

CHALMERS



Integrating heterogeneous ranked sources with different reliabilities

A case study of Gene-Disease associations

Master of Science in Bioinformatics and Systems Biology

SABER AHMAD AKHONDI

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
Göteborg, Sweden, March 2011

The Author grants to Chalmers University of Technology and University of Gothenburg the non-exclusive right to publish the Work electronically and in a non-commercial purpose make it accessible on the Internet.

The Author warrants that he/she is the author to the Work, and warrants that the Work does not contain text, pictures or other material that violates copyright law.

The Author shall, when transferring the rights of the Work to a third party (for example a publisher or a company), acknowledge the third party about this agreement. If the Author has signed a copyright agreement with a third party regarding the Work, the Author warrants hereby that he/she has obtained any necessary permission from this third party to let Chalmers University of Technology and University of Gothenburg store the Work electronically and make it accessible on the Internet.

Integrating heterogeneous ranked sources with different reliabilities
A case study of Gene-Disease associations

SABER AHMAD AKHONDI

© SABER AHMAD AKHONDI, March 2011.

Supervisors: MARCUS BJARELAND
DANIEL DALEVI

R &D Information
AstraZeneca
SE-431 83 Mölndal
Sweden
Telephone + 46 (0)31-776 1000

Examiner: GRAHAM J. L. KEMP

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
Göteborg, Sweden March 2011

Abstract

Identifying genes associated with a certain disease, bioprocess or pathway remains a big challenge in pharmaceutical industries, this process is time consuming and costly. To speed up the process candidate genes could be prioritized using ranked lists created by different methods and data sources. Each of these ranked lists comes with different reliabilities; integrating results of these methods are becoming necessary. Several methods have been proposed that can integrate these ranked lists but they do not take in to account the differences in reliability and they do not handle missing data satisfactorily.

In this project, we modified the Discounted Rating System. The MDRS method integrates multiple ranked lists with different reliabilities, regardless of their scoring function and their list size. The reliability of different data sources were chosen through expert knowledge. The method was applied on gene-disease relations. To evaluate the results gold standard gene sets were used and output was analyzed using enrichment plots. By the uses of enrichment plots the performances of different methods and data sources were also observed. To our understanding, the MDRS method is shown to outperform current methods.

The correlation of different data sources and methods were analyzed using Venn diagrams and hierarchical clustering. Distance matrices were created using Spearman's rank correlation method and percentage of data similarities. Finally a method was introduced that would help analysis of a set of genes to find the most relevant diseases to the set.

Acknowledgments

This project was carried out within Research and Development Information at AstraZeneca Mölndal. The project was part of an in-house integrated knowledge platform called PharmaConnect which extracts, integrates and analyses Knowledge to support systematic, evidence based decision making, regardless of discipline or content source.

I would like to express my deepest appreciation to those who supported me through this master thesis. I am indebted to my supervisors Marcus Bjärelund and Daniel Dalevi for overseeing my work, providing insights, ideas, feedback and a supportive supervision during the project. I owe my deepest gratitude to Daniel Dalevi for his guidance, ideas, encouragement and continuous support throughout the project which all lead the project towards a successful completion. I want to express my appreciation to Marcus Bjärelund for giving valuable insights while overseeing the project. I cannot find words to express my gratitude to both of them for reviewing the current paper and providing me feedbacks. Many thanks to Graham Kemp from the Chalmers side for his supervision and being the examiner for this project and providing feedback on this thesis.

I would like to thank Martin Johansson, Johanna Sagemark and the Topps team for providing valuable information throughout the project; also I should appreciate Jing Guo, Shanmukha Sampath and Sukanya Ramasamy for their help and support during the project.

Finally, I would like to dedicate this work to my parents Mehdi Akhondi, Zohreh Behjati and my love Bahareh Beyk, who all have always supported my decisions.

Table of Contents

1	Introduction	5
1.1	Overview of related work	7
1.2	Contribution	7
1.3	Overview of the thesis	8
2	Related literature review	9
2.1	Data sources	9
2.2	Previous approaches to integrate ranked lists	12
2.2.1	Kernel-based data fusion	13
2.2.2	Order statistics used in Endeavour	13
2.2.3	Optimal weight matrix	14
2.2.4	Discounted Rating System	16
3	Method	19
3.1	Modified Discounted Rating System (MDRS)	20
3.1.1	MDRS and multi-reliability in one data source	22
3.2	Data Selection	23
3.3	Method selection	23
3.4	Gold standard gene sets	23
3.4.1	Literature based	24
3.4.2	Based on scientific research at AstraZeneca	25
3.4.3	Based on launched drug phases	26
3.5	Enrichment plot	26
3.5.1	Truncate enrichment plot	27
3.6	Clustering	27
3.6.1	Spearman's rank correlation coefficient	27
3.6.2	Percentages of data similarities	28
3.7	Analysis of sets of genes or diseases	28
3.7.1	Method description	28
3.8	Implementation	31
4	Results	32
4.1	Weighing the sources	32
4.2	Correlations of different data sources	32

4.2.1	Venn diagram	32
4.2.2	Clustering	33
4.3	Performance of the MDRS method.....	35
4.4	Comparison between different sources	37
4.4.1	Comparison of literature based sources.....	37
4.5	Analysis of sets of genes	39
4.5.1	Analysis of gold standard sets.....	39
4.5.2	Disease-Pathway relation	40
4.6	Discussion.....	41
5	Conclusion and future work.....	42
5.1	Discussion.....	42
5.1.1	Method comparison.....	42
5.1.2	Data sources.....	43
5.2	Future work.....	43
5.3	Conclusion.....	43
	References	44
	Appendix	47

1 Introduction

Identifying genes associated with a certain disease, bioprocess or pathway remains a big challenge in pharmaceutical industries. Of more than 2500 human protein coding genes in the Entrez database fewer than 2000 have been listed with human disease phenotypes (Hamosh, *et al.*, 2005).

In general a list of candidate genes involved in a certain disease could be created by using different methods and a variety of data sources. Narrowing down this list to a few dozen by experiments and lab work would be too expensive and time consuming (Tranchevent, *et al.*, 2008). Data on gene-disease relationships are available through multiple data sources. Prioritization on these data would help the scientific community reduce cost and resources. Prioritizing candidate genes from most to least promising genes related to a biological process is called *gene prioritization* (De Bie, *et al.*, 2007).

During the past five years, several strategies to establish gene prioritization has been developed (Oti and Brunner, 2007; Tranchevent, *et al.*, 2008; Zhu and Zhao, 2007). So far candidate gene prioritization has been done by utilizing some of the available sources (Sun, *et al.*, 2009). However most studies only focus on one single type of data sources (Friedman, *et al.*, 2000) for example GAD or HuGE.

To deal with the large amounts of information available in a drug project, integrating results of different methods and different data sources is becoming necessary. Prioritizing the final output by providing a ranked list would speed up scientific research and reduce the cost of scientific experiment (Figure 1-1).

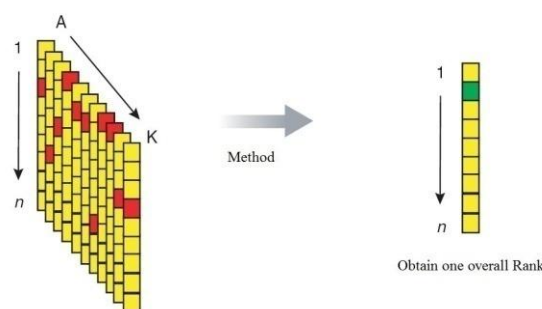


Figure 1-1

Gene prioritization through genomic data fusion (Aerts, *et al.*, 2006)

A variety of methods have been applied to integrate multiple data sources. Two of the major issues of these methods are:

- 1- They do not take differences in reliability of the data sources in to account.
- 2- They don't handle missing data satisfactorily.

As a motivational example we will first look at a gene-disease relation. Consider the Peroxisome Proliferator-Activated Receptor Gamma (PPARG) gene. It influences pancreatic beta-cell function resulting in changes in insulin secretion and insulin sensitivity of the peripheral tissues. This gene is currently known to be involved in obesity, diabetes, atherosclerosis and cancer (Ji and Huang, 2006;

Tanko, *et al.*, 2005). We looked up diseases related to PPARG in three different data sources. The results are present in Table 1.

DISEASEID	GAD	Rank	MTRA	Rank	HuGE	Rank
Metabolism system disease	1	45	1657	1	1	75
Glucose metabolism disorder	1	45	1507	2	3	38
Endocrine system disease	-	-	1240	3	-	-
Cancer	2	28	1135	4	3	38
Diabetes mellitus	8	9	1008	5	19	9
Cardiovascular system disease	4	18	888	6	16	11
Hyperinsulinism	2	28	781	7	4	29
Insulin Resistance	26	3	768	8	146	3
Non-insulin-dependent diabetes mellitus	79	1	730	9	169	2
Vascular disease	-	-	725	10	-	-
Nutrition disorder	-	-	570	11	-	-
Obesity	55	2	550	12	288	1
Brain disease	-	-	498	13	-	-
Colorectal Cancer	11	6	398	14	25	6
Inflammation	11	6	384	15	16	11
Digestive system cancer	-	-	362	16	-	-
Gastrointestinal disease	-	-	338	17	-	-
Nervous system disease	-	-	337	18	-	-
Neoplasms, Glandular and Epithelial	1	45	331	19	2	52
Arterial Occlusive disease	-	-	295	20	-	-

Table 1

Result of diseases most relevant to PPARG gene based on three different data sources, GAD (Section 2.1.1), MTRA (Section 2.1.3) and HuGE (Section 2.1.2).

As can be observed data is sorted based on the MTRA method.

Some of the diseases with known relations to PPARG are labeled red in the table.

Data looks messy and not easily usable.

Table 1 contains the most relevant diseases to PPARG gene based on three different data sources. GAD (Section 2.1.1), MTRA (Section 2.1.3) and HuGE (Section 2.1.2). Row one of the table shows “Metabolism system disease”. The score of this disease in GAD is 1 while its ranks position is 45. This disease is ranked in first position based on the MTRA method with the score of 1664.

Note the challenges in interpreting Table 1 due to the differences in rank from the sources. For example, compare “Inflammation” and “Diabetes Mellitus”. Two of the sources (MTRA, HuGE) agree on ranking “Inflammation” above “Diabetes Mellitus”, while the third disagrees. This is a very common situation.

Table 2 contains diseases related to PPARG gene based on the MDRS score. The MDRS method integrates these three sources considering different weights for each source based on their reliabilities. “Metabolism system disease” has lost its first rank position despite its high MTRA score and has moved to rank number six. “Non-insulin-dependent diabetes mellitus” has replaced first rank according to the MDRS score. Known diseases related to PPARG gene (colored in red) have also been found in top ranks meaning we are retrieving more reliable hits at earlier stages of the ranked list.

DISEASEID	GAD	Rank	MTRA	Rank	HuGE	Rank	Score
Non-insulin-dependent diabetes mellitus	79	1	730	9	169	2	7.14364759043436
Obesity	55	2	550	12	288	1	6.39771568642766
Insulin Resistance	26	3	768	8	146	3	3.59622073065477
Metabolic syndrome X	18	4	174	37	27	5	2.43986641141494
Hypertension	17	5	159	38	32	4	2.3231373568056
Metabolism system disease	1	45	1657	1	1	75	1.81053898947496
Colorectal Cancer	11	6	398	14	25	6	1.69749496695827
Diabetes mellitus	8	9	1008	5	19	9	1.64114445583446
Inflammation	11	6	384	15	16	11	1.56375621052194
Coronary heart disease	9	8	130	44	11	16	1.34815301255298
Weight gain	5	13	70	74	22	7	1.34596613101022
Obesity, Morbid	6	11	15	211	21	8	1.30939943734667
Glucose metabolism disorder	1	45	1507	2	3	38	1.2891830747697
Insulin-dependent diabetes mellitus	8	9	35	131	11	16	1.27862326776809
Cardiovascular system disease	4	18	888	6	16	11	1.23256205874184
Polycystic ovary syndrome	5	13	29	142	16	11	1.21464028984641
Diabetic nephropathy	6	11	49	97	10	19	1.2089695618289
Myocardial Infarction	5	13	43	107	12	14	1.20658800851179
Glucose intolerance	5	13	37	125	11	16	1.18381001424844
Diabetic retinopathy	5	13	18	187	6	22	1.11331163827965

Table 2

Diseases related to PPARG gene. These diseases are ranked based on Modified Discounted Rating System method.

MDRS method has integrated these data sources by considering different reliabilities for each ranked list.

We can see that “Metabolism system disease” has lost its first rank position despite its high MTRA score.

1.1 Overview of related work

Attempts to integrate ranked lists have been recently taken in methods such as Kernel-based data fusion (Section 2.2.1), Order statistics (Endeavour, Section 2.2.2), Optimal weight matrix (Section 2.2.3) and Discounted Rating System (DRS, Section 2.2.4).

Research has also been done in Information Retrieval studies (IR). Among the most cited are the work of Fox and Shaw (1994) and Freund, *et al.* (2003) where they present various methods for the problem. The majority of IR methods cannot be discussed since they are heavily dependent on IR similarities and therefore could not be generalized.

1.2 Contribution

The Discounted Rating System can integrate ranked list where the sources have different reliabilities. The method is fast and seems in general to produce better results compared to the other available methods studied in this project.

In this project we have modified the DRS method in order to integrate different sized ranked lists with missing data. The MDRS method produces more reliable results than the individual sources, i.e. we get better ranks by combining several data sources.

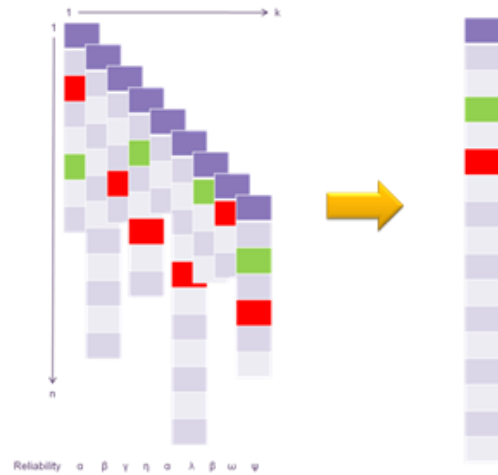


Figure 1-2

The MDRS method can integrate different sized ranked list with missing data. In MDRS different weights are assigned to ranked lists according to their reliability.

The final ranked list will contain all the elements in the other sources ranked from most relevant to least relevant.

1.3 Overview of the thesis

In the Related review section (Section 2.1) we will describe the available data sources from literature for gene-disease relations. We also describe previous approaches taken to integrate ranked list (Section 2.2).

The Method section (Section 3) describes the MDRS method together with the gold standard gene sets (Section 3.2). We will then describe enrichment plots (Section 3.5) and clustering methods (Section 3.6), which we later used to analyze our results. The Method section ends with defining a method for the analysis of sets of genes or diseases (Section 3.7). All results and future works are then described in the Results and Conclusions (Section 4 and 5).

2 Related literature review

In this section, we will first briefly discuss a few different data sources that produce ranked lists based on gene-disease and disease-gene relations. We will also review some of the previous approaches taken to integrate multiple ranked lists. One of these approaches is the Discounted Rating System (DRS). We have decided to improve the DRS method and use a modified version of the method (MDRS). (For this reason we will spend more time, and give more details about the DRS method.)

2.1 Data sources

Different data sources are used in scientific research to produce gene-disease and disease-gene ranked lists (Tiffin, *et al.*, 2006; Yu, *et al.*, 2008). Here we will describe the different data sources that have been used in this project. Some other data sources were also investigated but they were not used in the project (Appendix section A- I).

2.1.1 The Genetic Association Database (GAD)

The Genetic Association Database (GAD; <http://geneticassociationdb.nih.gov/>) is a literature based database. As described in Becker, *et al.* (2004), GAD is designed as a public repository of genetics association data. Queries can be written in a systematic manner. Analysis in GAD can be done in the context of modern high-throughput assay system and current annotated molecular nomenclature. "GAD aims to collect standardize and archive genetic association study data and make it easily accessible to the scientific community" (Becker, *et al.*, 2004).

GAD was designed to overcome OMIM's (Hamosh, *et al.*, 2002) drawback on comparing large set of molecular data and its lack of finding lower significant or negative findings (Becker, *et al.*, 2004). Relations in GAD are manually curated. Therefore, gene-disease and disease-gene relations provided by GAD come with high reliabilities.

For the purpose of this project, ranked lists are created from GAD data, based on the number of documents relating diseases and genes. "Each of the records in GAD is annotated with links to molecular databases (LocusLink, GeneCards, HapMap, etc.) and reference databases (PubMed, CDC)" (Becker, *et al.*, 2004). Therefore mapping of the data can be done using Entrez identification to ugene and udisease vocabularies available at AstraZeneca.

2.1.2 Human Genome Epidemiology Navigator (HuGE)

Due to the large scale genetic association studies and the rapid growth in the amount of publications in human health and disease, Human Genome Epidemiology Network (HuGENet) has maintained a database of published, population based epidemiologic studies of human genes extracted from PubMed, Yu, *et al.* (2008). HuGE provides information on "population prevalence of genetic variants, gene-disease association, Gene-Gene and Gene-Environment interaction and evaluation of genetic tests" (HuGE; <http://hugenavigator.net/>).

HuGE navigator is a literature based source that produces ranked lists of genes relevant to a certain disease or vice versa. Terms in HuGE are sorted on the number of documents relevant to each gene or

disease. HuGE data is also mapped using Entrez Identification to ugene and udisease data available at AstraZeneca.

2.1.3 MeSH Term Relevance Analysis (MTRA)

MeSH Term Relevance Analysis (MTRA) is a statistical machine learning method (Johansson, 2008) that takes into account the use of metadata. MTRA uses the chemical MeSH term in Medline publications associated with an entry in EntrezGene as training data. It uses these data to create a semantic fingerprint for each gene. A threshold is used to check the extent of the fingerprint. MTRA aims to improve the accuracy when searching for a gene in the literature.

MTRA is a literature based method which is available at AstraZeneca. Rank of the outputs in MTRA is based on the frequency of relevant documents. Entrez identification of a gene is used to map the data to the ugene, udisease database.

2.1.4 Peregrine

One of the services available in-house at AstraZeneca is Peregrine. Peregrine provides two different types of services. The first service is a text markup service which finds entities inside a text string. The second service is a synonym service which returns synonyms for a given term.

By using the text markup service provided by Peregrine we can retrieve entities with their synonyms and their place of occurrence. The Peregrine gene-disease relationship database was created by setting two parameters: entities to “disease and gene” and occurrence of the terms to “in one sentence”. The data is then ranked based on the frequency of the term.

2.1.5 MOAh, MOAm, MOAI

Production of a drug involves several years of study. During this time a drug should go through several phases. In order for a drug to be approved it should go through clinical trials involved in drug production. Clinical trials are classified into four different phases. Each phase is then divided into several sub phases. If a drug passes all four phases it would then be moved to phase IV (post approval studies). In each phase gene-disease relationships are present.

Through Mode of Action (MOA) database we can produce ranked lists with multiple reliabilities. The ranked lists are sorted based on the number of times a gene-disease relationship has been observed in different drugs. We have gathered the data from an in-house database available at AstraZeneca (Knowledge Engineering). We categorized clinical data into three grouped (high, medium, low) and ranked each group separately.

- MOAh contains the following phases: Launched, Registered, Pre-Registered, Phase IV, Phase III
- MOAm contains: Phase II/III, Phase II, Phase I/II, Phase I
- MOAI contains: Phase 0, Preclinical, No Development Reported

2.1.6 CoPub

CoPub (<http://services.nbic.nl/copub/portal/>) is a literature based method that uses co-occurrence of biomedical concepts in abstracts from Medline to establish relations between biomedical concepts. These biomedical concepts include “all human, mouse and rat genes, biological processes, molecular functions and cellular components from Gene Ontology, and also liver pathologies, diseases, drugs and pathways”(Frijters, *et al.*, 2008) .

CoPub (Frijters, *et al.*, 2010) has the ability of finding new biological relations. The basic idea behind the method is that there is a hidden relation between two biological concepts, called A and C, when there is a relation between concept A and an intermediate concept B, and also a relation between concept C and intermediate concept B. Concept B could be a known concept (Closed Discovery) or an unknown concept (Open Discovery).

CoPub uses R-scaled score adapted from Wren’s minimal MIM model to rank the output query (Wren, 2004).

CoPub is publically available. For the purpose of this project CoPub version 2.5 alpha is used. We set the parameters so that results would contain relations with more than or equal to one publications. Data is then ranked based on default R score setting. Mapping of the data is done using Entrez identification to ugene database available at AstraZeneca.

