Bioinformatics protocols for analysis of functional genomics data applied to neuropathy microarray datasets

Ilhem Diboun

Department of Structural and Molecular Biology
University College London

A thesis submitted to University College London in the Faculty of Science for the degree of Doctor of Philosophy

June 2009

Declaration

I, Ilhem Diboun confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Microarray technology allows the simultaneous measurement of the abundance of thousands of transcripts in living cells. The high-throughput nature of microarray technology means that automatic analytical procedures are required to handle the sheer amount of data, typically generated in a single microarray experiment. Along these lines, this work presents a contribution to the automatic analysis of microarray data by attempting to construct protocols for the validation of publicly available methods for microarray.

At the experimental level, an evaluation of amplification of RNA targets prior to hybridisation with the physical array was undertaken. This had the important consequence of revealing the extent to which the significance of intensity ratios between varying biological conditions may be compromised following amplification as well as identifying the underlying cause of this effect. On the basis of these findings, recommendations regarding the usability of RNA amplification protocols with microarray screening were drawn in the context of varying microarray experimental conditions.

On the data analysis side, this work has had the important outcome of developing an automatic framework for the validation of functional analysis methods for microarray. This is based on using a GO semantic similarity scoring metric to assess the similarity between functional terms found

enriched by functional analysis of a model dataset and those anticipated from prior knowledge of the biological phenomenon under study. Using such validation system, this work has shown, for the first time, that 'Catmap', an early functional analysis method performs better than the more recent and most popular methods of its kind. Crucially, the effectiveness of this validation system implies that such system may be reliably adopted for validation of newly developed functional analysis methods for microarray.

Acknowledgements

I would like to express my biggest gratitude to my supervisor Prof Christine Orengo for her support and guidance throughout this project and for giving me the opportunity to pursue a higher degree in science. I am very grateful to all the people in the Orengo lab for making my PhD a wonderful and enriching experience, with special mentions to Ollie Redfern and James Perkins who helped improve the quality of this manuscript.

I would like to thank the LPC and Wellcome Trust for generously funding this project. Special thanks go to all LPC principal investigators; in particular, Prof Martin Koltzenburg for help, support and illuminating discussions on the biology of neuropathic pain.

Finally, all my love and gratitude goes to my husband Dr Aghar Elrayess as without all his support and sacrifices all those years, this work would never have been completed. A special mention goes to my lovely daughter Arwa whose smiles, hugs and kisses kept me going during difficult times. Finally, I would like to pay special tribute to my mother and sisters whose moral support has been so important to completing this work.

Table of contents

List of Figures

List of tables

CHAPTER I: INTRODUCTION

1.1. The London Pain Consortium research mission

The work presented in this thesis is part of the ongoing collaborative efforts by members of the London pain Consortium (LPC) to achieve a better understanding of the origin of chronic pain, under neuropathological conditions. The LPC is a group of scientists that was formed in 2002 funded by the Wellcome Trust and has since undertaken exciting research to reveal mechanisms of chronic pain from a variety of different angles, including screening for gene expression regulation. Different animal models of painful neuropathies have been used by LPC experimentalists to generate large amount of gene expression data, with the hope of identifying common mechanisms of pain. In this project, our role as members of the LPC has been to assist with the analysis of these expression data using bioinformatics approaches, notably via integration with other useful types of data as will be discussed in the work chapters of this thesis.

1.2. Neuropathic pain mechanisms

Pain is usually the natural consequence of tissue injury that serves to trigger an appropriate defensive response and is therefore an important mechanism for survival. Normally, pain subsides as the healing process commences; this usual form of pain is known as *acute pain* and is distinct from the rather pathological long lasting pain known as *chronic pain*. Unlike acute pain that serves to promote healing and preserve tissue integrity, chronic pain has no physiological role as it is rather debilitating often causing depression and reducing the sufferers' quality of life. Chronic pain is hence a disease state that needs to be treated.

There are two forms of chronic pain: nociceptive and neuropathic pain; which are the products of different neuro-physiological processes. While nociceptive pain is caused by the continuous stimulation of pain receptor fibres by nerve sensitising substances (examples are inflammatory substances such as histamine, bradykinin and substance-P), neuropathic pain is caused by damage to or pathological changes in the peripheral or central nervous systems. Examples of nociceptive pain are post-operative pain, pain associated with trauma, and the chronic pain of arthritis. As for neuropathic pain, clinical examples are post herpetic neuralgia, reflex sympathetic dystrophy (nerve

trauma), entrapment neuropathy and peripheral neuropathy most commonly caused by diabetes or chronic alcohol use.

For years, the chronic pain of neuropathy has confounded scientists. Traditional pain treatments, including powerful medications of the last resort such as morphine, rarely help. To date, neuropathic pain has been the subject of much research in an attempt to shed more light on its mechanisms and develop more effective treatments. One important advance in the field of neuropathic pain research has been the development of animal models of painful neuropathies, whereby the occurrence of nociceptive behaviour such as agitation and avoidance is taken to indicate the presence of pain. The best-established types of these models involve a form of experimentally induced injury to the nervous system, either peripherally (involving nerves innervating parts of the body, notably the limbs) or centrally (consisting of the brain and the spinal cord). More recently, animal models of disease induced neuropathic pain have also been developed; examples are those mimicking human clinical conditions such as diabetic and cancer neuropathy.

Current knowledge of the mechanisms of neuropathic pain is limited and biased by a focus on the well-established animal models of peripheral neuropathy. Before describing these mechanisms, it is important to understand the nature of common injuries to the peripheral nerve involved in these

models, which entails the need for a brief description of the anatomy of the peripheral nervous system. Figure 1.2.1 illustrates a section of the peripheral nervous system, featuring the sciatic nerve that innervates most of the skin of the back paw. The sciatic nerve consists of a bundle of axons from sensory and motor neurons whose cell bodies lie in the dorsal root ganglions (DRG). A second branch of axons extends from the DRG sensory neurons upwards to synapse with the spinal horn neurons in the spinal cord, creating an interface with the central nervous system. The models of peripheral nerve injury shown on Figure 1.2.1 vary in the type of injury as well as the location of injury, affecting either the whole of the sciatic nerve or its branches distal or at close proximity from the DRG compartments.

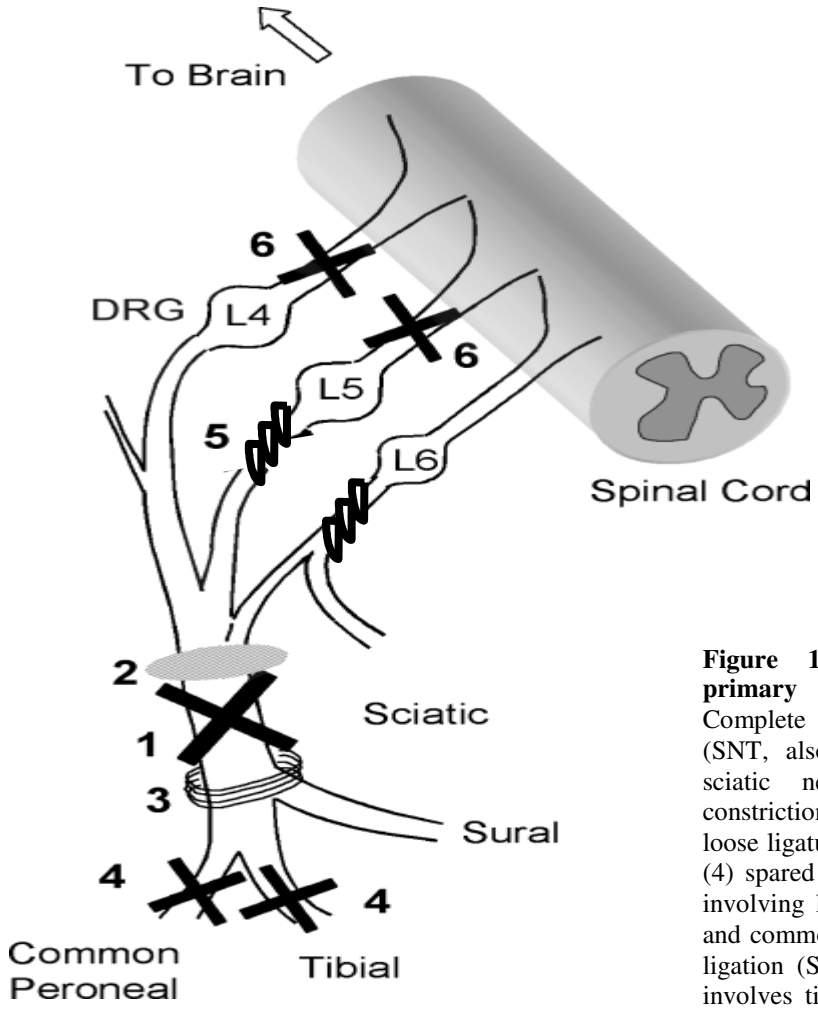


Figure 1.2.1. Injury models in primary sensory neurons. (1) Complete sciatic nerve transection (SNT, also known as axotomy); (2) sciatic nerve crush; (3) chronic constriction injury consisting of four loose ligatures around the sciatic nerve; (4) spared nerve injury model (SNI) involving ligation and section of tibial and common peroneal; (5) spinal nerve ligation (SNL, also known as Chung) involves tight ligations of L5/6 spinal nerves; (6) dorsal rhizotomy lesion involving transection of L4 and L5 DRGs. Based on the information and graphical images by (Ueda and Rashid, 2003)

With peripheral nerve injury, mechanisms of neuropathic pain involve sensitisation and neuronal plasticity of the peripheral nervous system, leading in turn to the recruitment of a more centralised nociceptive activity. Following peripheral nerve injury, an increased membrane density of Na^+ channels on injured fibres causes spontaneous discharge action potential to be generated in

the absence of any stimulation (Woolf, 2004). Such irregular discharge has been previously seen in injured and non-injured neighbouring fibres at the site of injury or in the dorsal root ganglion tissue (DRG) containing the cell bodies of injured neurons (Woolf, 2004). Evidence exists to suggest that increased levels of sympathetic activity at the site of injury may increase the ability of sprouting fibres to detect pain excitatory substances (Zimmermann, 2004). Indeed, sympathetic fibres may contribute to increased sensitisation of the growing fibres to inflammatory substances within the milieu of injury by releasing adrenaline and noradrenaline that modulate the activity of receptors on the growing fibres. Uninjured nerves adjacent to the site of injury may become involved as more central processes produce a localized release of sensitizing neurotransmitters (such as substance-P, glutamate, CRGP, and 5HT) into uninjured regions ultimately producing a self-sustained state of neurogenic inflammation.

Centrally, the continuous ectopic discharge by injured afferent fibres peripherally has the effect of sensitising post-synaptic neurons in the dorsal horn of the spinal cord. This sensitisation, also known as *wind-up,* has been interpreted as a system for the amplification of peripherally induced nociceptive signals in the spinal cord. Repetitive episodes of wind-up may precipitate long-term potentiation (LTP), which involves a long lasting increase in the efficacy of synaptic transmission. LTP is thought to be an

important mediator of *hyperalgesia*, an important landmark of neuropathic pain, which describes an exaggerated response to painful stimuli. Moreover, anatomical changes in the spinal cord have been observed following peripheral nerve injury (Zimmermann, 2004) whereby deep spinal neurons that normally receive and propagate non-noxious peripheral input sprout into superficial spinal regions involved in transmitting high intensity signals. This is thought to explain the origin of *allodynia*, another landmark of neuropathic pain, whereby painful sensations are caused by non-painful stimuli under a diseased state of the nervous system.

An important underlying mechanism to these changes affecting both the peripheral and central nervous systems; in particular, those of long lasting nature; consists of modification in gene expression at the cellular level (Woolf, 2004). For instance, the switch in the phenotype of neuronal subtypes centrally following neuropathy is thought to be largely mediated by a change in gene expression of affected nerve cells in the spinal dorsal horn. Importantly, much of these central effects are triggered by a change in the type and levels of neurotransmitters released by afferent fibers at the junction with the spinal cord as a result of much shift in gene expression activity in DRG nerve cells peripherally. Moreover, modification of gene expression at the peripheral level not only contributes to the establishment of neuropathic pain, but also supports its long lasting nature via endogenous synthesis of proinflammatory

substances that preserve the pathological conditions surrounding the nervous system. This acts as a feedback loop mechanism that ensures a prolonged state of reprogrammed gene expression at the cellular level and hence sustained shift in sensory neuron excitability both peripherally and centrally (Scholz and Woolf, 2007).

Owing to the central role of gene expression regulation in the development of chronic pain under neuropathological conditions, many studies have used microarray technology to characterise the global changes in gene expression in nerve tissue in animal models of neuropathy with painful phenotypes (Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002). However, a widely accepted view is that such an approach is limited since many processes other than those featuring direct relevance to pain are equally affected by gene expression regulation; examples are neuronal regeneration and immune/inflammatory processes that occur as a natural consequence of the injury to the nerve.

Nonetheless, there has been numerous attempts in literature to optimise the use of microarrays with animal models of painful neuropathies to detect changes in gene expression specific to pain sensation. The earliest of such attempts was made by Valder and colleagues (Valder et al., 2003) who analysed changes in DRG gene expression of two varying rat strains either sensitive or resistant to nerve injury-induced mechanical allodynia. By examining the injury-induced

strain-specific gene differential expression, Valder isolated genes that are directly relevant to the development of mechanical allodynia under neuropathic conditions. Along these lines, the LPC has recently developed an experimental strategy to identify pain specific changes in gene expression using microarray that relies on comparison of patterns of gene differential expression between painful neuropathies of different etiologies. One preliminary study was recently published by LPC member Maratou (Maratou et al., 2009), featuring a comparison of patterns of gene expression regulation between a model of HIV neuropathy and a model of traumatic nerve injury to isolate common pain meditating genes.

However, the current work has a slightly different focus in that it was aimed at developing reliable analysis protocols for microarray data that can be applied on individual datasets. Being one of the first projects undertaken in collaboration with the LPC, the project addressed the need for exploring with ways of performing basic analysis of microarray data; thereby, setting the scene for more sophisticated meta-analyses to be performed in the future from combining all LPC microarray datasets. Also, it was only recently that the LPC adopted the strategy of contrasting different pain models using microarrays and hence many of the currently existing datasets were not available to us at the time when this work was performed. Importantly, in this project, because the focus was on individual microarray datasets where

differential gene expression relates to a large number of processes other than pain, we were inclined to refer to LPC microarray data as neuropathy rather than pain expression data.

1.3. Microarray expression profiling

1.3.1. Microarray technology

For the remainder of this introduction, methods for microarray data analysis and microarray research applications are discussed in detail. Microarray technology allows simultaneous quantification of levels of expression for a large number of genes providing a way to study dependencies between their patterns of expression. A microarray is technically defined as a solid support onto which sequences from thousands of transcripts are immobilised, or attached at fixed locations. The supports themselves are usually glass microscope slides, but can also be silicon chips or nylon membranes. The sequences are printed, spotted or synthesised directly onto the support.

Microarray technology relies on the ability of a given mRNA molecule to bind the DNA template from which it originated. In a typical microarray experiment, labelled target mRNA isolated from a tissue of interest is hybridised with complementary array sequences and the amounts of fluorescence from double stranded hybrids are estimated using a scanner to determine the level of abundance of individual RNA targets in the original biological sample. Most commonly, microarrays are used to compare the gene

expression profiles of two biological tissues such as 'wild type' and 'diseased/treated'. Nowadays, microarray technology has evolved to cover a wide spectrum of research applications beyond detection of differential expression between varying biological conditions. For instance, SNP arrays are used to detect polymorphisms within and between populations whilst exon junction arrays are designed to assist in the measurement of alternatively spliced forms of transcripts.

The LPC is currently using DNA microarray technology in the traditional sense to detect changes in gene expression following painful neuropathies. Till now, the LPC has used oligonucleotide-based arrays manufactured by Affymetrix whereby 11 to 16 probes are selected among all possible 25-long oligonucleotides to represent each target transcript. The collection of these probes is known as a *probeset* and each probeset is given a unique identifier on the array. Importantly, a gene may be represented by more than one probeset on the same array. Within a probeset, each of the probes exists in two forms: a *perfect match (PM),* which perfectly aligns with the target sequence and a *mismatch* (*MM),* which has the same sequence as the PM except for the middle base which is made different. MMs are used by Affymetrix to provide an assessment for the level of non-specific hybridization. The probes are designed to bind to complementary RNA (cRNA) prepared from mRNA extracted from the biological tissue.

1.3.2. Microarray low level analysis

After hybridising labelled cRNA with Affymetrix array probes, a picture of the array is taken by the scanner and the individual intensities of all probes are estimated using the image scanning algorithm. Before analysing the data for differential expression, the individual intensities need to be calibrated in order to eliminate the experimental variation in the data. Calibration of microarray data proceeds through a number of different steps:

1.3.2.1. Background correction

The aim of this initial step is to subtract the contribution of non-specific binding from the overall intensity of each spot measured on the array. Probes may bind to sequences other than the target depending on their specificity and the conditions during the hybridisation step. Background fluorescence is another source of non-specificity. There exists a number of methods for correcting background. The method used by Affymetrix relies on using the area of the chip with the lowest fluorescence as an estimate of the background, whereas MM probe intensities are used to assess binding to non-targets. Other methods such as *RMA* (Irizarry et al., 2003) use a fitted stochastic model to the overall distribution of the PM probes (and sometimes the MMs) to estimate

background. Recently, a new background correction method *GeneChip RMA (GCRMA)* (Irizarry et al., 2003) was developed based on modelling the binding interactions between probes on the arrays and their target transcripts. GCRMA attempts to eliminate systematic contribution to noise from the sequence of the probe and labelled nucleotides in the target. GCRMA outperforms the other methods at the low end of the intensity scale where much of the signal is due to noise, therefore allowing changes in gene expression to be detected more reliably at this range of intensity (Irizarry et al., 2003).

1.3.2.2. Normalization

Normalization is then applied to compensate for systematic technical differences between arrays in order to emphasize real biological differences between samples. Most approaches to normalizing expression levels assume that the overall distribution of RNA abundance does not change much between samples, that is to say that most expressed genes maintain a constant expression level in the different biological states being investigated. The simplest approach to normalizing Affymetrix data is to re-scale each array in an experiment so that the average (or total) signal intensity across all arrays is equal. This linear scaling is generally criticized for failing to recognise that the array effect is not constant across all range of intensities. Numerous methods

implementing non-linear normalization of array data exist. One of such is *Quantiles normalization* which forces the intensity distribution on each chip to be identical by ranking the intensities, and resetting the intensity values in each rank across all arrays to the mean of the intensities at that rank. The rescaling is therefore different at each rank, which makes this normalization rather non-linear across the range of intensities.

1.3.2.3. Expression Summary

This step aims to reduce the 11-16 measures of probe intensity within a probeset into one value of expression that is indicative of the abundance of the corresponding RNA target. This is non-trivial given that individual probes show differences in binding affinities and it is typical to observe large discrepancies in the intensities of probes within the same probeset. Model based approaches for expression summary calculation explore the fact that the specific binding efficiency of each probe is inherent to its sequence and is constant across all arrays. Thus, using information from all arrays in an experiment, such methods fit models to the intensity data to estimate parameters such as probe specific effects and the level of mRNA bound to the probe on each array. These parameters are then used to derive a summary intensity value for each probeset on each array.

1.3.2.4. Statistical analysis of gene differential expression

After calibration comes the actual statistical analysis of the data that allows differentially expressed genes to be detected. In the simplest comparison of two biologically distinct conditions, t-statistics can be applied with multiple testing correction. This is necessary to account for the occurrences of false positives that are inevitable with the large number of genes tested for differential expression on the array. However, it is widely accepted that t-statistics may be inflated by the inevitable chance occurrences of very small variance with microarray data. That is because, typically with microarrays, only a handful of replicate measurements are available for each gene and furthermore, at low intensity levels, variation in intensity is usually minimal. This flaw has been addressed by many statistical methods specifically developed for differential expression analysis of microarray data, such as *Significance Analysis of Microarray (SAM)* (Tusher et al., 2001) and *Linear Models for Microarray data (Limma*) (Smyth, 2004).

SAM and Limma are fundamentally similar in that they are both based on a moderated t-statistic that features an optimized assessment of within group variation. The difference, however, lies in the mechanism used by either method for flooring such variance. With SAM and to make sure that the t-statistic based scores for genes are not inflated at low intensity levels due to

intrinsically low variation, an offset value is estimated that minimizes the variation in the t-statistic based scores as a function of variability from replicate gene expression measurements. Thus, the SAM statistics are simply t-statistics where the pooled standard deviation has been shifted systematically by a constant value for all genes. Unlike SAM, Limma fits a linear model on an individual gene basis using gene intensity data from all arrays to derive a gene-wise residual sample variance estimate that is more robust than ordinary variance.

- *Multiple testing correction*

As previously discussed, microarray differential expression analysis is a classical case of multiple testing problem. Thus, at p-value 0.01 and given an overall number of 10000 genes on the array, we may expect 100 genes to appear significant by chance. In classical statistics, there exists a number of methods for multiple testing correction that vary in stringency. These methods fall in two broad categories and are either based on controlling the *family wise error rate (FWE)* or the *false discovery rate (FDR)*. There exists a fundamental difference between the two approaches in that at any given significance level p-value=*P*, the FWE based methods operate by estimating the chance of occurrence of at least one false positive given the total number of hypotheses tested (meaning genes in the context of microarray). The FDR

based approach, on the other hand, is less stringent in that it gives an estimate of the expected proportion of false positives with p-value<*P* given the total number of hypotheses tested. In many ways, the FDR multiple testing correction approach is more practical with microarray data in that a protection against just one single false positive is far too stringent and does not justify the parallel loss in power. That is because even though the number of false positives is lower with the more stringent multiple testing correction procedures, there is an associated increase in the number of false negatives corresponding to a loss in statistical power.

Nonetheless, stringent multiple testing corrections have been incorporated with microarray data analysis such as the *Bonferroni* correction and the *Bonferroni Step-Down (Holm)* correction. In addition to being too stringent, these two FWE based multiple testing correction methods may be unsuitable for use with microarray data as they assume test independency; which is hardly true given that genes may be co-expressed. Dudoit and colleagues (Dudoit et al., 2004) were the first to use a more appropriate FWE based procedure: the *Westfall and Young step-down* approach that allows for test dependency by using a permutation type analysis to estimate the FWE.

However, because the FDR approach is least stringent and provides a good balance between discovery of statistically significant genes and limitation of

false positive occurrences, it has become more popular with microarray than the FEW-based methods. The original FDR based procedure by Benjamini and Hochberg (Benjamini and Hochberg, 1995) assumed test independency in a similar manner to the FWE controlling Bonferroni and Holms correction procedures. However, due to high interest by the microarray community, adaptations of the FDR based multiple testing correction to test dependency have been developed. Yekutieli and Benjamini (Benjamini and Yekutieli, 2001) introduced one early procedure to control the FDR, under test dependency, based on resampling. Tusher and colleagues (Tusher et al., 2001) similarly proposed a permutation based strategy for evaluation of the FDR to accompany their proposed algorithm for microarray differential expression analysis SAM (a moderated form of t-statistics, discussed earlier). More sophisticated ways for adaptation of FDR based multiple testing correction for microarrays have since been developed, such as those by Efron and colleagues (Efron and Tibshirani, 2002) and Storey (Storey, 2003) that use a Bayesian framework to achieve local FDR analysis.

1.3.3. Microarray datamining

Microarray data are best exploited when intelligently mined for biological information. There are two broad categories of datamining approaches for

microarray: unsupervised clustering and supervised classification. With clustering, coherent patterns of gene expression may be identified across a number of related biological conditions. Such a trend is biologically meaningful as co-expressed genes are likely to be involved in the same biological process. By contrast, classification approaches are supervised in that they are based on identification of marker genes that can distinguish between varying biological conditions.

However, the most illuminating form of microarray datamining is achieved by incorporation of other types of biological data, in order to achieve a system-wide view of biological phenomena. Owing to the complexity of living organisms and their pathological and diseased states, it is often necessary to combine data from different sources and disciplines to reach useful conclusions. Microarray data only provide insights into the transcriptional activity of living cells, which is a limited view to complex biological systems and activities involving other forms of key biological events such as protein-protein interactions and protein post-translational modification. Moreover, microarray data are inherently noisy. This means that even at the level of transcriptional activity, incorporating further information on gene expression from additional sources has the benefit of improving data quality.

One recent example of a successful attempt to integrate genomic data with literature mined protein-protein interaction data is the work by Li and colleagues (Li et al., 2006); which was aimed at characterizing the molecular mechanism of *angiogenesis:* a process that involves the growth of new capillary blood vessels in healthy organisms and is particularly important for the progression of cancer. Initially, text-mining approaches were used to search pubmed articles for gene/protein co-citations in the context of angiogenesis. Pairs of potentially interacting proteins were then analysed for gene co-regulation using angiogenesis related microarray expression data derived from comparison of wild-type endothelial cells with cells from solid tumours, available from the Standford Microarray Database (SMD). Finally, a refined network of angiogenesis was constructed revealing promising gene targets, defining potentially new venues for therapeutic treatment of cancer induced angiogenesis.

Another useful form of integrated datamining approaches for microarray is functional analysis, which requires the incorporation of functional information onto gene expression data. Functional analysis reveals the biological significance of gene expression regulation by exposing the functional categories most enriched among the differentially expressed genes. Functional analysis is the subject of chapter VI and is there discussed in more details.

CHAPTER II: MICROARRAY ANALYSIS AFTER T7 BASED RNA AMPLIFICATION CAN DETECT PRONOUNCED DIFFERENCES IN GENE EXPRESSION

2.1. Introduction

Microarray technology offers a high throughput approach to transcript profiling on a genomic scale thereby providing deeper insights into global gene interactions in complex biological networks. In Neuroscience, microarrays have contributed a great deal to correlating gene expression profiles with complex neurological behaviours such as learning, memory (Klur et al., 2004; Li et al., 2005; McClintick et al., 2003) and nociception processing. However, the complexity and versatility of the functions encoded in the nervous system dictates numerous specializations of neuronal cellular subtypes primarily dedicated to certain aspects of information processing. Efficient characterisation of transcriptional profiles underlying specific processes of scientific interest requires the ability to select the relevant cellular subtypes to enrich key signals otherwise concealed by irrelevant expression information.

2. Microarray analysis after T7 based amplification can detect pronounced differences in gene expression.

2.1. Introduction

One recent advance in single cell isolation that has revolutionised the efficiency of microarray screening is the *laser capture microdissection* (*LCM*) technique, which has been already applied in characterising single neuronal cells with considerable success (King et al., 2005; Paulsen et al., 2009). However, it has proved a major challenge to integrate single cell isolation technology with subsequent transcriptional profiling using microarrays, primarily due to the impracticality of isolating enough target cells to achieve an optimum yield of RNA sufficient for chip hybridisation. This limitation is further enlarged by the need for replicate samples, essential for statistical inference.

Parallel to technological advances in single cell excision, increasingly sophisticated approaches to RNA amplification from small tissue samples have been developed and enhanced continuously for use with microarrays. Of great concern to the credibility of information obtained from screening for transcriptional regulation is the ability of the amplification process to maintain faithful representation of the abundance of the individual transcripts in the original sample. From this prospective, the T7 based amplification approaches, with their linear characteristics, have gained more popularity than the exponential PCR based methods.

2. Microarray analysis after T7 based amplification can detect pronounced differences in gene expression.

2.1. Introduction

In brief, the first version of the T7 based amplification protocol was published in 1990 by van Gelder and colleagues (Van Gelder et al., 1990) and relied on the T7 based in-vitro transcription of cDNA strands obtained from reverse transcription of RNA target molecules from the original RNA sample. This became the basis for the Affymetrix standard labelling protocol (Fig 2.1.1). A greater fold increase in RNA concentration was the product of an additional round of T7 linear amplification as proposed by Eberwine and colleagues (Eberwine, 1996). This was later adapted by Affymetrix to formulate their small sample amplification protocol (Fig 2.1.1). Modifications of the T7 amplification protocol have been explored to improve the efficiency and quality of the amplified transcript. One of the most fruitful of such was the attempt by Baugh and colleagues (Baugh et al., 2001) to reduce template-independent product by reducing the amount of primer and overall reaction volume. Kenzelmann and colleagues (Kenzelmann et al., 2004) improved the sensitivity of the T7 linear protocol by increasing the temperature during the RT reaction.

2. Microarray analysis after T7 based amplification can detect pronounced differences in gene expression.

2.1. Introduction

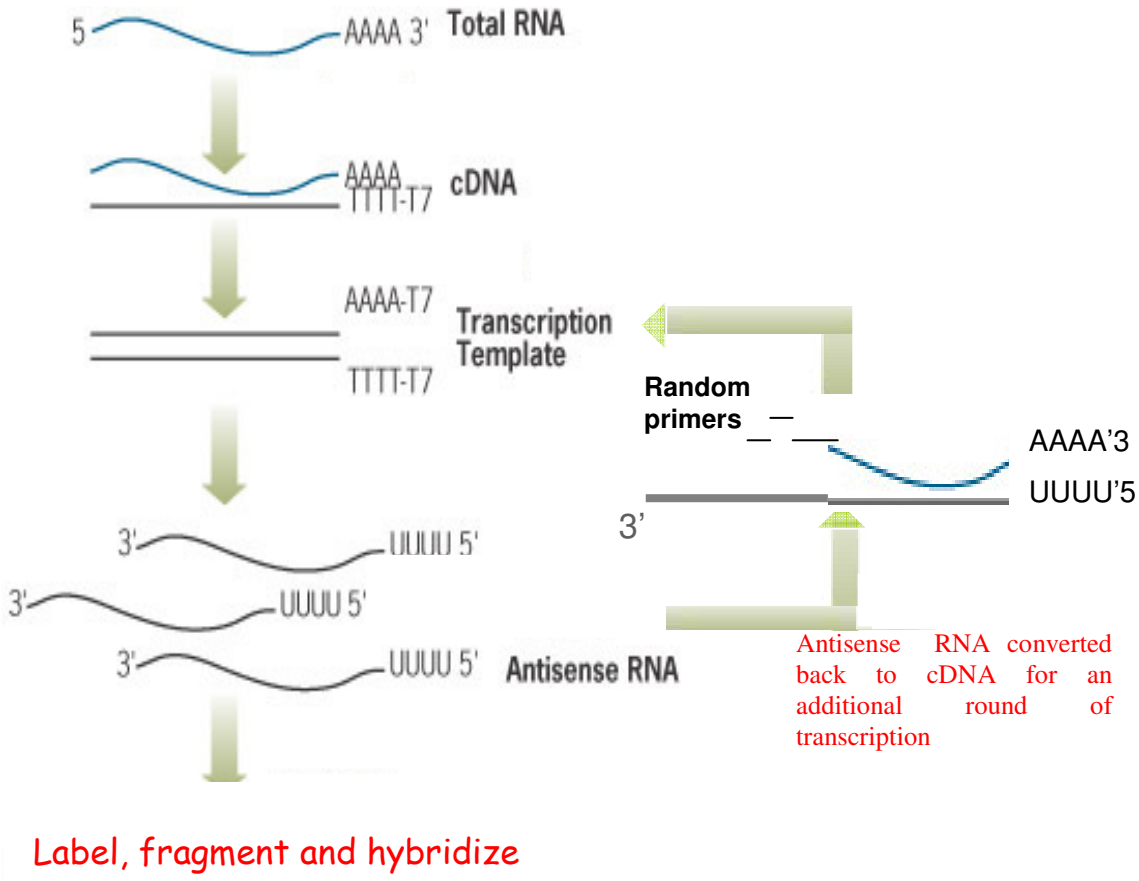


Figure 2.1.1. The Affymetrix T7 based small sample protocol. Briefly, RNA targets are first converted into cDNAs via a reverse transcription (RT) step using T7 promoter conjugated primers. The DNA strands complementary to the resulting cDNAs are taken through a transcription step to yield antisense RNA using the T7 polymerase. Together, these steps define the Affymetrix standard labelling protocol, which precedes RNA labelling, fragmentation and hybridision onto the arrays. With the small sample protocol, an additional round of transcription is performed to achieve higher-order amplification of the original sample. Hence, the antisense RNA from the previous round is converted back to cDNA via RT using random primers before a second transcription step is performed.

Despite the numerous benefits of the T7 based small sample amplification protocol, most notably its linearity and independence of transcript copy number in comparison to PCR based procedures, studies have reported

occasional failure to maintain the true abundance of targets, evidenced by a distortion in signal intensity with microarrays (Klur et al., 2004; Li et al., 2005; Wilson et al., 2004). This was attributed to a 3' bias effect that is thought to be related to the use of random hexamers to prime the reverse transcription (RT) reaction in the second round of transcription, corresponding to the additional round of amplification (Dumur et al., 2004; King et al., 2005; McClintick et al., 2003; Singh et al., 2005; Wilson et al., 2004). With priming that is remote from the 3' end of template antisense RNA (Fig 2.1.1), RT may not be successfully completed yielding truncated DNA strands that get lost in subsequent steps. This causes array probes originating from the 5' region of corresponding RNA templates (that is the 3' end of their antisense strands) to report artificially diminished intensity signals.

Importantly, a widely reported observation from studies featuring the assessment of the T7 small sample RNA amplification protocol for microarrays is the high reproducibility of the protocol. Signal intensities from independent amplifications of RNA samples from identical sources had proven highly correlated, implying that signal distortions were consistently reproduced by the protocol in biologically equivalent samples. The aim of this study is to address the question of whether such distortions are also reproducible in biologically distinct samples, which would imply that they

may be cancelled when taking the ratios. Thus, in this work, we take a different approach to the assessment of the T7 small sample protocol for microarrays by focussing on the intensity ratios instead of the absolute values of the intensities. This seems appropriate given that the usual prime target from microarray experiments is the analysis of the ratios to detect gene differential expression.

2.2. Methods

2.2.1. Microarray experiment design

In order to assess the effect of RNA amplification using the T7 Affymetrix small sample protocol on expression ratios, RNA samples were obtained from three biologically different tissue pools: the spinal dorsal horn tissue from naïve animals (SN), the spinal dorsal horn tissue from animals with axotomised sciatic nerve (SA) and the dorsal root ganglion (DRG) tissue from naïve animals. The Affymetrix standard protocol was used to generate three labelled cRNA samples from each tissue pool using 5 µg of total RNA as starting material whilst the T7 based small sample protocol was used to generate 4 labelled samples using 50 ng of starting material from each tissue pool (Fig 2.2.1). Material from the 21 RNA preparations was then hybridised to MOE430A arrays. For the rest of the article, we shall refer to the Affymetrix standard protocol and the small sample protocol as the *OneRA* (one round amplification) protocol and the *TwoRA* (two rounds amplification) protocol respectively, because the latter incorporates one additional round of amplification further to the initial round of amplification featured by the former (Fig 2.1.1). It is important to note that the experimental phase of this study including animal handling, tissue collection, RNA extraction, RNA

amplification, labelling and hybridization was exclusively performed by experimentalists from the London Pain Consortium (LPC) and that a detailed description of the experimental phase can be found in the published version of this work (Diboun et al., 2006).

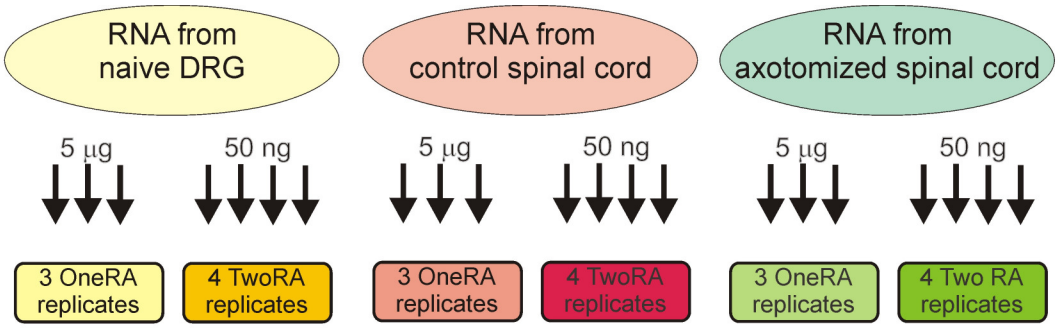


Figure 2.2.1. Experimental design. Three biologically distinct tissue pools were obtained. From each tissue pool, 3 RNA samples versus 4 RNA samples were obtained using the OneRA and the TwoRA protocols respectively.

2.2.2. Microarray data analysis

Feature intensity values from scanned arrays were background-corrected, normalised and reduced into expression summaries using the *GCRMA* algorithm implemented as a function in the *GCRMA* library of the *Bioconductor* package (Gentleman et al., 2004) of *R,* the open source environment for statistical analysis. Arrays were then inspected for quality

control using a variety of built-in QC tools from the Bioconductor *Affy* package. QC consisted of visual examination of probe array images, scatter plots from replicate arrays, hierarchical clustering of array hybridisations as well as RNA degradation plots performed on probe raw intensities. Detection calls indicating the presence or absence of signal from each probeset were obtained by processing the raw data with the Microarray Analysis Suite 5.0 (MAS5). To obtain a consensus detection call across replicate hybridizations, a probeset was considered to be present if it received a P (present) detection call from all replicates or n-1 replicates with an M (marginal) call from the remaining replicate. Consensus A (absent) detection calls across replicates were determined in the same way.

For further analysis investigating the 3' bias effect by the TwoRA protocol, probesets 3' locations were obtained by downloading the MOE430A probe annotation files made available by the Affymetrix online support at http://www.affymetrix.com/analysis/index.affx. A probeset location was considered equal to the 3' distance of the probe most distal from the 3' end of the corresponding RNA target in the set. To test for differential expression, we used the Bayesian adjusted t-statistics from the Bioconductor *Limma* (linear models for Micoarray data) package (Smyth, 2004), applied with an *FDR* multiple testing correction.

2.3. Results

In this study, we undertake a detailed analysis of RNA amplification for microarrays using the Affymetrix small sample protocol (TwoRA). This analysis was performed using control data from standard protocol (OneRA) preparations as reference. While, the main objective of this study is to assess the extent to which biologically relevant variations in gene expression can be detected in the TwoRA, we begin by confirming the reproducibility of the TwoRA protocol and show comprehensive evidence for the protocol 3' bias effect.

2.3.1. Reproducibility and fidelity in maintaining expression levels

Scatter plots of log2 intensities from paired TwoRA replicates from all three biological groups show expectedly high level of consistency similar to that observed with the OneRA replicates from all groups (Fig 2.3.1-A&B); with (r) ranging from 0.990 to 0.994. However, comparing the average log2 intensity values from the OneRA versus the TwoRA (Fig 2.3.1-C) for a single tissue,

we see evidence of variability confirming that the TwoRA protocol occasionally distorts the signal.

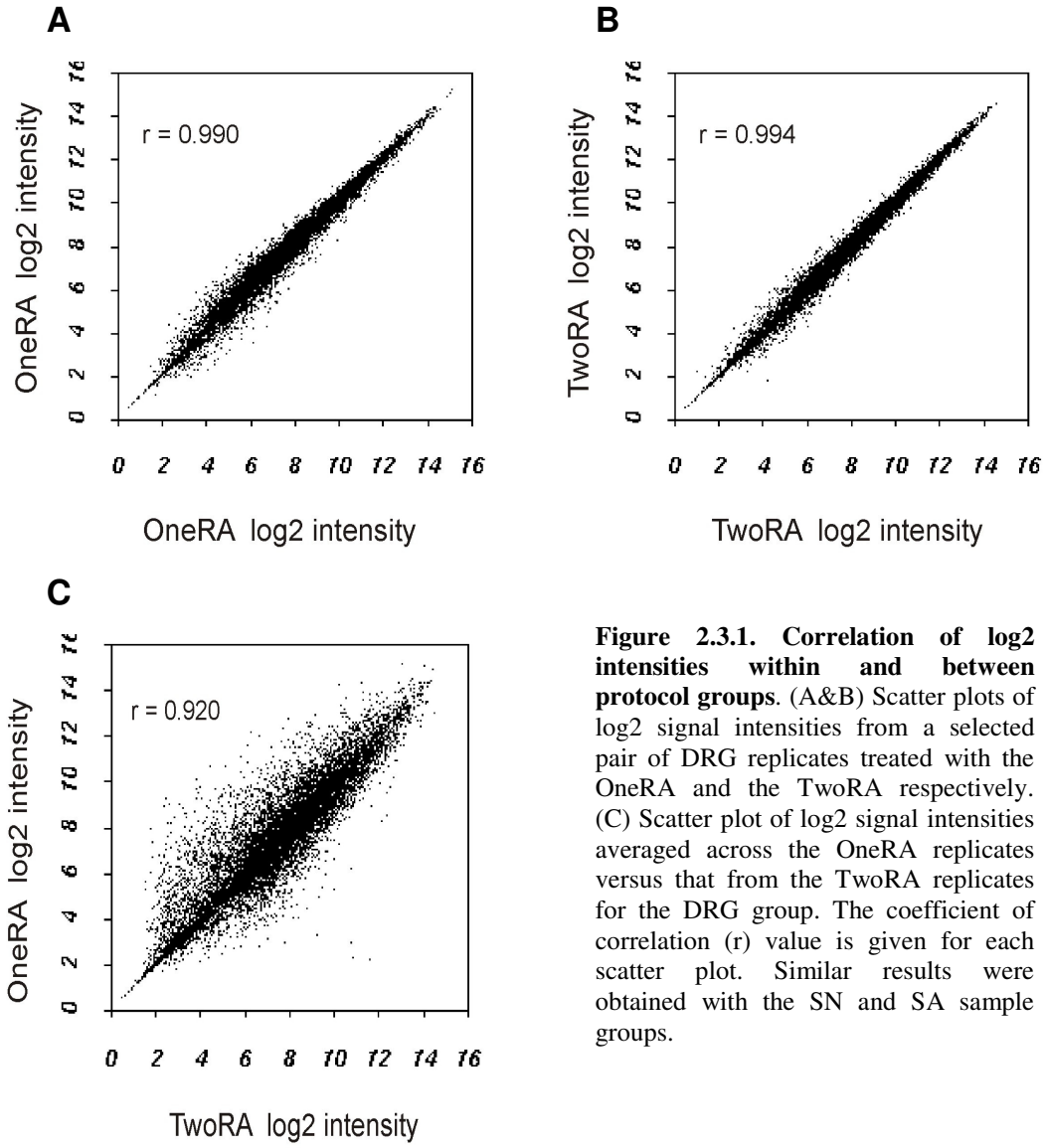


Figure 2.3.1. Correlation of log2 intensities within and between protocol groups. (A&B) Scatter plots of log2 signal intensities from a selected pair of DRG replicates treated with the OneRA and the TwoRA respectively. (C) Scatter plot of log2 signal intensities averaged across the OneRA replicates versus that from the TwoRA replicates for the DRG group. The coefficient of correlation (r) value is given for each scatter plot. Similar results were obtained with the SN and SA sample groups.

We used an ANOVA approach to confirm that the variability between protocol groups is greater than that among replicates within each group. In particular, a one-way two-levels ANOVA analysis was performed for each gene separately with 3 measurements from the OneRA (level1) and 4 measurements from the TwoRA (level 2). First, the between group mean sum of squares *MSA* as well as the mean residual sum of squares *MSE* were calculated. The median of the MSA (across the genes) was higher than the median of the MSE (given in parenthesis) in all biological groups: DRG 0.050 (0.023), SN 0.062 (0.016), SA 0.068 (0.02).

To test whether protocol variability is significantly greater than the residual variability, we derived p-values from the F-values (MSA/MSE) for each gene (using the upper tail of an F-distribution with 1 and 3 + 4 – 2 degrees of freedom). In fact, the p-values were far from uniformly distributed. Storey suggests the following estimate of the proportion of hypotheses from the null using p-values: the fraction of p-values above the median p-value m, divided by (1-m) (Storey and Tibshirani, 2003). This results in the following estimates of the proportion of genes with significantly higher amplification variability: DRG 47%, SN 50%, SA 41%. That is, in all cases at least 40% of genes show differences between protocols, which are not explained by variability within replicates. The ANOVA analysis was advised on by Prof Lorenz Wernish

from Birkbeck College, who jointly supervised this work with LPC principle investigator Prof Martin Koltzenburg.

Distortions in signal intensity following TwoRA are likely to be caused by the TwoRA protocol 3' bias effect whereby, as explained in the introduction, the use of random hexamers to prime the RT step during the second round of amplification favors the representation of parts of the RNA close to the 3' end. To affirm the 3' bias feature of the TwoRA protocol, individual array probes from each probeset were numbered 1 to 11 from the 5' end of corresponding transcripts. For each chip, raw intensities corresponding to the same probe number across all probesets were averaged. The resulting probe average intensities were correlated with the corresponding probe numbers. The results appear in Figure 2.3.2.

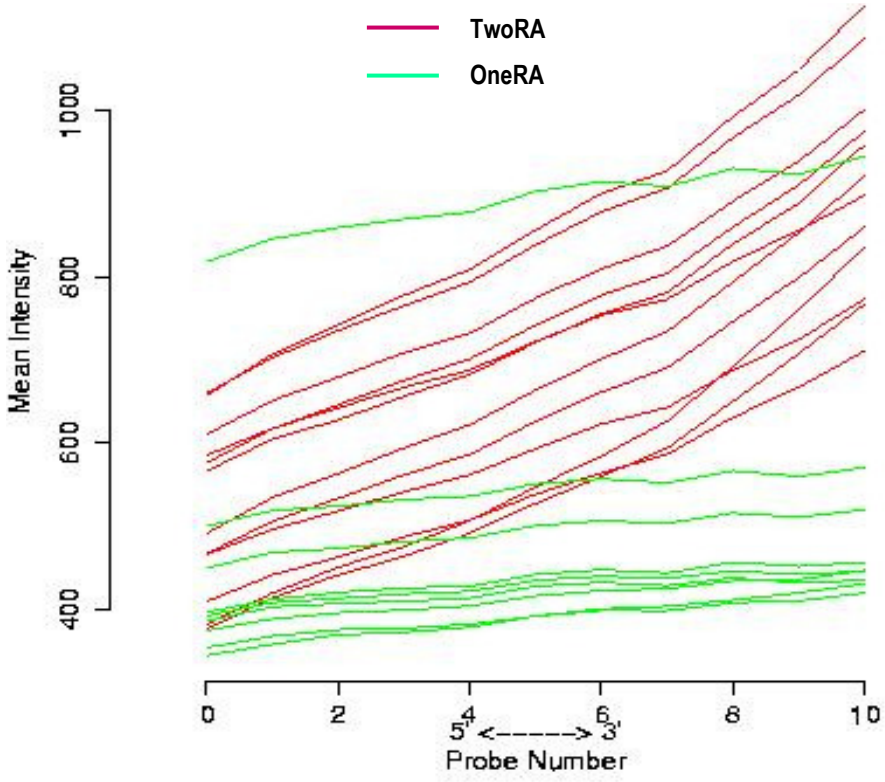


Figure 2.3.2. RNA digestion plot. Probes from each probeset were numbered by distance from the 5' end of target RNA (probe numbers shown on the x-axis). For each chip, the average raw intensity value from probes with the same probe number across all probesets were calculated (y-axis). Each line corresponds to a single chip.

The mean probe intensity from the OneRA target hybridisations seems to be fairly constant across the ranks of the various probes in Figure 2.3.2. In contrast, probe mean intensity from the TwoRA hybridisations is clearly dependent on probe location and is highest at close proximity from the 3' end. Importantly, array normalisation seemed to have no effect on the bias

observed (this was done by repeating the analysis on normalised probe intensities).

In this study, in addition to the RNA digestion plot (Fig 2.3.2) used frequently in the literature to highlight the TwoRA protocol 3' bias effect, we undertook a different analysis that associates, for the first time, distortions in the signal following TwoRA to probeset location on template RNA targets. This analysis was performed using data from the DRG tissue pool and similar results were obtained with the remaining tissue pools SA and SN. Thus, we correlated the differences in log2 intensity in the DRG samples following TwoRA ($\Delta log2IN = log2IN_{TwoRA} - log2IN_{OneRA}$) with the probesets 3' locations on corresponding targets (see methods for a description of how these locations were obtained) (Fig 2.3.3). The trend suggests that probesets distal from the 3' end are more likely to endure an attenuation of signal intensity following TwoRA whilst those close to the 3' end are likely to show intensification of signal.

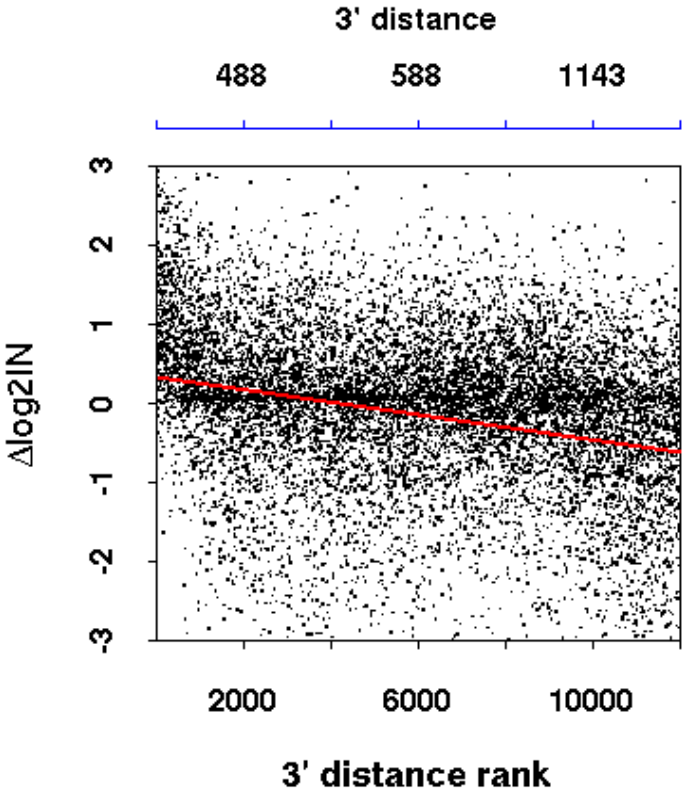


Figure 2.3.3. Deviation in log2 intensity following TwoRA (Δlog2IN) as a function of probeset 3' distance rank. Δlog2IN values on the y-axis were calculated by subtracting the mean OneRA from the mean TwoRA probeset log2 intensities. The x-axis shows the ranks of probesets locations. Probesets locations are relative to the 3' end of the transcripts. Since the probesets locations have a skewed distribution, their ranks were used instead of their absolute values; this allows dispersion of data points. The actual probesets locations that correspond to the rank intervals on the x-axis are shown on the blue horizontal axis on the top of the figure. The regression line is shown in red. Only data from the DRG preparation were used, similar results were obtained with the SN and SA groups.

In a separate but related analysis, probesets whose absolute Δlog2IN values were greater than 2 were reviewed for their 3' location distribution. This was compared to the distribution of 3' location of all probesets on the array (Fig 2.3.4). The latter appears to be skewed and peaks at around 600 bp (Fig 2.3.4). The distribution of 3' location from probesets with intensified signal following TwoRA (Δlog2IN>2) shows an additional peak to the left suggesting a distinct population of probesets closer than average to the 3' end of RNA targets (Fig 2.3.4-A). This is further highlighted by a decrease in the 25% quantile relative to the overall population of probesets in the boxplot on Figure 2.3.4-C.

Conversely, the 3' location distribution from probesets with attenuated signal shows a second peak to the right indicating an overrepresentation of more distal probesets relative to the 3' end of RNA targets (Fig 2.3.4-B). This corresponds to an increase in the 75% quantile (Fig 2.3.4-C).

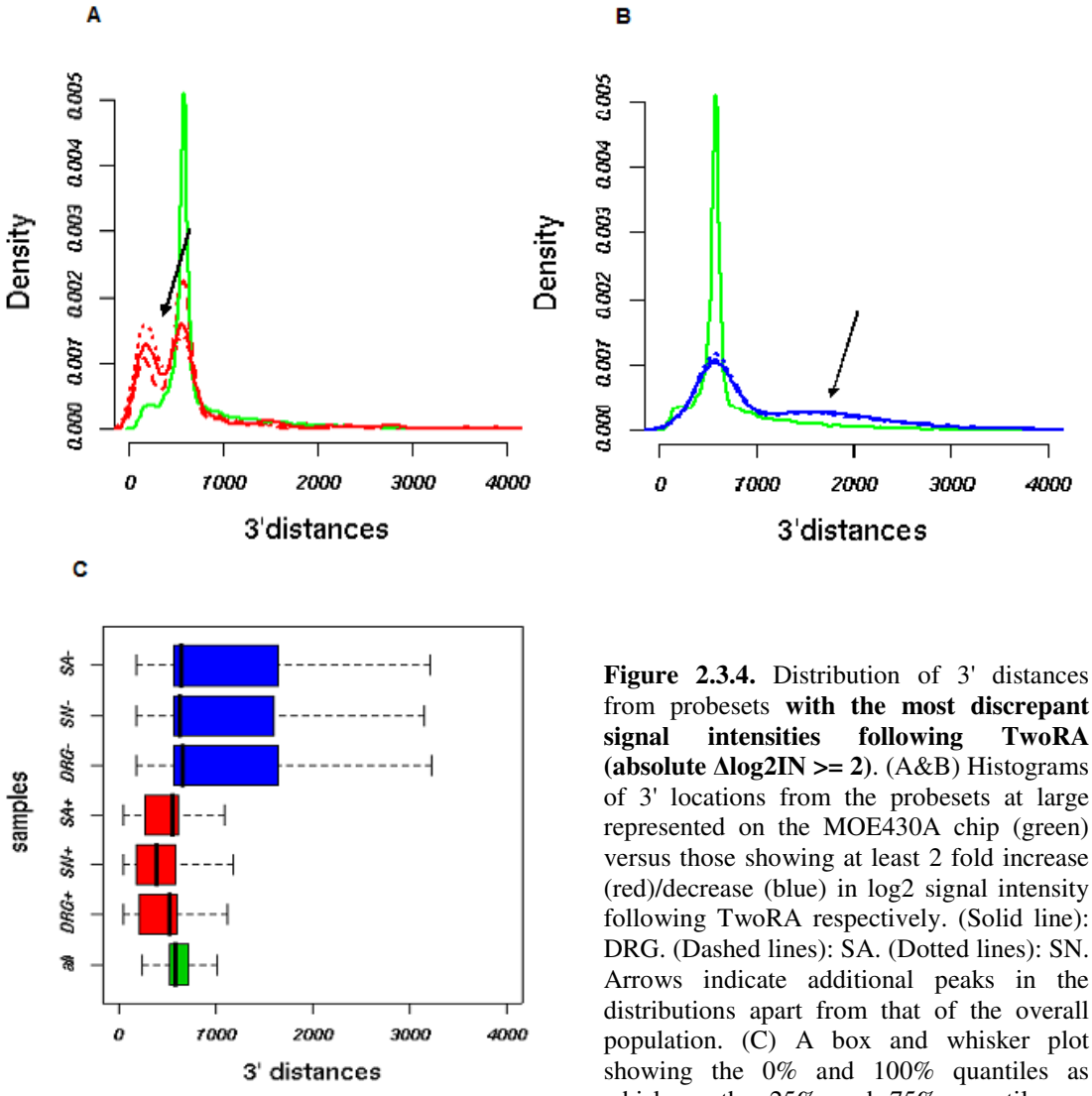


Figure 2.3.4. Distribution of 3' distances from probesets **with the most discrepant signal intensities following TwoRA (absolute Δlog2IN >= 2)**. (A&B) Histograms of 3' locations from the probesets at large represented on the MOE430A chip (green) versus those showing at least 2 fold increase (red)/decrease (blue) in log2 signal intensity following TwoRA respectively. (Solid line): DRG. (Dashed lines): SA. (Dotted lines): SN. Arrows indicate additional peaks in the distributions apart from that of the overall population. (C) A box and whisker plot showing the 0% and 100% quantiles as whiskers, the 25% and 75% quantiles as boxes and the 50% quantile as horizontal dash within the box. The plot summarises the distributions shown in A and B. On the y-axis, (+) indicates increase in signal intensity following TwoRA, (-) indicates decrease in signal intensity following TwoRA.

2.3.2. Fidelity in mainting expression ratios

The ultimate aim of microarrays is the identification of differential expression. Thus, a good amplification protocol should faithfully maintain expression ratios. To verify this, we cross-compared expression ratios from biologically distinct tissue samples treated with the OneRA and the TwoRA protocols.

First, we considered the (SA,SN) pair. Expression ratios on log2 scale from the OneRA samples were correlated with their equivalents from the TwoRA (Fig 2.3.5-A). The significant changes in expression, including the well established *activating transcription factor 3* (Wiggins et al., 2004) and *small proline-rich repeat protein 1A* (Wright and Snider, 1995) in the literature, seem to be consistent in the TwoRA and the OneRA groups (Fig 2.3.5-A). However, there are relatively few differences in gene expression between these two biological samples, probably due to the fact that the tissue from the injured animals included areas of the spinal cord not affected by the axotomy, which could have caused a dilution of effect in the relevant areas. To reliably evaluate the effect of the TwoRA protocol on ratios, a larger profile of differential expression is needed. This was possible with the (DRG,SN) pair. Thus, we decided to base our assessment of the effect of the TwoRA on ratios from the (DRG,SN) samples.

Encouragingly, log2 expression ratios from the (DRG,SN) treated with the OneRA and the TwoRA protocols are comparable (Fig 2.3.5-B); though they show more variability than their counterparts from the (SA,SN) pair (Fig 2.3.5-A). Moreover, the regression line (shown in blue, Fig 2.3.5-B) appears to be shifted from the diagonal in a way that suggests that the expression ratios are on average slightly lower in the TwoRA relative to the OneRA with the (DRG,SN) pair.

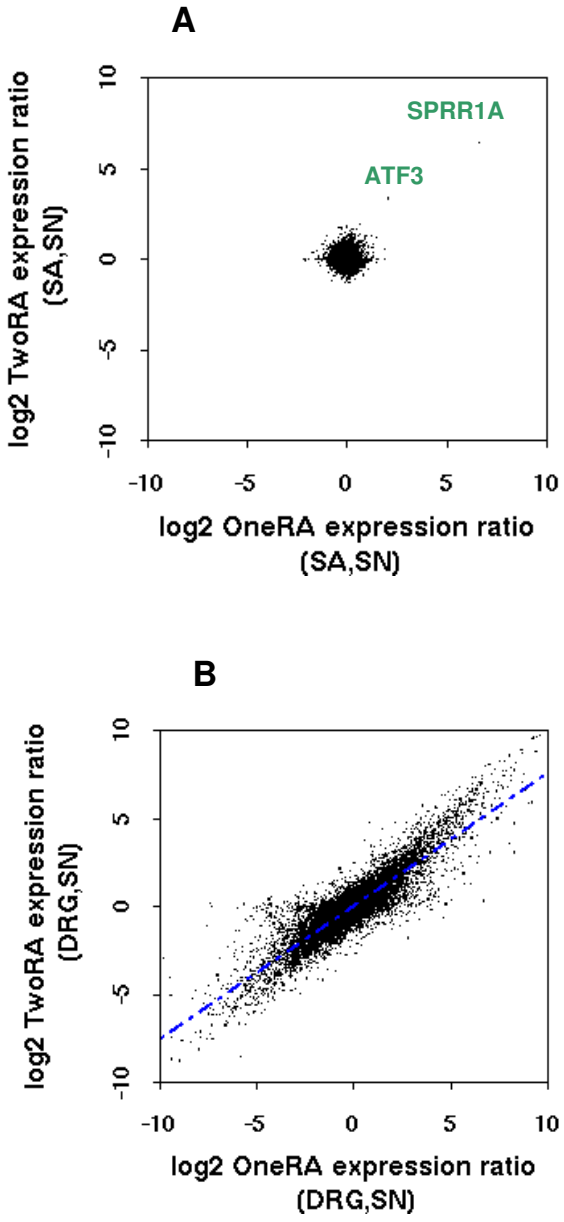


Figure 2.3.5. Correlation of log2 ratios from the OneRA and the TwoRA for the (SA,SN) and (DRG,SN) sample pairs, A&B respectively. Not many changes in gene expression are detected with the (SA,SN) pair in A. Many more changes in gene expression are observed with the (DRG,SN) pair in B. The regression line is shown in blue and indicates that the TwoRA ratios are overall smaller than their OneRA counterparts.

