Evaluating Markov Chain Ontology Analysis as a Standalone Enrichment Analysis Method

Alex Hawkins-Hooker

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2017

Abstract

Biological ontologies such as the Gene Ontology play an important role in organizing knowledge about the biological function of genes. Enrichment analysis of ontology terms is commonly performed to identify the functions most strongly associated with a list of genes. Enrichment analysis methods achieve this by using known associations of genes and ontology terms, registered within ontology databases as 'annotations' of genes to terms. The traditional 'over-representation analysis' approaches which treat each term as an independent gene set to be separately tested for enrichment using a statistical test ignore the structure of the ontology and are not designed to incorporate information about annotation confidence. Markov Chain Ontology Analysis (MCOA) is a recently proposed enrichment analysis method that promises to overcome these limitations by modelling the ontology together with annotations from a set of genes of interest as a Markov Chain (Frost and McCray, 2012). This project explores the use of MCOA as a standalone enrichment analysis method. Alternatives to the procedure for computing enrichment scores from the steady state distribution of the Markov Chain used by Frost and McCray (2012) are proposed and evaluated on simulated and real datasets. An implementation of MCOA using these scoring procedures was developed and shown to outperform over-representation analysis in finding enriched Gene Ontology terms on simulated datasets. The ability of MCOA to incorporate gene and annotation weights was explored using a set of real human disease datasets collected by Tarca et al. (2012) and a set of weighted annotations to the Human Disease Ontology (He, 2016). Incorporating annotation weights was found to improve the ability of MCOA to prioritize relevant disease terms on these datasets.

Acknowledgements

I am extremely grateful to my supervisor, Ian Simpson, for his generous support and suggestions throughout the project. Thanks also to Xin He for providing access to his database of annotations to the HDO.

Declaration

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

(*Alex Hawkins-Hooker*)

Table of Contents

1 Introduction **1**

2 Background **3**

2.1 Ontology Term Enrichment Analysis 3

2.1.1 Ontologies in biology 3

2.1.2 Gene lists for enrichment analysis 4

2.1.3 Over-representation analysis 4

2.1.4 Other Enrichment analysis approaches 5

2.2 Personalised Pagerank 7

3 MCOA **10**

3.1 MCOA as application of Personalised Pagerank to hybrid Gene-Ontology Graphs 10

3.1.1 The hybrid Gene-Ontology Graph 10

3.1.2 The personalization vector 11

3.1.3 Enrichment Scores 12

3.1.4 Incorporating annotation weights as weighted gene-term edges 12

3.1.5 Interpretation: random walker with restart 14

3.2 MCOA as a standalone method for enrichment analysis 15

3.2.1 Computing enrichment scores 15

3.2.2 Enrichment Scoring Using Null Models of the Target Graph . 17

3.2.3 The damping factor 19

3.3 Implementation Details 20

4 Evaluation **21**

4.1 Analysis of effect of model choices on enrichment ranking on a real dataset 21

4.1.1 Method 21

4.1.2 Results 22
4.1.3 Discussion 24
4.2 Benchmarking performance using Simulated Data 24
4.2.1 Unweighted genes and edges 25
4.2.2 Using KEGG Pathways to improve selection of enriched terms for simulation studies 31
4.2.3 Simulating datasets with weighted annotations 33
4.2.4 Discussion 36
4.3 Benchmarking performance on real human disease datasets using the Human Disease Ontology 38
4.3.1 Method 38
4.3.2 Results 40
4.3.3 Discussion 43

5 Conclusion **44**

Bibliography **47**

Chapter 1

Introduction

Advances in experimental techniques in the past few decades have made it possible to sequence entire genomes as well as to perform experiments which are capable of simultaneously measuring the expression levels of thousands of genes. Deriving biological insight from such experiments relies on interpreting the long lists of genes that result. One way of doing this is by trying to establish which biological functions and processes are most strongly associated with the genes in the list. Enrichment analysis is a set of statistical methods which uses known associations between genes and terms in biological knowledge bases such as the Gene Ontology (Ashburner et al., 2000) to detect those terms which are significantly 'enriched' in a list of genes. Although simple statistical tests have been used for this purpose, making best use of all the information available about the genes and their relations to the knowledge base terms of interest remains a challenging problem, for a number of reasons. The biology associated with genomes and genetic expression profiles is complex, and operates at many scales from the molecular to the phenotypic. As a result, expression levels of individual genes will typically be correlated to different degrees depending on the genes' involvement in multiple processes operating at a range of scales. Associations between genes and terms in knowledge-bases are uneven in coverage, not always reliable, and ambiguous in the sense that the nature or strength of the association may differ widely from case to case. Knowledge bases themselves are often highly structured leading to correlations between terms which may have an impact on any attempt to use statistical methods to perform enrichment analysis. Finally, it is hard to establish the superiority of one method over another, since the task of enrichment analysis - to identify the terms that are most descriptive of a list of genes - is inherently ambiguous, and there is no available 'gold standard' by which to benchmark the performance of different methods.

The result of this complexity is that a wide variety of enrichment analysis techniques have been proposed, with no clear consensus over which is the best.

Markov Chain Ontology Analysis (MCOA, Frost and McCray (2012)) is one of a number of more recent methods that seek to incorporate the hierarchical graph structure of biological ontologies into the enrichment analysis procedure. MCOA, inspired by the way personalised pagerank uses the graph structure of the web to retrieve topic-specific search results, accomplishes this by modelling the ontology together with the set of genes being investigated as a Markov Chain. This approach is promising because it allows the possibility of including not just the dependencies between terms encapsulated in the overall structure of the ontology, but also the confidence associated with gene-term annotations, which may well be important in the application of enrichment analysis to terms in ontologies that are less well established and less actively curated than the Gene Ontology, such as the Human Disease Ontology.

Despite the promise of MCOA, its capabilities are not well studied, and its use as a standalone enrichment method has not been explored or rigorously tested against other methods. We therefore propose to implement a standalone version of MCOA and test its capabilities. We are particularly interested in its possible advantages relative to other methods, in particular its use of the structure of the ontology, and its capability to handle weighted gene-term annotations.

At heart, the MCOA method introduced by Frost and McCray (2012) is an application of personalised pagerank (Haveliwala, 2003) to the enrichment analysis problem. I have developed an implementation of MCOA based on an open source implementation of the personalised pagerank algorithm (in the Python library NetworkX). In keeping with this way of understanding MCOA, chapter 2 will first provide an overview of enrichment analysis approaches, then describe personalised pagerank as a general method for finding node importance in graphs. Chapter 3 will introduce MCOA as an adaptation of personalised pagerank to enrichment analysis, and explore alternatives to the enrichment scoring procedure used by Frost and McCray (2012) in their original presentation of the algorithm. Chapter 4 seeks to evaluate the suitability of MCOA as a standalone method for enrichment analysis on real and simulated datasets.

Chapter 2

Background

2.1 Ontology Term Enrichment Analysis

2.1.1 Ontologies in biology

An ontology is a structured representation of entities in a particular domain and the relationships between them. The Gene Ontology (GO) is the most widely used ontology in biology, and is an attempt to systematically represent knowledge about the properties of gene products in a structured form. The ontology consists of a set of terms each of which represents either a molecular function, a biological processes, or a location in the cell, and a set of relations between these terms. Genes are associated with terms by means of 'annotations' supplied by either human curators or computer algorithms, based on research into gene function. Such a gene-term annotation represents the involvement of the gene or its products in the biological property represented by the term.

Although the GO is the best known example, there are a large number of biological ontologies representing knowledge in different domains. Gene enrichment analysis techniques can equally well be applied to any such knowledge base, as long as associations of genes to terms in the knowledge base are available. Although most of this report is concerned with enrichment analysis of GO terms, in chapter 4 I will also consider the use of MCOA to perform enrichment analysis of terms from the Human Disease Ontology (Schriml et al., 2012), which is an ontology consisting of terms relating to human diseases and their relationships.

2.1.2 Gene lists for enrichment analysis

The input to an enrichment analysis method is a list of genes of interest, typically found as the output of some kind of genomic profiling experiment. For example, differential expression profiling experiments typically measure the level of gene expression for thousands of genes across two biological conditions, such as diseased vs healthy tissue samples (Robinson and Bauer, 2011). The results of such experiments are average expression levels for each gene in each condition. Enrichment analysis is commonly employed to help interpret these results by first selecting some proportion of the genes which show the most significant variation in expression level across the two conditions as genes of interest, then running enrichment analysis of GO terms method to identify the terms most strongly associated with these differentially expressed genes.

2.1.3 Over-representation analysis

The classic approach to enrichment analysis, which is still commonly used, is over-representation analysis. In this approach, ontology terms are treated as independent and are individually tested for enrichment. A term is regarded as enriched with respect to a list of genes of interest if the number of genes of interest annotated to the term is higher than would be expected by chance, given the total number of genes annotated to the term. The significance of the enrichment is encapsulated in a p-value calculated using a statistical test, typically Fisher's exact test. These p-values are then adjusted to account for the fact that multiple terms are being tested, increasing the likelihood of a chance false positive. This adjustment is made by applying a standard multiple testing correction, such as Benjamini-Hochberg (Benjamini and Hochberg, 1995).

Using Fisher's exact test, the procedure for calculating the raw p-value for a term T is as follows. Let X be a random variable denoting the number of genes of interest belonging to T, and let K be the total number of genes belonging to T. Let n be the number of genes of interest and N be the total number of genes with annotations to any ontology term (this is the relevant population of genes for this test). The probability of X genes belonging to T out of a list of n genes chosen at random without replacement from the population is given by the hypergeometric distribution, with the initial probability of drawing a gene belonging to the term being $\frac{K}{N}$. That is:

$$X \sim hypergeom(N, K, n) \tag{2.1}$$

Then if a set of genes of interest of length n contains k members of T, T is found to be

significantly over-represented in the genes of interest if $P(X \geq k) \leq \alpha$, where α is the significance level of the test.

In what follows references will sometimes be made to 'the classic method' for enrichment analysis. This should be taken as shorthand for the over-representation analysis procedure using Fisher's exact test described here.

2.1.4 Other Enrichment analysis approaches

Huang et al. (2009) distinguish between three classes of method for enrichment analysis of gene lists. Singular enrichment analysis (SEA) methods test for enrichment of each term separately. The canonical example of this approach is the over-representation analysis method outlined above. There is, however, a whole family of methods which follow this approach, with different choices of statistical significance test, multiple-testing correction, and the use of further corrections to, for example, reduce the number of redundant results due to ontology terms being associated to highly overlapping sets of genes.

Gene set enrichment analysis (GSEA) methods seek to address some deficiences of SEA approaches. For example, when using an SEA approach to analyse the results of a differential expression profiling experiment, it is necessary to first select some number of the genes whose expression levels were profiled as being the most interesting, passing only the list of interesting genes as input to the SEA method. As described in section 2.1.2, this is usually done by identifying the most significantly differentially expressed genes. However, the cutoff between genes in this category and genes not deemed to be significantly differentially expressed is arbitrary, and discards information about the relative degrees of differential expression of all the genes. GSEA methods circumvent this problem by using the differential expression levels to rank the full list of profiled genes. Genes at the top and bottom of the list represent the genes most strongly over- and under- expressed in one condition relative to the other. The intuition behind GSEA is that ontology terms whose associated genes tend to appear at either of the extremes of this ranked list are likely to be of greater functional importance in explaining the biological differences between the two conditions that those whose genes tend to appear towards the middle, and therefore show little difference in expression.

In the canonical GSEA approach (Subramanian et al., 2005), for each ontology term to be tested, an enrichment score is computed which measures the tendency of

that term's genes to appear at the top or bottom of the list. This score is the maximum deviation from 0 of a cumulative statistic calculated while moving down the gene list, with the contribution from each gene being the correlation of the gene with the first condition if the gene is associated with the term being tested, and -1 otherwise. This score is tested for significance, producing p-values to which a multiple-testing correction is applied.

While GSEA's focus on sets of genes takes into account the fact that small differences in differential expression observed across a significant proportion of a set of correlated genes may provide a signal that is missed by SEA, both GSEA and SEA fail to take into account the fact that ontology terms are also strongly correlated, as expressed by the ontology's hierarchical structure. This structure constitutes an important additional source of information. Modular enrichment analysis (MEA) methods are those which seek to incorporate the dependencies between terms, such as those encapsulated in the structure of ontologies, into enrichment tests.

As an example of this kind of approach, Alexa et al. (2006) propose a method which uses the hierarchical structure of the ontology as the basis for a correction to the standard SEA approach designed to reduce redundancies owing to the fact that parent terms inherit all the annotations from their children, leading to highly overlapping and correlated sets of genes accruing to terms appearing in parent-child relationships in the ontology. This dependency between ontology terms leads to highly biased results if not integrated into the enrichment scoring process. Alexa et al. (2006) remove this problem by preventing parent terms from inheriting annotations from child terms that are found to be significantly enriched, using a standard SEA-style significance testing procedure. Doing this for all terms involves processing the ontology's graph representation in a bottom-up manner. Although this takes account of inter-term dependencies, it uses only local graph structure (the parent-child relationship) to do so, thereby overlooking long range dependencies. It also fails to include dependencies between terms that are not encapsulated in the graph structure, such as highly overlapping categories occuring in different parts of the graph.