2.1.7 Novo|Seek

Novo|Seek (<http://www.novoseek.com/>) is a free biomedical search engine that indexes PubMed, and U.S. Grants databases. Novo|Seek offers the user, information about concepts that are related to a given query (Allende, 2009). Novo|Seek uses intelligent agents to find biomedical concepts such as diseases, genes, drugs or chemical substances related to the search. It creates a ranked list based on a statistical measure of how unlikely it is to find a certain number of documents in the results of a search.

For the purpose of this project ranked lists are created based on the filters provided by Novo|Seek. These filters show how relevant a gene is to a certain disease. Entrez Identification is used to map these genes into ugene database available at AstraZeneca.

2.1.8 GeneRanker

GeneRanker (<http://cbioc.eas.asu.edu/generanker/>) is a publically available application used for mining gene-disease relations. GeneRanker uses Protein-Protein and gene-disease interactions available in the CBioC database to construct networks. GeneRanker uses graph theory to create a ranked list of the best candidate genes in the network (Gonzalez, *et al.*, 2008). CBioC database is located at Arizona State University. GeneRanker uses Entrez Identifiers which we used to map the data to ugene database available at AstraZeneca.

2.1.9 Endeavour

With the use of a training set, Endeavour (Aerts, *et al.*, 2006) takes a machine learning approach to rank genes based on their similarities to the training sets. To prioritize genes, Endeavour uses correlations for vector space data and BLAST for sequence data sets.

Endeavour integrates ranked data from microarray, InterPro, BIND, sequence, GO annotation, Motif, KEGG, EST and text mining using order statistics method. It uses order statistics of an N-dimensional distribution to calculate the final score of each candidate (Yu, *et al.*, 2008). In their last update (Tranchevent, *et al.*, 2008), Endeavour has made new developments which consist of additional data sources, new organisms (*Mus musculus*, *Rattus norvegicus* and etc.) and a web based user interface (Endeavour, <http://homes.esat.kuleuven.be/~bioiuser/endeavour/index.php>).

2.2 Previous approaches to integrate ranked lists

The sources described above produce ranked lists of genes or diseases based on the type of query that is applied to them. As previously discussed some of these data sources integrate multiple ranked lists to produce a final ranking. For example, Endeavour integrates data from microarrays, InterPro, Bind and Ext.

Research has also been going on to integrate multiple data sources in Information Retrieval (IR) studies. An example is the work of Fox and Shaw (1994) who described six different approaches to combine the similarity values in IR systems. Another example is the work of Freund, *et al.* (2003) who designed an algorithm for ranking when numerical weighting is not available. IR methods have also been analyzed by other researchers such as (Lee, 1997 [a]; Lee, 1997 [b]), but a large majority of these method are not applicable to this project since they are heavily dependent on IR similarities and therefore could not be generalized.

Table 3 lists, four different methods that have been previously applied to integrate ranked lists from different data sources. One of these methods is based on an IR approach. We will spend more time describing this method since we have decided to improve and use it for this project.

Table 3

Table presenting four methods taken to integrate ranked lists from different data sources.

Method	Need of training data	Ability to weigh data sources	Speed	Source
Kernel-based data fusion	Yes	Yes, expert knowledge	Fair	(De Bie, <i>et al.</i> , 2007)
Order statistics (Endeavour)	No	No	Fair	(Aerts, <i>et al.</i> , 2006)
Optimal weight matrix	Yes	Yes, randomly weighed	Slow	(Sun, <i>et al.</i> , 2009)
Discounted Rating System	No	Yes, expert knowledge	Fast	(Li and Patra, 2010)

2.2.1 Kernel-based data fusion

Kernel-based data fusion (De Bie, *et al.*, 2007) is a method that integrates different ranked lists using a set of training genes.

In this method relationships between genes and diseases are represented in space as vectors. Genes found by one data source are shown as open circles in space (Figure 2-1). A set of genes known to be related with the disease, known as training set are also represented in space in the same manner -filled circles in Figure 2-1.

The method then finds a hyperplane using a kernel function and separates all of the training genes from the origin. This is such that all training data lies on one side of the hyperplane while the distance between the hyperplane and the origin is maximized. Now using the distance of the genes to the hyperplane prioritization of genes is done (De Bie, *et al.*, 2007).

Integration of different data sources is done by combining different kernels, each representing different data sources. A kernel matrix is then designed in a way that the margin would be maximized to the origin. In the next step using one class support vector machine, prioritization of the candidate gene is done. The smaller the distance of the gene to the hyperplane the better position the gene would get in the final ranked list.

Weighing different data sources could also be done in the kernel combination stage, meaning that this method has the ability to weigh different data sources based on their reliabilities.

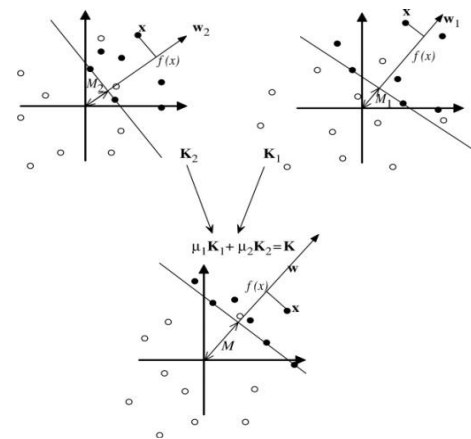


Figure 2-1

“Schematic representation of the hyperplane separating the training genes (filled circles) from the origin, along with the negative genes (open circles). Combining two kernels in an optimal way leads to a new space, where the distance of the positive genes to the origin is larger”. Figure and text taken from (De Bie, *et al.*, 2007).

2.2.2 Order statistics used in Endeavour

Endeavour, Aerts, *et al.* (2006) uses a machine learning approach to integrate ranked lists. Endeavour uses an order statistics which was previously discussed by Stuart, *et al.*(2003).

A Q-statistic is calculated for each element observed in multiple ranked lists (Equation 1). The scores are then fitted to a specific distribution. Elements are then ranked based on this score.

To calculate the Q-statistics ranked lists are first sorted according to their scores, so that the rank ratio of each element could be calculated. This ratio is based on the probability of observing the specific element of the ranked list by chance, in its current position (r_i) on data source i . This is done by dividing each rank by the total number of ranked elements in the data source, considering that if an element lacks a rank, it is not counted as an element and therefore it would be observed as a missing data. By

calculating the observed ratio on each data source now a Q statistic score for each element is calculated using the join cumulative distribution of N-dimensional order statistics (Equation 1).

Equation 1

$$Q(r_1, r_2, \dots, r_N) = N! \int_0^{r_1} \int_{s_1}^{r_2} \dots \int_{s_{N-1}}^{r_N} ds_N ds_{N-1} \dots ds_1$$

where N is the number of data sources and r_i is the rank ratio of the specific element in data source i .

An exact formula was used by Aerts, *et al.* (2006) with a complexity of $O(N^2)$. This is illustrated in (Equation 2). We derive the formula in Appendix section A- VIII.

Equation 2

$$Q(r_1, r_2, \dots, r_N) = N! V_N, \quad V_k = \sum_{i=1}^k (-1)^{i-1} \frac{V_{k-i}}{i!} r_{N-k+1}^i$$

where $V_0 = 1$ and $r_0 = 0$.

Aerts, *et al.* (2006), discovered that the Q statistic scores do not follow a uniform distribution but a beta distribution for less than five data sources and a gamma distribution for more than five data sources. Endeavour fits data to these distributions in the final step (Yu, *et al.*, 2010).

2.2.3 Optimal weight matrix

Another approach to integrate multiple data sources has been taken by Sun, *et al.* (2009). In their approach they assign random weights to different data sources, and produce a final ranking by integrating the ranked lists using these weights based on a gold standard. They choose the weights generating the best result. The optimal weight matrix approach consists of 8 steps that are described below:

- Step 1: All the ranked lists are sorted based on their scores.
- Step 2: All sources are aligned together in a table so that each element corresponds to a row and each source a column.
- Step 3: A weight vector with a given size is assigned to each data source. The weight vector consists of random weights.
- Step 4: A weight matrix containing the combined score of each element in the merged matrix is formed, for all the possible combinations of weights among different data sources. Each score in the weight matrix is calculated with the use of Equation 3.

Equation 3

$$S = \sum_{i=1}^N w_j * r_i$$

where N is the number of data sources, w_j is the element j of weight vector and r_i is the rank position of an element in the merged matrix in data source i .

With K different weights a K^N weight matrix is produced. So that the best weight vector could later be chosen from the weight matrix.

- Step 5: In order to choose the best weight matrix a set of known genes with relations to the disease is extracted from available resources (Gold Standard genes). This is based on the knowledge of the user according to the type of the ranked list.

In the original research Sun, *et al.*(2009) applied the method to Schizophrenia to find the best candidate genes. A set of known gene-diseases were chosen as core genes.

- Step 6: For each list in the weight matrix the final ranked list is produced by sorting the elements using the combined scores (Equation 3).
- Step 7: Two parameters are introduced for each ranked list in the weight matrix: proportion of core gene presented in the top elements ($m, \%$), and the proportion of all other candidate genes that are in the top s positions ($n, \%$).
- Step 8: The best weight vector is chosen in a way that m is maximized and n is minimized. Using this weight vector, the ranked list is created.

2.2.4 Discounted Rating System

In the work of Li, *et al.* (2010) a new strategy has been taken to combine ranked lists derived from multiple data sources with different reliabilities. The authors define a scoring system called Discounted Rating System (DRS), which is inspired by Discounted Cumulated (Kalervo Järvelin and Kekäläinen, 2000). The DCG score is widely used to evaluate results obtained from methods doing document retrieval.

The Discounted Rating System, designed by Li, *et al.* (2010), performs the rank-integration of data sources in several steps. The DRS method assumes that all of the elements present in any of the ranked lists are also present in all other ranked lists. We will later show how we can modify this method to overcome this restriction.

- Step 1: Ranked list for each data source is created. The ranked lists are produced by sorting the results using some scores which could be counts, percentages, P-values, etc. (Figure 2-2). Ranks are assigned from 1 to n and $r_i \neq r_{i+1}$, where r_i is the rank of element i , meaning that no duplication is allowed.
- Step 2: Each ranked list is divided in to five equally sized groups. Each group is then assigned a rate score from 1 to 5 depending on to which bin they belong (Figure 2-3).

Rank	count
1	45
2	30
3	20
4	11
5	11
6	8
7	4
8	4
9	1
10	1
11	1

Figure 2-2

Step 1-Ranked list are created based on scores generalized by the data source. In this table scores are based on number of documents or counts.

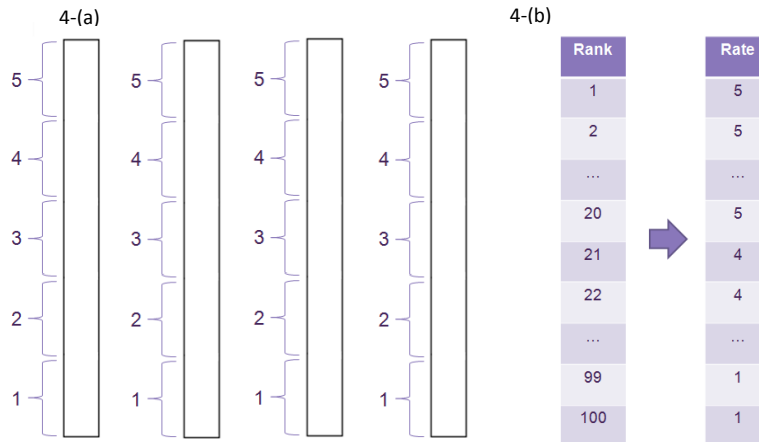


Figure 2-3

4-(a) Step 2-based on the amount of data which is equal in all data sources each ranked list is divided into five equally sized groups. Each group is assigned a rate score starting from 5 for the elements with r_i high ranks to 1 for the element with low ranks.

4-(b) Assigned rates for every element are stored in a rate list.

- Step 3: All sources are merged together so that for each element a ranking and rating is available for all sources (Figure 2-4).

element	Source 1	Source 2	...	Source i
1	47	15	...	51
2	8	43	...	58
3	1	65	...	5
4	45	55	...	38
5	14	19	...	55
...
100	16	1	...	34

element	Source 1	Source 2	...	Source i
1	3	5	...	3
2	5	3	...	3
3	5	2	...	7
4	3	3	...	4
5	5	5	...	3
...
100	5	5	...	4

Figure 2-4

Step 3- Two different matrices are created where the left contains merged rate lists and the right contains merged ranked lists.

Based on the formula discussed in Li, *et al.* (2010), the rank score and the rate score of elements of separate sources are now combined into one value called the Discounted Rating (dr):

Equation 4

$$dr_i = \frac{rating_i}{\log_2(r_i+1)}$$

where $rating_i$ is the rating of investigated element based on data source i and r_i is its rank position.

- Step 4: Using Equation 4, a new matrix is calculated which contains the discounted rating scores. We are hence creating a discounted rating matrix (Figure 2-5).
- Step 5: A combined scores of all data sources is calculated by taking the mean value of all DRS values.

element	Source 1	Source 2	...	Source i
1	0.53	1.25	...	0.53
2	1.58	0.55	...	0.51
3	5	0.33	...	2.70
4	0.54	0.52	...	0.75
5	1.28	1.16	...	0.52

Figure 2-5

Discounted Rating Score Matrix, derived from ranked list matrix and rate list matrix using discounted rating formula.


Equation 5

$$S_{dr} = \frac{1}{N} \sum_{i=1}^N dr_i$$

where N is the number of data sources.

- Step 7: Now a list is available with the d_r scores relative to each element. The final ranked list is created by sorting the list using the scores (Figure 2-6).

element	dr Score
1	0.65
2	0.74
3	2.31
4	0.53
5	0.94
...	...
100	1.83



Rank	Score
1	5
2	4.8
3	3
4	2.91
5	2
...	...
100	0.003

Figure 2-6

Step 6: The list is sorted using the dr Score and Final Ranked list is created.

2.2.4.1 Weighted DRS

Different data sources have different reliabilities. To weigh data sources a new variable was introduced called μ_i .

By introducing this variable we can replace Equation 5, with:

$$S_{wdr} = \frac{\sum_{i=1}^N (\mu_i dr_i)}{\sum_{i=1}^N \mu_i}$$

where μ_i is the weight of data source i . Note that if all the sources have the same weight $S_{wdr} = S_{dr}$.

3 Method

In this section we will discuss a modified version of the Discounted Rating System that can be used to integrate ranked lists from sources with different reliabilities. The method section is divided into seven subsections.

- 1) We provide a full description of the Modified Discounted Rating system method (MDRS). The original DRS method is described in Section 2.
- 2) We name the selected data sources and the selected method for this project.
- 3) We introduce gold standard gene sets for several diseases. These sets are genes with known relations to a disease. Gold standard gene sets are later used for the purpose of evaluating the MDRS method. Two different approaches were taken to produce gold standard gene sets. In the first approach we gathered data from literature studies. In the second approach, gold standard gene sets were created using launched drug phases.
- 4) We describe enrichment plots that are used to visualize and analyze the results from both individual and combined sources. These plots are later used to compare the performance of different methods.
- 5) The correlation between different data sources were analyzed using hierarchical clustering using either, Spearman's correlation distance or percentage similarity distance.
- 6) A method to analyze sets of genes or diseases is described. The basic idea is, given a set of genes, is there a way to find the most relevant disease to this set. This method is later applied to gene sets from pathways and diseases.
- 7) Implementation of the method is finally described at the end of this section.

3.1 Modified Discounted Rating System (MDRS)

The DRS assumption that all ranked lists are equally sized is not fulfilled for most gene disease data. A solution would be to only apply DRS to the intersection of the elements for all sources. This would seriously impact the utility of the method as we will see later on. Instead we will modify DRS to accommodate the differences in elements, as well as other specific features of the implied domain. We made the following modifications:

1. Deal with unequal ranked lists.
2. Give same ranks to same scores.
3. Give higher priority to top δ ranks in each list.
4. Assign rating of zero to missing data.
5. Move same ranks to same bins.

Modification on DRS gives it the ability of ranking different size lists with missing data. The MDRS method is done in seven steps that we will now describe:

- Step 1: Ranks are assigned from 1 to n based on a list sorted with repeat to some score, with the condition:

$$r_i = r_{i+1} \text{ if } element_i = element_{i+1}$$

where r_i is the rank position of element i in the ranked list (Figure 3-1).

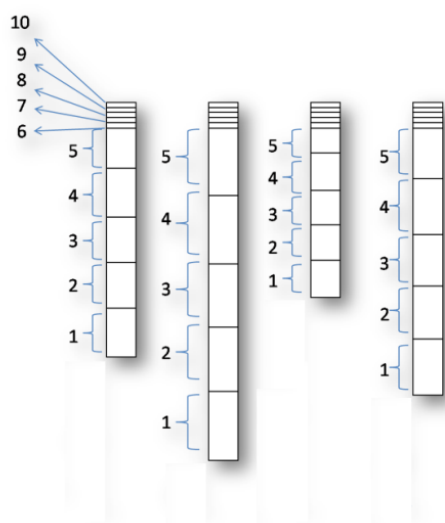
By taking this step, we have given same priority to elements with same scores. This will obviously increase the performance of the method.

- Step 2: Rate list are created out of each ranked list. This time by dividing the ranked list in to $5 + \delta$ groups and assigning a rate score from $5 + \delta$ to 1 to each group. For example for a case of $\delta = 5$, first five elements of each ranked list would get a rate score from 10 to 5 respectively, where the element with the highest rank would get the highest rate score of 10 and so on. Furthermore the remaining ranks with no rate score are then divided into five equally sized groups. The size of the groups is based on their own

Rank	Element Score
1	45
2	30
3	20
4	11
4	11
6	8
7	4
7	4
9	1
9	1
9	1

Figure 3-1

Elements in ranked lists are ranked, based on their score on a descending order while, elements that have the same score would get the same rank.



list size. A rate score of 5 to 1 is given to each group as illustrated in (Figure 3-2), where the group containing the highest rank will get a higher rate score.

Figure 3-2

Based on $\delta = 5$ rate lists are created by assigning rates from $10 \rightarrow 5$ to the first five elements of each ranked list and dividing the rest into five equal sized groups and assigning the rate of $5 \rightarrow 1$ to them.

The basic idea for this configuration is to lift up, the highest candidates of the ranked lists. We have chosen to lift up the first five candidates, while this number is chosen arbitrary. We have reached this decision after applying the method on different data sets.

- Step 3: Each ranked list is checked with its rate list respectively. Elements that have the same rank, but differ in rate, will be shifted so that the element with the lower rate score is assigned the higher rate score.

Our motivation for this step is to treat elements with the same rank equally.

In this step we will also assign a zero score to missing data in each ranked list.

We will consider missing data as elements that are present in any other ranked lists, but are missing in the given list (Figure 3-3).

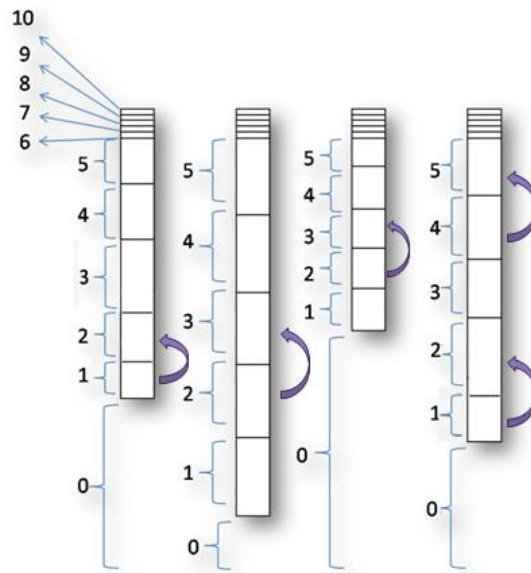


Figure 3-3

Based on $\delta = 5$, rate scores are updated by shifting elements with different rate score but having the same rank score to the group that contains the higher rate scores. Missing data are assigned a zero rate.

- Step 4: All sources are aligned together in a table so that each element corresponds to a row and each source a column (Figure 3-4).

Element	Source A	Source B	Source C	Source D
1	47	15	63	-
2	8	43	78	58
3	1	-	17	5
4	45	55	62	-
5	14	19	-	55
...				
100	16	1	66	34

Merged Rate list

Element	Source A	Source B	Source C	Source D
1	3	5	2	0
2	5	3	2	3
3	10	0	5	7
4	3	3	2	0
5	5	5	0	3
...				
100	5	10	2	4

Merged Rank list

Figure 3-4

Unique elements in all sources are found. Two tables are formed that show the rank and rate position of each unique element based on multiple data sources. If an element is missing in one ranked list it is treated as missing data and assigned a score of zero.

