Rule-Based Approaches for Large Biological Datasets Analysis

A Suite of Tools and Methods

MARCEL KRUCZYK
Abstract

This thesis is about new and improved computational methods to analyze complex biological data produced by advanced biotechnologies. Such data is not only very large but it also is characterized by very high numbers of features. Addressing these needs, we developed a set of methods and tools that are suitable to analyze large sets of data, including next generation sequencing data, and built transparent models that may be interpreted by researchers not necessarily expert in computing. We focused on brain related diseases.

The first aim of the thesis was to employ the meta-server approach to finding peaks in ChIP-seq data. Taking existing peak finders we created an algorithm that produces consensus results better than any single peak finder.

The second aim was to use supervised machine learning to identify features that are significant in predictive diagnosis of Alzheimer disease in patients with mild cognitive impairment. This experience led to a development of a better feature selection method for rough sets, a machine learning method.

The third aim was to deepen the understanding of the role that STAT3 transcription factor plays in gliomas. Interestingly, we found that STAT3 in addition to being an activator is also a repressor in certain glioma rat and human models. This was achieved by analyzing STAT3 binding sites in combination with epigenetic marks. STAT3 regulation was determined using expression data of untreated cells and cells after JAK2/STAT3 inhibition.

The four papers constituting the thesis are preceded by an exposition of the biological, biotechnological and computational background that provides foundations for the papers.

The overall results of this thesis are witness of the mutually beneficial relationship played by Bioinformatics in modern Life Sciences and Computer Science.

Keywords: Rough sets, peak finding, gliomas, Alzheimer disease, STAT3, machine learning, feature selection, next generation sequencing

Marcin Kruczyk, Uppsala University,

© Marcin Kruczyk 2013

ISSN 1651-6214
ISBN 978-91-554-8733-1
urn:nbn:se:uu:diva-206137 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-206137)
To my Family and Friends
who have always been there for me
through thick and thin
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the publishers.
## Contents

**Introduction** 9  
**Biological and Biotechnological Background** 11
- Chromatin structure ................................................. 11
- Chromatin immunoprecipitation and DNA sequencing ............ 13
**Computational Methods** 15
- Finding Peaks in ChIP-seq data ................................... 15
- Rough Sets in Bioinformatics ........................................ 15
- Monte Carlo Feature Selection ...................................... 26
**Aims** 28
**Results and Discussion** 29
- Paper I ................................................................. 29
- Paper II ...................................................................... 29
- Paper III ..................................................................... 29
- Paper IV .................................................................... 30
**Future research** 32  
**Concluding Remarks** 33  
**Sammanfattning på svenska** 34  
**Acknowledgements** 36  
**References** 38
Introduction

Recent developments in biotechnology have made it possible to collect unprecedentedly rich data about biological systems. New and multiple roles of various genes and proteins have been found to take part in the cellular processes and in some cases it has been explained how they work. Although the number of successes cannot be overestimated there still remains a huge number of challenges for today’s biology. As we have learnt during the last years many cellular processes cannot be explained with activity of one or two genes or proteins. They are very often complex and may involve various genes on different chromosomes and at first glance they seem not to be related. Some of the cellular processes cannot be explained on the genetic level and involve non-DNA signals. The discipline explaining the phenomena of changes in gene expression or cell phenotypes caused by mechanisms not included in the DNA code is called epigenetics. Moreover, the same genes or proteins may play various roles and functions depending on the type of the cell.

The mechanisms of various diseases are also believed to involve numerous components and thus are very difficult to discover and understand. The mystery of cancer has been investigated for several decades and we still do not know much about the disease. Many genes have been assigned as oncogenes, many as tumour suppressors, thousands of scientists work on the problem in many labs all over the world but the problem remains largely unsolved. There are a number of diseases that have not yet been explained such as Alzheimer disease, Parkinson disease not to mention diabetes type II.

Since the majority of the biological challenges seems to be very complex, a new broader perspective, the so-called whole-genome perspective, has been proposed as an approach to solving such problems. The last decade provided microarrays and Next Generation Sequencing (NGS) as whole-genome technologies. The introduction of microarrays on the market was a breakthrough for the genomic research. They provided a possibility to investigate expression of thousands of genes in one experiment. Also, whole-genome protein-DNA binding studies became possible. The whole-genome approach became a fact. NGS increased these possibilities even further enabling investigation of the DNA and protein sequences with up to 1-base pair resolution and thus leading to a design of new types of experiments.

It is here where new challenges for bioinformatics begin. The biological processes are often complex and multi-parametrical and so is the data describing them. The sizes of such datasets are measured in gigabytes. The number of parameters that are investigated at the same time is calculated in thousands
or sometimes in millions. In order to analyze the data and obtain *legible and easily interpretable models* of biological phenomena it is necessary to employ machine learning methods of a particular kind such as rule-based formalisms.

Today bioinformatics permeates almost all parts of biological experiments. They consist of the following steps:

1. Experiment design
2. Wet-lab experiment
3. Data collection
4. Data analysis and model construction
5. In-silico validation
6. Interpretation of the results
7. Wet-lab validation
8. Repeat the procedure, if necessary

The only two steps where bioinformaticians are not needed are Wet-lab experiments and Wet-lab validation. Data analysis and model construction are becoming a dominant part of biological experiments. Taking into account this step in the experiment design will likely enable or at least make it easier and faster to obtain the results. Data analysis and model construction are still in early stages and new methods often need to be developed. There exist pipelines for analyzing microarray data, but sequencing datasets still lack this kind of a standard. In-silico validation is a very important step of the analysis since it restricts hypothetical findings to those that are statistically correct. In practice, it prevents running unnecessary wet-lab validation. It is only after the wet-lab validation that the result has a biological value. We make a point that models should be legible and easily interpretable. They are computational or mathematical constructs that need to be easily accessible to non-experts of computer science or mathematics. In this study, we believe we developed such new methods and pipelines and employed them in biological projects.