A second class of MEA methods represent a departure from the shared framework of all the SEA, GSEA, and MEA methods discussed so far in which each ontology term is separately tested for significance, in favour of a global approach in which all terms are considered at the same time. One important advantage of treating all terms at once is that the effect of term overlaps can be included within the method rather than as a post-hoc correction. Term overlaps are a strong biasing factor in other methods as

discussed in Ballouz et al. (2017). The GenGO (Lu et al., 2008) and MGSA (Bauer et al., 2010) algorithms achieve this by inferring a probabilistic model of the process by which the gene lists which are the subject of the analysis are generated. Lu et al. (2008) achieve this using a simple model according to which in any given experiment a small set of terms is allowed to be active. The goal of the algorithm is to find the set of active terms which maximise a likelihood function which is contructed in such a way as to discourage the appearance of highly overlapping categories in the maximum likelihood term set. MGSA likewise assumes that the observed gene lists can be explained by a set of active terms, but instead of seeking the active set that minimizes a constructed loss function, models the generative process as a Bayesian network and attempts to infer the active set using approximate probabilistic inference.

Although the exact details of the MGSA and GenGO models differ, their fundamental motivation is the same: to identify, using a model-based approach, a small set of enriched terms, thereby minimizing the number of false positives returned as a result of term overlaps. One limitation that is shared by the two methods is that they do not incorporate the hierarchical structure of the ontology, thus treating the overlaps between related terms in the same way as those between unrelated ones. MCOA (Frost and McCray, 2012) was proposed as a modification to the GenGO algorithm designed to incorporate the ontology structure. MCOA relies on a graph representation of the ontology, with additional edges representing gene-term annotations. This graph is then used to define a Markov Chain, allowing a steady-state probability distribution to be computed over the nodes, representing the importance of those nodes relative to a gene set of interest given the structure of the ontology together with its annotations. These probabilities can either be used to adjust the scoring of another enrichment method, such as GenGO, as was originally proposed by Frost and McCray (2012), or on its own as a standalone measure of enrichment, which is the usage we will seek to explore in this project.

2.2 Personalised Pagerank

The MCOA approach to enrichment analysis is based on the personalised pagerank algorithm (Haveliwala, 2003). In this section I will introduce this algorithm as a general algorithm for computing node importance in a directed graph. In chapter 3 I will introduce MCOA as a variant of personalised pagerank adapted to the enrichment analysis problem.

Let D be a directed graph with N nodes $\{n_i\}$, and let A be its adjacency matrix, so that.

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge from node i to node j} \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

Further, let $L(i)$ be the total number of edges starting on node i. Then we can define a stochastic matrix M:

$$M_{ij} = \begin{cases} \frac{A_{ij}}{L(i)}, & \text{if } L(i) > 0 \\ \frac{1}{N}, & \text{otherwise} \end{cases} \tag{2.3}$$

Personalised pagerank works by using the graph to define a Markov Chain, with one state for each node in the graph, and an irreducible transition matrix, M' computed from the adjacency matrix. The transition matrix of a Markov Chain defines its dynamics, with a probability distribution vector u over states in the chain evolving according to (Meyer, 2000):

$$u^T(t+1) = u^T(t)M' \tag{2.4}$$

Given an irreducible transition matrix, arbitrary initial distributions $u(0)$ over the states of the chain allowed to evolve according to (2.4) are guaranteed to converge on a single stationary distribution v, satisfying (Meyer, 2000):

$$v^T = v^T M' \tag{2.5}$$

These steady-state probabilities can naturally be interpreted as a measure of the importance of states in the chain. In the personalized pagerank formulation, the transition matrix of the Markov Chain, M', is computed from the normalized adjacency matrix, M, of the graph as follows (Langville and Meyer, 2005):

$$M'_{ij} = \alpha M_{ij} + (1-\alpha)p_j \tag{2.6}$$

Where p is a 'personalization vector' with one value for each node, such that $\sum_i p_i = 1$. The output of personalised pagerank is the set of steady state node probabilities, or 'pageranks', v_i found by solving equation (2.5) numerically given the transition matrix defined in equation (2.3). Since the matrix M' represents the probabilities of transitions between states in the chain, the two terms in equation (2.6) represent two different types of permitted transition: transitions to adjacent states following the edge

structure of the graph D, as encoded in the normalized adjacency matrix M, and transitions to any other state made randomly according to the distribution represented by the personalization vector. The steady state distribution v represents the long run distribution over states in the chain given transitions of these two kinds between states, with the parameter α controlling the relative probability of each kind of transition. Thus a nonuniform personalization vector p can be used to bias the steady state distribution towards a particular set of nodes, by encouraging more frequent transitions to these nodes than would be achieved by following the structure of the graph alone. For example, in the original personalized pagerank approach, the personalization vector was used to bias web search results towards a particular topic, by setting the personalization vector entries of out-of-topic pages to zero (Haveliwala, 2003). This leads to an interpretation of the steady state probabilities, or 'pageranks' v_i of nodes as a measure of the importance of those nodes relative to the topic in question.

Chapter 3

MCOA

3.1 MCOA as application of Personalised Pagerank to hybrid Gene-Ontology Graphs

The output of personalised pagerank on a graph is a probability distribution over the nodes in the graph. The individual node probabilities can be interpreted as the importance or prominence of each node given the structure of the graph. The personalization vector can be used to bias these importances towards a set of 'seed' nodes of interest. This approach can be applied to enrichment analysis, by constructing a graph representing the set of genes of interest, the ontology terms to be tested for enrichment, and the gene-term annotations, and using the personalization vector to bias the pagerank calculation towards the gene nodes. The pageranks that are found as the solution to equation (2.5) on this graph can then be used to calculate enrichment scores for each term, which can be returned as the output of the enrichment analysis procedure. In this section, I will explain each of these aspects of the MCOA approach to enrichment analysis, as well as how annotation confidence scores can be incorporated.

3.1.1 The hybrid Gene-Ontology Graph

The first step in the MCOA algorithm is the construction of a graph whose nodes represent either ontology terms or genes of interest. I will use different notation in this section to distinguish between the two sets of nodes: term nodes $T = \{t_i\}$, and gene nodes $G = \{g_i\}$. The overall set of graph nodes is then $\{n_i \in G \bigcup T\}$. In the simplest version of the algorithm, two types of unweighted edge are allowed:

1. Term-term edges representing hierarchical ontology relationships (for example the Gene Ontology relationships is-a and part-of). These edges start on the child, and end on the parent.

2. Gene-term edges representing gene-term annotations. These edges start on the gene and end on the ontology term to which the gene is annotated.

A different graph is constructed for each run of the algorithm, since the list of genes of interest determines which genes and gene-term annotations should be included as nodes. In detail, given a particular list of genes of interest, a graph is constructed by taking the following steps. First, the graph is initialized simply as a tree representation of all terms in the ontology and their hierarchical relationships, with edges from child terms to parent terms as described above. Next, genes in the list of genes of interest without annotations to the ontology are discarded. The remaining genes of interest are added to the graph as nodes. For every annotation between a gene of interest and a term in the ontology, a corresponding unweighted edge is added to the graph. Finally, term nodes which are not the endpoint of at least one path starting on a gene node are discarded. This last step is necessary to ensure that the transition matrix computed from the graph according to the equations outlined in the previous section is irreducible. It has the additional benefit of reducing computation time.

3.1.2 The personalization vector

A nonuniform personalization vector is used to bias the calculation of ontology term importances towards the set of genes of interest.

$$p_j = \begin{cases} \frac{1}{|G|}, & \text{if } n_j \in G \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

Differences in importance amongst these genes can in turn be included by weighting the gene nodes. Gene weights can be used to incorporate continuous valued information about the importance of the gene in question, such as a t-statistic from a differential expression measurement (Frost and McCray, 2012). Then letting $w(n_j)$ denote the weight of a gene node n_j:

$$p_j = \begin{cases} \frac{w(n_j)}{\sum_{n_i \in G} w(n_i)}, & \text{if } n_j \in G \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

3.1.3 Enrichment Scores

Let v be the steady state probability vector, computed as the solution to equation (2.5). v's probability mass is distributed over both term and gene nodes. Let $v(t_i)$ denote the steady-state probability at the ith term node, t_i. Since in enrichment analysis we are only interested in computing the importance of the term nodes relative to the gene nodes, the first post-processing step used by Frost and McCray (2012) is to compute adjusted term probabilities $e(t_i)$, by normalizing over the terms (in what follows, I will refer to $e(t_i)$ as the 'termrank' of term t_i, to reflect the fact that these are pagerank scores normalized over ontology terms):

$$e(t_i) = \frac{v(t_i)}{\sum_k v(t_k)} \tag{3.3}$$

Instead of simply computing termranks $e(t_i)$ for all terms, and reporting these values as term enrichment scores, Frost and McCray (2012) compare the termranks computed on two graphs: a 'target' graph, in which the gene nodes are the gene nodes of interest, and a 'background' graph, in which the gene nodes constitute a background gene set (for example, all genes in the genome with at least one annotation to an ontology term). Let $e_{tar}(t_i)$ be the termrank, computed according to (14), on the target graph, and $e_{bgr}(t_i)$ be the termrank computed on the background graph. Using these adjusted term probabilities, Frost and McCray (2012) compute a score $s(t_i)$ for each ontology term:

$$s(t_i) = \frac{e^2_{tar}(t_i)}{e_{bgr}(t_i)} \tag{3.4}$$

In the approach of Frost and McCray (2012), this score is not reported directly as an enrichment score, but instead used to add a term to the cost function of the GenGO algorithm. In this project, I will consider the use of (3.4) as a final enrichment score, and will refer to this scoring procedure as 'MCOA background' from now on, to distinguish it from alternatives considered over the course of this project.

3.1.4 Incorporating annotation weights as weighted gene-term edges

Annotations of genes to ontology terms are made based on evidence of different types and reliability. While annotations to the Gene Ontology are actively curated, sourcing annotations to less well established ontologies such as the Human Disease Ontology (HDO) will typically require the use of text mining techniques with little scope for

manual curation. In this context it is natural to associate a confidence score to each annotation, as has been done by He (2016) in compiling the Human Disease Gene Database (HDGDB), a database of confidence-weighted annotations to HDO. The hybrid gene-term graph employed in MCOA means that annotation weights can naturally be incorporated into the algorithm as graph edge weights, in analogy with approaches to using pagerank with weighted edges (Xing and Ghorbani, 2004). This implies only a slight modification to the matrix M (equation (2.3)) that encodes the edge structure of the graph:

$$M_{ij} = \begin{cases} \frac{W_{ij}}{\sum_j W_{ij}}, & \text{if } \sum_j W_{ij} > 0 \\ \frac{1}{N}, & \text{otherwise} \end{cases} \tag{3.5}$$

where W is the weighted adjacency matrix of the graph, with W_{ij} representing the weight of an edge from node i to node j. The term-term edges all have weight 1, while the gene-term edges are weighted according to the annotation weights. Equation (3.5) handles annotation weights by normalising over the total outgoing weight at each node, to ensure that the total weight of all the outedges of a node i is 1. In practice, this means that the effect of the gene-term edge weights is relative at each gene: so a gene with two annotations each of weight 0.1 and a gene with two annotations each of weight 0.9 are indistinguishable in this model - the entries in matrix M for each of these edges will be 0.5. Thus simply adding weighted edges to the graph to represent annotation confidence fails to take into account differences across genes in the overall reliability of the annotations associated with them. In cases where these differences are large, therefore, it may be useful to encode these differences as gene weights . This can be achieved by associating with each gene a weight equal to the average weight of the annotations to that gene:

$$w(g_i) = \frac{1}{L(g_i)} \sum_j W_{ij} \tag{3.6}$$

$L(g_i)$ here is the number of terms annotated to gene g_i. In cases where another continuous value is associated with each gene to encode, for example, differential expression levels, as described above, the two values can simply be multiplied together.

3.1.5 Interpretation: random walker with restart

The probability distribution encoded in the pagerank vector v (equation (2.5)) represents the long-run distribution over states sampled by a random walker on the nodes in the gene-term graph, starting at an initial node picked from an arbitrary distribution $u(0)$, and making transitions between nodes according to the transition matrix M'. As explained in the presentation of personalised pagerank, this matrix encodes two types of transition: (i) transitions between nodes which are connected by a directed edge, such as between a gene and one of the terms to which it is annotated, or between a term and one of its parents, and (ii) 'restarts': transitions to non-connected nodes sampled from the distribution encoded in the personalization vector. As a result the role of the personalization vector is to bias the random walker towards a set of seed states, which are favoured by the personalization distribution. In the case of MCOA, these seed states are the gene nodes, so balancing the two types of transition means that typical behaviour for the random walker is to start on a gene node, follow one of its annotations to an ontology term, move up the ontology for a few steps, then jump back to another gene node, follow one of its annotations, and proceed as before. In this way, the pattern of ontology nodes sampled by the random walker will tend to indicate the strength of the association of those nodes with the genes of interest.