- Step 5: Discounted Rating score is calculated based on the discounted rating formula and the result is stored in the Modified Discounted Rating table:

$$dr_i = \frac{rating_i}{\log_2(r_i+1)}$$

where $rating_i$ is the rating of element based on data source i and r_i is its rank (Figure 3-5).

- Step 6: Weighted discounted rating score previously discussed by Li, *et al.* (2010) is used to create the final score for the unique elements in all ranked lists:

$$S_{wdr} = \frac{\sum_{i=1}^N (\mu_i dr_i)}{\sum_{i=1}^N \mu_i}$$

where μ_i is the weight of data source i , based on its reliability.

Element	Source A	Source B	Source C	Source D
1	0.53	1.25	0.33	0
2	1.58	0.55	0.33	0.51
3	10	0	1.20	2.70
4	0.54	0.52	0.33	0
5	1.28	1.16	0	0.52
...				
100	1.22	10	0.32	0.77

MDRS Score Matrix

Figure 3-5

MDRS Score Matrix is calculated using Discounted Rating formula based on merge rate matrix and merged rank matrix. By combining rank score and rate score MDRS score for missing data would be zero.

MDRS Score		Rank	MDRS Score
4.3		1	10
8.83		2	9.7
10		3	6.2
9.7	Sort	4	4.3
3.42		5	3.91
		...	
		100	0.003

Figure 3-6

A new ranking list is created by sorting on the MDRS scores.

- Step 7: A final ranked list is created by sorting the S_{dwr} scores (Figure 3-6).

3.1.1 MDRS and multi-reliability in one data source

It is possible that a data source can be used to generate ranked lists with different reliabilities. In such cases we can deal with each list separately, by assigning them different reliability weights (μ_i). Then we can use the MDRS method to integrate them together with other ranked lists.

3.2 Data Selection

Ranked lists were obtained from the Genetic Association Database (GAD), Human Genome Epidemiology navigator (HuGE), MeSH Term Relevance Analysis (MTRA) of PubMed (Yu, *et al.*, 2008) documents, Mode of Action High (MOAh), Mode of Action Medium (MOAm), Mode of Action Low (MOAl) and Peregrine. These ranked lists were combined using the MDRS method. We also imported ranked lists from CoPub, Novo|Seek and GeneSeeker using their online user interface (see Section 2.1).

3.3 Method selection

Integration of ranked lists is done using different methods. A description of some of these methods is given in section 2.2. Among the discussed methods are Kernel based data fusion method and optimal weight matrix. Calculating the final ranked list of both of these methods, needs training sets with known gene-disease relations. Our assumption for this project is that core genes or gold standard genes are not available for all of the disease, therefore these methods were not chosen. It should be added that both of these methods consume a longer time to produce their final results.

We chose and modified Discounted Rating System. This was done because of the advantages of this method over the order statistics method in Endeavour. Modified DRS has the ability of integrating ranked lists with missing data, meaning that the ranked lists would not be the same size. This fact is not through for the Order statistics method in Endeavour. The MDRS method has also the ability to weigh data sources with different reliabilities or even it can deal with data sources with multi reliabilities while the order statistics method discussed in Endeavour considers all the data sources as equally reliable.

The MDRS method performed better than the order statistics method even when the sources are weighed equally (Section 5.3A- V). The method would also produce the output 120 times faster than the order statistics method on equally weighed ranked lists of 20 sources. The MDRS method was applied on the whole Human Genome and it produced the outputs in less than 3 hours of CPU time on a desktop computer.

3.4 Gold standard gene sets

Sets of genes that are known to be involved in a certain disease are presented in this section. We will refer to these gene sets as gold standards. Reliable gold standard gene sets help us understand how well different methods can integrate data from multiple sources. Gold standard gene sets are derived using two different approaches: a literature based approach and a launched drug phases approach. As also discussed in Yu, *et al.* (2008) a perfect prioritization should rank the gene with the causal link to a biomedical concept, represented by the training set at the highest position. A summary of these gene sets are given in Table 4.

Table 4

Gold Standard Gene Sets

NAME	Disease	No. of genes	Source
Obesity 14g	Obesity	14	(Speliotes, Willer <i>et al.</i> 2010)
Obesity 13g	Obesity	13	(Bell, Walley <i>et al.</i> 2005)
Schizophrenia	Schizophrenia	38	(Sun, Jia <i>et al.</i> 2009)
Alzheimer 6g	Alzheimer's Disease	6	(Petteri Sevon , Lauri Eronen <i>et al.</i> 2006)
T2D	Type 2 Diabetes	9	(Petteri Sevon , Lauri Eronen <i>et al.</i> 2006)
Dyslipidemia	Dyslipidemia	91	(Johansson Personal Communication)
Hypertension 15g	Hypertension	15	(Johansson Personal Communication)
Alzheimer 63g	Alzheimer's Disease	63	Launched drug phases
Asthma	Asthma	64	Launched drug phases
Hypertension 55g	Hypertension	55	Launched drug phases
Multiple Sclerosis	Multiple Sclerosis	32	Launched drug phases
Myocardial Infection	Myocardial Infection	38	Launched drug phases
Pain Disorder	Pain Disorder	52	Launched drug phases
Rheumatoid Arthritis	Rheumatoid Arthritis	42	Launched drug phases

3.4.1 Literature based

Varieties of genes that have been proven to be involved in diseases have been chosen from different literature studies. Performance of different methods and ranked list could later be investigated using these sets. Here we discuss each of them separately.

3.4.1.1 Obesity13g, Obesity 14g

Obesity occurs when increase in body fat has passed a position that will cause health problems. The disease can lead to other diseases such as type 2 diabetes and cardiovascular disease (NIH-Publication, 1998; Speliotes, *et al.*, 2010). The disease is mostly based on two major factors, changes in human lifestyle for example fast foods and genetic causes (with heritability estimates (h^2) of ~40%–70). This differs in different regions of the world (Maes, *et al.*, 1997; Speliotes, *et al.*, 2010). Since a large portion of the population is dealing with obesity it is one of the mostly focused areas in Disease-Gene drug development studies.

We have obtained two different gold standard sets for obesity. The first one is from a recent study with 14 previously identified genes (Speliotes, *et al.*, 2010), and the second set is imported from Bell, *et al.* (2005) with a set of 13 novel genes.

3.4.1.2 Schizophrenia

"Schizophrenia is a major complex and debilitating psychiatric disorder with a life time prevalence of ~1% in world" (Irving I. Gottesman, *et al.*, 1991)." The disease originates from a complex combination of genetic effect and environmental factors, which has been strongly supported by families, twin and adoption studies" (Ross, *et al.*, 2006).

In a research article by Sun, *et al.* (2009) a set of 38 Schizophrenia genes are presented. We have used this set as gold standard for Schizophrenia disease. These genes are also accessible via SZGR, (<http://bioinfo.mc.vanderbilt.edu/SZGR/>).

3.4.1.3 Alzheimer 6g

“ Alzheimer disease is a neurodegenerative disorder that is characterized by progressive cognitive deterioration and associated decline in activities of daily living” (Razani, *et al.*, 2010).

A set of 6 genes was gathered from Petteri Sevon , *et al.* (2006) as a gold standard and has been used in this study.

3.4.1.4 Type 2 diabetes

The increasing availability of energy-dense food and the sedentary lifestyle that is becoming prevalent in both first world and developing nations have led to a worldwide epidemic in type 2 diabetes. Mellitus diabetes currently afflicts more than 220 million people and this will increase to more than 400 million by 2030 (Tiganis, 2010) .

A set of 9 genes was proven to be involved in Type 2 Diabetes in a research article by Petteri Sevon, *et al.* (2006) was used in this study as a gold standard gene for Type 2 Diabetes disease.

Table 5

Table containing the Gold standard gene sets for several diseases, full list of gold standard genes is presented in the appendix (section A- II).

Alzheimer		T2D		Obesity_13g		Obesity_14g		Hypertension	
Name	Entrez ID	Name	Entrez ID	Name	Entrez ID	Name	Entrez ID	Name	Entrez ID
APP	351	PPARG	5468	ADIPOQ	9370	FTO	79068	ALOX15	246
PSEN1	5663	GYS1	2997	ADRA2A	150	TMEM18	129787	ALOX12	239
AD5	8081	IRS1	3667	ADRA2B	151	MC4R	4160	ALOX5	240
COL25A1	84570	INS	3630	ADRB1	153	GNPDA2	132789	ALOX5AP	241
APOE	348	KCNJ11	3767	ADRB2	154	BDNF	627	LTA4H	4048
PSEN2	5664	ABCC8	6833	ADRB3	155	NEGR1	257194	LTB4R	1241
		SLC2A1	6513	LEP	3952	SH2B1	25970	MPO	4353
		PPARGC1A	10891	LEPR	3953	ETV5	2119	PLA2G2A	5320
		CAPN10	11132	NR3C1	2908	MTCH2	23788	PTGER1	5731
				PPARG	5468	KCTD15	79047	PTGER2	5732
				UCP1	7350	SEC16B	89866	PTGER3	5733
				UCP2	7351	TFAP2B	7021	PTGER4	5734
				UCP3	7352	FAIM2	23017	PTGIR	5739
						NRXN3	9369	PTGS2	5743
								TBXA2R	6915

3.4.2 Based on scientific research at AstraZeneca

Apart from gold standard genes derived from literature, several gold standard gene sets were also imported from scientific research done at AstraZeneca and other pharmaceutical companies. The diseases and their related gold standard genes are briefly described in the following section. The full sets are available in the Appendix (Section A- II).

3.4.2.1 Dyslipidemia, Hypertension 15g

A set of 91 genes was imported as gold standard genes from Johansson, (Personal Communication) for Dyslipidemia disease and a set of 15 genes was chosen from domestic research at AstraZeneca, Mölndal for Hypertension disease (Johansson, Personal Communication).

3.4.3 Based on launched drug phases

The development of a drug involves several years of research and investigations. If a particular drug has been used to treat a particular disease and also this drug affects the operation of a set of genes, we can assume that these genes are related to the disease. Based on this assumption we could also build gold standard gene sets for each disease. We used the relation between genes and drugs, available in Launched Phase, Phase III, Phase IV, Pre-Registered Phase and Registered phases of drug target database at AstraZeneca to produce gold standard gene sets for different diseases. The full set is presented in the Appendix (Section A- II) in Table 13.

3.4.3.1 Alzheimer 63g, Asthma, Hypertension 55g, Multiple Sclerosis, Myocardial Infection, Pain disorder and Rheumatoid Arthritis

Based on the launched drug phases gold standard sets were produced and 63 genes were selected for Alzheimer's Disease, 64 genes were related to Asthma, 55 genes related to Hypertension, 32 genes involved in Multiple Sclerosis, 38 genes related to Myocardial Infection, 52 genes involved in Pain disorder and 42 genes were selected as gold standard genes involved in Rheumatoid Arthritis.

3.5 Enrichment plot

We have introduced gold standard gene sets to analyze the performance of different methods. We expect these genes to end up at a high position in the final ranked list. A simple way to compare the performance between different methods would be to create a table (for each disease) where the rank of each gold standard gene is shown for each method. In Table 6 we can see the performance of different methods for Dyslipidemia genes in the top 20 list. The table contains 14 genes where 13 are in the top 20 of at least one method. If we consider all the gold standard genes for Dyslipidemia (91 genes; Table 4) this table will be messy.

Therefore we took another approach for visualizing the data. Enrichment plots (Figure 3-7) are plots mostly used in chemistry studies to evaluate results. We used the same plots to analyze our results. The X axis on the plot represents the top x ranked genes and the Y axis the percentage of gold standard genes found among these genes. To describe the plot consider the case that you have a final ranked list containing n elements, you also have a set of m gold standard genes, in a manner that $n > m$ while (*elements in $m \in$ elements in the ranked list*). The best case of finding the m elements would be to find them in the first m position of the ranked list. This means that when we

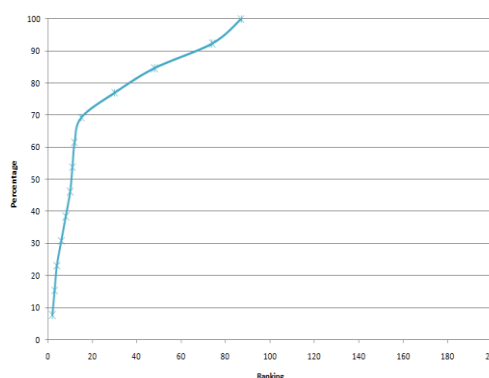


Figure 3-7

An example of an Enrichment plot

Table 6

The table shows the rank position of Gold standard genes found by different methods for Dyslipidemia disease. The table only contains genes that are present in the top 20 ranks of every method.

Genes	GAD	HuGE	MTRA	Prgrin	MDRS
CETP	1	3	16	9	2
APOA5	2	5	-	20	4
APOE	4	1	7	3	3
PPARA	5	8	20	1	5
EC_3-1-1-3	6	4	5	-	6
APOC3	7	7	17	-	7
PPARD	8	16	-	-	15
ABCA1	10	6	15	15	8
APOA1	12	-	10	5	13
PPARG	15	13	-	7	11
PLTP	-	18	-	-	-
EC_2-3-1-43	-	14	14	-	19
EC_1-14-13-17	-	-	-	-	-
APOH	-	-	2	-	17

reach the m position of the X axis of the plot we have found all the data we were looking for or 100% of the data are found (Y axis).

In (Figure 3-7) an enrichment plot is illustrated. The trend in the plot shows the percentages of the gold standard genes found by the given rank. For example we can see that the ranked list could find around 70% of the gold standard genes in the top 20 ranked genes. The trend increases rapidly between $x = 0$ and $x = 20$ meaning that many of the gold standard genes are among the top 20 genes in the result. For $x = 90$ we have found all of our gold standard genes.

3.5.1 Truncate enrichment plot

Now consider the case where we have 30 elements which we would like to find in our sorted list, but the plot is only showing 20 first ranks (x axis of the plot is from 1 to x and $20 > x$). The best case (100% percent on the Y axis), would be finding any 20 of the 30 elements of the gold standard genes by rank 20, meaning we have found 100% of the elements we could find by rank 20.

Take note that in conditions where gold standard genes are not in the ranked list the plot will never reach 100%.

3.6 Clustering

Clustering is done to analyze the correlation between different data sources. We cluster the methods in a way that methods with similar ranked lists are clustered next to each other. We have used two different distances: Spearman's correlation coefficient and percentage of data similarities to produce the clusters.

3.6.1 Spearman's rank correlation coefficient

Spearman's rank correlation coefficient, named after Charles Spearman, is a non-parameter statistical method measuring the dependence between two variables. The method shows us how closely two parameters are related to each other using a monotonic function (Gaylor, *et al.*, 2004). The method is mostly used for comparing the order of ranked lists. Spearman's method illustrates how well the relations between variables have been preserved. The method considers two ranked lists and uses (Equation 6) to find their correlations.

Equation 6

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the rank of an element in the first list and the second list and n is the number of genes shared between the two lists.

Spearman's method doesn't consider elements that are not present in both of the ranked lists. Since $-1 \leq \rho \leq 1$ we used the following formula to create the distance matrix:

$$D_{i,j} = 1 + \rho$$

where i and j corresponds to the pair of ranked lists being investigated and $0 \leq distance_{i,j} \leq 2$. The smaller the distance the more similar the ranking will be. Software package R (R Development Core Team, 2005), was used to create the hierarchical clustering of the given distance matrix.

3.6.2 Percentages of data similarities

Since we are focusing on ranked lists that are provided by different sources there is a high chance of having missing data between them. While Spearman's method focuses on the order of correlated data in a pair of ranked lists, percentage of data similarity method focuses on the similarity between the pairs. The method calculates the percentage of genes present among the top M candidates regardless of the order. The distance matrix is calculated using Equation 7 between every pair of data sources.

Equation 7

$$D_{i,j} = 1 - \frac{m}{n}$$

where m is the number of corresponding genes in both of the ranked lists and n is the number of unique elements in both sets, and i and j correspond to the pair of ranked lists that are being investigated and $0 \leq distance_{i,j} \leq 1$. The smaller the distance the more similar the data that has been found by both of the data sources.

A distance matrix is created by finding the similarities between the data sources. Hierarchical clustering is done using the software package R (R Development Core Team, 2005).

3.7 Analysis of sets of genes or diseases

Consider the case where a set of genes is available to the user, and the user wants to know which diseases are most relevant to this set. We tried to design a method which will rank diseases based on a set of genes. This problem was first encountered when we tried to establish a relation between pathways and diseases (Section 3.7). We will also use the same approach to examine how well the gold standard gene sets correspond to their diseases (i.e. testing their integrity). In the upcoming sections we will describe the method.

3.7.1 Method description

We designed a method to analyze a set of genes or diseases to find the most relevant diseases. The method consists of two steps:

- Step 1: For each gene in the set, the top 10 ranked diseases are investigated using the MDRS method. A matrix is then formed. Every row of the matrix contains the top 10 diseases of every gene with their S_{wdr} scores. To simplify the matrix we have only presented the best top three ranks for each gene in a small set in (Table 7) without their score.

Table 7

Gene set Matrix is created using the MDRS method.

The top 10 hits of each gene in the set is investigated using the MDRS method and are imported into a matrix with their scores. The matrix above shows the best three candidates just for simplifying. The matrix could be created for disease to gene relations with the same procedure.

Gene	Rank 1	Rank 2	Rank 3
Phosphoinositide-3-kinase, catalytic, gamma polypeptide	Obesity, Morbid	Insulin Resistance	Autistic disorder
Phosphoinositide-3-kinase, catalytic, delta polypeptide	Edema	Drug toxicity	Cardiovascular system disease
Protein kinase C, eta	Brain Infarction	Cancer	Rheumatoid arthritis
Mitogen-activated protein kinase 3	Asthma	Cancer	Neoplasm's, Glandular and Epithelial
Sphingomyelin phosphodiesterase 1, acid lysosomal	Coronary heart disease	Brain disease	Cardiovascular system disease
Solute carrier family 27 (fatty acid transporter), member 2	Alzheimer's disease	Brain disease	Nervous system disease
Solute carrier family 27 (fatty acid transporter), member 5	Non insulin dependent diabetes mellitus	Cancer	Brain disease
ATP-binding cassette, sub-family C (CFTR/MRP), member 8	Non insulin dependent diabetes mellitus	Glucose intolerance	Complications of Diabetes Mellitus

- Step 2: Based on this matrix and the different scores described below most related disease to the set is chosen. The best candidate could be chosen using any of these scores or by combining them together.

3.7.1.1 Observed score

The observed Score O_i , where i is the elements of the second column of the matrix, (named rank 1 in Table 7) is the number of times a certain disease has been observed as best candidate for the genes in the set. The disease with the highest score would simply be the best candidates for the list. For example, in the Table 7 “Non-insulin dependent diabetes mellitus”, is the best candidate.

3.7.1.2 Total observed score

Total observed Score TO_i is the number of times elements in the second column of the matrix have been repeated in the whole matrix. Total observed score is basically the number of repeats of disease in column one in the whole matrix. The Disease with the highest score is the best candidate chosen by total observed score.

3.7.1.3 R' score

R' Score is the average of S_{wdr} scores of repeated elements in second column of the matrix meaning that the S_{wdr} scores are added together for every repetition of a disease in column two of the matrix and then it is divided by the number of observation of that disease:

Equation 8

$$R'_i = \frac{\sum_{i=1}^n S_{wdr_i}}{O_i}$$

where n is the number of elements in the investigated set and S_{wdr_i} is the MDRS score of element i in the second column of the matrix. An element with the highest R' Score would be the most related

element to the gene set. To describe this consider Table 7, the R' score of “Non-insulin dependent diabetes mellitus” is the sum of S_{wdr_7} and S_{wdr_8} divided by 2. This is because this disease has been observed twice as the first rank for the set of genes.

3.7.1.4 S' score

For every element in the second column of the matrix (named Rank 1), there is a S_{wdr} score. S' Score basically adds the S_{wdr} scores of identical elements in the second column and divides it by the number of genes in the data set. The score is then assigned to all of the identical elements. To describe this consider Table 7, the S' score of “Non-insulin dependent diabetes mellitus” is the sum of S_{wdr_7} and S_{wdr_8} divided by 8.