Before presenting the results in the form of four papers we give a necessary background to our research and define problems that we wanted to solve in these studies. In what follows we provide a description of the biological background, biotechnological developments and a basic exposition of the computational methods applied in this thesis.
Biological and Biotechnological Background

Chromatin structure

Chromatin consists of DNA and proteins situated in the nucleus of a eukaryotic cell. The length of DNA in each human cell is around 3 m. Therefore, to fit the DNA particle in a cell nucleus, which on average has a diameter of 6 \( \mu \text{m} \), a special structure and a compact manner of packaging and organization are needed. The structure of DNA can be divided into three categories (Fig 1).

**Primary structure:** DNA is wrapped around proteins called histones forming chromatin units - nucleosomes.

**Secondary structure:** Multiple nucleosomes form 30 nm fibre creating the most compact form of chromatin called heterochromatin.

**Tertiary structure:** Higher, 3-dimensional levels of DNA packaging (e.g. chromosomes)

![Figure 1. Levels of DNA packaging in eukaryotic cells. Adopted from the webpage of University of Edinburgh (http://www.ccm.mvm.ed.ac.uk/content/molecular-mechanisms-stress).](image)

In this study we will focus on the first category since we investigated gene expression regulation using only the information from the primary structure of DNA.

Apart from DNA itself, histones are the most basic structural elements of chromatin. Histone octamers i.e. complexes of eight cooperating proteins form the main skeleton for the DNA. A histone octamer consists of pairs of the following histones: H2A, H2B, H3 and H4. The DNA double helix is wrapped around the core of eight histones making 2.25 loops. The length of DNA that is wrapped around a histone core is 147 base pairs. The histone core
with DNA wrapped around is a basic DNA unit and is called nucleosome (Fig. 2).

Histones H3 and H4 forming the nucleosome core have long tails that may be covalently modified (e.g. methylated, phosphorylated, acetylated). Moreover, the whole core may also be modified. It has been proven that histone modifications take part in various processes in the cell such as, for instance, gene expression regulation or DNA repair, and that different combinations may be associated with different functions. Combinations of histone modifications in the cell are called the histone code and are one of the most important branches of epigenetics, i.e. a discipline explaining phenomena of changes in gene expression or cell phenotypes caused by mechanisms not included in the DNA code. In our research, we focused on the association of two of the histone modifications, namely H3K4me3 (H3 fourth lysine tri-methylation) and H3ac (H3 acetylation), with gene expression regulation.
Chromatin immunoprecipitation (ChIP) is a modern experimental procedure for DNA-protein interaction investigation. It is often used for determining the actual transcription factor binding sites and localizing the modified histones. The ChIP procedure consists of several steps. Firstly, the proteins are cross-linked to DNA. Usually formaldehyde is used to achieve the association of the proteins with the DNA. Secondly, the DNA is sheared. It can be performed by employing restriction enzymes or sonication. Thirdly, the proteins of interest are immunoprecipitated i.e. an antibody coupled with a solid substrate binds to a protein and the protein-DNA complex is pulled out from the chromatin soup. The antibody is designed to bind only the proteins of interest. The solid substrate coupled with the antibody can be, for example, a superparamagnetic microbead, but there are also other techniques. Finally, the DNA is purified and the DNA fragments obtained in this way are considered to interact with the protein of interest in vivo.

After ChIP, the fragments of DNA may be analyzed with several techniques. We focus on two of them. The older one but still widely used is by means of microarrays. The DNA fragments are hybridized to a pre-designed microarray, and the signal from the microarray is read and mapped onto a reference genome. This combination of technologies is called ChIP-chip. The alternative to microarrays is sequencing. The DNA fragments are sequenced and the read sequences are mapped onto a reference genome. This technology is called ChIP-seq. Apart from microarrays and sequencing the ChIP procedure can be followed by PCR or qPCR. The procedures of ChIP-chip and ChIP-seq are illustrated in Fig 3.

ChIP-seq is the most recent technology employing ChIP. It is still under development. Nevertheless, it has already opened several paths of analysis that were not possible before. For instance, it is now possible to investigate the influence of single nucleotide polymorphisms (SNPs) i.e. one base pair mutation on binding of a protein to DNA and therefore on gene expression regulation.
Figure 3. Chromatin immunoprecipitation (ChIP) procedure followed by alternative technologies of DNA fragments analysis.
Computational Methods

Finding Peaks in ChIP-seq data

From a ChIP-seq experiment we get a number of fragments of DNA mapped to a reference genome with their coordinates in the genome, the so-called *reads*. The next challenge for the analysis is to find regions of DNA enriched in the ChIP-seq reads obtained from the experiment, the so-called *peaks*. There exist a number of applications for finding peaks in ChIP-seq data called *peak finders*. *Window-based, overlap-based* and *Hidden Markov Model-based* approaches emerged as the main approaches to finding peaks in such data.

In the window-based approach the program calculates a number of reads in *candidate regions*, which are moved over the entire chromosome. The start and stop locations of the regions and the threshold defining a peak are needed. In the overlap-based approach the program looks for local maxima over the read overlaps and on that basis it identifies the enriched regions. In the Hidden Markov Model-based approach the read accumulation along the genome is described as a sequence of two different states: peak and background; the definition includes also a transition function between these two states. More details about the functioning of popular peak finders can be found in an excellent review of Laajala and colleagues [17].

However, there is no golden standard for defining peaks and thus the results returned by the methods differ significantly [24]. Finding a consensus method for finding peaks in ChIP-seq data was one of the main aims of our studies.

Rough Sets in Bioinformatics

Since biological data is today usually very large and multi-parameter, and since we wanted to provide legible and easily interpretable models for such data, we had to turn to appropriate machine learning techniques such as *rough sets*. Rough set is a well defined concept developed by Zdzisław Pawlak [21], [22]. The theory is based on logics and Boolean reasoning and in particular on the upper and lower approximations of a set, the approximation space, and the modeling of sets [30]. The theory provides tools for solving classification challenges. The tools produce models the quality of which is comparable with other widely used methods. The added value, which makes rough sets a preferred tool for classification problems, is the structure of the classification model, that is minimal subsets of attributes sufficient for classification.
and their combinations. In certain cases we would also like an ordering of the attributes.

