UNIVERSITY OF TARTU
Institute of Computer Science
Software Engineering Curriculum

Vitalii Peretiatko

Using Robust Rank Aggregation for prioritising autoimmune targets on protein microarrays

Master's Thesis (30 ECTS)

Supervisor(s): Dmytro Fishman, M.Sc
Elena Sügis, M.Sc

Tartu 2016

Using Robust Rank Aggregation for prioritising autoimmune targets on protein microarrays

Abstract:

Autoimmune diseases are very common in the modern world. More and more diseases associated with an autoimmune process. Autoimmune reaction is a process in which the immune system produces antibodies (autoantibodies) that attack organism's own cells. Causes and mechanisms of autoimmune diseases are yet to be understood. One of the ways to study autoimmunity is to explore reasons why certain cells and particularly proteins were attacked by autoantibodies. To achieve this, many technologies have been developed and one of which is Protein microarray. This technology allows estimating the amount of autoantibodies in patient serum against 9000 unique human proteins. Consequently, applying methods of data analysis on this data, bioinformaticians might be able to identify proteins that attract prevalent amount of autoantibodies. Knowing these proteins, biologists could conduct experiments and formulate new hypotheses about mechanisms of work and appearance of autoimmune diseases. Common data analysis methods focused on how to select only the most reliably differing proteins between healthy and diseased groups. Moreover, ignoring the fact that in the case of an autoimmune disease - the repertoire of the affected proteins can differ greatly between patients. So even single cases of high protein reactivity may carry important information for understanding the mechanisms of disease. In this thesis, we propose to apply Robust Rank Aggregation algorithm as an adaptive method to identify a wide repertoire of reactive proteins. We compared expediency and effectiveness of the classical methods of analysis, method recently applied by biologists and RRA on synthetic and real data. Experiments on synthetic data sets with known reactive proteins show that RRA outperforms these methods while also being more robust to incorporated noise. Applying RRA on real data and conducting an enrichment analysis on lists of reactive proteins for each method, we got comparable numbers of proteins overrepresented in the classes associated with biological and immune responses.

Keywords:

Autoimmunity, autoimmune disease, biomarker, reactive protein, robust rank aggregation, enrichment analysis

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Robust Rank Aggregation meetodi rakendamine autoimmuunsete sihtmärkide prioritiseerimiseks valgukiipidel

Lühikokkuvõte:

Autoimmuunhaigused on tänapäeva maailmas väga sagedased. Üha enam ja enam haigusi on seotud autoimmuunsete protsessidega. Autoimmuunreaktsioon on protsess, mille käigus immuunsüsteem toodab antikehasid (autoantikehad) organismi enda rakkude vastu. Autoimmuunhaiguste põhjused ja mehhanismid on aga veel selgeks tegemata. Üheks võimaluseks, kuidas autoimmuunhaigusi õppida on välja selgitada, miks kindlad rakud ja iseäranis just valgud on autoantikehade märklauaks. Selle eesmärgi saavutamiseks on välja töötatud mitmesuguseid tehnoloogiaid, kuhu kuuluvad ka valgukiibid. See tehnoloogia võimaldab hinnata autoantikehade kogust patsiendi seerumis 9000 unikaalse inimese valgu vastu. Seega, rakendades andmeanalüüsi meetodeid on bioinformaatikud võimelised tuvastama autoantikehade märklaudvalke. Teades neid valke, saavad bioloogid läbi viia edasisi katseid ning formuleerida uusi hüpoteese autoimmuunhaiguste mehhanismide ja esinemise kohta. Traditsioonilised andmeanalüüsi meetodid keskenduvad ainult selliste valkude leidmisele, mis erinevad kõige kindlamalt tervete ja patsientide grupi vahel. Need meetodid aga jätavad kõrvale fakti, et märklaudvalkude repertuaar võib patsientide vahel oluliselt erineda. Seega võib isegi üksikjuhtum sisaldada olulist informatsiooni haiguse mehhanismide mõistmisel. Käesolevas lõputöös pakume välja, et Robust Rank Aggregation (RRA) algoritmi saab kasutada adaptiivse meetodina leidmaks reaktiivsete valkude (märklaudvalkude) laia repertuaari. Me võrdlesime klassikaliste analüüsimeetodite otstarbekust ja efektiivsust RRA-ga nii sünteetilistel kui ka pärisandmetel. Katsed sünteetilise andmehulgaga ehk andmehulgaga, mille puhul on reaktiivsed valgud teada näitavad, et RRA ületab teisi meetodeid olles samal ajal vähem mõjutatud "mürast". Rakendades RRA-d pärisandmetel ning viies läbi rikastusanalüüsi iga meetodi kohta saadud reaktiivsete valkude listidega, saime me sarnase arvu valke, mis olid bioloogilise ja immuunvastusega seotud klassides üleesindatud.

Võtmesõnad:

Autoimmuunsus, autoimmuunhaigus, biomarker, reaktiivne valk, robust rank aggregation, rikastusanalüüs

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetria

Acknowledgments

I would like to express my sincere gratitude to Dmytro Fishman. Firstly, for introducing me the world of Data Science as well as fostering interest in this field and its opportunities. Secondly, for his invaluable support, shared knowledge and experience, brilliant ideas, useful criticism and enduring patience alongside work on this thesis. I am thankful for his guidance and being constantly approachable throughout this research journey. Without him this work would not have been done.

I am truly grateful to Elena Sügis for her support and bright ideas used in the thesis.

I would like to sincerely thank Priit Adler for his wise and valuable comments.

I would like to show my gratitude to Institute of Computer Science of Tartu University, particularly, to all staff of Software Engineering program and BIIT research group for support, professionalism and inspiration during my studying.

Last but not least, I would like to thank Kateryna Shokalo and my relatives for being invaluably supportive during all the ups and downs.

Table of Contents

1 Introduction

1.1 Autoimmunity

Over the past decade, understanding of autoimmune disorders has been improved significantly [1]. Such diseases are caused by an autoimmune reaction process, which occurs when the human immune system starts to produce antibodies that attack body's own healthy cells and tissues[1]. These antibodies are called autoantibodies.

According to American Autoimmune Related Diseases Association, there are more than 80 distinct types of autoimmune disorders, which are associated with an autoimmune response[2]. Typically, women are more frequently affected by autoimmune diseases than men [2]. The number of registered cases throughout human population has been increased over past 30 years [1]. Multiple Sclerosis International Federation indicated that there are over 1.6 million cases of the eponymous disease[3]. Other autoimmune disorders such as Rheumatoid Arthritis, Celiac Disease and Autoimmune Hepatitis are also widely spread across the planet. However, causes of autoimmune reaction are still unexplored [2, 3].

To investigate the mechanisms of autoimmune disease as well as to obtain diagnosis indicators for the each of such disease, scientists must know and understand which human cells are exposed to attacks of the immune system. For identifying target proteins, high-throughput protein microarray technology[4] is used. Such technique allows getting a quantitative measure of the autoantibodies presence in human serum samples against each of approximately 9000 unique human proteins. These proteins are immobilized at the protein microarray surface. Subsequently, data analysis methods and techniques are applied to identify a list of significantly reactive proteins.

1.2 Problem and Motivation

Currently, to identify highly reactive proteins, it is a common practice to apply the following methods: t-statistics [4, 5], M-statistics [6] and method based on the standard score [7]. T-statistics is used in DNA microarray analysis to identify differentially expressed genes. M-statistics was specifically developed for protein microarrays. These methods are aimed to

[1] https://en.wikipedia.org/wiki/Autoimmunity
[2] http://www.aarda.org/
[3] https://www.msif.org
[4] https://en.wikipedia.org/wiki/Protein_microarray

discover statistically reliable autoimmune biomarkers, which show significantly bigger reactivity in disease patients on average than in healthy cohort. In other words, such methods might discard high protein reactivity shown in a small proportion of samples because the statistical significance of protein reactivity highly depends on average reactivity across sample groups. A different approach was applied by biologists to address this issue [7]. However, proposed method that is based on standard score[5] (also known as z-score) produces a large number of type I errors – false positives[6]. Using this approach, biologists get an extensive set of proteins. However, they compelled to conduct additional experiments to identify proteins to satisfy their biological hypotheses.

Overall, reactive proteins might help to provide biological knowledge to diagnose autoimmune diseases on early stages. To investigate underlying causes of autoimmune disorders, it is crucial to have a wider repertoire of such proteins. Therefore, in this thesis, we study the claim that common methods are not well suitable to identify proteins that attract autoantibodies only in few patient samples. Consequently, there is a need for a method that defines less false positives than z-score based method as well as a wider repertoire of reactive protein than t-test and M-statistics.

1.3 Contribution

To find significantly reactive proteins expressed in a very small proportion of samples we propose to apply Robust Rank Aggregation algorithm (RRA) [8]. This method is based on order statistics[7]. Therefore, it firstly treats investigated samples separately producing ranking lists. Afterwards, RRA aggregates all the ranking lists producing a final ranking that outputs statistical significance of rankings per protein. Thereby, it allows getting proteins that are expressed in few samples.

We conducted various experiments to show that RRA applies to the autoimmunity related protein microarray analysis. We compared RRA with differential analysis methods, such as t-test and m-statistics as well as with the approach based on the z-score. Moreover, we also picked for experiments another order statistics based – method proposed by Stuart et al. [9]. These tests are done on two different data sets: synthetic and real biological data. Also, we

[5] https://en.wikipedia.org/wiki/Standard_score
[6] https://en.wikipedia.org/wiki/Type_I_and_type_II_errors
[7] https://en.wikipedia.org/wiki/Order_statistic

experimented with RRA resistance to noise in different levels and compared results with other methods.

1.4 Outline

Background chapter provides a comprehensive description of the problem as well as biological insights and Protein Microarray data analysis pipeline. It describes full workflow to identify differentially reactive proteins using protein microarray data. Conventional solutions for biomarker discovery using differential expression analysis are also outlined.

Methods chapter includes a detailed overview of methods and algorithms for identifying reactive proteins. Also, we describe real biological and synthetically generated data sets that are used in the experiments. Enrichment analysis is described as a way to characterize and interpret obtained lists of reactive proteins.

Results chapter explains all obtained results of conducted experiments using the approaches and methods described in the previous section. Also, the process of characterising results using enrichment analysis is shown.

Discussion gives a brief overview of notable limitations during work on this thesis. We also outline the work that should be done in the future.

Conclusions chapter provides a summary of this thesis outcome.

2 Background

This chapter provides a brief introduction into the immunology and autoimmunity, which is necessary to understand the rest of the thesis. It also describes the main techniques and approaches used to analyse autoimmune-related protein microarray data.