-Variation in ratios

From our previous analysis, we know that the TwoRA protocol may shift the absolute intensity levels. However, this only affects expression ratios if the intensity is shifted unequally in the two biological samples. That is, deviations in intensity (Δlog2IN) following TwoRA, that differ in the two samples, can result in variability in the expression ratios from the OneRA and the TwoRA groups. To get further insights into how unequal shifts in the intensity level following amplification of different biological samples affect the expression ratios, we ranked probesets by the absolute difference in their OneRA and TwoRA log2 expression ratios in a descending order and selected the top 100 for further analysis. Specifically, we examined the average intensities from these selected probesets in all four groups: the OneRA and the TwoRA DRG, SN. The resulting intensity profiles were classified into four categories depending on the direction of change in intensity after TwoRA and the tissue where this change occurred (Fig 2.3.6). The most populated categories show a significant reduction in the intensity in one of the samples whilst the intensity in the other sample is minimally reduced (Fig 2.3.6-A&C). Less frequently, the intensity increases after TwoRA in one of the samples but not in the other sample (Fig 2.3.6-B&D).

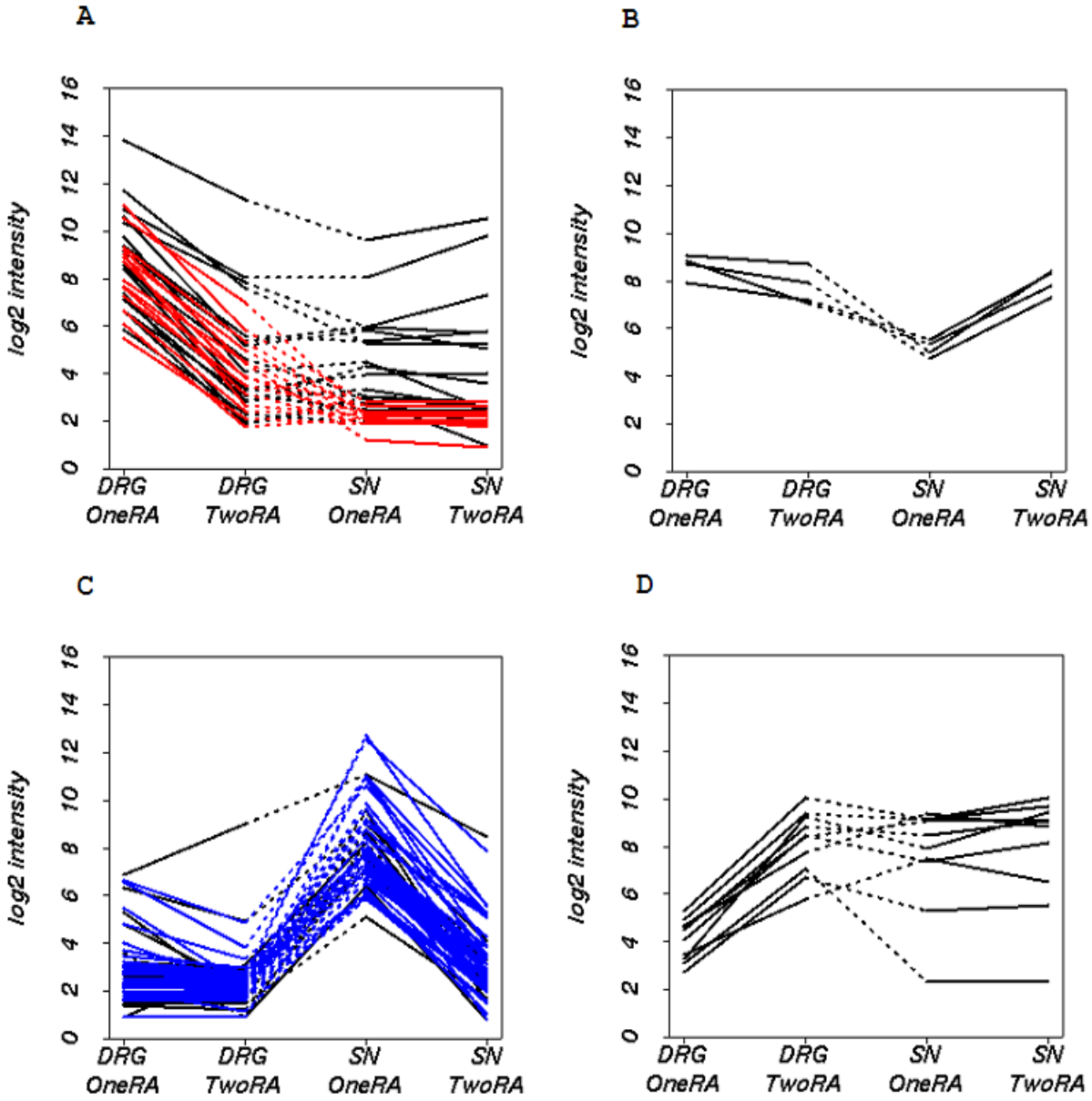


Figure 2.3.6. Intensity profiles of probesets with top 100 most deviant expression ratios following TwoRA. The profiles are classified into four categories: A&C, the intensity is reduced in the tissue sample where the gene is more expressed (DRG, SN respectively) following TwoRA. B&D, the intensity is increased in the sample where the gene is less expressed (SN, DRG respectively). Solid lines mark the shift in intensity from OneRA to TwoRA for one tissue sample. Dashed lines link the intensity data for equivalent probesets in the two biological samples. In colour are probesets with absent call in the SN (red) and DRG (blue).

Interestingly, with all four categories of deviant probesets shown on Figure 2.3.6, expression ratios appear to be reduced rather than inflated following TwoRA. Moreover, the majority of the selected probesets have varying intensity levels in the DRG versus SN, OneRA. Frequently these probesets have absent calls in one sample but are associated with high levels of expression in the other sample (shown as coloured lines in Figure 2.3.6); which may explain the deviation in expression ratios following TwoRA. If one takes the example of HipK2, the log2 intensity in the SN was reduced from 8.20 in the OneRA to 0.73 in the TwoRA. However, HipK2 is absent in the DRG (the OneRA log2 intensity is 0.87), thus an equivalent reduction in the intensity level in the DRG sample is not possible (floor effect). As such the log2 expression ratio for HipK2 is shifted from -7.33 in the OneRA to 0.15 in the TwoRA. Alternatively, in other cases, if amplification increases the intensity in one sample, an equal increase in the other sample would not be possible if the intensity was close to saturation (ceiling effect).

Thus, distortions in the expression ratios may occur when a shift in intensity (Δlog2IN) in one sample cannot be mirrored in the other sample because it would cause the intensity to fall outside the dynamic range of the scanner. To assess the extent to which this phenomenon explains the deviation in expression ratios between the OneRA and TwoRA for the (DRG, SN) pair, we

undertook the following analysis. We selected all probesets where a shift in intensity following TwoRA in one sample would cause the intensity in the other sample to fall outside the detectable intensity range, that is below the background noise or higher than the saturation level. These limits were chosen to be the 3% and 98% quantiles of the distribution of signal intensity from a randomly selected chip, respectively. The analysis was conducted by first determining the absolute Δlog2IN = $|$(log2 TwoRA – log2 OneRA)$|$ for each probeset from each biological group in the (DRG,SN) pair. Then, if the maximum shift in intensity ($|\Delta\text{log2IN}|$) is featured in the DRG group, we shift the corresponding OneRA log2 intensity from the SN group by the same amount and vice versa. If the resulting value is outside the chosen limits, the probeset is selected by our analysis.

Since the selected probesets show a **f**loor and **c**eiling **e**ffect, we shall refer to them as *FCE probesets* for the rest of the chapter. Interestingly, the FCE probesets correspond to those probesets showing the most pronounced variation in shifts in intensity following TwoRA, i.e. featuring the most varying $|\Delta\text{log2IN}|$ between the DRG and SN samples (colored in red, fig 2.3.7-A). Consequently, these same probesets show the most deviant (DRG,SN) expression ratios following TwoRA (colored in red, Fig 2.3.7-B). In fact, the correlation between (DRG,SN) expression ratios across protocols

(r) = 0.89 is improved to 0.93 when the FCE probesets are excluded. Interestingly, we found that the FCE probesets show consistent Δlog2IN following TwoRA with the (SA,SN) pair (in red, Fig 2.3.7-C). This is because unlike the (DRG,SN) pair, the FCE probesets have similar OneRA intensities in both biological groups SA and SN (recall, very little differential expression was observed between the SA and SN biological groups in Figure 2.3.5-A) and hence a shift in intensity in one biological sample should be possible in the other sample following TwoRA.

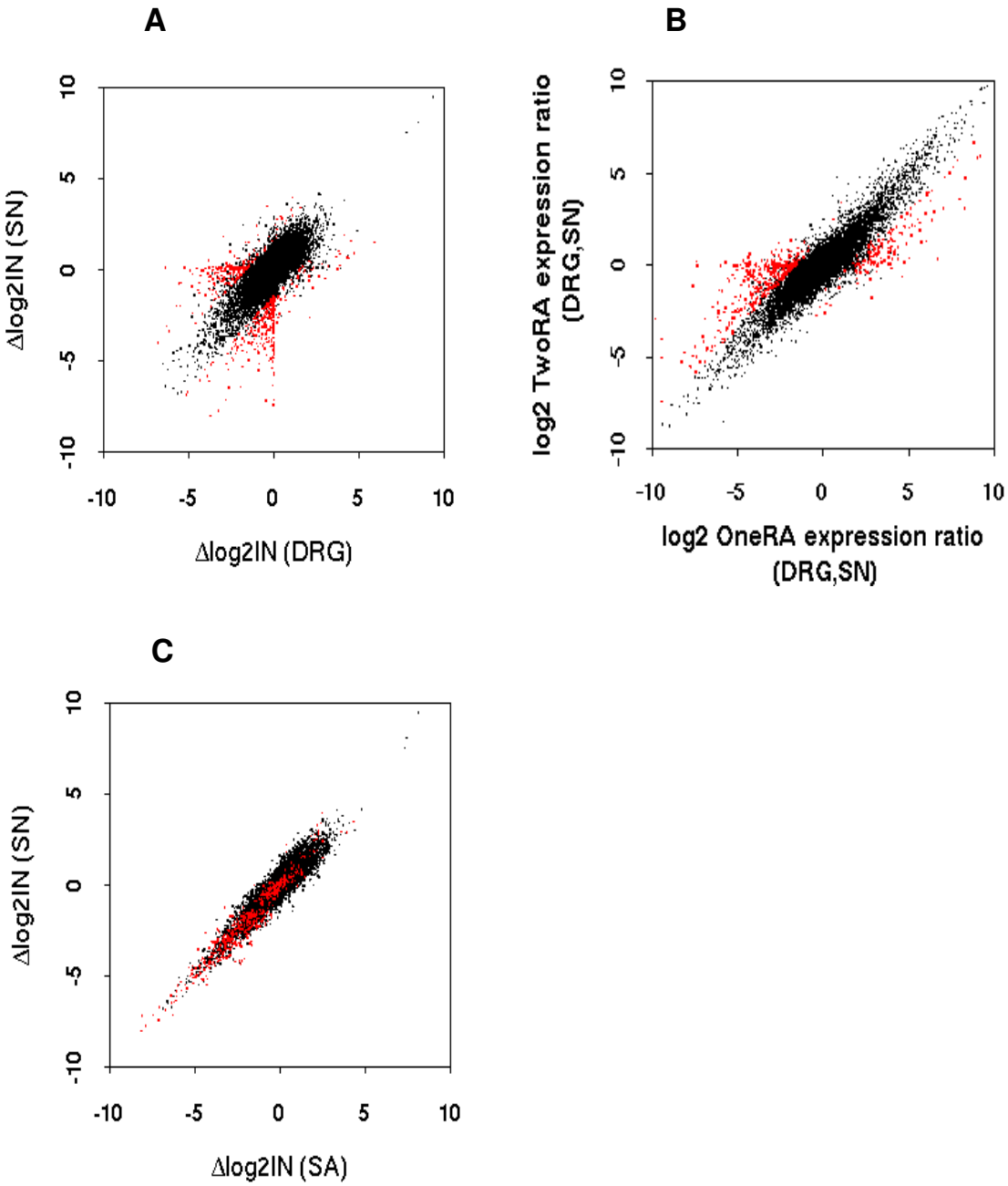


Figure 2.3.7. Deviation in log2 expression ratios (DRG,SN) following TwoRA and its origin. (A&C) Scatter plots of Δlog2IN values for the (DRG,SN) pair and the (SA,SN) pair respectively. (B) Scatter plots of log2 expression ratios from the OneRA and the TwoRA for the (DRG,SN) pair. For instance, the log2 OneRA expression ratio for the (DRG,SN) pair is log2 OneRA DRG – log2 OneRA SN. Δlog2IN in A&C were calculated by subtracting the log2 OneRA intensity from the log2 TwoRA intensity. Points in red in (A) are probesets where the intensity in one sample could not be shifted as much as in the other sample because the intensity cannot lie outside the dynamic range of the scanner. These are referred to as FCE (floor & ceiling effect) probesets and have varying expression ratios with TwoRA (colored in red, B). Though, for these same probesets, the Δlog2IN values in the SA and the SN groups are fairly consistent (points in red, C).

2.3.3. Maintaining the statistical significance of the expression ratios

The primary aim of a microarray experiment is to detect significant changes in gene expression. However, our results suggest that large ratios in the OneRA may get reduced following TwoRA, which may hinder the detection of differentially expressed genes. Indeed, we found good evidence from the literature to suggest that 9 from the 10 genes with the most severely reduced expression ratios following TwoRA are indeed differentially expressed between the SN and DRG.

Despite shifts in expression ratios, genes can remain significant following TwoRA if their ratios are still large relative to the average in the TwoRA. Moreover, among the population of genes with high expression ratios in the OneRA (Fig 2.3.5-B), many do maintain their ratios in the TwoRA, most likely due to a faithful two rounds amplification (TwoRA) of transcripts in the two biological samples.

We applied the limma statistical test to identify transcripts differentially expressed in the (DRG,SN) tissue samples prepared with both protocols (OneRA and TwoRA). An FDR based multiple testing correction was used

and genes were ranked by their FDR values in ascending order. 87% of the top 100, 300 and 500 most significant genes were consistently found common to the OneRA and the TwoRA comparisons.

For a more global assessment of the effect of distortions in expression ratios on their statistical significance, we used a scatter plot of **n**egated **l**n **p**-**v**alues (nlPv) from the limma analysis of the OneRA and the TwoRA (DRG,SN) (Fig 2.3.8). The FCE probesets are highlighted in red and it can be seen that their nlPv are least correlated between the two protocols, due to distortions in the expression ratios (scatter on Fig 2.3.7-B).

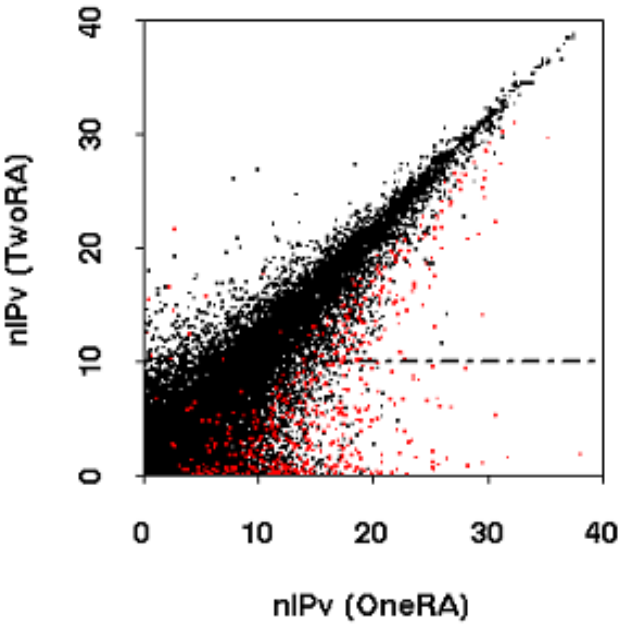


Figure 2.3.8. Effect of distortion in expression ratios on their statistical significance following TwoRA. Scatter plots of FDR corrected nlPv (negated log transformed p-value) from the Limma analysis of the OneRA and the TwoRA DRG and SN samples. As a result of negating the p-values, large nlPv indicate stronger evidence of differential expression. Data points in red represent the FCE probesets. The dashed line is at nlPv = 10 in the TwoRA, above which genes may be considered significant.

Amongst the FCE probesets on Figure 2.3.8, some still show reasonable nlPv following TwoRA (>10). Inspection of these genes revealed that they have large expression ratios in the OneRA and moderate ratios in the TwoRA (the median log2 expression ratios was 5.09, 2.52 respectively). By contrast, those FCE probesets with low nlPv (<10) in the TwoRA have had their log2 expression ratios reduced severely following TwoRA (median log2 ratio in the TwoRA = 0.43). Interestingly, the latter have on average moderate expression ratios in the OneRA (median log2 ratio in the OneRA = 2.8). This is expected since with moderate expression ratios, any reduction would have a greater impact on their statistical significance. Indeed, looking at the whole population of probesets, out of those with an nlPv between 10 and 20 in the OneRA, only 69% have an nlPv above 10 in the TwoRA, compared to probesets with high nlPv (> 20) in the OneRA where 87% of them have nlPv above 20 in the TwoRA. This suggests that the TwoRA protocol is more suitable with experiments where large differences in gene expression are occurring.

2.4. Discussion

Microarray technology is currently limited by the need for relatively large transcript quantities, which makes it incapable of handling small biological samples. The T7 in-vitro transcription has been widely explored to achieve a linear amplification of RNA targets for microarrays. Although, the reproducibility of such techniques and their fidelity in maintaining absolute levels of expression have been extensively analysed, much less is known about their ability to accurately reproduce differential expression in distinct biological samples; which we hope to have addressed in this study.

Our analysis confirms the high reproducibility of the small sample TwoRA protocol and the occasional failure in its fidelity to maintain the original levels of gene expression. In this study, robust analyses were used to confirm the 3' bias role in signal distortion. Importantly, the fact that the intensity range is limited by background noise on one end and saturation on the other end implies that intensity may only be shifted by a limited amount. This relationship bears important consequences on the consistency of the TwoRA protocol in amplifying targets with varying intensities across different samples. Thus, the shifts in intensity following amplification will not appear to be equivalent in two different biological samples if the shift in one sample is

limited by the range of the scanner. This has the consequence of distorting the expression ratios, as clearly demonstrated by our data.

Unsurprisingly, the statistical significance of expression ratios is only affected when the expression ratio in the TwoRA is reduced to the point where it can no longer be distinguished from noise. Importantly, large ratios are less likely to be critically diminished and more likely to remain significant following TwoRA. This explains why despite the distortions in ratios in our dataset, there was up to 87% agreement in the most significant genes (nlPv > 20) from the TwoRA and OneRA (DRG,SN). On the other hand, less agreement was observed among the less pronounced ratios (69%) since distortions are more critical. This leads us to the important conclusion that TwoRA may affect the statistical significance of genes with moderate expression ratios to a greater extent.

2.5. Conclusion

We conclude that the Affymetix small sample amplification protocol is useful with the following caveats: First, it should be only used when tissue homogeneity is a crucial factor and sufficient amounts of starting material cannot be obtained by any other means. Secondly, target amplification using the small sample protocol appears to be suitable in situations where big differences in gene expression are expected. Fortunately, it is reasonable to expect large differential expressions with experiments characterizing different cells within a mixed tissue where amplification of transcript is necessary. However, expression data obtained from amplified samples might be less suitable for more comprehensive numerical analysis, for example characterizing regulatory networks, due to the problems caused by possible shifts in signal and expression ratios.

CHAPTER III: A DATABASE OF GENE EXPRESSION DATA FROM ANIMAL MODELS OF PERIPHERAL NEUROPATHY

3.1. Introduction

3.1.1. Gene expression databases

Microarray databases are essential for effective management of microarray data. Besides storing raw and processed numerical data, various types of annotations need to be recorded that capture information on the scanning process of individual array hybridisations and downstream analysis steps leading to the data. Also, from earlier stages in the microarray experiment, annotations describing the origin, extraction and the manipulation of the biological material as well as the array platform used provide essential contextual information that is crucial for a correct biological interpretation of microarray data and integration of discrete microarray datasets. To this end, *MIAME* or the **m**inimum **i**nformation **a**bout a **m**icroarray **e**xperiment (Brazma et al., 2003), was developed as a data model standard for microarray data capture. The MIAME guidelines were later formally encapsulated within an

object model framework (MAGE-OM) and an accompanying data exchange format based on the XML language (MAGE-ML) (Spellman et al., 2002); an effort that was jointly coordinated by members of the MGED society .

The development of standards for microarray data annotation and exchange formats laid the ground for public microarray data repositories to be developed. Most popular among these are ArrayExpress (Brazma et al., 2003), GEO (Barrett et al., 2009), the Stanford Microarray Database (Demeter et al., 2007) and CIBEX (Ikeo et al., 2003). Furthermore, requirements were put in place for microarray studies to be made accessible in public microarray data repositories in the MIAME format by prominent scientific journals as part of the submission process. This allowed microarray data repositories to fulfil their maximum potential by leveraging the great amount of expression data produced worldwide in a standard format that is amenable to exchange.

However, the need for local microarray database facilities that serve the needs of small communities undertaking collaborative research projects was soon acknowledged. Tools emerged that distribute their full source code and provide built-in facilities for microarray data storage, data analysis and management of user accounts; providing an ideal easy to use platforms for specialised labs undertaking microarray work. One of the earliest of such tools

that is MIAME-compliant is the BioArray Software Environment (BASE) (Saal et al., 2002). BASE uses a server-client architecture framework that features a centralised database core for storing the data at the server side whilst allowing online user access to the database at the client side. BASE features an integrated framework for the storage and analysis of microarray data. Within BASE, analysis scenarios may be created that combine varying steps of data manipulation and further more explore variations at each analysis step. Results are stored in a hierarchical structure that reflects both the specificity and the timing of each analysis step in the workflow. Users may share data and analysis results between them according to well-enforced rules; thereby allowing management of microarray data at a laboratory/project scale.

Other free software microarray platform solutions also exist that feature varying points of focus. For instance, with many tools, the main aim was to provide a comprehensive built-in suite of analysis tools that is fully integrated with the internal microarray data structure; examples are TM4 and Gecko (Saeed et al., 2006; Theilhaber et al., 2004). More recently, more free software platform solutions have emerged that extend the classical set of analysis methods applicable to individual datasets to provide the necessary tools that allow disparate datasets, possibly originating from different array platforms, to be efficiently combined (WebArray, Xia et al., 2005). Other microarray data

management systems have focussed on usability with respect to data analysis (EzArray, Zhu et al., 2008), capture of the microarray data and experimental details in MIAME as well as ease of import/export of such standardised data (maxd, Hancock et al., 2005). Finally, systems have also been developed that address the need for fine-tuned user privileges that reflect the varying ways in which different types of users may wish to interact with the data (MiMiR, Tomlinson et al., 2008).

It is a fact that at the biological level, the potential of microarray technology is only fully realised when disparate microarray expression datasets pertaining to a common biological subject are combined together and furthermore integrated with other types of biologically relevant data. Indeed, there are many examples in the literature of biologically specialised microarray databases that were designed to serve research communities dedicated to a particular research subject in an effort to consolidate their data. Examples are the Genopolis Microarray Database specialised in immunopathology (Splendiani et al., 2007), the Gene Aging Nexus (GAN) database (Pan et al., 2007), the Cancer microarray database OncoMINE (Rhodes et al., 2004) and the Staphylococcus Aureus Microarray meta-database SAMMD (Nagarajan and Elasri, 2007).

Ironically, with respect to the microarray data capture model, many of such subject dedicated microarray databases, including GAN, OncoMINE, the Pancreatic Expression database (Chelala et al., 2007) and SAMMD, don't use MIAME. Rather, they tend to only capture essential information about the microarray experiments that are most relevant to the interpretation of the data relative to the key common biological topic. This is because, for most of these resources, the mission is to corroborate information on gene differential expression via combining biologically relevant datasets obtained from public repositories whilst the full MIAME specifications of the original microarray experiments are already defined in the source repository. Typically, these biologically specialised microarray databases tend to have their own data model and analysis tools and focus on methods that allow integrative analysis of disparate microarray datasets such as cross-platform analysis and normalisation.

On the other hand, there have been examples where research communities have successfully adopted free generic microarray software platforms to set up local microarray databases tailored to their specific research needs. The use of free software implies the chance to benefit from an already existing platform for data storage and analysis that can be further extended. For instance, in the Institute of Food Research (IFR), BASE was successfully used to create a

local database capturing more than 4600 prokaryote and eukaryote microarray hybridisations and is being continuously extended and optimised (Mark Alston et al.,2004).

The maxd software (Hancock et al., 2005) has had even more success among specialised microarray research communities. Already two major consortia: the Generation Challenge Programme (GCP) , an international consortium of agricultural research centres, and the Environmental Genomics Working Group (EGWG) have adapted customized versions of maxd to capture extended MIAME-based annotations of their microarray experiments (MIAME/plant, MIAME/Env) (Zimmermann et al., 2006; Morrison et al., 2006) that reflect the specificity of their respective biological topic of interest.

Maxd is a comprehensive free software environment that features three main components: maxdLoad2, maxdView and maxdBrowse. MaxdLoad2 sits at the core of maxd and features a friendly interface to an underlying relational database that allows data input, query searches and data editing. There are a handful of attractive features to maxdLoad2: first, the ability to handle formatted annotations of microarray experiments; most notably, in the form of spreadsheets and the ability to generate structured summary reports of these annotations. Second, and most importantly, the ability to customise the

annotations by extending the MIAME standards with domain specific information. This is a winning feature that has certainly contributed to the popularity of maxd among domain focussed microarray research communities. As a complement to maxdLoad2, maxdBrowse features a comprehensive web-server platform for browsing the content of maxdLoad2 in a multi-layer fashion that reflects the specific needs of various types of users. MaxdView, on the other hand, is the component of maxd that deals with data analysis and visualisation and is modular in nature allowing straightforward incorporation of additional functionality.

3.1.2. Functional annotation data

An essential part of setting up a gene expression database is to capture the biological role of the genes by associating them with their functional annotations. Luckily, gene functional annotation is a task that has been widely explored in bioinformatics and many public resources exist nowadays that offer functional annotations for complete genomes. Thus, incorporating functional data into locally established databases is often an operation that involves no more than mirroring gene functional associations from source databases, by establishing links between internal gene identifiers from the local database and the source annotation database. In the following, we give a

brief overview of the essence of gene functional annotation from a bioinformatics perspective.

3.1.2.1. Modelling of biological functions

Before genes may be associated with functional terms, a standardised vocabulary needs to be defined to formalise those terms that cover the range of known biological functions. This is a non-trivial task involving conceptualisation of domain knowledge and this has been appropriately resolved with the use of ontologies. The gene ontology (GO) initiative (The Gene Ontology Consortium, 2008) currently hosts the largest and most comprehensive set of gene functional concepts. Importantly, GO recognises three distinct components of a gene/protein function that are independent of each other: the molecular activity carried out by the protein, the broad biological process in which the protein performs this molecular activity and finally the site of action within the cell. Thus, a gene may be associated with one or more instances of biological processes, one or more instances of molecular functions and one or more instances of cellular locations. Importantly, these different aspects of function are independent of each other; thus, as an example, the receptor binding molecular activity mediates many

biological processes including signal transduction as well as the translocation of viruses into host cells.

Along these lines, a separate ontology was developed for each of these functional themes, as part of the GO database that categorises instances of the theme. Importantly, GO features a top-down categorisation approach that provides a step-wise specification of a concept semantics. Importantly, such a framework exposes similarities between concepts by revealing broad common functional themes that capture their semantics. An example GO subgraph is shown in Figure 3.1.1.

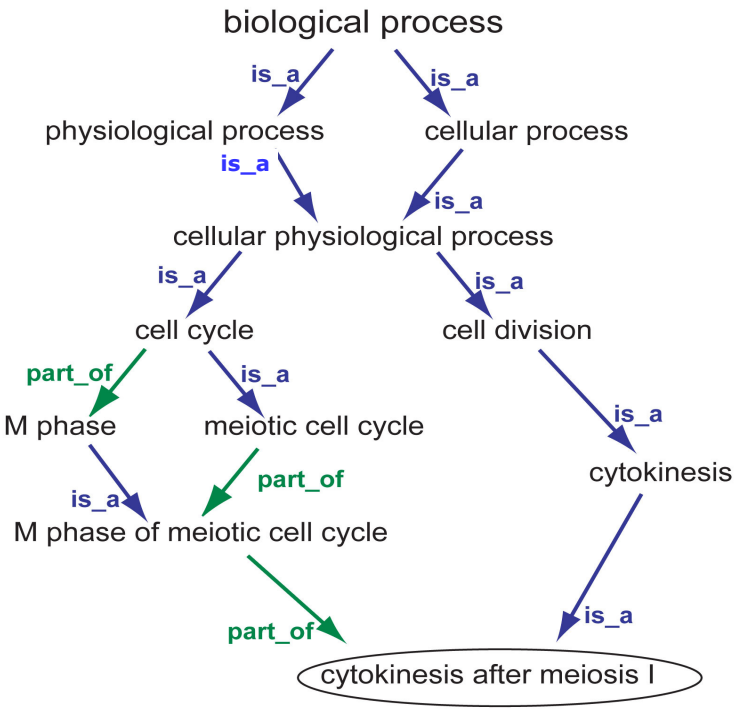


Figure 3.1.1. A model GO subgraph illustrating GO terms and relationships between them.

Thus, the ‘cytokinesis’ biological process is a type of ‘cell division’ process, which is in turn a type of ‘cellular physiological process’; whilst the biological processes ‘cytokinesis’ and ‘meiotic cell cycle’ have in common the fact they are both instances of ‘cellular physiological process’ (Fig 3.1.1). Importantly, beside the ‘is_a’ relationship that indicates that concepts provide an abstraction of the semantics of other concepts from lower levels in the hierarchy, the ‘part_of’ relationship is used by GO to reflect the fact that many low level biological processes may come together to give rise to higher level, more complex, biological systems within the cell (Fig 3.1.1).

Importantly, GO uses a directed acyclic graph structure to organise the set of terms from each ontology; with the main difference to tree structures being the possibility of having more than one parent term for a given child term. An important rule that applies to GO is the true path rule, stating that the meaning of a term implies the semantics of all its ancestor terms. This has important consequences at the level of gene annotations in that for any given gene-term association, every parent of the term is also a valid annotation for the gene. The GO vocabularies are constantly revised, with new terms and relationships being added by curators in consultation with biological experts.

Various other forms of structured vocabularies for biological functions exist in the public domain. Examples are the *Functional Catalogue (FunCat)* (Surmeli et al., 2008), which provides a tree-like categorisation of functions at varying levels of specificity and the *Kyoto Encyclopedia of Genes and Genomes (KEGG)* collection of biological pathways annotated with higher-order functions from *KEGG BRITE* supplement classification of biological systems (Okuda et al., 2008). Also, the *Enzyme Classification (EC)* providing a hierarchical classification of enzymatic reactions that is also used for enzyme nomenclature. However, GO remains the most comprehensive resource of biological functions and the most widely used in biological research applications. This despite many limitations, notably, the separation between the three different ontologies that hinders the appreciation of the multi-level nature of biological functions; in addition to the lack of consistency while defining relations between terms.

3.1.2.2. Methods for deriving gene function

Formalised functional vocabularies provide the mechanism for associating genes with functional terms that best describe their functions. There are broadly two main approaches for gene function discovery: *experimental*, based on laboratory direct assays and the *inference-based* approach that relies on

educated prediction from knowledge of functions of related genes. The former has the quality of rigor but the drawback of being slow whilst the latter is known to produce quick information, which can be at the expense of accuracy.

The notion of functional conservation in evolution has been a major principal in gene function prediction in bioinformatics. With the emergence of fully sequenced genomes from eukaryotic organisms, it became apparent that gene sequences, structures and functions are shared between species. Such similarity in genetic characteristics between species is due to shared ancestry, commonly referred to as *homology*. Homology comes in two flavours: orthology and paralogy. Orthologous sequences are sequences originating from a speciation event, which is when a species diverges in evolution to give rise to two separate species. Paralogous sequences on the other hand, are the result of a gene-duplication event in the same organism. With paralogy, the additional copy of the gene may acquire new functional characteristics because the availability of the original copy implies no constraint for functional diversion; as such, paralogous sequences tend to be functionally less similar than orthologous sequences.

In the context of exploring homology for function prediction, important bioinformatics research has identified thresholds of sequence similarity above

which function is likely to be conserved. For instance, the study by Todd and Orengo (Todd et al., 2001) indicated that EC numbers, consisting of a numerical code that expresses the enzymatic reaction class of enzymes by the Enzyme Classification database, rarely vary at sequence identity above 40%. This was reiterated by the more recent study by Tian and Skolnick (Tian and Skolnick, 2003) suggesting that the first digits of the EC numbers, corresponding to higher-order classes of enzymes, may be reliably transferred at sequence similarity above 40%. Inheriting functional information using homology is currently considered the most efficient way for characterising protein function and has proven wrong the long time assumption that a protein function may only be predicted when its three-dimensional structure is fully characterised.

To assist with homology based functional prediction, many public resource databases have arisen to provide family based classification of biological sequences across genomes. For instance, *PANTH* (Thomas et al., 2003) uses curated family and subfamily classification to organise known protein sequences and derives HMM profiles from the functionally distinct subfamilies to identify novel homologues from newly sequenced genomes. The *SYSTERS* database (Meinel et al., 2005) uses a two-tier family/superfamily clustering approach to organise proteins from the Swiss-

Prot/TrEMBL database and derives a list of key functional attributes for each family. *HAMAP* (Lima et al., 2009) is another family based database that hosts sequences from microbial genomes. Importantly, HAMAP protein families are curated manually and propagation of functional annotations to uncharacterised homologues is supervised with high level of care from template sequences from the family whose functions have been characterised by experimental means.

The *BioMap* database (also known as the *CATH-Gene3D* family/function database, (Maibaum, 2004)), used in this work, features a multi level classification of protein sequences originating from a large number of fully sequenced genomes. At the top of the classification hierarchy are protein families that define sets of evolutionary related proteins. The latter are formed using the PFScape protocol (Lee et al., 2005) that exploits the TribeMCL clustering algorithm (Enright et al., 2003; Enright et al., 2003). Protein sequences from the same family are subsequently grouped into clusters of sequences with at least 30% sequence identity; the latter are in turn partitioned into even finer clusters featuring 35% or more sequence identity. More granular clustering at increasing levels of sequence identity follows to yield clusters of increasingly similar sequences (Fig 3.1.2). Importantly, at each level of sequence identity, the resulting clusters are given unique numbers and

at the end of the clustering process, each protein sequence is assigned a cluster number by concatenating the numbers of the clusters featuring the sequence from consecutive rounds of clustering (illustrated in Fig 3.1.2).

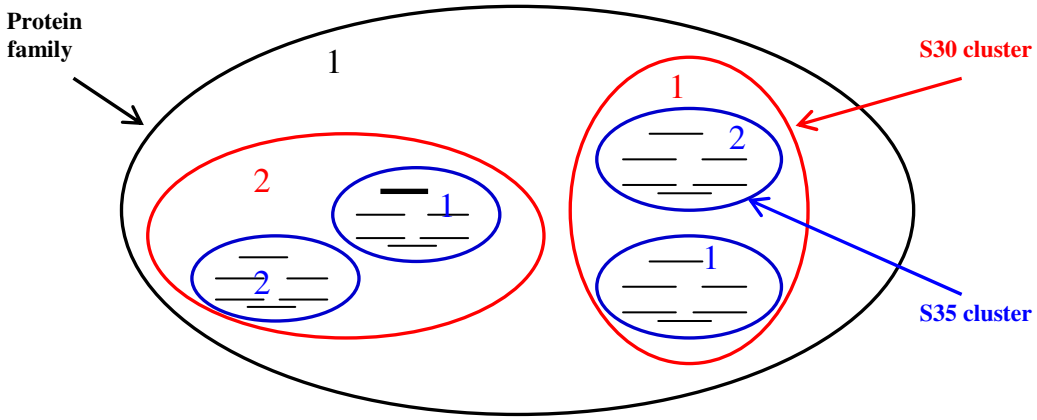


Figure 3.1.2. Diagram illustrating the nested homology based classification of sequences by BioMap. The outer circle in black delineates the protein family. Inner circles in red indicate the 30% sequence identity clusters whilst those in blue the nested 35% sequence identity clusters. The latter may then be divided into clusters of sequences featuring more than 40% sequence identity and likewise increasingly more granular clusters are formed at 50%, 60%, 70%, 80%, 90%, 95% and 100% sequence identity levels (not shown on the diagram). The numbering of clusters from each round of clustering is indicated. On the basis of the family->30%->35% classification illustrated in the diagram, the example sequence in bold may be assigned the cluster number 1.2.1.

With GO, the process of deriving functional information for genes is carried out by members of the GO consortium. These are research organisations that have committed to the sequencing and subsequent annotation of genomes from different organisms, such as the *Arabidopsis Information Resource (TAIR)* and

the *Mouse Genome Informatics (MGI)* database. Similarly, the *Gene Ontology Annotation (GOA)* initiative (Barrell et al., 2009) aims to annotate proteins from the *UniProt* database with GO terms thereby providing a comprehensive source of annotation of proteins from all species.

In order to capture the different ways in which functions of genes are identified by annotators from the GO consortium, GO provides a set of annotation evidence codes. The latter extend the broad experimental/inference-based classification of gene function discovery methods (discussed earlier) to account for the many practical details that arise during the process of gene functional annotation. For instance, functional information derived via a process of homology inference is classified differently by GO depending on whether such information was curated manually or generated purely via computational work.

The GO annotation evidence codes are fully described at http://www.geneontology.org/GO.evidence.shtml; but briefly, they fall into the following classes:

1. *Experimental*, involving direct experimental work.
2. *Computational*, where the information on function is derived on the basis of sequence or structural similarity but curated manually.

3. *From author statement*, usually from review type of articles where the evidence for the information is mentioned in the form of a reference to the original experimental work.
4. *Curatorial*, where the information was reasonably inferred by the curator, but for which no direct evidence is available.
5. *Electronic*, usually involving large-scale computational annotation of sequences and genomes featuring no manual curation.

3.1.3. Chapter aim

The main aim of this chapter is the setting up of a database of functionally integrated gene expression data from animal models of peripheral neuropathy. We refer to this database as the *LPD* standing for the 'London pain database' as from the LPC perspective, the primary objective of the database is the study of the pain aspect of neuropathy. The expression data in the LPD originate from microarray initiatives undertaken by the LPC, as well as published work. The functional annotations of the genes from the expression datasets were obtained by exploring various annotation pipelines, notably Biomap the family oriented functional database.

3. A database of gene expression data from animal models of peripheral neuropathy

3.1. Introduction

At the time this work was underway, only few free software microarray data storage platforms were available, namely BASE. Unfortunately, BASE did not offer much support for cross platform integration of microarray datasets and meta-analysis of lists of differentially expressed genes. Thus, similar to many biologically specialised databases such as SAMMD and GAN, the LPD was designed to use a simple, other than MIAME, model to store the data and annotations of experiments whilst featuring a greater focus on providing the type of exploratory tools that will allow efficient integration of different microarray datasets. The advantages and disadvantages of such an in-house system as oppose to adaptation of free software solutions are discussed in the conclusion section at the end of this chapter.

3.2. Data types and data acquisition

The LPD hosts three main types of data: microarray expression data, gene annotation and family data as well as biological domain data. These different types of data and methods for their acquisition are described below in detail:

- *Microarray expression data:* The primary source of expression data captured in the LPD consists of the set of microarray experiments run by the LPC. Additional microarray datasets were obtained from the following published studies (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002; Yang et al., 2004). These studies were selected on the basis of biological relevance to the animal models of neuropathy investigated by LPC microarray work. Datasets from these studies were not available for electronic download; instead, information on differentially expressed genes was identified in the corresponding articles and manually entered in the LPD. Table 3.2.1 gives a survey of all microarray expression studies captured in the LPD, listing the name of the main experimenter and the animal model investigated.

Microarray Study	Experimental model investigated
Tony Dickenson (unpublished)	*Murine model of bone cancer pain* (Schwei et al., 1999): Following injection of tumor cells in the femur bone, animals tend to guard the affected limb showing clear evidence of pain-related behaviour. The progression of bone destruction and consequent increase in pain is accompanied by clear neurochemical change in the spinal cord.
(Maratou et al., 2009; Valder et al., 2003; Wang et al., 2002; Yang et al., 2004)	*Selective nerve ligation*, also known as SNL/CHUNG (Kim and Chung, 1992): As illustrated in Figure 1.2.1, SNL involves unilaterally tying the L5 and L6 spinal segments of the sciatic nerve proximal to their DRGs.
Andrew Rice (unpublished)	*Rat model of Zoster-associated pain, or VZV* (Kim and Chung, 1992): Involves the subcutaneous injection of VZV-infected fibroblasts into the left hind foot. The virus then undergoes retrograde axonal transport along the sciatic nerve to establish a latent infection in the corresponding DRG.
(Maratou et al., 2009)	*HIV model of neuropathy* (Maratou et al., 2009): This model consists of injecting the HIV coat protein gp120 into the paw of the animal. Since the antiviral drug ddC is known to contribute to neuropathy in human subjects, the drug is also injected in the paw to fully mimic the neuro-pathology of HIV infection.
Maria Fitzgerald (unpublished)	*Spared Nerve Ligation (SNI)* (Decosterd and Woolf, 2000): As shown in Figure 1.2.1, the model consists of transection of common perineal and tibial branches of the sciatic nerve.
Geranton & Hunt (unpublished)	*Arthritis CFA-induced model* (Geranton et Hunt, unpublished): A model of inflammatory pain achieved by injection of the inflammatory substance CFA in the ankle joint.
John Wood (unpublished)	Nav1.7 knockout: featuring the knock-out of the sodium channel Nav 1.7.
John Wood (unpublished)	Nav1.8 knockout: featuring the knock-out of the sodium channel Nav 1.8.
John Wood (unpublished)	ASIC1 knockout: featuring the knock-out of the acid channel ASIC1.
(Rabert et al., 2004)	Brachial plexus spinal root avulsion: Avulsed DRG removed by surgery from patients suffering from brachial plexus lesions.
(Costigan et al., 2002; Xiao et al., 2002)	Involving sciatic nerve transection (Fig 1.2.1)

Table 3.2.1. Source microarray studies of the expression datasets stored in the LPD. Microarray studies by the LPC are indicated in red whilst those taken from literature are indicated in black**.** Because the LPC main research interest is the study of pain, expression datasets featuring animal models with phenotypes indicative of pain of non-neuropathic origin were also included in the LPD, such as models of inflammatory and cancer pain.

- *Annotation and family data:* Functional annotations of the genes in the LPD were derived from within a family based setting using BioMap, the Oracle implemented data warehouse. Additional functional annotations were obtained from Ensembl via the *EnsMart* (Kasprzyk et al., 2004) web facility and the array manufacturer online annotation centre *NetAffx* (Liu et al., 2003). Functional annotations from all these different sources consisted of GO and KEGG pathway information.

- *Domain related data*: The final type of data in the LPD consists of biological knowledge in relation to neuropathy and pain, mainly descriptions of animal models used to generate hosted expression data. Such information is crucial not only for documenting the type of pathology being investigated in individual experiments but also to assure that comparisons of separate microarray experiments are biologically sensible. Formalised descriptions of animal models of neuropathy and pain were obtained from the literature (Eaton, 2003; Wang and Wang, 2003) and via consultation with experimentalists from the LPC. In the future, the LPD may evolve to integrate additional neuropathy related data such as clinical data.

3.3. Data structure: the LPD schema

The different types of data in the LPD were used to derive a logical conceptual data model, which was implemented in a relational setting using the *MySQL* platform. Thus, major entities in the data were identified and captured in tabular structures that include a specification of the entity properties and attributes. Relationships between the entities were also modelled that indicate how instances from different entities relate to each other. The diagram on Figure 3.3.1 shows the LPD data structure. Importantly, tables from each data type (consisting of expression data, annotation and domain data) are shown in different colours. The LPD data model including entities and their relationships is discussed in full in the following paragraphs.

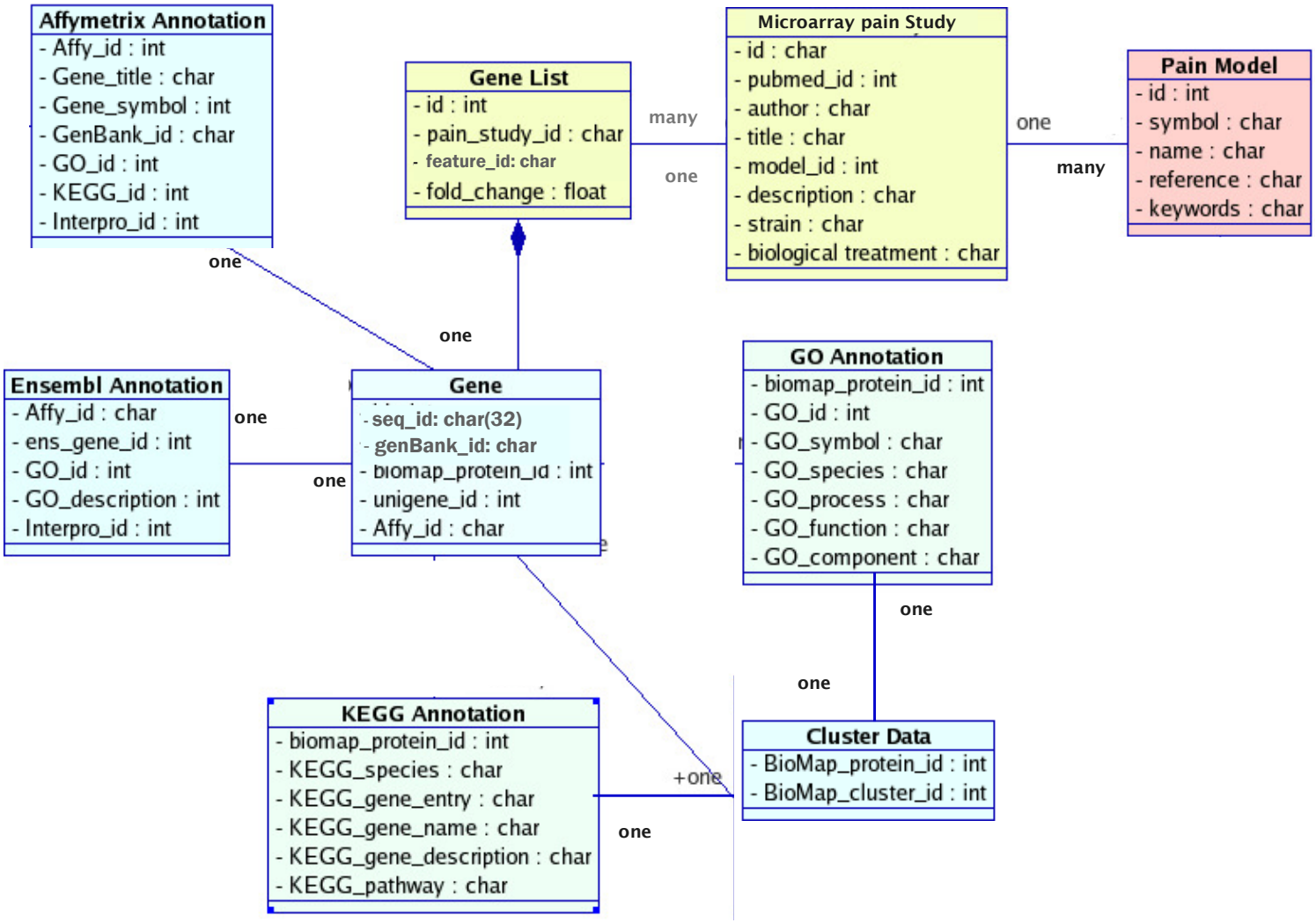


Figure 3.3.1. The LPD data structure. Green tables correspond to gene expression data including experiment annotations, blue tables store gene annotations while the pink table captures domain information consisting of definitions of experimental models of neuropathy and pain.

3.3.1. Domain data tables

Beginning with the biological domain data, the LPD schema features one unique entity: the *Pain Model* or perhaps more appropriately the *Experimental model* entity. Owing to variations in the experimental procedures used to realise these animal models, only basic but common features of the models were taken to define the attributes of the representative class *Pain Model.* These consisted of the model common names, the original study that first developed the model and keywords capturing the pathological and phenotypic characteristics of the model. The latter may be more formally expressed using the Mammalian phenotype ontology (Smith et al., 2005), part of the Open Biomedical Ontologies (OBO) .

3.3.2. Gene expression data tables

As for the gene expression data, two main entity classes were recognised: the *Microarray Pain Study* class and the *Gene List* class. The former class captures summaries of microarray experiments, including information on the experimenter and various useful experimental details such as the animal model investigated (hence the link to the *Pain Model* entity), species/strain information, array platform and handling of the RNA material. The *Gene List*

class on the other hand, captures the gene expression data outcome of the microarray study; in particular, genes found most differentially expressed and their fold changes. Importantly, with some array platforms such as Affymetrix, the expression measurement is identified with a probeset identifier instead of the gene identifier and many probesets may map to the same gene. Consequently, the *Gene List* entity features a generic *feature_identifier* attribute, which can take the value of an Affymetrix probeset identifier or a gene identifier (usually GenBank or UniGene).

3.3.3. Functional annotation data tables

A number of tables exist in the LPD that hold functional annotations of the array genes, corresponding to different sources of annotation. These include the *Affymetrix Annotation* table, the *Ensembl Annotation* table and the *GO/KEGG Annotation* tables derived from BioMap. Logically, functional annotations should be modelled as a single entity since the source of annotation is merely an attribute of the annotation. However, owing to the differences in the way functional information is encoded in each source database and also for ease of maintenance, it was decided to keep annotations from the different sources in separate tables. For instance, with Ensembl and unlike the rest of the source databases, the GO annotation terms from all three

ontologies: functional process, molecular function and cellular component are given together in a single string without indication of their ontology type. It is important to note that the reason why annotations from different sources were pulled together in the LPD is because it was noticed that they complemented each other and for many array genes, functional annotations were only available from one source and not the rest.

The BioMap functional annotation data have the special feature of being linked to protein family classification data and are arranged in a special data table structure that requires more explanation. One key table is *Cluster Data.* This table was mirrored from the BioMap database and hosts information on family classification of sequences by linking all BioMap proteins to their corresponding sequence cluster numbers. Each protein entry in *Cluster Data* is functionally annotated via association, where possible, with one or more entries from the *GO Annotation* and *KEGG Annotation* tables.

As an interface between the gene expression data and the functional annotation data, an additional table capturing the entity *Gene* was created. Importantly, the latter defines important information for each array feature, notably sequence and identifier attributes of the corresponding gene. This has the important consequence of revealing probeset association to identical genes

with Affymetrix based expression datasets (more details will follow in the next section).

Importantly, the association to BioMap protein identifiers in the *Gene* table allows each array feature (gene) to be linked to the corresponding BioMap cluster number, by reference to the *Cluster Data* table. Knowing the BioMap cluster number for a given array feature (gene) allows functional information to be retrieved from homologous BioMap proteins at a desired level of sequence identity. For instance, if an array feature/gene associated BioMap cluster number is 1.2.1.3.4.1.5.3.6.1.1, then all BioMap proteins with BioMap cluster numbers beginning with the same first four digits 1.2.1.3 in *Cluster Data,* are in the same S40 cluster; that is sharing at least 40% sequence with the array gene protein. Functional annotations may then be inherited from these homologs from the GO and KEGG annotation tables (for more details on BioMap cluster numbers, refer to Figure 3.1.2).

3.4. Data integration

The essence of the LPD is to store expression values of the genes as well as their functional annotations. However, because gene expression data were derived from a number of different sources (both in-house and from literature) utilising varying array platforms and similarly gene annotation data were obtained from various annotation databases, it was possible for the same gene to be referred to by different identifiers in the different datasets. Clearly, data integration was necessary to eliminate redundancy and promote data unity. It is worth noting that such mapping between identical entries from the various datasets is exclusively captured in the *Gene* table, as will be explained later.

In the following, the methodology used for integrating the different datasets in the LPD is summarised. We begin by describing our strategy for integrating expression data from the varying sources and proceed by examining the manner by which gene expression data were integrated with annotation data.

3.4.1. Integrating gene expression data

As mentioned before, the LPD expression datasets originated from two main sources: the in-house datasets derived from LPC microarray experiments were

based on Affymetrix arrays and feature Affymetrix probeset identifiers as primary identifiers. On the other hand, published expression data are typically identified by GenBank and UniGene identifiers. Luckily, the Affymetrix array manufacturer provides mappings of Affymetrix probeset identifiers to all common gene identifiers used by popular repositories of biological data including GenBank and UniGene. However, because UniGene provides an automated partitioning of GenBank sequences into non-redundant sets of gene-oriented clusters, it was deemed more appropriate to map all expression data to UniGene identifiers. Thus, entries from the published expression datasets that were only named with their GenBank identifiers were mapped to UniGene identifiers using the NCBI web service *Elink* (Baxevanis, 2008). *Elink* allows cross-linking of identifiers from various NCBI databases and in our case, it was used to map the GenBank identifiers to UniGene identifiers.

However, since not all GenBank identifiers from the LPD expression datasets were successfully mapped to UniGene identifiers, it was necessary to perform sequence comparison to identify additional identical entries between the various expression datasets. Thus, nucleotide sequences were obtained by querying the NCBI web service *Efetch* (Baxevanis, 2008) with the GenBank identifiers from the expression datasets. *Efetch* allows linking of various gene identifiers (including GenBank identifiers) with appropriate NCBI database

entries and the retrieval of useful information from the selected records including nucleotide and peptide sequences. Sequences from the various expression datasets showing 100% sequence identity revealed an additional set of identical entries between datasets, amounting to 10% of the overall number of gene entries in the LPD.

Importantly, such a sequence comparison based approach may fail when the sequences are partial i.e. not spanning the whole length of the gene, such as ESTs. The problem of EST mapping to genes is non-trivial, but luckily the many EST sequences submitted to GenBank are regularly classified into gene-centric clusters via robust EST annotation protocols by UniGene. Thus, our original mapping to UniGene identifiers may have been complementary to sequence comparison searches since the former is more robust at dealing with ESTs and partial matches than the latter. In the LPD and to keep track of equivalent entries between the various expression datasets, UniGene identifiers as well as nucleotide sequences MD5 digests (unique 32 character strings computed from the sequences) were captured in the *Gene* table in columns *unigene_id* and *seq_id* respectively (Fig 3.3.1). Table 3.4.1 shows examples where sequence and UniGene identifiers were instrumental to recognising identical entries from different expression datasets whilst Figure

3.4.1 shows a flowchart summarising the steps performed for integrating these datasets.

Microarray Study	GenBank identifier	Sequence MD5	UniGene ID
(Wang et al., 2002)	K02248	11eaacf2431bafb6ec80cec311d77b5f	Not known
(Xiao et al., 2002)	NM_012659	11eaacf2431bafb6ec80cec311d77b5f	395919
(Costigan et al., 2002)	X53054	Ytgrf5643ijnbf62as1qqkl90867fgvd	395454
(Valder et al., 2003)	AF084934	00lki87yhbfr5ffcdsnh8777maa520	395454

Table 3.4.1. Identical gene entries from different published expression datasets stored in the LPD. Genbank entries K02248 and NM_012659 were mapped to the same gene due to identical sequences (shown in blue) (sequences are denoted by unique 32 character long strings referred to as MD5), whilst Genbank entries X53054 and AF084934 were found biologically equivalent due to identical UniGene identifiers (shown in red).

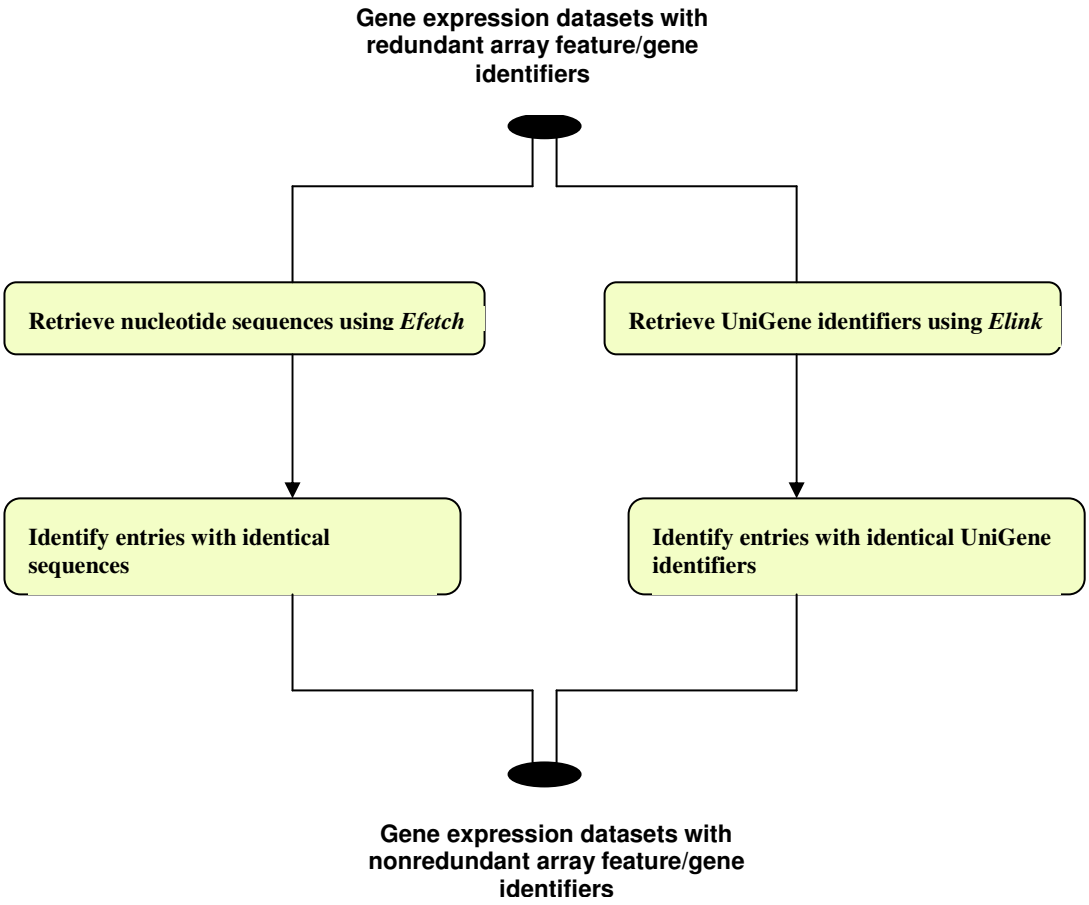


Figure 3.4.1. Flowchart showing the combined methodology used for identifying equivalent biological entries across the different LPD expression datasets. NCBI web services, *Elink* and *Efetch,* were used to retrieve UniGene identifiers and nucleotide sequences for array features using their GenBank identifiers. Equivalent biological entries across the different datasets were identified by means of identical UniGene identifiers and/or identical sequences. The two strategies complemented each other: UniGene mapping allows entries featuring partial sequences of the same gene to be identified while sequence matches are more appropriate when UniGene identifiers are unknown.

