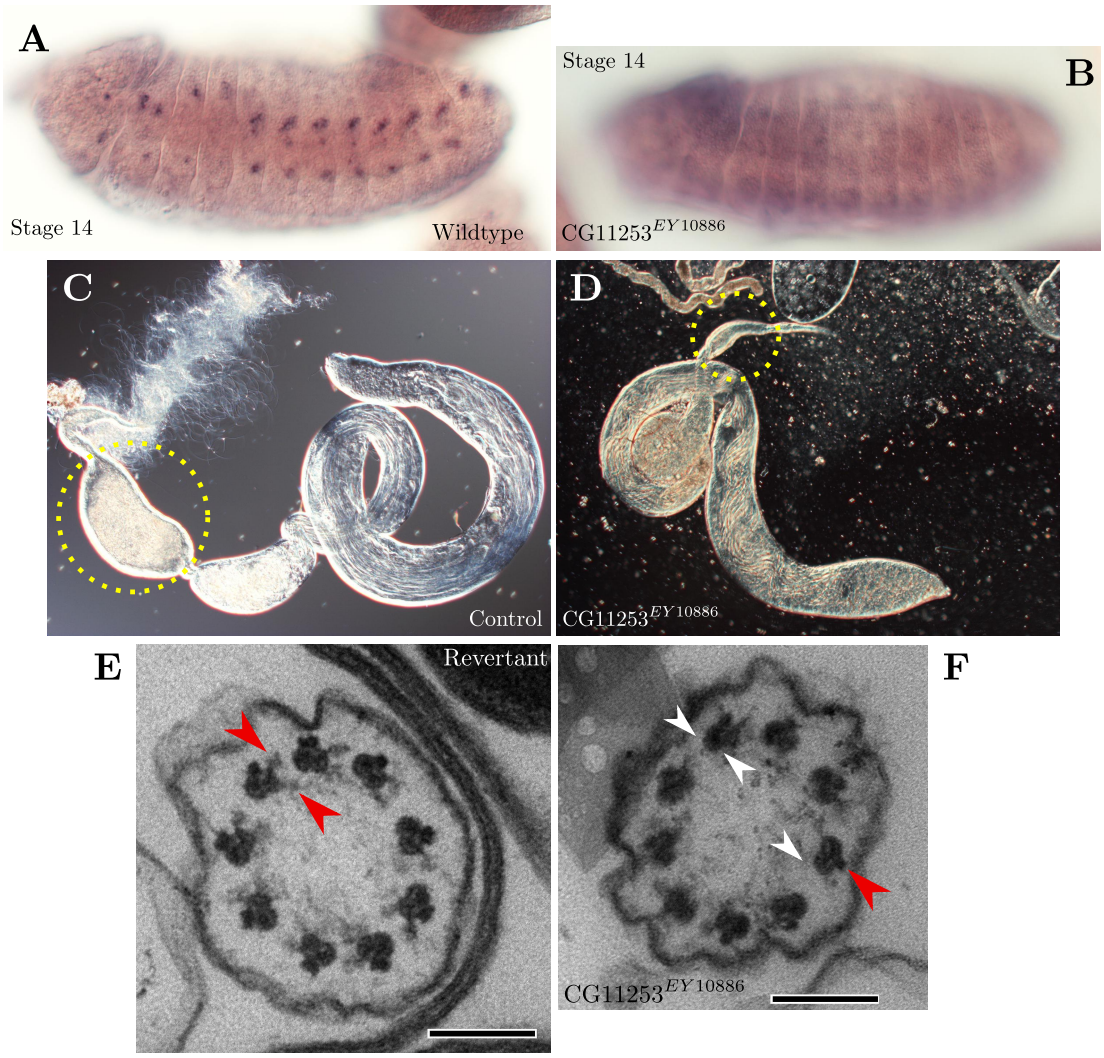

Figure 4.7: Experimental analysis of the ciliary motility gene CG11253 (Zmynd10). **A-B**: *In Situ* hybridisation shows expression knock-down in mutant obtained through p-element insertion (**B**) with respect to control (**A**). **C-D**: CG11253 mutant males are infertile with immotile sperm. In the control (**C**) the large seminal vesicle is clearly visible (yellow circle). In the mutant, (**D**) the seminal vesicle is empty and sperm are immotile. Sperm bundles appear to have formed normally in the mutant. **E-F**: axonemal dynein arm disruption in $CG11253^{EY10886}$. **E**: control (revertant). Inner and outer dynein arms present. **F**: $CG11253^{EY10886}$. Dynein arms disrupted or missing. Red arrows indicate presence of a dynein arm, white arrows indicate absence of dynein arm.

tal viewpoint, this well annotated candidate set was initially screened using a climbing assay to select genes whose disruption (obtained through P element insertions and genetically induced RNAi) would result in proprioceptive and gravitactic deficiency, indicating compromised chordotonal neuron function. 40 RNAi lines were assayed, covering 28 genes. This yielded 10 hits with statistically significant climbing deficiencies (data not shown). CG11253 was chosen based on the severity of the deficiencies highlighted by this assay, while participation in protein interactions and, specifically, participation in the ciliary motility community enriched in Fd3F targets represented another piece of evidence. Additionally, evidence proving it is a Fd3F target had been found in the lab [Newton et al., 2012]. However, additional work was needed to verify whether it would also have a role in determining ciliary motility. An *in situ* hybridisation confirmed that the gene is expressed only in regions containing the two motile ciliated cell types, chordotonal neurons and testis (Figure 4.7-A) and P-element insertion resulted in reduced expression of the gene (Figure 4.7-B).

Comparative analysis also gave promising cues: CG11253 is conserved among most ciliated eukaryotes, and its mouse orthologue, *Zinc finger MYND domain-containing protein 10* (Zmynd10, 33% identity, with 8/10 cysteine/histidines in the zinc finger domain being conserved) has been shown to be enriched in mouse tissues containing motile cilia [McClintock et al., 2008]. Phylogenetic distribution analysis additionally showed that CG11253 co-occurs with axonemal dyneins in ciliated organisms similarly to what shown for the gene tilB by Kavlie et al. [2010] (data not shown). The gene interacts with tilB in community B (Table 4.4, data from a HT Y2H screen in Giot et al. [2003]), which prompts a biological hypothesis for a role of the gene in the assembly of axonemal dyneins. The expression profile of the gene was found to be similar to ciliary genes known to be involved in Primary Ciliary Dyskinesia patients [Geremek et al., 2011].

In a second group of tests, the role of CG11253 for ciliary motility in sperm cells was assayed. Results show that CG11253^{EY10886} mutants are infertile: their seminal vescicles are empty and no motile sperms are observed following a testis squash assay (Figure 4.7-C,D), suggesting that CG11253 is crucial for sperm motility. To evaluate if CG11253 has a role in dynein arm morphogenesis, electron microscopy images of antennal chordotonal cilia were analysed (Figure 4.7-E,F). The results evidence disruption and partial loss of dynein arms, proving that CG11253 is required for normal localisation and structure of the axonemal dynein arms and ultimately for cilium motility. While these results need further validation (in particular, the Y2H tilB interaction will need to be reproduced and electron microscopy of the testis will be performed to further prove the absence of motility in CG11253^{EY10886} mutant sperm) these results show that an approach based on evidence supported by comparative information, transcriptional data and protein interaction-based predictions can lead to relatively small subsets of high quality hypotheses that can be selected for wet-lab based evaluation.

4.4.2 Term Enrichment Analysis

As a preliminary survey of publicly available functional information for the proteins in NET_CATOT4_union, I analysed available Gene Ontology (GO) functional annotation [Ashburner et al., 2000] for the genes in community A and community B. I did this to see if statistical analysis of curated functional data would offer additional insight into the function of the two communities (compared to the information obtained through manual research and discussed in the previous sections).

The test evaluated whether any GO terms are over-represented in the communities with reference to the full NET_CATOT4_union network. This was meant to gauge the depth of the GO functional annotation available and verify if it can discriminate between the network and the two communities. To evidence any over-representation of terms, I obtained the most recent gene ontology (OBO 1.0, v. 1.1.3384) and *Drosophila* annotation (07/17/2012) from the GO project's website[15]. Then, I tested over-representation through an hypergeometric test: given X genes (i.e. all genes in the community having GO annotation) extracted out of a set of N genes (i.e. all genes in NET_CATOT4_union having GO annotation), I calculated the probability that x or more of the X genes are labelled with a functional category C shared by n of the total N genes. This is the probability of seeing by chance x annotated genes in a cluster of X genes taken from a set of N genes of which n share the annotation C. I used the Bingo software [Maere et al., 2005] to compute the FDR-corrected p-values for any over-represented ($\alpha = 0.05$) GO terms, and carried out the test for both communities and for all the GO sub-domains: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC).

For community B, annotation is extremely sparse and no enrichment test is possible for BP and MF. For CC, the community is enriched with respect to NET_CATOT4_union for the term "cyclin-dependent protein kinase holoenzyme complex" (based on annotation for two genes, *cdk5* and *cdk5alpha*, p = 4.28E-2). Therefore a term enrichment analysis for Community B is inconclusive. Community A is composed of better annotated genes. MF and CC term enrichment is shown in Table 4.6 while the longer list of enriched BP terms is in Table A.7, Page 202. For Community A, term enrichment analysis summarises quite accurately the involvement of the bHLH factors in neurogenesis and the involvement of Enhancer of Split genes in negative repression during the Notch pathway. The results confirm that the genes are significantly more related to nucleic acid binding transcription factor function in processes of morphogenesis and organ development, compared with genes in the rest of the network. Overall, annotation re-confirms my previous observations that this cluster describes interactions between the products of genes implicated in neural fate commitment and co-repressors implicated in transcriptional inhibition to maintain progenitor cells in undifferentiated state through lateral inhibition mediated by the Notch pathway [Kageyama et al., 2007]. The lack of term over-representation for

[15] www.geneontology.org/

GO-ID	p-value	corr p-value	cluster freq	total freq	Description
				Molecular Function	
1071	2.4326E-12	1.5934E-10	19/38 (50.0%)	95/1190 (8.0%)	nucleic acid binding transcription factor activity
3700	2.4326E-12	1.5934E-10	19/38 (50.0%)	95/1190 (8.0%)	sequence-specific DNA binding transcription factor activity
43565	1.4262E-8	6.2277E-7	13/38 (34.2%)	63/1190 (5.3%)	sequence-specific DNA binding
3677	1.9652E-8	6.4361E-7	18/38 (47.4%)	137/1190 (11.5%)	DNA binding
46982	2.7183E-8	7.1221E-7	9/38 (23.7%)	26/1190 (2.2%)	protein heterodimerization activity
46983	2.2764E-7	4.9701E-6	10/38 (26.3%)	42/1190 (3.5%)	protein dimerization activity
982	2.7148E-5	4.2193E-4	5/38 (13.2%)	13/1190 (1.1%)	see note (1)
1078	2.8987E-5	4.2193E-4	4/38 (10.5%)	7/1190 (0.6%)	see note (2)
1227	2.8987E-5	4.2193E-4	4/38 (10.5%)	7/1190 (0.6%)	see note (3)
3676	4.3467E-5	5.6942E-4	19/38 (50.0%)	247/1190 (20.8%)	nucleic acid binding
42803	2.1305E-4	2.3258E-3	5/38 (13.2%)	19/1190 (1.6%)	protein homodimerization activity
42802	2.1305E-4	2.3258E-3	5/38 (13.2%)	19/1190 (1.6%)	identical protein binding
981	4.1620E-4	4.1940E-3	6/38 (15.8%)	33/1190 (2.8%)	sequence-specific DNA binding RNA polymerase II transcription factor activity
8134	1.0643E-3	9.9588E-3	6/38 (15.8%)	39/1190 (3.3%)	transcription factor binding
43492	2.9704E-3	2.2890E-2	4/38 (10.5%)	20/1190 (1.7%)	ATPase activity, coupled to movement of substances
16820	2.9704E-3	2.2890E-2	4/38 (10.5%)	20/1190 (1.7%)	see note (4)
42626	2.9704E-3	2.2890E-2	4/38 (10.5%)	20/1190 (1.7%)	ATPase activity, coupled to transmembrane movement of substances
15405	5.0660E-3	3.4929E-2	4/38 (10.5%)	23/1190 (1.9%)	P-P-bond-hydrolysis-driven transmembrane transporter activity
15399	5.0660E-3	3.4929E-2	4/38 (10.5%)	23/1190 (1.9%)	primary active transmembrane transporter activity
5515	7.6159E-3	4.9884E-2	18/38 (47.4%)	333/1190 (28.0%)	protein binding
				Cellular Component	
5634	1.0909E-5	1.0255E-3	23/34 (67.6%)	334/1064 (31.4%)	nucleus
43231	2.0803E-4	6.5183E-3	26/34 (76.5%)	486/1064 (45.7%)	intracellular membrane-bounded organelle
43227	2.0803E-4	6.5183E-3	26/34 (76.5%)	486/1064 (45.7%)	membrane-bounded organelle
33179	9.9201E-4	1.8650E-2	2/34 (5.9%)	2/1064 (0.2%)	proton-transporting V-type ATPase, V0 domain
220	9.9201E-4	1.8650E-2	2/34 (5.9%)	2/1064 (0.2%)	vacuolar proton-transporting V-type ATPase, V0 domain
43229	1.3454E-3	1.8717E-2	29/34 (85.3%)	642/1064 (60.3%)	intracellular organelle
43226	1.3938E-3	1.8717E-2	29/34 (85.3%)	643/1064 (60.4%)	organelle
33177	2.9163E-3	3.4266E-2	2/34 (5.9%)	3/1064 (0.3%)	proton-transporting two-sector ATPase complex, proton-transporting domain
5667	4.2337E-3	4.4218E-2	4/34 (11.8%)	22/1064 (2.1%)	transcription factor complex

Table 4.6: GO MF/CC Term Enrichment for Community A. (1) RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity. (2) RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription. (3) RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription. (4) hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances.

community B confirms the limited depth of the GO annotation: what sets apart community B from the rest of the network is too specific to make it stand out, from an annotation point of view, from the rest of the network.

In the next section I will propose an additional set of tests based on the analysis of publicly available annotation. These tests go beyond basic term enrichment evaluation: they study the relationship between available annotation and network topology. Looking for terms enriched in the communities with respect to the full network, as done in this section, does not take into account the network structure and the actual protein interactions. Instead, the physiological relevance of pairs of protein based on any annotation they share will now be assessed.

4.4.3 Annotation Similarity Analysis

In this section, I will test the consistency of functional annotation in relation to the protein interactions in the communities of `NET_CATOT4_union` (Figure 4.5). The rationale is that protein interactions in the same community should have more related functional annotation than protein interactions across communities. Proteins that interact are likely to be in similar locations or involved in similar biological processes compared to those that do not interact. If the obtained communities are meaningful, there should be a level of correlation between community membership and functional annotation of its protein interactions, whereas a meaningless community assignment would produce groups of interactions with unrelated or conflicting annotations. This would suggest poor interaction to community assignment. A clustering method producing topologically interesting but mostly functionally meaningless communities would be of limited interest and its results should be used with caution.

To see how pairs of interacting proteins map to functional annotation, one can again use information in the Gene Ontology [Ashburner et al., 2000]. The approach of course has drawbacks. As shown in the term enrichment analysis, GO functional annotation is far from complete, and additionally the information in the Gene Ontology is biased at several levels. Blind, exclusive reliance on GO functional information has led to erroneous conclusions (as reported, for example, by Thomas et al. [2012]). However, GO information can still be used to formulate open world hypotheses: if annotation is available, something can be stated about interaction to function mapping within a community; if no annotation is available for a group of interacting proteins, nothing can be stated about in terms of functional consistency — we cannot tell if these proteins are related via biological processes that have not been described or if on the other hand they are not functionally related at all.

4.4.3.1 A Semantic Similarity-based Approach

The first method I employed is based on the usage of semantic similarity within the Gene Ontology hierarchy [Jain and Bader, 2010]. This builds on a fairly well established class of

Interactions	**BP**	**MF**	**CC**
Scored (% total)	850 (41)	944 (45.5)	800 (38.6)
1-**scoring**	11	15	1
0-**scoring**	104	456	134

Table 4.7: TCSS – Results for `NET_CATOT4_union`. First row: total interactions for which a semantic similarity score was found. Second row: total maximum semantic similarity scores. Third row: total minimum semantic similarity scores.

methods using information theoretical metrics to obtain numerical values describing similarity between the annotation for pairs of genes [Resnik, 1995]. The idea of using semantic similarity to explore the information in an ontology is based on a principle similar to the one described in Chapter 2, where I obtained multi-level information within the HUPO PSI-MI ontology to compare experimental method and interaction type annotation for protein interactions obtained through IntAct (Section 2.2.3.4, Page 37). Here, I test the Topological Clustering Semantic Similarity (TCSS) algorithm proposed by Jain and Bader [2010], which improves on other semantic similarity methods by considering unequal depth of biological knowledge in GO in different branches of its directed acyclic graph[16]. The method will score interacting proteins higher if their annotation belongs to the same GO sub-graph (as compared to if the two annotations came from different sub-graphs).

For the evaluation, I used `NET_CATOT4_union` as the input set. I obtained the TCSS python scripts from the developer's website[17] and the most recent gene ontology (OBO 1.0, v. 1.1.3384) and *Drosophila* annotation (07/17/2012) from GO. I discarded GO IEA (Inferred from Electronic Annotations) annotations, and carried out the analysis for the three GO domains available (BP, MF, CC). Some figures about the results are shown in Table 4.7.

Results for the CC sub-domain are sparse, while molecular function and biological process annotation inform a similarly sized portion of `NET_CATOT4_union` (41% and 45.5% respectively). Figure 4.8 is a visualisation of the resulting interaction scores obtained for the BP domain. Communities A and B are again highlighted. Functional annotation for community B is scarce for BP and close to non-existent for the MF and CC domains (data not shown). Due to the open-ended nature of GO annotation, a protein functional similarity analysis involving this cluster is therefore not possible. On the other hand, the proteins in community A are described by highly semantically similar functional annotation in the BP (Figure 4.8) MF and CC domains. This community was discussed earlier because it links together several proneural transcription factors as well as HES factors and has been shown to be enriched for GO terms

[16] For example, the "intracellular" term has more depth than the extracellular term: there are many more biological terms for some concepts with respect to others.

[17] `baderlab.org/Software/TCSS`

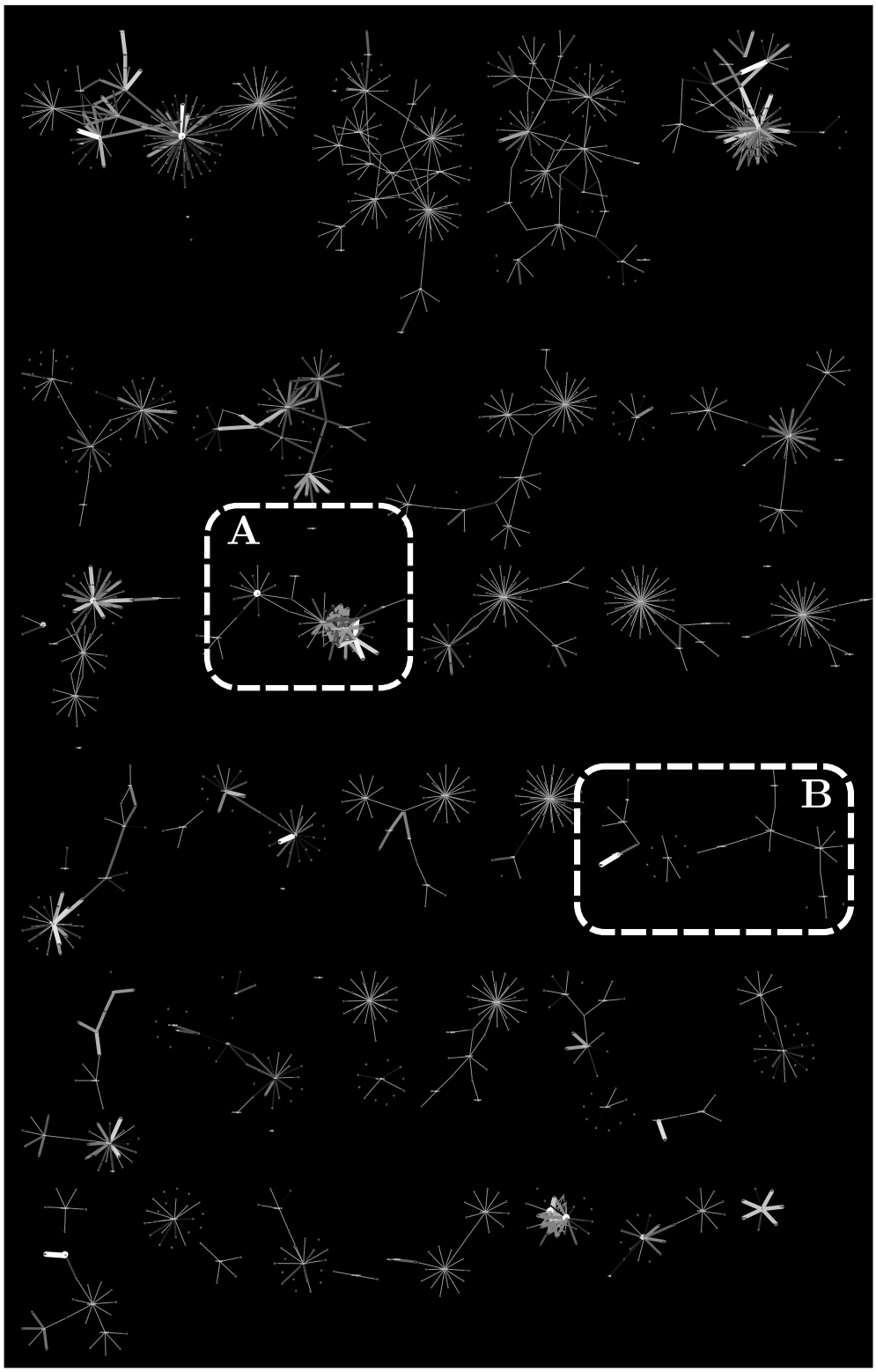

Figure 4.8: Network communities for `NET_CATOT4_union`, same as in Figure 4.5. Additionally, edge width and colour are mapped to semantic similarity score. Brighter and thicker edges correspond to larger similarity scores. No edge indicates no annotation was available. Mapping shown is for GO BP semantic similarity.

describing transcription factor activity, negative regulation and organ morphogenesis.

Since a reasonable amount of annotation for the genes in this community is present, I decided to carry out a statistical test to evaluate if protein functional similarity based on GO is stronger within the community than with the rest of `NET_CATOT4_union`. I collected all the semantic similarity scores for all the interacting pairs within community A. Next, I also collected all semantic similarity scores for all the interacting pairs for which one member is within community A and the other is outside. I then carried out a Mann-Whitney U-test on these two vectors. The null hypothesis H_0 is that the data in the two vectors are independent samples from identical continuous distributions with equal medians. The results ($p_{\mathrm{BP}} = 0.0468$, $p_{\mathrm{MF}} = 0.3989$, $p_{\mathrm{CC}} = 0.0016$) indicate that H_0 is rejected at the 5% significance level for the biological process and cellular component sub-domains. Based on the available data, the null hypothesis cannot be rejected for the molecular function sub-domain.

According to these results and to available GO annotation at the time of writing, the molecules in community A are significantly more related to one another (with respect to the rest of the protein network) in terms of their biological process and cellular component annotation. Based on available data, the genes in the community are not significantly more related to one another (with respect to the rest of the protein network) in terms of their molecular function annotation. One reason for this is probably the lower resolution of the MF annotation which is not granular enough to discriminate different function within `NET_CATOT4_union`, which is a sub-network of the complete interactome obtained by selecting *a priori* functionally related genes. Evidence supporting this conjecture would come from the high number of 0-scores in the semantic similarity MF results (Table 4.7) with respect to the total number of resulting MF scores. These could indicate that the algorithm was not able to build GO tree sub-graphs of unrelated groups of MF concepts, because the MF tree is not detailed enough for *Drosophila*. This possibility seems to be supported by the findings of several studies which employ all three sub-domains (BP, CC, MF) for protein interaction evaluation: in these, while CC and BP have a similar predictive power, the predictive power of MF annotation is considerably lower [Jain and Bader, 2010; Maetschke et al., 2012].

4.4.3.2 A Hybrid Semantic Similarity/Supervised Machine Learning Approach

Having looked at semantic similarity methods to evaluate functional relatedness of protein interactions through GO, I next surveyed another class of approaches to score protein interactions based on common function. These use GO annotation to inform supervised machine learning models, based on the following principle. Using a training set of known protein interactions together with their GO functional annotation, a statistical model is learnt. It is then possible to compute the probability that two proteins interact based on the probability of observing their interaction and related annotation in the model. These probability scores can be used in the

same way the semantic similarity scores are used, i.e. to obtain a numerical estimate of the possibility that a protein interaction exists based on any functional annotation shared by its participants. Supervised learning approaches potentially yield more accurate predictions than purely unsupervised methods like the one in Jain and Bader [2010]. However, unlike Jain and Bader [2010], most of these methods do not account for the multi-level semantic complexity of the Gene Ontology tree and do not exploit the Gene Ontology topology to discover more accurate relationships between terms.

Recently, a class of approaches combining the advantages of semantic and machine learning methods have surfaced. To see if a combined approach would lead to better functional prediction compared to a pure semantic-based one, I tested the hybrid machine learning/semantic similarity method recently proposed by Maetschke et al. [2012].

Given two proteins p_1 and p_2, the approach computes a mapping $(S_1, S_2) \rightarrow S$ of the GO term sets these proteins are annotated with (respectively, S_1 and S_2). The output of the mapping is a new, induced term set S which is then projected onto a binary feature vector. This is used as the input for a standard ML classifier. Maetschke et al. propose several alternative mapping functions and call them *inducers*. In their paper, inducers are benchmarked and ranked on the basis of their performance. In this analysis, I tested the highest-ranking inducer only, known as the ULCA inducer (Up to Lowest Common Ancestor). Figure 4.9 provides an example of the algorithm using an ULCA inducer.

For the analysis, I obtained version 1.02 of the *go2ppi* package from the official repository[18] and again used the most recent GO ontology (1.1.3384) and *Drosophila* annotation (1.222). I used the default machine learning classifier (Naive Bayes) on a training dataset composed of the full `NET_CATOT4_union` network plus GO annotation for all its nodes (BP, MF and CC), stripped out of electronically annotated (IEA) entries. In a separate experiment, I used a much larger training dataset corresponding to the full known *Drosophila* interactome obtained from IntAct. However, with this second training set the program failed to complete on all machines it was tested on[19] due probably to the size and complexity of the dataset and of the related annotation[20]. Back to the original experiment, 5-fold cross-validation yielded an AUC $= 0.72$ which excluded random guessing. The test dataset I used was composed of the full set of proteins in the `NET_CATOT4_union` network. The output of the classifier was a score in the $[0, 1]$ interval which, for every possible interaction between the nodes in `NET_CATOT4_union`, indicated the likelihood of an interaction existing based on functional annotation. I filtered this set to only retain scores for the actual interactions in `NET_CATOT4_union`. Figure 4.10 shows results for community B. These can be interpreted as follows.

[18] `acb.qfab.org/acb/go2ppi`.

[19] Including an 8-core latest-generation workstation kindly made available by the School of Informatics, Edinburgh. An updated release of *go2ppi* with support for parallelisation is available at the time of writing but was not available when the analysis was performed.

[20] The computational complexity of the algorithm grows quadratically with the annotation file size.

