Subgroup Discovery on Numeric and Ordinal Targets, with an Application to Biological Data Aggregation

Barbara F.I. Pieters

Technical Report UU-CS-2010-012
May 2010

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

Subgroup Discovery on Numeric and Ordinal Targets, with an Application to Biological Data Aggregation

Barbara F.I. Pieters
Department of Information and Computing Sciences, Utrecht University,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands.
e-mail: bpieters@cs.uu.nl

Abstract

Subgroup discovery can generate descriptive patterns given a nominal or binary target variable. To do so, subgroup discovery uses quality measures that define the quality of a subgroup given the target values of the subgroup. However, not all problems are nominal in nature. More specifically, data can be either ranked and/or the target can be continuous. In the past, non-nominal targets needed to be discretized. Discretization can lead to less powerful or even faulty patterns, due to loss of information. Quality measures capable of dealing with continuous and even ordinal targets directly can help to overcome these issues. In this research, such quality measures are investigated and tested on the problem of gene set enrichment. Here, the goal is to find common functional knowledge on ranked genes. In this case, the genes are ranked according to their relevance to neuroblastoma, the most common extracranial solid tumour found in children. The results of the experiments are promising and show that subgroup discovery can be an addition to conservative research methods in for instance biology.

Contents

1 Introduction

In the recent past, it has become immensely popular to store all kinds of data. Moreover, it has become much easier to store large amounts of data, due to the developments in the hardware industry: hard disks and internal memory have become relatively cheap, and computers have become very fast.

In the field of biology and biomedicine, the ability to store vast amounts of data has been warmly welcomed. Ever since the human genome is known, genetic data is used to understand the function of (parts of) the DNA, up to understanding which genes and cell processes are of particular interest considering the causation of diseases. Storing such amounts of data also gives rise to a new problem. Data potentially contains valuable information, but searching through large amounts of data to retrieve this information is not done easily by hand. The situation in biology is no different. Although the search through data is still partly done by hand, by looking at irregularities and patterns in the data, it can be argued that the human eye can not fully find all irregularities and patterns. Therefore, along with the growing popularity of data storage, data mining has become equally popular, either to fully take over the data mining from human experts, or to aid experts in their search for valuable information.

In this thesis research, the technique of data mining is used to search for interesting and possibly unknown information considering the cause of neuroblastoma, one of the most common tumours found in children. It is thought that data mining can provide us with valuable information on the causation of neuroblastoma, or at least can give a better understanding of the development of neuroblastoma. It is believed that a more thorough understanding can help to improve existing therapies or even help the search for new ways to treat neuroblastoma. To achieve a deeper understanding, the idea is to enrich genetic neuroblastoma data with other data sources on genetics in general. Therefore, the technique of aggregation through multi-relational data mining is used, in order to combine the different sources and to find patterns from the combined sources.

1.1 Data Storage and Data Mining

Stored data can take many shapes. The simplest representation is data stored in a text file, where each line represents a record. A more elaborate representation is when data is stored in a *database*, such as commonly used relational database management systems. No matter what shape the data is in, data is usually stored as a collection of individuals, where each individual is called a *record*. An *individual* is just a collection of attribute-value pairs, where in some cases one of the attributes is viewed as the *target* or *class* attribute. Stored data can hold valuable information, for instance through *patterns* (relations, dependencies), which are obscured by the vast amount of the data. The main idea of data mining is the retrieval of valuable information by means of identifying patterns, to either describe the data or to classify new data [37, 25].

1.1.1 Subgroup Discovery

One of the techniques with which one can mine data, is subgroup discovery. A *subgroup* is a subset of individuals in the database, where the individuals in the subgroup are set apart from all other individuals by the characteristics of their attributes. These characteristics ensure that a subgroup displays a different distribution on the target attribute, compared to the distribution on the target attribute in the complete dataset. The characteristics of a subgroup are captured by a *condition*, where only individuals meeting this condition are part of the subgroup. To make things more clear, let us look at an example, as shown in Table 1.

This is a very small dataset, with attributes gender, age and married. Consider the attribute *married* to be the target attribute, i.e. we aim to search for characteristics of individuals given that the individual is married ($married = true$). A condition on which to characterize the subgroup is the age of a person: when one is older than 30, one is more likely to be married. This *relationship* can be formalized into the *rule*: $age > 30 \rightarrow married = true$. The conditional part of the rule, $age > 30$, gives an interesting different distribution of the data. When the data is divided on the

Gender	Age	Married
M	33	true
M	27	false
F	32	true
F	40	true
M	25	false

Table 1: Exemplary dataset

basis of the age of persons, the data shows a different distribution on the class variable *married*. In the whole population (all records in the dataset), 60% of the individuals are married, whereas 100% of the individuals are married when condition $age > 30$ is met.

Subgroup discovery is a rule learner, but it is not the only algorithm which mines for interesting rules. Originally, rule learning is concerned with classification (learning predictive rules) or learning descriptive rules. Subgroup discovery is a supervised learning technique, like other classification algorithms. However, instead of learning predictive rules, it generates descriptive rules, like association rule learning and other non-classification rule induction techniques [23, 28, 37, 1, 2]. Furthermore, unlike these techniques, subgroup discovery generates the interesting rules by means of a *quality measure* (*utility function*), where different measures return different rules, thus giving the user the ability to adapt the behaviour of subgroup discovery in general. Moreover, the rules found by subgroup discovery are relatively simple, and thus easy to understand [23, 3, 28, 36]. These characteristics, i.e. supervised rule learning, learning descriptive rules and the ability to adapt subgroup discovery, have made this technique more and more popular over the years, especially in the field of bioinformatics.

1.1.2 Quality Measures

The strength of subgroup discovery is also its drawback. Although current quality measures like novelty (a.k.a. weighted relative accuracy [29]) and information gain, are highly functional and heavily used as utility functions, they are not able to deal with numeric (or even ordinal) targets, such as age. To make this possible, the most easy solution would be to discretize (or even binarize) the target itself [45], and thus lose important information that is captured by the target variable. Apart from that, there is the issue of where to place the cut-off value when binarizing the data. To decide on the cut-off value, a data analyst has to have proper knowledge of the domain, which is not always the case.

To address this problem, one needs to define quality measures that can deal with numeric or ordinal target values. There are only a few measures currently known (and used) to evaluate numeric targets. Most of these measures use the mean of a subgroup for evaluation [20, 45]. Ordinal targets, where individuals display a certain meaningful order, pose an even bigger problem. For ordinal targets, quality measures to define interestingness are rare. Most solutions are about manipulating the target itself, changing it into a numeric or discrete target, or limiting the number of possible class values [27, 26, 15]. Purely statistical evaluation functions for ordinal data are also not that common, although they can be found in the field of nonparametric statistics or statistics for categorical data. When the ordinal target is a discrete one, preferably with a small number of categories, statistical measures for categorical data can be found in the field of behavioral sciences [4, 19]. In other cases, when the ordinal target is a ranking or is even continuous, nonparametric statistics are a better choice, such as the Wilcoxon's Rank Sum test and the Mann-Whitney U test [10, 6]. In this thesis, new quality measures are proposed in order to apply subgroup discovery to numeric and ordinal targets.

1.2 EET Pipeline

The idea to define new quality measures to evaluate subgroups with numeric or ordinal targets stems from situations where data can be ordered and/or where the target is continuous. For instance, genetic data can display an ordering, such as a ranking. What does ranked genetic data mean? When biomedical experts try to find out which genes play a role in the development of a disease, they measure for instance the gene expression of the DNA of each patient. Using data mining or another processing tool, the genes whose expression stand out when compared to normal gene expression, are believed to be important for the disease under investigation. Given the irregular gene expressions, the genes can be ranked, where the gene with the most interesting differential expression is set to be the gene with the highest rank. For this thesis research, we were presented with such genetic data, both with unprocessed genetic data and processed data, i.e. a gene ranking. Our genetic data was made available by the European Embryonal Tumour Pipeline poject, EET Pipeline or EETP in short. Within this project, several research groups work together to get a better understanding of embryonal tumours, such as neuroblastoma, medulloblastoma and retinoblastoma. Of the available datasets, the data on neuroblastoma is largest, which is the reason why only neuroblastoma was chosen as a research topic for this thesis.

1.2.1 Neuroblastoma

Neuroblastoma is the most common extracranial tumour found in children younger than 15 years and originates from primitive neuroblasts [32, 9]. Most of the research on neuroblastoma is dedicated to achieve a better understanding of the functioning of genes and their signalling processes with respect to neuroblastoma [7, 32, 9, 43]. Hence, the focus in the EET Pipeline also lies on the analysis of genomic and gene expression data. Four datasets were made available, three of which contain information on the genetic disposition of the neuroblastoma patients under investigation. These three datasets contain DNA, mRNA and miRNA data. The fourth dataset contains important clinical information on the patients, such as age at diagnosis, whether or not a clinical event had taken place and the stage of the neuroblastoma, which is related to the type of tumour. The genomic and gene expression datasets, DNA, messenger RNA (*mRNA*) and microRNA (*miRNA*), are interrelated in the following way. RNA is produced from DNA, it is a 'working copy' of the DNA that can be used for further processes in the cell. After that, mRNA is transcribed from RNA in such a way that only selected parts of the DNA (and thus RNA) end up in the mRNA. miRNA's are copies of very small portions of DNA that regulate whether proteins are translated from mRNA. miRNA's thus have a hand in which genes of the DNA (mRNA) are translated into proteins. From these three datasets, gene rankings can be made. Each ranking tells us which genes play a role in neuroblastoma and how important they are compared to other genes.

1.2.2 Meta Information on Genes

Apart from finding lists of ranked genes, there is even more that data mining can offer. As briefly mentioned above, the data from the project is interconnected by genes and proteins. Next to the gene rankings, there is also meta information available on genes. For instance, to which proteins genes code, to which protein families proteins belong, and whether a protein interacts with other proteins in the cell. There is even more interesting knowledge to explore: the gene ontology (*GO*) [18]. The gene ontology is an ontology of concepts related to genes, which strives for the standardization of the representation of genes and their characteristics, functions and products. There are of course many other knowledge domains that have a relation with genes and can provide additional information. Although it might seem evident that there is a lot to gain by using all kinds of meta information, it is not done that often. Why? There are several reasons for this. The most important one is the difficulty to combine and mine all data, a task that can be done through aggregation. The more conventional ways to deal with data are not well suited for aggregation. Mining multi-relational data in a multi-relational way is one of the best options to deal with aggregation problems.

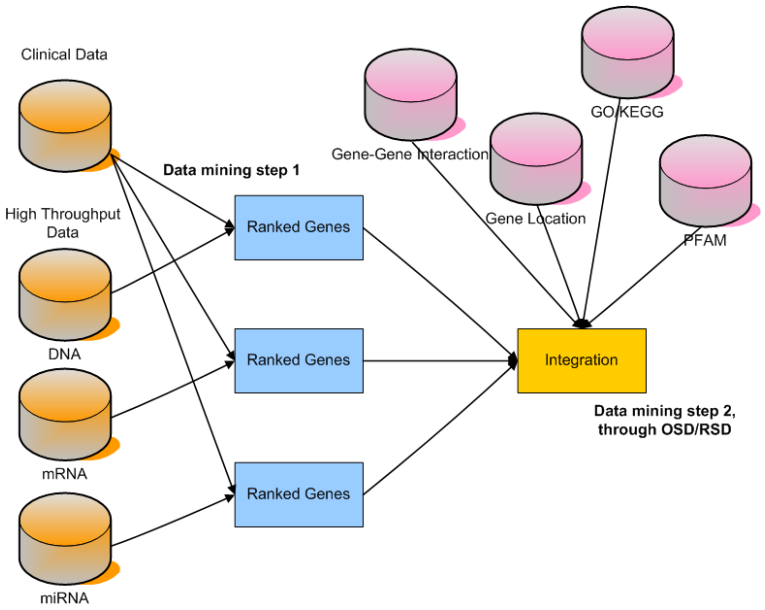


Figure 1: Data mining task: aggregating meta information and ranked data

1.3 Multi-Relational Data Mining

Currently, most data mining tools assume that data can be easily captured in a single file or a simple tabular structure. In real life however, data usually has a more complex structure, which is not easily captured in a single table. Even so, if it is possible to make the data tabular, this might clutter the data or information might become lost [11, 25]. For instance, for the aggregation problem described above, consider a single gene. We would have to store the gene, its position on the genome, interacting genes, protein families and GO-terms in one file. Although it is possible to store the data in such a way, it would be troublesome to decide on dimensionality and it would over-complicate the data file too much. Furthermore, relations between and knowledge of the different domains might get lost. Especially the tree structure of GO-terms is difficult to translate and lots of information can become lost after translation.

There are very few tools available that can fully tackle data mining problems in a multi-relational way. One tool that can mine multi-relational data properly is Safarii, a generic multi-relational data mining environment [25, 34]. Apart from that, subgroup discovery is available in Safarii, thus making it a logical choice to use Safarii as the tool for the mining tasks in this thesis. For the goals of this thesis, Safarii is enhanced so that it can perform subgroup discovery on numeric and ordinal targets too.

1.4 Combining All Issues

The enhancement of subgroup discovery and aggregating multiple data sources in order to aid in the neuroblastoma research, is done as follows. From the EET Pipeline data sources (the DNA, mRNA and miRNA data), gene rankings are made, using any (clinically interesting) target. Here, the stage of neuroblastoma tumour and whether a clinical event (such as a relapse) has taken place were identified as the most interesting targets. The obtained gene rankings are then enriched with additional knowledge domains, which are the location of the gene on the genome, interacting genes, GO-terms and protein families. This enrichment is done through aggregation: the gene ranking is mined multi-relationally with the domain knowledge data. The enrichment provides us with possibly valuable information and knowledge, in order to better understand which processes are involved in the development of neuroblastoma. Furthermore, the enrichment can help further

research on neuroblastoma, for instance to decide if the research focus needs to be adjusted.

In order to perform aggregation, specifically ordinal subgroup discovery is needed, since the targets, the gene ranking and the corresponding numeric attribute, are ordinal. Nevertheless, although the problem under investigation here is multi-relational, both Safarii and ordinal/numeric subgroup discovery can of course be applied to propositional data as well. Figure 1 shows how the research is performed. Here, OSD stands for ordinal subgroup discovery, and RSD stands for regressional subgroup discovery, which is subgroup discovery on numeric targets.

This thesis is structured as follows. In Chapter 2, a more extensive explanation of the EET Pipeline project and neuroblastoma is given. Following, Chapter 3 will thoroughly describe the data from the EET Pipeline project and the additional data sources that were used for this thesis. Also, preprocessing and alteration steps performed on the data, in order to obtain a gene ranking, are explained here. In Chapter 4, the concept of multi-relational data mining is described. In order to get a good understanding of subgroup discovery, Chapter 5 describes the concepts of this technique, and the subtypes of subgroup discovery are discussed here. Chapter 6 explains what kind of characteristics of (or *intuitions* on) subgroups are important when evaluating new subgroups. In Chapter 7, several (new) quality measures, which are capable to cope with numeric and ordinal targets, are described. Also, the measures are evaluated in terms of the intuitions from Chapter 6. In Chapter 8, experiments done on the neuroblastoma data enriched with the additional data are discussed. Finally, Chapter 9 concludes this thesis.

2 European Embryonal Tumour Pipeline Project

The European Embryonal Tumour Pipeline Project, *EET Pipeline* or *EETP* in short, is an EU-funded project that focuses on improving diagnostics and treatment for embryonal tumours. The tumours under investigation here include, among others, medulloblastoma, retinoblastoma and neuroblastoma (*nb*). For all these tumour types, the biologists and physicians involved in the EET Pipeline project are responsible for retrieving clinical and genetic data (meaning genomic and gene expression data) on patients. Apart from these researchers, also computer scientists are involved in the project. They are responsible for mining the available data in order to provide biologists and physicians with additional knowledge on the causation of the tumours considered in the project. The biologists and physicians are primarily located in Ghent, Belgium, and Essen and Heidelberg, Germany. The core of the computer science group resides in Ljubljana, Slovenia.

Considering all tumours under research in the EET Pipeline, most patients are diagnosed with neuroblastoma. Within the project, data from 101 patients diagnosed with neuroblastoma is available. Although a set of 101 individuals is small, seen from a data mining point of view, this is a rather large set of records in the opinion of biologists and physicians. Since for the other tumours the number of patients in the project is relatively small (around 30 or less), only the neuroblastoma data is used in this thesis.

2.1 Neuroblastoma

Neuroblastoma is the most common extracranial solid tumour found in children. It originates from neuroblasts, which are primitive cells of the sympathetic nervous system, mostly of the adrenal glands. The tumour can develop in nerve tissues in the neck, chest, abdomen and pelvis. The tumour is rare in older children or adults, only 10% of the cases occur in children of age>5. Of 4000 neuroblastoma cases only 2% of the patients were older than 18 [49].

Each case of neuroblastoma is classified into one of 5 (6) stages: 1, 2 (2a and 2b), 3, 4 and 4s, where classification is done upon diagnosis. Of all these stages, stages 4 and 4s are very important, and stage 2a and 2b are not distinguished in our data. Stage 4, metastatic neuroblastoma, is mostly found in children older than 1.5 years. Spontaneous regression or maturation of the tumour is frequently found in younger children, even when the disease is metastatic, which is the case for nb stage 4s. Spontaneous regression or maturation also occurs when young children (age $\leq$ 1.5 years) are diagnosed with stages 1 and 3 [32]. Stage 4 and 4s tumours look more or less the same. Both are metastatic, although in stage 4s the dissemination is still limited. The biggest difference between the two stages is that patients diagnosed with stage 4s tumours have a good survival rate, whereas patients diagnosed with stage 4 tumours have a high mortality rate: 80% of survival versus 30% [7].

The causality of neuroblastoma is not well understood. Neuroblastoma develops at an early age, even in embryos. A mass screening study in the industrialized world has shown that the incidence of neuroblastoma is fairly uniform. Furthermore, research into (environmental) risk factors is ongoing, but current results have been inconclusive. Taking all the insights in consideration, and especially the early onset, it seems unlikely that environmental factors play an important role [49, 7]. Thus, current research has focused on getting a better understanding of which genes and processes govern the disease [9]. The best option would be to compare normal – non-tumour – cells, i.e. neuroblasts, to neuroblastomas. The difficulty here is that neuroblasts are not detectable in postnatal life [9]. Despite this problem, previous research on neuroblastoma has shown that at least the status of gene *MYC-N* and chromosome 17 signal higher or lower risks, depending on the stage of the tumour and if MYC-N and/or chromosome 17 are amplified or deleted [7, 43, 32, 9].

The research on neuroblastoma is still ongoing, with an interest in the genetic, cellular and molecular processes and functions that are involved in the development of neuroblastoma. Some research using meta information is done already, for instance by De Preter et al. [9, 8].

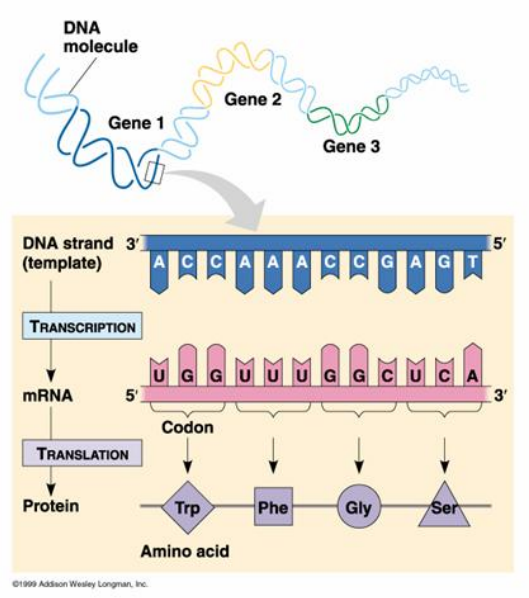


Figure 2: Relation between DNA, mRNA and proteins

2.2 EET Pipeline Data

During the lifespan of the EET Pipeline project several datasets have become available. At first, only a small set of (neuroblastoma) patients was examined. In the winter of 2009 a larger group was examined, and different types of data became available. All data that is of specific interest for this thesis is about genes in some way. To be more precise, DNA, messenger RNA and microRNA data is available, alongside clinical information on the patients. The three types of genetic data are interrelated in a specific way. The RNA is produced from the DNA in the cell, making RNA a 'working copy' of the DNA. In each cell the DNA has to perform different actions, depending on the cell in which the DNA resides. For the DNA(RNA) to behave differently, messenger RNA (*mRNA*) is produced from RNA through transcription. During transcription, bits of the actual RNA (thus DNA) are copied into mRNA, where the copied parts of the DNA are important to the cell at hand. The mRNA is then used to translate genes into proteins, through translation. The process of translating proteins from mRNA can be manipulated in several ways. MicroRNA, (*miRNA, μRNA*), which are very small pieces of RNA copied from the DNA strand, is one way how translation into proteins is regulated. For instance, miRNA can keep a gene from translating into a protein (lower expression), or miRNA gives genes a higher expression by allowing the gene to translate into a multitude of proteins (amplification). Figure 2 shows the relation between DNA, mRNA and proteins. How miRNA is copied and how it interacts with mRNA is not shown.

Each cell in an organism has the same DNA. Also, each cell in an organism has a specific function, and multiple cells can have the same function. For a cell to behave differently from other cells, certain parts are active where other parts are not. For instance, the cells in the ear that enable an organism to hear, need a different functionality from the cells in the eye, and vice versa. A cell can obtain its function through its DNA and the mRNA produced from the DNA. Depending on the function of the cell, parts of the DNA are of no use, where other parts are heavily important. The important parts are thus transcribed into mRNA. Thus, when investigating the role of genes in the development of tumours, it is vital to obtain genetic information from tumour cells and preferably their healthy counterparts. The genetic data sources (DNA, mRNA and miRNA data) available from the EET Pipeline are more fully described in Chapter 3.

2.2.1 Previous Work on EET Pipeline Data

A lot of research has already been done on older EET Pipeline data. In Van de Koppel et al. [42], initial data mining was performed on the data that was available to the project at that time. One of the goals was to get predictive models on different targets, such as the stage of the neuroblastoma. For this research, multiple datasets were used to combine information and to build the predictive model. This study suffered from a few problems. First, the number of records per dataset was very small (ranging from 19 to 63 neuroblastomas). Secondly, the intersection of the used datasets was even smaller than the individual datasets. These problems made it difficult to do a proper aggregation on the data and held back the accuracy of predictive models. In February 2009, a bigger sample set became available. This new dataset was used in the research of De Preter et al. [8]. The goal of this study was to find new therapeutic compounds to treat neuroblastoma. The search for new compounds was done through an integrative genomic meta-analysis of neuroblastoma cells and a comparison of these cells to neuroblast cells. For the full study, the reader is referred to [8].

Although further research upon the EET Pipeline data is currently conducted, none of it is published yet. One ongoing study performed at the Jožef Stefan Institute (IJS) is of particular interest. In this study, different datasets are combined in order to create rankings of differentially expressed genes. The focus of this study is *how* to create a proper ranking on genetic data, using different techniques and quality measures, such as the median value of the expression of individual genes. One of the rankings of their study was also used in this thesis study. For further information on the experiments conducted on the EET Pipeline data for this thesis, see Chapter 8.

2.2.2 Domain Knowledge

Since genes play an important role in regulating all kinds of processes in a cell, it is interesting to investigate which processes are regulated by which genes. Just looking at the genes that seem important for neuroblastoma is not enough. Specific domain knowledge is needed in order to understand which processes are involved. Such domain knowledge of course resides in biologists, physicians and several other specialists, but using human specialists for their knowledge on genetic processes in tumour research poses the same problems as leaving the data mining itself to human specialists. Thus, to be able to fully use domain knowledge on genetic processes, data mining can again be used, this time for enrichment. Still, specialists are needed to decide which knowledge is of importance. For the EET Pipeline, several sources were identified as interesting and easy to use, although in the future more sources can become interesting or usable due to ongoing research:

Protein Families Proteins, translated directly from genes on the mRNA, belong to protein families depending on the structure of the protein. Thus, proteins that look alike, or have the same function, belong to the same family. Families, in turn, are part of an even bigger structure: clans [13]. Information on protein families can be found at `http://pfam.sanger.ac.uk`.

RNA Families RNA, more specifically microRNA, can be categorized into families because of their similar structure and sequence. RNA families can be browsed at `http://rfam.sanger.ac.uk`.

Protein-Protein Interactions Proteins can interact with each other in order to change functionality and behaviour in the cell. Since proteins are translated from genes, protein-protein interactions can be viewed as gene-gene interactions. Protein-protein interactions can be found at the Human Protein Reference Database, `http://www.hprd.org`.

Gene Ontology - GO Terms The gene ontology is a structure in which genes are assigned to multiple terms. It was brought into life to provide consistent descriptions of gene products: the terms. There are three large subgroups of terms: cellular component, biological process and molecular function. Although GO is a highly interesting source of information, it has to be

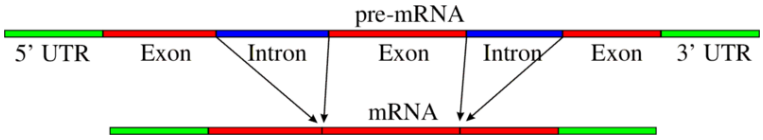


Figure 3: Splicing of RNA (pre-mRNA) to mRNA

viewed with a certain reservation considering its authority. The reason for this is the difficulty to empirically validate the GO terms assigned to genes, which can only be done through extensive research. Furthermore, defining the terms and their internal relationships is difficult. Information on the Gene Ontology can be found on `http://www.geneontology.org`.

Genetic Pathways - KEGG Genetic pathways are, like GO, a bit controversial considering their authority. Nevertheless, the information genetic pathways can provide is highly interesting. Genetic pathways are about the activation and signalling of genes/proteins through a (possibly large) network of genes. Thus, one gene at the top of the pathway can have an effect on a gene at the bottom, although there is no other way to tell these two genes have some effect on each other, except through the pathway. KEGG suffers from the same issues as GO. Information on genetic pathways, or KEGG, the Kyoto encyclopedia of genes and genomes, can be found at `http://www.genome.jp/kegg/`.

Other Sources The sources described above are highly informative and provide us with useful domain knowledge on genes and their functionality in cells. Of course, there is even more information available. For instance, when DNA is transcribed into mRNA, this process is done by *splicing*. During splicing, specific parts of the RNA, exons, end up in the mRNA. The exons which eventually end up in the mRNA are not always the same. Thus, determining which splice variants are coded, can be highly informative. The process of splicing, where exons are the parts that can end up in mRNA and introns signal when to start copying RNA to mRNA, is depicted in Figure 3.

Furthermore, DNA is a backbone to which nucleotides are attached, where the sequence of nucleotides gives the genetic makeup of an individual. DNA can have minor differences in those nucleotides when compared to the DNA of another individual. In particular cases, when such minor changes occur in a larger set of the population and when the change only involves one nucleotide in a larger sequence of nucleotides, they are referred to as *SNPs* (which stands for single-nucleotide polymorphisms, and is pronounced as *snips*). An example of a SNP: AAGC**C**TA versus AAGC**T**TA. Although this information can be very useful, it was not feasible to explore the possibility of including these extra sources, given the scope of this study.

Only a few of the additional information sources above are used in this thesis, primarily since some data sources lacked a proper representation. In these cases, it was too time-consuming to preprocess the data to achieve the proper format. Only for protein families (PFAM), gene-gene interactions (which are the same as the protein-protein interactions) and GO/KEGG, a proper representation was available or easily acquired. These data sources are further described in Chapter 3.

stage	event	deceased	total	stage total
1	no	no	22	
1	yes	no	1	23
1	yes	yes	0	
2	no	no	4	
2	yes	no	2	7
2	yes	yes	1	
3	no	no	4	
3	yes	no	4	11
3	yes	yes	3	
4	no	no	16	
4	yes	no	6	43
4	yes	yes	21	
4s	no	no	17	
4s	yes	no	0	17
4s	yes	yes	0	

Table 2: Distributions of patients on target types

3 Data

The data used in this thesis consists of data received directly from the EET Pipeline project and data retrieved from external sources on the internet. The data will be described more thoroughly in this chapter. Also, steps that were taken to (pre)process the data are discussed here.

3.1 Data from EETP

Preceding the winter of 2009, 101 neuroblastoma patients were examined. This data was made available in February 2009. This set is not so large seen from a data mining point of view, although it is a large set of patients by the opinion of biologists/physicians. Unfortunately, not all datasets have data on all 101 patients, and in some cases there are missing values. All data except the clinical data is recorded using *probes*. Each gene is covered by more than one probe and each probe can cover one or more genes. For these probes, their expression, which is a numeric value, is recorded. Thus, each probe shows the expression of (more than) one gene. One of the goals is to find genes that have a descriptive value to neuroblastoma, in other words, to find genes that are differentially expressed. To do this, the genetic probe data of the patients is mined using targets such as the stage of the tumour or the occurrence of an event. There are four datasets available: the clinical dataset and the DNA, mRNA and miRNA datasets. These datasets are described below.

3.1.1 Clinical Information

The clinical dataset contains important clinical information of the examined patients. Information is recorded on the stage of the neuroblastoma, the age of the patient at diagnosis, whether there has been some sort of event (such as a relapse of the tumour), if the patient is still alive, etc. From this data, multiple useful and interesting target attributes can be chosen to mine the data for differentially expressed genes. Of all these attributes, the stage of neuroblastoma and whether there has been an event (death or relapse) have been identified as highly interesting to use as target variables [8, 43]. Table 2 shows the distribution of patients according to these targets. The combined target $deceased = yes \wedge event = no$ is not shown, since $deceased = yes$ also sets *event* to *yes*. When *event* is used as the target attribute, there are 38 positive cases ($event = yes$), as opposed to 63 negative cases ($event = no$). When taking the NB stage as the target, this attribute is binarized as follows. $Stage = 4$, which gives a bad prognosis, is set as the positive case in terms

Dataset	Measuring	# Patients	# Probes	Gene mapping	Missing values
Array CGH	DNA	96	30813	Yes	Yes
Array CGH (CBS)	DNA	96	39573	Yes	Yes
Affymetrix	mRNA	101	284288	Yes	No
Affymetrix (core)	mRNA	101	22012	Yes	No
qPCR	miRNA	99	354	No	Yes

Table 3: Characteristics of genetic datasets

of data mining, resulting in 43 cases. The compound of NB stages 1, 2, 3, and 4s is set to be the negative class ($stage \neq 4$), adding up to 58 patients.

3.1.2 DNA

The DNA dataset is retrieved using the technique of array CGH [46]. Array CGH data shows whether a gene, (parts of) a chromosome or cytoband are amplified or deleted. In the case of amplification, a gene or (a region of) a chromosome is duplicated. The opposite is deletion, where the gene or (a region of) a chromosome no longer exists. Both amplification and deletion are believed to play an important role in the evolution and the occurrence of diseases. For all datasets described in this subsection, the DNA, mRNA and miRNA data all compare neuroblastomas to a control sample to compute the relative expression values of genes. In the case of array CGH, a healthy control sample was used. In other words, in array CGH the neuroblastoma cells are compared to neuroblast cells, which are the predecessors of neuroblastoma cells. The array CGH data is available in two flavours: normalized data and data preprocessed using the CBS algorithm [33, 44]. The CBS algorithm reduces the noise in a dataset, but can also modify the original values. Due to the data manipulation and loss of information that can occur when using CBS as a preprocessing algorithm, the normalized data is considered as best to use. Both datasets contain a fair amount of null values. The array CGH dataset contains 30813 probes, whereas the CBS variant contains 39573 probes. For both datasets there is a mapping from probes to genes available.

3.1.3 mRNA

The technique of gene expression profiling through DNA microarrays [47] is used to compute the mRNA data. Specifically, Affimetrix chips were used to profile the mRNA [48]. The mRNA dataset also needs a control sample in order to compare and compute the expression of genes. In contrast to the DNA data, mRNA uses a compound of 100 neuroblastoma samples as a control sample. There are two types of data available: the single probeset and the core probeset. In the case of the core probeset, each compound probe covers a larger part of DNA, usually one or more complete genes. The single probes mostly only cover a small part of a gene. Both mRNA datasets are normalized and have no missing values. The single probeset data is comprised of 284288 probes, making it the largest. The core probeset contains 22012 probes. For both datasets, a gene mapping is available.

3.1.4 miRNA

The miRNA dataset is made using the technique of qPCR [50]. As is the case with mRNA, miRNA uses a compound of 100 neuroblastoma samples as a control sample. Compared to the other two datasets, this dataset has a small number of probes, where each probe is one miRNA. The miRNA dataset only contains 354 probes, making this the smallest data set. The miRNA dataset is also normalized and contains a fair amount of null values. There was no mapping available from probes to genes.

Table 3 gives an schematic overview of the genetic datasets and their characteristics.