3.4.2. Integrating expression data with functional annotation data

As previously mentioned, the functional annotations by the Affymetrix array manufacturer and Ensembl stored in the LPD were originally tailored for

Affymetrix arrays and hence needed no further integration with the Affymetrix based expression datasets in the LPD. However, one important aim of the current work was to derive functional annotations for the genes from the various expression datasets by exploiting the BioMap family oriented annotation framework. Using BioMap, additional functional information for uncharacterised genes may be gained from other functionally characterised homologs. This was particularly important as the average functional coverage for the arrays, achieved by either annotation source (Affymetrix/Ensembl), was rather limited. Furthermore, functional information derived from BioMap may be assessed by considering the extent of functional variation within individual protein families. Finally, exploiting BioMap provided an opportunity to annotate the LPD expression datasets originating from literature, which were not based on Affymetrix arrays and needed to be explicitly annotated.

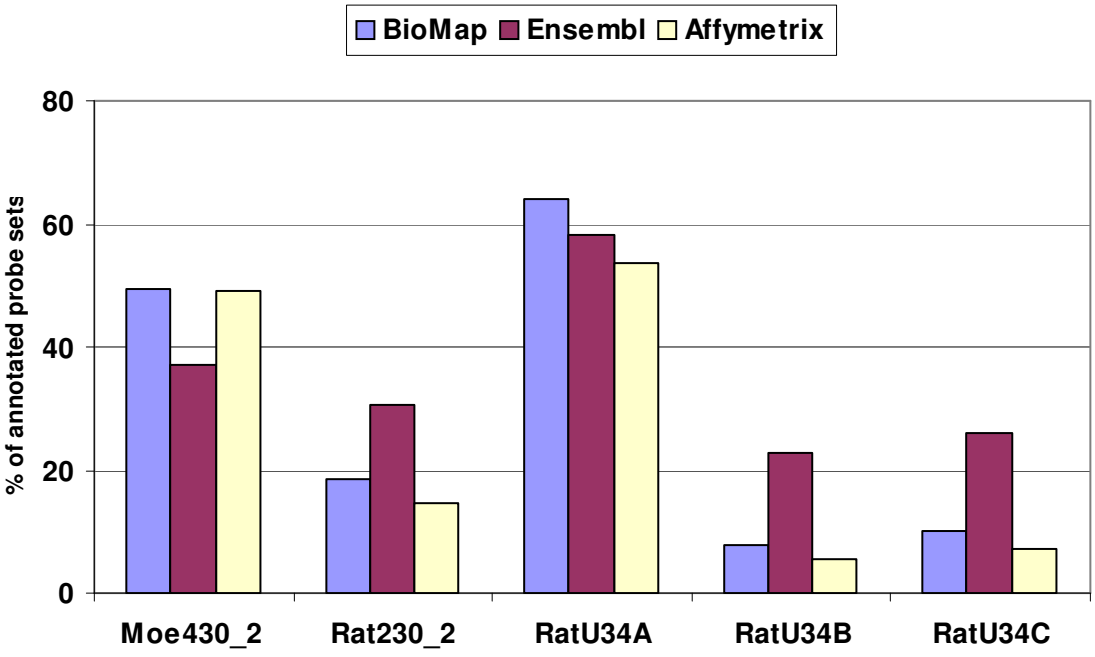
Initially, the protein sequences from LPD array features/genes were obtained by querying the NCBI *Efetch* web service with the corresponding GenBank identifiers. To check whether these protein sequences existed in BioMap and hence already classified in the appropriate BioMap sequence clusters, their MD5 digests were matched against BioMap protein identifiers based similarly on MD5 digests of corresponding sequences. Where no match was found, the

BioMap protocol for assigning new sequences to existing clusters was used. Finally, the updated *Cluster Data* table from BioMap containing mappings of all BioMap proteins (including LPD array protein sequences) to BioMap cluster numbers was mirrored in the LPD.

To assess the overall efficacy of the BioMap functional annotation of genes performed in this work, we compared the extent of functional coverage achieved with various Affymetrix arrays by BioMap, Ensembl and the Affymetrix array manufacturer. It is worth noting that with BioMap, functional information was inherited from related BioMap sequences at a sequence identity level greater or equal to 40%; that is functionally characterised homologs from S40 clusters.

The results are shown on Figure 3.4.2. Rather disappointingly, the BioMap based annotation seems to be only slightly better than that by the array manufacturer. Moreover, the Ensembl annotation appears to be more comprehensive for certain arrays, mainly the Rat230_2, RatU34B and the RatU34C. The explanation for this lies in the fact that these arrays feature a high percentage of EST sequences, meaning that the probesets in these arrays were mostly derived from short EST sequences instead of full-length genes (Fig 3.4.2). This is rather problematic with the BioMap annotation framework

as EST sequences are usually of unknown gene origin and it is hence difficult to obtain protein sequences for them that may be searched against BioMap protein sequences. By contrast, the annotation strategy used by Ensembl is based on nucleotide instead of protein sequence comparison, whereby probe sequences (including those derived from ESTs) may be mapped to genomic cDNA sequences from the appropriate organism according to well-defined rules.



Array EST content	47%	82%	11%	91%	91%

Figure 3.4.2. Percentage of functionally characterised probesets from various Affymetrix arrays by the different annotation approaches: BioMap, Ensembl and Affymetrix. Note that the percentages are relative to the total number of probesets on the arrays.

In Figure 3.4.3, the extent of functional annotation of Affymetrix arrays by BioMap at varying homology levels is shown. The analysis reveals that about 95% of functional assignments were derived from highly similar BioMap sequences with greater than 95% sequence identity, the majority of which featured exact matches. This implies that annotations inferred from homologous sequences at lower levels of sequence identity were not substantial; presumably, owing to the fact that the arrays subject to annotation in this work featured functionally well characterised genomes from the mouse and rat species. This seems to explain why the BioMap annotation pipeline did not perform better than the Ensembl and the array manufacturer annotations (Fig 3.4.2), as the former is based on exploiting homology to derive functional attributes for genes. However, despite the marginal gain in function assignment, the mappings between individual Affymetrix genes and the BioMap protein families achieved in this work can be used to inherit various other forms of useful information such as protein-protein interactions. Such data have been largely generated for yeast and are not directly available for the mouse and rat species except through family inheritance.

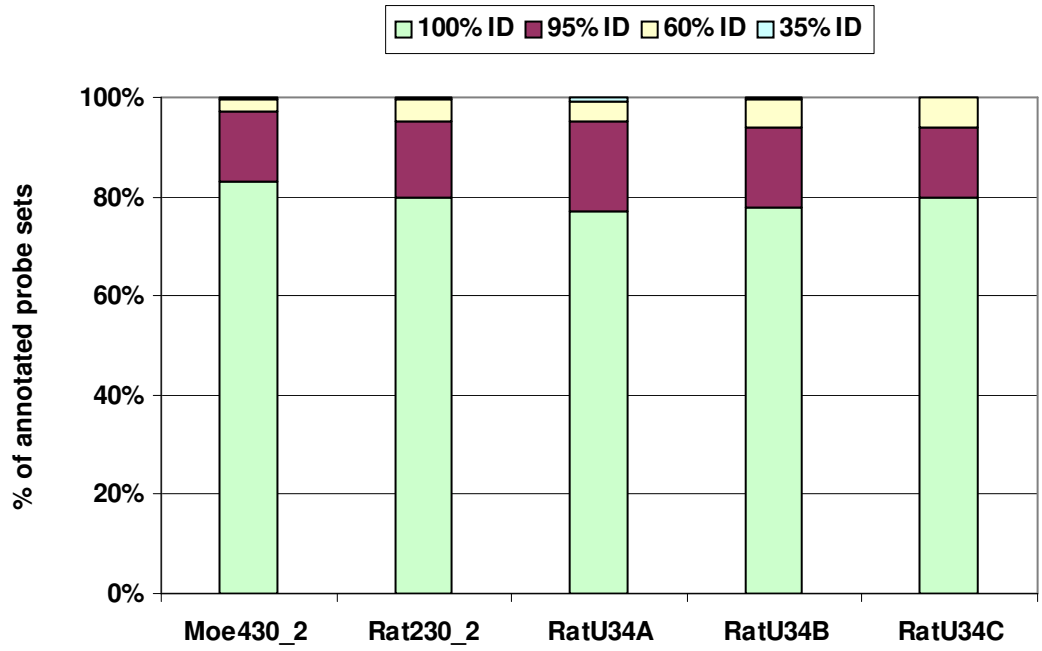
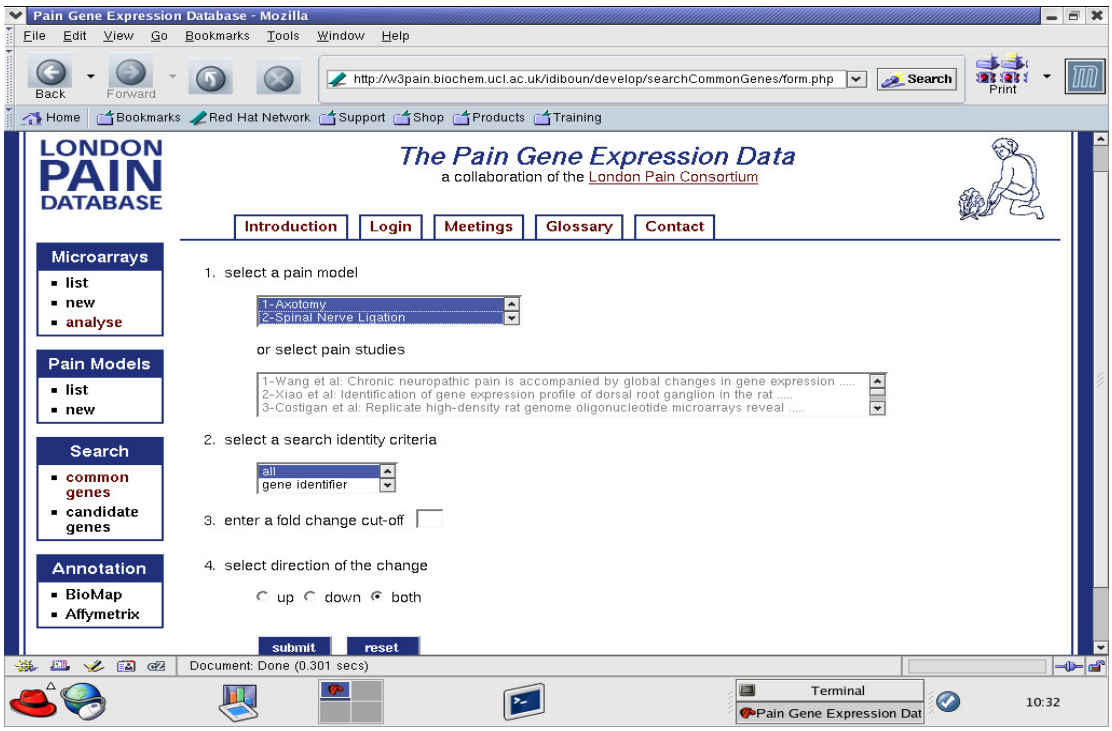


Figure 3.4.3. Number of annotated probesets at any given sequence similarity threshold expressed as a percentage from the total number of annotated probesets per array. Note that ID means sequence identity.

3.5. Data retrieval: the database web interface

A set of web pages were set up to allow a user-friendly interface to the LPD (Fig 3.5.1), which can be found at http://w3pain.biochem.ucl.ac.uk/idiboun/develop/search/searchCommonGenes/introduction.php. The web pages allow retrieval of various types of data from the LPD and were designed in accordance with a set of anticipated use cases specified by potential users from the LPC. One important use case was the possibility to retrieve genes showing a similar pattern of expression regulation across a number of microarray pain experiments. Figure 3.5.1 shows the form that allows this search to be conducted. Various drop-down menus and free-text fields are used to allow the user to specify the required search parameters. Among these, the pain model(s) of interest so that all microarray experiments featuring this model(s) are compared or alternatively, a subset of experiments that are of particular interest to the user. In addition, the desired fold change or significance value, allowing the most significant subset of the common genes to be filtered out. Importantly, the ability to identify common genes between different experiments is powered by the mapping between the heterogeneous gene identifiers from the different array platforms, discussed earlier.

A



B

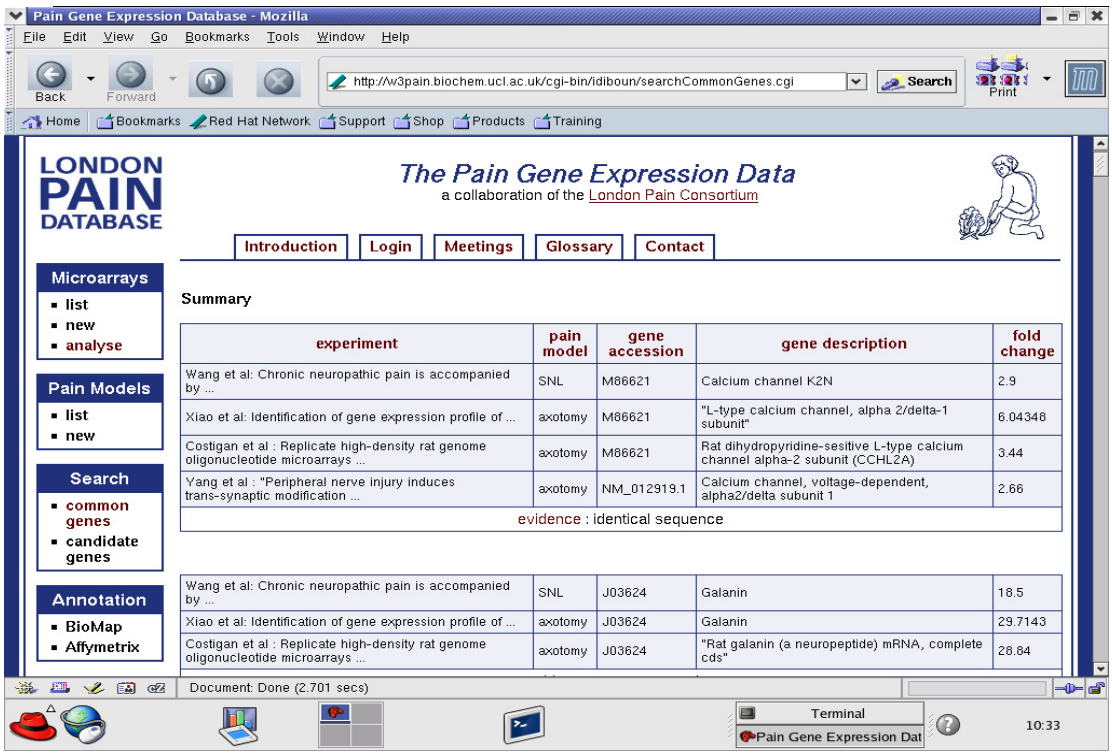


Figure 3.5.1. LPD meta-analysis web pages. Showing (A) the search form that allows genes commonly regulated in a number of selected expression studies or pain/neuropathy models to be retrieved, (B) the result from this search.

Figure 3.5.1-B shows the results from a search of commonly regulated genes across a number of randomly selected studies. The results for each gene are shown in a separate table. The rows of the table describe information about the gene as specified by each selected dataset; including the gene identifier, a textual description of the function of the gene and the fold change.

Further to searching for commonly regulated genes across varying microarray experiments, an important use case scenario consisted of the ability to browse functional information of lists of genes of interest; such as the ones obtained from cross-comparing microarray experiments. Figure 3.5.2 shows the LPD web interface that allows functional information for a given gene in a gene list to be broken down by homology to the protein annotation source as well as the type of annotation consisting of KEGG or GO.

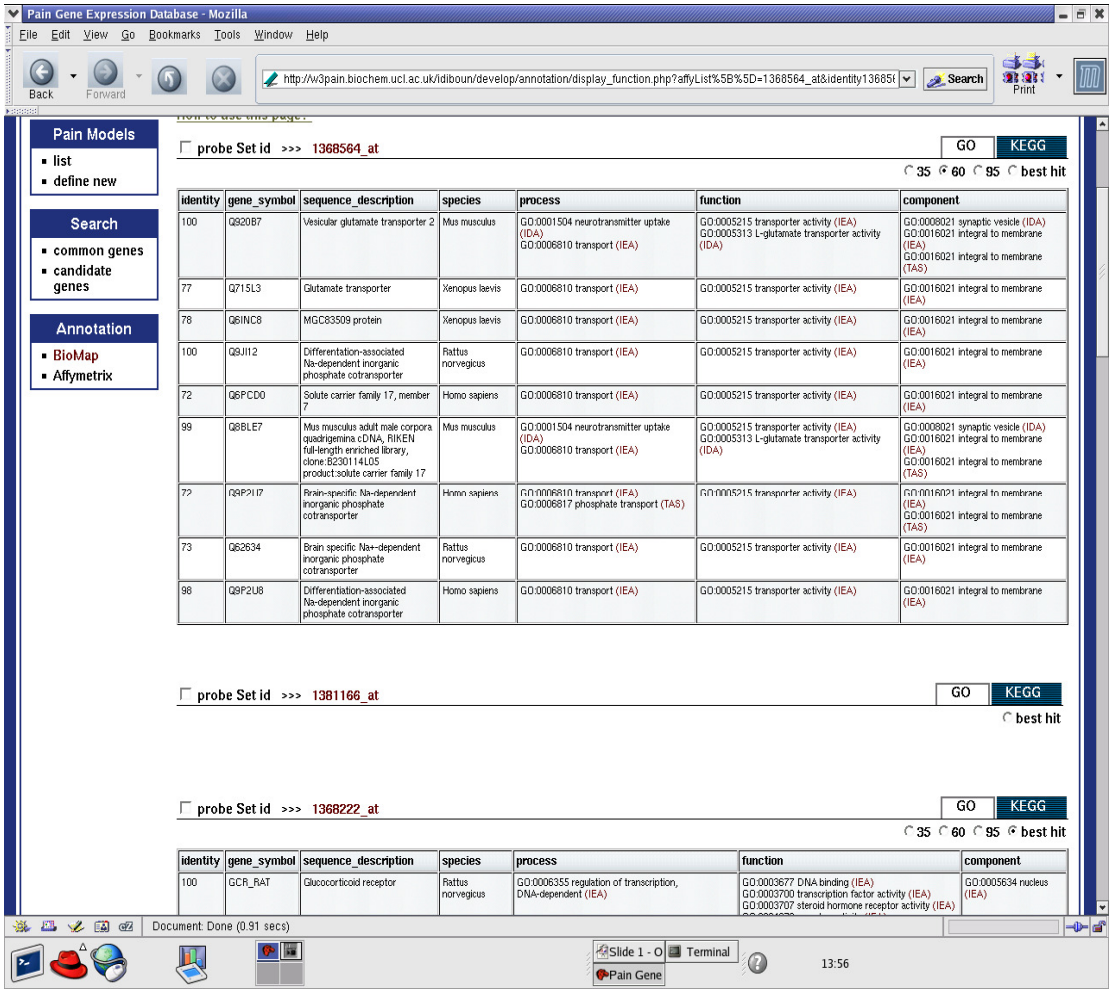


Figure 3.5.2. LPD functional annotation web pages. For each gene/probeset, GO and KEGG functional information are broken down by sequence identity to BioMap protein homologs serving as the source of annotation.

3.6. Conclusion

Microarray screening is characterised by a sheer genomic scale amount of data. Setting up a microarray database that is capable of handling such data efficiently is a non-trivial task and is further compounded by the need to project functional annotations on the gene expression data. The latter are heterogeneous in nature and often use different nomenclature schemes to refer to the same genes; which adds significantly to the complexity of the task involved. Furthermore, the need to capture information on the microarray experimental procedure implies an additional layer of data, leading to an even more complex underlying database schema.

The work presented in this chapter has certainly shed light on some of the overheads with the setting up of a microarray database. First, the integration of disparate gene expression and functional datasets proved rather challenging and is a process that requires considerable amount of time and resources to be maintained. Second, our choice to use a simplified data model than MIAME, although beneficial from the point of view of reducing the complexity of the data model, proved occasionally inefficient for failing to capture more complex microarray experimental designs such as time course experiments

and also for offering little assistance with constructing MIAME compliant descriptions of LPC microarray experiments.

In effect, many of these complex tasks such as the formalisation of descriptions of microarray experiments based on the MIAME standard and data integration are fairly non-specialised procedures that can be handled with generic software. This is because the MIAME data model was designed to be fairly general to accommodate all different microarray experimental designs that might be applied to study any biological phenomenon. Similarly, industry manufactured genomic-wide arrays, such as Affymetrix arrays, are becoming very popular among research communities undertaking microarray work. Because of their popularity, robust functional annotations for these arrays have already been assembled and are constantly revised by many independent sources; examples are the annotations by Ensembl and Bioconductor.

Microarray free software platforms are key to leveraging generic software solutions intended to serve routine handling of microarray data. For instance and as outlined in the introduction of this chapter, many provide user friendly tools for experimental data input in the MIAME format and deploy the logic of the MIAME model to support downstream statistical analysis of the data. Array probes functional annotations are provided built-in and additional

annotations may be easily incorporated, which also provides a mechanism for easy updates. Moreover, many free software microarray platforms provide generic tools for meta-analysis of the data; notably, cross comparisons of gene lists across different datasets of similar array platforms.

In effect, open source software systems constitute ideal microarray data management platforms. Thus, in addition to offering basic generic functionality for handling microarray data, these tools are often fully extendable; which allows them to harbour additional tools tailored to the specific needs of specialised research communities. In the future, the LPD will benefit from the open source software solution by adapting the maxd software (highlighted in the introduction section), for its numerous benefits. First, the fact that maxd accepts and assists in the development of customised MIAME data model is an attractive feature that, together with the use of ontologies, will help the LPD evolve into a pain knowledge-base repository. Second, maxd has a range of data browsing and analysis tools that would allow members of the LPD to conduct basic manipulations and searches of the data. Finally, maxd is configured to allow easy incorporation of additional functionality. This feature will be used to incorporate in-house analysis protocols as well as other free analysis software tools such as MatchMiner (Bussey et al., 2003). The latter is a tool that allows mapping of heterogeneous

gene identifiers, which is instrumental for cross-comparison of microarray results obtained with different array platforms.

CHAPTER IV: A GENE ONTOLOGY BASED MODEL OF THE FUNCTIONAL CHARACTERISTICS OF PERIPHERAL NERVE INJURY

4.1. Introduction

4.1.1. Aim of the chapter

The current chapter follows on from the previous chapter and aims to assemble a library of gene functions induced at the transcriptional level under the condition of peripheral neuropathy using the expression data from the LPD. This will be used in chapter VI as a gold standard to validate the efficacy of functional analysis methods applied to a spinal nerve transection (SNT) microarray dataset from LPC experimental work.

In addition to identifying this set of nerve injury related functions, one further aim to this chapter is to reveal the specific biological relevance of each function in the set to the biology of nerve injury. To substantiate this biological analysis and as an introduction to this chapter, the molecular

mechanisms underlying the physiological response to peripheral neuropathy are discussed. This is rather different to the material presented in the introduction chapter, which focussed primarily on the mechanisms of peripheral neuropathic pain. As for the GO functional paradigm, used extensively in this chapter, we feel that it has been adequately described in the introductory material of the previous chapter and needs no further explanation at this stage.

4.1.2. Pathophysiology of peripheral nerve injury: a molecular perspective

Peripheral neuropathy refers to the conditions that result when nerves that connect to the spinal cord from the rest of the body are damaged or diseased. Experimentally, the best studied form of peripheral neuropathy is that involving direct injury to the peripheral nerve as it is relatively easily mimicked in animal models than the more complex forms of peripheral neuropathies such as diabetic neuropathy. Despite the significant advances in understanding the molecular machinery deployed under the condition of nerve damage made with these models, the main challege remains to characterise these molecular changes in terms of cause and effect; in particular, in relation to the development of neuropathic pain. Examples of experimental models of

peripheral nerve injury were illustrated in the diagram in Figure 1.2.1 in the introduction chapter. This was used to give an overview of the anatomy of the peripheral nervous system which is essential for understanding the effect of nerve injury on DRG neurons in these models. This constitutes useful background for some of the material that follows.

In what follows, the pathophysiology and underlying molecular response to the most common form of experimentally induced nerve injury, involving nerve cut (axotomy), is discussed. Peripheral nerve axotomy is a significant occurrence to affected neurons that triggers a whole series of adaptive events, primarily aimed at extending the axon to regain contact with target territories (being the parts of the body innervated by the injured nerve). Maintaining contact with target territories is fundamental to the integrity of neurons since the latter depend on target-derived growth factors, also known as trophic factors, for normal function. Following injury, axonal regeneration leading to target reinnervation holds the key for neuronal survival, though this repair process is known to be limited and highly dependent on a number of factors such as the type and site of lesion. Moreover, reestablishment of connectivity with targets does not usually result in full recovery of lost sensory or motor functions as regrown axons may show poor target specificity and reinnervation adequacy.

To fully appreciate the reaction of neurons to axonal injury, it is important to consider the cascade of events first taking place at the site of the lesion. This is illustrated in Figure 4.1.1 (it is worth noting that most of the information presented in Figure 4.1.1 and discussed in the following paragraphs was taken from the following two reviews: (Navarro et al., 2007; Scholz and Woolf, 2007). Thus, upon injury, the axon is split into two parts: the part that loses contact with the cell body is called the '*distal part*' as opposed to the '*proximal part*' that stays attached (Fig 4.1.1). The axonal segment distal to the lesion begins to degenerate concurrently with the disintegration of surrounding myelin sheaths. This degenerative process results in the formation of debris that attracts the early immune cells, mainly local macrophages, causing Schwann cells to become reactive to injury. Active Schwann cells release cytokines such as leukemia inhibitory factor (LIF) and interleukin (IL)-1 (Tofaris et al., 2002) that further attract macrophages capable of phagocyting myelin and axonal debris. Cytokines are subsequently produced by the activated macrophages. The events that lead to the destruction of the distal stump are known as the *Wallarian degeneration* (Fig 4.1.1).

More important are the events taking place at the proximal end of the injured axon. Since the proximal stump remains attached to the cell body, it serves as a communication bridge between the site of injury and the cell body allowing

injury signals to be transduced to the inside of the cell, which causes the cell to respond to injury. Overall, the response may have one of two outcomes: cell growth and survival or cell death. There is a fine balance between the two opposing effects and much less is known about the pathway mechanisms contributing to neuronal death following injury, probably due to greater research interest in identifying growth promoting molecules. What is known though is that the same pathway mechanism could lead to either outcomes depending on the timing of the individual reactions and the pattern of cross-talking between the pathways.

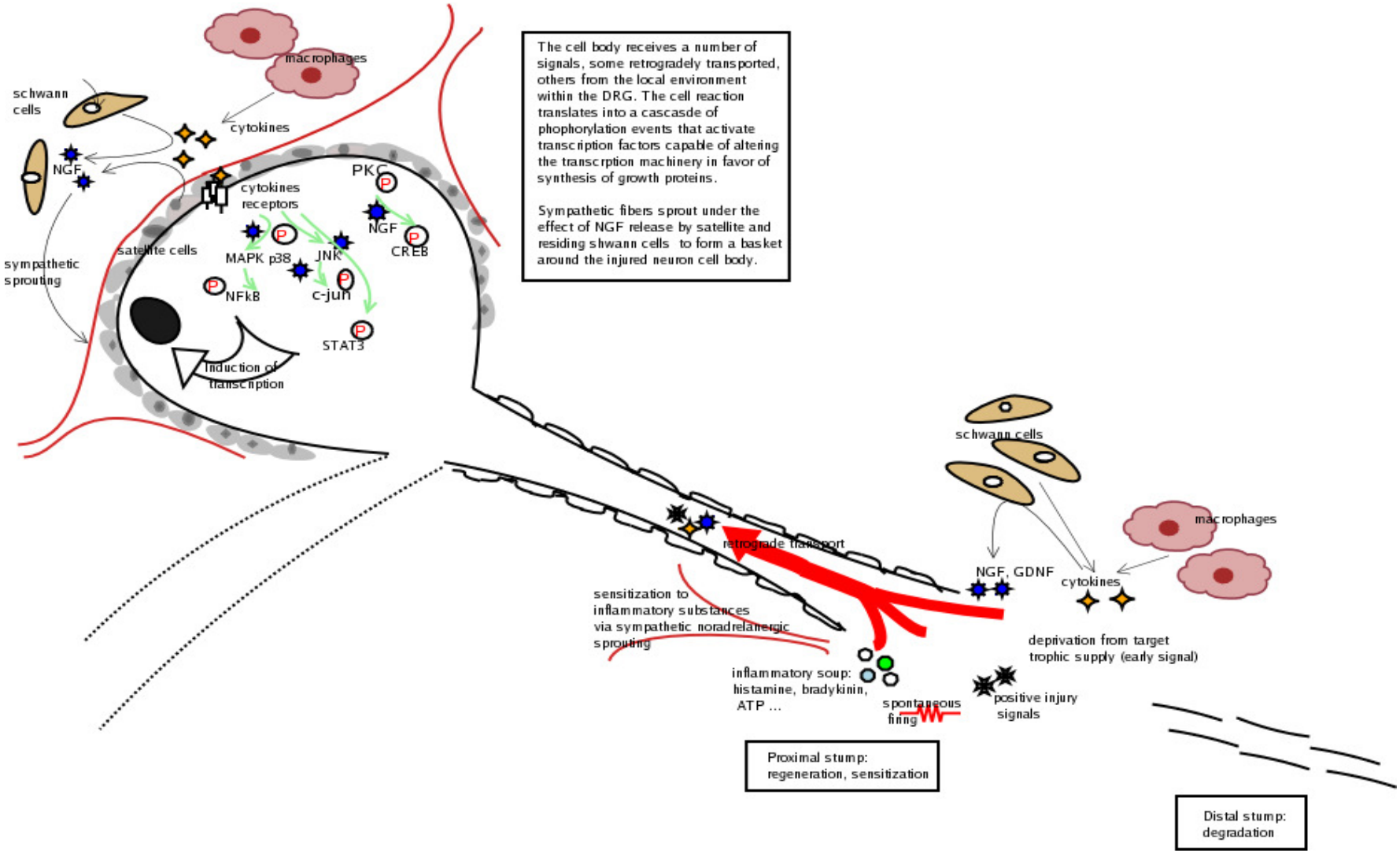


Figure 4.1.1. Schematic diagram showing the events that take place following peripheral nerve injury both at the lesion site and distal within the DRG where the injured nerve cell bodies reside. Note that the dotted lines represent the second axonal process that projects to the spinal cord, the latter is included in the figure for the sake of completion. The figure was based on the information in (Navarro et al., 2007 & Scholz and Woolf, 2007).

The first signal reaching the cell body of injured neurons is a burst of action potential resulting from a rapid depolarisation that occurs immediately after the axon is exposed to the extracellular medium following rupture of its axoplasmic membrane. Additional signals follow consisting of early deprivation from target trophic factors and later on partial compensation by retrograde transport of neurotrophins such as nerve growth factor (NGF), brain-derived neurotrophic factor (BDNF) and glial cell-line derived neurotrophic factor (GDNF) released by active Schwann cells at the site of injury. The cell body also comes under the influence of proinfammatory substances building up at the site of the lesion such as cytokines. Moreover, recent work has led to renewed interest in the axon endogenous proteins that undergo posttranslational shifts following injury, known as '*positive injury signals*', and their potential role in conveying the nociceptive message to the cell body. These signals originate from the site of injury and are transmitted to the cell body via the process of retrograde transport (Fig 4.1.1).

In addition to the lesion environment, injury to the axon is also signalled to the cell body by neighbouring non-neural cells within the DRG tissue. Following injury, macrophages invade the DRG and begin to release cytokines that in turn stimulate resident Schwann cells and glial satellite cells to produce neurotrophins. In addition to their effect on sensory neurons, these locally

produced growth molecules are thought to play a prominent role in stimulating sprouting of sympathetic fibres within the DRG into basket-like structures that surround neurons (Ramer et al., 1998) (Fig 4.1.1). Sympathetic input is one factor in establishing nociceptive sensitisation and neuropathic pain.

Cellular transduction of signals, originating from both the DRG local environment as well as the site of the lesion, involves the activation of many signalling pathway genes. For instance, recruitment of TRAF receptors by the proinflammatory cytokine TNF-α activates MAP kinases JNK and p38 while protein kinase A and B (PKA, PKC) may potentially be activated by the early influx of calcium upon injury. The downstream events consist of activation of potent transcription factors. Thus, taking the example of cytokine induced JNK, we find it associated with the expression and phosphorylation of c-Jun, a transcription factor with wide functionality following nerve injury. Active c-Jun has been implicated in nerve cell growth and survival; it was also associated with neuronal death (Elmquist et al., 1997) in conjunction with other key growth regulators following axonal injury. In addition, it appears to regulate the expression of a variety of neurotransmitters such as VIP and NPY (Son et al., 2007) as well as substance P and CGRP.

Similarly, phosphorylated p38 kinase activates the NFκβ transcription factor thought to promote neuronal growth (Aggarwal, 2003), though also implicated in neuronal death following transection of the optic nerve (Kikuchi et al., 2000). The significance of p38 phosphorylation lies furthermore in the resulting increase in the density of tetrodotoxin (TTX)-resistent sodium channel currents in nociceptors following injury (Jin and Gereau, 2006). STAT3 is another transcription factor that is thought to be induced by cytokines to promote neuron survival and regeneration (Lee et al., 2004).

Trophic factors play a prominent role in modulating intracellular signalling reactions in injured neurons. For instance, the early activation of survival inducing transcription factor ATF-3 is thought to be due to the early loss of target derived NGF and GDNF (Averill et al., 2004) whilst the phosphorylation of transcription factor CREB is dependent on the presence of compensatory neurotrophins (Miletic et al., 2004) released by active Schwann cells and DRG satellite cells following injury.

In surviving neurons, the functional outcome of promoting gene expression is the synthesis of molecules that support and stimulate axonal growth; among these are membrane lipids, adhesion molecules, growth associated proteins and cytoskeletal proteins that mediate the anterograde transport of growth

material to the growing end of the axon. On the other hand, neurotransmitter metabolism is given a lower priority, though with a marked plasticity following injury. Research has described a marked decrease in excitatory neurotransmitters content such as substance P and CGRP in small neurons (Butler et al., 1984) and an opposing increase in inhibitory neurotransmitters such as Galanin (Zhang et al., 1998). This, in addition to the upregulation of NPY, VIP and peptide histidine isoleucine, which are thought to play a role in communicating nociceptive injury signal to dorsal horn neurons, potentially contributing to neuropathic pain. Interestingly, the expression of excitatory neurotransmitters was found to be upregulated in large DRG neurons following injury suggesting a possible role in central sensitisation. Since large fibres are natural sensors of innocuous mechanical stimuli, it was speculated that they might be implicated in establishing mechanical allodynia (painful sensations caused by non-painful mechanical stimuli) following injury.

4.2. Methods

4.2.1. The gold standard term set

Published expression datasets from the LPD, featuring direct injury to the peripheral nerve, were selected in order to assemble a library of biological functions enriched at the transcriptional level during peripheral neuropathy. These included two SNL as well as two axotomy datasets (details about these animal models can be found in Figure 1.2.1) from the following published microarray studies: (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002). A fifth and final dataset was obtained from a literature survey conducted in the Costigan study of genes previously found to be regulated in animals with injured sciatic nerve by a variety of wet lab experimental techniques. Thus, the fifth dataset is not a microarray dataset, though, it was deemed worth including as it reported expression data that were validated experimentally.

Following dataset selection and by reference to the functional tables in the LPD, the most specific GO terms associated with each gene from the five chosen datasets were obtained. Since the ultimate goal in compiling this set of functional terms is to achieve a gold standard reference for validating

functional analysis of a nerve injury LPC microarray dataset (presented in chapter VI), we refer to this set as the *gold standard term set.*

Clearly, one important criterion for the gold standard terms is reliability. Thus, beyond ensuring the quality of individual datasets by only referring to published work, our approach of combining a number of expression datasets was meant to deal with the inherently noisy nature of microarray data. We thus look for commonalities between the different datasets following the logic that frequently occurring functional terms are likely to be the most believable.

To quantify the level of confidence associated with each term from the gold standard term set, we counted the number of studies featuring the term or its progeny as the term semantics are also implied by its descendents. We refer to this measure as the *term study occurrence measure*. We used functions from the GOstats package, an interface to GO from within the programming environment of Bioconductor, to identify the descendents of any given term from the gold standard term set.

4.2.2. Categorisation of the gold standard terms

In order to explore the biological significance of the gold standard terms, we sought to categorise them by the broad sense of their functions. This is particularly useful as the gold standard term set is relatively large. We used the *Gene Ontology Categoriser (GOC)* algorithm (Joslyn et al., 2004) to classify the gold standard terms into a handful of functional groups that are easier to study.

GOC comes as part of the software *POSOC* (Joslyn et al., 2004) designed to capture, manipulate and analyse the structures of graph based ontologies and is available at http://www.c3.lanl.gov/posoc/. The GOC algorithm is meant to provide a solution to the problem of categorising ontology terms: thus, given a set of terms of interest, what broad terms best summarise them in the ontology? In GO, parent terms are intrinsically an abstraction of the semantics of their children. As such, GOC considers all parents to the terms of interest as potential categorisation points. Among the many possible parents, the selection is made on the basis of the desired balance between coverage and specificity. Thus, taking the example of the model ontology graph shown on Figure 4.2.1, we find that the query terms (shown in green) ‘d’, ‘e’, ‘j’, ‘k’ and ‘l’ are all children of term ‘A’; as such, category ‘A’ shows the best level of

coverage. However, we may decide that 'A' is associated with a far too general meaning and decide to choose the more specialised term 'C' instead, despite the fact that the new category fails to include the query term 'd'.

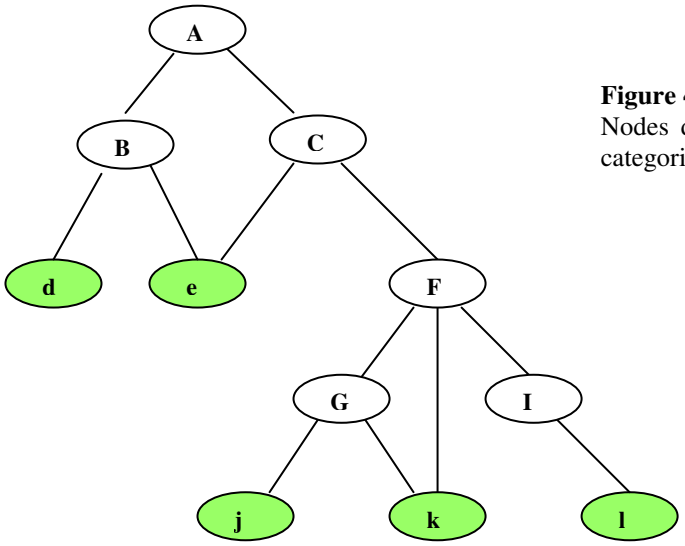


Figure 4.2.1. A model ontology graph. Nodes d, e, j, k, l are the targets for the categorisation process.

The GOC score (described in detail in appendix 4.5.1) for any given parent is a reflection of the parent's fitness to achieve the desired level of abstraction of the functions of the query terms it subsumes. In the GOC mathematical model, the desired level of specificity is set via parameter s. A positive s emphasises specificity and as such the highest scores are given to the most specialised parents. On the other hand, a negative s downweights specificity in favor of coverage and as such the top scores are granted to parents with broader

semantics. In appendix 4.5.1, we explain in detail the way parameter s modulates the dynamics of specificity and coverage in the GOC mathematical model.

In this work, we applied GOC to categorise the gold standard term set (which is the set of terms associated with the genes from the published microarray nerve injury studies). These are the so-called 'query terms' in the GOC vocabulary. The input to GOC consisted of a file listing the gold standard terms, a second file containing GO in XML file format as well as a chosen value for parameter s; all other parameters were set to default. Moreover, we experimented with varying the value of s; thus, we ran GOC with s set to one of three values –1, 1 and 2. Expectedly and as a general trend, the higher the s the more specialised were the resulting clusters. However, we noticed that at any given value of s, individual clusters may vary in their levels of specificity. This is because in the GOC model, specificity is expressed as a relative entity so that for any given parent, specificity is based on how far up in the GO graph the parent is from child query terms (more details in appendix 4.5.1). Since query terms from different clusters may be at different levels in the GO graph, so will the root terms for the clusters. Given this observation, we combined the results from different GOC runs featuring varying s values and selected a final set of clusters that featured a comparable level of semantic

specificity, on the basis of reasonable judgment. However, where clusters overlapped with each other, we felt that it was necessary to make the clusters slightly more specialised to cut down on the amount of overlap.

The previous analysis was largely done by manual inspection of the clusters, which depended on our ability to visualise the clusters. For that, we used the graph visualisation and manipulation tool *yEd* available from *www.yworks.com/products/yed***.** yEd is a java-based software that allows fine drawing of graphs using a variety of different layouts. Moreover, the graph images by yEd are dynamic and can be edited in a variety of ways. More importantly, yEd provides a wide selection of graph manipulation tools. For instance, for any given target node(s), it is possible to select predecessor or successor nodes or generally any node reachable from the target(s). All graph images presented in the work were generated using the yEd software.

yEd is designed to take in various file formats of graph structures such as *XML*, the *graph modelling language (GML)* and its XML derivative *XML-based GML.* Unfortunately, the clusters from the GOC output were given in a format that is not recognised by yEd: the 'dot file' format. Therefore, scripts were written to convert the output clusters from GOC to the GML file format to make them compatible with yEd.

4.3. Results & discussion

4.3.1. Reliability of the gold standard terms

In this chapter, we extracted the GO term annotations of genes from published studies featuring direct injury to the nerve, in an attempt to build a formal model of the functions enriched at the transcriptional level following peripheral nerve injury. The resulting set of terms, which we named the 'gold standard term set', is composed of 560 unique terms originating from 346 unique genes. Table 4.3.1 shows the count of genes and terms from each study.

Table 4.3.1. Genes and terms counts from all five selected studies.

	Wang et al	Xiao et al	Costigan et al	Valder et al	Literature survey
Genes	127	119	230	114	69
GO Terms	229	171	298	85	92

One important requirement for the gold standard term set is consistency with the biology of nerve injury. We use the term study occurrence (see methods, section 4.2.2) as a measure of confidence, following the logic that terms occurring most frequently across the studies are most believable.

Unfortunately, examining the distribution of term study occurrence values from all terms in the gold standard set, we find that half of the terms occur in only one of five selected studies (table 4.3.2).

Table 4.3.2. The distribution of term study occurrence values from all gold standard terms. The counts of terms scoring a term study occurrence value of 1,2,3,4 and 5 are given.

Study occurrence	1	2	3	4	5
Term count	278	118	65	50	49

Instead of seeking exact term matches between the studies, it is likely to be more efficient to look for similarities between the terms by exploiting the relationships between them in the ontology. This approach is more efficient for two main reasons: first, the fact that we are combining slightly different models of peripheral neuropathy (axotomy, SNT); second, functionally equivalent genes from different datasets may be annotated with terms capturing varying levels of the function semantics. It is important to note that this chapter simply discusses the idea of considering semantic relationships with terms from the gold standard set while evaluating the evidence for each individual term in this set. Chapter VI on the other hand, takes this concept further by incorporating it into the mathematical model used to benchmark the results from functional analysis of an LPC nerve injury dataset against the gold standard set of terms.

For now and in order to further study the semantic relationships between the gold standard terms, we will analyse their induced GO subgraph. This consists of the part of the GO graph that features all paths leading from the gold standard terms to the root term (Fig 4.3.1). The resulting subgraph has the benefit of encapsulating the set of gold standard terms within a unified ontology based structure that captures the logical relationships between them.

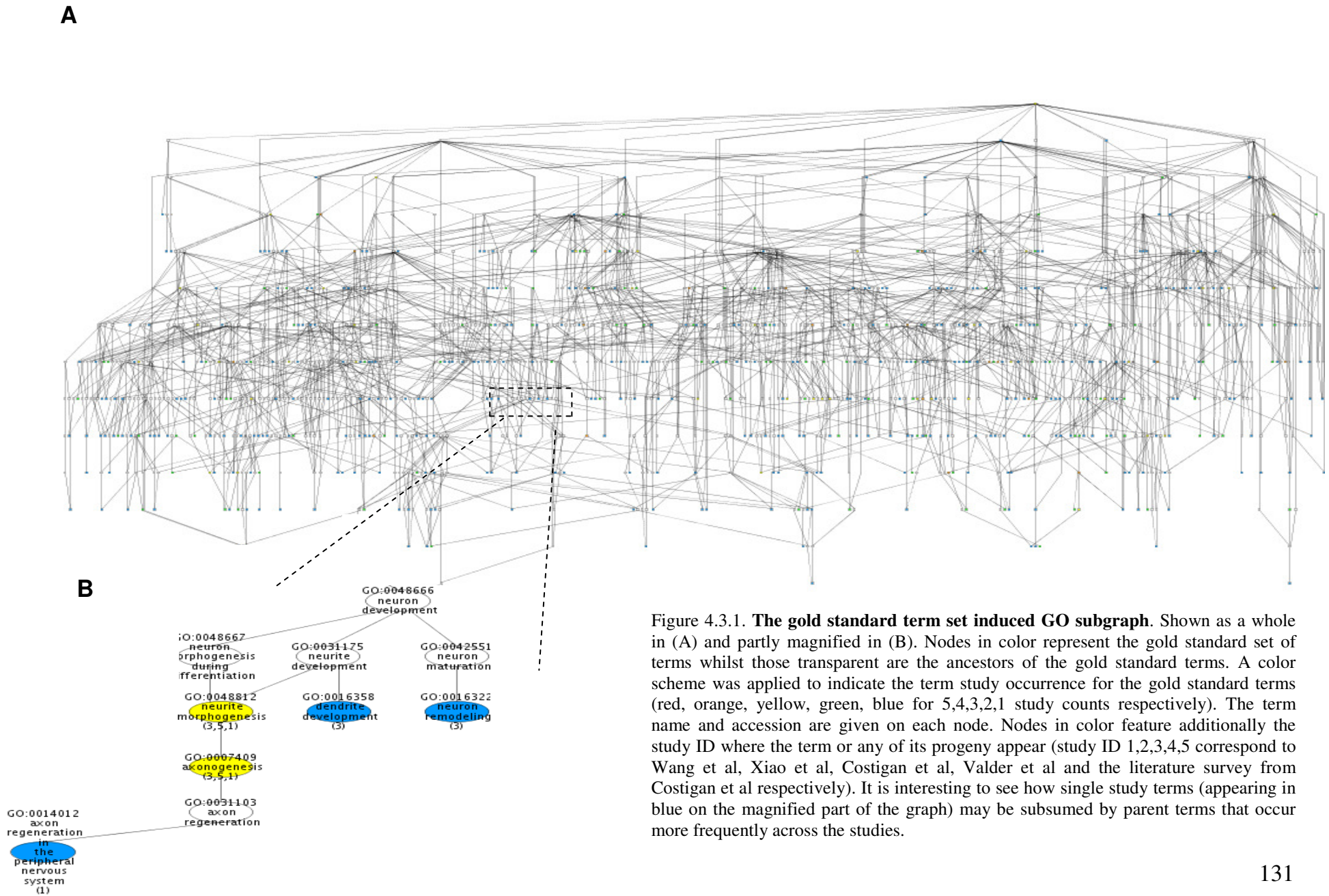


Figure 4.3.1. **The gold standard term set induced GO subgraph**. Shown as a whole in (A) and partly magnified in (B). Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to Wang et al, Xiao et al, Costigan et al, Valder et al and the literature survey from Costigan et al respectively). It is interesting to see how single study terms (appearing in blue on the magnified part of the graph) may be subsumed by parent terms that occur more frequently across the studies.

Close examination of the gold standard term induced subgraph (Fig 4.3.1) revealed that many terms in the gold standard term set are ancestors to other terms in the same set. More important is the observation that many of the subsumed terms are those single study occurring terms that account for half of the gold standard term set, while their subsumers appear to be more common across the studies. Arguably, frequent occurrences of parent terms could add to the confidence level of their children. In many cases, low study occurrence terms correspond to those featuring very specialised functions as it is generally a feature of the GO graph that the most specialised terms are the least populated with genes and as such least likely to be common across the studies. One example is the term 'axon regeneration in the peripheral nervous system (GO:0014012)' which only appears in the Wang study; going one level up, we find that the less specialised parent 'axonogenesis (GO:0007409)' is more common across the studies as it also features in the Xiao and Costigan studies (Fig 4.3.1-B). Naturally, our confidence about the child term increases when we consider association with the parent.

Subsumption by parent terms is not the only relationship observed in the gold standard term induced subgraph. Other, perhaps more distal relationships are also visible. For instance, some terms are cousins thus sharing common ancestors (Fig 4.3.1-B). Going one level higher from pair relationships, we

could consider concentrations of terms. These more complex relationships should also be explored to boost our confidence about participating terms. However, such inference has to be handled with care and should only be allowed in the presence of a strong semantic link. For instance, terms with a general meaning should not be used to reinforce our confidence about their progeny and a similar level of caution should be applied with distant cousins.

4.3.2. Analysis of clusters of gold standard terms

Semantically, concentrations of terms are biologically important as they define major functional themes that take part in the complex biological response to peripheral nerve injury. The complexity of this response is certainly visible on the gold standard term induced subgraph shown on Figure 4.3.1. Hence, it was considered useful to split the subgraph into major components. Splitting the subgraph is equivalent to categorising the gold standard terms under parents terms that provide an abstraction of their functions; a task that can be handled by the gene ontology categoriser GOC (described briefly in the methods section 4.2.3 and in detail in appendix 4.5.1). It is important to note that the purpose of this clustering analysis is to assist in the biological interpretation of the gold standard terms and not to provide a mechanism for exploiting the

relationships between them to evaluate their evidence, as this is rather the subject of chapter VI.

After a few GOC runs at varying values of specificity parameter s, the output was visualised with yEd and manual refinement was performed to yield 14 distinct clusters. The criterion for cluster selection was based on achieving the highest level of abstraction that preserves the essence of the function. Although, sometimes clusters were chosen to be more specialised in order to avoid extensive overlap. Out of the 560 terms in the gold standard term set, the clustering excluded 70 terms, of which 45 are singletons while the rest either corresponded to very general terms or formed small clusters, which were deemed insignificant. The clusters are referred to by the name of their root terms and are listed in table 4.3.3, together with the count of the gold standard terms and genes associated with them. It is important to note that although the clusters are referred to by their root terms, each cluster only contains the progeny of the root term that is part of the gold standard term induced subgraph.

Table 4.3.3. Clusters from the gold standard term induced subgraph obtained using GOC and further refined manually. The count of gold standard terms and genes associated with each cluster is given in absolute numbers as well as percentages with respect to the overall number of gold standard terms and their gene associates from the five published datasets respectively.

Cluster	**Term count (%)**	**gene count (%)**
Nervous system development (GO:0007399)	25 (4)	51 (11)
cell cycle process (GO:0022402)	8 (1)	24 (5)
Cellular component organization and biogenesis (GO:0016043)	56 (10)	66 (14)
cell adhesion (GO:0007155)	9 (2)	22 (5)
Inflammatory response (GO:0006954)	8 (1)	13 (3)
Metabolic process (GO:0008152)	140 (25)	177 (38)
Apoptosis (GO:0006915)	20 (4)	38 (8)
Immune system response (GO:0002376)	36 (7)	41 (9)
reproduction (GO:0000003)	18 (3)	19 (4)
Signal transduction (GO:0007165)	62 (11)	129 (28)
behavior (GO:0007610)	13 (2)	24 (5)
transport (GO:0006810)	54 (10)	103 (22)
Neurological system process (GO:0050877)	23 (4)	54 (12)
Organ development (GO:0048513)	31 (5)	35(7)

Inspection of the resulting clusters leads to some interesting observations. Satisfyingly, there are clusters that describe the changes to nerve cells and the neuronal processes they mediate following injury: mainly the ‘nervous system development (GO:0007399)’ and the ‘neurological system process (GO:0050877)’ clusters. Other specialised functions are also observed: the ‘immune system response (GO:0002376)’ and surprisingly ‘reproduction (GO:0000003)’.

By contrast to the immune response, justifiable by the invasion of the DRG tissue by immune cells following injury, the reproduction function is clearly

absent from the DRG tissue and so are the terms describing the development of other than neuronal or immune related organs in the 'organ development (GO:0048513)' cluster. Therefore, these clusters seemed to be false positives and were consequently discarded from the rest of the analysis. The explanation of their occurrence may lie in the versatility of gene function in different anatomical environments so that the same genes acting upon nerve injury could also be essential to sustaining other cell types residing in other organs. Taking the example of the FGF2 (fibroblast growth factor 2) gene that triggers the fibroblast growth factor receptor signalling pathway, this pathway is known to be critical for the development of many different tissues beyond neuronal ones, such as reproductive gonads, inner ear, lung and muscle tissues.

In addition to these biologically specialised clusters, we also note the presence of clusters featuring generic biological functions that seem applicable to all cell types. Examples are the 'cellular component organization and biogenesis (GO:0016043)', 'apoptosis (GO:0006915)' and 'transport (GO:0006810)' clusters. Further inspection of the clusters reveals that the more specialised clusters correspond to complex system processes such as 'nervous system development (GO:0007399)' whilst the generic ones encapsulate simpler biological processes which may be sorted by their level of granularity into

molecular, subcellular and cellular processes. The molecular processes are those involving the synthesis or manipulation of biological molecules such as metabolic processes, the subcellular class refers to processes that affect particular structures inside the cell such as organelles and finally the cellular processes are those altering the functioning of the cell as a whole such as apoptosis and the cell cycle. Table 4.3.4 organises the GOC clusters into the four classes of biological processes outlined above: system, cellular, subcellular and molecular.

Table 4.3.4. Classification of GOC clusters by increasing complexity of the biology process they encapsulate.

Biological process class	GOC clusters
Molecular	Metabolic process (GO:0008152) Transport (GO:0006810) Signal transduction (GO:0007165)
Subcellular	Cellular component organization and biogenesis (GO:0016043)
Cellular	Cell adhesion (GO:0007155) Cell cycle process (GO:0022402) Apoptosis (GO:0006915)
System	Nervous system development (GO:0007399) Neurological system process (GO:0050877) Immune system process (GO:0002376) Behavior (GO:0007610) Inflammatory response (GO:0006950)

The reason the clusters show varying levels of biological complexity is because the gold standard terms they include are also at varying levels of semantic granularity. This is because the gold standard terms were obtained from gene candidates and genes are usually annotated with terms of varying

granularity in an attempt to capture the semantic complexity of their mediated biological processes. Taking the example of the trophic fibroblast growth factor 2 (FGF2) from the Wang and Costigan studies, we find it associated with the following terms:

‘neurite morphogenesis (GO:0048812)’

‘activation of MAPK activity (GO:0000187)’

‘nuclear translocation of MAPK (GO:0000189)’

‘positive regulation of transcription (GO:0045941)’

The term ‘neurite morphogenesis (GO:0048812)’ specifies the type of cellular activity undertaken by FGF2 as part of the nervous system response to injury, presumably referring to the process of axonal elongation that allows injured neurons to regain contact with the target. The rest of the terms provide insights into the intracellular molecular processes that drive neurite morphogenesis. Thus, it appears that FGF2 acts by activating the key MAP kinase, which once transported to the nucleus induces the transcriptional activity within the cell body of injured neurons, presumably leading to the synthesis of essential growth material for the growing axon.

4.3.2.1. Cluster gene overlap analysis

As reflected by FGF2, the dependencies between biological processes from varying biological complexity levels are revealed in the context of gene function. Thus, we looked to find genes common between pairs of GOC clusters across the hierarchy of biological process classes outlined in table 4.3.4 in order to characterise the functional dependencies between them. In particular, by revealing how biological processes from the different levels in the hierarchy may in turn take part in more complex processes from higher levels, this analysis enabled us to reach a better understanding of the biological significance of the generic GOC clusters (from the molecular, subcellular and cellular levels) by ultimately associating them with either major system processes induced following injury to the peripheral nerve (being neuronal/neurological and inflammatory/immune systems).

Since genes may be associated with terms that are not functionally related; either due to erroneous annotations or because they capture different functions mediated by the same gene in different biological contexts, the occurrence of genes annotated with terms from two clusters may not necessarily imply a functional association between them. On the other hand, we would expect two functionally related clusters to show an amount of gene overlap that is

significantly higher than an unrelated pair of clusters. This was investigated by calculating the gene overlap from all possible pairs of clusters from within and across the different classes of biological processes outlined in table 4.3.4. Moreover, for the gene overlap value to be comparable across all cluster pairs, it was normalised for the sizes of clusters within the pairs. This was done by expressing the gene overlap as a fraction of the total gene count from both clusters in the pair. Examination of the resulting distribution of gene overlap values revealed that a value of 0.1 may be a reasonable significance threshold as only 20% of all possible cluster pairs scored a higher value. The results from the gene overlap analysis for all cluster pairs are shown in table 4.3.5.

	Transport 0006810	metabolic process 0008152	Signal transduction 0007165	Cellular component organization and biogenesis 0016043	Cell adhesion 0007155	cell cycle process 0022402	apoptosis 0006915	nervous system development 0007399	Neurological system process 0050877	immune system process 0002376	behavior 0007610	inflammatory Response 0006950
Transport 0006810	103	(0.05) 16	(0.06) 15	(0.15) 26	(0.01) 2	(0.02) 3	(0.04) 6	(0.13) 20	(0.14) 22	(0.07) 10	(0.04) 5	(0.01) 2
metabolic process 0008152		177	(0.10) 31	(0.03) 8	(0.02) 4	(0.02) 5	(0.07) 16	(0.12) 28	(0.03) 8	(0.12) 28	(0.05) 11	0
signal transduction 0007165			129	(0.04) 8	(0.03) 5	(0.07) 10	(0.14) 24	(0.11) 20	(0.11) 21	(0.07) 13	(0.08) 13	(0.02) 3
Cellular component organization and biogenesis 0016043				66	(0.07) 6	(0.05) 5	(0.14) 15	(0.12) 14	(0.04) 5	(0.03) 3	(0.03) 3	(0.03) 3
Cell adhesion 0007155					22	(0.02) 1	(0.03) 2	(0.11) 8	(0.01) 1	(0.16) 10	(0.04) 2	0
cell cycle process 0022402						24	(0.04) 3	(0.13) 10	0	(0.10) 7	(0.02) 1	0
Apoptosis 0006915							38	(0.11) 10	(0.03) 3	0	(0.05) 3	(0.02) 1
Nervous system development 0007399								51	(0.10) 10	(0.04) 4	(0.06) 5	(0.03) 2
Neurological system process 0050877									54	(0.01) 1	(0.13) 10	(0.03) 2
immune system process 0002376										41	(0.07) 5	(0.14) 8
Behavior 0007610											24	0
Inflammatory Response 0006950												13

Table 4.3.5. Gene overlap analysis. For each pair of clusters, the number of genes in common is given in absolute numbers and as a fraction of the total number of genes from both clusters (shown in between parentheses). A gene overlap amounting to a fraction that is greater or equal to 0.1 is considered significant (shown in red). For each cluster, the total number of genes is shown in green.

4.3.2.2. Cluster term overlap analysis

In addition to gene overlap analysis, an ontology term overlap analysis was also conducted, again to investigate the functional dependencies between the various GOC clusters. Here, we check whether two clusters share the same ontology terms. We use the diagram on Figure 4.3.2 to illustrate the difference between the gene and term overlap analyses. Thus, the two clusters of terms in Figure 4.3.2, delimited by blue and red dashed lines, feature two terms in common (shown in purple) which constitute their term overlap. As for the gene overlap, there are 5 genes in common to both clusters (shown in bold and underlined); these are NPY, FGF, GDNF, ATF3 and BDNF.