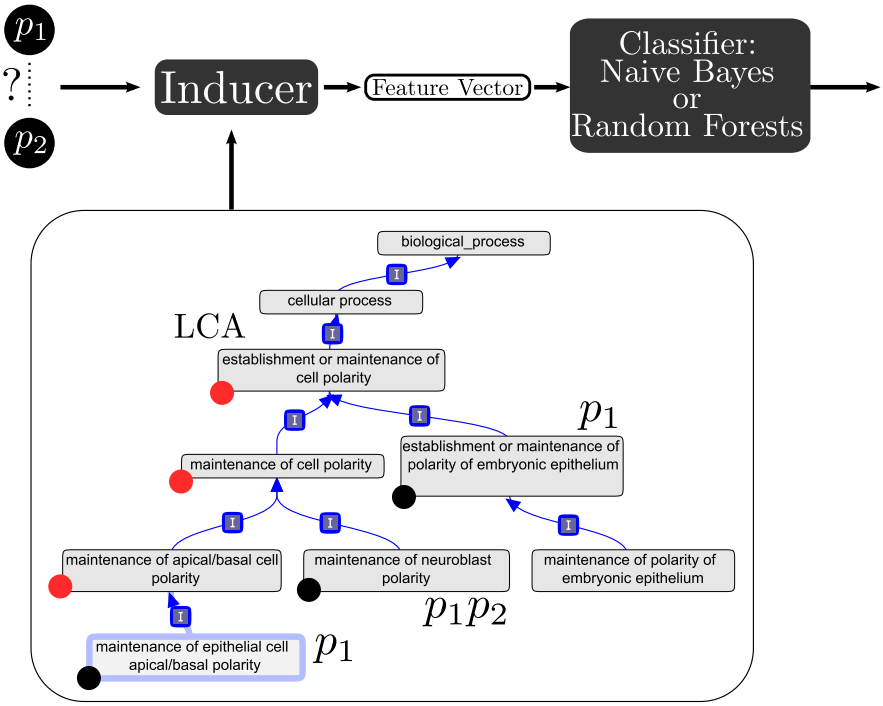

Figure 4.9: Hybrid semantic similarity/ML-based functional mapping using inducers. An inducer processes the Gene Ontology by mapping two term sets S_1 and S_2, assigned to two proteins, onto a feature vector that serves as the input to a machine learning classifier. The particular inducer shown, ULCA (Up to Lowest Common Ancestors) builds the feature vector using the GO concepts p_1 and p_2 are annotated with (black dot) plus all the nodes up to their lowest common ancestors (red dot) (*Adapted from* Maetschke et al. [2012]).

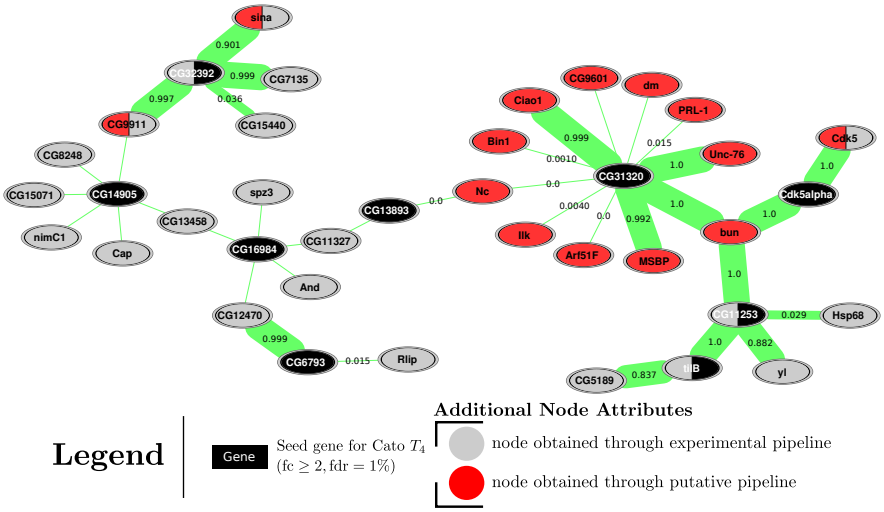

Figure 4.10: Inducer-based Classification - Results for Community B (Figure 4.6, Page 142). The `NET_CATO-T4_union` network and its related GO annotation were the input data for inducer method in Maetschke et al. [2012]. The ensuing model was tested on the node list of `NET_CATOT4_union`. The output was a score indicating, for each possible interaction between each two nodes in the list, the possibility of an interaction happening. In the figure above scores are shown over-imposed on the protein interactions in the set. Edge width is proportional to the corresponding score. Interactions showing no score values link nodes for which no GO functional annotation existed, and therefore no classification could be performed.

Given a model describing pairs of interacting proteins in `NET_CATOT4_union` and their associated functional annotation, new samples are classified based on the annotation they share and on the probability (described by the model) that two training proteins labelled with similar[21] annotation interact. It follows from this that no predictions are possible for proteins having no annotation in GO. This is the case for several nodes in community B (Figure 4.10).

In those instances where functional annotation is available, classification based on annotation seems to be in good agreement with interolog-based interaction data, although correlations between the data informing the two approaches cannot be excluded: it is a possibility that some of the *Drosophila* functional information is projected from other species without being explicitly tagged as such. Keeping this in mind, and assuming labelling errors of comparative annotations in GO are rare, it appears that the *go2ppi* classification identifies small areas in the community characterised by strong functional support.

In particular, a number of putative interolog predictions in community B (radial motif around CG31320) are highlighted by functional annotation: several putative interactions such as those between CG31320 and Ciao1, Unc-76, bun and MSBP (Table 4.5, Page 146) are reconfirmed through this analysis. The classifier's output is also useful to evaluate experimental interactions obtained from IntAct: the experimental interaction between CG11253 and tilB was discussed in Section 4.4.1.3 and represented one of supporting bits of evidence leading to an experimental evaluation of CG11253. The interaction, reported in a large Y2H study [Giot et al., 2003] is independently reconfirmed here. A number of other interactions involving CG31320 are instead strongly penalised.

It is important to remark, however, that a low score between two nodes does not mean their interaction cannot exists, but rather that in the training network used there is little or no evidence supporting an interaction between two proteins annotated with related GO terms. It is a possibility that a much larger training network would have provided a more precise prediction, however as stated this possibility was not tested due to technical reasons.

While a full comparison of this hybrid model versus the purely semantic one has not been carried out, the inducer method clearly provided a higher number of predictions compared to the semantic method, allowing partial analysis of community B (whereas the semantic analysis was not informative on the same community, as shown in Figure 4.8).

I performed a Mann-Whitney U test to evaluate the significance of the functional relatedness of the protein interactions in this cluster. This test compared the functional scores of the interactions within Community B with the functional scores of the interactions between nodes in community B and their neighbours outside the community. The null hypothesis H_0 is that the two groups are independent samples from distributions with equal medians and $\alpha = 0.05$. The results allow to reject H_0 at the α level, both when scores for all edges in community B are

[21] where the similarity is in terms of the UCLA inducer mapping.

considered, including 0 scores assigned for edges lacking annotation ($p = 4.7816E-04$) and when 0 scores are discarded, thus leaving only pure classifier's predictions ($p = 0.0184$). The result of the test indicates that the nodes within the community are linked by edges whose score distribution is significantly different from the score distribution of the edges between the nodes in the community and the outside nodes. This suggests that the meaningfulness of community B is supported by the functional relatedness of the protein interactions in it.

4.5 An Improved Approach to Sub-Network Selection

4.5.1 Motivation

In the network studies discussed so far I have mostly relied on the following approach:

1. Given a species of interest, I build a large network dataset[22] using experimental data mined from IntAct and augmented with putative interaction data obtained with `Bio::Homology::InterologWalk` (using an implementation presented in Chapter 2).

2. In parallel, I identify a small dataset of *interesting* genes: this can be a list of genes linked to one another due to proven evidence of operation in common pathways, similar functional annotation, evidence of co-expression in transcriptome enrichment analyses, and so on. A gene set obtained this way was known in my previous discussions as the *seed* set: a small core of interesting genes which I then mapped to an interaction network like those described in point (1).

3. I obtain all the direct neighbours of the seed set in the interaction network, and extract the sub-network data to carry out further analysis.

I used this approach because it is able to build hypotheses based on unrelated experimental evidence. A source of information is encoded in the seed set itself. A second source of information is represented by the protein interactions involving products of the seed genes and their direct interactors. The reason for this is that some of these protein interactions can *entangle* in a functional characterisation genes that escaped the characterisation that generated the seed set.

For instance, let us suppose the seed set is a list of very highly enriched genes in neuron cells with respect to the rest of an organism. If they are all enriched at a similar developmental time point, they are probably implicated in a shared process. However, genes implicated in the same process which show *ubiquitous* over-expression in the organism (at the same developmental time point) will not be detected by an enrichment study alone. However, a protein

[22] Its upper limit being the full interactome.

interaction study looking at the direct neighbours of the products of the enriched seed genes can detect these ubiquitously highly expressed genes through guilt by association.

The approach is successful in creating principled computational predictions which drive or support lab analysis. Still, the reason why I used only direct neighbours of the seed genes has not been sufficiently justified. Intuitively, there are many reasons for mining only the first neighbours of some seed genes: the resulting sub-networks are often manageable in terms of size, and it could be argued that, in a guilt-by-association perspective, the closest neighbours are "maximally guilty".

4.5.2 Methods

In this section, I would like to briefly introduce a more principled approach to build protein interaction sub-networks which is based on more rigorous statistical foundations. To summarise, the approach "grows" a protein interaction sub-network around seed genes by selecting, instead of their first neighbours, genes which preferentially attach to seed genes rather than other genes. The approach builds on work done by Cerami et al. [2010], who successfully used a statistics driven network attachment algorithm to identify core pathways involved with a brain tumour, the Glioblastoma multiforme. Cerami et al. implemented their method in the free software package called Netbox[23]. While I have not used this software to carry out my analysis[24], some of the ideas in Netbox have inspired my own implementation for the statistical network approach I shall now describe. The basic principle is illustrated in Figure 4.11.

The algorithm is based on generating a sub-network $\mathcal{N} \subset I$ of the full interactome where seed genes are connected either with other seed genes or with non-seed genes which *link* two or more seed genes to one another. The hope is that some of those non-seed genes (those that attach preferentially to seed genes) are biologically informative. Given a seed gene s_i, all first interactors of s_i are obtained. For each interaction, if the corresponding node x is also a seed gene, x is kept in the growing network, that is to say an interaction $(s_i, x = s_j)$ is added to $\mathcal{N}$. If, on the other hand, x is not a seed gene, the search on that path continues and two possibilities exist: a) at least one of the interactors of x is a seed gene[25], s_k. In this case, the path $\rho_x = (s_i, x, s_k)$ is added to $\mathcal{N}$. b) none of the interactors of x is a seed gene[26]: x is discarded and exploration on that path is terminated.

The output of this process is a sub-network $\mathcal{N}$ composed of a set of N seed genes $\mathcal{S} = \{s_i\}_{i=1}^{N}$ and a set of M candidate linking genes $\mathcal{G} = \{g_j\}_{j=1}^{M}$. In order to retain only information-rich linking genes, the algorithm carries out a statistical test for each node $\mathcal{G}$. For each generic

[23] cbio.mskcc.org/netbox

[24] It is written for human interactome analysis and furthermore the author, when contacted for discussions, was unresponsive.

[25] Excluding the original s_i.

[26] Again, excluding the original s_i.

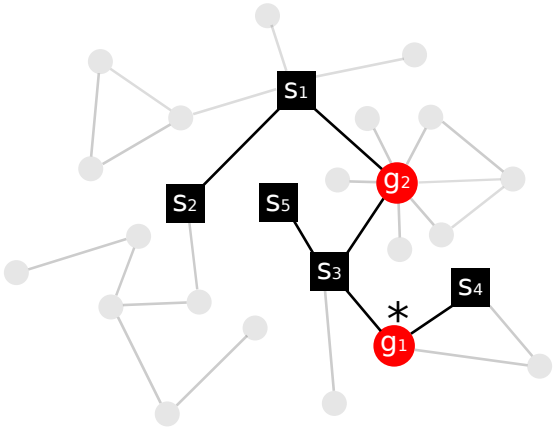

Figure 4.11: An improved method to mine sub-networks from seed genes. $s_i, i \in \{1,2,3,4,5\}$ are seed genes. Starting from the full interactome obtained using `Bio::Homology::InterologWalk`, all first neighbours of the $\mathcal{S} = \{s_i\}$ genes are obtained. For each seed gene, if the first neighbour is also a seed gene, the interaction is kept in the growing sub-network $\mathcal{N}$ (e.g. (s_1, s_2)). If not, a path of length 2 from the seed gene is calculated. If the second interactor is also a seed gene, the path (e.g. $s_1 - g_2 - s_3$) is provisionally kept in $\mathcal{N}$, otherwise it is discarded. Genes in the middle of a path of length 2 between two seed genes (g_1, g_2, red dots) are candidate linking genes. A statistical analysis is performed for each linking gene in the set $\mathcal{G} = \{g_i\}$ to prune out of $\mathcal{N}$ those linking genes and related interactions which are *not* statistically enriched for connections to seed genes. In the example in figure, g_2 would be more likely to be pruned because it attaches to many irrelevant genes in the interactome compared to g_1.

g_i, the test uses its global degree value in the interactome I and the hypergeometric test to assess the probability that g_i would connect to the observed number of seed genes by chance alone. Next, the algorithm applies FDR correction and obtains a new $\mathcal{G}^* \subseteq \mathcal{G}$, composed of all the linking genes which pass the FDR-corrected p-value threshold of 0.05.

Having obtained a sub-network $\mathcal{N}$ composed of N seed genes and of the statistically enriched linking genes, I carried out a series of global and local comparisons of this network with null models. In the global comparison, I compared the size[27] of the largest connected component of $\mathcal{N}$ to the size of the largest connected component obtained by running the algorithm described above using N randomly selected genes from the genome in place of the seed set. I repeated this process 1000 times. I used this test to evaluate if the real sub-network $\mathcal{N}$ was more connected than expected by chance. For the local null model comparisons, I partitioned $\mathcal{N}$ in communities using the NG algorithm and modularity maximisation (both introduced earlier in the chapter). Having calculated the modularity of $\mathcal{N}$, I obtained 10000 degree-conserving randomisation of it and recomputed the modularity for each. This was done to assess the statistical significance of the modularity of $\mathcal{N}$.

[27] I.e. number of nodes and edges

4.5.3 Set-Up and Results

For the analysis, I obtained the *Drosophila* interactome using `Bio::Homology::InterologWalk`, and augmented it using the putative pipeline as described in Chapters 2 and 3. I used the Ensemblgenomes pan-homology database (V.66) to obtain the homology predictions. As for the seed gene list, I obtained it from Cato T_4 data, as before (Section 4.2, Page 134). This means I used the same seed list of 285 genes enriched in Cato T_4 (Table A.6, Page 196). I set the p-value cut-off for the linking genes enrichment test to 0.05. Additionally, I set the number of repetitions for the global null model and the number of degree-conserving randomisations for the local null model to 1000, while I set the number of random-rewiring operations for $\mathcal{N}$ for each network to 20000. Data about the resulting analysis is shown in Table 4.8. 170 of the 285 seed genes participate in interactions, and the total number of linking genes found by the algorithm is 150. Of these, 75% are discarded after the hypergeometric test. These are linking genes for which the probability of connection to the observed number of seed genes (compared to the number of total interactome genes they connect to) by chance alone is over the statistical threshold of 0.05. Data regarding the 37 linking genes which do connect preferentially to seed genes is shown in Table 4.9.

The largest connected component of the final network, $\mathcal{N}$, is composed of 88 genes and 152 interactions and is shown in Figure A.7, Page 192. If we decompose, for clarity, this connected network using the fast greedy Newman-Girvan algorithm [Su et al., 2010] we obtain the communities shown in Figure 4.12.

A few observations can be made about the protein interactions isolated in Figure 4.12. The 10 linking genes marked by an asterisk in Figure 4.12 are the most highly expressed (in an absolute sense) within Cato-expressing cells at T_4 in the transcriptional study performed in the lab [Jarman *et al.*, manuscript in preparation]. Four of these very highly expressed linking genes, *extra macrochaetae* (emc), *daughterless* (da), *E(spl) region transcript mγ* (HLHmgamma) and *E(spl) region transcript m3* (HLHm3) appear in cluster 1, which is clearly recapitulating a group of neurogenesis-related, bHLH-domain co-transcription factors, including proneural factors (ato, da) and negative regulators of the Enhancer of split complex active at the end of the notch pathway to repress proneural identity in the developing peripheral nervous system [Baker et al., 2011]. Two of these linking genes (*da* and *emc*) are ubiquitously expressed throughout the embryo [Chintapalli et al., 2007] and show no enrichment in cato expressing cells (Table 4.9). This means they could not be implicated in a model like the one shown in Figure 4.12 through a pure enrichment study. Here, they have been captured as part of a network of enriched genes purely through computational prediction and through their statistically significant preference to connect to enriched genes, captured by the network building algorithm I have presented.

Community number 3 contains four linker genes. One of these is *shu* (shutdown), the fly

Parameter	Value
Interactome I	
#nodes	8397
#edges	31214
#seeds in list	285
#seeds w/ interactions	170
Network $\mathcal{N}$	
#seeds, pre-FDR-correction	117
#linkers, pre-FDR-correction	150
#linkers, pruned	113
#linkers, survived	37
#total nodes	154
#total edges	163
#nodes, largest connected component	88
#edges, largest connected component	152
Global Random Model	
(mean, 1000 reps)	
#nodes, largest connected component	12.470
#edges, largest connected component	13.184
p-value (nodes)	0.007
p-value (edges)	0.003
Local Random Model	
(1000 trials, 20000 random rewiring ops/network)	
Network Modularity, μ	0.524
Network Modularity, σ	0.016
Network Modularity, observed in $\mathcal{N}$	0.680
Network Modularity, Z-score	9.931

Table 4.8: Cato $T-4$ Network Identification - Results

ID	degree (seed)	degree (global)	*p*-val	*p*-val (FDR)	Gene Name	Abs Expr	FC
FBgn0082598	3	25	0.012	0.049	akirin	7914.659	1.061
FBgn0000575	2	8	0.010	0.042	emc	5085.078	0.849
FBgn0002609	3	14	0.002	0.020	HLHm3	3987.854	1.467
FBgn0000413	4	25	0.001	0.015	da	3369.735	1.114
FBgn0020510	3	18	0.005	0.027	Abi	3139.478	1.150
FBgn0002735	5	12	2.213e-06	0.000	HLHmgamma	3064.929	2.130
FBgn0024196	2	8	0.010	0.042	robl	2643.376	1.210
FBgn0039929	3	16	0.003	0.025	CG11076	2516.967	0.733
FBgn0037718	2	8	0.010	0.042	P58IPK	1999.736	0.952
FBgn0021967	2	6	0.005	0.029	Pdsw	1276.574	0.795
FBgn0032202	3	18	0.005	0.027	CG18619	1114.572	1.068
FBgn0001230	3	8	0.000	0.008	Hsp68	1072.737	1.892
FBgn0003076	2	4	0.002	0.020	Pgm	632.026	0.498
FBgn0040087	7	86	0.001	0.015	p115	591.699	1.055
FBgn0002734	3	16	0.003	0.025	HLHmdelta	573.104	0.965
FBgn0030710	3	15	0.003	0.023	CG8924	549.805	1.280
FBgn0039712	2	8	0.010	0.042	CG15514	479.521	1.083
FBgn0002631	4	15	0.000	0.007	HLHm5	398.158	1.303
FBgn0052179	3	7	0.000	0.007	Krn	343.874	0.981
FBgn0039125	2	6	0.006	0.029	CG5857	342.706	0.607
FBgn0040385	5	30	0.000	0.007	CG12496	122.320	1.937
FBgn0020236	2	7	0.008	0.036	ATPCL	119.973	1.018
FBgn0003401	2	4	0.002	0.019	shu	102.998	0.647
FBgn0004462	5	57	0.005	0.027	Pk17E	101.070	0.911
FBgn0036819	5	40	0.001	0.015	dysb (CG6856)	100.875	0.670
FBgn0036769	2	3	0.001	0.015	Tsp74F	87.262	0.933
FBgn0010433	2	3	0.001	0.015	ato	79.631	1.226
FBgn0029936	7	77	0.000	0.012	CG4617	71.949	0.808
FBgn0003366	2	5	0.004	0.026	sev	30.468	1.277
FBgn0262617 (FBgn0037806)	3	17	0.004	0.026	CG43143	25.758	0.959
FBgn0000022	4	14	0.000	0.007	ac	23.876	0.691
FBgn0262477 (FBgn0052937)	3	14	0.002	0.020	FoxP	23.814	1.174
FBgn0029747	4	13	9.695e-05	0.007	CG5062	19.783	1.004
FBgn0085197	5	55	0.004	0.026	CG34168	15.459	1.048
FBgn0033963	4	18	0.000	0.008	CG12857	16.285	1.038
FBgn0035657	2	7	0.008	0.036	alphaKap4	15.221	1.134
FBgn0263774 (FBgn0033831)	5	63	0.007	0.035	CG43691	12.834	1.057

Table 4.9: Cato T_4 Network Identification - List of the 37 linking genes survived after pruning through hypergeometric test. Pruning is based on FDR corrected *p*-value. The genes are ordered by their absolute expression value in Cato-expressing cells at T_4. Fold change of expression between Cato-expressing and Non-Cato-expressing cells is also shown.

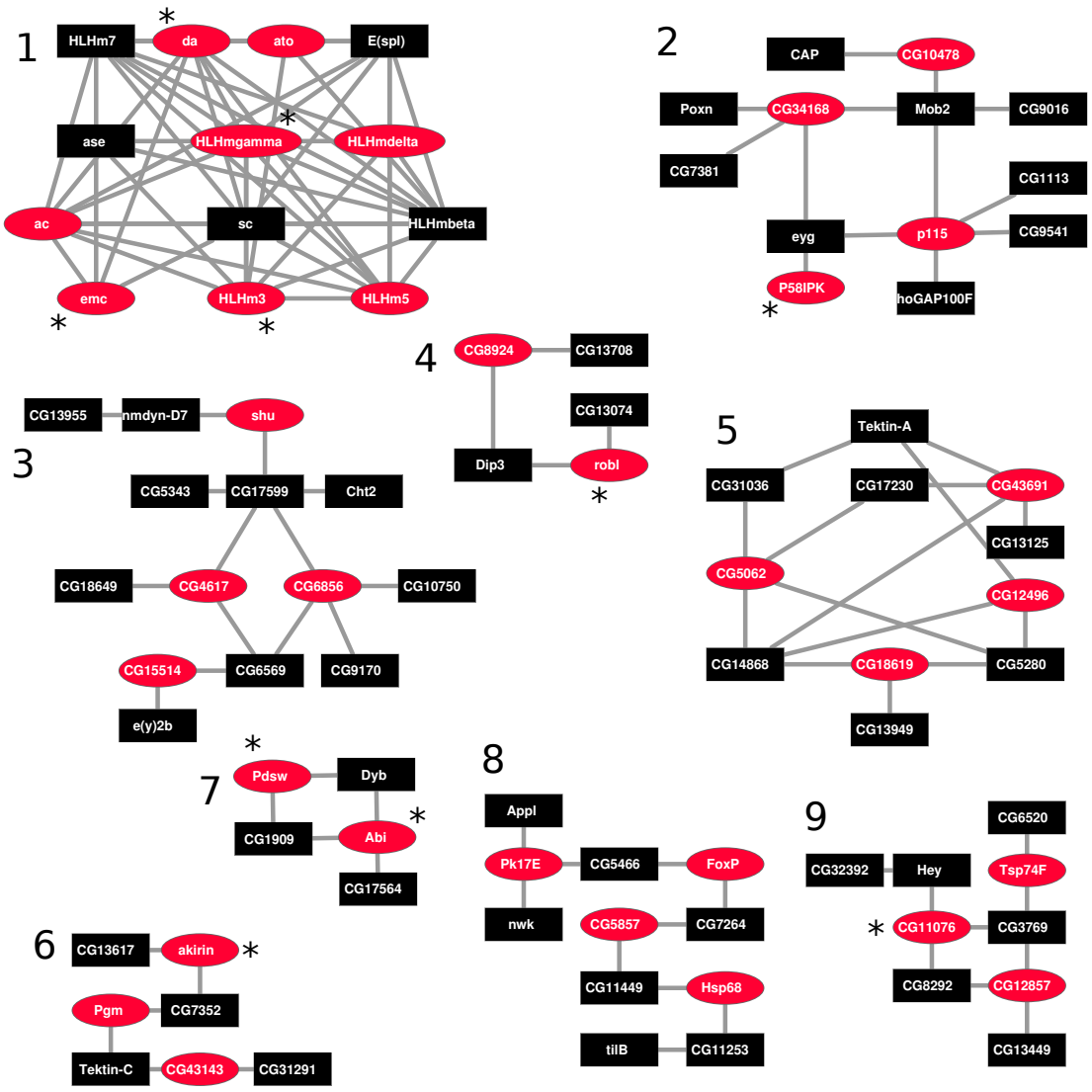

Figure 4.12: Communities obtained from the largest connected component of network $\mathcal{N}$. Black square nodes: seed genes. Round red nodes: linking genes. Asterisks indicate the 10 linking genes showing the highest values of absolute expression in Cato-expressing cells at T_4 in a fly embryo transcriptome analysis carried out in the lab [Jarman *et al.*, unpublished data].

homologue of the FK506 binding protein 6 (FKBP6), essential for homologous chromosome pairing in meiosis during spermatogenesis: mutations in this gene have been related to infertility phenotypes in humans [Zhang et al., 2007]. The gene is a linker for CG17599, one of the genes discussed in Chapter 3 (Section 3.4.2, Page 115). CG17599 is specifically expressed in chordotonal neurons and has been found to a have a role in sensory cilium formation and intraflagellar transport. *shu* also links to *nmdyn-D7*, a gene for which there is high evidence of expression in the adult testis [Wasbrough et al., 2010].

Several other genes in the cluster show evidence for moderate to very high expression in the adult testis: CG6856 (*dysbindin*) is a regulator of synaptic plasticity with moderate expression in adult spermathecae [Chintapalli et al., 2007], CG6569 and CG10750 both show high testis expression and so does the transcription factor *enhancer of yellow 2b* [Chintapalli et al., 2007]. Other genes in the cluster involve CG5343, a transcription factor involved in dendrite morphogenesis [Parrish et al., 2006a] and the homologue of a HMG-box-containing xenopus protein (HMGBX4), known to negatively regulate Wnt/beta-catenin signalling during development [Yamada et al., 2003]. CG9170 is the fly homologue of Centrosomal protein of 164 kDa (CEP164) implicated in primary cilium formation [Graser et al., 2007]. Overall, there is clear evidence that these genes are implicated in related ciliary differentiation processes. The presence of CG17599 and of several genes which are highly expressed in the testis suggest the possibility that the community contains members of a pathway involved in sensory cilium formation in sperm cells.

Overall, the obtained network appears to be able to recapitulate related processes through a combination of functionally annotated linking genes and enriched seeds, while a number of novel candidate linking genes with no known function are also introduced. Some of these could represent good candidates for exploratory studies. Based on the coherence of the communities described, I argue that a combination of transcriptional data together with a principled method to build protein interactions can lead to interesting insights and will increase its usefulness once large protein interaction studies will allow the description of more accurate interaction networks: clearly, a topology-based network algorithm is only as good as the data it is based on.

4.6 Conclusions

In this chapter, I continued the study of PNS differentiation genes initiated in Chapter 3, where I had used data produced as described in Chapter 2 and insights from time course microarray expression data to isolate novel genes with a role in ciliogenesis. The encouraging results obtained by other members of the lab who worked on the biological validation of some the computational hypotheses proposed prompted me to continue working in a similar direction.

Building on the insights gained from the work in Chapter 3, in this chapter I presented a more refined computational workflow and an array of statistical analyses to reach a more precise goal: using bioinformatics approaches to help elucidating some of the processes related with ciliary specialisation in *Drosophila*.

Fd3F has been shown to regulate genes required for the cilium motile segment: it regulates genes encoding Ch-specific axonemal dyneins and TRPV ion channels (required for sensory transduction) and retrograde transport genes (required to differentiate motile and sensory zones in cilia). Ongoing work in the lab is focusing on obtaining a more complete picture of the genes downstream of Fd3F to understand how ciliary structures arise, and what can go wrong during ciliogenesis. This latter point is particularly important because irregularities in ciliogenesis are the cause of a large array of developmental abnormalities known as ciliopathies [Pazour and Rosenbaum, 2002].

In Chapter 2 and 3 I discussed the simple option of obtaining small sub-networks from protein networks based on the selection of few functionally close genes and their first neighbours. Even when a specific subset of the interactome has been selected, the task of identifying interesting bits of information to follow up via experimental investigation can be a daunting one due to the number of interactors involved. A more principled way to isolate groups of related proteins in a network is to first use algorithms to decompose the network in communities, i.e. groups including nodes having more to do with other nodes in the group than with outsiders. I have tested two metrics for community decompositions on the data. One of them, based on the fast greedy Newman Girvan algorithm, provided a decomposition in clusters some of which immediately struck us for their ability to relate proteins known to be involved in related biological processes. I addressed two of these communities in further studies.