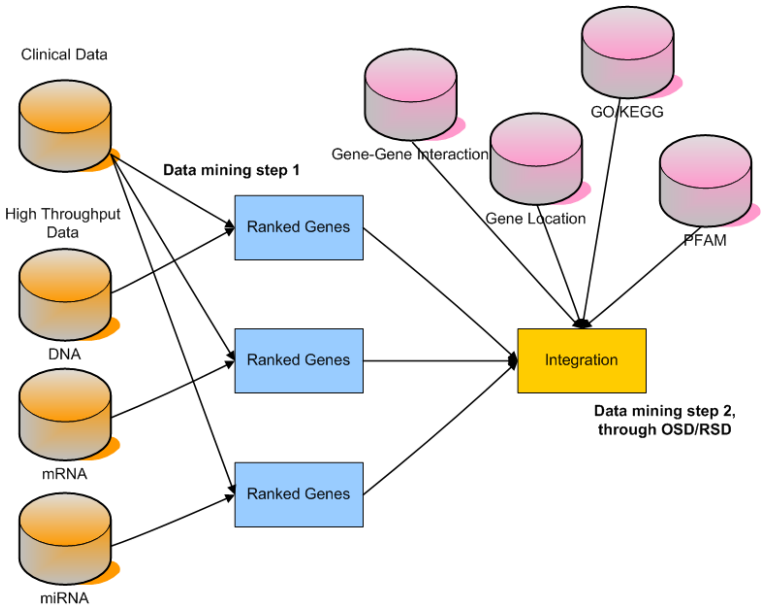


Figure 4: Data mining task: aggregating meta information and ranked data

3.2 Preprocessing EET Pipeline Data

The data described above has to be preprocessed before it can be properly used in the second data mining step, where aggregation takes place. For gene enrichment, a list of ranked genes is needed. To make a list of ranked genes, one genetic dataset is mined. There are two targets used, $event = yes$ and $stage = 4$. The goal was to make rankings for all types of genetic data, but only the mRNA dataset is chosen. Only for this dataset a proper mapping from (core) probes to genes is available, and there were no missing values. For the rankings, core probe expressions were mined. After mining the probes, the probes were mapped to genes.

Two rankings in this thesis were made using Safarii and the subgroup discovery algorithm in Safarii, one mRNA gene ranking with target $event = yes$ and one with target $stage = 4$. The quality measure used for subgroup discovery is *novelty*. Safarii and subgroup discovery are discussed in Chapters 4 and 5 respectively. The research group in Ljubljana provided the third ranking, a mRNA gene ranking with target $event = yes$.

3.3 Meta Information

Figure 4 shows how to aggregate the rankings with the meta information. As described in Chapter 2, there are many data sources interesting for aggregation. Of all these sources, only three were chosen for their ease of retrieval and their informative value. The only problem in adding new data sources is to get the data in a good format – a mapping to gene names is necessary. Furthermore, the extensive time needed to retrieve new data sources can be an issue, which was the case in this study.

Aggregating data is not new, especially in the field of bioinformatics. A method for multi-relational subgroup discovery and aggregation to search and enrich differentially expressed genes was developed by Trajkovski et al. in 2008 [41]. As opposed to the method of Trajkovski et al., it is very simple to add new data sources for aggregation in our approach, due to the generic multi-relational data mining tool Safarii.

3.3.1 GO/KEGG

The GO/KEGG dataset contains information on gene ontologies (*GO*) and genetic pathways (*KEGG*) in one. The dataset originally had a list of GO- and KEGG-terms per gene, where each gene, GO-term and KEGG-term was denoted by a numeric identifier. For the use in this research, the identifiers were changed to the gene names and GO/KEGG identifiers and names. Furthermore, genes are no longer stored with a list of GO- and KEGG-terms, but rather as gene-GO/KEGG-term pairs. Thus, for each gene, it is possible to have more than one pair in the database, although each pair is unique. This format was chosen in order to have no restrictions on the GO/KEGG terms selected during mining and to make proper multi-relational mining possible. There is only one drawback, resulting from the dataset itself. Both GO and KEGG have a tree-like structure. Thus, a term can have parent terms (and children terms). These parent terms, if any, are not all available. Availability depends on the structure of the tree, and whether a gene was set to both a GO term and its parent. One can argue that all parental terms can provide additional knowledge and thus should be accessible, but the data did not support this. The GO/KEGG dataset was made available by the Jožef Stefan Institute in Ljubljana, Slovenia, thanks to the research of Trajkovski [41, 22].

3.3.2 Gene to Gene Interaction

Genes can interact with each other through the proteins they translate into. The data for gene to gene interaction is stored as gene-gene tuples, where each tuple is unique. This dataset was also available with only numeric identifiers. The dataset was altered in such a way that each identifier was replaced by the corresponding gene name. Like the GO/KEGG dataset, for each gene in the dataset there was a list of interacting genes available, this was altered to obtain the gene-gene tuple format. This dataset too was received from the Jožef Stefan Institute in Ljubljana, Slovenia, thanks to the research of Trajkovski [41, 22].

3.3.3 Protein Families *PFAM*

The protein family dataset is publicly available from the PFAM website [13]. The mapping from protein family identifiers to genes was retrieved from Ensembl, a project to produce and maintain automatic annotations [12, 35]. This website can be used to retrieve many mappings from genes to a large range of other sources.

3.3.4 Gene Location

Since the mapping from genes to their (exact) location on the genome was not available, this mapping was added after retrieval from Ensembl [12, 35], using BioMart. This dataset contains information on the gene and its position on the chromosome. The position is available on chromosome and cytoband level. This data is useful since it is interesting to see whether differentially expressed genes lie on the same chromosome or on the same cytoband. Such information can let biologists and physicians decide to take a closer look on a specific chromosome or cytoband.

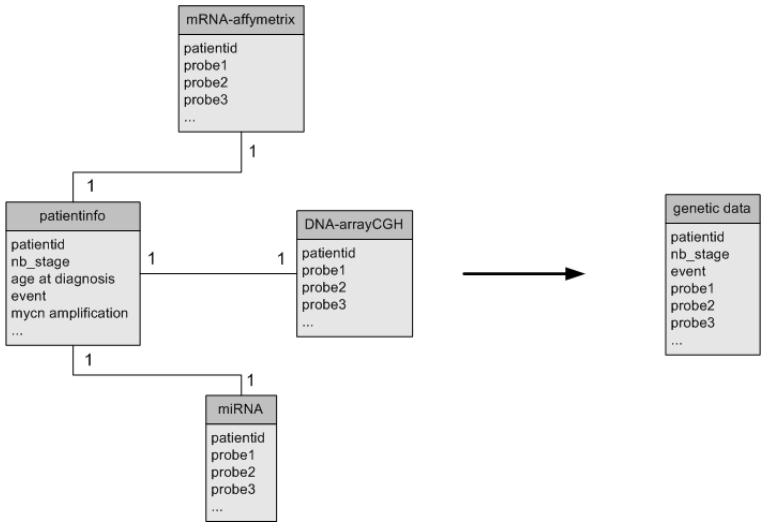


Figure 5: From multi-relational data to propositional data

4 Multi-Relational Data Mining

As explained earlier, the data used in this thesis is best represented in a multi-relational way, since the data is highly structured. However, in the case of the EET Pipeline data, it *is* possible and feasible to modify the data to make it propositional. Figure 5 shows the data for the EET Pipeline, divided into four separate datasets. From here, it can be seen that it is possible to combine the clinical information dataset with one of the high throughput data, such as the mRNA dataset. Although propositionalizing is feasible here, it is not necessary since Safarii can mine multi-relational data.

It is altogether different in the case of the meta information. Figure 6 shows which meta data is chosen for aggregation, and how difficult it is to propositionalize this data. For instance, in the case of gene interactions, should gene to gene interactions be stored as tuples? Or as an n-dimensional gene to genes tuple? Clearly, it is best to tackle the aggregation multi-relationally. Thus, the data will remain flexible and relations between different sources and in one data source are preserved.

Although many data mining techniques (or software, for that matter) focus on tabular data, data usually does not keep itself to such restrictions. Unfortunately, there are not that many tools available that can mine multi-relational data, without resorting to techniques to propositionalize the data. There are, however, a few tools that can mine data multi-relationally. For instance, MIDOS [51], a tool which can find subgroups in multi-relational data. Another example, although domain specific, is SEGS [41, 39], which is short for *Search for Enriched Gene Sets*. This tool also can perform multi-relational subgroup discovery on genes, gene-gene interactions, and GO/KEGG-terms simultaneously. MIDOS, however, is not domain specific. Most of these approaches, like SEGS and MIDOS follow the ideas of (Inductive) Logic Programming (*ILP*), thus restricting the data to be in a first order logic format. Furthermore, the results are also bound by constraints, usually set by the developer of the tool. Thus, a user can not easily adapt the representation of the results. More information on the history of multi-relational data mining, inductive logic programming, and tools created for the multi-relational data mining task can be found in [11, 14].

4.1 Safarii

Although most multi-relational data mining tools adopt the concepts of ILP, there is at least one that does not. Safarii, developed at Utrecht University by dr. A.Knobbe, is based on the concepts

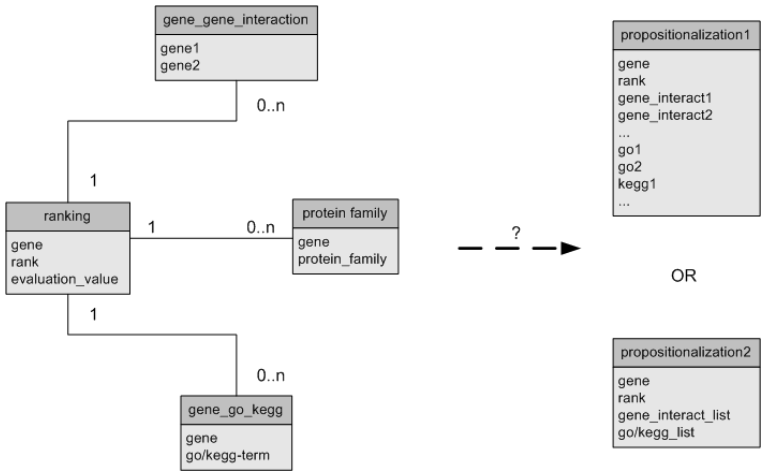


Figure 6: Fully multi-relational data

of the relational database model, which is still the dominant model for industrial database systems [25]. Safarii can mine data stored in a relational database management system, and is of course also capable of mining propositional data [34].

Safarii can be used to build a classifier on data or just to find interesting patterns (subgroups) in the data. The search for patterns can at least be done by the subgroup discovery algorithm, which was used in this thesis and will be discussed more thoroughly in Chapter 5. Safarii initially could only perform subgroup discovery on nominal targets. Furthermore, as is the case with other mining tools, even propositional mining tools, mining data with numeric targets or ordinal targets was not possible.

Given the multi-relational data, and the research question to mine lists of ranked genes, Safarii was enhanced so that it can find interesting patterns when dealing with numeric and ordinal targets whilst making use of the specific characteristics of these target types. This enhancement is called for, since former approaches to deal with numeric and ordinal targets mostly focused on discretizing or binarizing the target attributes, which is usually not done without hazard. Subgroup discovery and quality measures for ordinal and numeric targets are discussed in Chapters 5 and 7.

5 Subgroup Discovery

Subgroup discovery is a rule learning technique. *Rule learning* is usually divided into two separate techniques: classification rule learning and association rule learning. In the case of classification, the goal is to construct *predictive/classification* rules, and is a form of *supervised learning*. Association rule learning is about *descriptive rule induction*, a form of unsupervised learning, aiming to identify interesting patterns in the data [31, 37, 1, 2]. Thus, where the rules in classification are used to *predict* the class of a new individual, the rules in association rule learning are used to *describe* the data given the attributes of the data. Descriptive rules in this sense can thus be used to understand relations in the data, irrespective of some class (target) attribute.

Subgroup discovery is on the intersection of predictive and descriptive rule induction. The main goal of subgroup discovery is to find interesting patterns in the data given a target attribute. Thus, subgroup discovery can provide us with rules that describe the data given a target attribute, showing us underlying relations in the data. In contrast to classification rule induction, subgroup discovery does not build models or rules to maximize classification accuracy. Therefore, subgroup discovery can select rules more loosely, the only constraint is that rules should be interesting according to the user. Here, interestingness is usually defined in terms of the target attribute distribution. The rules found by subgroup discovery are usually simple in the sense that they are easy to understand by a user, especially a domain expert. Therefore, subgroup discovery has gained large interest from the field of expert guided data mining [17, 36, 31].

How does subgroup discovery work? Subgroup discovery finds groups of attributes, where the attributes render a different distribution on the target attribute, when compared to the distribution on the target given the complete set of attributes. In other words, subgroup discovery tries to find conditions on (a subset of all) attributes which divide the dataset into individuals belonging to the subgroup (these individuals meet the conditions) and individuals belonging to the complement of the subgroup: all other individuals. The individuals of the subgroup display a different and interesting distribution on the target attribute, compared to the distribution of the target attribute given the whole dataset, or the target attribute distribution of the individuals in the complement of the subgroup.

A rule in subgroup discovery is defined as follows:

$$rule = B \rightarrow H, \tag{1}$$

where B is named the body (*condition*), and H the head (*class, target*) of the rule. Subgroup discovery uses the condition of the rule (the body) to set individuals apart from the whole population. In other words, the individuals meeting the condition of the rule belong to the same subgroup. The class (*target*) values of the individuals are used to calculate the interestingness of the subgroup, by means of a utility function (*heuristic, quality measure*). In the case of nominal subgroup discovery, the head of a rule is the target value indicating the positive class, such as $target = 4$. The body of a rule defines the attributes and specific conditions on these attributes that specify a subgroup. Given the example from figure 1 with rule: $age > 30 \rightarrow married = true$, then $age > 30$ is the body of the rule, displaying the attribute and its condition. $Married = true$ is the head of the rule.

5.1 Target Attributes

The focus in subgroup discovery has been primarily on data where the target attribute is nominal. Of course, there are more types of target attributes, as there are more variants of attributes in general. When the target is nominal, it can only obtain a value from a (predefined) small set of values. The best known case is when the target attribute is *binary*. For nominal subgroup discovery (or data mining on nominal targets in general), there are many quality measures already available and well-researched. Nominal subgroup discovery is discussed further in Section 5.1.1 Apart from nominal targets, there are at least two other types of targets. One of them is the *numeric* target, where the target attribute can assume a range of values, particularly from a continuous interval of numbers. Understandably, deciding on whether a subgroup is interesting in the case of a numeric

id	target
1	5.01
2	3.27
3	0.98
4	1.25
5	2.89
6	0.01
7	0.25
8	0.26
9	4.28
10	7.65

Table 4: Numeric target

target is a bit different from handling nominal targets. Subgroup discovery on numeric targets, *regressional* subgroup discovery, is discussed in Section 5.1.2. A special subtype of both numeric and nominal targets is the *ordinal* target. In this case, the target attribute can pick a value from a range of discrete or continuous numbers or categories. Here, the numbers or categories display a certain order. For instance, the ranking of popular movies is an ordinal target. Because of the specific characteristics, ordinal targets should be handled appropriately, taking advantage of the characteristics. Ordinal subgroup discovery is further discussed in Section 5.1.3.

5.1.1 Nominal Subgroup Discovery

Subgroup discovery on nominal targets is the best researched variant of subgroup discovery. In this case, the target attribute can assume a value from a predefined finite range of values. Usually, the target is binarized, by taking one value as one class, and combining all other values into the complement class. An example is the neuroblastoma stage attribute, which is depicted in figure 2. Here, the stage can assume 5 values (1, 2, 3, 4 and 4s), but subgroup discovery is performed by comparing patients with stage 4 to all other patients, thus aggregating all patients with stages 1, 2, 3 and 4s into the compound class $stage \neq 4$.

As stated previously, subgroup discovery uses a *quality measure* in order to define whether a subgroup is interesting [3, 25, 28, 36]. One widely used quality measure for subgroup discovery in the field of bioinformatics and genetic research is the *novelty* (weighted relative accuracy, *WRAcc*) of a subgroup [23, 30, 36]. The novelty defines how different or *novel* the distribution of the target is given a rule, compared to the target distribution of the complete dataset. The novelty is defined as follows:

$$novelty(B \rightarrow H) = p(BH) - p(B)p(H), \tag{2}$$

where B and H again stand for the body and head of a rule, $p(BH)$ for the probability of B *and* H, also known as the *support* (correctly classified examples) of a rule, and $p(B)$ and $p(H)$ are the probabilities of the body and of the head of the rule respectively. The value of the novelty ranges from -0.25 to 0.25. A value of 0.25 indicates a strong relation between B and H, whereas -0.25 indicates a strong relation between B^C, the complement of B, and H. A value of 0 tells there is no relation between B and H, i.e. B and H are independent. A few other quality measures are for instance the entropy [5], and the χ^2 statistical test [10, 6].

5.1.2 Regressional Subgroup Discovery

Of course, not all possible target attributes are nominal. Target attributes can also be numeric, and continuous. Numeric targets can assume a large range of values, where the range can be either discrete or continuous. In the case of discrete numeric targets, the range of values is very large,

id	target	numeric	$\text{rank}_{\text{partial}}$	id	target	$\text{rank}_{\text{complete}}$
1	huge	5	1.5	1	0.15	1
2	huge	5	1.5	2	0.145	2
3	big	4	3.5	3	0.144	3
4	big	4	3.5	4	0.12	4
5	normal	3	6	5	0.118	5
6	normal	3	6	6	0.112	6
7	normal	3	6	7	0.11	7
8	small	2	8.5	8	0.09	8
9	small	2	8.5	9	0.086	9
10	tiny	1	10	10	0.08	10

Table 5: Ordinal targets with partial and complete rankings

possibly infinite. In the case of continuous attributes, the range is by definition infinite. Table 4 shows a dataset with a numeric (continuous) target.

One way to deal with numeric targets is to discretize them, preferably in such a way that the target becomes binary. Usually, this is done by discretizing the target into intervals [45]. Understandably, this can result in a loss of valuable information. Therefore, it is a better idea to use quality measures that can deal with numeric targets properly. Thus, a more appropriate approach is to use quality measures in which the distribution of the numeric target attribute is used. Logically, metrics such as mean and standard deviation are useful, which can then be compared to the metrics on the overall population or the complement of the subgroup. As is the case with nominal subgroup discovery, there are statistical measures available which can deal with continuous attributes, such as the mean itself, or the t statistic.

Previous research on regressional rule learning focused on finding new quality measures (for subgroup discovery) and dealing with numeric target attributes (and rule learning in general) can be found in [45, 41, 20, 24]. Quality measures for regressional subgroup discovery will be discussed in Chapter 7.

5.1.3 Ordinal Subgroup Discovery

Another interesting target type is the ordinal one. Ordinal targets are targets where the order of the target values captures information, consider for instance the ranking of athletes, where the top ranked athletes are the best athletes. Ordinal targets are usually numeric or can be represented numerically. Let's consider for example the fictitious datasets shown in Table 5. Here, we have two datasets containing 10 elements, where the target of each element is its size. On the left, the size is recorded by choosing from textual categories. Evidently, this textual category can be translated into a numeric one, which is already done in the example. The category "tiny" is represented by the number 1, whereas the highest category (huge) is represented by the number 5. This gives us a numeric order on the target, ranging from 1 to 5. Furthermore, let's assume that larger elements are preferred over smaller ones. This assumption can give us a *ranking* on the elements, as depicted in the column $rank_{partial}$. On the right, the size is recorded as a number, where the ranking of these elements is shown in column $rank_{complete}$. As can be seen, when a complete ranking can be constructed, there are no two elements with the same underlying numeric target. On the other hand, whenever there are individuals with equal target values, the ranking derived from the target is called an partial ranking.

Let's assume that ordinal targets are always numeric, since each type can be changed into a numeric one, whilst preserving the order. Then, interesting subgroups can be found through the techniques of regressional subgroup discovery. Although this is a good start, this approach plainly ignores the characteristics of the target, namely the ordering. It is not enough just to compare the distributions on the target of subgroups by the usual metrics, such as mean and standard deviation, like in the case of numeric targets. Especially if we are only interested in specific

individuals, such as top-ranked individuals. In the case of ordinal targets, no assumptions can be made on the distribution of the target attribute. This specifically calls for quality measures that can deal with a biased search, such as searching for the biggest elements or the best performing athletes.

Although ordinal targets and quality measures on such targets are not well-researched in the field of computer science, there is a strong interest from the field of behavioural sciences [4, 19]. The approaches presented there assume that the ordinal target only contains a (small) finite set of categories, or the target is modified into a finite set of categories [4, 26, 27]. Sometimes, the target is bluntly discretized or even binarized [15, 27].

Such alterations, like limiting the number of categories for the target, might not be justified. When considering the case of the ranking of athletes, where the ranking or even their running times can be used as a target attribute, the number of possible categories is infinite. This calls for a different approach on ordinal targets, where quality measures can deal with both finite and infinite ordinal targets. The discussion on quality measures for ordinal targets is continued in chapter 7.

$\mathbf{Rank_{complete}}$	$\mathbf{Rank_{partial}}$	**Target**	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
1	1.5	0.150	1	0	1	0	1	1	0	1	0	1
2	1.5	0.150	1	0	1	0	1	0	1	1	0	1
3	3	0.140	1	1	1	0	1	1	0	1	0	1
4	4.5	0.130	1	1	0	0	1	0	1	1	0	1
5	4.5	0.130	1	1	0	0	1	1	0	1	1	1
6	6	0.110	1	1	1	1	0	0	1	1	1	0
7	7	0.100	1	1	1	1	0	1	0	1	1	0
8	9	0.090	1	1	0	1	0	0	1	0	1	0
9	9	0.090	1	1	1	1	0	1	0	0	1	0
10	9	0.090	1	1	0	1	0	0	1	0	1	0
11	11.5	0.070	0	1	0	1	0	1	0	0	1	0
12	11.5	0.070	0	1	0	1	0	0	1	0	0	0
13	13.5	0.035	0	0	1	1	1	1	0	0	0	0
14	13.5	0.035	0	0	0	1	1	0	1	1	0	0
15	15	0.001	0	0	1	1	1	1	0	1	0	0
subgroup size			10	10	8	10	8	8	7	9	7	5

Table 6: Subgroups in a dataset (including the auxiliary complete ranking)

6 Intuitions on Subgroups

As explained in Chapter 5, subgroup discovery uses quality measures to calculate the quality of the subgroup. Subgroups, like quality measures, have certain characteristics, where the characteristics determine which subgroup is better, i.e. which subgroup should receive a higher evaluation value. The characteristics of quality measures are not always the same. Thus, it can be the case that the subgroup defined as best by one quality measure is not classified as such when using a different quality measure.

Furthermore, an analyst performing the subgroup discovery might also have certain ideas on what kind of characteristics good subgroups have. In other words, a user can have wishes on the characteristics of subgroups that should be generated. Such wishes can be translated into *quality intuitions* on subgroups. Since the characteristics of quality measures also determine what kind of subgroups are found, quality intuitions and quality measure characteristics are strongly related.

Each quality measure has a way to calculate the target attribute distribution of the subgroup, and some also have a way to calculate how different the target distribution of the subgroup is compared to the target distribution of the whole population, i.e. the dataset. Nevertheless, the calculation differs for each quality measure. The measures use certain *factors* (characteristics) of the subgroups, such as the subgroup size. Thus, quality intuitions are derived from the factors of a subgroup.

6.1 Preliminaries

Quality intuitions and quality measures on subgroups are best explained through examples. For this reason, a fictitious dataset is created, and depicted in Table 6. This dataset will also be used in further chapters. As can be seen from Table 6, our exemplary dataset consists of only 15 individuals. These individuals, and to which of the 10 subgroups they belong, are depicted on the right side of the table. The first three columns denote the different target types, where *target* stands for the original numeric target and $rank_{partial}$ denotes the partial ranking made from the original target. $Rank_{complete}$ is the complete ranking produced from the data, where ties in the partial ranking are cut arbitrarily. Although the order and ties of the original ranking should always be captured in the artificial ranking, the complete ranking is added since it is helpful to get a better understanding of quality intuitions and quality measures.

The subgroups in the table are denoted by s_i, where i is the identifier of the subgroup. The individuals covered by a subgroup are denoted by the value '1', whereas '0' means that the individual is not present in the subgroup. As explained earlier, certain factors can be taken into account in order to define the quality of a subgroup. One such factor is the subgroup size. The subgroup

size of s_i is denoted by n_i. For instance, the size of subgroup s_1 is $n_1 = 10$. Moreover, there are several factors that can be used to calculate for instance the target distribution of the subgroup, or to compare this distribution to the target distribution of the population. In the case of nominal targets, the distribution of the subgroup can be defined as the probability that an individual in the subgroup has the desired target value. This can be done by counting the individuals with the desired target value and dividing this number by the total number of individuals present in the subgroup. For nominal targets it is of course clear what kind of target value is desired, thus calculating the distribution is not that difficult. However, when considering numeric and ordinal targets, just counting individuals is not enough, since it is not clear how we should define which target values are desirable. For this purpose, standard statistical metrics of distributions are used, such as the mean and variance (standard deviation) of a subgroup. Hence, these metrics are defined as follows. The mean of subgroup s_i is denoted by μ_i, and the standard deviation, which is the square root of the variance, is depicted by σ_i. For instance, for subgroup s_1 with the original target, the mean is $\mu_1 \approx 0.118$, and the standard deviation is $\sigma_1 \approx 0.025$.

The goal here is to qualify subgroups, where the quality depends on intuitions. Therefore, we define the quality of subgroup s_i as q_i, where the quality is solely dependent on the intuition at hand. All intuitions presented here are applicable to both numeric and ordinal target types, unless stated otherwise. Although the focus here is on intuitions for numeric and ordinal targets, they are in essence also applicable to nominal targets.

6.2 Defining Quality Intuitions

One of the most interesting intuitions considers the size of the subgroup. Experts can for instance search for descriptive patterns, where the patterns cover many individuals in the dataset. One reason to search for large subgroups, is that the patterns accompanying these subgroups are very generic. In such a case, the pattern can reveal a very common relation in the data. This intuition can be stated as follows: if two subgroups, s_i and s_j, are completely equal except for their size, then the subgroup with the biggest size has the highest quality. To put it more formally:

Intuition 1 (Subgroup size maximization) *Given subgroups s_i, s_j, for which the following holds: $\mu_i = \mu_j$, $\sigma_i = \sigma_j$, and $n_i > n_j$, then $q_i > q_j$*

In some cases, an expert might be interested in relatively small subgroups. Such subgroups render patterns that are highly specific. For these subgroups the opposite of the subgroup size maximization is desired: the minimization of the subgroup size. The intuition can be changed accordingly: given two subgroups which are exactly the same except for their size, then the smallest subgroup is of a better quality.

Intuition 2 (Subgroup size minimization) *Given subgroups s_i, s_j, for which the following holds: $\mu_i = \mu_j$, $\sigma_i = \sigma_j$, and $n_i < n_j$, then $q_i > q_j$*

As can be seen, the second intuition, subgroup size minimization, is the exact opposite of the subgroup size maximization.

Apart from the subgroup size, there are several other important characteristics of subgroups. For instance, how the individuals of a subgroup are spread throughout the population, in other words, how the individuals are clustered. If individuals are evenly distributed in the population with respect to their target attribute values, the subgroup itself is not considered to be very different from the population. Consequently, the interestingness of a subgroup becomes highly questionable whenever the individuals of a subgroup are evenly distributed (i.e. the individuals are loosely clustered). The variance or standard deviation of the subgroup is a good metric to calculate the spread of subgroup individuals, since the variance generally tells us how closely the subgroup individuals are to the subgroup mean. Thus, given that two subgroups have equal size and have the same mean, but a different standard deviation, then the individuals of the subgroup with the highest deviation are more evenly spread throughout the population. This subgroup is then considered to be of lesser quality. Formally:

Intuition 3 (Spread of individuals (*Deviation*)) *Given subgroups s_i, s_j, for which the following holds: $n_i = n_j$, $\mu_i = \mu_j$, then $q_i > q_j$ iff $\sigma_i < \sigma_j$*

Consider for instance subgroups s_7 and s_9 in Table 6, with the complete ranking as the target. Both have equal sizes and equal means: $n_7 = n_9 = 7$ and $\mu_7 = \mu_9 = 8$. Their standard deviations are $\sigma_7 \approx 4.32$ and $\sigma_9 \approx 2.16$. Thus, subgroup s_9, which has a viewable smaller spread of individuals (the individuals are more tightly clustered) than subgroup s_7, is better compared to subgroup s_7 given Intuition 3.

The position of the subgroup individuals, i.e. the position of the cluster, is another interesting factor. Instead of determining that each tight cluster is equally good, the analyst might also have a preference for a certain position of the cluster. For instance, any cluster is good, as long as it is not clustered around the population mean. Thus, if there are two subgroups for which the individuals have an equal spread throughout the population (i.e. they have the same standard deviation), but the clusters of individuals have a different mean, then the subgroup with the best mean is considered the best subgroup. Take for instance the complete ranking as the target, where individuals with a top rank (small ranking number) are desired over individuals with bottom ranks. Then, the mean should be small for a subgroup to be qualified as a better subgroup.

Intuition 4 (Cluster position) *Given subgroups s_i, s_j, with equal sizes $n_i = n_j$ and equal standard deviations $\sigma_i = \sigma_j$. Then, $q_i > q_j$ iff $\mu_i < \mu_j$ in the case of mean minimization. Consequently, $q_i > q_j$ iff $\mu_i > \mu_j$ in the case of mean maximization.*

Let's consider subgroups s_1 and s_4 from Table 6, given the complete ranking as the target. For this target, we wish to minimize the mean, since individuals with top ranks are considered to be better. Both subgroups have an equal standard deviation: $\sigma_1 = \sigma_4 \approx 3.028$. The means of these subgroups are $\mu_1 = 5.5$ and $\mu_4 = 10.5$. Then, subgroup s_1 is considered to be better given Intuition 4. Due to the definition of this intuition, it is specifically applicable to ordinal targets. For ordinal targets it is our prime goal to find subgroups with a preference toward a cluster position, such as the minimization of the mean when a ranking is used as the target.

Analogous to the cluster position intuition, we can formulate an intuition about the difference in target attribute distribution. To be more precise, subgroups are generally considered more interesting whenever their target attribute distribution is *different* from the target attribute distribution given the whole population. Consequently, if the target distribution of one subgroup differs more from the population target distribution than the target distribution of another subgroup, then the first subgroup is considered to be better. Whether two distributions differ, can be calculated by substracting their target means with the population target mean.

Intuition 5 (Distribution difference) *Given subgroups s_i, s_j, with equal sizes $n_i = n_j$, standard deviations $\sigma_i = \sigma_j$, and unequal means $\mu_i \neq \mu_j$. Consider population p with mean μ_p. Then $q_i > q_j$ iff $|\mu_i - \mu_p| > |\mu_j - \mu_p|$*

Let's again consider subgroups s_1 and s_4 in Table 6, together with subgroup s_2, given the complete ranking as the target. All three have the same standard deviation ($\sigma \approx 3.028$), but have different means: $\mu_1 = 5.5$, $\mu_2 = 7.5$, $\mu_4 = 10.5$. The population mean is $\mu_p = 8$. Subgroups s_1 and s_4 are of equal quality, their difference is $|\mu_1 - \mu_p| = |\mu_4 - \mu_p| = 2.5$. Subgroup s_2 however, is of lesser quality: $|\mu_2 - \mu_p| = 0.5$.

6.3 Quality Intuitions versus Quality Measures

As explained earlier, the quality intuitions presented here are derivations of wishes an analyst might have considering the quality of the subgroups that are found. Of course, since these intuitions are used to describe characteristics of subgroups, they can be used to describe characteristics of quality measures as well. Nevertheless, one has to keep in mind that not all intuitions necessarily

have to be applicable to a quality measure or a subgroup at once. The quality intuitions can also be viewed as *features* that might hold for a quality measure, up to a certain degree. For most quality measures, several of the intuitions are applicable at once, although not with equal weight. Thus, defining which quality measures to use given the wish list of a user, is a matter of deciding which wishes are more important. To what extent the intuitions are applicable to the measures, is discussed in Chapter 7.

7 Quality Measures

The evaluation values calculated by the quality measures define the quality of subgroups. To calculate the evaluation values, several *(statistical) metrics* can be used. These metrics tell something about the characteristics of the subgroup, such as the deviation of the subgroup individuals. Although quality measures are based on statistical metrics and statistical tests, they are not completely the same. The quality measures have to be looked upon as *heuristics* and are only usable for evaluation.

As such, several requirements and assumptions accompanying the statistical metrics and tests do not have to be met. Consequently, if a data analyst wishes to make statistical inferences based on the metrics and test statistics, this should be done with great caution. For instance, for some quality measures, the evaluation values can be used to tell how significant a subgroup is. During mining, none of the requirements needed to obtain a confidence level, such as hypothesis testing and correction upon multiple hypothesis testing, are met. The approach to treat measures as heuristics is not novel, but has a rich background in the field of subgroup discovery and rule evaluation methods [24, 16, 3].