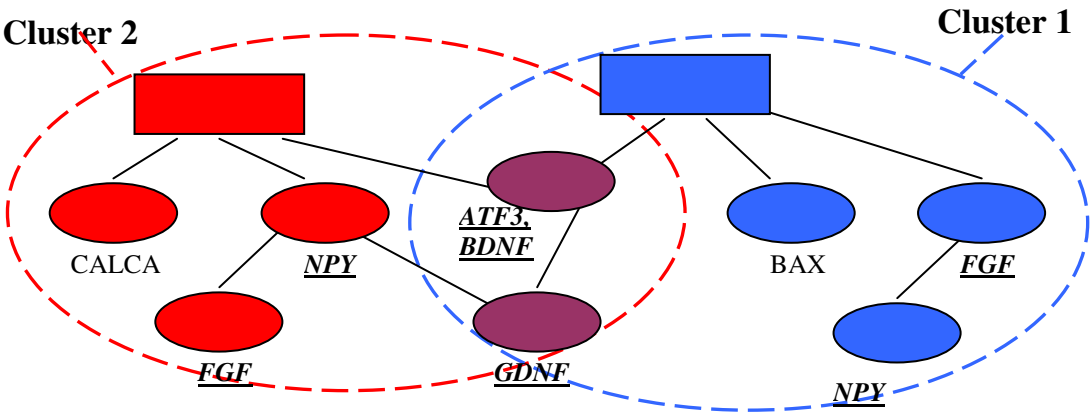


Figure 4.3.2. Diagram illustrating the gene and term overlap analyses between clusters of terms. Two clusters are visible on the diagram: clusters 1 and 2 delimited by dashed lines in blue and red respectively. Terms in blue correspond to cluster 1 whilst those in red correspond to cluster 2. Terms in purple are shared between the two clusters. Genes are shown below the terms that annotate them. Importantly, a gene may be annotated with two different terms from different clusters. Genes likewise shared between clusters are indicated in bold and underlined.

Importantly, overlap in terms between clusters implies overlap in genes as the terms and the genes annotated with them are collectively shared by the clusters. The opposite is not true since the same genes could be associated with different terms from two clusters. The reason why we opted to use the term overlap analysis in addition to the gene overlap analysis despite the fact that the latter is implied by the former is that where the gene overlap between two clusters falls below the significance threshold, the existence of a common term would re-establish the evidence for a functional association between the two clusters.

The rational behind using the term overlap analysis to trace functional relationships between different GOC clusters is that ontology terms from deeper levels in the GO graph are more granular, reflecting additional functional details that may uncover unanticipated links with higher-order functions. For instance, the term ‘Notch signalling pathway involved in neuron fate commitment (GO:0021880)’ depicts the involvement of the Notch signalling pathway in the process of neuron fate commitment. The term in question is common to the ‘signal transduction (GO:0007165)’ and the ‘nervous system development (GO:0007399)’ GOC clusters (Fig 4.3.3) from the molecular and system classes respectively. Importantly, the term appears to relate to the root term from the ‘signal transduction (GO:0007165)’ cluster

via a chain of 'is a' type of relationships while it links to the 'nervous system development (GO:0007399)' cluster via a 'part of' relationship. This illustrates the essence of the term overlap analysis, whereby functional associations between GOC clusters from varying levels of biological complexity (outlined in table 4.3.4) are revealed by means of identifying terms from clusters from low complexity levels whose functionality is inherently partial to higher-order biological processes from clusters from higher levels.

The term overlap analysis was based on identifying gold standard terms common to pairs of clusters but could have been also targeted at the overlap in the progeny of gold standard terms from the two clusters, since child terms are semantically indicative of their parents in the gene ontology. This applies to the previous example: term 'Notch signalling pathway involved in neuron fate commitment (GO:0021880)', which is not a gold standard term itself but which inherits two gold standard parent terms: the 'Notch signalling pathway (GO:0007219)' and the 'neuron differentiation (GO:0030182)' from the 'signal transduction (GO:0007165)' and the 'nervous system development (GO:0007399)' clusters respectively (Fig 4.3.3).

The occurrence of a common term between clusters can only arise from a functional link between them. As such, unlike the gene overlap analysis, we

did not need to infer any significance from the number of common terms. It follows that the term overlap measure is expressed in an absolute rather than a relative fashion. The results from all cluster pairs are shown in table 4.3.6.

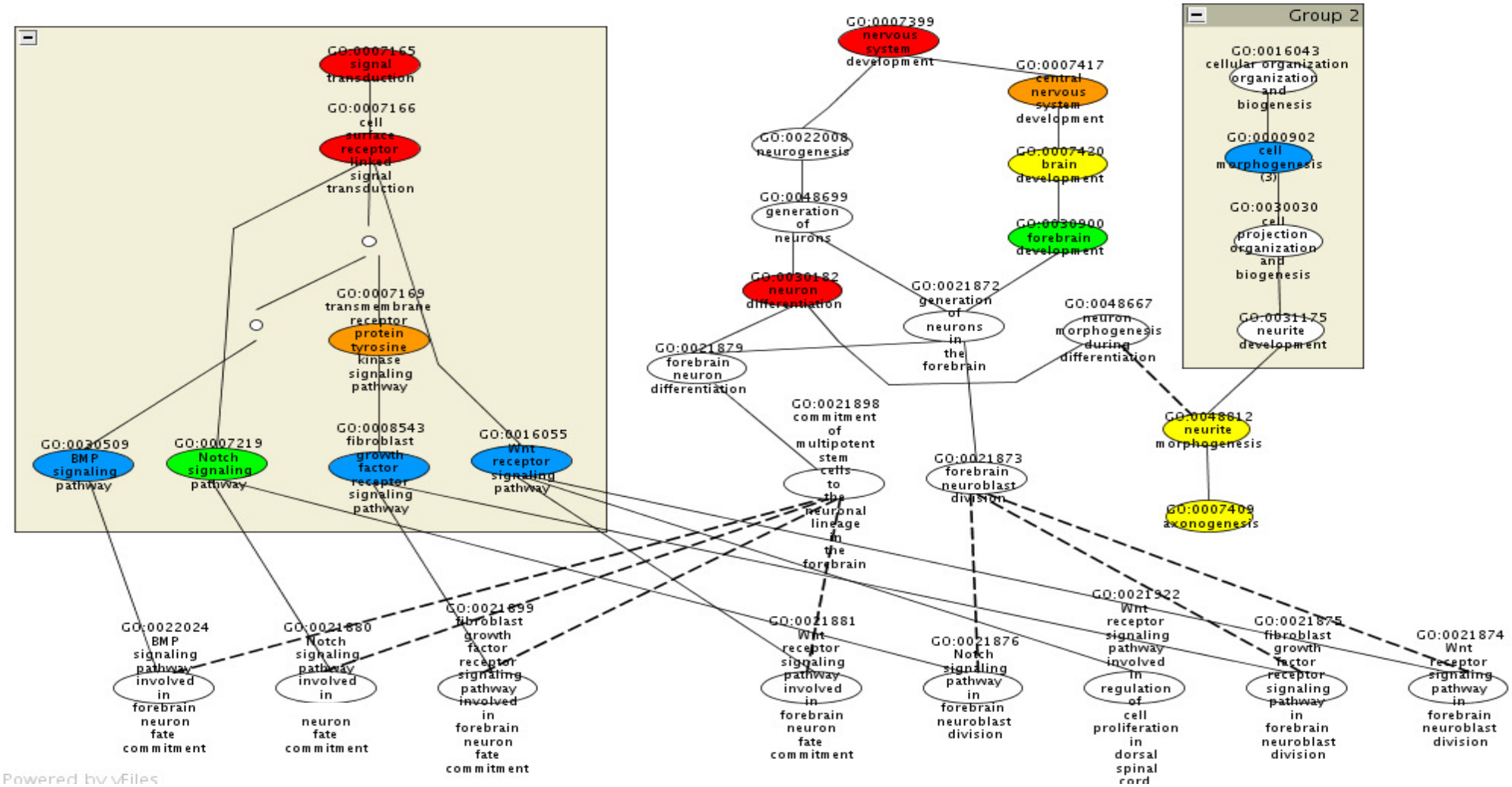


Figure 4.3.3. Relationships between low and high level biological processes captured as 'part-of' relationships in GO. Child terms common to the 'signal transduction (GO:0007165)' cluster, the 'cellular component organization & biogenesis (GO:0016043)' cluster (both clusters marked in grey boxes) and the 'nervous system development (GO:0007399)' cluster from the more complex biological system class are shown. Importantly, these common children terms are associated with the higher order nervous system development process via 'part-of' relationships (shown in dashed lines). Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Clusters were truncated to show only parents to common child terms.

	transport 0006810	metabolic process 0008152	signal transduction 0007165	Cellular component organization and biogenesis 0016043	cell adhesion 0007155	cell cycle process 0022402	apoptosis 0006915	nervous system development 0007399	neurological process 0050877	immune system process 0002376	behavior 0007610	inflammatory Response 0006950
Transport 0006810	54	0	0	16	0	0	0	0	4	0	0	0
metabolic process 0008152		140	0	10	0	0	1	0	4	5	0	0
signal transduction 0007165			62	0	0	0	8	0	0	0	0	0
Cellular component organization and biogenesis 0016043				56	0	1	6	2	0	0	0	0
Cell adhesion 0007155					9	0	0	0	0	0	0	0
cell cycle process 0022402						8	0	0	0	0	0	0
Apoptosis 0006915							20	0	0	0	0	0
Nervous system development 0007399								25	2	0	0	0
Neurological process 0050877									27	0	0	0
immune system process 0002376										39	0	4
Behavior 0007610											17	0
Inflammatory Response 0006950												8

Table 4.3.6. Term overlap analysis. The table shows the number of gold standard terms shared by pairs of clusters. The occurrence of term overlap is indicated in red. The total number of gold standard terms from each cluster is shown in green.

The results from the term and gene overlap analyses complemented each other in a variety of ways. Where there were term overlap and significant gene overlap between two clusters from the varying biological process classes outlined in table 4.3.4, the ontological terms in common were examined to reveal details about the nature of functional association between the clusters in the pair. Taking the example of the 'signal transduction (GO:0007165)' and the 'nervous system development (GO:0007399)' clusters, a significant proportion of genes seems to be in common between them indicating a functional interrelationship. Exactly which signalling pathways are involved in which neuronal processes is partly revealed by the terms common to both clusters. Thus, as shown in Figure 4.3.3, a number of signalling pathways seem to be involved in the process of neuron differentiation that occurs following nerve injury including the BMP, Notch, Wnt and the fibroblast growth factor signalling pathways.

Sometimes, two clusters may show an overlap in gene content that is significant enough to suggest a functional link between their encapsulated functions, yet no terms are found in common between them. In other words, the two clusters show a significant gene overlap but no term overlap. In this case, the functions of the genes in common are examined to determine the nature of functional relationships between the clusters. The opposing scenario

is where clusters show an overlap in constituent terms, but score no significant gene overlap. This occurs when the number of genes annotated to the common terms amounts to a minor fraction of the clusters total gene count. Here the functional link between clusters is evident from the term overlap analysis alone.

4.3.2.3. Interpretation of cluster biological significance

As mentioned before, the purpose of the gene and term overlap analyses was to expose the relationships between GOC clusters of processes featuring varying levels of biological complexity and ultimately associate the generic clusters from lower complexity levels with clusters encapsulating complex system processes that are biologically specialised. Interestingly, the gene and term overlap analyses also indicate relationships between clusters from the same biological class. From a biological point of view, the relationships among the system processes clusters are important as they highlight the functional integration of varying biological systems during the response to peripheral nerve injury. One example is how the inflammatory state that builds up shortly following injury triggers and maintains the immune response.

In the following sections, we review the functional significance of the GOC clusters while highlighting the functional relationships between them as revealed by the gene and term overlap analyses. We follow a top to bottom approach: clusters from the system processes class are discussed first, followed by those from the cellular, subcellular and finally the molecular class range.

4.3.2.3.1. System process clusters

Among the GOC clusters featuring system processes, we begin with those underlying the neuronal response to injury: the 'nervous system development (GO:0007399)' and the 'neurological system process (GO:0050877)' clusters. There is a tight relationship between the two functions as revealed by the gene and term overlap analyses (tables 4.3.5 & 4.3.6); which is logical in the sense that changes to nerve cells have direct consequences on the signalling processes they mediate.

The 'nervous system development (GO:0007399)' cluster (Fig 4.3.4-A) captures the changes that affect the varying cell types within the DRG tissue following injury. Thus, for the injured neurons, we find terms involved in repair activities whereby the lost part of the axon is replaced in order to regain

contact with target territories: example terms are ‘axonogenesis GO:0007409’ referring to the process of axonal growth and ‘axon ensheathment GO:0008366’ whereby the growing axon is covered with structural myelin from differentiated schwann cells. Other types of cells include Schwann and satellite glial cells, which seem to undergo differentiation following injury as revealed by the term ‘glial cell differentiation (GO:0010001)’. Indeed, the differentiation of Schwann cells is an essential part in the process of myelin formation whereas glial satellite cells that move to surround injured neurons in the DRG following injury are thought to differentiate into neurons to replace those lost by apoptosis (Scholz and Woolf, 2007).

The ‘neurological system process (GO:0050877)’ cluster describes alterations to signal transmission processes following injury to the axon and the resulting effects on sensory perception functions. The plasticity in synaptic transmission underlined in part by a change in the level and type of neurotransmitters and their receptors peripherally following injury (captured in the ‘neurological system process (GO:0050877)’ cluster, Fig 4.3.4-B) serves to sensitise the central nervous system resulting in a net enhancement in sensory functions to noxious and non-noxious stimuli as well as spontaneous aberrant sensations such as neuropathic pain.

Pain is likely to affect certain aspects of behaviour in the injured animal such as sleep, feeding, mobility and social behavior; these processes are all captured under the 'behavior (GO:0007610)' cluster (Fig 4.3.4-C). The relationship between the 'neurological system process (GO:0050877)' cluster and the 'behavior (GO:0007610)' cluster is confirmed by the gene overlap analysis (table 4.3.5).

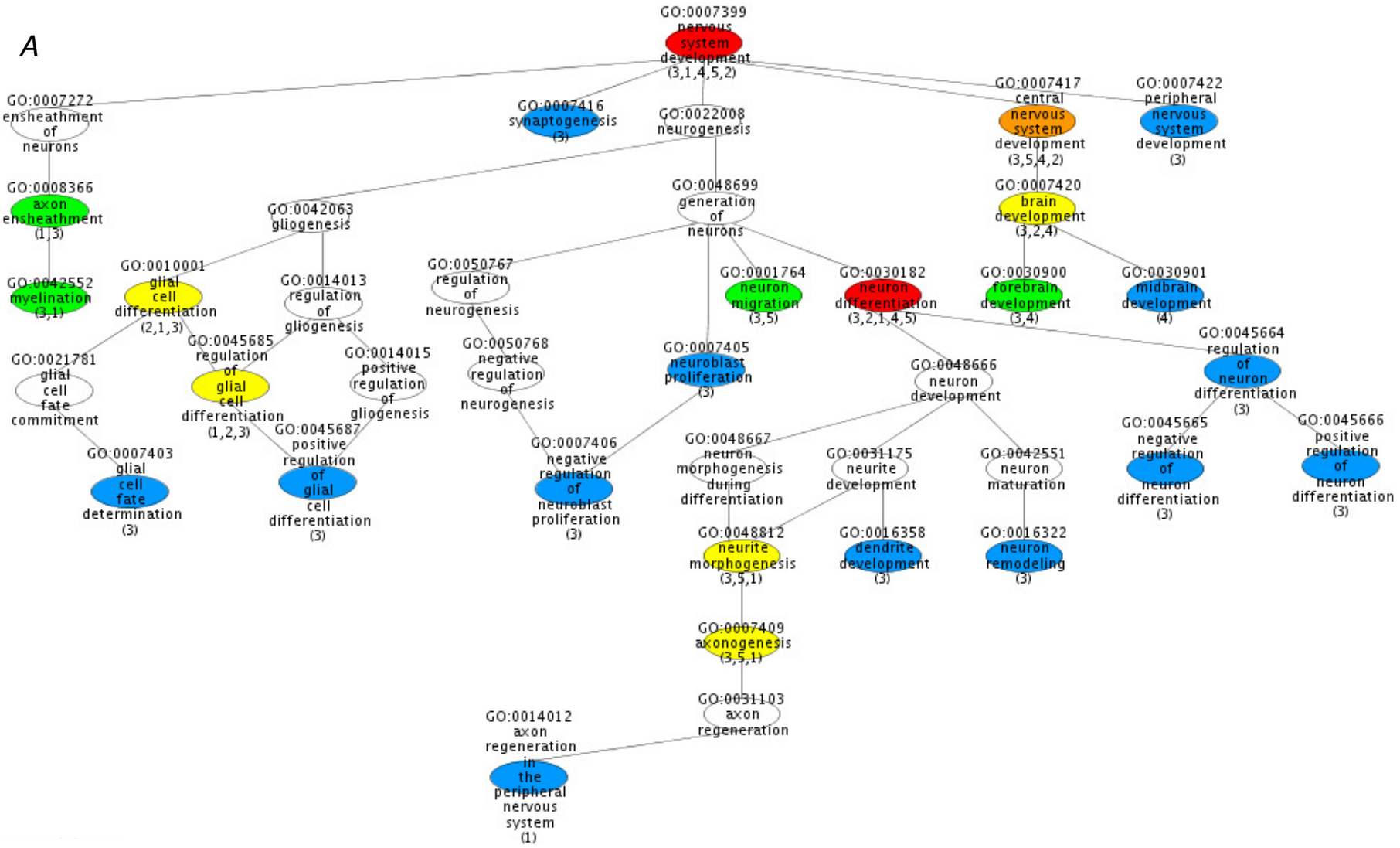


Figure 4.3.4-A. Clusters from the gold standard terms induced subgraph: 'nervous system development (GO:0007399)'. Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

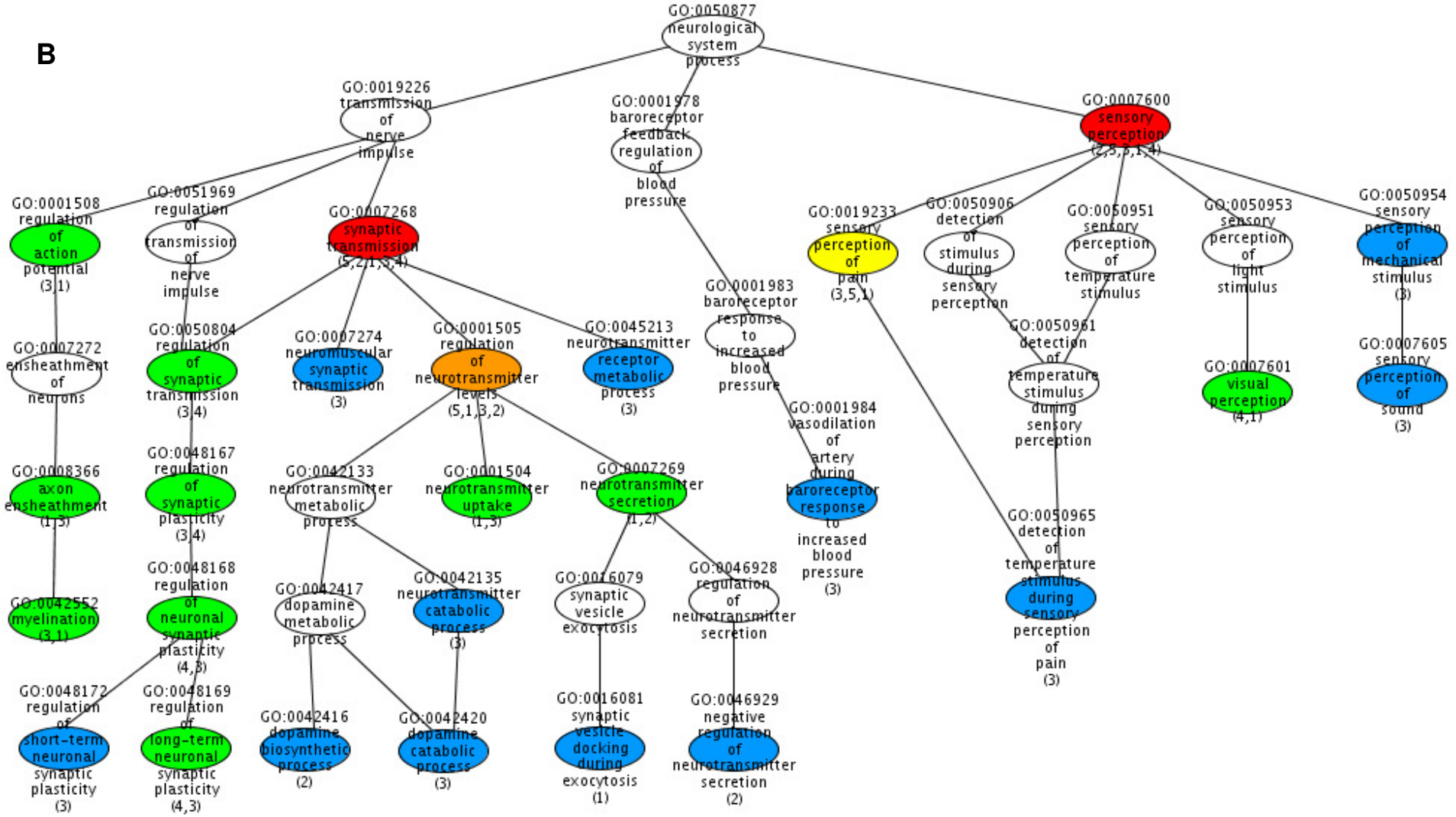


Figure 4.3.4-B. Clusters from the gold standard terms induced subgraph: 'neurological system process (GO:0050877)'. Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

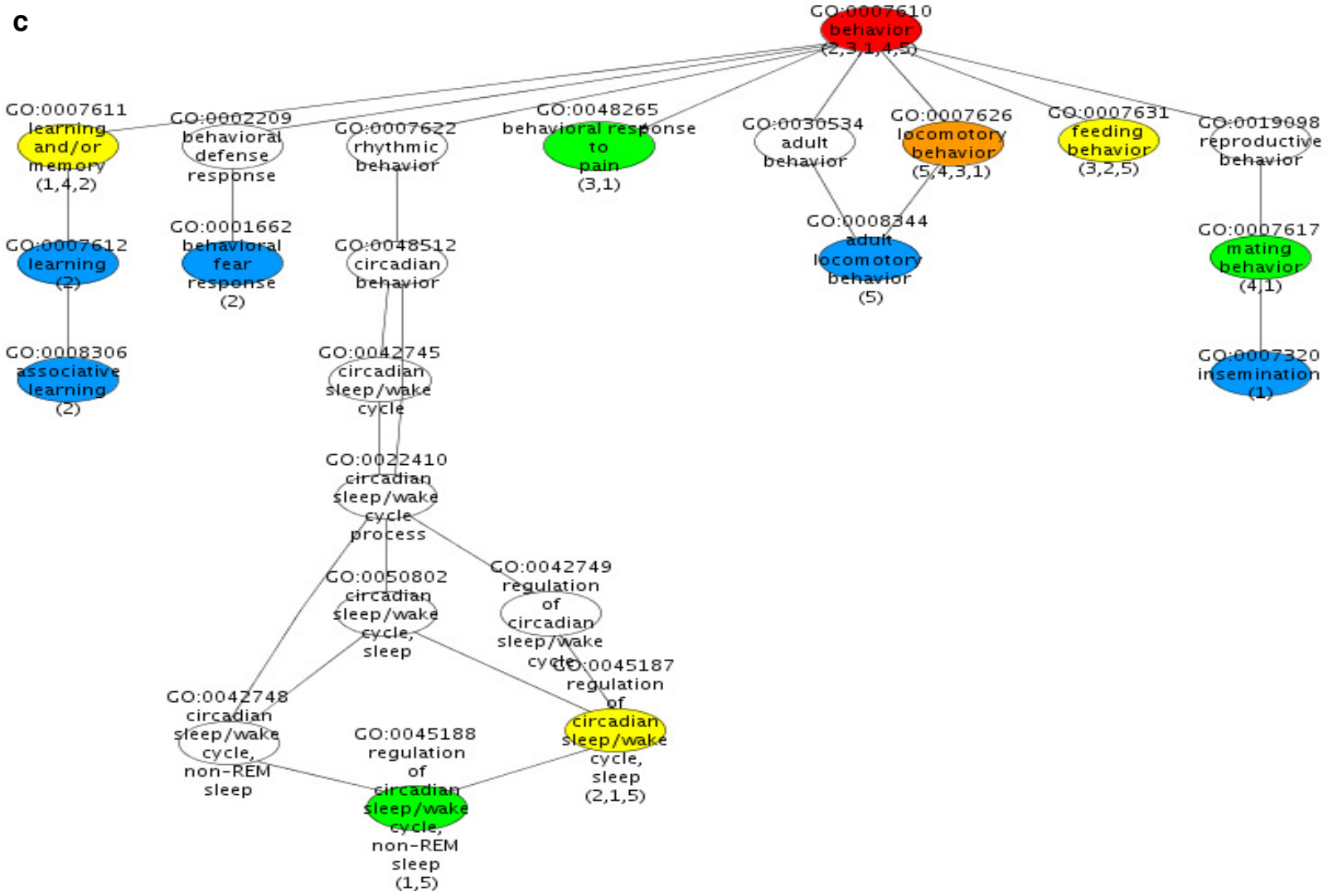


Figure 4.3.4-C: Clusters from the gold standard terms induced subgraph: 'behavior (GO:0007610)'. Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

Beside neuronal cell types, the DRG tissue contains immune cells, which tend to increase in number following injury to the nerve. The bulk of immune processes induced within the DRG following peripheral nerve injury is captured under the 'immune system process (GO:0002376)' cluster (Fig 4.3.4-D). Such processes consist of differentiation and proliferation of immune cells such as T-cells and macrophages as well as antigen processing and presentation, complement activation and immunoglobulin deployment.

The immune response local to the DRG is sustained by an inflammatory state induced by the release of proinflammatory cytokines by invading macrophages following injury. The inflammatory process is captured under the 'inflammatory response (GO:0006954)' cluster (Fig 4.3.4-E). The interplay between the inflammatory and immune processes is well manifested by the gene and term extent of overlap (tables 4.3.5 & 4.3.6) between the 'inflammatory response (GO:0006954)' and the 'immune system process (GO:0002376)' clusters. Among the genes in common to the 'inflammatory response (GO:0006954)' and the 'immune system process (GO:0002376)' clusters are, of course, key cytokines.

Interestingly, proinflammatory cytokines have a well-established role in signalling injury to neurons via activation of numerous intracellular signalling

pathways ultimately altering the transcriptional activity in favour of growth and repair. In the reverse direction, there is evidence in the literature that suggests induction of expression of cytokine interleukin-6 in sensory neurons following injury, which serves to sustain the inflammatory state and related immune processes in the DRG tissue, in what could constitute a feedback loop mechanism. However, such intermingling of neuronal and inflammatory/immune processes is not captured by the gene and term overlap analyses, probably because it only occurs under abnormal pathological conditions which is outside the scope of the GO.

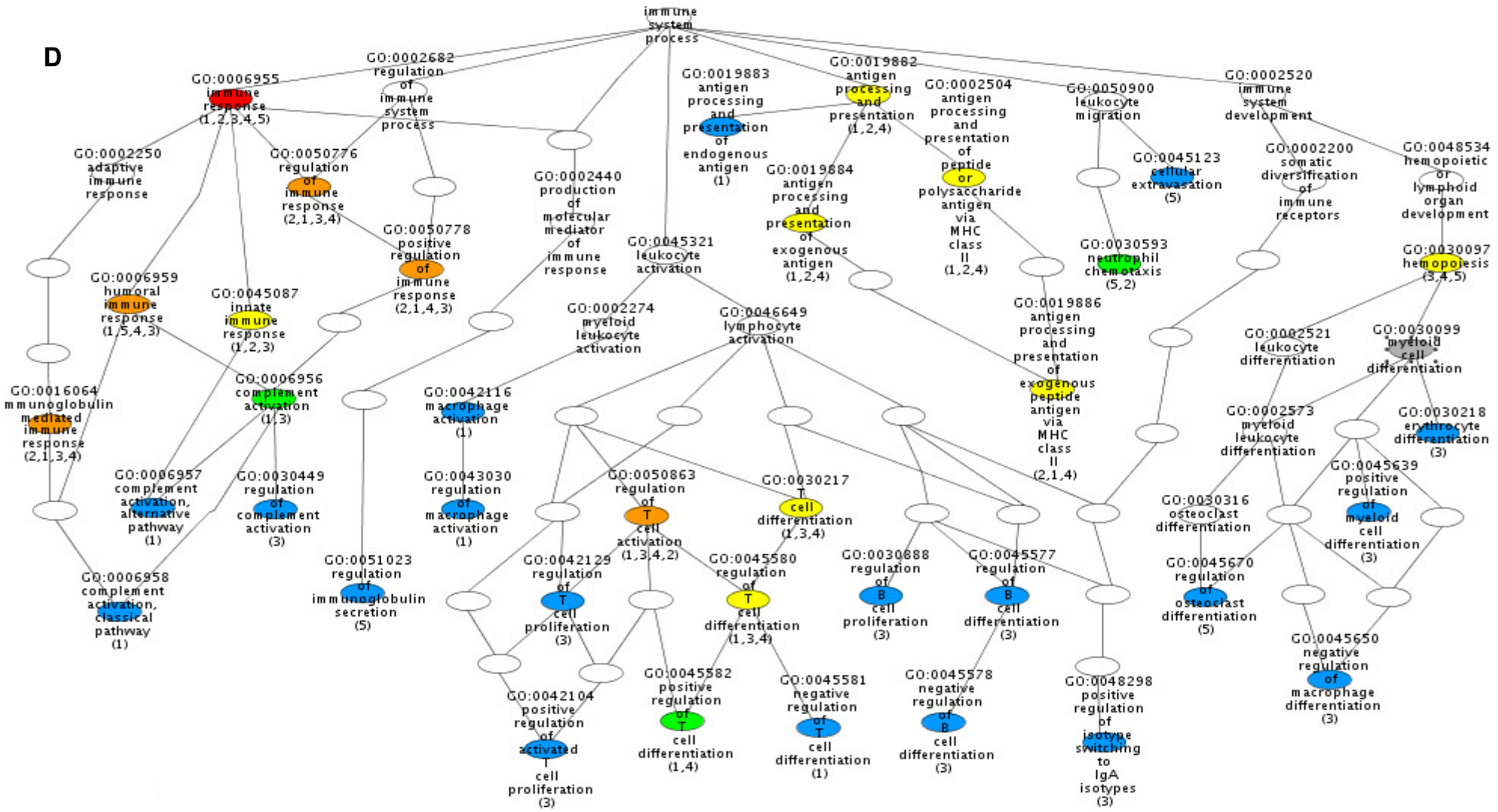


Figure 4.3.4-D. Clusters from the gold standard terms induced subgraph: 'immune system process (GO:0002376)'. Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively). Occasionally, labels from transparent nodes were hidden for picture clarity.

E

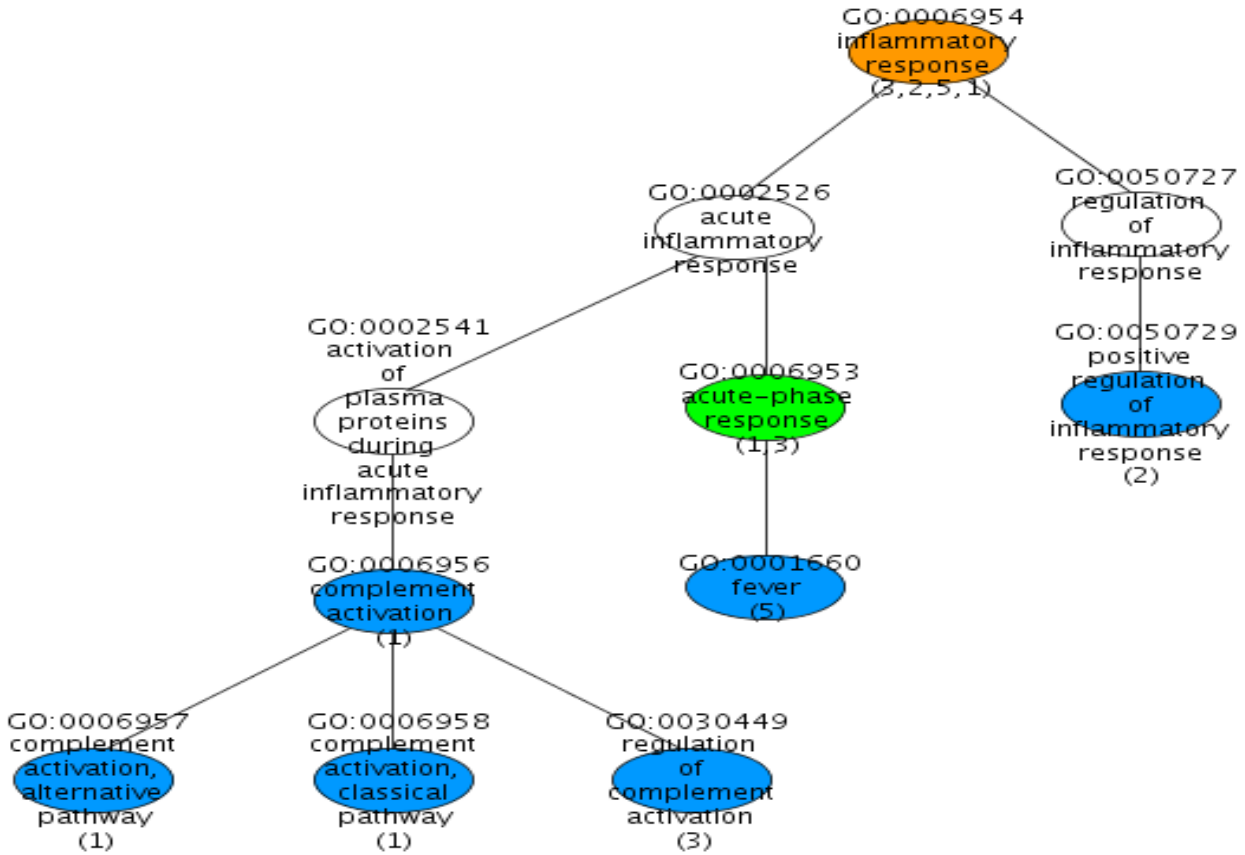


Figure 4.3.4-E. Clusters from the gold standard terms induced subgraph: 'Inflammatory reponse(GO:0006954)'. Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

4.3.2.3.2. Cellular process clusters

Among the 14 GOC clusters, 3 were representative of cellular processes: these are the 'apoptosis (GO:0006915)', 'cell cycle (GO:0022402)' and the 'cell adhesion (GO:0007155)' clusters (Fig 4.3.4-F&G&H). Intuitively, cellular processes are indicative of the changes affecting the varying cell types within the DRG tissue following injury to the nerve. For us to understand the significance of these cellular processes in the context of the biology of nerve injury, we refer to the results from the term and gene overlap analyses for pairs of clusters from the cellular and system classes. For instance, there appears to be a significant proportion of genes common to the 'apoptosis (GO:0006915)' and the 'nervous system development (GO:0007399)' clusters (table 4.3.5). One example is the BAXA_RAT (apoptosis regulator BAX, membrane isoform alpha) gene annotated with both terms 'apoptosis (GO:0006915)' and 'neuron fate determination (GO:0048664)'. As such, we conclude that the apoptotic process is associated with the neuronal cell type, which may be a biologically valid statement since it has been postulated in the literature that a proportion of DRG neurons undergo apoptosis following axonal damage when failing to mount an effective repair reaction to injury.

With the 'cell cycle (GO:0022402)' cluster, the gene overlap analysis (table 4.3.5) indicates a link to the 'nervous system development (GO:0007399)' as well as the 'immune system process (GO:0002376)' clusters, suggesting that cells from both the nervous and immune systems show increased cell cycle activity following nerve damage. This is plausible in the view that the cell cycle is at the heart of cell proliferation and differentiation processes important for both the maintenance of the immune response as well as the repair activities mounted by the nervous system following injury, mainly the differentiation of Schwann cells to form myelin and the probable differentiation of satellite cells into neurons.

Similarly, the cell adhesion function appears to be adopted by both immune and nerve cell types. The significance of the cell adhesion process to the physiology of the nervous system following nerve injury is captured by the gene overlap analysis (table 4.3.5) and can be illustrated by the example of the TSP4_RAT (Thrombospondin 4 precursor) gene, which encodes an adhesive glycoprotein that mediates cell to cell matrix interaction, a process that is vital for axonal pathfinding during neurite growth. As for the immune system, the term 'leukocyte adhesion (GO:0007159)' illustrates the applicability of the cell adhesion function to immune cell types. This is further demonstrated by the gene overlap analysis where a significant proportion of genes appears to be

common to the ‘cell adhesion (GO:0007155)’ and the ‘immune system process (GO:0002376)’ clusters (table 4.3.5).

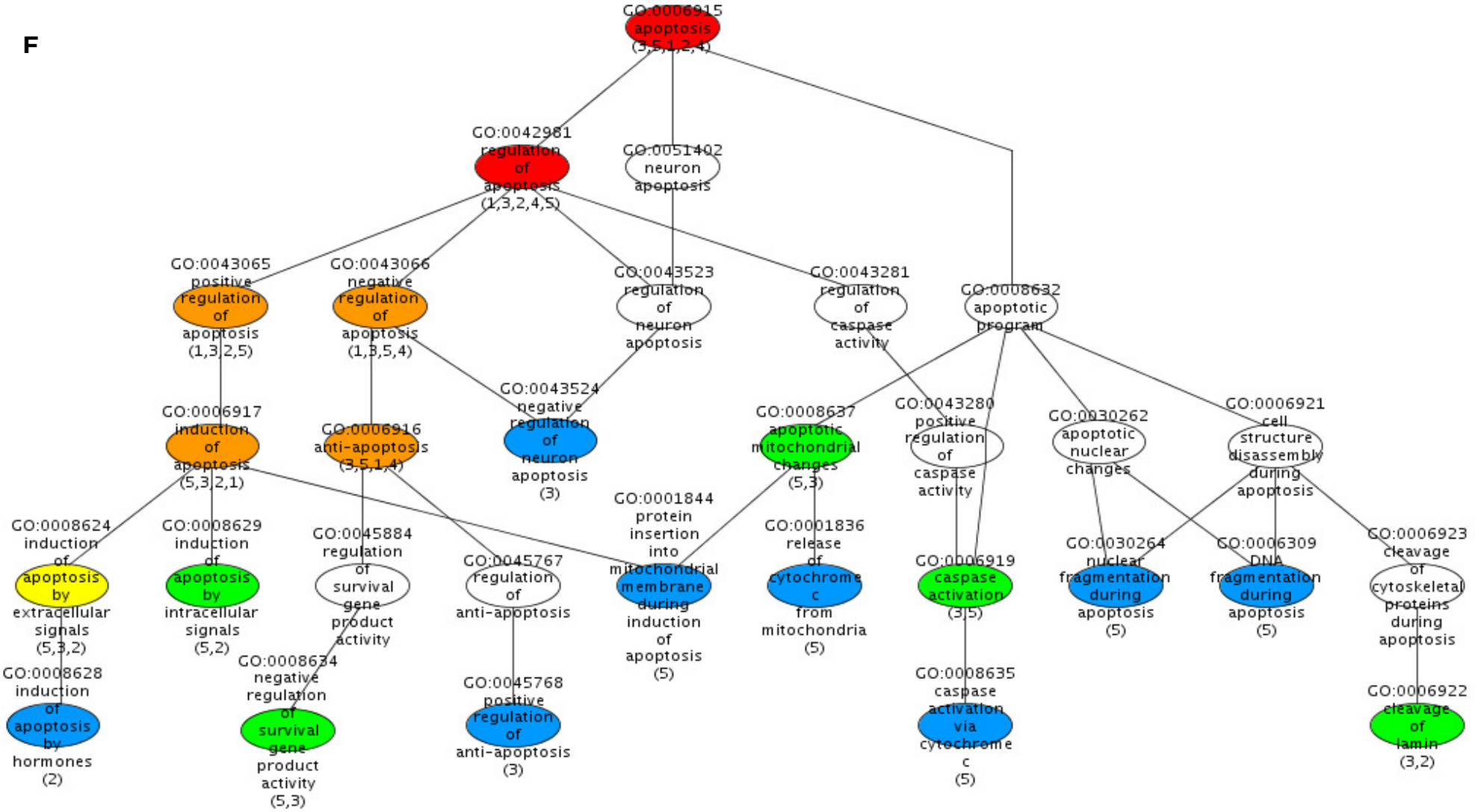


Figure 4.3.4-F. Clusters from the gold standard terms induced subgraph: apoptosis (GO:0006915). Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

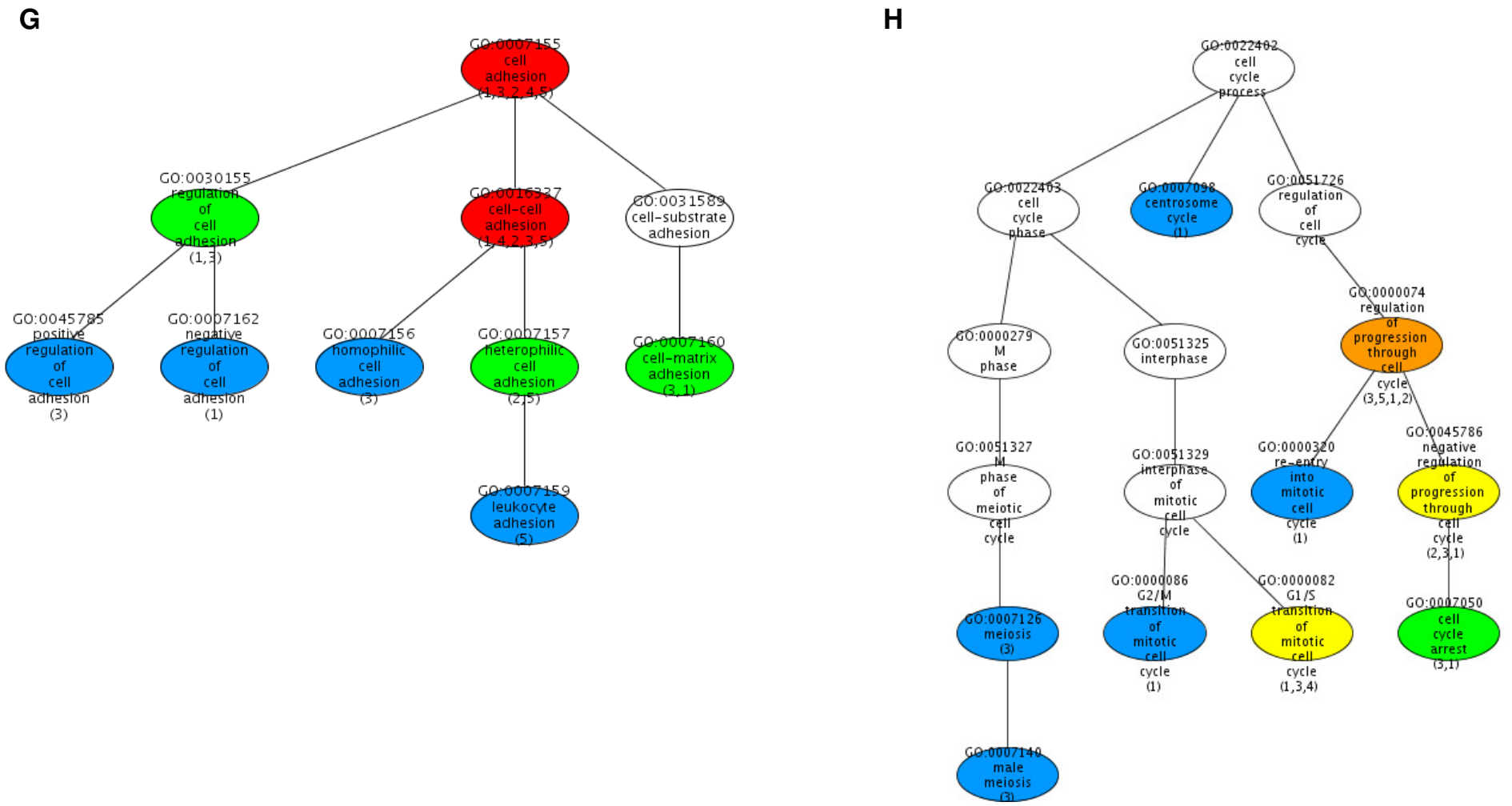


Figure 4.3.4-G&H. Clusters from the gold standard terms induced subgraph: (G) cell adhesion (GO:0007155), (H) cell cycle process (GO:0022402). Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

4.3.2.3.3. Subcellular process clusters

The next level in the classification (table 4.3.4) is that of subcellular processes. The 'cellular component organization and biogenesis (GO:0016043)' cluster (Fig 4.3.4-I) was alone affiliated to this class. In the gene ontology, the 'cellular component organization and biogenesis (GO:0016043)' term refers to the processes that lead to the formation, arrangement of constituent parts, or disassembly of cellular components. Both the term and gene overlap analyses (table 4.3.5 & 4.3.6) indicate a strong association between the 'cellular component organization and biogenesis (GO:0016043)' and the 'nervous system development (GO:0007399)' clusters. Importantly, the process of axonal elongation following injury entails a morphological change that involves membrane biogenesis and organisation of membrane proteins and channels. Furthermore, the retrograde transport of signalling molecules to the nucleus as well as the opposite anterograde transport of axonal growth substances towards the growing end of the axon require cytoskeletal organisation and biogenesis. As for neurons that commit to apoptosis, the cellular component structural diassembly as well as the apoptotic mitochondrial changes are all a form of subcellular processes; hence, the significant gene overlap with the 'apoptosis (GO:0006915)' cluster (table 4.3.5).

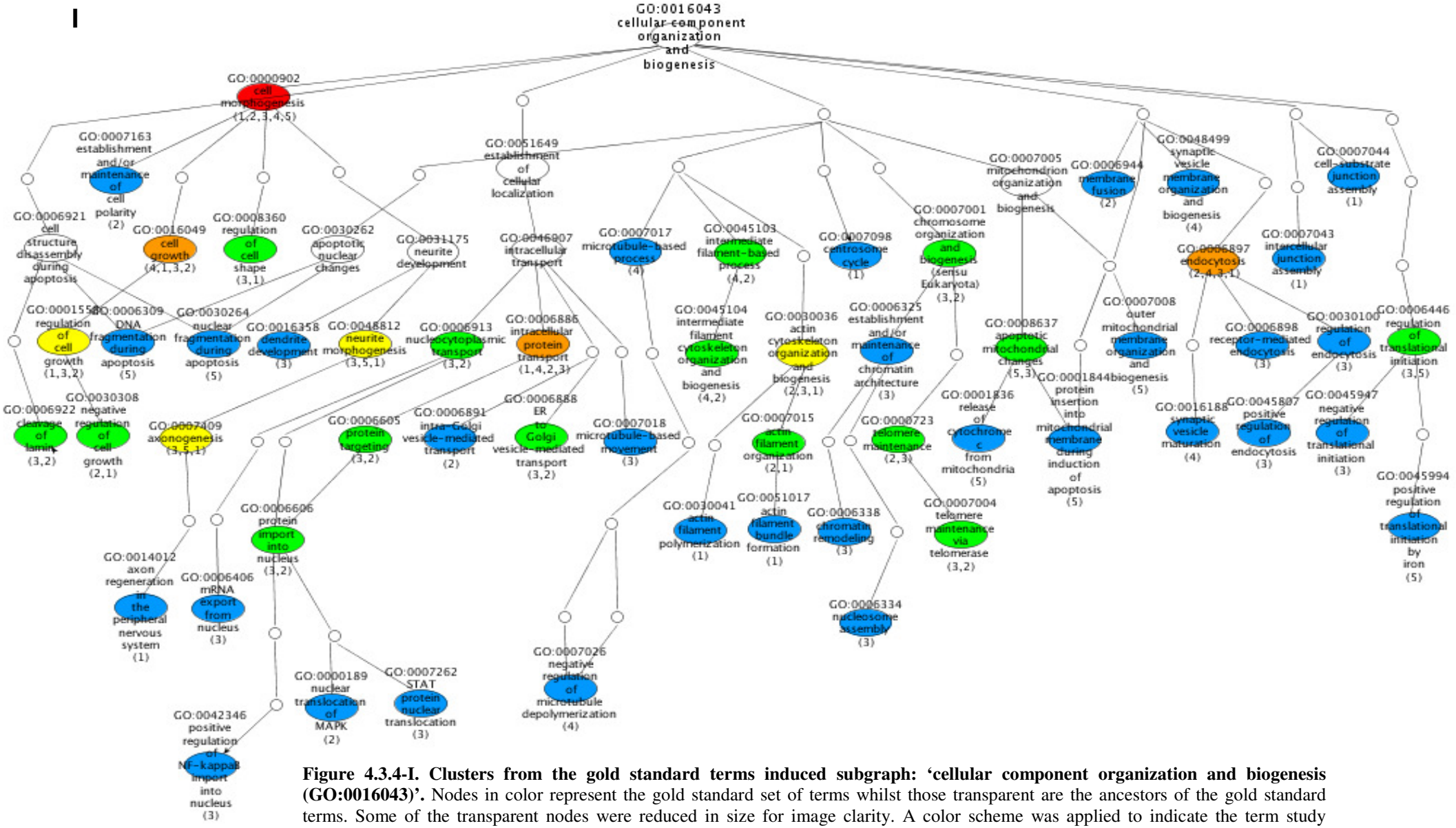


Figure 4.3.4-I. Clusters from the gold standard terms induced subgraph: 'cellular component organization and biogenesis (GO:0016043)'. Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. Some of the transparent nodes were reduced in size for image clarity. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

4.3.2.3.4. Molecular process clusters.

At the fourth tier of our cluster classification, outlined in table 4.3.4, lie those clusters representing core molecular functions that serve to support the cellular and higher system processes induced following nerve injury. These core functions appear in the following clusters: the 'signal transduction (GO:0007165)', 'transport (GO:0006810)' and 'metabolic process (GO:0008152)' clusters.

The 'signal transduction (GO:0007165)' cluster, shown in Figure 4.3.4-J, is an encapsulation of the chain reaction initiated by the interaction of an outside signal with membrane receptors, which causes a change in the level or activity of a second messenger or other downstream target, ultimately effecting a change in the functioning of the cell. In the context of nerve injury, the variety of signals that build up at the site of the lesion and locally within the DRG are transduced to neuronal and non-neuronal cell bodies via a number of intracellular cascades ultimately inducing a change in the cell transcriptional activity. Examples are the JAK-STAT cascade, the MAPKKK cascade, the NF-kappaB cascade and the cytokine/chemokine mediated signalling pathways, all captured under the 'signal transduction (GO:0007165)' cluster.

As elaborated in the introduction section, the transduction of injury related signals may result in the induction of apoptotic signalling cascades; hence the link to the 'apoptosis (GO:0006915)' cluster. Indeed, the term 'apoptotic process (GO:0008632)' and descendents are common to both the 'signal transduction (GO:0007165)' and 'apoptosis (GO:0006915)' clusters as revealed by the term overlap analysis (table 4.3.6) whilst the gene overlap analysis indicates a significant fraction of genes in common to both clusters (table 4.3.5).

Interestingly, the shifts in cellular transcriptional activity that result from transduction of injury related signals lead to de novo or increased synthesis of additional signalling molecules that help recruit further signalling pathways. One example is BMP or bone morphogenesis protein whose pathway appears to be critical for the generation of neurons during development (Fig 4.3.3), but which may also be involved in generating neurons, following injury, to replace those lost by apoptosis. Other signalling metabolites include neurotransmitters such as glutamate and tachykinin that play a role in enhancing synaptic transmission at the junction with the dorsal horn, leading to central sensitisation mechanisms that underlie many of the abnormal sensations following nerve injury such as hyperalgesia, allodynia and chronic pain. In accordance with these observations, there exists significant gene overlap

between the 'signal transduction (GO:0007165)' cluster and the 'nervous system development (GO:0007399)' as well as the 'neurological system process (GO:0050877) clusters' from the system process class (table 4.3.5).

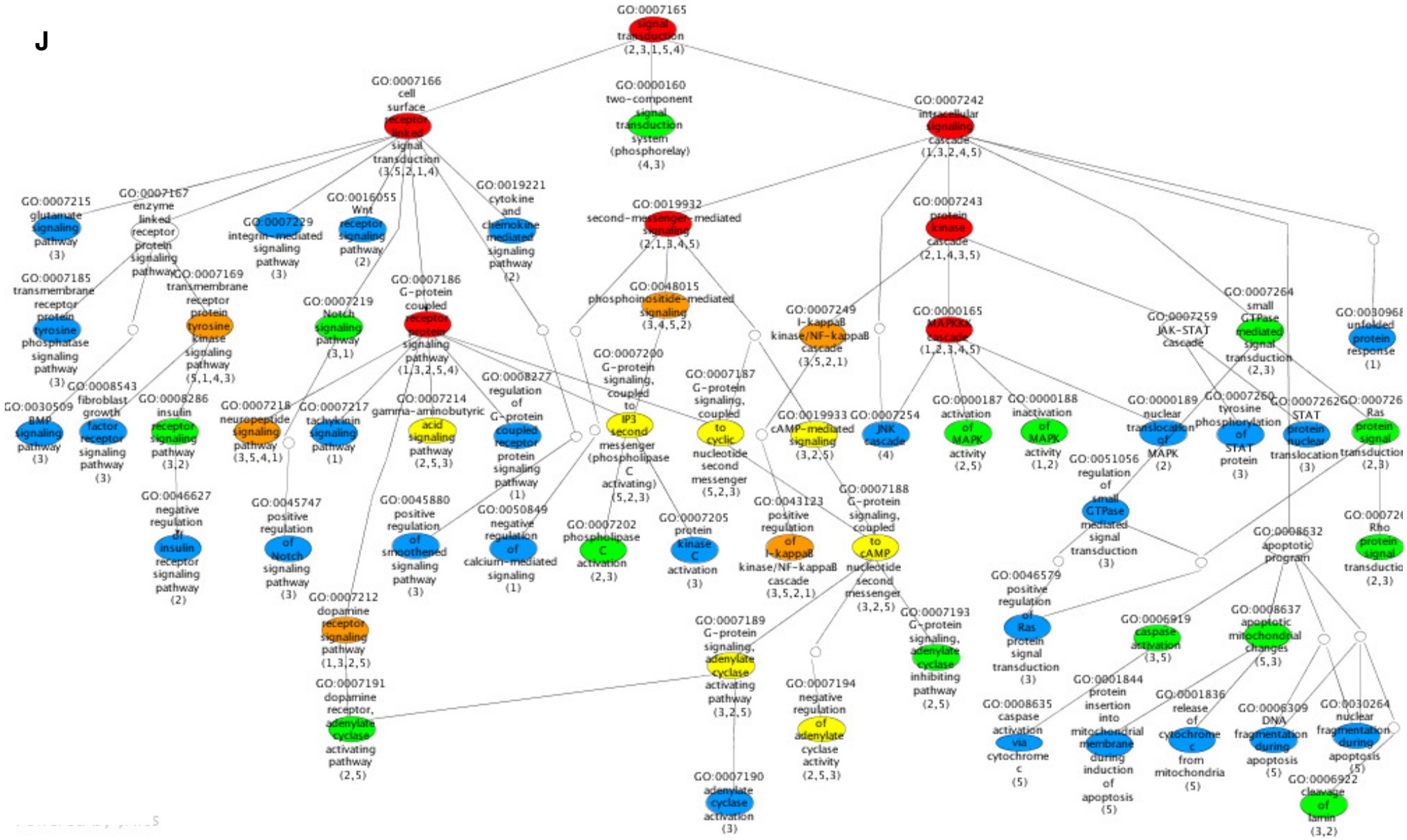


Figure 4.3.4-J. Clusters from the gold standard terms induced subgraph: signal transduction (GO:0007165). Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. Some of the transparent nodes were reduced in size for image clarity. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

One further GOC cluster from the molecular process class is the 'transport (GO:0006810)' cluster (Fig 4.3.4-K). From the gene and term overlap analyses (tables 4.3.5 & 4.3.6 respectively), we find evidence of a functional link with the 'neurological process (GO:0050877)' cluster. Indeed, following injury, ion transport mechanisms are enhanced as well as the uptake and secretion of neurotransmitters, which has a profound impact on the excitability of nerve cells peripherally and centrally.

In addition, the transport function appears to play a role in processes affecting the nervous tissue following injury as there appears to be a significant number of genes in common to the 'transport GO:0006810' and the 'nervous system development (GO:0007399)' clusters. Effectively, the retrograde transport of signalling molecules from the site of the lesion to the nucleus is the primary mechanism for altering the transcriptional activity in the cell of injured neurons to assist with growth and repair whilst the anterograde transport guarantees the supply of growth material to the growing end of the axon.

The gene and term overlap analyses also reveal an association between the 'transport GO:0006810' cluster and the 'cellular component organization and biogenesis (GO:0016043)' cluster from the subcellular process class, which is

only logical given that the transport function is fundamental for the organization and localisation of cellular components.

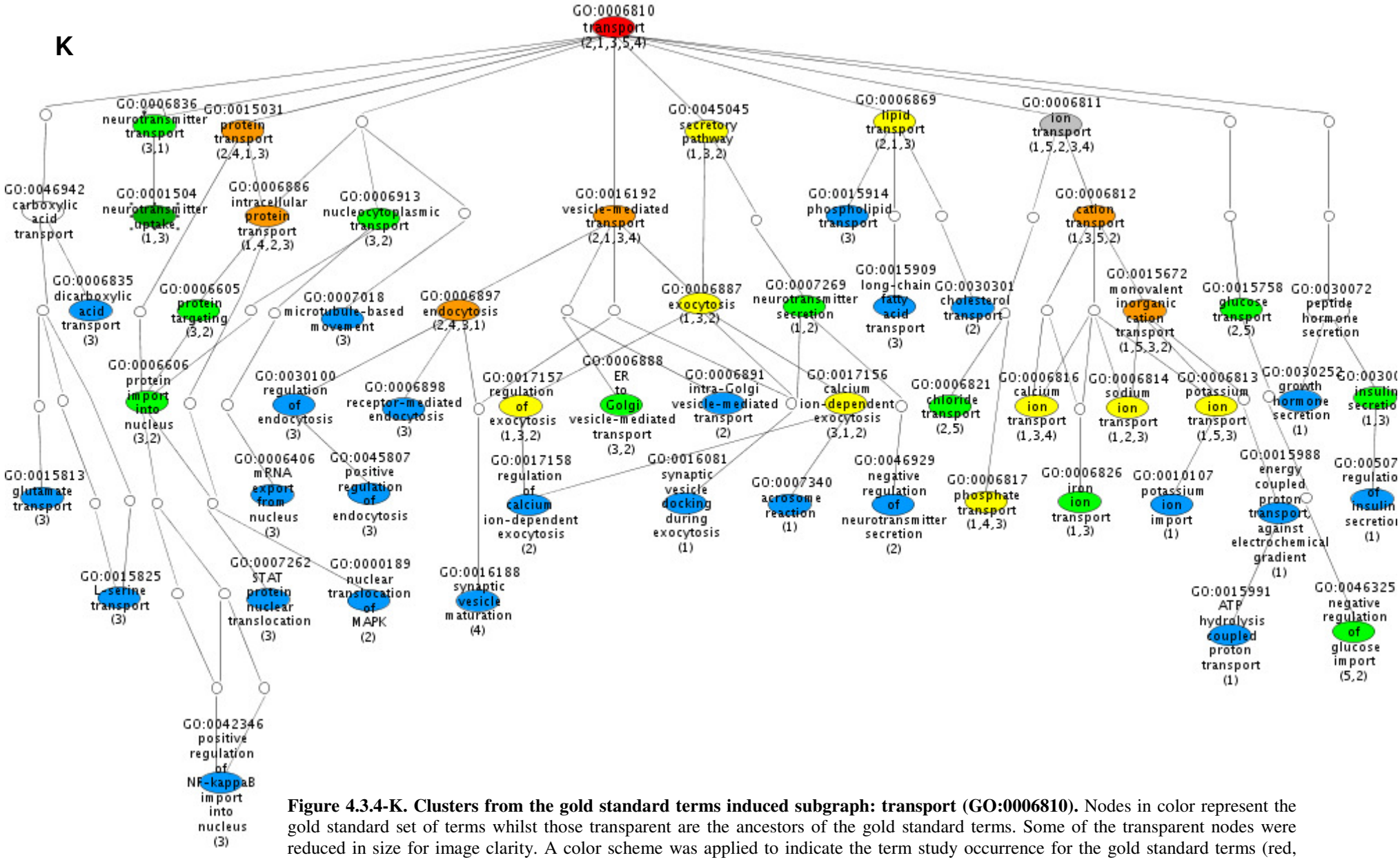


Figure 4.3.4-K. Clusters from the gold standard terms induced subgraph: transport (GO:0006810). Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. Some of the transparent nodes were reduced in size for image clarity. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

The last cluster in the molecular process class is the 'metabolic process (GO:0008152)' cluster. Upon damage to the axon, neurons shift their metabolism to achieve the molecular repertoire that can support the nature of the response to injury. These shifts affect a wide range of biological molecules: lipids, nucleobase and nucleic acid, proteins, amino acids and carbohydrates. From the gene and term overlap analyses (tables 4.3.5 & 4.3.6 respectively), the metabolic function appears to be associated with most higher levels processes including the 'nervous system development (GO:0007399)' process, the 'neurological process (GO:00550877)' and the 'immune response (GO:0002376)' process.