Equation 9

$$S'_i = \frac{\sum_{i=1}^n S_{wdr_i}}{n}$$

where n is the number of elements in the gene set and S_{wdr_i} is the MDRS score of element i in the second column of the matrix. An element with the highest S' Score would be the best candidate for the set.

3.7.1.5 Z score

There is a strong bias in the number of gene-disease relations from the sources. We have seen that about 25% of the gene-disease relations are to cancer. This means that for any random gene list it is likely that cancer will show up in the top ranked diseases. One simple way to incorporate this information is to estimate the probability of sampling a disease at random by its frequency, i.e. $\hat{p} = f_D/N$, where f_D is the number of relations between any gene to the disease D . If observing O_i diseases in a list of n diseases, we would expect to find $E = np$ disease by chance. The Z score is a normal approximation (a rough approximation) that will measure how unlikely it is to make this observation by chance (Equation 10).

Equation 10

$$Z_i = \frac{O_i - E_i}{\sqrt{E_i(1 - E_i)}}$$

3.8 Implementation

All the implementation of the MDRS method was done using the Perl programming language on a PC with an Intel dual core processor. The script has the ability of integrating multiple ranked lists where different weights are given as input for the different data sources. The number of data sources may vary. The numbers of bins is by default set to 5 but can be changed with a parameter. It is also possible to change the number of top ranked elements to give them a higher priority; this is by default set to five. The main inputs of the program are ranked lists (text or Excel).

As well as ranked list, gold standard genes could also be uploaded to the script so that analysis based on the gold standard set could be available as output. The script is written in a way that different types of analysis can be printed. Output of the script could be:

- 1- Top n ranked elements of MDRS as well as other methods.
- 2- The result of the MDRS method containing the MDRS score and ranking together with previous ranks in individual sources.
- 3- The position of gold standard genes in the MDRS output or any of the ranked lists. This can later be used to create enrichment plots.
- 4- Calculate the Order statistic methods (used in Endeavour) and produce ranked lists from it.

Mapping of different data sources and calculating distance matrices based on Spearman rank correlation and percentage of data similarities were also done using additional Perl scripts. Enrichment plots were created using Microsoft Excel and clustering was done using the hclust package in the R programming language.

4 Results

When integrated different data sources we first analyzed the correlation between them by Venn diagrams and clustering. This provides us with useful details about the data and the overlap of different sources. We used the MDRS method to integrate the ranked data and analyzed the performance using enrichment plots. We finally designed a method to analysis sets of genes to find relevant diseases to the set. This was done to produce disease-pathway relations.

4.1 Weighing the sources

How much confidence do we put in the data? Upon the availability of various sources it is always a challenge to find good weights. In this work we assume that the weights will be added by the scientist in the field who has knowledge about the sources. Furthermore it is important to know the correlation between ranked lists produced by different methods. This would increase the knowledge of the user on data sources and hence give him more information so that better decisions would be taken on weighing the data sources. We applied different weights to each of the data sources to produces the integrated ranked list. The final weights agreed with expectations by the experts. In summary, we assigned the weights of the gene-disease sources as follows: $\mu_{GAD} = 3$, $\mu_{HuGE} = 2$, $\mu_{MTRA} = 1$, $\mu_{Peregrine} = 1$. That is, we are most confident GAD, then in HuGE and least confident in MTRA and Peregrine. The other sources where not used in MDRS but are present in the plot for comparisons of performance.

4.2 Correlations of different data sources

The correlations between different data sources may provide the user with valuable data on how to weigh them and how similar rankings the data may provide. Here we will present two types of analyzes we have taken to estimate the correlation. We will first discuss Venn diagram and then clustering with either a simple similarity distance or a rank-correlation measure.

4.2.1 Venn diagram

We analyze the intersection, overlap and the distribution of data between ranked lists using Venn diagrams. We created these for different gold standard gene sets (Section 3.2) of several diseases.

Figure 4-1 shows a Venn diagram based on the top 50 ranks in the gold standard genes of Pain disorder (which contains 52 genes). The data is based on launched drug phases. Among those genes HuGE finds 7 of the gold standard genes while GAD only finds 3. They find two in common but the remaining six are not present in both. By combining HuGE and GAD it should be possible to find all eight genes, which is not possible when considering them separately. This shows us that there is additional information and that MDRS may improve our results.

On the other hand, NovoSeek only finds a single gene that is picked up by both GAD and HuGE. In this case, it is not evident that if adding NovoSeek we will enrich the analysis.

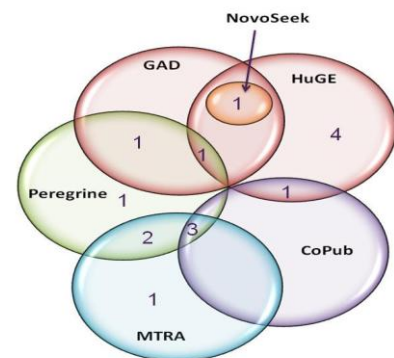


Figure 4-1

Venn diagram produced on Pain disorder disease based on the top 50 ranks. The figure illustrates, overlaps of different methods on gold standard genes produced in the top 50 positions of the ranked lists.

4.2.2 Clustering

Another approach to gain knowledge about the correlation between different data sources is to cluster the elements in their ranked lists based on a measure of similarity. Similar ranked lists will have a small distance and dissimilar ones a large distance. The distance matrices were formed using either Spearman's rank correlation coefficient or the percentage data similarity distance.

4.2.2.1 Spearman's rank correlation coefficient method

In Table 8 and Figure 4-2 Spearman's correlation distance matrix and corresponding hierarchical clustering of the top 50 ranked genes in Alzheimer's disease are presented.

The dendrogram shows that HuGE and GAD give similar order among their top 50 genes. Therefore since we have the knowledge that GAD and HuGE are both trustable sources they should probably be weighed similarly. We can also see that MTRA, MOAI and Peregrine give similar results and that these are different from GAD and HuGE. Also, NovoSeek and GeneRanker seem to be identical so maybe only one of them should be included. CoPub provides lists that do not have any similarities with others. This may either be good (it finds new true genes) or it may be bad (the results are not reliable). The distances are shown in Table 8 where we can see that CoPub has a distance of two to all other sources, i.e. it produces completely different ranked list which does not have any correlation with the others. All other clusters are available in the Appendix (section A- IV).

Alzheimer - Top 50 Spearman's Rank Correlation

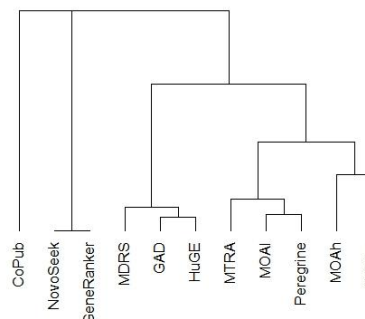


Figure 4-2

Clustering of different data sources for Alzheimer's disease.

	GAD	HuGE	MTRA	MOAh	MOAm	MOAI	Peregrine	MDRS	NovoSeek	GeneRanker
HuGE	0.116									
MTRA	0.652	0.947								
MOAh	0.2	0.6	0.4							
MOAm	0.2	0.971	0.35	0.503						
MOAI	0.714	1.333	0.281	0.436	0.564					
Peregrine	0.631	0.683	0.149	0.8	0.476	0.1428				
MDRS	0.178	0.212	0.455	0.502	0.654	0.557	0.375			
NOVO/Seek	0.482	0.589	0.2	0.4	0.809	0.433	0.071	0.328		
GeneRanker	1.8	0.5	2	2	2	2	2	2	0	
CoPub	2	2	2	2	2	2	2	2	2	2

Table 8

Distance Matrix for Alzheimer's disease created using Spearman's method.

One of the biggest drawbacks of Spearman's method is that it only considers the order of the genes that are present in both ranked lists and does not deal with missing data. In cases where same set of genes are present in both lists and in the same order but in different rank positions, the method considers the lists as completely correlated.

4.2.2.2 Percentages of data similarities

A naïve way to incorporate missing data is to study the percentage of data similarities in the top elements (for example 50) of the ranking (Section 3.6.2). This method clusters data sources next to each other if their ranked lists contain similar elements. Since it is based on the content of the elements their order will not contribute to the distance.

Table 9

Percentage of data similarity distance matrix for Alzheimer's disease based on the top 50 ranks of the different data sources.

	GAD	HuGE	MTRA	MOAh	MOAm	MOAI	Peregrine	MDRS	Novo Seek	GeneRanker
HuGE	0.1									
MTRA	0.68	0.68								
MOAh	0.92	0.92	0.9							
MOAm	0.9	0.88	0.82	0.62						
MOAI	0.86	0.84	0.78	0.6	0.28					
Peregrine	0.64	0.64	0.66	0.92	0.84	0.86				
MDRS	0.3	0.28	0.68	0.66	0.72	0.7	0.64			
NOVO Seek	0.66	0.62	0.78	0.88	0.84	0.82	0.84	0.66		
GeneRanker	0.92	0.94	0.98	1	1	0.98	1	0.98	0.94	
CoPub	1	1	1	1	1	1	1	1	1	1

Alzheimer- Top 50 Percentage Similarites Genes

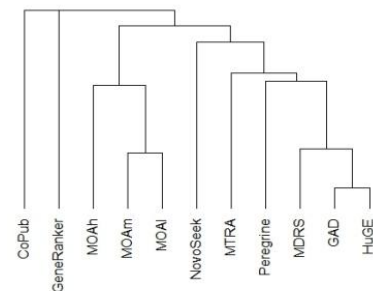


Figure 4-3

The dendrogram of Alzheimer's disease based on the top 50 ranks of the different data sources using the percentage similarity method.

In Table 9 and Figure 4-3 a distance matrix and a dendrogram of Alzheimer's disease are presented based on percentage similarities. Similar conclusions can be made here. Note, however, that GeneRanker and Novo|Seek are no longer identical. This was not captured by the rank correlation distance. The reason is that we removed all the missing elements which here contribute to differences.

4.3 Performance of the MDRS method

The MDRS method was applied on ranked lists created by GAD, HUGE, MTRA and Peregrine, with the weights: $\mu_{GAD} = 3, \mu_{HUGE} = 2, \mu_{MTRA} = 1, \mu_{Peregrine} = 1$. The performance of the methods was evaluated using enrichment plots which are described in Section 3.5.

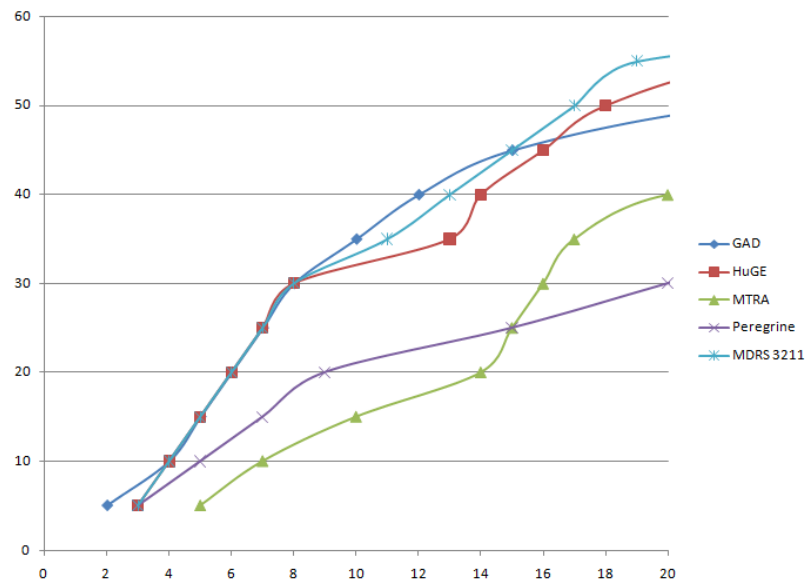


Figure 4-4

The performance of MDRS and other sources for Dyslipidemia disease. Total number of unique genes in all ranked lists equals 1776 genes.

In Figure 4-4 an enrichment plot of Dyslipidemia disease is presented. The plot shows four trends each representing different ranked lists produced by different methods. There are 91 genes present in the Dyslipidemia set (Table 4). The plot only shows the top 20 ranks (see the **X** axis) which means that it is only possible to find 20 of these 91 genes. If all genes are found we will reach a peak of 100% by rank 20 – the trend would be a straight line with slope five ($y = 5x$).

We can observe that GAD is the first method that picks up a gold standard gene at position 2. Peregrine, HuGE and MDRS pick up their first gene at rank 3 while MTRA does worse than all other sources with the first gene at rank 5. MTRA will improve its performance between rank 14 and 18 when the trend dramatically increases and bypasses Peregrine at rank 15.

The MDRS method finds around 55% of the gold standard genes by rank 20, performing better than any other source until this rank. We can observe that by combining the data sources we have improved the ranking. The MDRS trend completely depends on the performance of the other sources. As can be seen in Figure 4-4, from rank 8 to 16 the MDRS does not perform as well as GAD. During these ranks, HuGE retrieves genes that are not presented in the gold standard set, thus although HuGE has low weight, but it still decreases the trend for a short period.

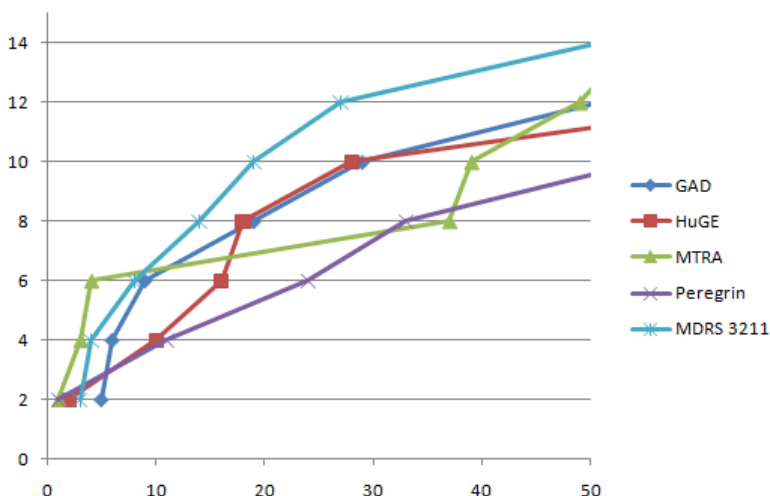


Figure 4-5

Enrichment plot of Alzheimer disease. The gold standard set contains 63 genes chosen from the Launched drug phases. Total number of unique genes in the whole ranked lists equals 3844 genes.

In our second case study (Figure 4-5), Alzheimer's disease was ranked using GAD, HuGE, MTRA and Peregrine and the results integrated using the MDRS method. The MDRS method performs better than all other sources and improves the performance of the ranked lists dramatically when the data is distributed between ranked lists. Figure 4-5, shows us that MDRS is helping us pick up more related genes, at earlier ranks.

In our third case study we looked at Type 2 diabetes disease (Figure 4-6). The trends show that MDRS picks 10% of the gold standard genes, faster than any other method in the early ranks. In this case, using

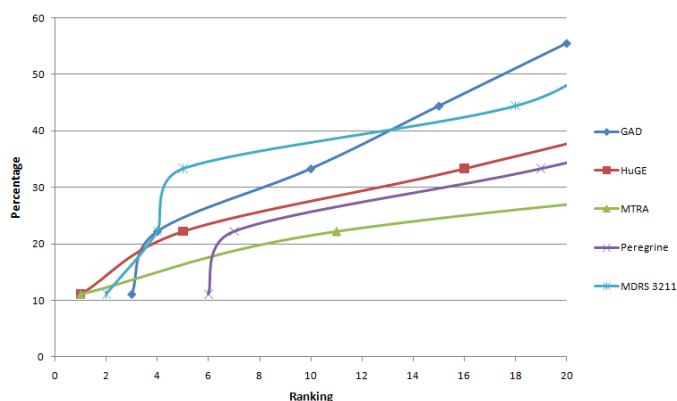


Figure 4-6

Enrichment plot of MDRS method applied on four different data sources to retrieve ranked list on Type 2 diabetes disease. Gold standard gene set contains 9 genes. The number of unique genes was 3189 genes.

GAD alone would have been the best choice. MDRS will be affected by the bad performance of MTRA. In conclusion, if one source is much better than all the others, MDRS will most likely not be as good as that source. Other examples motivating the use of the MDRS method are available in the Appendix (Section A- III).

4.4 Comparison between different sources

Enrichment plot gives you the possibility of comparing different data sources. One of our interests was to investigate the performance of different public literature sources. We ran queries on Novo|Seek, GeneRanker and CoPub on all of the gold standard diseases (Table 4). The retrieved ranked lists were then mapped to our in-house vocabulary using Entrez ID. We then produced enrichment plots based on different gold standard gene sets.

4.4.1 Comparison of literature based sources

Based on a large set of enrichment plots (available in the Appendix, Section A- III) of gold standard genes from literature, we observe that in most of the cases GAD and HuGE outperform the other sources. Further, GAD performs better than HuGE, while Peregrine and MTRA pick up a lot of noise. However, by looking at the plots we realize that, MTRA and Peregrine have a better chance of picking up the whole gold standard sets at the end. This is simply because these methods contain a lot more genes.

This conclusion could change in a few years. The performances of these methods are completely dependent on the amount of research done in the field. In literature based methods a gene would gain higher rank, if more research is applied on its relation with the disease. Figure 4-7, shows analyses done on Obesity disease using Obesity 13g gene set (Table 4).

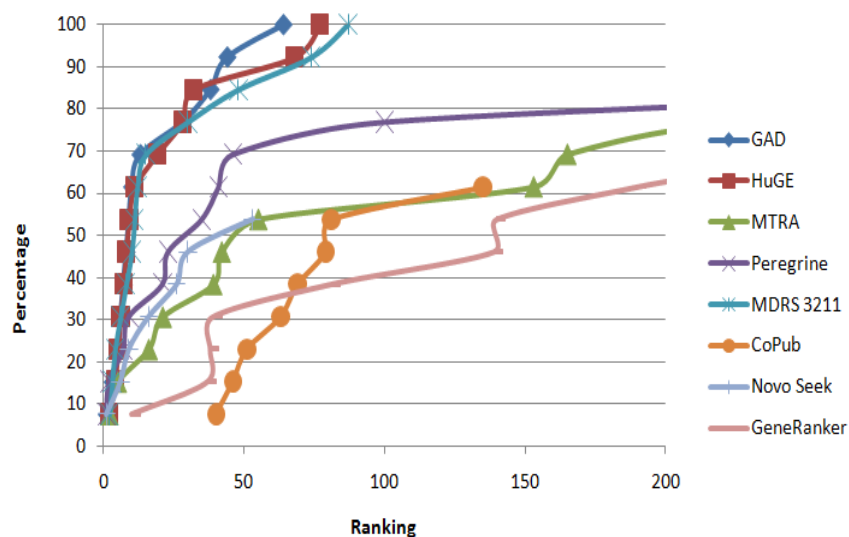


Figure 4-7

Obesity Disease, Gold standard contained 13 genes picked up from (Bell, *et al.*, 2005). The whole data set contains 3224 unique genes. GAD, HuGE and the MDRS method are performing dramatically better than other sources.

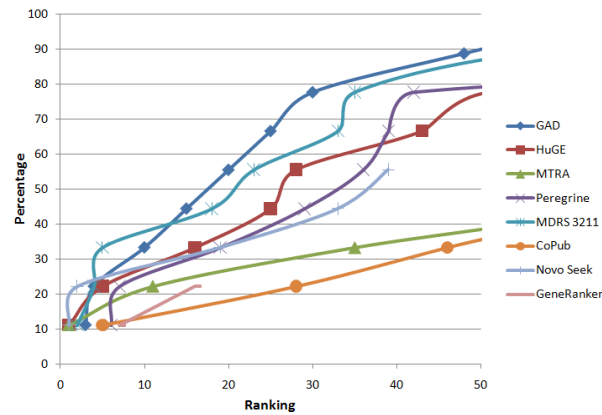


Figure 4-8

Enrichment plot of Obesity created using 7 different data sources. GAD data source provides better ranked list than other methods this is followed by MDRS and HuGE.

On the other hand when analyses were done on gold standard sets chosen from launched drug phases (MoAh, Section 3.4.3), methods such as MTRA and Peregrine performed better than GAD and HUGES (Figure 4-9).