The 'damping factor', α, controls the relative likelihood of these two types of transition. In the limit $\alpha = 1$, the random walker only makes transitions according to the graph structure. Given that the ontology-term graph is a tree, this means that the random walker will always end up at the root, since all paths lead to this node, and none lead out of it. In the other limit $\alpha = 0$, the random walker simply samples states according to the personalization vector p, and as a result the steady state probability distribution of random walker states is trivially the personalization vector, $v = p$.

It is important to note that, while the random walker interpretation is useful for understanding the calculation, there is no need to actually simulate the behaviour of an ensemble of random walkers to calculate the distribution v: this distribution can be found more quickly by solving (2.5) numerically, but would be the distribution that a hypothetical ensemble of random walkers converged on.

3.2 MCOA as a standalone method for enrichment analysis

In Frost and McCray (2012)'s presentation of the MCOA algorithm, the output of the Markov Chain computation was a score for each ontology term designed to be included in the cost function of the GenGO model (Lu et al., 2008). This section addresses two algorithmic issues raised by instead treating MCOA as a standalone enrichment analysis method. First, whether alternatives to the term scoring procedure used by Frost and McCray (2012) might better suit the requirements of a final enrichment score; second, the effect of the choice of the main model parameter, the 'damping factor' α.

3.2.1 Computing enrichment scores

A central requirement of an enrichment scoring procedure is that it should make it possible to distinguish between ontology terms that score highly because of a genuine association with the genes of interest, and ontology terms that score highly because of other factors, such as the structure of the ontology, or the total number of annotations to the term. In the classic approach, which uses a Fisher's exact test to test for overrepresentation of individual ontology terms in a particular set of genes, this is achieved by asking whether the number of annotations to a term in a query gene set is higher than would be expected at random, given the total number of genes annotated to the term, and the size of the gene set.

The raw steady state term probabilities, however, are naturally biased towards terms with more total (direct and indirect) annotations. These terms will tend to occur higher up the ontology, with many incoming paths from the genes and few outgoing paths. As a result a random walker following the structure of the ontology will tend to sample these terms more often than terms lower in the hierarchy with fewer paths leading to them. Although the specific preference for high level (low depth) terms can be mitigated by using a lower value of α, even for low values of α a leaf term which is annotated to most genes will tend to score highly for most input gene lists, even if a relatively small proportion of its total genes are present in any one.

Figure 3.1 highlights this bias, by visualizing the distribution of term probabilities as a function of total annotations when MCOA was run on a real dataset, a list of genes whose products were found to be more abundant in human as opposed to mouse post-synaptic density (Bayes et al., 2012). There is a clear correlation between high

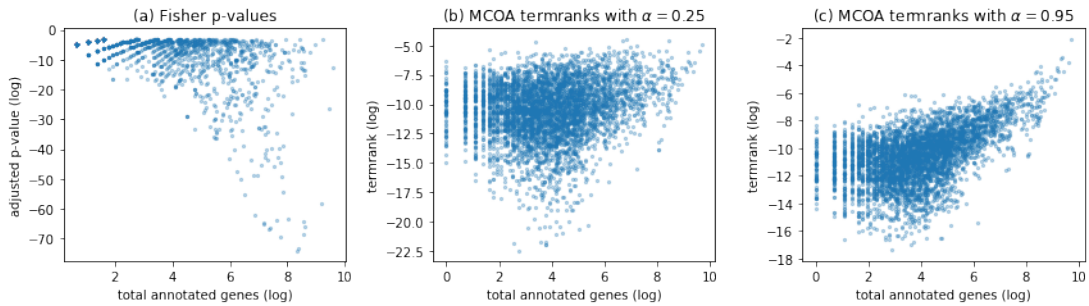

Figure 3.1: Comparison of distribution of GO biological process term Fisher p-values and termranks as function of total number of genes annotated to the term. All three plots shown here were obtained by running enrichment analysis on the same dataset, a set of 186 genes whose products were found to be relatively more abundant in the human post synaptic density (Bayes et al., 2012). a) shows adjusted p-values obtained by running Fisher's exact test on all terms, then correcting for multiple testing by applying a Benjamini-Hochberg correction. Only terms found to be significantly enriched ($p < 0.05$) are shown. b) and c) show distributions of MCOA termranks obtained using $\alpha = 0.25$ and $\alpha = 0.95$ respectively.

annotation count and high term probability, with the terms with more than around 400 annotations all having above average term probability. Clearly not all terms with more than 400 annotations are enriched, so this plot makes plain the unsuitability of the raw term probability as an enrichment score. In contrast, the classic method (left panel) shows a far more desirable-looking distribution, with a spread of p values at all total annotation counts. Although having more annotations obviously shouldn't imply enrichment, it does make more confident assessments of the likelihood of enrichment possible, since the higher the number of total annotations to a term, the higher the statistical power of a Fisher test on that term. This explains the greater spread along the y axis with increasing total number of annotated genes.

A crucial question for an algorithm based on the Markov Chain computation that outputs these probabilities is, therefore, how to correct for the bias towards high-annotation-count terms. What is missing from the calculation is a measure of the typicality of a particular term's steady state probability relative to an input gene list. If a particular term tends to score highly irrespective of the input, then a high score on a particular set of genes of interest is not attributable to any special association between the term and the genes in question, and the term should probably not be reported as enriched. An unbiased enrichment analysis method needs to take account of this, so that terms with relatively few annotations but several of those annotations present can be retrieved as enriched as well as terms with larger numbers of overall annotations.

The scoring procedure 'MCOA background' introduced in section 3.1.3 represents the approach taken by Frost and McCray (2012) to address this problem. This is not the only possible approach, however, and it is of interest to explore its merits compared to other options, especially since it was used to generate scores to be input to a second algorithm, and as such perhaps did not have to fully satisfy the requirement of good enrichment scores by itself. This section considers two alternative scoring procedures, 'MCOA ratio' and 'MCOA z' explored over the course of the project. Both are based on the idea of comparing the termranks obtained from a given input gene list to a distribution of termranks found by running the Markov Chain computation on randomised gene lists.

3.2.2 Enrichment Scoring Using Null Models of the Target Graph

A general approach to estimating the typicality of some feature of the nodes in a particular graph is to construct a 'null distribution' over the feature, by constructing randomised versions of the graph with similar structural features (Maslov et al., 2004). These null distributions then provide the requisite measure of the typicality of features in graphs of particular types.

In the enrichment case, two primary structural features of Gene-Term graphs that may explain distributions of termranks are the number of gene nodes and the structure of the ontology. Accordingly, we can assess the typicality of a particular termrank by comparing it to the distribution of termranks for the same term in Gene-Term graphs with the same ontology structure, but randomised sets of gene nodes. As an example of how this works, consider the case of performing enrichment analysis of Gene Ontology terms on an input list of 300 human genes. In this case, generating a null sample involves randomly selecting 300 human genes from the set of all human genes with annotations to at least one Gene Ontology term, constructing a graph from this random set of genes according to the procedure outlined in 3.1, finding the Markov Chain steady state distribution for this graph, and then using this to compute a termrank for each term according to equation (3.3). Having constructed a null distribution of termranks for each term by collecting the results of many such null samples, we can assess the typicality of the observed termrank for a given term on the 'target graph' constructed from the true input by comparing it to the expected termrank for that term across our set of null samples. One possible approach is to simply take the ratio of the observed term probability and its null mean:

$$\text{'MCOA ratio':} \qquad s_{\mu}(t_i) = \frac{e_{tar}(t_i)}{E_{null}[e(t_i)]} \tag{3.7}$$

We could also consider incorporating information about the variance of the term probability, by using a statistic such as the z-score rather than a simple ratio:

$$\text{'MCOA z':} \qquad s_z(t_i) = \frac{e_{tar}(t_i) - E_{null}[e(t_i)]}{\sigma_{null}[e(t_i)]} \tag{3.8}$$

where $\sigma_{null}[e(t_i)]$ is the standard deviation of the termrank for term t_i across the null samples.

3.2.2.1 Null Models for Weighted Gene Lists

The assumption behind a null modelling strategy based on the randomisation of input gene lists is that it is the identities of the genes in the input list which encode the crucial information about which biological functions might be enriched. With weighted gene lists this is not necessarily the case, since it may be the weights, rather than the identities of the genes in the list which are important. For example, the results of differential expression profiling studies can be summarized as a list of genes each associated with a continuous value encoding the degree and direction of differential expression of that gene between two conditions. The identities of the genes in this list are determined by the nature of the experimental equipment used to measure gene expression levels, and may include most or all of the genes in the genome. It is the differential expression levels of these genes which encode the biologically relevant information about biological differences between the two conditions. As a result, if the full list of genes and weights based on differential expression levels is to be supplied as input to MCOA, it makes sense to construct null samples not by randomising the identities of genes in the list, but instead by randomising the gene weights.

Given such a weighted input gene list, then, each null sample will be generated by retaining the same list of genes, but randomising the gene weights. To achieve this the following procedure will be followed. First, the gene weights on the true input list will be calculated. Then each null sample will be constructed by randomly redistributing this set of gene weights across the genes, and calculating the termranks given these randomised weights. Thus the distribution of gene weights is the same in each null sample: it is the assignment of weights to genes that is randomised.

As before, the null distribution of termranks generated by running the Markov Chain computation on these sets of random gene-weight assignments can then be used

to compute enrichment scores, according to either of the scoring functions 'MCOA ratio' and 'MCOA z' defined above.

3.2.3 The damping factor

The role of the parameter α in personalized pagerank is most easily interpreted as the probability that a random walker on a graph makes a transition to its next state by following an out edge from its current state rather than taking a random jump to another state according to the distribution encoded in the personalization vector. In the application of personalized pagerank to the gene-term graph, then, α controls the number of edges to parent terms that will be followed by a random walker before it jumps back to another gene node. Alternatively, it controls the form of the drop off in the weight accorded to an indirect annotation to a term relative to a direct annotation, as a function of the distance between the term and its ancestor to which the direct annotation is associated. This is a potentially attractive feature built into the MCOA method: Frost and McCray (2012) speak of the method's accommodating a notion of 'semantic distance' which is ignored by standard enrichment analysis approaches. As a result, the choice of α is an important question to be decided. Frost and McCray (2012) use a value of $\alpha = 0.85$, possibly because this is the value originally used by Google when calculating pagerank on the web graph (Brin and Page, 1998). However, this value was justified based on the specific requirements of the web search case. In general, higher values of α lead to stronger dependence on the link structure of the graph and also slower convergence, while lower values lead to stronger dependence on the distribution encoded in the personalization vector (Ding et al., 2009).

Since α directly controls the number of steps up the ontology random walkers will tend to take, a heuristic approach to choosing α might simply be based on a consideration of the distribution of this number of steps under different values of α, and a judgement of what seems appropriate given the ontology in question. At every transition, the random walker has the same fixed probability $1-\alpha$ of jumping back to one of the starting states, rather than progressing up the ontology in accordance with the gene-term graph structure. Therefore the number of steps up the graph before a restart follows a geometric distribution with mean $\frac{1}{1-\alpha}$ (White and Smyth, 2003). So a value of $\alpha = 0.85$, as chosen by Frost and McCray (2012), means that, on average, a walker will take 6.667 steps up the gene-term graph before restarting.

3.3 Implementation Details

I have developed a Python implementation of MCOA, with the option to select any of the three scoring functions described above. The implementation is based on the Python graph library NetworkX: first a NetworkX graph object is constructed to represent the hybrid gene-ontology graph, then NetworkX's personalized pagerank function is run on that graph to compute pageranks for each node, from which the enrichment scores can be computed. The implementation works with two ontologies: the Gene Ontology, and the Human Disease Ontology. The Python library goenrich is used to parse the Gene Ontology file and convert it into a NetworkX graph. Computation of null distributions is parallelized using the multiprocessing library.

Chapter 4

Evaluation

4.1 Analysis of effect of model choices on enrichment ranking on a real dataset

In order to develop an understanding of the behaviour of the MCOA algorithm under the various possible choices of scoring function and damping factor, it is useful to look at the results produced on some real input. For this purpose, I used the set of genes found by Bayes et al. (2012) to be differentially expressed in human vs mouse post-synaptic density. This section compares the results produced by the different scoring functions with a particular focus on the distributions of high-scoring terms. These distributions can reveal the kinds of terms that tend to be favoured by a particular enrichment scoring procedure, and therefore serve as a useful means of exploring the characteristics of enrichment methods. In particular this section looks at how the different scoring functions treat terms of different depth, and of different annotation count.

4.1.1 Method

Bayes et al. (2012) report average average relative log-expression levels in mouse and human PSD for 831 gene products, together with corresponding t-test p-values. We focus on the set of genes with higher expression level in humans than mice for which the difference in expression is found to be significant by the t-test at a significance level of 0.05. This results in a list of 186 genes, identified by Entrez ID. 180 of the 186 significantly differentially expressed genes are found to have annotations to terms in the biological process section of the Gene Ontology, and these genes are used as (unweighted) inputs to the enrichment methods. The set of annotations used is the set

of human annotations downloaded from the Gene Ontology website, cross-referenced with the Uniprot Knowledge Base ID mapping to create a set of annotations searchable by Entrez ID.