2.1 Autoimmune Response

The human immune system is defined as a network-like structure of different cells, tissues and organs that collaborate to prevent an organism from the influence of malignant foreign substances. These substances are various bacteria, viruses and parasites. Our immune system has an ability to recognise and eliminate millions of different types of intruders. To be able to carry out its function, the immune system has to distinguish individual's cells and "enemy" substances – antigens [10]. The immune system uses special cells – antibodies to detect and eliminate antigens. Under certain circumstances, it produces antibodies that attack body's own normal cells and tissues. These antibodies are called autoantibodies. Such abnormal immune process triggers autoimmune diseases [2, 3, 10].

It is known that autoimmune diseases can affect almost any organ and tissue of the human body. Moreover, many of them affect more than one organ during disease development. Aaron Lerner et al. [1] described four groups of autoimmune diseases based on their origin and an area of their activity. These are neurological, gastrointestinal, endocrinological and rheumatic. Also, some disorders behave themselves very individually. These factors make it almost impossible to develop general diagnosis and treatment strategies.

However, as was previously mentioned, identification of reactive proteins may contribute substantially to discovering of autoimmune disease biomarkers candidates. Such biomarkers may give extra knowledge to early disease diagnosis as well as may play a key role in accurate distinguishing autoimmune diseases with identical clinical symptoms. Another application area of biomarkers is monitoring of illness progressions [11, 12].

Identification of reactive molecules can be possible analysing specific microarrays with incubated serum samples of diseased and healthy people.

2.2 Protein Microarrays

Nucleic acid arrays (DNA arrays) were invented at the end of the previous century and had become a standard technology in the gene expression analysis [13]. Initially, such approach

defines a process of printing DNA fragments on glass microscope slides in duplicates. Nowadays, it is possible to insert almost 2 million DNA sequences into a single microarray. Fragments of DNA themselves are marked with fluorescent tags. Such labelling gives a possibility for specific scanners to read light intensities and convert them into quantitative measures [12, 13, 14].

However, DNA microarray approach is not suitable for the analysis of complex protein-based reactions as DNA microarray structure is too simple to meet protein requirements for triggering biochemical activities [15]. Therefore, protein-based microarray technology was developed with a focus on a retaining protein assembly and related changes. It allows simultaneously processing many thousands of proteins in considerably short time [16]. Three different classes of protein microarrays exist that allow investigating various types of autoimmunity. They are analytical microarrays, functional protein microarrays and reverse phase protein microarrays. Functional protein microarrays are commonly used in autoimmunity-related researches. The idea of usage exactly this type of protein microarrays is determined by microarray architecture [15].

In this thesis, we use ProtoArray® Human Protein Microarray[8] developed by Thermo Fisher Scientific Inc.[9] Such microarray consists of 9k unique functional protein peptides that are immobilized on the array's surface. Protein microarray is incubated with human serum or other body liquid that contain interactive biomolecules. Zhu et al. state that "when an autoantibody presented in human serum associated with an autoimmune diseases recognizes a human protein spotted on the array, it can be readily detected with fluorescently labelled anti-human immunoglobulin antibodies (e.g., anti-IgG) and a profile of autoantibodies associated with a disease thus created." [17] Presence of these labelled auxiliary antibodies creates intensity signals that are read by specific microarray scanners producing high-resolution array images.

Thermo Fisher states that ProtoArrays support a broad range of research applications such as immune response biomarker discovery, enzyme substrate profiling, protein-protein interactions and others[10]. Its main application focus is immune response biomarker discovery as it is indicated that this product provides high sensitivity level of candidate detection and

[8] http://www.thermofisher.com/ee/en/home/life-science/protein-biology/protein-assays-analysis/protein-microarrays.html
[9] http://www.thermofisher.com
[10] https://www.thermofisher.com/ua/en/home/life-science/protein-biology/protein-assays-analysis/protein-microarrays/human-protein-microarrays-overview.html

availability to reproduce experiments in addition to low requirements for biological material[11].

The overall quality of protein microarrays and reducing possible sources of defects and variabilities of their production are critical for the analysis. Being sure that scanning results are reproducible is essential [12].

2.3 Protein Microarray Data Analysis Pipeline

This section gives an overview of data analysis pipeline used in this thesis for autoimmune profiling study starting from data acquisition methods, going through data pre-processing and ending with a brief differential expression analysis overview.

2.3.1 Data Acquisition

After incubation of a protein microarray and succeeding scanning of fluorescence intensities high-resolution images in TIFF format for each protein array are acquired. For distinguishing between printed spots and array background intensity, specific image segmentation software is used for each spot delimitation and labelling. To do this, we need exact information about spot locations, sizes, distances between them and protein identification numbers. This information is provided by GenePix Array List files (stored in GAL format). After loading these supplementary files, segmentation software outputs quantitative measures of fluorescent signal intensities as well as intensity signals of array background. Obtained results for each protein spot in each protein array are stored in GenePix Results files (GPR file format). For next steps of data analysis, these files are uploaded into the statistical environment. One GPR file represents raw data of one sample of protein array [12].

[11] https://www.thermofisher.com/content/dam/LifeTech/migration/en/filelibrary/protein-expression/pdfs.par.16180.file.dat/protoarray%20v5%20irbp-fhr.pdf

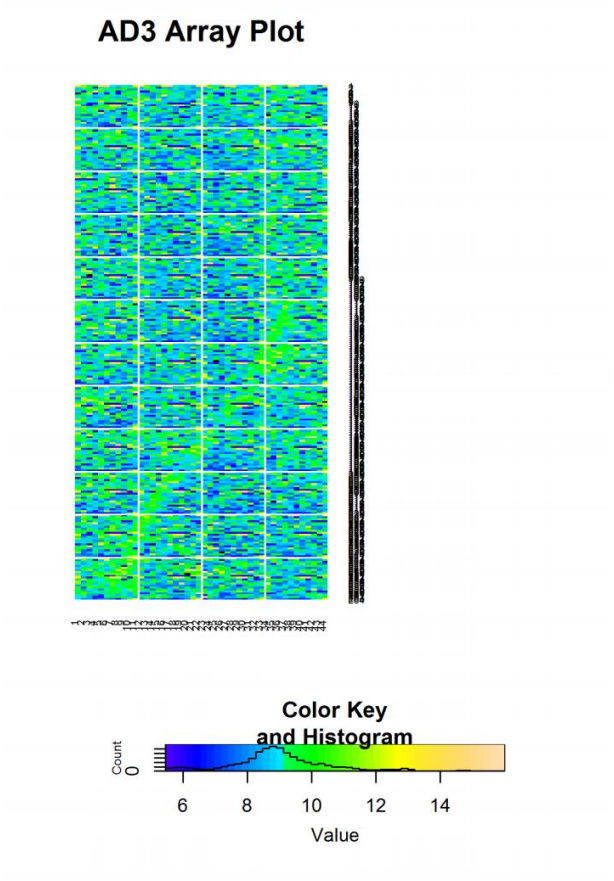

Figure 1. Visual representation of protein microarray by PAA package [18]

Each protein is immobilised twice on protein microarray surface, therefore during data importing process signal per particular protein is obtained by taking a minimum or mean signal value between duplicated proteins. Also, all subsequent analysis manipulations are done on binary logarithmic transformed signal values.

2.3.2 Data Pre-processing and Normalization

To make sure that signal intensities in each data sample are from statistically similar distributions across all of the provided GPR files and can be subsequently processed, it is necessary to eliminate the effects that could appear due to different technical, chemical and physical external factors. However, it is crucial to preserve biological differences between the samples. Applying combinations of pre-processing and cleaning techniques facilitate reduction of side effects. These practices are focused on detecting and removing outliers as well as correcting array background influence. All these techniques were applied using the functionality of the corresponding publicly available R software package limma [19].

Data normalization is another important step in reducing variabilities between samples preserving the real biological difference. These variabilities are measured in comparison with intensities of so-called control proteins – which are also spotted on the protein arrays. Control proteins are not expected to react with the autoantibodies. Therefore, intensity signals of such proteins should be constant across and within arrays and can serve as estimators of the normalization process. The most common methods used to normalize signals across protein microarrays are quantile[12], cyclic loess, variance stabilizing normalization (VSN) and robust linear model normalization [6, 20, 21, 22]. Cyclic loess normalization elaborates on an idea of MA plot[13] to compare log-transformed values and their mean. VSN is aimed to scale signal values while the variance tends to be independent of the mean values. In this work, Robust Linear Model Normalization (RLM) technique is used. It was reported that RLM normalization based on specific control proteins outperforms more traditional methods such as global and quantile normalization in terms of resistance to outliers as well as outlines decreased inter and intra-array variabilities [6]. Furthermore, to reduce possible high skewness of the signals' distribution and even make data more approachable to visualisation we use a log transformation of signal intensities. In addition, it helps to approximate obtained distribution to the normal one that is required for methods of differential expression analysis.

2.3.2.1 Robust Linear Model Normalization

Robust Linear Model normalization is built under a process of fitting a linear model using robust regression on mean values of weighted least-squares based on the median. Weighting and using median instead of the mean estimator provides high robustness to outliers. General formula for training robust linear model is:

$$y_{ijkr} = a_i + b_j + \tau_k + \varepsilon_r$$

Where a_i - effect of array i, b_j - effect of subarray or block j, τ_k - effect of the protein feature k and ε_r - normally distributed around zero random error, r – protein spot in range from 1 to 2, which states for spot replicate, y_{ijkr} – log-transformed initial signal intensity for given spot r within feature k in subarray j of array i.

[12] https://en.wikipedia.org/wiki/Quantile_normalization
[13] https://en.wikipedia.org/wiki/MA_plot

Normalization itself is conducted by subtracting values of array and subarray effects from signal:

$$y'_{ijkr} = y_{ijkr} - a_i - b_j$$

Implementation of this method is available in Prospector software[14] and R package MASS[15].

2.3.3 Differential Expression Analysis for Biomarkers Identification

Normalized and pre-processed data obtained from protein microarray experiments allows identifying reactive protein signals to distinguish between groups of healthy and autoimmune disease samples. Such proteins are called biomarkers.

Univariate filter methods are considered for biomarker selection such as t-tests and M-statistic [4, 5, 6], which give us a statistical measure – p-value for each protein feature. As a result, the list of differentially expressed proteins sorted by p-values with specified threshold is obtained. However, these methods focus on identifying biomarkers for groups distinguishing and might drop proteins that show signal reactivities that are not statistically significant to distinguish healthy and patient samples.

2.4 Existing tools

This section describes existing technologies for analysing protein microarray data. These approaches are focused only on identification of candidates for autoimmune diseases biomarkers.