Rough sets can be applied to incomplete, inconsistent and noisy data. Rough sets models consist of sets of minimal rules, which can be easily interpreted by the user.

Following the tutorial by Komorowski et al. [16] rough sets usually operate on decision systems. A decision system is a table where each row represents an object and each column represents an attribute (also known as a feature) and the last attribute is called the decision attribute. Formally, we have a pair \( A = (U, A \cup \{d\}); d \notin A \), where \( U \) is the universe of objects, \( A \) is a set of attributes called conditions and \( d \) is the decision attribute (also known as the classification outcome). Both \( U \) and \( A \) are non-empty and finite. Each object from \( U \) has a value (can be a missing value) for each attribute from \( A \) and for the decision attribute \( d \).

\[
\forall a \in A : U \rightarrow V_a
\]

where:
\( V_a \) - a set of possible values for attribute \( a \), i.e. the value set of \( a \).

\[
d : U \rightarrow V_d
\]

where:
\( V_d \) - a set of possible values for decision attribute \( d \), i.e. the value set of \( d \).

**Table 1. An example of decision system.**

<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>Height</th>
<th>Weight</th>
<th>APOE expression</th>
<th>Alzheimer risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>M</td>
<td>≥ 50</td>
<td>&lt; 170</td>
<td>&lt; 60</td>
<td>↑</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>F</td>
<td>&lt; 50</td>
<td>&lt; 170</td>
<td>≥ 60</td>
<td>0</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>M</td>
<td>&lt; 50</td>
<td>≥ 170</td>
<td>≥ 60</td>
<td>↑</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>M</td>
<td>≥ 50</td>
<td>&lt; 170</td>
<td>&lt; 60</td>
<td>↑</td>
</tr>
<tr>
<td>( x_5 )</td>
<td>F</td>
<td>≥ 50</td>
<td>≥ 170</td>
<td>&lt; 60</td>
<td>0</td>
</tr>
<tr>
<td>( x_6 )</td>
<td>F</td>
<td>≥ 50</td>
<td>≥ 170</td>
<td>≥ 60</td>
<td>0</td>
</tr>
<tr>
<td>( x_7 )</td>
<td>F</td>
<td>&lt; 50</td>
<td>&lt; 170</td>
<td>≥ 60</td>
<td>↑</td>
</tr>
<tr>
<td>( x_8 )</td>
<td>F</td>
<td>≥ 50</td>
<td>≥ 170</td>
<td>&lt; 60</td>
<td>0</td>
</tr>
</tbody>
</table>

Many biological tasks lend themselves to machine learning-driven classification. Aspects of this approach have been described in [1]. Here, we focused on legible and interpretable models that received less attention there. Frequently, the classification outcome is known and thus we can use the approach of supervised learning. Once such outcomes are available, it is possible to
create a *decision system*. Table 1 shows a sample decision system. The system contains information about patients such as *Sex* or *Age* and a decision attribute - *Alzheimer risk*.

A decision system can carry redundant information. There may be many indiscernible objects and there can be many attributes not necessary from the classification point of view. Let us start from defining the *equivalence relation*. 

\[ R \subseteq X \times X \text{ is is called an equivalence relation if and only if it is reflexive, symmetrical and transitive.} \]

The *equivalence class* of an element \( y \) denoted \( [y] \) is defined as the set \( [y] = \{ x \in X \mid yRx \} \).

As one may notice, in our example the objects \( x_5 \) and \( x_8 \) are indiscernible, while objects \( x_1 \) and \( x_4 \) are also indiscernible but have different decision values. To illustrate the ability of rough sets to handle uncertainty explicitly we introduce for the purpose of this summary presentation a new class *HighOrLow* for the indiscernible objects with different outcomes, \( x_1 \) and \( x_4 \) (Table 2).

**Table 2. An example of a decision system with unique outcomes.**

<table>
<thead>
<tr>
<th>Sex</th>
<th>Age</th>
<th>Height</th>
<th>Weight</th>
<th>APOE expression</th>
<th>Alzheimer risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>M</td>
<td>≥ 50</td>
<td>&lt; 170</td>
<td>&lt; 60</td>
<td>↑</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>F</td>
<td>&lt; 50</td>
<td>&lt; 170</td>
<td>≥ 60</td>
<td>0</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>M</td>
<td>&lt; 50</td>
<td>≥ 170</td>
<td>≥ 60</td>
<td>↑</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>M</td>
<td>≥ 50</td>
<td>&lt; 170</td>
<td>&lt; 60</td>
<td>↑</td>
</tr>
<tr>
<td>( x_5 )</td>
<td>F</td>
<td>≥ 50</td>
<td>≥ 170</td>
<td>&lt; 60</td>
<td>0</td>
</tr>
<tr>
<td>( x_6 )</td>
<td>F</td>
<td>≥ 50</td>
<td>≥ 170</td>
<td>≥ 60</td>
<td>0</td>
</tr>
<tr>
<td>( x_7 )</td>
<td>F</td>
<td>&lt; 50</td>
<td>&lt; 170</td>
<td>≥ 60</td>
<td>↑</td>
</tr>
<tr>
<td>( x_8 )</td>
<td>F</td>
<td>≥ 50</td>
<td>≥ 170</td>
<td>&lt; 60</td>
<td>0</td>
</tr>
</tbody>
</table>

Given a decision system \( \mathcal{A} = (U, A \cup \{d\}) \) we define *C-indiscernibility* relation \( IND_{\mathcal{A}}(C) \) for any subset of attributes \( C \subseteq A \): 

\[
IND_{\mathcal{A}}(C) = \left\{ (x, x') \in U^2 \mid \forall a \in C \ a(x) = a(x'), \ d(x) = d(x') \right\}
\]  

We call \( x \) and \( x' \) indiscernible using attributes \( C \), if \( (x, x') \in IND_{\mathcal{A}}(C) \). We denote the equivalence classes of the above relation as \( [x]_C \).