The first (named Community A) accurately models a group of several basic Helix-Loop-Helix proneural factors, of bHLH transcriptional co-repressors implicated in neurogenesis segmentation and sex determination and Zinc finger domain proteins. It describes interactions between the products of genes implicated in neural fate commitment within proneural clusters and co-repressors implicated in transcriptional inhibition. This community assignment suggested that the algorithm can capture functionally related groups of genes through network topology analysis alone. The second community (named Community B) stood out for containing five poorly known or putative interacting Fd3F targets. For some of these, other members of the Jarman lab were already building evidence for a role in cilia. One had been implicated in ciliary differentiation after the computational analysis in Chapter 3. Other ones had already a faint trace of functional role in cilia, however this by itself did not make them especially worthy of lab study compared to several other proteins showing similar amounts of functional evidence. The protein interaction evidence presented here shifted the balance in favour of them and motivated lab tests.

A large part of the chapter addresses the necessity to evaluate these communities through several types of significance studies. First of all, I tested the significance of the communities from a purely topological point of view, proving that the density of internal interactions is different from the density of interactions with the background network, which means the Newman Girvan algorithm is producing structurally meaningful partitions.

Using more advanced techniques borrowed from physics research, I assessed the significance of these communities comparing them with the best communities that could have been obtained from degree-conserving randomisations of the same network. This test produced significance for some of the communities in the partition, however not for A and B. I identified as possible reasons for this result the novelty of the testing framework and potential biases (the algorithm is very recent and lacks external validation on biological data at the time of writing), slight technical differences between the community-finding algorithm I used and the one hypothesised by the model or the incompleteness of the network being partitioned. It would be interesting to test this significance algorithm on an experimental dataset including the complete *Drosophila* interactome, using the standard non-greedy Newman Girvan algorithm for community finding.

For the last test of the group, I focused on Community B, and looked at the probability of discovering a community of 5 interacting Fd3F target genes by chance. The results of this test allowed me to say that a community with so many targets cannot be a result of chance alone. The generally encouraging results obtained from these tests prompted me to follow up my evaluations of the communities from a functional point of view.

I utilised the Gene Ontology resource to run a series of functional enrichment experiments. In the first of these, I assessed GO term enrichment. The results provided further evidence for the relatedness of the proteins in the "control" community A. For community B, the tests were inconclusive, due to the scarcity of functional annotation at the depth necessary to discriminate poorly annotated ciliary specialisation roles in community B from background functional annotation in the rest of the network.

Having looked at term over-representation, I next analysed the consistency of the functional annotation available in relation to the protein interactions in the communities. The reason why I did this is that protein interactions in the same community should have more related functional annotation than protein interactions across communities. I used two approaches to score protein interactions based on functional annotations: an unsupervised one and a supervised, machine learning-based one.

Using the first approach, I determined that the protein interactions in community A are described by highly semantically similar functional annotation, and proved that the semantic similarity within the community is different than between community genes and background genes. This showed that, in addition to being topologically plausible, this community is func-

tionally plausible. The first approach was inconclusive for Community B, due to the scarcity of annotation and the open ended nature of GO annotation.

The supervised approach was more powerful and allowed partial evaluation of Community B. One interesting result obtained with this methods is that the test reconfirmed as high-scoring some of the putative interactions in the community. This means that comparative protein hypotheses obtained through interolog projections are reconfirmed through functional annotation similarity prediction. It could be argued that these are two independent sources of biological data. However, I do not exclude the possibility of correlations between the two sources: GO could contain annotations obtained through comparative analysis and not correctly labelled to indicate this, and a separate group of experiments and evaluations would be needed to address this point. Better still, it would be interesting to repeat this test designing an approach using pure sequences (instead of GO annotation) for supervised functional similarity prediction.

An additional result from this evaluation is that the predictions provided by the machine learning-based method for community B allowed me to test the significance of the functional relatedness of its protein interactions, similar to what I have done with community A earlier. The result of the test indicates that the nodes within community B are linked by edges whose score distribution is significantly different from the score distribution of the edges between the nodes in the community and the outside nodes. This suggests that the meaningfulness of community B is supported by the functional relatedness of the protein interactions in it.

The results of these tests provide enough justification to motivate the study of some of the poorly studied genes in the output communities. We selected one of the genes in Community B, *CG11253*, for experimental validation. The gene is being investigated by another member of the lab, Daniel Moore. I briefly summarised his work in one of the sections of this chapter, and presented the evidence he found to implicate the gene with ciliary motility roles. Another gene extracted from the cluster, *CG16984*, has been experimentally identified as an Fd3F target based on this protein interaction analysis. Yet another one, *CG14905*, was being studied due to promising evidence making it a good potential target, and this analysis provided further elements to substantiate the decision of pursuing this gene further.

Taken together, the information emerging about this group of Fd3F targets from community A allows the formulation of a simple speculative model. Fd3F regulates genes needed for the motility apparatus in cilia. One of the 5 targets in community B, *tilB*, is reasonably well characterised: its phenotype is missing motility apparatus in Ch neurons. In zebrafish [Serluca et al., 2009] tilB plays a crucial role in regulating cilia motility because it is required for the assembly of the axonemal dynein motors in the cytoplasm before their transport into the cilia. CG14905 is another promising candidate: its orthologue in *Chlamydomonas reinhardtii*, *flagellar outer dynein arm-docking complex protein 2* (ODA1) is required for motor assembly: it binds outer arm dyneins to the axonemal microtubules in the cilia [Takada et al., 2002].

CG11253 interacts with tilB and is required for the normal localisation of the axonemal dynein arms, from which it follows that it is also necessary for cilium motility. Overall, the simplest model is that all these genes are involved in the ordered assembly of the motility apparatus during cilium differentiation. This model is currently being tested in the lab.

The community-finding experiment has revealed that many Fd3F targets are connected in a single dense group. This suggests the interesting idea that the target genes of proneural transcription factors might encode proteins that tend to interact. The over-representation of interacting targets in community B provides evidence in favour of this hypothesis, however to prove this conjecture a full evaluation using several genomes and high-confidence protein interaction data would be needed.

It follows from these results that the approach consisting in building protein networks based on a set of interesting genes augmented with their first neighbour is, arguably, useful for creating computational predictions which can drive or support lab analysis. However, the reason why only direct neighbours of the seed genes are used had not been sufficiently justified.

Intuitively, there are many reasons for mining only the first neighbours of some seed genes: the resulting sub-networks are often manageable in terms of size, and it could be argued that, from a guilt-by-association perspective, the closest neighbours are "maximally guilty". Still, the approach is not proven to be the best option for the selection of interesting sub-networks from an interactome. Therefore, in the final section I introduced a more principled approach to building protein interaction sub-networks, based on more rigorous statistical foundations and on the ideas proposed by Cerami et al. [2010], who used a similar principle to identify network modules implicated in the brain tumour Glioblastoma Multiforme.

The approach "grows" a protein interaction network around seed genes by selecting, instead of their first neighbours, either other seed genes, or genes which preferentially attach to seed genes rather than other genes, up to a depth of 2 neighbours from the original protein. The algorithm is more specific than the direct selection of direct neighbours because it admits only two kinds of proteins to be added to a network: 1) seed genes 2) "linking" genes statistically enriched for connections to seed genes. This is an ideal framework to pursue the task outlined earlier in this section: to identify potentially important pan-neural ciliogenesis genes which cannot be highlighted by direct analysis of enrichment data in cells expressing proneural genes.

For the test in this chapter, the seed list was a list of genes enriched in cells expressing the Cato proneural factor during late embryonal PNS development. Using the algorithm proposed, I was able to build a network including, alongside these seed genes, 37 linker genes enriched for seed gene connectivity. As expected, these have transcripts which are not enriched in *cato*-GFP cells, however some of them show extremely high values of absolute expression at the same time points. Some of the communities formed by these seeds and linker genes are extremely interesting. The core of Community A, discussed earlier, appears in this analysis as well. I also

found a novel community containing 3 linker genes: this is particularly interesting because most of its nodes, while poorly studied, appear to be highly expressed in testis, suggesting the possibility that the community contains members of a pathway involved in sensory cilium formation in sperm cells.

CHAPTER 5

Conclusions

High throughput technology has had a significant impact on biomedical research, and has led from studies of individual gene function to studies of pathway activity and organismal physiology. The development of large scale technologies and their application to the study of mRNA levels or protein interactions immediately yielded interesting information: it produced large amounts of new types of data, whose analysis has in turn spurred the development of whole new branches of computational research and ultimately pushed biology to a new era.

However, the analysis of microarray and protein interaction data has also taught a valuable lesson: it has led us to realise that "horizontally exhaustive" data collection techniques are not the holy grail. Real increased understanding of biological systems has happened when horizontally exhaustive datasets have been supported by analyses providing "vertical" views on the systems. Unbiased large scale accumulation of data still needs clear biological questions to be posed and clear hypotheses to be formulated, and the data ensuing from these exhaustive screens is only useful when more detailed, small scale studies are carried out to answer particular questions. In other words, I believe that in order to be successful, any high throughput analysis should be seamlessly integrated in a preexisting research agenda so that its results can be sifted to extract relevant insight. As it stands, pure large scale data collection cannot drive biological research.

In addition to the importance of integrating hypothesis-driven and data-driven research, I am convinced that one way to evaluate large, often unspecialised datasets to obtain useful predictions is to cross them with other large datasets. In this thesis, I attempted to suggest a few scenarios for large scale data integration, showing how highly supported "orthogonal" datasets can lead to biological discoveries. In Chapter 2, I proposed the integration of protein interaction data and sequence homology data to build augmented protein interaction datasets in fly PNS sensory neurons. In Chapter 3 and Chapter 4, I showed that these enhanced putative interaction datasets can be sifted to produce highly relevant hypotheses if other large scale data is employed, this time describing transcriptional levels in maturing sensory neurons.

I proposed that a combination of novel computational methods and of insights from other technical scientific disciplines represents one way to drive the "vertical" biological discovery

process that I was referring to earlier. Clearly, my particular approach was driven by the particular data available and by the problem under scrutiny. In this concluding chapter, I will discuss alternative approaches to computational data analysis. This includes speculative discussion on possible scenarios for additional computational work on the same biological problem, as well as alternative methods that could be explored to reach similar goals, given that suitable alternative data sets were at my disposal.

5.1 Further Work

5.1.1 Overlapping Communities and Line Graphs

In Chapter 4 I introduced an approach to the study of protein networks in sensory neurons which employs community-finding algorithms to identify functionally related communities. The results suggest that fast greedy Newman Girvan community-finding provides meaningful functional hypotheses and is ultimately useful as a tool to organise information within a dataset. However, node partitioning algorithms like those I discussed have a drawback: each node is assigned to one community only. This may be an undesirable constraint for networks composed of nodes belonging to highly overlapping or nested communities. This is the case for many social networks [Arora et al., 2012] where individuals typically belong to several communities at once (e.g. school, church, family, workplace), scientific collaboration networks, where individuals doing interdisciplinary work may belong to different research communities [Newman, 2001], and clearly also metabolic networks and protein networks, where molecules can take on different roles or work at the interface of multiple pathways and biological processes [Spirin and Mirny, 2003]. Such "inter-community" individuals (Figure 5.1-A) can provide interesting insights on how the communities interface with one another and on the high level organisation of pathways.

Several algorithms have been proposed to overcome the problem of exclusive group assignment in community-finding algorithms. Given a network G, one way to obtain a partition which allows communities to overlap is to cluster its edges instead of its nodes. This can be done via a *line graph* transformation [Whitney, 1932], which creates a new network $\mathcal{L}(G)$ as follows:

1. Each edge of G is a node of $\mathcal{L}(G)$;
2. Two nodes of $\mathcal{L}(G)$ are connected by an edge if and only if their corresponding edges are adjacent in G.

The idea behind line graphs is sketched in Figure 5.1-B. Finding communities in the line-graph allows the assignment of the original nodes to multiple communities, in the sense that a node is now part of all the communities which its adjacent edges belong to. The approach has attracted

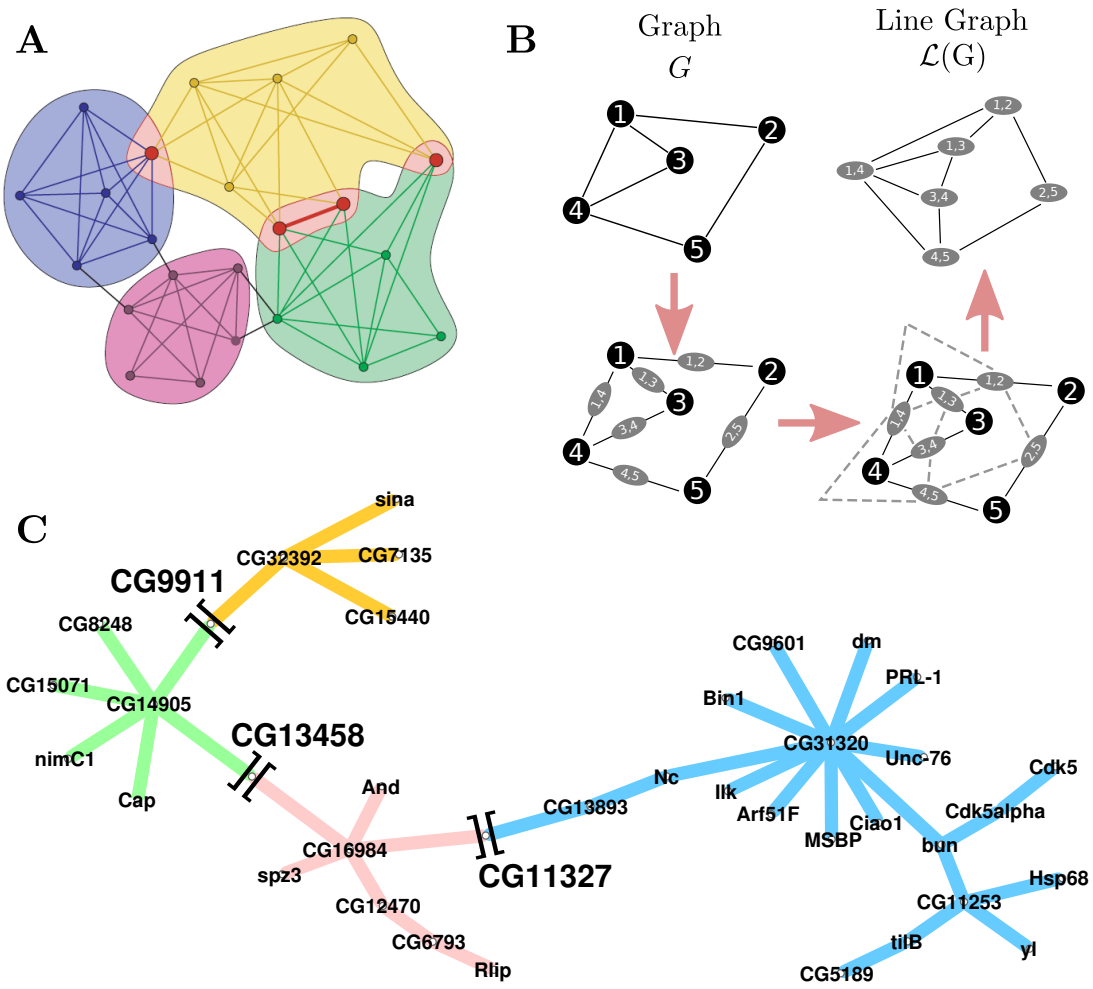

Figure 5.1: Overlapping communities and line graphs. **A** An example of a hypothetical network partitioned in four overlapping communities. Four inter-community nodes are highlighted. (*adapted from* Palla et al. [2005]) **B** Line graph transformation: each edge in G becomes a node in $\mathcal{L}(G)$, and two nodes in $\mathcal{L}(G)$ are connected by an edge if their corresponding edges in G are adjacent. **C**: community B (Section 4.4.1, Page 141). I had obtained Community B via greedy fast Newman Girvan partitioning of `NET_CATOT4_union`. For this visualisation, I derived the line graph transformation for `NET_CATOT4_union` and partitioned it again using the fast greedy NG algorithm. I then remapped the line-graph community assignment on the original community B shown here. The result suggests that line graph partitioning allows for a more fine grained cluster allocation which decomposes the original community in 3 sub-communities, linked by the three inter-community nodes CG9911, CG13458 and CG11327.

the attention of physicists mostly [Evans and Lambiotte, 2009] and applications in network biology are still scarce: some work on this path has been done by Pereira-Leal et al. [2004] and Ahn et al. [2010].

I carried out a preliminary line-graph analysis of the network data in Chapter 4 to illustrate the potential of this approach and to observe any agreements with standard clustering methods, as well as any potential benefits. I derived an unweighted line-graph representation of `NET_CATOT4_union` (Figure A.8, Page 194), which I then partitioned using the greedy fast NG algorithm. Finally, I remapped the ensuing edge cluster assignments onto the original community partition of `NET_CATOT4_union` (Figure 4.5, Page 140). Results for Community B, extensively studied in Chapter 4, are shown in Figure 5.1-C. The line-graph approach partitions Community B in three overlapping sub-clusters. The existence of three inter-community nodes, CG9911, CG13458 and CG11327, suggests the possibility that this kind of analysis might provide higher level of detail compared to a standard clustering approach. It would be interesting to carry out a network-wide analysis of inter-community nodes and look for GO annotation enrichment with respect to intra-community nodes. Another option would be to observe which inter-community nodes are also linking genes (the idea of linking gene has been introduced in the last section of Chapter 4).

5.1.2 Differential Network Analysis

In this thesis, I gained an understanding on some of the dynamics in PNS development by combining evidence from protein interaction networks as well as transcriptional data. The protein networks I discussed share one property: they are static "snapshots" of some underlying — and unknown — dynamic molecular processes. The approach used in Chapter 3 and 4 sought to extract, from static protein interaction networks, some molecules that appeared to be active under the experimental conditions for the transcriptional data.

However, static protein network datasets cannot provide, by definition, information on how the protein interactions occur in time. They cannot reveal the sequence organisation of molecular interaction events. Consequently, it is not possible to use protein networks alone to understand how molecular interaction activity changes as a consequence of environmental and genetic changes. Additionally, we cannot use static network information to identify molecules, interactions or pathways that are condition-specific.

In order to address these shortcomings, differential protein interaction studies have been proposed [Ideker and Krogan, 2012]. These are based on the idea that multiple condition-specific large scale protein interaction datasets could be obtained in a given species. The ensuing network datasets would then be termed *dynamic protein interaction networks*. A differential study would explore the quantitative differences between the condition-specific networks composing one such dynamic network (Figure 5.2). One example of the usage of dynamic network

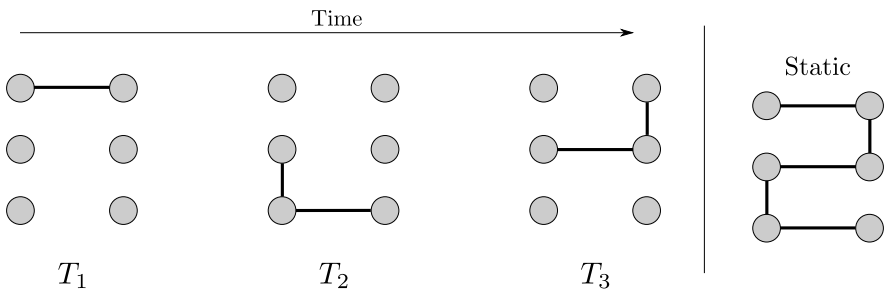

Figure 5.2: Sample temporal graphs and corresponding collapsed static graph.

is in Bisson et al. [2011], who investigate the kinetics of protein interaction networks looking at how they assemble and dissolve to generate specific cellular responses. Using data collected via a novel mass spectrometry method called Affinity Purification–Selected Reaction Monitoring (AP-SRM), the study is successful in elucidating context specific and time-dependent formation of complexes comprising 90 proteins in Human Embryonic Kidney HEK293T cells.

A number of approaches have started to surface to extend to dynamic network data the topological network analysis techniques consolidated for static interaction networks. For instance, Tang et al. [2010b] propose an extension to the concepts of characteristic path length and small world behaviour for the case of dynamic protein networks. Specifically, they introduce the notion of temporal path (a quantification of the temporal distance between the nodes) and a measure of time persistence for the interactions. These insights lead the authors to define a concept of small world behaviour in time-dependent networks. Tests of the related metrics on biological networks are currently non existent — Tang et al. briefly show the existence of dynamic small world properties in a cortical network of 16 neurons, but no molecular interaction data is analysed. Other extensions include dynamic network versions of betweenness and closeness centrality measures [Tang et al., 2010a] and conceptual models like *Flow Graphs* [Lambiotte et al., 2011], weighted networks where dynamical flow is embedded in the interaction weights.

Although these techniques imply the existence of *ad hoc* condition-dependent data, it would be interesting to see if any useful insights could be obtained by approximating condition-dependent networks with specific subsets of static protein interaction data. For example, a possibility would be to use transcriptional time course data from Cachero et al. [2011] to create sub-networks of proteins for genes enriched at T_1, T_2 and T_3 and compare their differences. Better yet, the network growing method proposed in the last section of Chapter 4 could be applied to obtain three networks, one per time point. A differential analysis could then potentially reveal clusters of seeds and linking genes active only during one specific phase of embryonal development in *Drosophila* PNS.

5.1.3 Regulatory Network Inference

A branch of computational systems biology research attempts to use machine learning techniques to automatically infer regulatory network structures from transcriptome data. This is known as the problem of *reverse engineering* gene regulatory networks: given n genes and m (possibly noisy) transcriptional level observations (where typically $n >> m$) for each gene, can we reconstruct the unknown data-originating process? A large number of approaches have been proposed (some have been reviewed by De Smet and Marchal [2010]). Popular theoretical frameworks include conditional correlation analysis [Rice et al., 2005], Graphical Gaussian models [Schäfer and Strimmer, 2005] and Bayesian Networks [Friedman et al., 2000]. Bayesian Networks are flexible probabilistic models that represent dependencies between variables in a directed acyclic graph, capturing properties of conditional independence between the variables [Ghahramani, 1998]. These have been rather popular over the past years, and favoured over other competing techniques for their ability to provide built-in ways to deal with noisy data and for the analytical tractability of some of the underlying model's components. Additionally, while most of these methods are suited for condition-based microarray data (e.g. control vs mutation), Bayesian Networks have been extended to time course-based microarray data inference, via so-called Dynamic Bayesian Networks (DBNs) [Friedman et al., 1998].

5.1.3.1 Bayesian Networks

The Bayesian Network framework provides a rigorous learning paradigm for network structure inference. Given the space of all possible Bayesian Network models that fit some given microarray data, $\mathcal{D}$, we can define the conditional probability of the generic model $\mathcal{G}$ given the data

$$m(\mathcal{G}) = P(\mathcal{G}|\mathcal{D}). \tag{5.1}$$

We wish to find the model $\mathcal{G}^\star$ that maximises this conditional probability,

$$\mathcal{G}^\star = \arg\max_{\mathcal{G}}\{P(\mathcal{G}|\mathcal{D})\}, \tag{5.2}$$

that is, we wish to find the BN that is best supported by the data. Applying Bayes' rule,

$$P(\mathcal{G}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{G})P(\mathcal{G})}{P(\mathcal{D})}. \tag{5.3}$$

Since $P(\mathcal{D})$ is a constant depending only on the data and not on the particular model, we can also state that

$$P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G})P(\mathcal{G}). \tag{5.4}$$

Equation 5.4 states that the probability of any network given the microarray data, $P(\mathcal{G}|\mathcal{D})$ (the *posterior distribution*), is proportional to (and ultimately can be factored into) a term made of

two sub-models: the first, $P(\mathcal{D}|\mathcal{G})$, is usually known as the *marginal likelihood* of the data given the model, and depends on the microarray data $\mathcal{D}$; the second, $P(\mathcal{G})$, is called *prior* distribution, and encapsulates any beliefs and biases existing towards a model $\mathcal{G}$ *before* the data $\mathcal{D}$ is taken into account to inform the model.

This learning framework can be used for principled integration of heterogeneous data sources into a Bayesian network model, resulting in potentially improved predictions. An interesting approach in this sense is the one by Imoto et al. [2003], later extended by Werhli and Husmeier [2007]. Their method integrates biological prior knowledge with transcriptional data by modelling any supporting datasets into the prior distribution $P(\mathcal{G})$ using *Energy functions*[1]. Given, for instance, two prior sources of supporting data (for example, protein interactions and transcription factor binding data), Imoto et al. [2003] define the prior over the network structures using the Gibbs distribution and let

$$P(\mathcal{G}) = P(\mathcal{G}|\beta_1,\beta_2) = \frac{e^{\{-\beta_1 E_1(\mathcal{G})-\beta_2 E_2(\mathcal{G})\}}}{Z(\beta_1,\beta_2)}. \tag{5.5}$$

In this expression, $Z()$ is a normalisation constant known as the *partition function*, while β_1 and β_2 are two hyperparameters which indicate the influence strength of each of the two prior sources of biological information relative to the data. The terms E_1 and E_2 are the two energy functions: each measures the agreement of one source of prior with the expression data. The generic energy function E is given by

$$E(\mathcal{G}) = \sum_{i,j=1}^{N} |p_{i,j} - g_{i,j}| \tag{5.6}$$

where N is the total number of nodes (i.e. genes in the microarray data). The term $g_{i,j} \in \{0,1\}$ is the i,j-th term of the adjacency matrix for the network $\mathcal{G}$, while $p_{ij} \in [0,1]$ is the i,j-th term for the biological prior knowledge matrix $\mathcal{P}$. For each pair of genes, these two terms define the agreement between data and prior information: if $E = 0$, there is perfect agreement between data and prior, while increasing values of E indicate increasing mismatch.

In this approach, an informative source of biological information reshapes the posterior distribution on the basis of *a priori* evidence. In other terms, information provided by the prior *seeds* the search for the optimal network by acting as a soft constraint and driving the optimisation procedure towards biologically-supported networks (Figure 5.3-A).

5.1.3.2 Dynamic Bayesian Networks

Plain Bayesian Networks have some shortcomings which make them unsuitable to modelling time course microarray data. One of these is the impossibility of modelling feedback loops. In order to overcome this and other limitations, Dynamic Bayesian Networks (DBNs) have been

[1] Using theoretical results from statistical mechanics.

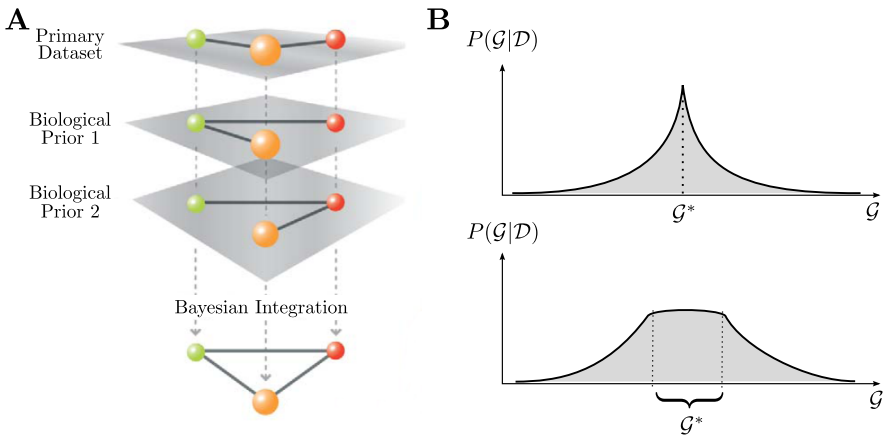

Figure 5.3: **A**: Bayesian Data Integration and the contribution of prior data sources (*Adapted from* Franke et al. [2006]). **B**: Inference Uncertainty. In these hypothetical distributions, the vertical axis shows the posterior probability $P(\mathcal{G}|\mathcal{D})$ while the horizontal axis represents all graph structures $\mathcal{G}$. Given an informative dataset (top) the best structure G^* is clearly defined. Sparse or insufficient data (bottom) leads to uncertainty in the prediction (*Adapted from* Werhli [2007]).

used in transcriptional time course data inference. DBNs factor the probability of the data given the model based on the following Markov expansion [Grzegorczyk and Husmeier, 2011]:

$$P(\mathcal{D}|\mathcal{G},\theta) = \prod_{n=1}^{N}\prod_{t=2}^{m} P(X_n(t) = \mathcal{D}_{n,t}|\pi_n(t-1) = \mathcal{D}_{\pi_n,t-1},\theta_n). \tag{5.7}$$

In this expression, $\mathcal{D}$ is the transcriptional data, $X_1,\ldots,X_N$ are the variables (e.g. gene expression values) measured at equidistant time points $t_1,\ldots,t_m$, and θ_n are the distribution parameters. The term π_n denotes the *parent set* of node X_n, i.e. the nodes from which an edge points to X_n in $\mathcal{G}$. Finally, $\mathcal{D}_{n,t}$ and $\mathcal{D}_{\pi_n,t}$ denote, respectively, the t-th realisations $X_n(t)$ and $\pi_n(t)$ of X_n and π_n.