2.4.1 Thermo Fisher ProtoArray Prospector

Thermo Fisher Scientific Inc. provides a range of services related to life science field, especially for the protein biology sector. One of their main assets is an analysis solution ProtoArray Prospector software[16]. In case if commercial ProtoArray Human Protein Array technology is acquired, limited version of Prospector software can be used free of charge. It works only with data obtained from ProtoArray microarray. Full workflow of data analysis with front-end interface is provided and consists of data acquisition and loading, pre-pro-

[14] https://www.thermofisher.com/ua/en/home/life-science/protein-biology/protein-assays-analysis/protein-microarrays/technical-resources/data-analysis.html
[15] https://cran.r-project.org/web/packages/MASS/
[16] https://www.thermofisher.com/ua/en/home/life-science/protein-biology/protein-assays-analysis/protein-microarrays/technical-resources/data-analysis.html

cessing and normalization, selection of algorithms and methods to apply and reporting analysis results by displaying them throughout application interface and exporting to Microsoft Excel file. Data acquisition is represented by loading GPR files. Pre-processing takes into account background intensities and conducts normalization, where Robust Linear Model is also available as one of the normalization techniques. The next step is dedicated to comparison of two groups of samples, healthy and diseased ones. Particularly, it focuses on identification of proteins, which show off increased signal intensities in one group compare to another.

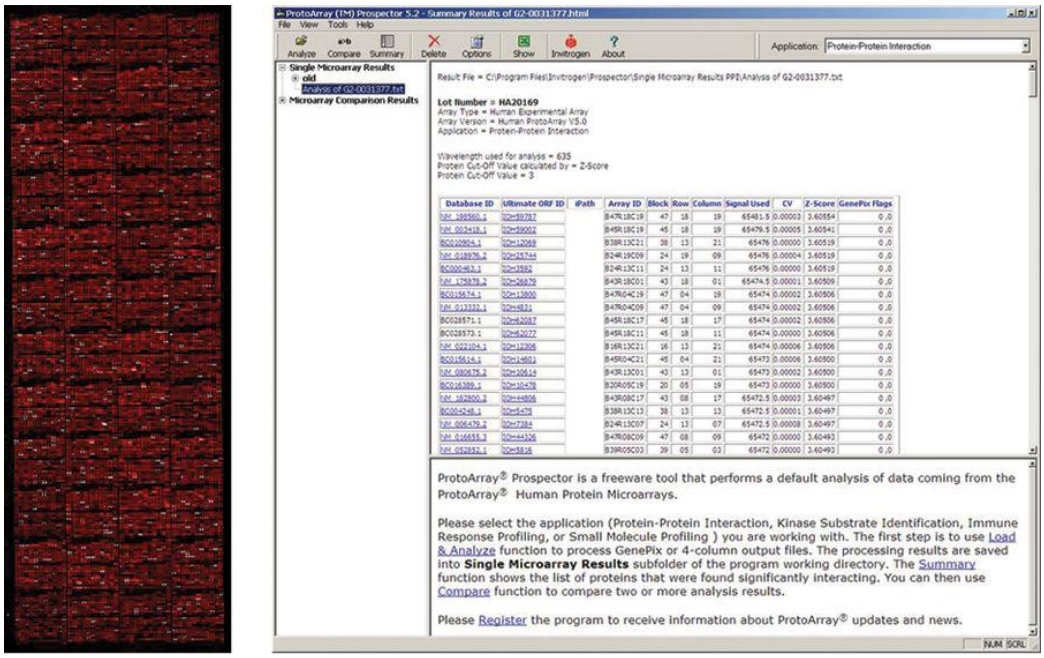

Figure 2. Screenshots of the ProtoArray® Human Protein Microarray (left) and ProtoArray® Prospector software interface (right)

Turewicz M. et al. [22] proposed a set of improvements for Prospector software workflow after analysing large datasets to identify potential diagnosis biomarkers. Suggested improvements cover almost all aspects of default Prospector workflow. Particularly, improvements concern the acquisition of raw data methodology, correction of batch effects, available normalization techniques and feature preselection as well as selection methods and feature validation. An improvement regarding data acquisition is the replacement of GenePix Pro workflow by fully automatic StrixAluco (Strix Diagnostics, Berlin, Germany) software. GenePix Pro is not effective in terms of time needed to handle variance issues. It requires manual steps of array partitioning (need of GAL files). The second improvement concerns the lack of a 64-bit version of the software, and this affects the performance of processing data sets with a large number of samples. The third problem is a lack of the solution for

elimination of batch effects. Batch effects occur due to inconsistencies in spotting conditions such as spotting in different labs. The author proposed solutions for this problem relying on a particular study described in the article [22]. Therefore, applied approaches cannot be used as a general methodology, but the problem is defined and should be taken into account. Next addressed issue lies in biomarker candidate selection via M-statistics. The main concern about this method provides p-values that are not adjusted for multiple testing, and there is no clear statement how unbiased threshold should be selected. As a solution, it was suggested to apply multivariate feature selection and manual selective methods in addition to M-statistics. Also, it was stated that the combination of these approaches deprives the batch effect issue. The last improvement proposed unbiased validation of selected potential biomarkers using classification algorithm Random Forest on randomly split samples from study data into independent training and test sets. Random forest model was trained on the training set and consequently tested on the test set. In the result, classification accuracies of potential biomarkers were obtained and estimated [22]. These improvements have been included in the new versions of Prospector software.

2.4.2 Protein Array Analyser Package

Another alternative software tool that analyses ProtoArrays and reports potential biomarker candidates is Protein Array Analyzer (PAA) [18]. PAA is provided via the open-source Bioconductor [23] software for bioinformatics and is used via R programming language. The default pipeline of the package is very similar to the one that is used in Prospector software. Moreover, authors state that PAA inherits main Prospector methods and approaches while being more customizable and fully free of charge.

The pipeline of PAA contains six main steps. The first part is data importing. Input for the analysis represents microarray data in GPR formatted files. Importing itself is done running function *loadGPR* is written and executable in R software environment. This function is based on the function *read.maimages* from limma package [19] imported as a dependency for PAA. Additional parameters could be specified if different kind of protein arrays is used or other relevant data should be considered. The data is loaded into a specific class of expression lists *EListRaw*.

Regarding pre-processing steps, it provides various functionality for the background correction, batch effect elimination and normalization. The main distinction (except functionality for batch filtering) from the Prospector software on this stage is a precise visualization of

the results. All the sub-steps of pre-processing come alongside with suited visualization functionality. For instance, the package provides the ability to assist in a choice of normalization technique comparing visualization of raw data and normalized by different techniques showing MA plots or boxplots. Available normalization methods are cyclic loess, quantile, Variance Stabilizing Normalization and Robust Linear Model.

For the differential expression analysis, the package provides two strategies: ordinal t-test and M-statistics. Choice of the appropriate method can be made using provided Volcano plot visualization of differential features. Such visualization helps to select an appropriate threshold for p-values and the computational method itself. Another visualization is provided to look at the p-values themselves.

Afterwards, classification methods are used to validate proteins-features that are identified as ones capable of distinguishing disease and patient samples. The following step is a feature pre-selection that provides functionality to get rid of the distinctively not differential features. Subsequently, multivariate feature selection functionality is presented. It can be done via following available algorithms: k-cross validation using Random Forest, Random Jungle and Support Vector Machine. In the result, package functionality provides an output list of differential features with reported classification accuracies. In addition, the package provides visualization functionality for manual validation of selected features, particularly intensity plots and heatmaps.

We used PAA package in experiments on synthetic and real data due to its low computational costs; we used M-statistic implementation as it is developed in C++ and ported to R afterwards.

3 Data and Methods

This chapter gives a detailed overview of methods and algorithms for identifying reactive proteins. In addition, we describe biological data and generation of synthetic data sets that are used in the experiments.

During work on this thesis, GNU R is used for all steps of data analysis [24]. R is a programming language and free of charge, highly extensible software environment for statistical computing, graphics and statistical software development. In addition, RStudio as an R environment and an open-source set of various tools for data analysis is also used [25].

3.1 Data

3.1.1 Synthetic data

The idea behind the need for synthetic data is that the ground truth about all reactive proteins or biomarkers is not known in biological data. To solve this issue, we generated synthetic data sets, which are statistically similar to the available biological data. Also, we inserted different levels of noise to estimate the robustness of the methods. Datasets with extreme noise level were also generated to conduct experiments to define "performance" cut-offs of each method.

Workflow of synthetic data generation starts from picking distribution parameters. We determine the mean and standard deviation of to-be-generated datasets. Eventually, all of the samples in this data at this step are synthetic healthy samples. The next step is to define dataset dimensions. We decided to generate 100 samples with 1000 protein each. This number of proteins is chosen to speed up computations. We generated the dataset with given dimensions and normally distributed signal values with arbitrary chosen mean and standard deviations of healthy samples from the real biological data [26]. Next, 50% of the samples are selected to represent patients and rest are a healthy cohort. Insertion of additional signal values into patient samples provides us with the ground truth of reactive proteins.

To prevent reactive signals generation from being biased, we followed the procedure: firstly, we randomly chose 5% of proteins to serve as differentially reactive proteins; they represent true positives that we are trying to identify. Secondly, we randomly defined the number of samples per each reactive protein where extra signal value had been inserted. Thirdly, we added to the existing signal randomly generated value drawn from the normal distribution with specified mean and standard deviation parameters. Parameters are chosen based on

obtained distribution of just-generated healthy group samples. The mean value is three standard deviations, and standard deviation equals to one standard deviation of healthy samples signals.

Consequently, we got 50 patients and 50 healthy samples of 1000 proteins where 5% of proteins were randomly selected to be reactive proteins in patient samples. For experiments, we generated 1000 such datasets.

3.1.1.1 Noise insertion

Moreover, to observe how well methods are resistant to noise for each of the previously generated datasets, we decided to insert artificial noise in two dimensions. Particularly, we vary noise coverage on both number of proteins (protein coverage) and number of samples across only "patient" group (sample coverage). Moreover, noise can also be inserted to the proteins that were previously chosen to be reactive. The noise itself represents addition signal value added on top of already generated signal and equals to a value drawn from the same normal distribution that was used to generate reactive protein signal.