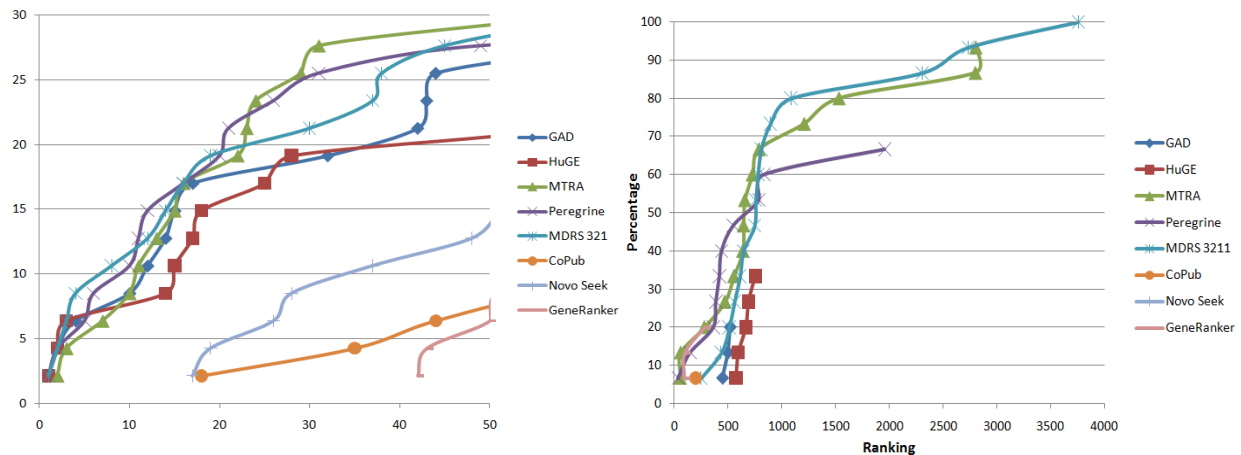


Figure 4-9

Enrichment plots for Hypertension (55 genes, left plot) with 3181 unique genes and for Asthma (64 genes, right plot). The number of unique genes in all data sources is 2533 genes. MTRA and Peregrine performs better on these data than GAD and HuGE.

4.5 Analysis of sets of genes

Given a set of genes (or diseases), it is sometimes desirable to output the most relevant diseases (or genes). In this section we present results from gold standard gene sets and pathways.

4.5.1 Analysis of gold standard sets

In the Methods (Section 3.7) we describe three measures – O score, Z' score and S' score – that are applicable to sets of genes (or diseases). Here we evaluate these using our gold-standard gene sets. We expect the disease of the gold standard to appear highest in the ranking. All results are shown in the Appendix (Section A- II).

As an example we show the output from Asthma. Table 10-(a) shows the observed number of diseases that are found in the highest ranked position of the MDRS results of all the genes in the set (O score). We can see that Asthma is the primary disease found and then comes inflammation. Table 10-(b) shows the Z -score that is normalized with respect to the underlying distribution of gene-disease terms (See Methods for explanation, Section 3.7.1.5). Also here Asthma is the best candidate but on second place we get “Ossification of posterior longitudinal ligament” which is probably a false positive. Table 10-(c) shows the S' score which is based on the MDRS score (see Section 3.7.1.4). Also here Asthma appears at the top followed by Inflammation (same as for the observed score).

In summary, all of the scores presented in the table had the ability to identify Asthma as their number one choice. In general, however, looking at all results (Appendix, Section A- VI), the S' score had the best performance. Further, we believe the Z score should be used in combination with the S' score to assign confidence to it (how likely is this observation). We should hence rank on S' not Z .

Table 10

Analysis of sets of Asthma disease genes. The MDRS method has been applied and the tables show the top 5 diseases obtained by the three resulting scores (see Methods). Asthma got the best score in all the methods which is the expected candidate.

10-(a)

	Disease	Observed
1	Asthma	11
2	Inflammation	6
3	Hypertension	4
4	Malignant neoplasm of breast	3
5	Obesity	2

10-(b)

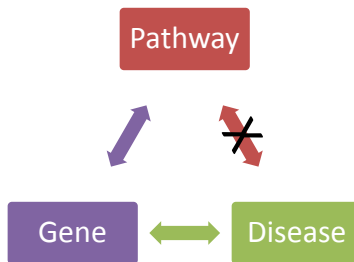
	Disease	Z Score
1	Asthma	1.43608206314716
2	Ossification of Posterior Longitudinal Ligament	1.13986504696089
3	Hyperuricemia	-0.50398579862325
4	Habitual abortion	-0.63302029756289
5	Celiac disease	-1.48223813484271

10-(c)

	Disease	S'
1	Asthma	2.30954735418977
2	Inflammation	1.67783770268185
3	Malignant neoplasm of breast	1.33981498376288
4	Hypertension	1.32495549898104
5	Obesity	1.27679666795499

4.5.2 Disease-Pathway relation

One aim of the project was to predict pathways-diseases relations as we do not have a reliable source for this relation. Several databases were inspected but none had this information. We therefore tried to infer these indirectly by using the set scores and the gene-disease relations we already described.



Relations between gene and pathways are present in IPA (IPA, 1998) where each pathway has a list of genes. We looked at the sets of genes in the pathways and used our method (Section 3.7) to predict the disease. Our (naïve) assumption was that if a disease is related to the set of genes in the pathway, then this pathway might play a role in the disease. We picked some of the pathways that were annotated with a disease and analyzed the output of the scores for the gene sets. All results are shown in the Appendix (Section 5.3A- VII).

Table 11 shows the results of the method based on two different pathways: Leptin signaling in obesity pathway and Parkinson's signaling pathway. The table shows the top five diseases related to the gene set based on the S' scores (Section 3.7) and we can see that in both these cases the correct disease term is found. These results, however, are not representative for all pathways we studied. Often the correct disease cannot be inferred using this approach. One problem may be that the genes are not specific to the pathway but involved in several pathways and functions.

We found that diseases such as cancer are heavily overrepresented among the gene-disease relations. Approximately 25% of all relations from a given gene are to cancer. The Z score will correct for this but we never rank on the Z score but only use it in combination with the S' score. In general, it is not easy to use this approach to go from gene to pathway. More advanced models might be necessary but these are out of scope for the current project.

Table 11

The table shows analyses done on two sets of genes related to two different pathways. The first annotated with Obesity and the second with Parkinson's disease.

Leptin Signaling in Obesity		
Score	S'	Z
Obesity	1.41	1.78
Cancer	1.03	0.35
Non-insulin-dependent diabetes mellitus	0.89	1.25
Insulin Resistance	0.65	-2.9
Cardiovascular system disease	0.47	-1.5

Parkinson's Signaling		
Score	S'	Z
Parkinson disease	2.88	3.88
Parkinsonian disorder	1.12	4.06
Cardiovascular system disease	0.53	-0.9
Malignant neoplasm of breast	0.40	-1.5
Inflammation	0.30	-0.3

4.6 Discussion

The MDRS method performance can be influenced by many aspects but overall the following statements can be drawn from its performance:

- 1- The MDRS method performs at its best when we have disjoint ranking of true elements between the set of ranked lists when the remaining elements have little or no correlations between them. The MDRS method also performs well when the ranked lists share most of their true elements while the remaining elements have low or no correlation. In this way ranked list are highly correlated assuming that the data sources are independent.
- 2- The MDRS method performs equally if data sources are highly correlated (e.g. identical copies of the same ranked list).
- 3- The MDRS method performs worse when one ranking list is correct and the other is exactly the opposite (reversed order).

We have also observed that in most cases genetic data sources such as GAD or HuGE perform better than any other type of investigated sources.

5 Conclusion and future work

Identifying gene-disease relation is a problem of primary importance in biomedical research. Biologists usually take two steps to solve this problem. They first look for candidate genes through different processes such as high throughput genomic techniques and in the second step they evaluate their result using wet lab techniques (De Bie, *et al.*, 2007). This process is time consuming and costly. To speed up the process candidate genes could be prioritized using different methods. Several databases exist that provides gene-disease prioritization through ranked lists. Each of these ranked lists comes with different reliabilities. Several methods have been proposed that can integrate these ranked lists. Among these is the Discounted Rating System.

During this project, we modified the Discounted Rating System (Method Section). The MDRS method integrates multiple ranked lists with different reliabilities into a final ranked list. The reliability of different data sources were chosen through expert knowledge. The MDRS method can be applied to any type of ranked lists regardless of the scoring function and the list size.

The method was applied on gene-disease relations. To evaluate the results gold standard gene sets were used and output was analyzed using enrichment plots. By the use of enrichment plots the performance of different methods and data sources was also observed.

The correlation of different data sources and methods were analyzed using Venn diagrams and hierarchical clustering. Distance matrices were created using Spearman's rank correlation method and percentage of data similarities method. Finally a method was introduced that would help analysis of a set of genes to find the most relevant diseases to the set.

5.1 Discussion

Among the four methods described, the Order statistics method used in Endeavour (Section 2.2.2) and the DRS method (Section 2.2.4), were chosen to integrate ranked lists. We modified the DRS method and used the MDRS method to integrate our data sources considering their reliabilities.

5.1.1 Method comparison

Kernel based data fusion method (Section 2.2.1) and optimal weight matrix (Section 2.2.3), need training sets with known gene-disease relations (Gold Standard sets). Since our assumption is that core genes are not available for all of the disease, we did not choose these methods. Optimal weight matrix method weighs different data sources randomly. The method chooses the best weight regardless of the reliability on the data sources.

Modified DRS has the ability to integrate ranked lists with missing data. This means that ranked lists should not be the same size while the order statistics method considers all ranked lists equally sized. The MDRS method can also weigh different data sources based on their reliabilities. The MDRS method can deal with data sources with multiple reliabilities while order statistics method assumes that data sources are equally reliable. The MDRS method performed better than the order statistics method even when the sources are weighed equally (Appendix Section 5.3A- V).

To integrate different data sources using the MDRS method we found out that giving the first five elements of each ranked list better priority provides us best results. This decision could differ on other data sources. The performance of the MDRS method is influenced by the quality of each data source and the correlation between data sources. If a ranked list brings a huge amount of noise, it is not reasonable to integrate the rank list with trusted ranked lists. The MDRS method performs best when databases are partially correlated.

5.1.2 Data sources

Using enrichment plots comparison of different literature based data sources was implemented (Appendix A- III). In most cases, GAD and HuGE performed better than other sources when using literature based gold standard genes (Section 3.4.1), while Peregrine (Section 2.1.4) and MTRA (Section 2.1.3) picked up noise. By looking at the plots we realize that MTRA and Peregrine would have a better chance of picking up the whole gold standard genes at the end. Performance of other method varied from case to case while mostly Peregrine and MTRA performed similarly. We can observe that these two data sources were also clustered near each other since they were correlated (Appendix 5.3A-IV). Since the performances of these methods are completely dependent on the amount of research done in the field, these conclusion could change in a few years.

5.2 Future work

By the use of enrichment plots we can now compare a variety of data sources. The focus of this project was on literature based studies. While gathering data from literature based studies we should consider that newly found relations are not well discussed in different literature and therefore they wouldn't get a good ranking position. We would like to continue by looking at other gene-disease relation methods which will focus on other aspects. Meanwhile more work could be done on establishing disease-pathway relationship. By combining different scores described in section 3.7.1 we have reduced the noise, but more could be done in this field.

5.3 Conclusion

We have presented a new approach to integration ranked lists with different reliabilities. To our understanding, the MDRS method is shown to outperform current methods that are used to integrate data sources. The success of the method is because of its ability to integrate different sized data sources while considering their reliabilities. Weighting the data sources is done using expert knowledge. The DRS method can be applied to any type of ranked lists regardless of the scoring function of the ranked lists. The method was validated using gold standard gene sets and enrichment plots.

References

- [1] Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. and Pickard, B.S. (2005) Speeding disease gene discovery by sequence based candidate prioritization, *BMC Bioinformatics*, **6**, 55.
- [2] Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. and Pickard, B.S. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates, *Bioinformatics*, **22**, 773-774.
- [3] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B., *et al.* (2006) Gene prioritization through genomic data fusion, *Nat Biotechnol*, **24**, 537-544.
- [4] Allende, R.A. (2009) Accelerating searches of research grants and scientific literature with novo|seek, *Nature Methods*, **6**.
- [5] Becker, K.G., Barnes, K.C., Bright, T.J. and Wang, S.A. (2004) The genetic association database, *Nat Genet*, **36**, 431-432.
- [6] Bell, C.G., Walley, A.J. and Froguel, P. (2005) The genetics of human obesity, *Nat Rev Genet*, **6**, 221-234.
- [7] De Bie, T., Tranchevent, L.C., van Oeffelen, L.M. and Moreau, Y. (2007) Kernel-based data fusion for gene prioritization, *Bioinformatics*, **23**, i125-132.
- [8] Fox, E.A. and Shaw, J.A. (1994) Combination of Multiple Searches *The Second Text REtrieval Conference*.
- [9] Freudenberg, J. and Propping, P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes, *Bioinformatics*, **18 Suppl 2**, S110-115.
- [10] Freund, Y., Iyer, R., Schapire, R.E. and Singer, Y. (2003) An Efficient Boosting Algorithm for Combining Preferences, *The Journal of Machine Learning Research*, **4**.
- [11] Friedman, N., Linial, M., Nachman, I. and Pe'e, D. (2000) Using Bayesian Networks to Analyze Expression Data, *J.Comput. Biol.*, **7**, 601-620.
- [12] Frijters, R., Heupers, B., van Beek, P., Bouwhuis, M., van Schaik, R., de Vlieg, J., Polman, J. and Alkema, W. (2008) CoPub: a literature-based keyword enrichment tool for microarray data analysis, *Nucleic Acids Res*, **36**, W406-410.
- [13] Frijters, R., van Vugt, M., Smeets, R., van Schaik, R., de Vlieg, J. and Alkema, W. (2010) Literature mining for the discovery of hidden connections between drugs, genes and diseases, *PLoS Comput Biol*, **6**.
- [14] Gaylor, D.W., Lutz, W.K. and Conolly, R.B. (2004) Statistical analysis of nonmonotonic dose-response relationships: research design and analysis of nasal cell proliferation in rats exposed to formaldehyde, *Toxicol Sci*, **77**, 158-164.
- [15] Gonzalez, G., JC, U., B., A., W., M. and M.E., B. (2008) GeneRanker: An Online System for Predicting Gene-Disease Associations for Translational Research, *Translational Bioinformatics*.
- [16] Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D. and McKusick, V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Res*, **30**, 52-55.
- [17] Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Res*, **33**, D514-517.
- [18] IPA (1998). Ingenuity Systems, Redwood City, California.
- [19] Irving I. Gottesman, Wolfgram, D.L. and Freeman, W.H. (1991) Schizophrenia genesis: The origins of madness, *Behavior Genetics*, **23**, xiii + 296.
- [20] Ji, S.L. and Huang, Q.Y. (2006) [PPARgamma variants and complex diseases], *Yi Chuan*, **28**, 993-1001.

- [21] Johansson, M. (2008) Calculating MeSH-term relevance for genes by combining EntrezGene, Mesh and Medline data, Method description and evaluation report. AstraZeneca Discovery Information, Mölndal, Sweden
- [22] Johansson, M. (Personal Communication). Research and Discovery, AstraZeneca. Mölndal, Sweden.
- [23] Kalervo Järvelin and Kekäläinen, J. (2000) IR evaluation methods for retrieving highly relevant documents, *In SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 41-48. .
- [24] Lee, J.H. (1997 [a]) Combining Multiple Evidence from Different Relevance Feedback Methods, *DATABASE SYSTEMS FOR ADVANCED APPLICATIONS*, pp 421-430.
- [25] Lee, J.H. (1997 [b]) Analyses of multiple evidence combination, *SIGIR '97 Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, **31**.
- [26] Li, Y. and Patra, J.C. (2010) Integration of multiple data sources to prioritize candidate genes using discounted rating system, *BMC Bioinformatics*, **11 Suppl 1**, S20.
- [27] López-Bigas, N. and Ouzounis, C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease, *Nucleic Acids Research*, **32**, Pp. 3108-3114.
- [28] Maes, H.H., Neale, M.C. and Eaves, L.J. (1997) Genetic and environmental factors in relative body weight and human adiposity, *Behav Genet*, **27**, 325-351.
- [29] NIH-Publication (1998) *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults--The Evidence Report*. National Institutes of Health. *Obes Res*.
- [30] Oti, M. and Brunner, H.G. (2007) The modular nature of genetic diseases, *Clin Genet*, **71**, 1-11.
- [31] Perez-Iratxeta C, Bork P and MA, A. (2002) Association of genes to genetically inherited diseases using data mining, *Nature Genetics*, **31**, 316-319
- [32] Petteri Sevon , Lauri Eronen , Petteri Hintsanen , Kimmo Kulovesi and Toivonen, H. (2006) Link Discovery in Graphs Derived from Biological Databases, *HIIT Basic Research Unit, Department of Computer Science, University of Helsinki, Finland*.
- [33] R Development Core Team (2005) A language and environment for statistical computing. In Computing, F.f.S. (ed), *BMC Bioinformatics*. Vienna, Austria.
- [34] Razani, J., Bayan, S., Funes, C., Mahmoud, N., Torrence, N., Wong, J., Alessi, C. and Josephson, K. (2010) Patterns of Deficits in Daily Functioning and Cognitive Performance of Patients With Alzheimer Disease, *J Geriatr Psychiatry Neurol*.
- [35] Ross, C.A., Margolis, R.L., Reading, S.A., Pletnikov, M. and Coyle, J.T. (2006) Neurobiology of schizophrenia, *Neuron*, **52**, 139-153.
- [36] Smith, N.G. and Eyre-Walker, A. (2003) Human disease genes: patterns and predictions, *Gene*, **318**, 169-175.
- [37] Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L., Lindgren, C.M., Luan, J., Magi, R., *et al.* (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index, *Nat Genet*, **42**, 937-948.
- [38] Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules, *Science*, **302**, 249-255.
- [39] Sun, J., Jia, P., Fanous, A.H., Webb, B.T., van den Oord, E.J., Chen, X., Bukszar, J., Kendler, K.S. and Zhao, Z. (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case, *Bioinformatics*, **25**, 2595-6602.
- [40] Tanko, L.B., Siddiq, A., Lecoœur, C., Larsen, P.J., Christiansen, C., Walley, A. and Froguel, P. (2005) ACDC/adiponectin and PPAR-gamma gene polymorphisms: implications for features of obesity, *Obes Res*, **13**, 2113-2121.

- [41] Tiffin, N., Adie, E., Turner, F., Brunner, H.G., van Driel, M.A., Oti, M., Lopez-Bigas, N., Ouzounis, C., Perez-Iratxeta, C., Andrade-Navarro, M.A., *et al.* (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes, *Nucleic Acids Res*, **34**, 3067-3081.
- [42] Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B. and Hide, W.A. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates, *Nucleic Acids Res*, **33**, 1544-1552.
- [43] Tiganis, T. (2010) Reactive oxygen species and insulin resistance: the good, the bad and the ugly, *Trends Pharmacol Sci*.
- [44] Tranchevent, L.C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S. and Moreau, Y. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species, *Nucleic Acids Res*, **36**, W377-384.
- [45] Turner, F.S., Clutterbuck, D.R. and Semple, C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes, *Genome Biol*, **4**, R75.
- [46] Wren, J.D. (2004) Extending the mutual information measure to rank inferred literature relationships, *BMC Bioinformatics*, **5**, 145.
- [47] Yu, S., Tranchevent, L.C., De Moor, B. and Moreau, Y. (2010) Gene prioritization and clustering by multi-view text mining, *BMC Bioinformatics*, **11**, 28.
- [48] Yu, S., Van Vooren, S., Tranchevent, L.C., De Moor, B. and Moreau, Y. (2008) Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining, *Bioinformatics*, **24**, i119-125.
- [49] Yu, W., Gwinn, M., Clyne, M., Yesupriya, A. and Khoury, M.J. (2008) A navigator for human genome epidemiology, *Nat Genet*, **40**, 124-125.
- [50] Zhu, M. and Zhao, S. (2007) Candidate gene identification approach: progress and challenges, *Int J Biol Sci*, **3**, 420-427.

Appendix

A- I Additional data source

In this section additional data sources that retrieve ranked lists for gene-disease and disease-gene relations are described.

A- I.I Freudenberg and Propping

Freudenberg and Propping prioritize disease relevant to human genes by clustering diseases of known genetic origin based on their phenotypic similarities consisting of their episodic, etiology, tissue, onset and inheritance. The prioritization is done in a way that each cluster contains the disease and related genes to that disease. Afterwards in each cluster potential disease genes are then scored based on their functional similarities to known genes in the cluster with phenotypically similarity to the disease that has been queried. Validation of the result is then done using OMIM database.(Freudenberg and Propping, 2002)

A- I.II Disease Gene Prediction (DGP)

Disease Gene Prediction uses the specific sequence probability patterns of the genes known to be involved in monogenic hereditary disease to detect the key features shared in the group, these features consist of protein length, degree of conservation, phylogenetic extend and paralogy patterns. By using these patterns, DGP assigns probabilities to the genes that have the potential to only mutate based on their sequence probabilities. (López-Bigas and Ouzounis, 2004) Based on their probability DGP produces a ranked list that can then be used for the purpose of this project although the ranked list is only created for one disease.