Upregulation of lipid synthesis serves in part to supply the growing axonal membrane with lipid structural constituents in addition to other types of lipids such as steroids and prostaglandins associated with the inflammatory/immune response (Fig 4.3.4-L). At the DNA level, injury results in a net enhancement in transcriptional activity through activation of transcription factors such as NFκB. Furthermore, the DNA replication machinery is also induced to assist with the proliferation of glial and immune cells. Apoptosis on the other hand entails metabolic fragmentation of the DNA (Fig 4.3.4-L).

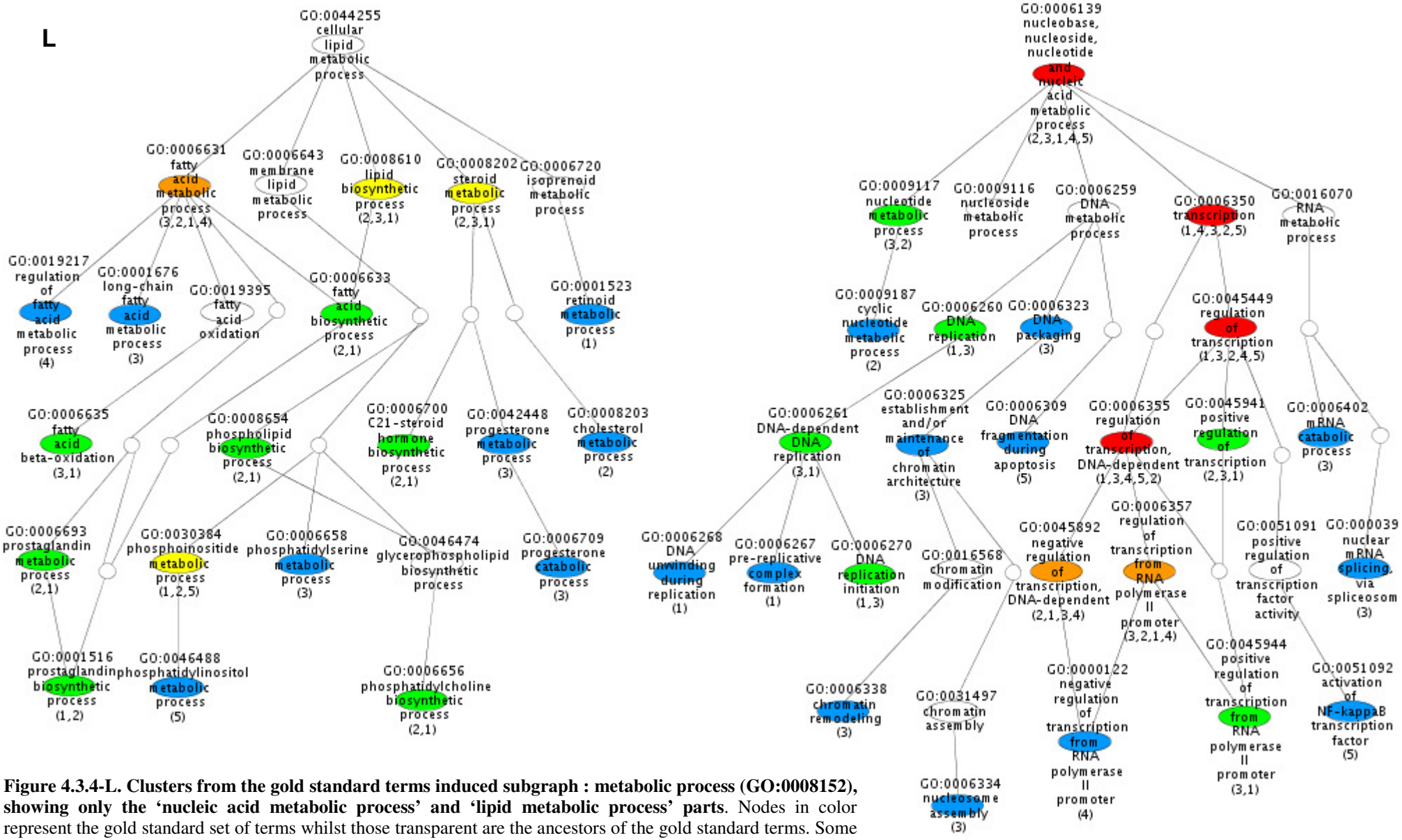


Figure 4.3.4-L. Clusters from the gold standard terms induced subgraph : metabolic process (GO:0008152), showing only the 'nucleic acid metabolic process' and 'lipid metabolic process' parts. Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. Some of the transparent nodes were reduced in size for image clarity. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

Another category of metabolic changes induced upon axonal injury (Fig 4.3.4-M) is that affecting proteins. Growth associated proteins, neurotransmitters and cytokines are all examples of proteins that get overexpressed following injury, in addition to amino acid derivative neurotransmitters (Fig 4.3.4-N). The activity of proteins is modulated by upregulation of posttranslational modification machinery within the cell. An example is the process of phosphorylation that serves to activate key signalling kinases (Fig 4.3.4-M). Furthermore, there are changes to the metabolism of carbohydrates (Fig 4.3.4-N). Such changes are required to support the energy-consuming processes that are induced upon axonal injury, such as the cell cycle as well as the antero-retrograde forms of molecular transport that occur across the proximal part of the axon.

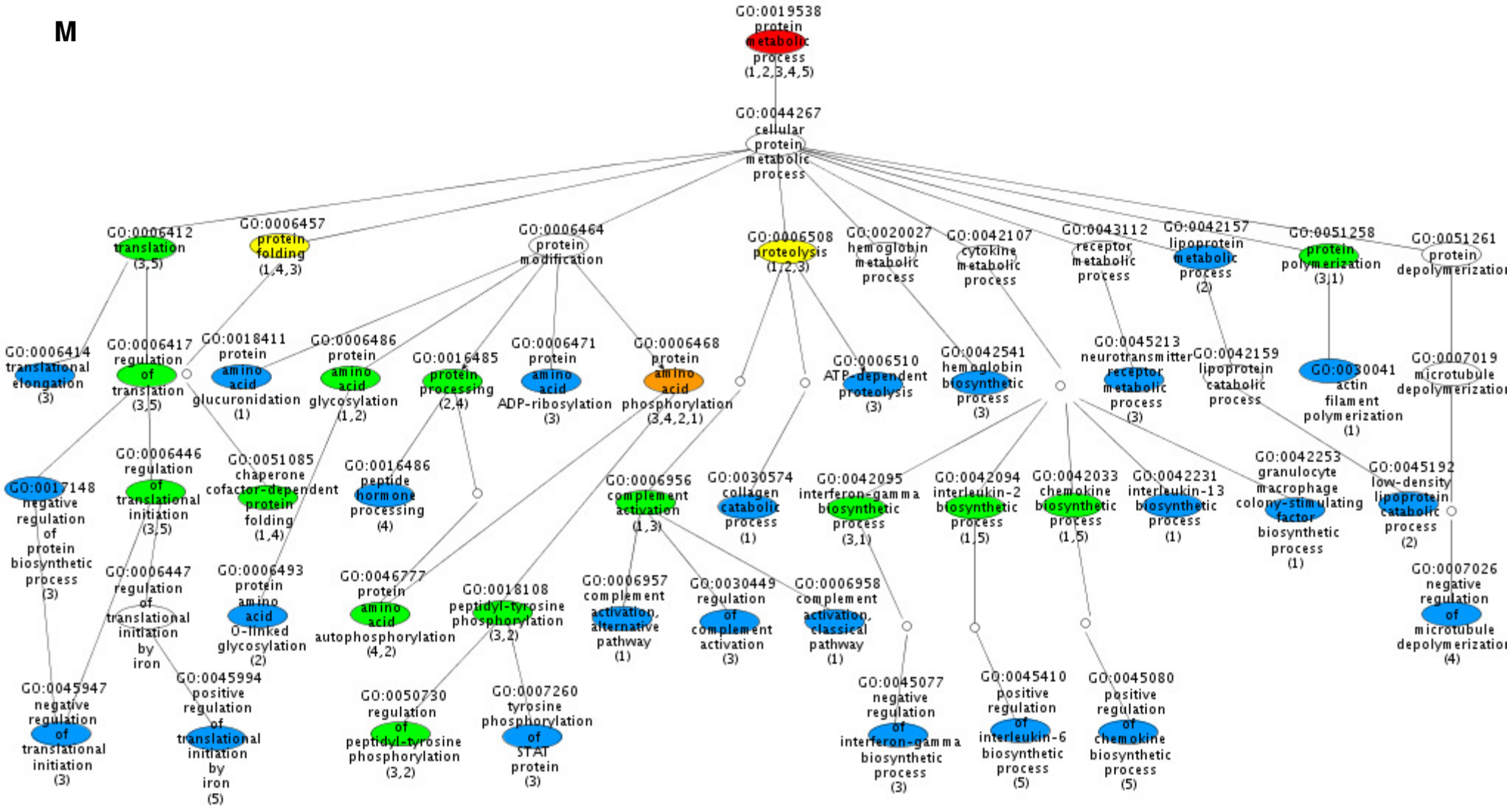


Figure 4.3.4-M. Clusters from the gold standard functional dataset: metabolic process (GO:0008152), showing the 'protein metabolic process' part only. Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. Some of the transparent nodes were reduced in size for image clarity. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) al respectively).

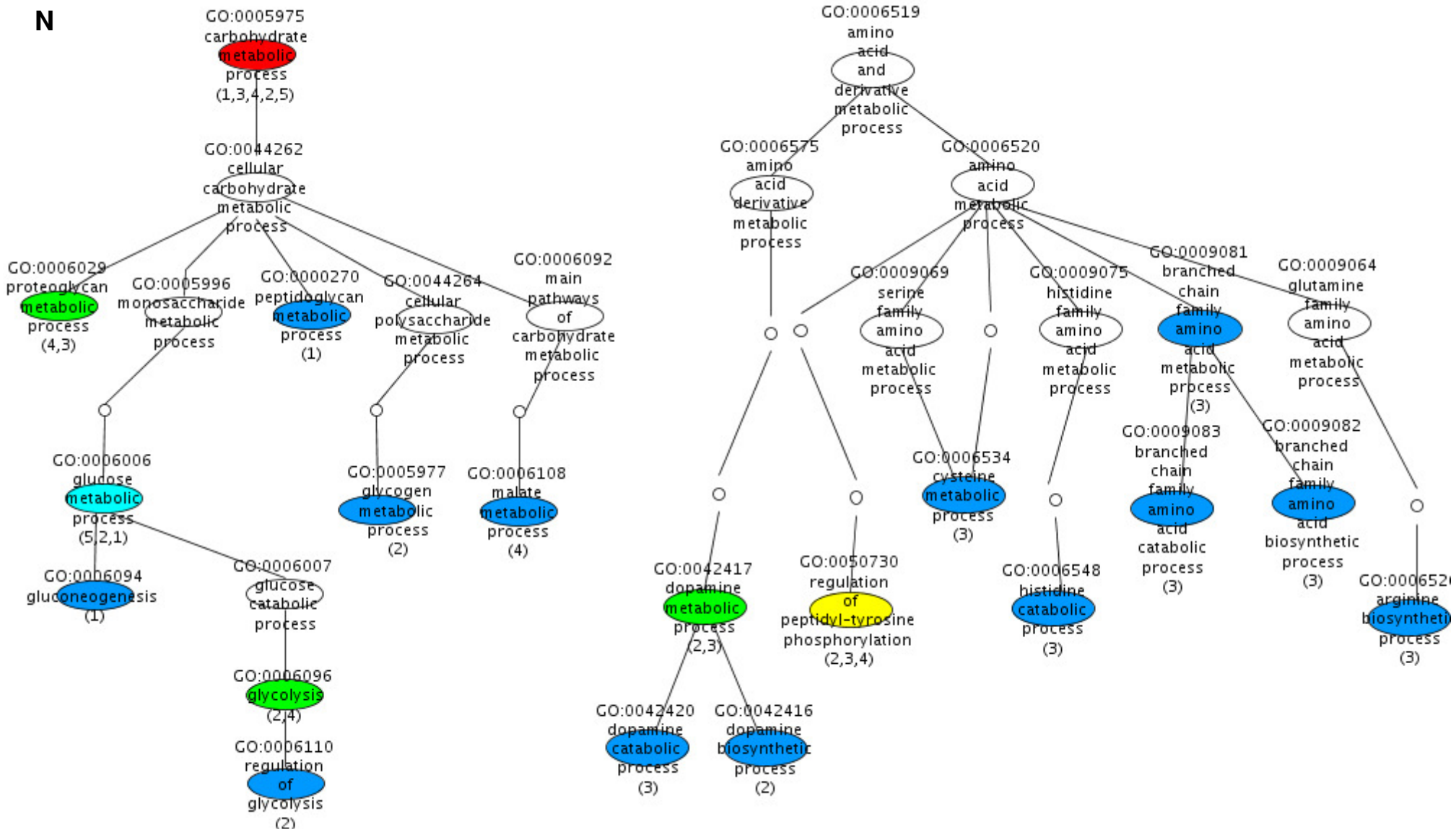


Figure 4.3.4-N. Clusters from the gold standard functional dataset: metabolic process (GO:0008152), showing the 'carbohydrate metabolic process' and the 'amino acid and derivatives metabolic process' parts only. Nodes in color represent the gold standard set of terms whilst those transparent are the ancestors of the gold standard terms. Some of the transparent nodes were reduced in size for image clarity. A color scheme was applied to indicate the term study occurrence for the gold standard terms (red, orange, yellow, green, blue for 5,4,3,2,1 study counts respectively). The term name and accession are given on each node. Nodes in color feature additionally the study ID where the term or any of its progeny appear (study ID 1,2,3,4,5 correspond to (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2002; Xiao et al., 2002) and the literature survey from (Costigan et al., 2002) respectively).

The above discussion of functional links between the GOC clusters has certainly not captured the full extent of the functional intermingling between the varying biological processes induced upon damage to the peripheral nerve. However, it does have the benefit of hinting at the significance of each cluster of processes with respect to the overall response. A summary of the relationships between the GOC clusters from the designated classes of biological processes, outlined in table 4.3.4, is presented in Figure 4.3.5.

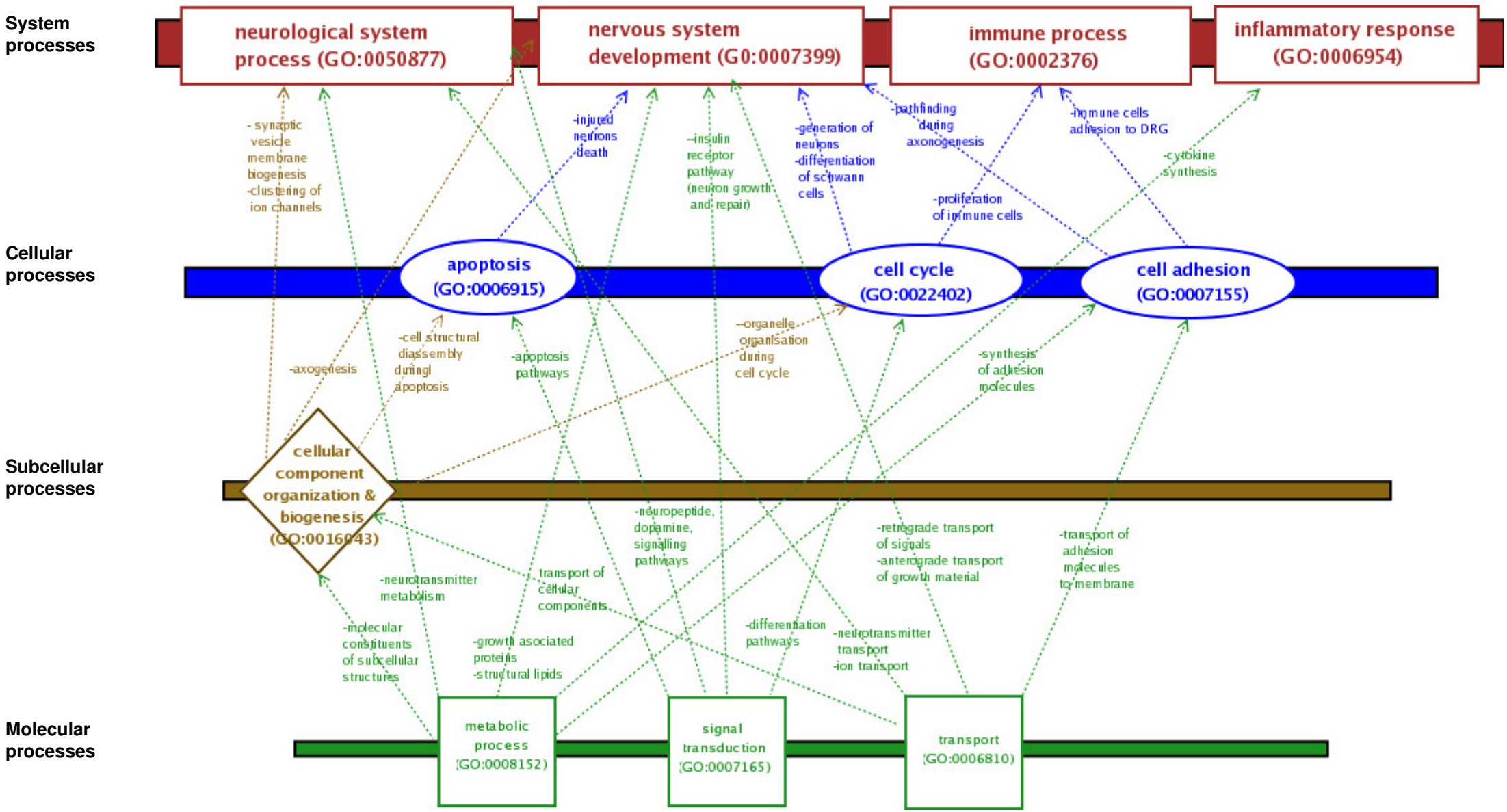


Figure 4.3.5. Relationships between GOC clusters from the varying biological process classes.

4.4. Conclusion

To summarise, this chapter used the GO framework to capture knowledge of functions that show enrichment following peripheral nerve injury. This was achieved by deriving the set of GO term annotations of genes that were reported to show a change in expression following injury to the peripheral nerve in a number of published studies. This set of terms, which we refer to as the *gold standard set of terms*, is of particular importance to this work, as it will be used in chapter VI to evaluate the results from functional analysis of a spinal nerve transection expression dataset by the LPC.

Because genes are often annotated with a number of GO terms in order to capture the full extent of their functions, stripping terms of their genes has the drawback of flattening the association between them that arise in the context of gene function. In this work and in order to reveal the biological significance of the gold standard terms, originally derived from candidate genes from published studies, we used the gene and term overlap analyses to trace the associations between clusters of these gold standard terms.

From a biological perspective, it was interesting to note how the reprogramming of the transcriptional activity within the DRG tissue, following

injury to the nerve, affects a complex network of functions, some of which are non-neuronal in origin. This is because the DRG tissue compromises a number of different cell types: neurons, glial and immune cells. This suggests that using microarray expression profiling technology with animal models of neuropathy, in the traditional sense, to study particular pathological aspects such as pain is rather limited. But this can be optimised with intelligent experimental design and powerful datamining approaches to allow the most relevant information to be obtained.

4.5. Appendices

4.5.1. The gene ontology categoriser: GOC

In this work, we used the GOC algorithm to categorise the gold standard terms by the broad sense of their encapsulated functions. In the context of GO, this translates into finding the most appropriate parent term for a subset of functionally related gold standard terms that preserves the essence of their functions. In effect, the process of ontology term categorisation is governed by two opposing criteria: specificity and coverage. For instance; considering the model graph shown below (Fig 4.5.1): the parent term 'A' is the most representative of all query terms (shown in green), yet semantically it is less specialised than parent 'F', which in turn covers less query terms than its predecessor 'A'.

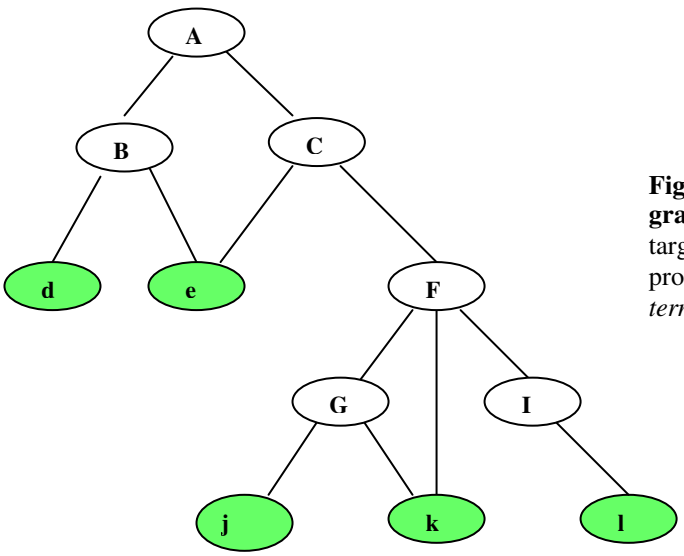


Figure 4.5.1. A model ontology graph. Nodes d, e, j, k, l are the targets for the categorisation process; in other words, the *query terms.*

The GOC mathematical model for ontology term categorisation captures this interplay between coverage and specificity and is here described. During the categorisation process, for any given parent term, GOC measures the distance to each of the query child terms. In its simplest form, the distance is taken to be equal to the number of edges connecting the parent to the child term via the shortest path. The distance measure, given by the symbol δ in the GOC equation (shown below), is taken by GOC as an indication of the parent specificity as the more distal the parent is from query child terms, the closest it gets to the root hence the least specific it becomes. As such, δ is inversely correlated with specificity, which explains the use of the reciprocal of δ in the GOC equation:

$$S(p) = \sum_{c' \in C} 1/(\delta^{2^s}(c', p) + 1) \quad (1)$$

Essentially, for a given parent term p and the set of query terms it subsumes C, a score S(p) is given based on the sum of the reciprocal of δ from each query term c' belonging to the set C raised to power 2^s; where s is a user-defined parameter. The significance of power s is that by altering the magnitude of the specificity indicator δ, it provides a mechanism to adjust the balance between specificity and coverage.

Consider the case of parent 'F' and the more distal parent 'A' from the model graph, the common query child 'j' is further away from 'A' than 'F' (and so are 'k' and 'l'); hence, $\delta_{(A,\ j)}$ is greater than $\delta_{(F,\ j)}$. A positive power s inflates $\delta_{(A,\ j)}$ further with respect to $\delta_{(F,\ j)}$ causing the reciprocal $1/\delta_{(A,j)}{}^{s}$ to be even smaller than $1/\delta_{(F,j)}{}^{s}$ the larger s gets. As such, a positive power s has the effect of amplifying variation in δ; the effect is more dramatic when using 2^s, as featured in the GOC equation.

A negative value of s has the opposite effect in that it acts to suppress the differences in δ. This is because raising δ to 2^s where s is negative, is mathematically equivalent to taking the $(2^{|s|})$th root of δ where $|s|$ is the absolute value of s. Contrary to power transformation, a root transformation causes data to shrink, reducing larger data to greater extents; thereby minimising the gap between large and small values. As such, $\delta_{(A,j)}{}^{2^{s}}$ is closer to $\delta_{(F,j)}{}^{2^{s}}$ the more negative is the value of s.

Just how the power transformation of δ serves to adjust the balance between coverage and specificity needs further clarification. Going back to the case of the general parent 'A' and the more specialised 'F' from the model graph, the overall scores for both parents are the following respectively:

$$S(A) = \sum_{c \in \{d,e,j,k,l\}} 1/(\delta^{2^{s}}(c, A) + 1)$$

$$S(F) = \sum_{c \in \{j,k,l\}} 1/\delta^{2^{s}}(c, F) + 1)$$

The more negative is s, the smaller are the $1/\delta^{2^{s}}$ from any parent-child pair causing the sum; in other words, the final score to become increasingly governed by the number of individual $1/\delta^{2^{s}}$ contributions from query child terms; rather to the advantage of parent 'A' as it subsumes more query terms than 'F'. As such, negative s emphasizes coverage. On the other hand, the more positive is s, the larger the $1/\delta^{2^{s}}$ from common children 'j', 'k' and 'l' for parent 'F' than 'A'; ultimately, overcoming the additional contributions from children 'd' and 'e' exclusive to 'A' thereby causing the score from 'F' to rise above that from 'A'. As such, positive s emphasizes specificity.

Indeed, in table 4.5.1, we see the actual GOC scores for parents 'A', 'C' and 'F' from GOC analysis of the model graph shown above for a range of values s = {-1,0,1,2}. The very general parent 'A' scores the best when s is set to a negative value. Moving to positive values of s, there is a shift towards more specialised parents beginning by 'C' at s = 1 and finishing with the most specific parent 'F' at the highest value s = 3.

Table 4.5.1. Highlighting the different clustering results by GOC while varying 's'. Results obtained from running GOC on the model graph on Fig 4.5.1

	s = -1	s =0	S =1	s =2
'A'	1.84	1.27	0.61	0.14
'C'	1.63	1.30	0.9	0.58
'F'	1.32	1.16	0.9	0.61

The top GOC score from each round is shown in red

CHAPTER V: A GO SEMANTIC SIMILARITY METRIC TO MEASURE THE SIMILARITY BETWEEN GO TERMS

5.1. Aim of the chapter

This chapter aims to introduce some of the aspects of the methodology used in the following chapter to validate the GO functions found enriched in a spinal nerve transection (SNT) microarray dataset against the set of gold standard terms discussed in the previous chapter. In particular, the chapter explores ways for comparing these two sets of GO functions by means of deriving a measure that expresses the semantic similarity between terms in the GO graph.

The outline of the chapter is as follows: First, a review of existing theories for measuring the semantic similarity between GO terms is presented. Then, we introduce a novel approach, developed as part of this work, that expresses the similarity level between two GO terms based on the ontological 'records' of their immediate common ancestor. The last part of the chapter evaluates the

performance of the proposed method against one widely used GO semantic similarity approach in the literature.

5.2. Introduction

The GO ontology is a mesh of interconnected terms representing biological functions organized into a hierarchical structure, similar to a taxonomy, whereby each term is in one or more parent-child relationships to other terms in the ontology. In GO, parent terms provide an abstraction of the meaning of their child terms. For any given term in the ontology, a series of increasingly general abstractions of the term's semantics is reflected by consecutively occurring ancestor terms on the paths leading from the term to the root of the ontology. Such decomposition of function semantics by GO offers the opportunity to capture similarities between the various functional terms in a measurable format.

The notion of semantic similarity was originally developed for taxonomies. For example, the earliest studies looking at quantifying conceptual semantic similarity were mostly targeted at the WordNet (Fellbaum, 1998), which is a lexical taxonomy for the English language. Two major approaches for estimating semantic similarity were soon presented, one that explored the hierarchical structure of the taxonomy and one based on the idea of information content.

Rada and colleagues presented one of the first instances of semantic similarity measures based on the structure of a medical taxonomy (Rada and Bicknell, 1989). In their work, the similarity between two terms was based on the distance in edges linking them along the shortest path, where the smaller the distance the higher the similarity. A major drawback from this approach is that it makes the assumption that edges denote equal semantic distances, which seems to be a poor assumption with taxonomies. Resnik and colleagues pointed out this problem and proposed an alternative method to quantify semantic similarity (Resnik, 1995). The new approach was based on the concept of information content whereby the usage frequency of a term's semantics is evaluated within a corpus, which implies counting the occurrences of the term and its children. The ratio of this occurrence value to the total number of occurrences of all terms in the taxonomy indicates the term's probability of occurrence. The term's information content value is defined as the negative log of its probability of occurrence value ($-\log P$).

The Resnik conceptualisation of information content is intuitive in that frequent terms with high probability of occurrence feature small information content values, capturing the fact that they are least informative. Also, it logically follows the structure of the taxonomy in that the further down in the tree the higher the information content value; owing to the fact that the

probability of occurrence value from children terms can only be smaller or equal to that from their parents.

In the work by Resnik, the idea of using information content to measure semantic similarity is based on the assumption that two concepts are most similar if they share much information between them, which is essentially the information content of their immediate common parent. Thus, given two terms c1 and c2, the similarity between them is given by the information content of the lowest common parent C0 that subsumes them both:

$$\text{sim}(c1,c2) = -\log P(C0) \quad (1)$$

In 1998, Lin suggested an alternative for incorporating information content into a semantic similarity metric (Lin, 1998). The new theory was that the extent of similarity between two concepts is better evaluated when considering the differences between them. In the formal model by Lin, the semantic similarity measure is defined as the ratio between the information in common to the two concepts (which expresses the similarity between them) and the bulk of information needed to describe each of them as a whole (which accounts for the differences in addition to the similarities in their semantics). In a taxonomy domain, it is defined as the ratio between the information

content of the lowest common parent and the sum of information content values from the two terms being compared:

$$\text{sim}(c1,c2) = \frac{2\log P(C0)}{\log P(c1) + \log P(c2)} \quad (2)$$

One further contribution to the theory of semantic similarity was made by Jiang (Jiang and Conrath, 1997). The Jiang model followed a combinatorial approach that uses both information content as well as path distances within the taxonomy structure. The idea was that both approaches have strengths and weaknesses and could consolidate each other if used in a complementary fashion. Thus, the information content approach, although theoretically plausible, shows a strong dependency on the chosen corpus and may display poor sensitivity at the very bottom of the taxonomy tree. This is because highly specialised terms may not occur in the corpus, which implies that an information content value may not possibly be derived for such terms. The distance approach on the other hand is intuitive and is equally applicable to all nodes in the tree structure, though it is sensitive to the problem of varying edge weights.

Essentially, the Jiang method is an optimisation of the shortest path distance metric whereby a mechanism is devised to adjust for variable edge weight. Thus, instead of treating edges homogenously by simply adding their number along the path, the Jiang method assigns a weight to each edge on the path based on the difference in information content values of the parent and child node linked by that edge (Fig 5.2.1). This is rather intuitive in the sense that if a parent and a child are semantically close to each other, the difference in their information content values should be small. The overall distance between two concepts is given as the summation of edge weights along the shortest path separating the corresponding nodes in the tree structure (illustrated in Fig 5.2.1); which after mathematical simplification is reduced to equation (3). The smaller the distance, the higher the similarity between the terms.

$$\text{Dist}(c1,c2) = 2 * \log P(C0) - (\log P(c1) + \log P(c2)) \quad (3)$$

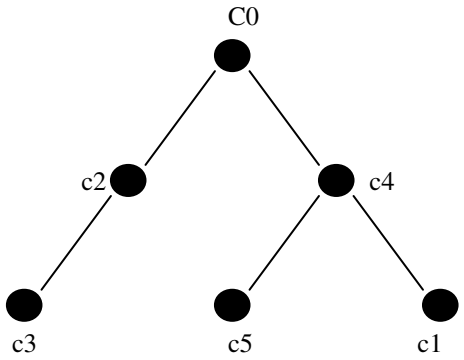


Fig 5.2.1. Illustration of the Jiang similarity metric. The weight of an edge is defined as the difference between the information content values of the parent term and the child term connected by the edge. Thus, the weight from edge (c1->c4) is given as (-log P(c1) + log P(c4)). On the other hand, the distance between two terms is given as the sum of edge weights along the shortest path. Thus, for the pair of terms (c1 & c2) featuring the shortest path (c1->c4->C0->c2), dist(c1,c2) = (-log P(c1) + log P(c4)) + (-log P(c4) + log P(C0)) + (-log P(c2) + log P(C0)), which after simplification reduces to equation (3) from above.

One common problem to the Lin and Jiang models is that the similarity score between two terms is based on assessing their information content values relative to that from their common ancestor while taking no account of the location of the common ancestor on the graph. This could be problematic at close proximity to the root. This is because at high levels in the GO graph, terms have general meanings and hence small information content values. This implies that the information content values from a pair of terms and their common ancestor at this level in the GO graph are equally small and may not significantly differ from each other, which results in an artificially high similarity value. This is rather misleading as semantically broad terms cannot possibly be similar to each other.

Lord and colleagues were the first to use the information content approach to measure semantic similarity between GO terms (Lord, 2003). In their study, the Resnik, Lin and Jiang metrics were used to compute semantic similarities between GO terms, based on the occurrence frequency of individual GO terms in the SwissProt database. To validate the suitability of these similarity measures for GO, Lord explored one important tenet of biology, which is the association between sequence similarity and functional conservation. Thus, for all pairs of proteins from the SwissProt database, sequence similarity scores were obtained using BLAST. As for the functional similarity scores, because

proteins are often annotated with a number of terms, for each protein pair, an average similarity score was derived from all possible pairs of terms from the two proteins. Importantly, Lord concluded that the Resnik's functional similarity scores were the most correlated with the sequence similarity scores for pairs of proteins, though the correlation level was no higher than 0.577.

In the recent evaluation study by Pesquita and colleagues (Pesquita et al., 2008), the performance of the Resnik, Lin and Jiang similarity measures was re-evaluated, similarly by examining the correlation with sequence similarity scores from pairs of proteins. Importantly, the study explored different ways of deriving a unique similarity value for each pair of proteins; including taking the average similarity value from all combinations of terms from the two proteins (Lord, 2003) as well as considering the maximal similarity value (Sevilla et al., 2005). A third approach, also known as the *best-match average* approach, consisted of pairing each term from the first protein with its best match from the second protein and vice versa, then deriving an average similarity value (Couto et al., 2007; Schlicker et al., 2006).

In agreement with the conclusion by Lord (Lord, 2003), the Resnik measure proved the best, in particular when used with the best-match average summary approach for pairs of proteins. The same study investigated two other

similarity measures that have the distinctive feature of directly calculating a unique similarity value for a given pair of proteins. The first of these measures is the *simUI* by Gentleman (Gentleman, 2005) where for any given pair of proteins and associated terms including ancestral terms, the similarity value is simply the ratio between the number of terms in common and the overall number of terms from both proteins. This novel approach was extended by Pesquita to include the information content values of individual terms (Pesquita et al., 2008). The new method or *simGIC,* is based on calculating the ratio between the sum of information content values for the terms in common to the sum of information content values from all terms from both proteins. Pesquita reports an improvement using simGIC in comparison to Resnik/best-match average approach.

The use of information content has certainly improved our ability to measure semantic similarity between concepts including GO functional terms. Though, one popular view is that the ontology structure is equally relevant and should also be considered. Beyond using path distances, more successful methods have recently emerged that deploy the structure of the ontology to measure semantic similarity between GO terms. One such method was proposed by Wang and colleagues (Wang et al., 2007). The method defines the similarity between two terms by the extent of contribution of common ancestral terms to

their semantics whilst the contribution of an ancestor term to a child term semantics is defined as the product of weights of edges in the path to the child. Edge weights are generally uniform throughout the GO graph but are slightly higher for 'is a' than 'part of' relationships. In cases where there are multiple paths to the child, the maximum score from all paths is taken to indicate the contribution of the ancestor to the child semantics.

Thus, taking the example of the child 'GO:0043231' from the GO cellular component subgraph shown in Figure 5.2.2, choosing an edge weight value of 0.8 for 'is a' (solid lines) versus 0.6 for 'part of' (dashed lines) types of relations and measuring the semantic contributions from parents 'GO:0043229' and 'GO:0005623', we find that the former scores 0.8 whereas the latter scores 0.288 = (0.8*0.6*0.6); conforming to the fact that the latter is an earlier ancestor and hence contributes less to the semantics of the child. Thus, this model features the basic idea that the more edges separating the child term from its ancestor, the less the contribution of the ancestor to the child's semantics.

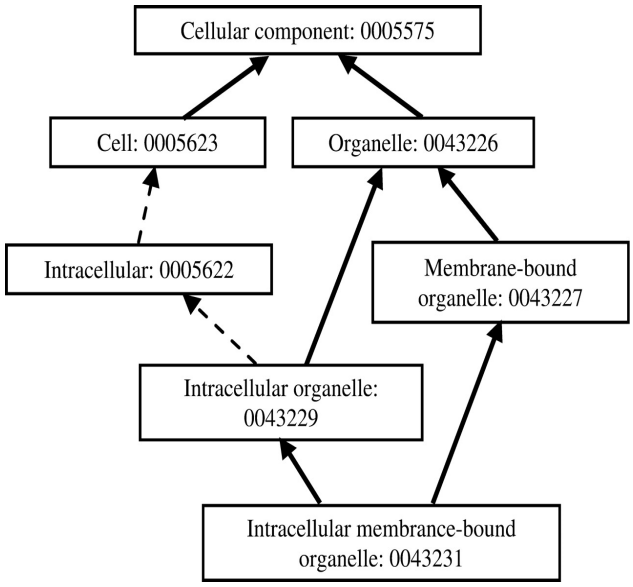


Figure 5.2.2. A model GO subgraph. Edges in solid and dashed lines represent 'is a' and 'part of' relationships respectively.

In the Wang method, the similarity between two terms is given as the ratio between the sum of semantic contributions from their common ancestors to the sum of semantic contributions from all ancestors of both term. This captures the logic that the more representative are the common ancestors of the terms' semantics, the more similar the terms are. Thus, given two terms A and B, and their set of ancestor terms T_A and T_B respectively, the similarity between them is given as:

$$sim_{(A,B)} = \frac{\sum_{(t \in T_A \cap T_B)} (S_A(t) + S_B(t))}{\sum_{(t \in T_A)} S_A(t) + \sum_{(t \in T_b)} S_B(t)} \quad (4)$$

where $S_A(t)$ denotes the contribution of ancestor 't' to the semantics of child A.

To adapt their similarity metric to proteins, Wang and colleagues adopted the best-match average approach. Using pathway gene annotations, they demonstrated how their similarity measure correlated better than Resnik's with human perception of the extent of functional association between the varying reactions in a pathway. For instance, genes mediating the same reaction in a pathway are expected to be annotated with more similar terms than those taking part in parallel or alternative reactions in the pathway.

The advantage of the Wang similarity approach is that it takes account of the terms strength of relationships with their ancestors. Thus, unlike the Resnik method where different pairs of terms would score an identical similarity value if they share the same most specialised ancestor, the Wang score is sensitive to the location of each term in the pair on the GO graph. One limitation to the Wang approach, also common to the Lin and Jiang methods, is the artificially high similarity values at close proximity from the root. This

is because at high levels in the GO graph, terms are usually subsumed by a small number of fairly general ancestors. Thus, for any pair of terms, common ancestors (although with broad semantics) may appear to contribute significantly to the semantics of the terms relative to the remaining contributions by few other general ancestors unique to each term (equation 4). Perhaps, one extreme example is that of terms ‘Organelle GO:0043226’ and ‘Cell GO:0005623’ from the model graph shown in Figure 5.2.2. The two terms have a unique common ancestor ‘GO:0005575’, which is also the only ancestor for each of them. As such, their similarity value using the Wang metric (equation 4) would be equal to 1, which is the highest possible value. Thus, although these two terms are clearly distinct from each other, they turn out to be highly similar according to the Wang metric.

5.3. The GOTrim similarity measure

5.3.1. Theoretical basis

In this work, we propose a new strategy that explores the GO structure to quantify the semantic similarity between GO terms. Our method is fundamentally similar to the Resnik approach in that the similarity between two terms is indicated by the level of specificity of their common most specialised ancestor. Though, instead of using information content as an indicator of the semantic specificity of the common ancestor, we derive such value from the structure of the ontology.

Our idea for measuring specificity is based on the fact that the semantic granularity of an ontology term is the result of a gradual semantic specialisation process effected by the chain of consecutive ancestor terms on the path(s) from the root to the term in the ontology. As such, the specificity of a term can be estimated by combining the amount of semantic specialisation contributed by each of its ancestor terms.

We define the extent of semantic specialisation by an ancestor term relative to the total semantic space captured in the whole of the ontology as the ratio of

the ancestor progeny terms to the total number of terms in the ontology. This follows the logic that at each ancestor term, the whole range of biological functions in GO is restricted to the semantics of the function of the ancestor term, which include functional subtypes expressed by all of its descendent terms. Moreover, along any given path from the root, moving from one ancestor to the next one down features the selection of increasingly smaller subsets of progeny terms, which allows our method to capture the increase in semantic specialisation by each consecutive ancestor in turn along the path. This is illustrated in the diagram in Figure 5.3.1.

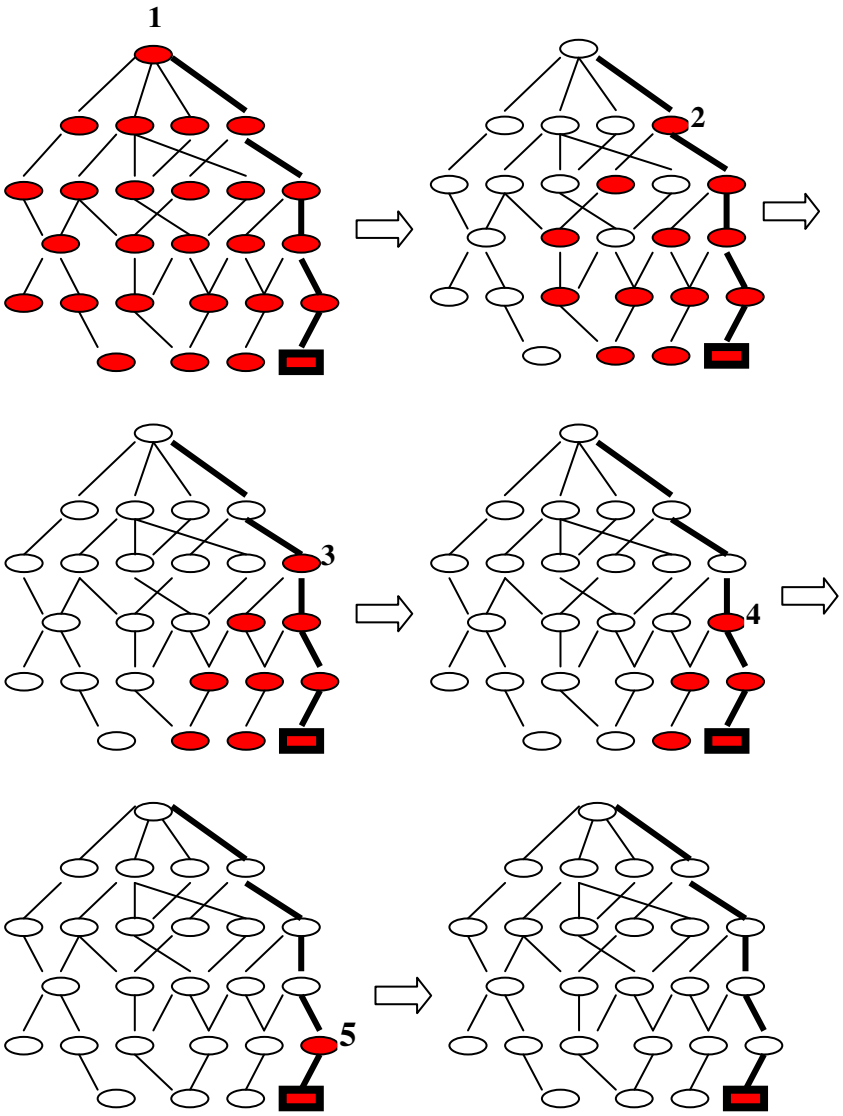


Figure 5.3.1. Diagram illustrating the process of gradual semantic specialisation effected by the chain of ancestral terms of a given child term. The latter is indicated by a black rectangular box. The path from the root to the child term is indicated with bold edges. Moving from one graph to the next corresponds to selecting consecutive ancestor terms along the path and their progenies (in red). Progressively smaller subsets of progeny terms appear for sequential ancestors reflecting the progressive increase in semantic specialisation along the path. Numbers indicate the order of ancestor terms beginning by the root.

The specificity of a given term is then given as the aggregate of the extent of semantic specialisation at each ancestor term in the path from the root to the term. Thus, given a term t and the set of all its ancestors (together referred to as A_t), the specificity *spf* of term t is the cube root of the sum of the reciprocals of the ratios of the number of progeny terms to the total number of ontology terms for all ancestral terms in set A_t:

$$spf_t = \left(\sum_{(a_t \in A_t)} (1/(n_{at}/N)) \right)^{1/3} \quad (5)$$

where n_{at} is the number of progeny terms for ancestor a_t and N is the total number of terms in GO. Ratios are inversed so that ancestors furthest from the root featuring smaller subsets of progeny terms contribute more weight to the final specificity score. The cube root transformation is applied to shrink the overall sum so that the specificity scores from all terms cover a confined range of values (we chose to use a root transformation as oppose to a log transformation because the former is more linear than the latter and appeared to yield similarity values for randomly selected portions of the GO graph that were most intuitive).

Our proposed semantic specificity metric from equation (5) features a number of important characteristics. First, by combining the ratios from all ancestor

terms by taking their sum, the final score is influenced by the individual semantic specialisation 'merits' of ancestor terms and not solemnly governed by the number of these ancestor terms, as would have been the case with taking the product of these ratios. Thus, analysing the relationship between specificity scores calculated, according to equation 5, for all terms from the GO biological process ontology and the number of their ancestor terms (Fig 5.3.2), we find indeed that terms with a similar number of ancestors may show a broad range of specificity scores, as indicated by the scatter.

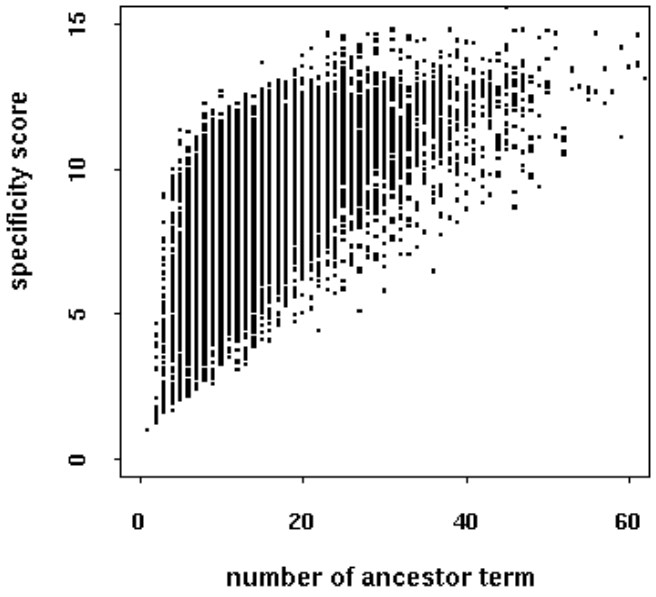


Figure 5.3.2. The relationship between the specificity scores of terms from the GO biological process ontology and the number of their ancestor terms. The scatter indicates that the specificity scores are not solely determined by the number of ancestor terms.

Another important feature of our proposed semantic specificity metric is that the specificity score from any child term in the ontology can only be higher than the score from any of its parent terms. This is because the child term would feature at

least one more ancestor than its parents. Furthermore, our specificity metric has the benefit of allowing specificity scores from all terms in the ontology to be compared in a meaningful way. This is so because the specificity scores for all terms are derived from measuring the extent of semantic specialisation by ancestor terms relative to the same point of reference; that is the overall semantic space given by the total number of terms in the ontology.

To further illustrate the GO semantic specificity metric proposed in this work, we present a snapshot of a portion of the GO graph and label the terms therein with their calculated specificity scores (Fig 5.3.3). As expected, along any given path, the specificity score increases the further we get from the root, reaching a maximum value of 9.39 at the leaf term ‘axonogenesis (GO:0007409)’ at the bottom of the graph.

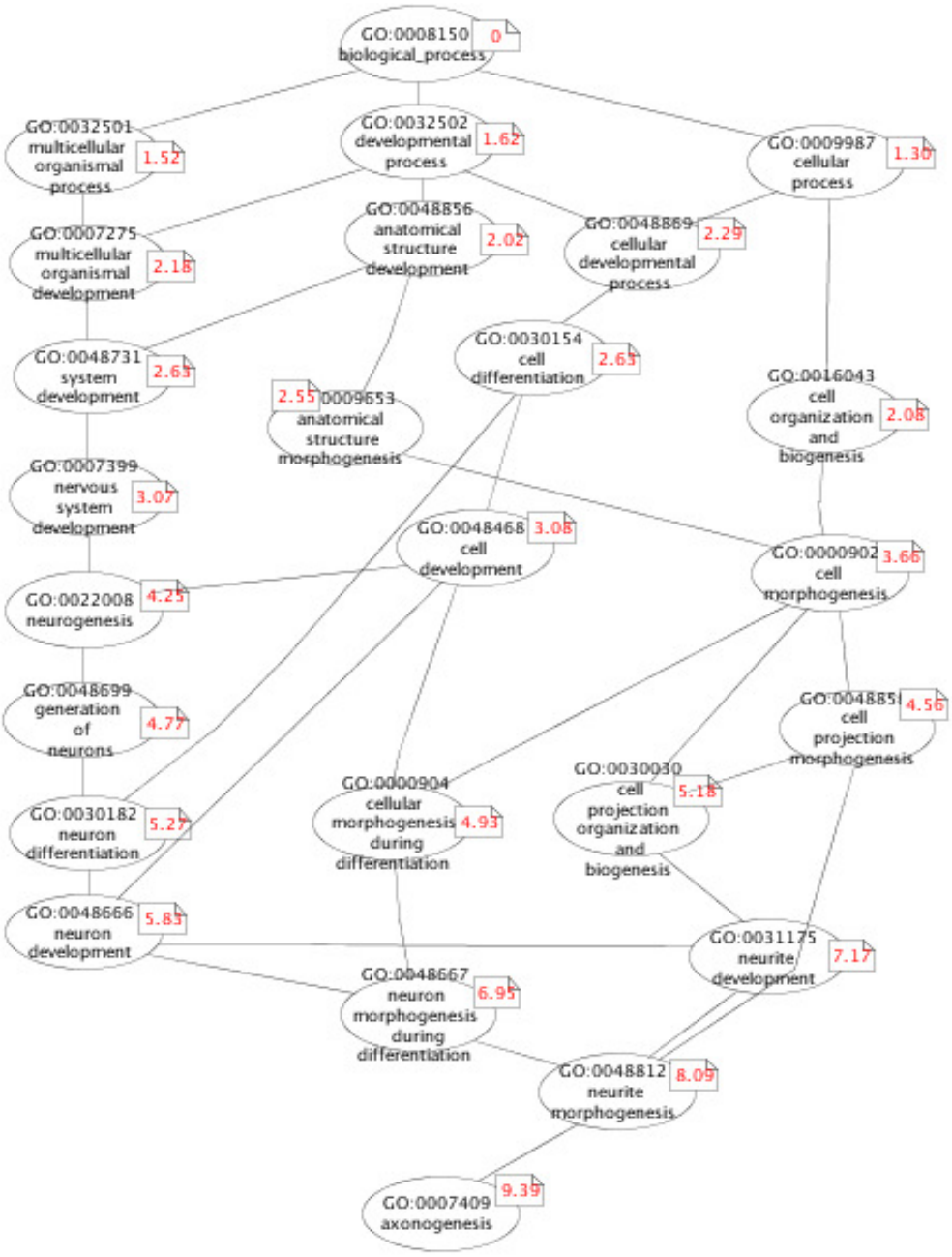


Figure 5.3.3. A portion of the GO graph featuring the specificity scores of the terms attached as labels to the nodes representing the terms.

We now explain how the proposed specificity scores are used to derive the semantic similarity between GO terms. As previously mentioned and rather similar to Resnik's, we define the similarity between two terms as the maximal specificity value from their common ancestors. Thus, given two terms *a* and *b,* the similarity between them *Sim* is the specificity score *Spf* of their common ancestor *C:*

$$Sim_{(a,b)} = Spf(C) \quad (6)$$

where *C* is the common ancestor with the highest specificity score.

We shall refer to the proposed similarity measure as the *GOTrim* similarity measure since the specificity of the common ancestor is derived on the basis of accumulating the extents of successive trimming of GO by the set of predecessor terms of the common ancestor.

5.3.2. Evaluation of the GOTrim method

To evaluate the performance of the GOTrim similarity method developed as part of this work, we compared it to the Resnik method, which formed the

current ‘state of the art’ method at the time the work was being performed. This follows the benchmarking strategy adopted by methods developed subsequent to Resnik, such as the more highly performing Wang and the simGIC similarity methods outlined in the introduction section.

Thus, we compared the semantic similarity scores obtained with each method for pairs of GO biological process terms. However, since with both methods, the similarity score for a pair of terms is derived on the basis of the specificity score of the terms’ immediate common ancestor, we first compared the specificity scores from all terms by each method. To obtain the Resnik information content-based specificity scores, we used the yeast genome database as the body of information and measured the frequency of individual GO biological process terms associated with the gene entries in the database. This frequency value was transformed into an information content (IC) value via a log transformation, as described in the introduction section. Although a log2 transformation was used instead of the natural log to make sure that the Resnik (IC) and GOTrim specificity scores span similar ranges of values and may hence be compared against each other.

The result from comparing the Resnik (IC) and GOTrim specificity scores is shown on Figure 5.3.4-a. The regression line suggests that two scores are

reasonably correlated in the sense that they tend to increase together. However, the fact that the regression line appears to be shifted to the left indicates that the IC specificity scores are on average higher than their GOTrim counterparts. This is also evident from examining their distributions (Fig 5.3.4-C&D), where there appears to be a strong skewing in the distribution of the IC specificity scores towards higher values.

Next, the similarity scores for pairs of terms were obtained with each method. The slightly higher specificity scores by the Resnik method result in similarity scores that are occasionally higher than those obtained with the GOTrim metric, as indicated by the scatter in Figure 5.3.4-b. Although, the regression analysis suggests that the bias is less dramatic with the pairwise similarity scores than with the specificity scores calculated for individual terms using this approach.

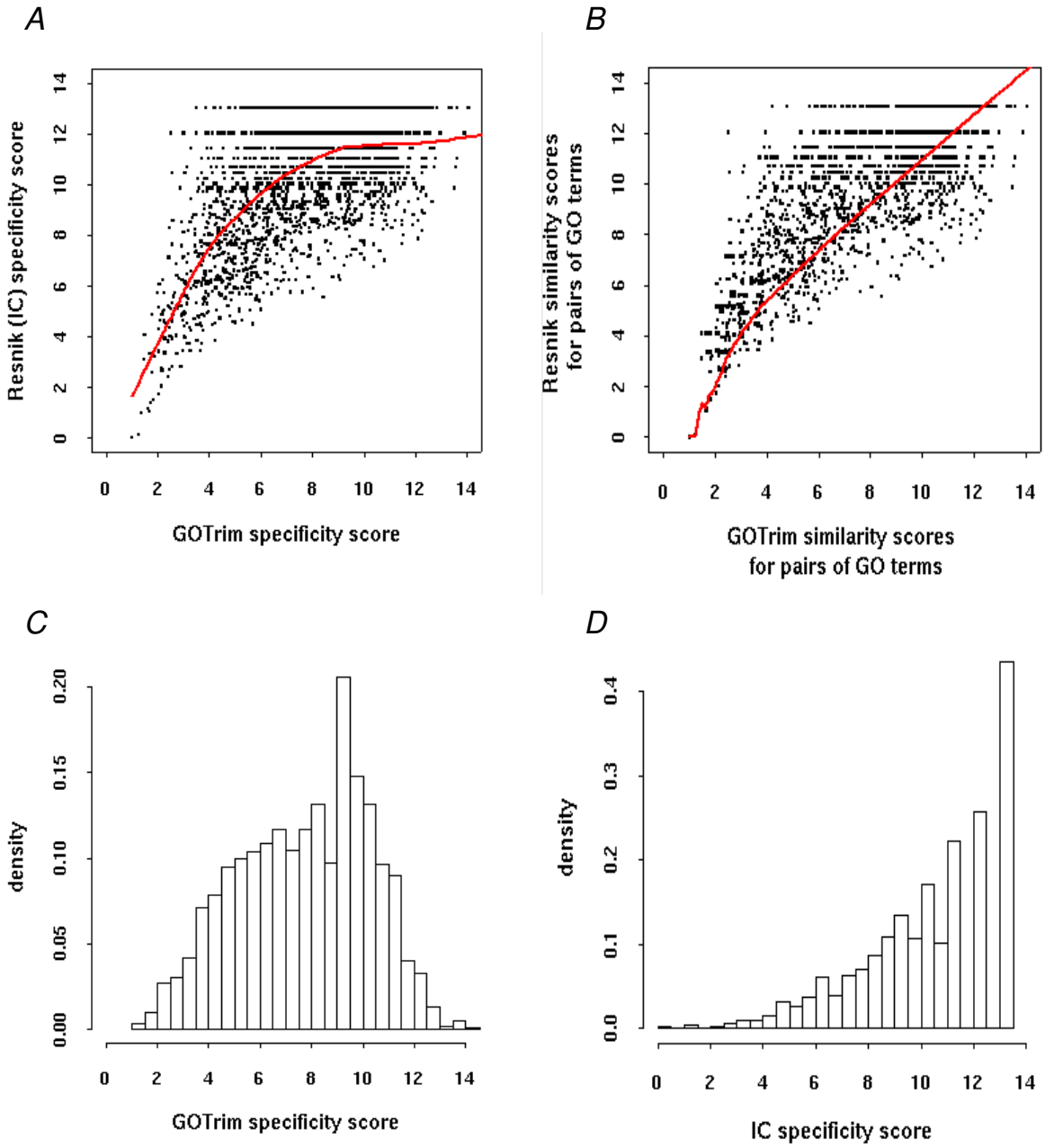


Figure 5.3.4. Comparison of Resnik (IC) and GOTrim methods. (A) A scatter plot of specificity scores. (B) A scatter plot of similarity scores. The lowess regression is shown in red. (C) & (D) Show the distributions of the GOTrim and IC specificity scores respectively.

To investigate the source of bias observed with the IC specificity scores, we manually examined cases showing major discrepancies between the two methods, which corresponded mostly to situations where the IC score indicated high specificity while the GOTrim score indicated low specificity. The term instances we looked at appeared to reflect a level of semantics that is better captured by the GOTrim scores while the IC scores appeared to be rather exaggerated. For instance, the term "negative regulation of neurological process (GO:0031645)" appears to have a significant IC score (=13.28 out of a range of 1 to 14.5) when it clearly has broad semantics; on the other hand, its GOTrim score (=5.01) is certainly more believable. One other example is the term ‘regulation of cell projection organisation and biogenesis (GO:0031344)’ which is given a high specificity score by the IC approach (=13.08) and a more reasonable score of 5.96 by the GOTrim method.

To further confirm the exaggeration in the specificity scores by the Resnik IC approach, we correlated the terms specificity scores obtained with each method with their shortest path distances from the root. Whilst the latter is no accurate measure of specificity since, as discussed in the introduction section, it ignores the problem of varying edge weights at different levels in the GO graph, it may serve as a rough indicator of specificity. Thus, terms only few edges away from the root can only have broad semantics whilst those furthest

from the root are likely to be more specialised. The results appear in Figure 5.3.5.