Noise coverage of proteins is represented by five levels, which are denoted as coefficients: 0.2, 0.4, 0.6, 0.8 and 1. Iteratively multiplying them by a total number of proteins in the synthetic dataset (N_P), we get proteins where noise signals are added. Indexes of these proteins are identified at random. So, the number of proteins with added noise (n_P) is denoted as:

$$n_P \in (\lfloor 0.2 N_P i \rfloor)_{i=1}^{5}, \text{ where } N_P = 1000$$

We defined 20 noise levels of sample coverage per each protein coverage. These 20 levels are coefficients to identify the number of samples across synthetic patient samples (N_S). Moreover, we randomly set which samples per predefined protein are going to be enriched with noise. Corresponding number of samples with inserted noise (n_S) per noise level is defined as:

$$n_S \in (\lfloor 0.05 N_S i \rfloor)_{i=0}^{20}, \text{ where } N_S = 50$$

We assume that in such kind of experiments, control samples usually have signals with low noise level or are not affected by noise at all. Therefore, we did not insert noise signals to synthetically generated health samples.

Overall, we generated 100K datasets (1000 datasets per each of 100 noise levels) out of them 1000 datasets without any noise added per each protein noise level. Later, t-statistics, M-statistics, z-score-based method, Stuart method and RRA are applied on this data to identify synthetically generated reactive proteins and to assess methods' overall performance and resistance to noise.

3.1.2 Real biological data

To assess relevance and efficiency of Robust Rank Aggregation we used Nagele et al. [25] research on Alzheimer's disease using protein microarray data as a case study. Nagele and colleagues intended to identify significantly differentially reactive proteins that are present in patient serum samples and determine them as potential biomarkers for Alzheimer's disease. Authors reported ten potential biomarkers - proteins that have shown high-level diagnostic accuracy score of more than 90%. Differential reactivity analysis through the research was done using previously described Prospector Software.

The case study aimed to replicate data analysis steps of protein microarray data and to apply RRA to obtain lists of reactive proteins. Particularly, we implemented almost all algorithmic, processing and functional features of Prospector software that were reported in the corresponding research article. The main goal for manual implementation of all functionality was to construct utilizable code scripts designed for analysing autoimmune reactivities of protein microarray data.

Biological data has been extracted from Gene Expression Omnibus database[17] that contains samples with scanned fluorescence signals of Alzheimer's disease patients serum samples (AD) and non-demented disease control samples (NDC) of younger and older population separately. The format of data is Genepix resulting files (GPR). Overall, we got 90 samples: 50 AD and 40 NDC instances.

After raw data had been acquired, we implemented a workflow for training Robust Linear Model using functionality from MASS package [17]. Afterwards, we got pre-processed and normalized data that had been ready for protein reactivity analysis using M-statistics, RRA and other methods.

[17] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29676

3.2 Methods

This section describes methods that are used in this thesis to identify a wide range of reactive proteins. Description of their performance estimation is presented in the next chapter - Results.

3.2.1 T-statistics

One of the methods for the identification of differentially expressed genes or proteins in microarray data is a moderated t-statistics [27]. It is evolved from an ordinary t-statistics [28]. Under the null hypothesis – no changes in gene expression are observed, and an assumption that the data is normally distributed it is defined as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE(_{\bar{X}_1 - \bar{X}_2})}$$

Where the upper part of the equation stands for the difference in mean log-transformed gene signal values and lower part equals to the standard error of the expression above. Substantially, the p-value is identified as a probability that t value is as extreme as the observed value satisfying the null hypothesis or even bigger.

However, ordinary t-statistics is poorly efficient in a domain of differential expression analysis of genes as ordinary version uses data for analysis from only one gene at a time. In addition, there is a possibility of the unstable variance of separate genes, which significantly affects resulting value. Moreover, the main weakness of the ordinary t-statistics is a need of big sample sizes that are commonly small for such kind of research. Therefore, Smyth et.al [27] introduced a moderated t-statistics based on fitting a linear model with empirically obtained Bayesian meta-parameters. Particularly, it replaces gene-specific standard error with moderated standard error across all genes. According to [27], for gene j it is defined as:

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g \sqrt{v_{gj}}}$$

Where $\hat{\beta}_{gj}$ – contrast estimator – the difference in mean signal values of comparing genes ($\bar{X}_1 - \bar{X}_2$), v_{gj} – unscaled standard deviation, s_g – posterior residual sample variances that depend on prior estimators with degrees of freedom. These degrees of freedom are delivered from the assumptions of normal distribution of contrast estimator $\hat{\beta}_{gj}$ and scaled

chi-square distribution of residual variances s_g. Delivered p-values are different from ordinary t-statistics p-values by the increase in degrees of freedom amount.

Full implementation of moderated t-statistics is collected in differential expression analysis R/Bioconductor package limma [19]. Particularly limma provides functionality for fitting model objects, multiple testing corrections across contrast estimators and outputting summary tables with the results of the linear model fitting, hypothesis testing and p-values adjustment for multiple testing itself. This package was used in our real data and synthetic experiments.

3.2.2 M-statistics

Identification of biomarkers candidates using M-statistical analysis was initially provided by Prospector software. Concisely, M-statistics is an iterative comparison of all signal values in one group with signals in another group for a certain protein. Then M-value is calculated, which is the number of times when signal values from the second group are higher than the maximal value in the first group. The number of values with signals bigger than the second (third, fourth and so on) largest signal in another group are also calculated. The p-value is defined as a probability of having M-value greater than obtained M-value, which is done using hypergeometric distribution[18]. In result, p-values give a list of differentially expressed proteins between two groups of healthy people and patients.

A cutoff value (or threshold) is described as the comparison that returns the lowest p-value. This lowest p-value is the significance of the certain protein. Afterwards, proteins are sorted by the difference in prevalence between patient and control groups, and top proteins are reported as reactive proteins. Prevalence is defined as:

$$E(P) = \frac{M_{max} + 1}{n_y + 2}$$

Where M_{max}- M-value of the corresponding lowest p-value, n_y - group size

M-statistics from PAA package was used in all experiments.

[18] https://en.wikipedia.org/wiki/Hypergeometric_distribution

3.2.3 Z-statistics as a method for identification of reactive proteins

Some autoimmune diseases affect a very diverse range of proteins. Therefore, consistently reliable mean signal reactivity between 2 cohorts of samples might not occur [29]. To address this issue, in recent autoimmunity-related studies scientists have started to apply z-statistics [7, 29, 30].

For each signal, z-score tells how *far* in terms of standard deviations the mean of disease group from the mean of signal values of healthy cohort and it is denoted as:

$$z = \frac{X - \mu}{\sigma}$$

Where z is the z-score itself, X states for the value of the signal intensity, μ is the signal mean of healthy samples and σ is the healthy cohort's standard deviation.

Z-statistics identifies the high amount of false positives as its output list of reactive proteins linearly depends on the number of patient samples. Consequently, researchers commonly define significance threshold number of patients manually. In [7] and [29], authors identified significance thresholds relying on their biological hypothesis. That might negatively affect impartiality aspect of the research.

However, we decided to apply z-statistics as a method for detection of reactive proteins to compare them with more statistics-based methods. Let us denote the case when z-score ≥ 3 in at least one patient as a p1z3 method.

3.2.4 Robust Rank Aggregation

It is well known that data from biological experiments comes from different origins and is obtained in various conditions as well as in the presence of noise sources. That could lead to incomparability of the researched data material itself. Therefore, strategies for analysing datasets from different sources separately and aggregating the results together are defined to be relevant for such spectre of researches. However, aggregation into single result unit is very sensitive to assessing aggregation "correctness", and its interpretation concerns results significance for different quality of the data.

To address described issues, Robust Rank Aggregation (RRA) algorithm was developed by R.Kolde et al. as an unbiased and noise-robust method based on the order statistics for analysis in genomic data applications [8]. Particularly, it focuses on a data aggregation from

high-throughput biological experiments while input data structures themselves are represented as prioritized lists of genes.

In the domain of autoimmune researches, this method ranks proteins by their signal intensity for each sample and outputs results as ranked lists. The next step is a comparison of protein position in the ranked list to baseline scenario when all lists are randomly shuffled. In the result, it gives a p-value for each protein. These p-values describe the probability to see certain protein rank higher in the ranked list rather than in the randomly ordered. P-values are also used for re-ranking proteins in the final list and assigning significance score.

However, to obtain the final ranking list, RRA takes a minimum of p-values across all the ranking lists. It means that these p-values are not explicitly p-values themselves. Consequently, there is a need to apply multiple hypothesis testing correction. As the process of obtaining exact p-values is computationally expensive and the same to the one developed for Stuart method improvement [31], authors of RRA proposed to use multiple testing correction technique per each obtained p-value. Consequently, newly obtained p-values are corrected for multiple testing to find statistically significant protein ranks.

Robust Rank Aggregation algorithm is designed to be resistant to outliers and noise. Another advantage is that its computational costs are low due to linear complexity in respect of input size of proteins in the sample [8].

Method implementation and data pre-processing for its execution are provided and used within R/Bioconductor package RobustRankAggreg [32].

3.2.5 Stuart method

The method by Stuart et al. [9] was chosen for experiments as an another ranking method and treated as a direct competitor of Robust Rank Aggregation.

Stuart method is based on the joint cumulative distribution of the order statistic[19] to calculate all rank ratios. Initially, it is denoted as:

$$P(r_1 r_2, \dots, r_n) = n! \int_0^{r_1} \int_{s_1}^{r_2} \cdots \int_{s_{n-1}}^{r_n} ds_1 ds_2 \cdots ds_n$$

Where, r – rank ratio of sample s, n – number of samples.

[19] https://en.wikipedia.org/wiki/Order_statistic

Formula to compute the integral above is proposed as:

$$P(r_1 r_2, \dots, r_n) = n! \sum_{i=1}^{n} (r_{n-i+1} - r_{n-1}) P(r_1 r_2, \dots, r_{n-i}, r_{n-i+2}, \dots, r_n)$$

Where r_i – rank ratio of data sample i and $r_0 = 0$.

However, Aerts et al. [31] firstly reports inefficiency of this formula because of its complexity $O(n!)$. Authors proposed faster version with quadratic complexity $O(n^2)$:

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

Where $P(r_1 r_2, \dots, r_n) = n!\, V_n$, $V_0 = 1$, r_i – rank ratio of data sample i.

Secondly, authors discovered that joint cumulative distribution calculation outputs values that cannot be used as p-values but rather as p-scores. Under the null hypothesis, these result values are not uniformly distributed. Therefore, they proposed to fit a distribution for every possible number of ranks and use such distribution to calculate an approximate p-value.

Stuart method with improvements by Aerts et al. [31] is implemented in RobustRankAggreg package [32] and was used in this thesis.

3.2.6 Enrichment analysis

When analysis of microarray protein data discovers statistically reactive proteins, there is a need to extract biological knowledge from the results. For this purpose, enrichment analysis is used to find groups of proteins that are statistically over-represented in the list of identified reactive proteins [33]. The aim of this is to find connections between classes of obtained proteins and different diseases. Protein classes are pre-defined by prior biological knowledge and describe protein or gene groups that share similar biological functions. Later biologists use identified enriched terms in their experiments in laboratories.