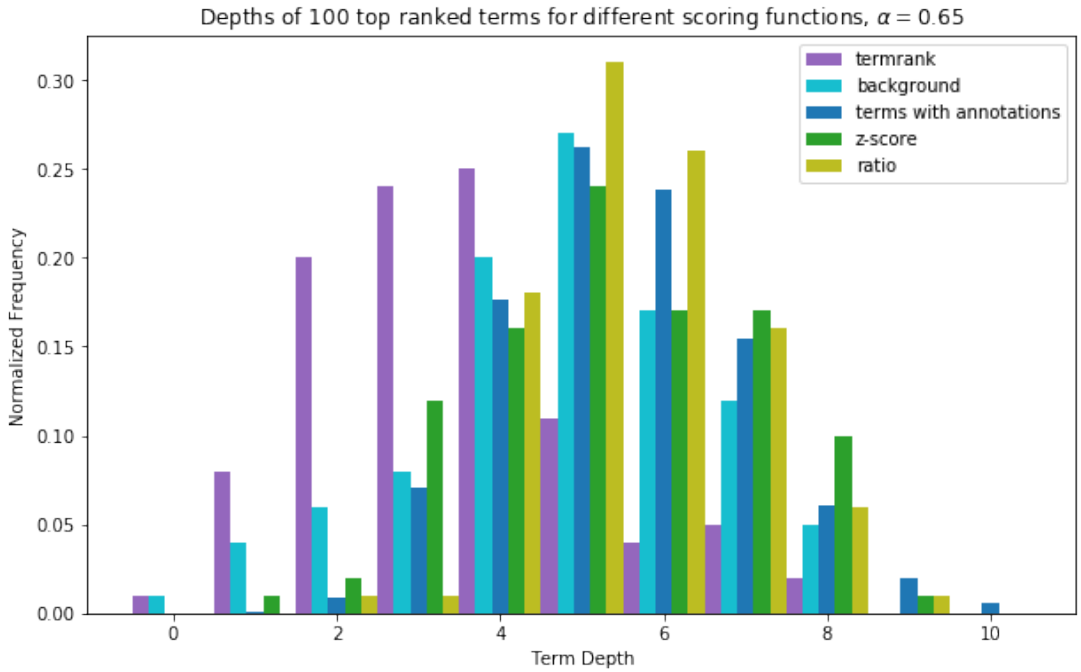

Figure 4.1: Normalised histograms of term depth among the top 100 biological process GO terms by enrichment score for various methods when run on the set of 186 genes found to be significantly differentially expressed in the human post synaptic density at a significance level of 0.05 using data reported by Bayes et al. (2012). For comparison the histogram of the depths of all terms with at least one annotated gene from the input gene list is also shown (blue). A value of $\alpha = 0.65$ was used for all methods.

4.1.2 Results

Dependence of ranking on depth for different scoring functions The raw termrank tends to favour low depth terms, since these have more incoming paths and fewer outgoing paths. One advantage of the scoring functions considered in this previous section is that they correct for this bias. This can be seen by looking at how the distributions of the 100 top ranked terms for each of the scoring methods compare to the overall distribution of depths for all terms with at least one annotation. This is done in figure 4.1, which shows a clear bias of the raw termrank (green bars) towards low depth terms as compared to the overall distribution of term depths (blue bars). Both the 'MCOA z' (green) and the 'MCOA background' (cyan) scores seem to reasonably effectively

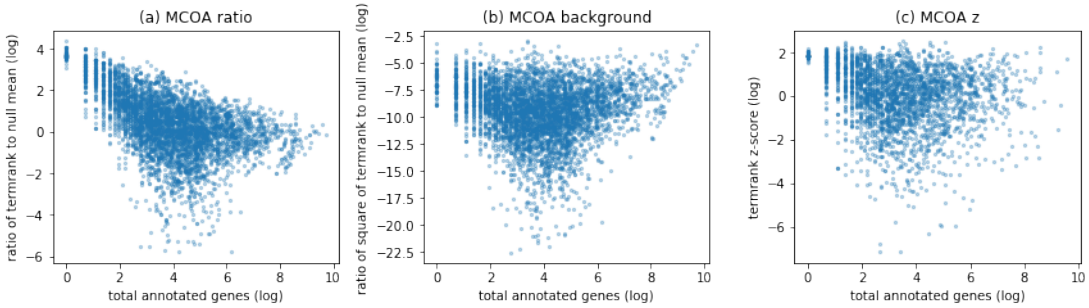

Figure 4.2: Distributions of enrichment scores and total annotation counts for all GO biological process terms with at least one annotation from the set of genes differentially expressed in the human PSD, when that set of genes was used as input to the various MCOA scoring functions. A value of $\alpha = 0.65$ was used for all methods.

correct for this bias, with depth distributions much closer to the underlying depth distribution, as would be expected for an unbiased enrichment scoring procedure. Although for lower values of α the skew towards low depths is less pronounced, it is still clear.

Dependence of ranking on annotation count for different scoring functions Another undesirable bias in the raw termranks is towards terms with high numbers of total annotations, as explained in section 3.2.1. To check whether the scoring functions are redressing this bias in a satisfactory way it is useful to look at how the distribution of term scores varies with the number of annotations to the term. Although there is no obviously correct form for this distribution, for an enrichment analysis procedure to be able to retrieve enriched terms of varying degrees of specificity, the highest scoring terms should be distributed across varying numbers of annotations. In particular, it is desirable to avoid a preference for terms with either very low or very high annotation counts, since these are likely to be the least informative terms in general. Figure 4.2 shows these distributions for the 'MCOA ratio', 'MCOA z' and 'MCOA background' scoring methods. 'MCOA ratio' is strongly biased to terms with small numbers of annotated genes, meaning it will struggle to retrieve as enriched terms with medium or large numbers of annotations. On the other hand, neither 'MCOA z' nor 'MCOA background' scores seem to be so strongly determined by the number of genes annotated to a term, with high-scoring terms coming from the full spread of the overall distribution of total annotation counts.

4.1.3 Discussion

The results of this section help provide some insight into the characteristics of the different scoring functions when used to measure ontology term enrichment on a real dataset. The scoring functions can be differentiated by the types of terms they tend to favour. These differences reflect different approaches among the scoring functions to attempting to correct for the bias in the raw termranks to terms of low depth and high annotation count, which naturally results from modelling a tree-structured ontology graph as a Markov Chain. 'MCOA ratio' is strongly biased towards high depth terms with low annotation counts, suggesting that it is a poor choice of scoring method. 'MCOA z' and 'MCOA background', are less clearly biased, with the top of their enrichment rankings featuring terms of a range of depths and annotation counts, as would be expected from an unbiased enrichment analysis procedure. However, the two methods differ in the way that they correct the raw termranks and as a result do show different tendencies. 'MCOA background', which divides the square of an ontology term's termrank by its 'background' termrank, inherits some of the bias inherent in the original termrank, with evidence of a slight preference for low depth terms with high annotation counts. Meanwhile, 'MCOA z' makes it more easy for ontology terms with low termranks to score highly if they have a small null standard deviation. This can be the case for terms with very small numbers of annotations, which appear to perform disproportionately well under this scoring procedure. If these scoring procedures are to be used for enrichment analysis, it is important to be aware of these tendencies. For example, if 'MCOA z' is being used it may often be useful to filter the terms based on the total number of annotations before reporting results, to prevent terms with small numbers of annotations from dominating the results, since these terms are often of less interest.

4.2 Benchmarking performance using Simulated Data

Evaluation of enrichment analysis methods is challenging because of the lack of any ground truth against which the performance of various approaches can be benchmarked. To facilitate comparison of the sensitivity of different methods it is therefore fairly common to generate gene lists based on the genes annotated to a set of 'target' terms, which should be returned as enriched (Alexa et al., 2006; Bauer et al., 2010; Frost and McCray, 2012; Ballouz et al., 2017) . These simulated datasets can be

created with varying amounts of noise and varying numbers of target terms to probe the ability of the enrichment analysis methods to retrieve the desired terms in various conditions. The advantage of this kind of simulation study is that it allows some quantitative comparison to be made of different enrichment analysis methods. Precision and recall are particularly useful metrics, even in this artificial setting, since better-performing methods will tend to prioritise the relevant results and not throw up too many irrelevant results, whereas weaker methods will struggle to retrieve results with high precision even in this relatively straightforward scenario.

4.2.1 Unweighted genes and edges

The simplest application of the MCOA method involves the case in which the inputs are an unweighted gene list, and a set of unweighted annotations to an ontology of interest. This is the type of situation in which classic over-representation analysis can be employed, and as a result offers the possibility of comparing MCOA to the classic method. In order to do this, we generate lists of genes according to the following procedure. First, 5 target terms are selected at random from amongst the Gene Ontology's 'biological process' terms. Following Frost and McCray (2012), these genes are filtered based on the number of direct annotations, with only terms with at least 5 direct annotations considered. Terms with more than 1000 total (direct and indirect) annotations are also discarded. The set of all genes annotated to the 5 selected target terms constitute a 'signal' gene list, from which enrichment methods should be expected to recover the target terms with little difficulty. However, such strong signal is unrealistic, so 40% of these genes are randomly replaced with other genes from a background set consisting of all genes with at least one annotation to a biological process term not in the original signal set. Following this procedure, we generate 100 gene lists each with 5 target enriched terms to be used as tests of the sensitivity of MCOA under different model choices and in comparison to the classic method using Fisher's exact test. For the purpose of comparison, the output of each model on a given gene list is treated as a list of ontology terms ranked by enrichment score. For the classic Fisher's exact test method, we use the adjusted p-value as the ranking criterion, with the top ranked term being the one with the lowest p-value. We then compute the top-k precision and recall of the ranking for different values of k between 1 and 100 to generate a set of precision and recall values. For example, at $k = 3$, the precision of a method is the fraction of true positives amongst the top 3 most enriched terms according to that method, and the

recall is the number of true positives amongst the top 3 divided by the total number (5) of true positives. Finally, the precision and recall at each k are averaged over all of the gene lists to generate a single set of precision-recall datapoints for each method.

When performing tests involving null models (section 3.2.2), this procedure is slightly modified. Computing a null termrank distribution for a given input gene list requires sampling several hundred random gene lists of the same size, and running the Markov chain computation on these lists. Because of the computational demands of doing this for 100 different input gene lists, we instead constrain gene lists to have size $100 \times n, n = 1, 2, .., 10$ by either randomly adding or removing genes to get the length of the gene list to the nearest hundred, and discarding gene lists with length > 1050. Restricting the possible lengths of the genes allows the null distributions for the target lengths to be precomputed, significantly reducing the computational demands of running the simulation experiment on 100 gene lists. The combination of this random addition or deletion to meet the length restriction, and the random replacement of genes from the original 'signal' gene list can be seen as a corruption process with a variable noise level.

4.2.1.1 Results

The simulation dataset can be used to investigate the questions relating to model choice raised in the previous section: the dependence of the null model based methods on the number of null samples generated, the role of the damping factor α, and the effect of choosing one or other of the enrichment scoring methods outlined in section 3.1.3 and 3.2.2.