Employing equivalence classes and set theoretic operations we may attempt to construct sets such as Low Alzheimer Risk or High Alzheimer Risk. Sometimes such sets will be exact, at other times they will be approximate. We have examples of such classes in Table 1. As we noticed previously objects \( x_1 \) and \( x_4 \) cannot be unambiguously assigned to any of these classes. Other objects in the universe can be assigned strictly to one class. For example, objects \( x_5, x_8 \) having the same set of attribute values, both belong to class *Low*.  

17
Given a decision system $\mathcal{A} = (U, A \cup \{d\})$, a subset of attributes $C \subseteq A$ and a set of objects $X \subseteq U$ $C$-lower and $C$-upper approximations are constructed, respectively, as follows:

$$C^X = \{x \mid [x]_C \subseteq X\} \quad (4)$$

and

$$\overline{C}^X = \{x \mid [x]_C \cap x \neq \emptyset\} \quad (5)$$

The boundary region is now defined to be:

$$BN_C(X) = \overline{C}X - C^X \quad (6)$$

It means that objects that cannot have a uniquely assigned class label belong to a boundary region. The boundary region covers the new decision class $\text{HighOrLow}$ shown in Table 2. If the boundary region for a decision system is not empty, the set defined by the system is rough. In other words, if $BN_C(X) = \emptyset$, $X$ is crisp, otherwise it is rough. An example of a rough set is illustrated in Fig. 4.

![Figure 4. Visualization of a rough system. Example of partitioning of a universe into classes High and Low and a boundary region marked as light grey (HighOrLow). The area marked dark grey is the C-lower approximation area and the area marked with light grey is the C-upper approximation area.](image)
The decision system from Table 1 is indeed rough. Let \( H = \{ x \mid \text{Alzheimer risk}(x) = \text{High} \} \). Approximations that we obtain will be as follow:

\[
\overline{A}H = \{ x_1, x_3, x_4, x_7 \}
\]

and

\[
\underline{A}H = \{ x_3, x_7 \}
\]

Hence, the boundary region is:

\[
BN_A(H) = \{ x_1, x_4 \}
\]

The boundary region is not empty and concept \textit{High} is rough.

The most important construct in rough sets theory is \textit{reduct}. It is a minimal subset of features of the original dataset that preserves the information needed for classification i.e. it preserves discernibility. Of course, there may be many different reducts. Computing equivalence classes is a conceptually simple task. However, it has been proven that finding all \textit{reducts} is an NP-hard problem [25] and is one of the challenges for computing with rough sets.

The same concept has been used in relational databases under the name of \textit{functional dependency}, but the methods reducing and minimizing similar functions have been originally explored in the context of designing logical circuits in electronics. This logic is based on the \textit{Boolean variables} i.e. variables that can take only logical values \textit{True} (usually marked by 1) and \textit{False} (usually marked by 0). The construction of Boolean expressions is based on \textit{truth tables}, i.e. decision systems where all condition attributes and the decision attribute are Boolean variables. There are several methods of minimizing truth tables, for example, Karnaugh Method [12], Quine-McCluskey algorithm [18], or Espresso heuristic logic minimizer [3]. The logic of Boolean equations is funded in the ‘algebra of logic’ created in 1847 by George Boole [2].

Reducts are constructed as follows. Given a decision system as defined in Table 2, \( \mathcal{A} = (U, A \cup \{d\}) \), we first find the equivalence classes for all objects in \( U \) using all attributes in \( A \).

\[
IND_{\mathcal{A}}(A) = \{ \{ x_1, x_4 \}, \{ x_2 \}, \{ x_3 \}, \{ x_5, x_8 \}, \{ x_6 \}, \{ x_7 \} \}
\]

There are the following equivalence classes:

\[
\begin{align*}
[x_1] &= \{ x_1, x_4 \} \\
[x_2] &= \{ x_2 \} \\
[x_3] &= \{ x_3 \} \\
[x_5] &= \{ x_5, x_8 \} \\
[x_6] &= \{ x_6 \} \\
[x_7] &= \{ x_7 \}
\end{align*}
\]
Now we can construct a *discernibility matrix* i.e. a $n \times n$ matrix, where $n$ is the number of equivalence classes, containing disjunctions of attributes that discern objects from pairs of equivalence classes with different decision values. In each cell $c_{ij}$ of the table there is a set of attributes which discern between equivalence classes $i$ and $j$. If objects from classes $i$ and $j$ have the same decision attribute, then it is the empty set which is put into such cell of the matrix i.e. $c_{ij} = \emptyset$.

**Table 3. A discernibility matrix for the decision system from Table 2.**

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$X_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>\emptyset</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td>Sex, Age, Weight, APOE</td>
<td>\emptyset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td>Age, Height, Weight</td>
<td>\emptyset</td>
<td>\emptyset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_5$</td>
<td>Sex, Height, APOE</td>
<td>\emptyset</td>
<td>\emptyset</td>
<td>\emptyset</td>
<td></td>
</tr>
<tr>
<td>$X_6$</td>
<td>Sex, Height, Weight, APOE</td>
<td>Age, Height</td>
<td>Sex, Age, APOE</td>
<td>Weight</td>
<td>\emptyset</td>
</tr>
<tr>
<td>$X_7$</td>
<td>Sex, Age, Weight</td>
<td>APOE</td>
<td>Sex, Height</td>
<td>Age, Height, Weight, APOE</td>
<td>\emptyset</td>
</tr>
</tbody>
</table>