Only quality measures which are capable to deal with ordinal or numeric target attributes are presented here, since subgroup discovery is enhanced to find subgroups with such target attributes. For quality measures on nominal target attributes, the reader is referred to Chapter 5 and articles on subgroup discovery on nominal target attributes, such as [3, 30].

7.1 Preliminaries

Most quality measures make use of standard statistical metrics on a subgroup s and the population p. The population meant here is the complete dataset that is available for mining [24, 20, 38, 41, 3]. As such, the definition of the population is by no means equal to a population seen from a statistical point of view. The dataset is just treated as if it is a population from which a random sample – the subgroup – is drawn. Strictly speaking, the dataset itself is also just a random sample. Here, the statistics on the dataset are used as population estimates. This is statistically somewhat problematic. It would be more sound not to treat the dataset as the population, but to divide the dataset into all individuals covered by the subgroup and all individuals not covered by the subgroup: the complement of the subgroup. This calls for *two-sample* tests on subgroups (and their complements) whenever two distributions are compared. It is currently unclear whether treating the dataset in a proper way would result in better subgroups. Nevertheless, it is believed that, as long as the subgroups are viewed as being highly informative and are not used for statistical inferences, the current approach is not problematic. One benefit of this assumption is that the computation of statistics (and tests) can be done relatively easy, as opposed to when the subgroup and the dataset are treated properly from a statistical point of view. Furthermore, this approach enables subgroup discovery to produce rules that might not be statistically highly interesting, although they can be informative to the user.

One of the most important metrics used, is the size of either the subgroup and the population. These sizes are denoted by n_s and n_p respectively. Some standard statistical metrics used are the average and standard deviation. The estimated mean of the target values of the subgroup is denoted by $\mu_s = \frac{\sum_{i=1}^{n_s} t_i}{n_s}$, the estimated mean of the target values of the population is $\mu_p = \frac{\sum_{i=1}^{n_p} t_i}{n_p}$. Whenever the standard deviation is mentioned, the standard deviation from the estimated variance is meant, despite the use of the σ for these deviations. Although there are several estimators known, we chose to use the unbiased estimator of the variance for both the population and the subgroup target attribute standard deviations: $\hat{\sigma}_t^2 = \frac{1}{n_t - 1} \sum_{i=1}^{n} (t_i - \mu_t)$. The standard deviation is then just the square root of the variance estimator: $\sigma_t = \sqrt{\hat{\sigma}_t^2}$. Thus, σ_p and σ_s stand for the estimated standard deviations of the population target values and the subgroup target values.

$\mathbf{Rank_{complete}}$	$\mathbf{Rank_{partial}}$	**Target**	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
1	1.5	0.150	1	0	1	0	1	1	0	1	0	1
2	1.5	0.150	1	0	1	0	1	0	1	1	0	1
3	3	0.140	1	1	1	0	1	1	0	1	0	1
4	4.5	0.130	1	1	0	0	1	0	1	1	0	1
5	4.5	0.130	1	1	0	0	1	1	0	1	1	1
6	6	0.110	1	1	1	1	0	0	1	1	1	0
7	7	0.100	1	1	1	1	0	1	0	1	1	0
8	9	0.090	1	1	0	1	0	0	1	0	1	0
9	9	0.090	1	1	1	1	0	1	0	0	1	0
10	9	0.090	1	1	0	1	0	0	1	0	1	0
11	11.5	0.070	0	1	0	1	0	1	0	0	1	0
12	11.5	0.070	0	1	0	1	0	0	1	0	0	0
13	13.5	0.035	0	0	1	1	1	1	0	0	0	0
14	13.5	0.035	0	0	0	1	1	0	1	1	0	0
15	15	0.001	0	0	1	1	1	1	0	1	0	0
subgroup size			10	10	8	10	8	8	7	9	7	5

Table 7: Subgroups in a datase (including the auxiliary complete ranking)

In the case of quality measures for ordinal targets, some metrics need to be redefined. Here, the means of a subgroup or the population μ_s and μ_p are the means of the ranks, unless stated otherwise. Accordingly for the standard deviation, which becomes the estimated standard deviation of the ranks.

Some measures do not compare the distribution of the subgroup to the distribution of the population, but to the distribution of the complement of the subgroup: $c = p - s$. The size of the complement is thus $n_c = n_p - n_s$.

If needed, the dataset from Chapter 6, depicted here in Table 7, is used to gain a better understanding of quality measures.

7.2 Quality Measures for Regressional Subgroup Discovery

Most quality measures currently available and used for numeric target attributes are derived from statistical measures or tests, which are applicable to numeric attributes in general. Thus, the measures presented here are also capable of dealing with complete and partial rankings, although they are not specifically designed for ordinal numeric attributes.

7.2.1 Average

A relatively simple and effective quality measure is the average μ_s [10, 6] of a subgroup. Depending on the subgroup search objectives, a maximum of all averages or minimum of all averages is best. For instance, if the ranking on the original targets is used as target for subgroup discovery, then usually the top ranked individuals are considered best. In this case, the subgroup discovery algorithm using the average as quality measure should return subgroups with the lowest average on top (*minimization*). Given the list of subgroup target attribute values t_s with size n_s, then the mean is calculated as follows:

Definition 1 (Average) $\varphi_{avg}(s) = \frac{\sum_{i=1}^{n_s} t_i}{n_s}$

7.2.2 Mean Test

A more complex measure is the mean test. This measure was introduced by Klösgen [24], and adopted by Grosskreutz [20]. The latter has applied the mean test as a quality measure in subgroup discovery on numeric targets. The mean test is capable to compare the distribution of the target attribute in the subgroup to the distribution in the whole population, as opposed to the mean itself. Furthermore, it also takes the size of the subgroup into account.

Definition 2 (Mean test) *Given subgroup size n_s, subgroup mean μ_s and population mean μ_p, then $\varphi_{mt}(s)= \sqrt{n_s}(\mu_s - \mu_p)$*

7.2.3 Z-Score

The z-score [10, 6] is a metric that measures how many standard deviations an individual is away from the population mean. Here, we are not interested in the z-score of just one individual, but in the z-score of the whole subgroup, which is a set of individuals. The z-score for a group of individuals can be calculated by the standardized version of the z-score [38]:

Definition 3 (Standardized z-score) *Given subgroup mean μ_s, population mean μ_p, population standard deviation σ_p and subgroup size n_s, $\varphi_z(s) = \frac{\mu_s - \mu_p}{(\frac{\sigma_p}{\sqrt{n_s}})} = \frac{\sqrt{n_s}(\mu_s - \mu_p)}{\sigma_p}$*

The standardized z-score measures how far the mean of the subgroup is away from the mean of the population, in terms of standard deviations. The bigger the value for the z-score, the bigger the difference between the population and the subgroup. The z-score has a strong background in the normalization of data.

$\varphi_z(s)$ and $\varphi_{mt}(s)$ are strongly related in the sense that they are order equivalent: $\varphi_{mt}(s) \sim \varphi_z(s)$. Order equivalence is defined as follows:

Definition 4 (Order equivalence) *Two functions $\varphi_1(s)$ and $\varphi_2(s)$ are order equivalent $\varphi_1(s) \sim \varphi_2(s)$, iff $\varphi_1(s_1) > \varphi_1(s_2) \rightarrow \varphi_2(s_1) > \varphi_2(s_2) \wedge \varphi_1(s_1) = \varphi_1(s_2) \rightarrow \varphi_2(s_1) = \varphi_2(s_2)\ \forall s_1, s_2 \in s.$*

$\varphi_{mt}(s)$ and $\varphi_z(s)$ are order equivalent given $\varphi_z(s) = \frac{\varphi_{mt}(s)}{\sigma_p}$. Dividing a quality measure by any constant, such as σ_p, does not affect the internal order of subgroups, given the old and new quality measures. If $\varphi_z(s_1) > \varphi_z(s_2)$, then $\frac{\varphi_{mt}(s_1)}{\sigma_p} > \frac{\varphi_{mt}(s_2)}{\sigma_p} \equiv \varphi_{mt}(s_1) > \varphi_{mt}(s_2)$. Equivalently, if $\varphi_z(s_1) = \varphi_z(s_2)$ then $\frac{\varphi_{mt}(s_1)}{\sigma_p} = \frac{\varphi_{mt}(s_2)}{\sigma_p} \equiv \varphi_{mt}(s_1) = \varphi_{mt}(s_2)$. When two quality measures are order equivalent, the underlying order of subgroups is equal for both functions, although evaluation values may differ.

Although the value of the $\varphi_z(s)$ itself is already highly interesting, it is also interesting to see whether subgroups are *significant* and to what level. Furthermore, although the value returned by Safarii is not suitable to determine significance levels, it can be used to approximate the significance[1]. To obtain the significance level, the p-value can be looked up using the $\varphi_z(s)$-value. The p-value for a certain $\varphi_z(s)$-value can be found in the table of z-values of the normal distribution, which is depicted in Appendix C.1. For example, let's consider subgroup s_4 from table 7, with the original target. For this subgroup, the metrics needed for the z-score are $\mu_s \approx 0.069$, $\mu_p \approx 0.093$, $\sigma_p \approx 0.045$, and $n_s = 10$, given the raw target as target attribute. This results in $\varphi_z(s_4) = \frac{\sqrt{10}(0.069 - 0.093)}{0.045} \approx -1.687 \approx -1.69$. To look up the p-value, this number has to be chopped in two portions: 1.6 and 0.09. The p-value can be found at the intersection of the row of 1.6 with the column of 0.09. Thus, the p-value of s_4 is 0.9545. In other words, subgroup s_4 and the population are distributed differently with a confidence of 95%. A good rule of thumb for using the $\varphi_z(s)$ is that the further the value is away from 0, the more significant the subgroup is. The z-score is also one of the tests the SEGS tool to find enriched gene sets uses, in order to classify gene sets [38, 40]. For more information on the z-score, the reader is referred to any statistics handbook, such as [6, 10].

7.2.4 t Statistic

A somewhat different statistic than the z-score is the t statistic that is used in the Student's t test [10, 6]. The t statistic is much more accurate for smaller sample sizes, and thus more suited when subgroup sizes can or should be small. To obtain a higher accuracy, the statistic uses the subgroup

[1]For all quality measures directly derived from statistical tests, it holds that values returned by Safarii *can* be used to approximate the significance level, although the obtained level does not have the proper statistical validation.

deviation instead of the population deviation. This makes the t statistic also more sensitive to differences in variances in subgroups.

Definition 5 (t-Statistic) *Given subgroup mean μ_s, population mean μ_p, subgroup standard deviation σ_s and subgroup size n_s, then* $\varphi_t(s) = \frac{\mu_s - \mu_p}{(\frac{\sigma_s}{\sqrt{n_s}})} = \frac{\sqrt{n_s}(\mu_s - \mu_p)}{\sigma_s}$

Like in the case of the z-score, the higher the value for the t statistic (or lower, for a negative difference), the more significant the difference is, and the more interesting the subgroup is. Let's take a look at subgroup s_4 with the original target. It has the following metrics: $\mu_s \approx 0.069$, $\mu_p \approx 0.093$, $\sigma_s \approx 0.035$, and $n_s = 10$. Then, $\varphi_t(s_4) \approx \frac{\sqrt{10}(0.069-0.093)}{0.035} = -2.168$. To obtain the p-value for this $\varphi_t(s_4)$ value, one has to look up the p-value in the table of the t distribution, such as the one in Appendix C.2 [6, 10]. To do so, the degrees of freedom is needed. The degrees of freedom is the sample size (here: the subgroup size) minus one. For instance, for the example shown above, the degrees of freedom is $df = 10 - 1 = 9$. Then, with the $\varphi_t(s)$-value for subgroup s_4 and $df = 9$, the p-value lies between 0.95 and 0.975, but is more close to the latter. Subgroup s_4 therefore has an approximated significance level of at least 95%. For more information on the Student's t-test and the t statistic, the reader is referred to any statistics handbook, such as [6, 10].

7.2.5 Median χ^2 Statistic

The median χ^2 statistic [10, 6] does not define the difference in distributions through the mean of either the subgroup or the population, but uses the median of the population instead. The median is a more robust metric than the mean, since it is less sensitive to outliers in the data. Seen from a statistical point of view, no assumptions on the underlying distribution have to be met, it is thus a *nonparametric* test.

The test works as follows. The median χ^2 statistic takes the individuals in both the subgroup and the population, and divides them in whether the target attribute value of the individual lies above or below the population median med_p. If the target value of the individual is equal to the population median, it is grouped with the individuals whose value lie below the population median. A schematic view of the information needed by the median χ^2 statistic is given in table 8.

	Above med_p	**At or below** med_p
Subgroup	S_a	S_b
Population	P_a	P_b

Table 8: Counts of individuals for the median χ^2 statistic

In table 8, S_a and P_a represent the number of individuals in both the subgroup and the population whose target values lie above the population median. Accordingly, S_b and P_b are the numbers of records (individuals) whose target value is at or below the population median, for the subgroup and the population respectively. Using these counts, statistical tests can be used to tell whether the distribution of the subgroup is (significantly) different from the population distribution. One such test, although not the most sensitive one, is the χ^2 test. Despite the insensitivity of this test, it is easily implemented and gains sensitivity with sufficiently large record counts.

Definition 6 (Median χ^2 statistic) *Given subgroup and population counts S_a, S_b and P_a, P_b, then* $\varphi_{\chi^2}(s) = \frac{(S_a - P_a)^2}{P_a} + \frac{(S_b - P_b)^2}{P_b}$

Although the degrees of freedom *df* is not compulsory for subgroup discovery, it is needed if one wishes to get an approximation of the significance of the subgroup. The degrees of freedom here is solely dependent on the number of categories being compared, in this case whether the

value of an individual lies above or below the population median. Thus, there are only 2 categories here, resulting in a degrees of freedom $df = 2 - 1 = 1$.

Let's again look at subgroup s_4, using the original target as target attribute. The population median is $med_p = 0.09$, resulting in $P_a = 7$ and $P_b = 8$. The counts for s_4 are $S_a = 2$ and $S_b = 8$. Thus, after substitution in the formula for the median χ^2 statistic $\varphi_{\chi^2}(s_4) = \frac{(2-7)^2}{7} + \frac{(8-8)^2}{8} \approx 3.571$. If one wishes to retrieve an approximation of the significance of the subgroup, one has to look up the p-value in the table of the χ^2-distribution, such as the one in Appendix C.3. Looking up the value for $\varphi_{\chi^2}(s_4)$, gives a p-value which lies between 0.9 and 0.95, although it is more close to 0.95. Thus, the significance of subgroup s_4 is at least 90%.

Unfortunately, this measure does not make a difference between subgroups where the majority of individuals have a target value above the median or below the median. Therefore, when such subgroups should be treated as distinct subgroups, this measure is not the best to use. On the positive side, this measure can be used given any target type, numeric or ordinal. Still, as already mentioned, this measure is not very sensitive. It only works well if the counts of each cell in table 8 are at least 1.5, although a minimum of 5 is preferred [6, 10].

7.3 Quality Measures for Ordinal Subgroup Discovery

As explained previously, ordinal target attributes display a certain ordering, and usually the best individual has the top rank, i.e. 1. As stated earlier, there are two types of ranking, the complete and the partial ranking. A complete ranking is a ranking where all underlying target values are different. In the partial ranking, there exist ties between the target values of individuals. The first and second column of Table 7 give examples of a complete and partial ranking, remember that the complete ranking here is auxiliary.

From a statistical point of view, ordinal data is data for which it is not known from what kind of distribution the data originates. More specifically, it is assumed that such data does not even follow a distribution. Therefore, in the field of statistics, nonparametric tests are used to make inferences on ordinal data. Thus, the quality measures used for subgroup discovery on ordinal data are either based on nonparametric tests or inspired by them.

Nonparametric tests, and therefore the measures derived by and inspired on them, are less sensitive than their parametric counterparts, such as the t-test and z-score. This is due to the fact that for nonparametric statistics no assumptions on the underlying data are made, resulting in less specific information (and inferences) on the data at hand. Usually a drop in sensitivity is accompanied by an increase in robustness: due to the absence of specific information on the data, the inferences on the data should be more general to be powerful, resulting in more robust tests. Accordingly, the measures presented here are less sensitive, but more robust.

For all measures in this section and in order to use them in Safarii, it is assumed that the ranking of the target attribute is added to the dataset by the data analyst as a preprocessing step. Moreover, it is assumed that rankings are in ascending order, rendering the top ranked individuals to be the more desirable ones. Furthermore, the measures presented here only work on rankings, both partial and complete, unless stated otherwise.

7.3.1 AUC of ROC

The area under the Receiver Operating Characteristic (ROC) curve [21, 16] is traditionally a metric to compare the performance of classifiers. In such classification tasks, the target attribute variable is binary: there are only two class types considered, $class = 1$ and $class = 0$ [21]. Ordinal target attributes, however, do not separate individuals into classes 0 and l. The AUC of ROC can be modified in such a way that it can measure how interspersed the individuals of a subgroup are in the overall population. In other words, this measure is very useful in order to define the position of the subgroup individuals in the population and if they are grouped together or more spread out. To do so, the AUC of ROC divides the individuals into 'belonging to the subgroup' and 'not belonging to the subgroup (thus belonging to the complement)':

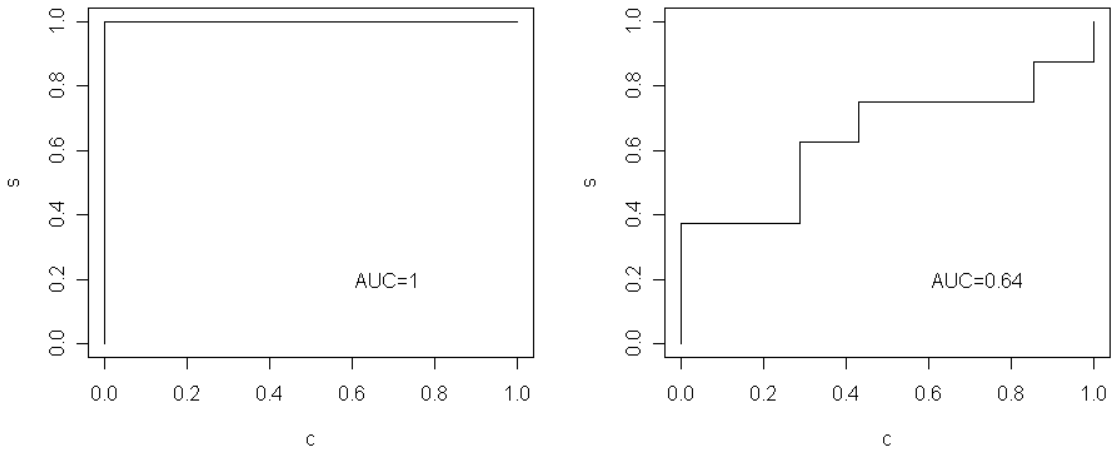


Figure 7: $\varphi_{roc}(s)$ scores for subgroup s_1 and s_3 respectively

Definition 7 (AUC of ROC for ordinal targets) *Given subgroup s, its complement c, the sum of ranks of the complement of the subgroup ς_c, subgroup size n_s and complement size n_c, then* $\varphi_{roc}(s) = \frac{\varsigma_c - \frac{n_c(n_c+1)}{2}}{n_c n_s}$

The $\varphi_{roc}(s)$ measure can only be used on complete rankings. The highest possible result is 1 for the best subgroup, whereas the worst subgroup will receive a 0. For a subgroup to return a 1, all individuals of the subgroup should be grouped together, and the first individual is the top ranked individual of the population. If the measure returns a 0, this means again that all individuals are grouped together, but now the individual with the worst rank also has the worst rank in the population ranking. All other values indicate that the individuals are either not closely packed and/or the best (or worst) individual is not present in the subgroup. The size of the subgroup does not affect the value of $\varphi_{roc}(s)$.

In short, this measure does the following. It starts on point $(0, 0)$ on the ROC curve, and for every individual it takes a $\frac{1}{n_s}$ step up (denoted by a $+$) if it is present in the subgroup, and a $\frac{1}{n_c}$ step to the right (denoted by a $-$) if it is not. Let's consider subgroups s_1 and s_3 from table 7. For subgroup s_1, there are 10 individuals in the subgroup, all grouped together, resulting in the following walking pattern: $(+,+,+,+,+,+,+,+,+,+,-,-,-,-,-)$. Subgroup s_3 shows a different walking pattern: $(+,+,+,-,-,+,+,-,+,-,-,-,+,-,+)$. The two walking patterns result in the following areas under the curve: $\varphi_{roc}(s_1) = \frac{65 - \frac{5(5+1)}{2}}{5 \cdot 10} = 1$, and $\varphi_{roc}(s_3) = \frac{64 - \frac{7(7+1)}{2}}{7 \cdot 8} \approx 0.64$. The walking patterns and the values of $\varphi_{roc}(s_1)$ and $\varphi_{roc}(s_3)$ are shown in figure 7.

7.3.2 Wilcoxon-Mann-Whitney Ranks statistic

The Wilcoxon-Mann-Whitney Ranks (*wmw*) statistic [10, 6], is derived from the nonparametric Wilcoxon-Mann-Whitney Ranks test. It has a strong relation to the z-score, since it calculates the difference of the means of the ranks through the z-statistic. Instead of comparing the population distribution directly with the subgroup distribution, the distribution of the subgroup is compared to the distribution of the complement of the subgroup. If both distributions are the same, i.e. the wmw statistic is near 0, then the population distribution is equal to the subgroup distribution. If the distribution of the subgroup and its complement are not the same, the individuals of the subgroup are differently dispersed throughout the population distribution than the individuals of the complement [6, 10].

Definition 8 (Wilcoxon-Mann-Whitney Ranks statistic) *Given subgroup size n_s and subgroup complement size n_c, sum of ranks of the subgroup ς_s, rank mean of the population $\mu_p = \frac{n_s(n_p+1)}{2}$ and rank deviation of the population $\sigma_p = \sqrt{\frac{n_s n_c(n_p+1)}{12}}$, then $\varphi_{wmw}(s) = \frac{\varsigma_s - \mu_p}{\sigma_p}$*

Like in the case of the $\varphi_z(s)$ and $\varphi_t(s)$ measures, the value returned by this metric can be either positive or negative. When the value is positive, the subgroup mean of ranks is larger than the subgroup complement mean (and thus the population mean), indicating that the individuals of the subgroup are concentrated among the bottom ranks. When this value is negative, the majority of the individuals of the subgroup are grouped near the top ranks. And similar to the $\varphi_z(s)$ and $\varphi_t(s)$, the further away the returned value is from 0, the more significant the found subgroup is. To check upon an approximation of the significance level, the p-value can be looked up in Appendix C.2.

Let's again look at an example, such as subgroup s_1 from table 6, with target attribute $\text{rank}_{\text{partial}}$. Here, $n_s = 10$, $n_c = 5$, $\varsigma_s = 55$, $\mu_p = 80$, and $\sigma_p \approx 8.16$. Then, $\varphi_{wmw}(s) = \frac{55-80}{8.16} \approx -3.06$, with $df = n_s - 9$. Given the Appendix C.2, the (approximated) p-value is at least $p = 0.99$.

7.3.3 Median MAD Metric

Apart from the tests described above, a new metric was developed. This new metric maximizes on the subgroup size and minimizes on the median and median absolute deviation, the *mad*. This test is strictly only applicable to both complete and partial rankings. The test does not compare the subgroup distribution to the population distribution, but just calculates a ratio for the subgroup size and the position of the individuals (cluster) in the subgroup.

Definition 9 (Median MAD metric) *Given subgroup median size n_s, subgroup median m_s and subgroup median absolute deviation mad_s, then $\varphi_{mmad}(s) = \frac{n_s}{2 \cdot m_s + mad_s}$*

The median absolute deviation [10] is defined as follows:

Definition 10 (Median Absolute Deviation (*mad*)) *Given the target attribute list $t_s = t_1, t_2, \ldots, t_k$ of the subgroup with median m_s, then $mad_s(t_s) = median(y)$, where $y = \{|t_1 - m_s|, |t_2 - m_s|, \ldots, |t_k - m_s|\}$*

As stated previously, the quality measures for ordinal targets are usually somehow derived from or inspired by nonparametric tests. One of the characteristics of such tests is that they are usually more robust to anomalies in the data. Apart from robust tests, the field of statistics also has some standard metrics which are more robust, metrics that are less sensitive to anomalies such as outliers. Two of such metrics are the median and the median absolute deviation, where the latter is similar to the standard deviation of the mean. One of the reasons to develop a whole new quality measure, is that there are not many quality measures currently available which sufficiently take the size of the subgroup into account, even though the coverage of a subgroup can be an important characteristic of the subgroup. Apart from that, it can be argued that for the sake of generality, the qualification of a subgroup should not suffer too much from a few anomalies in the data, especially if the subgroup is considerably large. All these considerations call for a different heuristic than the ones presented earlier. The new heuristic is specifically designed to give a higher qualification to larger subgroups, hence the factor n_s in the numerator. To make sure that subgroups where the majority of individuals (despite some anomalies) are highly ranked, are preferred over other subgroups, the median and median absolute deviation of the subgroup are calculated. Of course, the median and median absolute deviation should be as small as possible. The median is given a higher weight than the median standard deviation for obvious reasons, the requirement that the majority of the individuals should be among the top ranks is stronger than whether these individuals are too dispersed throughout the population. Part of the latter requirement is also met by the median itself. Hence, to minimize on the median and deviation of the median, they are grouped together in the denominator of the equation. One can wonder why the deviation is added to the median instead of multiplied by it. The reason for this is that it is

	φ_{avg}	φ_{mt}	φ_z	φ_t	φ_{χ^2}	φ_{roc}	φ_{wmw}	φ_{mmad}
Numeric/Ordinal targets	both	both	both	both	both	ordinal	ordinal	ordinal
Complete/Partial ranking	both	both	both	both	both	complete	both	both
Symmetric	no	yes	yes	yes	no	no	yes	no
Distribution information	s	s&p	s&p	s&p	s&p	s&c	s&c	s
p-value approximation	no	no	yes	yes	yes	no	yes	no

Table 9: Quality measures and their characteristics

relatively common to obtain a 0 for the median deviation. For instance, consider the situation in which a subgroup only contains four elements with the following ranks: $(2, 2, 2, 4)$. The median of this subgroup would then of course be 2. The vector for the absolute deviations would thus be $(0, 0, 0, 2)$, for which the median is 0. Due to the characteristics of the median and its deviation, it was chosen to add the median and the deviation instead of multiplying them, to avoid a division by 0.

All in all, the $\varphi_{mmad}(s)$ maximizes on the size of the subgroup, and minimizes on the median and deviation of the subgroup, hence showing a bias toward subgroups where the majority of the individuals have top ranks. Since this measure is completely new, it is important to get a feeling of the performance of this measure. Consider subgroups s_1 and s_2 from table 7 with $\text{rank}_{\text{partial}}$ as the target. Subgroups s_1 and s_2 have the following target values: $t_{s_1} = 1.5, 1.5, 3, 4.5, 4.5, 6, 7, 9, 9, 9$ and $t_{s_2} = 3, 4.5, 4.5, 6, 7, 9, 9, 9, 11.5, 11.5$. The medians of these subgroups are $m_{s_1} = 5.25$ and $m_{s_2} = 8$. Then, the deviations are: $y_{s_1} = 3.75, 3.75, 2.25, 0.75, 0.75, 0.75, 1.75, 3.75, 3.75, 3.75$ and $y_{s_2} = 5, 3.5, 3.5, 2, 1, 1, 1, 1, 3.5, 3.5$, resulting in $mad_{s_1} = 3$ and $mad_{s_2} = 2.75$ as median absolute deviation values. Both subgroups are of equal size: $n_{s_1} = n_{s_2} = 10$. The evaluation values for the subgroups are $\varphi_{mmad}(s_1) = \frac{10}{2 \cdot 5.25 + 3} = \approx 0.741$ and $\varphi_{mmad}(s_2) = \frac{10}{2 \cdot 8 + 2.75} \approx 0.533$, thus subgroup s_1 is the better one.

7.4 On Quality Intuitions and Quality Measures

In order to use the presented quality measures properly, it is important to get a good understanding of the characteristics of the quality measures. Table 9 lists the characteristics of the quality measures. This table shows what kind of targets the measures can deal with (ordinal or numeric). Also, whether a measure can deal with partial or complete rankings is shown here. Symmetry is a characteristic that needs a little more explanation. Symmetry means that the values returned by the measure are grouped around 0. Moreover, if the target distribution of the individuals in a subgroup is symmetric to the distribution of individuals in another subgroup, then the two subgroups would be qualified with the same value. Although the values would be symmetric, they are distinguishable by their sign: one of the subgroup values is positive, the other is negative. When the evaluation value of a symmetric measure is 0, the evaluated subgroup has the same target distribution as the population. The fourth characteristic is about the distribution of the subgroup, the population and/or the subgroup complement. It tells what kind of distribution information the quality measure uses to evaluate a subgroup. The fifth characteristic, the possibility to retrieve an approximation of the p-value, might seem a bit strange. It does not mean that the measure itself returns the (approximation of the) p-value, since Safarii can not calculate the approximation. Nevertheless, for the quality measures for which an approximation of the p-value can be made, this approximation can be looked up in distribution tables, such as the distribution tables given in Appendix C.

It is important to get a feeling of the performance of all quality measures, and to understand which intuitions are covered by which measures. The understanding of the quality measures helps to make an educated choice about which measures are suitable to use in a data mining task. The small exemplary dataset from table 7 is used for this purpose. All subgroups in this dataset are evaluated using the quality measures presented here. The results are depicted in table 11. As can be seen, the quality measures were used on all targets, depending on whether the measure can deal

	φ_{avg}	φ_{mt}	φ_z	φ_t	φ_{χ^2}	φ_{roc}	φ_{wmw}	φ_{mmad}
Configurations (max/min/abs)	max&min	all	all	all	max	max	all	max
Original target	ND	ND	ND	ND	ND	NA	NA	NA
Ranking (1 is best)	min	min	min	min	max	max	min	max
Ranking (max is best)	max	max	max	max	max	NA	max	NA

Table 10: Maximizing or minimizing on evaluation values

with the target type. The best evaluation values are in **bold**. Whether the best evaluation values of the subgroups are the maximum or minimum values, depends heavily on the target attribute and the search objective. Table 10 can be used to understand how to use the quality measures, i.e. whether the evaluation values should be maximized or minimized. NA means that evaluation is not possible for a given target, thus a definition of whether the measure should minimize or maximize is not available. ND stands for not defined, i.e. it is impossible to decide whether the measure should maximize or minimize without knowing the properties of the target attribute. For quality measures φ_{mt}, φ_z, φ_t and φ_{wmw}, it is also possible to take the absolute value. In this case, one is only interested in whether the individuals of a subgroup are differently distributed given their target values, compared to the distribution of the population target values. Thus, it is unimportant where the majority of the individuals of the subgroup lies, either left or right of the population mean.

Below, all quality measures are informally qualified given the intuitions from Chapter 6: Intuitions 1 (size), 3 (spread of individuals), 4 (cluster position), and 5 (distribution difference). Intuition 2 is left out, since this is the exact opposite of Intuition 1.

φ_{avg}: Only Intuition 4 holds for this measure, for obvious reasons. It can be argued that this measure favours smaller subgroups where the individuals are grouped together, to ensure that the mean of the subgroup sets the subgroup apart. Therefore, Intuitions 3 and 5 both hold partly. Consequently, Intuition 1 does not hold: the larger the subgroup, the more the mean is likely to move toward the population mean.

φ_{mt}: This measure is an improvement over φ_{avg}, in the sense that the distributions of the subgroup and the population are compared. Equivalently, Intuitions 4 and 3 are partially applicable. Intuitions 5 and 1 are completely applicable, since this measure maximizes both the size of the subgroup and the difference in target distributions to evaluate the subgroup.

φ_z: φ_z and φ_{mt} are order equivalent, as argued before. Hence, φ_z is not particularly an improvement over φ_{mt}, and the same intuitions are applicable here. The only improvement is that φ_z enables the user to make an approximation of the significance of a subgroup and subgroup evaluations are comparable irrespective of the target that is used. Like φ_{mt}, this measure is also available in other tools as well [38, 41].

φ_t: φ_t does take the deviation of the subgroup into account and is thus more sensitive to changes in the spread of a subgroup with the same size. However, this measure is more likely to favour smaller subgroups over larger onces, since smaller subgroups tend to have a smaller deviation. Look for instance at subgroups s_2 and s_3, given the $\text{rank}_{\text{partial}}$ as target attribute. φ_z qualifies s_3 as a better subgroup, although the individuals of this subgroup show a larger deviation. Subgroup s_2, in which the individuals are closely packed, is qualified as better. Intuitions 3, 4 and 5 all hold. Intuition 1 is not fully applicable to this measure.