This model overcomes the feedback loop constraint of static Bayesian Networks, and therefore allows modelling cycles in regulatory networks. However, it makes a *stationarity assumption*: it assumes that the time series are generated from a homogeneous Markov process, which is almost never the case in practice. More recent DBN approaches have overcome this limitation [Robinson and Hartemink, 2010; Dondelinger et al., 2010; Grzegorczyk and Husmeier, 2011].

Structure inference using Bayesian Networks and informative biological priors has great potential to reveal hidden patterns in transcriptional data. Unfortunately, this class of approaches is unsuitable to study time series datasets like the transcriptional data in Cachero et al. [2011]. The high number of variables (thousands of genes for each time point) and the availability of only 3 time points mean that DBN inference would hardly be informative. Most of the successful applications of these methods are based on datasets one or even two orders

of magnitude smaller, with dozens or even hundreds of time points[2]. The problem is basically indetermination: the larger the number of variables, the larger the complexity of the inference problem, the more difficult it is to find a unique solution to the inference problem (Figure 5.3-B).

It would be interesting to test approaches to reduce the data space by, for instance, removing inactive genes — quiet genes showing negligible changes in mRNA concentration over the time point — although care should be taken in the process, because this operation could also be likely to compromise the results of the entire inference. Another possibility would be to reduce the complexity of the problem by binning the transcriptional profiles according to similarity, i.e. a few signature "model profiles" could be determined and profiles could be assigned to these according to their characteristics. This would represent an improvement over standard expression clustering, which usually hypothesizes conditional independence of variables across time points. An approach of this kind has been proposed for instance by Ernst et al. [2005], leading to interesting results on a simulated dataset of 5000 variables and 5 time points. A more recent review of simplification strategies is provided by Wang et al. [2008]. Better yet, it would be interesting to attempt a Bayesian analysis given additional microarray data.

5.1.4 Transcription Factor Target Inference

Although progress is being made, most regulatory structure inference methods are still inadequate when it comes to modelling real world short time series transcriptional datasets [De Smet and Marchal, 2010]. An additional group of approaches to transcriptional inference try to address a simpler question: instead of attempting to infer the full GRN structure, is it possible, given a transcription factor and its profile in short time course data, to tell which genes are regulated by this factor?

A class of recent studies has attempted to answer this question using linear ordinary differential equation (ODE) models of TF protein translation and transcriptional regulation [Barenco et al., 2006; Lawrence et al., 2007; Rogers et al., 2007; Sanguinetti et al., 2009]. In general, given one transcription factor and one target gene to test, TF translation and target transcriptional activation are modelled using a linear system of ODEs, e.g.

$$\frac{\mathrm{d}p(t)}{\mathrm{d}t} = f(t) - \delta p(t), \tag{5.8}$$

$$\frac{\mathrm{d}m_j(t)}{\mathrm{d}t} = B_j + S_j p(t) - D_j m_j(t), \tag{5.9}$$

where $p(t)$ is the transcription factor protein concentration at time t, $f(t)$ is the transcription factor mRNA concentration profile and $m_j(t)$ is the j-th target mRNA concentration at time t.

[2] One biological dataset often used to benchmark structure inference algorithms contains expression time series taken during the whole life cycle of *Drosophila* [Arbeitman et al., 2002]. Usually, data for a subset of around 10-20 genes over 60 time points is considered for the inference.

The parameters B_j and S_j are, respectively, the baseline transcription rate and the sensitivity of gene j to the TF activity; D_j is the decay rate of the protein encoded by gene j, while δ is the decay rate of the TF protein.

One recent version of this ODE-based approach attempts to use the information in short time series data to provide additional support for hypothesized targets of a specific transcription factor, using the ODE model likelihood as a score to rank targets [Honkela et al., 2010]. This method models the TF mRNA concentration as a realisation drawn from a Gaussian process prior distribution [Rasmussen and Williams, 2006]. Gaussian processes are the functional equivalent of the Gaussian distribution, are fully specified by a mean function, $\mu(t)$, and a covariance function, $k(t,t')$ and have interesting analytical properties: realisations from Gaussian processes with square exponential covariance are smooth, infinitely differentiable and stationary functions. Also, any linear operation applied to a function drawn from a Gaussian process leads to a function that is drawn from a related Gaussian process, which allows analytical tractability within the ODE model.

It would appear this approach is especially suitable for validating some of the PNS differentiation targets discussed in this thesis. I have carried out a preliminary experiment to illustrate the potential of the algorithm in Honkela et al. [2010], which has been made available through an excellent R-Bioconductor implementation called *tigre* [Honkela et al., 2011]. In this test run I have performed a model fit using atoGFP+ data from Cachero et al. [2011]: the transcription factor chosen was *ato*, while the targets were three known *ato* targets, *Rfx*, *fd3F* and *dila*. This was done to evaluate potential gross errors in the fit of three known *ato* targets. A joint model for all the targets was learnt (Figure A.9, Page 195). The algorithm ranking gives pretty homogeneous log-likelihood values for the three targets, with *Rfx* coming out on top as the most likely *ato* factor according to the data and the model. This anecdotal evidence suggests that transcription factor inference might potentially reveal interesting patterns in the data. However there are many potential pitfalls to consider: some depend on the structure of our data, some depend on the actual model.

Firstly, three time points might not be sufficient to obtain informative predictions — the shortest time course data used in Honkela et al. [2010] consisted of a set of 12 points subsampled to 7. Moreover, the three time points in Cachero et al. [2011] model a developmental process involving molecules whose decay rates are possibly different from what Honkela et al. hypothesized for their model. Additionally, the algorithm works with absolute expression values, not enrichment values. For my test, affymetrix CEL files for atoGFP+ cells have been used. These might not be informative enough as this data had been designed for joint analysis of enrichment using companion data from atoGFP- cells. It would be interesting to carry out further model tests employing enrichment values instead of absolute expression values.

In addition to potential data incompatibilities, some of the assumptions made in the model

by Honkela et al. [2010] might render it not suitable for our purpose. A critical assumption is that the model cannot deal with cofactors, i.e. it is assumed that each target can only be regulated by one transcription factor[3]. This is clearly not the case for differentiation genes in fly PNS (Figure 4.1, Page 130). Approaches like the one suggested by Opper and Sanguinetti [2010], who propose a methodology to simultaneously infer the activities of multiple interacting TFs, might be more suitable in this sense.

A related source of potential problems is the smoothness of the Gaussian functional priors. Using a GP prior for TF activity introduces a continuity constraint [Sanguinetti et al., 2009]. Honkela et al. [2010] assume that the microarray data can be appropriately modelled by a smooth process modelled by a squared exponential covariance, however we cannot exclude the possibility that single-cell mRNA counts may go up and down in a highly stochastic manner. One way to address this would be to test methods which model the latent process as a Markovian stochastic dynamical process, as suggested for example by Sanguinetti et al. [2009].

As a final note, it would be interesting to evaluate the possibility of reinforcing the model proposed by Honkela et al. [2010] using additional data sources as suggested earlier in the thesis. For example, assuming that products of co-regulated genes tend to organise in tightly connected protein interaction clusters, then protein interaction information could once again be used to further reward or penalise the log likelihood scores proposed through the ODE inference method thus providing further support to the ranked lists proposed by the inference.

5.2 Final Remarks

The advent of functional genomics has allowed the characterisation of the molecular constituents of organisms. However, having understood single molecules, an even bigger challenge consists in understanding how these act. How do proteins orchestrate the large variety of processes that enable a living organism to function? What are the signalling and regulatory interactions in control of the observed changes in cell state or organism state? How is this control exerted?

Large-scale proteomics and expression profiling are two developments of modern biology that have driven our current understanding of how cells work. The analysis of data describing organismal dynamics can increase our comprehension of how living beings develop, acquire, store and use energy, defend against pathogens, adapt to day-night cycles, and die [Altman and Raychaudhuri, 2001]. One class of poorly understood processes drives the development of nervous systems.

In this thesis, I attempted to enrich current understanding of one interesting conserved

[3] An assumption know as the single-input motif (SIM) scenario.

model of nervous system development, neural cell fate specification in *Drosophila* sensory neurons. I demonstrated that it is possible to generate concrete hypotheses for the underlying mechanisms governing PNS differentiation by integrating data at two levels.

The first level of integration happens across organisms. Systematic two-hybrid screens and TAP tag experiments are populating the public databases with a wealth of protein interaction data. I show that some of this data can be used to inform research across organisms, provided it is carefully selected, ordered in classes on the basis of its trustworthiness and analysed to evidence anomalies which would render it harmful rather than helpful.

The second level of integration happens across data describing different stages of protein biosynthesis within an organism. In the late 2000s, when this project started, the relatively new microarray technology allowed large scale quantitative information to be gathered at the transcriptional level. Meanwhile, high throughput Y2H and TAP were allowing the collection of large scale information about protein interaction. Based on previous evidence, I show that the two data sources can be used in conjunction with functional annotation data to build smaller hypothesis sets characterised by high biological support at different levels.

This work shows that data prioritisation and computational integration of information obtained from different experiments at different molecular levels are essential for the understanding of complex biological systems. I argue that this will remain true as new high throughput experimental techniques are proposed to shed light on aspects of cellular regulation, transcription, translation, protein modification and protein action. Several data analysis techniques are being developed to elucidate this wider range of cellular dynamics (for instance, ChIP-seq [Johnson et al., 2007] and RNA-seq [Wang et al., 2009]). Many of these methods allow to collect data at a much higher-resolution than the older generation of methods permitted. I believe that the increased detail and heterogeneity of perspectives will certainly increase our understanding of biological dynamics. However this will only happen if new suitable analytical techniques able to make sense of this ever-growing complexity continue to be developed.

APPENDIX A

Additional Data

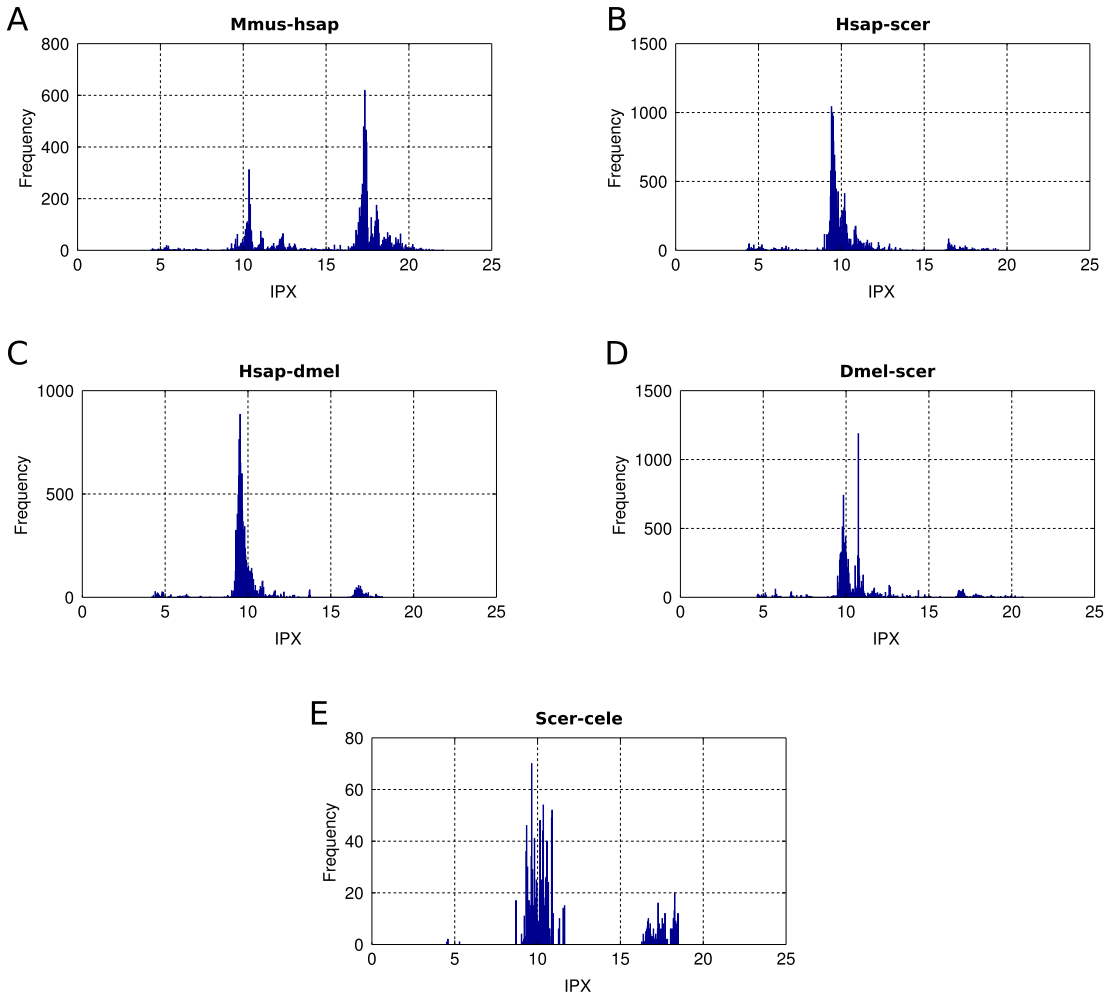

Figure A.1: IPX Histograms for the five putative PPI datasets built from the Positive datasets. The tri-modal shape for the IPX distribution (Figure 2.11, page 49) is evident for all datasets.

	FEPT		Chi-Square	
Contingency Table	*P* (one-tailed)	*P* (two-tailed)	Yates	Pearsons
$\mathcal{F}_1$	1.69e-41	2.43e-41	170.7	172.91
$\mathcal{F}_2$	3.90e-66	3.90e-66	284.29	286.88
$\mathcal{F}_3$	1.38e-59	1.38e-59	251.9	254.37
$\mathcal{F}_4$	4.16e-20	4.16e-20	83.01	84.97

Table A.1: Fisher Exact Probability Test/Chi-Square Test — Results.

$\mathcal{F}_1$	RP	NRP	*total*
Mmus-Hsap	216	56	272
Dmel-Scer	69	220	289
total	285	276	561

$\mathcal{F}_2$	RP	NRP	*total*
Mmus-Hsap	216	56	272
Hsap-Dmel	95	436	531
total	311	492	803

$\mathcal{F}_3$	RP	NRP	*total*
Mmus-Hsap	216	56	272
Hsap-Scer	89	372	461
total	305	428	733

$\mathcal{F}_4$	RP	NRP	*total*
Mmus-Hsap	216	56	272
Scer-Cele	49	96	145
total	265	152	417

Table A.2: 2X2 contingency tables for the Fisher Exact Probability Test/Chi-Square Test. Category X (columns): Known Positive Data Retrieval Capability at $\mathrm{IPX}_{\mathrm{thr}} = 15$. Category Y (rows): Known Positive Dataset. RP: Retrieved Known Positive. NRP: Known Positive Not Retrieved.

Dataset	Low	Mid	High	**Total**
`Mmus-Hsap`	202	3119	7601	10922
			69.6%	
`Hsap-Scer`	667	16800	880	18347
		91.6%		
`Hsap-Dmel`	298	13175	913	14386
		91.6%		
`Dmel-Scer`	553	11177	819	12549
		90%		
`Scer-Cele`	4	1191	311	1506
		79.1%		

Table A.3: Distribution statistics for the known TP samples in the IPX histograms in Figure A.1

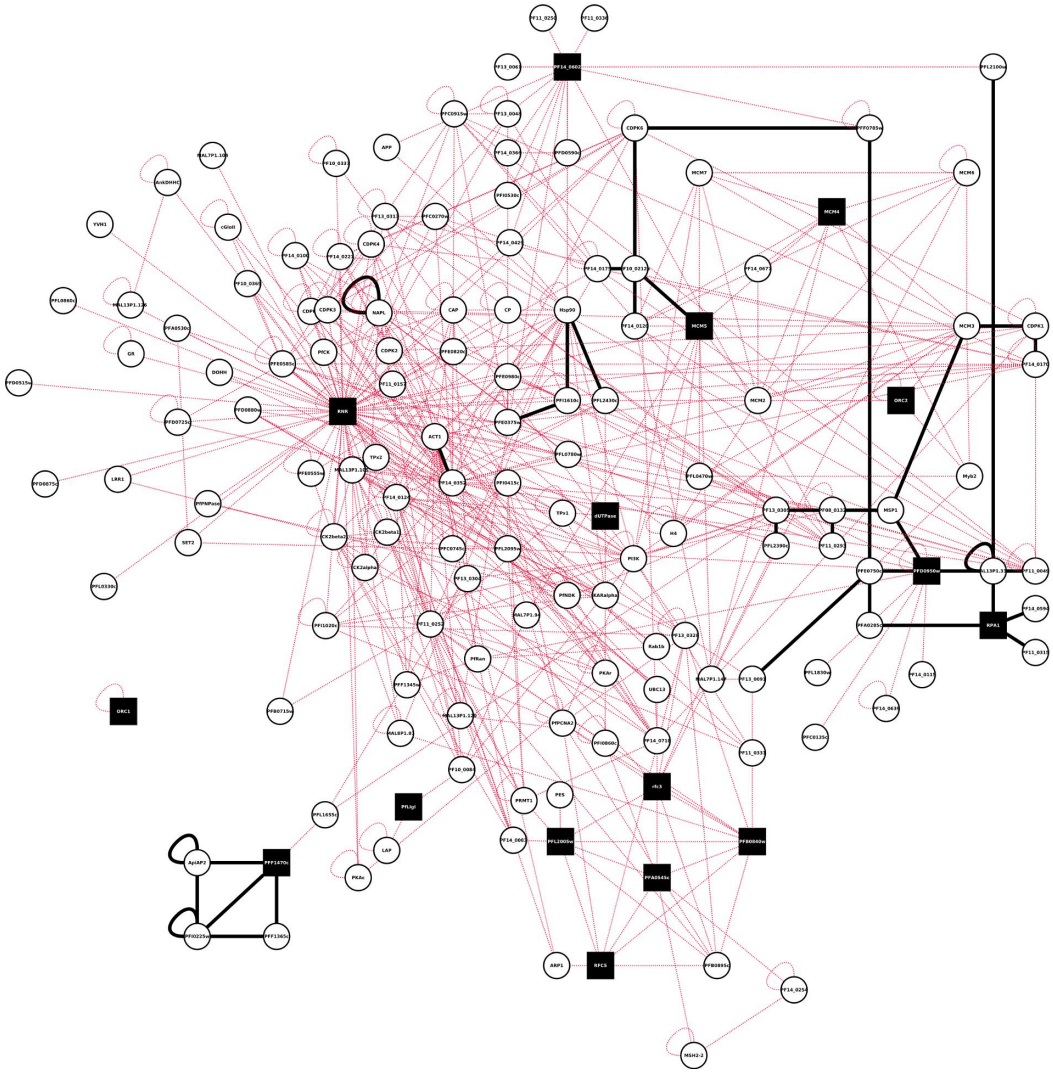

Figure A.2: *Plasmodium falciparum* DNA replication protein interaction network model obtained with `Bio::Homology::InterologWalk`. Data extracted from `NET_DS_PFAL_known` as follows: 1. select all genes annotated with `DNA Replication` GO biological process (16 genes, black nodes) 2. select all their nearest neighbours (white nodes). Solid connections (black) are EBI-Intact experimental protein-protein interactions originally in `NET_DS_PFAL_known`, dotted connections (red) are putative predictions originally in `NET_DS_PFAL_putative`.

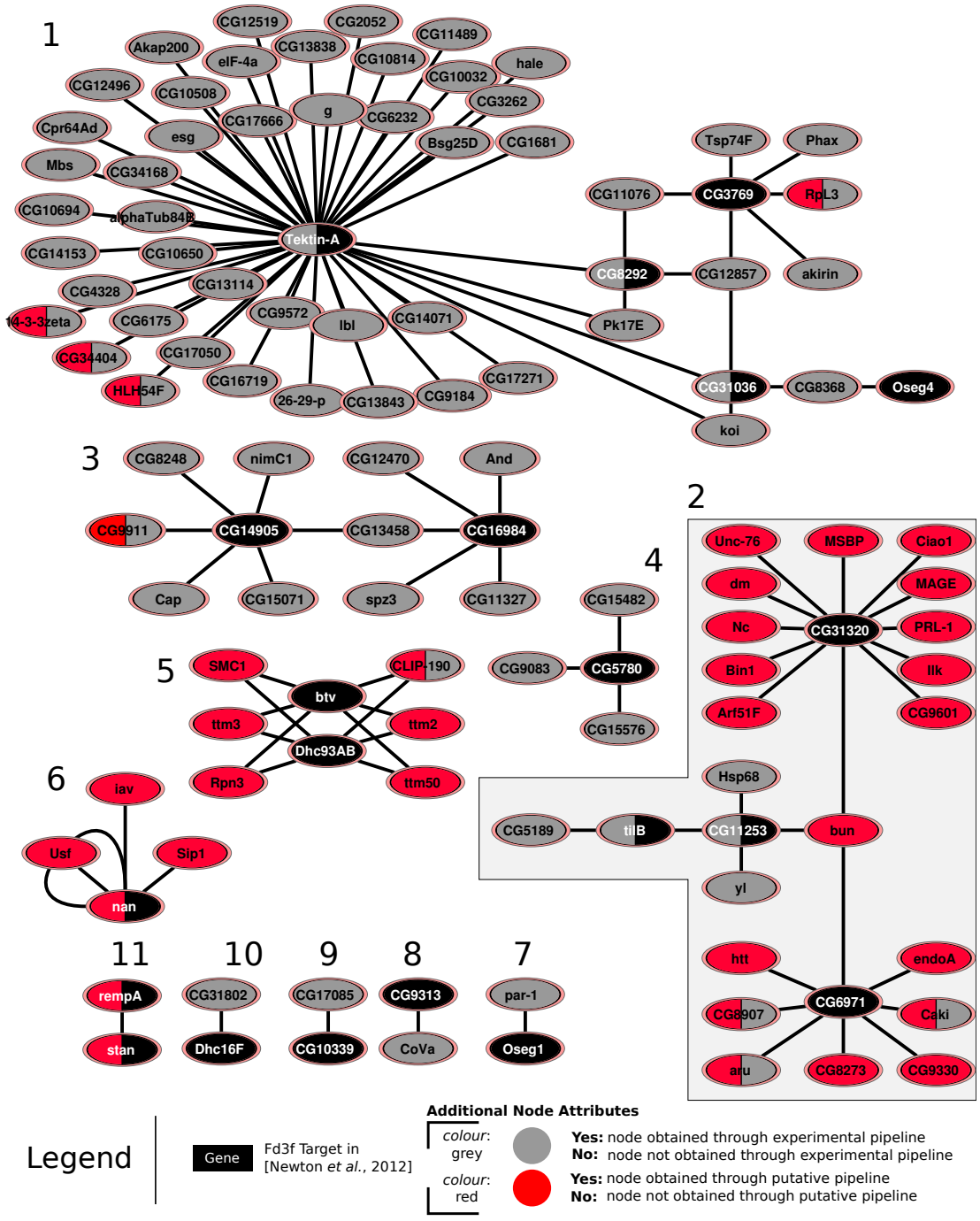

Figure A.3: NET_CATOT4_NN — *Drosophila melanogaster* Cato T_4 nearest-neighbour sub-network. The black nodes are Fd3F targets listed in Newton et al. [2012].

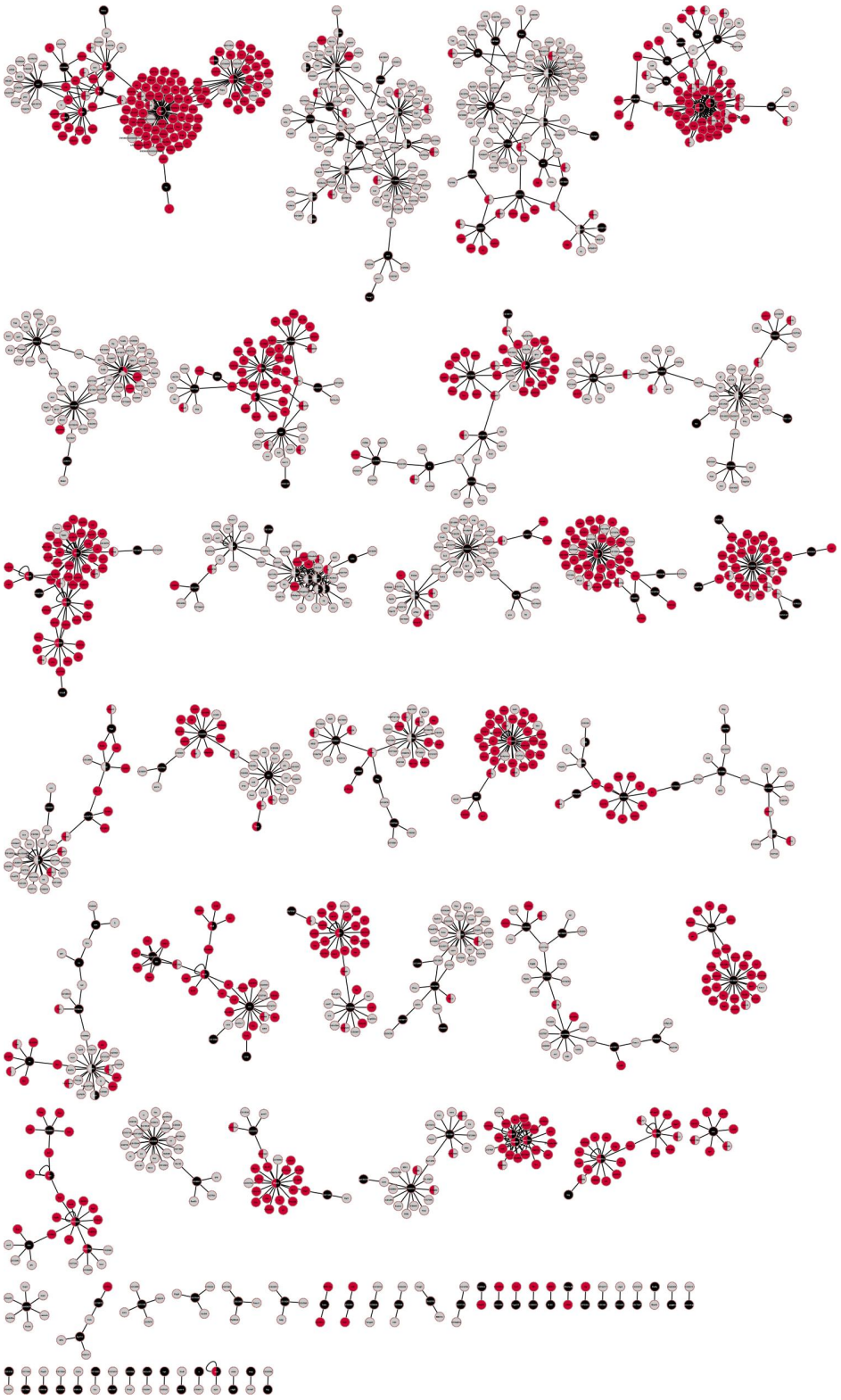

Figure A.4: `NET_CATOT4_union` — Communities found using the clusterMaker Cytoscape plug-in, using the optimised greedy Newman-Girvan proposed in Glay [Su et al., 2010]. Colour coding as in Figure A.3, save for black nodes which are Cato T_4 seed genes.