A- I.III PROSPECTR

Genes involved in a disease would share a certain pattern on their sequences through evolution (Smith and Eyre-Walker, 2003). PROSPECTR uses basic sequence information and a machine learning approach to take sequenced based features such as gene length into account. PROSPECTOR classifies genes into genes likely or unlikely to be involved in a certain disease. These classifications are ranked based on a score calculated by the method which ranges from zero to one (Adie, *et al.*, 2005)

A- I.IV SUSPECTS

SUSPECTS (Adie, *et al.*, 2006) uses annotation data from Gene Ontology, InterPro and expression libraries together with the scores created from PROSPECTR. SUSPECT compares annotations with a set of genes that are known to be involved in that certain disease. Ranking of the genes is based on the likelihood of them being involved in a particular disorder. (Tiffin, *et al.*, 2006)

A- I.V Gene2Diseases

Gene2Diseases (G2D; http://www.ogic.ca/projects/g2d_2/) uses data mining algorithms to rank genes based on their phenotype of the disorder. G2D also ranks genes based on their similarity with a known disease gene on the chromosomal region where the disease is mapped. Candidate genes are scored through a BLASTX search on reference sequence.(Perez-Iratxeta C, *et al.*, 2002)

A- I.VI Prioritization of Candidate Genes Using Statistics (POCUS)

POCUS (<http://www.hgu.mrc.ac.uk/Users/Colin.Semple>) uses identifiable similarities such as share GO annotations, share InterPro domain and similar expression data to score a gene within loci. POCUS method of ranking is based on the idea of over representation of functional annotation between loci within the same loci. POCUS takes into account the prior knowledge of a disease gene such as preferred genes or known genes for a certain disease.(Turner, *et al.*, 2003)

A- I.VII eVOC

In the work of (Tiffin, *et al.*, 2005) they use eVOC anatomical ontology as a control vocabulary to integrate text mining of MEDLINE abstracts and data mining of available human gene expression data. eVOC uses the co-occurrence of genes to a specific disease to create ranked list of the best possible genes for a certain disease.

A- II Gold Standard Gene Sets

In this section all of the gold standard genes sets that were used in this project are identified with their Entrez identification. Table 12 shows genes that are found during literature search and data collections within AstraZeneca. This table is additional to Table 5 in Section 3.2.

Table 13 illustrates gold standard genes found by launched drug phases with their Entrez identification (section 3.4.3).

Table 12

Gold Standard gene sets created using literature search and data collected within AstraZeneca. These sets are described in method (Section 3.2).

Schizophrenia			Dyslipidemia		
MTHFR	4524	NR1H3	10062	HDLBP	3069
RGS4	5999	CES1	1066	HMGA1	3159
PLXNA2	5362	CETP	1071	APOF	319
DISC1	27185	CYP46A1	10858	APOA1	335
TPH1	7166	SAAL1	113174	APOA2	336
DRD4	1815	APOA5	116519	APOA4	337
GRIK4	2900	CLU	1191	APOC1	341
DRD2	1813	CYP2R1	120227	APOC3	345
FEZ1	9638	SGPP2	130367	APOD	347
OPCML	4978	CP	1356	APOE	348
GRIN2B	2904	CYP7A1	1581	APOH	350
DAO	1610	CYP11B1	1584	ENTPD8	377841
HTR2A	3356	CYP11B2	1585	ACAT2	39
DAOA	267012	CYP17A1	1586	LCAT	3931
NPAS3	64067	CYP27A1	1593	LIPA	3988
AKT1	207	AGT	183	LIPC	3990
CHRNA7	1139	AGTR1	185	LRP2	4036
RPGRIP1L	23322	ABCA1	19	MPO	4353
HP	3240	S1PR1	1901	PAFAH1B1	5048
TP53	7157	S1PR3	1903	PAFAH1B2	5049
SLC6A4	6532	EDN1	1906	ATP5B	506
APOE	348	EDNRA	1909	CYP39A1	51302
IL1B	3553	A2M	2	PLA2G5	5322
GAD1	2571	EP300	2033	PLTP	5360
ZNF804A	91752	EPHX2	2053	S1PR5	53637
ERBB4	2066	GPX6	257202	P2RY13	53829
PRODH	5625	PCOLCE2	26577	PON1	5444
MTHFR	4524	ANGPTL3	27329	PON2	5445
COMT	1312	C9orf47	286223	PON3	5446
GABRB2	2561	GPX1	2876	PPARA	5465
DRD1	1812	GPX2	2877	PPARD	5467
DTNBP1	84062	GPX3	2878	PPARG	5468
OFCC1	266553	GPX4	2879	APOM	55937
MUTED	63915	GPX5	2880	NCLN	56926
GRM3	2913	GPX7	2882	SAA1	6288
RELN	5649				
NRG1	3084				
PPP3CC	5533				
SLC18A1	6570				

Table 13

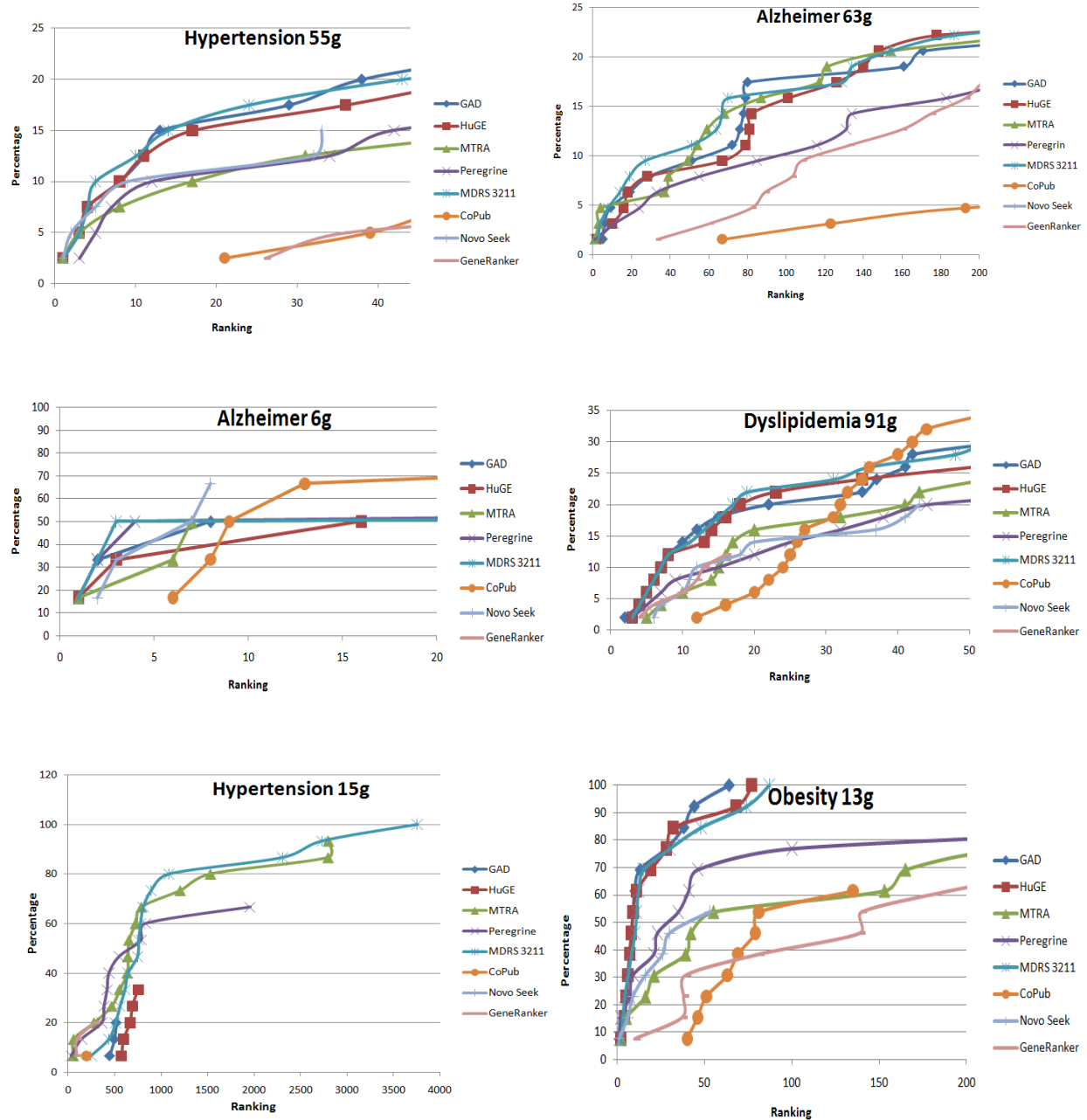
Gold Standard gene sets created using launched drug phases described in (section 3.4.3) with their Entrez identification.

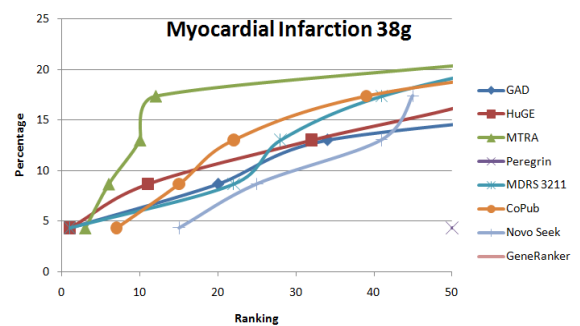
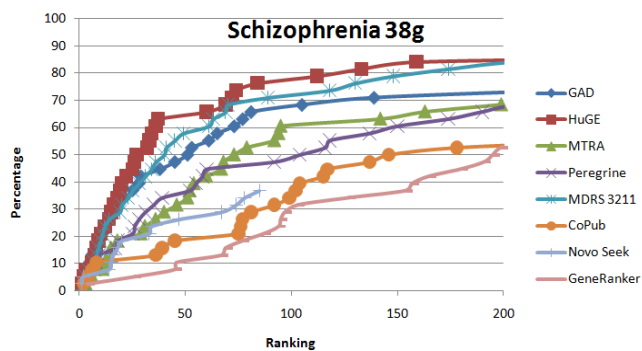
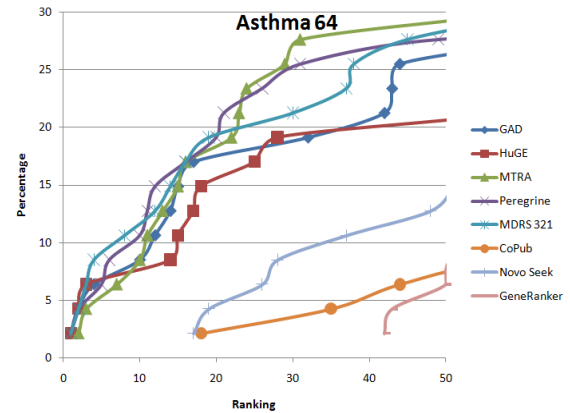
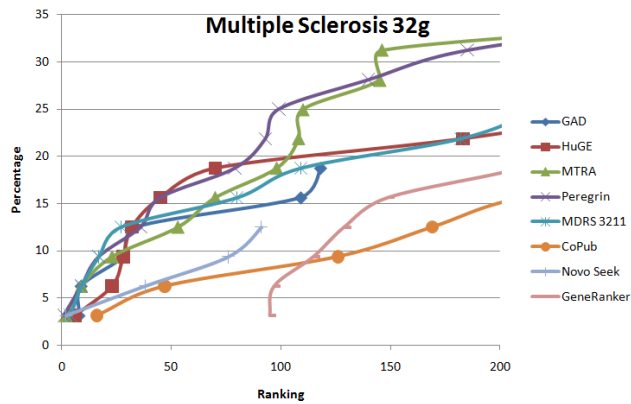
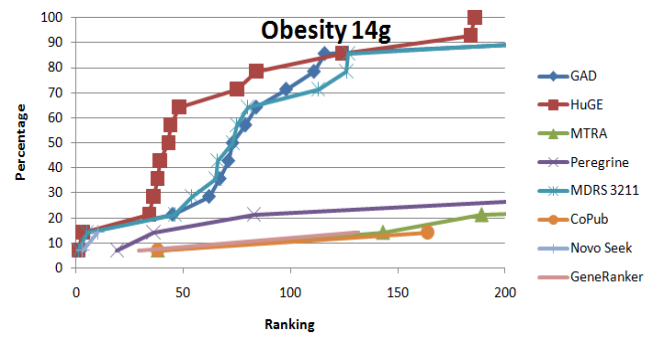
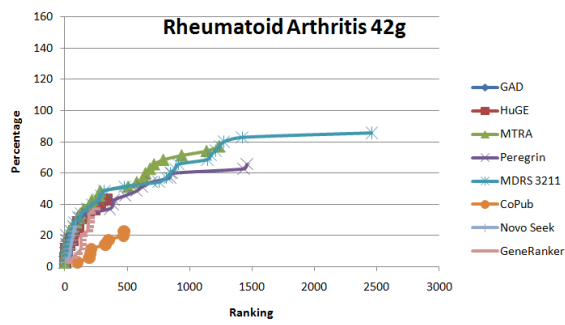
Multiple Sclerosis		Myocardial Infarction		Rheumatoid Arthritis		Pain disorder		Hypertension		Alzheimer		Asthma	
HMGR	3156	CACN	N/A	EC_1-1-1-205	N/A	CACN	N/A	CACN	N/A	CACN	N/A	ALOX5	240
DHODH	1723	HMGCR	3156	ALOX5	240	ALOX5	240	HMGCR	3156	ADH4	127	IDO1	3620
ADA	100	ALOX5	240	IDO1	3620	XDH	7498	PDE1A	5136	HMGCR	3156	PTGDS	5730
TOP2A	7153	MGAT1	4245	XDH	7498	XO	7498	CNP	1267	XDH	7498	XDH	7498
p38	7789	MGAT5	4249	XO	7498	AASDH	132949	ACE	1636	XO	7498	XO	7498
BCAM	4059	ACE	1636	AASDH	132949	SPAM1	6677	REN	5972	AASDH	132949	PARP4	143
CD52	1043	F2	2147	DHODH	1723	CASP3	836	MME	4311	MAOB	4129	PDE1A	5136
CNR1	1268	PLAT	5327	DHFR	1719	2000066	N/A	NPR1	4881	EC_2	N/A	CHIA	27159
CNR2	1269	PLAU	5328	PPP4C	5531	2000068	7498	2000071	N/A	SOAT1	6646	ADCY10	55811
ESR1	2099	HSD3B1	3283	CASP3	836	2006059	N/A	2000253	N/A	ACAT1	38	TBXAS1	6916
IL2	3558	2000071	N/A	2005834	N/A	ABCC1	4363	2005786	N/A	ACHE	43	TOP1	7150
IL2RA	3559	2005787	N/A	2006059	N/A	ABCC3	8714	2005787	N/A	BCHE	590	TOP2A	7153
ITGA4	3676	2006059	N/A	2006069	N/A	ADRA1A	148	ACCN5	51802	PDE1A	5136	2000021	N/A
KCNA2	3737	ADORA2A	135	ABCC1	4363	ADRA2A	150	ADRA1A	148	ACE	1636	2000233	N/A
MAPK14	1432	ADRA2B	151	ABCC3	8714	B3GAT1	27087	ADRA1D	146	CASP3	836	2005787	N/A
MBP	4155	ALOX5AP	241	CD4	920	CACNA1B	774	ADRA2B	151	FKSG2	59347	2005847	N/A
MS4A1	931	C20orf181	100128998	HSD17B6	8630	CALCA	796	ADRB2	154	2000021	N/A	2006152	N/A
NFE2L2	4780	C5AR1	728	IL10	3586	CALCR	10203	AGTR1	185	2005849	N/A	2006309	N/A
NFKB1	4790	CALCA	796	IL1B	3553	CD160	11126	CA2	760	2006309	N/A	2006502	N/A
NFKBI	4792	DRD2	1813	IL2	3558	CNR1	1268	CA7	766	ADRA1A	148	ADORA1	134
NFKBI	4795	EPO	2056	IL6	3569	CNR2	1269	CACNA1B	774	AKR1B1	231	ADRB2	154
NOS2	4843	P2RY12	64805	IL6R	3570	DRD2	1813	CDKL1	8814	ALDH7A1	501	ALDH7A1	501
S1PR1	1901	PDPK1	5170	INS	3630	GPR44	11251	DRD1	1812	APP	351	ALOX5AP	241
S1PR3	1903	PTAFR	5724	ITGAL	3683	GRIN1	2902	DRD2	1813	BBC3	27113	CCL2	6347
S1PR4	8698	PTGS1	5742	LCK	3932	HRH1	3269	FASLG	356	CALCA	796	CHIT1	1118
S1PR5	5363	SLC9A1	6548	MS4A1	931	HRH2	3274	GABARAP	11337	CES1	1066	CHRM3	1131
TOP2B	7155	TNXA	7146	NFKB1	4790	HSD17B6	8630	GRIA2	2891	CHRM2	1129	COL11A2	1302
Bdnf	12064	TP53	7157	NFKBI	4792	HTR2A	3356	GRIN1	2902	CHRM3	1131	CYSLTR1	10800
TRGT-00323	N/A	TRGT-00145	N/A	NFKBI	4795	ICAM1	3383	HTR2A	3356	CHRM4	1132	GPR44	11251
TRGT-01166	N/A	TRGT-00157	N/A	NIACR1	338442	KIR2DS2	100132285	INSR	3643	EGF	1950	GTGT-01529	N/A
TRGT-01167	N/A	TRGT-00229	N/A	NOS2	4843	NFKB1	4790	KCNA2	3737	ESR1	2099	GTGT-01530	N/A
TRGT-01258	N/A	TRGT-00301	N/A	NR3C1	2908	NFKBIA	4792	KCNA5	3741	FASLG	356	HRH1	3269
		TRGT-00302	N/A	PRDX5	25824	NFKBIL1	4795	KCND3	3752	GABARAP	11337	HSD11B1	3290
		TRGT-01244	N/A	PRKCI	5584	NR3C1	2908	LACTB	114294	GNRH1	2796	HSD11B2	3291
		TRGT-01738	N/A	PTAFR	5724	OPRD1	4985	MBLAC2	153364	GRIN1	2902	HSD17B3	3293
		TRGT-01743	N/A	PTGS1	5742	OPRL1	4987	NANOS3	342977	HSD17B6	8630	ICAM1	3383
		TRGT-02300	N/A	PTGS2	5743	PGR	5241	NOS3	4846	HTR6	3362	IFNG	3458
		PK	N/A	PTPN3	5774	PTGS1	5742	NR3C2	4306	JUN	3725	IGH@	3492
				TNF	7124	PTGS2	5743	PDE3B	5140	MAOA	4128	IL10	3586
				TRGT-	N/A	S100A12	6283	PPARG	5468	NFKB1	4790	IL2	3558