φ_{χ^2}: φ_{χ^2} only measures to what extent the subgroup contains individuals which are unevenly distributed throughout the whole population. This measure does not differentiate between individuals being above or below the population median. The problem with this quality measure is, that it does not matter where the individuals reside. Thus, two subgroups with a different

	φ_{avg}	φ_{mt}	φ_{z}	φ_{t}	φ_{χ^2}	φ_{roc}	φ_{wmw}	φ_{mmad}
				target=Rank$_{complete}$				
s_1	5.5	-7.906	-1.768	-2.611	3.571	**1**	**-3.062**	0.741
s_2	7.5	-1.581	-0.354	-0.522	1.786	0.6	-0.612	0.571
s_3	7	-2.828	-0.632	-0.555	3.411	0.64	-0.926	0.471
s_4	10.5	7.906	1.768	2.611	3.125	0	3.062	0.426
s_5	7.125	-2.475	-0.553	-0.424	3.411	0.63	-0.81	0.667
s_6	8	0	0	0	3.286	0.5	0	0.4
s_7	8	0	0	0	4.286	0.5	0	0.35
s_8	6.333	-5.001	-1.118	-1	3.696	0.78	-1.768	**0.75**
s_9	8	0	0	0	4.286	0.5	0	0.389
s_{10}	**3**	**-11.18**	**-2.5**	**-7.072**	**8.125**	**1**	**-3.062**	0.714
				target=Rank$_{partial}$				
s_1	5.5	-7.906	-1.781	-2.66	5		**-3.062**	0.741
s_2	7.5	-1.581	-0.356	-0.532	2.2		-0.612	0.533
s_3	7.063	-2.65	-0.597	-0.512	3.4		-0.868	0.464
s_4	10.5	7.906	1.781	2.66	2.5		3.062	0.44
s_5	7.125	-2.475	-0.557	-0.425	3.3		-0.81	0.667
s_6	8.125	0.354	0.08	0.071	3.3		0.116	0.395
s_7	7.857	-0.378	-0.085	-0.091	4.3		-0.116	0.333
s_8	6.278	-5.166	-1.164	-1.056	2.7		-1.827	**0.783**
s_9	8	0	0	0	4.6		0	0.35
s_{10}	**3**	**-11.18**	**-2.518**	**-7.454**	**7.5**		**-3.062**	0.667
				target=Target$_{raw}$				
s_1	0.118	0.079	1.757	3.162	3.125			
s_2	0.102	0.028	0.632	1.138	1.696			
s_3	0.097	0.011	0.251	0.21	3.696			
s_4	0.069	-0.076	-1.687	-2.168	3.571			
s_5	0.096	0.008	0.189	0.137	3.696			
s_6	0.09	-0.008	-0.189	-0.163	3.286			
s_7	0.096	0.008	0.176	0.209	4.286			
s_8	0.105	0.036	0.8	0.679	4.5			
s_9	0.097	0.011	0.235	0.557	4.286			
s_{10}	**0.14**	**0.105**	**2.335**	**10.51**	**8.571**			

Table 11: Evaluation values on example database. Best evaluation values are in **bold**

	φ_{avg}	φ_{mt}	φ_{z}	φ_{t}	φ_{χ^2}	φ_{roc}	φ_{wmw}	φ_{mmad}
I1: Size	−−	++	++	+	−−	−−	++	++
I3: Spread of individuals	+	+	+	++	−−	++	+	+
I4: Cluster position	++	+	+	+	−−	++	+	+
I5: Distribution difference	+	++	++	++	++	−−	++	+

Table 12: Informal qualification of quality measures given intuitions

distribution of individuals can be qualified as equal. Look for instance at subgroups s_7 and s_9, given the original target. φ_{χ^2} qualifies them as being equal, since they have the same number of individuals whose values lie above the median and below (or at) the median. When looking at these subgroups, one can see that the deviation of the individuals in subgroup s_7 is larger than the deviation in subgroup s_9. Consequently, only Intuition 5 holds for this measure.

φ_{roc}: As was already mentioned at the presentation of this evaluation measure, this measure shows a big tendency toward topmost individuals in the subgroup. This is shown by the evaluation of subgroups s_1, s_{10}, and s_8, where these subgroups contain individuals mostly or exclusively from the top ranks. Looking at the evaluation values, ROC does not value bigger subgroups over smaller ones, something that becomes clear when looking at subgroup s_1, with size 10 and subgroup s_{10} with size 5, where both are evaluated equally. Due to the preference of having individuals in one block, especially when the block covers the top individuals, Intuitions 3 and 4, are accounted for. Unfortunately, this does not count for Intuitions 1 and 5.

φ_{wmw}: φ_{wmw} calculates the subgroup mean and deviation in such a way that it becomes highly dependent on the subgroup, and its complement. Furthermore, both mean and standard deviation are dependent on the sizes of the subgroup and the subgroup complement. All this ensures that Intuition 1 is covered by φ_{wmw}, whereas 3 is covered partly. When looking at the evaluation of the subgroups from the example dataset, this quality measure works fairly well. It is the only one in which subgroups s_1 and s_{10} tie.

φ_{mmad}: φ_{mmad} is especially designed to favour a bigger subgroup size over a small difference in median. Furthermore, since both the median and the MAD are used, this measure has a bias toward topmost individuals, with preferably a small deviation in the target distribution of the individuals. Thus, this metric covers intuitions 1, and 4 and 3. When looking at subgroups s_1 and s_8 for instance, it can be seen that both subgroups have almost the same size, 10 and 9 respectively. Due to the smaller number of individuals with top rankings in subgroup s_8, this subgroup gets assigned a smaller median and median absolute deviation. Thus, φ_{mmad} favours s_8 over s_1. Both subgroups s_1 and s_8 get a better qualification than subgroup s_{10}, since the size of subgroup s_{10} is very small compared to the other two subgroups. One has to note, however, that the median and mad metrics are robust metrics. Thus, φ_{mmad} accepts subgroups where there are a few 'bad' individuals, individuals which reside in the bottom of the population. Hence, although φ_{mmad} does take the variance of the target distribution of the individuals into account, it is insensitive to a small number of outliers.

Table 12 informally classifies the quality measures given the quality intuitions. The exemplary dataset shows that all measures for numeric targets are heavily in favour of subgroup s_{10}. Only the ordinal quality measures show a different picture, where subgroups s_1 and s_8 are also categorized as important subgroups. All in all, the choice upon quality measures for mining should be guided by the characteristics of the quality measures and their behaviour according to the intuitions as defined in Chapter 6. Since the choice upon quality measures also heavily depends on the data at hand and the research questions, this discussion is continued in Chapter 8.

8 Experiments & Results

As set out in previous chapters, the objective of this thesis study was to enhance subgroup discovery in order to be able to perform subgroup discovery on numeric and ordinal targets. The reason for this enhancement stems directly from the EET Pipeline project, although there are many more applications thinkable where ordinal or numeric subgroup discovery are very useful.

The second objective of this thesis study is aggregation. More precisely, the aggregation of genes with meta information on genes. Before aggregation, the genes are ranked according to their expression given a neuroblastoma target, such as $nb_{stage} = 4$ and $event = 1$. The ranking of the genes is then used for aggregation.

The background on the aggregation and the enhancement of subgroup discovery is given in previous chapters. Let's now put things in perspective and take a look at the behaviour and performance of the aggregation. Furthermore, it is important to take a closer look at the behaviour of the new quality measures with respect to the EET Pipeline data. First, the rankings used for aggregation are discussed, together with how they were produced. After that, the experiments done on aggregation are discussed.

8.1 Ranking the Genes

As described in Chapter 3, the EET Pipeline project provided us with four datasets, of which one contains clinical information and the other three contain information on gene expressions. Of the three genomic datasets, only the mRNA dataset was chosen for this thesis study, due to the ease with which the probes in this dataset could be mapped to genes. Two of the clinical attributes were classified as being good targets to search for interesting genes considering neuroblastoma. One of them is the *status:* $event = 1$. If an event has taken place, the patient has had a relapse of the tumour, or is deceased. The other attribute is the *stage* of the tumour: $stage = 4$. Stage 4 tumours are most severe, and patients diagnosed with this type of neuroblastoma are most likely to suffer from a relapse or death [7, 32, 9, 43].

The rankings are made using the core probeset of the mRNA data. The raw mRNA data is acquired by measuring the expression of a large number of genetic probes, where each probe covers only part of a gene, or in some cases, parts of more than one gene. This raw data can then be compounded into a smaller datafile, the core probeset. In the core probeset, each probe covers at least one gene, sometimes more than one. Instead of covering only a small part of a gene (as is the case in the single probeset), the core probeset covers complete genes. The core probeset data, in other words the compounded raw data, was made available by the research group in Ghent, Belgium.

Although the raw dataset is usable for mining, it was chosen not to use the raw dataset for further mining steps. This is due to some issues that accompany this dataset. For instance, it is the goal to map the ranked probes to genes, but it is difficult to decide how. Before ranking? After ranking? When done before ranking, the same compounding step is performed as described above, and is thus obsolete. When the mapping is done after ranking, a new evaluation value has to be chosen to replace the values from the individual probes. A good option is to take the median (or the mean), but this approach can result in values such that genes are no longer properly distinguishable from each other. This is undesirable, since the second data mining step, the aggregation, relies heavily on the evaluation values for a good performance. Thus, we chose to mine the core probeset with targets $event = 1$ and $stage = 4$ subsequently, using the Safarii tool. An additional ranking was received from the research group from the Jozef Stefan Institute [22], Ljubljana, Slovenia. This ranking was made using $event = 1$ as the target.

8.2 Mining Meta Information

The gene rankings and meta information are aggregated (see Chapter 3), in order to find interesting genes and additional information on genes for neuroblastoma. Three types of experiments on aggregation are performed here. The first compares the meta information domains. The layout

of the experiment and the results are described in Section 8.2.1. Secondly, the performance of each quality measure is monitored. This experiment is done on only one ranking, using multiple targets, depending on the quality measure at hand. Further in-depth information and results are presented in Section 8.2.2. Lastly, two mining tools are compared, namely Safarii (using regressional subgroup discovery) and SEGS [38, 39], which stands for Search for Enriched Gene Sets tool and was developed at the Jozef Stefan Institute in Ljubljana, Slovenia. The details of this experiment can be found in Section 8.2.3.

8.2.1 Comparison of Knowledge Domains

The first experiment compares the various domains of meta information, i.e. GO/KEGG terms (GO in short), gene-to-gene interactions (abbreviated by gene2gene or G2G), protein families (PFAM in short), and gene locations (abbreviated by LOC). The idea is to compare these domains in order to see how the domains perform considering subgroup size and evaluation values. Moreover, the differences between the three rankings are investigated, together with the differences in targets. For this experiment, only one quality measure was used, the z-score, φ_z. This quality measure was chosen since it has a good background in subgroup discovery, either as the z-score itself or through the order equivalent mean test [24, 20, 38]. Furthermore, preliminary tests have shown that φ_z performs reasonably well in terms of subgroup size. Although the sizes of subgroups tend to be on the larger side, the sizes still vary, as opposed to for instance when φ_t is used as the quality measure, which has a preference to smaller subgroups (see the experiments in Section 8.2.2 for further detail). φ_z can be used on both target types (numeric or ordinal), whilst preserving the ability to compare evaluation values, irrespective of the target type. Lastly, φ_z can be used to give an *indication* of the p-value of a subgroup, although the reader has to remember that subgroup discovery and accompanying quality measures (in Safarii) are not designed for significance testing, but rather for exploratory data analysis.

For the experiment, all three rankings are used, i.e. the Safarii $event = 1$ and $stage = 4$ rankings, and the IJS $event = 1$ ranking. Each ranking is aggregated with GO/KEGG terms, gene2gene interactions, PFAMs and gene locations subsequently. For aggregation, two targets were used subsequently: the novelty (measure for the IJS ranking), which is a numeric target, and the partial ranking, which is of course ordinal.

The top-25 patterns for all aggregations can be found in Appendix A. The progression of the evaluation values over subgroup ids is shown in Figures 8(a), 8(b) and 9(a). Tables 13 and 14 show the averages and standard deviations for the φ_z evaluation values and subgroup sizes for each ranking, domain and target. The evaluation values show that the best performing domains are GO/KEGG terms and gene2gene interactions, followed by the gene locations and protein families. In some cases, a domain might have a better start, but the subgroup evaluations devaluate more rapidly. This happens for instance with the gene2gene domain compared to the gene locations domain, for the Safarii $stage = 4$ ranking, both with target novelty. Here, the gene locations domain performs much better until subgroup 17, where the gene2gene domain starts performing better. Furthermore, the gene2gene domain devaluates less rapidly and thus has a lower standard deviation: an average standard deviation of 1.53 for the GO/KEGG terms domain, as opposed to an average standard deviation of 0.68 for the gene2gene domain. Although the GO/KEGG terms domain performs better on the first 25 subgroups, the gene2gene domain probably performs better down the list, since the evaluation values of this domain degrades with a smaller factor than the values of the GO/KEGG domain. Figures 8(a), 8(b) and 9(a) and Table 13 indicate this, since the evaluation values of the gene2gene and GO/KEGG terms domains converge to each other toward subgroups further down the lists. It also has to be noted that for all domains the drop in evaluation values stabilizes when moving down the list of subgroups.

On choosing a target for aggregation, the experiments suggest that the novelty is the best to use for the Safarii rankings. This is mainly due to the high evaluation values at the start, the data suggest that the partial rank results in a smaller drop (i.e. smaller standard deviation) of the evaluation values (see for instance the gene2gene and GO/KEGG terms domains for both targets for the Safarii $event = 1$ ranking, in Figure 8(b)). For the IJS ranking, however, the partial

ranking is definitely the best choice for the target, since this target results in better evaluation values and a smaller drop of evaluation values over subgroup ids.

Apart from looking at the evaluation values themselves, it is also interesting to consider the sizes of the subgroups that the various domains return. The sizes of the found subgroups are depicted in Figures 10(a), 10(b) and 9(b). The averages and standard deviations of the top-25 subgroups are depicted in Table 14. In general, the GO/KEGG terms and gene locations domains return relatively large subgroups, as compared to the gene2gene and PFAM domains. Only the subgroup sizes of the gene2gene domain portray a relatively small standard deviation, as opposed to the subgroup sizes of the other domains. Here, the standard deviations are always bigger than the average of the subgroup sizes itself. This indicates that all domains except the gene2gene domain can return both large and small subgroups, whereas subgroups are generally small when the gene2gene domain is used. This suggests that genes do not interact with a very large number of other genes, at least not the most interesting genes for neuroblastoma. Please note that the φ_z quality measure itself has a small bias toward larger subgroups.

It is also important to investigate what kind of subgroups the aggregations produce, to be more precise, to look at the conditions of the subgroups. The subgroups and their conditions can be found in Appendix A.

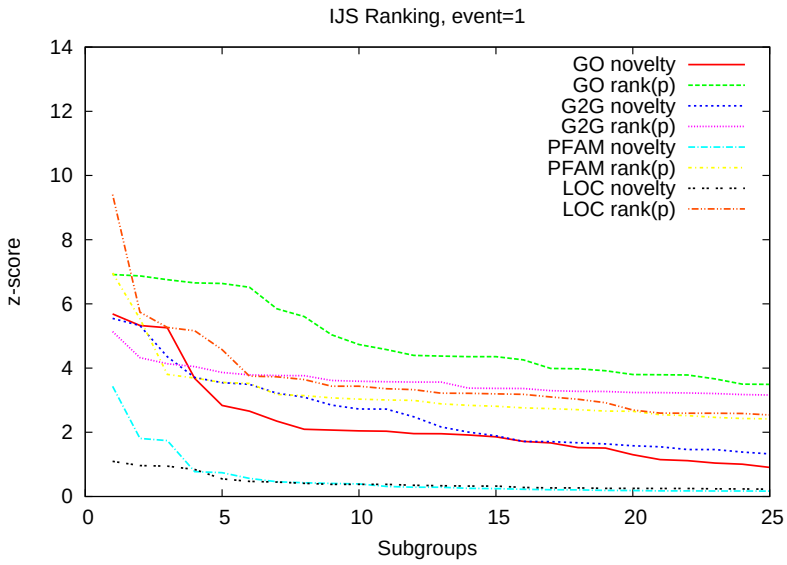
GO/KEGG Terms First of all, the absolute top of the subgroups are generally equal, except for the top of the subgroups given the IJS $event = 1$ ranking, with the measure as the target. GO/KEGG terms that are of high interest are, amongst others, DNA replication, cell cycle, mitosis and nucleus. The GO/KEGG terms generated from the different rankings do not differ much. Only the IJS $event = 1$ ranking, with $t = measure$ shows many different GO/KEGG terms at the top, such as 3-chloroallyl aldehyde dehydrogenase activity and chromatin assembly complex. At least the GO/KEGG terms DNA replication, cell cycle and DNA replication initiation can be found in the literature [9].

Gene-to-gene Interactions Again, as is the case for the GO/KEGG terms, the topmost genes returned by the aggregation are more or less equal, independent of ranking and target type. High scoring interacting genes are, amongst others, BIRC5, RAD51, CDC2, CDC7, CDC6, E2F4, MCM2, MCM3 and BRCA1, of which at least BIRC5 can be found in the literature [9]. Moreover, genes from the MCM group and the CDC group are found both in the literature and in the aggregation results.

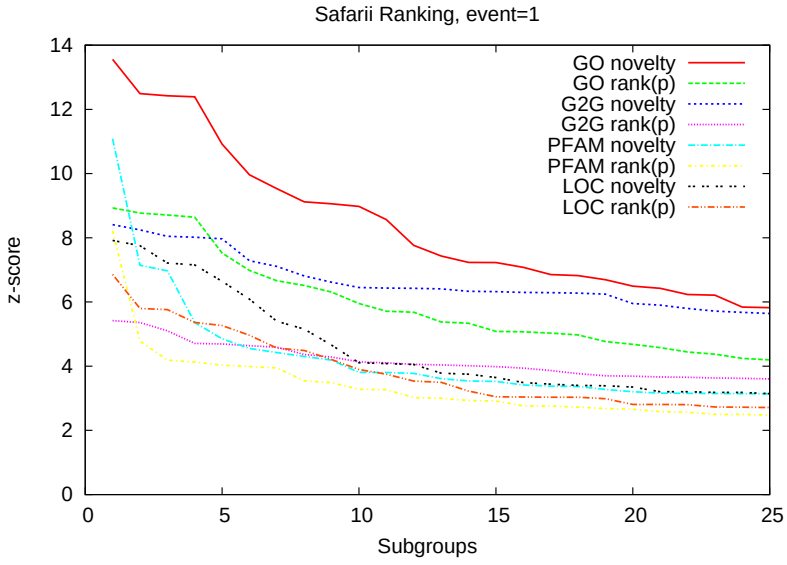
Protein Families Analogously to the found genes in the gene2gene aggregation, the PFAM aggregation returns MCM, Rad51 and E2F_TDP as high scoring protein families, amongst other high scoring protein families such as Cadherin, Kinesin and Histone. None of the families are found in literature, this is mainly due to the fact that the protein families knowledge domain has not been used in other studies yet.

Gene Locations When searching through the literature, it becomes evident that there are many parts of the chromosome that play a role in the development of neuroblastoma, either through the deletion or the gain of (parts of) the chromosome. However, whether (a part of) a chromosome is deleted or gained, is no longer clear during aggregation. Nonetheless, many chromosome regions found in literature are also found in our analysis, with the following difference. In literature, the chromosome regions are usually only specified until the chromosome arm, whereas in our analysis, the chromosome regions can be specified much further. Compare for instance region 17p, as found in [9]. In our analysis, chromosome 17 and region 17p11.2 are both presented as interesting considering an unfavorable nb stage ($stage = 4$). Other interesting chromosomes and regions that were found are chromosomes X, 6 and 2 and regions Xq28, 6p22.1.

Concluding Remarks The results of this experiment show that there are several differences in the knowledge domains. First and foremost, the GO/KEGG terms and the gene2gene interactions perform well, together with the gene locations domain. Other research on neuroblastoma and meta information has focused on these domains [9, 7, 38] as well. Given our results, the protein family domain also seems highly interesting, performing reasonably well compared to the other domains. Furthermore, this domain can give information on groups of genes that might be interesting, such as the MCM family, harboring genes MCM2, MCM3, MCM7 and so forth. Also, the analysis presented here shows that an automated informative analysis on neuroblastoma data and meta information can give information that can also be found in other research (think of the found GO/KEGG terms, the gene2gene interactions and gene locations). This gives rise to the idea that the information presented here that is *not* found in the literature, is still of value to domain experts. In other words, an automated analysis and aggregation of the neuroblastoma data can aid domain experts in their search of relevant processes considering neuroblastoma.

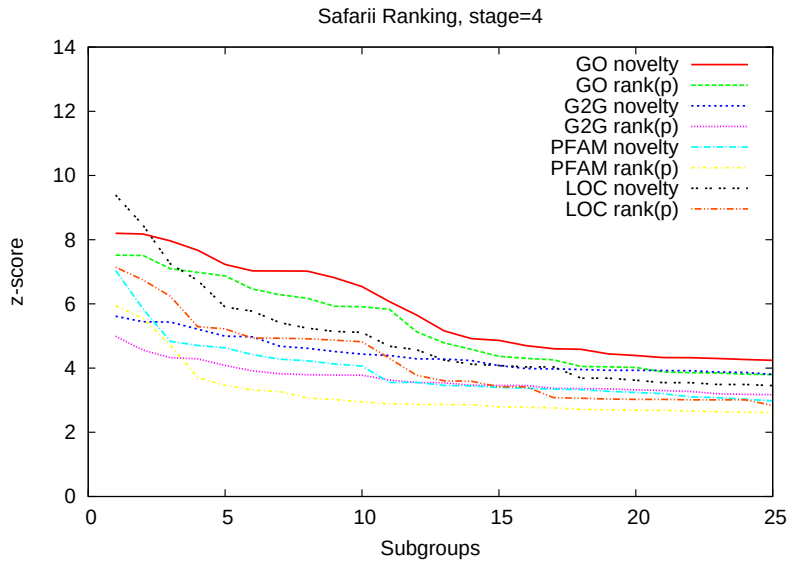


(a)

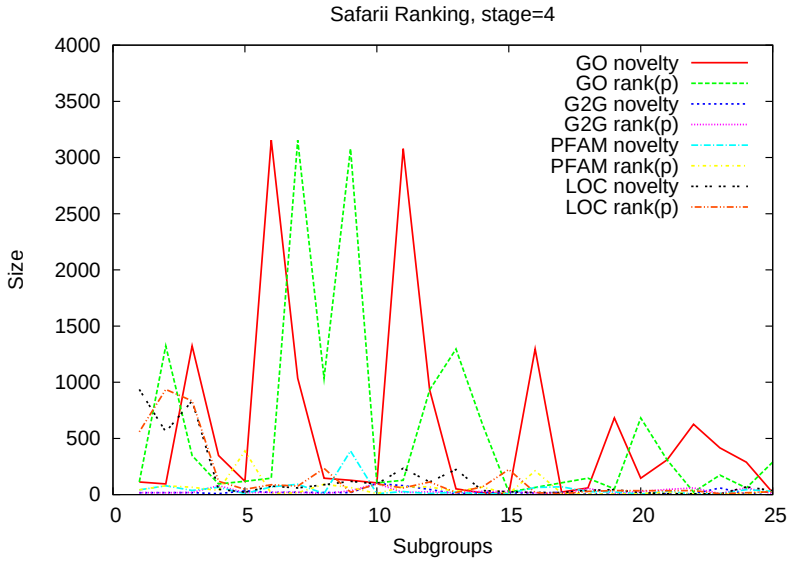


(b)

Figure 8: φ_z evaluation values for $event = 1$ rankings

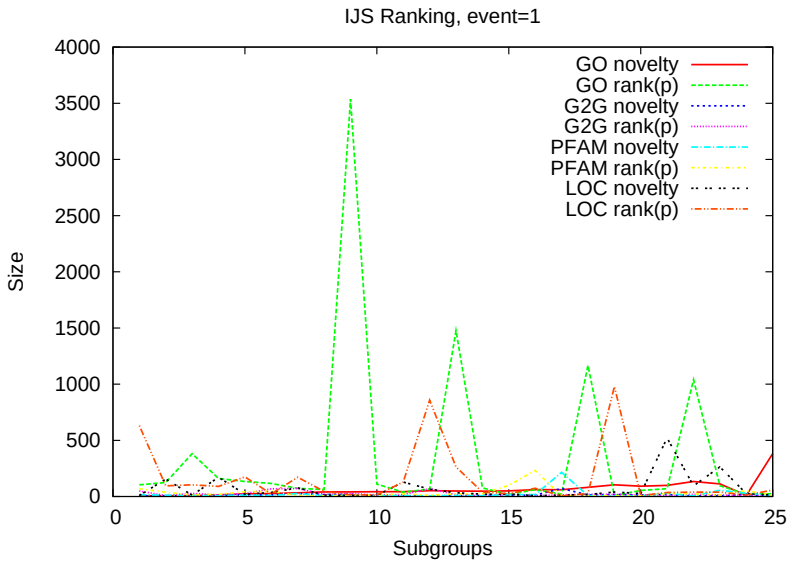


(a)

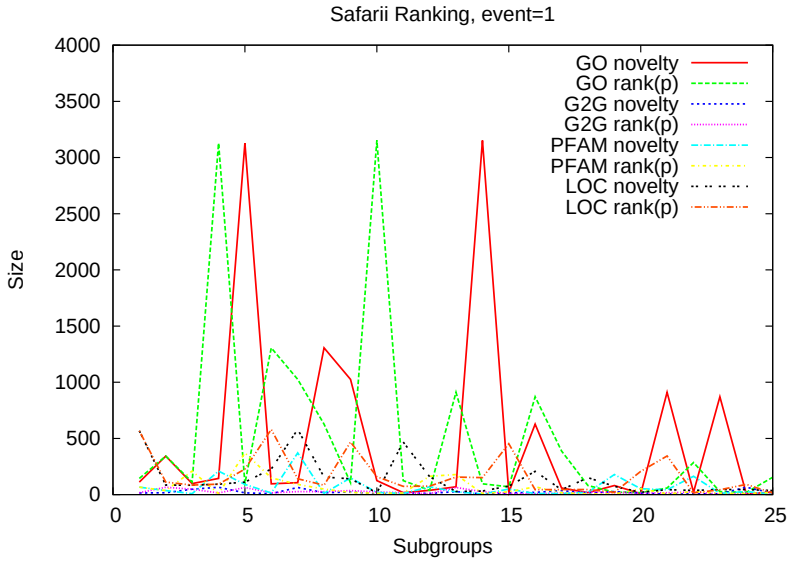


(b)

Figure 9: φ_z evaluation values and subgroup sizes for $stage = 4$ ranking



(a)



(b)

Figure 10: Subgroup sizes for $event = 1$ rankings

$\varphi_z(s)$ **values**	$\mu_{\text{GO/KEGG}}$	$\sigma_{\text{GO/KEGG}}$	$\mu_{\text{gene2gene}}$	$\sigma_{\text{gene2gene}}$	μ_{PFAM}	σ_{PFAM}	μ_{loc}	σ_{loc}	μ_μ	μ_σ
IJS *event* $= 1$, **t=measure**	2.27	1.34	2.58	1.21	0.56	0.74	0.43	0.25	1.46	0.89
IJS *event* $= 1$, **t=rank$_{\text{partial}}$**	4.87	1.21	3.6	0.45	3.2	1.02	3.69	1.47	3.84	1.04
Safarii *event* $= 1$, **t=novelty**	8.44	2.33	6.67	0.85	4.29	1.79	4.58	1.6	5.99	1.64
Safarii *event* $= 1$, **t=rank$_{\text{partial}}$**	5.94	1.53	4.19	0.54	3.4	1.2	3.88	1.19	4.35	1.12
Safarii *stage* $= 4$, **t=novelty**	5.78	1.43	4.41	0.56	3.9	0.96	4.91	1.6	4.75	1.14
Safarii *stage* $= 4$, **t=rank$_{\text{partial}}$**	5.25	1.32	3.68	0.47	3.21	0.89	4.17	1.27	4.08	0.99
μ	5.43	1.53	4.19	0.68	3.09	1.1	3.61	1.23		

Table 13: Averages and standard deviations of φ_z evaluation values, for all rankings and targets

subgroup sizes	$\mu_{\text{GO/KEGG}}$	$\sigma_{\text{GO/KEGG}}$	$\mu_{\text{gene2gene}}$	$\sigma_{\text{gene2gene}}$	μ_{PFAM}	σ_{PFAM}	μ_{loc}	σ_{loc}	μ_μ	μ_σ
IJS *event* $= 1$, **t=measure**	64.64	74.7	13.92	13.22	21.6	41.84	71.84	113.31	43	60.77
IJS *event* $= 1$, **t=rank$_{\text{partial}}$**	361.44	767.73	22.8	19.76	32.96	47.15	158	262.92	143.8	274.39
Safarii *event* $= 1$, **t=novelty**	495.88	878.5	22.04	18.8	65.92	87	140.92	161.68	181.19	286.5
Safarii *event* $= 1$, **t=rank$_{\text{partial}}$**	527.84	865.04	23.76	17.78	64.72	87.55	171.64	172.59	196.99	285.74
Safarii *stage* $= 4$, **t=novelty**	580.32	861.59	23.4	23.86	46.8	75.74	148.8	250.16	199.83	302.84
Safarii *stage* $= 4$, **t=rank$_{\text{partial}}$**	576.6	861.09	29.56	24	54.08	82.88	149.24	249.87	202.37	304.46
μ	434.45	718.11	22.58	19.57	47.68	70.36	140.07	201.76		

Table 14: Averages and standard deviations of subgroup sizes, for all rankings and targets

Safarii $event = 1$ **ranking, target=novelty**						
normalized φ **values**	**d=3**	$\sigma_{\mathbf{d=3}}$	**d=4**	$\sigma_{\mathbf{d=4}}$	μ	μ_σ
φ_{avg}	0.89	0.05	0.95	0.02	0.92	0.03
φ_{mt} **and** φ_z	0.67	0.14	0.86	0.06	0.77	0.1
φ_t	0.59	0.14	0.66	0.11	0.62	0.13
φ_{χ^2}	1	0	1	0	1	0
μ	0.76	0.09	0.87	0.05		
Safarii $event = 1$ **ranking, target=rank**$_{\mathbf{partial}}$						
normalized φ **values**	**d=3**	$\sigma_{\mathbf{d=3}}$	**d=4**	$\sigma_{\mathbf{d=4}}$	μ	μ_σ
φ_{avg}	0.36	0.2	0.36	0.15	0.36	0.17
φ_{mt} **and** φ_z	0.72	0.14	0.86	0.07	0.79	0.11
φ_t	0.42	0.22	0.38	0.22	0.40	0.22
φ_{χ^2}	1	0	1	0	1	0
φ_{roc}	0.93	0.03	0.98	0.01	0.96	0.02
φ_{wmw}	0.69	0.14	0.85	0.07	0.77	0.11
φ_{mmad}	0.42	0.24	0.46	0.22	0.44	0.23
μ	0.66	0.14	0.72	0.1		

Table 15: Averages and standard deviations of normalized evaluation values, for Safarii $event = 1$ ranking, all targets and all quality measures

8.2.2 Performance of Quality Measures

Of course, φ_z is not the only quality measure available. Moreover, the different quality measures that are present in Safarii portray different preferences considering the subgroups that are returned. Therefore, the second experiment compares the quality measures on one ranking, given both target types. The IJS ranking performed not as good as the two Safarii rankings, thus, this ranking was decided against. Both Safarii rankings would have been a good choice, but mainly because the performance on the topmost subgroups of the $event = 1$ ranking is quite good and better than the performance of the $stage = 4$ ranking, the $event = 1$ ranking was chosen (see Section 8.2.1 for further details).

The experiment is set up as follows. For the Safarii $event = 1$ ranking, aggregation was done with $target = novelty$ using all regressional quality measures, i.e. φ_{avg}, φ_{mt}, φ_z, φ_t and φ_{χ^2}, and using all meta information. Different search depths were also chosen. Search depth here denotes how many domains (database tables) Safarii is allowed to combine in its search for interesting patterns. Depth $d = 3$ renders primarily subgroups with only one condition in the pattern. Depth $d = 4$ also returns patterns with more conditions in the pattern, usually two. The latter variant can give interesting combinations of meta information. When two conditions are combined in one pattern, this is denoted by $\wedge$. Furthermore, the ranking was also aggregated using the partial rank as the target, for all quality measures, including the ordinal ones: φ_{avg}, φ_{mt}, φ_z, φ_t and φ_{χ^2}, φ_{roc}, φ_{wmw} and φ_{mmad}. Please note that the φ_{roc} quality measure can not be used on partial ranks, thus, only for this quality measure, a complete rank was produced, deciding upon ties arbitrarily. Again, this aggregation is performed on two search depths, $d = 3$ and $d = 4$.