Application of enrichment analysis in this thesis is described in more details in Results section.

4 Results

This chapter presents results of experiments using RRA and other methods aimed to identify reactive proteins. We conducted experiments on two datasets – synthetically generated and real biological protein microarray data. Firstly, we tested methods on synthetic data without any noise and with various levels of inserted noise. Later, protein microarray dataset from [26] was used to assess the performance on real data. As a result, lists of reactive proteins were identified. Enrichment analysis has been performed on these lists in order to interpret proteins identified by each method. Enrichment analysis has provided us with lists of significantly enriched terms provided as an input for further analysis from the biology point of interest. Enrichment analysis was conducted by uploading lists of reactive proteins obtained from experiments on real data to g:Profiler [34].

4.1 Experiments on the synthetic data

Having generated synthetic data sets and knowing the ground truth of reactive proteins, we assessed each method performance and compared them.

Each tested method outputs p-value for each protein in the synthetic dataset. Multiple testing correction was applied on obtained p-values to control the cases where null-hypotheses (protein is not expressed) were incorrectly rejected. For all the tested methods, we applied Bonferroni[20] correction and proteins were sorted by adjusted p-values. Significance threshold for all the experiments equals to 0.01.

We used F1 score as the main method performance metric. It does not take into account true negatives in which we are not interested. The F1 measure is based on precision and recall metrics.

Precision in the context of conducted experiments is defined as the number of proteins correctly identified as reactive (Tp), divided by the sum of Tp and Fp, where Fp – the number of proteins incorrectly identified as reactive. It can be represented by the formula:

$$Precision = \frac{Tp}{Tp + Fp}$$

Recall is denoted as:

[20] https://en.wikipedia.org/wiki/Bonferroni_correction

$$Recall = \frac{Tp}{Tp + Fn}$$

Where Fn - the number of inserted reactive proteins but not identified as reactive.

Thus, F1 score describes harmonic mean of precision and recall, and it is defined as

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Score value is in the range between [0, 1] range, where 1 means the best possible performance.

Experiments were conducted on the data with inserted different levels of noise. Results of the experiments on the noise-free synthetic datasets report average F1 score value of 1000 experiment runs per each of the used methods. Results on the noisy data sets are obtained in the same way for varying number of samples and proteins with inserted noise signals.

Figure 3A and 3B present comparisons of the methods on the data with 20% of proteins with added noise. Each facet represents results on the data with the corresponding percentage of samples with added noise.

The first facet stands for the results on synthetic datasets without any noise.

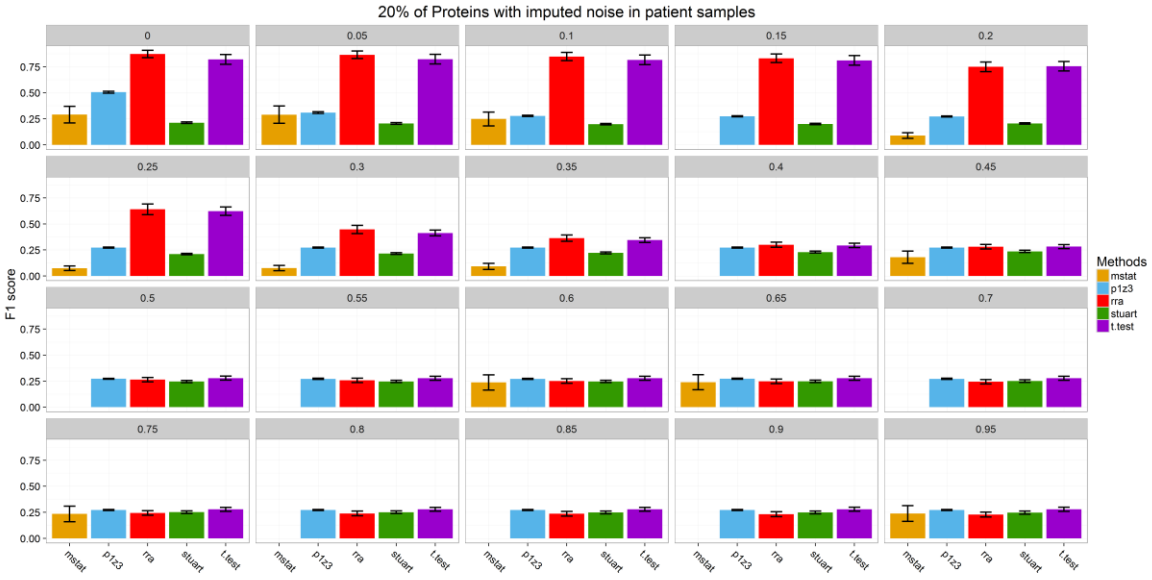

Figure 3A. 20% of proteins got noise insertion per 20 levels of noise in samples. Applied methods on the x-axis, reported F1 scores on the y-axis. First facet represents zero noise dataset. Each error bar represents one standard deviation of F1 scores uncertainty across all 1000 experiments per method and noise in sample.

The first facet on this plot shows that Robust Rank Aggregation identified reactive proteins with the higher F1 score than all other methods. Such trend is preserved on the second facet where noise is present in only two to three patient sample of 200 proteins. Already at this point as Figure 3B shows, p1z3 identifies an extremely wide range of proteins - both noise signals and all true reactive ones with three standard deviations away from the mean. Therefore, F1 score for p1z3 remains on the same level for other facets. Though RRA continues clearly outperforming other methods until the noise is added in at least 35% of samples per protein. Afterwards, all methods show the relatively poor F1 score.

However, Figure 3A shows the behaviour of M-statistics that cannot be described as empirically logical. On noise-free dataset, M-statistics shows F1 score near 0.25, slightly losing it once few noise signals are added. However, at some point of sample coverage, F-score is approaching zero. Using PAA package, we discovered that M-statistics produces resulting lists with too few identified proteins. Such lists do not outline true positives in a significant proportion of experiments. It leads to zeros in recall and precision, consequently making F1 score undefined.

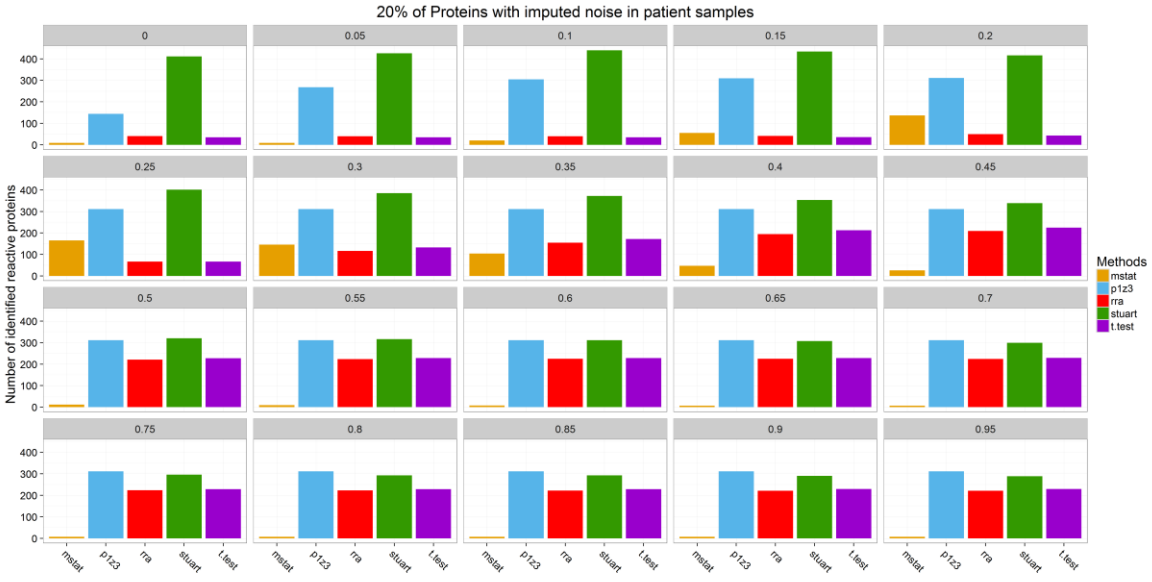

Figure 3B. 20% of proteins got noise insertion per 20 levels of noise in samples. Applied methods on the x-axis, reported number of identified reactive proteins on the y-axis. First facet represents zero noise dataset.

Stuart method shows low performance across all the facets illustrated in Figure 3A, while Figure 3B shows that the reason is in the extremely big amount of false positives – true signals lie in 50 proteins while noise is inserted in 200 and number of identified proteins is

bigger than 400. However, we can observe that Stuart identifies fewer proteins as reactive on higher noise levels in samples.

Evaluating next level of noise coverage in proteins, Figure 4A shows that Robust Rank Aggregation loses its F1 score across the noise more gradually than t-test and p1z3. This is due to the order statistics algorithm behind RRA. Particularly, composed ranking lists are relatively independent of each other before aggregation to the final list. Another order-statistic based method – Stuart shows low F1 score across all the noisy and noise-free data sets and identifies very many false positives that follow from Figure 4B.

Z-score based method starts to identify the enormous amount of false positives - proteins with noise - very rapidly as Figure 4B illustrates. While t-test shows relatively similar performance as RRA on the small amount of noise, but higher noise level negatively affects differential aspect of t-test much more significantly than RRA by increasing the number of identified false positives.

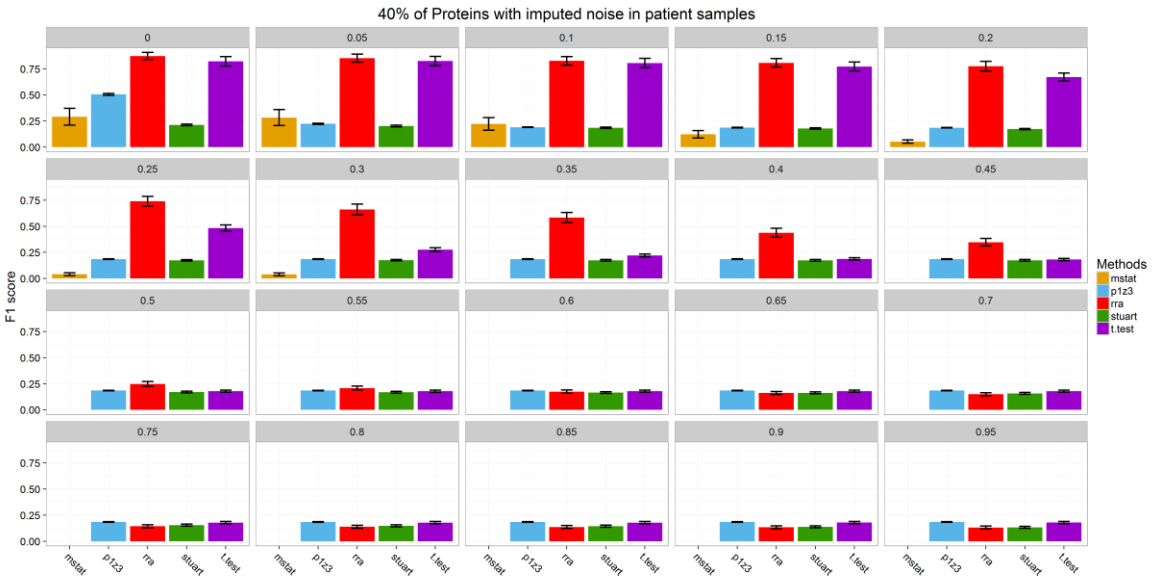

Figure 4A. 40% of proteins got noise inserted into 20 randomly chosen sets of samples. Applied methods on the x-axis, reported F1 scores on the y-axis. First facet represents zero noise dataset. Each error bar represents one standard deviation of F1 scores uncertainty across all 1000 experiments per method and noise in sample.