Figure 4.3 shows a precision-recall plot for various choices of enrichment scoring function. The raw termrank performs very poorly, with a maximum recall of just over 0.5 indicating that on average only around half of the truly enriched terms were amongst the top 100 terms by termrank (the last datapoint for each method corresponds to the precision and recall within the top 100 terms returned by that method). 'MCOA ratio' also fares poorly: although its recall within the top 100 is high, its precision is always low indicating that it fails to effectively distinguish between true and false positives. The two scoring methods which are designed to correct for the bias in 'MCOA ratio', 'MCOA background' and 'MCOA z' both perform well. The first few datapoints, which correspond to precision and recall within the top 1,3 and 5 results respectively, show that these MCOA methods are as good as or slightly better than the classic method at identifying relevant results within the top few terms returned, while

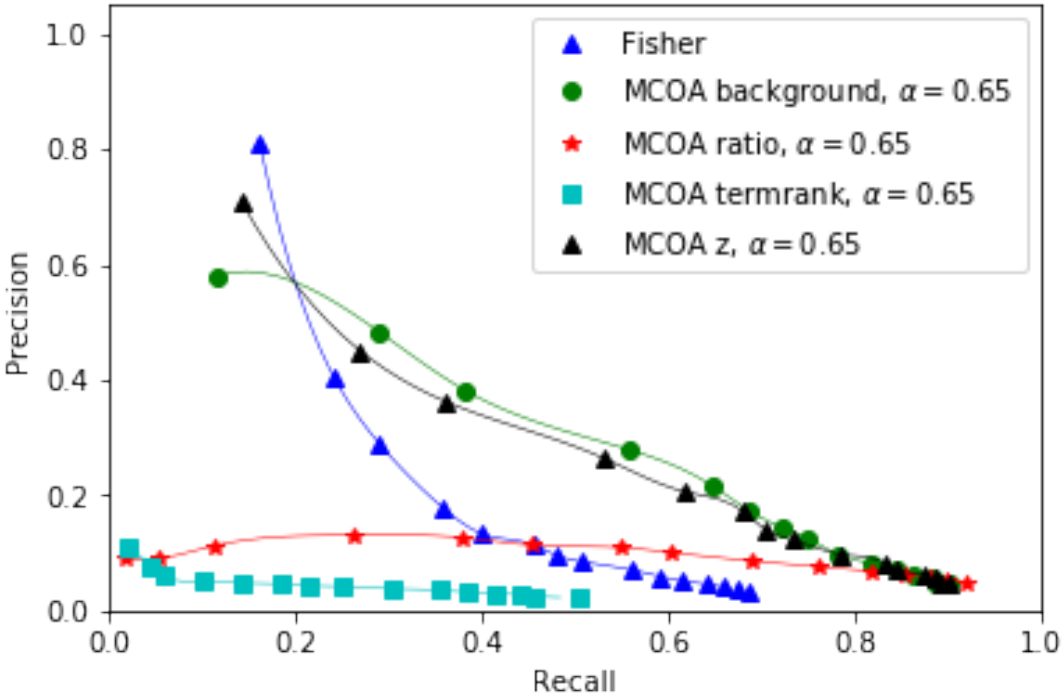

Figure 4.3: Average precision-recall curves for different scoring functions over 100 simulated gene lists, each with 5 'true positive' enriched terms. Each marker represents the average precision and recall within the top k ranking of most enriched terms according to the enrichment methods. The values of k used were 1,3,5,10,15,20,25,30,40,50,60,70,80,90 and 100. The scoring functions used were: the p-value from Fisher's exact test ('Fisher'), 'MCOA background', 'MCOA ratio' and 'MCOA z'. For the latter two scoring functions, 1000 null samples were used to create the null distribution.

the slower drop off in precision of these methods than the classic method indicates that the most of less easily retrieved terms are still reliably identified as amongst the top results, whereas the classic method has more difficulty retrieving these terms. Finally, all three of the MCOA methods that attempt to correct for the bias in termrank show much better maximum recall than the classic method. The classic method on average returns only around 70% of the truly enriched terms within the top 100 results, whereas the MCOA methods return close to 100% within the top 100.

Figure 4.4 shows a precision-recall plot for the classic method as well as three different versions of the z-score based method, distinguished by the number of null samples used to compute the mean and standard deviation of the termranks in the null model (equation (3.8)). Using only 100 samples leads to significantly worse performance than using 300 or 1000 samples, confirming the importance of generating

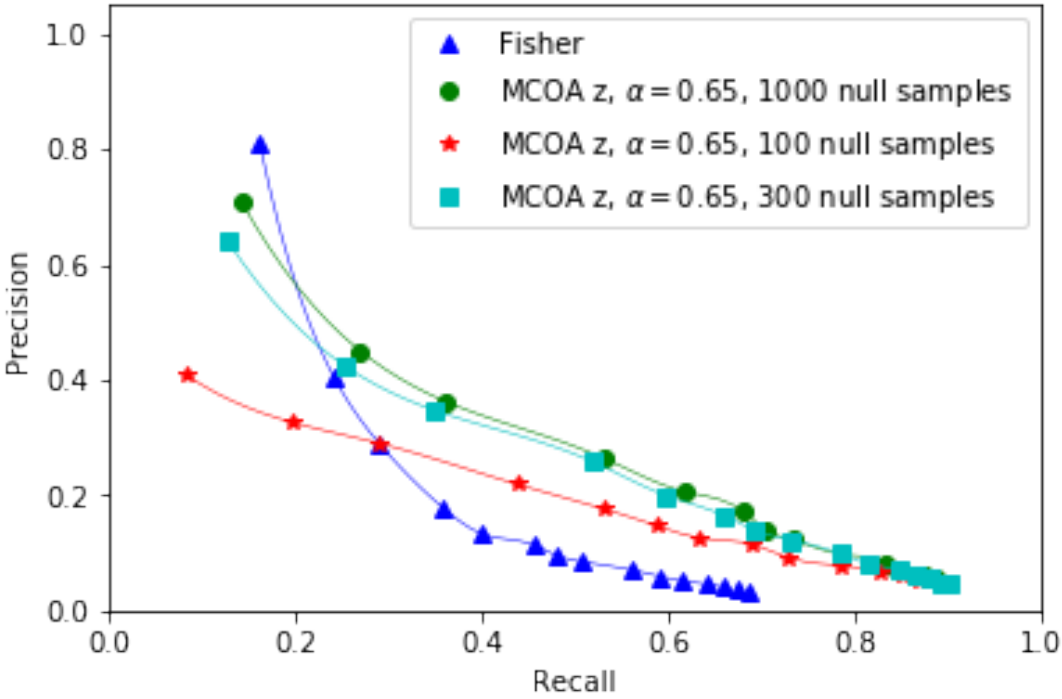

Figure 4.4: Precision-recall plot for the simulated data experiment showing the effect of using different numbers of null samples to form the null distribution used by the 'MCOA z' scoring procedure.

enough null samples to allow the null statistics to converge when using this scoring procedure.

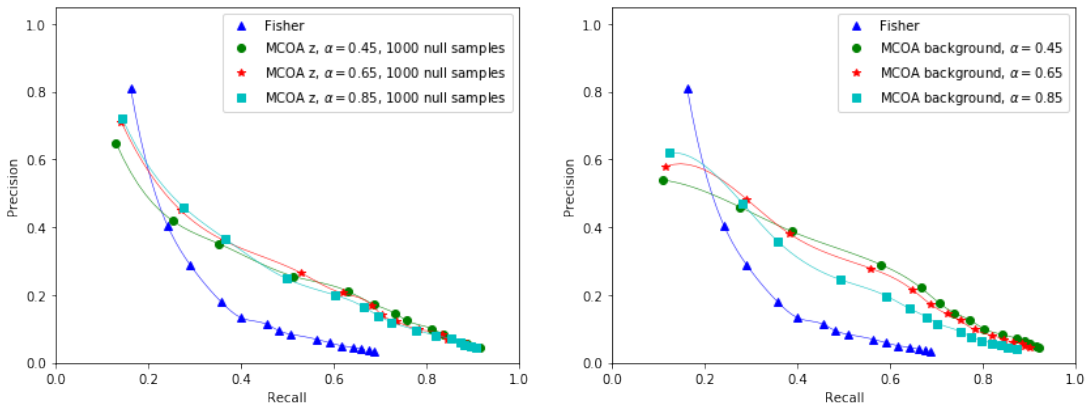

Figure 4.5: Precision-recall plots for the simulated data experiment showing the effect of various values of α on 'MCOA z' (left panel) and 'MCOA background' (right panel)

Figure 4.5 shows precision-recall plots for 'MCOA-z' (left panel) and 'MCOA background' (right panel), run on the same input lists using different values of α. The

effect of α on performance does not seem to be large, although it is stronger in the case of 'MCOA background'.

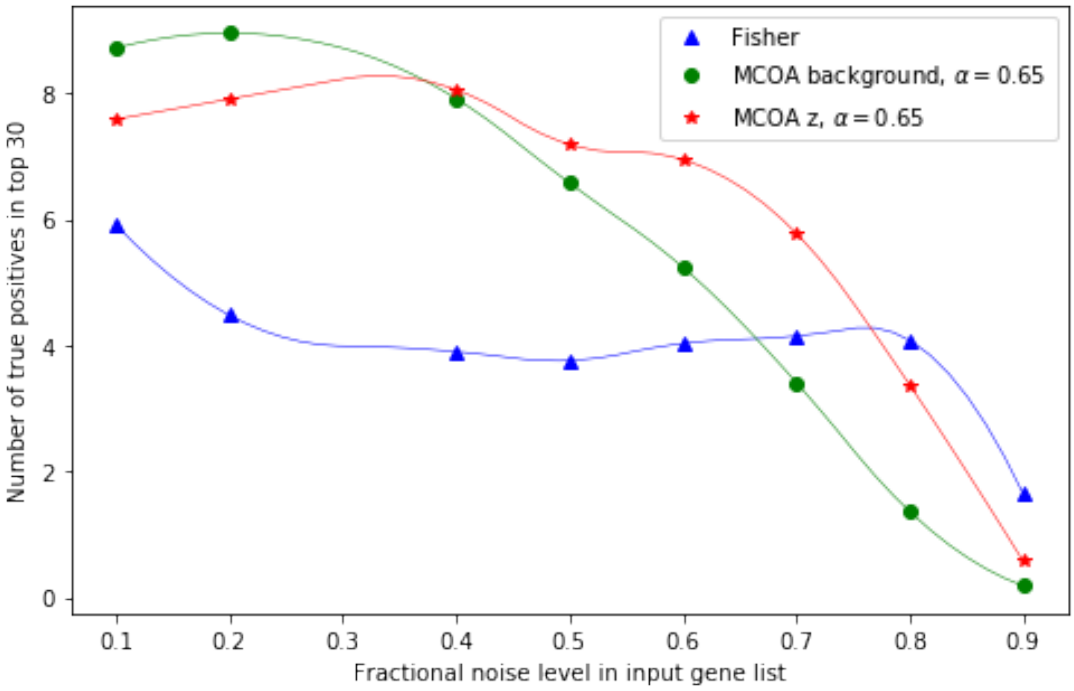

Figure 4.6: Average number of target terms ranked in top 30 most enriched terms by classic method, 'MCOA background' and 'MCOA z', as function of the fraction of genes from the signal list replaced at random with 'noise' genes. The average is taken across 100 simulated datasets with 20 target terms each, generated according to the procedure outlined in (section 4.2.1).

Effect of noise level The precision-recall plots displayed above were obtained by generating noisy datasets, where the amount of noise was controlled by two factors: the fraction of the 'signal' list chosen to be randomly replaced, and the number of genes added or removed to round the list to one of the permitted lengths. The ability of the enrichment methods to successfully retrieve enriched terms in the presence of noise is important because biological datasets are inherently noisy. It is therefore desirable that adding a reasonable amount of noise to the input should not drastically change the result. Although there is no strong requirement on the amount of noise an enrichment method is able to handle, it is interesting to examine how the results change as the noise level is varied. To do this, the fraction of the 'signal' genes to be randomly replaced was varied between 0 and 1. For each of these noise levels, 100 were generated using the procedure of section 4.2.1, with 20 target terms per gene list. To compare results at

different noise levels, for each method we calculate the average number of target terms retrieved within the top 30 most enriched terms across the 100 gene lists at each noise level. Figure 4.6 is a plot of the results, which reveals some interesting differences between the methods. The performance of the classic method is almost independent of the level of noise up to a noise level of between 0.8 and 0.9, when it rapidly degrades. At low noise levels, 'MCOA background' outperforms 'MCOA z', but at higher noise levels the reverse is true. It makes sense for 'MCOA z' to be overall more robust to noise, since the whole purpose of the calculation of the null distributions in this scoring procedure is to compare the termranks on the input gene lists to those on random gene lists, and as the noise level increases the gene lists are closer to random.

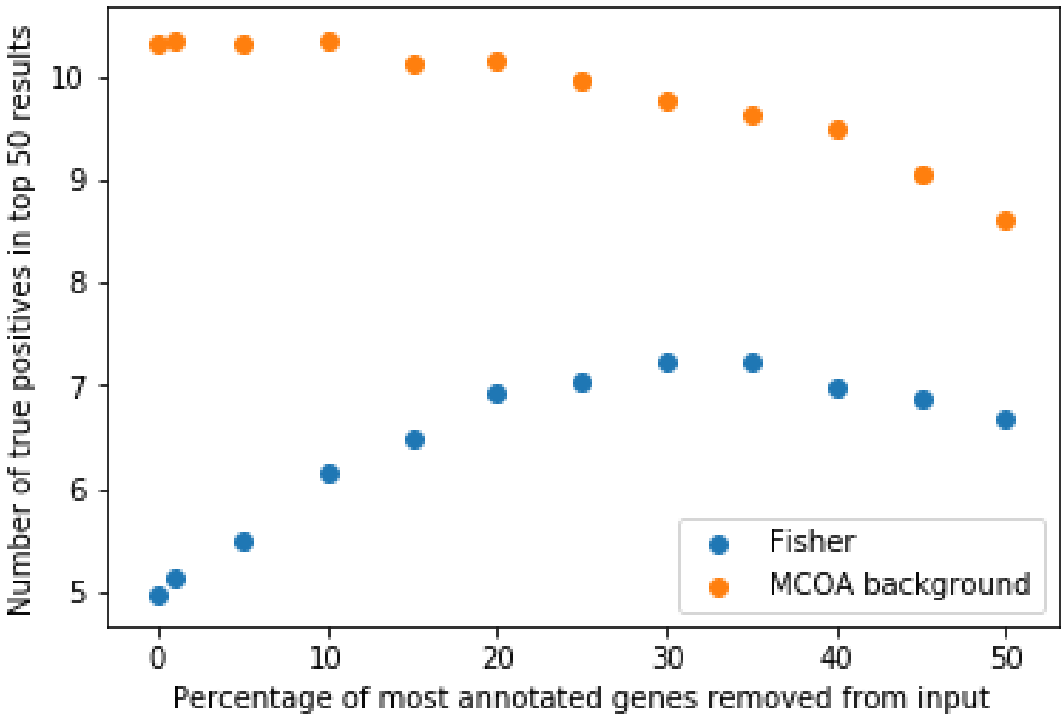

Figure 4.7: Average number of target terms ranked in top 50 most enriched terms by classic method and MCOA background, as function of percentage of top genes by annotation count removed from input list. The average is taken across 100 simulated datasets with 20 target terms each, generated according to the procedure outlined in (section 4.2.1).