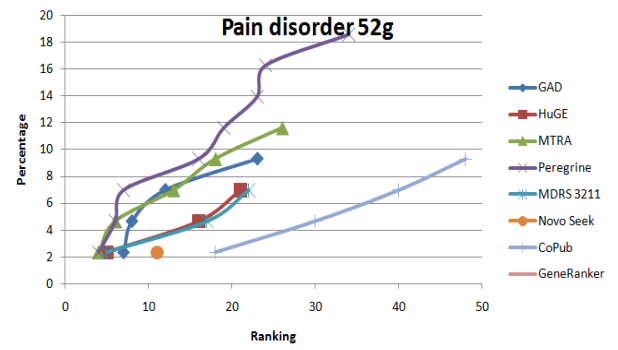
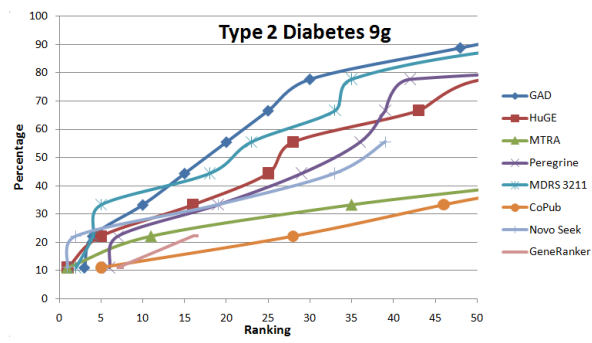
	01428									
	TRGT-04503	N/A	SCN4A	6329	SCN2A	6326	NFKBIA	4792	IL4	3565
	TXNRD1	7296	SLC6A2	6530	SCN4A	6329	NFKBIL1	4795	IL5	3567
			SLC6A4	6532	SCN5A	6331	PGR	5241	IL8	3576
			TACR1	6869	SCN7A	6332	PPARG	5468	KCNA2	3737
			TNF	7124	TRGT-00020	N/A	PTAFR	5724	KIR2DS2	100132285
			TP53	7157	TRGT-00157	N/A	PTGS1	5742	NFKB1	4790
			TRGT-00020	N/A	TRGT-00158	N/A	PTGS2	5743	NFKB1A	4792
			TRGT-00301	N/A	TRGT-00229	N/A	PTP4A2	8073	NFKBIL1	4795
			TRGT-00323	N/A	TRGT-00301	N/A	PTPN1	5770	NR3C1	2908
			TXNRD1	7296	TRGT-00302	N/A	SCN5A	6331	PARP1	142
			VCAM1	7412	TRGT-00304	N/A	SIRT1	23411	PTAFR	5724
			SCN	N/A	TRGT-03332	N/A	SLC18A2	6571	PTPRC	5788
					TRGT-03347	N/A	SLC6A2	6530	SLC22A12	116085
					TRH	7200	SSTR2	6752	TLR4	7099
					SCN	N/A	TNF	7124	TNF	7124
							TRGT-00012	N/A	TRGT-00229	N/A
							TRGT-00031	N/A	TRGT-00322	N/A
							TRGT-00051	N/A	TRGT-01431	N/A
							TRGT-00301	N/A	TRGT-01738	N/A
							TRGT-01743	N/A	TRGT-01739	N/A
							TRH	7200	TRGT-04503	N/A
							TXNRD1	7296	VCAM1	7412
							PDE	N/A	VDR	7421
									PDE	N/A

A- III Enrichment Plots on Gold Standard Genes

In this section enrichment plot of different data sources is given. The MDRS method was applied to GAD, HuGE, MTRA and Peregrine with weights assigned as $\mu_{GAD} = 1, \mu_{HuGE} = 1, \mu_{MTRA} = 1, \mu_{Peregrine} = 1$



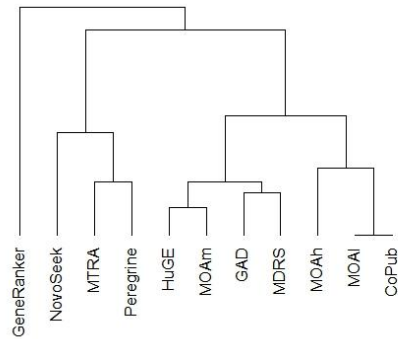




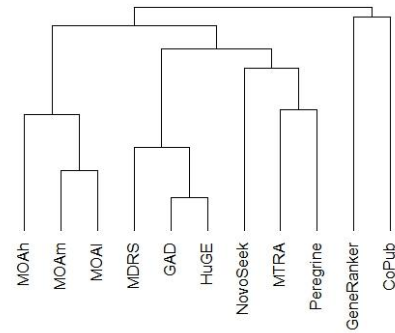
A- IV Clustering

Clustering of different data sources discussed in (section 3.6) is shown below. Two different types of clustering were applied for each disease: Spearman's rank correlation and Percentage similarities.

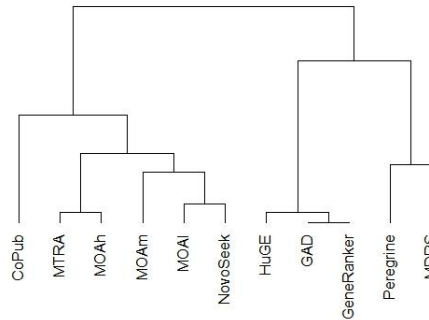
Asthma - Top 50 Spearsman's Rank Correlation



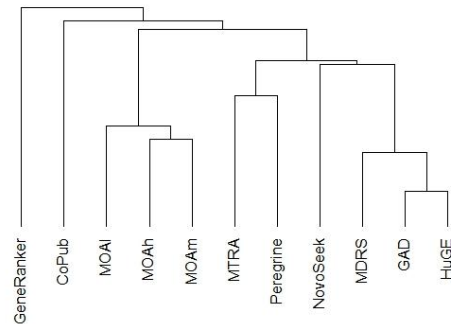
Asthma- Top 50 Percentage Similarites Genes



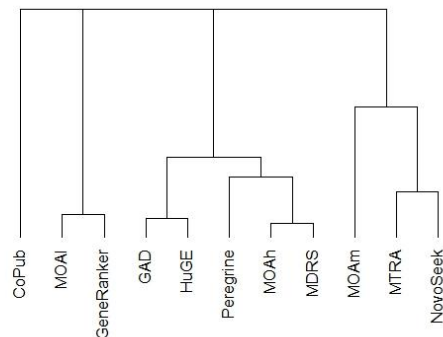
Hypertension - Top 50 Spearsman's Rank Correlation



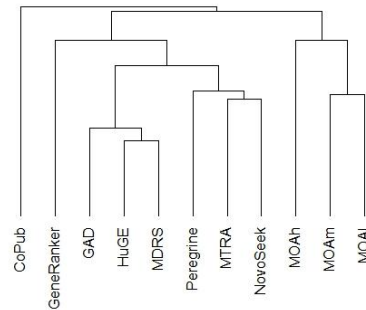
Hypertension - Top 50 Percentage Similarites Genes



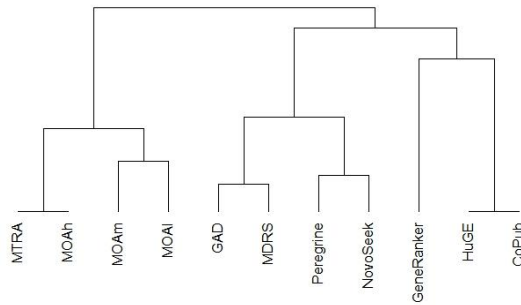
Dyslipidemia - Top 50 Spearsman's Rank Correlation



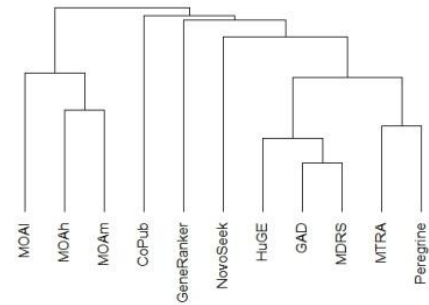
Dyslipidemia- Top 50 Percentage Similarites Genes



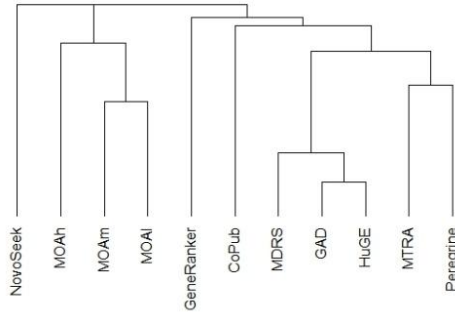
Multiple Sclerosis - Top 50 Spearsman's Rank Correlation



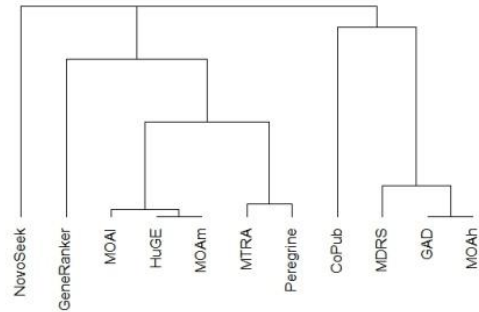
Multiple Sclerosis - Top 50 Percentage Similarites Genes



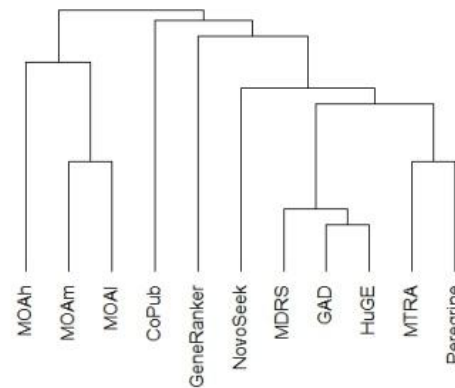
Myocardial Infarction - Top 50 Percentage Similarites Genes



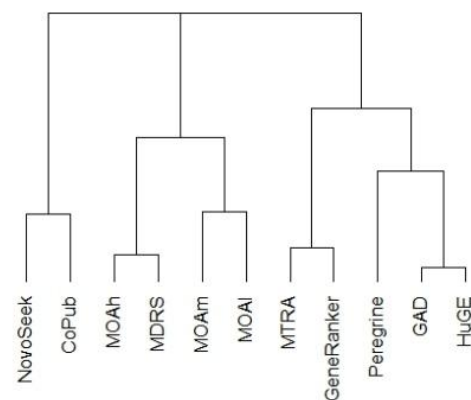
Myocardial Infarction - Top 50 Spearsman's Rank Correlation



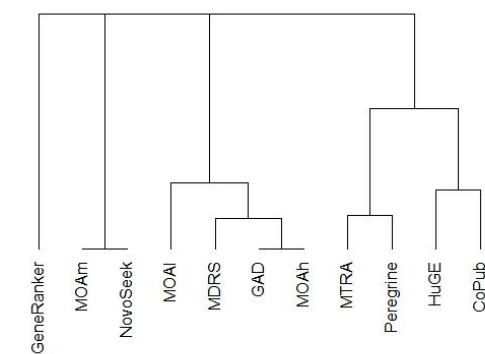
Obesity - Top 50 Percentage Similarites Genes



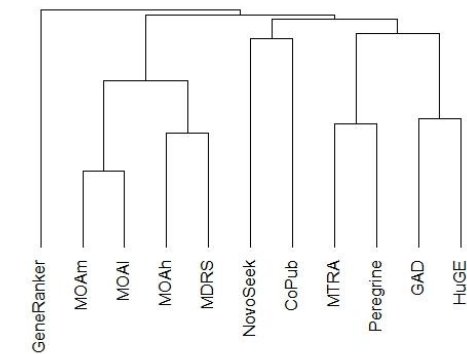
Obesity - Top 50 Spearsman's Rank Correlation



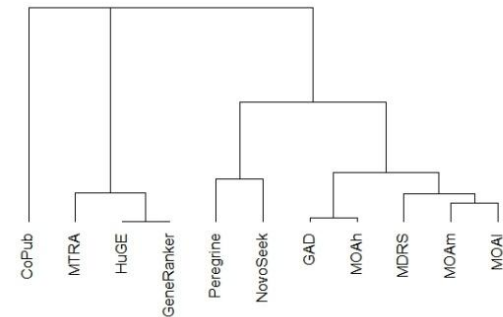
Pain Disorder - Top 50 Spearsman's Rank Correllation



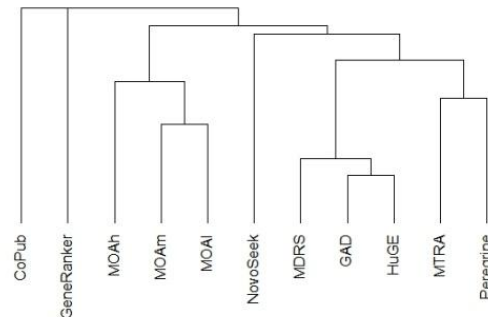
Pain Disorder- Top 50 Percentage Similarites Genes



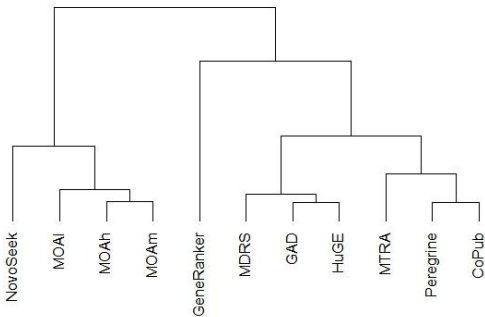
Rheumatoid Arthritis - Top 50 Spearsman's Rank Correllation



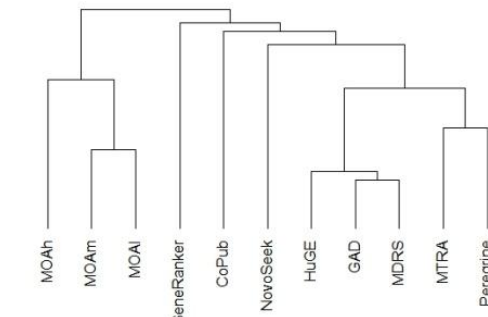
Rheumatoid Arthritis- Top 50 Percentage Similarites Genes



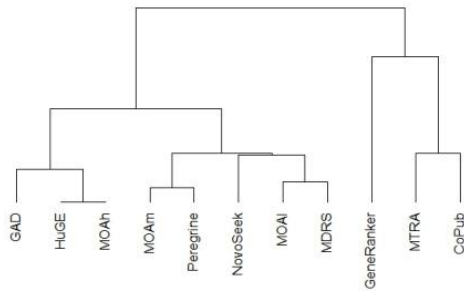
Schizophrenia - Top 50 Spearsman's Rank Correllation



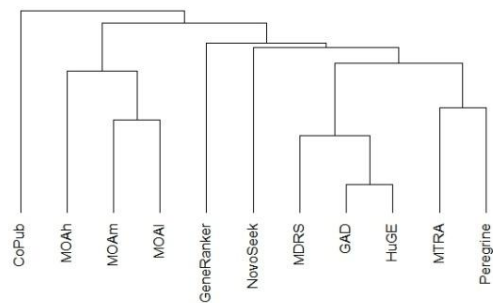
Schizophrenia - Top 50 Percentage Similarites Genes



Diabetes Mellitus Type 2 - Top 50 Spearsman's Rank Correlation

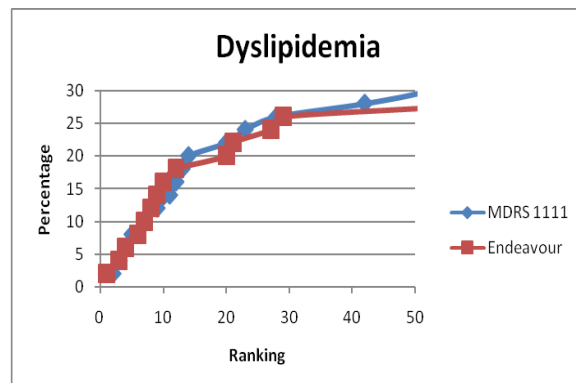
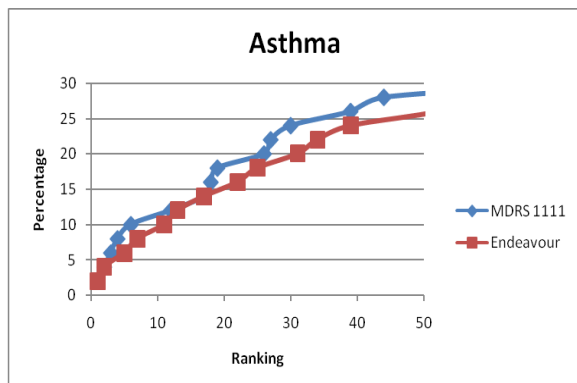
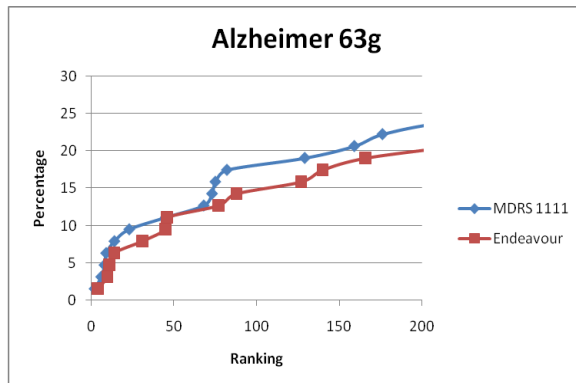
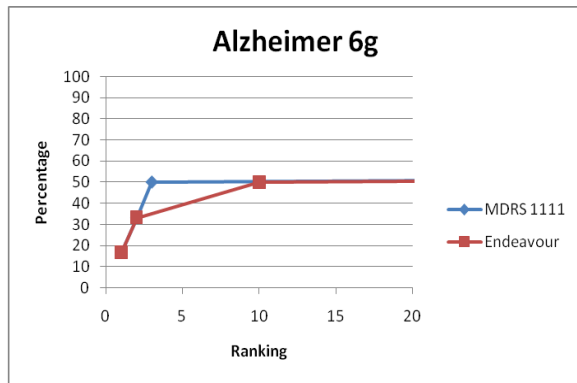


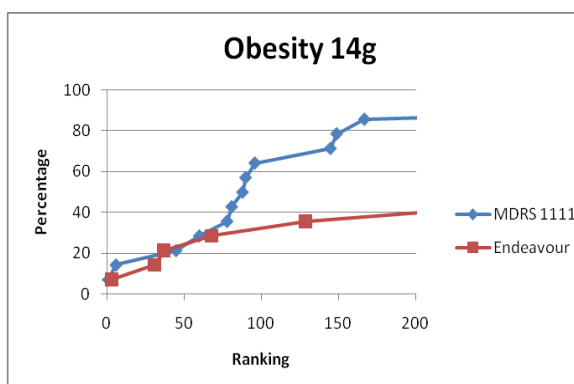
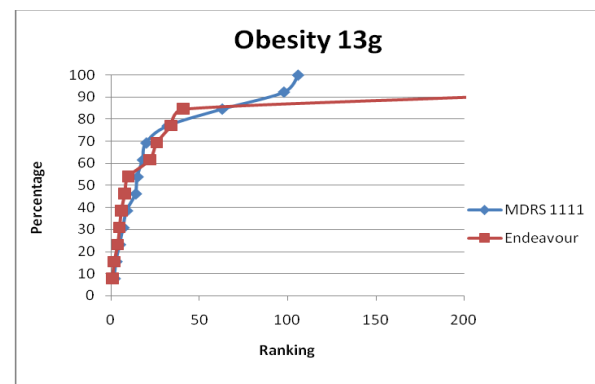
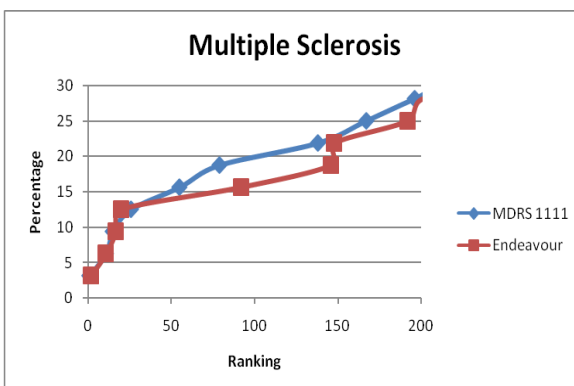
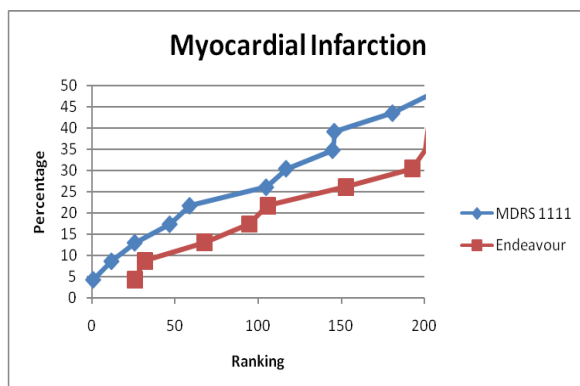
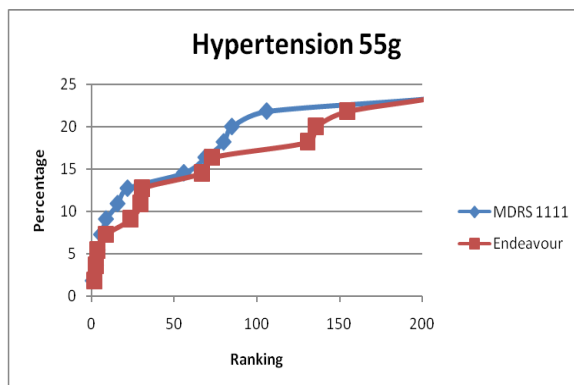
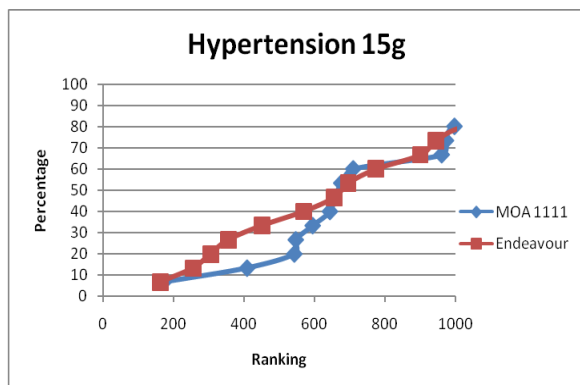
Diabetes Mellitus Type 2 - Top 50 Percentage Similarites Genes