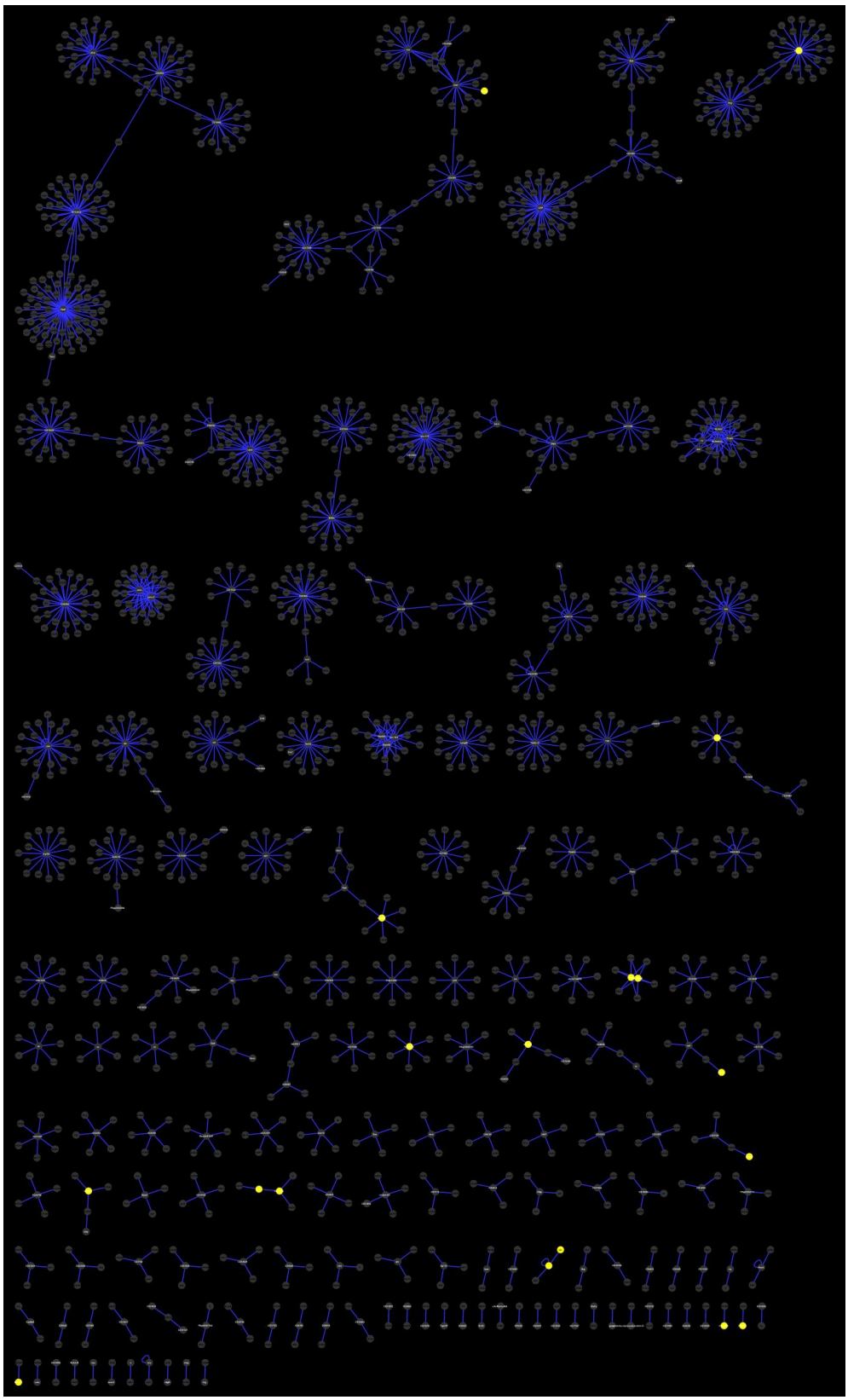

Figure A.5: NET_CATOT4_union — Communities found using MCL [van Dongen, 2000]. Inflation parameter $I = 1.7$. Yellow nodes are fd3F target genes (Table 4.1, Page 135).

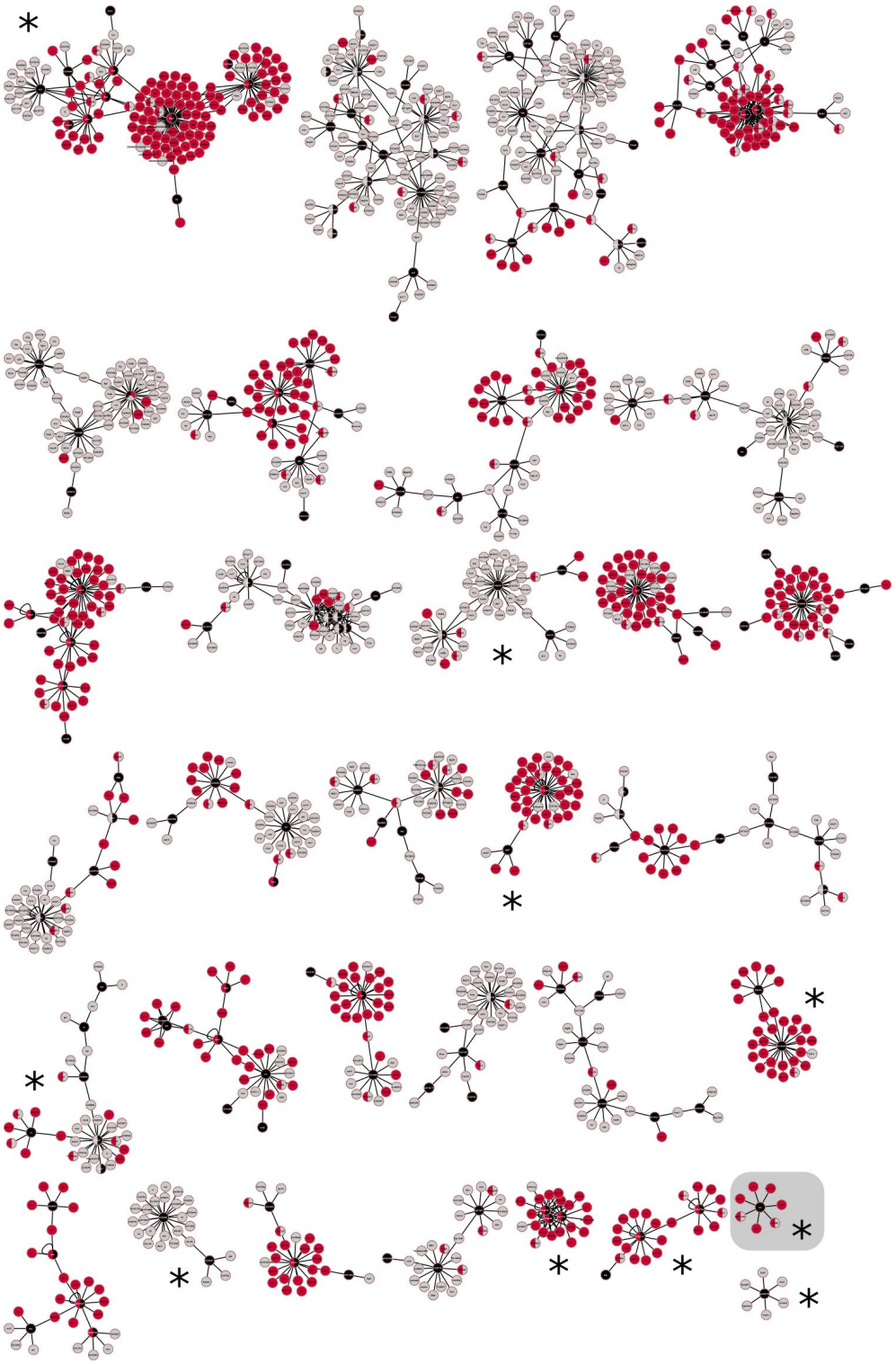

Figure A.6: `NET_CATOT4_union` – $\mathcal{C}$ and $\mathcal{B}$- score community significance. Communities marked by an * have been found to have a B-score ≤ 0.05. One community (shaded square) additionally satisfies $C \leq 0.05$. Results are based on the significance metrics in Lancichinetti et al. [2010].

Parameter	X	μ	σ	z-score	p-val
D	9.000000	9.142000	0.676974	-0.209757	0.8339
CPL	4.009462	3.705467	0.010824	28.084422	1.4823E-173
C_G	0.015709	0.013380	0.000582	3.998426	6.2884E-05
$\overline{C_L}$	0.083513	0.044560	0.002735	14.244060	4.9971E-46

Table A.4: Results of `NET_DCBB_putative` two-tailed randomization z-tests based on four global network parameters: D: network diameter. CPL: Characteristic (Average) Path Legth. C_G: Global Clustering Coefficient (also known as Global Transitivity). $\overline{C_L}$: Average of Local Clustering Coefficients (Average of Local Transitivities). X is observed value in `NET_DCBB_putative`. μ and σ are mean and standard deviation for 1000 degree-conserving randomisations of `NET_DCBB_putative`. z-score indicates distance of X from μ in standard deviation units.

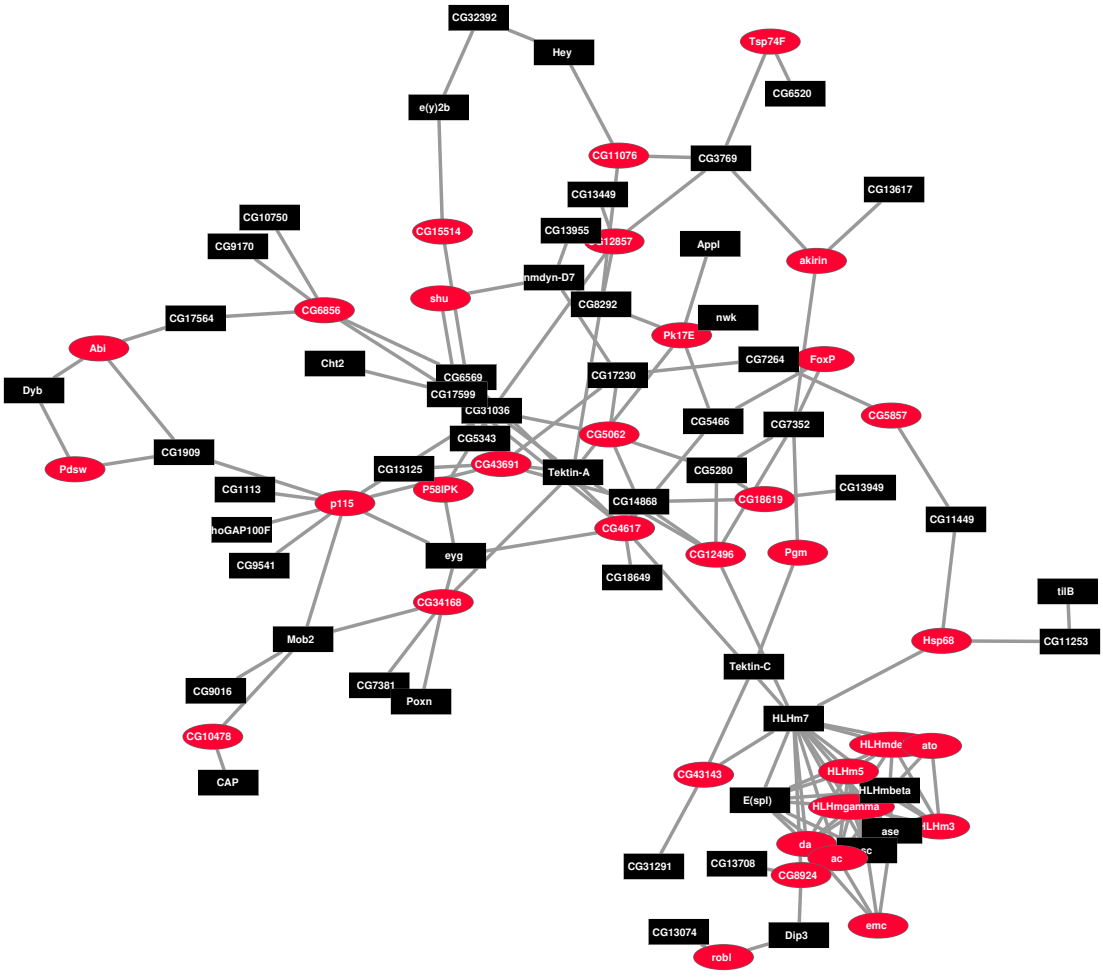

Figure A.7: Largest connected component in network $\mathcal{N}$. Black square nodes: seed genes. Round red nodes: linking genes.

Community	**#Nodes**	$\mathcal{C}$-**score**	$\mathcal{B}$-**score**	$\mathcal{C}$-5% core	$\mathcal{B}$-5% core
1	39	0.811715	0.690059	4	33
2	60	0.666154	0.0793163	0	60
3	112	0.741559	0.306806	0	0
4	107	0.947741	0.246031	0	0
5	65	0.81059	0.136376	0	0
6	33	0.810749	0.075619	0	0
7	75	0.945228	0.974459	26	45
8	37	0.715475	**0.0428175**	0	37
9	50	0.614645	**0.0263295**	0	50
10	61	0.769956	0.473828	7	8
11	23	0.650648	**0.024394**	13	23
12	53	0.686551	0.416859	19	21
13	149	0.816544	**0.000632341**	8	149
14	60	0.827508	0.0780219	0	4
$\longrightarrow$15	38	0.699956	0.570861	0	0
16	36	0.627135	0.369447	8	8
17	28	0.484607	0.0690712	0	28
18	57	0.813273	0.312328	0	9
19	39	0.876164	0.277697	0	0
20	26	0.67565	0.0738685	0	0
21	39	0.838933	0.659665	0	36
22	35	0.532984	0.0790726	0	0
23	39	0.664719	**0.00779402**	0	39
24	28	0.716743	**0.0133392**	0	27
25	34	0.669112	0.32906	0	0
26	40	0.760963	0.324416	0	0
27	49	0.463168	0.229191	0	0
28	6	0.0883773	0.0716907	0	6
29	23	0.391485	**0.0351638**	0	23
30	26	0.699649	0.0987836	0	0
31	4	0.0874859	**0.0388548**	0	4
32	5	0.0753102	**0.0238327**	0	5
33	32	0.480963	**0.000251818**	0	32
*34	8	**0.049533**	**0.00809862**	8	8
35	4	0.0888873	**0.0378414**	0	4
36	7	0.0514943	**0.0106353**	7	7
37	4	0.0902695	0.0588346	0	4

Table A.5: `NET_CATOT4_union` – $\mathcal{C}$- and $\mathcal{B}$- score community significance. The $\mathcal{C}$- and $\mathcal{B}$- q cores at the $q = 0.05$ significance level are also shown. Arrow indicates results for Community B.

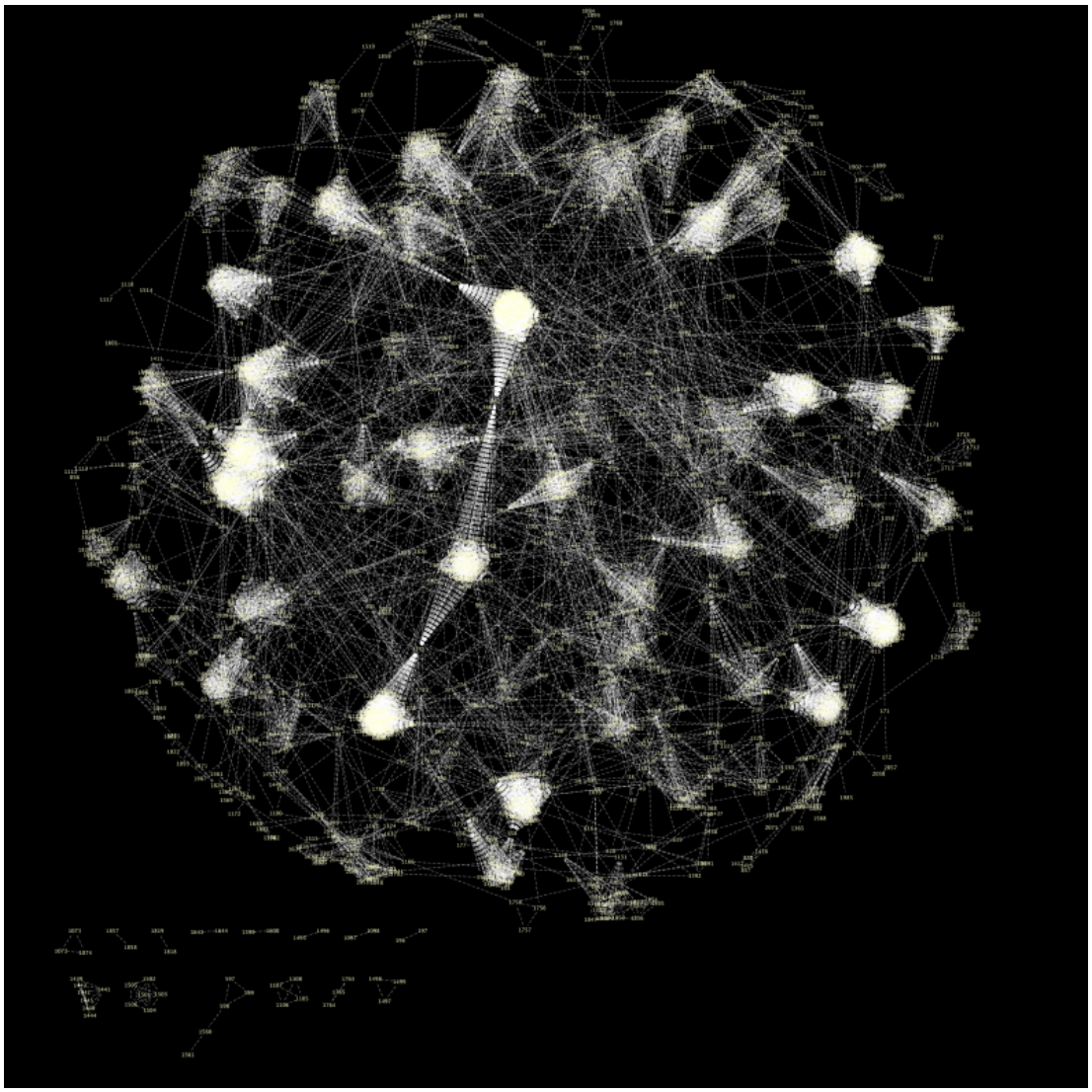

Figure A.8: Line Graph Transformation of `NET_CATOT4_union` ($N = 2044$, $E = 27725$).

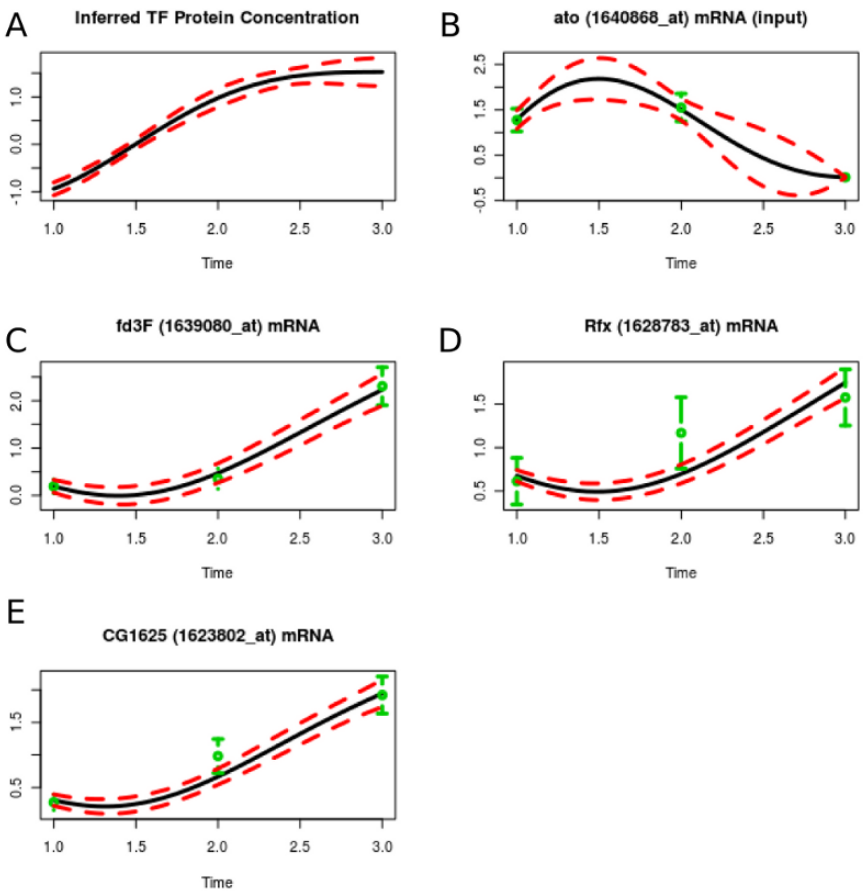

Figure A.9: Test run of `tigre` using the linear ODE model with Gaussian Process Prior. Dataset is atoGFP+ [Cachero et al., 2011], 3 time points $\times$ 4 replicates. **A**: predicted protein concentration for *ato*; **B**: predicted expression level for *ato*; **C**: predicted expression level for *fd3F* (LL score $= -37.24$); **D**: predicted expression level for *Rfx* (LL score $= -35.88$); **E**: predicted expression level for CG1625 (*dila*) (LL score $= -47.85$). Solid lines represent the mean inference, dashed lines show the 95% credible intervals. Green crosses are the observed gene expression data with error bars showing the technical error from each individual Affymetrix microarray processed using the puma package [Liu et al., 2005]. Data and reconstructed profiles are shown on an unlogged normalised scale. Time is measured in hours.

Table A.6: **Top 285 *cato*-correlated genes at time point T4** - List of genes ranked by fold change (FC) (i.e., ratio of expression in catoGFP cells versus the rest of the embryo) (1% FDR).

Rank	Flybase ID	Gene Name	AffyDros2 ID	FC
1	FBgn0039228	CG6980	1640684_at	27.203
2	FBgn0038452	CG14905	1631651_at	25.898
3	FBgn0032768	CG17564	1637094_at	23.523
4	FBgn0085326	CG34297	1637157_at	23.129
5	FBgn0003710	tipE	1626200_s_at	22.273
6	FBgn0038358	CG4525	1635131_at	19.313
7	FBgn0036338	CG11253	1636602_at	19.186
8	FBgn0061173	**fd3F**	1639080_at	18.563
9	FBgn0039152	CG6129	1634341_a_at	17.785
10	FBgn0032163	CG13125	1636760_a_at	17.572
11	FBgn0038079	CG14394	1633481_at	17.529
12	FBgn0052006	CG32006	1625727_at	17.256
13	FBgn0051291	CG31291	1624002_a_at	17.189
14	FBgn0000206	boss	1628369_at	17.026
15	FBgn0031783	tectonic	1625411_at	16.474
16	FBgn0032470	CG5142	1632369_at	16.213
17	FBgn0036437	CG5048	1629835_at	15.910
18	FBgn0031829	osm-6	1640025_at	15.551
19	FBgn0034972	CG10339	1637886_at	15.435
20	FBgn0016047	nompA	1641377_a_at	15.418
21	FBgn0023096	btv	1638983_at	15.364
22	FBgn0033447	dila	1623802_at	15.217
23	FBgn0035724	CG10064	1631786_at	15.172
24	FBgn0038029	CG17639	1628094_at	15.120
25	FBgn0034352	CG17669	1623294_at	14.950
26	FBgn0032119	CG3769	1634763_at	14.899
27	FBgn0033628	CG13203	1628436_at	14.127
28	FBgn0032692	CG15161	1632286_at	13.774
29	FBgn0062517	CG16984	1637969_at	13.557
30	FBgn0085211	RpS28-like	1629350_at	13.150
31	FBgn0052703	CG32703	1636384_at	12.955
32	FBgn0036687	CG6652	1624197_a_at	11.612
33	FBgn0035168	CG13889	1641002_at	11.592
34	FBgn0039467	CG14253	1625563_s_at	10.682
35	FBgn0003513	ss	1625198_at	10.477
36	FBgn0032891	Oseg5	1640383_at	10.071
37	FBgn0034095	CG15701	1639248_at	9.864
38	FBgn0039201	CG13617	1630908_at	9.764
39	FBgn0051216	Naam	1640214_at	9.704
40	FBgn0037280	CG1126	1627157_at	9.625
41	FBgn0038909	CG6569	1639974_a_at	9.358

Continued on next page

Table A.6 –continued from previous page

Rank	Flybase ID	Gene Name	AffyDros2 ID	FC
42	FBgn0003295	ru	1632183_at	8.995
43	FBgn0051036	CG31036	1633971_at	8.758
44	FBgn0023090	dtr	1635400_at	8.538
45	FBgn0028902	Tektin-A	1625669_at	8.295
46	FBgn0031255	BBS8	1627476_at	8.186
47	FBgn0020379	Rfx	1628783_at	7.632
48	FBgn0031196	CG17599	1628749_at	7.464
49	FBgn0035317	osm-1	1633181_at	7.445
50	FBgn0034239	CG43370	1640213_at	7.424
51	FBgn0031634	Ir25a	1640674_at	7.326
52	FBgn0036206	CG5964	1635108_at	7.256
53	FBgn0033629	Tsp47F	1629058_a_at	7.255
54	FBgn0034452	Oseg6	1628733_at	7.244
55	FBgn0034037	CG8214	1630315_at	7.147
56	FBgn0029656	CG10793	1627457_at	7.117
57	FBgn0033578	BBS4	1636372_at	7.110
58	FBgn0032083	CG9541	1629772_at	7.030
59	FBgn0051320	CG31320	1632590_at	6.882
60	FBgn0038342	CG14870	1630162_at	6.814
61	FBgn0034106	CG9068	1636873_at	6.728
62	FBgn0039463	CG18472	1637320_at	6.687
63	FBgn0033943	CG12869	1629232_at	6.600
64	FBgn0036115	CG6327	1623632_s_at	6.566
65	FBgn0037712	CG16789	1629421_at	6.546
66	FBgn0036935	CG14186	1624772_at	6.399
67	FBgn0039408	CG14551	1638911_at	6.238
68	FBgn0031496	CG17258	1641706_at	6.228
69	FBgn0033412	CG13955	1631706_at	6.022
70	FBgn0034446	CG7735	1624234_at	5.978
71	FBgn0013811	Dhc62B	1637896_at	5.963
72	FBgn0035256	CG13930	1640436_at	5.752
73	FBgn0037962	CG6971	1627798_at	5.717
74	FBgn0032004	CG8292	1636724_at	5.663
75	FBgn0034278	CG14488	1637274_at	5.637
76	FBgn0000591	E(spl)	1629966_at	5.549
77	FBgn0030634	CG9164	1627024_s_at	5.549
78	FBgn0038221	CG3259	1641499_at	5.544
79	FBgn0035577	CG13708	1640312_at	5.488
80	FBgn0043550	Tsp68C	1623096_a_at	5.404
81	FBgn0039522	CG13972	1635516_at	5.375
82	FBgn0004118	nAcRbeta-96A	1635349_a_at	5.308
83	FBgn0031707	CG14020	1634371_at	5.291
84	FBgn0038098	CG7381	1635467_a_at	5.135
85	FBgn0031288	CG13949	1627467_at	5.037
86	FBgn0033054	CG14591	1640767_s_at	5.015
87	FBgn0035952	CG5280	1638851_at	4.900
88	FBgn0036567	CG13074	1632535_at	4.869
89	FBgn0027550	CG6495	1626911_at	4.864
90	FBgn0038579	CG14313	1626939_at	4.752

Continued on next page

Table A.6 –continued from previous page

Rank	Flybase ID	Gene Name	AffyDros2 ID	FC
91	FBgn0052392	CG32392	1640789_a_at	4.710
92	FBgn0017590	klg	1635144_at	4.696
93	FBgn0067317	Cby	1625135_at	4.692
94	FBgn0035710	SP1173	1633553_s_at	4.620
95	FBgn0037162	CG11449	1637426_at	4.598
96	FBgn0015721	king-tubby	1639525_at	4.493
97	FBgn0260933	rempA	1632552_at	4.442
98	FBgn0033504	CAP	1633353_s_at	4.412
99	FBgn0036520	CG13449	1637521_at	4.400
100	FBgn0000303	Cha	1640284_at	4.328
101	FBgn0030395	CG15730	1625536_at	4.295
102	FBgn0034920	CG5597	1638697_at	4.276
103	FBgn0031550	CG8853	1639369_at	4.239
104	FBgn0032428	CG6405	1628525_at	4.224
105	FBgn0038916	dnd	1638592_at	4.191
106	FBgn0032345	CG14921	1638032_at	4.190
107	FBgn0030004	CG10958	1637127_at	4.180
108	FBgn0050259	CG30259	1638823_at	4.109
109	FBgn0036214	CG7264	1624173_at	4.057
110	FBgn0032225	CG5022	1625191_at	4.057
111	FBgn0030716	CG9170	1641400_at	4.053
112	FBgn0004170	sc	1625273_at	4.044
113	FBgn0038641	CG7708	1624129_s_at	3.995
114	FBgn0022702	Cht2	1637421_at	3.954
115	FBgn0085221	robls54B	1640171_at	3.928
116	FBgn0014395	tilB	1640477_at	3.881
117	FBgn0033288	pdm3	1631222_at	3.853
118	FBgn0028997	nmdyn-D7	1640512_at	3.841
119	FBgn0039203	CG13618	1626789_at	3.834
120	FBgn0037076	ebd2	1630696_at	3.820
121	FBgn0030485	CG1998	1632313_at	3.815
122	FBgn0013812	Dhc93AB	1630304_at	3.778
123	FBgn0047330	CG32235	1640919_at	3.706
124	FBgn0038330	CG14868	1625817_at	3.699
125	FBgn0053200	ventrally-expressed-protein-D	1639902_at	3.678
126	FBgn0036725	CG18265	1632141_at	3.673
127	FBgn0051790	CG43338	1639122_at	3.658
128	FBgn0016032	lbm	1639406_at	3.644
129	FBgn0032446	CG5780	1637810_at	3.642
130	FBgn0034070	SP2353	1625982_at	3.579
131	FBgn0032002	CG8353	1632345_at	3.569
132	FBgn0036771	CG14353	1629820_at	3.550
133	FBgn0032769	CG10750	1626622_at	3.515
134	FBgn0034103	CG15704	1628430_at	3.499
135	FBgn0035967	CG4641	1637294_at	3.496
136	FBgn0052137	CG32137	1625087_a_at	3.493
137	FBgn0004381	Klp68D	1639576_at	3.479
138	FBgn0052085	CG32085	1628373_at	3.448
139	FBgn0034566	CG9313	1640970_at	3.443