Effect of multiply-annotated genes An important shortcoming of the classic method is its failure to account for differences in the number of annotations to individual genes (Ballouz et al., 2017). Because each term is separately tested for enrichment, the contribution of genes which are annotated to many terms is effectively counted many times

over. As a result adding or removing only a small number of such multiply-annotated genes from input gene lists can significantly alter enrichment results. MCOA promises to avoid this problem by computing enrichment scores in a global rather than term-by-term manner. Because the total weight of the outedges from any gene node is 1 (section 3.1), edges from multiply annotated genes are effectively downweighted in proportion to the number of annotations to that gene. Intuitively this seems a more satisfactory approach: if a gene has only a few annotations, then those annotations are probably individually more informative than individual annotations to a gene which has hundreds of annotations.

Ballouz et al. (2017) showed that the classic method's treatment of multiply annotated genes lead to degraded performance on simulation studies such as the one explored here, to the extent that removing a fraction of the most heavily annotated genes from the input list before running the analysis actually *improved* the ability of the classic method to prioritise relevant results towards the top of the rankings. Following this example, I compared the behaviour of the MCOA methods and the classic method in response to the removal of a varying number of heavily annotated genes. To do this, each enrichment method was run multiple times on each of 100 test datasets generated as in section 4.2.1 but using 20 target terms, with a different percentage of the most heavily annotated genes removed before each run. This enabled the calculation of average number of target terms returned by each method within the top 50 as a function of the fraction of genes removed from the input. The results are shown in figure 4.7. The difference in the way multiply-annotated genes are handled by the two algorithms is clear. Removing genes from the input to the MCOA has little effect initially, then starts to decrease performance as the fraction of genes removed becomes more significant. In contrast, removing multiply annotated genes from the input to the classic method significantly improves its performance, to the extent that when the top 30% of genes by annotation count are removed, the number of target terms retrieved by the classic method is over 1.4 times that retrieved when no genes are removed.

4.2.2 Using KEGG Pathways to improve selection of enriched terms for simulation studies

Although the sort of simulation study conducted in the previous section is valuable as a means of comparing different enrichment analysis approaches, it is unrealistic in a number of ways. One of these is that the target enriched terms are selected at ran-

dom from across the ontology. The lists of genes generated by collecting the genes annotated to these target terms will therefore be collections of unrelated gene sets that could not plausibly have any coherent biological meaning. By contrast, in real applications enrichment analysis is usually applied to lists of genes collectively associated with some biological function or condition of interest. As a result many of the ontology terms most associated with these lists of genes might be expected to be related or overlapping.

A slightly more realistic approach might involve a way of selecting target terms that tended to lead more often to the selection of related terms amongst the target set. We choose to do this by cross-referencing with the pathway database KEGG (Kanehisa and Goto, 2000). Pathway databases such as KEGG store lists of genes annotated with particular biological pathways. We then select target terms based on the terms associated with the pathway genes, leading to sets of target terms that have a greater degree of relatedness than they would were they to be selected completely at random.

The details of the procedure are as follows. KEGG contains 320 human pathways, and each pathway contains a list of genes involved in that pathway. For each set of pathway genes, all gene ontology terms with at least 2 direct annotations from the pathway genes are selected as target terms. Target terms with fewer than 5 or more than 500 annotations overall are discarded (small terms are not of interest, and large terms will lead to bloated gene lists). Signal gene lists are then generated from each set of target terms as before, and as before 40% of the set of signal genes are replaced at random. Gene lists with more than 4500 or fewer than 50 genes are discarded, and then the lengths of the remaining gene lists are adjusted by random addition or removal of genes. Permitted gene list lengths in this case are 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 4000, 5000. As before the purpose of this constraint is to allow null distributions to be precomputed for the permitted lengths to restrict the number of null distributions that must be computed. The result is a set of 264 input gene lists of varying lengths and corresponding sets of target terms of varying sizes. Figure 4.8 shows histograms of the number of enriched terms and the lengths of the corresponding gene lists. As can be seen, this collection represents a wide range of gene list sizes and numbers of enriched terms.

To verify that the procedure for selecting enriched terms followed here is having the desired effect of increasing the relatedness of the enriched terms it is instructive to look at the average overlaps of the target terms selected according to this procedure and according to the random procedure used previously. The overlap in the sets of genes

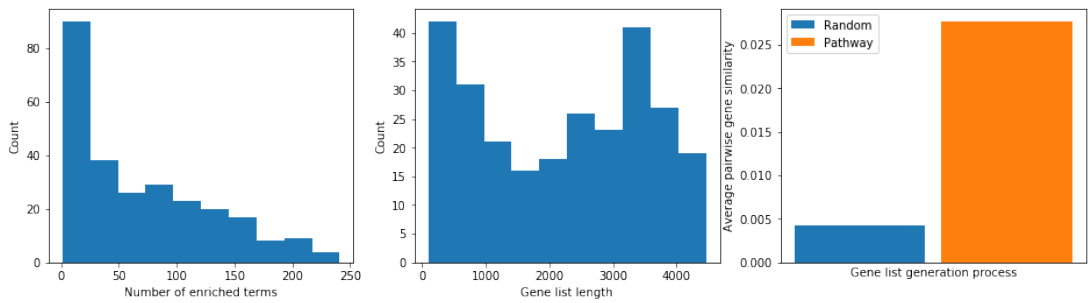

Figure 4.8: Histograms of number of target terms (left panel), and length of simulated gene list (centre panel) for gene lists generated from ontology terms associated with pathway genes, for all 264 human pathways used in the study. Right panel: bar chart showing average pairwise overlap between annotations of pairs of genes in lists generated according to the random method (section 4.2.1) and the pathway method.

T_1 and T_2 annotated to a pair of terms can be quantified using the Jaccard similarity coefficient:

$$J(T_1, T_2) = \frac{T_1 \cap T_2}{T_1 \cup T_2} \tag{4.1}$$

For each set E of target enriched terms we compute the average pairwise overlap, and call this the relatedness $rel(E)$ of the target terms in E. The average relatedness of the pathway target term sets is much higher than that of the randomly selected target term sets (figure 4.8, right panel).

4.2.2.1 Results

To evaluate performance across the whole range of generated datasets, average precision and recall are computed in the same way as before, with each data point corresponding to the top-k precision and recall of a ranking of terms by their level of enrichment according to a given method. Values of k again range between 1 and 100. The resulting average precision recall curves for the three top performing methods from the previous simulation experiment are shown in figure 4.9. Again the MCOA methods comfortably outperform the classic method in terms of average precision and recall statistics across these datasets.

4.2.3 Simulating datasets with weighted annotations

One of the particularly appealing features of MCOA is its ability to naturally incorporate annotation confidence scores, by adding weights to the gene-term edges in the

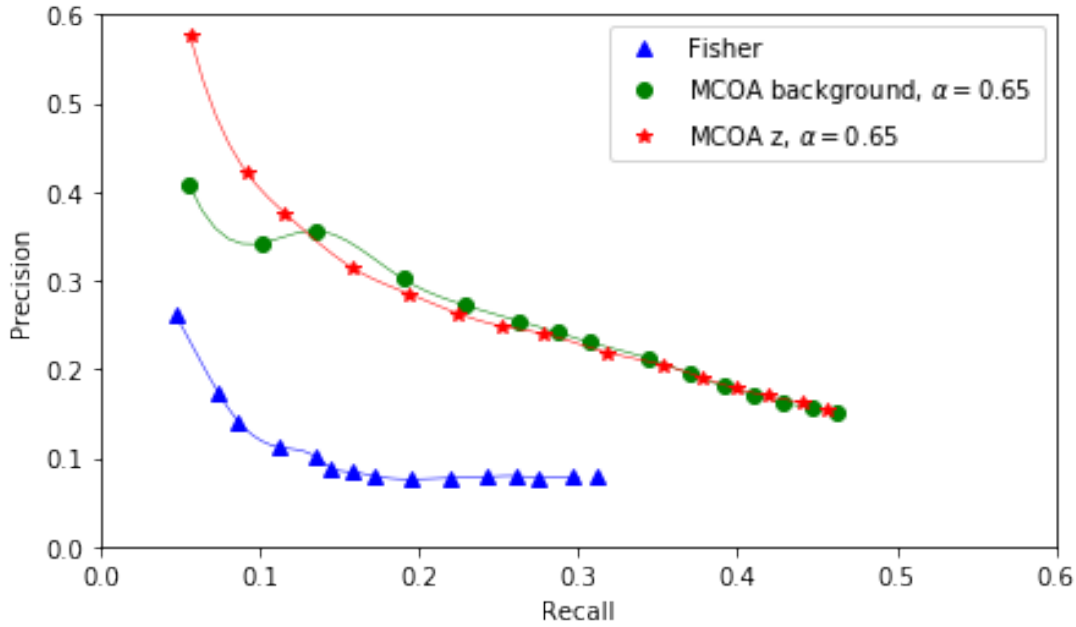

Figure 4.9: Average precision and recall of different enrichment methods on the datasets generated according to the pathway-based selection procedure of section 4.2.2.

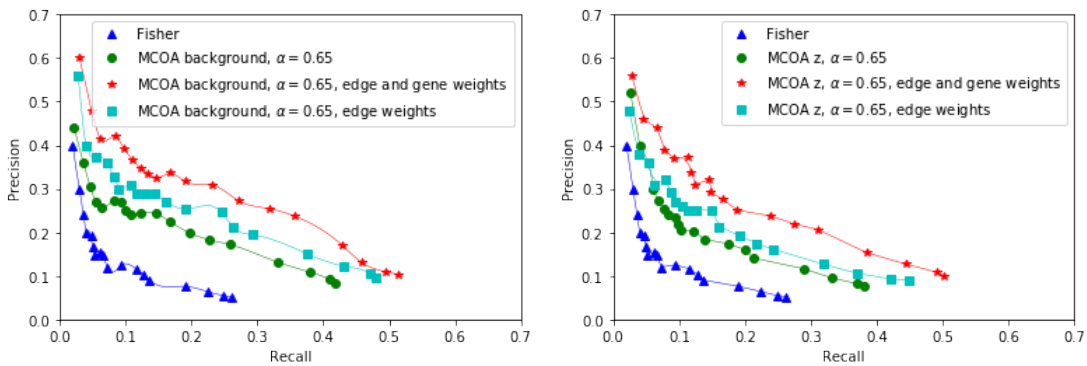

Figure 4.10: Average precision and recall of different enrichment methods on the datasets with noisy annotations generated according to the procedure outlined in section 4.2.3. MCOA methods were either run with no annotation weights, with annotation weights incorporated as weighted gene-term edges ('edge weights'), or with annotation weights incorporated both as weighted gene-term edges and by weighting genes by the mean weight of their annotations, as described in section 3.1.4 ('edge and gene weights').

graph (section 3.1.4). Simulated datasets can be used to test the usefulness of this ability in situations where not all annotations are equally useful, as is undoubtedly the case in real datasets. To do this we will use a simple noisy annotation model, which effectively acts as an extension to the testing procedure used in the previous section. The generation of test gene lists proceeds exactly as before: a set of target terms is selected, the genes annotated to these target terms are found and treated as a 'signal' gene list, which is then corrupted with noise resulting in the test input list. In the unweighted tests, the annotations to these input lists were then used to run each of the enrichment analysis methods and attempt to retrieve the target terms. Here we instead generate a noisy set of annotations to the genes in the input list, consisting of two classes of annotation: 'confident' and 'possible', with each class half the size of the original set. The set of 'confident' annotations is generated by first randomly selecting half of the true annotations, then randomly swapping the terms for a fraction η_c of the true annotations to the input genes with the terms from randomly selected annotations to genes not in the input list. The set of 'possible' annotations is generated in the same way but with a fraction $\eta_p > \eta_c$ created by random swaps. The result is a new set of annotations to the input list, of the same size as the original set, of which a fraction $\frac{\eta_c+\eta_p}{2}$ are guaranteed be unhelpful in finding relevant terms. The division of the annotation set into helpful (true) and unhelpful (false) annotations is intended as a simplified version of the way in which real annotations to biological ontologies will tend to have a wide variation in reliability, depending for example on the type and strength of evidence available for each annotation. In general, enrichment methods should be more useful if able to discount relatively unreliable annotations. To test this in this situation, we associate different confidence scores with the two classes of annotation, representing the relative unreliability of the 'possible' annotations with respect to the 'confident' ones. Each confident annotation is assigned a weight of 2, and each possible annotation a weight of 1. These scores can then be used as additional inputs to the MCOA method, representing the confidence that a given annotation is true.