Figures 11, 12, 13, 14 and Tables 15 and 16 show the results of this experiment. Full results for the top-25 subgroups can be found in Appendix B. Due to the characteristics of the different quality measures and evaluation values, the evaluation values are also *normalized*, to enable a quality measure comparison. For normalization, the best scoring value, which is the value of the first found subgroup, is set to be the maximum, and gets assigned 1. All other evaluation values are divided by this maximum. This ensures that all evaluation values obtain a value between 0 and 1, where 1 is the new maximum evaluation value. If a subgroup is evaluated to 0, it also gets assigned 0 when normalized. The normalization gives us the evaluation *trends* of the quality measures, and also enables us to compare the behaviour of the quality measures. The normalized values in Figures 11, 12 and Table 15 show that building slightly more complex patterns stabilizes the devaluation of the quality measures. Not only that, the original evaluation values also show that the performance of the quality measures is better with $d = 4$. The original evaluation values can be found in Appendix B.

Safarii $event = 1$ **ranking, target=novelty**						
subgroup sizes	**d=3**	$\sigma_{\mathbf{d=3}}$	**d=4**	$\sigma_{\mathbf{d=4}}$	μ	μ_σ
φ_{avg}	7.68	3.79	5.48	1.08	6.58	2.44
φ_{mt} **and** φ_z	429.8	875.81	284.56	630.34	357.18	753.07
φ_t	479.4	877.58	208.68	649.46	344.04	763.52
φ_{χ^2}	5	0	5	0	5	0
μ	270.34	526.6	157.66	382.24		
Safarii $event = 1$ **ranking, target=rank**$_{\mathbf{partial}}$						
subgroup sizes	**d=3**	$\sigma_{\mathbf{d=3}}$	**d=4**	$\sigma_{\mathbf{d=4}}$	μ	μ_σ
φ_{avg}	6.72	2.78	5.48	1.12	6.1	1.95
φ_{mt} **and** φ_z	525.88	856.66	332.72	661.11	429.3	758.89
φ_t	156.92	623.33	5.84	1.07	81.38	312.2
φ_{χ^2}	5	0	5	0	5	0
φ_{roc}	6.72	2.78	5.44	1.08	6.08	1.93
φ_{wmw}	536.36	852.16	280.48	631.85	408.42	742.01
φ_{mmad}	1400.48	799.84	1524.12	761.95	1462.3	780.9
μ	395.50	499.28	311.48	339.91		

Table 16: Averages and standard deviations of subgroup sizes, for Safarii $event = 1$ ranking, all targets and all quality measures

φ_{χ^2} seems to be the best performing measure, but unluckily, this measure only produces very small subgroups (all of size 5), and the top-25 consists solely of protein families. The data shows that φ_{mt}, φ_z and φ_{wmw} behave similarily compared to each other, considering the progression of the evaluation values, and also considering subgroup sizes. This is very logical, since φ_{mt} and φ_z are order equivalent. φ_{wmw}, on the other hand, is not strictly order equivalent, but does calculate the z-statistic, which is also calculated by the φ_z.

When looking at Figure 12, it becomes apparent that φ_{avg}, φ_t and φ_{mmad} are the worst performing measures, although φ_{mmad} still performs better than φ_{avg} and sometimes even better than φ_t (especially in the case of more complex patterns, with depth $d = 4$). However, when looking at the sizes of the subgroups in Figures 13 and 14, φ_{mmad} performs very well on large subgroup sizes: it is the only quality measure that steadily finds large subgroups. Not surprisingly, φ_{χ^2}, φ_{avg}, φ_{roc} and occasionally φ_t perform worse on the subgroup size, they all show a bias toward smaller subgroups. Furthermore, φ_{mt}, φ_z and φ_{wmw} do not seem to have a preference in either very large or very small subgroups, although the tendency in this data is toward larger subgroups. Despite this tendency, the variance in subgroup size is high.

Choosing a Measure How to choose a quality measure then? This all depends on the objective of the researcher, although trying out at least two different types of quality measures is good and can be very informative. When several quality measures are tried, it is obviously best to try measures that do not belong to the same performance group, so to say. For instance, trying φ_{mt}, φ_z and φ_{wmw} in one go is rather uncalled for if one wishes to obtain different patterns. In such a case it is better to try out φ_z together with for instance the φ_{roc} and φ_{mmad}. Then, how to choose one quality measure from a set of likewise performing measures? This depends on the targets at hand and if a ranking can be made. Of course, when no ranking can be produced, the choice is very limited. If a choice has to be made between φ_{mt} and φ_z, φ_z should be probably favoured over φ_{mt}. φ_z and φ_{mt} perform equally, due to their order equivalence. The scores for φ_z can always be easily compared with one another and is also easily understandable, due to the statistical background. The trouble with φ_z though, is that the evaluation values can be easily misused for significance testing. Furthermore, φ_z is computationally a little more complex, but this complexity is in most cases no issue. The reasoning for φ_{wmw} is equivalent to the reasoning for φ_z, although one has to keep in mind that φ_{wmw} can only be used on a ranking. The φ_{avg}, φ_t, φ_{χ^2} and φ_{roc} measures should only be considered when the subgroup size should be reasonably small, with a probability to fairly large subgroups, for instance when using φ_t. There are no real reasons to favour one over the other, although it can be said that φ_{χ^2} has such a preference toward small subgroups, that

the patterns returned are not extremely interesting, especially when a multitude of domains are used in aggregation. If a ranking can be made, φ_{roc} seems a better choice than φ_{avg}. They both produce almost identical patterns, but φ_{roc} has a smaller slope of quality devaluation. Then, last but not least, φ_{mmad}. This measure is the measure to use when one wishes to obtain really big subgroups, containing patterns that are very generic.

Choosing a Target Of course, not only the quality measures are important to the results of the data mining exercise. The target which is used for the data mining is also very important. How to choose a target depends first of all on what kind of information the analyst seeks. If an ordinal target is chosen, or when the numeric target is in essence also ordinal, then both numeric and ordinal measures can be used. For this, a small experiment can be done, using for instance two quality measures on the raw numeric target and the ranking. The results from our experiment show us that, no matter what quality measure is chosen, *all* quality measures show the same preference to some target. In our case, the measures work better on the novelty, and the novelty was a proper target to begin with. One of the reasons for this behaviour can be that the novelty can show a more 'rough' pattern in value assignment, whereas the ranks for the individuals increase evenly. In other words, the novelty can make differences between individuals more extreme (or, in some cases, less extreme), whereas these differences are approximately equal for the rankings. In some cases, an ordinal numeric target can not be properly used for data mining. In such a case, the complete or partial ranking should always be used.

On Search Depths Our experiment shows that the search depth influences the performance of the quality measures heavily, either by higher subgroup qualities or by subgroup sizes. This is very logical. In theory, deepening the search (allowing more domains to be combined, even one domain being combined with itself) always ensures that the performance of the quality measures becomes better. When the search is performed too deep, subgroup discovery tends to overfit the data with its patterns. Overfitting [5, 37] is a hazard of data mining in general, and subgroup discovery is no exception. The main problem with overfitting lies in the generality of the patterns. Once the patterns overfit the data, they are no longer generally applicable and their informative value becomes questionable. Thus, deepening the search should only be done with great caution.

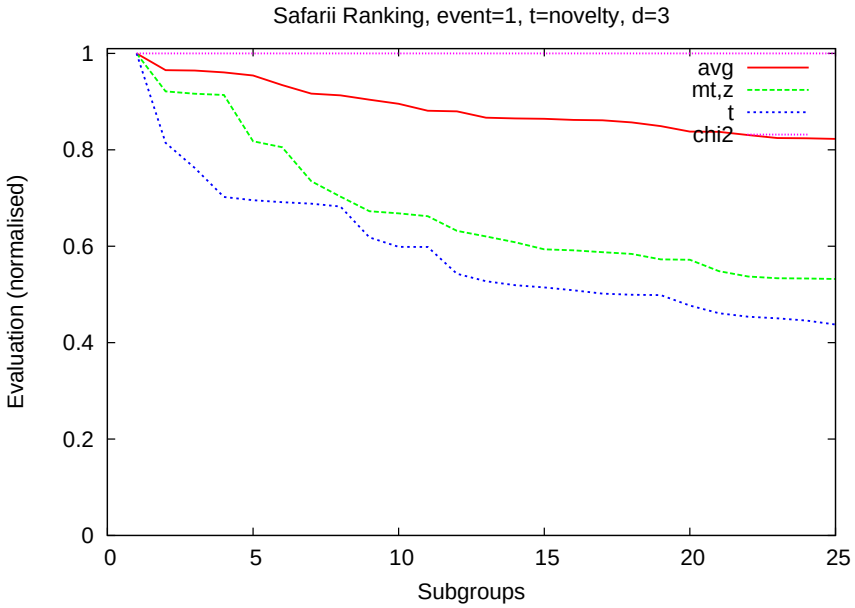
Patterns Found It is also interesting and important to consider the differences in patterns returned by the different aggregation methods. The first thing to note is that the bigger the subgroup, the more likely it becomes that the pattern holds one or more GO/KEGG terms or gene locations. Furthermore, some patterns that scored high in the previous experiment (Section 8.2.1) also score high in the more extensive aggregations, such as GO terms cell cycle, nucleus and mitosis, or even combinations of these GO terms (i.e. cell cycle $\wedge$ nucleus). Some patterns pop up regularly, such as PFAM=Histone, sometimes together with chromosome region 6p22.1, or PFAM=MCM, coupled with different interacting genes of this family: MCM6. Also, genes from the CDC range occur generally, such as CDC7 together with MCM6 or several GO/KEGG terms. Genes of the CDC range also often occur together with genes from the CDK range: interacting genes CDK3 and CDC2 are found on a regular basis together. The largest subgroups, subgroups covering a multitude of differentially expressed genes, can be found using conditions such as GO terms nucleus, membrane, protein binding and metal ion binding, or on chromosomes (or regions) such as chromosomes 1, 11 and 19.

Concluding Remarks First of all, this experiment supports the finding that the novelty is the best target to use for our data. Which quality measure to use, depends on the objective of the researcher. If, for instance, subgroup sizes have to be very large, φ_{mmad} is the best option. Furthermore, if a domain expert prefers smaller subgroup sizes or has no specific preference, other quality measures are better. All in all, all quality measures perform reasonably well on the data, the patterns have reasonably high scores. Also, the patterns returned by the different

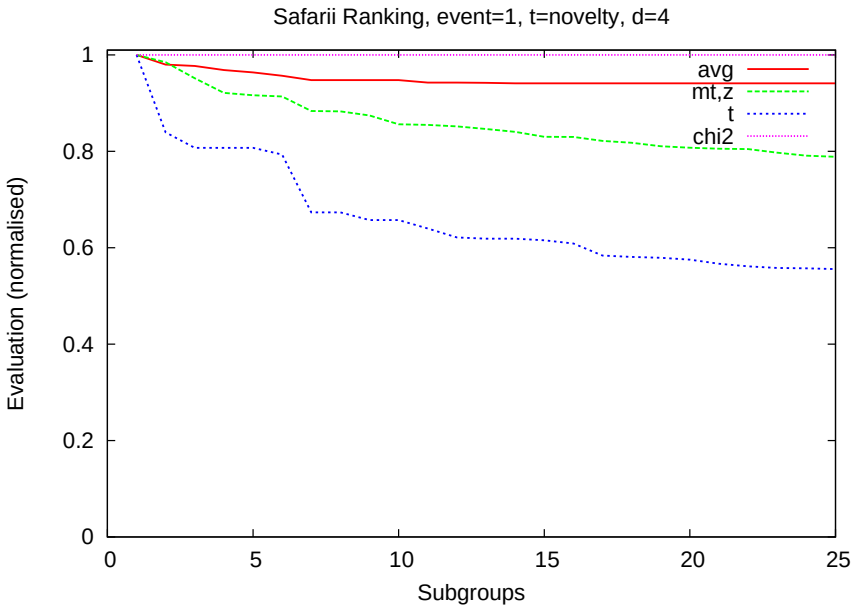
quality measures seem highly interesting, considering that some of the patterns can be found in the literature.

8.2.3 Safarii vs. SEGS

SEGS, like Safarii, is a multi-relational subgroup discovery tool. However, SEGS can do relational subgroup discovery solely on gene rankings. Furthermore, SEGS is only capable to aggregate with gene-to-gene interactions and GO terms. Thus, to make the comparison between Safarii and SEGS honest, rankings are mined with a smaller set of meta information in Safarii, namely only gene2gene interactions and GO terms. Unfortunately, the gene2gene interactions are not available as such in SEGS, SEGS tries to couple the interacting genes with GO terms, whereas Safarii does not. Furthermore, SEGS uses several (complicated) quality measures for subgroup discovery, of which only the Z-Score was available to Safarii as φ_z. Additionally, SEGS also calculates the p-values for each found subgroup, whereas Safarii is not able to calculate p-values.

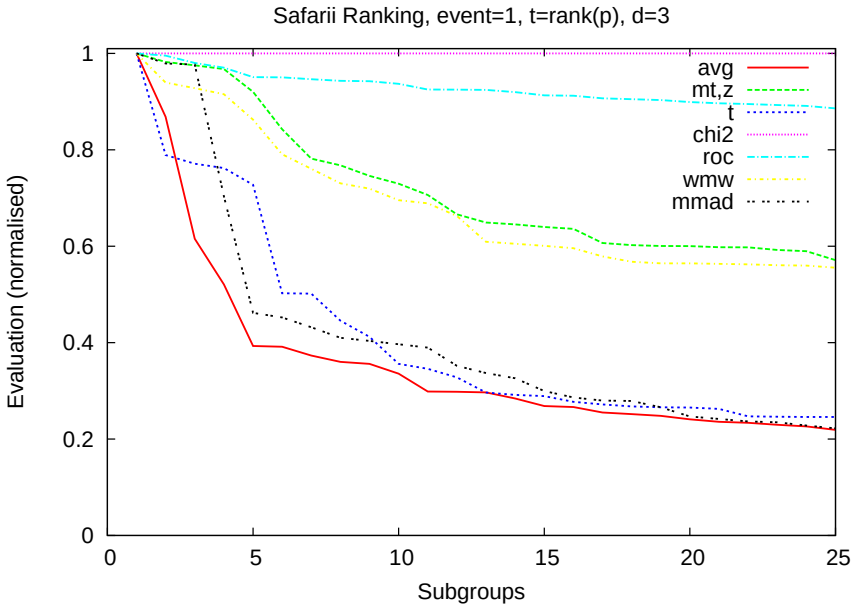


(a)

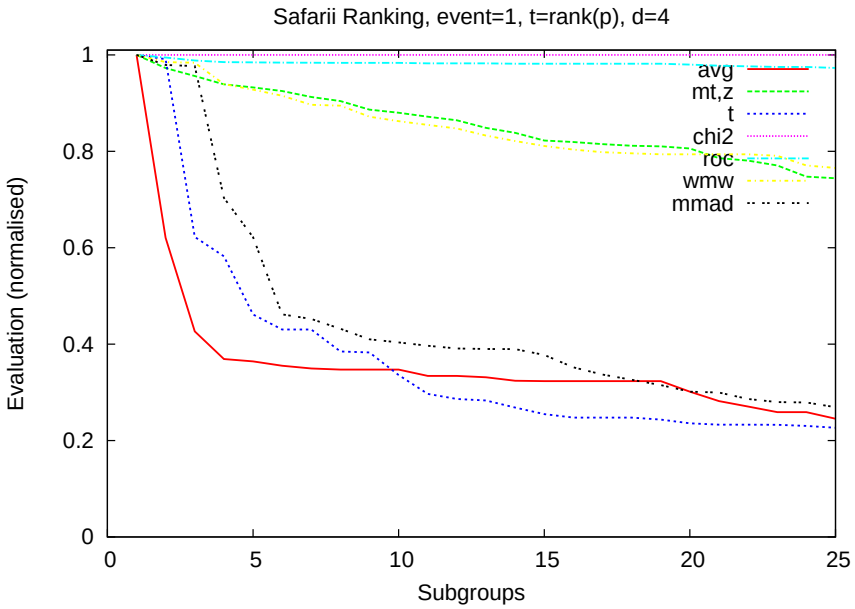


(b)

Figure 11: Quality measures evaluation values for RSD, normalized

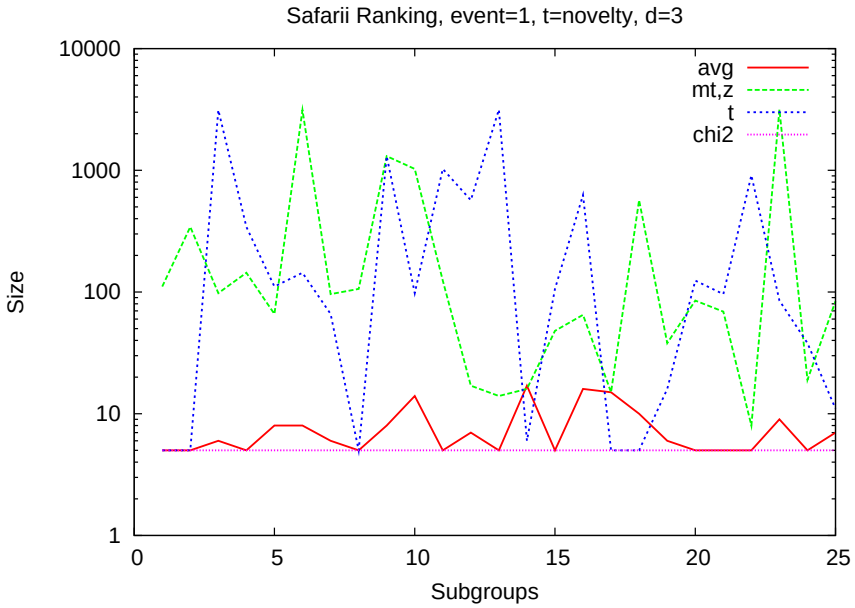


(a)

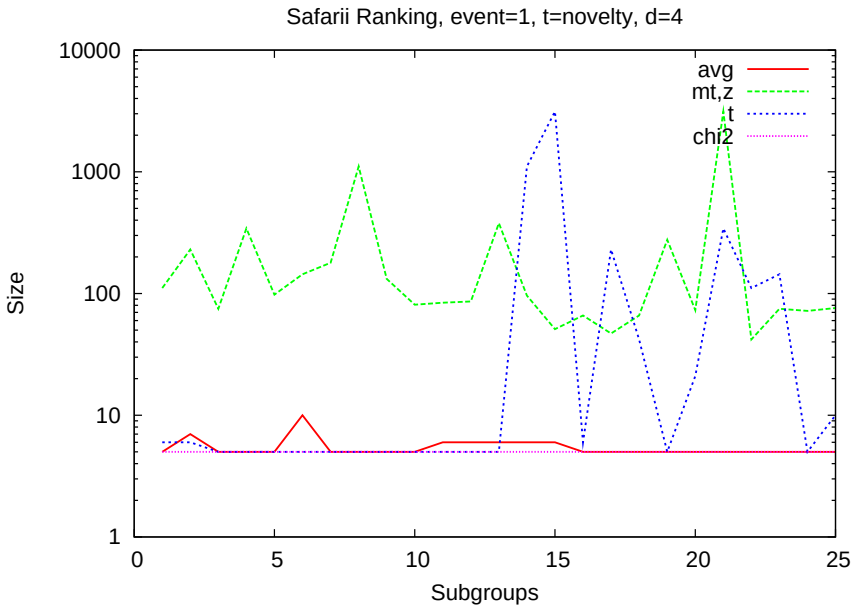


(b)

Figure 12: Quality measures evaluation values for OSD, normalized

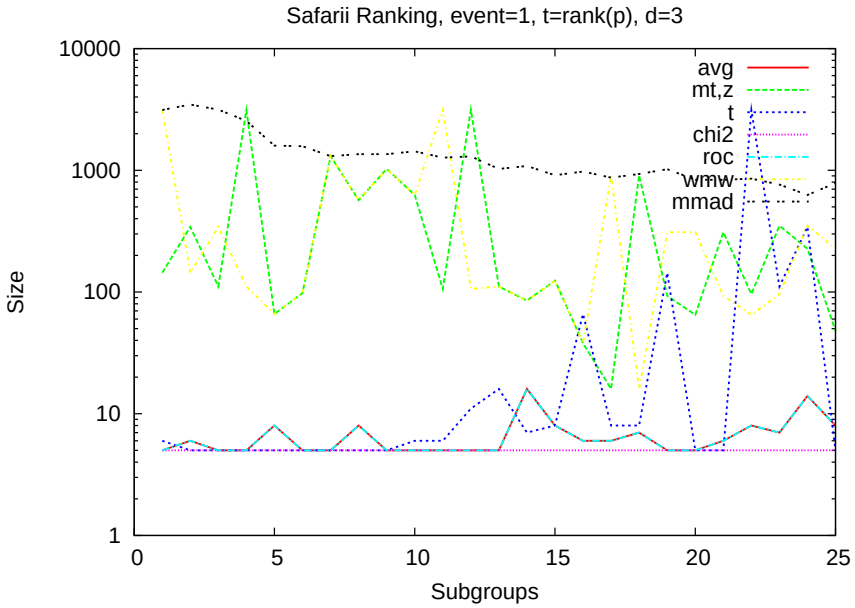


(a)

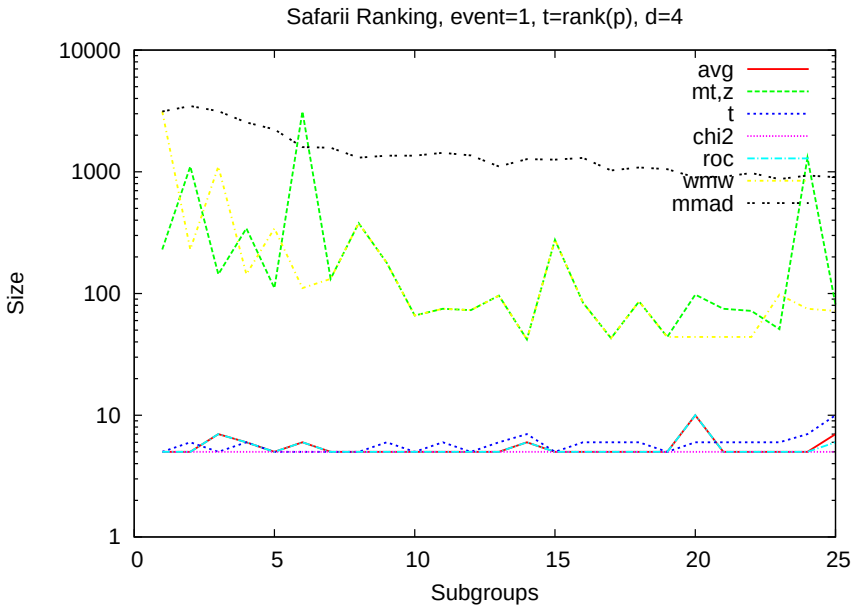


(b)

Figure 13: Subgroup sizes for RSD



(a)



(b)

Figure 14: Subgroup sizes for OSD

pattern	**size**	**Z-Score**
nucleus ∧ MutLalpha complex binding	85	12.607
nucleus ∧ mismatch repair complex binding	94	11.937
cell development ∧ intracellular organelle part ∧ sequence-specific DNA binding	90	11.802
response to DNA damage stimulus ∧ nuclear chromosome part	50	11.482
regulation of transcription ∧ DNA-dependent ∧ chromosome ∧ pericentric region	94	11.447
cell development ∧ nucleus ∧ chromatin binding	106	11.184
mismatch repair complex binding	107	11.175
cell differentiation ∧ intracellular organelle part ∧ sequence-specific DNA binding	102	11.033
nucleoplasm ∧ protein localization ∧ intracellular transport	106	10.776
DNA metabolic process ∧ nucleus ∧ MutLalpha complex binding	59	10.612
nuclear part ∧ DNA-dependent DNA replication	120	10.496
DNA metabolic process ∧ nucleus ∧ DNA-directed DNA polymerase activity	70	10.487
nucleoplasm part ∧ transport	111	10.35
response to stress ∧ nuclear chromosome part	62	10.292
cellular component organization and biogenesis ∧ nuclear part ∧ sequence-specific DNA binding	85	10.292
nucleus ∧ base-excision repair ∧ gap-filling	67	10.269
nuclear part ∧ in utero embryonic development	111	10.1
response to DNA damage stimulus ∧ nuclear chromosome	64	10.074
cell development ∧ intracellular organelle part ∧ transcription cofactor activity	127	10.069
nuclear lumen ∧ intracellular transport	121	10.063
cell differentiation ∧ nucleus ∧ chromatin binding	130	9.993
nuclear lumen ∧ protein localization	122	9.965
DNA metabolic process ∧ cyclin-dependent protein kinase holoenzyme complex	74	9.965
cell development ∧ chromatin binding	132	9.939
cell development ∧ nuclear part ∧ DNA binding	134	9.912
...	...	...

Table 17: Patterns found by SEGS, IJS ranking, target $event = 1$

pattern	**size**	φ_z
gene2gene = CBX5 ∧ GO:0005634: nucleus	26	18.845
gene2gene = CBX5	27	18.480
GO:0006260: DNA replication ∧ GO:0006355: regulation of transcription, DNA-dependent	22	17.962
GO:0006260: DNA replication ∧ GO:0006350: transcription	22	17.962
GO:0008152: metabolic process ∧ KEGG:00310: Lysine degradation	21	17.787
KEGG:00280: Valine, leucine ∧ isoleucine degradation ∧ KEGG:00071: Fatty acid metabolism	20	17.547
GO:0016491: oxidoreductase activity ∧ KEGG:00330: Arginine ∧ proline metabolism	20	17.525
GO:0008152: metabolic process ∧ KEGG:00380: Tryptophan metabolism	22	17.362
GO:0007049: cell cycle ∧ GO:0006260: DNA replication	22	17.147
GO:0016491: oxidoreductase activity ∧ KEGG:00380: Tryptophan metabolism	21	17.116
KEGG:00280: Valine, leucine ∧ isoleucine degradation ∧ KEGG:00650: Butanoate metabolism	21	17.093
GO:0008152: metabolic process ∧ KEGG:00640: Propanoate metabolism	21	17.064
GO:0016491: oxidoreductase activity ∧ KEGG:00620: Pyruvate metabolism	22	16.891
GO:0005739: mitochondrion ∧ KEGG:00650: Butanoate metabolism	22	16.787
GO:0008152: metabolic process ∧ KEGG:00010: Glycolysis/Gluconeogenesis	22	16.76
KEGG:00310: Lysine degradation ∧ KEGG:00650: Butanoate metabolism	22	16.698
GO:0006281: DNA repair ∧ gene2gene = PCNA	20	16.654
gene2gene = TCERG1	20	16.355
KEGG:00410: beta-Alanine metabolism	23	16.304
GO:0016491: oxidoreductase activity ∧ KEGG:00120: Bile acid biosynthesis	23	16.283
GO:0005739: mitochondrion ∧ KEGG:00071: Fatty acid metabolism	25	16.274
KEGG:00010: Glycolysis/Gluconeogenesis ∧ KEGG:00620: Pyruvate metabolism	24	16.154
GO:0008152: metabolic process ∧ KEGG:00280: Valine, leucine ∧ isoleucine degradation	24	15.951
GO:0016491: oxidoreductase activity ∧ KEGG:00010: Glycolysis/Gluconeogenesis	25	15.799
GO:0008152: metabolic process ∧ KEGG:00071: Fatty acid metabolism	27	15.629
...	...	...

Table 18: Patterns found by Safarii, IJS ranking, target $event = 1$

pattern	**size**	**Z-Score**
nuclear part ∧ Cell cycle ∧ regulation of progression through mitotic cell cycle	21	12.702
DNA binding ∧ Cell cycle ∧ DNA replication initiation	21	12.344
nuclear part ∧ Cell cycle ∧ traversing start control point of mitotic cell cycle	20	12.339
intracellular organelle part ∧ Cell cycle ∧ regulation of progression through mitotic cell cycle	26	12.242
DNA binding ∧ Cell cycle ∧ traversing start control point of mitotic cell cycle	21	12.156
nuclear part ∧ Cell Growth and Death ∧ regulation of progression through mitotic cell cycle	23	12.115
nuclear part ∧ Cell cycle ∧ protein serine/threonine kinase activity	23	12.114
intracellular organelle part ∧ Cell cycle ∧ traversing start control point of mitotic cell cycle	24	12.069
DNA binding ∧ Cell cycle ∧ regulation of progression through mitotic cell cycle	22	12.052
nuclear part ∧ Cell cycle ∧ protein amino acid phosphorylation	23	12.009
nuclear part ∧ Cell cycle ∧ protein serine/threonine kinase activity ∧ phosphate metabolic process	24	11.848
nuclear part ∧ Cell cycle ∧ phosphorylation	24	11.789
nuclear part ∧ Cell cycle ∧ DNA-dependent DNA replication	22	11.787
DNA binding ∧ Cell Growth and Death ∧ regulation of progression through mitotic cell cycle	23	11.776
DNA binding ∧ Cell cycle ∧ G1 phase of mitotic cell cycle	23	11.591
intracellular organelle part ∧ Cell Growth and Death ∧ traversing start control point of mitotic cell cycle	26	11.576
nuclear part ∧ Cell Growth and Death ∧ DNA-dependent DNA replication	23	11.517
nuclear part ∧ Cell cycle ∧ sequence-specific DNA binding	23	11.516

Table 19: Patterns found by SEGS, Safarii ranking, target $event = 1$

pattern	**size**	φ_z
GO:0005792: microsome ∧ KEGG:00150: Androgen ∧ estrogen metabolism	23	20.961
GO:0008152: metabolic process ∧ KEGG:00150: Androgen ∧ estrogen	26	19.686
GO:0008152: metabolic process ∧ GO:0005792: microsome	32	17.785
KEGG:00150: Androgen ∧ estrogen metabolism	49	14.493
GO:0005515: protein binding ∧ GO:0006260: DNA replication	47	11.854
GO:0008152: metabolic process ∧ GO:0016491: oxidoreductase activity	96	10.493
GO:0005634: nucleus ∧ GO:0006260: DNA replication	81	10.412
GO:0006260: DNA replication	98	10.071
GO:0005792: microsome	113	9.35
GO:0003677: DNA binding ∧ GO:0006260: DNA replication	40	8.908
GO:0005515: protein binding ∧ KEGG:04110: Cell cycle	66	8.834
GO:0005634: nucleus ∧ KEGG:04110: Cell cycle	76	8.237
KEGG:04110: Cell cycle	96	7.448
GO:0007049: cell cycle ∧ GO:0006350: transcription	51	7.319
GO:0007049: cell cycle ∧ GO:0006355: regulation of transcription, DNA-dependent	57	6.138
GO:0005634: nucleus ∧ GO:0008283: cell proliferation	82	5.712
GO:0006281: DNA repair ∧ GO:0006355: regulation of transcription, DNA-dependent	21	5.558
GO:0006260: DNA replication ∧ GO:0006355: regulation of transcription, DNA-dependent	20	5.542
GO:0006260: DNA replication ∧ GO:0006350: transcription	20	5.542
GO:0016491: oxidoreductase activity	331	5.43
GO:0005634: nucleus ∧ GO:0004519: endonuclease activity	22	5.419
KEGG:04110: Cell cycle ∧ GO:0006350: transcription	21	5.387
GO:0005730: nucleolus ∧ GO:0003723: RNA binding	24	5.256
GO:0005634: nucleus ∧ gene2gene = PCNA	58	5.211
GO:0006281: DNA repair ∧ GO:0006350: transcription	24	5.188
GO:0005634: nucleus ∧ GO:0007049: cell cycle	230	5.107
...	...	...

Table 20: Patterns found by Safarii, Safarii ranking, target $event = 1$

For this experiment, two rankings are used and compared, namely the IJS $event = 1$ ranking and the Safarii $event = 1$ ranking. Tables 17 to 20 show the results of this experiment. In the tables, a $\wedge$ stands for AND, thus $\wedge$ denotes that terms or interactions are combined together in the pattern. From the top GO terms (and for SEGS, indirectly interactions), there are not many terms that overlap. However, some (partial) patterns occur using both rankings, and both tools. These partial patterns are for instance DNA binding, nucleus and DNA repair. Although the results are a bit saddening, it does not necessarily mean that either SEGS or Safarii is doing poorly. One of the main reasons why the two tools present us with different patterns, is the way they work. The underlying algorithms performing the subgroup discovery are quite different, this is likely to be a reason for the mismatch between results.

Concluding Remarks All in all this small experiment suggests that both mining methods work reasonably well. However, when a choice has to be made between one of the two tools, Safarii comes out as the better one of the two. This is mainly because Safarii is generic, and it is thus possible to incorporate more meta information or even tweak the meta information such that it has the best representation. The drawback of Safarii over SEGS is that SEGS also calculates p-values, thus giving information on the significance of patterns. On the other hand, subgroup discovery is mainly used for descriptive purposes, for giving informative patterns, not significant ones. In the case of gathering informative patterns, Safarii does a much better job, since it is not restricted by rendering only significant patterns, but is able to render any pattern that is classified as interesting by the quality measure used. However, it is probably best to have the best of both worlds: easily adding and tweaking meta information for aggregation and the *possibility* of significance testing per pattern.