Continued on next page

Table A.6 –continued from previous page

Rank	Flybase ID	Gene Name	AffyDros2 ID	FC
140	FBgn0034136	DAT	1633205_at	3.419
141	FBgn0036202	CG6024	1636848_at	3.369
142	FBgn0028572	qtc	1638464_a_at	3.355
143	FBgn0033174	CG11125	1626167_at	3.327
144	FBgn0035170	dpr20	1623063_at	3.324
145	FBgn0034622	CG15666	1641154_at	3.316
146	FBgn0030230	Rph	1639837_at	3.276
147	FBgn0050090	CG30090	1636154_at	3.264
148	FBgn0004880	scrt	1630860_at	3.259
149	FBgn0034225	veil	1635447_at	3.240
150	FBgn0040491	Buffy	1623425_at	3.202
151	FBgn0036993	CG5910	1623928_at	3.168
152	FBgn0034451	CG11242	1639253_at	3.167
153	FBgn0039916	CG9935	1626574_at	3.163
154	FBgn0019940	Rh6	1640642_at	3.150
155	FBgn0038607	CG7669	1637784_at	3.111
156	FBgn0035146	CG13893	1636193_at	3.099
157	FBgn0033960	CG10151	1630624_s_at	3.085
158	FBgn0038247	Cad88C	1637488_at	3.070
159	FBgn0037276	CG17387	1636580_at	3.063
160	FBgn0035891	Oseg1	1627247_at	3.019
161	FBgn0027376	rha	1625115_at	2.998
162	FBgn0033774	CG12374	1638361_at	2.988
163	FBgn0020248	stet	1641652_a_at	2.984
164	FBgn0038114	CG11670	1626623_at	2.984
165	FBgn0013725	phyl	1624262_at	2.975
166	FBgn0024249	cato	1623462_at	2.943
167	FBgn0040670	e(y)2b	1626717_at	2.927
168	FBgn0036469	CG18649	1631350_at	2.925
169	FBgn0000108	Appl	1624033_at	2.922
170	FBgn0003950	unc	1637035_at	2.910
171	FBgn0032800	CG10137	1641339_at	2.906
172	FBgn0035246	CG13928	1637150_at	2.894
173	FBgn0005561	sv	1636090_a_at	2.880
174	FBgn0032084	CG13101	1636168_s_at	2.880
175	FBgn0030742	CG9919	1626619_at	2.877
176	FBgn0013809	Dhc16F	1637895_at	2.872
177	FBgn0011701	repo	1625431_at	2.831
178	FBgn0036285	toe	1633383_at	2.830
179	FBgn0030120	CG17440	1631166_at	2.816
180	FBgn0000179	bi	1637049_at	2.813
181	FBgn0010407	Ror	1626152_at	2.783
182	FBgn0000625	eyg	1632742_at	2.783
183	FBgn0031762	CG9098	1631188_a_at	2.768
184	FBgn0026403	Ndg	1632082_at	2.754
185	FBgn0038047	CG5245	1637467_at	2.753
186	FBgn0038815	CG5466	1633986_at	2.750
187	FBgn0036242	CG6793	1628115_at	2.748
188	FBgn0039734	Tace	1639144_a_at	2.741

Continued on next page

Table A.6 –continued from previous page

Rank	Flybase ID	Gene Name	AffyDros2 ID	FC
189	FBgn0000527	e	1639823_at	2.728
190	FBgn0039202	CG13622	1623500_at	2.717
191	FBgn0028550	Atf3	1637792_at	2.708
192	FBgn0002633	HLHm7	1625493_at	2.707
193	FBgn0030600	hiw	1632637_at	2.697
194	FBgn0032248	CG5343	1632413_at	2.694
195	FBgn0051072	Lerp	1632908_s_at	2.691
196	FBgn0028494	CG6424	1632465_s_at	2.657
197	FBgn0032447	PICK1	1623746_a_at	2.652
198	FBgn0031191	Cp110	1631024_s_at	2.628
199	FBgn0033983	CG10253	1629034_at	2.627
200	FBgn0035542	DOR	1623536_s_at	2.601
201	FBgn0027571	CG3523	1624549_at	2.598
202	FBgn0002573	sens	1632294_at	2.594
203	FBgn0032926	-	1632466_at	2.584
204	FBgn0028642	esn	1632679_s_at	2.558
205	FBgn0005613	Sox15	1633526_at	2.557
206	FBgn0031257	CG4133	1631360_at	2.556
207	FBgn0024836	stan	1626087_at	2.550
208	FBgn0002937	ninaB	1625743_at	2.548
209	FBgn0037304	CG1113	1625769_at	2.543
210	FBgn0032749	Phlpp	1623550_at	2.539
211	FBgn0031005	Hs3st-B	1638753_at	2.537
212	FBgn0034224	CG6520	1627238_at	2.533
213	FBgn0036348	CG17687	1626868_at	2.522
214	FBgn0086370	sra	1639162_at	2.518
215	FBgn0033408	CG8800	1624820_at	2.509
216	FBgn0035743	CG15829	1633251_at	2.502
217	FBgn0015001	iotaTry	1632540_at	2.498
218	FBgn0037727	CG8358	1626377_at	2.468
219	FBgn0051118	RabX4	1623261_at	2.464
220	FBgn0000634	Fas1	1624183_a_at	2.446
221	FBgn0036626	CG13036	1638786_at	2.439
222	FBgn0029663	CG10804	1632629_a_at	2.434
223	FBgn0052458	nrm	1635083_at	2.431
224	FBgn0004618	gl	1623923_a_at	2.424
225	FBgn0052365	CG32365	1628913_at	2.421
226	FBgn0037581	CG7352	1632855_at	2.405
227	FBgn0259182	CG42286	1626627_at	2.390
228	FBgn0051125	CG31125	1641314_at	2.374
229	FBgn0027788	Hey	1624527_at	2.369
230	FBgn0002733	HLHmbeta	1639900_at	2.369
231	FBgn0040465	Dip3	1631536_at	2.362
232	FBgn0038202	CG12402	1634709_at	2.359
233	FBgn0004380	Klp64D	1628601_at	2.353
234	FBgn0039808	CG12071	1624143_a_at	2.338
235	FBgn0036273	CG10426	1639095_at	2.334
236	FBgn0034158	CG5522	1632545_s_at	2.325
237	FBgn0004054	zen2	1630133_at	2.322

Continued on next page

Table A.6 – concluded from previous page

Rank	Flybase ID	Gene Name	AffyDros2 ID	FC
238	FBgn0051632	sens-2	1628245_at	2.320
239	FBgn0032876	CG1962	1639807_s_at	2.320
240	FBgn0003886	alphaTub85E	1623910_at	2.318
241	FBgn0000346	comt	1639825_at	2.316
242	FBgn0031065	CG14234	1632391_at	2.303
243	FBgn0037838	CG4089	1636309_at	2.302
244	FBgn0021738	Crg-1	1624373_at	2.295
245	FBgn0038256	CG7530	1628081_s_at	2.260
246	FBgn0259481	Mob2	1640774_a_at	2.244
247	FBgn0033739	Dyb	1630118_s_at	2.240
248	FBgn0039883	RhoGAP100F	1630285_at	2.233
249	FBgn0032429	CG5446	1641334_at	2.232
250	FBgn0003053	peb	1622949_at	2.229
251	FBgn0038926	CG13409	1639758_at	2.221
252	FBgn0035092	Nplp1	1628450_at	2.217
253	FBgn0004854	B-H2	1640139_at	2.209
254	FBgn0034155	unc-104	1637684_at	2.207
255	FBgn0033072	CG17994	1638728_at	2.203
256	FBgn0032897	CG9336	1627590_at	2.199
257	FBgn0036859	CG14085	1634752_a_at	2.195
258	FBgn0035264	Oseg4	1629688_at	2.185
259	FBgn0052187	CG32187	1625582_at	2.169
260	FBgn0010114	hig	1636585_a_at	2.168
261	FBgn0034184	CG9646	1632187_at	2.160
262	FBgn0036414	nan	1640192_at	2.134
263	FBgn0025549	unc-119	1628148_at	2.129
264	FBgn0000137	ase	1635124_at	2.112
265	FBgn0034493	CG8908	1640231_a_at	2.109
266	FBgn0003130	Poxn	1632977_at	2.108
267	FBgn0000630	f	1625621_s_at	2.098
268	FBgn0031751	CG9016	1631128_s_at	2.097
269	FBgn0035085	CG3770	1627872_at	2.097
270	FBgn0031596	CG15429	1628693_at	2.095
271	FBgn0035164	CG13901	1635239_at	2.093
272	FBgn0036962	CG17122	1634050_at	2.085
273	FBgn0035638	Tektin-C	1628238_at	2.081
274	FBgn0032019	mtsh	1633211_a_at	2.078
275	FBgn0038039	CG5196	1623448_at	2.070
276	FBgn0020391	Nrk	1635246_at	2.055
277	FBgn0000414	Dab	1629243_at	2.043
278	FBgn0036236	CG6931	1639782_at	2.040
279	FBgn0035903	CG6765	1639157_at	2.036
280	FBgn0028996	onecut	1630376_at	2.033
281	FBgn0039911	CG1909	1632023_s_at	2.025
282	FBgn0030847	CG12991	1635980_s_at	2.013
283	FBgn0035521	VhaM9.7-a	1635300_at	2.005
284	FBgn0051660	pog	1635488_at	2.004
285	FBgn0046225	CG17230	1625201_s_at	2.004

Table A.7: **Analysis of BP-GO terms enriched in Community A**

GO-ID	*p*-value	corr *p*-value	cluster freq	total freq	Desc
2001141	4.0071E-10	1.2782E-7	21/36 (58.3%)	169/1192 (14.2%)	regulation of RNA biosynthetic process
6355	4.0071E-10	1.2782E-7	21/36 (58.3%)	169/1192 (14.2%)	regulation of transcription, DNA-dependent
6357	5.3556E-10	1.2782E-7	16/36 (44.4%)	89/1192 (7.5%)	regulation of transcription from RNA polymerase II promoter
2000112	1.2564E-9	1.7992E-7	21/36 (58.3%)	179/1192 (15.0%)	regulation of cellular macromolecule biosynthetic process
10556	1.2564E-9	1.7992E-7	21/36 (58.3%)	179/1192 (15.0%)	regulation of macromolecule biosynthetic process
9889	2.6723E-9	2.3917E-7	21/36 (58.3%)	186/1192 (15.6%)	regulation of biosynthetic process
31326	2.6723E-9	2.3917E-7	21/36 (58.3%)	186/1192 (15.6%)	regulation of cellular biosynthetic process
51252	2.6723E-9	2.3917E-7	21/36 (58.3%)	186/1192 (15.6%)	regulation of RNA metabolic process
10468	3.3100E-9	2.6333E-7	22/36 (61.1%)	209/1192 (17.5%)	regulation of gene expression
19219	1.0945E-8	7.8363E-7	21/36 (58.3%)	200/1192 (16.8%)	regulation of nucleobase-containing compound metabolic process
51171	1.2048E-8	7.8419E-7	21/36 (58.3%)	201/1192 (16.9%)	regulation of nitrogen compound metabolic process
7423	2.0268E-8	1.2093E-6	17/36 (47.2%)	129/1192 (10.8%)	sensory organ development
80090	2.4805E-8	1.3662E-6	22/36 (61.1%)	231/1192 (19.4%)	regulation of primary metabolic process
31323	4.1229E-8	2.1086E-6	22/36 (61.1%)	237/1192 (19.9%)	regulation of cellular metabolic process
60255	4.4802E-8	2.1386E-6	22/36 (61.1%)	238/1192 (20.0%)	regulation of macromolecule metabolic process
19222	7.2209E-8	3.2313E-6	23/36 (63.9%)	268/1192 (22.5%)	regulation of metabolic process
48513	1.5206E-7	6.4043E-6	23/36 (63.9%)	278/1192 (23.3%)	organ development
61382	3.8097E-7	1.5154E-5	5/36 (13.9%)	7/1192 (0.6%)	Malpighian tubule tip cell differentiation
122	5.0207E-7	1.8920E-5	9/36 (25.0%)	37/1192 (3.1%)	negative regulation of transcription from RNA polymerase II promoter
45892	9.1408E-7	3.2724E-5	10/36 (27.8%)	51/1192 (4.3%)	negative regulation of transcription, DNA-dependent
51253	1.6088E-6	5.4852E-5	10/36 (27.8%)	54/1192 (4.5%)	negative regulation of RNA metabolic process
45934	2.2964E-6	7.4739E-5	10/36 (27.8%)	56/1192 (4.7%)	negative regulation of nucleobase-containing compound metabolic process
7422	2.4941E-6	7.5120E-5	8/36 (22.2%)	33/1192 (2.8%)	peripheral nervous system development
2000113	2.7278E-6	7.5120E-5	10/36 (27.8%)	57/1192 (4.8%)	negative regulation of cellular macromolecule biosynthetic process
51172	2.7278E-6	7.5120E-5	10/36 (27.8%)	57/1192 (4.8%)	negative regulation of nitrogen compound metabolic process
10558	2.7278E-6	7.5120E-5	10/36 (27.8%)	57/1192 (4.8%)	negative regulation of macromolecule biosynthetic process
9890	3.2283E-6	8.2553E-5	10/36 (27.8%)	58/1192 (4.9%)	negative regulation of biosynthetic process
31327	3.2283E-6	8.2553E-5	10/36 (27.8%)	58/1192 (4.9%)	negative regulation of cellular biosynthetic process
31324	3.7504E-6	9.2597E-5	11/36 (30.6%)	73/1192 (6.1%)	negative regulation of cellular metabolic process
50789	5.3001E-6	1.2650E-4	29/36 (80.6%)	522/1192 (43.8%)	regulation of biological process
10629	6.1209E-6	1.4137E-4	10/36 (27.8%)	62/1192 (5.2%)	negative regulation of gene expression
10605	8.4261E-6	1.8853E-4	11/36 (30.6%)	79/1192 (6.6%)	negative regulation of macromolecule metabolic process
9892	1.7591E-5	3.8166E-4	11/36 (30.6%)	85/1192 (7.1%)	negative regulation of metabolic process
1654	1.9895E-5	4.1897E-4	12/36 (33.3%)	103/1192 (8.6%)	eye development
7539	2.3075E-5	4.5893E-4	4/36 (11.1%)	7/1192 (0.6%)	primary sex determination, soma
7460	2.3075E-5	4.5893E-4	4/36 (11.1%)	7/1192 (0.6%)	R8 cell fate commitment
50794	2.3756E-5	4.5972E-4	27/36 (75.0%)	485/1192 (40.7%)	regulation of cellular process
7541	2.5358E-5	4.7724E-4	3/36 (8.3%)	3/1192 (0.3%)	sex determination, primary response to X:A ratio
8407	2.5995E-5	4.7724E-4	6/36 (16.7%)	22/1192 (1.8%)	chaeta morphogenesis
48731	3.3886E-5	6.0657E-4	25/36 (69.4%)	428/1192 (35.9%)	system development
45165	3.6120E-5	6.3078E-4	12/36 (33.3%)	109/1192 (9.1%)	cell fate commitment
7538	4.5156E-5	7.3481E-4	4/36 (11.1%)	8/1192 (0.7%)	primary sex determination
18993	4.5156E-5	7.3481E-4	4/36 (11.1%)	8/1192 (0.7%)	somatic sex determination
45465	4.5156E-5	7.3481E-4	4/36 (11.1%)	8/1192 (0.7%)	R8 cell differentiation
7219	6.4999E-5	1.0342E-3	5/36 (13.9%)	16/1192 (1.3%)	Notch signaling pathway
65007	6.7421E-5	1.0365E-3	29/36 (80.6%)	580/1192 (48.7%)	biological regulation
48699	6.8037E-5	1.0365E-3	15/36 (41.7%)	177/1192 (14.8%)	generation of neurons
48749	7.0656E-5	1.0540E-3	11/36 (30.6%)	98/1192 (8.2%)	compound eye development
7530	7.9531E-5	1.1165E-3	4/36 (11.1%)	9/1192 (0.8%)	sex determination
7400	7.9531E-5	1.1165E-3	4/36 (11.1%)	9/1192 (0.8%)	neuroblast fate determination
14017	7.9531E-5	1.1165E-3	4/36 (11.1%)	9/1192 (0.8%)	neuroblast fate commitment
22416	1.1501E-4	1.5836E-3	6/36 (16.7%)	28/1192 (2.3%)	chaeta development
14016	1.2970E-4	1.7521E-3	4/36 (11.1%)	10/1192 (0.8%)	neuroblast differentiation
48523	1.3316E-4	1.7656E-3	14/36 (38.9%)	165/1192 (13.8%)	negative regulation of cellular process
7275	1.4711E-4	1.9151E-3	26/36 (72.2%)	494/1192 (41.4%)	multicellular organismal development
61326	1.6197E-4	2.0346E-3	5/36 (13.9%)	19/1192 (1.6%)	renal tubule development
72002	1.6197E-4	2.0346E-3	5/36 (13.9%)	19/1192 (1.6%)	Malpighian tubule development
35295	1.8416E-4	2.2734E-3	7/36 (19.4%)	43/1192 (3.6%)	tube development
1655	2.1126E-4	2.5210E-3	5/36 (13.9%)	20/1192 (1.7%)	urogenital system development
72001	2.1126E-4	2.5210E-3	5/36 (13.9%)	20/1192 (1.7%)	renal system development
7417	2.5632E-4	3.0086E-3	8/36 (22.2%)	60/1192 (5.0%)	central nervous system development
48519	3.7339E-4	4.3121E-3	14/36 (38.9%)	181/1192 (15.2%)	negative regulation of biological process
7399	4.3357E-4	4.9276E-3	19/36 (52.8%)	309/1192 (25.9%)	nervous system development
48856	4.5403E-4	5.0794E-3	26/36 (72.2%)	523/1192 (43.9%)	anatomical structure development
32502	5.7537E-4	6.3379E-3	27/36 (75.0%)	564/1192 (47.3%)	developmental process
50673	8.1598E-4	8.7201E-3	3/36 (8.3%)	7/1192 (0.6%)	epithelial cell proliferation
61331	8.1598E-4	8.7201E-3	3/36 (8.3%)	7/1192 (0.6%)	epithelial cell proliferation involved in Malpighian tubule morphogenesis
1709	1.0303E-3	1.0849E-2	6/36 (16.7%)	41/1192 (3.4%)	cell fate determination
45893	1.1107E-3	1.1361E-2	7/36 (19.4%)	57/1192 (4.8%)	positive regulation of transcription, DNA-dependent
51254	1.1107E-3	1.1361E-2	7/36 (19.4%)	57/1192 (4.8%)	positive regulation of RNA metabolic process
10628	1.2356E-3	1.2461E-2	7/36 (19.4%)	58/1192 (4.9%)	positive regulation of gene expression
10557	1.3714E-3	1.3638E-2	7/36 (19.4%)	59/1192 (4.9%)	positive regulation of macromolecule biosynthetic process

Continued on next page

Table A.7 – concluded from previous page

GO-ID	*p*-value	corr *p*-value	cluster freq	total freq	Description
45944	1.5138E-3	1.4498E-2	6/36 (16.7%)	44/1192 (3.7%)	positive regulation of transcription from RNA polymerase II promoter
45935	1.5187E-3	1.4498E-2	7/36 (19.4%)	60/1192 (5.0%)	positive regulation of nucleobase-containing compound metabolic process
51173	1.5187E-3	1.4498E-2	7/36 (19.4%)	60/1192 (5.0%)	positive regulation of nitrogen compound metabolic process
48813	1.7085E-3	1.5887E-2	6/36 (16.7%)	45/1192 (3.8%)	dendrite morphogenesis
16358	1.7085E-3	1.5887E-2	6/36 (16.7%)	45/1192 (3.8%)	dendrite development
32501	1.9488E-3	1.7889E-2	27/36 (75.0%)	600/1192 (50.3%)	multicellular organismal process
45464	2.6119E-3	2.3672E-2	2/36 (5.6%)	3/1192 (0.3%)	R8 cell fate specification
31328	2.9301E-3	2.5901E-2	7/36 (19.4%)	67/1192 (5.6%)	positive regulation of cellular biosynthetic process
10604	2.9301E-3	2.5901E-2	7/36 (19.4%)	67/1192 (5.6%)	positive regulation of macromolecule metabolic process
9891	3.1950E-3	2.7898E-2	7/36 (19.4%)	68/1192 (5.7%)	positive regulation of biosynthetic process
48565	3.4813E-3	2.9674E-2	4/36 (11.1%)	22/1192 (1.8%)	digestive tract development
55123	3.4813E-3	2.9674E-2	4/36 (11.1%)	22/1192 (1.8%)	digestive system development
7419	4.6193E-3	3.8911E-2	3/36 (8.3%)	12/1192 (1.0%)	ventral cord development
35239	5.1632E-3	4.2986E-2	5/36 (13.9%)	39/1192 (3.3%)	tube morphogenesis
31325	5.6031E-3	4.6113E-2	7/36 (19.4%)	75/1192 (6.3%)	positive regulation of cellular metabolic process

Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y.-H. C., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Miklos, Abril, J. F., Agbayani, A., An, H.-J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., Pablos, B. d., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M.-H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D. C., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z.-Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage, T., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R.-F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). The Genome Sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–2195.

Aerts, S., Quan, X.-J., Claeys, A., Naval Sanchez, M., Tate, P., Yan, J., and Hassan, B. A. (2010). Robust Target Gene Discovery through Transcriptome Perturbations and Genome-Wide Enhancer Predictions in *Drosophila* Uncovers a Regulatory Basis for Sensory Specification. *PLoS Biol*, 8(7):e1000435.

Ahn, Y.-Y., Bagrow, J. P., and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764.

Albert, R., Jeong, H., and Barabasi, A.-L. (1999). Internet: Diameter of the World-Wide Web. *Nature*, 401(6749):130–131.

Allen, R. D. (1967). Fine Structure, Reconstruction and Possible Functions of Components of the Cortex of *Tetrahymena Pyriformis*. *J. Protozool.*, 14:553–565.

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., and Dessimoz, C. (2012). Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. *PLoS Comput Biol*, 8(5):e1002514.

Altman, R. B. and Raychaudhuri, S. (2001). Whole-genome expression analysis: challenges beyond clustering. *Current Opinion in Structural Biology*, 11(3):340 – 347.

Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. *Nature*, 426(6966):570–574.

Andrade, M. A. and Bork, P. (1995). HEAT repeats in the Huntington's disease protein. *Nat Genet*, 11(2):115–116.

Arama, E., Agapite, J., and Steller, H. (2003). Caspase Activity and a Specific Cytochrome C Are Required for Sperm Differentiation in *Drosophila*. *Dev Cell*, 4(5):687–697.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, 38(suppl 1):D525–D531.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S. L., Ceol, A., Chautard, E., Dana, J. M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R. E. W., Isserlin, R., Jimenez, R. C., Kerssemakers, J., Khadake, J., Lynn, D. J., Michaut, M., O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G. D., Cesareni, G., Donaldson, I. M., Eisenberg, D., Kleywegt, G. J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, M., and Hermjakob, H. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Meth*, 8(7):528–529.

Arbeitman, M. N., Furlong, E. E. M., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., and White, K. P. (2002). Gene Expression During the Life Cycle of *Drosophila melanogaster*. *Science*, 297(5590):2270–2275.

Arora, S., Ge, R., Sachdeva, S., and Schoenebeck, G. (2012). Finding overlapping communities in social networks: toward a rigorous approach. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 37–54, New York, NY, USA. ACM.

Artavanis-Tsakonas, S., Rand, M. D., and Lake, R. J. (1999). Notch signaling: Cell fate control and signal integration in development. *Science*, 284(5415):770–776.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284.

Avidor-Reiss, T., Maer, A. M., Koundakjian, E., Polyanovsky, A., Keil, T., Subramaniam, S., and Zuker, C. S. (2004). Decoding cilia function: Defining specialized genes required for compartmentalized cilia biogenesis. *Cell*, 117(4):527 – 539.

Baker, R., Kuehl, J., and Wilkinson, G. (2011). The Enhancer of split complex arose prior to the diversification of schizophoran flies and is strongly conserved between *Drosophila* and stalk-eyed flies (*Diopsidae*). *BMC Evolutionary Biology*, 11(1):354.

Baker, S. A., Freeman, K., Luby-Phelps, K., Pazour, G. J., and Besharse, J. C. (2003). Ift20 links kinesin ii with a mammalian intraflagellar transport complex that is conserved in motile flagella and sensory cilia. *Journal of Biological Chemistry*, 278(36):34211–34218.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Barenco, M., Tomescu, D., Brewer, D., Callard, R., Stark, J., and Hubank, M. (2006). Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman & Hall, London.

Basto, R., Lau, J., Vinogradova, T., Gardiol, A., Woods, C., Khodjakov, A., and Raff, J. (2006). Flies without centrioles. *Cell*, 125(4):1375–1386.

Battiti, R. and Protasi, M. (2001). Reactive local search for the maximum clique problem. *Algorithmica*, 29(4):610–637.

Bendixen, C., Gangloff, S., and Rothstein, R. (1994). A yeast mating-selection scheme for detection of protein-protein interactions. *Nucleic Acids Res.*, 11;22(9)(0305-1048 (Linking)):1778–9.

Berg, J., Lassig, M., and Wagner, A. (2004). Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 4(1):51.

Berggård, T., Linse, S., and James, P. (2007). Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7(16):2833–2842.

Bertrand, N., Castro, D. S., and Guillemot, F. (2002). Proneural genes and the specification of neural cell types. *Nat Rev Neurosci*, 3(7):517–530.

Bier, E. (2005). *Drosophila*, the golden bug, emerges as a tool for human genetics. *Nat Rev Genet*, 6(1):9–23.

Bisson, N., James, D. A., Ivosev, G., Tate, S. A., Bonner, R., Taylor, L., and Pawson, T. (2011). Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the grb2 adaptor. *Nat Biotech*, 29(7):653–658.

Blachon, S., Gopalakrishnan, J., Omori, Y., Polyanovsky, A., Church, A., Nicastro, D., Malicki, J., and Avidor-Reiss, T. (2008). *Drosophila* asterless and vertebrate Cep152 are orthologs essential for centriole duplication. *Genetics*, 180(4):2081–2094.

Blacque, O. E., Perens, E. A., Boroevich, K. A., Inglis, P. N., Li, C., Warner, A., Khattra, J., Holt, R. A., Ou, G., Mah, A. K., McKay, S. J., Huang, P., Swoboda, P., Jones, S. J., Marra, M. A., Baillie, D. L., Moerman, D. G., Shaham, S., and Leroux, M. R. (2005). Functional genomics of the cilium, a sensory organelle. *Curr Biol*, 15(10):935–941.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *Journal of Molecular Biology*, 283(4):707–725.

Borlado, L. R. and Méndez, J. (2008). CDC6: from DNA replication to cell cycle checkpoints and oncogenesis. *Carcinogenesis*, 29(2):237–243.

Bosco, G., Du, W., and Orr-Weaver, T. (2001). DNA replication control through interaction of E2F-RB and the origin recognition complex. *Nature Cell Biology*, 3(3):289–295.

Brand, A. and Perrimon, N. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development*, 118(2):401–415.

Bray, D. (2003). Molecular networks: The top-down view. *Science*, 301(5641):1864–1865.

Breitkreutz, B.-J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bähler, J., Wood, V., Dolinski, K., and Tyers, M. (2008). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Research*, 36(suppl 1):D637–D640.