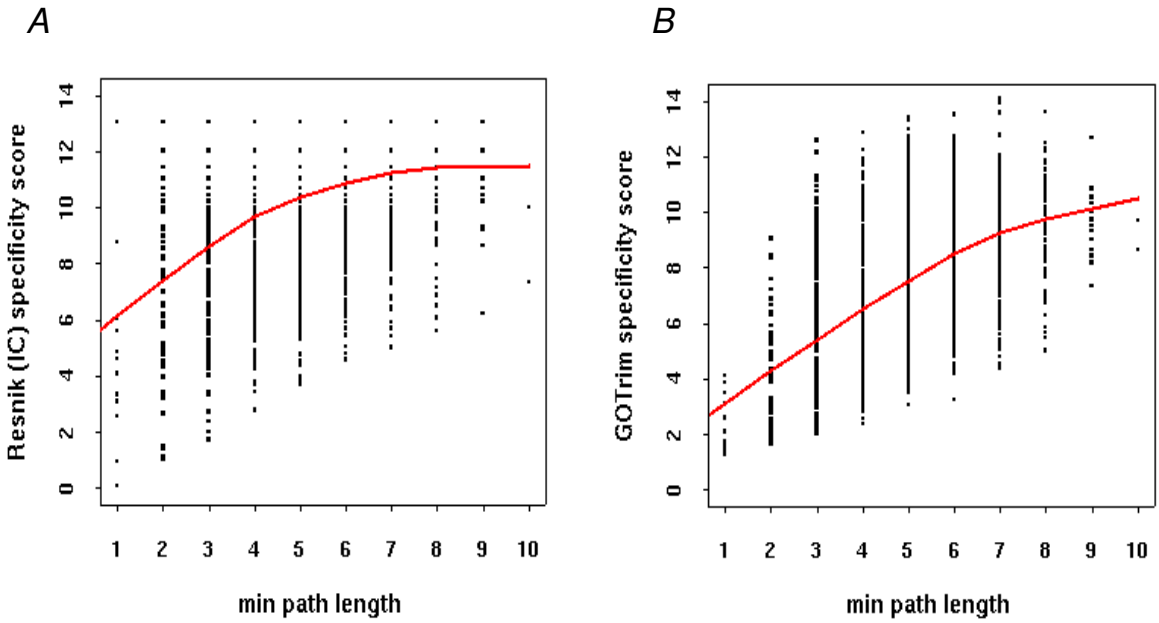


Figure 5.3.5. **Terms specificity scores versus the length of their shortest paths to the root**. (A) Resnik (IC). (B) GOTrim. Lowess regression lines appear in red.

As a general trend, the GOTrim specificity scores correlate better with the shortest path lengths for the varying terms than the IC-based specificity scores (Fig 5.3.5). This may seem rather expected since the GOTrim specificity scores are partly influenced by the number of ancestral terms and hence the number of edges on the path(s) to the root. However, one striking observation is that some of the closest terms to the root appear to have extremely high IC-based specificity scores (corresponding to the data points on the top left hand

corner of the plot in Fig 5.3.5-A); which is rather illogical as these terms can only have broad semantics. This further confirms the occasional exaggeration of term specificity scores by the Resnik IC-based approach.

The explanation for the occasional flaws with the IC approach for measuring GO terms semantic specificity may lie in the fact that the level of representation of a term in a corpus functional database is not an absolute attribute of the term specificity as it may potentially be influenced by a number of other factors. For instance, the extent of scientific interest in characterising the molecular basis of varying biological functions differs and as such, a term may be associated with a relatively significant number of genes because the function it embodies has been of general interest and hence widely investigated. By contrast, terms encapsulating fairly general functions may turn out to be associated with fewer numbers of genes because their functions have not yet been studied adequately. In addition, we know that some biological functions utilise multiple mechanisms and would naturally employ a larger set of genes unlike other functions that are effected by a fewer number of genes. This all suggests that our knowledge of gene associations with the various terms in GO does not always truly reflect the terms' level of semantic specificity.

So far, we have evaluated the GOTrim method with respect to the Resnik approach and attempted to link variations between the two methods to possible flaws with the Resnik metric. However, in order to rigorously validate the performance of the GOTrim method, we studied the correlation between the method's derived similarity scores for pairs of yeast proteins and their sequence similarity levels, thereby exploiting the tight relationship between sequence identity and functional conservation. This approach has traditionally been used for validation of most GO semantic similarity methods, as pointed out in the introduction section.

Thus, protein sequences from the yeast genome were compared using BLAST to obtain sequence identity scores. The latter were based on the log reciprocal of the blast bit scores (LRBS) similar to the study by Pesquita and colleagues (Pesquita et al., 2008). Thus, for each pair of proteins A&B, the LRBS is the log of the average of the bit score from comparing A against B and B against A. It is probably worth mentioning that the BLAST bit score for a pair of sequences is a measure of significance that takes account of the gaps and substitutions in the query sequence when aligned against the target sequence and the higher the bit score, the higher the significance of the alignment from the two sequences. As such, the bit score for a pair of sequences changes

depending on which sequence in the pair is treated as the query and which is treated as the target.

Because proteins are often annotated with more than one GO biological process term, we used the best-match average approach to deduce a summary GOTrim similarity value for each protein pair. As explained in the introduction section, with this approach, each term from one protein is matched with its most similar term from the other protein and vice versa; an average similarity value is then calculated to denote the summary similarity value for the protein pair.

For the sake of comparison, the Resnik semantic similarity scores for the same protein pairs were also calculated, similarly using the best-match approach, and correlated in a similar fashion with the corresponding LRBS. For both methods, the scatter of points from the correlation analysis was summarised by applying a lowess regression (Fig 5.3.6-A&B). The correlation coefficients were found to be equal to 0.68 and 0.59 with the GOTrim and Resnik (IC) metrics respectively; thus indicating the superiority of the GOTrim approach. However, the shape of the regression line indicates that the relationship between the average semantic similarity and sequence similarity scores is not linear, which makes the use of correlation coefficients non-optimal.

This was also noted by Pesquita and colleagues (Pesquita et al., 2008) who suggested an alternative assessment criterion based on sensitivity, that is the extent to which, on average, variations in the sequence similarity scale are translated into the semantic similarity scale. This relationship is precisely modelled by the regression analysis featuring in Figure 5.3.6-A&B. In Figure 5.3.6-C, the regression lines from the sequence similarity and semantic similarity correlation analysis, by both methods, are superimposed. Clearly, the range of sequence similarity detected by either method is the same (roughly 2-3 LRBS). By contrast, this same range of sequence similarity is resolved into a higher range of semantic similarity by GOTrim than Resnik, (roughly between 2-9 as oppose to 2-7 for each method respectively); indicating higher sensitivity by the GOTrim method.

To understand the significance of the range sensitivity criterion, one needs to consider a few important facts. First that the similarity scores by both methods span a similar range of values (between 0 and 12). Also, from previous analysis (Fig 5.3.4-B), the similarity scores by Resnik were shown to be on average slightly higher than those by the GOTrim method for pairs of GO terms. Despite that, the average sequence similarity score by the GOTrim method appear to show steadily higher values than the Resnik method with bins of increasing LRBS scores. This indicates that the GOTrim similarity

scores are generally more accurate reflecting more consistently the true extent of semantic similarity between pairs of proteins than the Resnik IC similarity scores.

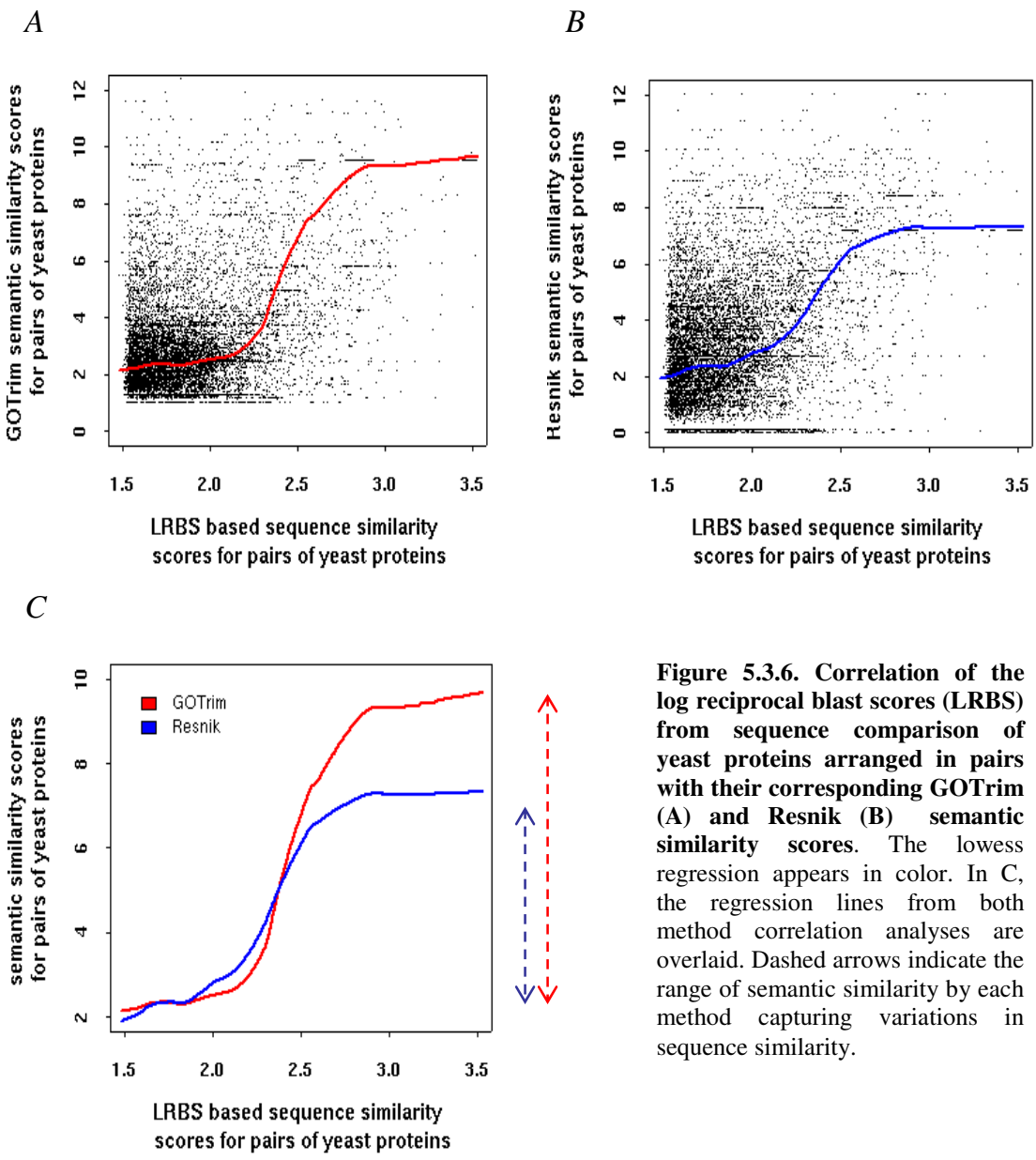


Figure 5.3.6. Correlation of the log reciprocal blast scores (LRBS) from sequence comparison of yeast proteins arranged in pairs with their corresponding GOTrim (A) and Resnik (B) semantic similarity scores. The lowess regression appears in color. In C, the regression lines from both method correlation analyses are overlaid. Dashed arrows indicate the range of semantic similarity by each method capturing variations in sequence similarity.

5.4. Discussion

In this work, a GO semantic similarity measure was developed that explores the structure of the GO ontology. The basic idea behind the method is that the semantic similarity between GO terms may be measured on the basis of the extent of information in common between them, captured by the semantic specificity of their immediate common ancestor. The more information shared between terms, that is the more specialised the common ancestor, the higher the similarity between them. The method's strategy for measuring the semantic specificity of the common ancestor was designed to explore the 'history' of semantic specialisation by predecessor terms from high up in the GO graph.

The advantage of the GOTrim method is that because it uses a different approach to information content to measure the level of informativeness of individual terms, it avoids potential flaws with the information content based approach. In particular, in case of GO, the additional factors that could influence a term's level of association with genes in a corpus database other than the extent of specialisation of its encapsulated function. For example, the complexity of the mechanism involved at the molecular level and the rigorousness with which this mechanism has been investigated. Indeed, we

were able to show that the GOTrim method performs better than the IC based Resnik method.

However, the GOTrim method has a number of drawbacks. First, its usage of the ontology structure makes it susceptible to the problem of varying term density across different parts of the GO graph. This is because the ontology is under a constant process of refinement and at any given point in time, some functions may be better annotated with terms than other functions. Two more disadvantages of the method proposed, also shared by the Resnik approach, relate to the fact that the similarity between two terms is taken to be the specificity of their most specialised common ancestor. Thus, if the two terms being compared are identical, their similarity score would reduce to the specificity value of the immediate parent; the latter is a random value that gives no indication of the fact that the two terms are in fact identical.

The other problem is that for any ancestor term, all pairs of terms for which this same ancestor is the most specialised ancestor would have the same similarity value regardless of how deep down they come from in the GO graph. This could lead to a loss of useful information. For instance, looking at Figure 5.3.6 showing the correlation between semantic and sequence similarity for pairs of yeast proteins, where highly similar proteins display

unexpectedly low semantic similarity scores, we cannot be sure whether this is due to the proteins being associated with completely different terms or to one of the proteins being annotated with a rather general term where the common ancestor is bound to have a low specificity value.

Ironically, existing GO semantic similarity approaches that provide solutions to this last problem, which both our method as well as the Resnik metric appear to suffer from, have their own different limitations. Thus, the Lin, Jiang and Wang approaches are sensitive to the location of the terms being compared on the GO graph; yet, they all suffer from the problem of artificially high similarity values at close proximity from the root, as outlined in the introduction chapter.

CHAPTER VI: A GO BASED FRAMEWORK FOR AUTOMATIC BIOLOGICAL ASSESSMENT OF MICROARRAY FUNCTIONAL ANALYSIS METHODS.

6.1. Introduction

6.1.1. Microarray functional analysis

Following the low level analysis of microarrays expression data whereby probeset intensities are processed to eliminate noise and inter-chip variation to yield the most optimal expression levels for the genes, statistical analysis usually follows to determine whether the expression levels of genes show any significant change between biological conditions. The outcome from such analysis is usually a substantial list of genes ranked by the statistical evidence for their differential expression. To extract useful biological information from such lists of genes, higher-level analyses can be applied. For example, by clustering the genes over a number of experimental conditions and reconstructing transcriptional networks to help identify key transcription factor coding genes.

One important form of high level microarray analysis is functional analysis. With functional analysis, the relevance of a functional theme with respect to the biological phenomenon investigated in the microarray study is determined on the basis of a coordinated behavior of mediating genes in response to the phenomenon. Such coordinated behavior is versatile and common examples are a concerted change in expression with respect to the normal state or a common expression profile over a number of experimental conditions exploring variations in a particular aspect of the phenomenon, usually detected by clustering analysis. In simple words, functional analysis may be defined as the study of enrichment of particular functional annotations among selected subsets of genes grouped on the basis of a common biologically relevant feature such as differential expression, correlation with a phenotype of interest or common regulatory patterns.

When applied to the list of ranked genes from statistical analysis, functional analysis is particularly useful as it helps reduce the resulting sheer amount of information to a more manageable list of functional categories, while revealing the functional properties of the biological reaction involved. Also, functional analysis represents an improvement from traditional gene-based statistical analysis approaches in that it helps highlight instances where genes exhibit individually modest changes in expression, but tend to change in a

coordinated manner as a group of genes with similar function. This is important in situations where changes in gene expression are confined to a specific subtype of cells within the sampled tissue and may be compromised by dilution effects.

One universal advantage of functional analysis is the fact that it helps suppress experimental variation. Microarray data are known to be influenced by a variety of experimental factors such as laboratory equipment, the experimenter's handling of the experiment, the design of the chip and it is now well established in the literature that microarray experiments addressing the same biological question often show little gene-specific expression changes in common. By contrast, functional analysis is capable of highlighting similarities between independently generated, yet biologically equivalent microarray datasets, by focussing on functional groups of genes instead of the genes per se. Also, from the analysis point of view, it has been shown that the additional variability introduced by the choice of the low-level analysis methods for microarray data may be suppressed by functional analysis (Hosack et al., 2003).

In the last ten years, many functional analysis tools have been made available to the microarray research community. The main difference between these

tools lies in the statistical framework employed by each method to test for enrichment of gene functional categories. In addition, these methods may display varying additional features in order to enhance their usability. For instance, many provide built-in functional groupings of genes and can handle a variety of gene identifiers. One example is *FatiGO* (Al Shahrour et al., 2004) that uses GO as a basis for classifying genes while maintaining links to major sequence databases such as GenBank, Unigene, Ensembl and Swissprot/TrEMBL. In addition, many functional analysis tools have evolved to allow grouping of genes on the basis of functional vocabularies other than GO, common key terms and common biological properties. For instance, the revised version of the functional analysis tool *GSEA* (Subramanian et al., 2005) features an integrated database, the Molecular Signatures Database *MSigDB*, containing gene sets derived from common regulatory motifs, chromosomal locations, functional attributes, in addition to common relevance to distinct biological states as postulated in literature and from knowledge of domain experts.

At the statistical level, assessing the level of representation of a functional category c among a subset of selected genes g involves taking into account the full size of gene list G from which g was selected as well as the overall level of occurrence of category c in G given as C. The hypergeometric distribution

(Tavazoie et al., 1999) appropriately models the probability of occurrence of c in g by chance given G and C and has been largely used by pioneer functional analysis tools such as *CLENCH* (Shah and Fedoroff, 2004), *Onto-Express* (Khatri et al., 2002), *FunSpec* (Robinson et al., 2002) and *FuncAssociate* (Berriz et al., 2003). With many densely populated arrays, the hypergeometric probability is computationally expensive to calculate and an approximation to the binomial probability distribution was used additionally by a number of functional analysis methods such as CLENCH and Onto-Express.

Good alternatives to the hypergeometric and binomial distribution statistics are the $\chi 2$ test for equality of proportions used by CLENCH and Onto-Express, as well as the Fisher's Exact test used by FatiGO, *GOstats* (Falcon and Gentleman, 2007), *GOminer* (Zeeberg et al., 2003) and *EASE* (Hosack et al., 2003). These tests are based on a 2 x 2 contingency table specifying the observed proportion of genes attributed to c as well as those not attributed to c from the chosen subset of genes g and the remaining genes in the list G-g. The counts in the table are combined to yield the $\chi 2$ statistics for the $\chi 2$ test whilst the Fisher's test operates by means of calculating the hypergeometric probability of observing these counts.

Perhaps, two major advances in developing a statistical framework for functional analysis have been the introduction of rank based procedures to evaluate enrichment in functions based on the ordering of genes according to a chosen ranking metric, as well as multiple testing correction. We shall discuss the latter before reverting to the former. Because many functional categories are typically tested at once during functional analysis of microarray data, some of them will score low p-values by chance alone. By applying multiple testing correction, the statistical significance of individual categories is adjusted for the size of the database of functional categories tested. Most of the earliest functional analysis methods such as Onto-Express and FunSpec paid no attention to the problem of multiple testing. However, more recent tools adopted some form of multiple testing correction procedures, such as controlling the FDR (Catmap, FuncAssociate..) and/or calculating the FEW family wise error rate as with GOCluster and GOstats (more details on multiple testing and correction procedures are available in the introduction chapter).

The incentive for introducing rank based statistics in functional analysis has been the recognition that with many pre-analyses, the relevance of genes is indicated by ranking them according to a chosen metric. Thus, unlike with many clustering approaches where subsets of jointly regulated genes are

readily defined in a deterministic way, differential expression analysis often results in a list of ranked genes where the change in expression is most important at the top of the list. With the earliest functional analysis approaches employing classical statistics, such as the Fisher's test and the hypergeometric probability, the subset of relevant genes was typically filtered by applying an arbitrary cut-off on the list of ranked genes and choosing all genes above that cut-off. Such an approach is known to be limited because information from the list below the cut-off is typically lost and the choice of the cut-off is not obvious, in particular with noisy data.

To address this issue, functional analysis procedures were designed to make direct use of gene ranks in their statistical framework and avoid selection of relevant subsets of genes beforehand. The first of such methods to have emerged were FuncAssociate and GSEA. FuncAssociate, similar to *IGA* (Breitling et al., 2004) developed a year later, was designed around the concept of minimizing a hypergeometric based probability of enrichment for individual categories by means of identifying the subset of genes in a category that cluster high at the top of the list. Effectively, these methods strive to optimize a cut-off for each category individually (a full description of IGA will follow in section 6.1.2).

Along these lines, GSEA uses the Kolmogorov Smirnov rank based statistics to assess the observed versus expected ranks of category members in the ranked list of genes. The test is performed by traversing the list from top to bottom and estimating, at each current position, the cumulative fraction of genes that are members of the category while separately calculating the cumulative fraction of genes that are not members of the category. A score is then derived on the basis of the maximal difference between the two running fractions (see Fig 6.1.1).

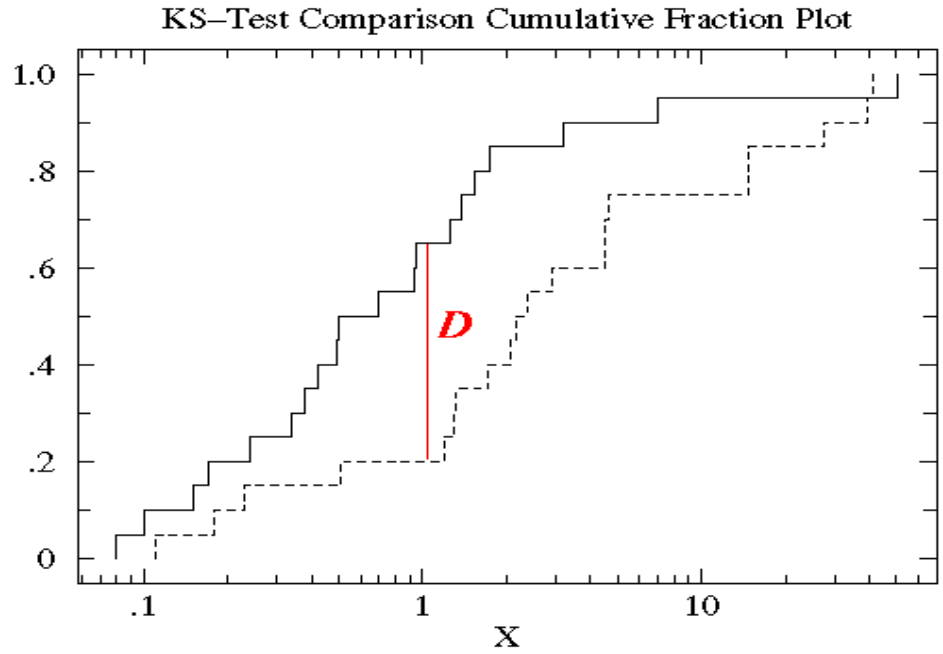


Figure 6.1.1. Picture taken from http://www.physics.csbsju.edu/stats/KS-test.html illustrating the Kolmogorov Smirnov test. In broad terms, the non-parametric KS-test is used to determine whether two datasets differ significantly (treatment versus control or in terms of functional analysis category membership versus non-membership) on the basis of cumulative fractions. This is illustrated in the plot above. The x-axis shows the actual data values from both datasets ranked on a log scale (in the case of functional analysis, the value of the gene ranking metric can be used). At any given rank, the cumulative fractions; in other words, the fractions of data from both datasets are given on the y-axis. A score D is defined where the cumulative fractions from both datasets are most different from each other.

By using the Kolmogorov-Smirnov statistics, GSEA may thus be regarded as adopting a strategy of score optimization, similar to IGA and FunAssociate, since a category score may only be drawn at the point in the list where the cumulative fractions of category member genes and that of non-member genes are most distinct from each other.

A rather different functional analysis rank based approach was later introduced by Breslin and colleagues (Breslin et al., 2004) in their functional analysis tool *Catmap*. Catmap does not select a subset of optimally ranked genes in a category; rather, it uses a comprehensively derived score that combines the ranks of all genes in the category. The score on its own does not reflect the significance of the category but is assigned a significance level, at a later stage, using permutation analysis.

In this work, we assess the performance of the two major rank-based statistical approaches for functional analysis. We use IGA and GSEA as examples of the reductionist approach, whereby ranks from only a subset of genes in a category are used to derive the score for the category. In addition, Catmap is taken to represent the more global approach, which uses rank information from all genes in the category. In the following, we discuss all three methods in more detail.

6.1.1.1. Catmap

To assess the significance of the ranks of all gene members in a category, Catmap calculates a summary score based on their sum, also known as the *Wilcoxon rank sum*. The significance of the score is then calculated as the probability p-value of obtaining a lower score for the category assuming the null hypothesis. The simplest null hypothesis is one based on randomly shuffling the genes in the gene list; which is equivalent to having the genes in the category assigned random ranks.

However, this null hypothesis assumes independency in gene expression level in individual biological samples; which is incorrect considering that genes may be co-expressed. The Catmap algorithm recommends a different null hypothesis based on sample label permutation, whereby random gene lists are obtained from fold change statistical analysis of randomly labelled samples. Such null hypothesis is considered more suitable because it conserves the dependencies between co-expressed genes at the sample level.

With Catmap, the choice of the null hypothesis is specified by the user since, although the sample label permutation is statistically more robust, in experiments where few replicates exist for each phenotype, the sample label

permutation may not be rigorous enough making the use of randomised gene lists inevitable.

Once the choice of the null hypothesis is determined, for each category a distribution of scores is obtained by Catmap by summing up the ranks of its gene members as they occur in each of the random gene lists. A p-value for the category is then calculated by counting the number of times a lower score than the category actual score is encountered (on average) under the null hypothesis. Categories are then ranked by their p-values.

The last step in the Catmap algorithm is that of multiple testing correction. Catmap uses two different approaches for multiple testing correction, based on controlling the family-wise error rate (FWE) and the false discovery rate (FDR). In this work, we are mostly concerned with the FDR, as it can be compared across the different functional analysis methods in a meaningful way. Thus, at any given category rank, the method with the lowest FDR is the best performing.

With Catmap, the FDR is derived from permutation analysis. First, category scores obtained from the random gene lists are assigned p-values similar to the way those from the real gene list were given p-values. Thus, for category X

with score s from a given random gene list, a p-value is obtained by calculating the number of times X has a lower score than s in the remaining random lists of genes divided by their number. This allows Catmap to determine the range of p-values possibly obtained under the null hypothesis, given the number of categories tested. Table 6.1.1 illustrates how the p-value for a category is obtained by Catmap for a small number of random gene lists as well as the real gene list.

	Real gene list	Random Gene list 1	Random Gene list 2	Random Gene list 3	Random Gene list 4	Random Gene list 5
Gene ranks	{309,567,1098 ,14657,20009}	{5670,8937, 10987,16789, 23456}	{4567,5678, 9013,13478, 18976}	{65,4576, 8769,14578, 20980}	{45,176,457 9456,15670}	{13658,16793, 20987,26009, 27430}
Wilcoxon rank sum	36640	65839	51712	48968	25804	104877
p-values	1/5 = 0.2	3/5 = 0.6	2/5 = 0.4	1/5 = 0.2	0/5 = 0	4/5 = 0.8

Table 6.1.1. Illustrating the calculation of the p-value for a category on the basis of biologically meaningful gene ranks (in red) and random gene ranks (in black).

Next, given the list C of categories ordered by p-values from the real gene list, the FDR for category c with p-value P at rank J is calculated by Catmap as the number of times a p-value from any category under the null hypothesis is smaller than P divided by the number of random gene lists divided by J. The first division serves to obtain an average count of categories scoring a better p-value than P over the randomised gene lists whilst the second division aims to

express such count as a fraction of categories from C at higher rank than J. In other words, the FDR gives an estimate of the proportion of categories above category c in the ordered list of categories by Catmap expected to occur by chance. The diagram on Figure 6.1.2 summarises the different steps in the Catmap algorithm.

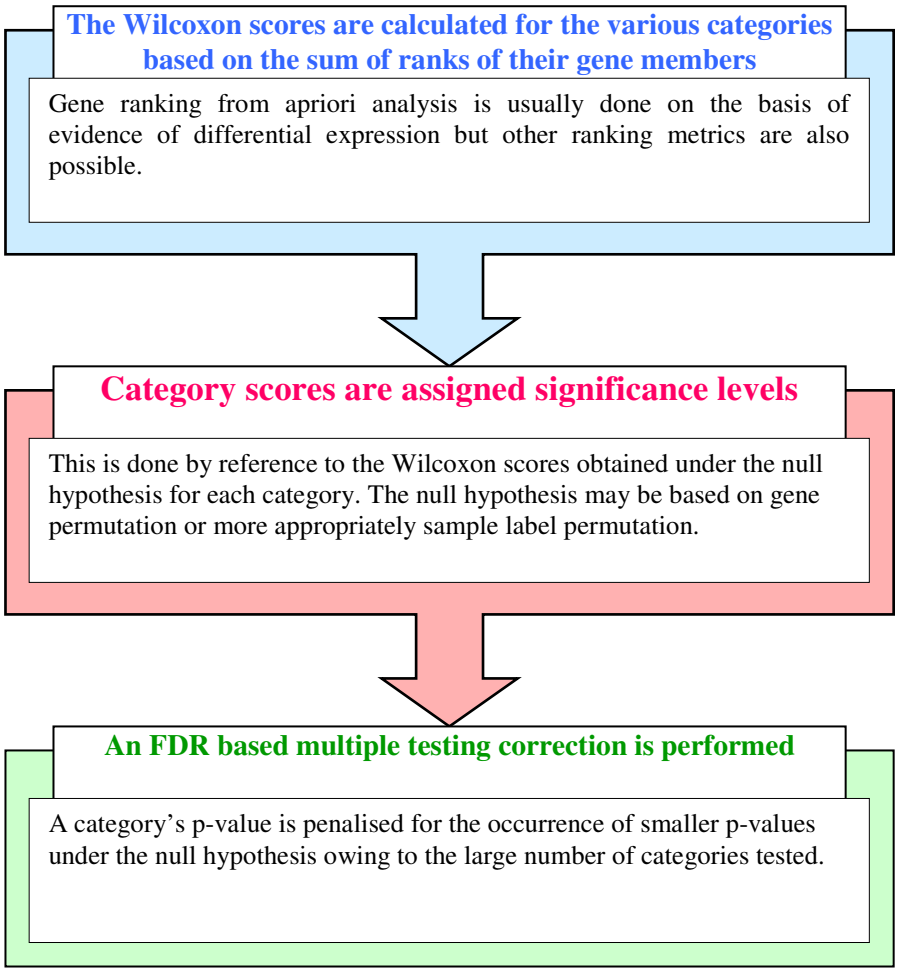


Figure 6.1.2. A schematic diagram illustrating the Catmap algorithm

6.1.1.2. IGA (Iterative Group Analysis)

In contrast to the Catmap method where a category is assessed on the basis of information from all its member genes, IGA analysis of gene categories is based on identifying the subset of genes in a category that prove most relevant to the biological question investigated in the microarray study. For example, in experiments aiming to reveal gene expression regulation in diseased or treated biological states with respect to 'normal', only genes in each category with high evidence of differential expression are typically considered. This reflects the view that for a given biological event, not all genes in a functional class undergo necessarily a change in expression and it seems more effective from a functional analysis point of view to ignore the unaffected genes in a category.

In order to identify the set of potentially important genes in a category, IGA uses an iterative approach whereby the ranked list of genes is scanned from top to bottom and a summary statistic is recalculated each time a new member of the category is found. The summary statistic is calculated using the hypergeometric probability (p) of encountering that many member genes, including the currently identified member, at that point in the list by chance given the total number of genes in the category.

The occurrence of the last category member in the list marks the end of the iterative process. The result is a profile of p-values over the numbered occurrences of category gene member in the list. An example profile is shown on Figure 6.1.3. The profile indicates that the p-value improves with each occurrence of gene members from the top of the list, but deteriorates gradually when including members from further down the list. This constitutes the basis for identifying the potentially important genes in the category and a cut-off may be set for the category at the point in the profile where the p-value reaches its minimal value. Such a value is taken to define the category score and is referred to by IGA as the *probability of change* value or the '*PC value*'.

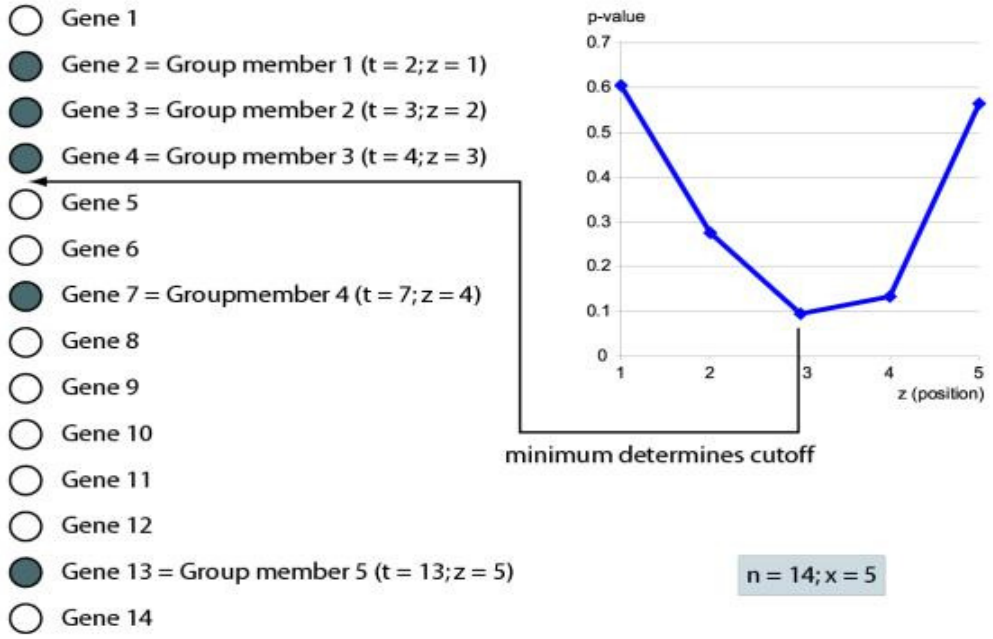


Figure 6.1.3. Principle of Iterative Group Analysis (figure from the IGA paper by (Breitling et al., 2004)). The list of genes ranked by differential expression is shown on the left. The genes member to the category, scored here by IGA, are indicated by black circles. Parameter 't' indicates the ranks of the category genes in the list whilst 'z' numbers them in the order in which they appear in the list. A profile of p-values over z is shown on the right.

Unlike Catmap, IGA does not attempt to derive statistical significance for each category score (or PC value) because the latter is based on p-value. This assumption is rather debatable, as will be discussed later. Instead, categories are sorted by their PC values and a multiple testing correction step follows, based on controlling the FDR in a similar fashion to Catmap. To do that, IGA employs a null model based on randomly shuffling the order of the genes in the gene list. For the list of categories ordered by the actual PC values from IGA analysis of the real gene list, the FDR at rank J corresponding to PC value P is given as the number of times a PC value from the null distribution is smaller than P divided by the number of random gene lists divided by J.

6.1.1.3. GSEA (Gene set enrichment analysis)

Similar to IGA, the GSEA statistics are based on identifying the subset of genes in a category that cluster at the top of the list more than expected by chance. In the original version of GSEA (Mootha et al., 2003), a category is scored using the Kolmogorov-Smirnov statistic (illustrated earlier, Fig 6.1.1), whereby walking down the ordered list of genes, a running score is incremented by a constant at the occurrence of a category member gene and decremented at the occurrence of a gene not a member of the category. The

maximal absolute value of the running score denotes the category enrichment score (ES).

When the genes in a given category are randomly dispersed along the original list of genes, the running score will tend to fluctuate around 0. By contrast, if a set of genes in a category cluster higher in the list than expected by chance, the running score will tend to rise above its background level, giving an enrichment in the observed fraction of member genes. However, with this original scoring scheme by GSEA, a marked concentration of category members anywhere in the list (not just at the top) would also cause the running score to shift from its background level. This is all illustrated in Figure 6.1.4.

Figure 6.1.4-A shows the profile of the running score for a category that is truly enriched. The occurrence of many member genes at the top of the list is reflected by an increase in the running score, which then decreases gradually as lower gene ranks get explored by the scoring process. This effect is better appreciated when considering the profile in Figure 6.1.4-B, corresponding to a category whose member genes occupy random ranks in the list of genes. Figure 6.1.4-C illustrates the weakness of this scoring process, where an increase in the running score appears to be triggered by the occurrence of category members at the middle of the list more frequently than can be

accounted for by chance. Such occurrence has no biological significance, as only genes from the top of the list are important.

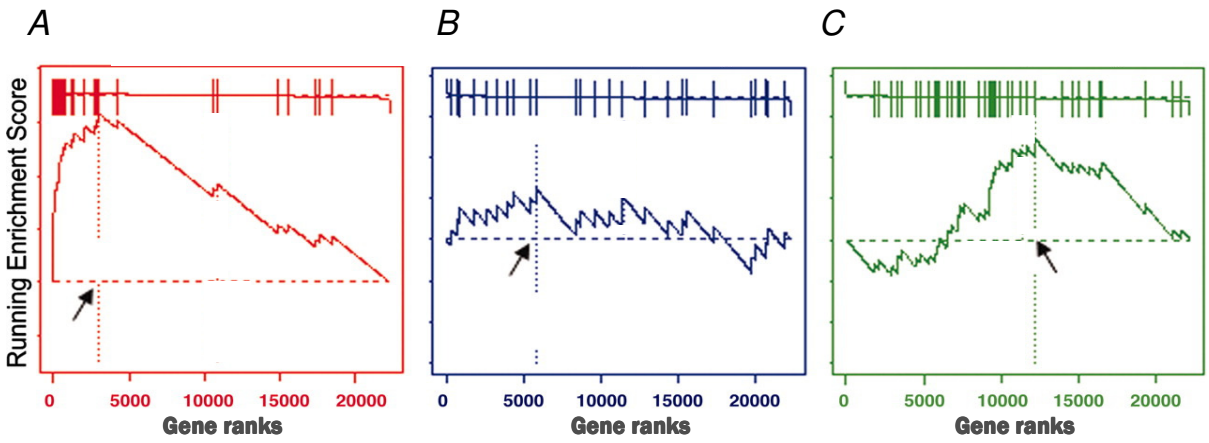


Figure 6.1.4. Enrichment score profiles by the original version of GSEA. (A) An enriched category, (B) a category showing no enrichment (C) a category showing an enrichment in member genes around the middle part of the list. Each category member ranks are indicated on a bar at the top of the plot. The arrow indicates the point where the running enrichment score features the maximal deviation from its background level. This value denotes the category ES score.

The new implementation of GSEA (Subramanian et al., 2005) was intended to deal with this problem. This was done by updating the original GSEA scoring scheme by applying a weight on the increment, so that the running enrichment score is increased to a larger extent when encountering a gene member from the top of the list than from lower parts in the list. The exact mathematical model used by the new version of GSEA to score gene classes is described in equations (1) & (2). Thus, for category S of N_H genes and at any position i in the list of genes L,

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R} \quad \text{, where } N_R = \sum_{g_j \in S} |r_j|^p \tag{1}$$

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{1}{(N - N_H)} \tag{2}$$

and the ES for category S is the maximum deviation from 0 of $P_{hit} - P_{miss.}$

g_j is the gene at position j, N is the total number of genes in L and r_j is the value of the ranking metric at position j (which could be based on the correlation to a phenotype of interest, fold change or a significance value). Thus, setting the power parameter p to a value equal to or greater than 1 causes the increment to be weighted by the value of the ranking metric of category gene members whilst setting p to 0 causes the GSEA algorithm to simply count the occurrences of category members in the list, thereby reverting to its original version.

Following the calculation of enrichment scores (ES) for gene categories, the next step in the GSEA algorithm is to infer statistical significance from these scores using permutation analysis. Similar to Catmap, GSEA recommends the

sample label permutation null hypothesis to generate random lists of genes that preserve the dependencies in expression patterns of the varying genes. For each category, a distribution of ES is obtained under the null hypothesis and a p-value is derived for the category based on the number of ES from the null distribution found higher than the category actual ES (from analysis of the real gene list) divided by the number of randomised lists of genes analysed.

The last step in the GSEA algorithm is that of multiple testing correction whereby the significance of category scores is re-evaluated given the large number of categories tested and the inevitable margin of error. However, unlike most functional analysis methods, GSEA argues that the multiple testing correction should not be applied on the p-values as the latter are not adjusted for category size. This is important, because when correcting for multiple testing, a category's score is assessed with respect to all scores from all categories under the null hypothesis and as such any bias in the scores owing to the size of the categories needs to be eliminated beforehand.

Instead, and as a preliminary step to multiple testing correction, GSEA calculates a normalised version of the ES or 'NES', obtained by dividing the actual ES for a given category by the mean expected ES for the category obtained under the null hypothesis. This allows the value of the observed ES

to be evaluated against the expected range of ES values for the category given its size; although, from a different angle to classical p-values.

The GSEA procedure for estimating the FDR from category NES is here outlined. Given a set of randomised gene lists obtained under the sample label permutation null hypothesis, a null distribution of ES is obtained for each category (an ES from each randomised list). A null distribution of NES is then obtained for the category by diving the ES from each randomised gene list by the mean ES from the rest of the randomised lists. Gene categories are ranked by descending order of their observed NES from analysis of the real list of genes and walking down the resulting list of categories, the FDR at rank j corresponding to NES n is the number of times an NES from any category NES null distribution is greater than n divided by the number of randomised gene lists, divided by j. The flow chart in Figure 6.1.5 summarises the various steps of the GSEA algorithm:

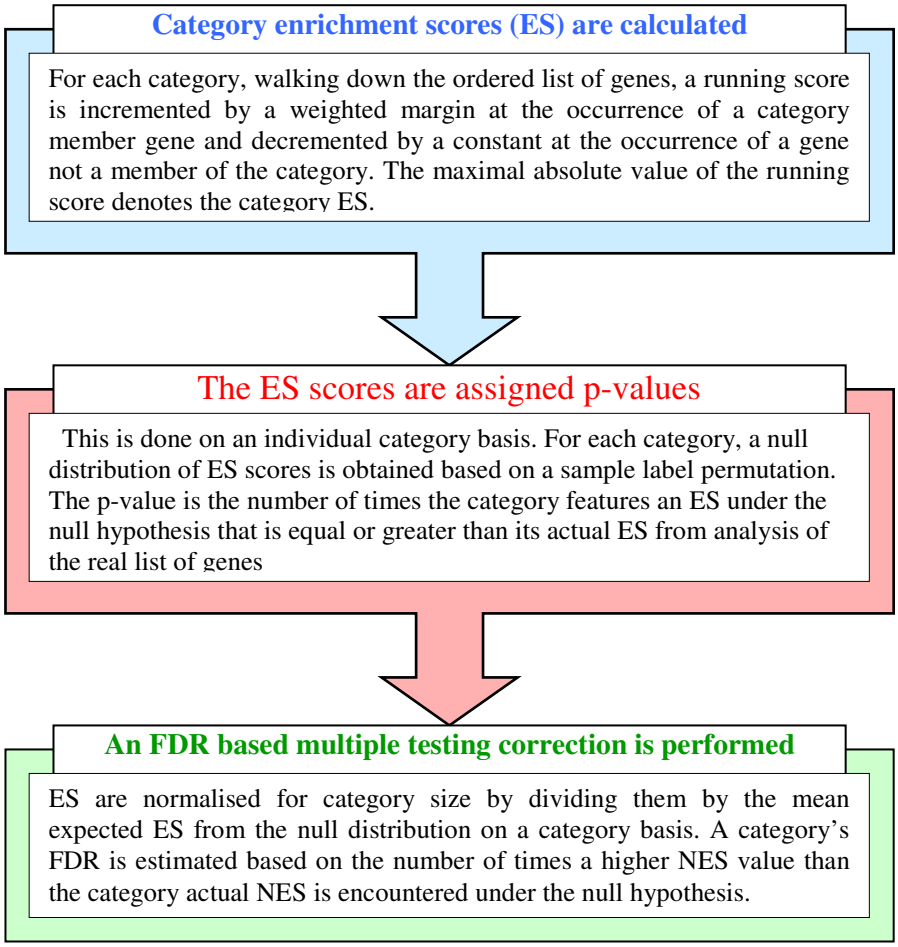


Figure 6.1.5. A schematic diagram illustrating the GSEA algorithm

6.1.1.4. Validation of functional analysis methods

Beyond advocating their statistical theoretical basis, functional analysis tools are ultimately judged on the biological validity of their results. Typically, this has been performed by means of a test expression dataset whose functional properties are well characterised. For instance, a number of publicly available

cancer microarray datasets have been used to validate a significant proportion of functional analysis methods in literature such as the Van't Veer et al (van de Vijver et al., 2002) and AML/ALL leukemia datasets (Dazzi et al., 1995). Thus, the observation of functions such as cell proliferation, apoptosis, p53 related pathways and cell cycle control among the best scoring categories is a positive indication of the validity of the functional analysis approach used to generate these results.

With some functional analysis tools, simpler test datasets with more easily anticipated functional outcomes have been used. For instance, with goCluster and GSEA, validation has been performed on expression profiles of male versus female germ and lymphoblastoid cells, where sets of genes mediating gender specific functions or showing a regulation pattern linked to the Y or X chromosome were expected to show a change in expression. Furthermore, microarray experiments featuring the knock-out of a well-characterized gene may also constitute an ideal setting for validation as it is relatively easy to trace the functional implications of the absent gene. For instance, a p53 knock-out microarray dataset was used among other datasets for the validation of the updated version of GSEA. However, the use of such simple test datasets may not allow a rigorous validation, which is why they are often used together with more complex datasets for the validation process.

Other less compelling forms of biological assessment of functional analysis methods exist in the literature. For instance, IGA used a semi-blind approach where microarray datasets of unknown biology were obtained from collaborators and the results from functional analysis of these datasets were used to predict their unknown physiological states (Breitling et al., 2004). With GOminer, the assessment took the form of validating functions found enriched by functional analysis that were not previously linked to the biology of the analysed test dataset using wet lab experimental techniques. Further evidence of biological validity was sometimes obtained by showing that the results from analysing two biologically identical datasets (but independently generated) were similar, as used in the GSEA and IGA studies.

A valid point of criticism for these varying forms of biological validation for functional analysis is that they are all fairly subjective, requiring human input to trace the link between the observed results and the expected outcome. Moreover, with the more common and most convincing form of validation featuring the use of a functionally well-characterised dataset as a test case for analysis, the results are simply surveyed for biological relevance but not quantitatively assessed, giving no estimate for the proportion of true and false hits among the top results. This is justified by the difficulty in accounting for all possible effects occurring at the level of function in the test dataset and

while the relevance of certain functions in the results may seem clear, the implication of other functions may be rather subtle. The subjective and informal nature of this validation approach makes it unsuitable for use in a comparative context, where the performances of a number of functional analysis approaches may be desirably compared.

6.1.2. Aim of the chapter

In this work, we propose an improved strategy for the biological validation of functional analysis methods and demonstrate its effectiveness by using it to pinpoint differences in the performance of publicly available functional analysis tools. Our validation strategy is based, similar to the traditional approach, on a test microarray dataset that is biologically well-characterised. However, our method has the additional feature of using a fully automated protocol to capture the similarity between functions known to be induced in the test dataset and the results from functional analysis of this dataset. This is achieved by annotating both sets of anticipated and observed functions with GO terms whilst using the semantic categorization by GO to capture the semantic similarity between them.

By using an automated procedure to evaluate the link between the observed results and anticipated outcome, our method provides an alternative to the subjective nature of the traditional approach. Moreover, the proposed evaluation method has the additional advantage of yielding useful numerical estimates. Thus, while acknowledging the difficulty to derive absolute statistics regarding the number of true and false hits among the observed results, primarily owing to the difficulty to account for the full range of truly affected functions, we find that a satisfactory solution lies in gathering information on all functions possibly implicated while providing a mechanism to weight the evidence supporting them. A score may then be derived to express the overall level of confidence in the observed results.

Our strategy to reveal potentially affected functions in the test dataset is based on identifying functions associated with genes found differentially expressed in a number of published microarray studies, obtained under similar biological conditions to the test dataset. A confidence level is then derived for each function on the basis of its frequency among these chosen studies as well as the frequency of closely related functions. Importantly, the optimized validation strategy proposed in this work has one additional advantage, which is the relative ease with which the anticipated set of functions can be compiled from literature without the need to refer to expert biologists.

To our knowledge, this work constitutes the first attempt to construct a quantitative and automated framework that uses real data for the biological validation of functional analysis methods. The only earlier attempt to perform such automated validation, made by Alexa et al (Alexa et al., 2006), used simulated data. This consisted of a population of random GO categories with a small number of deliberately enriched categories. Alexa and colleagues conducted their evaluation of their proposed functional analysis tool on the basis of the fraction of correctly identified GO terms in the simulated category dataset. However, this approach is not optimal as simulated data are idealistic in comparison to real data and may cause the performance of functional analysis methods to be overestimated.

Finally and further to developing the methodology for an automatic biological assessment of functional analysis tools, one important aim of this chapter is to run a comparison of publicly available functional analysis methods utilising varying rank based statistics: notably, IGA based on the minimized p-value metric, GSEA based on a weighted Kolmogorov-Smirnov statistics and Catmap featuring the Wilcoxon sum of ranks. Importantly, we hope to address the question of whether the reductionist approach employed by IGA and GSEA, that derives a category score on the basis of the ranks of a handful of

its member genes, has any advantage at the performance level in comparison to the more global approach employed by Catmap.

6.2. Methods

6.2.1. The test expression dataset

To evaluate the performance of all three functional analysis approaches IGA, GSEA and Catmap, a test microarray expression dataset was analysed for functional enrichment by all three methods. This dataset was obtained by microarray profiling of DRG tissue from animals that have been subjected to the spinal nerve transection procedure (SNT) by LPC experimentalists. The dataset was considered suitable because a number of expression datasets featuring the same or biologically related nerve injury models were available from the literature and conveniently integrated in the LPD. In other words, there was considerable knowledge about its functional properties.

The test dataset will be referred to as the *SNT dataset* for the rest of the chapter, owing to the spinal nerve transection (SNT) procedure performed during the experimental phase. Whilst the experimental details of the original study are described in full in (Maratou et al., 2009), here we give a brief overview. As described in (Bridges et al., 2001), the SNT procedure was performed by first exposing the L5 segment through the paraspinal muscle sheath, ligating tightly the L5 with a silk suture then cutting a few millimeters

away from the suture. Only SNT submitted animals exhibiting significant levels of mechanical hypersensitivity at day 14 post surgery were included in the experiment. In parallel, sham (control) animals were obtained by similarly exposing the L5 while keeping it undamaged. The SNT and sham animals constitute the two varying biological conditions subject to comparison in this microarray study.

For both conditions, mRNA was pooled from ipsilateral L5 DRG tissue from three animals and 200 ng was sampled for amplification using the Affymetrix small sample protocol VII (http://affymetrix.com). Four replicate hybridizations were obtained for each condition using the Affymetrix GeneChip Rat Genome 230 2.0 arrays (Santa Clara, CA, USA). After staining and washing, the arrays were scanned and CEL files containing probe raw intensity values were obtained using the Affymetrix Microarray Suite software.

6.2.2. Low level analysis of the SNT test dataset.

The raw data from the Affymetrix CEL files were processed with a range of microarray low level analysis functions from several Bioconductor packages accessible from R, the programming environment for statistical analysis. The

first step consisted of quality control (QC), in which outlier arrays were identified. QC functions from the *affy* package were used to generate intensity scatter plots from all possible pairs of arrays within and across conditions as well as clustering arrays on the basis of similarity in gene intensity in a hierarchical setting.

The gcrma function from the *GCRMA* package was then applied on the raw data from quality arrays to achieve background correction, normalization and calculation of probeset expression summaries (these steps are described in the introduction chapter). Statistical analysis of differential expression (sham versus SNT) was performed using the *lmFit* and *eBayes* functions from the *limma* package as detailed in the limma vignette, which can be found on http://www.statsci.org/smyth/pubs/limma-biocbook-reprint.pdf. The result was a list of probesets ranked by the estimated evidence of differential expression by limma.

6.2.3. Functional analysis of the SNT dataset.

Two main data files are required for functional analysis by each of the three methods being compared GSEA, IGA and Catmap: the gene list file providing a list of genes ranked by a chosen metric and the gene annotation file listing

associations between genes and annotations of interest. In this work, the ranking of genes was on the basis of significance of differential expression as determined by limma log odds values. Though, since the limma output list features probeset identifiers instead of gene identifiers and because the gene to probesets mapping on Affymetrix arrays is typically that of one to many, further processing of the list was necessary to remove redundant probesets. This is important to assure that the enrichment score for each functional theme is based on single reading from individual member genes. To do this, information on probesets from the RAT230 array was obtained from the LPD database, including UniGene identifiers and sequence MD5 digests and where two probesets were found mapping to the same gene (on the basis of identical UniGene IDs or sequence MD5s), the probeset with the best rank was retained. This meant that the limma list, originally containing 31100 probeset entries, was reduced to 23943 gene entries. It is worth noting though that the list may still contain some information redundancy because the less well-annotated EST probesets, originating from identical genes, may not have all been detected.

As for the gene annotation file, the GO biological process annotations for the Rat 230 array were obtained from the LPD database and further processing of these annotations proved necessary before they were used to construct the

annotation file. This is because while examining the frequency of gene annotation co-occurrence, it was found that most of the GO terms were associated with only a small number of genes (this is explored further and illustrated with a figure in appendix 6.5.1). This may compromise the power of statistics while assessing the likelihood of enrichment of these categories using functional analysis.

A common solution to this problem, featuring in many published functional analysis studies, consists of back-propagating genes from associated terms to all ancestor terms; justified by the fact that the semantics of a parent term are applicable to all its progeny terms in GO. This was achieved by means of an R script that makes use of the *GOBPANCESTOR* environment object: a precompiled look-up table that links all terms in the GO biological process ontology to their ancestral terms from the Bioconductor GOstats package. After the back-propagation of genes, categories with a consolidated gene count greater than 112 were eliminated because the calculation of IGA and GSEA statistics for such large categories proved rather unfeasible. Moreover, these categories were too general to be useful. Singleton categories associated with single genes were also removed. Also, category terms with the same gene content as any of their child terms were removed; in other words, a parent category term was only retained if it featured at least one additional gene

association in comparison to all it child terms (more details, including a discussion of the advantages and disadvantages of the back-propagation of genes in the GO graph are discussed in appendix 6.5.1).

It is important to note that the different functional analysis methods Catmap, IGA and GSEA require different file formats for the gene list and gene annotation files. Thus, although the information content of these files was strictly identical with all three method analyses to ensure a fair comparison, separate files were created for each method that adhered to the recommended file formats. In the following, we give details of how the individual methods were run on the SNT dataset.

6.2.3.1. Functional analysis by Catmap

The Catmap script was download as part of a Perl package accessible from http://bioinfo.thep.lu.se/Catmap, which also features help files giving instructions on how to run the Catmap script and details of the required file formats. Importantly, the Catmap script was run using the --randomnull option to indicate the randomized gene list permutation null hypothesis as oppose to the recommended sample label permutation null hypothesis (all other options were set as recommended). Our choice of null hypothesis was

based on the fact that fewer than four replicate hybridizations remained after quality control (QC) for each phenotype (results from QC analysis of the SNT dataset arrays are presented in appendix 6.5.2); which limits the number of sample label permutations possible. The number of gene list permutations was set via the --multiple option to 5000.

6.2.3.2. Functional analysis by IGA

The IGA perl script was obtained from the supplementary material that accompanies the IGA paper, together with a helper file giving useful notes on the script and required file formats. A distinct feature of the IGA program is that permutation analysis may only be performed at a separate run following an initial run during which categories are assessed for enrichment. Thus, in the first instance, the gene list and gene annotation files are submitted to IGA for analysis while specifying the value of the sensitivity threshold T. The value of T has the range of 1 to n, where n is the total number of categories and only categories scoring a PC-value (the probability of change based on a minimized p-value) less than T/n are included in the results. Thus, setting T equal to n implies that the results from all categories are shown in the output file; although, more commonly, T is given a smaller value so that only the most significant categories are returned.

In this initial run, we have no estimate for the proportion of false positives among the returned categories featuring PC-values less than T/n. To get such estimate, IGA may be run in permutation mode by specifying the option –R while fixing the value of T. The results from this second run show instances of categories scoring PC-values less than T/n from analysis of a set of permuted gene lists. The FDR value at the significance level T/n can then be estimated by dividing the number of 'false' hits from the second run by the number of true hits from the first run.

In effect, IGA in its original form is more suitable for use by biologists who are only interested in identifying the most significant categories. Notably, a biologist may wish to vary the value of the T parameter a few times until a satisfactory FDR value is obtained. However, for the sake of our evaluation study, calculating the FDR at each possible value T separately is tedious and it was deemed far more efficient to change the IGA script to allow permutation analysis to be performed on the fly. This modified version of IGA was run on the SNT dataset and FDR values were obtained using 5000 permuted gene lists, similar to Catmap.

6.2.3.3. Functional analysis by GSEA

With GSEA, Java files were downloaded from the GSEA website at http://www.broad.mit.edu/gsea/, corresponding to the updated version of the GSEA algorithm that uses a weighted Kolmogorov-Smirnov statistics to score gene categories. Documentation and help pages are available at the same address; in particular, information on input file formats may be accessed at http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Data_formats.

An important feature of the GSEA software, in comparison to the two other methods, is that it was designed to accept probeset expression summary data by default while offering a range of different analyses from which various meaningful statistics may be used to rank the genes. Thus, effectively, GSEA assists the user in ranking the genes prior to performing category enrichment analysis. The standard ranking metric by GSEA is the signal to noise ratio, explained in details in appendix 6.5.3.

In this work and in order to assure a fair comparison of all three methods, the input for each method has had to be the same. For this particular reason, we chose to run the *GSEAPreranked* tool of the GSEA software that is suitable for use with pre-ranked lists of genes. Likewise, the limma ranked list of genes

could be used with GSEA just like with Catmap and IGA. More information on the GSEAPreranked tool can be found at http://www.broad.mit.edu/gsea/doc/GSEAUserGuideFrame.html.

In accordance with the notes on GSEAPreranked required file formats at http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Data_formats, a '.rnk' gene list file was created that contained two columns: one listing the Affymetrix probeset identifiers and one listing the corresponding values of the ranking metric. The latter was set to the negated log transformed p-values from limma because the GSEAPreranked algorithm automatically ranks the gene entries in the first column in descending order of the ranking metric in the second column during run time. Likewise, the most significantly differentially expressed genes will be positioned at the top of the list and GO categories enriched among the highly ranked genes will be assigned positive ES values.

As the GSEA algorithm regards category gene enrichment at the top and bottom parts of the list as equally important (in accordance with its default signal to noise ranking metric outlined in appendix 6.5.3), in this work and since we have chosen instead to rank the genes by differential expression, all categories found significant by GSEA for being enriched at the bottom part of

the list were discarded (appendix 6.5.3). This explains why some of the GSEA plots in the result section feature less categories than the total number of categories tested.

In addition to the '.rnk' gene list file, a '.gmt' annotation file was created that captured the GO annotations for the genes from the RAT230 array into a GSEA suitable file format (refer to http://www.broad.mit.edu/cancer/software/gsea/wiki/index.php/Data_formats for more details). It is important to recall that the GSEA analysis package has a built-in database of gene sets known as the MsigDB, but we chose not to use these gene sets and use our own set of GO categories for consistency with Catmap and IGA. Also for compatibility with the model GO category set against which the results from analysing the SNT test dataset by all three functional analysis methods, including GSEA, will be validated.

The GSEAPreranked java tool was run from the command line with the most basic parameters set to their default values, except for the --nperm option which was set to 5000; thus, requesting 5000 gene list permutations for the calculation of p-values, NES and multiple testing correction. It is worth mentioning that with GSEAPreranked, the choice of the null hypothesis is restricted to gene list permutations.