9 Conclusions and Future Work

The main objective of this thesis was to enhance the subgroup discovery algorithm in such a way that it could deal with numeric and ordinal target types. Furthermore, the goal was to apply the new features of subgroup discovery to biological data. More specifically, the goal was to perform subgroup discovery on a ranking of genes.

For this thesis, the genes are ranked according to their differential expression considering neuroblastoma, a tumour found in children. Not only this, the ranking of genes has been aggregated with several other knowledge domains to be able to capture existing and possibly new relations considering the causality of neuroblastoma. The new domains used here are the GO/KEGG terms and gene to gene interactions, which are domains that have been used earlier in automated searches to interesting genetic patterns. Furthermore, gene locations and protein families have also been used here, both have shown their added value in our experiments.

Several new quality measures were implemented and tested, and all measures performed well on the task set to them: find interesting patterns regarding neuroblastoma. Not all measures provided us with the same patterns. Regarding the specific characteristics of the quality measures, they returned patterns that fitted their characteristics and also fitted intuitions (wishes) on subgroups that a user might have. For instance, some quality measures enable us to find fairly large subgroups (the φ_{mmad} quality measure), whereas others return patterns for which the target attribute distribution is very different from the population target attribute distribution (φ_z, φ_t). Thus, not only is Safarii capable of performing subgroup discovery given numeric and ordinal targets, the user also has the ability to choose what kind of patterns Safarii will return. Still, the topic of subgroup discovery on numeric and ordinal targets is relatively new, and not much research has been done. Thus, much work still has to be done on this topic.

The performed experiments show that the quality measures can return known patterns of neuroblastoma, such as the importance of several genetic locations (chromosomes 1, 6, 17, X) or chromosome regions (6p22.1, 17p11.2) or GO terms (cell cycle, DNA replication). This at least validates that automated data mining can find interesting patterns. Due to this conclusion, it seems logical that the found patterns with no background in the literature, are also highly informative and thus can aid researchers in their research on neuroblastoma. One of the benefits of using data mining as an aid, is that mining data itself can be performed at a relatively low cost. Moreover, with data mining, cross references can be easily made using different meta information sources. For the future, it would be highly interesting to add even more meta data to the current data set, preferably data that can render patterns which are not easy to find without automated mining.

Acknowledgements

Many people have helped me in some way throughout my thesis research, some of which I would like to thank specifically. First and foremost, I would like to thank Arno Knobbe for his guidance, patience and understanding throughout my thesis research. Furthermore, I am thankful for the nice and constructive talks I had with Ad Feelders, and for him reviewing my thesis. Also, I would like to thank the research group from the Department of Pediatrics and medical genetics at Ghent University, Belgium. I especially want to express my gratitude toward Katleen De Preter, Filip Pattyn, Steve Lefever and Candy Kumps, for teaching me so much about neuroblastoma and genetics, as well as helping me out with data issues I encountered. Lastly, I would like to thank the research group from the Department of Knowledge Technologies at the Jožef Stefan Institue in Ljubljana, Slovenia. Special thanks goes to both Sašo Džeroski and Ivica Slavkov, for the nice and intensive cooperation throughout my stay in Ljubljana.

A Results Knowledge Domain Comparison

This appendix contains the results for the knowledge domain comparison. The results are categorized as follows. First, the results of the Safarii $event = 1$ ranking are presented, where the top-25 patterns obtained with the novelty as the target are situated on the left, and the top-25 patterns using the partial rank as the target are on the right. Secondly, the results for the IJS $event = 1$ ranking are shown, followed by the results for the Safarii $stage = 4$ ranking. For all rankings, first the results of the aggregation with the GO/KEGG terms domain are shown, followed by the results of the gene-to-gene domain and the PFAM domain, and concluded by the results of the aggregation with the gene locations domain.

Safarii $event = 1$ **ranking, target=novelty, domain: GO**

pattern	**size**	φ_z
GO:0007067: mitosis	111	13.557
GO:0007049: cell cycle	343	12.489
GO:0006260: DNA replication	98	12.423
GO:0051301: cell division	144	12.389
GO:0005634: nucleus	3128	10.917
KEGG:04110: Cell cycle	96	9.962
GO:0005694: chromosome	106	9.529
GO:0000166: nucleotide binding	1306	9.119
GO:0005524: ATP binding	1025	9.058
GO:0006281: DNA repair	124	8.977
GO:0048015: phosphoinositide-mediated signaling	17	8.565
GO:0000775: chromosome, centromeric region	38	7.764
GO:0006334: nucleosome assembly	69	7.432
GO:0005515: protein binding	3152	7.232
GO:0006270: DNA replication initiation	19	7.228
GO:0005739: mitochondrion	627	7.078
GO:0000786: nucleosome	58	6.854
GO:0000776: kinetochore	17	6.822
KEGG:00240: Pyrimidine metabolism	81	6.692
GO:0007051: spindle organization	8	6.494
GO:0016740: transferase activity	910	6.426
GO:0008094: DNA-dependent ATPase activity	21	6.232
GO:0003677: DNA binding	871	6.210
GO:0004523: ribonuclease H activity	5	5.840
KEGG:03030: DNA replication	23	5.824
$\mu_{\textbf{top-10}}$	648.1	10.842
$\mu_{\textbf{top-25}}$	495.88	8.445

Safarii $event = 1$ **ranking, target=rank$_{\text{partial}}$, domain: GO**

pattern	**size**	φ_z
GO:0051301: cell division	144	8.930
GO:0007049: cell cycle	343	8.770
GO:0007067: mitosis	111	8.710
GO:0005634: nucleus	3128	8.641
GO:0006260: DNA replication	98	7.526
GO:0000166: nucleotide binding	1306	6.981
GO:0005524: ATP binding	1025	6.662
GO:0005739: mitochondrion	627	6.516
GO:0005694: chromosome	106	6.308
GO:0005515: protein binding	3152	5.950
GO:0006281: DNA repair	124	5.713
GO:0000775: chromosome, centromeric region	38	5.681
GO:0016740: transferase activity	910	5.378
KEGG:04110: Cell cycle	96	5.337
GO:0006334: nucleosome assembly	69	5.082
GO:0003677: DNA binding	871	5.070
GO:0003723: RNA binding	387	5.030
KEGG:00240: Pyrimidine metabolism	81	4.974
KEGG:00710: Carbon fixation in photosynthetic organisms	21	4.766
GO:0048015: phosphoinositide-mediated signaling	17	4.682
GO:0000786: nucleosome	58	4.577
GO:0004674: protein serine/threonine kinase activity	285	4.441
GO:0006333: chromatin assembly or disassembly	26	4.373
GO:0006270: DNA replication initiation	19	4.241
GO:0008380: RNA splicing	154	4.195
$\mu_{\textbf{top-10}}$	1004	7.499
$\mu_{\textbf{top-25}}$	527.84	5.941

Safarii $event = 1$ **ranking, target=novelty, domain: gene2gene**

pattern	**size**	φ_z
gene2gene = CDC7	14	8.409
gene2gene = CDC25A	16	8.246
gene2gene = CDC2	48	8.046
gene2gene = PCNA	65	8.018
gene2gene = CDC6	15	7.966
gene2gene = CDK3	8	7.284
gene2gene = E2F4	63	7.105
gene2gene = BIRC5	16	6.812
gene2gene = MCM2	23	6.614
gene2gene = DBF4	6	6.452
gene2gene = MCM6	16	6.432
gene2gene = ORC3L	10	6.424
gene2gene = PTMA	27	6.409
gene2gene = MCM4	5	6.334
gene2gene = MCM3	15	6.323
gene2gene = ORC2L	19	6.296
gene2gene = CDK2	36	6.292
gene2gene = CDC20	13	6.277
gene2gene = RBL2	27	6.244
gene2gene = MSH2	13	5.952
gene2gene = HAUS1	5	5.899
gene2gene = CHAF1B	6	5.796
gene2gene = ORC4L	7	5.714
gene2gene = E2F1	65	5.677
gene2gene = ASF1A	13	5.642
$\mu_{\textbf{top-10}}$	27.4	7.495
$\mu_{\textbf{top-25}}$	22.04	6.666

Safarii $event = 1$ **ranking, target=rank$_{\text{partial}}$, domain: gene2gene**

pattern	**size**	φ_z
gene2gene = CDC25A	16	5.418
gene2gene = PCNA	65	5.359
gene2gene = CDC2	48	5.101
gene2gene = CDC7	14	4.708
gene2gene = YWHAG	63	4.693
gene2gene = CDC6	15	4.639
gene2gene = PTMA	27	4.591
gene2gene = RBL2	27	4.364
gene2gene = CDK2	36	4.286
gene2gene = SMN1	23	4.135
gene2gene = RAD51	17	4.102
gene2gene = CDK3	8	4.057
gene2gene = E2F4	63	4.037
gene2gene = CBX5	25	4.014
gene2gene = DDX20	11	3.985
gene2gene = DBF4	6	3.940
gene2gene = CBX1	13	3.865
gene2gene = MCM2	23	3.769
gene2gene = MCM6	16	3.701
gene2gene = ASF1A	13	3.689
gene2gene = BIRC5	16	3.663
gene2gene = MCM10	22	3.654
gene2gene = MCM4	5	3.633
gene2gene = MSH2	13	3.622
gene2gene = CTDP1	9	3.604
$\mu_{\textbf{top-10}}$	33.4	4.729
$\mu_{\textbf{top-25}}$	23.76	4.185

Safarii $event = 1$ **ranking, target=novelty, domain: PFAM**		
pattern	**size**	φ_z
PFAM = Histone	66	11.085
PFAM = Kinesin	36	7.141
PFAM = MCM	8	6.973
PFAM = WD40	209	5.360
PFAM = Helicase_C	91	4.851
PFAM = Cyclin_C	14	4.553
PFAM = Pkinase	371	4.424
PFAM = Linker_histone	11	4.301
PFAM = RRM_1	149	4.190
PFAM = Chromo	19	3.807
PFAM = E2F_TDP	10	3.802
PFAM = PHD	69	3.773
PFAM = Cyclin_N	24	3.612
PFAM = GAF	5	3.538
PFAM = RhoGEF	49	3.527
PFAM = BAH	9	3.415
PFAM = Na_trans_assoc	9	3.379
PFAM = FHA	20	3.367
PFAM = Pkinase_Tyr	178	3.276
PFAM = CH	52	3.202
PFAM = AAA	43	3.161
PFAM = PH	163	3.154
PFAM = FA	11	3.150
PFAM = SNF2_N	27	3.142
PFAM = Anticodon_1	5	3.140
$\mu_{\textbf{top-10}}$	97.4	5.669
$\mu_{\textbf{top-25}}$	65.92	4.293

Safarii $event = 1$ **ranking, target=rank$_{\textbf{partial}}$, domain: PFAM**		
pattern	**size**	φ_z
PFAM = Histone	66	8.218
PFAM = Kinesin	36	4.780
PFAM = WD40	209	4.190
PFAM = MCM	8	4.127
PFAM = Pkinase	371	4.032
PFAM = RRM_1	149	3.990
PFAM = Helicase_C	91	3.947
PFAM = RhoGEF	49	3.547
PFAM = Chromo	19	3.487
PFAM = Linker_histone	11	3.280
PFAM = Cyclin_C	14	3.270
PFAM = PH	163	3.024
PFAM = Pkinase_Tyr	178	3.000
PFAM = OATP	10	2.927
PFAM = Cyclin_N	24	2.913
PFAM = PHD	69	2.763
PFAM = PDEase_I	16	2.757
PFAM = NTF2	6	2.724
PFAM = RNase_PH_C	5	2.681
PFAM = CH	52	2.656
PFAM = RNase_PH	6	2.582
PFAM = SNF2_N	27	2.566
PFAM = LSM	12	2.498
PFAM = THAP	7	2.498
PFAM = PAS_3	20	2.472
$\mu_{\textbf{top-10}}$	100.9	4.360
$\mu_{\textbf{top-25}}$	64.72	3.397

Safarii $event = 1$ **ranking, target=novelty, domain: gene location**		
pattern	**size**	φ_z
chromosome = X	568	7.917
cytoband = Xq28	85	7.753
cytoband = 6p22.1	84	7.212
cytoband = q28	93	7.153
cytoband = 11q13.1	111	6.642
cytoband = q13.1	227	6.088
chromosome = 16	580	5.400
cytoband = p22.1	156	5.157
cytoband = 16p13.3	141	4.656
cytoband = 2q24.3	15	4.107
cytoband = p13.3	468	4.083
cytoband = 19p13.3	159	4.052
cytoband = 11q12.2	28	3.780
cytoband = Xq13.1	32	3.751
cytoband = 16p11.2	75	3.647
cytoband = q24.3	207	3.489
cytoband = p16.3	43	3.438
cytoband = p13	151	3.399
cytoband = p11.23	73	3.390
cytoband = Xp11.22	20	3.351
cytoband = Xp11.23	44	3.204
cytoband = 4p16.3	37	3.204
cytoband = q15.1	45	3.181
cytoband = 15q15.1	45	3.181
cytoband = 11q12.3	36	3.142
$\mu_{\textbf{top-10}}$	206	6.208
$\mu_{\textbf{top-25}}$	140.92	4.575

Safarii $event = 1$ **ranking, target=rank$_{\textbf{partial}}$, domain: gene location**		
pattern	**size**	φ_z
chromosome = X	568	6.857
cytoband = 11q13.1	111	5.797
cytoband = Xq28	85	5.761
cytoband = q28	93	5.363
cytoband = q13.1	227	5.267
chromosome = 16	580	4.962
cytoband = 16p13.3	141	4.562
cytoband = 6p22.1	84	4.489
cytoband = p13.3	468	4.207
cytoband = 19p13.3	159	3.897
cytoband = p11.23	73	3.750
cytoband = 16p11.2	75	3.534
cytoband = p22.1	156	3.496
cytoband = p13	151	3.225
chromosome = 14	453	3.048
cytoband = Xp11.23	44	3.041
cytoband = q15.1	45	3.032
cytoband = 15q15.1	45	3.032
cytoband = 5q33.3	20	2.987
cytoband = q24.3	207	2.810
cytoband = q22.1	344	2.807
cytoband = 2q24.3	15	2.802
cytoband = p16.3	43	2.728
cytoband = p13.11	91	2.721
cytoband = p22.11	13	2.712
$\mu_{\textbf{top-10}}$	251.6	5.116
$\mu_{\textbf{top-25}}$	171.64	3.876

IJS $event = 1$ ranking, target=measure, domain: GO

pattern	size	φ_z
GO:0004028: 3-chloroallyl aldehyde dehydrogenase activity	6	5.687
GO:0005678: chromatin assembly complex	6	5.327
GO:0004029: aldehyde dehydrogenase (NAD) activity	7	5.254
KEGG:00053: Ascorbate and aldarate metabolism	14	3.678
KEGG:00410: beta-Alanine metabolism	23	2.836
KEGG:00903: Limonene and pinene degradation	26	2.657
KEGG:00640: Propanoate metabolism	33	2.348
KEGG:00340: Histidine metabolism	41	2.094
KEGG:00120: Bile acid biosynthesis	41	2.069
KEGG:00620: Pyruvate metabolism	43	2.043
KEGG:00650: Butanoate metabolism	43	2.034
KEGG:00310: Lysine degradation	52	1.958
KEGG:00071: Fatty acid metabolism	49	1.953
KEGG:00280: Valine, leucine and isoleucine degradation	47	1.916
KEGG:00330: Arginine and proline metabolism	50	1.860
KEGG:00010: Glycolysis / Gluconeogenesis	61	1.716
KEGG:00561: Glycerolipid metabolism	60	1.668
KEGG:00380: Tryptophan metabolism	82	1.525
GO:0006260: DNA replication	104	1.509
GO:0003682: chromatin binding	92	1.302
GO:0051082: unfolded protein binding	99	1.147
GO:0006281: DNA repair	134	1.117
GO:0006461: protein complex assembly	112	1.041
GO:0000080: G1 phase of mitotic cell cycle	10	1.005
GO:0007049: cell cycle	381	0.907
$\mu_{\textbf{top-10}}$	24	3.399
$\mu_{\textbf{top-25}}$	64.64	2.266

IJS $event = 1$ ranking, target=$\text{rank}_{\text{partial}}$, domain: GO

pattern	size	φ_z
GO:0006260: DNA replication	104	6.914
GO:0007067: mitosis	124	6.871
GO:0007049: cell cycle	381	6.754
GO:0051301: cell division	160	6.654
GO:0006281: DNA repair	134	6.636
GO:0005694: chromosome	116	6.518
GO:0006334: nucleosome assembly	72	5.846
GO:0000786: nucleosome	62	5.604
GO:0005634: nucleus	3536	5.037
KEGG:04110: Cell cycle	110	4.735
GO:0000775: chromosome, centromeric region	39	4.577
GO:0003777: microtubule motor activity	66	4.395
GO:0000166: nucleotide binding	1477	4.376
GO:0007018: microtubule-based movement	76	4.357
GO:0048015: phosphoinositide-mediated signaling	17	4.356
GO:0000785: chromatin	61	4.256
GO:0008094: DNA-dependent ATPase activity	22	3.988
GO:0005524: ATP binding	1167	3.978
GO:0007051: spindle organization	8	3.917
GO:0005875: microtubule associated complex	54	3.801
KEGG:04610: Complement and coagulation cascades	69	3.792
GO:0016740: transferase activity	1040	3.786
GO:0005840: ribosome	96	3.664
GO:0006270: DNA replication initiation	20	3.499
KEGG:03030: DNA replication	25	3.492
$\mu_{\textbf{top-10}}$	479.9	6.157
$\mu_{\textbf{top-25}}$	361.44	4.872

IJS $event = 1$ ranking, target=measure, domain: gene2gene

pattern	size	φ_z
gene2gene = CHAF1B	6	5.548
gene2gene = MBD1	6	5.323
gene2gene = ASF1B	9	4.353
gene2gene = SETDB1	12	3.706
gene2gene = CBX1	14	3.545
gene2gene = ASF1A	14	3.494
gene2gene = CBX5	27	3.217
gene2gene = BAZ1B	17	3.094
gene2gene = TCERG1	20	2.848
gene2gene = TFDP2	5	2.724
gene2gene = JMY	5	2.722
gene2gene = TP53INP1	6	2.483
gene2gene = CDK3	8	2.167
gene2gene = TFDP1	9	2.007
gene2gene = MT1G	10	1.889
gene2gene = UXT	12	1.717
gene2gene = PCNA	71	1.709
gene2gene = MSH2	14	1.674
gene2gene = BIRC5	18	1.639
gene2gene = TP53BP2	14	1.581
gene2gene = CDC6	17	1.547
gene2gene = PURA	5	1.463
gene2gene = RECQL	5	1.461
gene2gene = GAB2	18	1.383
gene2gene = CSDA	6	1.328
$\mu_{\textbf{top-10}}$	13	3.785
$\mu_{\textbf{top-25}}$	13.92	2.585

IJS $event = 1$ ranking, target=$\text{rank}_{\text{partial}}$, domain: gene2gene

pattern	size	φ_z
gene2gene = CDC2	49	5.139
gene2gene = CDK3	8	4.325
gene2gene = PTMA	28	4.136
gene2gene = BRCA2	13	4.048
gene2gene = MCM2	25	3.865
gene2gene = E2F4	68	3.787
gene2gene = PCNA	71	3.767
gene2gene = RAD51	18	3.764
gene2gene = CCNB1	25	3.613
gene2gene = TAF9	14	3.589
gene2gene = MCM6	16	3.574
gene2gene = YWHAG	69	3.564
gene2gene = MCM3	15	3.563
gene2gene = MCM4	5	3.377
gene2gene = ORC2L	20	3.369
gene2gene = MDC1	8	3.365
gene2gene = CDC7	14	3.296
gene2gene = RPA1	19	3.276
gene2gene = CDC6	17	3.270
gene2gene = CD19	14	3.238
gene2gene = CTDP1	9	3.235
gene2gene = CCNA1	20	3.224
gene2gene = CBX1	14	3.209
gene2gene = TP53INP1	6	3.175
gene2gene = RAD51L3	5	3.161
$\mu_{\textbf{top-10}}$	31.9	4.003
$\mu_{\textbf{top-25}}$	22.8	3.597

IJS $event = 1$ ranking, target=measure, domain: PFAM

pattern	size	φ_z
PFAM = Aldedh	16	3.434
PFAM = E2F_TDP	11	1.809
PFAM = Macro	8	1.741
PFAM = NTF2	8	0.773
PFAM = FA	13	0.741
PFAM = MBT	8	0.562
PFAM = FERM_M	27	0.460
PFAM = Peptidase_M28	6	0.430
PFAM = MORN	13	0.408
PFAM = MCM	8	0.392
PFAM = Piwi	8	0.313
PFAM = Anticodon_1	5	0.290
PFAM = PAZ	9	0.288
PFAM = DUF1669	6	0.254
PFAM = LSM	16	0.248
PFAM = ResIII	16	0.224
PFAM = LRR_1	215	0.211
PFAM = tRNA-synt_1g	8	0.208
PFAM = DMAP_binding	5	0.190
PFAM = FAT	5	0.186
PFAM = SNF2_N	29	0.176
PFAM = FATC	5	0.175
PFAM = RhoGEF	55	0.172
PFAM = FYVE	30	0.170
PFAM = Na_trans_assoc	10	0.167
$\mu_{\textbf{top-10}}$	11.8	1.075
$\mu_{\textbf{top-25}}$	21.6	0.561

IJS $event = 1$ ranking, target=$\text{rank}_{\text{partial}}$, domain: PFAM

pattern	size	φ_z
PFAM = Histone	69	6.963
PFAM = Kinesin	38	5.533
PFAM = Cyclin_C	14	3.803
PFAM = Chromo	19	3.689
PFAM = Cadherin_2	42	3.553
PFAM = Sushi	47	3.520
PFAM = MCM	8	3.180
PFAM = Rad51	5	3.142
PFAM = E2F_TDP	11	3.073
PFAM = Linker_histone	12	3.034
PFAM = Methyltransf_11	23	3.008
PFAM = Na_trans_assoc	10	2.993
PFAM = PAS	23	2.887
PFAM = MAGE	25	2.839
PFAM = Helicase_C	99	2.810
PFAM = WD40	231	2.762
PFAM = Anticodon_1	5	2.736
PFAM = GATase	5	2.703
PFAM = ATP-grasp	5	2.662
PFAM = Piwi	8	2.653
PFAM = Cyclin_N	27	2.549
PFAM = Spectrin	24	2.517
PFAM = BAR	12	2.466
PFAM = adh_short	53	2.429
PFAM = Band_3_cyto	9	2.421
$\mu_{\textbf{top-10}}$	26.5	3.949
$\mu_{\textbf{top-25}}$	32.96	3.197

IJS $event = 1$ ranking, target=measure, domain: gene location

pattern	size	φ_z
cytoband = 18p11.31	9	1.096
cytoband = p13.1	155	0.962
cytoband = p11.31	12	0.945
cytoband = 19p13.3	171	0.841
cytoband = 6p22.3	20	0.550
cytoband = 20q11.22	36	0.469
cytoband = 8q24.3	77	0.448
cytoband = 8q21.11	13	0.410
cytoband = p22.13	9	0.380
cytoband = Xp22.13	9	0.380
cytoband = 11q13.1	128	0.378
cytoband = q11.22	65	0.350
cytoband = Xq22.3	34	0.336
cytoband = p36.31	20	0.325
cytoband = 1p36.31	20	0.325
cytoband = 6q24.3	7	0.283
cytoband = 2p22.2	12	0.267
cytoband = 12q22	17	0.266
cytoband = 11q12.3	40	0.253
cytoband = 22q13.31	26	0.252
cytoband = p13.3	521	0.250
cytoband = Xq28	95	0.249
cytoband = q13.1	263	0.238
cytoband = 5q33.3	25	0.234
cytoband = p11.1	12	0.227
$\mu_{\textbf{top-10}}$	51.1	0.648
$\mu_{\textbf{top-25}}$	71.84	0.429

IJS $event = 1$ ranking, target=$\text{rank}_{\text{partial}}$, domain: gene location

pattern	size	φ_z
chromosome = X	631	9.406
cytoband = Xq28	95	5.739
cytoband = q28	104	5.267
cytoband = 6p22.1	90	5.162
cytoband = p22.1	174	4.572
cytoband = Xp11.22	23	3.751
cytoband = 19p13.3	171	3.729
cytoband = Xp11.23	46	3.642
cytoband = p22.11	14	3.434
cytoband = Xp22.11	14	3.434
cytoband = 11q13.1	128	3.357
chromosome = 6	857	3.327
cytoband = q13.1	263	3.218
cytoband = Xq13.1	39	3.216
cytoband = 5q33.3	25	3.198
cytoband = 17q25.3	73	3.181
cytoband = 5q21.3	5	3.100
cytoband = 11q13.2	27	3.028
chromosome = 2	979	2.922
cytoband = 2p16.2	8	2.692
cytoband = Xq22.1	36	2.598
cytoband = p31.1	39	2.593
cytoband = 1p31.1	39	2.593
cytoband = 5q14.3	13	2.587
cytoband = q24	57	2.540
$\mu_{\textbf{top-10}}$	136.2	4.814
$\mu_{\textbf{top-25}}$	158	3.692

Safarii $stage = 4$ ranking, target=novelty, domain: GO

pattern	size	φ_z
GO:0007156: homophilic cell adhesion	112	8.199
GO:0006260: DNA replication	95	8.177
GO:0000166: nucleotide binding	1323	7.962
GO:0007049: cell cycle	348	7.670
GO:0007067: mitosis	118	7.230
GO:0005515: protein binding	3154	7.029
GO:0005524: ATP binding	1036	7.024
GO:0051301: cell division	146	7.023
GO:0006281: DNA repair	127	6.813
KEGG:04110: Cell cycle	104	6.539
GO:0005634: nucleus	3078	6.071
GO:0016740: transferase activity	932	5.650
GO:0005875: microtubule associated complex	51	5.159
GO:0009116: nucleoside metabolic process	15	4.915
GO:0030496: midbody	8	4.864
GO:0005737: cytoplasm	1296	4.699
GO:0008094: DNA-dependent ATPase activity	22	4.603
GO:0003777: microtubule motor activity	59	4.583
GO:0016787: hydrolase activity	682	4.444
GO:0005874: microtubule	146	4.392
GO:0004674: protein serine/threonine kinase activity	309	4.329
GO:0005739: mitochondrion	626	4.326
GO:0006468: protein amino acid phosphorylation	416	4.301
GO:0000287: magnesium ion binding	288	4.266
GO:0006270: DNA replication initiation	17	4.244
$\mu_{\textbf{top-10}}$	656.3	7.367
$\mu_{\textbf{top-25}}$	580.32	5.780

Safarii $stage = 4$ ranking, target=$\text{rank}_{\text{partial}}$, domain: GO

pattern	size	φ_z
GO:0007156: homophilic cell adhesion	112	7.518
GO:0000166: nucleotide binding	1323	7.512
GO:0007049: cell cycle	348	7.097
GO:0006260: DNA replication	95	6.982
GO:0007067: mitosis	118	6.873
GO:0051301: cell division	146	6.460
GO:0005515: protein binding	3154	6.286
GO:0005524: ATP binding	1036	6.176
GO:0005634: nucleus	3078	5.925
KEGG:04110: Cell cycle	104	5.912
GO:0006281: DNA repair	127	5.830
GO:0016740: transferase activity	932	5.130
GO:0005737: cytoplasm	1296	4.784
GO:0005739: mitochondrion	626	4.580
GO:0008094: DNA-dependent ATPase activity	22	4.369
GO:0006334: nucleosome assembly	63	4.303
GO:0005694: chromosome	105	4.256
GO:0005874: microtubule	146	4.051
GO:0005875: microtubule associated complex	51	4.038
GO:0016787: hydrolase activity	682	4.020
GO:0004674: protein serine/threonine kinase activity	309	3.882
GO:0000910: cytokinesis	22	3.855
GO:0016874: ligase activity	173	3.846
GO:0017111: nucleoside-triphosphatase activity	59	3.810
GO:0000287: magnesium ion binding	288	3.797
$\mu_{\textbf{top-10}}$	951.4	6.674
$\mu_{\textbf{top-25}}$	576.6	5.252

Safarii $stage = 4$ ranking, target=novelty, domain: gene2gene

pattern	size	φ_z
gene2gene = CDC6	16	5.614
gene2gene = BIRC5	17	5.443
gene2gene = RAD51	17	5.433
gene2gene = CDK3	8	5.212
gene2gene = CDK2	34	4.994
gene2gene = ORC2L	18	4.975
gene2gene = CDC25A	20	4.679
gene2gene = SKP2	18	4.616
gene2gene = CCNA1	17	4.517
gene2gene = RB1	99	4.433
gene2gene = BRCA1	80	4.391
gene2gene = CDC2	46	4.287
gene2gene = DIAPH1	6	4.282
gene2gene = RICS	18	4.234
gene2gene = MCM3	15	4.078
gene2gene = BTRC	15	3.984
gene2gene = CCNA2	19	3.974
gene2gene = HABP4	5	3.955
gene2gene = JMY	5	3.933
gene2gene = SPTB	6	3.932
gene2gene = NCAM1	8	3.924
gene2gene = TP53INP1	6	3.917
gene2gene = YWHAG	59	3.879
gene2gene = TFDP2	5	3.866
gene2gene = RBL2	28	3.804
$\mu_{\textbf{top-10}}$	26.4	4.992
$\mu_{\textbf{top-25}}$	23.4	4.414

Safarii $stage = 4$ ranking, target=$\text{rank}_{\text{partial}}$, domain: gene2gene

pattern	size	φ_z
gene2gene = RAD51	17	4.990
gene2gene = CDC6	16	4.560
gene2gene = ORC2L	18	4.324
gene2gene = BRCA1	80	4.288
gene2gene = BIRC5	17	4.080
gene2gene = CCNA2	19	3.913
gene2gene = CDC25A	20	3.826
gene2gene = CDK3	8	3.791
gene2gene = CDK2	34	3.786
gene2gene = RB1	99	3.778
gene2gene = SKP2	18	3.627
gene2gene = MCM2	24	3.552
gene2gene = MCM3	15	3.543
gene2gene = MCM7	30	3.466
gene2gene = RBL1	28	3.461
gene2gene = BTRC	15	3.458
gene2gene = DIAPH1	6	3.370
gene2gene = CDKN1A	55	3.359
gene2gene = CHEK1	11	3.354
gene2gene = RBL2	28	3.322
gene2gene = CDC2	46	3.300
gene2gene = YWHAG	59	3.271
gene2gene = ADRB1	9	3.201
gene2gene = E2F4	60	3.182
gene2gene = RAD51L1	7	3.172
$\mu_{\textbf{top-10}}$	32.8	4.134
$\mu_{\textbf{top-25}}$	29.56	3.679

Safarii $stage = 4$ ranking, target=novelty, domain: PFAM

pattern	size	φ_z
PFAM = Cadherin_2	42	7.040
PFAM = Cadherin	78	5.847
PFAM = Kinesin	37	4.835
PFAM = HEAT	65	4.706
PFAM = FA	9	4.631
PFAM = PHD	68	4.420
PFAM = Helicase_C	92	4.279
PFAM = Pribosyltran	9	4.228
PFAM = Pkinase	388	4.127
PFAM = Na_trans_assoc	8	4.062
PFAM = SNF2_N	27	3.554
PFAM = DnaJ_C	9	3.554
PFAM = FERM_M	21	3.462
PFAM = zf-Tim10_DDP	5	3.444
PFAM = E2F_TDP	9	3.396
PFAM = SAM_2	63	3.374
PFAM = SAM_1	69	3.345
PFAM = TIG	27	3.332
PFAM = FHA	23	3.275
PFAM = HhH-GPD	5	3.240
PFAM = UQ_con	29	3.205
PFAM = KA1	5	3.103
PFAM = BAH	8	3.079
PFAM = RhoGAP	44	3.030
PFAM = Myb_DNA-binding	30	2.978
$\mu_{\textbf{top-10}}$	79.6	4.817
$\mu_{\textbf{top-25}}$	46.8	3.902

Safarii $stage = 4$ ranking, target=$\text{rank}_{\text{partial}}$, domain: PFAM

pattern	size	φ_z
PFAM = Cadherin_2	42	5.945
PFAM = Cadherin	78	5.547
PFAM = HEAT	65	4.710
PFAM = Kinesin	37	3.708
PFAM = Pkinase	388	3.461
PFAM = Pribosyltran	9	3.318
PFAM = Methyltransf_12	16	3.268
PFAM = Helicase_C	92	3.073
PFAM = SAM_2	63	3.019
PFAM = DnaJ_C	9	2.947
PFAM = SAM_1	69	2.890
PFAM = PHD	68	2.866
PFAM = UQ_con	29	2.864
PFAM = Histone	66	2.854
PFAM = Na_trans_assoc	8	2.794
PFAM = WD40	211	2.781
PFAM = BAR	12	2.757
PFAM = Rad51	5	2.714
PFAM = zf-Tim10_DDP	5	2.697
PFAM = MCM	8	2.690
PFAM = Methyltransf_11	20	2.684
PFAM = Arfaptin	6	2.664
PFAM = KA1	5	2.643
PFAM = SMC_hinge	6	2.623
PFAM = Pkinase_C	35	2.611
$\mu_{\textbf{top-10}}$	79.9	3.900
$\mu_{\textbf{top-25}}$	54.08	3.205

Safarii $stage = 4$ ranking, target=novelty, domain: gene location

pattern	size	φ_z
chromosome = 11	936	9.391
chromosome = X	559	8.454
chromosome = 17	834	7.247
cytoband = 17p11.2	47	6.728
cytoband = 11q21	21	5.900
cytoband = q21	75	5.771
cytoband = 5q31.3	59	5.415
cytoband = Xq28	88	5.246
cytoband = 11q13.1	118	5.141
cytoband = q28	95	5.112
cytoband = q13.1	233	4.680
cytoband = q31.3	112	4.570
cytoband = p11.2	223	4.249
cytoband = 11p15.1	36	4.119
cytoband = 11q12.2	25	4.083
cytoband = p36.31	18	4.037
cytoband = 1p36.31	18	4.037
cytoband = 11q12.3	38	3.687
cytoband = 11p11.2	44	3.687
cytoband = 6q14.1	18	3.622
cytoband = q31.21	7	3.545
cytoband = 4q31.21	7	3.545
cytoband = 9q22.2	5	3.491
cytoband = 17q25.1	68	3.491
cytoband = 17q22	36	3.455
$\mu_{\textbf{top-10}}$	283.2	6.440
$\mu_{\textbf{top-25}}$	148.8	4.908

Safarii $stage = 4$ ranking, target=$\text{rank}_{\text{partial}}$, domain: gene location

pattern	size	φ_z
chromosome = X	559	7.142
chromosome = 11	936	6.745
chromosome = 17	834	6.237
cytoband = 11q13.1	118	5.293
cytoband = 17p11.2	47	5.219
cytoband = Xq28	88	4.934
cytoband = q21	75	4.929
cytoband = q13.1	233	4.914
cytoband = 11q21	21	4.868
cytoband = q28	95	4.819
cytoband = 5q31.3	59	4.308
cytoband = q31.3	112	3.773
cytoband = 6q23.3	18	3.604
cytoband = 17q25.1	68	3.590
cytoband = p11.2	223	3.419
cytoband = 6q14.1	18	3.415
cytoband = 17q23.1	7	3.078
cytoband = 16p12.1	29	3.059
cytoband = 11p15.1	36	3.035
cytoband = q32.33	35	3.027
cytoband = 14q32.33	35	3.027
cytoband = 17q22	36	3.012
cytoband = 17q24.1	8	3.011
cytoband = 2q24.3	16	3.010
cytoband = 11q12.2	25	2.835
$\mu_{\textbf{top-10}}$	300.6	5.510
$\mu_{\textbf{top-25}}$	149.24	4.172

B Results Quality Measure Performance

In this appendix, the results considering the quality measure performance can be found. The results are organized as follows. First, the top-25 patterns of the aggregation with search depth 3 are presented. They are ordered by quality measures as follows: φ_{avg}, φ_{mt}, φ_z, φ_t, φ_{χ^2}, φ_{roc}, φ_{wmw} and φ_{mmad}. If possible, the results using the novelty as the target are presented first, followed by the results using the (partial) rank as the target. Of course, this is only possible for the quality measures for regressional subgroup discovery. Next, the results generated with search depth 4 are presented. When conditions are combined for a pattern, the combination is denoted by $\wedge$, which stands for AND. Furthermore, 'norm. φ_x' stands for the normalized evaluation values, where the maximum evaluation value is set to 1, and all values get assigned a number between 0 and 1.