Bridges, C. B. and Morgan, T. H. (1923). *The third-chromosome group of mutant characters in Drosophila melanogaster*. Carnegie Institution of Washington.

Broadhead, R., Dawe, H. R., Farr, H., Griffiths, S., Hart, S. R., Portman, N., Shaw, M. K., Ginger, M. L., Gaskell, S. J., McKean, P. G., and Gull, K. (2006). Flagellar motility is required for the viability of the bloodstream *trypanosome*. *Nature*, 440(7081):224–227.

Brohee, S. and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7(1):488.

Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082.

Brunato, M., Hoos, H. H., and Battiti, R. (2008). On effectively finding maximal quasi-cliques in graphs. In Maniezzo, V., Battiti, R., and Watson, J.-P., editors, *Learning and Intelligent Optimization*, pages 41–55. Springer-Verlag, Berlin, Heidelberg.

Brunet, J.-F. and Ghysen, A. (1999). Deconstructing cell determination: proneural genes and neuronal identity. *BioEssays*, 21(4):313–318.

Bulman, M. P., Kusumi, K., Frayling, T. M., McKeown, C., Garrett, C., Lander, E. S., Krumlauf, R., Hattersley, A. T., Ellard, S., and Turnpenny, P. D. (2000). Mutations in the human Delta homologue, DLL3, cause axial skeletal defects in spondylocostal dysostosis. *Nat Genet*, 24(4):438–441.

Cachero, S., Simpson, T. I., zur Lage, P. I., Ma, L., Newton, F. G., Holohan, E. E., Armstrong, J. D., and Jarman, A. P. (2011). The Gene Regulatory Cascade Linking Proneural Specification with Differentiation in *Drosophila* Sensory Neurons. *PLoS Biol*, 9(1):e1000568.

Camargo, L. M., Collura, V., Rain, J.-C., Mizuguchi, K., Hermjakob, H., Kerrien, S., Bonnert, T. P., Whiting, P. J., and Brandon, N. J. (2006). Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia. *Mol Psychiatry*, 12(1):74–86.

Campuzano, S. and Modolell, J. (1992). Patterning of the *Drosophila* nervous system: the achaete-scute gene complex. *Trends in Genetics*, 8(6):202 – 208.

Carreira, V., Soto, I., Mensch, J., and Fanara, J. (2011). Genetic basis of wing morphogenesis in *Drosophila*: sexual dimorphism and non-allometric effects of shape variation. *BMC Developmental Biology*, 11:32–.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38(suppl 1):D532–D539.

Cerami, E., Demir, E., Schultz, N., Taylor, B. S., and Sander, C. (2010). Automated Network Analysis Identifies Core Pathways in Glioblastoma. *PLoS ONE*, 5(2):e8918.

Cesareni, G., Ceol, A., Gavrila, C., Palazzi, L. M., Persico, M., and Schneider, M. V. (2005). Comparative interactomics. *FEBS Letters*, 579(8):1828–1833.

Chang, P.-J., Hsiao, Y.-L., Tien, A.-C., Li, Y.-C., and Pi, H. (2008). Negative-feedback regulation of proneural proteins controls the timing of neural precursor division. *Development*, 135(18):3021–3030.

Chautard, E., Ballut, L., Thierry-Mieg, N., and Ricard-Blum, S. (2009). MatrixDB, a database focused on extracellular protein–protein and protein–carbohydrate interactions. *Bioinformatics*, 25(5):690–691.

Chintapalli, V., Wang, J., and Dow, J. (2007). Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nature Genetics*, 39(6):715–720.

Chodhari, R., Mitchison, H., and Meeks, M. (2004). Cilia, primary ciliary dyskinesia and molecular genetics. *Paediatric Respiratory Reviews*, 5(1):69 – 76.

Christensen, T. and Tye, B. (2003). *Drosophila* MCM10 interacts with members of the prereplication complex and is required for proper chromosome condensation. *Molecular Biology of the Cell*, 14(6):2206–2215.

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6)(6):066111.

Clauset, A., Shalizi, C., and Newman, M. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703.

Coates, P. and Hall, P. (2003). The yeast two-hybrid system for identifying protein–protein interactions. *The Journal of Pathology*, 199(1):4–7.

Collins, KL nd Russo, A., Tseng, B., and Kelly, T. (1993). The role of the 70 kDa subunit of human DNA polymerase alpha in DNA replication. *EMBO J.*, 12(12):4555–66.

Cruz, C., Ribes, V., Kutejova, E., Cayuso, J., Lawson, V., Norris, D., Stevens, J., Davey, M., Blight, K., Bangs, F., Mynett, A., Hirst, E., Chung, R., Balaskas, N., Brody, S. L., Marti, E., and Briscoe, J. (2010). Foxj1 regulates floor plate cilia architecture and modifies the response of cells to sonic hedgehog signalling. *Development*, 137(24):4271–4282.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Cullmann, G., Fien, K., Kobayashi, R., and Stillman, B. (1995). Characterization of the five replication factor C genes of *Saccharomyces cerevisiae*. *Molecular and Cellular Biology*, 15(9):4661–71.

Cusick, M. E., Klitgord, N., Vidal, M., and Hill, D. E. (2005). Interactome: gateway into systems biology. *Human Molecular Genetics*, 14(suppl 2):R171–R181.

Date, S. V. and Stoeckert, C. J. (2006). Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale. *Genome Research*, 16(4):542–549.

de Paulo Castro Teixeira, V., Blattner, S. M., Li, M., Anders, H.-J., Cohen, C. D., Edenhofer, I., Calvaresi, N., Merkle, M., Rastaldi, M. P., and Kretzler, M. (2005). Functional consequences of integrin-linked kinase activation in podocyte damage. *Kidney Int*, 67(2):514–523.

De Smet, R. and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat Rev Micro*, 8(10):717–729.

Ding, C., He, X., and Peng, H. (2005). Finding cliques in protein interaction networks via transitive closure of a weighted graph. In *Proceedings of the 5th international workshop on Bioinformatics*, BIOKDD '05, pages 69–75, New York, NY, USA. ACM.

Dokucu, M., Zipursky, S., and Cagan, R. (1996). Atonal, rough and the resolution of proneural clusters in the developing *Drosophila* retina. *Development*, 122(12):4139–4147.

Dondelinger, F., Lebre, S., and Husmeier, D. (2010). Heterogeneous Continuous Dynamic Bayesian Networks with Flexible Structure and Inter-Time Segment Information Sharing. In Furnkranz J., J. T., editor, *International Conference on Machine Learning (ICML)*, pages 303–310.

Dorstyn, L., Colussi, P. A., Quinn, L. M., Richardson, H., and Kumar, S. (1999). DRONC, an ecdysone-inducible *Drosophila* caspase. *Proceedings of the National Academy of Sciences*, 96(8):4307–4312.

Dubruille, R., Laurençon, A., Vandaele, C., Shishido, E., Coulon-Bublex, M., Swoboda, P., Couble, P., Kernan, M., and Durand, B. (2002). *Drosophila* Regulatory factor X is necessary for ciliated sensory neuron differentiation. *Development*, 129(23):5487–5498.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., and Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–749.

Dunne, J. A., Williams, R. J., and Martinez, N. D. (2002). Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917–12922.

Efimenko, E., Bubb, K., Mak, H. Y., Holzman, T., Leroux, M. R., Ruvkun, G., Thomas, J. H., and Swoboda, P. (2005). Analysis of xbx genes in *C. elegans*. *Development*, 132(8):1923–1934.

Emery, P., Durand, B., Mach, B., and Reith, W. (1996). RFX Proteins, a Novel Family of DNA Binding Proteins Conserved in the Eukaryotic Kingdom. *Nucleic Acids Research*, 24(5):803–807.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584.

Erdos, P. and Renyi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61.

Ernst, J., Nau, G. J., and Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, 21(suppl 1):i159–i168.

Evans, T. S. and Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Phys. Rev. E*, 80:016105.

Everitt, B. S., Landau, S., and Leese, M. (2009). *Cluster Analysis*. Wiley.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*, 3:–.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the Internet topology. *SIGCOMM Comput. Commun. Rev.*, 29(4):251–262.

Fiedler, M. (1973). Algebraic Connectivity of Graphs. *Czechoslovak Mathematical Journal, Praha*, 23 (98):298–305.

Fielding, R. T. and Taylor, R. N. (2000). Principled design of the modern Web architecture. In *ICSE '00: Proceedings of the 22nd international conference on Software engineering*, pages 407–416, New York, NY, USA. ACM.

Figeys, D., McBroom, L. D., and Moran, M. F. (2001). Mass Spectrometry for the Study of Protein-Protein Interactions. *Methods*, 24(3):230 – 239.

Fletcher, R. J., Bishop, B. E., Leon, R. P., Sclafani, R. A., Ogata, C. M., and Chen, X. S. (2003). The structure and function of MCM from archaeal *M.Thermoautotrophicum*. *Nat Struct Mol Biol*, 10(3):160–167.

Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Koscielny, G., Kulesha, E., Lawson, D., Longden, I., Massingham, T., McLaren, W., Megy, K., Overduin, B., Pritchard, B., Rios, D., Ruffier, M., Schuster, M., Slater, G., Smedley, D., Spudich, G., Tang, Y. A., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S. P., Zadissa, A., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Smith, J., and Searle, S. M. J. (2010). Ensembl's 10th year. *Nucleic Acids Research*, 38(suppl 1):D557–D562.

FlyBase Curators, Members, S.-P. P., and Members, I. P. (2004). Gene Ontology annotation in FlyBase through association of InterPro records with GO terms.

Follit, J. A., Tuft, R. A., Fogarty, K. E., and Pazour, G. J. (2006). The Intraflagellar Transport Protein IFT20 Is Associated with the Golgi Complex and Is Required for Cilia Assembly. *Molecular Biology of the Cell*, 17(9):3781–3792.

Foote, M., Hunter, J. P., Janis, C. M., and Sepkoski Jr., J. J. (1999). Evolutionary and Preservational Constraints on Origins of Biologic Groups: Divergence Times of Eutherian Mammals. *Science*, 283(5406):1310–1314.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., Jacq, B., Arpin, M., Bellaiche, Y., Bellusci, S., Benaroch, P., Bornens, M., Chanet, R., Chavrier, P., Delattre, O., Doye, V., Fehon, R., Faye, G., Galli, T., Girault, J.-A., Goud, B., de Gunzburg, J., Johannes, L., Junier, M.-P., Mirouse, V., Mukherjee, A., Papadopoulo, D., Perez, F., Plessis, A., Rossé, C., Saule, S., Stoppa-Lyonnet, D., Vincent, A., White, M., Legrain, P., Wojcik, J., Camonis, J., and Daviet, L. (2005). Protein interaction mapping: A *Drosophila* case study. *Genome Research*, 15(3):376–384.

Forslund, K., Pekkari, I., and Sonnhammer, E. (2011). Domain architecture conservation in orthologs. *BMC Bioinformatics*, 12(1):326.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3–5):75 – 174.

Franke, L., van Bakel, H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M., and Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78(6):1011–1025.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary Rate in the Protein Interaction Network. *Science*, 296(5568):750–752.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4)(1066-5277 (Linking)):601–620.

Friedman, N., Murphy, K., and Russell, S. (1998). Learning the structure of dynamic probabilistic networks. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, UAI'98, pages 139–147, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gajiwala, K. S., Chen, H., Cornille, F., Roques, B. P., Reith, W., Mach, B., and Burley, S. K. (2000). Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. *Nature*, 403(6772):916–921.

Gallone, G., Simpson, T. I., Armstrong, J. D., and Jarman, A. (2011). `Bio::Homology::InterologWalk` - A Perl module to build putative protein-protein interaction networks through interolog mapping. *BMC Bioinformatics*, 12(1):289.

Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–293.

Garcia-Bellido, A. (1979). Genetic Analysis of the Achaete-Scute System of *Drosophila melanogaster*. *Genetics*, 91(3):491–520.

Garcia-Bellido, A. and Santamaria, P. (1978). Developmental Analysis of the Achaete-Scute System of *Drosophila melanogaster*. *Genetics*, 88(3):469–486.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M. A., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M., and Barrell, B. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511.

Gaudet, R. (2008). TRP channels entering the structural era. *The Journal of Physiology*, 586(15):3565–3575.

Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet*, 29(4):482–486.

Gees, M., Colsoul, B., and Nilius, B. (2010). The Role of Transient Receptor Potential Cation Channels in Ca2+ Signaling. *Cold Spring Harbor Perspectives in Biology*, 2(10):1–31.

Gerdes, J. M., Davis, E. E., and Katsanis, N. (2009). The Vertebrate Primary Cilium in Development, Homeostasis, and Disease. *Cell*, 137(1):32–45.

Geremek, M., Bruinenberg, M., Ziętkiewicz, E., Pogorzelski, A., Witt, M., and Wijmenga, C. (2011). Gene expression studies in cells from primary ciliary dyskinesia patients identify 208 potential ciliary genes. *Human Genetics*, 129:283–293.

Ghahramani, Z. (1998). Learning dynamic Bayesian networks. In *Adaptive Processing of Sequences and Data Structures*, pages 168–197. Springer-Verlag.

Gherman, A., Davis, E. E., and Katsanis, N. (2006). The ciliary proteome database: an integrated community resource for the genetic and functional dissection of cilia. *Nat Genet*, 38(9):961–962.

Ghysen, A. and Dambly-Chaudiere, C. (1988). From DNA to form: the achaete-scute complex. *Genes & Development*, 2(5):495–501.

Ghysen, A. and Dambly-Chaudiere, C. (1989). Genesis of the *Drosophila* peripheral nervous system. *Trends Genet*, 5(8):251–255.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., J., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A Protein Interaction Map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826.

Gleiser, P. and Danon, L. (2003). Community Structure in Jazz. *Advances in Complex Systems*, 6:565–573.

Goetz, M. P., Toft, D. O., Ames, M. M., and Erlichman, C. (2003). The Hsp90 chaperone complex as a novel target for cancer therapy. *Annals of Oncology*, 14(8):1169–1176.

Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H., Lamb, B. T., and Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics*, 24(15):1743–1744.

Gong, Z., Son, W., Doo Chung, Y., Kim, J., Shin, D. W., McClung, C. A., Lee, Y., Lee, H. W., Chang, D.-J., Kaang, B.-K., Cho, H., Oh, U., Hirsh, J., Kernan, M. J., and Kim, C. (2004). Two Interdependent TRPV Channel Subunits, Inactive and Nanchung, Mediate Hearing in *Drosophila*. *The Journal of Neuroscience*, 24(41):9059–9066.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Systematic Zoology*, 28(2):132–163.

Gopfert, M. C., Albert, J. T., Nadrowski, B., and Kamikouchi, A. (2006). Specification of auditory sensitivity by *Drosophila* TRP channels. *Nat Neurosci*, 9(8):999–1000.

Goulding, S. E., White, N. M., and Jarman, A. P. (2000a). *cato* Encodes a Basic Helix-Loop-Helix Transcription Factor Implicated in the Correct Differentiation of *Drosophila* Sense Organs. *Developmental Biology*, 221(1):120 – 131.

Goulding, S. E., zur Lage, P., and Jarman, A. P. (2000b). *amos*, a Proneural Gene for *Drosophila* Olfactory Sense Organs that Is Regulated by *lozenge*. *Neuron*, 25(1):69–78.

Graser, S., Stierhof, Y.-D., Lavoie, S. B., Gassner, O. S., Lamla, S., Le Clech, M., and Nigg, E. A. (2007). Cep164, a novel centriole appendage protein required for primary cilium formation. *The Journal of Cell Biology*, 179(2):321–330.

Grzegorczyk, M. and Husmeier, D. (2011). Non-homogeneous dynamic Bayesian networks for continuous data. *Mach. Learn.*, 83(3):355–419.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press, Cambridge, MA.

Guindon, S. and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5):696–704.

Hahn, A., Rahnenfuhrer, J., Talwar, P., and Lengauer, T. (2005). Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics*, 6(1):112.

Hakes, L., Pinney, J. W., Robertson, D. L., and Lovell, S. C. (2008). Protein-protein interaction networks and biology — what's the connection? *Nat Biotech*, 26(1):69–72.

Hall, N., Pain, A., Berriman, M., Churcher, C., Harris, B., Harris, D., Mungall, K., Bowman, S., Atkin, R., Baker, S., Barron, A., Brooks, K., Buckee, C. O., Burrows, C., Cherevach, I., Chillingworth, C., Chillingworth, T., Christodoulou, Z., Clark, L., Clark, R., Corton, C., Cronin, A., Davies, R., Davis, P., Dear, P., Dearden, F., Doggett, J., Feltwell, T., Goble, A., Goodhead, I., Gwilliam, R., Hamlin, N., Hance, Z., Harper, D., Hauser, H., Hornsby, T., Holroyd, S., Horrocks, P., Humphray, S., Jagels, K., James, K. D., Johnson, D., Kerhornou, A., Knights, A., Konfortov, B., Kyes, S., Larke, N., Lawson, D., Lennard, N., Line, A., Maddison, M., McLean, J., Mooney, P., Moule, S., Murphy, L., Oliver, K., Ormond, D., Price, C.,

Quail, M. A., Rabbinowitsch, E., Rajandream, M.-A., Rutter, S., Rutherford, K. M., Sanders, M., Simmonds, M., Seeger, K., Sharp, S., Smith, R., Squares, R., Squares, S., Stevens, K., Taylor, K., Tivey, A., Unwin, L., Whitehead, S., Woodward, J., Sulston, J. E., Craig, A., Newbold, C., and Barrell, B. G. (2002). Sequence of *Plasmodium falciparum* chromosomes 1, 3-9 and 13. *Nature*, 419(6906):527–531.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P., and Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93.

Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotech*, 23(7):839–844.

Han, Y.-G., Kwok, B. H., and Kernan, M. J. (2003). Intraflagellar Transport Is Required in *Drosophila* to Differentiate Sensory Cilia but Not Sperm. *Curr Biol*, 13(19):1679–1686.

Hannigan, G., Troussard, A. A., and Dedhar, S. (2005). Integrin-linked kinase: a cancer therapeutic target unique among its ILK. *Nat Rev Cancer*, 5(1):51–63.

Harr, B. (2001). *Drosophila melanogaster* strain pe1 cramped (*crm*) gene, complete cds.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402:C47–C52.

He, F., Zhang, Y., Chen, H., Zhang, Z., and Peng, Y.-L. (2008). The prediction of protein-protein interaction networks in rice blast fungus. *BMC Genomics*, 9(1):519.

He, X. and Zhang, J. (2005). Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. *Genetics*, 169(2):1157–1164.

Hegyi, H. and Gerstein, M. (2001). Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins. *Genome Research*, 11(10):1632–1640.

Henricson, A., Forslund, K., and Sonnhammer, E. (2010). Orthology confers intron position conservation. *BMC Genomics*, 11(1):412.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004). The HUPO PSI's Molecular Interaction format — a community standard for the representation of protein interaction data. *Nat Biotech*, 22(2):177–183.

Hildebrandt, F., Benzing, T., and Katsanis, N. (2011). Ciliopathies. *New England Journal of Medicine*, 364(16):1533–1543.

Hittinger, C. T. and Carroll, S. B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, 449(7163):677–681.

Holland, P. W. and Leinhardt, S. (1971). Transitivity in Structural Models of Small Groups. *Small Group Research*, 2(2):107–124.

Honkela, A., Gao, P., Ropponen, J., Rattray, M., and Lawrence, N. D. (2011). *tigre*: Transcription factor inference through Gaussian process reconstruction of expression for bioconductor. *Bioinformatics*, 27(7):1026–1027.

Honkela, A., Girardot, C., Gustafson, E. H., Liu, Y.-H., Furlong, E. E. M., Lawrence, N. D., and Rattray, M. (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences*, 107(17):7793–7798.

Huang, T.-W., Lin, C.-Y., and Kao, C.-Y. (2007). Reconstruction of human protein interolog network using evolutionary conserved network. *BMC Bioinformatics*, 8(1):152.

Huang, T.-W., Tien, A.-C., Huang, W.-S., Lee, Y.-C. G., Peng, C.-L., Tseng, H.-H., Kao, C.-Y., and Huang, C.-Y. F. (2004). POINT: a database for the prediction of protein–protein interactions based on the orthologous interactome. *Bioinformatics*, 20(17):3273–3276.

Hulsen, T., Huynen, M., de Vlieg, J., and Groenen, P. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, 7(4):R31.

Ideker, T. and Krogan, N. J. (2012). Differential network biology. *Mol Syst Biol*, 8:–.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233–S240.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science*, 292(5518):929–934.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks. In *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, CSB '03, pages 104–, Washington, DC, USA. IEEE Computer Society.

Ishikawa, H. and Marshall, W. F. (2011). Ciliogenesis: building the cell's antenna. *Nat Rev Mol Cell Biol*, 12(4):222–234.

Ispolatov, I., Krapivsky, P. L., and Yuryev, A. (2005). Duplication-divergence model of protein interaction network. *Phys. Rev. E*, 71:061911.

Jacquet, B. V., Salinas-Mondragon, R., Liang, H., Therit, B., Buie, J. D., Dykstra, M., Campbell, K., Ostrowski, L. E., Brody, S. L., and Ghashghaei, H. T. (2009). FoxJ1-dependent gene expression is required for differentiation of radial glia into ependymal cells and a subset of astrocytes in the postnatal brain. *Development*, 136(23):4021–4031.

Jain, S. and Bader, G. (2010). An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, 11(1):562.

Jana, S. C., Girotra, M., and Ray, K. (2011). Heterotrimeric kinesin-II is necessary and sufficient to promote different stepwise assembly of morphologically distinct bipartite cilia in *Drosophila* antenna. *Molecular Biology of the Cell*, 22(6):769–781.

Janke, W., Berg, B. A., and Billoire, A. (2003). Extreme order statistics. *Nuclear Physics B - Proceedings Supplements*, 119(0):867 – 869. <ce:title>Proceedings of the XXth International Symposium on Lattice Field Theory</ce:title>.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating Whole-Genome Expression Data with Protein-Protein Interactions. *Genome Research*, 12(1):37–46.

Jarman, A., Sun, Y., Jan, L., and Jan, Y. (1995). Role of the proneural gene, *atonal*, in formation of *Drosophila* chordotonal organs and photoreceptors. *Development*, 121(7):2019–2030.

Jarman, A. P. and Ahmed, I. (1998). The specificity of proneural genes in determining *Drosophila* sense organ identity. *Mechanisms of Development*, 76(1-2):117–125.

Jarman, A. P., Grau, Y., Jan, L. Y., and Jan, Y. N. (1993). *atonal* is a proneural gene that directs chordotonal organ formation in the *Drosophila* peripheral nervous system. *Cell*, 73(7):1307–1321.

Jarman, A. P., Grell, E. H., Ackerman, L., Jan, L. Y., and Jan, Y. N. (1994). *atonal* is the proneural gene for *Drosophila* photoreceptors. *Nature*, 369(6479):398–400.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl 1):D412–D416.

Jeong, H., Mason, S. P., Barabasi, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830):1497–1502.

Kafri, R., Levy, M., and Pilpel, Y. (2006). The regulatory utilization of genetic redundancy through responsive backup circuits. *Proceedings of the National Academy of Sciences*, 103(31):11653–11658.

Kageyama, R., Ohtsuka, T., and Kobayashi, T. (2007). The Hes gene family: repressors and oscillators that orchestrate embryogenesis. *Development*, 134(7):1243–1251.

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database*, 2011:1–3.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theoret. Biol.*, 22(0022-5193 (Linking)):437–467.

Kavlie, R. G., Kernan, M. J., and Eberl, D. F. (May 2010). Hearing in *Drosophila* Requires TilB, a Conserved Protein Associated With Ciliary Motility. *Genetics*, 185(1):177–188.

Kemmer, D., Huang, Y., Shah, S., Lim, J., Brumm, J., Yuen, M., Ling, J., Xu, T., Wasserman, W., and Ouellette, B. F. (2005). Ulysses — an application for the projection of molecular interactions across species. *Genome Biology*, 6(12):R106.

Kernighan, B. W. and Lin, S. (1970). An Efficient Heuristic Procedure for Partitioning Graphs. *The Bell system technical journal*, 49(1):291–307.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A., Vinod, N., Bader, G., Xenarios, I., Wojcik, J., Sherman, D., Tyers, M., Salama, J., Moore, S., Ceol, A., Chatraryamontri, A., Oesterheld, M., Stumpflen, V., Salwinski, L., Nerothin, J., Cerami, E., Cusick, M., Vidal, M., Gilson, M., Armstrong, J., Woollard, P., Hogue, C., Eisenberg, D., Cesareni, G., Apweiler, R., and Hermjakob, H. (2007). Broadening the horizon — level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biology*, 5(1):44.

Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I., and Apweiler, R. (2005). Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Research*, 33(suppl 1):D297–D302.

Kersey, P. J., Lawson, D., Birney, E., Derwent, P. S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kähäri, A., Kinsella, R. J., Kulesha, E., Maheswari, U., Megy, K., Nuhn, M., Proctor, G., Staines, D., Valentin, F., Vilella, A. J., and Yates, A. (2010). Ensembl Genomes: Extending Ensembl across the taxonomic space. *Nucleic Acids Research*, 38(suppl 1):D563–D569.

Kim, J., Chung, Y. D., Park, D.-y., Choi, S., Shin, D. W., Soh, H., Lee, H. W., Son, W., Yim, J., Park, C.-S., Kernan, M. J., and Kim, C. (2003). A TRPV family ion channel required for hearing in *Drosophila*. *Nature*, 424(6944):81–84.

Kim, J. C., Badano, J. L., Sibold, S., Esmail, M. A., Hill, J., Hoskins, B. E., Leitch, C. C., Venner, K., Ansley, S. J., Ross, A. J., Leroux, M. R., Katsanis, N., and Beales, P. L. (2004). The Bardet-Biedl protein BBS4 targets cargo to the pericentriolar region and is required for microtubule anchoring and cell cycle progression. *Nat Genet*, 36(5):462–470.

Kimura, M. (1991). Recent development of the neutral theory viewed from the wrightian tradition of theoretical population genetics. *Proceedings of the National Academy of Sciences of the United States of America*, 88(14):5969–5973.

Kishimoto, N., Cao, Y., Park, A., and Sun, Z. (2008). Cystic Kidney Gene *seahorse* Regulates Cilia-Mediated Processes and Wnt Pathways. *Dev Cell*, 14(6):954–961.

Klemm, R. D., Austin, R. J., and Bell, S. P. (1997). Coordinate Binding of ATP and Origin DNA Regulates the ATPase Activity of the Origin Recognition Complex. *Cell*, 88(4):493–502.

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics*, 39(1):309–338.

Kosman, D., Ip, Y., Levine, M., and Arora, K. (1991). Establishment of the mesoderm-neuroectoderm boundary in the *Drosophila* embryo. *Science*, 254(5028):118–122.

Kozminski, K. G., Johnson, K. A., Forscher, P., and Rosenbaum, J. L. (1993). A motility in the eukaryotic flagellum unrelated to flagellar beating. *Proceedings of the National Academy of Sciences*, 90(12):5519–5523.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S.,

Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440(7084):637–643.

Kuechle, M., Predd, H., Fleckman, P., Dale, B., and Presland, R. (2001). Caspase-14, a keratinocyte specific caspase: mRNA splice variants and expression pattern in embryonic and adult mouse. *Cell Death Differ.*, 8(1350-9047 (Linking)):868–870.

Kumar, R., Musiyenko, A., and Barik, S. (2003). The heat shock protein 90 of *Plasmodium falciparum* and antimalarial activity of its inhibitor, geldanamycin. *Malaria Journal*, 2(1):30.

Kusumi, K., Sun, E., Kerrebrock, A. W., Bronson, R. T., Chi, D.-C., Bulotsky, M., Spencer, J. B., Birren, B. W., Frankel, W. N., and Lander, E. S. (1998). The mouse pudgy mutation disrupts Delta homologue Dll3 and initiation of early somite boundaries. *Nat Genet*, 19(3):274–278.

LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S., and Hughes, R. E. (2005). A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature*, 438(7064):103–107.

Lambiotte, R., Sinatra, R., Delvenne, J.-C., Evans, T. S., Barahona, M., and Latora, V. (2011). Flow graphs: Interweaving dynamics and structure. *Phys. Rev. E*, 84(1):017102–.

Lancichinetti, A., Radicchi, F., and Ramasco, J. J. (2010). Statistical significance of communities in networks. *Phys. Rev. E*, 81:046110.