6.2.4. Validation of functional analysis results

To validate the results from Catmap, IGA and GSEA analyses of the SNT dataset, a set of functional categories was assembled using the GO annotations of genes found differentially expressed in similar models of peripheral nerve injury in a number of microarray published studies; including (Costigan et al., 2002; Valder et al., 2003; Wang et al., 2007; Xiao et al., 2002) and one final dataset consisting of genes found regulated in a number of sciatic nerve injury models using a variety of wet lab techniques compiled by the Costigan study. This set of GO functional terms is what was referred to as the 'gold standard set of terms' in chapter IV and we shall refer to some of the observations from this chapter whilst deploying this functional set to validate the results from functional analysis methods in the current chapter.

Importantly, the identification of these published datasets was done in liaison with LPC experimentalists to ensure biological relevance to the test dataset. Thus, in addition to exploring similar peripheral nerve injuries, all datasets were derived from analysis of the expression profile of the DRG tissue ipsilateral to injury. Moreover, the period of time elapsing the nerve injury procedure and the extraction of tissue is consistent for the Wang, Xiao, Valder and our test SNT dataset and consists of two weeks; with the exception of the

Costigan dataset featuring 3 days elapse time and varying times for the dataset compiled from experimental work.

As explained in chapter III, with the four microarray datasets by Valder, Wang, Xiao and Costigan, the raw probeset intensity values were not available and the lists of significantly regulated genes were obtained from the published versions of these studies. This meant that we could not ensure that these varying lists of genes reflected a similar level of statistical significance, because they were derived independently and often using varying statistics. Consequently, the relevance of the GO annotations of these genes (together forming the gold standard term set) to the biology of nerve injury was not certain.

In chapter IV, we explored ways in which a confidence level may be derived for individual categories from the gold standard set, notably via the use of the term study occurrence measure. This was done by first back-propagating genes from terms to their parent terms from the gold standard set and then deriving a confidence measure for each term based on combining the number of genes associated with the term and the number of different studies featuring these associated genes. Such approach was found rather inefficient and a more robust alternative was discussed based on pooling evidence for closely related

terms. In this chapter, we incorporate this concept into a mathematical model that evaluates the collective evidence from groups of gold standard terms while assessing their level of similarity with the results from functional analysis of the SNT test dataset, as will be shown in the result section.

The Results from Catmap, IGA and GSEA analyses were captured in table structures in R and the top scoring categories from each analysis were selected for validation against the gold standard set of terms. Before performing the validation, these top scoring categories were processed to remove subsuming ancestral categories: thus, if a category and its child are both among these top categories, the former is discarded. This was done via an R function that scans the ranked list of categories from each functional analysis top to bottom and evaluates the number of non-subsuming categories from the top and up to each subsequent position in the list, until X number of non-subsuming categories is achieved. Using this function, the 50 top most specialized categories were distilled from the top results of each analysis. These will be referred to as the 'query categories' that we wish to validate against the gold standard set or 'target categories' during the validation process.

Our comparison of query and target categories was optimized so that in addition to identifying exact matches across the two sets of terms, the

semantic relationships between closely related terms were also captured. To this end, the GOTrim semantic similarity metric (described in details in chapter V) was used. The GOTrim method was implemented as an R script and used to derive the similarity value for each pair of query and target categories.

Since it is the aim of this work to develop a scoring protocol to capture the level of agreement between top scoring categories from each functional analysis method (i.e. the query terms) and the set of gold standard terms (i.e. the target terms), further details on the scoring process are given in the results section.

6.3. Results & discussion

The results in this chapter are given in two main parts: the first part compares the results from functional analysis of the SNT dataset by all three methods Catmap, IGA and GSEA and evaluates them from a purely statistical perspective. The second part describes the biological validation of these results, which is the prime aim of this chapter, and features both a description of the methodology used for the validation as well as the outcome from applying this methodology to the top results from each method analysis.

6.3.1. Comparison of functional analysis results by Catmap, IGA and GSEA

Following the low level analysis of the SNT microarray dataset (outlined in Appendix 6.5.2), enrichment of gene functional categories was assessed by means of three different functional analysis methods: Catmap, IGA and GSEA. Whilst the exact implementation details of these analyses are presented in full in the method section; here, we examine and compare their results. First, we look at the distribution of resulting p-values for all categories, which reflects on the ability of each method to identify enriched categories and

second, we assess the performance of each method by analysing its profile of FDR corrected p-values.

6.3.1.1. The distribution of p-values

The distribution of p-values by Catmap, GSEA and that of the minimised p-values (or PC-values, explained in details in section 6.1.1.2) by IGA for all categories are shown in Figure 6.3.1. As it can be seen, IGA has a greater peak at the low end of the scale, followed by catmap then GSEA. This is better shown in Figure 6.3.1-D, where the distributions from all three method analyses are overlayed. This sugests that many more categories were assigned small p-values by IGA than the rest of the methods.

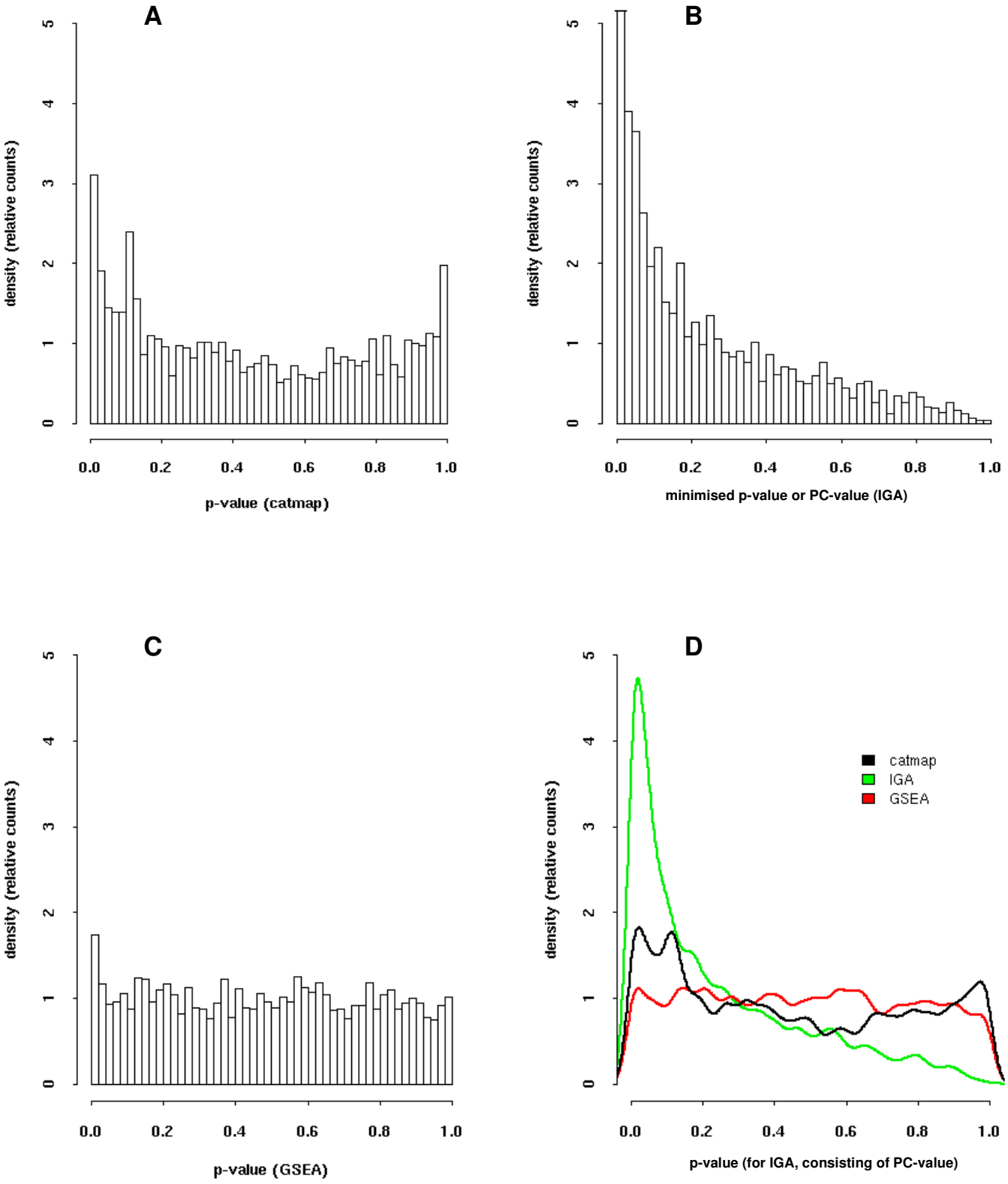


Figure 6.3.1: Histograms showing the distribution of p-values from (A) catmap, (B) IGA, (C) GSEA. The plot in D is a summary of the three previous plots, the only difference is that it uses lines to show the counts of categories over the p-value range instead of bars.

6.3.1.2. The FDR profile

In statistics, hypothesis multiplicity is characterised by the problem of inevitable occurrences of small p-values purely due to chance. One common and least stringent form of multiple testing correction is based on estimating the false discovery rate (FDR), expressing the percentage of categories at any given level of statistical significance expected to occur by chance, usually estimated by permutation analysis.

In this work, an FDR based multiple testing correction was used with all three functional analysis methods. Importantly, the FDR may be used as a basis to compare the performance of the methods, whereby at any given rank in the resulting lists of categories ordered by evidence of enrichment, the method with the smallest FDR is the best performing.

In Figure 6.3.2-A, the FDR profile over the range of p-values by Catmap and that of the minimised p-values by IGA is shown. With GSEA, because the FDR is derived on the basis of category NES instead of p-values, the FDR profile is shown separately on Figure 6.3.2-B (more details about the GSEA algorithm may be found in section 6.1.1.3; but briefly, GSEA justifies its use of NES for the derivation of the FDR on the basis that the latter accounts for

category size as oppose to p-values). In both plots, the effect of multiple testing correction is evident in that the FDR appears to deteriorate a lot faster than the original significance values, reflecting the effected penalisation of the latter for random effects. What is interesting though is that the FDR increases more sharply with IGA than Catmap (Fig 6.3.2-A) in that generally speaking, the FDR value by IGA is higher than that by Catmap at any given p-value. This indicates that IGA statistics are characterised by a higher rate of false positives than Catmap.

Importantly, it is possible to compare the FDR from all three method analyses by considering the ranks of category significance values (p-values by Catmap, minimised p-values by IGA and NES by GSEA), which masks variations in the nature of these values across the methods (Fig 6.3.2-C). Importantly, Catmap appears to perform the best; for example, if one selects the top 50 categories from each analysis, the FDR is 0.02, 0.22 and 0.4 for Catmap, IGA and GSEA respectively (Fig 6.3.2-C&D). The rather poor FDR profile by GSEA may not be surprising given that the p-value distribution by GSEA indicated a modest peak at the low p-value end of the scale (Fig 6.3.1-C&D); implying the inability of GSEA to find much statistical significance among the individual categories tested.

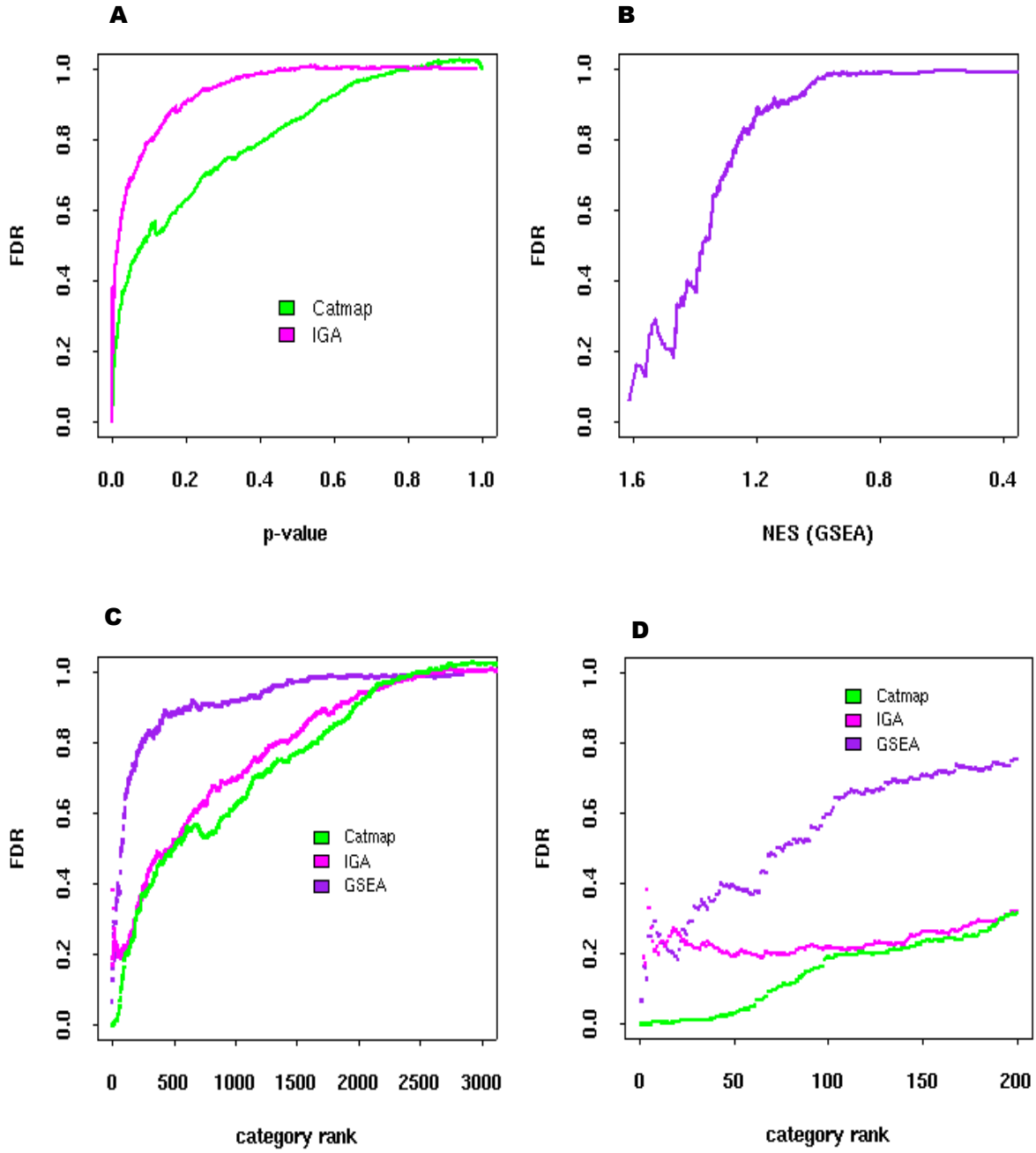


Figure 6.3.2. Assessment of method performances based on false discovery rate (FDR) profiles. (A&B) FDR versus significance values: p-value/minimised p-value by Catmap and IGA respectively (A) and NES by GSEA (B). (C) FDR versus category rank by significance for all three methods. (D) A zoomed version of the plot in C, only showing the FDR for the top 200 categories from each method.

The FDR results from IGA are worthy of more discussion. From the previous analysis of the distribution of category p-values from each method analysis (Fig 6.3.1), it appeared that IGA finds the highest number of categories with small p-values; which suggested at the time a good level of performance. However, from the current analysis, we know that IGA statistics are characterised with a higher FDR than Catmap and thus, many of the putative significant categories from previous analysis may simply be false positives.

The explanation for this phenomenon lies in the nature of the IGA statistics that operate by scoring categories on the basis of minimised p-values (or PC-values) and unlike the rest of the methods, no significance is derived from such category scores on the basis that they are based on p-values. Thus, as featured in the IGA paper by Breitling et al '…the PC-values may occasionally be underestimating the true probability of changes because they are based on determining the minimum p-value within each class (category)…'. Moreover and as suggested by Breslin et al, authors of the Catmap study, these PC-values should not be interpreted as p-values because they are biased by the minimisation process and should rather be thought of as scores from which statistical significance still needs to be inferred.

In fact, the bias in the IGA PC-values is further confirmed by examining the distribution of PC-values under the null hypothesis from IGA analysis of categories with random gene ranks, shown in Figure 6.3.3-A. Thus, whereas the distribution of p-values from Catmap analysis of categories with similarly randomised gene ranks (Fig 6.3.3-B) is uniform as expected under the null hypothesis, that of the minimised p-values (or PC-values) by IGA is skewed towards the low end of the scale; evidencing an overall underestimation of the categories true level of significance.

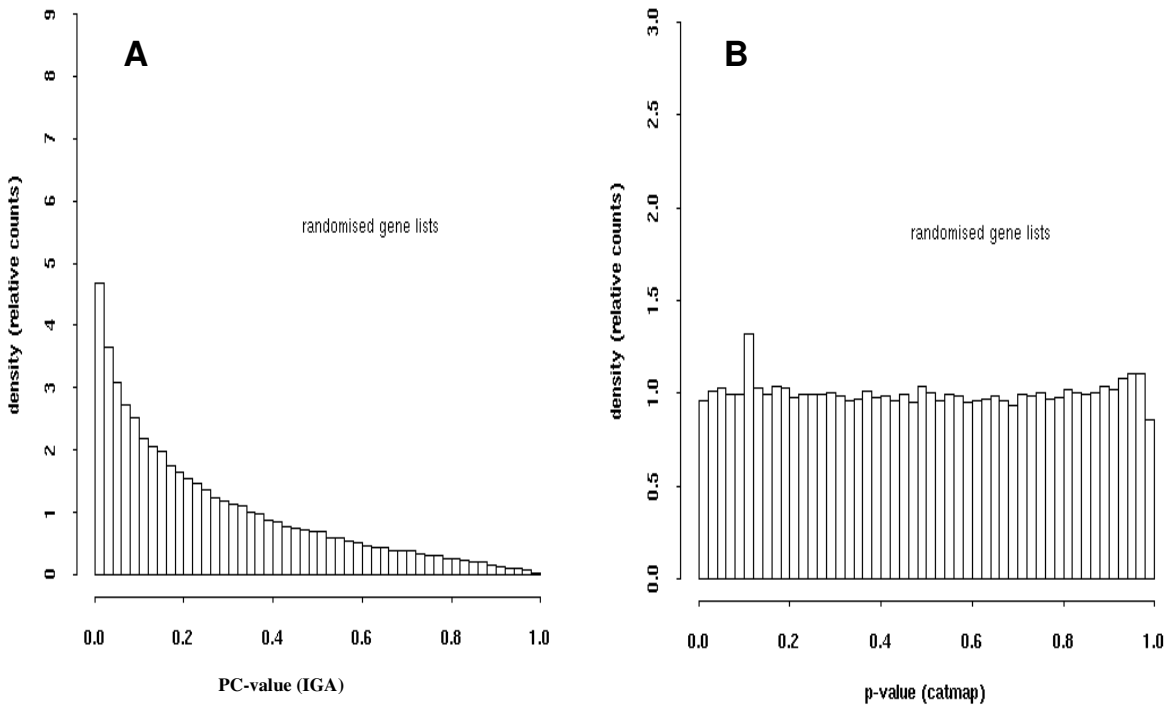


Figure 6.3.3. Histogram of PC-values/p-values from IGA and Catmap analysis of randomised gene lists, A&B respectively. The skewed nature of the distribution by IGA confirms the presence of bias in the minimised p-values (also referred to as the PC-values by IGA).

6.3.1.3. Correlation in category ranks

Previous results indicate that there are clear differences in the statistical properties of Catmap, IGA and GSEA; which suggests in turn that the ranking of categories from analysis of the SNT dataset by all three methods is likely to differ. Indeed, there seems to be some discrepancies in category ranks, with GSEA showing the least level of agreement with the two other methods Catmap and IGA (Fig 6.3.4-B&C), consistent with the observation that GSEA features the highest FDR (Fig 6.3.2-C&D) and is thus least capable of detecting true hits. On the other hand, the category ranks by Catmap and IGA appear to be more correlated (Fig 6.3.4-A). Interestingly, the fact that the most pronounced discrepancies in ranks between Catmap and IGA correspond to instances where categories were ranked lower by IGA than Catmap (top left corner of the correlation plot, Fig 6.3.4-A) supports the hypothesis that IGA statistics are characterised by a tendency to underestimate the true probability of category enrichment.

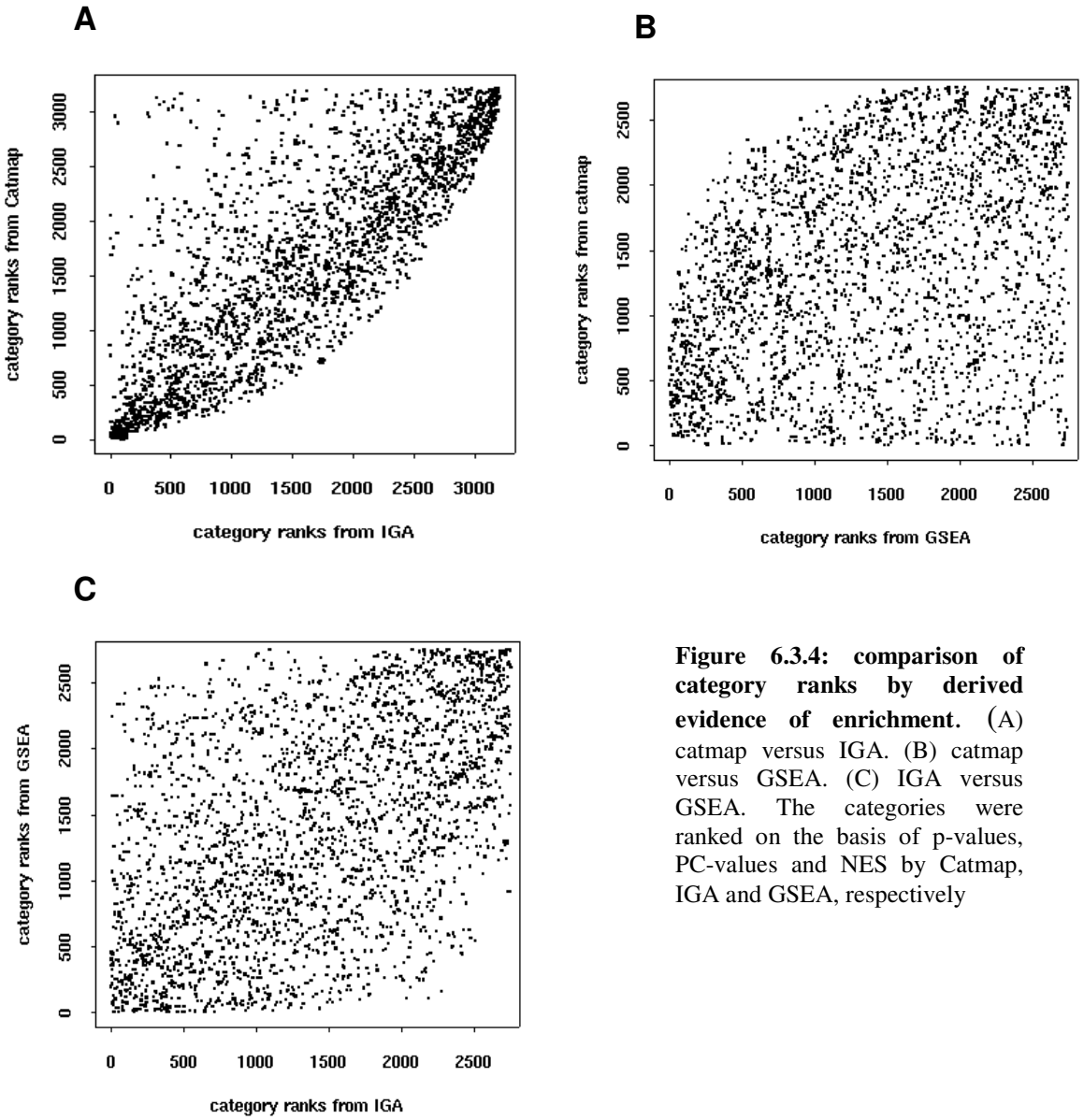


Figure 6.3.4: comparison of category ranks by derived evidence of enrichment. (A) catmap versus IGA. (B) catmap versus GSEA. (C) IGA versus GSEA. The categories were ranked on the basis of p-values, PC-values and NES by Catmap, IGA and GSEA, respectively

The functional analysis of gene categories performed in this study revealed important information on the statistical properties of functional analysis methods used. Thus, GSEA appears to perform least well as it showed the highest FDR and identified the lowest number of significant categories (Fig 6.3.1-C). IGA statistics, on the other hand, appear to have more potential (on the basis of showing a smaller FDR than GSEA) but are nonetheless limited by the tendency to underestimate the category true probability of enrichement. This is due to the nature of the IGA statistics that use minimised p-values as the ultimate significance scores for the categories. Finally, the best performance was revealed by Catmap owing to the small FDR among its top results.

6.3.2. Biological validation of Catmap, IGA and GSEA

In this work, our main aim was to undertake an evaluation of functional analysis methods from a biological perspective, as biological validity is the ultimate criterion for quality. We anticipated the results from the biological assessment to further confirm the previous conclusions about the performance of each method at the statistical level.

To assess the biological validity of the results from functional analysis of the SNT dataset by each method (denoting the query categories), we compared them to the gold standard set of terms (or the target categories) derived from GO annotations of genes reported differentially expressed in a number of microarray published studies investigating similar neuropathy models to the SNT. Thus, whilst the functional analysis of the SNT dataset identifies potentially enriched categories on the basis of a concerted change in expression of member genes in this unique dataset, the gold standard target categories were derived on the basis of occurrence of member genes across a number of published datasets; which makes them more believable from a human perspective and justifies their use as a model answer.

However and as shown in chapter IV, the different target categories from the gold standard set are representative of the published studies to varying extents and are thus associated with varying levels of confidence. This was taken into account while developing a scoring protocol to capture the level of similarity between query and target categories in this chapter, which is described in full in the following section.

6.3.2.1. A scoring protocol to assess the results from functional analysis using prior knowledge.

As already explained, two main factors are meant to be captured during the scoring process of query categories from functional analysis of the SNT dataset: the similarity to the target categories and the evidence supporting these target categories. We use the GOTrim scores (discussed in chapter V) to denote the similarity between categories from the query and target sets. However, since the similarity to a target category is given by the GOTrim method as the specificity of the most specialised ancestor shared with the query category and since many target categories may share the same most specialised ancestor with the query category, it is more efficient to simply consider the specificity of ancestors shared by groups of target categories with the query category term. This is illustrated in Figure 6.3.5.

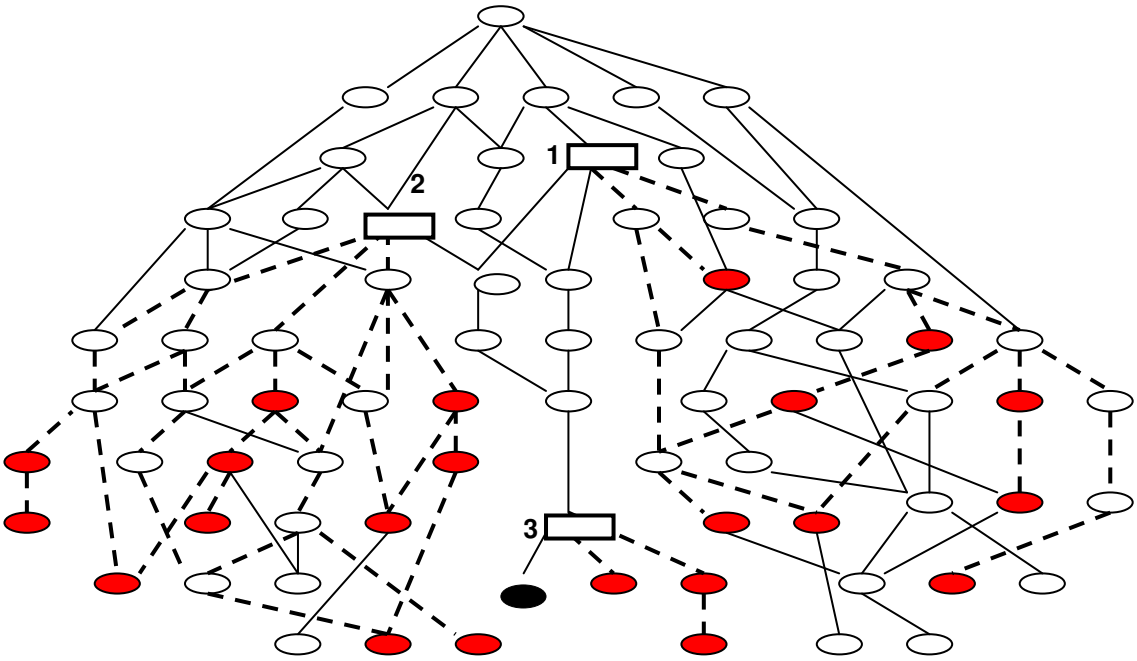


Figure 6.3.5. Diagram illustrating how target categories (corresponding to nodes filled in red) may be organised into clusters during the scoring process of a query category (node filled in black) on the basis of the same most specialised ancestors (shown as rectangular nodes) shared with the query category. Three of such clusters are visible on the diagram and numbered. Paths from the target category terms to the shared common ancestor in each cluster are indicated by dashed lines. More distant ancestors are able to capture larger sets of functionally distinct target categories to the query category (groups 1&2) whilst groups of closely related target categories are generally smaller in size (group 3).

Moreover and beyond simplifying the scoring process, such clustering of target categories has the important advantage of providing a mechanism for pooling evidence across defined sets of target categories. Thus, in chapter IV, we came to the conclusion that a large fraction of target categories feature in only one published dataset but may have related functions to other more highly represented target categories across the different datasets. This indicated the importance of exploring the relationships between target

categories, possibly by means of consolidating the evidence from groups of related targets. However, in this chapter because the ultimate aim from assessing the evidence from the target categories is to evaluate the relevance of query categories that match to these target categories, we have opted to consolidate the evidence from groups of target categories at the same level of similarity with the query (Fig 6.3.6).

In order to derive an evidence measure from groups of target categories, we pool the genes from all target categories in the group. Importantly, we slightly modify the term occurrence evidence measure used in chapter V so that in addition to calculating the number of unique studies featuring this pooled set of genes, we also take account of the number of genes in this set (we refer to these two values as the *study count* and the *gene count* respectively). Importantly, the study count and the gene count values, illustrated in Figure 6.3.6, express two different logical entities and may differ from each other. This is because more than one gene may be reported by the same study. The new measure, which combines the study and gene counts, is referred to as the *gene/study* or *'GS'* measure and is defined in equation 3.

$$GS = \text{study count} + (\log(\text{gene count} / \text{study count})) \quad (3)$$

Importantly, with the GS measure, the study count is still emphasised to a larger extent than the gene count. This reflects our view that the evidence for a query category is most strongly reflected by the level of representation of matching target categories across the selected published studies, rather than the count of their associated genes from these studies. In equation 3, we have minimised the contribution of the gene count to the GS value by first estimating its average value per study and second by log transforming it. The reason why we chose these transformations is because they always yield a value of less than 1. This implies that a group of target categories reported by x number of unique studies (study count = x) may only score a GS value from the range [x, x+1]; meaning that its GS value will always be less than that by any group of targets with a study count greater than x, regardless of the corresponding gene counts. On the other hand, the gene count would have a decisive role in establishing the evidence for groups of target categories with similar study counts; which is why we have chosen to include it in the GS measure in the first place.

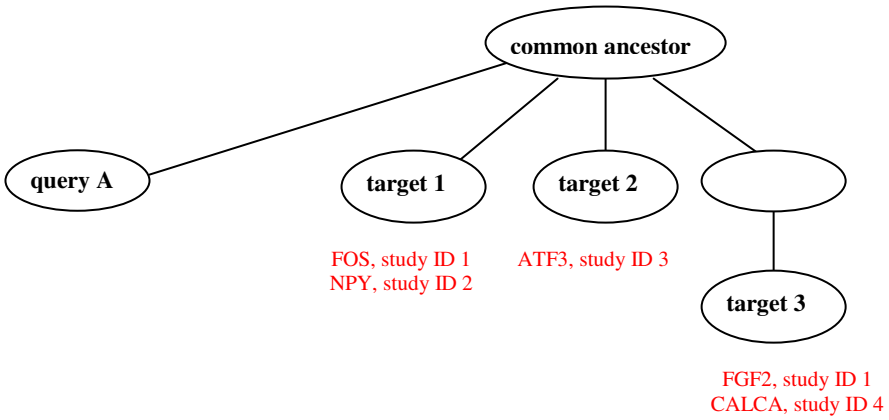


Figure 6.3.6. Diagram illustrating the process of deriving an evidence estimate for a set of target categories grouped on the basis of being at the same level of similarity with the query category. On the diagram, under each target category is a listing of associated genes together with the ID of the studies where these genes appear. The evidence for such group of target categories is based on deriving the number of genes from all target categories in the group (gene count = 5, including FOS, NPY, ATF3, FGF2 and CALCA) and the number of unique studies in which these genes appear (study count = 4, consisting of study ID 1, 2, 3, 4).

The flow chart in Figure 6.3.7 summarises the various steps of the scoring protocol developed in this work to validate the set of query categories from functional analysis of the SNT test dataset against expectedly enriched target categories from published work, featuring similar models of peripheral neuropathy. So far, we have covered the first two steps of the protocol whilst the third and last step remains to be explained. This will be the topic of the following section.

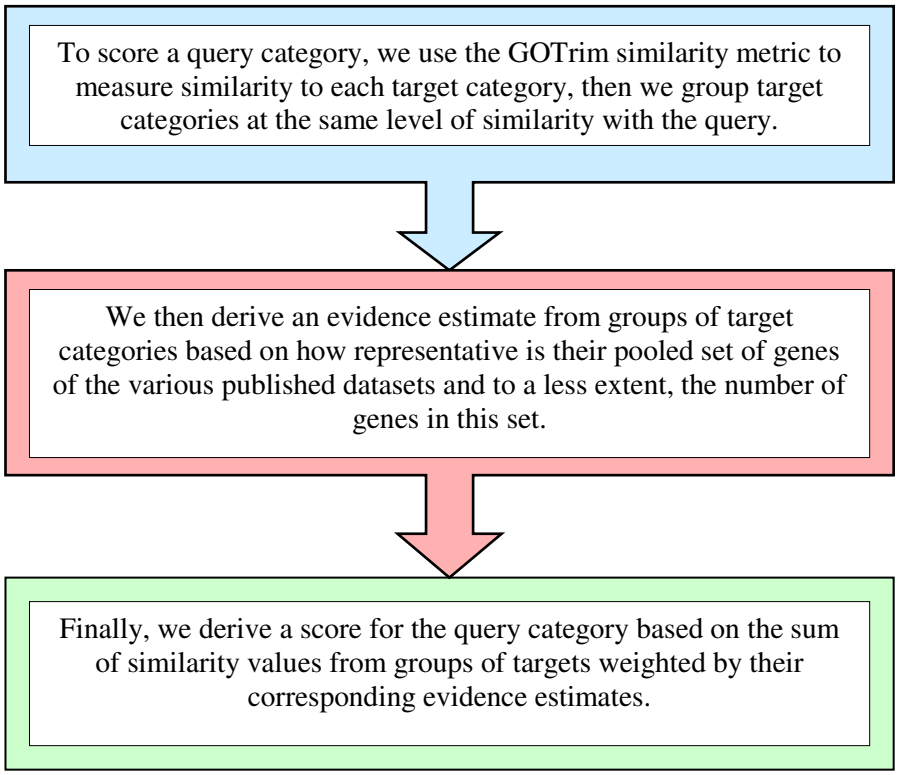


Figure 6.3.7. A schematic diagram illustrating the milestones of the scoring protocol designed to capture the biological relevance of each category promoted by functional analysis (or query category).

Upon defining groups of target categories at the same level of similarity with the query and deriving their GS evidence estimates, the next step in the protocol is to combine the similarity and GS values for groups of target categories by raising the former to the power of the latter. This allows us to weigh out the significance of associations between the query and pre-defined sets of target categories. Finally, the cube root of the sum of calculated $similarity^{GS}$ values from groups of target categories is taken to define the

query final score (we have chosen to apply a root transformation as opposed to a log transformation because the former has desirable linear characteristics). Thus, the score S for query category q is:

$$S_q = \sqrt[3]{\sum_{g:1\rightarrow n} Sim_g^{GSg}} \quad (4)$$

where g is one group of target categories among n groups defined for category q during the scoring process and Sim_g is the similarity value to group g given by the GOTrim method as the specificity of the most immediate ancestor of the query category that is also an ancestor of the target categories in group g.

Using the scoring metric shown in equation 4, the most evidenced and highly similar groups of target categories to the query category are set to have the highest contributions to the query final score, whilst weakly related groups of targets would only contribute minimally. Also, the more frequently a query category is associated with groups of well substantiated target categories, the higher the final score for the category. This implies that by using our proposed

scoring protocol, we are able to detect varying levels of likelihood of biological relevance of query categories promoted by functional analysis.

However, further analysis revealed a possible flaw: arising from the fact that the GS value tends to reach its maximal levels at low similarity with the query. This can cause contributions from weakly related targets to grow artificially high (equation 4); causing in turn a loss of protocol sensitivity. The explanation of this phenomenon may be logically attributed to the complex nature of the biological response to nerve injury at the DRG level (analysed in depth in chapter IV). Thus, among the 560 target categories, reported in literature to be associated with peripheral neuropathy, is a wide spectrum of biological functions ranging from a diverse range of neuronal processes to inflammatory and immune functions. As such, for any given query category, only few target categories may be closely related whilst the vast majority will entail distinct functions. Thus, whilst grouping target categories by the level of similarity to the query category, groups of dissimilar targets are likely to be bigger, leveraging a substantially higher number of genes from published studies; hence a larger GS. This is captured in Figure 6.3.5.

To counteract this effect, the original scoring metric (shown in equation 4) was modified in a way that reduces the effect of the GS values at weak

similarity levels with the query. But, before this was possible, there was a need to define a similarity cut-off below which such a modification may take effect. Whilst no absolute rule exists to define the boundary between strong and weak semantic associations between GO categories, in this work we sought to identify a distinct level of similarity between query and target categories by examining the frequency of similarity values at the background level. This was done by comparing the original population of GO categories, from which the query categories were drawn by functional analysis, against the set of target categories. Such original population consisted of the set of GO categories associated with all genes on the array and for each category in this set, we obtained the maximal similarity value from comparing it to all target categories. The resulting distribution is shown in Figure 6.3.8.

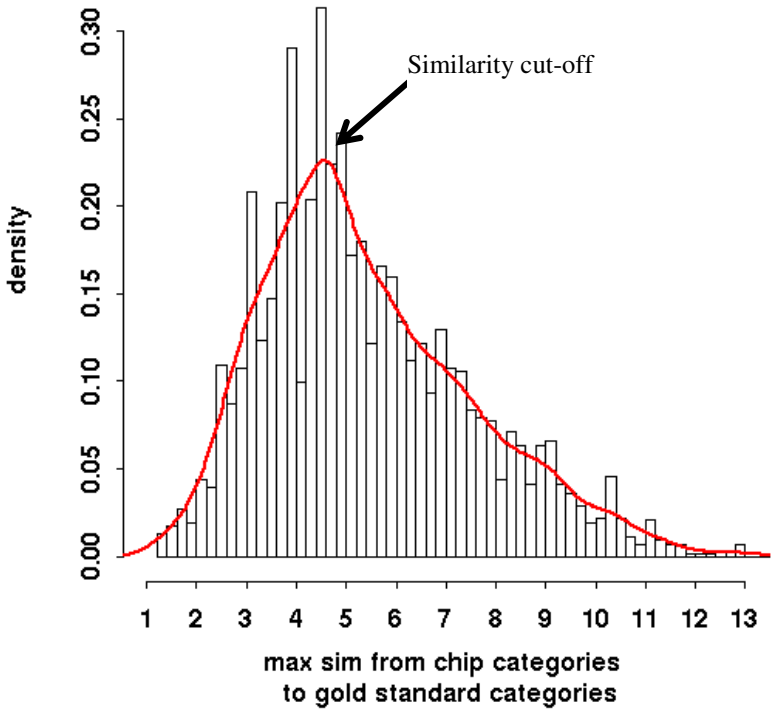


Figure 6.3.8. The distribution of maximal similarity value from comparison of each chip-represented category and the gold standard set of target categories.

Thus, it appears that a considerable proportion of array-associated categories feature a level of similarity to the target categories that is at best below 4.7, which is the value at the peak of the distribution (Fig 6.3.8). These categories may plausibly be taken to represent the substantial proportion of categories on the array expected to entail genomic functions not part of the functional response to nerve injury and hence the weak association with the target categories. On the other hand, the gradual decrease in the frequency of array categories at higher similarity values indicates the significance of this range of similarity. On this basis, we set our similarity cut-off to the value of 4.8 just above 4.7.

After identifying the similarity cut-off and in order to marginalise the effect of increase in GS at low similarity with the query on the query final score (equation 4), a function was developed that reduces the similarity values from groups of weakly related target categories (from below the threshold) to small fractions of less than 1. Likewise, while adding up the similarity^{GS} terms from groups of target categories, those weakly related to the query will have minor contributions to the sum, since in maths, raising a fractional value to any power (no matter how large) always returns a smaller fraction. Moreover, this function was optimised to ensure that such minor contributions from weakly related groups of targets never add up to a value higher than 1. This allows

query categories with no significant (above threshold) association to any of the target categories to be characterised by scores of less than 1 whilst those showing at least one significant association to be distinguished by scores of higher than 1. This is important as it makes sure that the additive effect from weakly related targets may never grow to exert a similar impact on the final score as a contribution from a group of strongly related targets. We refer to this function as the *similarity transformation function* and we explain it in detail in Appendix 6.5.4. For now, we incorporate the transformation function in equation 4 to obtain:

$$S_q = \sqrt[3]{\sum_{g:1\rightarrow n} Tr(Sim)_g{}^{GSg}} \quad (5)$$

S_q is the score for query category q, g is one target group among n groups defined for category q during the scoring process and Tr(Sim) is the transformation function applied on the similarity values Sim.

Of course, one other possibility to suppress the effect of high GS at low similarity with the query is simply by discarding groups of targets at a similarity level below the threshold during the scoring process of the query

category. Although providing a straight forward solution to the problem, such strategy is limited in the case of query categories featuring no similarity relationship with any groups of targets above the threshold, as it will results in these categories receiving no scores. Alternatively, instead of completely discarding groups of target categories from below the threshold, we could have set their similarity values to 0; likewise, any query category not showing a significant association (from above the threshold) to any group of target categories will be given a score based on a sum of 0s amounting to a value of 0. One apparent drawback from this approach is that a row of null scores will be obtained for query categories only weakly related to the target categories; which hinders the derivation of a continuous distribution of scores from a potentially mixed population of query categories, showing both strong and weak associations with the target categories. In this work, we have chosen to implement a more elegant solution that allows a continuous range of scores to be generated for the query categories across the whole range of similarity to the target categories. Importantly, owing to the similarity transformation function incorporated in our scoring metric (equation 5), although the scores from query categories showing both strong and weak association(s) with the target categories run in a continuous range, they segregate into two disparate range of values (above and below 1 respectively); which makes them easily distinguishable.

6.3.2.2. Assessment of the scoring protocol

In order to demonstrate the effectiveness of our scoring protocol to reliably detect optimal similarities with the target categories, we performed the following test. On one hand, we used the scoring protocol to score the set of gold standard target categories against themselves and on the other hand, we used the protocol to score the overall population of array associated GO categories against this set of target categories.

Thus, if we were to compare the distributions of resulting scores from both comparisons and had our scoring protocol been sensitive enough, we would anticipate the former to show high scores (above 1) only whilst the latter to yield a combination of high scores as well as low scores (below 1). This is because in the first comparison, the target categories are being compared to themselves and each category is logically 'highly similar' or more precisely 'identical' to itself (we talk about high similarity instead of identity because the GOTRIM similarity value for a pair of categories is based on the specificity level of the immediate common parent, even when the categories are identical). As for the second comparison, among the set of categories represented on the array, only a fraction will be related to any of the target categories (which we hope to capture with functional analysis) because the

chip is meant to cover the whole set of expressed genes on the rat genome, thereby capturing the whole spectrum of biological functions known to this organism.

Figure 6.3.9 shows the distributions of the resulting scores from both comparisons (red, blue respectively). If we first concentrate on the distribution in blue where the categories from the array were scored against the target categories from the gold standard set, we find that this distribution is bimodal and features two distinct population of scores: one at the low range of below 1 and one at a range higher than 1; in other words, a mixture of low and high scores as anticipated. By contrast, the distribution of scores from the target categories self-comparison shows a slightly different pattern to that expected. Thus, although the vast majority of scores are high (≈ 75%), the remaining population of scores (≈ 25%) are from the low range and the question is how could this possibly occur given that each target category should be at least highly similar to itself? Examination of some of these low self-scoring categories revealed that their semantics are rather general as genes may occasionally be associated with terms that lack precision. Thus, while scoring these categories against themselves, which entails taking the specificity value of their immediate parents, we are bound to drop below the similarity cut-off value of 4.8.

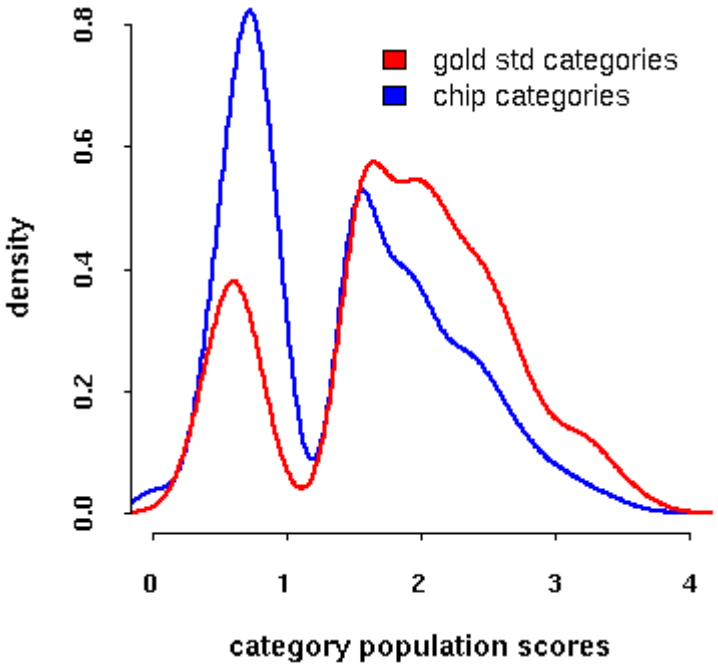


Figure 6.3.9. Distribution of scores from a cross-comparison of target categories from the gold standard set (red) and a comparison of chip associated categories against the latter (blue).

Despite this effect, our scoring protocol is capable of revealing a difference between the two distributions, which is reflected in the larger proportion of high scores from the target categories self-comparison as oppose to when scoring the set of array represented categories against the target categories (75% and 40% respectively).

To highlight the importance of the transformation function to the sensitivity of our scoring protocol, we repeat the same comparisons but this time omitting the transformation function from our scoring metric; in other words, reverting

back to equation 4. The resulting distributions are shown on Figure 6.3.10. Clearly, the distributions of scores from the two comparisons appear to be alike and show no real difference between them, which suggests a loss of sensitivity. This is because, in the absence of the transformation function, the fact that the GS shows typically high values at low similarity with the query causes the similarity^{GS} values from groups of targets weakly and strongly associated with the query to have comparable weights in the final score.

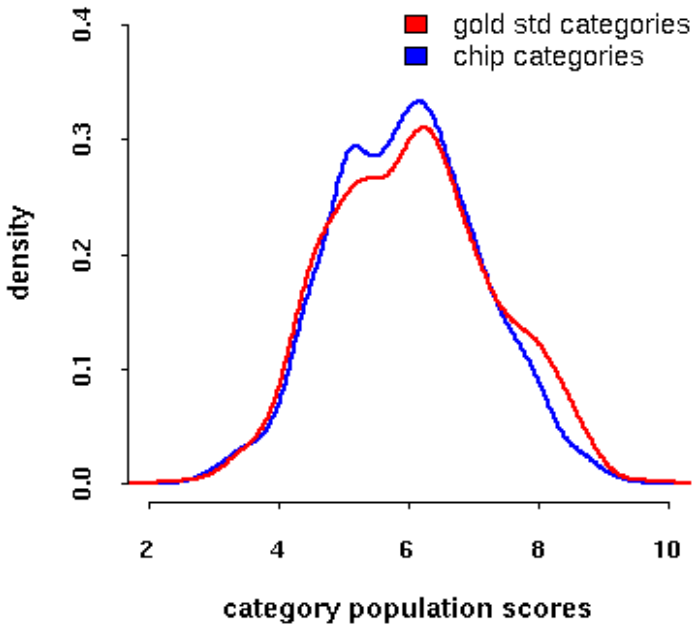


Figure 6.3.10. Distribution of scores from a cross-comparison of target categories from the gold standard set (red) and a comparison of the chip represented categories against the latter (blue), but this time omitting the transformation function from our scoring metric.

6.3.3. Applying the scoring protocol to the results from Catmap, IGA and GSEA functional analysis.

In this section, we present the results from scoring the categories from Catmap, IGA and GSEA functional analysis of the SNT dataset against the set of target categories reported to be enriched, in literature, under similar conditions of peripheral nerve injury, using our scoring protocol. Initially, top categories from each method analysis were processed to remove ancestral categories in order to avoid information redundancy. This was done, as described in the method section, by scanning the list of ranked categories from each analysis top to bottom, each time accummulating the category at the current position or eliminating it if it proves to be a predecessor of a category from higher ranks, until 50 categories were obtained. This meant that 122, 110 and 77 categories were filtered in this manner from Catmap, IGA and GSEA top results before the desired number of categories was obtained for each analysis.

Applying the scoring protocol on this set of top 50 most specialised query categories from each functional analysis returned a distribution of scores for each analysis, shown in Figure 6.3.11. To interpret these distributions, it is important to recall that all scores below 1 correspond to query categories

showing no optimal similarity relationship to any of the target categories whilst higher scores correspond to query categories at an optimal similarity level to one or more groups of target categories.

Clearly, Catmap shows the best performance by yielding the largest proportion of high scores, followed by IGA then GSEA. This effectively means that the top 50 most specialised categories from Catmap show the highest proportion of biologically relevant categories in comparison to the two other methods. Interestingly, Catmap's leading performance has previously been characterised, from a statistical perspective, from examination of the FDR (section 6.3.1.2).

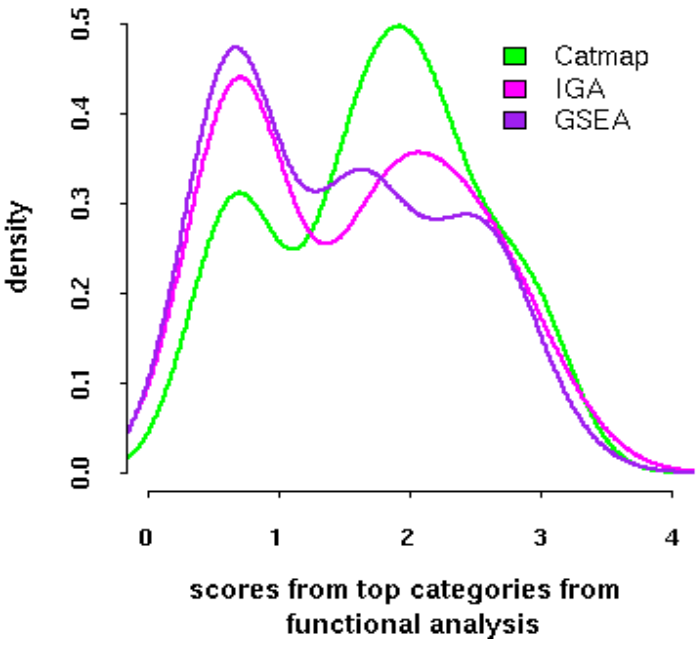


Figure 6.3.11. Distribution of scores from the top 50 most specialised categories from Catmap, IGA and GSEA analysis of the SNT dataset, obtained using our scoring protocol.

Even though the bimodality of the resulting distributions makes it difficult to derive reliable summary statistics for these distributions, we have chosen to use the mean value for the following reasons: There are two main criteria from these distributions that reflect on the performances of the methods and ought to be captured by the summary statistics: one is the proportion of high scoring categories and second is the magnitude of their scores. Thus, beside simply counting the number of query categories from functional analysis showing a similarity level to one (or more) group of target categories above threshold (hence featuring scores greater than 1), it is important to consider the strength of these associations as well as the evidence supporting the target categories involved, both captured by the magnitude of these scores (refer to equation 5 for information on how the scores are derived). Importantly, choosing to use the mean as the summary statistics for the distributions in question allows both criteria to be captured. This is because the mean would tend to increase if the distribution features a higher proportion of the significant scores (meaning scores greater than 1). Moreover, unlike the median, the mean is sensitive to the magnitude of individual scores in the distribution.

The mean scores from Catmap, IGA and GSEA distributions of scores (from Figure 6.3.11) were found equal to 1.74, 1.57 and 1.47 respectively; thus, capturing the anticipated differences in the performances of all three methods.

However, despite their informativeness, these mean score values imply no significance in their own right and one can not tell whether the mean score by Catmap, being the highest, is any different to what could have been obtained by chance. In order to estimate the significance of these mean scores, we examined the range of mean scores obtained under the null hypothesis by applying our scoring protocol on the top 50 categories (processed similarly by removing ancestral categories) from 5000 random lists of categories. A p-value was then derived for each functional analysis method based on the average number of times, the mean score value under the null hypothesis is higher or equal to that from the actual list of categories ranked by the method by evidence of enrichment. The following p-values 0.023, 0.28 and 0.59 were obtained for Catmap, IGA and GSEA respectively; indicating statistical significance for Catmap only.

It is important to note that the random lists of categories used for this analysis were not generated by randomly shuffling the order of the categories in the lists. That is because such null hypothesis would have been inapropriate as it makes the assumption of independency between categories; which is untrue given the fact that a substantial overlap in genes between categories may cause them to show a similar pattern of enrichment. Instead, we used the lists of categories obtained from analysis of permuted list of genes by Catmap, IGA

and GSEA. Importantly, with these lists, the order of categories is biologically insignificant but the dependency between them is certainly maintained.

To further illustrate the difference between these lists of categories randomised in such 'supervised' manner and those obtained from pure random shuffling of categories, we generated 5000 lists of randomly shuffled categories, then applied the scoring protocol on the top 50 most specialised categories from each list. A distribution of 5000 mean score values (a mean score from the distribution of 50 scores from each randomised list) was obtained. This was compared to the distribution of 5000 mean scores obtained from the top 50 most specialised categories from lists of categories by Catmap, IGA and GSEA analyses of permuted gene lists. The results are shown on Figure 6.3.12. There is a clear difference in the mean score value distributions from both types of random lists. Importantly, had we chosen to use lists of randomly shuffled categories to assess the significance of our mean scores from top Catmap, IGA and GSEA results, the p-values would have been underestimated.

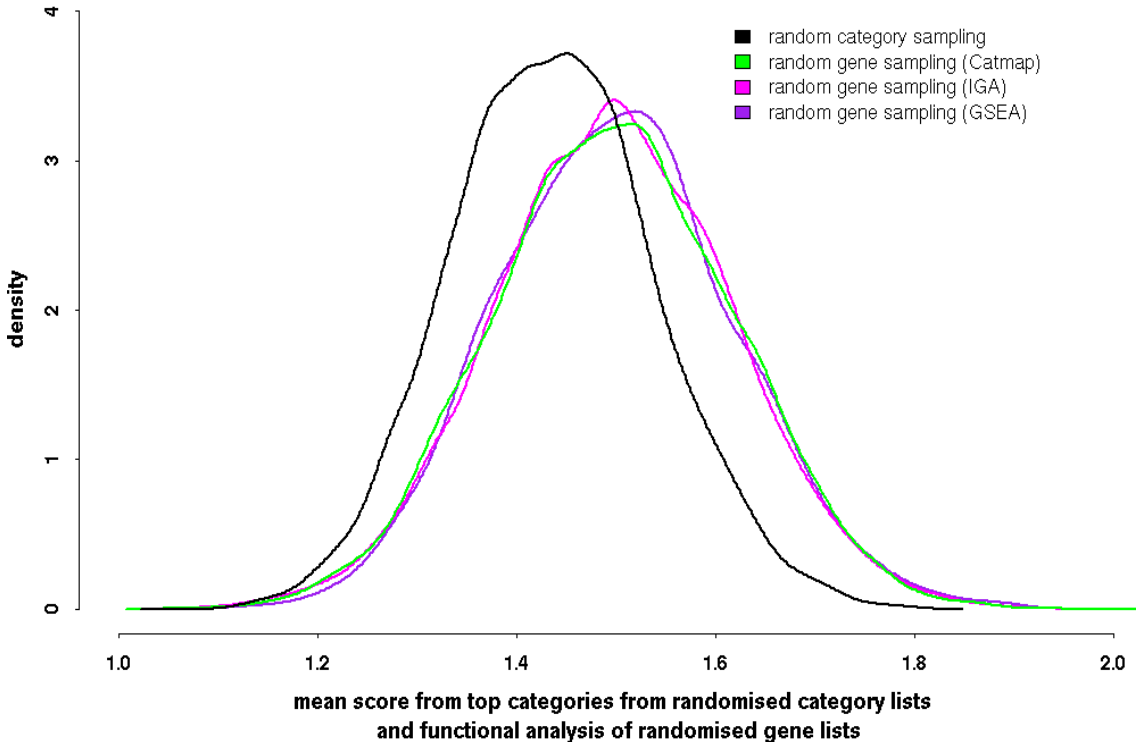


Figure 6.3.12. The distribution of mean scores from running top categories from random category lists, generated either via random shuffling of categories or from functional analysis of randomly permuted gene lists, through our scoring protocol.

6.4. Conclusion

This work has had the important outcome of successfully automating the biological validation of functional analysis methods thereby eliminating the need for human judgement to trace the link between the observed and expected patterns of functional enrichments. There are three main dimensions to our automated validation approach: a test case dataset featuring a well characterised biological phenomenon, a set of manually curated functional categories capturing the range of functions know to be key to the phenomenon under study, serving as 'model answer' and a scoring protocol aiming to capture the level of concordance between the results from functional analysis of the test dataset and the model answer.

Importantly, the proposed automatic validation was used to successfully point out variations in the performance of three different functional analysis methods, some of which are widely used by the microarray reseach community. Further confirmation of the credibility of these conclusions comes from the observation that Catmap's leading performance was also indicated by the FDR, from a statistical perspective. Interestingly, a similar evaluation has previously appeared in the Catmap study (Breslin et al., 2004) whereby the top 10 categories from analysis of the cancer dataset of Van't Veer by Catmap

Wilcoxon statistics, GSEA Kolmogorov statistics and minimized p-values derived in a similar fashion to IGA, were compared. The conclusion was that there was a substantial overlap between the top 10 categories from all three functional analyses, which suggested that they behaved similarly. However, a valid point of criticism for this evaluation is that the range of results considered was marginal and cannot possibly lead to a solid conclusion. Moreover, the model dataset used had many ideal features that are far from common in ordinary expression datasets, most notably the remarkably high number of replicates (51 x 46) and the homogeneity of the profiled tissue. Within such optimal experimental conditions, functional differences are striking, which makes their detection by all three functional analysis methods rather expected and not necessarily indicative of high performance. In this work, we used a rather noisy dataset to achieve a more rigorous evaluation, which has indeed exposed variations between the methods.

One further important conclusion from this work is that in contrast to expectation, the reductionist approach employed by IGA and GSEA, that strives to optimize the score for a category based on selecting the subset of genes in the category most likely to have endured a change in expression, is less robust than the more comprehensive approach employed by Catmap that uses the ranks from all genes member to the category. This could be justified

by the fact that gene expression data are inherently noisy and consequently it may be best to consider all evidence available for the category then to try to distill the most probable piece of evidence for the category.