Given this experimental setup, we compare the average precision and recall of MCOA when including annotation confidence scores versus not over a number of sets of test genes, annotations and target terms, to verify that the additional information is being used in a sensible way by the algorithm. Both ways of incorporating edge weights introduced in section 3.1 are tested: simply adding the scores as edge weights to the gene term edges (signified in plot legends as 'with edge weights'), and weighting genes by mean annotation score as well as using the edge weights ('with gene and

edge weights'). Figure 4.10 shows the results of such a test, with ten sets of test data generated using $\eta_c = 0$ and $\eta_p = 1$. For this extreme case, the best strategy would be simply to ignore the 'possible' annotations completely, as they are guaranteed to be unhelpful. As a result it is to be expected that the methods which discount the 'possible' edges should fare significantly better than those which do not, and this is indeed the case, with 'MCOA background' using edge weights and using edge and gene weights solidly outperforming both MCOA background without edge or gene weights and the classic method.

4.2.4 Discussion

Both 'MCOA background' and 'MCOA z' perform comfortably better than the classic approach to enrichment analysis on all the types of simulated data explored in this section. Despite the artificiality of the situation this suggests that MCOA methods are better able to prioritise relevant ontology terms than the classic approach. More broadly, it offers vindication for the graph-based, global approach to the enrichment analysis problem represented by MCOA, and characteristic of Modular Enrichment Analysis approaches, by comparison to the term-by-term Singular Enrichment Analysis type approach exemplified by the classic method. An important difference between the MCOA methods and the classic method in this regard is the fact that whereas the annotations to multiply annotated genes are naturally discounted in MCOA, they tend to disproportionately influence the output of the classic method. The significance of this difference is highlighted by the finding that whereas removing significant percentages of the most multiply annotated genes from input gene lists in the sort of simulation studies explored here has little effect on the performance of the MCOA methods, it significantly *improves* the performance of the classic method. This observation suggests that this way of handling multiply annotated genes is an important strength of the MCOA approach, in two respects. First, and most obviously, not according disproportionate weight to the annotations to these genes leads to more balanced output of enrichment scores, contributing to MCOA's significantly better precision-recall performance across the simulation studies. Second, it makes the MCOA methods more robust to small changes in the input than the classic method – whereas removing just a small percentage of multiply annotated genes significantly changes the output of the classic method (as measured by its recall within the top 50 results), it barely affects the MCOA method. It is also noteworthy that the downweighting of multiply-annotated

genes that is built into MCOA has been proposed as an ad-hoc correction to Gene Set Enrichment Analysis methods, such as in the Pathway Analysis by Down-weighting Overlapping Genes approach of Tarca et al. (2012). Moreover, MCOA's ability to handle annotation confidences allows it to find more relevant enrichment results in situations in which relative differences in annotation reliability can be quantified, as shown by the results of the study using simulated gene lists with noisy annotation sets from section 4.2.3.

Although suggestive, the results of such simulation analysis should not be exclusively relied on to determine the best enrichment method, since the generated datasets are unrealistic in some important ways. The use of a random sample of unrelated terms from across the ontology does not reflect the fact that in real enrichment analysis scenarios, there is likely to be a high degree of correlation between relevant biological functions. In the same way, using unrelated target terms to generate signal gene lists leads to gene lists that are unlikely to resemble gene lists that could be generated by any biological experiment, owing to the genes forming largely non-overlapping clusters associated with separate, unrelated functions. This undesirably artificial feature of the simulated data sets is partially mitigated by using pathways to control the selection of target terms, as explored in 4.2.2. However this sort of term selection procedure still of course comes nowhere close to imitating the biological processes that underlie real gene lists. The simulated sets of noisy annotations used in 4.2.3 represent a considerable oversimplification of the variability in reliability that may actually be expected to be found in real annotations. Finally, though the notion of true and false positive enriched term relied on in the analysis of these experiments is a useful way of measuring the relevance of enrichment output, in truth relevance is not binary and will typically depend as much on the ultimate ends served by the enrichment analysis as the data itself. This makes flexibility and interpretability desirable features of enrichment methods, since they allow researchers to handle results according to their needs. These are features that should not be ignored but cannot be captured by the sort of simulated data experiment performed here.

4.3 Benchmarking performance on real human disease datasets using the Human Disease Ontology

Although it is in general difficult to use real datasets to benchmark enrichment analysis performance, owing to the lack of known relevant terms to treat as true positives, there are some datasets for which at least some terms that are definitely relevant can be identified. For example, many expression profiling studies have been performed to compare the expression levels of genes across the human genome between groups of patients with a particular disease, and groups of healthy patients. Performing enrichment analysis of Human Disease Ontology (HDO) terms on such a dataset, we would expect to find terms corresponding to the disease in question being ranked highly by good enrichment analysis methods. Tarca et al. (2012) collected and labelled a set of 24 human disease datasets for the purpose of comparing different methods for enrichment analysis of terms from the KEGG pathway database. In this section, I propose to use the same set of disease datasets to compare the performance of the MCOA methods with the classic method when performing enrichment analysis of HDO terms. The reason for using HDO terms rather than KEGG pathway terms, as was originally done by Tarca et al. (2012) is that the KEGG pathway database is not structured hierarchically like an ontology, which makes the graph-based approach MCOA takes to computing enrichment scores less appropriate. Using HDO terms also offers the possibility of incorporating annotation weights retrieved from the Human Disease Gene Database (HDGDB) database developed by He (2016).

4.3.1 Method

Dataset 24 human disease differential expression datasets obtained from the Gene Expression Omnibus were collected by Tarca et al. (2012). The datasets together with KEGG pathway labels were retrieved through the R package KEGGdzPathwaysGEO. For each dataset, the HDO term relevant to the disease was found by using the cross reference metadata on individual HDO terms to find the HDO term corresponding to the KEGG pathway with which the dataset was labelled. For example, one of the datasets, GSE9348, contains expression data from 12 normal samples and 70 colorectal cancer samples reported by Hong et al. (2010), with normalised expression level measurements for 20514 genes. In KEGGdzPathwaysGEO, this dataset is labelled with the KEGG Colorectal Cancer pathway (hsa05210). The corresponding HDO term is

Colorectal Cancer (DOID:9256). Annotations of genes to HDO were retrieved from HDGDB. Each annotation in HDGDB has an associated confidence score: these scores were used as annotation weights in the MCOA methods.

Data Analysis Classic over-representation analysis and 'MCOA z' and 'MCOA background' methods were run on each dataset. To simplify the analysis, only the 16 datasets with unpaired experimental design were used. 9 different diseases are represented across these 16 datasets. To select input gene lists for the classic method, the following procedure was followed. First a two-sided t-test was performed on the normalised control and disease differential expression levels for each gene to test for significant differential expression between the control and disease samples. The resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure, with any genes with an adjusted p-value of < 0.01 being selected as members of the input list. If the resulting input list consisted of less than 200 genes, an adjusted p-value of 0.05 was instead used as the threshold for the input genes. If this still yielded fewer than 200 genes, the top 1% of most differentially expressed genes (ranked by p-value) were selected. These selection criteria were chosen to match those used by Tarca et al. (2012).

'MCOA z' and 'MCOA background', by contrast, used the full list of genes as input, with absolute values of the t-score from the t-test used as gene weights. The 'MCOA z' method used the null sampling procedure based on randomising gene weights rather than randomising members of the input gene list, as discussed in 3.2.2.1. For each dataset, 150 null samples were used to calculate null means and null standard deviations for the termranks. Both 'MCOA z' and 'MCOA background' were run separately with and without annotation weights on each dataset.

The HDO terms corresponding to the 9 target diseases are all quite highly annotated, with the smallest number of total (indirect and direct) annotations to a target disease being 222 (Huntington's disease), and the highest being 2430 (colorectal cancer). The target disease terms are all within the top 11% by annotation count of HDO terms with at least one annotation. In order to encourage the enrichment methods to focus on these sort of terms, therefore, the outputs of each method are filtered so that enrichment scores are only reported for terms with at least 100 total annotations, of which there are 548 in the HDO.

4.3.2 Results

4.3.2.1 Initial analysis of scoring functions on a single Colorectal Cancer dataset

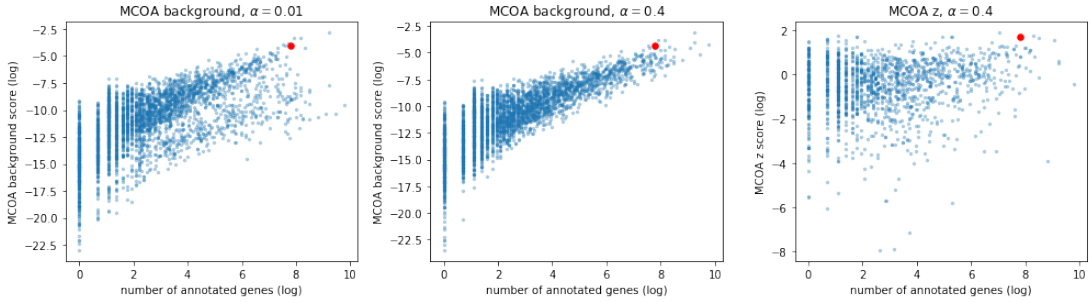

Figure 4.11: Distributions of enrichment scores and total number of annotated terms for HDO terms with at least one annotated gene, when the MCOA methods were run on the colorectal cancer dataset GSE9348. The red dot marks the target colorectal cancer HDO term.

To get an idea of the behaviour of the MCOA methods when using full lists of weighted genes and weighted HDO annotations, both 'MCOA background' and 'MCOA z' were run on GSE9348, the colorectal cancer dataset described above. Figure 4.11 shows the distribution of enrichment scores and total annotation counts for the two methods. The red dots represent the target HDO term, 'Colorectal Cancer', which both methods return in the top ten results. It is clear from the distributions that the MCOA background scores (left panel with $\alpha = 0.01$, centre panel with $\alpha = 0.4$) are very strongly biased towards terms with high annotation counts, with all terms with high annotation counts clearly scoring higher than most other terms. Although choosing a much lower value of α (left panel) seems to reduce this bias somewhat, it is still clear. 'MCOA z' on the other hand seems to be relatively unbiased, with high scoring terms coming from the full range of annotation counts, and highly annotated terms receiving a wide range of enrichment scores.

4.3.2.2 Comparison of MCOA and Classic Methods on 16 Disease Datasets

'MCOA background', 'MCOA z' and classic methods were run on the 16 disease datasets, and on each the rank of the target disease term was recorded. For 'MCOA background' a value of $\alpha = 0.01$ was used to try to reduce the bias towards highly annotated terms identified above. 'MCOA z' was used with $\alpha = 0.4$. Figure 4.12 shows boxplots of the ranks of the target disease terms on each dataset for the various methods. All methods tend succesfully retrieve the target terms towards the top of

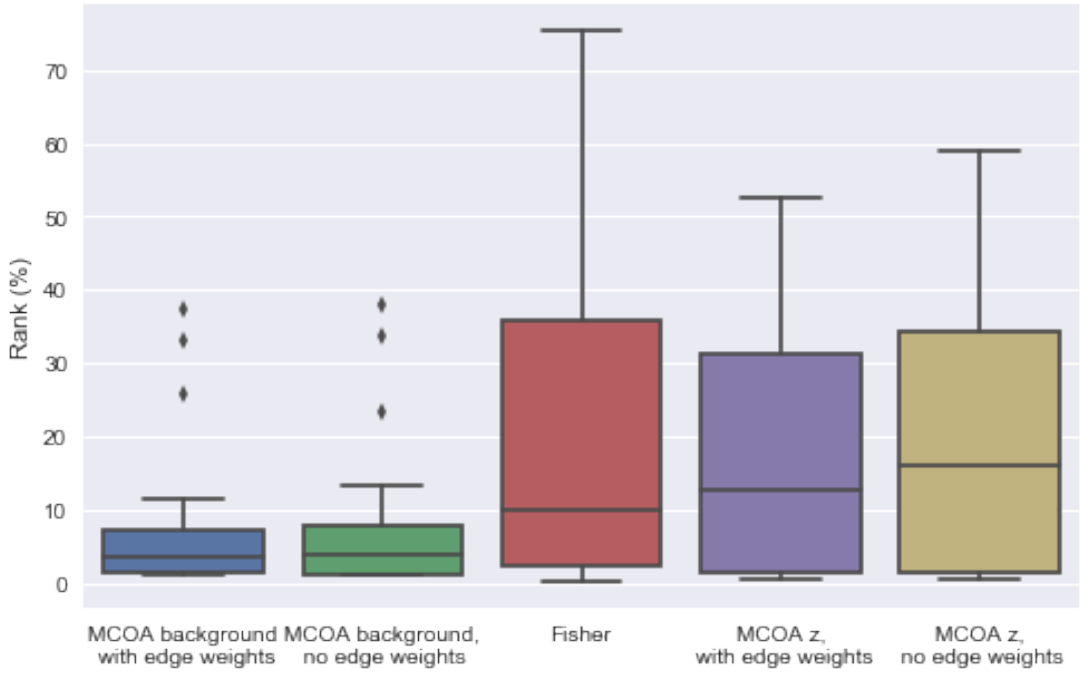

Figure 4.12: Boxplots of the percentage rank of the target disease terms across the 16 datasets for various methods. Lower values mean that the target term was ranked closer to the top of the ranking.

their enrichment rankings. Both MCOA methods are slightly better at prioritizing the relevant term when incorporating annotation weights than not. The target terms tend to be closer to the top of the rankings for 'MCOA background' than for the other two methods.