It was discovered that on higher noise levels while p1z3, Stuart, M-statistics and t-test are steadily losing their performance indicators, RRA starts to identify less amount of reactive proteins, which eventually slightly increases F1 score.

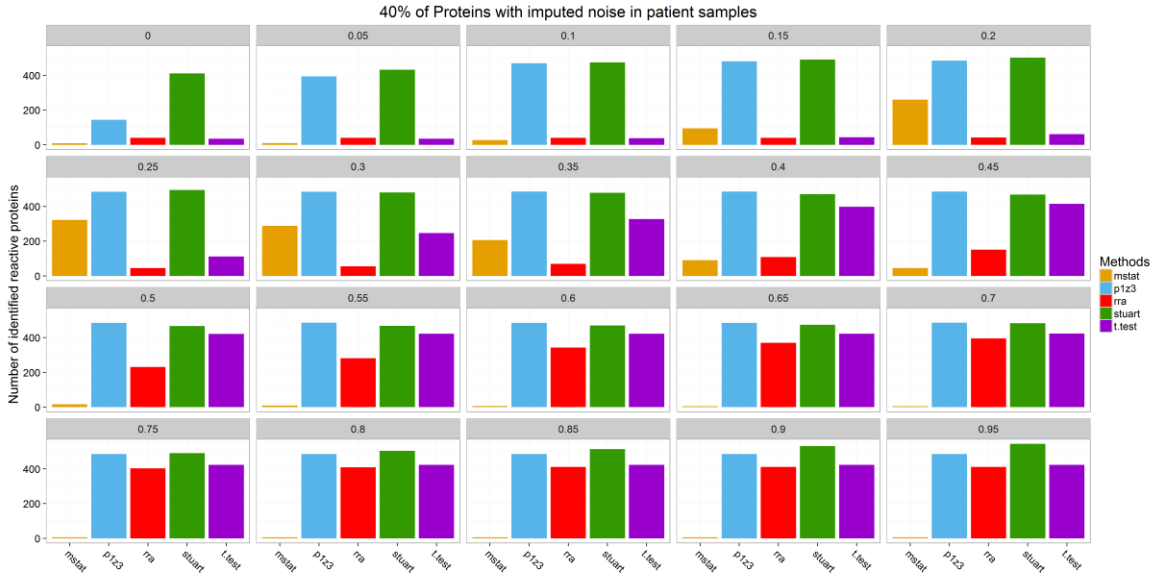

Figure 4B. 40% of proteins got noise insertion per 20 levels of noise in samples. Applied methods on the x-axis, reported number of identified reactive proteins on the y-axis. First facet represents zero noise dataset.

Plots with higher noise levels can be observed in Appendix A.

Overall, we can see that in perfect conditions when there is no noise, RRA performs slightly better than t-statistics. M-statistics, p1z3 and Stuart are significantly outperformed. Addressing real-life problems, where data could not be obtained without being subjected to any external influences and perfectly pre-processed and normalised, RRA might be a better choice showing substantially stronger resistance to noise than other tested methods.

4.2 Experiments on the real biological data

Experiments on the biological data are not focused on biomarker identification and validation of results reported in the related study but rather on the identification of a wide repertoire of reactive proteins. Nevertheless, given that we do not know the ground truth of this data, we conducted an enrichment analysis to identify significantly enriched terms that are present in the obtained lists of reactive proteins and compared the number of identified enriched terms between all methods. It is done by uploading lists of reactive proteins and all proteins presented in the data to g:Profiler [34]. It is a web tool that calculates a probability of having proteins of the specific class (enriched term) overrepresented in the list of proteins identified as significantly reactive (query), given all proteins presented in the data (background) comparing this probability to random chance.

Obtained numbers of identified reactive proteins in biological data and enriched terms per method are shown in the table below.

Table 1. Reactive proteins and enriched terms per method

Method	# of identified reactive proteins	# of identified enriched terms
RRA	1782	155
T-test	360	99
M-statistics	1261	91
P1Z3	1210	1

Method Stuart identified more than 7k reactive proteins (out of 9k proteins overall) and wasn't taken into consideration further.

However, there is not enough knowledge to say that RRA performs better than T-statistics or other methods especially taking into account very large number of identified reactive proteins. We also cannot state that the number of identified enriched terms is the appropriate measure to estimate the methods.

On the other hand, Meyer et al. [7] discovered approximately 3000 reactive proteins that showed high reactivity in at least one sample and stated that patients with autoimmune diseases could have a wide variety of reactive proteins. That also means that reactivity profiles of autoimmune disorders can be highly diverse.

In the related study [22], authors outlined ten proteins, which they chose as biomarkers candidates of Alzheimer's disease. On Figure 5, we show the intersection of these proteins with the identified reactive ones by the tested methods.

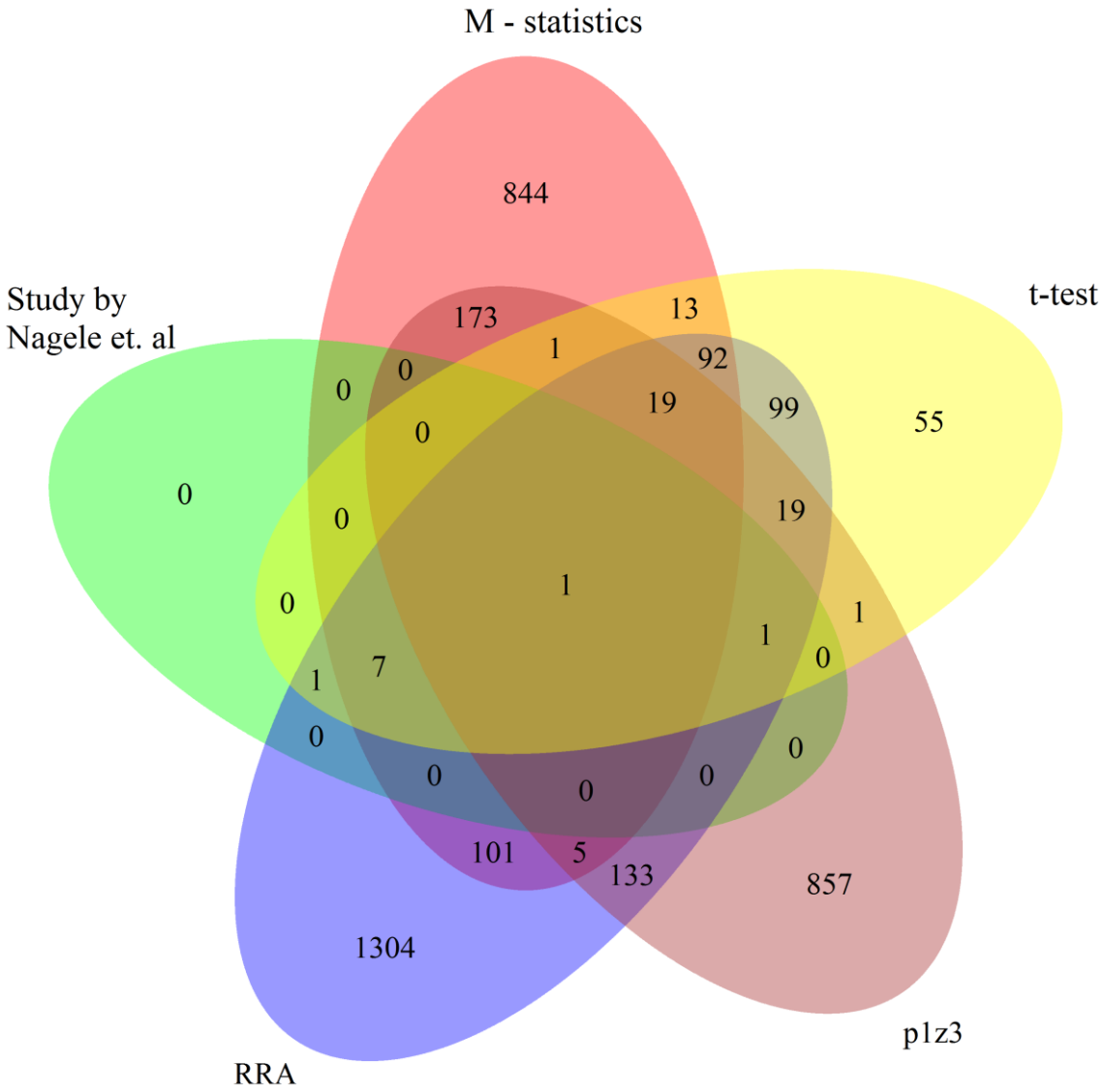

Figure 5. Venn diagram illustration of amounts of identified reactive proteins per each tested method and intersection between reactive proteins.

Venn diagram shows that all ten proteins from the study were found in RRA and t-test resulting lists of reactive proteins, while M-statistics found nine and p1z3 managed to outline only two proteins.

However, verification of top reactive proteins reported by RRA is not possible as it is impossible to get original blood samples of patients and controls that are used in the related research.

5 Discussion

When we have shown that Robust Rank Aggregation applies to identify reactive protein within the study of autoimmunity, there is room for the experiments using other novel methods and approaches. Particularly, researchers have recently introduced a number of rank aggregation techniques. For example, a network-based method for aggregation of multiple ranks for Gene Ontology prediction [35]. Authors of this approach proposed to use network information incorporating it into ranking genes or proteins before the aggregation to a final list. RRA with integrated network information was also tested against "original" RRA and encouraging results were reported in the corresponding article. Another novel approach uses multi-objective genetic algorithm [36]. Authors compared it to Stuart method, but proposed method have not succeeded in outperforming it across all experiments. However, it might also be interesting to test it against RRA. The third approach, which caught our attention, is Hybrid Bayesian-rank integration [37]. Authors developed an approach for rank aggregation, which works with data that is statistically unreliable by using Bayesian reasoning. This method is already implemented in R and freely available on the authors' webpage[21]. Overall, new methods are developing, and they should be tested as approaches to identify lists of reactive proteins in the domain of autoimmune disorders.