Safarii $event = 1$ **ranking, target=novelty, measure=**φ_{avg}**, depth=3**

pattern	**size**	φ_{avg}	**norm.** φ_{avg}
gene2gene = MCM4	5	0.1079	1.000
gene2gene = HAUS1	5	0.1041	0.965
gene2gene = DBF4	6	0.1040	0.965
GO/KEGG = GO:0004523: ribonuclease H activity	5	0.1036	0.961
gene2gene = CDK3	8	0.1029	0.954
pfam = MCM	8	0.1008	0.934
gene2gene = CHAF1B	6	0.0989	0.917
GO/KEGG = GO:0005658: alpha DNA polymerase:primase complex	5	0.0985	0.913
GO/KEGG = GO:0007051: spindle organization	8	0.0975	0.904
gene2gene = CDC7	14	0.0966	0.895
gene2gene = CKS2	5	0.0950	0.881
gene2gene = ORC4L	7	0.0949	0.880
GO/KEGG = GO:0000076: DNA replication checkpoint	5	0.0935	0.867
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling	17	0.0933	0.865
gene2gene = TFDP2	5	0.0932	0.864
gene2gene = CDC25A	16	0.0930	0.862
gene2gene = CDC6	15	0.0929	0.861
gene2gene = ORC3L	10	0.0924	0.857
GO/KEGG = GO:0000307: cyclin-dependent protein kinase holoenzyme complex	6	0.0916	0.849
gene2gene = JMY	5	0.0904	0.838
gene2gene = ATF5	5	0.0903	0.837
gene2gene = UNC13B	5	0.0896	0.831
GO/KEGG = GO:0006268: DNA unwinding during replication	9	0.0889	0.824
gene2gene = CENPF	5	0.0889	0.824
GO/KEGG = GO:0000070: mitotic sister chromatid segregation	7	0.0887	0.823
μ**top-10**	7	0.1015	0.941
μ**top-25**	7.68	0.0956	0.887

Safarii $event = 1$ **ranking, target=rank**$_{\textbf{partial}}$**, measure=**φ_{avg}**, depth=3**

pattern	**size**	φ_{avg}	**norm.** φ_{avg}
gene2gene = MCM4	5	-536.2000	1.000
gene2gene = DBF4	6	-617.4167	0.868
GO/KEGG = GO:0004523: ribonuclease H activity	5	-871.4000	0.615
GO/KEGG = GO:0005658: alpha DNA polymerase:primase complex	5	-1029.9000	0.521
pfam = MCM	8	-1364.0625	0.393
gene2gene = UNC13B	5	-1369.8000	0.391
gene2gene = ATF5	5	-1436.8000	0.373
gene2gene = CDK3	8	-1488.9375	0.360
gene2gene = POLR1A	5	-1506.0000	0.356
gene2gene = HAUS1	5	-1597.7000	0.336
gene2gene = AURKB	5	-1795.4000	0.299
gene2gene = NAP1L4	5	-1798.3000	0.298
GO/KEGG = GO:0005844: polysome	5	-1806.5000	0.297
gene2gene = CDC25A	16	-1887.6563	0.284
GO/KEGG = GO:0007051: spindle organization	8	-1997.5000	0.268
GO/KEGG = GO:0000307: cyclin-dependent protein kinase holoenzyme complex	6	-2013.3333	0.266
gene2gene = CHAF1B	6	-2101.9167	0.255
gene2gene = NCAM1	7	-2129.3571	0.252
GO/KEGG = GO:0006221: pyrimidine nucleotide biosynthetic process	5	-2160.5000	0.248
gene2gene = HABP4	5	-2224.5000	0.241
gene2gene = LBR	6	-2275.2500	0.236
GO/KEGG = GO:0005881: cytoplasmic microtubule	8	-2295.4375	0.234
gene2gene = MIS12	7	-2336.6429	0.229
gene2gene = CDC7	14	-2368.8929	0.226
GO/KEGG = GO:0005762: mitochondrial large ribosomal subunit	8	-2446.8750	0.219
μ**top-10**	5.7	-1181.8217	0.521
μ**top-25**	6.72	-1738.2511	0.363

Safarii $event = 1$ **ranking, target=novelty, measure=**φ_{mt} **and** φ_z**, depth=3**			
pattern	**size**	φ_{mt}/φ_z	**norm.** φ_{mt}/φ_z
GO/KEGG = GO:0007067: mitosis	111	0.2615/13.5572	1.000
GO/KEGG = GO:0007049: cell cycle	343	0.2409/12.4894	0.921
GO/KEGG = GO:0006260: DNA replication	98	0.2396/12.4231	0.916
GO/KEGG = GO:0051301: cell division	144	0.2389/12.3895	0.914
pfam = Histone	66	0.2138/11.0849	0.818
GO/KEGG = GO:0005634: nucleus	3128	0.2105/10.9170	0.805
GO/KEGG = KEGG:04110: Cell cycle	96	0.1921/9.9621	0.735
GO/KEGG = GO:0005694: chromosome	106	0.1838/9.5289	0.703
GO/KEGG = GO:0000166: nucleotide binding	1306	0.1759/9.1186	0.673
GO/KEGG = GO:0005524: ATP binding	1025	0.1747/9.0582	0.668
GO/KEGG = GO:0006281: DNA repair	124	0.1731/8.9770	0.662
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling	17	0.1652/8.5653	0.632
gene2gene = CDC7	14	0.1622/8.4085	0.620
gene2gene = CDC25A	16	0.1590/8.2455	0.608
gene2gene = CDC2	48	0.1552/8.0457	0.593
gene2gene = PCNA	65	0.1546/8.0185	0.591
gene2gene = CDC6	15	0.1536/7.9660	0.588
chromosome = X	568	0.1527/7.9174	0.584
GO/KEGG = GO:0000775: chromosome, centromeric region	38	0.1497/7.7640	0.573
cytoband = Xq28	85	0.1495/7.7527	0.572
GO/KEGG = GO:0006334: nucleosome assembly	69	0.1433/7.4321	0.548
gene2gene = CDK3	8	0.1405/7.2843	0.537
GO/KEGG = GO:0005515: protein binding	3152	0.1395/7.2323	0.533
GO/KEGG = GO:0006270: DNA replication initiation	19	0.1394/7.2281	0.533
cytoband = 6p22.1	84	0.1391/7.2125	0.532
μ**top-10**	642.3	0.2132/11.0529	0.815
μ**top-25**	429.8	0.1763/9.1431	0.674

Safarii $event = 1$ **ranking, target=rank**$_{\textbf{partial}}$**, measure=**φ_{mt} **and** φ_z**, depth=3**			
pattern	**size**	φ_{mt}/φ_z	**norm.** φ_{mt}/φ_z
GO/KEGG = GO:0051301: cell division	144	44642.8750/8.9302	1.000
GO/KEGG = GO:0007049: cell cycle	343	43843.9599/8.7704	0.982
GO/KEGG = GO:0007067: mitosis	111	43541.5790/8.7099	0.975
GO/KEGG = GO:0005634: nucleus	3128	43195.5870/8.6407	0.968
pfam = Histone	66	41084.4316/8.2184	0.920
GO/KEGG = GO:0006260: DNA replication	98	37623.3841/7.5260	0.843
GO/KEGG = GO:0000166: nucleotide binding	1306	34897.4015/6.9807	0.782
chromosome = X	568	34279.2997/6.8571	0.768
GO/KEGG = GO:0005524: ATP binding	1025	33302.9459/6.6618	0.746
GO/KEGG = GO:0005739: mitochondrion	627	32575.7604/6.5163	0.730
GO/KEGG = GO:0005694: chromosome	106	31535.0295/6.3081	0.706
GO/KEGG = GO:0005515: protein binding	3152	29742.3482/5.9495	0.666
cytoband = 11q13.1	111	28979.9292/5.7970	0.649
cytoband = Xq28	85	28801.9111/5.7614	0.645
GO/KEGG = GO:0006281: DNA repair	124	28561.4637/5.7133	0.640
GO/KEGG = GO:0000775: chromosome,centromeric region	38	28398.8875/5.6808	0.636
gene2gene = CDC25A	16	27085.3750/5.4180	0.607
GO/KEGG = GO:0016740: transferase activity	910	26883.3274/5.3776	0.602
cytoband = q28	93	26810.7490/5.3631	0.601
gene2gene = PCNA	65	26791.2546/5.3592	0.600
cytoband = 16p1	310	26693.2691/5.3396	0.598
GO/KEGG = KEGG:04110: Cell cycle	96	26693.2691/5.3367	0.598
cytoband = Xq	350	26678.5155/5.2890	0.592
cytoband = q13.1	227	26440.1012/5.2671	0.590
gene2gene = CDC2	48	25498.4581/5.1006	0.571
μ**top-10**	741.6	38898.7224/7.7811	0.871
μ**top-25**	525.88	32183.2445/6.4349	0.721

Safarii *event* = 1 **ranking, target=novelty, measure=φ_t, depth=3**

pattern	**size**	φ_t	**norm.** φ_t
gene2gene = HABP4	5	12.5422	1.000
GO/KEGG = GO:0006221: pyrimidine nucleotide biosynthetic process	5	10.2137	0.814
GO/KEGG = GO:0005634: nucleus	3128	9.5629	0.762
GO/KEGG = GO:0007049: cell cycle	343	8.8068	0.702
GO/KEGG = GO:0007067: mitosis	111	8.7241	0.696
GO/KEGG = GO:0051301: cell division	144	8.6704	0.691
pfam = Histone	66	8.6318	0.688
gene2gene = POLR1A	5	8.5592	0.682
GO/KEGG = GO:0000166: nucleotide binding	1306	7.7537	0.618
GO/KEGG = GO:0006260: DNA replication	98	7.5075	0.599
GO/KEGG = GO:0005524: ATP binding	1025	7.5068	0.599
chromosome = X	568	6.8082	0.543
GO/KEGG = GO:0005515: protein binding	3152	6.6157	0.527
gene2gene = DBF4	6	6.5099	0.519
GO/KEGG = GO:0005694: chromosome	106	6.4533	0.515
GO/KEGG = GO:0005739: mitochondrion	627	6.3805	0.509
gene2gene = MCM4	5	6.2914	0.502
GO/KEGG = GO:0005844: polysome	5	6.2618	0.499
gene2gene = CDC25A	16	6.2518	0.498
GO/KEGG = GO:0006281: DNA repair	124	5.9820	0.477
GO/KEGG = KEGG:04110: Cell cycle	96	5.7825	0.461
GO/KEGG = GO:0016740: transferase activity	910	5.6922	0.454
cytoband = Xq28	85	5.6508	0.451
GO/KEGG = GO:0000775: chromosome, centromeric region	38	5.5917	0.446
gene2gene = DDX20	11	5.4875	0.438
μtop-10	521.1	9.0972	0.725
μtop-25	479.4	7.3695	0.588

Safarii *event* = 1 **ranking, target=rank$_{\text{partial}}$, measure=φ_t, depth=3**

pattern	**size**	φ_t	**norm.** φ_t
gene2gene = DBF4	6	34.0101	1.000
gene2gene = HABP4	5	26.8436	0.789
gene2gene = MCM4	5	26.2291	0.771
gene2gene = POLR1A	5	25.9197	0.762
GO/KEGG = GO:0006221: pyrimidine nucleotide biosynthetic process	5	24.7377	0.727
GO/KEGG = GO:0005844: polysome	5	17.0839	0.502
GO/KEGG = GO:0004523: ribonuclease H activity	5	17.0613	0.502
gene2gene = UNC13B	5	15.1527	0.446
GO/KEGG = GO:0005658: alpha DNA polymerase:primase complex	5	14.0118	0.412
gene2gene = POLR2K	6	12.1077	0.356
gene2gene = LBR	6	11.7541	0.346
gene2gene = DDX20	11	11.1524	0.328
gene2gene = CDC25A	16	10.0867	0.297
gene2gene = NCAM1	7	9.9208	0.292
GO/KEGG = GO:0005881: cytoplasmic microtubule	8	9.8310	0.289
pfam = Histone	66	9.4289	0.277
pfam = MCM	8	9.2470	0.272
GO/KEGG = GO:0005762: mitochondrial large ribosomal subunit	8	9.1039	0.268
GO/KEGG = GO:0051301: cell division	144	9.0426	0.266
gene2gene = AURKB	5	9.0238	0.265
gene2gene = ATF5	5	8.9310	0.263
GO/KEGG = GO:0005634: nucleus	3128	8.3998	0.247
GO/KEGG = GO:0007067: mitosis	111	8.3733	0.246
GO/KEGG = GO:0007049: cell cycle	343	8.3611	0.246
gene2gene = SEC61B	5	8.3561	0.246
μtop-10	5.2	21.3158	0.627
μtop-25	156.92	14.1668	0.417

Safarii *event* = 1 **ranking, target=novelty, measure=φ_{χ^2}, depth=3**

pattern	**size**	φ_{χ^2}	**norm.** φ_{χ^2}
pfam = zf-UBR	5	17307	1
pfam = zf-C2HC	5	17307	1
pfam = zf-AN1	5	17307	1
pfam = WSC	5	17307	1
pfam = Vps4_C	5	17307	1
pfam = UPF0020	5	17307	1
pfam = Tub	5	17307	1
pfam = tRNA-synt_2	5	17307	1
pfam = Trefoil	5	17307	1
pfam = TRAM_LAG1_CLN8	5	17307	1
pfam = TPR_3	5	17307	1
pfam = Tim17	5	17307	1
pfam = TFR_dimer	5	17307	1
pfam = TFIIS_M	5	17307	1
pfam = TFIIS	5	17307	1
pfam = Tektin	5	17307	1
pfam = TB	5	17307	1
pfam = SWIRM	5	17307	1
pfam = Surp	5	17307	1
pfam = Sulfotransfer_2	5	17307	1
pfam = SRF-TF	5	17307	1
pfam = SPAN-X	5	17307	1
pfam = Somatomedin_B	5	17307	1
pfam = Sel1	5	17307	1
pfam = SCP2	5	17307	1
μtop-10	5	17307	1
μtop-25	5	17307	1

Safarii *event* = 1 **ranking, target=rank$_{\text{partial}}$, measure=φ_{χ^2}, depth=3**

pattern	**size**	φ_{χ^2}	**norm.** φ_{χ^2}
pfam = zf-UBR	5	17307	1
pfam = zf-C2HC	5	17307	1
pfam = zf-AN1	5	17307	1
pfam = WSC	5	17307	1
pfam = Vps4_C	5	17307	1
pfam = UPF0020	5	17307	1
pfam = Tub	5	17307	1
pfam = tRNA-synt_2	5	17307	1
pfam = Trefoil	5	17307	1
pfam = TRAM_LAG1_CLN8	5	17307	1
pfam = TPR_3	5	17307	1
pfam = Tim17	5	17307	1
pfam = TFR_dimer	5	17307	1
pfam = TFIIS_M	5	17307	1
pfam = TFIIS	5	17307	1
pfam = Tektin	5	17307	1
pfam = TB	5	17307	1
pfam = SWIRM	5	17307	1
pfam = Surp	5	17307	1
pfam = Sulfotransfer_2	5	17307	1
pfam = SRF-TF	5	17307	1
pfam = SPAN-X	5	17307	1
pfam = Somatomedin_B	5	17307	1
pfam = Sel1	5	17307	1
pfam = SCP2	5	17307	1
μtop-10	5	17307	1
μtop-25	5	17307	1

Safarii *event* = 1 **ranking, target=rank, measure=φ_{roc}, depth=3**			
pattern	**size**	φ_{roc}	**norm.** φ_{roc}
gene2gene = MCM4	5	0.9692	1.000
gene2gene = DBF4	6	0.9646	0.995
GO/KEGG = GO:0004523: ribonuclease H activity	5	0.9499	0.980
GO/KEGG = GO:0005658: alpha DNA polymerase:primase complex	5	0.9406	0.970
pfam = MCM	8	0.9214	0.951
gene2gene = UNC13B	5	0.9211	0.950
gene2gene = ATF5	5	0.9173	0.946
gene2gene = CDK3	8	0.9142	0.943
gene2gene = POLR1A	5	0.9132	0.942
gene2gene = HAUS1	5	0.9081	0.937
gene2gene = AURKB	5	0.8967	0.925
gene2gene = NAP1L4	5	0.8963	0.925
GO/KEGG = GO:0005844: polysome	5	0.8958	0.924
gene2gene = CDC25A	16	0.8914	0.920
GO/KEGG = GO:0007051: spindle organization	8	0.8849	0.913
GO/KEGG = GO:0000307: cyclin-dependent protein kinase holoenzyme complex	6	0.8840	0.912
gene2gene = CHAF1B	6	0.8790	0.907
gene2gene = NCAM1	7	0.8772	0.905
GO/KEGG = GO:0006221: pyrimidine nucleotide biosynthetic process	5	0.8754	0.903
gene2gene = HABP4	5	0.8715	0.899
gene2gene = LBR	6	0.8690	0.897
GO/KEGG = GO:0005881: cytoplasmic microtubule	8	0.8674	0.895
gene2gene = MIS12	7	0.8653	0.893
gene2gene = CDC7	14	0.8636	0.891
GO/KEGG = GO:0005762: mitochondrial large ribosomal subunit	8	0.8588	0.886
μtop-10	5.7	0.9320	0.962
μtop-25	6.72	0.8998	0.928

Safarii *event* = 1 **ranking, target=rank$_{\text{partial}}$, measure=φ_{wmw}, depth=3**			
pattern	**size**	φ_{wmw}	**norm.** φ_{wmw}
GO/KEGG = GO:0005634: nucleus	3128	9.5456	1.000
GO/KEGG = GO:0051301: cell division	144	8.9675	0.939
GO/KEGG = GO:0007049: cell cycle	343	8.8585	0.928
GO/KEGG = GO:0007067: mitosis	111	8.7379	0.915
pfam = Histone	66	8.2340	0.863
GO/KEGG = GO:0006260: DNA replication	98	7.5474	0.791
GO/KEGG = GO:0000166: nucleotide binding	1306	7.2598	0.761
chromosome = X	568	6.9724	0.730
GO/KEGG = GO:0005524: ATP binding	1025	6.8681	0.720
GO/KEGG = GO:0005739: mitochondrion	627	6.6376	0.695
GO/KEGG = GO:0005515: protein binding	3152	6.5782	0.689
GO/KEGG = GO:0005694: chromosome	106	6.3275	0.663
cytoband = 11q13.1	111	5.8157	0.609
cytoband = Xq28	85	5.7756	0.605
GO/KEGG = GO:0006281: DNA repair	124	5.7339	0.601
GO/KEGG = GO:0000775: chromosome, centromeric region	38	5.6870	0.596
GO/KEGG = GO:0016740: transferase activity	910	5.5247	0.579
gene2gene = CDC25A	16	5.4205	0.568
cytoband = 16p1	310	5.3880	0.564
cytoband = 16p	310	5.3880	0.564
cytoband = q28	93	5.3775	0.563
gene2gene = PCNA	65	5.3693	0.562
GO/KEGG = KEGG:04110: Cell cycle	96	5.3515	0.561
cytoband = Xq	350	5.3432	0.560
cytoband = q13.1	227	5.3019	0.555
μtop-10	741.6	7.9629	0.834
μtop-25	536.36	6.5605	0.687

Safarii $event = 1$ **ranking, target=rank$_{\text{partial}}$, measure=φ_{mmad}, depth=3**			
pattern	**size**	φ_{mmad}	**norm.** φ_{mmad}
GO/KEGG = GO:0005634: nucleus	3128	0.1577	1.000
GO/KEGG = GO:0016020: membrane	3466	0.1544	0.979
GO/KEGG = GO:0005515: protein binding	3152	0.1540	0.977
GO/KEGG = GO:0016021: integral to membrane	2543	0.1109	0.704
GO/KEGG = GO:0046872: metal ion binding	1597	0.0728	0.462
GO/KEGG = GO:0008270: zinc ion binding	1575	0.0713	0.452
GO/KEGG = GO:0000166: nucleotide binding	1306	0.0681	0.432
GO/KEGG = GO:0006355: regulation of transcription, DNA-dependent	1358	0.0646	0.410
GO/KEGG = GO:0005622: intracellular	1358	0.0636	0.404
chromosome = 1	1431	0.0625	0.397
GO/KEGG = GO:0005737: cytoplasm	1271	0.0615	0.390
GO/KEGG = GO:0007165: signal transduction	1302	0.0555	0.352
GO/KEGG = GO:0005524: ATP binding	1025	0.0531	0.337
GO/KEGG = GO:0006350: transcription	1086	0.0514	0.326
GO/KEGG = GO:0016740: transferase activity	910	0.0472	0.300
chromosome = 19	977	0.0451	0.286
GO/KEGG = GO:0003677: DNA binding	871	0.0441	0.280
chromosome = 11	932	0.0440	0.279
GO/KEGG = GO:0004872: receptor activity	1022	0.0419	0.266
chromosome = 17	837	0.0389	0.247
chromosome = 2	834	0.0381	0.242
GO/KEGG = GO:0005887: integral to plasma membrane	855	0.0373	0.237
chromosome = 6	764	0.0370	0.235
GO/KEGG = GO:0005739: mitochondrion	627	0.0360	0.228
chromosome = 3	785	0.0350	0.222
μtop-10	2091.4	0.0980	0.622
μtop-25	1400.48	0.0658	0.418

Safarii $event = 1$ **ranking, target=novelty, measure=**φ_{avg}**, depth=4**

pattern	**size**	φ_{avg}	**norm.** φ_{avg}
gene2gene = CDK3 ∧ gene2gene = CDC2	5	0.1190	1.000
gene2gene = CDC7 ∧ GO/KEGG = GO:0003677: DNA binding	7	0.1166	0.980
gene2gene = CDC25A ∧ gene2gene = CDK3	5	0.1162	0.977
gene2gene = MCM2 ∧ gene2gene = MCM6	5	0.1152	0.969
gene2gene = CDC7 ∧ gene2gene = MCM6	5	0.1147	0.964
gene2gene = CDC7 ∧ GO/KEGG = GO:0005515: protein binding	10	0.1139	0.957
GO/KEGG = GO:0006268: DNA unwinding during replication ∧ GO/KEGG = GO:0008094: DNA-dependent ATPase activity	5	0.1128	0.948
GO/KEGG = GO:0006268: DNA unwinding during replication ∧ GO/KEGG = GO:0005524: ATP binding	5	0.1128	0.948
GO/KEGG = GO:0006268: DNA unwinding during replication ∧ GO/KEGG = GO:0000166: nucleotide binding	5	0.1128	0.948
GO/KEGG = GO:0008094: DNA-dependent ATPase activity ∧ GO/KEGG = GO:0006268: DNA unwinding during replication	5	0.1128	0.948
pfam = MCM ∧ GO/KEGG = KEGG:04110: Cell cycle	6	0.1122	0.943
GO/KEGG = GO:0008094: DNA-dependent ATPase activity ∧ GO/KEGG = KEGG:04110: Cell cycle	6	0.1122	0.943
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling ∧ GO/KEGG = GO:0006260: DNA replication	6	0.1121	0.942
gene2gene = MCM6 ∧ GO/KEGG = GO:0006260: DNA replication	6	0.1120	0.941
gene2gene = MCM6 ∧ GO/KEGG = GO:0003677: DNA binding	6	0.1120	0.941
pfam = MCM ∧ gene2gene = MCM6	5	0.1120	0.941
pfam = MCM ∧ gene2gene = MCM10	5	0.1120	0.941
GO/KEGG = GO:0008094: DNA-dependent ATPase activity ∧ gene2gene = MCM6	5	0.1120	0.941
GO/KEGG = GO:0008094: DNA-dependent ATPase activity ∧ gene2gene = MCM10	5	0.1120	0.941
gene2gene = MCM6 ∧ pfam = MCM	5	0.1120	0.941
gene2gene = MCM6 ∧ GO/KEGG = GO:0008094: DNA-dependent ATPase activity	5	0.1120	0.941
gene2gene = MCM6 ∧ GO/KEGG = GO:0006270: DNA replication initiation	5	0.1120	0.941
gene2gene = MCM6 ∧ GO/KEGG = GO:0005524: ATP binding	5	0.1120	0.941
gene2gene = MCM6 ∧ GO/KEGG = GO:0000166: nucleotide binding	5	0.1120	0.941
gene2gene = MCM6 ∧ gene2gene = MCM10	5	0.1120	0.941
μ**top-10**	5.7	0.1147	0.964
μ**top-25**	5.48	0.1131	0.950

Safarii $event = 1$ **ranking, target=**$\text{rank}_{\text{partial}}$**, measure=**$\varphi_{avg}$**, depth=4**

pattern	**size**	φ_{avg}	**norm.** φ_{avg}
gene2gene = CDK3 ∧ gene2gene = CDC2	5	-151.1000	1.000
gene2gene = CDC25A ∧ gene2gene = CDK3	5	-243.5000	0.621
gene2gene = CDC7 ∧ GO/KEGG = GO:0003677: DNA binding	7	-354.2143	0.427
pfam = MCM ∧ GO/KEGG = KEGG:04110: Cell cycle	6	-409.3333	0.369
gene2gene = CDK3 ∧ gene2gene = CDK2	5	-414.8000	0.364
gene2gene = CDK3 ∧ GO/KEGG = GO:0005515: protein binding	6	-425.3333	0.355
gene2gene = CDC7 ∧ gene2gene = MCM6	5	-432.2000	0.350
GO/KEGG = GO:0006268: DNA unwinding during replication ∧ GO/KEGG = GO:0008094: DNA-dependent ATPase activity	5	-435.2000	0.347
GO/KEGG = GO:0006268: DNA unwinding during replication ∧ GO/KEGG = GO:0005524: ATP binding	5	-435.2000	0.347
GO/KEGG = GO:0006268: DNA unwinding during replication ∧ GO/KEGG = GO:0000166: nucleotide binding	5	-435.2000	0.347
pfam = MCM ∧ gene2gene = MCM6	5	-452.2000	0.334
pfam = MCM ∧ gene2gene = MCM10	5	-452.2000	0.334
pfam = MCM ∧ gene2gene = MCM2	5	-456.2000	0.331
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling ∧ GO/KEGG = GO:0006260: DNA replication	6	-466.2500	0.324
gene2gene = CDC7 ∧ pfam = MCM	5	-467.3000	0.323
gene2gene = CDC7 ∧ GO/KEGG = GO:0008094: DNA-dependent ATPase activity	5	-467.3000	0.323
gene2gene = CDC7 ∧ GO/KEGG = GO:0006355: regulation of transcription, DNA-dependent	5	-467.3000	0.323
gene2gene = CDC7 ∧ GO/KEGG = GO:0006350: transcription	5	-467.3000	0.323
pfam = MCM ∧ GO/KEGG = GO:0005515: protein binding	5	-467.3000	0.323
gene2gene = CDC7 ∧ GO/KEGG = GO:0005515: protein binding	10	-501.3000	0.301
gene2gene = MCM4	5	-536.2000	0.282
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling ∧ GO/KEGG = GO:0006281: DNA repair	5	-558.3000	0.271
GO/KEGG = GO:0007051: spindle organization ∧ GO/KEGG = GO:0007067: mitosis	5	-583.3000	0.259
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling ∧ GO/KEGG = GO:0007067: mitosis	5	-583.3000	0.259
GO/KEGG = GO:0007051: spindle organization ∧ GO/KEGG = GO:0048015: phosphoinositide-mediated signaling	7	-616.2143	0.245
μ**top-10**	5.4	-373.6081	0.453
μ**top-25**	5.48	-451.1218	0.363

Safarii $event = 1$ **ranking, target=novelty, measure=**φ_{mt} **and** φ_z**, depth=4**			
pattern	**size**	φ_{mt}/φ_z	**norm.** φ_{mt}/φ_z
GO/KEGG = GO:0007067: mitosis	111	0.2615/13.5572	1.000
GO/KEGG = GO:0005634: nucleus AND GO/KEGG = GO:0007049: cell cycle	230	0.2575/13.3507	0.985
GO/KEGG = GO:0005634: nucleus AND GO/KEGG = GO:0007067: mitosis	75	0.2488/12.9000	0.952
GO/KEGG = GO:0007049: cell cycle	343	0.2409/12.4894	0.921
GO/KEGG = GO:0006260: DNA replication	98	0.2396/12.4231	0.916
GO/KEGG = GO:0051301: cell division	144	0.2389/12.3895	0.914
GO/KEGG = GO:0007049: cell cycle AND GO/KEGG = GO:0005515: protein binding	179	0.2310/11.9791	0.884
GO/KEGG = GO:0005634: nucleus AND GO/KEGG = GO:0005515: protein binding	1105	0.2308/11.9693	0.883
GO/KEGG = GO:0051301: cell division AND GO/KEGG = GO:0007049: cell cycle	132	0.2286/11.8526	0.874
GO/KEGG = GO:0005634: nucleus AND GO/KEGG = GO:0006260: DNA replication	81	0.2238/11.6067	0.856
GO/KEGG = GO:0051301: cell division AND GO/KEGG = GO:0007067: mitosis	84	0.2235/11.5873	0.855
GO/KEGG = GO:0007049: cell cycle AND GO/KEGG = GO:0007067: mitosis	86	0.2227/11.5483	0.852
GO/KEGG = GO:0000166: nucleotide binding AND GO/KEGG = GO:0005634: nucleus	378	0.2213/11.4730	0.846
GO/KEGG = GO:0051301: cell division AND GO/KEGG = GO:0005634: nucleus	96	0.2198/11.3949	0.841
GO/KEGG = GO:0007067: mitosis AND GO/KEGG = GO:0005515: protein binding	51	0.2170/11.2525	0.830
GO/KEGG = KEGG:04110: Cell cycle AND GO/KEGG = GO:0005515: protein binding	66	0.2170/11.2497	0.830
GO/KEGG = GO:0006260: DNA replication AND GO/KEGG = GO:0005515: protein binding	47	0.2148/11.1372	0.821
pfam = Histone	66	0.2138/11.0849	0.818
GO/KEGG = GO:0005524: ATP binding AND GO/KEGG = GO:0005634: nucleus	276	0.2119/10.9885	0.811
GO/KEGG = GO:0051301: cell division AND GO/KEGG = GO:0005515: protein binding	73	0.2111/10.9457	0.807
GO/KEGG = GO:0005634: nucleus	3128	0.2105/10.9170	0.805
pfam = Histone AND cytoband = 6p22.1	42	0.2104/10.9095	0.805
GO/KEGG = GO:0000166: nucleotide binding AND GO/KEGG = GO:0007049: cell cycle	75	0.2084/10.8067	0.797
GO/KEGG = GO:0005524: ATP binding AND GO/KEGG = GO:0007049: cell cycle	72	0.2068/10.7216	0.791
GO/KEGG = KEGG:04110: Cell cycle AND GO/KEGG = GO:0005634: nucleus	76	0.2062/10.6902	0.789
μ**top-10**	249.8	0.2401/12.4518	0.918
μ**top-25**	284.56	0.2247/11.6490	0.859