6.5. Appendices

6.5.1. Functional Category dataset

This appendix explores the process of functional categorisation of genes, which preceded the functional analysis performed in this work. In this work, we used the GO biological process ontology terms as a basis for this categorisation. Thus, genes attributed to the same GO term are members of the same category denoted by the term. Figure 6.4.1 shows the cumulative distribution of the size of obtained categories, whereby for each value x denoting size n, the corresponding y value expresses the proportion of categories with size less or equal to n. Clearly, categories with low gene count are mostly dominant as 75% of all categories appear to have at most 5 gene members.

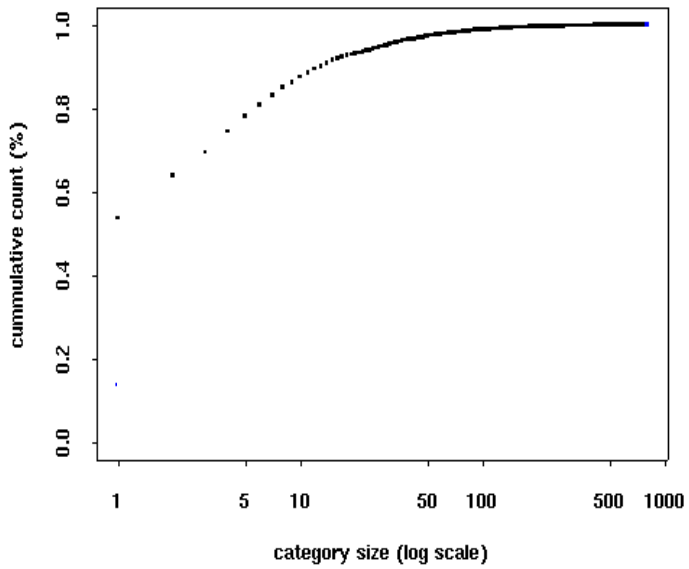


Figure 6.4.1. **Analysis of the gene category dataset**. The cumulative count of category size is shown in black whilst the cumulative count of genes over increasing category size is shown in blue.

The fact that the gene categories are dominated by small categories hinders their use with functional analysis, as it is difficult to deduce any reliable statistics from categories with low gene count. To overcome this problem, we adopted a strategy of back propagating genes to ancestor category terms. Thus, where categories A and B and C have a gene each, their ancestor D would feature a total gene count of three after the back-propagation. As a result of such back-propagation, the average category gene count improved and 75% of all categories were found associated with at least 20 gene members as oppose to 5 originally. However, the inclusion of ancestral term categories as the

result of the back-propagation of genes resulted in a net increase in the overall number of gene categories (from 2907 to 3202), which may exacerbate the problem of multiple testing with functional analysis.

6.5.2. Overview of the low level analysis of the SNT dataset

In order to derive a list of genes ranked by the evidence of differential expression for functional analysis, a series of low-level processing steps were performed on the SNT dataset including quality assessment of the individual arrays, data normalisation, calculation of summary intensity values for individual probesets and finally limma significance analysis of differential expression. In the following, we concentrate on the early step of quality control (QC) and discuss the logic that led to the exclusion of certain array hybridisations from the dataset at this stage of analysis. This is important because the size of the dataset has a major influence on the choice of the null hypothesis during functional analysis. Thus, where only few replicates exist for each condition, the choice of the gene list permutation null hypothesis is inevitable as oppose to the sample label permutation null hypothesis, which is statistically more robust.

Figure 6.4.2 shows scatter plots of raw probe intensity from SNT arrays across and within biological conditions, arranged in a matrix of rows and columns. Each slot in the matrix shows the scatter plot from the pair of arrays indicated on the labels of the column and row defining the slot in the matrix. Using this type of analysis, it was possible to simultaneously assess the level of

consistency between replicate arrays from each group condition during the QC step.

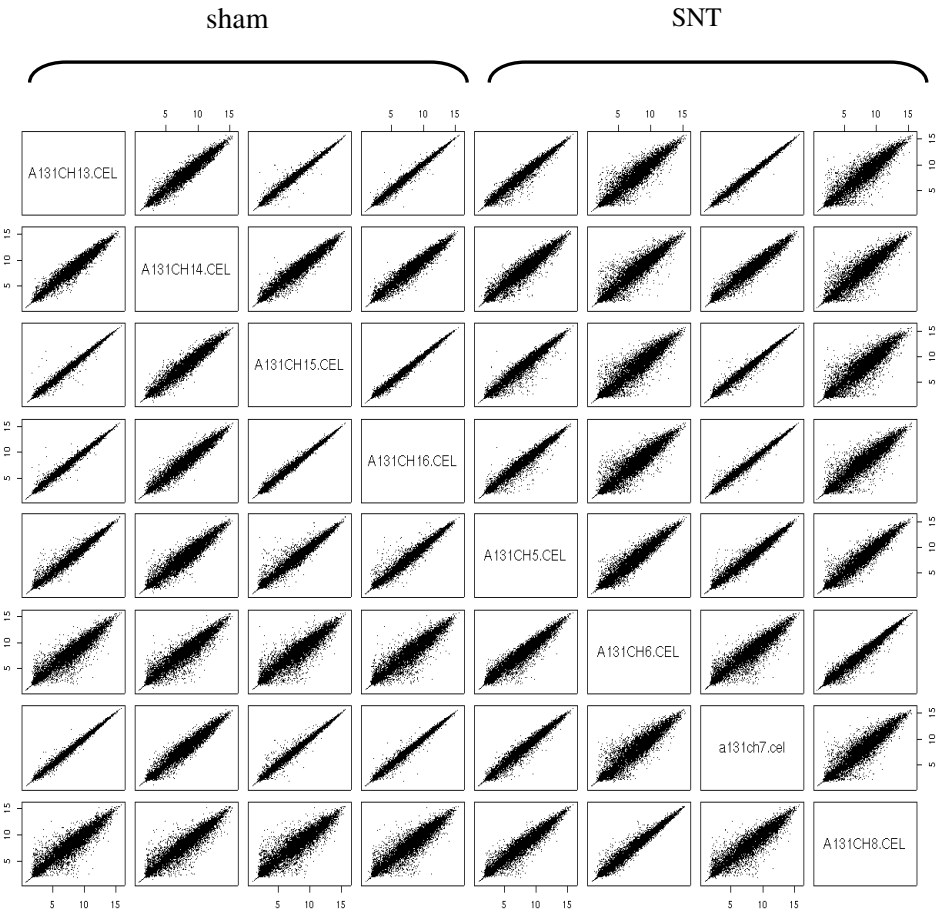


Figure 6.4.2 Quality control of the SNT microarray dataset. Showing scatter plots of raw probe intensities from all pairs of arrays from within and across conditions. Each scatter plot corresponds to the pair of arrays indicated on the labels of the column and row defining the slot featuring the plot. Arrays A131CH5,6,7&8 correspond to SNT whilst A131CH13,14,15&16 correspond to sham.

Visual inspection of the resulting scatter plots (Fig 6.4.2) revealed that with the sham group (columns 1-4), arrays 13, 15 and 16 show a good level of similarity between them but seem to be poorly correlated with array 14 from the same group. On the other hand, the results from the SNT replicate arrays (columns 5-8) indicated an overall lower level of consistency, with array 5 appearing to correlate best with array 7 and array 6 being most similar to array 8. The less perfect nature of the data from the SNT hybridisations is probably justified by the additional variability introduced by the injury to the nerve during the SNT procedure that is absent in the sham.

The above observations were further confirmed using array clustering analysis, which operates by iteratively clustering arrays by decreasing level of similarity in gene intensity. The result is a dendrogram (Fig 6.4.3) where each successive round of coarser clustering corresponds to moving one level up in the dendrogram. The similarity between two arrays is indicated by the lowest level in the dendrogram structure at which they cluster together, which corresponds to the point of fusion of their corresponding branches. The lower the fusion point in the dendrogram, the higher the similarity between arrays.

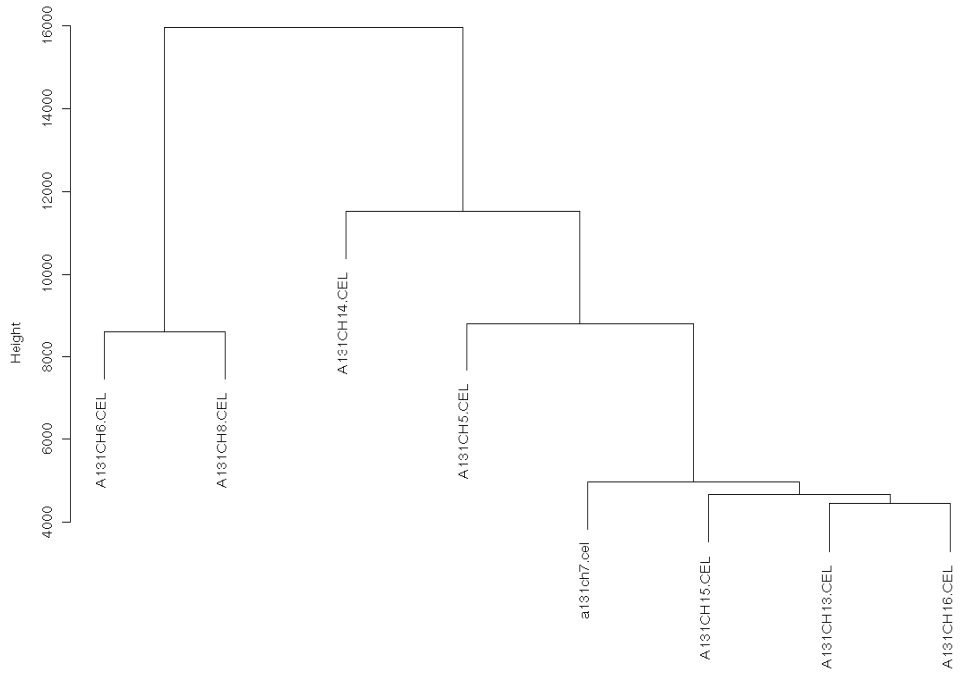


Figure 6.4.3. **Quality control of the SNT microarray dataset featuring array clustering analysis**. Arrays A131CH5,6,7&8 correspond to SNT whilst A131CH13,14,15&16 correspond to sham.

Thus, as before, we note that with the sham replicates, array 14 seems to be more of an outlier. As for the SNT group, we confirm the discrete similarities between arrays 5 & 7 and 6 & 8 respectively. However, we also make the important observation that the former pair of arrays seems to be more closely related to the shams than the latter pair of arrays (also visible but less markedly on Fig 6.4.2 from previous analysis). This suggests that the SNT procedure performed on animals used with arrays 6 and 8 may have had more

pronounced pathological effects, which is desirable from an experimental point of view.

On the basis of these observations from QC, it seemed sensible to ignore array 14 from the sham group and arrays 5 & 7 from the SNT group. To further substantiate these preliminary decisions, we proceeded further in the analysis and performed limma significance analysis of differential expression, both using the whole set of arrays (4 from each condition) and when excluding potentially outlier arrays 14, 5 and 7. In particular, with both analysis scenarios, we studied the correlation between the SNT/sham intensity ratios for individual genes and their corresponding log-odds significance values by limma. Ideally, the more positive and negative the log intensity ratios the more significant they are found, that is the higher the log odds. However, this correlation may be less optimal in situations where the within-group variations are high, thereby masking the significance of the inter-group variations corresponding to intensity ratios. This could be used as a basis to detect potentially noisy arrays in a microarray dataset as removal of such arrays should lead to an improvement in the correlation between the intensity ratios and their log-odds significance values by limma.

The correlation between the intensity ratios and the limma log-odds values from analysing the whole SNT dataset as well as when excluding outlier arrays is shown on Figures 6.4.4-A&B respectively. Clearly, an improvement in the significance levels of the most pronounced intensity ratios (in particular, the negative ones) is achieved via exclusion of outlier arrays; which further justifies the need for their elimination from the dataset.

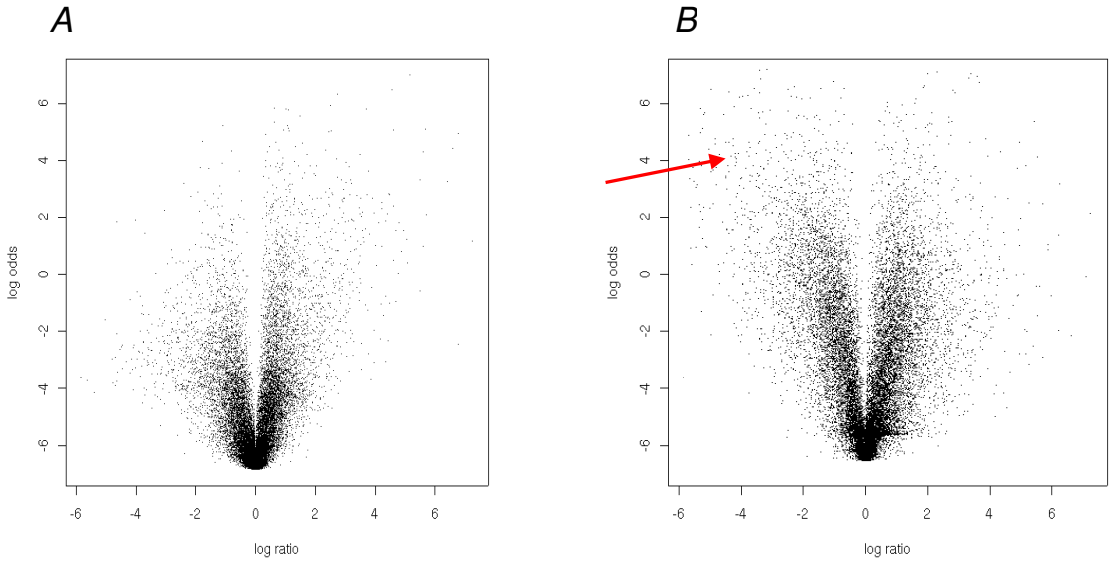


Figure 6.4.4. Plots correlating log intensity ratios with limma log odds significance values. From analysing all arrays in the SNT dataset (A) and when excluding arrays identified as sub-optimal during QC analysis (B). In plot B, the negative fold changes are given higher log-odds significance values (indicated with red arrow) than in plot A.

6.5.3. Additional notes on the GSEA algorithm: GSEA ranking metric

In the standard mode, the GSEA software is designed to accept probeset intensity values and offer a range of statistical metrics with which these probesets may be ranked in a biologically relevant manner. The standard ranking metric by GSEA is the signal to noise ratio consisting of the ratio between the difference in the gene mean expression in phenotypes A and B and the sum of the standard deviation in gene expression from each phenotype replicate samples, that is:

$$\text{Signal-to-noise} = (\mu_A - \mu_B) / (\delta_A + \delta_B)$$

In effect, the signal to noise ratio expresses the correlation between a gene level of expression and either phenotype: the more positive the signal-to-noise ratio the stronger the correlation with phenotype A and the more negative the signal-to-noise ratio the stronger the correlation with phenotype B. Using this metric, the genes are typically ranked by GSEA in a descending order and genes on the top of the list may be considered markers of phenotype A whilst those at the bottom of the list markers of phenotype B.

An important feature of this ranking scheme is that genes from the top as well as the bottom of the list are equally important and categories where member genes cluster at the bottom of the list are as important as categories whose member genes cluster at the top of the list. Remarkably, the GSEA algorithm detects category enrichment in either case whereby the stronger the enrichment at the top of the list, the more positive is the ES whilst the stronger the enrichment at the bottom of the list the more negative is the ES (more details on GSEA ES are available in the introduction part of the chapter, section 6.1.1.3). Importantly, both most positive and most negative ES(s) will be given small p-values; in other words, identified as significant by the GSEA algorithm.

In this work, because the GSEAPreranked option was used to allow our limma ranked gene list to be directly used by GSEA, categories with negative ES were discarded from the analysis because they would only correspond to instances where the category genes are enriched at the bottom of the list for being least differentially expressed.

6.5.4. The similarity transformation function

In this appendix, details of the similarity transformation function, part of the scoring protocol developed in chapter VI for evaluation of functional analysis results (or query categories) against a set of expectedly enriched categories (or target categories), are shown. As explained in chapter VI, the main aim of the transformation function is to shrink the similarity values below the similarity threshold to fractional values of less than 1. Likewise, whilst scoring each query category, the $similarity^{GS}$ terms from groups of target categories showing weak similarity to the query and typically high GS, have minor contributions to the query final score, which is based on summing up these $similarity^{GS}$ terms from all groups of target categories (equation 4 and 5). One way by which this may be achieved is simply by dividing the similarity values by the cut-off value. However, further analysis indicated that such a simple transformation might not be totally suitable for the task at hand.

To illustrate this, we shall examine the following example taken from real data. The tables below correspond to two different query categories and show the original similarity values, the GS values, the results of dividing the similarity values by the cut-off value of 4.8 as well as raising the resulting

values to the corresponding GS values for groups of target categories defined for each query category:

Query 1 'GO:0022402'

Target categories Group no	**Similarity value**	**GS**	**Similarity value / 4.8**	**(Similarity value / 4.8)^{GS}**
Group 1	2.40	4.67	0.52	0.04
Group 2	3.65	5.56	0.79	0.27
Group 3	3.89	3.90	0.84	0.52
Group 4	4.01	3.46	0.87	0.62
Group 5	4.35	2.78	0.94	0.85

Sum = 2.3

Query 2 'GO:0050767'

Target categories Group no	**Similarity value**	**GS**	**Similarity value / 4.8**	**(Similarity value / 4.8)^{GS}**
Group 1	2.40	4.67	0.52	0.05
Group 2	3.65	5.56	0.79	0.27
Group 3	5.03	2.08	1.09	1.19

Sum = 1.51

Thus, unlike query 1, query 2 shows one significant similarity relationship that is just above the threshold of 4.8 (indicated in red). However, attempting to derive a final score for each query category based on summing up the cut-off divided similarity values raised to the GS-th from all groups of target categories returns inappropriately a higher score for query 1 than query 2 (2.3 and 1.51 respectively). Therefore, even though dividing the similarity values by the cut-off value appears to suppress the individual effects from groups of

target categories from below the threshold, together they may add up to a significant value in the final score; possibly superseding the effect from groups of targets at an optimal similarity level with the query, in particular, if just above threshold.

To make sure that our protocol is sensitive enough to distinguish query categories just about significantly related to any of the target categories from those showing no significant similarity to any of the target categories, the transformation function needs to introduce a gap between similarity values above and below the cut-off. This may be achieved by reducing the similarity values below the cut-off to even smaller fractions with many decimal places to make sure that their sum is never going to be above 1, which would guarantee that a category may only receive a score higher than 1 if it features at least one optimal similarity relationship with a group of target categories. Alternatively, similarity values above the cut-off may be set to a higher range of values to make sure that the occurrence of at least one significant similarity, even if borderline, would significantly increase the value of the final score to an extent that is never matched by the added contributions from groups of targets only weakly related to the query.

A search of the literature did not reveal any mathematical function that precisely fits this purpose and no function was found able to create a gap in a continuous range of values as desired. Thus, it became clear that such function ought to be developed as part of this work. After trial and error and by exploring the properties of certain mathematical operations, we reached the mathematical function shown below:

$$Tr(Sim) = X^{[\ log2\ (\ Sim\ )\ -\ log2\ (\ cut\text{-}off\ )\ +\ K\ ]} \qquad (6)$$

where *Tr(Sim)* is the transformed similarity function, S*im* is the original similarity value, *cut-off* is the similarity cut-off which has the value of 4.8, *K* is a parameter with an absolute constant value (later explained) whilst the base *X* serves to optimise the final range of the transformed similarity values, as will be explained later.

Essentially, the transformation function first subtracts the log2 similarity values by the log2 similarity cut-off value. The resulting values, which we would refer to as the '*cut-off subtracted log2 similarity values*' run in a continuous range and shift from negative to positive values at the cut-off point. To create a discontinuity at the point of cut-off, the function adds a

value *K* that has a constant magnitude but variable sign shifting from negative to positive at the cut-off point, similar to the cut-off subtracted log2 similarity values. Effectively, adding *K* to the cut-off subtracted log2 similarity values shifts the positives ones higher in the positive range while causing the negative ones to plunge further in the negative range thereby achieving the desired separation.

- Derivation of the K-factor

K is obtained using the following formula:

$$K = 3 * [\log2 (sim) - \log2(cut\text{-}off)]^{(1 / Klp)} \qquad (7)$$

If we ignore the *Klp* parameter for a moment, we find that *K* is itself derived from the cut-off subtracted log2 similarity values, which explains why *K* has the same sign as these values. However, in this format, *K* is dependent on the original similarity values and is not constant. The reason we want *K* to have a constant magnitude is to make sure that while shifting the cut-off subtracted log2 similarity values by *K*, their order is maintained. This is all achieved by introducing the *K*-factor linearisation parameter *Klp.*

To understand the significance of the power parameter *Klp*, we shall refer to equation (7) and basic mathematics. The fact that the cut-off subtracted log2 similarity values are raised to the inverse of *Klp* implies that we are effectively taking their *Klp*-th root. Recalling the properties of power transformations in maths, we find that the nth root of x tends to 1 as n tends to infinity. As such, increasing the value of *Klp* significantly would cause the cut-off subtracted log2 similarity values to approach unity. Moreover, setting the *Klp* value to an odd number guarantees that the *Klp-root* of the negative cut-off subtracted log2 similarity values approximates –1, thus remaining negative.

In summary, equation (7) operates by borrowing the sign of the cut-off subtracted log2 similarity values and subsequently dumping them by deriving their lowest possible root. The resulting values approximate 1 and -1 where the similarity values are above and below the cut-off respectively and are scaled up via multiplication by 3 (equation 2) to enhance the magnitude of the end result *K*. Figure 6.4.5 illustrates how *K* becomes gradually more linear over the range of similarity values as we increase the value of *Klp*. We fix *Klp* to the value of 301 in equation (7) as it appears to linearise *K* to satisfaction (Fig 6.4.5-D).

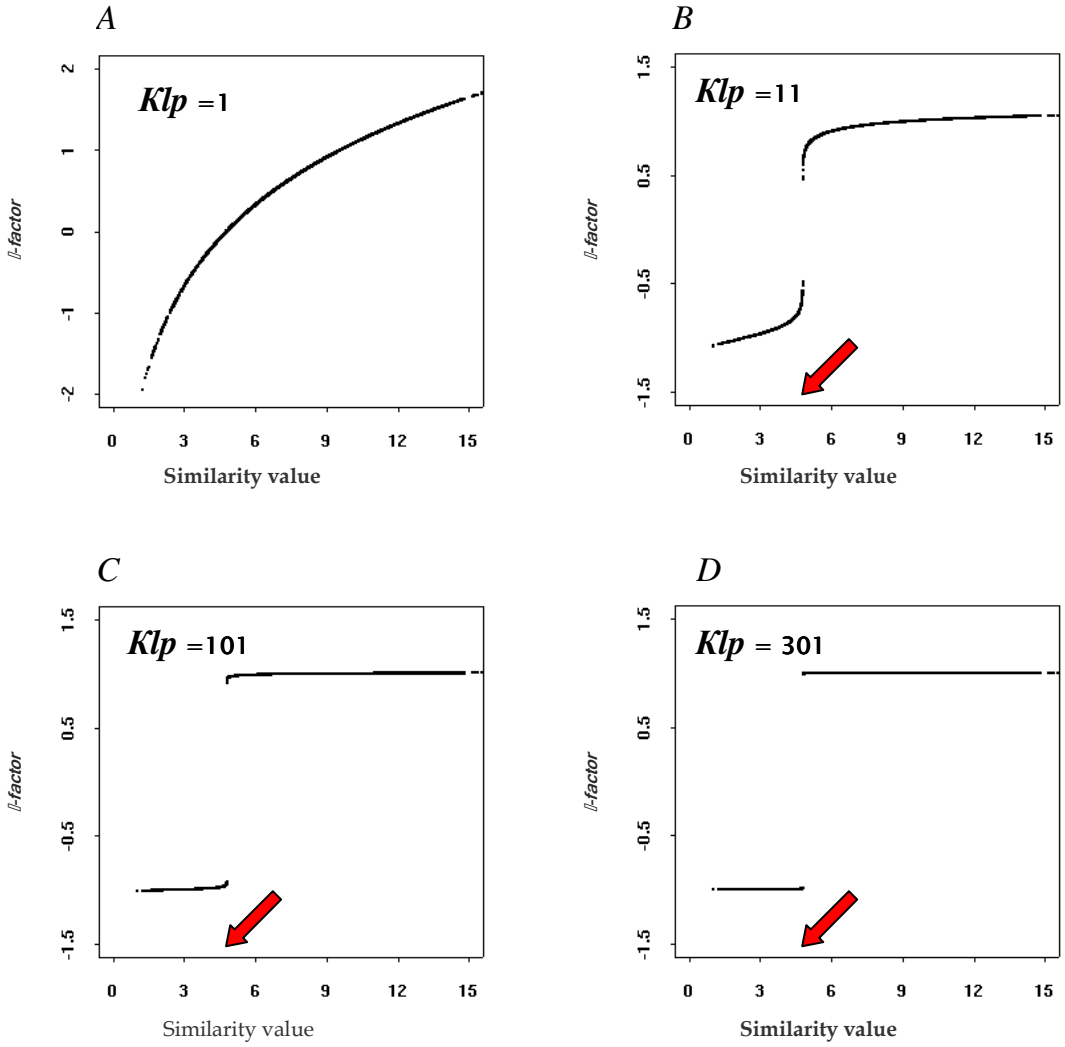


Figure 6.4.5. *K*-factor profile over the range of similarity values for varying *Klp* values. The similarity cut-off is shown with a red arrow.

Referring back to the similarity transformation function outlined in equation (6), it turns out that one more parameter needs to be explained, which is the base X. From what has been discussed sofar, the transformation function operates by subtracting the log2 cut-off value from the log2 similarity values

then shifting them by a constant *K* with the net result of creating two distinct sets of values separated by a gap: one lying in the positive range and one lying in the negative range, corresponding to similarity values above and below the cut-off repectively (Fig 6.4.6-A).

However, the purpose from the transformation function was not only to create a separation between similarity values above and below the cut-off but most importantly, ensure that the similarity values below the cut-off are reduced to the window of fractional values between 0 and 1. This is needed so that during the scoring process of query categories, the similarity^{GS} contributions from groups of targets at a level of similarity with the query below the threshold are marginal and exert no major influence on the final score.

Thus, there is a need to shift the 'cut-off subtracted log2 similarity values + *K',* corresponding to similarity values below threshold, from the negative range to positive fractions of less than 1. To do that, we make use of one more mathematical property, which is the fact that raising a number to a negative value always yields a fractional value. This justifies the use of base *X* in equation 6. In Figure 6.4.6, we correlate the 'cut-off subtracted log2 similarity values + *K'* before and after using them as a power to which *X* is raised with the original similarity values (Fig 6.4.6-A&B, respectively). We would refer to

the latter as the transformed similarity values, as these constitute the end product of the transformation function shown in equation 6. The effect of introducing base *X* is certainly visible on Figure 6.4.6-B in that all similarity values below the cut-off are transformed into fractions of less than 1 and importantly, these are still a distance away from the transformed similarity values from above the cut-off.

Interestingly, increasing the value of *X* helps expand the range of the transformed similarity values most notably those originally above the cut-off. This may be beneficial during the scoring process of query categories as it would mean that contributions from target groups at varying similarity levels with the query from above the threshold are weighted more finely. We set parameter *X* to the value of 1.45 as it appears to expand the range of the transformed similarity values to a reasonable level.

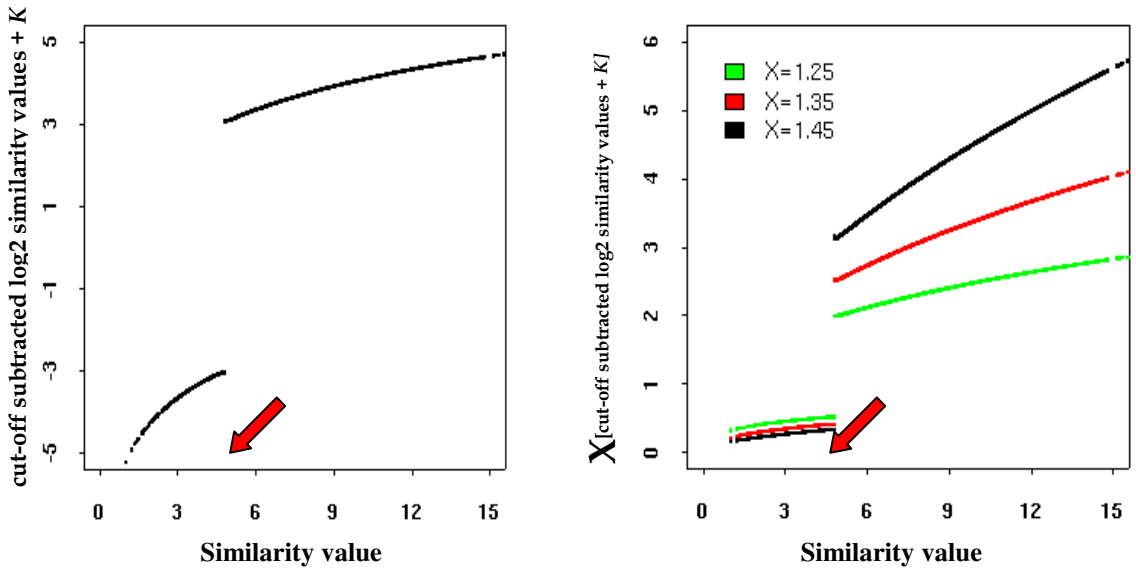


Figure 6.4.6. Highlighting varying steps of the transformation function. (A) The log2 similarity values are subtracted by the log2 similarity cut-off value of 4.8 then shifted by K, which introduces a gap separating similarity values above and below the cut-off. (B) Shows the next and final step in the transformation function where the resulting values from the previous steps are used as a power to which parameter X is raised. This causes the similarity values below cut-off to be confined to the window of fractional values between 0 and 1. The similarity cut-off is shown with a red arrow

Finally, we show a simplified version of the transformation function after incorporating into it the equation of parameter *K*. Thus given equation (6) from above showing the transformation function:

$$TrSim = X^{[\ log2\ (\ sim\ )\ -\ log2\ (\ cut\text{-}off\ )\ +\ K\ ]} \qquad (6)$$

and equation (7) specifying how parameter *K* is derived:

$$K = [\ 1/\ (\ \log2\ (\ sim\ ) - \log2(\ cut\text{-}off\ )\ )\ ]^{\ (\ 1\ /\ K\mathrm{lp}\ )} \qquad (7)$$

incorporating (7) into (6) and simplifying gives:

$$TrSim = X^{[\ y + y^{\ -(\ 1\ /\ K\mathrm{lp}\ )}\ ]} \qquad (8)$$

where $y = \log2\ (\ sim\ ) - \log2(\ cut\text{-}off\ )$

CHAPTER VII: CONCLUSION

Microarray technology offers a fruitful approach to study gene expression patterns in biological systems owing to its high-throughput screening ability. In practise, a microarray study runs through two major phases: experimental, involving the handling of the biological material in its varying forms and its hybridisation onto the physical array and analytical, during which array intensity data are analysed to yield useful biological information. This work has shed light on some important aspects of the technology relating to both implementation phases.

At the experimental level, this work has revealed the consequences of amplifying RNA targets prior to their hybridisation to array probes, which is instrumental in situations where the quantity of the starting biological material is small (chapter I). Importantly, we concluded that amplification can distort the expression ratios between two biological tissues. This was found to happen when distortions in the signal owing to amplification were inconsistent in the two tissues because the intensity falls outside the dynamic range of the scanner. Important conclusions were extrapolated regarding the suitability of the T7 based amplification protocol for microarrays that tie in with the specific experimental design of the microarray experiment.

Following the experimental phase, intensity data are typically run through a pipeline of analytical procedures to extract meaningful biological knowledge. This work has explored one important type of high level analysis methods for microarray data that strives to identify functions potentially enriched in a microarray expression dataset (chapter VI). Crucially, this work has exposed previously unknown variation in the performance of existing methods for microarray functional analysis.

The most striking outcome was the observation that GSEA, the most widely used functional analysis method, performs less well than Catmap: a less popular functional analysis method that has long been undermined by the microarray community; primarily, owing to poor usability. This highlights the importance of robust evaluation protocols in bioinformatics to objectively identify the true merits of methods and algorithms that may correct preconceived notions. On another hand, this work may have the important outcome of promoting Catmap among the microarray research community thereby encouraging efforts by the community to address Catmap usability issues.

As a future aim, the evaluation protocol for functional analysis methods will be applied on a more robust microarray dataset with a higher number of replicates than the SNT dataset used in this work. The reproducibility of

effect, consisting of the superiority of Catmap functional analysis, will further validate the effectiveness and robustness of the evaluation protocol proposed and may encourage its use in the future by the wider microarray bioinformatics community as a standard to assess newly developed functional analysis methods.

In addition to the most forthcoming outcomes of this work, other less apparent, though interesting, conclusions were also made. First, the functional analysis has by large illustrated the positive effect of integrating microarray expression data with other types of useful biological data to yield a more comprehensive picture of the biological phenomenon under investigation. Indeed, it was only by incorporating functional information onto the genes found differentially expressed in our test SNT microarray dataset that the functional consequences of gene expression regulation following nerve lesion in this dataset could be revealed (chapter VI).

Moreover, such integrative approach has the advantage of improving the quality of the biological information gained from microarray experiments. This is particularly important as microarray technology is characterised by a high level of variability owing to its multi-step nature. For example, the SNT microarray dataset used in this work was obtained with amplified RNA material and yet despite the occasional distortions in gene expression

regulation (anticipated from extrapolating the findings from chapter II), following integration with GO functional terms, the picture at the functional level proved rather consistent with what is known in the literature.

At the biological level, this work has had the contribution of highlighting the limitations from screening for pain related genes in animal models of peripheral neuropathy using microarray technology. The complex nature of the molecular response to nerve injury at the level of gene expression makes it difficult to identify pain specific effects. Nonetheless, these limitations can be addressed with good experimental design and the use of tailored downstream datamining approaches. The latter defines research venues that may be appropriately pursued in the future as part of ongoing collaboration with the London Pain Consortium.

Reference List

Aggarwal,B.B. (2003). Signalling pathways of the TNF superfamily: a double-edged sword. Nat. Rev. Immunol. *3*, 745-756.

Al Shahrour,F., Diaz-Uriarte,R., and Dopazo,J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics. *20*, 578-580.

Alexa,A., Rahnenfuhrer,J., and Lengauer,T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. *22*, 1600-1607.

Averill,S., Michael,G.J., Shortland,P.J., Leavesley,R.C., King,V.R., Bradbury,E.J., McMahon,S.B., and Priestley,J.V. (2004). NGF and GDNF ameliorate the increase in ATF3 expression which occurs in dorsal root ganglion cells in response to peripheral nerve injury. Eur. J. Neurosci. *19*, 1437-1445.

Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C., and Apweiler,R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic Acids Res. *37*, D396-D403.

Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Muertter,R.N., and Edgar,R. (2009). NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. *37*, D885-D890.

Baugh,L.R., Hill,A.A., Brown,E.L., and Hunter,C.P. (2001). Quantitative analysis of mRNA amplification by in vitro transcription. Nucleic Acids Res. *29*, E29.

Baxevanis,A.D. (2008). Searching NCBI databases using Entrez. Curr. Protoc. Bioinformatics. *Chapter 1*, Unit.

Benjamini,Y. and Hochberg,Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society *Ser. B 57* , 289-300.

Benjamini,Y. and Yekutieli,D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics ***29***, 1165-1188.

Berriz,G.F., King,O.D., Bryant,B., Sander,C., and Roth,F.P. (2003). Characterizing gene sets with FuncAssociate. Bioinformatics. *19*, 2502-2504.

Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G., Oezcimen,A., Rocca-

Serra,P., and Sansone,S.A. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. *31*, 68-71.

Breitling,R., Amtmann,A., and Herzyk,P. (2004). Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC. Bioinformatics. *5*, 34.

Breslin,T., Eden,P., and Krogh,M. (2004). Comparing functional annotation analyses with Catmap. BMC. Bioinformatics. *5*, 193.

Bridges,D., Ahmad,K., and Rice,A.S. (2001). The synthetic cannabinoid WIN55,212-2 attenuates hyperalgesia and allodynia in a rat model of neuropathic pain. Br. J. Pharmacol. *133*, 586-594.

Bussey,K.J., Kane,D., Sunshine,M., Narasimhan,S., Nishizuka,S., Reinhold,W.C., Zeeberg,B., Ajay,W., and Weinstein,J.N. (2003). MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome Biol. *4*, R27.

Butler,J.M., Ruskell,G.L., Cole,D.F., Unger,W.G., Zhang,S.Q., Blank,M.A., McGregor,G.P., and Bloom,S.R. (1984). Effects of VIIth (facial) nerve degeneration on vasoactive intestinal polypeptide and substance P levels in ocular and orbital tissues of the rabbit. Exp. Eye Res. *39*, 523-532.

Chelala,C., Hahn,S.A., Whiteman,H.J., Barry,S., Hariharan,D., Radon,T.P., Lemoine,N.R., and Crnogorac-Jurcevic,T. (2007). Pancreatic Expression database: a generic model for the organization, integration and mining of complex cancer datasets. BMC. Genomics *8*, 439.

Costigan,M., Befort,K., Karchewski,L., Griffin,R.S., D'Urso,D., Allchorne,A., Sitarski,J., Mannion,J.W., Pratt,R.E., and Woolf,C.J. (2002). Replicate high-density rat genome oligonucleotide microarrays reveal hundreds of regulated genes in the dorsal root ganglion after peripheral nerve injury. BMC. Neurosci. *3*, 16.

Couto,B.R., Ladeira,A.P., and Santos,M.A. (2007). Application of latent semantic indexing to evaluate the similarity of sets of sequences without multiple alignments character-by-character. Genet. Mol. Res. *6*, 983-999.

Dazzi,F., D'Andrea,E., Biasi,G., De Silvestro,G., Gaidano,G., Schena,M., Tison,T., Vianello,F., Girolami,A., and Caligaris-Cappio,F. (1995). Failure of B cells of chronic lymphocytic leukemia in presenting soluble and alloantigens. Clin. Immunol. Immunopathol. *75*, 26-32.

Decosterd,I. and Woolf,C.J. (2000). Spared nerve injury: an animal model of persistent peripheral neuropathic pain. Pain *87*, 149-158.

Demeter,J., Beauheim,C., Gollub,J., Hernandez-Boussard,T., Jin,H., Maier,D., Matese,J.C., Nitzberg,M., Wymore,F., Zachariah,Z.K., Brown,P.O., Sherlock,G., and Ball,C.A. (2007). The Stanford Microarray Database: implementation of new analysis tools and open source release of software. Nucleic Acids Res. *35*, D766-D770.

Diboun,I., Wernisch,L., Orengo,C.A., and Koltzenburg,M. (2006). Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. BMC. Genomics *7*, 252.

Dudoit,S., van der Laan,M.J., and Pollard,K.S. (2004). Multiple testing. Part I. Single-step procedures for control of general type I error rates. Stat. Appl. Genet. Mol. Biol. *3*, Article13.

Dumur,C.I., Garrett,C.T., Archer,K.J., Nasim,S., Wilkinson,D.S., and Ferreira-Gonzalez,A. (2004). Evaluation of a linear amplification method for small samples used on high-density oligonucleotide microarray analysis. Anal. Biochem. *331*, 314-321.

Eaton,M. (2003). Common animal models for spasticity and pain. J. Rehabil. Res. Dev. *40*, 41-54.

Eberwine,J. (1996). Amplification of mRNA populations using aRNA generated from immobilized oligo(dT)-T7 primed cDNA. Biotechniques *20*, 584-591.

Efron,B. and Tibshirani,R. (2002). Empirical bayes methods and false discovery rates for microarrays. Genet. Epidemiol. *23*, 70-86.

Elmquist,J.K., Scammell,T.E., and Saper,C.B. (1997). Mechanisms of CNS response to systemic immune challenge: the febrile response. Trends Neurosci. *20*, 565-570.

Enright,A.J., Kunin,V., and Ouzounis,C.A. (2003). Protein families and TRIBES in genome sequence space. Nucleic Acids Res. *31*, 4632-4638.

Falcon,S. and Gentleman,R. (2007). Using GOstats to test gene lists for GO term association. Bioinformatics. *23*, 257-258.

Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press . 1998.

Gentleman, R. Visualising and distances using GO. http://www.bioconductor.org/repository/devel/vignette/GOvis.pdf . 2005.

Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J., Hornik,K., Hothorn,T., Huber,W., Iacus,S., Irizarry,R., Leisch,F., Li,C., Maechler,M., Rossini,A.J., Sawitzki,G., Smith,C., Smyth,G., Tierney,L., Yang,J.Y., and Zhang,J. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

Hancock,D., Wilson,M., Velarde,G., Morrison,N., Hayes,A., Hulme,H., Wood,A.J., Nashar,K., Kell,D.B., and Brass,A. (2005). maxdLoad2 and maxdBrowse: standards-compliant tools for microarray experimental annotation, data management and dissemination. BMC. Bioinformatics. *6*, 264.

Hosack,D.A., Dennis,G., Jr., Sherman,B.T., Lane,H.C., and Lempicki,R.A. (2003). Identifying biological themes within lists of genes with EASE. Genome Biol. *4*, R70.

Ikeo,K., Ishi-i J, Tamura,T., Gojobori,T., and Tateno,Y. (2003). CIBEX: center for information biology gene expression database. C. R. Biol. *326*, 1079-1082.

Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U., and Speed,T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. *4*, 249-264.

Jiang, J. and Conrath, D. W. Semantic similarity based on Corpus statistics and lexical taxonomy. 1997. In *Proceedings of International Conference Research on Computational Linguistics*.

Jin,X. and Gereau,R.W. (2006). Acute p38-mediated modulation of tetrodotoxin-resistant sodium channels in mouse sensory neurons by tumor necrosis factor-alpha. J. Neurosci. *26*, 246-255.

Joslyn,C.A., Mniszewski,S.M., Fulmer,A., and Heaton,G. (2004). The gene ontology categorizer. Bioinformatics. *20 Suppl 1*, i169-i177.

Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T., and Birney,E. (2004). EnsMart: a generic system for fast and flexible access to biological data. Genome Res. *14*, 160-169.

Kenzelmann,M., Klaren,R., Hergenhahn,M., Bonrouhi,M., Grone,H.J., Schmid,W., and Schutz,G. (2004). High-accuracy amplification of nanogram total RNA amounts for gene profiling. Genomics *83*, 550-558.

Khatri,P., Draghici,S., Ostermeier,G.C., and Krawetz,S.A. (2002). Profiling gene expression using onto-express. Genomics *79*, 266-270.

Kikuchi,M., Tenneti,L., and Lipton,S.A. (2000). Role of p38 mitogen-activated protein kinase in axotomy-induced apoptosis of rat retinal ganglion cells. J. Neurosci. *20*, 5037-5044.

Kim,S.H. and Chung,J.M. (1992). An experimental model for peripheral neuropathy produced by segmental spinal nerve ligation in the rat. Pain *50*, 355-363.

King,C., Guo,N., Frampton,G.M., Gerry,N.P., Lenburg,M.E., and Rosenberg,C.L. (2005). Reliability and reproducibility of gene expression measurements using amplified RNA from laser-microdissected primary breast tissue with oligonucleotide arrays. J. Mol. Diagn. *7*, 57-64.

Klur,S., Toy,K., Williams,M.P., and Certa,U. (2004). Evaluation of procedures for amplification of small-size samples for hybridization on microarrays. Genomics *83*, 508-517.

Lee,D., Grant,A., Marsden,R.L., and Orengo,C. (2005). Identification and distribution of protein families in 120 completed genomes using Gene3D. Proteins *59*, 603-615.

Lee,N., Neitzel,K.L., Devlin,B.K., and MacLennan,A.J. (2004). STAT3 phosphorylation in injured axons before sensory and motor neuron nuclei: potential

role for STAT3 as a retrograde signaling transcription factor. J. Comp Neurol. *474*, 535-545.

Li,L., Roden,J., Shapiro,B.E., Wold,B.J., Bhatia,S., Forman,S.J., and Bhatia,R. (2005). Reproducibility, fidelity, and discriminant validity of mRNA amplification for microarray analysis from primary hematopoietic cells. J. Mol. Diagn. *7*, 48-56.

Li,S., Wu,L., and Zhang,Z. (2006). Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics. *22*, 2143-2150.

Lima,T., Auchincloss,A.H., Coudert,E., Keller,G., Michoud,K., Rivoire,C., Bulliard,V., de Castro,E., Lachaize,C., Baratin,D., Phan,I., Bougueleret,L., and Bairoch,A. (2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res. *37*, D471-D478.

Lin. An information-theoretic definition of similarity . 1998. fifteenth International Conference on Machine Learning.

Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D., and Siani-Rose,M.A. (2003). NetAffx: Affymetrix probesets and annotations. Nucleic Acids Res. *31*, 82-86.

Lord, P. W. Stevens R. D. Brass A. Goble C. A. Semantic Similarity Measures as Tools for Exploring the Gene Ontology. 2003. Proceedings of the 8th Pacific Symposium on Biocomputing (PSB 2003).

Maibaum, M. Rimon G. Orengo C. Martin A Poulovassilis A. BioMap: gene family based integration of heterogeneous biological databases using AutoMed metadata. 2004. Database and Expert Systems Applications.

Maratou,K., Wallace,V.C., Hasnie,F.S., Okuse,K., Hosseini,R., Jina,N., Blackbeard,J., Pheby,T., Orengo,C., Dickenson,A.H., McMahon,S.B., and Rice,A.S. (2009). Comparison of dorsal root ganglion gene expression in rat models of traumatic and HIV-associated neuropathic pain. Eur. J. Pain *13*, 387-398.

Mark Alston, Mark Fernandes, Gary Barker, and Jay Hinton (2004). Handling the Data Deluge: Setting up an Open Source Microarray Database. http://www.rothamsted.bbsrc.ac.uk/bab/conf/array/abs2.php

McClintick,J.N., Jerome,R.E., Nicholson,C.R., Crabb,D.W., and Edenberg,H.J. (2003). Reproducibility of oligonucleotide arrays using small samples. BMC. Genomics *4*, 4.

Meinel,T., Krause,A., Luz,H., Vingron,M., and Staub,E. (2005). The SYSTERS Protein Family Database in 2005. Nucleic Acids Res. *33*, D226-D229.

Miletic,G., Hanson,E.N., and Miletic,V. (2004). Brain-derived neurotrophic factor-elicited or sciatic ligation-associated phosphorylation of cyclic AMP response

element binding protein in the rat spinal dorsal horn is reduced by block of tyrosine kinase receptors. Neurosci. Lett. *361*, 269-271.

Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E., Houstis,N., Daly,M.J., Patterson,N., Mesirov,J.P., Golub,T.R., Tamayo,P., Spiegelman,B., Lander,E.S., Hirschhorn,J.N., Altshuler,D., and Groop,L.C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267-273.

Morrison,N., Wood,A.J., Hancock,D., Shah,S., Hakes,L., Gray,T., Tiwari,B., Kille,P., Cossins,A., Hegarty,M., Allen,M.J., Wilson,W.H., Olive,P., Last,K., Kramer,C., Bailhache,T., Reeves,J., Pallett,D., Warne,J., Nashar,K., Parkinson,H., Sansone,S.A., Rocca-Serra,P., Stevens,R., Snape,J., Brass,A., and Field,D. (2006). Annotation of environmental OMICS data: application to the transcriptomics domain. OMICS. *10*, 172-178.

Nagarajan,V. and Elasri,M.O. (2007). SAMMD: Staphylococcus aureus microarray meta-database. BMC. Genomics *8*, 351.

Navarro,X., Vivo,M., and Valero-Cabre,A. (2007). Neural plasticity after peripheral nerve injury and regeneration. Prog. Neurobiol. *82*, 163-201.

Okuda,S., Yamada,T., Hamajima,M., Itoh,M., Katayama,T., Bork,P., Goto,S., and Kanehisa,M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. *36*, W423-W426.

Pan,F., Chiu,C.H., Pulapura,S., Mehan,M.R., Nunez-Iglesias,J., Zhang,K., Kamath,K., Waterman,M.S., Finch,C.E., and Zhou,X.J. (2007). Gene Aging Nexus: a web database and data mining platform for microarray data on aging. Nucleic Acids Res. *35*, D756-D759.

Paulsen,S.J., Larsen,L.K., Jelsing,J., Janssen,U., Gerstmayer,B., and Vrang,N. (2009). Gene expression profiling of individual hypothalamic nuclei from single animals using laser capture microdissection and microarrays. J. Neurosci. Methods *177*, 87-93.

Pesquita,C., Faria,D., Bastos,H., Ferreira,A.E., Falcao,A.O., and Couto,F.M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. BMC. Bioinformatics. *9 Suppl 5*, S4.

Rabert,D., Xiao,Y., Yiangou,Y., Kreder,D., Sangameswaran,L., Segal,M.R., Hunt,C.A., Birch,R., and Anand,P. (2004). Plasticity of gene expression in injured human dorsal root ganglia revealed by GeneChip oligonucleotide microarrays. J. Clin. Neurosci. *11*, 289-299.

Rada,R. and Bicknell,E. (1989). Ranking documents with a thesaurus. J. Am. Soc. Inf. Sci. *40*, 304-310.

Ramer,M.S., Murphy,P.G., Richardson,P.M., and Bisby,M.A. (1998). Spinal nerve lesion-induced mechanoallodynia and adrenergic sprouting in sensory ganglia are attenuated in interleukin-6 knockout mice. Pain *78*, 115-121.

Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. International Joint Conference for Artificial Intelligence. 1995.

Rhodes,D.R., Yu,J., Shanker,K., Deshpande,N., Varambally,R., Ghosh,D., Barrette,T., Pandey,A., and Chinnaiyan,A.M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. *6*, 1-6.

Robinson,M.D., Grigull,J., Mohammad,N., and Hughes,T.R. (2002). FunSpec: a web-based cluster interpreter for yeast. BMC. Bioinformatics. *3*, 35.

Saal,L.H., Troein,C., Vallon-Christersson,J., Gruvberger,S., Borg,A., and Peterson,C. (2002). BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol. *3*, SOFTWARE0003.

Saeed,A.I., Bhagabati,N.K., Braisted,J.C., Liang,W., Sharov,V., Howe,E.A., Li,J., Thiagarajan,M., White,J.A., and Quackenbush,J. (2006). TM4 microarray software suite. Methods Enzymol. *411*, 134-193.

Schlicker,A., Domingues,F.S., Rahnenfuhrer,J., and Lengauer,T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. BMC. Bioinformatics. *7*, 302.

Scholz,J. and Woolf,C.J. (2007). The neuropathic pain triad: neurons, immune cells and glia. Nat. Neurosci. *10*, 1361-1368.

Schwei,M.J., Honore,P., Rogers,S.D., Salak-Johnson,J.L., Finke,M.P., Ramnaraine,M.L., Clohisy,D.R., and Mantyh,P.W. (1999). Neurochemical and cellular reorganization of the spinal cord in a murine model of bone cancer pain. J. Neurosci. *19*, 10886-10897.

Sevilla,J.L., Segura,V., Podhorski,A., Guruceaga,E., Mato,J.M., Martinez-Cruz,L.A., Corrales,F.J., and Rubio,A. (2005). Correlation between gene expression and GO semantic similarity. IEEE/ACM. Trans. Comput. Biol. Bioinform. *2*, 330-338.

Shah,N.H. and Fedoroff,N.V. (2004). CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. Bioinformatics. *20*, 1196-1197.

Singh,R., Maganti,R.J., Jabba,S.V., Wang,M., Deng,G., Heath,J.D., Kurn,N., and Wangemann,P. (2005). Microarray-based comparison of three amplification methods for nanogram amounts of total RNA. Am. J. Physiol Cell Physiol *288*, C1179-C1189.

Smith,C.L., Goldsmith,C.A., and Eppig,J.T. (2005). The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol. *6*, R7.

Smyth,G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Son,S.J., Lee,K.M., Jeon,S.M., Park,E.S., Park,K.M., and Cho,H.J. (2007). Activation of transcription factor c-jun in dorsal root ganglia induces VIP and NPY upregulation and contributes to the pathogenesis of neuropathic pain. Exp. Neurol. *204*, 467-472.

Spellman,P.T., Miller,M., Stewart,J., Troup,C., Sarkans,U., Chervitz,S., Bernhart,D., Sherlock,G., Ball,C., Lepage,M., Swiatek,M., Marks,W.L., Goncalves,J., Markel,S., Iordan,D., Shojatalab,M., Pizarro,A., White,J., Hubley,R., Deutsch,E., Senger,M., Aronow,B.J., Robinson,A., Bassett,D., Stoeckert,C.J., Jr., and Brazma,A. (2002). Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. *3*, RESEARCH0046.

Splendiani,A., Brandizi,M., Even,G., Beretta,O., Pavelka,N., Pelizzola,M., Mayhaus,M., Foti,M., Mauri,G., and Ricciardi-Castagnoli,P. (2007). The genopolis microarray database. BMC. Bioinformatics. *8 Suppl 1*, S21.

Storey,J.D. (2003). The positive false discovery rate: A Bayesian interpretation and the *q*-value. Annals of Statistics *31*, 2013-2035.

Storey,J.D. and Tibshirani,R. (2003). Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A *100*, 9440-9445.

Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., and Mesirov,J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A *102*, 15545-15550.

Surmeli,D., Ratmann,O., Mewes,H.W., and Tetko,I.V. (2008). FunCat functional inference with belief propagation and feature integration. Comput. Biol. Chem. *32*, 375-377.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J., and Church,G.M. (1999). Systematic determination of genetic network architecture. Nat. Genet. *22*, 281-285.

The Gene Ontology Consortium (2008). The Gene Ontology project in 2008. Nucleic Acids Res. *36*, D440-D444.

Theilhaber,J., Ulyanov,A., Malanthara,A., Cole,J., Xu,D., Nahf,R., Heuer,M., Brockel,C., and Bushnell,S. (2004). GECKO: a complete large-scale gene expression analysis platform. BMC. Bioinformatics. *5*, 195.

Thomas,P.D., Kejariwal,A., Campbell,M.J., Mi,H., Diemer,K., Guo,N., Ladunga,I., Ulitsky-Lazareva,B., Muruganujan,A., Rabkin,S., Vandergriff,J.A., and Doremieux,O. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res. *31*, 334-341.

Tian,W. and Skolnick,J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. *333*, 863-882.

Todd,A.E., Orengo,C.A., and Thornton,J.M. (2001). Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. *307*, 1113-1143.

Tofaris,G.K., Patterson,P.H., Jessen,K.R., and Mirsky,R. (2002). Denervated Schwann cells attract macrophages by secretion of leukemia inhibitory factor (LIF) and monocyte chemoattractant protein-1 in a process regulated by interleukin-6 and LIF. J. Neurosci. *22*, 6696-6703.

Tomlinson,C., Thimma,M., Alexandrakis,S., Castillo,T., Dennis,J.L., Brooks,A., Bradley,T., Turnbull,C., Blaveri,E., Barton,G., Chiba,N., Maratou,K., Soutter,P., Aitman,T., and Game,L. (2008). MiMiR--an integrated platform for microarray data sharing, mining and analysis. BMC. Bioinformatics. *9*, 379.

Tusher,V.G., Tibshirani,R., and Chu,G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A *98*, 5116-5121.

Ueda,H. and Rashid,M.H. (2003). Molecular mechanism of neuropathic pain. Drug News Perspect. *16*, 605-613.

Valder,C.R., Liu,J.J., Song,Y.H., and Luo,Z.D. (2003). Coupling gene chip analyses and rat genetic variances in identifying potential target genes that may contribute to neuropathic allodynia development. J. Neurochem. *87*, 560-573.

van de Vijver,M.J., He,Y.D., van't Veer,L.J., Dai,H., Hart,A.A., Voskuil,D.W., Schreiber,G.J., Peterse,J.L., Roberts,C., Marton,M.J., Parrish,M., Atsma,D., Witteveen,A., Glas,A., Delahaye,L., van,d., V, Bartelink,H., Rodenhuis,S., Rutgers,E.T., Friend,S.H., and Bernards,R. (2002). A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. *347*, 1999-2009.

Van Gelder,R.N., von Zastrow,M.E., Yool,A., Dement,W.C., Barchas,J.D., and Eberwine,J.H. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc. Natl. Acad. Sci. U. S. A *87*, 1663-1667.

Wang,H., Sun,H., Della,P.K., Benz,R.J., Xu,J., Gerhold,D.L., Holder,D.J., and Koblan,K.S. (2002). Chronic neuropathic pain is accompanied by global changes in gene expression and shares pathobiology with neurodegenerative diseases. Neuroscience *114*, 529-546.

Wang,J.Z., Du,Z., Payattakool,R., Yu,P.S., and Chen,C.F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics. *23*, 1274-1281.

Wang,L.X. and Wang,Z.J. (2003). Animal and cellular models of chronic pain. Adv. Drug Deliv. Rev. *55*, 949-965.

Wiggins,A.K., Wei,G., Doxakis,E., Wong,C., Tang,A.A., Zang,K., Luo,E.J., Neve,R.L., Reichardt,L.F., and Huang,E.J. (2004). Interaction of Brn3a and HIPK2

mediates transcriptional repression of sensory neuron survival. J. Cell Biol. *167*, 257-267.

Wilson,C.L., Pepper,S.D., Hey,Y., and Miller,C.J. (2004). Amplification protocols introduce systematic but reproducible errors into gene expression studies. Biotechniques *36*, 498-506.

Woolf,C.J. (2004). Dissecting out mechanisms responsible for peripheral neuropathic pain: implications for diagnosis and therapy. Life Sci. *74*, 2605-2610.

Wright,D.E. and Snider,W.D. (1995). Neurotrophin receptor mRNA expression defines distinct populations of neurons in rat dorsal root ganglia. J. Comp Neurol. *351*, 329-338.

Xia,X., McClelland,M., and Wang,Y. (2005). WebArray: an online platform for microarray data analysis. BMC. Bioinformatics. *6*, 306.

Xiao,H.S., Huang,Q.H., Zhang,F.X., Bao,L., Lu,Y.J., Guo,C., Yang,L., Huang,W.J., Fu,G., Xu,S.H., Cheng,X.P., Yan,Q., Zhu,Z.D., Zhang,X., Chen,Z., Han,Z.G., and Zhang,X. (2002). Identification of gene expression profile of dorsal root ganglion in the rat peripheral axotomy model of neuropathic pain. Proc. Natl. Acad. Sci. U. S. A *99*, 8360-8365.

Yang,L., Zhang,F.X., Huang,F., Lu,Y.J., Li,G.D., Bao,L., Xiao,H.S., and Zhang,X. (2004). Peripheral nerve injury induces trans-synaptic modification of channels, receptors and signal pathways in rat dorsal spinal cord. Eur. J. Neurosci. *19*, 871-883.

Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S., Bussey,K.J., Riss,J., Barrett,J.C., and Weinstein,J.N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. *4*, R28.

Zhang,X., Xu,Z.O., Shi,T.J., Landry,M., Holmberg,K., Ju,G., Tong,Y.G., Bao,L., Cheng,X.P., Wiesenfeld-Hallin,Z., Lozano,A., Dostrovsky,J., and Hokfelt,T. (1998). Regulation of expression of galanin and galanin receptors in dorsal root ganglia and spinal cord after axotomy and inflammation. Ann. N. Y. Acad. Sci. *863*, 402-413.

Zhu,Y., Zhu,Y., and Xu,W. (2008). EzArray: a web-based highly automated Affymetrix expression array data management and analysis system. BMC. Bioinformatics. *9*, 46.

Zimmermann,M. (2004). [Neuronal mechanisms of chronic pain]. Orthopade *33*, 515-524.

Zimmermann,P., Schildknecht,B., Craigon,D., Garcia-Hernandez,M., Gruissem,W., May,S., Mukherjee,G., Parkinson,H., Rhee,S., Wagner,U., and Hennig,L. (2006). MIAME/Plant - adding value to plant microarrray experiments. Plant Methods *2*, 1.