To combine into a single workflow all methods that are identified as suitable for the problem described in this thesis, as well as to provide an implementation of corresponding methods and justify their choice, we see a need in R package. It should provide functionality to apply methods that satisfy different prior biological knowledge as well as statistical expectations and requirements. Besides pre-processing and cleaning techniques, a choice of normalization approaches should be available. Moreover, enrichment analysis should also be integrated to provide input for biologists immediately. All the functionality should come alongside with a visualisation of each decision to be made. All these steps could be included in a continuous research project.

[21] http://www.pitt.edu/~mchikina/BIRRA/

6 Conclusions

Currently, there is no specific approach to identifying wide range of reactive proteins analysing protein microarray data. Such reactive proteins might help biologists to understand true causes of autoimmunity and to discover possible therapy strategies or to restrain the development of autoimmune diseases. In this thesis, we aimed to show that Robust Rank Aggregation method is applicable for identification of wide repertoire of reactive proteins. We evaluated the approaches meant to discover differentially expressed proteins and compared them with Robust Rank Aggregation.

We used synthetically generated data sets that resemble real biological data. This data was used to support the claim that Robust Rank Aggregation is an alternative approach to t-test and M-statistics as well as to p1z3 – z-score based method used by biologists. We showed how well methods identify reactive proteins if different set ups of noise signals were added.

Our results using synthetic data show that RRA performs better than t-test in terms of F1 score metric. While both t-test and RRA outperform M-statistics, p1z3 and Stuart methods on the data without noise. Moreover, experiments with different noise levels reveal impressive resistance of RRA, while t-test started to lose its performance quite early within the noise coverage. Other methods show quite poor performances initially, but they are preserved closely on the same level across the experiments with noise.

Furthermore, we experimented with publicly available protein microarray data. Having no ground truth about reactive proteins, we cannot assess methods' performance. However, we applied enrichment analysis on the lists of reactive proteins outlined by methods to find relevant biological terms, associated with these proteins. Immunology experts can estimate biological sense of these findings.

Nevertheless, we compared numbers of obtained enriched terms. Even though RRA identified the most proteins as reactive, it also identifies more enriched terms than other methods did. Also, a number of enriched terms found by t-test and M-statistics is almost the same regardless that M-statistics determined almost 4 times more proteins. These experiments were done under the assumption that more enriched terms are identified - the better method performs. However, we cannot affirm that such interpretation is correct. In order to estimate, which method works, we have to know which enriched terms are correct – but this information is unknown. Therefore, there is no clear way to say which method performed better from the biology point of view.

7 References

[1] A. Lerner, P. Jeremias and T. Matthias, "The World Incidence and Prevalence of Autoimmune Diseases is Increasing," *International Journal of Celiac Disease. Vol. 3, No. 4,* pp. pp 151-155, 2015.

[2] M. Gutierrez-Arcelus, S. S. Rich and S. Raychaudhur, "Autoimmune diseases — connecting risk alleles with molecular," *Nat Rev Genet,* vol. 3, no. 17, pp. 160-174, 2016.

[3] M. Roederer, L. Quaye, M. Mangino, M. H. Beddall, Y. Mahnke, P. Chattopadhyay, I. Tosi, L. Napolitano, M. Terranova, Barberio, C. Menni, F. Villanova, P. D. Meglio, T. D. Spector and F. O. Nestle, "The Genetic Architecture of the Human Immune System: A Bioresource for Autoimmunity and Disease Pathogenesis," *Cell,* vol. 161, no. 2, pp. 387-403, 2015.

[4] G.-M. Han, S.-L. Chen, N. Shen, S. Ye, C.-D. Bao and Y.-Y. Gu, "Analysis of gene expression profiles in human systemic lupus erythematosus using oligonucleotide microarray," *Genes and Immunity,* vol. 4, pp. 177-186, 2003.

[5] M. Crow and J. Wohlgemuth, "Microarray analysis of gene expression in lupus.," *Arthritis Research & Therapy.,* vol. 5, no. 6, pp. 279-287, 2003.

[6] A. Sboner, A. Karpikov, G. Chen and M. Smith, "Robust-linear-model normalization to reduce technical variability in functional protein microarrays," *Journal of Proteom Research,* vol. 8, no. 12, p. 5451–5464, 2009.

[7] S. Meyer, M. Woodward, C. Hertel, P. Vlaicu, Y. Haque, J. Kärner, A. Macagno, S. C. Onuoha, D. Fishman, H. Peterson, K. Metsküla, R. Uibo, K. Jäntti, K. Hokynar and A. S. Wolff, "AIRE-Deficient Patients Harbor Unique High-Affinity Disease-Ameliorating Autoantibodies," *Cell,* vol. 166, no. 3, pp. 582-595, 2016.

[8] R. Kolde, S. Laur, P. Adler and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis," *Bioinformatics 28 (4):,* pp. 573-580, 2012.

[9] J. M. Stuart, E. Segal, K. Daphne and S. K. Kim, "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules," *Science,* vol. 302, no. 5643, pp. 249-255, 2003.

[10] National Institute of Allergy and Infectious Diseases (U.S.); National Cancer Institute (U.S.), "Understanding the immune system and how it works," 2003.

[11] A. Mange, J. Lacombe, C. Bascoul-Mollevi and M. Jarlier, "Serum autoantibody signature of ductal carcinoma in situ progression to invasive breast cancer," *Clinical Canser Research,* vol. 18, pp. 1992-2000, 2012.

[12] L. Abel, S. Kutschki, M. Turewicz, M. Eisenacher, J. Stoutjesdijk, H. E. Meyer, D. Woitalla and C. May, "Autoimmune profiling with protein microarrays in clinical applications," *BBA Proteins and Proteomics,* vol. 1844, pp. 977-987, 2014.

[13] R. Bumgarner, "DNA microarrays: Types, Applications and their future," *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.],* vol. 0 22, no. 22.1, 2013.

[14] M. Smith, G. Jona, J. Ptacek, G. Devgan, H. Zhu, X. Zhu and M. Snyder, "Global analysis of protein function using protein microarrays.," *Mech. Ageing Dev,* pp. 126-171, 2005.

[15] H. Zhu and J. Qian, "Applications of Functional Protein Microarrays in Basic and Clinical Research," *Advances in genetics,* vol. 79, pp. 123-155, 2012.

[16] F. Reymond Sutandy, J. Qian, C.-S. Chen and H. Zhu, "Overview of Protein Microarrays," *Current protocols in protein science / editorial board, John E Coligan . [et al].*, vol. 0 27, no. 27.1, 2013.

[17] Z. H and Q. J., "Applications of Functional Protein Microarrays in Basic and Clinical Research," *Advances in genetics,* vol. 79, pp. 123-155, 2012.

[18] M. Turewicz, M. Ahrens, C. May, K. Marcus and M. Eisenacher, "PAA: An R/Bioconductor package for biomarker discovery with protein microarrays.," *Bioinformatics,* vol. 32, no. 10, pp. 1577-1579, 2016.

[19] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research,* vol. 43, no. 7, p. e47, 2015.

[20] B.Love, "Functional Protein Microarrays in Drug Discovery," *CRC Press,* pp. 381 - 402, 2007.

[21] Y. H. Yang, M. J. Buckley and T. P. Speed, "Analysis of cDNA microarray images," *Brief Bioinform,* vol. 2, no. 4, pp. 341-349, 2001.

[22] M. Turewicz, C. May, M. Ahrens, D. Woitalla, R. Gold, S. Casjens, B. Pesch, T. Bruning, H. Meyer, E. Nordhoff, M. Bockmann, C. Stephan and M. Eisenacher, "Improving the default data analysis workflow for large autoimmune biomarker discovery studies with ProtoArrays.," *Proteomics,* vol. 13(14), no. 2083-7, 2013.

[23] W. Huber, V. Carey, R. Gentleman and M. Morgan, "Orchestrating high-throughput genomic analysis with Bioconductor," *Nature Methods,* vol. 12, no. 2, pp. 115-121, 2015.

[24] R Core Team, "R: A Language and Environment for Statistical Computing," Vienna, Austria, 2016.

[25] RStudio Team (2015), "RStudio: Integrated Development Environment for R," RStudio, Inc., Boston, MA, 2015.

[26] E. Nagele, M. Han, C. DeMarshall, B. Belinka and R. Nagele, "Diagnosis of Alzheimer's Disease Based on Disease Specific Autoantibody Profiles in Human Sera," *PLoS One,* vol. 6(8), no. e23112, 2011.

[27] S. GK, "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Statistical Applications in Genetics and Molecular Biology,* vol. 3, no. 1, 2004.

[28] J. G. Thomas, J. M. Olson, S. J. Tapscott and L. P. Zhao, "An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles," *Genome Research,* vol. 11, no. 7, pp. 1227-1236, 2001.

[29] E. Nagele, M. Han, N. Acharya, C. DeMarshall, M. Kosciuk and R. Nagele, "Natural IgG Autoantibodies Are Abundant and Ubiquitous in Human Sera, and Their Number Is Influenced By Age, Gender, and Disease," *PLoS ONE,* vol. 8(4): e60726, 2013.

[30] N. Landegren, D. Sharon, E. Freyhult, Å. Hallgren, D. Eriksson, P.-H. Edqvist, S. Bensing, J. Wahlberg, L. M. Nelson, J. Gustafsson, E. S. Husebye, M. S. Anderson, M. Snyder and O. Kämpe, "Proteome-wide survey of the autoimmune target repertoire in autoimmune polyendocrine syndrome type 1," *Scientific Reports,* vol. 6, no. 20104, 2016.

[31] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet and Y. Moreau, "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, no. 5, pp. 537 - 544, 2006.

[32] R. Kolde and S. Laur, "RobustRankAggreg: Methods for robust rank aggregation," 2013. [Online]. Available: https://CRAN.R-project.org/package=RobustRankAggreg. [Accessed 11 August 2016].

[33] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *PNAS*, vol. 102, no. 43, p. 15545–15550, 2005.

[34] J. Reimand, M. Kull, H. Peterson and J. Vilo, "g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments," *Nucleic Acids Res*, vol. 35, p. W193–W200, 2007.