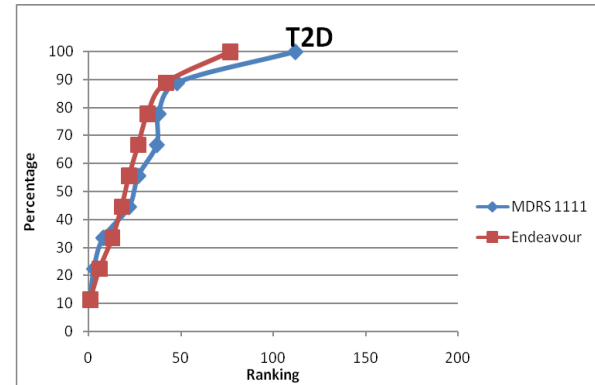
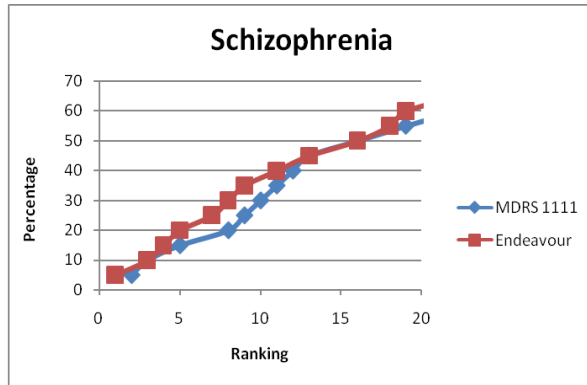
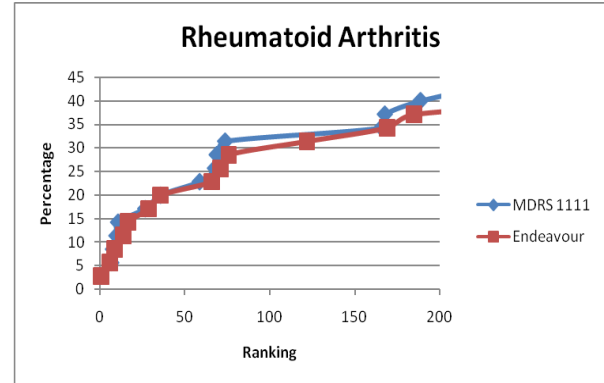
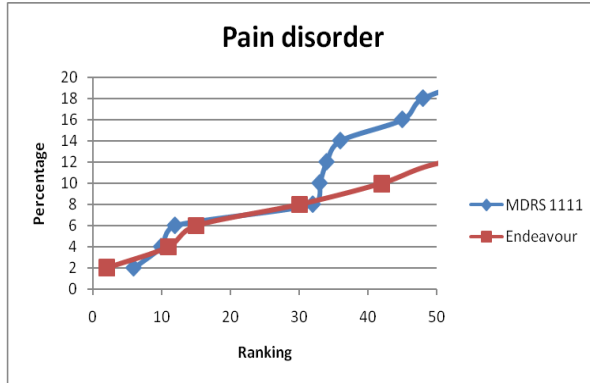


A- V MDRS vs. Order statistics in equal condition

Enrichment plots below shows the performance of Modified Discounted Rating System, with weights, $\mu_{GAD} = 1, \mu_{HUGe} = 1, \mu_{MTRA} = 1, \mu_{Peregrine} = 1$, and Order statistics method used in Endeavour (Section 2.2.2). MDRS is weighted equally so that the methods could be compared in an equal condition.







A- VI Analysis of a set of genes (Gold Standard Sets)

In Table 14, analysis of set of gene is done for gold standard genes. Analyses were done using method described in (Section 3.7) and gene sets given in (Section 3.2). The table shows the top five diseases related to a gold standard set sorted by their *Obs* score. Table also includes the *S'* score and *Z* score of each disease.

Table 14

Table showing the top five diseases related to each gold standard set sorted by their *Obs* score. Table also includes the *S'* score and *Z* score of each disease.

Gold Standard Set	GS Gene in set	Disease	<i>S'</i>	<i>Z</i>	<i>Obs</i>
Alzheimer	6	Alzheimer's disease	8.518	3.92	5
Hypertension	15	Asthma	4.086	2.574	6
		Colorectal Cancer	2.567	-1.173	3
		Premature Birth	1.790	1.699	2
		Lung neoplasms	1.517	-0.858	1
		Atherosclerosis	1.397	-0.858	1
Obesity	13	Obesity	7.566	4.326	9
		Hypertension	2.266	-0.452	2
		Attention deficit hyperactivity disorder	1.678	0.612	1
		Asthma	1.547	-0.799	1
Obesity	14	Obesity	9.597	6.673	13

		Alzheimer's disease	1.670	-1.510	1
Schizophrenia	38	Schizophrenia	6.137	8.201	9
		Chronic alcoholic intoxication	1.433	-1.004	2
		Attention deficit hyperactivity disorder	1.253	-0.497	
		Alzheimer's disease	1.237	-3.135	
		Unipolar depression	1.213	-0.812	1
Type 2 Diabetes	9	Non-insulin-dependent diabetes mellitus	7.284	3.439	6
		Insulin-dependent diabetes mellitus	1.972	-0.492	1
		Insulin Resistance	1.893	-0.365	1
		Obesity	1.864	-0.707	1
Dyslipidemia	91	Coronary heart disease	2.512	1.455	19
		Alzheimer's disease	2.181	-0.940	18
		Hypertension	1.587	-3.205	6
		Lung neoplasms	1.332	-4.913	8
		Malignant neoplasm of breast	1.262	-7.027	
Alzheimer	63	Alzheimer's disease	1.674	-2.056	6
		Malignant neoplasm of breast	1.598	-4.020	6
		Chronic alcoholic intoxication	1.432	-0.512	4
		Colorectal Cancer	1.374	-3.929	4
		Hypertension	1.364	-2.181	4
Asthma	64	Asthma	2.309	1.436	11
		Inflammation	1.677	-2.729	6
		Malignant neoplasm of breast	1.339	4.684	3
		Hypertension	1.324	-2.028	4
		Obesity	1.276	-2.776	2
Hypertension	55	Hypertension	2.490	1.106	11
		Schizophrenia	1.666	-0.489	5
		Long QT syndrome	1.309	0.705	2
		Atrial Fibrillation	1.166	-0.734	
		premature ovarian failure	1.161	0.081	
Multiple Sclerosis	32	Asthma	1.566	-0.591	3
		Malignant neoplasm of breast	1.527	-3.348	
		Rheumatoid arthritis	1.501	-1.554	2
		Multiple Sclerosis	1.465	-0.368	3
		Insulin-dependent diabetes mellitus	1.419	-1.297	2
Myocardial Infection	38	Asthma	1.602	-0.358	3
		Coronary heart disease	1.466	-0.353	3
		Hypertension	1.405	-1.396	2
		Cardiovascular system disease	1.231	-3.957	
		Chronic alcoholic intoxication	1.220	-0.926	
Pain disorder	52	Malignant neoplasm of breast	1.467	-4.401	3
		Asthma	1.431	-1.582	3
		Schizophrenia	1.402	-1.513	3
		Colorectal Cancer	1.326	-4.106	
		Attention deficit hyperactivity disorder	1.318	0.126	
		Pain Disorder	1.141	-2.003	2
		Inflammation	1.655	-2.526	4
Rheumatoid Arthritis	42	Asthma	1.589	-0.704	4
		Insulin-dependent diabetes mellitus	1.446	-1.368	3
		Colorectal Cancer	1.406	-3.572	2
		Rheumatoid arthritis	1.379	-2.096	2

A- VII Analysis of a set of genes (Pathway Analyze)

Table 15 shows the score of sets of genes related to different pathways. Method was described in (Section 3.7).

Table 15

Table showing the scores of gene sets related to a certain pathway described in (Section 3.7).

Airway Inflammation in Asthma									
Best candidates based on <i>TO</i> Score					Best candidates based on <i>O</i> Score				
	<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1 Asthma	4	4.62	10.5	7.89	Asthma	3	3.26	8.79	8.79
2 Inflammation	4	2.96	5.73	4.30	Z	1	-0.2	1.85	5.57
Airway Pathology in Chronic Obstructive Pulmonary Disease									
Best candidates based on <i>TO</i> Score					Best candidates based on <i>O</i> Score				
	<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1 Lung neoplasms	4	0.98	2.08	3.65	Inflammation	2	-0.3	1.96	6.86
2 Inflammation	4	1.14	2.96	5.18	Malignant neoplasm of breast	1	-1.5	1.04	7.28
3 Malignant neoplasm of breast	2	-0.8	1.44	5.06	Lung neoplasms	1	-1.2	1.03	7.26
4 Coronary heart disease	2	0.72	1.36	4.76	Coronary heart disease	1	-0.2	1.03	7.27
5 Bronchiolitis	2	3.11	0.65	2.28	Bronchiolitis	1	1.27	0.61	4.28
Cardiac Hypertrophy Signaling									
Best candidates based on <i>TO</i> Score					Best candidates based on <i>O</i> Score				
	<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1 Cancer	140	-1.2	0.84	0.84	Cardiovascular system disease	18	-11.3	0.23	0.23
2 Cardiovascular system disease	95	-0.9	0.63	0.63	Cancer	13	-19	0.10	0.10
3 Brain disease	68	-3.2	0.33	0.29	Malignant neoplasm of breast	12	-0.4	0.29	0.29
4 Nervous system disease	60	-7.6	0.17	0.33	Hypertension	11	-5.7	0.25	0.25
5 Malignant neoplasm of breast	54	-4.8	0.49	0.49	Schizophrenia	8	-4.8	0.17	0.17
Hepatic Fibrosis / Hepatic Stellate Cell Activation									
Best candidates based on <i>TO</i> Score					Best candidates based on <i>O</i> Score				
	<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1 Cardiovascular system disease	43	-2.5	0.59	1.74	Malignant neoplasm of breast	10	-7.3	0.51	6.48
2 Inflammation	42	0.51	1.30	3.92	Inflammation	10	-5.6	0.59	7.54
3 Malignant neoplasm of breast	39	-2.0	0.89	2.90	Lung neoplasms	6	-6.8	0.31	6.59
4 Lung neoplasms	30	-2.3	0.73	3.08	Asthma	6	-3.3	0.36	7.66
5 Alzheimer's disease	30	-0.1	0.77	3.24	Hypertension	6	-4.3	0.37	7.96
Leptin Signaling in Obesity									
Best candidates based on <i>TO</i> Score					Best candidates based on <i>O</i> Score				
	<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1 Cancer	49	0.35	1.03	1.5	Obesity	10	-1.4	1.06	7.53
2 Cardiovascular system disease	26	-1.5	0.47	1.29	Asthma	6	-1.7	0.40	4.79
3 Obesity	21	1.78	1.41	4.77	Alzheimer's disease	5	-3.4	0.24	3.46
4 Non-insulin-dependent diabetes mellitus	19	1.25	0.89	3.33	Non-insulin-dependent diabetes mellitus	4	-3.1	0.38	6.86
5 Insulin Resistance	15	-2.9	0.65	3.10	Edema	3	-1.5	0.07	1.87
Parkinson's Signaling									
Best candidates based on <i>TO</i> Score					Best candidates based on <i>O</i> Score				
	<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1 Parkinsonian disorder	7	4.06	1.12	2.57	Parkinson disease	6	3.11	2.84	7.59

2	Parkinson disease	7	3.88	2.88	6.58	Lung neoplasms	2	-1.6	0.17	7.05	
3	Cardiovascular system disease	5	-0.9	0.53	1.71	Immunologic deficiency syndrome	1	-1.3	0.08	1.42	
4	Inflammation	4	-0.3	0.30	1.22	Schizophrenia	1	-0.9	0.18	2.94	
5	Malignant neoplasm of breast	3	-1.5	0.40	2.14	Parkinsonian disorder	1	-0.6	2.84	1.71	
Prostate Cancer Signaling											
Best candidates based on <i>TO</i> Score		<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>	Best candidates based on <i>O</i> Score		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1	Cancer	96	6.66	1.69	1.70	Malignant neoplasm of ovary	11	-3.1	0.44	3.93	
2	Malignant neoplasm of breast	53	2.88	1.50	2.75	Prostatic neoplasm	10	-4.1	0.48	4.73	
3	Colorectal Cancer	31	-0.5	0.73	2.29	Malignant neoplasm of breast	10	-5.9	0.66	6.43	
4	Prostatic neoplasm	28	-0.1	0.67	2.33	Cancer	10	-11	0.17	1.73	
5	Malignant neoplasm of ovary	27	0.58	0.72	2.58	Asthma	5	-2.8	0.22	4.27	
Type I Diabetes Mellitus Signaling											
Best candidates based on <i>TO</i> Score		<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>	Best candidates based on <i>O</i> Score		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1	Cancer	72	0.53	1.03	1.51	Systemic lupus erythematosus	9	-1.9	0.35	4.08	
2	Inflammation	46	2.78	1.19	2.73	Insulin-dependent diabetes mellitus	8	-2.6	0.42	5.63	
3	Immune System disease	28	-3.2	0.26	0.99	Rheumatoid arthritis	8	-3.4	0.46	6.08	
4	Rheumatoid arthritis	27	1.50	0.86	3.37	Inflammation	8	-5.2	0.43	5.74	
5	Insulin-dependent diabetes mellitus	26	2.04	0.74	3.02	Asthma	7	-2.5	0.43	6.55	
Type II Diabetes Mellitus Signaling											
Best candidates based on <i>TO</i> Score		<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>	Best candidates based on <i>O</i> Score		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1	Cancer	90	-0.7	0.85	1.37	Non-insulin dependent diabetes mellitus	15	-2.9	0.72	6.96	
2	Brain disease	47	-1.9	0.31	0.96	Cardiovascular system disease	10	-9.2	0.19	2.89	
3	Insulin Resistance	45	5.26	0.86	2.78	Rheumatoid Arthritis	8	-4.2	0.32	5.95	
4	Cardiovascular system disease	45	-3.3	0.43	1.41	Insulin Resistance	8	-3.2	0.28	5.11	
5	Non-insulin dependent diabetes mellitus	44	3.06	1.24	4.09	Prostatic neoplasm	5	-6.7	0.10	3.02	
Arachidonic Acid Metabolism											
Best candidates based on <i>TO</i> Score		<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>	Best candidates based on <i>O</i> Score		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1	Cancer	70	-0.2	0.94	1.51	Malignant neoplasm of breast	13	-5.9	0.81	6.98	
2	Cardiovascular system disease	41	-1.5	0.50	1.37	Lung neoplasms	13	-4.7	0.57	4.94	
3	Lung neoplasms	36	-0.1	1.01	3.14	Colorectal Cancer	6	-6.3	0.35	6.58	
4	Colorectal Cancer	33	-0.8	0.71	2.43	Coronary heart disease	6	-2.9	0.33	6.26	
5	Malignant neoplasm of breast	32	-2.1	1.16	4.07	Asthma	5	-3.7	0.30	6.73	
Renin-Angiotensin Signaling											
Best candidates based on <i>TO</i> Score		<i>TO</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>	Best candidates based on <i>Obs</i> Score		<i>O</i>	<i>Z</i>	<i>S'</i>	<i>R'</i>
1	Cancer	84	3.01	1.24	1.54	Malignant neoplasm of breast	7	-6.9	0.32	4.78	
2	Cardiovascular system disease	37	-2.1	0.55	1.57	Hypertension	7	-3.5	0.47	7.09	
3	Cell Transformation, Neoplastic	27	-1.8	0.31	1.20	Alzheimer's disease	6	-4.4	0.19	3.45	
4	Malignant neoplasm of breast	27	-2.9	0.56	2.15	HIV infection	5	-3.5	0.21	4.27	
5	Non-insulin-dependent diabetes mellitus	21	-0.1	0.56	2.79	Asthma	5	-3.1	0.20	4.28	

VIII Deriving order statistics recursive formula:

Starting from Equation 1 with $N = k$:

$$\begin{aligned}
 \int_0^{r_1} \int_{S_1}^{r_2} \dots \int_{S_{N-1}}^{r_N} d_{S_N} d_{S_{N-1}} \dots d_{S_1} &= \int_0^{r_1} \int_{S_1}^{r_2} \dots \int_{S_{N-2}}^{r_{N-1}} [S_N]_{S_{N-1}}^{r_N} d_{S_{N-1}} d_{S_{N-2}} \dots d_{S_1} \\
 &= \int_0^{r_1} \int_{S_1}^{r_2} \dots \int_{S_{N-2}}^{r_{N-1}} (r_N - S_{N-1}) d_{S_{N-1}} d_{S_{N-2}} \dots d_{S_1} = \begin{cases} V_1 := r_N \\ = \text{const} \end{cases} \\
 &= \int_0^{r_1} \int_{S_1}^{r_2} \dots \int_{S_{N-2}}^{r_{N-1}} (V_1 - S_{N-1}) d_{S_{N-1}} d_{S_{N-2}} \dots d_{S_1} \\
 &= \int_0^{r_1} \int_{S_1}^{r_2} \dots \int_{S_{N-3}}^{r_{N-2}} \left[V_1 S_{N-1} - \frac{S_{N-1}^2}{2} \right]_{S_{N-2}}^{r_{N-1}} d_{S_{N-2}} d_{S_{N-3}} \dots d_{S_1} = \begin{cases} V_2 := V_1 r_{N-1} - \frac{r_{N-1}^2}{2} \end{cases} \\
 &= \int_0^{r_1} \int_{S_1}^{r_2} \dots \int_{S_{N-3}}^{r_{N-2}} \left(V_2 - V_1 S_{N-2} + \frac{S_{N-2}^2}{2} \right) d_{S_{N-2}} d_{S_{N-3}} \dots d_{S_1} \\
 &= \left\{ V_3 := V_2 r_{N-2} - V_1 \frac{r_{N-2}^2}{2} + \frac{r_{N-2}^3}{3!} \right\} = \dots \\
 &= \int_0^{r_1} \left(V_{k-1} - V_{k-2} S_1 + V_{k-3} \frac{S_1^2}{2} - V_{k-4} \frac{S_1^3}{3!} + \dots \pm \frac{S_1^{k-1}}{(k-1)!} \right) d_{r_1} = \overbrace{V_{k-1} r_1}^{i=1} - \overbrace{V_{k-2} \frac{r_1^2}{2}}^{i=2} \\
 &\quad + \overbrace{V_{k-3} \frac{r_1^3}{3!}}^{i=3} - \overbrace{V_{k-4} \frac{r_1^4}{4!}}^{i=4} + \dots \pm \overbrace{V_{k-k} \frac{r_1^k}{k!}}^{i=k} = \sum_{i=1}^k (-1)^{i-1} V_{k-i} \frac{r_1^i}{i!} = V_N
 \end{aligned}$$

where (for $k' < N$):

$$\begin{aligned}
 V_1 &= r_1 \\
 V_2 &= V_1 r_{N-1} - \frac{r_{N-1}^2}{2} \\
 V_3 &= V_2 r_{N-2} - V_1 \frac{r_{N-2}^2}{2} + \frac{r_{N-2}^3}{3!} \\
 V_4 &= V_3 r_{N-3} - V_2 \frac{r_{N-3}^2}{2} + V_1 \frac{r_{N-3}^3}{3!} - \frac{r_{N-3}^4}{4!} \\
 &\vdots \\
 V_{k'} &= V_{k'-1} r_{N-k'+1} - V_{k'-2} \frac{r_{N-k'+1}^2}{2!} + V_{k'-3} \frac{r_{N-k'+1}^3}{3!} - \dots \pm \frac{r_{N-k'+1}^{k'}}{k'!} \\
 &= \sum_{i=1}^{k'} (-1)^{i-1} V_{k'-i} \frac{r_{N-k'+1}^i}{i!}
 \end{aligned}$$

Then for $k' = k \leq N$ we have

$$V_k = \sum_{i=1}^k (-1)^{i-1} \frac{V_{k-i}}{i!} r_{N-k+1}^i$$

which equals Equation 2.