Laurencon, A., Dubruille, R., Efimenko, E., Grenier, G., Bissett, R., Cortier, E., Rolland, V., Swoboda, P., and Durand, B. (2007). Identification of novel regulatory factor X (RFX) target genes by comparative genomics in *Drosophila* species. *Genome Biology*, 8(9):R195.

Lawrence, N. D., Sanguinetti, G., and Rattray, M. (2007). Modelling transcriptional regulation using Gaussian processes. In *Advances in Neural Information Processing Systems*.

Lee, E., Sivan-Loukianova, E., Eberl, D. F., and Kernan, M. J. (2008). An IFT-A Protein Is Required to Delimit Functionally Distinct Zones in Mechanosensory Cilia. *Current Biology*, 18(24):1899 – 1906.

Lee, J. E. (1997). Basic helix-loop-helix genes in neural development. *Current Opinion in Neurobiology*, 7(1):13–20.

Lee, M. S. (1999). Molecular clock calibrations and metazoan divergence dates. *J Mol Evol*, 49(3)(0022-2844 (Linking)):385–91.

Lehner, B. and Fraser, A. (2004). A first-draft human protein-interaction map. *Genome Biology*, 5(9):R63.

Li, J. B., Gerdes, J. M., Haycraft, C. J., Fan, Y., Teslovich, T. M., May-Simera, H., Li, H., Blacque, O. E., Li, L., Leitch, C. C., Lewis, R. A., Green, J. S., Parfrey, P. S., Leroux, M. R., Davidson, W. S., Beales, P. L., Guay-Woodford, L. M., Yoder, B. K., Stormo, G. D., Katsanis, N., and Dutcher, S. K. (2004a). Comparative Genomics Identifies a Flagellar and Basal Body Proteome that Includes the BBS5 Human Disease Gene. *Cell*, 117(4):541 – 552.

Li, L., Krantz, I. D., Deng, Y., Genin, A., Banta, A. B., Collins, C. C., Qi, M., Trask, B. J., Kuo, W. L., Cochran, J., Costa, T., Pierpont, M. E. M., Rand, E. B., Piccoli, D. A., Hood, L., and Spinner, N. B. (1997). Alagille syndrome is caused by mutations in human Jagged1, which encodes a ligand for Notch1. *Nat Genet*, 16(3):243–251.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13(9):2178–2189.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., van den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004b). A Map of the Interactome Network of the Metazoan *C. elegans*. *Science*, 303(5657):540–543.

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., and Aberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907–908.

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabo, G., Rual, J.-F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., Barabasi, A.-L., Vidal, M., and Zoghbi, H. Y. (2006). A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration. *Cell*, 125(4):801–814.

Liu, X., Milo, M., Lawrence, N. D., and Rattray, M. (2005). A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644.

Lonetto, M., Gribskov, M., and Gross, C. A. (1992). The sigma 70 family: sequence conservation and evolutionary relationships. *Journal of Bacteriology*, 174(12):3843–3849.

Lynch, M. and Conery, J. S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, 290(5494):1151–1155.

Lynn, D. J., Winsor, G. L., Chan, C., Richard, N., Laird, M. R., Barsky, A., Gardy, J. L., Roche, F. M., Chan, T. H. W., Shah, N., Lo, R., Naseer, M., Que, J., Yau, M., Acab, M., Tulpan, D., Whiteside, M. D., Chikatamarla, A., Mah, B., Munzner, T., Hokamp, K., Hancock, R. E. W., and Brinkman, F. S. L. (2008). InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol Syst Biol*, 4:–.

Ma, L. and Jarman, A. P. (2011). Dilatory is a *Drosophila* protein related to AZI1 (CEP131) that is located at the ciliary base and required for cilium formation. *Journal of Cell Science*, 124(15):2622–2630.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, 21(16):3448–3449.

Maetschke, S. R., Simonsen, M., Davis, M. J., and Ragan, M. A. (2012). Gene ontology-driven inference of protein–protein interactions using inducers. *Bioinformatics*, 28(1):69–75.

Makunin, I., Volkova, E., Belyaeva, E., Nabirochkina, E., Pirrotta, V., and Zhimulev, I. (2002). The *Drosophila* suppressor of underreplication protein binds to late-replicating regions of polytene chromosomes. *Genetics*, 160(3):1023–1034.

Maslov, S. and Sneppen, K. (2002). Specificity and Stability in Topology of Protein Networks. *Science*, 296(5569):910–913.

Matsuda, K., Makise, M., Sueyasu, Y., Takehara, M., Asano, T., and Mizushima, T. (2007). Yeast two-hybrid analysis of the origin recognition complex of *Saccharomyces cerevisiae*: interaction between subunits and identification of binding proteins. *FEMS Yeast Research*, 7(8):1263–1269.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Research*, 37(suppl 1):D619–D622.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or "Interologs". *Genome Research*, 11(12):2120–2126.

McClintock, T. S., Glasser, C. E., Bose, S. C., and Bergman, D. A. (2008). Tissue expression patterns identify mouse cilia genes. *Physiological Genomics*, 32(2):198–206.

Michaut, M., Kerrien, S., Montecchi-Palazzi, L., Chauvat, F., Cassier-Chauvat, C., Aude, J.-C., Legrain, P., and Hermjakob, H. (2008). InteroPORC: automated inference of highly conserved protein interaction networks. *Bioinformatics*, 24(14):1625–1631.

Mika, S. and Rost, B. (2006). Protein–Protein Interactions More Conserved within Species than across Species. *PLoS Comput Biol*, 2(7):e79.

Minke, B. (2010). The History of the *Drosophila* TRP Channel: The Birth of a New Channel Superfamily. *Journal of Neurogenetics*, 24(4):216–233.

Mirshahvalad, A., Lindholm, J., Derlén, M., and Rosvall, M. (2012). Significant Communities in Large Sparse Networks. *PLoS ONE*, 7(3):e33721.

Mitsis, P., Kowalczykowski, S., and Lehman, I. (1993). A single-stranded DNA binding protein from *Drosophila melanogaster*: characterization of the heterotrimeric protein and its interaction with single-stranded DNA. *Biochemistry*, 32(19):5257–5266.

Morgan, G. W., Denny, P. W., Vaughan, S., Goulding, D., Jeffries, T. R., Smith, D. F., Gull, K., and Field, M. C. (2005). An Evolutionarily Conserved Coiled-Coil Protein Implicated in Polycystic Kidney Disease Is Involved in Basal Body Duplication and Flagellar Biogenesis in *Trypanosoma brucei*. *Molecular and Cellular Biology*, 25(9):3774–3783.

Morgan, T. H. (1917). The Theory of the Gene. *The American Naturalist*, 51(609):513–544.

Muqit, M. M. K. and Feany, M. B. (2002). Modelling neurodegenerative diseases in *Drosophila*: a fruitful approach? *Nat Rev Neurosci*, 3(3):237–243.

Murali, T., Pacifico, S., Yu, J., Guest, S., Roberts, G. G., and Finley, R. L. (2011). DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Research*, 39(suppl 1):D736–D743.

Murayama, T., Toh, Y., Ohshima, Y., and Koga, M. (2005). The *dyf-3* Gene Encodes a Novel Protein Required for Sensory Cilium Formation in *Caenorhabditis elegans*. *Journal of Molecular Biology*, 346(3):677 – 687.

Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., Jan, L. Y., Jan, Y. N., Cabrera, C. V., Buskin, J. N., Hauschka, S. D., Lassar, A. B., Weintraub, H., and Baltimore, D. (1989). Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell*, 58(3):537–544.

Nachury, M. V., Loktev, A. V., Zhang, Q., Westlake, C. J., Peränen, J., Merdes, A., Slusarski, D. C., Scheller, R. H., Bazan, J. F., Sheffield, V. C., and Jackson, P. K. (2007). A Core Complex of BBS Proteins Cooperates with the GTPase Rab8 to Promote Ciliary Membrane Biogenesis. *Cell*, 129(6):1201–1213.

Nehrt, N. L., Clark, W. T., Radivojac, P., and Hahn, M. W. (2011). Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS Comput Biol*, 7(6):e1002073.

Newman, M. E. J. (2001). The Structure of Scientific Collaboration Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):pp. 404–409.

Newman, M. E. J. (2004a). Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133.

Newman, M. J. (2004b). Detecting community structure in networks. *Eur. Phys. J. B*, 38:321–330.

Newton, F. G., zur Lage, P. I., Karak, S., Moore, D. J., Göpfert, M. C., and Jarman, A. P. (2012). Forkhead Transcription Factor Fd3F Cooperates with Rfx to Regulate a Gene Expression Program for Mechanosensory Cilia Specialization. *Dev Cell*, 22(6):1221–1233.

Oeffner, F., Moch, C., Neundorf, A., Hofmann, J., Koch, M., and Grzeschik, K. H. (2008). Novel interaction partners of Bardet-Biedl syndrome proteins. *Cell Motility and the Cytoskeleton*, 65(2):143–155.

Ohno, S. (1970). *Evolution by gene duplication.* Springer-Verlag, Berlin.

Oliver, S. (2000). Proteomics: Guilt-by-association goes global. *Nature*, 403(6770):601–603.

Opper, M. and Sanguinetti, G. (2010). Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics*, 26(13):1623–1629.

Ostrin, E. J., Li, Y., Hoffman, K., Liu, J., Wang, K., Zhang, L., Mardon, G., and Chen, R. (2006). Genome-wide identification of direct targets of the *Drosophila* retinal determination protein Eyeless. *Genome Research*, 16(4):466–476.

Ostrowski, L. E., Blackburn, K., Radde, K. M., Moyer, M. B., Schlatzer, D. M., Moseley, A., and Boucher, R. C. (2002). A Proteomic Analysis of Human Cilia. *Molecular & Cellular Proteomics*, 1(6):451–465.

Page, R. D. M. (1994). Maps Between Trees and Cladistic Analysis of Historical Associations among Genes,Organisms, and Areas. *Systematic Biology*, 43(1):58–77.

Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Paroush, Z., Finley, R. L., Kidd, T., Wainwright, S., Ingham, P. W., Brent, R., and Ish-Horowicz, D. (1994). Groucho is required for *Drosophila* neurogenesis, segmentation, and sex determination and interacts directly with hairy-related bHLH proteins. *Cell*, 79(5):805–815.

Parrish, J., Kim, M., Jan, L., and Jan, Y. (2006a). Genome-wide analyses identify transcription factors required for proper morphogenesis of *Drosophila* sensory neuron dendrites. *Genes & Development*, 20(7):820–835.

Parrish, J., Yu, J., Liu, G., Hines, J., Chan, J., Mangiola, B., Zhang, H., Pacifico, S., Fotouhi, F., DiRita, V., Ideker, T., Andrews, P., and Finley, R. (2007). A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biology*, 8(7):R130.

Parrish, J. R., Gulyas, K. D., and Finley, Jr, R. L. (2006b). Yeast two-hybrid contributions to interactome mapping. *Current Opinion in Biotechnology*, 17(4):387–393.

Pazour, G. J., Agrin, N., Leszyk, J., and Witman, G. B. (2005). Proteomic analysis of a eukaryotic cilium. *The Journal of Cell Biology*, 170(1):103–113.

Pazour, G. J. and Rosenbaum, J. L. (2002). Intraflagellar transport and cilia-dependent diseases. *Trends Cell Biol*, 12(12):551–555.

Pazour, G. J. and Witman, G. B. (2003). The vertebrate primary cilium is a sensory organelle. *Current Opinion in Cell Biology*, 15(1):105 – 110.

Pedamallu, C. S. and Posfai, J. (2010). Open source tool for prediction of genome wide protein-protein interaction network based on ortholog information. *Source Code for Biology and Medicine*, 5(1):8.

Pedersen, S. F., Owsianik, G., and Nilius, B. (2005). TRP channels: An overview. *Cell Calcium*, 38(3–4):233 – 252. <ce:title>Frontiers in calcium signalling</ce:title>.

Pei, J., Jiang, D., and Zhang, A. (2005). On mining cross-graph quasi-cliques. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 228–238, New York, NY, USA. ACM.

Pereira-Leal, J. B., Enright, A. J., and Ouzounis, C. A. (2004). Detection of functional modules from protein interaction networks. *Proteins: Structure, Function, and Bioinformatics*, 54(1):49–57.

Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A., and Cesareni, G. (2005). HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, 6(Suppl 4):S21.

Peterson, M. E., Chen, F., Saven, J. G., Roos, D. S., Babbitt, P. C., and Sali, A. (2009). Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Science*, 18(6):1306–1315.

Pi, H. and Chien, C.-T. (2007). Getting the edge: neural precursor selection. *Journal of Biomedical Science*, 14(4):467–473.

Pothen, A., Simon, H., and Liou, K. (1990). Partitioning Sparse Matrices with Eigenvectors of Graphs. *SIAM Journal on Matrix Analysis and Applications*, 11(3):430–452.

Powell, L. M. and Jarman, A. P. (2008). Context dependence of proneural bHLH proteins. *Current Opinion in Genetics & Development*, 18(5):411 – 417. <ce:title>Differentiation and gene regulation</ce:title>.

Powell, L. M., zur Lage, P. I., Prentice, D. R. A., Senthinathan, B., and Jarman, A. P. (2004). The Proneural Proteins Atonal and Scute Regulate Neural Target Genes through Different E-Box Binding Sites. *Mol. Cell. Biol.*, 24(21):9517–9526.

Prieto, C. and De Las Rivas, J. (2006). APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Research*, 34(suppl 2):298–302.

Pržulj, N., Wigle, D., and Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification. *Methods*, 24(3):218 – 229.

Pullan, W. and Hoos, H. H. (2006). Dynamic local search for the maximum clique problem. *J. Artif. Int. Res.*, 25(1):159–185.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rastogi, S. and Liberles, D. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology*, 5(1):28.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A.-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297(5586):1551–1555.

Razick, S., Magklaras, G., and Donaldson, I. (2008). iRefIndex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9(1):405.

Reeves, N. and Posakony, J. W. (2005). Genetic Programs Activated by Proneural Proteins in the Developing *Drosophila PNS*. *Developmental Cell*, 8(3):413–425.

Reiter, L. T., Potocki, L., Chien, S., Gribskov, M., and Bier, E. (2001). A Systematic Analysis of Human Disease-Associated Gene Sequences In *Drosophila melanogaster*. *Genome Research*, 11(6):1114–1125.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041–1052.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Rice, D., Aberg, T., Chan, Y., Tang, Z., Kettunen, P., Pakarinen, L., Maxson, R., and Thesleff, I. (2000). Integration of FGF and TWIST in calvarial bone and suture development. *Development*, 127(9):1845–1855.

Rice, J. J., Tu, Y., and Stolovitzky, G. (2005). Reconstructing biological networks using conditional correlation analysis. *Bioinformatics*, 21(6):765–773.

Robbertse, B., Reeves, J. B., Schoch, C. L., and Spatafora, J. W. (2006). A phylogenomic analysis of the *Ascomycota*. *Fungal Genetics and Biology*, 43(10):715 – 725.

Robinson, J. W. and Hartemink, A. J. (2010). Learning Non-Stationary Dynamic Bayesian Networks. *J. Mach. Learn. Res.*, 11:3647–3680.

Rogers, S., Khanin, R., and Girolami, M. (2007). Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, 8(Suppl 2):S2.

Rosenbaum, J. L. and Witman, G. B. (2002). Intraflagellar transport. *Nat Rev Mol Cell Biol*, 3(11):813–825.

Rouault, H., Mazouni, K., Couturier, L., Hakim, V., and Schweisguth, F. (2010). Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. *Proceedings of the National Academy of Sciences*, 107(33):14615–14620.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.

Sanguinetti, G., Ruttor, A., Opper, M., and Archambeau, C. (2009). Switching regulatory models of cellular stress response. *Bioinformatics*, 25(10):1280–1286.

Sarpal, R., Todi, S. V., Sivan-Loukianova, E., Shirolikar, S., Subramanian, N., Raff, E. C., Erickson, J. W., Ray, K., and Eberl, D. F. (2003). *Drosophila* KAP Interacts with the Kinesin II Motor Subunit KLP64D to Assemble Chordotonal Sensory Cilia, but Not Sperm Tails. *Curr Biol*, 13(19):1687–1696.

Schwed, G., May, N., Pechersky, Y., and Calvi, B. (2002). *Drosophila* minichromosome maintenance 6 is required for chorion gene amplification and genomic replication. *Molecular Biology of the Cell*, 13(2):607–620.

Schäfer, J. and Strimmer, K. (2005). Learning Large-Scale Graphical Gaussian Models from Genomic Data. *AIP Conference Proceedings*, 776(1):263 – 276.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003a). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–176.

Segal, E., Wang, H., and Koller, D. (2003b). Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(suppl 1):i264–i272.

Serluca, F. C., Xu, B., Okabe, N., Baker, K., Lin, S.-Y., Sullivan-Brown, J., Konieczkowski, D. J., Jaffe, K. M., Bradner, J. M., Fishman, M. C., and Burdine, R. D. (2009). Mutations in zebrafish leucine-rich repeat-containing six-like affect cilia motility and result in pronephric cysts, but have variable effects on left-right patterning. *Development*, 136(10):1621–1631.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504.

Sharan, R., Ideker, T., Kelley, B., Shamir, R., and Karp, R. M. (2005). Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data. *Journal of Computational Biology*, 12(6):835–846.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol Syst Biol*, 3:–.

Silverman, M. A. and Leroux, M. R. (2009). Intraflagellar transport and the generation of dynamic, structurally and functionally diverse cilia. *Trends in Cell Biology*, 19(7):306 – 316.

Sokolowski, M. (2001). *Drosophila*: genetics meets behaviour. *Nature Reviews. Genetics*, 2(11):879–890.

Sonnhammer, E. L. L. and Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, 18(12):619–620.

Sonnichsen, B., Koski, L. B., Walsh, A., Marschall, P., Neumann, B., Brehm, M., Alleaume, A.-M., Artelt, J., Bettencourt, P., Cassin, E., Hewitson, M., Holz, C., Khan, M., Lazik, S., Martin, C., Nitzsche, B., Ruer, M., Stamford, J., Winzi, M., Heinkel, R., Roder, M., Finell, J., Hantsch, H., Jones, S. J. M., Jones, M., Piano, F., Gunsalus, K. C., Oegema, K., Gonczy, P., Coulson, A., Hyman, A. A., and Echeverri, C. J. (2005). Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature*, 434(7032):462–469.

Spirin, V. and Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21):12123–12128.

Sprinzak, E., Sattath, S., and Margalit, H. (2003). How Reliable are Experimental Protein-Protein Interaction Data? *Journal of Molecular Biology*, 327(5):919–923.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, 12(10):1611–1618.

Stanyon, C., Liu, G., Mangiola, B., Patel, N., Giot, L., Kuang, B., Zhang, H., Zhong, J., and Finley, R. (2004). A *Drosophila* protein-interaction map centered on cell-cycle regulators. *Genome Biology*, 5(12):R96.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell*, 122(6):957 – 968.

Stolc, V., Samanta, M. P., Tongprasit, W., and Marshall, W. F. (2005). Genome-wide transcriptional analysis of flagellar regeneration in *Chlamydomonas reinhardtii* identifies orthologs of ciliary disease genes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3703–3707.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

Storm, C. E. V. and Sonnhammer, E. L. L. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–99.

Stubbs, J. L., Oishi, I., Izpisua Belmonte, J. C., and Kintner, C. (2008). The forkhead protein Foxj1 specifies node-like cilia in *Xenopus* and zebrafish embryos. *Nat Genet*, 40(12):1454–1460.

Studer, R. A. and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, 25(5):210 – 216.

Stumpf, M. P. H. and Wiuf, C. (2005). Sampling properties of random graphs: The degree distribution. *Phys. Rev. E*, 72:036118.

Stumpf, M. P. H., Wiuf, C., and May, R. M. (2005). Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4221–4224.

Su, G., Kuchinsky, A., Morris, J. H., States, D. J., and Meng, F. (2010). GLay: community structure analysis of biological networks. *Bioinformatics*, 26(24):3135–3137.

Suthram, S., Sittler, T., and Ideker, T. (2005). The *Plasmodium* protein network diverges from those of other eukaryotes. *Nature*, 438(7064):108–112.

Sutton, K. A., Jungnickel, M. K., Wang, Y., Cullen, K., Lambert, S., and Florman, H. M. (2004). Enkurin is a novel calmodulin and TRPC channel binding protein in sperm. *Developmental Biology*, 274(2):426 – 435.

Swoboda, P., Adler, H. T., and Thomas, J. H. (2000). The RFX-Type Transcription Factor DAF-19 Regulates Sensory Neuron Cilium Formation in *C. elegans*. *Mol Cell*, 5(3):411–421.

Takada, S., Wilkerson, C. G., Wakabayashi, K.-i., Kamiya, R., and Witman, G. B. (2002). The Outer Dynein Arm-Docking Complex: Composition and Characterization of a Subunit (Oda1) Necessary for Outer Arm Assembly. *Molecular Biology of the Cell*, 13(3):1015–1029.

Tanaka, R., Yi, T.-M., and Doyle, J. (2005). Some protein interaction data do not exhibit power law statistics. *FEBS Letters*, 579(23):5140 – 5144.

Tang, J., Musolesi, M., Mascolo, C., Latora, V., and Nicosia, V. (2010a). Analysing information flows and key mediators through temporal centrality metrics. In *Proceedings of the 3rd Workshop on Social Network Systems*, SNS '10, pages 3:1–3:6, New York, NY, USA. ACM.

Tang, J., Scellato, S., Musolesi, M., Mascolo, C., and Latora, V. (2010b). Small-world behavior in time-varying graphs. *Phys. Rev. E*, 81:055101.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A Genomic Perspective on Protein Families. *Science*, 278(5338):631–637.

Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., Blake, J. A., and on behalf of the Gene Ontology Consortium (2012). On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS Comput Biol*, 8(2):e1002386.

Thornton, J. W. and DeSalle, R. (2000). Gene Family Evolution and Homology: Genomics Meets Phylogenetics. *Annual Review of Genomics and Human Genetics*, 1(1):41–73.

Treisman, J., Lai, Z., and Rubin, G. (1995). Shortsighted acts in the decapentaplegic pathway in *Drosophila* eye development and has homology to a mouse TGF-responsive gene. *Development*, 121(9):2835–2845.

Trucco, C., Fernandez-Reyes, D., Howell, S., Stafford, W. H., Scott-Finnigan, T. J., Grainger, M., Ogun, S. A., Taylor, W. R., and Holder, A. A. (2001). The merozoite surface protein 6 gene codes for a 36 kDa protein associated with the *Plasmodium falciparum* merozoite surface protein-1 complex. *Molecular and Biochemical Parasitology*, 112(1):91 – 101.

Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12(3):368–373.

van Dongen, S. (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335.

Vlasblom, J. and Wodak, S. (2009). Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*, 10(1):99.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(suppl 1):D433–D437.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403.

Voss, T. S., Mini, T., Jenoe, P., and Beck, H.-P. (2002). *Plasmodium falciparum* Possesses a Cell Cycle-regulated Short Type Replication Protein A Large Subunit Encoded by an Unusual Transcript. *Journal of Biological Chemistry*, 277(20):17493–17501.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333.

Wakita, K. and Tsurumi, T. (2007). Finding community structure in mega-scale social networks: [extended abstract]. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 1275–1276, New York, NY, USA. ACM.

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000). Protein Interaction Mapping in *C. elegans* Using Proteins Involved in Vulval Development. *Science*, 287(5450):116–122.

Wallace, I. M., O'Sullivan, O., Higgins, D. G., and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, 34(6):1692–1699.

Wang, X., Wu, M., Li, Z., and Chan, C. (2008). Short time-series microarray analysis: Methods and challenges. *BMC Systems Biology*, 2(1):58.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.

Wasbrough, E., Dorus, S., Hester, S., Howard-Murkin, J., Lilley, K., Wilkin, E., Polpitiya, A., Petritis, K., and Karr, T. (2010). Supplemental table 1. mass spectrometry data.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440–442.

Werhli, A. and Husmeier, D. (2007). Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6(1):15.

Werhli, A. V. (2007). *Reconstruction of gene regulatory networks from postgenomic data*. PhD thesis, University of Edinburgh.

White, J. and Kilbey, B. (1996). DNA replication in the malaria parasite. *Parasitology Today*, 12(4):151 – 155.

Whitney, H. (1932). Congruent Graphs and the Connectivity of Graphs. *American Journal of Mathematics*, 54(1):pp. 150–168.

Wickstead, B. and Gull, K. (2007). Dyneins Across Eukaryotes: A Comparative Genomic Analysis. *Traffic*, 8(12):1708–1721.

Wiles, A., Doderer, M., Ruan, J., Gu, T.-T., Ravi, D., Blackman, B., and Bishop, A. (2010). Building and analyzing protein interactome networks by cross-species comparisons. *BMC Systems Biology*, 4(1):36.

Wilkinson, D. M. and Huberman, B. A. (2004). A method for finding communities of related genes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5241–5248.

Wojcik, J., Boneca, I. G., and Legrain, P. (2002). Prediction, Assessment and Validation of Protein Interaction Maps in Bacteria. *Journal of Molecular Biology*, 323(4):763–770.

Wold, M. S. (1997). Replication Protein A: A Heterotrimeric, Single-Stranded DNA-Binding Protein Required for Eukaryotic DNA Metabolism. *Annual Review of Biochemistry*, 66(1):61–92.

Wu, C. and Dedhar, S. (2001). Integrin-linked kinase (ILK) and its interactors. *The Journal of Cell Biology*, 155(4):505–510.

Wuchty, S., Adams, J. H., and Ferdig, M. T. (2009). A comprehensive *Plasmodium falciparum* protein interaction map reveals a distinct architecture of a core interactome. *PROTEOMICS*, 9(7):1841–1849.

Wuchty, S. and Ipsaro, J. J. (2007). A Draft of Protein Interactions in the Malaria Parasite *P. falciparum*. *Journal of Proteome Research*, 6(4):1461–1470.

Wuchty, S., Oltvai, Z. N., and Barabasi, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet*, 35(2):176–179.

Xu, K., Bezakova, I., Bunimovich, L., and Yi, S. V. (2011). Path lengths in protein–protein interaction networks and biological complexity. *PROTEOMICS*, 11(10):1857–1867.

Xue, J.-C. and Goldberg, E. (2000). Identification of a Novel Testis-Specific Leucine-Rich Protein in Humans and Mice. *Biology of Reproduction*, 62(5):1278–1284.

Yamada, M., Ohkawara, B., Ichimura, N., Hyodo-Miura, J., Urushiyama, S., Shirakabe, K., and Shibuya, H. (2003). Negative regulation of Wnt signalling by HMG2L1, a novel NLK-binding protein. *Genes to Cells*, 8(8):677–684.

Yang, Z. and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12):496–503.

Yin, G., Dai, J., Ji, C., Ni, X., Shu, G., Ye, X., Dai, J., Wu, Q., Gu, S., Xie, Y., Chunhua Zhao, R., and Mao, Y. (2003). Cloning and characterization of the human IFT20 gene. *Molecular Biology Reports*, 30:255–260. 10.1023/A:1026365124176.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabasi, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008a). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science*, 322(5898):104–110.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs. *Genome Research*, 14(6):1107–1118.

Yu, X., Ng, C. P., Habacher, H., and Roy, S. (2008b). Foxj1 transcription factors are master regulators of the motile ciliogenic program. *Nat Genet*, 40(12):1445–1453.

Zariwala, M. A., Knowles, M. R., and Omran, H. (2007). Genetic defects in ciliary structure and function. *Annual Review of Physiology*, 69:423–450.

Zhang, W., Zhang, S., Xiao, C., Yang, Y., and Zhoucun, A. (2007). Mutation screening of the FKBP6 gene and its association study with spermatogenic impairment in idiopathic infertile men. *Reproduction*, 133(2):511–516.

Zhang, Z. and Zhang, J. (2009). A Big World Inside Small-World Networks. *PLoS ONE*, 4(5):e5686.

Zmasek, C. M. and Eddy, S. R. (2001). A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, 17(9):821–828.

zur Lage, P. I. and Jarman, A. (2010). The function and regulation of the bHLH gene, *cato*, in *Drosophila* neurogenesis. *BMC Developmental Biology*, 10(1):34.

zur Lage, P. I., Prentice, D. R. A., Holohan, E. E., and Jarman, A. P. (2003). The *Drosophila* proneural gene *amos* promotes olfactory sensillum formation and suppresses bristle formation. *Development*, 130(19):4683–4693.

zur Lage, P. I., Simpson, T. I., and Jarman, A. (2011). Linking specification to differentiation: From proneural genes to the regulation of ciliogenesis. *fly*, 5(1933-6934):322–326.