The next step in reduct calculation is calculation of *discernibility function*. *Discernibility function* is a Boolean function of $m$ variables $a_1, ..., a_m$

$$f_{\mathcal{A}}(a_1, ..., a_m) = \bigwedge \left\{ \bigvee c_{ij}^* \mid 1 \leq i \leq j \leq n, c_{ij}^* \neq \emptyset, d(i) \neq d(j) \right\}$$ \hspace{1cm} (11)

where:

$$c_{ij}^* = \{ a \in c_{ij} \}$$ \hspace{1cm} (12)

Our discernibility function is:

$$f_{\mathcal{A}}(A) = (\text{Sex} \lor \text{Age} \lor \text{Weight} \lor \text{APOE}) \land (\text{Age} \lor \text{Height} \lor \text{Weight})$$
$$\land (\text{Sex} \lor \text{Height} \lor \text{APOE}) \land (\text{Sex} \lor \text{Height} \lor \text{Weight} \lor \text{APOE})$$
$$\land (\text{Age} \lor \text{Height}) \land (\text{Sex} \lor \text{Age} \lor \text{APOE}) \land (\text{Weight})$$
$$\land (\text{Sex} \lor \text{Age} \lor \text{Weight}) \land (\text{APOE}) \land (\text{Sex} \lor \text{Height})$$
$$\land (\text{Age} \lor \text{Height} \lor \text{Weight} \lor \text{APOE})$$ \hspace{1cm} (13)
To simplify the function we use the following Boolean law:

\[ A \land (A \lor B) = A \] (14)

And the function after simplification is:

\[ f_{\text{IF}}(A) = (\text{APOE} \land \text{Weight} \land \text{Height}) \lor (\text{APOE} \land \text{Weight} \land \text{Sex} \land \text{Age}) \] (15)

So:

\[ \text{IND}_{\text{IF}}(A) = \text{IND}_{\text{IF}}(\{\text{APOE}, \text{Weight}, \text{Height}\}) \]
\[ = \text{IND}_{\text{IF}}(\{\text{APOE}, \text{Weight}, \text{Sex}, \text{Age}\}) \] (16)

This computational problem NP-hard, as previously stated, and heuristics must be used to find reducts in real life applications. The publicly available ROSETTA [20] may be used to compute reducts.

Attributes APOE and Weight appear in both reducts. Therefore, they seem to be important factors in predicting the risk of Alzheimer disease. On the level of reducts users have the attributes important for classification already selected. However, there is no ranking or any other measure of importance of the attributes. The process of reduction of the attributes is very important to keep the rough set models transparent and legible. In the real datasets we rarely deal with a situation where there is a sharp border between the classes and where we can generate reasonably short reducts. Therefore, some approximate solutions have been developed which make the rough sets applications more universal and applicable to any kinds of decision systems. Concepts of approximate reducts have been developed by Ślęzak [26]. More details about approximate reducts can be found in [16].

By overlaying objects over the reducts we generate the IF-THEN rules. Rough set models are legible and transparent as mentioned before. The rules are very intuitive and are easily interpreted by non-experts. The syntax of the rules is as follows:

\[ \text{IF condition1 AND condition2 AND ... AND conditionN THEN Outcome} \]

The part of the rule that is placed between 'IF' and 'THEN' is called the condition part of the rule, while the part of the rule that goes after 'THEN' is called the decision part of the rule. The decision assigns objects to one of possible classification classes. The model consists of the rules produced from all detected reducts. Examples of rules in the model based on Table 1 are:

```
IF Height < 170 AND Weight < 60 AND APOE = ↑ THEN AD = Low OR High
IF Height ≥ 170 AND Weight ≥ 60 AND APOE = ↑ THEN AD = Low
IF Height ≥ 170 AND Weight ≥ 60 AND APOE = 0 THEN AD = High
IF Sex = M AND Age ≥ 50 AND Weight < 60 AND APOE = ↑ THEN AD = Low OR High
IF Sex = F AND Age < 50 AND Weight ≥ 60 AND APOE = 0 THEN AD = Low
IF Sex = F AND Age < 50 AND Weight ≥ 60 AND APOE = ↑ THEN AD = High
```
In our case the model consists of 12 rules, but not all of them are equally important. For each rule we define its **Support** and **Accuracy**. **Support** is the number of objects fulfilling the condition part of the rule, or firing the rule. **Accuracy** is a vector of parameters calculated for each outcome returned by the rule, therefore the length of the accuracy vector is equal to the number of the rule outcomes. For each rule, **Accuracy** is defined as follows:

\[
\text{Accuracy}_{d} = \frac{\left| \{x \in U : f(x) = d\} \right|}{\text{Support}}
\]  

\[(17)\]

where:

- \( f(x) \) - classification outcome for object \( x \).
- **Support** - the value of support parameter for the rule.

It is important to verify the quality of the constructed model. If we have two different datasets, we can construct a model on one of them, called the **training set**, and test it on the other - the **testing set**. We customarily construct a **confusion matrix** i.e. a table juxtaposing the true classification of objects from the testing dataset with the predicted classification returned by the model. An example of a hypothetical **confusion matrix** is given in Table 4.

**Table 4. A confusion matrix.**

<table>
<thead>
<tr>
<th>Actual classes</th>
<th>Predicted classes</th>
<th>Low</th>
<th>High</th>
<th>Undefined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>22</td>
<td>2</td>
<td>1</td>
<td>88.00%</td>
</tr>
<tr>
<td>High</td>
<td>3</td>
<td>28</td>
<td>4</td>
<td>80.00%</td>
</tr>
<tr>
<td>Undefined</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Undefined</td>
</tr>
</tbody>
</table>