Just looking at the average rank of the target terms is not sufficient however, since one way to perform well by this metric would be to simply rank all the 9 target disease terms highly regardless of the input. In order to distinguish methods which perform well by doing this from methods able to actually differentiate between enriched and non-enriched datasets, we can look at the difference in ranks for individual disease terms between truly enriched and non-enriched datasets. For example, 2 of the 16 datasets are colorectal cancer datasets. We would expect the rank of the disease term colorectal cancer to be lower on these 2 datasets than on the other 14 datasets, if the enrichment method is actually differentiating successfully between cases of colorectal cancer enrichment and non-enrichment rather than simply scoring colorectal cancer as highly enriched across the board. The left panel of figure 4.13 compares the rank of colorectal cancer on the two colorectal cancer datasets versus on the other 14 non-colorectal cancer datasets for 'MCOA z', 'MCOA background' and the classic method.

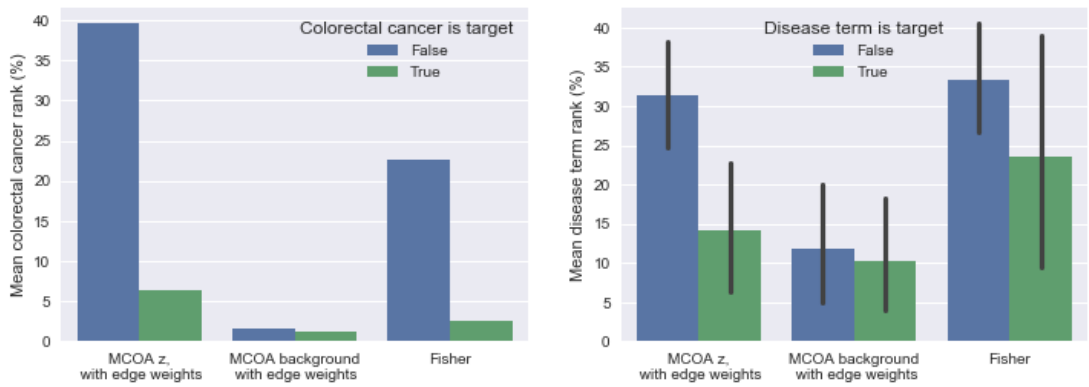

Figure 4.13: Barplots showing the difference in average percentage rankings (lower is better) of disease terms on datasets for which the disease term is the target, and datasets on which it is not the target. The left panel shows these rankings for a single disease term, colorectal cancer, which is the target on 2 datasets, and is not on the remaining 14. The right panel displays the same information, averaged across all 9 diseases.

While all of the methods successfully identify colorectal cancer as among the most enriched terms on the 2 relevant datasets, 'MCOA background' also tends to rate it as among the most enriched terms on all the other datasets, whereas classic and 'MCOA z' do not. This is not specific to colorectal cancer: the right panel of figure 4.13 compares the average rank across all 9 diseases of the disease term in on datasets where the disease is the target and where it is not. As expected, all the methods tend on average to rank the disease terms as more enriched on datasets for which they are actually enriched than on datasets for which they are not. However, the difference is much smaller in the case of 'MCOA background' than in the case of the other two methods, with 'MCOA z' showing by far the strongest ability to discriminate between cases of true enrichment and non enrichment (as measured by the gap between the average rankings of disease terms in relevant vs non relevant datasets). This suggests that the fact that 'MCOA background' prioritizes the target diseases towards the top of the rankings the most consistently is an artefact of the bias of 'MCOA background', which leads it to rank all the disease terms highly whether they are actually enriched or not. So of the three methods, 'MCOA background' is the least effective at actually distinguishing between true and spurious cases of enrichment on these datasets, with 'MCOA z' being the best.

4.3.3 Discussion

The labelled human disease datasets provide a useful way of comparing enrichment analysis algorithms on real data. It is notable that the MCOA methods are better able to prioritize relevant disease terms when incorporating weighted annotations than not, even if only slightly. It is also encouraging that 'MCOA z' performs competitively with the classic method, indicating that the approach of including all the full list of genes and associated weights as input is useful.

However, these results should still be treated with some caution. One problematic feature of using this as a benchmarking task is the fact that the relevant terms are relatively similar, in that they come from similar levels in the ontology (as general rather than specific disease terms), and are all amongst the top 11% of HDO terms with at least one annotation by total number of genes annotated. As a result, enrichment analysis methods with a bias towards such terms will tend to perform better on this benchmark as measured by the average rank of the target terms, which makes it important to also look at the difference in rank of disease terms across datasets with which they are associated compared to across unrelated datasets. Doing this revealed that 'MCOA background' was barely discriminating between cases of real and spurious enrichment, suggesting that 'MCOA z' is in fact the better of the MCOA methods at assessing enrichment on these datasets. Further, only 16 datasets associated with 9 different diseases were used, and the degree of variability in the rankings is quite high. It would be desirable to include further datasets to increase confidence in any conclusions drawn.

Chapter 5

Conclusion

MCOA is an approach to enrichment analysis of ontology terms that promises to overcome several of the limitations of traditional approaches. By explicitly modelling the graph structure of the ontology, it incorporates topological information that is ignored by over-representation analysis, but has been found to be useful by alternative methods (Alexa et al., 2006). By incorporating gene weights, it circumvents the need for an arbitrary cutoff of the input gene list, which was the primary motivation for the class of methods in the GSEA family. Multiply-annotated genes are naturally discounted in the MCOA framework, avoiding the disproportionate influence of these genes on results that can affect methods from both the SEA and GSEA families. Finally, by offering the possibility of including annotation weights as weighted gene-term edges, MCOA is able to include information about the reliability of annotations that is ignored by most traditional methods, but could be important especially given the widespread use of automated methods to generate annotations.

The first set of questions addressed in this project related to methodological issues raised by the use of MCOA as a standalone method for enrichment analysis (chapter 3). I proposed two methods, 'MCOA z' and 'MCOA ratio' for generating enrichment scores based on creating a family of null models of the hybrid gene-ontology graph, and compared them to the scoring procedure, 'MCOA background' used by Frost and McCray (2012). 'MCOA ratio' was found to be too heavily biased towards terms with low annotation counts, but both 'MCOA z' and 'MCOA background' showed the desirable property of returning terms with a range of depths and annotation counts amongst their top rankings when run on real data.

Comparison of these two MCOA scoring procedures with classic over-representation analysis on simulated datasets found that the MCOA methods prioritised the target

terms more successfully than the classic method across three different types of simulation study. The downweighting of multiply annotated genes was found to be a contributing factor in the superior performance of the MCOA methods on unweighted datasets, while the ability of the MCOA methods to incorporate edge weights and thereby downweight unreliable annotations allowed for much better performance on datasets in which a fraction of the annotations were replaced at random.

Finally, the disease benchmark introduced by Tarca et al. (2012) was used to compare the ability of the MCOA methods and the classic method to retrieve relevant HDO terms on human disease datasets. 'MCOA background' was found to be strongly biased towards highly annotated HDO terms, diminishing its ability to discriminate between enriched and non-enriched disease terms. 'MCOA z', in contrast, successfully made this distinction and performed competitively with the classic method. It was also found that including annotation weights helped the MCOA methods prioritise the target disease terms.

The success of the MCOA methods on the benchmarking tasks suggests that MCOA provides a useful framework for enrichment analysis, however they do not prove that MCOA is guaranteed to be better than the classic approach in all circumstances. Performance on simulated datasets may not be a reliable guide to the relevance of results on real datasets, and all of the simulated datasets were constructed using the Gene Ontology and human annotations, so there may be differences when using different ontologies and different annotation sets. The MCOA methods also have their own shortcomings. Throughout this report, I have treated the output of enrichment analysis methods as a list of ontology terms, ranked by enrichment score. However, one of the attractive features of the classic over-representation analysis approach is the fact that it outputs p-values, which can then be used to select only a fraction of terms to regard as significantly enriched. This offers two important advantages: the significance level can be chosen to control the false positive rate, and p-values can be compared across different experiments. In contrast 'MCOA background' scores are not easily compared from one experiment to another, meaning there is no comparable notion of a threshold value which can be used to separate significantly enriched terms from the rest. 'MCOA z' is perhaps better in this regard, since the z-scores could in principle be converted into p-values. However, doing this would involve the assumption that termranks were normally distributed, which is a good approximation only in some cases. Both 'MCOA z' and 'MCOA background' show biases in their overall distributions of enrichment scores, with 'MCOA z' disproportionately favouring terms with low anno-

tation counts and 'MCOA background' disproportionately favouring terms with high annotation counts. The bias in 'MCOA background' was especially pronounced on the HDO, highlighting the fact that the performance of enrichment analysis methods depends on the nature of the ontology being analysed. 'MCOA z' is also relatively expensive computationally, owing to the need to compute null termrank distributions. The computations involved in analysing a single input gene list for this project took on the order of tens of minutes when run in parallel on a laptop, compared to the matter of seconds required to perform the classic and 'MCOA background' analyses. Finally, when using the MCOA methods with gene weights to analyse data from genome-wide expression profiling experiments, as in the disease benchmark experiments of section 4.3, no distinction is made between up- and down-regulated genes, in contrast to GSEA methods.

Addressing some of these shortcomings might be a useful objective for further work. For example, different scoring functions could be considered to correct for the biases of 'MCOA background' and 'MCOA z'. The implementation of 'MCOA z' could be optimized to reduce computation time, or it might perhaps be possible to find approximations to the desired statistics that circumvented the need to compute the full distributions each time. A second interesting direction for further work would involve extensions of the evaluations performed here. I have only compared the MCOA methods to the classic method, but as discussed in section 2.1.4, there are a wide variety of different enrichment analysis methods to which MCOA's performance on the benchmark tasks could be compared. Since these methods tend to address different limitations of the classic method, this sort of comparison could be a useful way to gauge the MCOA methods' strengths.

Finally, one of MCOA's appeals is its capability of handling many different kinds of data. This project has explored using datasets with continuous gene values and annotation confidence scores. MCOA could also naturally incorporate data about relationships between pairs of genes, such as protein-protein interactions or gene co-expression levels. These pairwise scores could be used to add optionally weighted gene-gene edges to the graph used in the MCOA analysis (Frost and McCray, 2012). The full graph would then represent information about the relationships between terms in the ontology, between genes and terms, and between genes, which could prove a powerful basis for analysis.

Bibliography

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.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29.

Ballouz, S., Pavlidis, P., and Gillis, J. (2017). Using predictive specificity to determine when gene set analysis is biologically meaningful. *Nucleic Acids Research*, 45(4):e20.

Bauer, S., Gagneur, J., and Robinson, P. N. (2010). Going bayesian: model-based gene set analysis of genome-scale data. In *Nucleic acids research*.

Bayes, A., Collins, M. O., Croning, M. D. R., van de Lagemaat, L. N., Choudhary, J. S., and Grant, S. G. N. (2012). Comparative study of human and mouse postsynaptic protemoes finds high compositional conservation and abundance difference for key synaptic proteins. *PLoS One*.

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 Series B (Methodological)*, 57(1):289–300.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Ding, Y., Yan, E., Frazho, A., and Caverlee, J. (2009). Pagerank for ranking authors in co-citation networks. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2229–2243.

Frost, H. R. and McCray, A. T. (2012). Markov chain ontology analysis (mcoa). In *BMC Bioinformatics*.

Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. Knowl. Data Eng.*, 15:784–796.

He, X. (2016). *A Semi-automated Framework for the Analytical Use of Gene-centric Data with Biological Ontologies.* PhD thesis, University of Edinburgh, Institute for Adaptive and Neural Computation.

Hong, Y., Downey, T., Eu, K. W., Koh, P. K., and Cheah, P. Y. (2010). A 'metastasis-prone' signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. *Clinical & Experimental Metastasis*, 27(2):83–90.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30.

Lu, Y., Rosenfeld, R., Simon, I., Nau, G. J., and Bar-Joseph, Z. (2008). A probabilistic generative model for go enrichment analysis. In *Nucleic acids research*.

Maslov, S., Sneppen, K., and Zaliznyak, A. (2004). Detection of topological patterns in complex networks: Correlation profile of the internet. *Physica A: Statistical Mechanics and its Applications*, 333(1-4):529–540.

Meyer, C. D., editor (2000). *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Robinson, P. N. and Bauer, S. (2011). *Introduction to Bio-ontologies*. Taylor & Francis US.

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. (2012). Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946.

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. *Proceedings of the National Academy of Sciences of the United States of America*, 102 43:15545–50.

Tarca, A. L., Draghici, S., Bhatti, G., and Romero, R. (2012). Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):136.

White, S. and Smyth, P. (2003). Algorithms for estimating relative importance in networks. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 266–275, New York, NY, USA. ACM.

Xing, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305 – 314.