Safarii $event = 1$ **ranking, target=rank**$_{\textbf{partial}}$**, measure=**φ_{mt} **and** φ_z**, depth=4**			
pattern	**size**	φ_{mt}/φ_z	**norm.** φ_{mt}/φ_z
GO/KEGG = GO:0005634: nucleus AND GO/KEGG = GO:0007049: cell cycle	230	46692.4457/9.3402	1.000
GO/KEGG = GO:0005515: protein binding AND GO/KEGG = GO:0005634: nucleus	1105	45416.0664/9.0848	0.973
GO/KEGG = GO:0051301: cell division	144	44642.8750/8.9302	0.956
GO/KEGG = GO:0007049: cell cycle	343	43843.9599/8.7704	0.939
GO/KEGG = GO:0007067: mitosis	111	43541.5790/8.7099	0.933
GO/KEGG = GO:0005634: nucleus	3128	43195.5870/8.6407	0.925
GO/KEGG = GO:0007049: cell cycle AND GO/KEGG = GO:0051301: cell division	132	42622.4789/8.5260	0.913
GO/KEGG = GO:0000166: nucleotide binding AND GO/KEGG = GO:0005634: nucleus	378	42224.9831/8.4465	0.904
GO/KEGG = GO:0005515: protein binding AND GO/KEGG = GO:0007049: cell cycle	179	41384.8084/8.2784	0.886
pfam = Histone	66	41084.4316/8.2184	0.880
GO/KEGG = GO:0005634: nucleus AND GO/KEGG = GO:0007067: mitosis	75	40701.2887/8.1417	0.872
GO/KEGG = GO:0005515: protein binding AND GO/KEGG = GO:0051301: cell division	73	40352.8615/8.0720	0.864
GO/KEGG = GO:0005634: nucleus AND GO/KEGG = GO:0051301: cell division	96	39622.4868/7.9259	0.849
pfam = Histone AND cytoband = 6p22.1	42	39142.9795/7.8300	0.838
GO/KEGG = GO:0005524: ATP binding AND GO/KEGG = GO:0005634: nucleus	276	38396.3756/7.6807	0.822
GO/KEGG = GO:0007067: mitosis AND GO/KEGG = GO:0051301: cell division	84	38253.8689/7.6521	0.819
pfam = Histone AND cytoband = p22.1	43	38043.9708/7.6102	0.815
GO/KEGG = GO:0007067: mitosis AND GO/KEGG = GO:0007049: cell cycle	86	37889.4172/7.5792	0.811
pfam = Histone AND cytoband = 6p22	44	37836.2064/7.5686	0.810
GO/KEGG = GO:0006260: DNA replication	98	37623.3841/7.5260	0.806
GO/KEGG = GO:0000166: nucleotide binding AND GO/KEGG = GO:0007049: cell cycle	75	36690.8983/7.3395	0.786
GO/KEGG = GO:0005524: ATP binding AND GO/KEGG = GO:0007049: cell cycle	72	36448.8208/7.2911	0.781
GO/KEGG = GO:0005515: protein binding AND GO/KEGG = GO:0007067: mitosis	51	35983.2774/7.1979	0.771
GO/KEGG = GO:0000166: nucleotide binding	1306	34897.4015/6.9807	0.747
GO/KEGG = GO:0006260: DNA replication AND GO/KEGG = GO:0005634: nucleus	81	34742.0000/6.9496	0.744
μ**top-10**	581.6	43464.9215/8.6945	0.931
μ**top-25**	332.72	40711.4442/8.0116	0.858

Safarii $event = 1$ **ranking, target=novelty, measure=**φ_t**, depth=4**

pattern	**size**	φ_t	**norm.** φ_t
GO/KEGG = GO:0007067: mitosis ∧ gene2gene = CDC20	6	15.5384	1.000
GO/KEGG = GO:0007067: mitosis ∧ gene2gene = E2F4	6	13.0422	0.839
gene2gene = HABP4	5	12.5422	0.807
GO/KEGG = GO:0005524: ATP binding ∧ gene2gene = HABP4	5	12.5422	0.807
GO/KEGG = GO:0000166: nucleotide binding ∧ gene2gene = HABP4	5	12.5422	0.807
gene2gene = CDC2 ∧ gene2gene = CDK3	5	12.3245	0.793
GO/KEGG = GO:0051301: cell division ∧ GO/KEGG = GO:0007089: traversing start control point of mitotic cell cycle	5	10.4610	0.673
GO/KEGG = GO:0007049: cell cycle ∧ GO/KEGG = GO:0007089: traversing start control point of mitotic cell cycle	5	10.4610	0.673
GO/KEGG = GO:0006221: pyrimidine nucleotide biosynthetic process	5	10.2137	0.657
GO/KEGG = KEGG:00240: Pyrimidine metabolism ∧ GO/KEGG = GO:0006221: pyrimidine nucleotide biosynthetic process	5	10.2137	0.657
GO/KEGG = GO:0016740: transferase activity ∧ GO/KEGG = GO:0000785: chromatin	5	9.9429	0.640
GO/KEGG = GO:0005634: nucleus ∧ pfam = Kinesin	5	9.6554	0.621
GO/KEGG = GO:0007067: mitosis ∧ gene2gene = PTMA	5	9.6149	0.619
GO/KEGG = GO:0005515: protein binding ∧ GO/KEGG = GO:0005634: nucleus	1105	9.6137	0.619
GO/KEGG = GO:0005634: nucleus	3128	9.5629	0.615
GO/KEGG = GO:0051301: cell division ∧ GO/KEGG = GO:0000776: kinetochore	6	9.4605	0.609
GO/KEGG = GO:0007049: cell cycle ∧ GO/KEGG = GO:0005634: nucleus	230	9.0725	0.584
pfam = Histone ∧ cytoband = 6p22.1	42	9.0257	0.581
gene2gene = CDC25A ∧ gene2gene = CDK3	5	8.9999	0.579
GO/KEGG = GO:0005694: chromosome ∧ cytoband = 6p22.1	21	8.9385	0.575
GO/KEGG = GO:0007049: cell cycle	343	8.8068	0.567
GO/KEGG = GO:0007067: mitosis	111	8.7241	0.561
GO/KEGG = GO:0051301: cell division	144	8.6704	0.558
GO/KEGG = GO:0005739: mitochondrion ∧ GO/KEGG = GO:0006730: one-carbon compound metabolic process	5	8.6606	0.557
GO/KEGG = GO:0005515: protein binding ∧ gene2gene = CDC7	10	8.6361	0.556
μ**top-10**	5.2	11.9881	0.772
μ**top-25**	208.68	10.2906	0.662

Safarii $event = 1$ **ranking, target=rank**$_{\textbf{partial}}$**, measure=**φ_t**, depth=4**

pattern	**size**	φ_t	**norm.** φ_t
gene2gene = CDC2 ∧ gene2gene = CDK3	5	137.3102	1.000
GO/KEGG = GO:0007067: mitosis ∧ gene2gene = CDC20	6	136.0930	0.991
GO/KEGG = GO:0005634: nucleus ∧ pfam = Kinesin	5	85.4684	0.622
GO/KEGG = GO:0007067: mitosis ∧ gene2gene = E2F4	6	79.9513	0.582
gene2gene = CDC25A ∧ gene2gene = CDK3	5	63.3638	0.461
GO/KEGG = GO:0007067: mitosis ∧ pfam = Kinesin	5	59.0795	0.430
GO/KEGG = GO:0007067: mitosis ∧ GO/KEGG = GO:0007018: microtubule-based movement	5	59.0795	0.430
GO/KEGG = GO:0006260: DNA replication ∧ gene2gene = CDK2	5	52.8357	0.385
GO/KEGG = GO:0051301: cell division ∧ GO/KEGG = GO:0000776: kinetochore	6	52.5605	0.383
GO/KEGG = GO:0007049: cell cycle ∧ GO/KEGG = GO:0000776: kinetochore	5	46.1258	0.336
GO/KEGG = GO:0006260: DNA replication ∧ chromosome = 11	6	40.7649	0.297
GO/KEGG = GO:0007067: mitosis ∧ gene2gene = PTMA	5	39.2927	0.286
GO/KEGG = GO:0005515: protein binding ∧ gene2gene = CDK3	6	38.8905	0.283
GO/KEGG = GO:0003677: DNA binding ∧ gene2gene = CDC7	7	36.8713	0.269
gene2gene = CDK3 ∧ gene2gene = CDK2	5	34.9745	0.255
gene2gene = DBF4	6	34.0101	0.248
GO/KEGG = GO:0005634: nucleus ∧ gene2gene = DBF4	6	34.0101	0.248
GO/KEGG = GO:0005515: protein binding ∧ gene2gene = DBF4	6	34.0101	0.248
GO/KEGG = GO:0007067: mitosis ∧ gene2gene = BIRC5	5	33.4434	0.244
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling ∧ GO/KEGG = GO:0006260: DNA replication	6	32.3642	0.236
GO/KEGG = GO:0003677: DNA binding ∧ gene2gene = MCM6	6	31.9903	0.233
GO/KEGG = GO:0006260: DNA replication ∧ gene2gene = MCM6	6	31.9903	0.233
pfam = MCM ∧ GO/KEGG = KEGG:04110: Cell cycle	6	31.9618	0.233
GO/KEGG = GO:0007051: spindle organization ∧ GO/KEGG = GO:0048015: phosphoinositide-mediated signaling	7	31.6426	0.230
GO/KEGG = GO:0005515: protein binding ∧ gene2gene = CDC7	10	31.1060	0.227
μ**top-10**	5.3	77.1868	0.562
μ**top-25**	5.84	51.5676	0.376

Safarii *event* = 1 **ranking, target=novelty, measure=φ_{χ^2}, depth=4**			
pattern	**size**	φ_{χ^2}	**norm.** φ_{χ^2}
pfam = zf-UBR	5	17307	1
pfam = zf-C2HC	5	17307	1
pfam = zf-AN1	5	17307	1
pfam = WSC	5	17307	1
pfam = Vps4_C	5	17307	1
pfam = UPF0020	5	17307	1
pfam = Tub	5	17307	1
pfam = tRNA-synt_2	5	17307	1
pfam = Trefoil	5	17307	1
pfam = TRAM_LAG1_CLN8	5	17307	1
pfam = TPR_3	5	17307	1
pfam = Tim17	5	17307	1
pfam = TFR_dimer	5	17307	1
pfam = TFIIS_M	5	17307	1
pfam = TFIIS	5	17307	1
pfam = Tektin	5	17307	1
pfam = TB	5	17307	1
pfam = SWIRM	5	17307	1
pfam = Surp	5	17307	1
pfam = Sulfotransfer_2	5	17307	1
pfam = SRF-TF	5	17307	1
pfam = SPAN-X	5	17307	1
pfam = Somatomedin_B	5	17307	1
pfam = Sel1	5	17307	1
pfam = SCP2	5	17307	1
μtop-10	5	17307	1
μtop-25	5	17307	1

Safarii *event* = 1 **ranking, target=rank$_{\text{partial}}$, measure=φ_{χ^2}, depth=4**			
pattern	**size**	φ_{χ^2}	**norm.** φ_{χ^2}
pfam = zf-UBR	5	17307	1
pfam = zf-C2HC	5	17307	1
pfam = zf-AN1	5	17307	1
pfam = WSC	5	17307	1
pfam = Vps4_C	5	17307	1
pfam = UPF0020	5	17307	1
pfam = Tub	5	17307	1
pfam = tRNA-synt_2	5	17307	1
pfam = Trefoil	5	17307	1
pfam = TRAM_LAG1_CLN8	5	17307	1
pfam = TPR_3	5	17307	1
pfam = Tim17	5	17307	1
pfam = TFR_dimer	5	17307	1
pfam = TFIIS_M	5	17307	1
pfam = TFIIS	5	17307	1
pfam = Tektin	5	17307	1
pfam = TB	5	17307	1
pfam = SWIRM	5	17307	1
pfam = Surp	5	17307	1
pfam = Sulfotransfer_2	5	17307	1
pfam = SRF-TF	5	17307	1
pfam = SPAN-X	5	17307	1
pfam = Somatomedin_B	5	17307	1
pfam = Sel1	5	17307	1
pfam = SCP2	5	17307	1
μtop-10	5	17307	1
μtop-25	5	17307	1

Safarii *event* = 1 **ranking, target=rank, measure=φ_{roc}, depth=4**

pattern	**size**	φ_{roc}	**norm.** φ_{roc}
gene2gene = CDK3 ∧ gene2gene = CDC2	5	0.9914	1.000
gene2gene = CDC25A ∧ gene2gene = CDK3	5	0.9861	0.995
gene2gene = CDC7 ∧ GO/KEGG = GO:0003677: DNA binding	7	0.9798	0.988
pfam = MCM ∧ GO/KEGG = KEGG:04110: Cell cycle	6	0.9766	0.985
gene2gene = CDK3 ∧ gene2gene = CDK2	5	0.9761	0.985
gene2gene = CDK3 ∧ GO/KEGG = GO:0005515: protein binding	6	0.9756	0.984
gene2gene = CDC7 ∧ gene2gene = MCM6	5	0.9753	0.984
GO/KEGG = GO:0006268: DNA unwinding during replication ∧ GO/KEGG = GO:0008094: DNA-dependent ATPase activity	5	0.9751	0.984
GO/KEGG = GO:0006268: DNA unwinding during replication ∧ GO/KEGG = GO:0005524: ATP binding	5	0.9751	0.984
GO/KEGG = GO:0006268: DNA unwinding during replication ∧ GO/KEGG = GO:0000166: nucleotide binding	5	0.9751	0.984
pfam = MCM ∧ gene2gene = MCM6	5	0.9741	0.983
pfam = MCM ∧ gene2gene = MCM10	5	0.9741	0.983
pfam = MCM ∧ gene2gene = MCM2	5	0.9739	0.982
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling ∧ GO/KEGG = GO:0006260: DNA replication	6	0.9733	0.982
gene2gene = CDC7 ∧ pfam = MCM	5	0.9732	0.982
gene2gene = CDC7 ∧ GO/KEGG = GO:0008094: DNA-dependent ATPase activity	5	0.9732	0.982
gene2gene = CDC7 ∧ GO/KEGG = GO:0006355: regulation of transcription, DNA-dependent	5	0.9732	0.982
gene2gene = CDC7 ∧ GO/KEGG = GO:0006350: transcription	5	0.9732	0.982
pfam = MCM ∧ GO/KEGG = GO:0005515: protein binding	5	0.9732	0.982
gene2gene = CDC7 ∧ GO/KEGG = GO:0005515: protein binding	10	0.9715	0.980
gene2gene = MCM4	5	0.9692	0.978
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling ∧ GO/KEGG = GO:0006281: DNA repair	5	0.9680	0.976
GO/KEGG = GO:0007051: spindle organization ∧ GO/KEGG = GO:0007067: mitosis	5	0.9665	0.975
GO/KEGG = GO:0048015: phosphoinositide-mediated signaling ∧ GO/KEGG = GO:0007067: mitosis	5	0.9665	0.975
gene2gene = DBF4	6	0.9646	0.973
μtop-10	5.4	0.9786	0.987
μtop-25	5.44	0.9742	0.983

Safarii *event* = 1 **ranking, target=rank$_{\text{partial}}$, measure=φ_{wmw}, depth=4**

pattern	**size**	φ_{wmw}	**norm.** φ_{wmw}
GO/KEGG = GO:0005634: nucleus	3128	9.5456	1.000
GO/KEGG = GO:0007049: cell cycle ∧ GO/KEGG = GO:0005634: nucleus	230	9.4028	0.985
GO/KEGG = GO:0005515: protein binding ∧ GO/KEGG = GO:0005634: nucleus	1105	9.3893	0.984
GO/KEGG = GO:0051301: cell division	144	8.9675	0.939
GO/KEGG = GO:0007049: cell cycle	343	8.8585	0.928
GO/KEGG = GO:0007067: mitosis	111	8.7379	0.915
GO/KEGG = GO:0007049: cell cycle ∧ GO/KEGG = GO:0051301: cell division	132	8.5587	0.897
GO/KEGG = GO:0000166: nucleotide binding ∧ GO/KEGG = GO:0005634: nucleus	378	8.5402	0.895
GO/KEGG = GO:0005515: protein binding ∧ GO/KEGG = GO:0007049: cell cycle	179	8.3215	0.872
pfam = Histone	66	8.2340	0.863
GO/KEGG = GO:0007067: mitosis ∧ GO/KEGG = GO:0005634: nucleus	75	8.1594	0.855
GO/KEGG = GO:0005515: protein binding ∧ GO/KEGG = GO:0051301: cell division	73	8.0890	0.847
GO/KEGG = GO:0051301: cell division ∧ GO/KEGG = GO:0005634: nucleus	96	7.9479	0.833
pfam = Histone ∧ cytoband = 6p22.1	42	7.8395	0.821
GO/KEGG = GO:0005524: ATP binding ∧ GO/KEGG = GO:0005634: nucleus	276	7.7426	0.811
GO/KEGG = GO:0007067: mitosis ∧ GO/KEGG = GO:0051301: cell division	84	7.6707	0.804
pfam = Histone ∧ cytoband = p22.1	43	7.6196	0.798
GO/KEGG = GO:0007067: mitosis ∧ GO/KEGG = GO:0007049: cell cycle	86	7.5981	0.796
pfam = Histone ∧ cytoband = 6p22	44	7.5782	0.794
pfam = Histone ∧ cytoband = 6p2	44	7.5782	0.794
pfam = Histone ∧ cytoband = 6p	44	7.5782	0.794
pfam = Histone ∧ chromosome = 6	44	7.5782	0.794
GO/KEGG = GO:0006260: DNA replication	98	7.5474	0.791
GO/KEGG = GO:0000166: nucleotide binding ∧ GO/KEGG = GO:0007049: cell cycle	75	7.3554	0.771
GO/KEGG = GO:0005524: ATP binding ∧ GO/KEGG = GO:0007049: cell cycle	72	7.3062	0.765
μtop-10	581.6	8.8556	0.928
μtop-25	280.48	8.1498	0.854

Safarii $event = 1$ **ranking, target=rank$_{\text{partial}}$, measure=φ_{mmad}, depth=4**			
pattern	**size**	φ_{mmad}	**norm.** φ_{mmad}
GO/KEGG = GO:0005634: nucleus	3128	0.1577	1.000
GO/KEGG = GO:0016020: membrane	3466	0.1544	0.979
GO/KEGG = GO:0005515: protein binding	3152	0.1540	0.977
GO/KEGG = GO:0016021: integral to membrane	2543	0.1109	0.704
GO/KEGG = GO:0016021: integral to membrane ∧ GO/KEGG = GO:0016020: membrane	2234	0.0980	0.622
GO/KEGG = GO:0046872: metal ion binding	1597	0.0728	0.462
GO/KEGG = GO:0008270: zinc ion binding	1575	0.0713	0.452
GO/KEGG = GO:0000166: nucleotide binding	1306	0.0681	0.432
GO/KEGG = GO:0006355: regulation of transcription, DNA-dependent	1358	0.0646	0.410
GO/KEGG = GO:0005622: intracellular	1358	0.0636	0.404
chromosome = 1	1431	0.0625	0.397
GO/KEGG = GO:0008270: zinc ion binding ∧ GO/KEGG = GO:0046872: metal ion binding	1363	0.0616	0.391
GO/KEGG = GO:0005515: protein binding ∧ GO/KEGG = GO:0005634: nucleus	1105	0.0615	0.390
GO/KEGG = GO:0005737: cytoplasm	1271	0.0615	0.390
GO/KEGG = GO:0006355: regulation of transcription, DNA-dependent ∧ GO/KEGG = GO:0005634: nucleus	1259	0.0595	0.378
GO/KEGG = GO:0007165: signal transduction	1302	0.0555	0.352
GO/KEGG = GO:0005524: ATP binding	1025	0.0531	0.337
GO/KEGG = GO:0006350: transcription	1086	0.0514	0.326
GO/KEGG = GO:0006350: transcription ∧ GO/KEGG = GO:0005634: nucleus	1050	0.0497	0.315
GO/KEGG = GO:0005524: ATP binding ∧ GO/KEGG = GO:0000166: nucleotide binding	899	0.0475	0.301
GO/KEGG = GO:0016740: transferase activity	910	0.0472	0.300
chromosome = 19	977	0.0451	0.286
GO/KEGG = GO:0003677: DNA binding	871	0.0441	0.280
chromosome = 11	932	0.0440	0.279
GO/KEGG = GO:0006350: transcription ∧ GO/KEGG = GO:0006355: regulation of transcription, DNA-dependent	905	0.0425	0.269
μ**top-10**	2171.7	0.1015	0.644
μ**top-25**	1524.12	0.0721	0.457

C Tables of Distributions

This appendix contains several probability distributions, all of which can be used to compute the level of significance α_0. For all the distributions, the value of the statistic (z-score, t-score, χ^2-test) is required. Safarii can calculate these statistics, given the appropriate quality measures. In Chapter 7, it is defined which distribution to use lookup the level of significance, given a quality measure. How do these tables work? The statistic gives some value X, and the distributions can tell which p-value belongs to this value such that: $P(X < x) = p$. The significance value α_0 can then be calculated as follows: $\alpha_0 = 1 - p$. Given hypothesis H_0, which states that two distributions, such as the subgroup distribution and the population distribution, are the same, then H_0 will be rejected if $|X| > x$, with significance level α_0.

C.1 Table of the Normal Distribution (Z-Values)

The table of the normal distribution gives the p-value given the z-value. To find the p-value, split the z-value in two between the first and second decimal. The p-value can be found at the intersection of the row of the first two digits and the column of the third (rounded) digit.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

C.2 Table of the t Distribution

If X has a t distribution with degrees of freedom df (depicted in the first column), then the table gives the value of x, such that probability of $X < x$ is p: $P(X < x) = p$.

df	p=0.90	0.95	0.975	0.99	0.995	0.999	0.9995
1	3.078	6.314	12.71	31.82	63.66	318.3	637
2	1.886	2.920	4.303	6.965	9.925	22.330	31.6
3	1.638	2.353	3.182	4.541	5.841	10.210	12.92
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
32	1.309	1.694	2.037	2.449	2.738	3.365	3.622
34	1.307	1.691	2.032	2.441	2.728	3.348	3.601
36	1.306	1.688	2.028	2.434	2.719	3.333	3.582
38	1.304	1.686	2.024	2.429	2.712	3.319	3.566
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
42	1.302	1.682	2.018	2.418	2.698	3.296	3.538
44	1.301	1.680	2.015	2.414	2.692	3.286	3.526
46	1.300	1.679	2.013	2.410	2.687	3.277	3.515
48	1.299	1.677	2.011	2.407	2.682	3.269	3.505
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
55	1.297	1.673	2.004	2.396	2.668	3.245	3.476
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
65	1.295	1.669	1.997	2.385	2.654	3.220	3.447
70	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
150	1.287	1.655	1.976	2.351	2.609	3.145	3.357
200	1.286	1.653	1.972	2.345	2.601	3.131	3.340
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

C.3 Table of the χ^2 Distribution

If X has a χ^2 distribution with degrees of freedom df (depicted in the first column), then the table gives the value of x, such that probability of $X < x$ is p: $P(X < x) = p$. For the use in this thesis, the χ^2 distribution always has df 1, thus, only the top row is needed. The rest of the table is given for the sake of completeness.

df	p=0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

References

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, New York, NY, USA, 1993. ACM.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.

[3] M. Atzmueller. Subgroup discovery. *Künstliche Intelligenz*, (4):52–53, 2005. `http://ki.informatik.uni-wuerzburg.de/papers/atzmueller/2005/2005-SDSchlagwortKI_AtzmuellerM.pdf`.

[4] J. H. Beder and R. C. Heim. On the use of ridit analysis. volume 55, pages 603–616. Springer New York, 1990.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, chapter 1. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[6] S. Boslaugh and D. P. A. Watters. *Statistics in a nutshell*, chapter 7,8,10,11. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2008.

[7] G. M. Brodeur. Neuroblastoma: Biological insights into a clinical enigma. *Nature Reviews*, 3:203–216, 2003.

[8] K. De Preter, S. De Brouwer, T. Van Maerken, F. Pattyn, A. Schramm, A. Eggert, J. Vandesompele, and F. Speleman. Meta-mining of neuroblastoma and neuroblast gene expression profiles reveals candidate therapeutic compounds. 15:3690 – 3696, 2009.

[9] K. De Preter, J. Vandesompele, P. Heimann, N. Yigit, S. Beckman, A. Schramm, A. Eggert, R. L. Stallings, Y. Benoit, M. Renard, A. De Paepe, G. Laureys, S. Phlman, and F. Speleman. Human fetal neuroblast and neuroblastoma transcriptome analysis confirms neuroblast origin and highlights neuroblastoma candidate genes. *Genome Biology*, 7:R84, 2006. `http://genomebiology.com/2006/7/9/R84`.

[10] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*, chapter 8,9. Addison-Wesley, 2002.

[11] S. Džeroski. Multi-relational data mining: an introduction. *SIGKDD Explor. Newsl.*, 5(1):1–16, 2003.

[12] Ensembl, 2009. `http://www.ensembl.org`.

[13] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, and A. Bateman. The pfam protein families database, 2008. `http://pfam.sanger.ac.uk`.

[14] P. Flach and N. Lavrač. Rule induction. pages 229–267, 2003.

[15] E. Frank and M. Hall. *A Simple Approach to Ordinal Classification*, volume 2167/2001, pages 145–156. Springer Berlin/Heidelberg, 2001.

[16] J. Fürnkranz and P. A. Flach. Roc 'n' rule learning: towards a better understanding of covering algorithms. *Mach. Learn.*, 58(1):39–77, 2005.

[17] D. Gamberger and N. Lavrač. Expert-guided subgroup discovery: methodology and application. *J. Artif. Int. Res.*, 17(1):501–527, 2002.

[18] The gene ontology, 2009. `http://www.geneontology.org`.

[19] R. Göb, C. McCollin, and M. F. Ramalhoto. Ordinal methodology in the analysis of likert scales. *Quality and Quantity*, 41(5):601–626, 2007.

[20] H. Grosskreutz. Cascaded subgroups discovery with an application to regression. In *ECML PKDD '08: Proceedings of the 19th European Conference on Machine Learning and 12th European Symposium on Principles and Practice of Knowledge Discovery in Databases*, 2008.

[21] D. J. Hand and R. J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.*, 45(2):171–186, 2001.

[22] Institute jožef stefan. `http://www.ijs.si`.

[23] B. Kavšek, N. Lavrač, and V. Jovanoski. *APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery*, volume 2779/2003, pages 230–241. Springer Berlin/Heidelberg, 2003.

[24] W. Klösgen. Explora: a multipattern and multistrategy discovery assistant. pages 249–271, 1996.

[25] A. Knobbe. *Multi-Relational Data Mining*. PhD thesis, Utrecht University, 2004. `http://www.kiminkii.com/thesis.pdf`.

[26] W. Kotlowski, K. Dembczynski, S. Greco, and R. Slowinski. Stochastic dominance-based rough set model for ordinal classification. *Information Sciences*, 178(21):4019–4037, 2008.

[27] S. Kramer, G. Widmer, B. Pfahringer, and M. De Groeve. Prediction of ordinal classes using regression trees. *Fundam. Inf.*, 47(1-2):1–13, 2001.

[28] N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach. Decision support through subgroup discovery: Three case studies and the lessons learned. *Mach. Learn.*, 57(1-2):115–143, 2004.

[29] N. Lavrač, P. Flach, B. Kavšek, and L. Todorovski. Rule induction for subgroup discovery with cn2-sd. In *Proc. Integrating Aspects of Data Mining, Decision Support and Meta-Learning, Workshop at the ECML/PKDD-2002 Conference*, 2002.

[30] N. Lavrač, P. A. Flach, and B. Zupan. Rule evaluation measures: A unifying view. In *ILP '99: Proceedings of the 9th International Workshop on Inductive Logic Programming*, pages 174–185, London, UK, 1999. Springer-Verlag.

[31] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with cn2-sd. *J. Mach. Learn. Res.*, 5:153–188, 2004.

[32] A. Oberthuer, F. Berthold, P. Warnat, B. Hero, Y. Kahlert, R. Spitz, K. Ernestus, R. König, S. Haas, R. Eils, M. Schwab, B. Brors, F. Westermann, and M. Fischer. Customized oligonucleotide microarray geneexpression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology*, 24:5070–5078, 2006.

[33] A. B. Olshen, E. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572.

[34] Safarii multi relational data mining environment, 2008. `http://www.kiminkii.com/safarii.html`.

[35] P. Schattner. *Genomes, Browsers & Databases: Data-Mining Tools for Integrated Genomic Databases*. Cambridge University Press, 2008.

[36] M. Scholz. *Knowledge-Based Sampling for Subgroup Discovery*, volume 3539/2005, pages 171–189. Springer Berlin / Heidelberg, 2005.

[37] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*, chapter 4,6, pages 327–403. Pearson Education Inc., Boston, 2006.

[38] I. Trajkovski. *Functional Interpretation of Gene Expression Data*. PhD thesis, Jožef Stefan International Postgraduate School, 2007. http://cs.nyus.edu.mk/trajkovski/data/phd_thesis.html.

[39] I. Trajkovski. Search for enriched gene sets, 2007. http://kt.ijs.si/software/SEGS.

[40] I. Trajkovski, N. Lavrač, and J. Tolar. Segs: Search for enriched gene sets in microarray data. *J. of Biomedical Informatics*, 41(4):588–601, 2008.

[41] I. Trajkovski, F. Železný, N. Lavrač, and J. Tolar. Learning relational descriptions of differentially expressed gene groups. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(1):16–25, 2008.

[42] E. van de Koppel, I. Slavkov, K. Astrahantseff, A. Schramm, J. Schulte, J. Vandesompele, E. de Jong, S. Džeroski, and A. Knobbe. Knowledge discovery in neuroblastoma-related biological data. In *PKDD '07: Proceedings of the 11th European Symposium on Principles and Practice of Knowledge Discovery in Databases*, pages 45 – 56, 2007.

[43] J. Vandesompele, M. Baudis, K. De Preter, N. Van Roy, P. Ambros, N. Bown, C. Brinkschmidt, H. Christiansen, V. Combaret, M. Lastowska, J. Nicholson, A. O'Meara, D. Plantaz, R. Stallings, B. Brichard, C. Van den Broecke, S. De Bie, A. De Paepe, G. Laureys, and F. Speleman. Unequivocal delineation of clinicogenetic subgroups and development of a new model for improved outcome prediction in neuroblastoma. *Journal of Clinical Oncology*, 23:2280–2299, 2005.

[44] E. Venkatraman and A. B. Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663.

[45] G. I. Webb. Discovering associations with numeric variables. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 383–388, New York, NY, USA, 2001. ACM.

[46] Wikipedia. Array comparative genomic hybridization, 2008. http://en.wikipedia.org/wiki/Array_comparative_genomic_hybridization.

[47] Wikipedia. Dna microarray, 2008. http://en.wikipedia.org/wiki/DNA_Microarray.

[48] Wikipedia. Affymetrix – dna microarrays, 2009. http://en.wikipedia.org/wiki/Affymetrix.

[49] Wikipedia. Neuroblastoma, 2009. http://en.wikipedia.org/wiki/Neuroblastoma.

[50] Wikipedia. qpcr – real-time polymerase chain reaction, 2009. http://en.wikipedia.org/wiki/Real-time_polymerase_chain_reaction.

[51] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *PKDD '97: Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87, London, UK, 1997. Springer-Verlag.