88.00% 93.33% 0.00% 83.33%

In the rows of the **confusion matrix** we have the actual number of the objects in each class. For example, in Table 4 there were 22 + 2 + 1 = 25 objects belonging to class **Low**. 22 of them were classified as **Low**, 2 as **High** and 1 was **Undefined**, that is, there was no rule that fired for that object. The last column contains the accuracy of the model for objects from this class (88%). In the columns we have the number of objects assigned to a class by the model. For example, the model assigned 2 + 28 = 30 objects to class **High**, 28 of them were indeed from class **High**. In the last row we can see how many objects assigned to a class by the model truly belong to this class (for example 93.33% of objects classified as **High** were indeed **High**). In the bottom right cell of the matrix we can find the overall accuracy of the model i.e. the ratio of accurately classified objects to all objects classified by the model. In our case there were 22 + 28 = 50 objects assigned to correct classes out of 22 + 2 + 1 + 3 + 28 + 4 = 60 objects classified by the model. Hence, the ratio is 83.33%.
Another measure of the quality of classification of the elements from one of the classes is the Area Under the Receiver Operating Characteristic (AUC). First, we construct the ROC curve. This curve is a graphical representation of the quality of classification of a binary classifier. Before plotting the ROC curve we need to calculate True Positive Rate (TPR) and False Positive Rate (FPR). TPR is the ratio of number of correctly assigned objects from the class of interest to all classified objects from the class (Undefined objects do not count as classified and are not included in these calculations). FPR is the ratio of number of incorrectly assigned objects from different classes to the class of interest to all classified objects assigned to that class. The ROC graph is created by plotting TPR vs. FPR. In case of a single rough set model there will be only one break point on the curve. In the example from Table 4 TPR for class Low \((TPRLow)\) is

\[
TPRLow = \frac{22}{22 + 2} = 0.9167
\]

and FPR for class Low \((FPR_{Low})\) is

\[
FPR_{Low} = \frac{3}{28 + 3} = 0.0967
\]

Now we can plot the ROC curve (see Fig. 5).

\[\text{Figure 5. ROC curve for class Low for the model from Table 4. The marked area is the representation of } AUC_{Low}.\]
Having the ROC curve, we can calculate the area under the curve (the marked polygon on the plot). In our case
\[
AUC_{Low} = (1 - 0.0967) \times 0.9167 + 0.5 \times 0.9167 \times 0.0967 \\
+ 0.5 \times (1 - 0.0967) \times (1 - 0.9167) = 0.91
\]

Of course, the higher the AUC parameter the higher the quality of the model, since the model detects most of the True Positives i.e. has high sensitivity and includes a low number of False Positives in the results i.e. it has high specificity. Usually the ROC curves are used to compare various classifiers. In such case each classifier is represented by one point on a ROC curve and we can choose the trade between sensitivity and specificity.

A rough set model assigns objects to classes using a voting procedure. Each rule in the model casts a number of votes equal to its Support. By parametrizing the fraction of votes needed to assign an object to a class of interest we can modify the model. The higher is the predefined fraction, the higher the specificity but the lower the sensitivity. It is possible to repeat the calculations for different fractions and generate a ROC curve for them to choose the most convenient value of this parameter. If the decision classes are of equal interest we usually use the standard voting procedure. It assigns each object to the class that received the majority of the votes.

Not always a test set is available. Where it is not, testing the quality of the model is done with the m-fold cross-validation procedure. We split the dataset into m subsets, m times we train the model (build a model) on m-1 merged subsets and test it on the remaining one, so that each of the m subsets is used once as the testing set. Usually m is set to 10. Other schemata are also employed. For each of the iterations we compute a confusion matrix and obtain the accuracy of the model. AUC can be also computed in these iterations.

In order to check if the model is not an artifact of some random properties a permutation test is employed. The permutation test is a method for an experimental evaluation of statistical significance of a model. Firstly, we compute the model on the original data and find its accuracy or AUC. Then the values in the attribute columns or in the decision column are randomly reshuffled k times and a new model is constructed from the new decision system. n accuracy or AUC values are collected. The standard quality measure is an empirical p-value i.e. the ratio of models built on permuted data with the quality measure higher or equal to the original quality measure to n + 1. More formally, p-value of this test is calculated as follows:
\[ p \text{-value} = \frac{\sum_{k=1}^{n} I[Q(M_k) > Q(M_{\text{org}})]}{n + 1} \] (18)

where:
- \( n \) - is the number of permutations.
- \( Q() \) - is the quality measure of a model.
- \( M_k \) - is the model built in iteration \( k \) of the permutation test.
- \( M_{\text{org}} \) - is the original model built on unbiased data.

Rough set theory has been developed for over 30 years. The information included in this section is just a basic introduction to the concept of rough sets. There are many more issues and challenges about rough sets. Several of them are discussed in [16]. Rough sets have been used to solve many problems, including bioinformatic ones. These include modeling HIV-1 Reverse Transcriptase Resistome [14], Analysis of Molecular Interaction Networks Underlying Change of HIV-1 Resistance to Selected Reverse Transcriptase Inhibitors [13], cancer research (eg. [5, 8, 19, 33]) and profiling gene expression [4, 9, 11, 32].

This theory has been used to model a broad spectrum of problem starting from the molecular level and ending on a macro scale. An example of the latter is the diagnosis and prognosis in medicine including identification of hospice candidates in terminally ill patients [10]. Other examples of rough sets applied to diagnosis was a model for early diagnosis of congestive heart failure [28] or better applicability of sagittal abdominal diameter in identifying high risk patients. Earlier applications include [15, 27, 31].

The role of rough sets theory in bioinformatics is steadily increasing and every year there appear new successful applications of rough sets in Life Sciences.