[35] W. Wang, X. J. Zhou, Z. Liu and F. Sun, "Network tuned multiple rank aggregation and applications to gene ranking," *BMC Bioinformatics*, vol. 16, no. S6, 2015.

[36] M. Kaur, P. Kaur and M. Singh, "Rank Aggregation using Multi Objective Genetic Algorithm," in *2015 1st International Conference on Next Generation Computing Technologies (NGCT)* , Dehradun, 2015.

[37] M. A. Badgeley, S. C. Sealfon and M. D. Chikina, "Hybrid Bayesian-rank integration approach improves the predictive power of genomic dataset aggregation," *Bioinformatics*, vol. 2, no. 31, pp. 209-2015, 2015.

Appendix

I. Illustrations

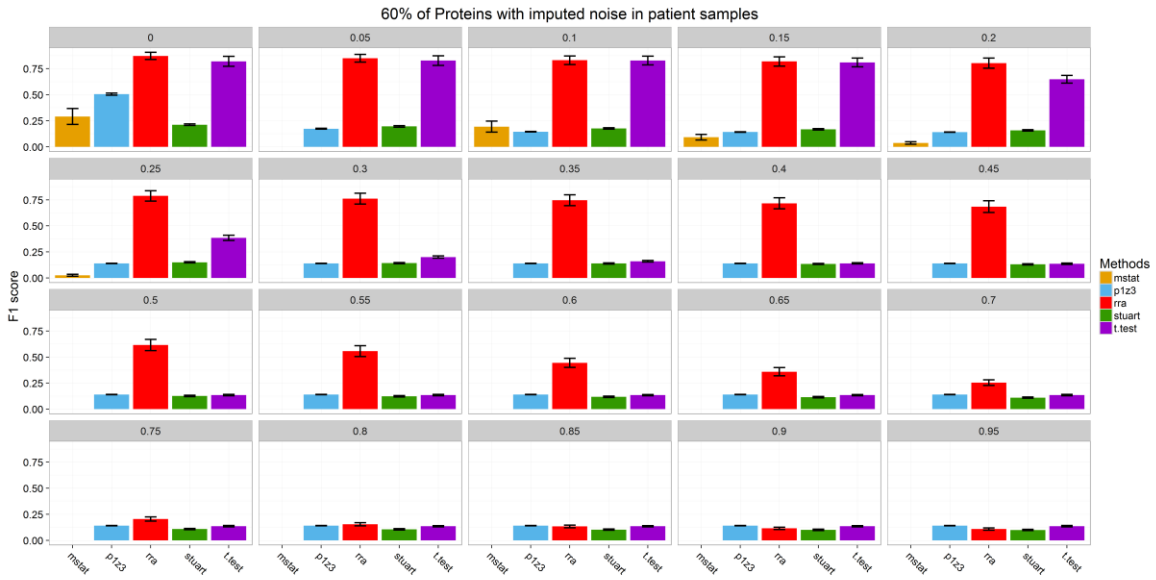

Figure 6A. 60% of proteins got noise inserted into 20 randomly chosen sets of samples. Applied methods on the x-axis, reported F1 scores on the y-axis. First facet represents zero noise dataset. Each error bar represents one standard deviation of F1 scores uncertain

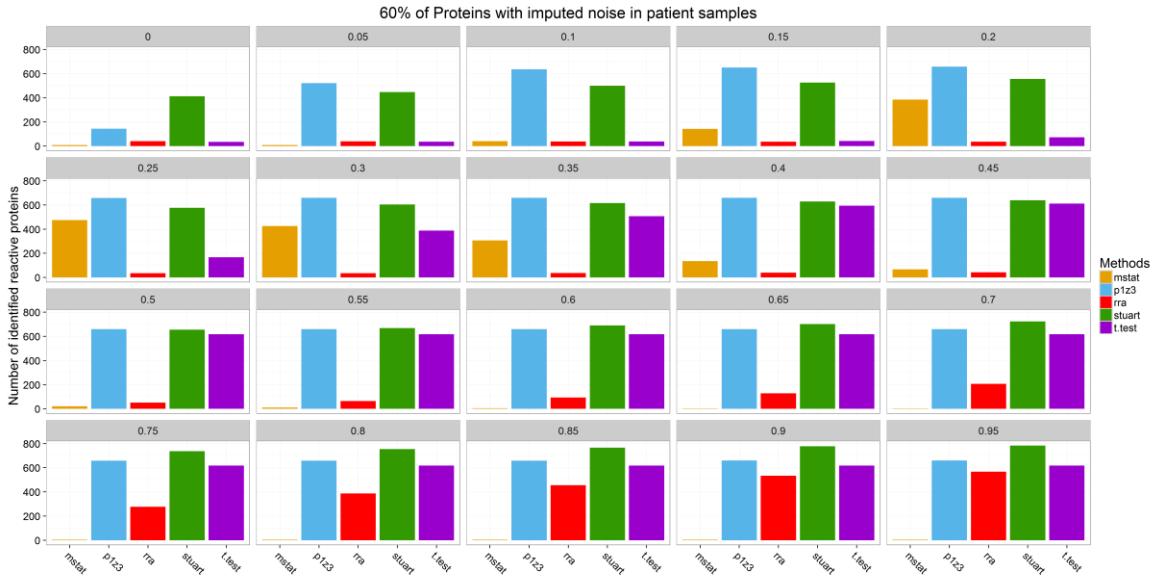

Figure 6B. 60% of proteins got noise insertion per 20 levels of noise in samples. Applied methods on the x-axis, reported number of identified reactive proteins on the y-axis. First facet represents zero noise dataset.

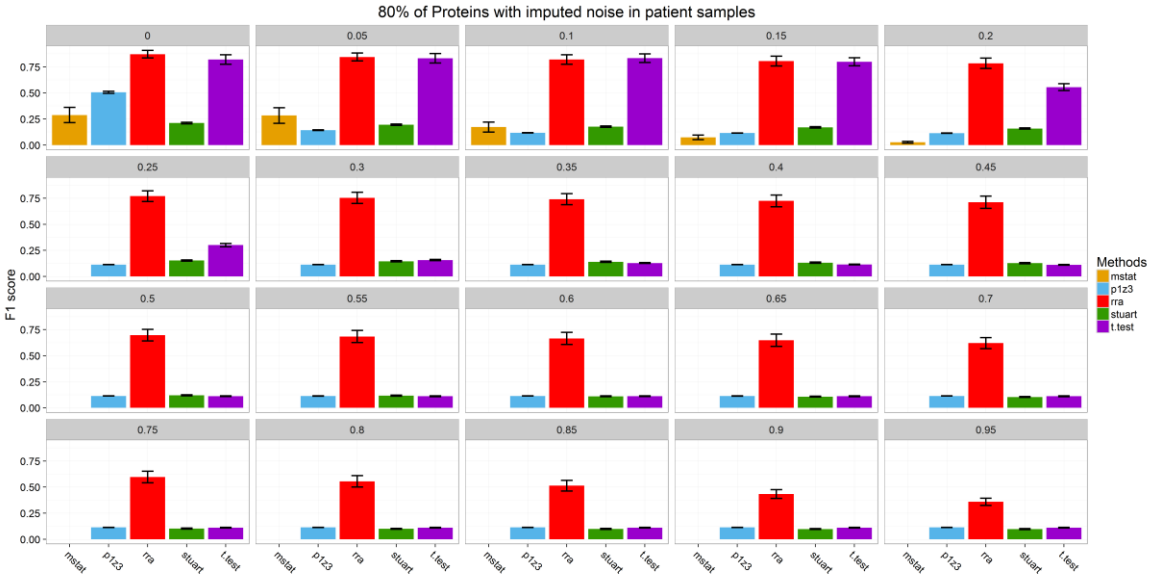

Figure 7A. 80% of proteins got noise inserted into 20 randomly chosen sets of samples. Applied methods on the x-axis, reported F1 scores on the y-axis. First facet represents zero noise dataset. Each error bar represents one standard deviation of F1 scores uncertain

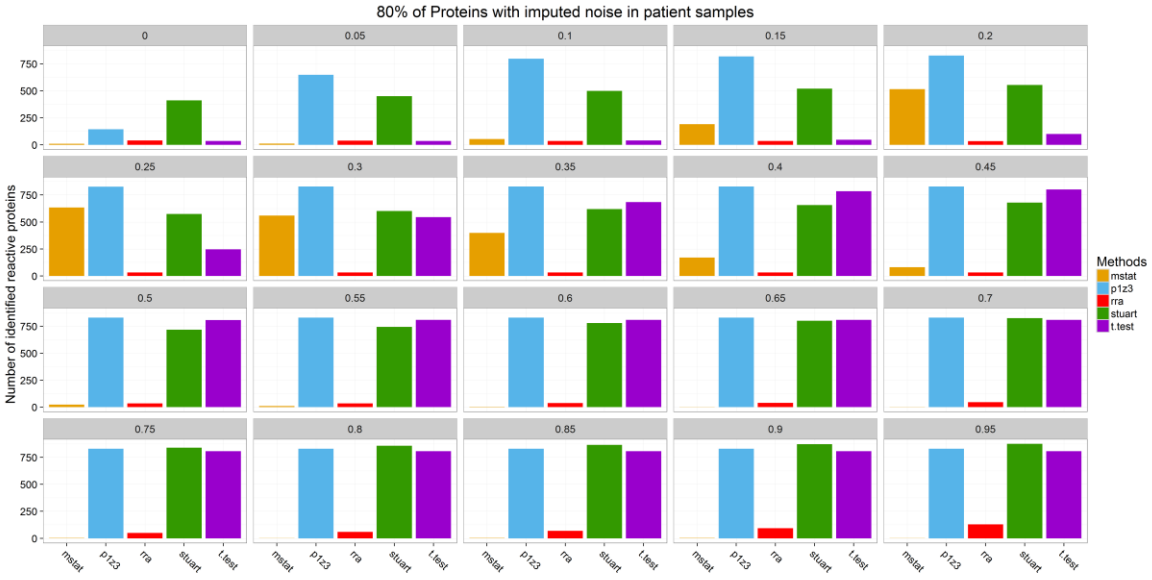

Figure 7B. 80% of proteins got noise insertion per 20 levels of noise in samples. Applied methods on the x-axis, reported number of identified reactive proteins on the y-axis. First facet represents zero noise dataset.

II. License

Non-exclusive licence to reproduce thesis and make thesis public

I, **Vitalii Peretiatko** (date of birth: 25th of May 1993),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:
   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and
   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

of my thesis

Using Robust Rank Aggregation for prioritising autoimmune targets on protein microarrays,

supervised by Dmytro Fishman and Elena Sügis,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, **03.11.2016**