Despite the advantages of rough sets not all issues have found good solutions yet. One of them is the need to rank attributes of a decision system with respect to their significance to the classification power. With the increasing size of data the existing algorithms and implementations of rough sets are reaching their practical computational limits. One remedy is to move to multi-thread computations. To address the first problem we turned to Monte Carlo Feature Selection (MCFS) [7]. The second issue was outside the scope of this thesis.
Monte Carlo Feature Selection

MCFS is an efficient method of finding features informative for classification including features that discriminate between classes only in combination with some other features. It is a heuristic dedicated to very large datasets that returns a ranked subset of features. The ranking is scored in terms of utility for classification. MCFS works as follows. Firstly, $s$ subsets of $m$ features randomly selected from the original data set are created. The subsets are then split $t$ times into training and testing sets. For each of the $t$ splits, a decision tree classifier is built on the training set and tested on the testing set and is called \textit{t-fold cross-validation}, see Fig. 6 adapted from [7].

\begin{equation}
\text{wAcc} = \frac{1}{c} \sum_{i=1}^{c} \frac{n_{ii}}{n_{i1} + n_{i2} + ... + n_{ic}}
\end{equation}

where $c$ is the number of classes and $n_{ij}$ the number of samples from class $i$ classified as those from class $j$. These scores are then used to calculate \textit{Relative Importance (RI)} of a feature.

\begin{equation}
RI_{g_k} = \sum_{\tau=1}^{s t} (\text{wAcc}_{\tau})^u \sum_{n_{g_k}} IG(n_{g_k}(\tau)) \left( \frac{\text{no.in} n_{g_k}(\tau)}{\text{no.in} \tau} \right)^v
\end{equation}

where summation is over all the $s \cdot t$ trees and, within each $\tau$-th tree, overall nodes $n_{g_k}(\tau)$ of the tree on which the split is made on feature $g_k$, $IG(n_{g_k}(\tau))$ stands for information gain for node $n_{g_k}(\tau)$, $\text{no.in} n_{g_k}(\tau)$ denotes the number of samples in node $n_{g_k}(\tau)$, $\text{no.in} \tau$ denotes the number of samples in the root of the $\tau$-th tree, and $u$ and $v$ are fixed positive reals (usually set by default to one).

Another challenge in the feature selection process is a calculation of the \textit{cut-off point} for inclusion of significant features. To this end Dramiński and
colleagues employed 30 permutation tests. Using the $RI$ parameter for a feature in the original unbiased dataset the authors obtain a $p$-value. Although no statistical proof can be provided, the authors observed that the distribution of $RI$'s can be approximated with a Gaussian distribution and 30 permutations were essentially sufficient to obtain a stable distribution (private communication).

MCFS is a successful approach to feature selection, yet it suffers from some inconveniences. The method is fast, but for the continuously growing sizes of biological datasets not always fast enough. Moreover, MCFS does not return shadowed features i.e. features carrying the same or partially the same information, important from the classification point of view. Such features need not be important to classification itself, but may be useful in the interpretation of the model. It is possible to find the shadowed features by repeating the MCFS procedure several times and removing the top features from the ranking after each MCFS run [6], but the method is time consuming and somewhat inconvenient.
Aims

Without doubt, in recent years informatics has had a great impact on Life Sciences and many discoveries would not be possible without advanced algorithms. Likewise, biology has significantly influenced Computer Science. This observation is not new. Already in mid 1970’s Stanisław Ulam expressed this opinion. This thesis attempts to make such a two-sided contribution. Specifically our aims were:

i. To create methods dedicated to ChiP-seq data analysis
ii. To contribute to diagnosis of Alzheimer disease
iii. To contribute to improved understanding of gene regulation in gliomas
Results and Discussion

Paper I

Our main interest focused on improving peak finding in ChIP-seq data. To this end we developed Peak Finder Metaserver (PFMS), which to the best of our knowledge is the first application of a meta-server concept to finding peaks. PFMS combines results from currently up to seven known peak finders and produces consensus results. The approach has been tested on three transcription factors datasets published in [24] and outperformed any single peak finder. We calibrated PFMS on the data and proposed a new standard for finding peaks in transcription factor datasets. In order to find a standard for other types of data such as histone modification ChIP-seq validated datasets are needed. The results of our work on PFMS are presented in Paper I.

Paper II

In Paper II we set on enriching rough set theory with an algorithm for feature selection and ranking. Using the framework of MCFS we developed Random Reducts (RR). In contrast to MCFS, instead of using decision trees as a classifier, the rough sets were used. This change allowed a simplification in the validation steps of the MCFS algorithm and resulted in RR being much faster than the original MCFS. The rankings returned by RR have similar quality as the MCFS ones. In addition, RR includes shadowed features. Experiments with RR suggest that it has a potential for finding interacting features.

Paper III

In Paper III we constructed a diagnostic model of Alzheimer disease in patients with mild cognitive impairment (MCI). This model was a basis for comprehensive risk estimation for future Alzheimer disease. Risk estimates were generated on the basis of age, gender, Mini-Mental State Examination scores, apolipoprotein E genotype and the cerebrospinal fluid biomarkers: total tau, phospho-tau181 and the 42 amino acid form of amyloid b in two sets of longitudinally followed MCI patients. The analysis of the importance of attributes showed that the biochemical CSF markers contributed the most to the predictions, and that added value was gained by combining several biochemical markers. Despite a correlation with the biochemical markers, the genetic
marker APOE ε4 did not contribute to the predictive power of the model. The model proposed by us has had very high accuracy and was easy to interpret. The results from this study not only present a high quality tool for Alzheimer disease prediction but also explain which factors in which combinations are associated with development of the disease. These results legitimate our approach as a very powerful tool for biological knowledge discovery. The traditional methods give also a very strong classification but they do not give an interpretable model. Rough sets allowed to interpret the classifier and to conclude which attributes were important to predictive diagnosis. The domain specialists may now further investigate the most promising attributes and their combinations. Such conclusions would be hard to make using traditional statistical methods.

**Paper IV**

For Paper IV we investigated the role of STAT3 transcription factor in gliomas. Data from ChIP-seq, ChIP-chip and microarray expression data were analyzed. We compared distributions of gene expression changes before and after the treatment with INH2 and cumulative distributions of gene expression changes for genes with and without STAT3 binding sites in the promoter. We observed that the expression of some genes increased after treatment with INH2 suggesting the repressive role of STAT3 in glioma cells.

The next step of the analysis was looking for STAT3, H3K4me3 and H3ac gene expression regulatory patterns. We clustered the distances of STAT3 binding sites and histone modifications occurrences from the nearest transcription start site (TSS) and constructed a decision system where presence or absence of STAT3 binding site and histone modification occurrences were features and genes were the objects, while changes in expression constituted the outcome. Using RR and MCFS significant features were selected and ROSETTA was used to construct a rough set model. Although we did not obtain a good accuracy of the model we still could analyze the rules of the model and found some potentially interesting properties and patterns (unpublished). Our main findings show that STAT3 usually regulates a target gene if it is bound farther than 1500bp upstream the TSS. We also noticed a strong relation of the activating histone modification occurrence with changes in gene expression. H3K4me3 regulates a target gene if it occurs in the body of the gene (most likely in introns) and H3ac when it occurs in the target gene’s promoter, not farther that 1500bp upstream from TSS. The probability of changing the expression is highest when all three factors co-occur together. The last step of the analysis was to select the most interesting genes for the validation and perform biological experiments verifying our findings. Using R statistical environment we selected those genes from our list of potential STAT3 targets that co-occurred in Pubmed database with expressions ‘oncogene’ and
We chose seven genes for biological validation: Ank1, Lta, Txnip, Peli1, Trib3, Gadd45g and Usf1. The first four changed expression significantly, as was expected. The remaining three did not change the expression. This finding suggests that for the first group STAT3 is indeed the key regulator and in fact is a repressor of transcription. Until now STAT3 has been considered to be rather an activator of transcription than a repressor.
Future research

An obvious extension to this research is repeating the analysis from Paper IV on a larger number of histone modifications. It would be intriguing to know whether with the help of rough sets we could find combinatorial epigenetic patterns of gene regulation.

RR will be further developed in the direction of detecting interactions between the features in decision systems.

Another possible line of research that grows out of the presented work is the possibility of automatic discovery of enhancer regions in the genome. By employing the methods developed in this study as well as the ROSETTA system we want to identify new enhancers and try to identify factors or combinations of factors marking enhancers. We also want to investigate the rules in the obtained model and try to interpret the role of individual TFs and histone modifications in defining enhancer regions.
Concluding Remarks

The methods presented in this study are a response to the problems with finding peaks in ChIP-seq data and selecting significant features in large decision systems. PFMS is currently the best method for peak finding in ChIP-seq data. Random reducts appeared to be a significant new addition to rough set theory with potentially many applications. The analysis of the Alzheimer disease dataset confirmed the claims that feature selection and legible models are useful for biomedical researchers. The glioma study added a new and unknown role of STAT3 transcription factor.
Denna avhandling presenterar ett bidrag till maskin-inlärning som förbättrar vår förståelse av komplexa sjukdomar med fokus påhjärn-relaterade sjukdomar.


Random Reducts (RR) var en ny utveckling av Monte Carlo Feature Selection (MCFS) där besluts-träden ersatts med grova mängder (rough sets). RR visade en förbättring jämfört med MCFS i hastighet och förmåga att upptäcka skuggade attribut. Skuggade attribut definieras som attribut som innehåller fullt eller delvis samma information som det skuggande attributet med respekt till synbarhet.

Analysen av vårt STAT3 experiment identifierade 1200 gener vilkas promotorer har både STAT3 inbindnings platser och epigenetiska märken karaktäristiska för aktivt transkriberande gener. STAT3 och H3K4me ChIP datan

Acknowledgements

I have been very privileged to get to know and to collaborate with many great people during my Phd studies. I would like to express my deep appreciation and gratitude to my main supervisor, Jan Komorowski, for giving me the opportunity to start my career in science and for his confidence in my ability to work independently and realize my own ideas.

Furthermore, I am grateful to my co-supervisor Bożena Kamińska for her disinterested support and the patient guidance she provided to me. In a similar vein, I would like to recognize my co-supervisor Claes Wadelius for his instructive comments, thought provoking suggestions and interesting educational discussions every week.

I owe thanks to my co-workers and former members of Jan’s group for creating a great environment to work in and for the general collegiality they offered to me. In particular, I would like to thank Husen Umer, with whom I tended to burn the midnight oil while testing PFMS. I must mention Susanne Bornelöv, who always supported me with ROSETTA, Nicholas Baltzer, with whom I spent hours improving RR and who was so kind to translate the executive summary of this thesis into Swedish. Tallyho! Special thanks go to Michał Dąbrowski and Klev Diamanti for long hours of constructive discussions and for helping me find accommodation in Uppsala, which was one of the major challenges during the Phd studies. I am grateful to Piotr Przanowski and Jakub Mieczkowski for patiently explaining to me the secrets of cell biology and statistics, respectively.

I also want to mention my friends, who put up with me through the Phd process and who stood behind me along this fulfilling road. In particular Marzena Poruszek and Maciek Kubiać, who are always there for me, Wojtek Chaładaj, whose insane ideas have been an inspiration and an impulse to self-development and Konrad Lisowski, who has taught me how to make a fortune and afterwards work in science just for pleasure. ;-)

I am indebted to my parents, Józef and Barbara and my sisters, Kasia and Monika, who believe in me more than I do and who support all my undertakings as long as what I undertake is safe. Thank you for not supporting the remaining, usually crazy ideas! ;-)

36
I would be remiss if I didn’t acknowledge the support of Magda, my beloved wife, who is always by my side and who walked with me through all the doubts and moments of crisis during the Phd studies. Thank you for your encouragement and unwavering love. It is a great joy to know that you decided to take this journey with me and it is undeniably comforting to know that you are not going to disembark until the end of the cruise.

Finally, many thanks to The Foundation for Polish Science, which has provided me with financial backing from the sources of International PhD Projects Programme (MPD).
References


[28] Chang-Sik Son, Yoon-Nyun Kim, Hyung-Seop Kim, Hyoung-Seob Park, and Min-Soo Kim. Decision-making model for early diagnosis of congestive heart


A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology.