BREAST CANCER CLASSIFICATION *with* SUPPORT VECTOR MACHINES

Analysis of the classification performance of recursive feature elimination methods incorporating protein-complex information

Moritz Angermann

Supervisor: Florian Blöchl
Examiner: Prof. Dr. Fabian J. Theis

April 2010

Submitted in partial fulfillment of the requirements for the degree of Diploma in Mathematics

to the

Department of Biomathematics
Faculty of Mathematics
Technische Universität München

ABSTRACT

Microarrays are high through-put, high sensitivity tools to measure thousands of gene expressions simultaneously. However, experiments are still quite expensive and are known to produce relatively noise measurements. Hence, their high number of features for relatively few samples makes them a particular challenge for data analysis. We propose a data integration method to compute virtual protein-complex expressions from the microarray gene expressions. With this, the number of correlated features and the total size of the data set can be reduced. The study of multi-class problems is less frequent, but support vector machines with recursive feature elimination (SVM-RFE) offer a possible solution. We verify our approach by training a SVM-RFE on a large breast-cancer data set, where we determine the classification performance on a second, independent data set from another study. We have found that the classification performance did not improve on protein-complex expression, but nevertheless biologically interesting and interpretable protein-complexes have been extracted by our recursive feature elimination method. Our data integration method is therefore a new and interesting approach to interpret such data sets, and to gain new insights into the role and funcions of protein-complexes.

STATEMENT OF AUTHORSHIP

I pledge that this project report is the result of my own work. Material from other published or unpublished works of others, which is referred to in the project report, is credited to the author in the text. No other sources or unauthorized aid has been used.

CONTENTS

PREFACE

This is a project report in the field of bioinformatics. It has been conducted at the department of biomathematics at the Technische Universität München and the Helmholz Zentrum Munich, the German Research Center for Environmental Health.

The style of this report has been crafted by Eivind Uggedal. I have received written permission to use the style, and made minor modifications and adoptions to fit the needs of this report.

All images have been used with permission. The graphs have been created using PGF/TikZ by Till Tantau. Inspiration has been taken from the PGF/TikZ manual as well as the TEXample website (http://www.texample.net/). The plots have been created using the R package `ggplot2` by Hadley Wickham (http://had.co.nz/ggplot2/).

Finally, I would like to thank Prof. Dr. Fabian J. Theis for offering me the opportunity to do this project, Florian Blöchl for supervising it, Andreas Kowarsch for help with the enrichment analysis and Andreas Ruepp for the project idea and help with biology related questions. I also want to thank Michelle Lam, Ralf Sangl, Thomas Roche, Sa Wu, Dinh Thi Anh Thu and my mother for proofreading my report or parts of it, and the emotional support they gave to me throughout the whole time.

Moritz Angermann
Garching b. München, Germany
April 2010

INTRODUCTION

1

The analysis of gene expressions has been an important topic in Molecular Biology. A gene expression is the process by which a functional gene product (e.g. protein) is synthesized using information encoded in a gene. Microarrays are high through-put, high sensitivity tools that allow the measurement of thousands of genes simultaneously. The conduction of microarray experiments is expensive. Therefore, measurements often exist only for very few samples. This leads to data analysis problems in very high features spaces with relatively few samples, which are an interesting challenge and have been subject to multiple machine learning approaches.

The CORUM database contains manually annotated protein complexes from mammalian organisms and is maintained by the Munich information center for protein sequences (MIPS) at the Helmholz Zentrum Munich. Its annotations for protein-complexes include protein-complex function, localization, subunit composition as well as literature references and more.

We propose a new data integration method using the protein-complex information from the CORUM database to compute virtual protein-complex expressions. While the study of multi-class problems is less frequent, we believe that support vector machines with recursive feature elimination (SVM-RFE) in combination with our new data integration method offer a potent solution.

To verify our approach, we have trained a multi-class support vector machine with recursive feature elimination on a large breast-cancer data set and have determined its performance on a smaller breast-cancer data set from a different study. Even though we were unable to increase classification performance on the computed virtual protein-complex expressions over gene-expressions, our novel data integration method coupled with multi-class SVM-RFE has extracted biologically interesting and interpretable protein complexes. Therefore, our new approach can be seen as an interesting method to analyse and interpret data sets of this kind. Furthermore, this can lead to new insights into the role and function of protein-complexes by using virtual protein-complex expressions from gene expression microarray data sets.

BACKGROUND

The first chapter will be a short introduction to the relevant background. It will start with molecular biology, continued by Support Vector Machines (SVMs) and possible derived approaches to solve multi-class problems. It will end with a brief introduction to breast cancer.

2

2.1 THE CENTRAL DOGMA OF MOLECULAR BIOLOGY

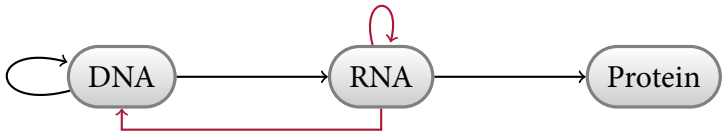


Figure 2.1: Diagram of the information flow between DNA, RNA and Protein according to the central dogma of molecular biology. Black arrows denote flow of the general group and red arrows denote flow of the special group.

In his article "Central dogma of molecular biology", Crick wrote 1970:

"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid",

and elaborates this idea through-out the article. The central dogma of molecular biology deals with the information transfer, which can be split into three groups: general, special and unknown. The general group, which occurs in all cells with minor exceptions, consists of the following three transports:

- deoxyribonucleic acid (DNA) → DNA,
- DNA → ribonucleic acid (RNA),
- RNA → Protein.

The special group does not occur in most cells, but has been shown to be present in virus infected cells with flows from: RNA → RNA and RNA → DNA. A third case for the special group: DNA → Protein flow has only been seen in cell-free systems containing neomycin. The other

information flows are postulated by the central dogma to never occur; these are Protein → Protein, Protein → DNA and Protein → RNA, and belong to the unknown group (Crick, 1970).

2.1.1 Genes

We divide all cells into two basic types: *prokaryotes* and *eukaryotes*. The defining structure that sets eukaryotic cells apart from prokaryotic cells is the nucleus. Prokaryotic cells have cell walls containing glycopeptides. Whereas eukaryotic organisms are made from cells that are organized by complex structures within membranes (Okafur, 2007, p.17).

Human beings are eukaryotes and contain billions of individual cells. Almost all of these cells contain, within each nucleus, the complete hereditary information for the organism in form of the genome. The genome consists of DNA and can be seen as the blueprint for all structures of the organism. The human genome is made from 23 pairs of chromosomes, where each pair is based on the chromosome pairs from their biological parents. The chromosomes contain chains of DNA, which consist of two polymers. These polymers are wrapped around each other and form a structure known as double helix. The polymer strands are held together by hydrogen bonds. They are large molecules of repeating monomers, which are called nucleotides. Each nucleotide is made from deoxyribose sugar, a phosphate group and one of the following four nitrogen bases: adenine, cytosine, guanine and thymine, usually represented by their first letter A, C, G and T. Due to a property of the nitrogen bases (called complementary base pairing) one can deduct from one strand of DNA the other complementary strand; in particular, adenine can only form hydrogen bonds with cytosine, and guanine with thymine. The sequence of these nucleotides in the double helix encode the hereditary genetic information. Genes are, as Pearson puts it in 2006:

"A locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions",

and can therefore be described by their ordered sequence of nitrogen bases. The length of these sequences can be hundreds of thousands of bases. The sequences encode particular patterns, of which the exact number in the human genome is unknown. The number of protein-coding genes is estimated between 20,000 and 25,000 genes. To have the hereditary information available within all cells, the DNA has to reside in each cell. Therefore, the DNA needs to be replicated to create new cells, which is essential to multi-cell organisms. The so-called DNA replication is the first of the general information flows described by the central dogma of molecular biology. To form a protein from a protein-coding gene, the gene information has to flow from the gene to the messenger RNA (mRNA) using a process called transcription, and from

the mRNA to the final protein through the so-called translation. These are the second and third information flows as postulated by the central dogma.

DNA replication

Because complete DNA is contained within each cell of the organism, it needs to be replicated upon cell division (cytokinesis). In eukaryotes, the replication is triggered at the end of the interphase. The interphase is followed by the separation of the chromosomes (mitosis), and the immediate cytokinesis. During the DNA replication process the double helix is split into its strands and each strand's complementary pair is synthesized using an enzyme called DNA-polymerase.

Transcription

Pearson (2006) noted that the genes consist of regions that are regulatory as well as regions that explicitly code for a protein. One of these regulatory regions is known as promoter, and is used by the RNA-polymerase that drives the RNA synthesis. The RNA synthesis is similar to the DNA synthesis, with the notable difference that only one strand is copied. Eukaryotic genes consist of exons and introns, which are DNA regions within a gene. The difference between exons and introns is that the final mRNA only represents the exons. During transcription, first precursor mRNA (pre-mRNA) is transcribed from the DNA strand[1]. A process called splicing later removes the introns from the pre-mRNA to yield the final one stranded mRNA, which only consists of exons. The mRNA leaves the nucleus via the nuclear membrane.

1. The reverse way is used in biotechnology as well as retroviruses, which include the HI / AIDS virus, where a single-stranded RNA is transcribed to a single-stranded DNA. (Okafur, 2007, p.36)

Translation

Outside of the nucleus the mRNA is used as a template for the synthesis of proteins, which is called translation. It contains the nucleotide uracil (U) instead of thymine. The translation is done by ribosomes, which are large complexes of proteins. These ribosomes read the genetic information carried by the mRNA molecules in triplets of nucleotides, and combine any of the 20 amino acids in the human body into complex polypeptide chains through chemical reactions. These triplets are called codons. The translation will start on the codon AUG and select phenylalanine if it reads the codon UUU, or glycine on GGG[2]; but if it reads the codon UAA, UAG or UGA the translation will be terminated (Okafur, 2007, p.38). The resulting polypeptide chains then form the protein.

2. These are examples, a complete list can be found in Okafur (2007, p.38, Table 3.1)

The process, in which the information from a gene is used to synthesize a functional gene product or protein via transcription and translation, is known as gene expression.

2.1.2 *Proteins*

Proteins are macromolecules; they form the building blocks of the organism and are responsible for numerous functions inside the living organism. Proteins are important for the metabolism, which is the maintaining of life in living things through chemical reactions. As hormones, proteins transport messages through the body. For example, the protein hemoglobin transports oxygen. Proteins are also the basis for enzymes, catalysts for chemical reactions. Proteins form 3-dimensional structures by folding their amino acid backbone. The protein's shape depends on the process it guides (Coen, 1999).

The linear chain of amino acids resulting from the translation phase is a so-called random-coil, a simple polypeptide. To form a proper protein, the random-coil needs to fold into a well-defined three-dimensional structure, which defines its characteristic and function. This process is driven by the interaction of the amino-acids from the random-coils during and after the protein synthesis. The correct fold is essential for proteins to function correctly, and failure results in inactive proteins or toxins.

Inside the human body, there are many different types of proteins. When several proteins with different polypeptide chains form a complex where each polypeptide chain contains different protein domains, the result can have multiple catalytic functions, and is called a *multi-protein complex* or *protein-complex*.

2.1.3 *Multiprotein Complexes*

Today, we know that the cell's dry mass consists mostly of proteins. These protein molecules form protein assemblies that carry out most of the major processes in the cell. These little machines of ten or more protein molecules perform complex biological functions where each assembly interacts with other protein-complexes (assemblies). These functions include cell cycle, protein degradation and protein folding (Alberts, 1998).

A similar definition is given by Ruepp et al. (2007, 2010) in *CORUM: the Comprehensive Resource of Mammalian Protein Complexes* with a slightly stronger emphasis on gene dependence:

"Protein complexes are key molecular entities that integrate multiple gene products to perform cellular functions."

The field of proteomics, the large-scale study of proteins, can be divided into cell-map and expression proteomics. The large-scale, quantitative study of protein-protein interactions through their isolation of protein complexes is called cell-map proteomics. It studies in particular the structure and function of the proteins contained within the protein-complexes. The study of protein expression changes is

called expression proteomics. The availability of complete sequences of the genome has shifted the focus towards functional interpretation of genomics (Blackstock and Weir, 1999). This has led to the creation of large scale databases containing protein-complexes, subunits and functional description. One of the largest, freely accessible databases is the CORUM database maintained as part of the Munich Information Center for Protein Sequences (MIPS), and is available at http://mips.helmholtz-muenchen.de/genre/proj/corum/. It is the comprehensive resource of mammalian protein-complexes, and contains mainly human (64%), mouse (16%) and rat (12%) protein-complexes, which have been experimentally verified. In 2007, the database contained more than 1,750 protein-complexes composed of 2,400 different genes, representing ~12% of the protein coding genes in humans. It has grown to more than 2,850 protein-complexes from 3,198 different genes by 2009, with the release of CORUM 2.0. It now represents ~16% of the protein coding genes in humans (Ruepp et al., 2007, 2010).

2.1.4 *Microarray Technology*

The conversion of information encoded in a gene into a functional gene product is the gene expression. It has become an important measurement in biology. A high through-put technology to measure gene expressions are microarrays. This high sensitivity tool can be used to measure the expression levels of thousands of genes; in some cases, even the expressions of the whole genome of an organism simultaneously (Stekel, 2003). They have become a very common method, and can measure the relative activity (gene expression) of multiple genes at once.

In their beginning, microarrays had been made using larger glass slides and nylon filters. Today, they consist of a $1 - 2cm^2$ solid surface, usually glass, silicon or plastic, which has been divided into many tiny spots. These spots are of the size of a few μm^2 arranged in an array-like structure, and contain labeled DNA molecules (probes). These probes consist of immobilized single stranded DNA (ssDNA) molecules that have been attached to the surface. Each probe contains ssDNA with the same nucleotide sequence, and represents a single gene. Probes contain either synthetic oligonucleotide, short nucleic acid polymers of 20 to 60 bases or complementary DNA (cDNA) that has been synthesized from mRNA using the enzyme: reverse transcriptase. Each spot targets a specific mRNA molecule, which is said to correspond to a certain gene. The idea behind this setup is that an mRNA molecule should hybridize to its complementary DNA, and form a strong mRNA-DNA bond. For later evaluation the ssDNAs have been labeled using fluorescent dye. These are Cyanine 3 (Cy3), a green-fluorescent dye, and Cyanine 5 (Cy5), a red-fluorescent dye. The level of fluorescence of the dye is later read by a laser-scanner. Stronger photon responses from the fluorescent dye suggest more hybridized mRNA, and therefore a higher expression level of the targeted gene.

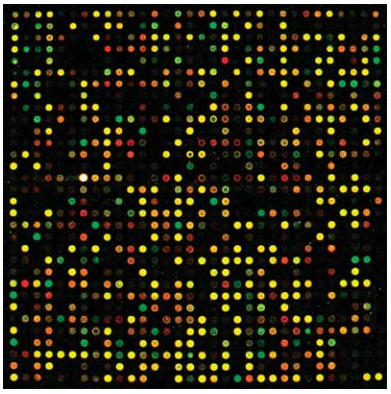


Figure 2.2: Cy®3- and Cy®5-labeled cDNA hybridized to a DNA microarray.
Retrieved March 10, 2010, used with permission from Promega Corporation (http://www.promega.com)

Microarrays are fabricated using two different methods: so-called spotted and in-situ microarrays.

Spotted microarrays

The first technology available to create microarrays were so-called spotted microarrays. At first, the probes are synthesized by amplifying cDNA using Polymerase Chain Reaction (PCR), and purifying the result or using pre-synthesized cDNA oligonucleotides. Afterwards, these probes are attached to the microarray using a spotting robot, which utilizes pins to transport the cDNA or oligonucleotide onto the array. After each step the pins are carefully washed to make sure that no contamination takes place (Stekel, 2003).

In-situ microarrays

With *in-situ* methods the ssDNAs synthesis takes place directly on the surface of the microarray, and is therefore fundamentally different from spotted microarrays. One approach used by Rosetta, Agilent and Oxford Gene Technology is to synthesize the probes by using a printer to print the ssDNAs with ink jet like nozzles. These have been modified to fire drops of A, C, G and T nucleotide bases instead of colour to produce the desired probes. Another approach is the use of lithographic technology similar to semiconductor manufacturing techniques. As a base the microarray is coated with silane (Si) molecules. A second photosensitive layer (called linker) is applied. Its purpose is to block the attachment of nucleotides. Using masks to selectively destroy the photosensitive layer allows to apply nucleotides very granular onto specific areas on the microarray chip. The repeated process of coating the chip with the linker, selective destruction of certain regions on the microarray and the application of nucleotides produces the desired probes. Occasionally, the nucleotides may not bind. To prevent the construction of ill-sequenced probes, a capping agent is applied after each coating with nucleotides. The capping agent seals the unprotected strands, onto which no nucleotide has bound, and consequently terminates all further binding of nucleotides onto this particular strand. The fabrication of the required masks is quite expensive, and can only be used for a specific set of microarrays. This is known as the Affymetrix technology. A similar method employed by Nimblegen and Febit uses digital micromirror arrays. These digital micromirror arrays can be used instead of the masks. The mirrors of the micromirror arrays are computer controlled. These are used instead of the masks to direct the ultra violet (UV) light to the appropriated parts of the chip (Stekel, 2003, p. 6).

The data-sets for the project described in chapter 3 (p. 23) stem from Affymetrix GeneChips.

The Affymetrix GeneChip

The GeneChip, as produced by Affymetrix, is an in-situ microarray. Its probes are grouped in so-called probe-sets, which are sets of ~20 different probes, which correspond to the same cDNA sequence. Cross hybridization describes the annealing of ssDNA to only partially complementary target ssDNA. This is an unwanted side effect, and can lead to distorted results. While it is common to add a repetitive DNA to the solution to reduce cross-hybridisation (Stekel, 2003, p. 13), Affymetrix reduces cross-hybridisation further by using only short polymers with 25 bases instead of the longer (> 1,000) ssDNAs used in spotted microarrays. Each probe of the probe-set then targets a different portion of the cDNA in question. For further accuracy each probe is synthesized in pairs of perfect match probes (PM) and miss match probes (MM). The so called perfect match probe (PM) is the complementary of the targeted polymer in question. The MM probe is identical to the PM probe, with the only difference that its middle base, the 13th, is swapped with the complementary base. This is meant to estimate the non-specific binding, when mRNA binds to the PM even though it's not targeted, and thus can be used to reduce the background noise of the measurement (Herold, 2007, p. 12). The Affymetrix technology allows a very dense placement of the probes, and creates rectangular regions. The use of light and the inherent interference results in leakage from one probe to its neighbours. This effect is compensated by the Affymetrix image-processing software, which uses only the inner portions of the probe's region to compute the expression (Stekel, 2003, p. 9).

Figure 2.3: An Affymetrix® GeneChip®
Retrieved March 10, 2010, from http://www.wikipedia.com

2.2 MACHINE LEARNING

The abundant amount of information resulting from microarray experiments requires algorithms that can deal with the high amount of data. The mathematical field of machine learning studies the classification or recognition of patterns in vast amounts of data.

Mathura and Kangueane (2009) describe this in chapter 3 as:

"For instance, cancer type classification based on microarray expression or determining whether a protein binds to DNA or not based on sequence and structural motifs are good examples of classification problems."

Classification is a subtopic of the supervised learning category as outlined in chapter 1 by Hastie et al. (2009):

"The learning problems that we consider can be roughly categorized as either supervised *or* unsupervised. *In supervised learning, the goal is to predict the value of an outcome measure based on a number of input measures; in unsupervised learning, there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures."*

Different methods have been applied to classification tasks in bioinformatics. Widely used are Decision Trees, Support Vector Machines, Bayesian Classifiers, Neural Networks and many more (Mathura and Kangueane, 2009, ch. 3). The following will introduce Support Vector Machines, which will be used extensively throughout the following chapters.

2.2.1 *Support Vector Machines*

Support Vector Machines (SVMs) are large margin classifiers. The margin can be understood as the distance of the example to the separation boundary. The large margin classifier generates a decision boundary with a large margin to almost all training examples. SVMs have been introduced by Boser et al. (1992), and can be used for binary classification problems. SVMs are no machines in the classical sense, and do not consist of any tangible parts. SVMs are merely mathematical algorithms for pattern-matching. To understand the mathematical model, a few definitions are needed.

Notation 1 (Scalar Product). Let V^n be an n-dimensional vector space, for $x, y \in V^n$ the scalar product will be denoted by:

$$\langle x, y \rangle := \sum_i x_i \cdot y_i. \tag{2.1}$$

Notation 2 (Euclidean Norm). Let $x \in \mathbb{R}^n$, the euclidean norm will be

written as follows:

$$||x|| := \sqrt{\langle x, x \rangle}. \tag{2.2}$$

A hyperplane is an $(n-1)$-dimensional object in the same sense that in a three dimensional space a two dimensional object can be seen as a plane. The formal definition follows:

Definition 2.2.1 (Hyperplane)**.** A hyperplane $h \subset \mathbb{R}^n$ can be explicitly defined by its perpendicular vector $w \in \mathbb{R}^n$ and its distance $b \in \mathbb{R}$ from the origin:

$$h := \{x \in \mathbb{R}^n : \langle x, w \rangle = b\}. \tag{2.3}$$

If $b = 0$ the hyperplane is said to be unbiased. The distance of the hyperplane to the origin is $\frac{b}{||w||}$ units.

Definition 2.2.2 (Linearly Separable)**.** A set of instance-label pairs

$$S := \{(x_i, y_i) : x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}, i = 1, \ldots, l\} \subset \mathbb{R}^n \times \{-1, 1\} \tag{2.4}$$

is said to be linearly separable, if there exist $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ for a hyperplane

$$h = \{x \in \mathbb{R}^n : \langle x, w \rangle = b\}, \tag{2.5}$$

such that

$$\forall (x_i, y_i) \in S : y_i(\langle x_i, w \rangle - b) > 0. \tag{2.6}$$

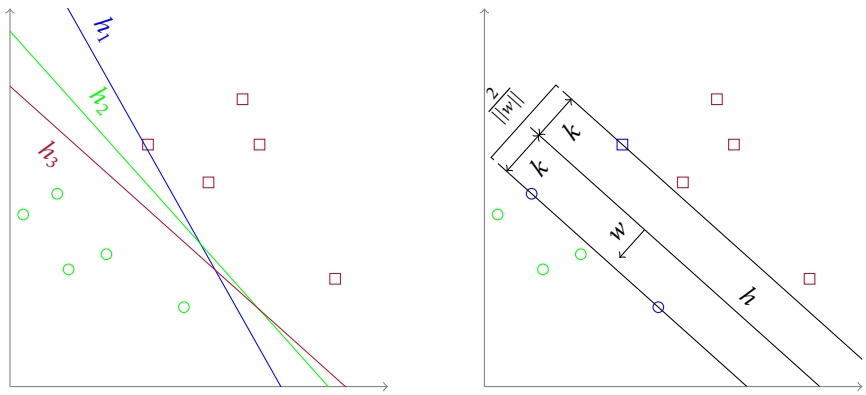
The binary classification problem can therefore be written as the following:

Definition 2.2.3 (Binary Classification Problem)**.** Assume a set of instance-label pairs

$$S = \{(x_i, y_i) : x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}, i = 1, \ldots, l\}. \tag{2.7}$$

The objective is to find a prediction function $f : \mathbb{R}^n \mapsto \{-1, 1\}$, which satisfies

$$\forall (x_i, y_i) \in S : f(x_i) = y_i. \tag{2.8}$$



(a) multiple separating hyperplanes (b) optimal hyperplane with margin

Figure 2.4: Hyperplanes in a two-dimensional space separating a linearly separable set of instance-label pairs.

Given a linearly separable binary classification problem, a hyperplane $h = \{x \in \mathbb{R}^n : \langle x, w\rangle = b\}$ exists that clearly separates the set $\{(x_i, y_i) \in S : y_i = -1\}$ from $\{(x_i, y_i) \in S : y_i = 1\}$. The prediction function $f : \mathbb{R}^n \mapsto \{-1, 1\}$ can then be derived from the hyperplane:

$$f(x) = \text{sign}(\langle x, w\rangle - b). \tag{2.9}$$

The hyperplane is not unique as figure 2.4a shows; but only one hyperplane has the broadest margin. The margin describes the shortest distance from any point in S to h as depicted in figure 2.4b. This leads to the formulation of the Support Vector Machine:

Definition 2.2.4 (Support Vector Machine)**.** The support vector machine finds the hyperplane h with the broadest margin for a linearly separable binary classification problem as the solution to

$$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{subject to} \quad \forall (x_i, y_i) \in S : y_i(\langle x_i, w\rangle - b) \geq 1. \tag{2.10}$$

This optimization problem is known as the primal form of the SVM. It can also be solved in its dual form.

Theorem 2.2.1 (Dual Form of the Support Vector Machine)**.** *The support vector machine can also be solved in its dual form, and the optimal solution of the primal and dual form coincide.*

Proof. Using the generalized Lagrangian, the primal form can be written as:

$$\min_{w,b} \max_{\alpha} \underbrace{\frac{1}{2}||w||^2 - \sum_i \alpha_i \left(y_i(\langle x_i, w\rangle - b) - 1\right)}_{=:L(w,b,\alpha)} \quad \text{subject to} \quad 0 \leq \alpha. \tag{2.11}$$

For which the dual form is defined as:

$$\max_{\alpha} \min_{w,b} \frac{1}{2}||w||^2 - \sum_i \alpha_i \left(y_i(\langle x_i, w\rangle - b) - 1\right) \quad \text{subject to} \quad 0 \leq \alpha. \tag{2.12}$$

Computing the partial derivatives for the primal variables, and equating them to 0 gives:

$$\frac{\partial}{\partial w} L(w, b, \alpha) = w - \sum_i \alpha_i y_i x_i \stackrel{!}{=} 0 \Rightarrow w = \sum_i \alpha_i y_i x_i \tag{2.13}$$

$$\frac{\partial}{\partial b} L(w, b, \alpha) = \langle \alpha, y\rangle \stackrel{!}{=} 0 \tag{2.14}$$

Substituting equations 2.13 and 2.14 in 2.12 leads to the dual form:

$$\max_{\alpha} \sum_i \left(\alpha_i - \frac{1}{2} \sum_j \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j \right\rangle\right) \quad \text{subject to} \quad 0 \leq \alpha \wedge \langle \alpha, y\rangle = 0. \tag{2.15}$$

Because the optimization problem is convex and Slater's condition is satisfied, strong duality holds:

$$\max_{0\leq\alpha}\min_{w,b} L(w,b,\alpha) = \min_{w,b}\max_{0\leq\alpha} L(w,b,\alpha). \tag{2.16}$$

This implies that the optimal solution from the primal and dual form coincide. From the optimal solution a^* of the dual form, the primal variables w^* and b^* can be computed. From the Karush-Kuhn-Tucker (KKT) conditions, the complementary slackness:

$$\forall i : \alpha_i^*(1 - y_i(\langle x_i, w^*\rangle - b^*)) = 0 \tag{2.17}$$

is used to compute b^*. It implies that $\forall i : \alpha_i^* > 0$ the point x_i lies on the margin: $y_i(\langle x_i, w^*\rangle = 1)$. This can also be seen in figure 2.5b.

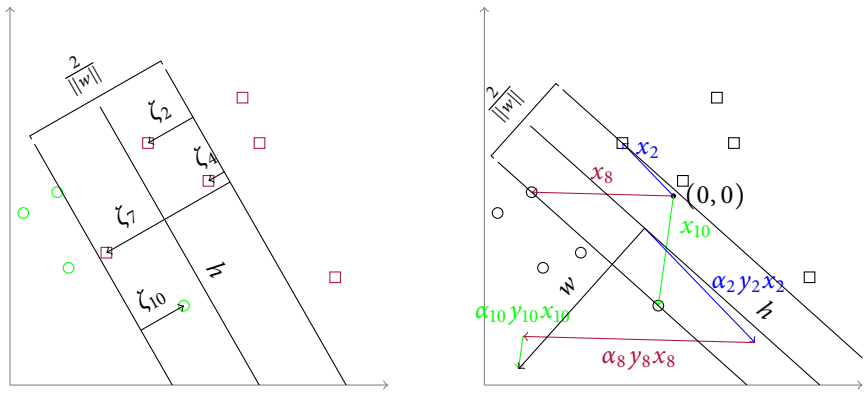
$$w^* = \sum_i \alpha_i^* y_i x_i \tag{2.18}$$

$$b^* = \frac{1}{2}\left(\max_{i:y_i=-1\ \wedge\ \alpha_i^*>0} \langle w^*, x_i\rangle + \min_{i:y_i=1\ \wedge\ \alpha_i^*>0} \langle w^*, x_i\rangle\right) \tag{2.19}$$

□

In order to remove the restriction on linearly separable classification problems, a penalty parameter[3] $C \in \mathbb{R}_+$ as well as slack variables $\zeta_i \in \mathbb{R}_+$ for each point in S are introduced. ζ will denote the vector $(\zeta_1, \zeta_2, \ldots)^t$. The slack variables are used to soften the constraints on the optimization problem while the penalty parameter allows to adjust the impact of the slack variables on the objective function.

3. Note on notation: if a variable $x \in \mathbb{R}$ is used where a vector is expected, x implicitly represents the vector $(x, x, \ldots, x)^t$ of the required dimension.



(a) penalized points inside the margin

(b) construction of the perpendicular vector w from its dual representation ($w = \sum_i \alpha_i y_i x_i$)

Figure 2.5: Hyperplanes in a two-dimensional space with soft-margins.

Definition 2.2.5 (Support Vector Machine with soft margins)**.** The support vector machine with soft margins finds the hyperplane h with the

broadest margin by solving

$$\min_{w,b,\zeta} \frac{1}{2}||w||^2 + \langle C, \zeta \rangle \quad \text{s.t.} \quad \forall (x_i, y_i) \in S : y_i(\langle x_i, w \rangle - b) \geq 1 - \zeta_i \wedge 0 \leq \zeta. \tag{2.20}$$

This allows a certain set of points to lie within the margin, or on the wrong side of the hyperplane.

At this point, a hyperplane can be found for arbitrary binary classification problems using the support vector machine with soft margins. Often, algorithms solve the dual form of the optimization problem. The results for the dual form can be obtained analogous to the dual form for the support vector machine without soft margins (cf. theorem 2.2.1). The dual form is given below.

Definition 2.2.6 (Dual form of the Support Vector Machine with soft margins)**.**

$$\max_{\alpha} \sum_i \left(\alpha_i - \frac{1}{2} \sum_j \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j \right\rangle \right) \text{ subject to } 0 \leq \alpha \leq C \wedge \langle \alpha, y \rangle = 0. \tag{2.21}$$

A notable feature of the dual form of the SVM with soft margins is that the slack variables have been eliminated from the optimization problem. Its only difference to the dual form for the SVM without soft margins is the additional constraint $\alpha \leq C$.

The prediction function f can therefore be written as

$$f(x) = \text{sign}(\langle x, w \rangle - b) = \text{sign}\left(\sum_i \alpha_i y_i \langle x_i, x \rangle - b \right)$$

utilizing the primal or dual form respectively. For a more detailed introduction to support vector machines and their extensions to the non-linear separable case, see e.g. Schölkopf and Smola (2002).

2.2.2 *Multi-class Support Vector Machines*

Multi-class support vector machines (sometimes abbreviated as MSVMs) aim to provide extensions for multi-categorical data to the standard SVM, which solves only the binrary case. These approaches can be grouped into three groups. The first group tries to solve the multi-class problem by extension of the SVM model, while the other two use the binary SVM for classification, partitioning the input data differently and using different heuristics for the class prediction.

To demonstrate the different approaches, toy-data have been generated. The three posed sample problems are visualized in figure 2.6. All tests have been performed using the default settings. The sample data has been rescaled according to the suggested ranges for each software package.

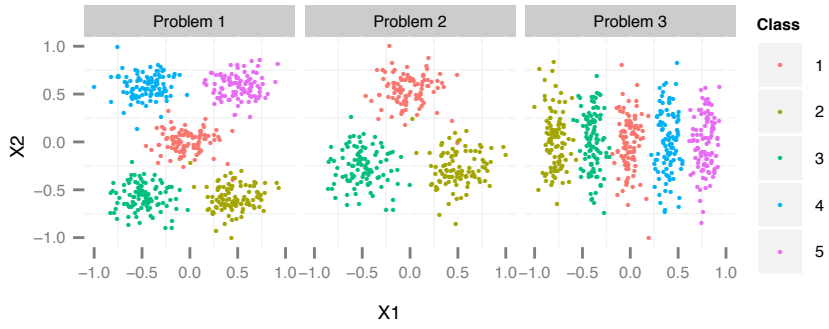


Figure 2.6: Three sample problems using two-dimensional gaussians with different centroid arrangements. The middle problem is a portion of the first. It contains only the lower three classes. All sample problems have been scaled to $[-1, 1]^2$.

All-In-One Multi-Class SVM

Direct extension of the SVM to tackle multi-class problems can lead to the addition of further constraints for each class onto the optimization problem. Their advantage is their ability to take inter-class correlation into account. This is lost with approaches relying on binary SVMs because they assume independence of the binary problems. An approach chosen by Crammer and Singer (2001) starts from a generalized notion of separating hyperplanes. They solve the compact quadratic optimization problem from the dual form of the problem using a fixed point algorithm.

An implementation of the multi-class SVM proposed by Crammer and Singer (2001) using a different algorithm to solve the inherent optimization problem can be found in the software package SVM^{multiclass} by Thorsten Joachims and is available from http://svmlight.joachims.org/svm_multiclass.html for free for non-commercial use.

Figure 2.7 indicates that SVM^{multiclass} will find decision boundaries through the origin, and therefore perform poorly on problems 1 and 3. The classification details for the prediction of the training data can be

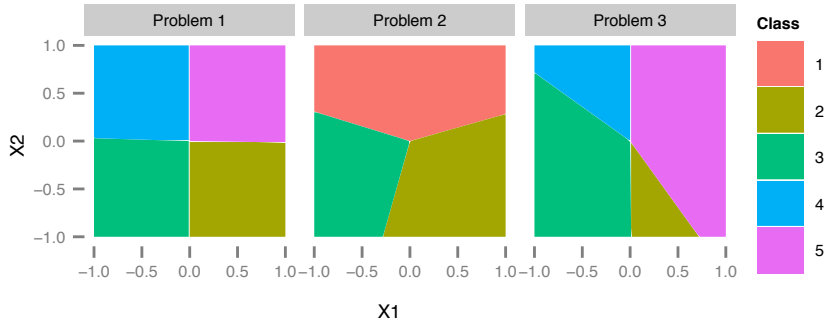


Figure 2.7: Empirical decision boundaries for the three sample problems using the SVM^{multiclass} software by Thorsten Joachims.

found in table 2.1. For problem 1, the model has virtually ignored class 1 in favour of the other classes. With problem 2, the SVM^{multiclass} software has found a very good model resulting in a training-performance of ~98% correct classification. By far the worst training-performance was achieved for problem 3.

		Predicted Class										
		Problem 1				Problem 2			Problem 3			
		2	3	4	5	1	2	3	2	3	4	5
Actual Class	1	15	33	28	24	99	1	0	14	35	25	26
	2	100	0	0	0	1	99	0	0	94	6	0
	3	0	100	0	0	1	2	97	0	80	20	0
	4	0	0	100	0				5	0	0	95
	5	0	0	0	100				0	0	0	100
Perf.		400/500 (80%)				295/300 (~98%)			180/500 (36%)			

Table 2.1: Class-prediction and performance for the three sample problems using the SVM^{multiclass} software by Thorsten Joachims.

All-Against-One (AAO) SVM

Other approaches decompose the multi-class training data into multiple binary problems. The All-Against-One (AAO) strategy constructs k different sets, where k is the number of different classes in the multi-class training data. Each set consists of all the data-points from the training data, but the points have been reassigned to different classes to form a binary classification problem. For each class $j : 1 \leq j \leq k$, the class labels of the corresponding set have been changed to form a binary classification problem. All data-points in the set that correspond to the class j are assigned the class-label 1; all other the class-label −1. This has been implemented by the author on top of a binary SVM. The SVM

implementation chosen was the LSVM package by Mangasarian and Musicant, which implements a Lagrangian approach in a very succinct manner as detailed in their paper *LSVM Software: Active Set Support Vector Machine Classification Software* (Mangasarian and Musicant, 2000a). The LSVM package is available as a MATLAB file from the authors' website http://www.cs.wisc.edu/dmi/lsvm/. A free license is granted for academic and research purposes.

The resulting k models with their hyperplanes bear the question of the heuristic to be used in the prediction function. Two heuristics will be presented.

The first heuristic will apply the point in question to all k trained models, and chooses the class, for which the point lies on the positive side. Two obvious issues arise: the data-point may lie on the positive side of two or more hyperplanes or possibly none at all. These points will be assigned the classes −1 and 0 respectively. The decision boundaries found using this heuristic are depicted in figure 2.8. Table 2.2 shows the prediction details and classification performance on the test set for the strict heuristic.

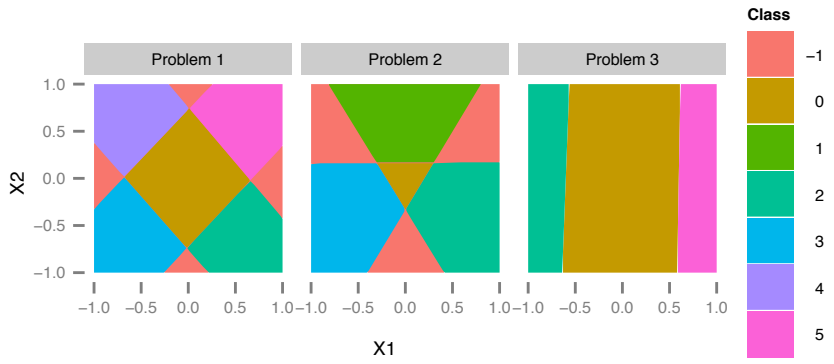


Figure 2.8: Empirical decision boundaries for the three sample problems using the LSVM algorithm and a strict classification heuristic.

		Predicted Class												
		Problem 1					Problem 2					Problem 3		
		0	2	3	4	5	-1	0	1	2	3	0	2	5
Actual Class	1	99	1	0	0	0	1	0	98	1	0	100	0	0
	2	4	96	0	0	0	0	1	1	98	0	2	98	0
	3	7	0	93	0	0	3	1	0	96	99	1	0	
	4	6	0	0	94	0						100	0	0
	5	1	0	0	0	99						0	0	100
Perf.		382/500 (~76%)					292/300 (~97%)					198/500 (~40%)		

Table 2.2: Class-prediction and performance for the three sample problems using the LSVM by Mangasarian and Musicant with a strict classification heuristic.

The second heuristic aims to eliminate the regions where the first heuristic would assign the classes −1 and 0. To achieve this, the hyperplanes from the models are not interpreted as strict boundaries. A data-point is assigned to the class, for which the corresponding model's objective function yields the largest value. When multiple objective functions return the same maximum, the class with the lowest index is chosen. The empirical decision boundaries are visualized in figure 2.9 with the detailed classification information in table 2.3.

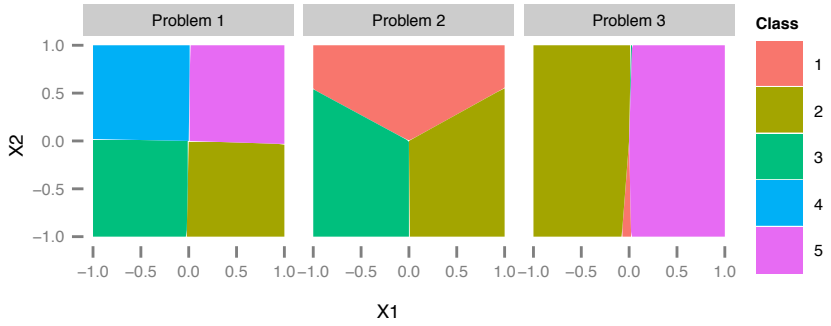


Figure 2.9: Empirical decision boundaries for the three sample problems using the LSVM algorithm to train the models and a continuous prediction heuristic

		Predicted Class										
		Problem 1				Problem 2			Problem 3			
		2	3	4	5	1	2	3	1	2	3	5
Actual Class	1	15	33	28	24	98	2	0	5	55	1	39
	2	100	0	0	0	1	99	0	0	100	0	0
	3	0	100	0	0	0	0	100	0	100	0	0
	4	0	0	100	0				0	0	0	100
	5	0	0	0	100				0	0	0	100
Perf.		400/500 (80%)				297/300 (99%)			205/500 (41%)			

Table 2.3: Prediction for the three sample problems using the LSVM software by Mangasarian and Musicant and a continuous prediction heuristic.

Comparing table 2.2 and table 2.3 as well as figure 2.8 and figure 2.9, one can see that both heuristics result in very different classifications. Notably, problem 3 shows very different classification results for identical models. The performance on this training set is similar for both heuristics, but the continuous heuristic outperforms the strict heuristic for the simple training set.

Pairwise Voting SVM

The third approach draws all possible pairs from the original multi-class classification problem. This results in $\frac{k \cdot (k-1)}{2}$ binary classification sub-problems for a k-class classification problem. For each sub-problem, a model is trained using a binary SVM. The trained models' hyperplanes therefore separate only two classes as opposed to the one-against-all strategy. Consequently, the objective functions of the models predict the class based on the two classes used during training. These predictions are understood as votes. For each data-point in question, the votes from all models are counted, and the class with most votes wins. This approach has been implemented for the multi-class classification support in the LIBSVM software package. LIBSVM is a library for SVMs developed by Chang and Lin and available from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. A problem arising from the voting comes with the fact that two or more classes may have obtained the same number of votes. In this case, the LIBSVM prediction heuristic will choose the class with the lowest index.

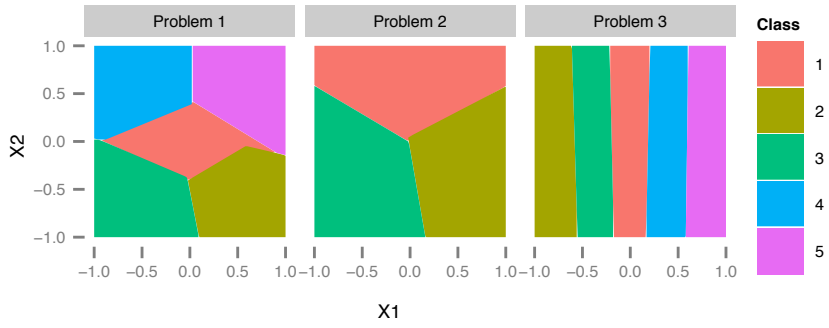


Figure 2.10: Decision boundaries using the pairwise-voting approach chosen by LIBSVM from Chang and Lin.

		Predicted Class												
		Problem 1					Problem 2			Problem 3				
		1	2	3	4	5	1	2	3	1	2	3	4	5
Actual Class	1	97	2	0	1	0	98	2	0	99	0	0	1	0
	2	1	99	0	0	0	1	99	0	0	99	1	0	0
	3	0	0	100	0	0	0	0	100	0	1	99	0	0
	4	1	0	0	99	0				0	0	0	100	0
	5	0	0	0	0	100				0	0	0	0	100
Perf.		495/500 (99%)					297/300 (99%)			497/500 (~99%)				

Table 2.4: Class-prediction and performance for the three sample problems using the LIBSVM software package by Chang and Lin.

Obviously, the pairwise voting methodology computes more models, if the number of classes exceeds three. The pairwise strategy permits to keep linear separable classes linear separable, which can be seen in the very good training-performance for problem 3 in figure 2.10 compared to the poor performance of the one-against-all heuristics in figure 2.8 and 2.9. The one-against-all heuristics have virtually no option to find a separating hyperplane for any of the three inner classes of problem 3.

Figure 2.7, figure 2.9 and figure 2.10 indicate that for problem 2, the LIBSVM's pairwise voting heuristic yields a very similar classification to the classification from the SVM^{multiclass} and continuous classification heuristic on top of the LSVM. The continuous classification heuristic and the LIBSVM even perform identical on problem 2's training set as table 2.3 and table 2.4 show.

2.3 BREAST CANCER

Cancer (malignant neoplasm) is in most cases a malicious form of invasive, uncontrolled cell growth, which in some cases may develop metastases[4]. Cancer can form tumors, though this is not necessarily the case as blood cancer (leukemia) shows. At any age, one is exposed to develop cancer; the risk increases with age, exposure to carcinogens[5] and genetic inheritance.

Cancer can be seen as a result of a defective cell as it stems from abnormalities in the genetic material, which can be caused by erroneous replication of DNA. Often, cancer-promoting oncogenes are highly expressed genes resulting in tumor cells, programmed cell death prevention and hyperactive growth. Also, reduced expression of tumor suppressor-genes can result in abnormal cell functions like inaccurate DNA replication and loss of cell cycle control.

The diagnosis of cancer is often performed using the biopsy[6] of suspected tissue and the histologic examination[7].

Breast cancer (breast carcinoma) is one of the most frequent malicious tumors of milk ducts, and has occasionally been found to metastasise to bone, liver, lung and brain. It is dominant in women, where it is found a hundred times more often than in men. It develops sporadically, and is the most likely cancer in the western hemisphere for women to develop; being the highest cause of death for women between 30 and 60 years of age, with a lethality rate of approximately 30%.

The common breast cancer treatment consist of a combination of surgical removal of the affected tissue, hormone and chemotherapy as well as radioactive therapy. Some subtypes of breast cancer depend on the oestrogen and progesterone hormones to grow. Drugs which block the production of the aforementioned hormones exist and are used after surgical removal of the cancer tissue to prevent a second onset and further spreading of the disease. These drugs are likely to result in death of the ovaries[8]and subsequent infertility.

The risk to develop breast cancer due to genetic inheritance is only at 5%, though carriers of the breast cancer susceptibility gen mutations BRCA1 or BRCA2 have a very high (five-fold increase) risk to develop breast or ovarian cancer during their lifetime. While abortion, contrary to the abortion-breast cancer hypothesis, does not increase the risk, the lack of childbearing or breastfeeding as well as high hormone levels are primal risk factors.

Early diagnosis of breast cancer can be done using screening methods, for example mammography, where the breast is x-rayed to detect irregular spots. The final diagnosis, given a suspected diagnosis, is often done using a needle or surgical biopsy.

There are multiple schemata to classify breast cancer. These schemata differ in their criteria as they target different purposes and are determined by different examiners. The most common schema is the TNM classification, which uses different stages for the breast cancer classifi-

4. Metastasis is the spread of malicious cells from one part of an organ to another part of the same (or completely different) organ through lymph or blood vessels.

5. Carcinogens are agents, which are directly involved in the increase, propagation or exacerbation of cancer. These include tobacco smoke, radiation, chemicals and diseases.

6. Biopsy is a medical test including the removal and analysis of cells and tissue; often performed with a needle, but larger tissue lumps may be surgically removed.

7. A histologic examination is the analysis of a thin slice of cells or tissue using a light- or electron microscope to gain insight into the microscopic structure of the retrieved specimen.

8. Ovaries are primary genitalia of females, where the ovocytes are produced. Ovaries are the source for the secretion of the hormones oestrogen and progesterone, and consequently play an important role in the hormone balance. Oestrogen is the hormone, which is responsible for the development of the secondary sex characteristics and reproductive organs as well as the maintenance of their mature functional state. Progesterone controls the cyclic change and the health of the endometrium.

cation. TNM has five major categories for the tumor size (*Tis*, *T1*, *T2*, *T3* and *T4*), four categories to classify lymph nodes (*N1* to *N4*) and a binary category for the existence of metastases outside of the breast and lymph nodes. The binary category describes the absence or presence of metastases with *Mo* and *M1* respectively.

Another rather simple schema uses the histological appearance of the examined tissue. The classification uses the Bloom-Richardson grade system. *Well-differentiated* in this context means that the tissue in question looks like ordinary tissue, *moderately-differentiated* means that it neither looks ordinary nor extremely disorganized. The final *poorly-differentiated* grade corresponds to tissue that looks nothing like ordinary tissue. This schema can be extended into the pathology schema, which uses in addition to the appearance additional criteria and leads to a very rich list of different breast cancer types.

A final schema to be presented uses gene and protein expressions to describe the breast cancer cells. As some breast cancer types rely on oestrogen and progesterone to grow, one can test for *oestrogen receptors (ER)* and *progesterone receptors (PR)*. Additionally, the *protein HER2/neu* (also known as *ERBB2*) is also used for classification. If none of these three is expressed, the cancer is said to be *triple negative* breast cancer, which is a subgroup of the so called *basal-type* breast cancers and are more aggressive and less responsive to standard treatment, and therefore leaving the patient with a poorer prognosis.

Sørlie et al. (2001) have found that they could make out six breast cancer subtypes using a hierarchical clustering approach. They have found gene expression patterns for the following subtypes: *Basal-Like*, *ERBB2+*, *Normal (Breast-) like*, *Luminal A*, *Luminal B* and *Luminal C*. From their hierarchical clustering they identified the tumor subtypes as follows: with low to absent gene expressions of ER and several additional transcriptional factors, the first three subtypes were categorized. The *basal-like* subtype had high expressions of keratins 5 and 17 as well as laminin and fatty acid binding protein 7. The *ERBB2+* subtype had among others high expressions of ERBB2 and GRB7 genes. The highest expression of many genes that are known to be expressed by adipose tissue as well as nonepithelial cell types was found in the *normal (breast-) like* subtype.

The specimen with high expressions of ER and several transcriptional factors were clustered into the subtypes *Luminal A*, *Luminal B* and *Luminal C*. The highest expressions of the ER α gene, GATA binding protein 3, X-box binding protein 1, trefoil factor 3, hepatocyte nuclear factor 3 α and estrogen-regulated LIV-1 were found in the subtype *Luminal A*. *Luminal B* and *Luminal C* were only marginally different and showed low to moderate expressions of luminal-specific genes including the ER cluster. A high expression of a novel set of genes with an unknown function, differentiated subtype *Luminal C* from *Luminal A* and *Luminal B* (Sørlie et al., 2001).

THE PROJECT

3

After the first part laid the very basis, this part will describe the conducted project in detail. At first, the data-sets will be introduced, followed by the results obtained during the conduction of the correlation analysis and recursive feature extraction.

3.1 DATA SETS

This project is based on the microarray breast-cancer data-sets of frozen breast-cancer tumors from Sweden. The data stems from two different cohorts. The data-sets can be obtained in the form of `.Rdata` files from the website of Prof. Yudi Pawitan at http://www.mep.ki.se/~yudpaw/. Furthermore, the author has been supplied with `.mat` files containing Gene-Symbol to protein mappings and a snapshot of the CORUM database's protein to protein-complex mappings.

3.1.1 Breast Cancer Data

The breast-cancer data is composed of two cohorts (Stockholm and Uppsala) with a total of 412 quality controlled RNA expressions. These expressions have been obtained using Affymetrix GeneChips. The Stockholm cohort consists of the expressions of tumor-cells from 159 patients. Their tumor-cell samples have been collected by the Karolinska Hospital in Stockholm, Sweden, from January 1st 1994 to December 31st 1996 from patients who have been operated for primary breast-cancer.

The Uppsala cohort consists of the expression values of tumor-cells from 253 patients. The tumor-cells have been collected by the Uppsala University Hospital, Uppsala, Sweden from January 1st 1987 to December 31st 1989.

Both cohorts have been subject to previous research. For a detailed account see Pawitan et al. (2005), Miller et al. (2005), Calza et al. (2006).

	Uppsala	Stockholm
Probe-Set IDs	44928	44928
Experiments	253	159
Unique PIDs	44760	44760

Table 3.1: Details of the cohorts size.

The Stockholm and Uppsala cohort contain expression values for 44,928 Probe-Set Identifiers (Probe-Set IDs or PIDs), of which only 44,760 Probe-Set IDs are unique as table 3.1 shows. Both cohorts' samples came annotated with a breast-cancer subtype schema as found and described by Sørlie et al. (2001). Figure 3.2 shows the different distribution of the subtypes within the cohorts.

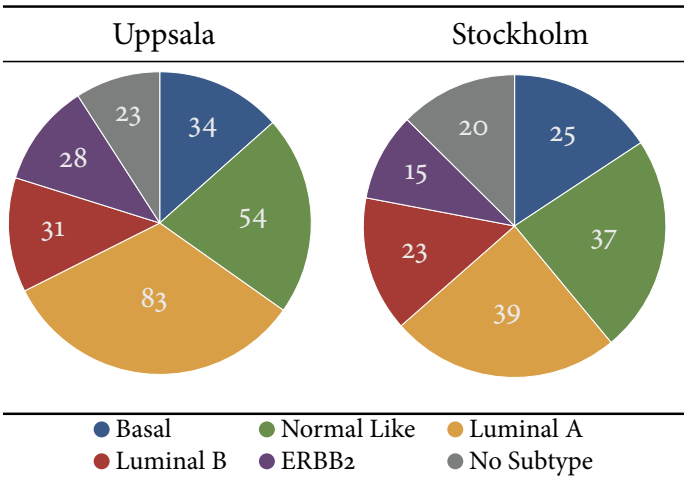


Table 3.2: Breast-cancer subtype distribution within the Stockholm and Uppsala cohort.

3.1.2 *Affymetrix GeneChip Array Information*

The array information for the Affymetrix GeneChips is provided by Affymetrix and contains additional information for the probe-sets on the GeneChip. They include multiple fields, of which three have been used: Probe-Set ID, Gene-Symbol and Swiss-Prot.

The Probe-Set ID is the identifier for the probe set, for which the expression has been measured. For example, these look like `200007_at`, `200011_s_at` or `200012_x_at`.

The Gene-Symbol can come from different organizations, which assign them to different species. Some of the possible databases, from which the information has been sourced include: The Human Genome Organization (HUGO), The Rat Genome Database (RGD), Mouse Genome Database Project (MGD) at the Mouse Genome Informatics (MGI) as well as the SubtiList, which analyses the genome of Bacillus subtilis (Affymetrix, 2010).

The Swiss-Prot entry contains the Swiss-Prot accession number. Examples are `P11474`, `Q8N4S8`, `Q96F89` or `Q96I02`. Swiss-Prot is a protein knowledge-base of protein sequences and part of the UniProt (http://www.uniprot.org) Knowledgebase (UniProtKB).

The number of entries available in the data-set that has been used, has been summarized in table 3.3.

Type	#Entries
Probeset IDs	54675
Gene Symbols	21373
Unique Gene Symbols	21372
Swiss Protein	44713

Table 3.3: Size of the Affymetrix GeneChip Array Information. Here unique means invariance under capitalization.

3.1.3 Additional Gene-Symbol To Protein Mappings

The additional Gene-Symbol to protein synonyms mapping contained protein mappings for 5,556 Gene-Symbols of which 4,940 were invariant under capitalization. The statistic can be found in table 3.4.

Type	#Entries
Gene-Symbols	5556
Unique Gene-Symbols	4940
Proteins	7152

Table 3.4: Size of mappable elements in the additional Gene-Symbol to Protein Mapping.

3.1.4 CORUM Protein To Protein-Complex Mappings

The CORUM database as it has been mentioned in chapter 2.1.3 (p. 6) is a database of mammalian protein-complexes from mainly human (64%), mouse (16%) and rat (12%) that have been experimentally verified. The database is freely accessible, and maintained as part of the Munich Information Center for Protein Sequences (MIPS) at the Helmholz Zentrum Munich, the German Research Center for Environmental Health.

The snapshot that has been used for this project containes proteins associated with protein-complex IDs and protein-complex names. Out of all the different protein-complex names 35 were identical after capitalization. The numbers of entries for each type have been summarized in table 3.5.

Type	#Entries
Proteins	3642
Protein-Complex IDs	2104
Protein-Complex Names	1880
Unique Protein-Complex Names	1845

Table 3.5: Size of mappable elements in the snapshot of the CORUM database that was used.

3.2 CORRELATION ANALYSIS

In the first part of this project the question was whether or not the superimposition of the structural information provided by the CORUM database would make a significant difference. As an indicator, the mean correlation coefficient has been selected.

A statistical test was performed. The null-hypothesis chosen is as follows: the mean correlation coefficient of expressions of Probe-Set IDs is the same as the mean correlation coefficient of expressions of Probe-Set IDs within protein-complexes.

It has been performed using a two-tailed Welch's t-test, which is similar to the students t-test and allows the two samples to have unequal variance. The significance level was set to 5%.

It defines the statistic t for two samples X_1 and X_2 as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

where $\bar{X}$ denotes the sample mean, s^2 the sample variance and N the sample size. Together with the estimated degrees of freedom ν, which can be computed using the Welch-Shatterthwaite equation:

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}}$$

the null-hypothesis can be tested using the t-distribution.

Three distributions of the mean correlation coefficients have been computed:

1 for randomly chosen pairs from all available Probe-Set IDs,
2 for randomly chosen pairs from all Probe-Set IDs, for which a mapping to a protein-complex was available,
3 for randomly chosen pairs of Probe-Set IDs that belong to a randomly chosen protein-complexe.

For each distribution, the mean value of 1,000 correlation coefficients of Probe-Set ID pairs have been sampled another 1,000 times. Figure 3.1 shows plots for all three distributions using a gaussian kernel density estimator. For some Probe-Set IDs, the microarray data contains two expressions. Therefore, these have been aggregated using minimum (min), average (avg) and maximum (max). Consequently, each distribution has been computed separately for each aggregation.

Table 3.6 lists the distributions' mean values and table 3.7 the p-values for the performed Welch's t-test of distribution 1 against distribution 2 and distribution 1 against distribution 3.

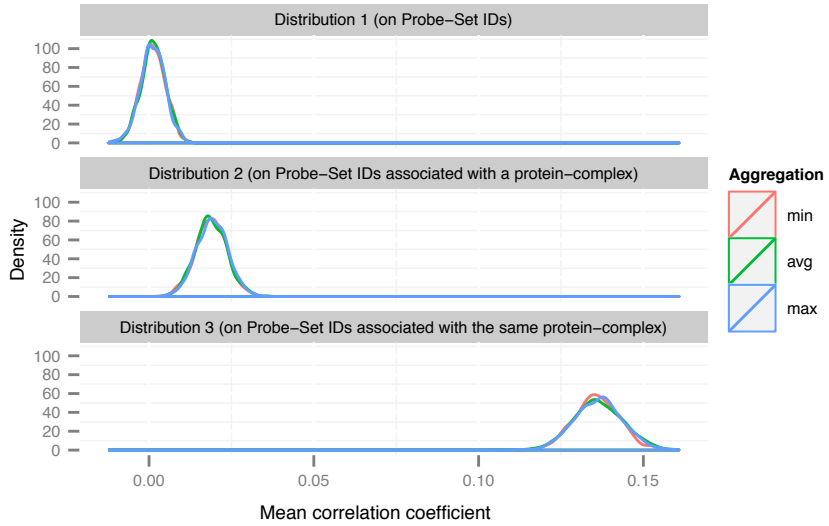


Figure 3.1: Kernel density estimates for the mean correlation coefficient distribution using a gaussian kernel.

	Distribution 1	Distribution 2	Distribution 3
min	0.001225644	0.018802678	0.135954911
avg	0.001337716	0.018857328	0.136323111
max	0.001110031	0.019165409	0.136386125

Table 3.6: The mean values for each distribution and aggregation.

	Distribution 1 against 2	Distribution 1 against 3
min	< 2.2e-16	< 2.2e-16
avg	< 2.2e-16	< 2.2e-16
max	< 2.2e-16	< 2.2e-16

Table 3.7: p-values from the Welch's two sample t-test.

The p-values in table 3.7 show that the null-hypothesis had to be rejected. This means that the distributions that have been tested against each other did not have equal mean.

Furthermore, table 3.6 shows that the mean of distribution 1 and 2 are close to 0, while the mean of distribution 3 is ~14%. It also indicates that the Probe-Set ID expressions aggregated using the maximum increase the spread between the mean of distribution 1 and 2 as well as the spread between the mean of distribution 1 and 3.

The performed analysis suggests that the superimposition of the CORUM database results in a valuable structural gain in information on the microarray breast-cancer data from the Uppsala and Stockholm cohorts.

3.3 RECURSIVE FEATURE ELIMINATION

The second part of this project deals with the inquiry whether the use of structural information could reduce the dimensionality of the classification problem and retain or even enhance the classification performance.

3.3.1 Methods

Recursive Feature Elimination (RFE) is a method, which tries to reduce the number of features in the classification problem to the most relevant features. The method has been described by Guyon et al. (2002) in combination with SVMs (SVM-RFE). SVM-RFE has initially been proposed for binary problems. Let p be the number of features. The algorithm works by removing the features $i = 1, \ldots, p$ with the smallest coefficient of the squared weight vector w_i^2 from the binary SVM. Let l be the number of binary SVMs in a multi-class SVM. The extension to a multi-class SVMs has been done similar to the multi-class algorithm proposed in Zhou and Tuck (2007). In the multi-class SVM-RFE, the features $i = 1, \ldots, p$ with the smallest mean over the corresponding coefficients of the squared weight vectors w_{ji}^2 $(j = 1, \ldots, l)$ are removed.

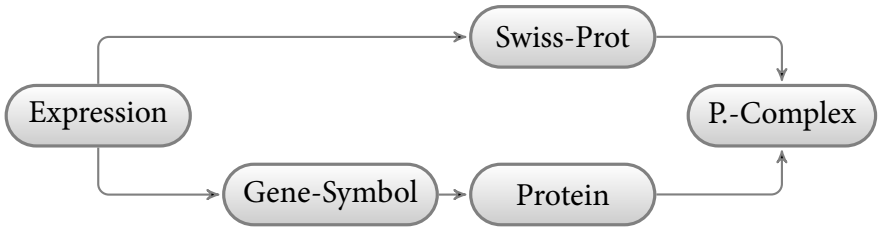


Figure 3.2: Pipelines that were used to compute the virtual expressions for Gene-Symbols, proteins and protein-complexes from the microarray expressions of the Uppsala and Stockholm cohorts.

SVM-RFE can be used to eliminate multiple features at once. For the application of the SVM-RFE on the breast-cancer data-set, piplelines had to be devised to describe the aggregation of expressions for the Probe-Set IDs and Gene-Symbols as well as the proteins and protein-complexes. The two pipelines are visualized in figure 3.2. The upper pipeline uses the Affymetrix GeneChip Information (see § 3.1.2 (p. 24)) to connect Probe-Set IDs to proteins from the Swiss-Prot database. A second step connects these proteins to their corresponding protein-complexes, according to the information from the CORUM database (see § 3.1.4 (p. 25)). The lower pipeline uses the Affymetrix GeneChip information (see § 3.1.2 (p. 24)) to find mapping paths to Gene-Symbols. And from there, paths to proteins through the Additional Gene-Symbol to protein mappings (see § 3.1.3 (p. 25)). The CORUM database is used again in the last step to find connections between proteins and their protein-complexes. These

two pipelines have been implicitly merged for the computation of protein and protein-complex expressions.

The expression data from the microarrays is only annotated with the corresponding Probe-Set IDs. For some Probe-Set IDs, not one, but two expressions are present. These have been aggregated using the aggregation functions: minimum (min), average (avg) and maximum (max). The proteins can then be described as sets of Probe-Set IDs. Because a protein can map to multiple Probe-Set IDs, another aggregation function for the calculation of the protein expressions is introduced: median. The protein-complex expressions have finally been computed using all the aforementioned aggregation function min, avg, median and max on the protein expressions. These computational paths are depicted in figure 3.3.

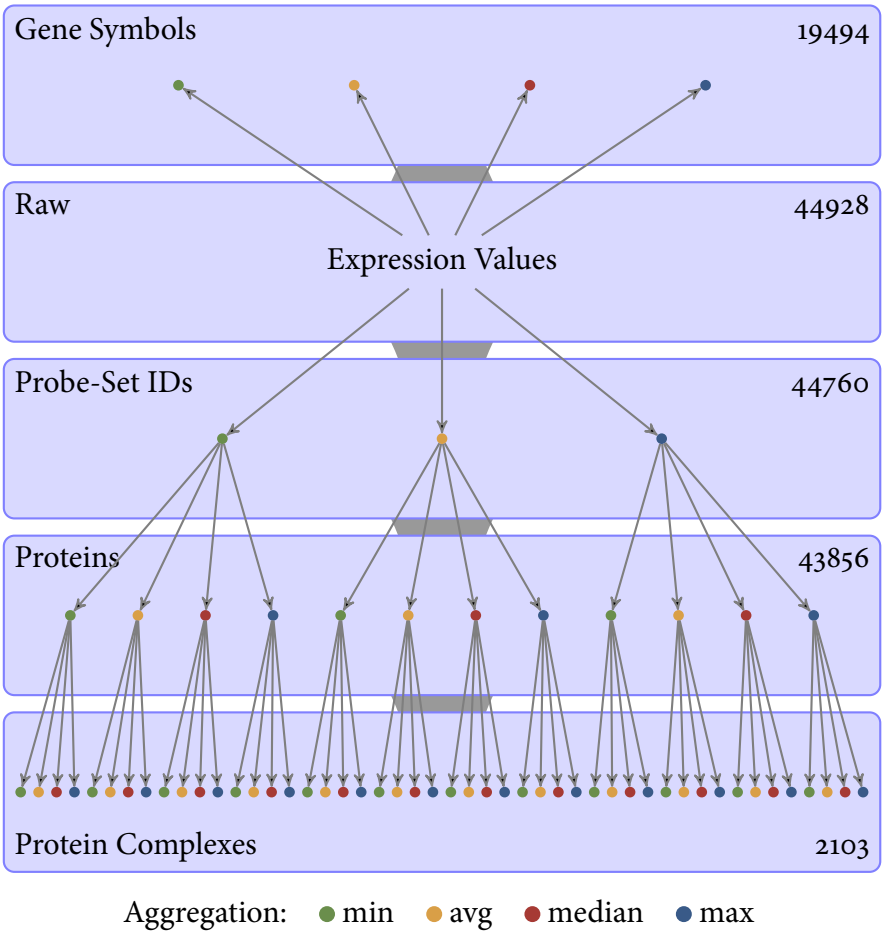


Figure 3.3: The flow of the expression computation. Each blue box represents one step. The number on the right side of each box denotes the number of features, for which expressions could be computed.

Initially, the mapping had been hand coded in MATLAB. It proved to be hard to verify and maintain. Furthermore, it was rather slow, and the adoption to new data or small changes in the pipelines was complicated. A different approach was chosen, and the expression data, accompanied

with the mapping information transferred into a relational database management system (RDBMS). PostgreSQL was chosen because it was readily available and provided the ability to use R functions through the embedded PL/R language. The mapping of the expressions was done in the database with a few Structured Query Language (SQL) statements and subsequently written back into MATLAB files.

The SVM has been trained using the expressions computed from the Uppsala data-set, and the performance has been measured on the expressions computed from the Stockholm data-set. The SVM-RFE that was initially used was based on the LSVM software. The multi-class and RFE logic was custom coded as a layer on top of the LSVM. The multi-class approach that was used was a one-against-all approach with a continuous prediction heuristic as described in § 2.2.2 (p. 16). An issue arose from the Out-of-Memory Exceptions due to the size of the problem. The problem size of the classification of Probe-Set ID expression values from the Uppsala data source required more memory than the available systems provided. As a solution, the LSVM algorithm accompanied with the multi-class and RFE logic was rewritten in Python using NumPy and SciPy to allow it to run on systems that were unable to run MATLAB (e.g. the computer barn at the institute of informatics of the Technische Universität München). Another motivation for this rewrite was that the use of Python would make it easy to run the computation in parallel. But the python version ran out of memory as well.

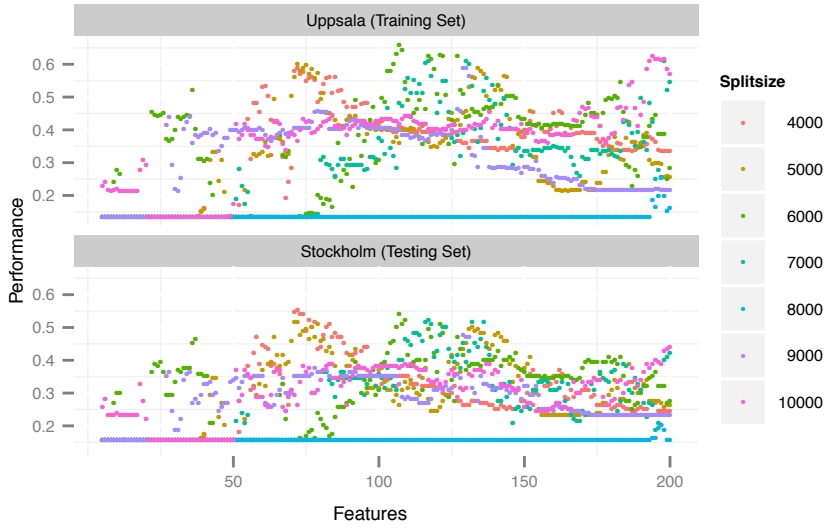


Figure 3.4: The classification performance of the LSVM algorithm with different split-sizes for the last 195 iterations from features 200 to 5.

An idea to circumvent the out of memory problem was to split the problem along the Probe-Set IDs (features) into sub-problems of smaller size. Even though this approach was questionable, it at least allowed to compute a model for the data. A few tests reviled that the computed

model depended heavily on the chosen size of the sub-problem (split-size). The dependence on the splitsize can be seen in the classification performance of the training-set (Uppsala) and the testing-set (Stockholm) for the max-aggregated Probe-Set IDs in figure 3.4.

With all the issues of the LSVM software, a different SVM was sought and the LIBSVM software package was selected. Its excellent support for R with the e1071 Package provided a tuning facility to find the optimal penalty parameter C for the linear SVM using a grid search approach. The LIBSVM has shown to be faster, and able to compute a model for the Probe-Set IDs' expressions in a fixed size of memory. The LIBSVM supports multi-class problems using a pairwise voting strategy (see § 2.2.2 (p. 19)). On top of the LIBSVM, the RFE algorithm has been implemented, based on a similar implementation by Fernández (2008).

3.3.2 *Results*

In this section, classification results using the LIBSVM with the RFE implementation will be presented on Gene-Symbols, Probe-Set IDs, proteins and protein-complexes. For the last fifty features, their Gene Ontology and Functional Catalogue information for the corresponding Gene-Symbols, Probe-Set IDs and Protein-Complexes can be found in appendix A (p. 61).

Classification on Gene-Symbols

Using only the Affymetrix GeneChip information for 19,494 different Gene-Symbols, expressions could be computed. These represent the expressions associated with 34,553 (~77%) of the available Probe-Set-IDs, which are implicit in the mapping[1]. The SVM-RFE performance on the min, avg, median and max aggregated expressions can be seen in figure 3.5. The RFE has been configured to reduce the features by 10% until 200 features are reached. From there on, one feature is removed during each iteration until five features are left and the RFE terminates.

1. The raw Uppsala and Stockholm microarray breast-cancer data is only annotated with Probe-Set IDs, and the Affymetrix GeneChip Information only contains associations between Probe-Set IDs and Gene-Symbols.

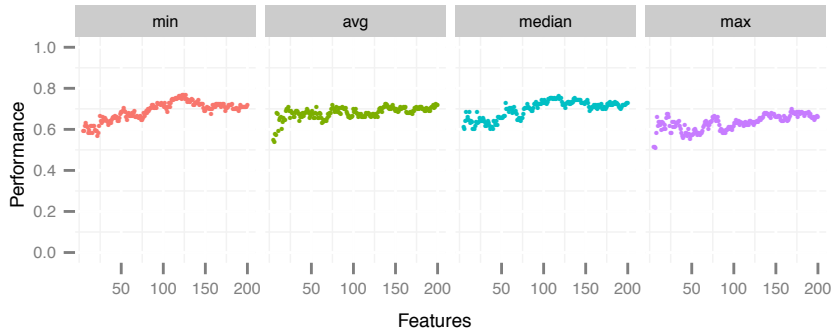


Figure 3.5: Classification-performance using Gene-Symbols on the testing-set (Stockholm) for the models, which have been trained on the Uppsala data-set.

Figure 3.5 only shows the performance of the last 195 iterations (features 200 – 5). The box-plots in figure 3.6 suggest that the expression aggregation with the median yields the best results, but figure 3.5 and figure 3.6 clearly show that an aggregation using the average of the expressions results in a more stable performance.

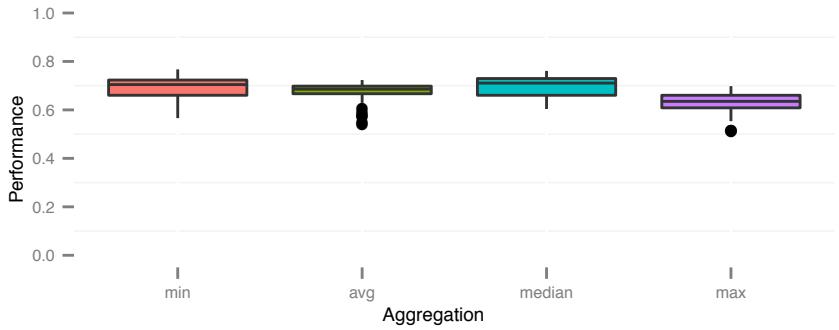


Figure 3.6: Box-plots for the classification performance on Gene-Symbols for the different aggregation functions during the last 195 iterations of the RFE.

When looking at the subtype performance in figure 3.7, one can see that the classification-performance for the average (avg) aggregation decreases for the *No Subtype* subtype, while it even improves for the *ERBB2* subtype with fewer features.

Finally, figure 3.8 shows the mean and standard deviation (SD) of the classification for the 195 last iterations of the RFE. It shows that using Gene-Symbols the classification of *Basal*, *Luminal A*, *Luminal B* and *Normal Like* subtypes is on average easier than the classification of *ERBB2* and *No Subtype* subtypes, which are often assigned to the wrong class.

Figure 3.9 depicts the number of common features that the SVM-RFE found for the different expression aggregations. While the intersection of all different aggregation paths shows only very few common features, the intersection of the two better performing aggregation paths (avg and median) shows more common features. This suggests that these two paths have found Gene-Symbols, which are more relevant for the classification-performance.

The Gene-Symbols at fifty features from the average path are the listed in table 3.8.

A **strong** annotation indicates that the Gene-Symbol is part of all paths, and an *emphasised* annotation indicates that the Gene-Symbol is also part of the fifty features of the median path. The number in parentheses is the cardinality of the subset of features, in which this particular feature was last seen.

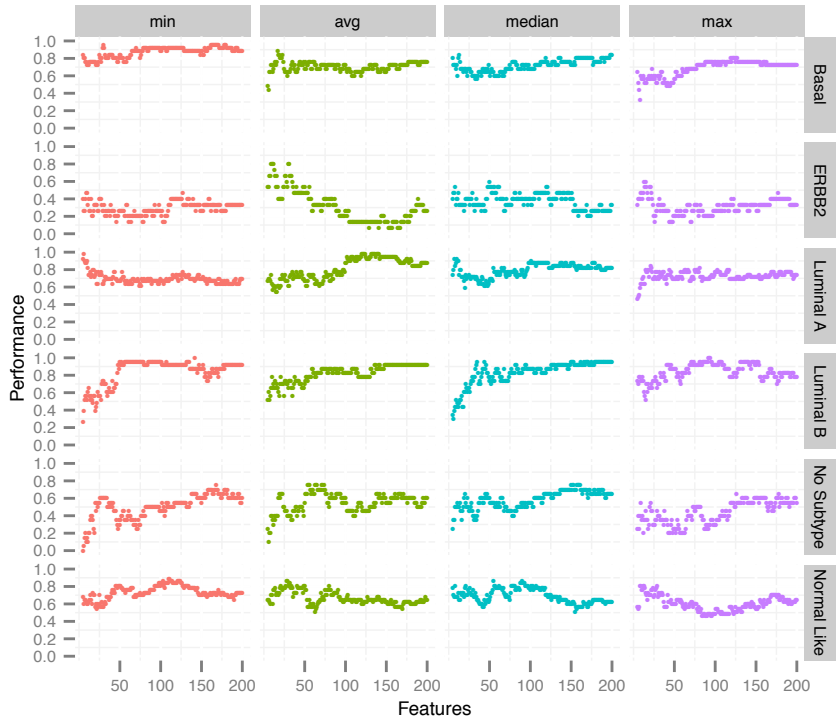


Figure 3.7: Classification-performance on Gene-Symbols; broken down into the six different subtypes.

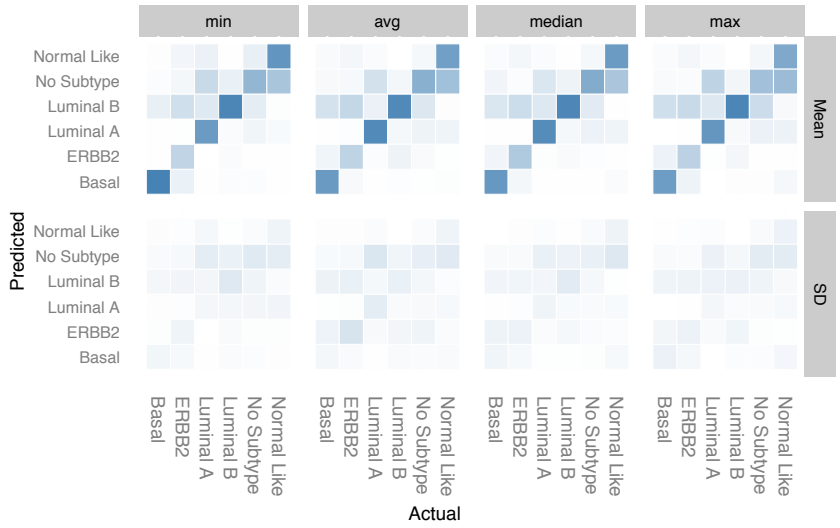


Figure 3.8: Classification for each subtype throughout the last 195 iterations; visualized as a heat-map.

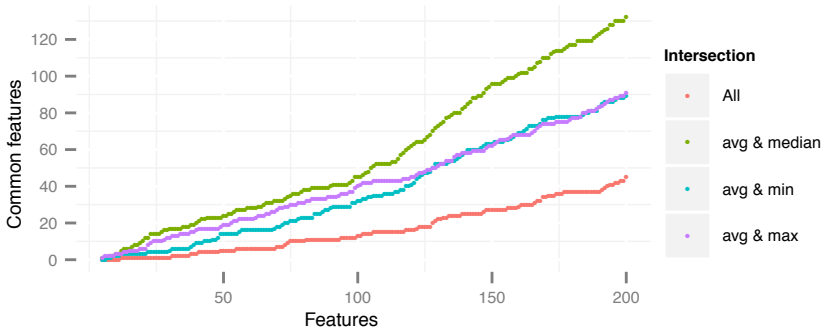


Figure 3.9: The number of common features of the aggregation paths. E.g. avg & median represent the cardinality of the intersection of the Gene-Symbols for the average and median aggregation.

(50) TMEM34	(49) TAS2R5	(48) TMEM91	(47) HN1
(46) ERBB2	(45) COL17A1	(44) GPR87	(43) TBC1D7
(42) CHI3L1	(41) ECHDC2	*(40) ANKRD5*	(39) CYP20A1
(38) LRP2	(37) SP8	(36) EXOSC3	*(35) PTN*
(34) ZNF780B	(33) GRTP1	(32) REPS2	(31) SQLE
(30) GSTM2	(29) NUCB2	*(28) FRAG1*	*(27) MPP7*
(26) MTF1	(25) KLHL20	(24) PPM1J	*(23) SCUBE2*
(22) FAM134B	***(21) CNTD2***	*(20) KRT14*	*(19) ELOVL6*
(18) NOL5A	*(17) DACH1*	*(16) CAPN10*	***(15) KRT223P***
(14) MSI2	*(13) HMGB3*	*(12) SAMD5*	*(11) ZNF684*
(10) ESR1	(9) RDHE2	***(8) DEFB1***	*(7) STARD3*
(6) CBX2	(5) MAPT	*(5) RABEP1*	*(5) SOX10*
(5) TCTA	(5) YWHAZ		

Table 3.8: The Gene-Symbols at the 50 feature step of the SVM-RFE. The number in parentheses is RFE step, in which the feature is eliminated.

Classification on Probe-Set IDs

Similar to the Gene-Symbol classification results, the results for the classification using Probe-Set IDs are presented. They have been computed using the same settings for the RFE that have been used with the Gene-Symbols. Figure 3.10 together with figure 3.11 suggest that the aggregation using the maximum (max) of the expressions yields a good path with respect to the classification performance.

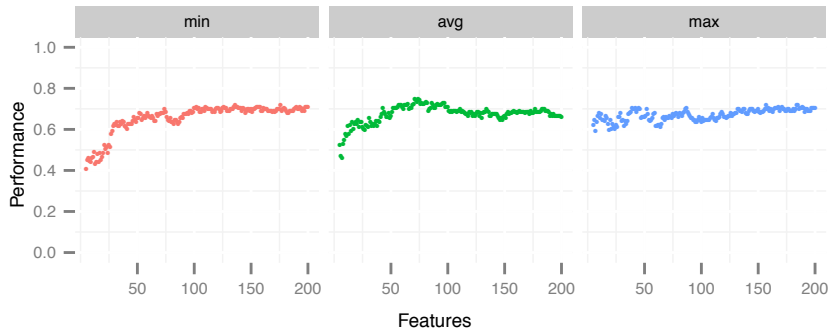


Figure 3.10: Classification-performance using expressions for the Probe-Set IDs. Models have been trained using the Uppsala cohort, and the classification-performance has been measured using the Stockholm cohort.

Using the average to compute the expressions results in a better classification-performance (around 75 features). The plots in figure 3.10 and figure 3.11 suggest that using the minimum to aggregate the expressions for the Probe-Set IDs is suboptimal.

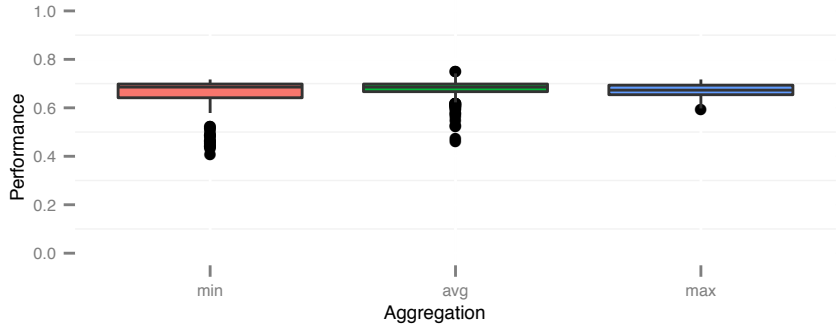


Figure 3.11: Boxplots for the classification-performance on Probe-Set IDs for different aggregation functions.

This is reinforced in figure 3.12 and figure 3.13, which show that the minimum aggregation under-performs for the *ERBB2* subtype.

Figure 3.12 further indicates that the SVM-RFE maximum aggregation improves the classification-performance for the *Luminal A* and *ERBB2* subtypes in favour of the *Luminal B* subtype.

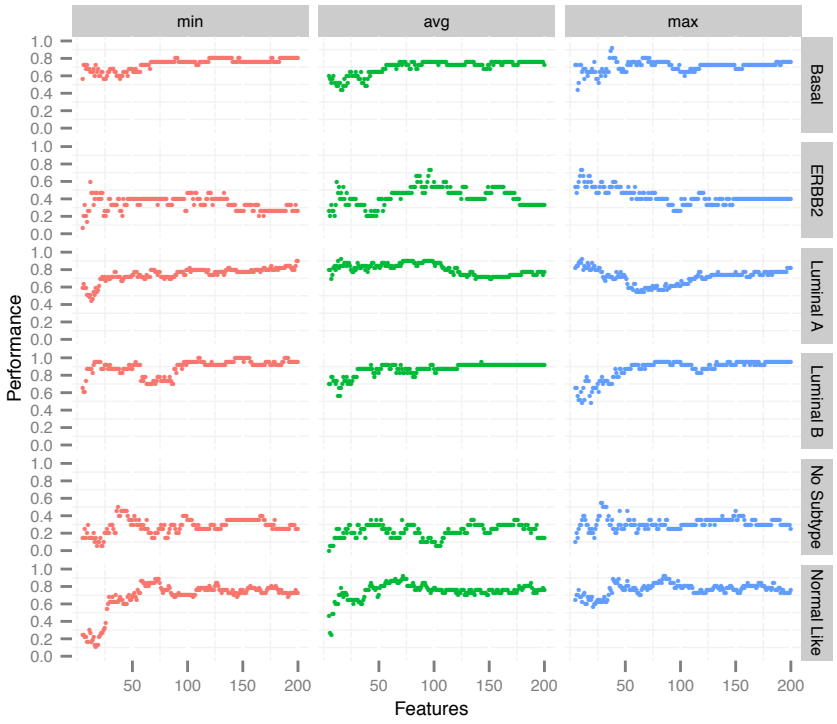


Figure 3.12: Classification-performance on Probe-Set IDs; broken down into the six different subtypes.

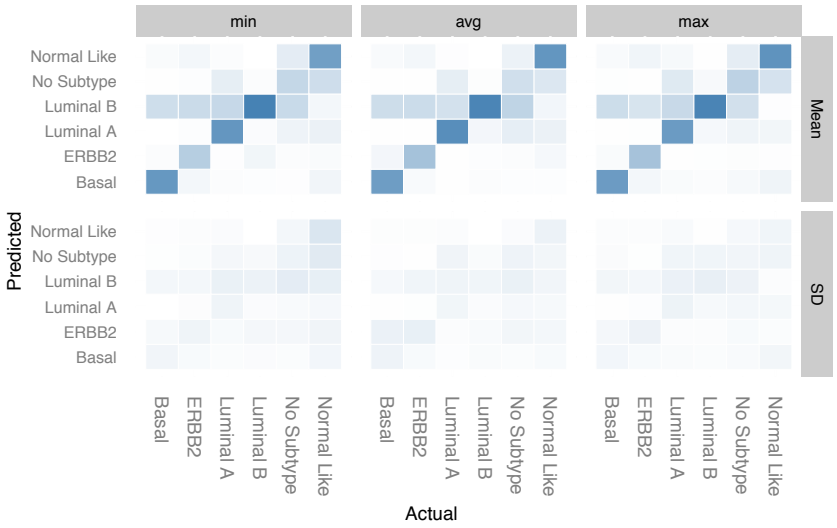


Figure 3.13: Classification for each subtype on Probe-Set IDs throughout the last 195 iterations. Visualized as heatmap.

That the subtype *Luminal B* is very hard to classify correctly using the expressions computed for the Probe-Set IDs, can be clearly seen in figure 3.13. While specimen with *Luminal A* are rarely assigned a different subtype, other subtypes are often miss-classified as *Luminal B* using the models trained during the iterations.

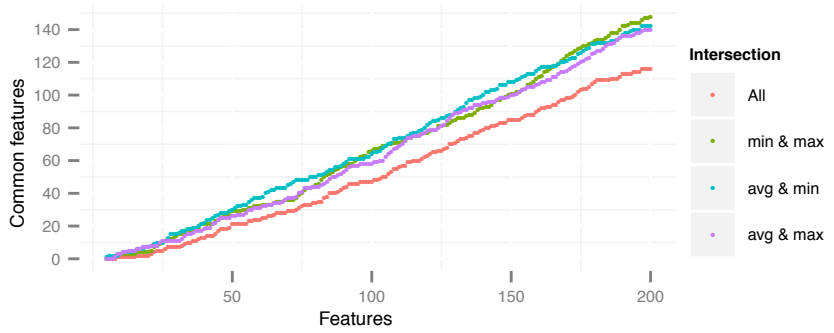


Figure 3.14: The number of common features of selected aggregated paths that have been used to compute the expressions for the Probe-Set IDs.

Figure 3.14 shows that the different aggregation paths often share the same features during the RFE.

Classification on Proteins

Similar to the Gene-Symbols, expression values for 43,856 different proteins could be computed from sets out of 33,735 different Probe-Set IDs by using both pipelines. Thus, for the proteins only ~75% of the available expressions could be utilized. Again, the RFE has been set to reduce the features by 10% until 200 features are reached. Upon which one feature is discarded during each iteration.

The classification-performance on the Probe-Set IDs has not been conclusive to which aggregation should be used on the Probe-Set IDs. Therefore, the results for all twelve computed expression paths are presented.

The best performance (hovering around 70%) could be achieved using the average aggregation on the minimum aggregated Probe-Set IDs (average-on-minimum) according to figure 3.15. This is also the same path that figure 3.16 suggests due to its suggested stability.

The minimum-on-average aggregation, average-on-minimum aggregation as well as the medium-on-maximum aggregation show good performance in both, figure 3.15 and figure 3.16. These paths have been broken down into their subtype performances in figure 3.17.

The heat-map in figure 3.18 indicates that for the average-on-minimum classification the *ERBB2* subtype is classified more often correctly, while the *No Subtype* subtype is less often classified correctly.

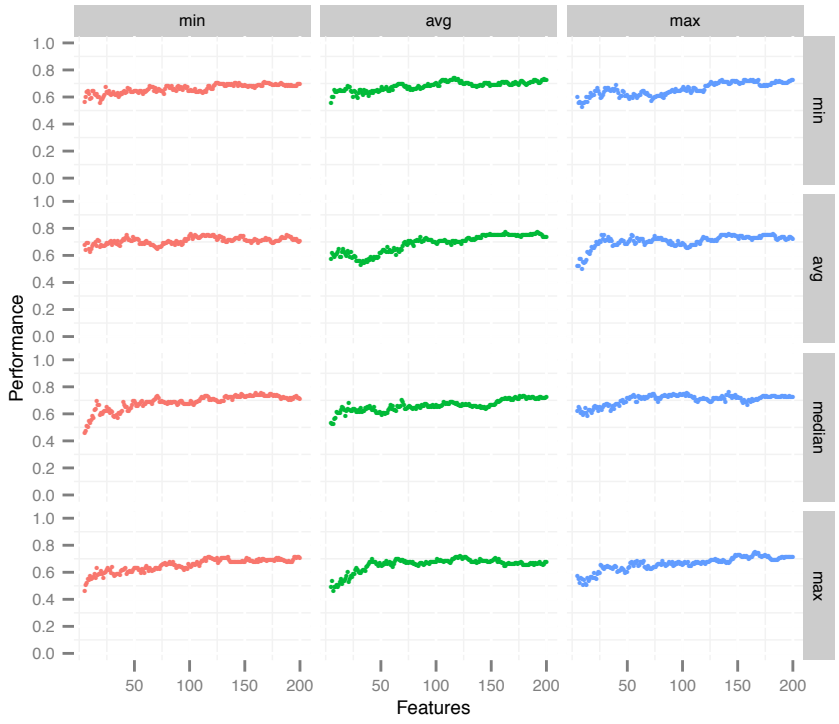


Figure 3.15: Classification-performance on proteins for all available paths that lead to expressions for the proteins.

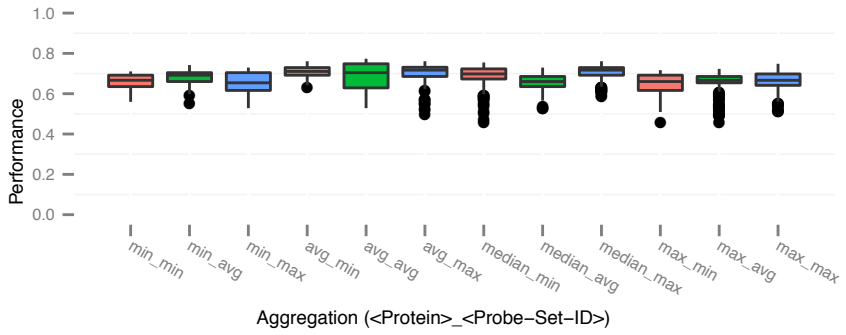


Figure 3.16: Box-plots of the classification-performance on proteins for all available aggregation paths.

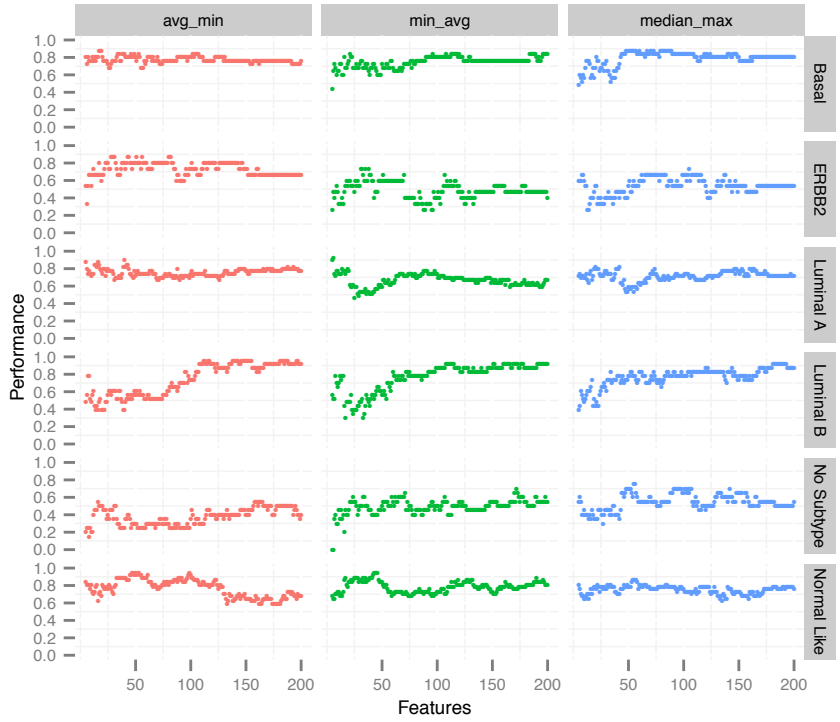


Figure 3.17: Classification-performance on selected expression paths for the proteins; broken down into their six different subtypes.

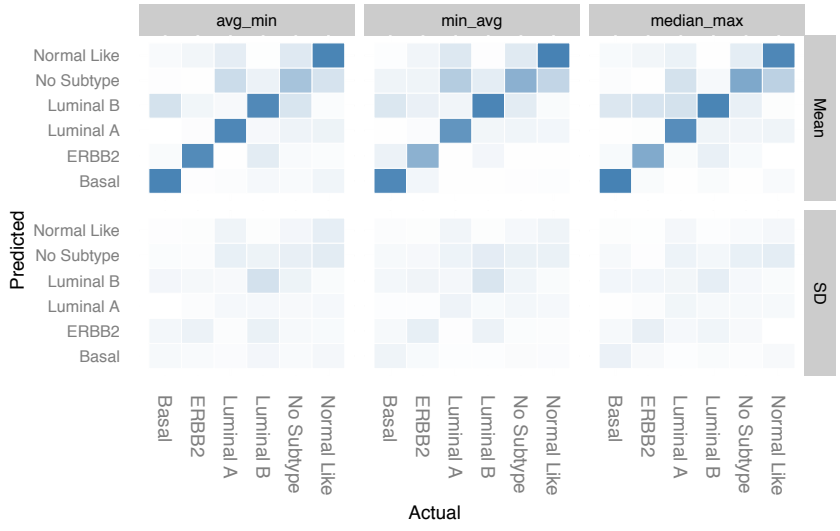


Figure 3.18: Classification for each subtype on selected expression paths for the proteins throughout the last 195 iterations. Visualized as a heatmap for the mean and standard deviation (SD).

Figure 3.19 shows that for the selected subset of paths the median-on-maximum and average-on-minimum paths share the most common features during the RFE.

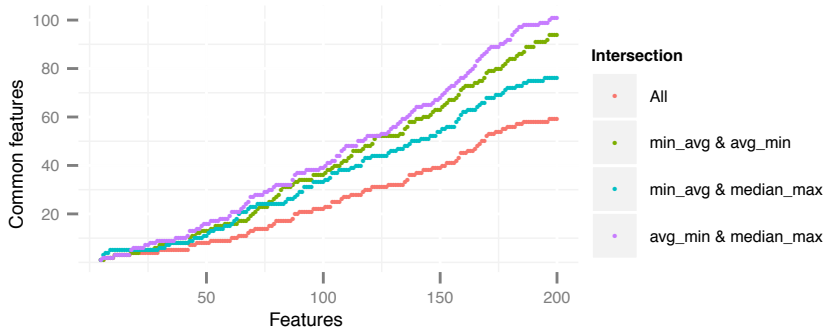


Figure 3.19: The numbers of common features of selected aggregation paths that have been used to compute the expressions for the proteins.

The last fifty remaining proteins for the average-on-minimum path are listed in table 3.9.

(50) Q59F11	(49) Q9BXG2	*(48) Q53G41*	(47) Q6ZVB0
(46) Q9HC96	(45) Q49A35	*(44) Q14508*	***(43) Q59EN9***
(42) Q15276	*(41) Q9NQF3*	(40) Q5U0B0	(39) Q7Z5Q5
(38) Q6FGL5	(37) Q9H5J4	(36) P57738	(35) Q5TC63
(34) Q9GZU0	***(33) Q6PJW8***	(32) Q9BX40	(31) Q99712
(30) Q9UI36	(29) Q8NFH8	(28) Q5T9H0	(27) O00148
(26) Q8N589	(25) Q14781	(24) Q9NYW4	*(23) Q0D2I8*
(22) Q9UNR6	(21) Q7Z5L7	(20) P21246	(19) Q8IY28
(18) Q7Z5C1	(17) Q6ZRF6	(16) Q9NVA4	(15) Q96C34
(14) Q9JIV6	(13) Q9BVG4	(12) Q96LX1	***(11) P60022***
(10) Q7Z684	(9) Q9NR86	(8) P05090	(7) Q9UK79
(6) Q6I9U4	(5) O35551	(5) Q13751	(5) Q6ZT07
(5) Q9H8F1	***(5) Q9UK76***		

Table 3.9: The protein accession numbers at the 50 feature step of the SVM-RFE using the average-on-minimum path. The number in parentheses is RFE step in which the feature is eliminated.

A **strong** annotation indicates that the protein is part of all paths, and an *emphasised* annotation indicates that the protein is also part of the fifty features of the median-on-maximum path. The number in parentheses is the cardinality of the subset of features, in which this particular feature was last seen.

Classification on Protein-Complexes

The protein-complexes represent the final mapping target. As described in figure 3.2, expressions could be computed from the microarray breast-cancer data using the pipelines for 2,103 different protein-complexes. These 2,103 protein-complexes are based on sets out of 3,519 different proteins, which in turn are based on sets out of 6,169 different Probe-Set IDs. Thus, the protein-complexes only represent ~8% of the expressions for proteins.

Based on the results from the previous section on the protein performance, only the path of the average-on-minimum aggregated expression values will be followed. For this sub-tree, the performance results will be presented. The RFE has been set to reduce the features by 10% till 200 features are reached. Upon which one feature is eliminated during each iteration.

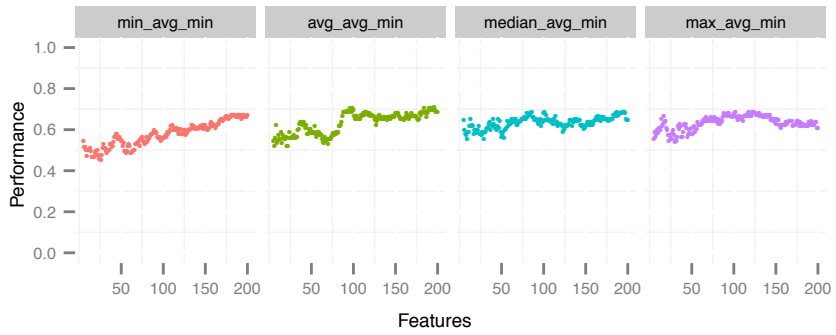


Figure 3.20: Classification-performance for the four differently aggregated protein-complex expressions following the average-on-minimum path.

After the average-on-minimum aggregation, figure 3.20 and figure 3.21 show that a very good aggregation is the median and very bad aggregation is the minimum.

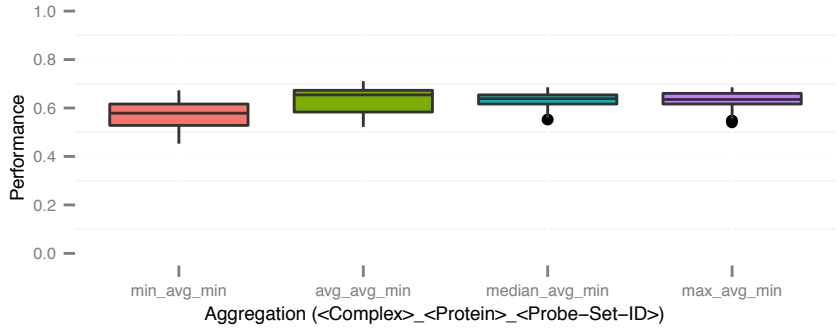


Figure 3.21: Box-plots for the classification-performance for protein-complex expressions following the average-on-minimum path.

Given that the median has been used to compute the protein-complex expression from its corresponding protein-set, one can see from the subtype-performance in figure 3.22 that the classification-performance for the *Basal* and *ERBB2* subtype improve, while it decreases for the subtype *Luminal B* at the same time.

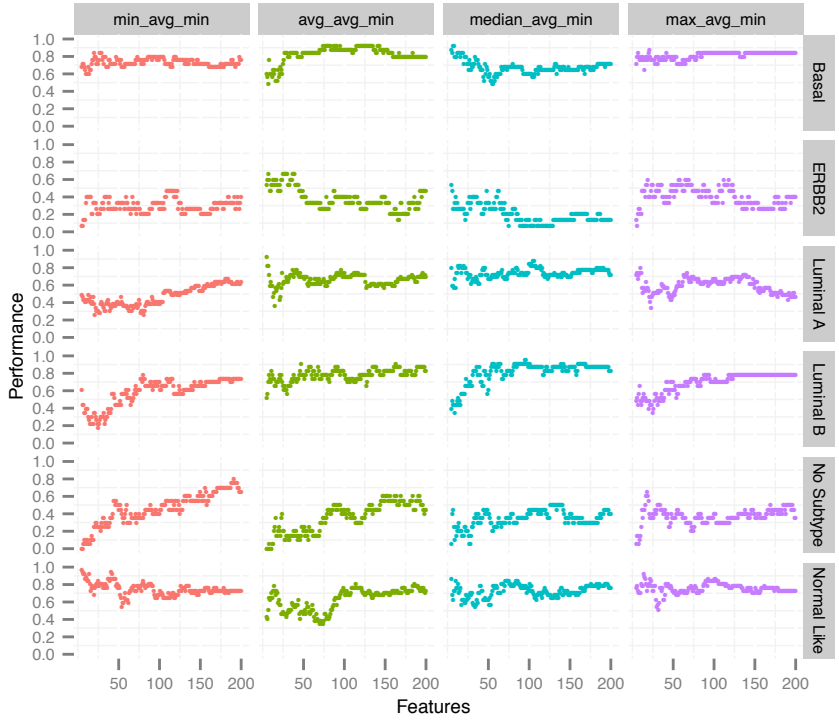


Figure 3.22: Classification-performance on protein-complexes following the average-on-minimum path for the proteins; broken down into the six different subtypes.

In figure 3.23 the heat map shows that, when using the median to aggregate protein-set expressions, the subtypes *Basal*, *Luminal A*, *Luminal B* and *Normal Like* often get classified correctly, while the subtypes *ERBB2* and *No Subtype* get miss-classified frequently. It should be noted that this is measured over the last 195 RFE iteration steps.

Figure 3.24 finally indicates that all four aggregation paths result in quite different feature-sets during the RFE, and that the most feature-similar paths are those using median and avg on the protein-sets.

As presented during the results for the Gene-Symbols and proteins, this chapter will end with the protein-complexes for the iteration step at fifty features for the median-on-average-on-minimum protein-sets. These can be found in table 3.10.

Again, protein-complexes shared with the average aggregated path are denoted with *emphasis*. At fifty features, no protein-complexes are shared among all four aggregation paths as figure 3.24 shows. The number in parentheses is the cardinality of the last feature-set in which the protein-complex was seen.

(47) Polyadenylation complex (CSTF1, CSTF2, CSTF3, SYMPK CPSF1, CPSF2, CPSF3)	
(45) Rab5 GDP/GTP exchange factor complex	
(40) Phosphatidylinositol 3-kinase complex (PIK3CA, PIK3R1)	
(39) Transcription elongation factor complex (SUPT5H, CDK9, CCNT1)	
(38) Kinase-scaffold-phosphatase complex, PKA-AKAP79-CaN	
(32) Chromosomal passenger complex CPC (INCENP, CDCA8, BIRC5, AURKB)	
(30) Prolactin (PRL) - PRL receptor (PRLR) complex	
(28) hMediator complex (MED23, CDK8, CCNC, MED7)	
(24) RICH1/AMOT polarity complex, Flag-Rich1 precipitated	
(19) nephrin-cadherin complex (Nphs1, Ctnnd1, Cdh3, Cd2ap)	
(18) NMDA receptor complex (NR2A, NR2B, NR1, PSD-95)	
(14) Ubiquitin E3 ligase (CRY1, SKP1A, CUL1, FBXL3)	
(5) Ubiquitin-protein ligase (UBE2N, UBE2V2/MMS2)	
(5) Nephrin-cadherin complex (Nphs1, Ctnnd1, Cdh3, Cd2ap)	
(50) SCF subcomplex (WEE1, SKP2, BTRC)	(49) PGC-1-SRp40-SRp55-SRp75 complex
(48) MAML3-RBP-Jkappa-Notch4 complex	*(46) AR-AKT-APPL complex*
(44) PAR-3-PKCz-Tiam2 complex	(43) Eps15-stonin2 complex
(42) HOXA9-PBX2-MEIS1 complex	(41) IL12B-IL12RB1-IL12RB2 complex
(37) Mi-2/NuRD-MTA2 complex	(36) SMG-1-Upf1-eRF1-eRF3 complex (SURF)
(35) GluR1-GluR2 heteromer complex	*(34) ERBB2-MEMO-SHC complex*
(33) p27-cyclinD2-Cdk4 complex	(31) Psd3-Actn1 complex
(29) Glur4-cadherin-catenin complex	(27) Rab11-Fip2-Reps1 complex
(26) ITGAV-ITGB3-NOV complex	(25) REST-CoREST-mSIN3A complex
(23) Acinar cell-specific C complex	(22) ITGA2-ITGB1-COL6A3 complex
(21) c-Src-Muc1 complex	(20) Ecsit complex (Ecsit2-Smad1)
(17) GIPC1-NTRK1-RGS19 complex	(16) SHARP-CtBP1-CtIP complex
(15) Ric-8A G alpha 13 complex	(13) ER-alpha-GRIP1-c-Jun complex
(12) Pparalpha-Pric320 complex	(11) Smooth muscle dystroglycan complex
(10) Survivin homodimer complex	*(9) ActRIIA-ActRIB-Smad3-Arip1 complex*
(8) SMAD3-VDR complex	(7) TRF1-TIN2 complex
(6) PYR complex	(5) APLG1-Rababtin5 complex
(5) EGFR-containing signaling complex	*(5) BLM-TRF2 complex*

Table 3.10: The Protein-Complexes at the 50 feature step of the SVM-RFE. The number in parentheses is the RFE step, in which the feature is eliminated.

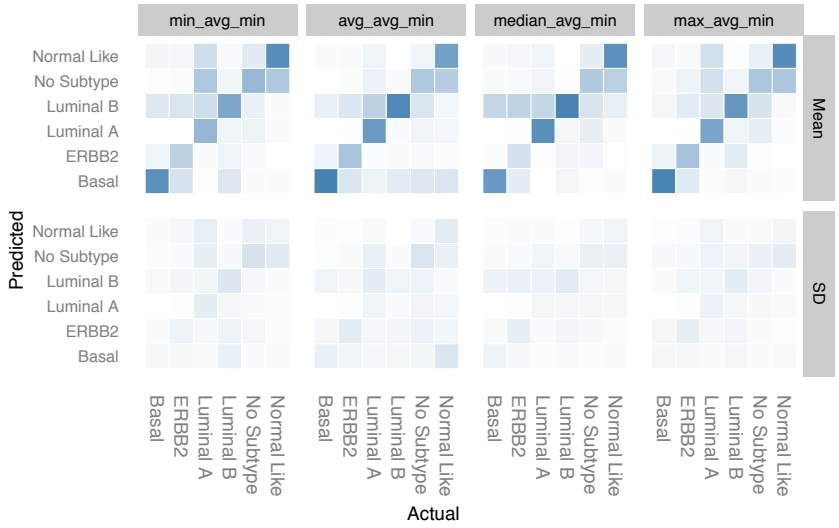


Figure 3.23: Classification for each subtype on selected expression paths for the protein-complexes throughout the last 195 iterations of the RFE. Visualized as a heatmap for the mean and standard deviation (SD).

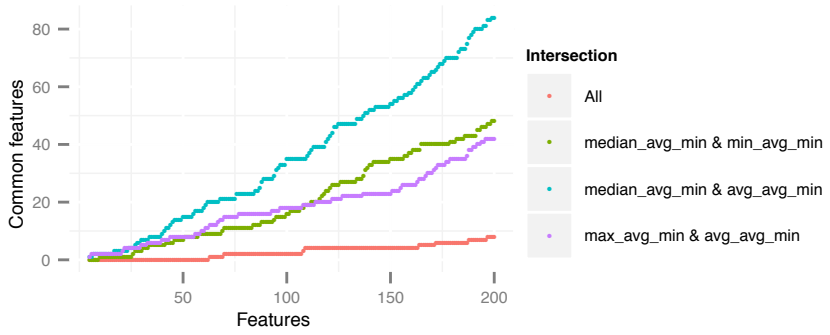


Figure 3.24: The numbers of common features of selected aggregation paths that have been used to compute the expressions for the protein-complexes.

3.3.3 *Discussion*

The results from the recursive feature elimination using support vector machines have shown that during the RFE, the classification performance fluctuates and depends on the aggregation paths used to compute the expressions.

The classification-performance on Gene-Symbol expressions representing ~77% of the available expressions for Probe-Set IDs shows to be around 70% for most of the last 195 iterations during the recursive feature elimination. This is similar to the classification performance using Probe-Set IDs. The classification on protein expressions representing ~75% of the Probe-Set ID expressions yields a similar performance, while the classification performance on protein-complexes is slightly below ~70%.

The number of common features for the top-performing paths on Gene-Symbol expressions is ~130 when the RFE is at 200 features. This is topped by the number of common features for the top-performing paths on Probe-Set ID expressions, which lie at ~150 common features. The number of common features for the protein expressions and protein-complex expressions are significantly lower at ~110 and ~90 common features respectively. These represent ~55% and ~45% of the 200 features for each path at this point during the RFE.

Additionally, the spread between the number of common features for the top-performing paths and all paths increases the longer the aggregation paths get. This is a further indication that the choice of aggregation functions for the computation of the expressions for the Gene-Symbols, Probe-Set IDs, proteins and protein-complexes plays a major role in the feature subset that the recursive feature eliminations selects.

Gene-Symbols

Some of the last fifty remaining Gene-Symbols play a role in cancer. This is according to their EntrezGene summaries.

- The expression of the *dachshund homolog 1 (DACH1)* is lost in some forms of metastatic cancer.
- The *v-erb-b2 erythroblastic leukemia viral oncogene homolog 2 (ERBB2)* has been reported to be over-expressed in numerous cancers, including breast and ovarian tumors (Lisa et al., 1999).
- The *estrogen receptor 1 (ESR1)* encodes an estrogen receptor. Estrogen receptors are known to be involved in pathological processes including breast-cancer.
- The *TBC1 domain family member 7 (TBC1D7)* belongs to a family of proteins, which are presumed to play a role in cell growth and differentiation.
- Finally, the Gene-Symbol *TCTA* is also part of the last 50 features. This

Gene-Symbol corresponds to the *T-cell leukemia translocation altered gene*.

The most obvious breast-cancer related Gene-Symbols are ERBB2 and ESR1. ERBB2 is already eliminated during the RFE from features 46 to 45, but ESR1 is kept as one of the last Gene-Symbols used for classification.

Further genes have already been found in other studies to be cancer related.

- The *transmembrane proteins 34 (TMEM34)* has been found to be down-regulated in breast-cancer specimens, compared to normal specimens (Mehta, 2009, p. 285)
- The *hematological and neurological expressed 1 (HN1)* has been one of four genes distinguishing tumor samples from normal ovarian surface epithelial cells (Lu et al., 2004).
- The *G protein-coupled receptor 87 (GPR87)* is essential for p53-dependent cell survival upon DNA damage, and may be utilized for cancer treatment and prevention (Zhang et al., 2009).
- The secretion of *pleiotrophin (PTN)* has been found to remodel the tumor's micro-environment, and stimulates breast-cancer (Chang et al., 2007).
- Down-regulated *nucleobindin 2 (NUCB2)* was recently found in gastic cancer cells, and is a potential tumor antigen (Kalnina et al., 2009).
- Links between the induction of apoptosis and the down-regulation of *FGF receptor activating protein 1 (FRAG1)* have been observed (Ishii et al., 2005).
- Finally, the Gene-Symbol *SCUBE2* corresponding to the *signal peptide, CUB domain, EGF-like 2* has been expressed in invasive breast carcinomas. It has recently been found that the alteration of the expression of *SCUBE2* is important in breast cancer progression (Cheng et al., 2009).

The gene set has also been imported into the Genomatrix Bibliosphere (Scherf et al., 2005). The top 10 results from the Gene Ontology (GO) enrichment and the Medical Subject Headings (MeSH) disease analysis are listed in table 3.11 and 3.12.

The Gene Ontology enrichment is not very conclusive, but the MeSH disease analysis shows that the gene set found is very well related to breast diseases and carcinomas.

Term	Z-Score
cellular component biogenesis	2.98
tissue development	2.92
nervous system development	2.89
organ development	2.72
system development	2.44
multicellular organismal development	2.22
lipid metabolic process	2.21
positive regulation of cellular process	2.19
anatomical structure development	2.19

Table 3.11: Gene Ontology (GO) enrichment results for the last 50 surviving Genes (using average aggregation).

Term	Z-Score
Breast Neoplasms [C04.588.180]	184.03
Breast Neoplasms [C17.800.090.500]	184.03
Breast Diseases [C17.800.090]	182.71
Skin Diseases [C17.800]	128.10
Skin and Connective Tissue Diseases [C17]	112.12
Neoplasms by Site [C04.588]	95.95
Carcinoma, Ductal, Breast [C04.588.180.390]	77.35
Carcinoma, Ductal, Breast [C17.800.090.500.390]	77.35
Carcinoma, Ductal, Breast [C04.557.470.615.132.500]	77.35

Table 3.12: Results from the Medical Subject Headings (MeSH) disease analysis for the last 50 surviving Genes (using average aggregation).

Proteins

Among the last fifty proteins that the RFE on the multi-class SVM left, are the following cancer related proteins:

- The *WAP four-disulfide core domain protein 2 (Q14508)*, which is said to be highly expressed in tumor cell lines, including breast and ovarian cancer.
- The isoform 2 of the *RalBP1-associated Eps domain-containing protein 2 (Q8NFH8)* is down-regulated during the progession of prostate cancer.
- The *transmembrane protein 184C (Q0NVA4)* is a possible tumor suppresor, which may play a role in cell growth.
- Finally, the *dachshund homolog 1 (Q9UI36)* as well as the *T-cell leukemia translocation-altered gene protein (P57738; associated with T-cell accute lymphoblasic leukemia)* are present among the last 50 proteins used for classification. These two were also found for the Gene-Symbol expressions.

A notably feature of the set of fifty proteins is that for ~50% (24/50) of the proteins their associated Gene-Symbols have been part of the set of fifty selected Gene-Symbols. For each of the last six proteins their corresponding Gene-Symbol was found in the set. These include the Gene-Symbols *TMEM34*, *DACH1*, *HN1* and *ERBB2*, which have been found to be related to cancer. Almost all of the above mentioned proteins are gone when the subset's cardinality is 35. Only the *transmembrane protein 184C (Q0NVA4)* is kept until the elimination passes 16 features. The information for the proteins has been sourced from the UniProt KnowledgeBase (UniProtKB).

Protein-Complexes

The last targets in the expression computation were the protein-complexes with the most intermediate steps. For each of the fifty last protein-complexes, their complex details were looked up through the MIPS Genome Research Environment (GenRE) web interface of the CORUM database at http://mips.helmholz-muenchen.de/genre/proj/corum.

From the categorization in the MIPS Functional Catalogue (FunCat) (http://mips.helmholz-muenchen.de/proj/funcatDB), which are embedded in the protein-complex details through the web interface, one could see that out of the fifty protein-complexes, 18 had their sub-cellular localization in the nucleus, and 13 took part in the cellular communication. Furthermore, 12 were involved in mRNA transcriptional control, and six were annotated with the cell-cycle. Another three of them are associated with proteins that have also been part of the last fifty proteins. This is in line with the ratio of the number of all proteins to the number of proteins, from which the protein-complexes have been computed (~8%). Further protein-complexes are discussed in more detail.

A critical regulator for the chromosomal segregation is the *chromosomal passenger complex (CPC)* that was among the last fifty protein-complexes. It corrects the nonbipolar microtubule-kinetochore interaction. Therefore, the CPC is presumed to prevent premature mitotic exit through the creation of unattended kinetochores. These are then sensed by the spindle assembly checkpoint (SAC) where the proper chromosome attachment to spindle microtubules is monitored (Rudner and Murray, 1996; Vader et al., 2007). The proper connection of microtubules to kinochores are essential for the chromosome segregation. Many cancer cells are associated with a property called chromosome instability. This is the loss or gain of chromosomes, and is suspected to arise from a lesion within the chromosomal segregation machine (Tanaka and Hirota, 2009).

Another complex has been found that is closely related to the CPC. The *Survivin homodimer complex* has been found to be over-expressed in tumor cells. As a part of the inhibitor of apoptosis protein (IAP) family, the protein *Survivin* (in its monomeric form) plays a role in the Chromosomal Passenger Complex (CPC). Its dimerization usually occurs in the form of a bow tie-shaped homodimer, in which it can result in the proliferation of tumor cells due to its suppression of apoptosis (Park and Li, 2010).

According to the comments on the protein-complex details page, the *HOXA9-PBX2-MEIS1 complex* is associated with leukemia, and is present in myoloid leukemia (Shen et al., 1999). It has been shown that the involvement of homeobox (HOX) proteins and the so called three amino acid loop extension (i.e., PBX and MEIS) are present in leukemia. Especially HOXA9 plays a key role for the characteristics of leukemia, and is involved in the morphogenesis. The onset of the leukemic transformation is accelerated by MEIS1, which functions as a cofactor for HOXA9 (Thorsteinsdottir et al., 2001; Sitwala et al., 2008).

The *MAML3-RBP-Jkappa-Notch4 complex* contains the *Neurogenic locus notch homolog protein 4 (Notch4)* protein. The notch family controls cell fate decisions and takes part in many developmental processes. Together with *mastermind-like (MAML)* transcriptional co-activators, which are important for the Notch signaling pathway, the Notch family is known to be related to breast-cancer (Wu and Griffin, 2004; Harrison et al., 2010). It has been shown that increased *RBP-Jkappa* dependent Notch signaling transforms normal breast epithelial cells through the suppression of apoptosis (Stylianou et al., 2006).

According to its FunCat annotation, the *Polyadenylation complex (CSTF1, CSTF2, CSTF3, SYMPK CPSF1, CPSF2, CPSF3)* is involved in the 3'-end processing. The enzyme polyadenylate polymerase (PAP) is a catalyst for polyadenylation and important in the determination of the stability of mRNA. Measurements of PAP are even part of the biological profile of tumor cells definitions. They have been found to be a possible prognostic factor in leukemia and breast-cancer (Scorilas, 2002).

Over-expression of the human growth hormone (HGH) has been

linked to the development of breast-cancer. Wennbo et al. (1997) have shown that in transgenic mice the activation of the Prolactin (PRL) receptor (PRLR) is sufficient for the induction of breast-cancer. An important signaling mechanism for tumor cells is the autocrine/paracrine loop. Reynolds et al. (1997) have shown evidence for the autocrine/-paracrine loop for the *Prolactin (PRL) - PRL receptor (PRLR) complex* within human breast tissues.

An important protein in multicellular organisms is the *protein 53 (p53)*, where it is a regulator for the cell cycle. It is involved in the prevention of cancer due to its tumor suppressor functions. One out of the last five surviving protein-complexes is the *BLM-TRF2 complex*. It was found that for the survival of cells without functional p53 this complex is particularly important (Kim, 2008). It is also related to the *Bloom syndrome* according to the comments on the protein-complex page. The *Bloom syndrome* is a rare autosomal recessive chromosomal disorder and poses a high risk to develop a broad spectrum of cancers for its carrier (German, 1997).

ERBB2 is known to play an important role in breast-cancer cell mortality and metastases formation. Its over-expression is specific to tumor cells. For ERBB2s ability to inhibit apoptosis the SHC signaling protein is required. Furthermore, SHC acts as a signaling pathway between ERBB2 and the Mediator of ERBB2 driven cell Mortality (MEMO). Therefore, the *ERBB2-MEMO-SHC complex* is highly related to breast-cancer (Lucs et al., 2009; Marone et al., 2004).

Another important protein is the *Transforming growth factor beta (TGF-β)* protein. It also plays an important role as a regulator during cell cycle, and can induct apoptosis via the SMAD pathway. In cancer cells the TGF-β pathways are mutated, and prohibit TGF-β to control the cell. Two directly related protein-complexes were identifiable within the last fifty using their FunCat annotation: the *Ecsit complex (Ecsit2-Smad1)* and the *SMAD3-VDR complex*. Both are annotated with the *TGF-beta-receptor signalling pathway*.

Finally, a FunCat-enrichment was performed with a hyper-geometric test on the last fifty protein-complexes. The results are summarized in table 3.13. From this we can see that some cell communication related FunCat entries are more frequent in our set of fifty Protein-Complexes than they would be in a random set of fifty Protein-Complexes (e.g. ligand-dependent nuclear receptors and regulation of signal transduction).

Factional Catalogue Entry	p-Value	Last 50 Protein-Complexs	All Protein-Complexes
nucleic acid binding	0.0450	1	2
ligand-dependent nuclear receptors	0.0390	3	33
regulation of signal transduction	0.0108	3	20
cell-cell adhesion	0.0365	2	13
rhythm (e.g. circadian, ultradian)	0.0453	1	2
learning and memory	0.0447	1	2
embryogenesis	0.0040	3	14
microtubule cytoskeleton	0.0187	2	9
blood cell	0.0139	2	8
nervous tissue	0.0042	2	5
intercellular junction (gap junction/adherens junction)	0.0112	2	7
lymphatic system	0.0487	1	2

Table 3.13: Results from the FunCat-enrichment on the last fifty protein-complexes, restricted to those categories with a p-value below 0.05.

3.3.4 Conclusion

We have shown that the SVM-RFE has selected breast-cancer related features using the virtual protein-complex expression we computed from the Uppsala microarray breast-cancer data set using our novel data integration method. The classification performance measured on a microarray breast-cancer data set from a different study from Stockholm have shown to be similar for the classification on Gene-Symbol expressions and protein expressions. Although the classification performance on protein-complex expressions did not improve, the extracted protein-complexes are biologically interesting and interpretable as the discussion shows. Our proposed data integration method has shown to be an interesting new approach for the extraction of cancer related protein-complexes, and can be used to gain new insights into the role and function of protein-complexes.

BIBLIOGRAPHY

Affymetrix. 2010. *Affymetrix — Manual, Probe Set Data in Tabular Format*. URL http://www.affymetrix.com/support/technical/manual/taf_manual.affx. [Online; accessed 30-March-2010]. Cited on p. 24.

Alberts, B. 1998. *The cell as a collection of protein machines: preparing the next generation of molecular biologists*. In Cell, vol. 92, no. 3, pp. 291–294. Cited on p. 6.

Blackstock, W.P. and Weir, M.P. 1999. *Proteomics: quantitative and physical mapping of cellular proteins*. In Trends in Biotechnology, vol. 17, no. 3, pp. 121–127. Cited on p. 7.

Boser, B.E.; Guyon, I.M.; and Vapnik, V.N. 1992. *A training algorithm for optimal margin classifiers*. In Proceedings of the fifth annual workshop on Computational learning theory, pp. 144–152. ACM. Cited on p. 10.

Calza, S.; Hall, P.; Auer, G.; Bjöhle, J.; Klaar, S.; Kronenwett, U.; Liu, E.; Miller, L.; Ploner, A.; Smeds, J.; et al. 2006. *Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients*. In Breast Cancer Research, vol. 8, no. 4, p. R34. Cited on p. 23.

Chang, Chih-Chung and Lin, Chih-Jen. 2001. *LIBSVM: a library for support vector machines*. Cited on p. 19.

Chang, Y.; Zuka, M.; Perez-Pinera, P.; Astudillo, A.; Mortimer, J.; Berenson, J.R.; and Deuel, T.F. 2007. *Secretion of pleiotrophin stimulates breast cancer progression through remodeling of the tumor microenvironment*. In Proceedings of the National Academy of Sciences, vol. 104, no. 26, p. 10888. Cited on p. 46.

Cheng, C.J.; Lin, Y.C.; Tsai, M.T.; Chen, C.S.; Hsieh, M.C.; Chen, C.L.; and Yang, R.B. 2009. *SCUBE2 suppresses breast tumor cell proliferation and confers a favorable prognosis in invasive breast cancer*. In Cancer Research, vol. 69, no. 8, p. 3634. Cited on p. 46.

Coen, E. 1999. *The art of genes: How organisms make themselves*. Oxford University Press, USA. Cited on p. 6.

Crammer, K. and Singer, Y. 2001. *On the algorithmic implementation of multi-class svms*. In Journal of Machine Learning Research, vol. 2, pp. 265–292. Cited on p. 15.

Crick, F. 1970. *Central dogma of molecular biology*. In Nature, vol. 227, no. 5258, pp. 561–563. Cited on pp. 3 and 4.

Fernández, Elmer Andrés. 2008. *R implementation of the Support Vector Mahine Recursive Feature Extraction (SVM-RFE) Algorithm*. In Elmer Andrés Fernández, vol. 6, no. 6, pp. 515–522. Cited on p. 31.

German, J. 1997. *Bloom's syndrome. XX. The first 100 cancers*. In Cancer genetics and cytogenetics, vol. 93, no. 1, pp. 100–106. Cited on p. 50.

Guyon, I.; Weston, J.; Barnhill, S.; and Vapnik, V. 2002. *Gene selection for cancer classification using support vector machines*. In Machine learning, vol. 46, no. 1, pp. 389–422. Cited on p. 28.

Harrison, H.; Farnie, G.; Howell, S.J.; Rock, R.E.; Stylianou, S.; Brennan, K.R.; Bundred, N.J.; and Clarke, R.B. 2010. *Regulation of Breast Cancer Stem Cell Activity by Signaling through the Notch4 Receptor*. In Cancer Research, vol. 70, no. 2, p. 709. Cited on p. 49.

Hastie, T.; Tibshirani, R.; and Friedman, JH. dec 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer New York. Cited on p. 10.

Herold, Daniela. 2007. *Nonlinear Machine Learning Approaches To Analyze Microarray Data*. Master's thesis, Universität Regensburg. Cited on p. 9.

Ishii, H.; Inageta, T.; Mimori, K.; Saito, T.; Sasaki, H.; Isobe, M.; Mori, M.; Croce, C.M.; Huebner, K.; Ozawa, K.; et al. 2005. *Frag1, a homolog of alternative replication factor C subunits, links replication stress surveillance with apoptosis*. In Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 27, p. 9655. Cited on p. 46.

Kalnina, Z.; Silina, K.; Bruvere, R.; Gabruseva, N.; Stengrevics, A.; Barnikol-Watanabe, S.; Leja, M.; and Line, A. 2009. *Molecular characterisation and expression analysis of SEREX-defined antigen NUCB2 in gastric epithelium, gastritis and gastric cancer*. In European Journal of Histochemistry, vol. 53, no. 1, p. 7. Cited on p. 46.

Kim, Sahn-ho. 2008. *Telomere dysfunction and cell survival: Roles for distinct TIN2-containing complexes*. In Lawrence Berkeley National Laboratory. Cited on p. 50.

Lisa, J.I.; Chu, L.; Devries, Y.; Matsumura, K.; Chew, K.; Ljung, B.M.; and Waldman, F.M. 1999. *Genetic Alterations in ERBB2-amplified Breast Carcinomas*. In . Cited on p. 45.

Lu, K.H.; Patterson, A.P.; Wang, L.; Marquez, R.T.; Atkinson, E.N.; Baggerly, K.A.; Ramoth, L.R.; Rosen, D.G.; Liu, J.; Hellstrom, I.; et al. 2004. *Selection of potential markers for epithelial ovarian cancer with*

gene expression arrays and recursive descent partition analysis. In Clinical Cancer Research, vol. 10, no. 10, p. 3291. Cited on p. 46.

Lucs, AV; Muller, WJ; and Muthuswamy, SK. 2009. *Shc is required for ErbB2-induced inhibition of apoptosis but is dispensable for cell proliferation and disruption of cell polarity*. In Oncogene, vol. 29, no. 2, pp. 174–187. Cited on p. 50.

Mangasarian, O. L. and Musicant, David R. June 2000a. *Lagrangian Support Vector Machine Classification*. Tech. Rep. 00-06, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin. Ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-06.ps. Cited on p. 17.

Mangasarian, O.L. and Musicant, D. R. 2000b. *LSVM Software: Active Set Support Vector Machine Classification Software*. Www.cs.wisc.edu/dmi/lsvm/. Cited on pp. 17 and 18.

Marone, R.; Hess, D.; Dankort, D.; Muller, W.J.; Hynes, N.E.; and Badache, A. 2004. *Memo mediates ErbB2-driven cell motility*. In Nature cell biology, vol. 6, no. 6, pp. 515–522. Cited on p. 50.

Mathura, Venkatarajan Subramanian and Kangueane, Pandjassarame. 2009. *Bioinformatics: A Concept-Based Introduction*. Springer. Cited on p. 10.

Mehta, Jai Prakash. 2009. *Gene expression analysis in breast cancer*. Ph.D. thesis, Dublin City University. School of Biotechnology. Cited on p. 46.

Miller, L.D.; Smeds, J.; George, J.; Vega, V.B.; Vergara, L.; Ploner, A.; Pawitan, Y.; Hall, P.; Klaar, S.; Liu, E.T.; et al. 2005. *An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival*. In Proceedings of the National Academy of Sciences, vol. 102, no. 38, p. 13550. Cited on p. 23.

Okafur, N. 2007. *Modern industrial microbiology and biotechnology*. Science Publishers, Enfield,(NH). Cited on pp. 4 and 5.

Park, I.H. and Li, C. 2010. *Dynamic Ligand-Induced-Fit Simulation via Enhanced Conformational Samplings and Ensemble Dockings: A Survivin Example*. In The Journal of Physical Chemistry B, pp. 183–189. Cited on p. 49.

Pawitan, Y.; Bjöhle, J.; Amler, L.; Borg, A.L.; Egyhazi, S.; Hall, P.; Han, X.; Holmberg, L.; Huang, F.; Klaar, S.; et al. 2005. *Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts*. In Breast Cancer Research, vol. 7, no. 6, pp. R953–R964. Cited on p. 23.

Pearson, H. 2006. *Genetics: what is a gene?* In Nature, vol. 441, pp. 398–401. Cited on pp. 4 and 5.

Reynolds, C.; Montone, K.T.; Powell, C.M.; Tomaszewski, J.E.; and Clevenger, C.V. 1997. *Expression of prolactin and its receptor in human breast carcinoma*. In Endocrinology, vol. 138, no. 12, p. 5555. Cited on p. 50.

Rudner, A.D. and Murray, A.W. 1996. *The spindle assembly checkpoint*. In Current opinion in cell biology, vol. 8, no. 6, pp. 773–780. Cited on p. 49.

Ruepp, A.; Brauner, B.; Dunger-Kaltenbach, I.; Frishman, G.; Montrone, C.; Stransky, M.; Waegele, B.; Schmidt, T.; Doudieu, O.N.; Stumpflen, V.; et al. 2007. *CORUM: the comprehensive resource of mammalian protein complexes*. In Nucleic Acids Research. Cited on pp. 6 and 7.

Ruepp, A.; Waegele, B.; Lechner, M.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; and Mewes, H.W. 2010. *CORUM: the comprehensive resource of mammalian protein complexes–2009*. In Nucleic Acids Research, vol. 38, no. Database issue, pp. D497–D501. Cited on pp. 6 and 7.

Scherf, M.; Epple, A.; and Werner, T. 2005. *The next generation of literature analysis: integration of genomic analysis into text mining*. In Briefings in bioinformatics, vol. 6, no. 3, p. 287. Cited on p. 46.

Schölkopf, B. and Smola, A.J. 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. the MIT Press. Cited on p. 14.

Scorilas, A. 2002. *Polyadenylate Polymerase (PAP) and 3' End pre-mRNA Processing: Function, Assays, and Association with Disease*. In Critical reviews in clinical laboratory sciences, vol. 39, no. 3, pp. 193–224. Cited on p. 49.

Shen, Wei-Fang; Rozenfeld, Sophia; Kwong, Angela; Kömüves, Laszlo G.; Lawrence, H. Jeffrey; and Largman, Corey. 1999. *HOXA9 Forms Triple Complexes with PBX2 and MEIS1 in Myeloid Cells*. In Mol. Cell. Biol, vol. 19, no. 4, pp. 3051–3061. Cited on p. 49.

Sitwala, K.V.; Dandekar, M.N.; and Hess, J.L. 2008. *HOX proteins and leukemia*. In International Journal of Clinical and Experimental Pathology, vol. 1, no. 6, p. 461. Cited on p. 49.

Sørlie, T.; Perou, C.M.; Tibshirani, R.; Aas, T.; Geisler, S.; Johnsen, H.; Hastie, T.; Eisen, M.B.; Van De Rijn, M.; Jeffrey, S.S.; et al. 2001. *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*. In Proceedings of the National Academy of Sciences of the United States of America, vol. 98, no. 19, p. 10869. Cited on pp. 22 and 24.

Stekel, D. 2003. *Microarray bioinformatics*. Cambridge Univ Press. Cited on pp. 7, 8, and 9.

Stylianou, S.; Clarke, R.B.; and Brennan, K. 2006. *Aberrant activation of notch signaling in human breast cancer*. In Cancer research, vol. 66, no. 3, p. 1517. Cited on p. 49.

Tanaka, Kozo and Hirota, Toru. 2009. *Chromosome segregation machinery and cancer*. In Cancer Science, vol. 100, no. 7, pp. 1158–1165. Cited on p. 49.

Thorsteinsdottir, U.; Kroon, E.; Jerome, L.; Blasi, F.; and Sauvageau, G. 2001. *Defining roles for HOX and MEIS1 genes in induction of acute myeloid leukemia*. In Molecular and Cellular Biology, vol. 21, no. 1, p. 224. Cited on p. 49.

Vader, G.; Cruijsen, C.W.A.; Van Harn, T.; Vromans, M.J.M.; Medema, R.H.; and Lens, S. 2007. *The chromosomal passenger complex controls spindle checkpoint function independent from its role in correcting microtubule kinetochore interactions*. In Molecular biology of the cell, vol. 18, no. 11, p. 4553. Cited on p. 49.

Wennbo, H.; Gebre-Medhin, M.; Gritli-Linde, A.; Ohlsson, C.; Isaksson, O.G.P.; and Tornell, J. 1997. *Activation of the prolactin receptor but not the growth hormone receptor is important for induction of mammary tumors in transgenic mice*. In Journal of Clinical Investigation, vol. 100, no. 11, pp. 2744–2751. Cited on p. 50.

Wu, L. and Griffin, J.D. 2004. *Modulation of Notch signaling by mastermind-like (MAML) transcriptional co-activators and their involvement in tumorigenesis*. In Seminars in cancer biology, vol. 14, pp. 348–356. Elsevier. Cited on p. 49.

Zhang, Y.; Qian, Y.; Lu, W.; and Chen, X. 2009. *The G Protein-Coupled Receptor 87 Is Necessary for p53-Dependent Cell Survival in Response to Genotoxic Stress*. In Cancer research, vol. 69, no. 15, p. 6049. Cited on p. 46.

Zhou, X. and Tuck, D.P. 2007. *MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data*. In Bioinformatics, vol. 23, no. 9, p. 1106. Cited on p. 28.

APPENDICES

ANNOTATED TABLES

The following tables list the last 50 Gene-Symbols, protein accession numbers and protein-complexes resulting from the training of the multi-class support vector machine with recursive feature elimination on the Uppsala microarray breast-cancer expressions. The Gene-Symbols, protein accession numbers and protein-complexes have been listed together with their respective name and additional annotation data. They have been sorted in descending order according to their elimination during the SVM-RFE.

A.1 GENE-SYMBOLS

The following table lists the Gene-Symbols with their Gene Ontology annotation. The data has been sourced from the Affymetrix Information file.

		Gene Ontology (GO)		
Gene-Symbol	Title	Cellular Component	Biological Process	Molecular Function
TMEM34	transmembrane protein 34	membrane, integral to membrane		
TAS2R5	taste receptor, type 2, member 5	membrane, integral to membrane, integral to membrane	signal transduction, signal transduction, G-protein coupled receptor protein signaling pathway, chemosensory behavior, response to stimulus, sensory perception of taste, sensory perception of taste	signal transducer activity, receptor activity, G-protein coupled receptor activity, taste receptor activity
TMEM91	transmembrane protein 91	membrane, integral to membrane		
HN1	hematological and neurological expressed 1	nucleus		

ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	extracellular region, cytoplasm, plasma membrane, membrane, integral to membrane, integral to membrane, apical plasma membrane	electron transport, protein amino acid phosphorylation, protein amino acid phosphorylation, signal transduction, cell surface receptor linked signal transduction, enzyme linked receptor protein signaling pathway, transmembrane receptor protein tyrosine kinase signaling pathway, transmembrane receptor protein tyrosine kinase signaling pathway, transmembrane receptor protein tyrosine kinase signaling pathway, nervous system development, nervous system development, peripheral nervous system development, heart development, heart development, cell proliferation, mammary gland development, myelination, positive regulation of MAPK activity, regulation of angiogenesis, phosphoinositide-mediated signaling, positive regulation of epithelial cell proliferation	nucleotide binding, protein kinase activity, protein-tyrosine kinase activity, transmembrane receptor protein tyrosine kinase activity, transmembrane receptor protein tyrosine kinase activity, non-membrane spanning protein tyrosine kinase activity, receptor signaling protein tyrosine kinase activity, receptor activity, epidermal growth factor receptor activity, calcium ion binding, protein binding, protein binding, ATP binding, electron carrier activity, kinase activity, transferase activity, identical protein binding, ErbB-3 class receptor binding, protein heterodimerization activity, protein heterodimerization activity, protein dimerization activity, iron-sulfur cluster binding
COL17A1	collagen, type XVII, alpha 1	proteinaceous extracellular matrix, cytoplasm, integral to plasma membrane, intercellular junction, membrane, integral to membrane	phosphate transport, cell-matrix adhesion, epidermis development	structural molecule activity

GPR87	G protein-coupled receptor 87	membrane, integral to membrane, integral to membrane	signal transduction, G-protein coupled receptor protein signaling pathway	rhodopsin-like receptor activity, signal transducer activity, receptor activity, G-protein coupled receptor activity, purinergic nucleotide receptor activity, G-protein coupled
TBC1D7	TBC1 domain family, member 7	intracellular	regulation of Rab GTPase activity	GTPase activator activity, Rab GTPase activator activity
CHI3L1	chitinase 3-like 1 (cartilage glycoprotein-39)	proteinaceous extracellular matrix, extracellular space	carbohydrate metabolic process, chitin catabolic process	catalytic activity, hydrolase activity, hydrolyzing O-glycosyl compounds, chitinase activity, extracellular matrix structural constituent, sugar binding, cation binding
ECHDC2	enoyl Coenzyme A hydratase domain containing 2		metabolic process	catalytic activity, 3-hydroxybutyryl-CoA dehydratase activity
ANKRD5	ankyrin repeat domain 5			calcium ion binding
CYP20A1	cytochrome P450, family 20, subfamily A, polypeptide 1	membrane	electron transport	monooxygenase activity, iron ion binding, oxidoreductase activity, heme binding, metal ion binding
LRP2	low density lipoprotein-related protein 2	lysosome, endosome, brush border, coated pit, membrane, integral to membrane	protein amino acid glycosylation, lipid metabolic process, vitamin metabolic process, endocytosis, endocytosis, receptor-mediated endocytosis, receptor-mediated endocytosis	receptor activity, calcium ion binding, protein binding
SP8	Sp8 transcription factor	intracellular, nucleus		nucleic acid binding, DNA binding, zinc ion binding, metal ion binding

EXOSC3	exosome component 3	nuclear exosome (RNase complex), cytoplasmic exosome (RNase complex), exosome (RNase complex), nucleus	rRNA processing, rRNA processing	3'-5'-exoribonuclease activity, RNA binding, nuclease activity, exonuclease activity, protein binding, hydrolase activity
PTN	pleiotrophin (heparin binding growth factor 8, neurite growth-promoting factor 1)	extracellular space, endoplasmic reticulum	regulation of progression through cell cycle, ossification, transmembrane receptor protein tyrosine phosphatase signaling pathway, nervous system development, learning, cell proliferation, positive regulation of cell proliferation, bone mineralization	protein phosphatase inhibitor activity, cytokine activity, growth factor activity, heparin binding
ZNF780B	zinc finger protein 780B	intracellular, nucleus	transcription, regulation of transcription, DNA-dependent	nucleic acid binding, DNA binding, zinc ion binding, metal ion binding
GRTP1	growth hormone regulated TBC protein 1	intracellular	regulation of Rab GTPase activity	GTPase activator activity, Rab GTPase activator activity
RESP2				
SQLE	Squalene epoxidase	microsome, membrane, integral to membrane	electron transport, metabolic process, sterol biosynthetic process	squalene monooxygenase activity, squalene monooxygenase activity, oxidoreductase activity
GSTM2	glutathione S-transferase M2 (muscle)	cytoplasm	metabolic process	glutathione transferase activity, glutathione transferase activity, transferase activity

NUCB2	nucleobindin 2	extracellular space, nucleus, nuclear outer membrane, cytoplasm, endoplasmic reticulum, ER-Golgi intermediate compartment, cytosol, plasma membrane, membrane		DNA binding, DNA binding, calcium ion binding, calcium ion binding
FRAG1	FGF receptor activating protein 1	integral to membrane		receptor activity
MPP7	membrane protein, palmitoylated 7 (MAGUK p55 subfamily member 7)			protein binding
MTF1	metal-regulatory transcription factor 1	intracellular, nucleus, nucleus	transcription, regulation of transcription, DNA-dependent, regulation of transcription from RNA polymerase II promoter, response to metal ion	nucleic acid binding, DNA binding, DNA binding, transcription factor activity, transcription coactivator activity, zinc ion binding, metal ion binding
KLHL20	kelch-like 20 (Drosophila)	cell surface, actin cytoskeleton	cytoskeleton organization and biogenesis	actin binding, protein binding
PPM1J	protein phosphatase 1J (PP2C domain containing)			catalytic activity, phosphoprotein phosphatase activity, protein phosphatase type 2C activity, hydrolase activity
SCUBE2	signal peptide, CUB domain, EGF-like 2			calcium ion binding

FAM134B	family with sequence similarity 134, member B	membrane, integral to membrane		
CNTD2	cyclin N-terminal domain containing 2		regulation of progression through cell cycle	
KRT14	keratin 14 (epidermolysis bullosa simplex, Dowling-Meara, Koebner)	intermediate filament, intermediate filament, keratin filament	epidermis development	structural molecule activity, structural constituent of cytoskeleton, protein binding, structural constituent of epidermis
ELOVL6	ELOVL family member 6, elongation of long chain fatty acids (FEN1/Elo2, SUR4/Elo3-like, yeast)	mitochondrion, endoplasmic reticulum, membrane, integral to membrane	fatty acid biosynthetic process, lipid biosynthetic process	
NOL5A	nucleolar protein 5A (56kDa with KKE/D repeat)	nucleus, nucleolus	rRNA processing	RNA binding
DACH1	dachshund homolog 1 (Drosophila)	nucleus	transcription, regulation of transcription, DNA-dependent, multicellular organismal development	DNA binding, protein binding
CAPN10	calpain 10	intracellular, cytoplasm	proteolysis, proteolysis	cysteine-type endopeptidase activity, calpain activity, calpain activity, peptidase activity, cysteine-type peptidase activity, hydrolase activity
KRT223P	keratin 223 pseudogene			

MSI2	musashi homolog 2 (Drosophila)			nucleotide binding, nucleic acid binding, RNA binding
HMGB3	high-mobility group box 3	chromatin, nucleus, chromosome	regulation of transcription, DNA-dependent, multicellular organismal development	DNA binding, DNA bending activity
SAMD5	sterile alpha motif domain containing 5			
ZNF684	zinc finger protein 684	intracellular, nucleus	transcription, regulation of transcription, DNA-dependent	nucleic acid binding, DNA binding, zinc ion binding, metal ion binding
ESR1	estrogen receptor 1	nucleus, nucleus, membrane, membrane, chromatin remodeling complex, chromatin remodeling complex	transcription, transcription, DNA-dependent, regulation of transcription, DNA-dependent, regulation of transcription, DNA-dependent, regulation of transcription, DNA-dependent, signal transduction, cell growth, estrogen receptor signaling pathway, estrogen receptor signaling pathway, negative regulation of mitosis	DNA binding, transcription factor activity, steroid hormone receptor activity, steroid hormone receptor activity, steroid hormone receptor activity, receptor activity, ligand-dependent nuclear receptor activity, steroid binding, protein binding, zinc ion binding, lipid binding, nitric-oxide synthase regulator activity, nitric-oxide synthase regulator activity, estrogen receptor activity, estrogen receptor activity, sequence-specific DNA binding, metal ion binding
RDHE2	epidermal retinal dehydrogenase 2		metabolic process	oxidoreductase activity
DEFB1	defensin, beta 1	extracellular region	chemotaxis, defense response, immune response, G-protein coupled receptor protein signaling pathway, response to bacterium, response to bacterium, defense response to bacterium, innate immune response, innate immune response	

STARD3	START domain containing 3	cytoplasm, membrane, integral to membrane	lipid metabolic process, steroid biosynthetic process, C21-steroid hormone biosynthetic process, transport, mitochondrial transport, lipid transport, steroid metabolic process, cholesterol metabolic process	lipid binding, cholesterol binding, cholesterol binding, cholesterol transporter activity
CBX2	chromobox homolog 2 (Pc class homolog, Drosophila)	chromatin, nucleus, nucleus	chromatin assembly or disassembly, transcription, regulation of transcription, DNA-dependent, negative regulation of transcription, chromatin modification, chromatin modification	DNA binding, chromatin binding, transcriptional repressor activity
MAPT	microtubule-associated protein tau	cytosol, cytosol, cytoskeleton, microtubule, microtubule associated complex, microtubule associated complex, plasma membrane, plasma membrane, plasma membrane, axon, axon, growth cone, growth cone, tubulin complex, tubulin complex	microtubule cytoskeleton organization and biogenesis, microtubule cytoskeleton organization and biogenesis, negative regulation of microtubule depolymerization, regulation of microtubule polymerization, positive regulation of microtubule polymerization, positive regulation of microtubule polymerization, positive regulation of axon extension, positive regulation of axon extension, generation of neurons, generation of neurons	structural constituent of cytoskeleton, structural constituent of cytoskeleton, protein binding, microtubule binding, microtubule binding, lipoprotein binding, SH3 domain binding, enzyme binding
RABEP1	rabaptin, RAB GTPase binding effector protein 1	early endosome	transport, endocytosis, endocytosis, apoptosis, membrane fusion, protein transport	GTPase activator activity, protein binding, growth factor activity

SOX10	SRY (sex determining region Y)-box 10	nucleus	transcription, regulation of transcription, DNA-dependent, regulation of transcription from RNA polymerase II promoter, sensory perception of sound, anatomical structure morphogenesis, cell differentiation, positive regulation of transcription from RNA polymerase II promoter, cell maturation	DNA binding, chromatin binding, RNA polymerase II transcription factor activity, RNA polymerase II transcription factor activity, enhancer binding, transcription coactivator activity, protein binding
TCTA	T-cell leukemia translocation altered gene			
YWHAZ	tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, zeta polypeptide	nucleus, cytoplasm, mitochondrion	protein targeting, anti-apoptosis, signal transduction	monooxygenase activity, protein binding, transcription factor binding, protein domain specific binding

A.2 PROTEINS

The following table lists the Protein Accession Numbers with the corresponding protein and Gene-Symbol(s). The Gene Ontology annotation has been sourced from the UniProt website (http://www.uniprot.org/). Protein Accession Numbers in parentheses indicate that the old accession number redirects to the one in parentheses. If the protein's name is listed as (deleted) it does not exist in the UniProt database anymore.

			Gene Ontology (GO)		
Accession	Protein	Genes	Cellular Component	Biological Process	Molecular Function
O00148	ATP-dependent RNA helicase DDX39	RPL35A, DDX39	nucleus	nuclear mRNA splicing, via spliceosome, mRNA export from nucleus	ATP binding, ATP-dependent helicase activity, nucleic acid binding, protein binding
O35551	Rab GTPase-binding effector protein 1	Rabep1	early endosome, recycling endosome	apoptosis, endocytosis, protein transport	growth factor activity, GTPase activator activity
P05090	Apolipoprotein D	APOD	extracellular space	lipid metabolic process, transport	lipid transporter activity, protein binding, retinoid binding
P21246	Pleiotrophin	PTN	endoplasmic reticulum, extracellular space	nervous system development, positive regulation of cell division, positive regulation of cell proliferation, transmembrane receptor protein tyrosine phosphatase signaling pathway	growth factor activity, heparin binding, phosphoprotein phosphatase inhibitor activity
P57738	T-cell leukemia translocation-altered gene protein	TCTA			
P60022	Beta-defensin 1	DEFB1	extracellular region	chemotaxis, defense response to bacterium, G-protein coupled receptor protein signaling pathway, innate immune response	
Q0D2I8	Glutathione S-transferase mu 2 (Muscle)	GSTM2		metabolic process	glutathione transferase activity
Q13751	Laminin subunit beta-3	LAMB3	basement membrane	cell adhesion, epidermis development, hemidesmosome assembly	structural molecule activity
Q14508	WAP four-disulfide core domain protein 2	WFDC2	extracellular space	proteolysis, spermatogenesis	serine-type endopeptidase inhibitor activity

Q14781	Chromobox protein homolog 2	CBX2	chromatin	cellular transcription, chromatin assembly or disassembly, chromatin modification	DNA binding, protein binding, transcription repressor activity
Q15276	Rab GTPase-binding effector protein 1	RABEP1	early endosome, recycling endosome	apoptosis, cellular membrane fusion, endocytosis, protein transport	growth factor activity, GTPase activator activity, protein homodimerization activity
Q49A35	(deleted)	LOC374491			
Q53G41	Tripartite motif protein TRIM29 isoform alpha variant	TRIM29			zinc ion binding
Q59EN9	Steroidogenic acute regulatory protein related variant	STARD3		steroid biosynthetic process	cholesterol binding, cholesterol transporter activity
Q59F11	Solute carrier family 11 (Proton-coupled divalent metal ion transporters), member 1 variant	SLC11A1	membrane	transport	transporter activity
Q5T9H0	Inhibitor of growth family, member 1	ING1			protein binding, zinc ion binding
Q5TC63	Growth hormone-regulated TBC protein 1	GRTP1	intracellular	regulation of Rab GTPase activity	Rab GTPase activator activity
Q5U0B0 (P21246)	Pleiotrophin	PTN	endoplasmic reticulum, extracellular space	nervous system development, positive regulation of cell division, positive regulation of cell proliferation, transmembrane receptor protein tyrosine phosphatase signaling pathway	growth factor activity, heparin binding, phosphoprotein phosphatase inhibitor activity
Q6FGL5	LCN2 protein	LCN2		transport	binding, transporter activity
Q6I9U4 (P57738)	T-cell leukemia translocation-altered gene protein	TCTA			

Q6PJW8	Consortin	C1orf71	integral to membrane, plasma membrane, protein complex, secretory granule, trans-Golgi network, transport vesicle	positive regulation of Golgi to plasma membrane protein transport	connexin binding
Q6ZRF6	(deleted)	MSI2			
Q6ZT07	TBC1 domain family member 9	TBC1D9	intracellular	regulation of Rab GTPase activity	calcium ion binding, Rab GTPase activator activity
Q6ZVB0	cDNA FLJ42812 fis, clone BRCAN2011602	FXYD1			
Q7Z5C1	Glycoprotein receptor gp330/megalin	LRP2			calcium ion binding, receptor activity
Q7Z5L7	Podocan	PODN	proteinaceous extracellular matrix		protein binding
Q7Z5Q5	DNA polymerase nu	POLN, C4orf15	nucleus	cellular DNA replication, DNA repair	DNA binding, DNA-directed DNA polymerase activity
Q7Z684	Putative uncharacterized protein DKFZp779G118	HMGB3	nucleus		DNA binding
Q8IY28 (Q5T2T1)	MAGUK p55 subfamily member 7	MPP7	adherens junction, tight junction		protein binding
Q8N589	PMAIP1 protein	PMAIP1			
Q8NFH8	RalBP1-associated Eps domain-containing protein 2	REPS2	cytoplasm	epidermal growth factor receptor signaling pathway, protein complex assembly	calcium ion binding, protein binding
Q96C34	RUN domain-containing protein 1	RUNDC1			

Q96LX1 (Q8N3Y7)	Epidermal retinol dehydrogenase 2	RDHE2	integral to membrane, integral to membrane of membrane fraction, endoplasmic reticulum membrane	oxidation reduction, retinal metabolic process, retinol metabolic process, detection of light stimulus involved in visual perception, keratinocyte proliferation	binding, retinol dehydrogenase activity
Q99712	ATP-sensitive inward rectifier potassium channel 15	KCNJ15	cytoplasm, focal adhesion, integral to plasma membrane	potassium ion transport	inward rectifier potassium channel activity, potassium ion binding
Q9BVG4	UPF0368 protein Cxorf26	CXorf26			
Q9BX40	Protein LSM14 homolog B	LSM14B			
Q9BXG2	P25	—			
Q9GZU0	Uncharacterized protein C6orf62	C6orf62	intracellular		
Q9H5J4	Elongation of very long chain fatty acids protein 6	ELOVL6	mitochondrion		
Q9H8F1 (Q9C091)	GREB1-like protein	KIAA1772	integral to membrane		
Q9HC96	Calpain-10	CAPN10	cytosol, plasma membrane	actin cytoskeleton reorganization, cellular response to insulin stimulus, positive regulation of apoptosis, positive regulation of glucose import, positive regulation of insulin secretion, positive regulation of intracellular transport, proteolysis	calcium-dependent cysteine-type endopeptidase activity, cytoskeletal protein binding, SNARE binding
Q9JIV6	P-cadherin	Cdh3	cytoplasm	homophilic cell adhesion, wound healing, response to drug	calcium ion binding
Q9NQF3	Serine hydrolase-like protein	SERHL2, SERHL			hydrolase activity
Q9NR86 (Q8N6W0)	CUG-BP- and ETR-3-like factor 5	BRUNOL5	cytoplasm, nucleus	mRNA processing	nucleotide binding, RNA binding

Q9NVA4	Transmembrane protein 184C	TMEM34	integral to membrane		
Q9NYW4	Taste receptor type 2 member 5	TAS2R5	integral to membrane	chemosensory behavior, G-protein coupled receptor protein signaling pathway, sensory perception of taste, signal transduction	taste receptor activity
Q9UI36	Dachshund homolog 1	DACH1		cellular transcription, multicellular organismal development, regulation of cellular transcription	nucleotide binding, protein binding
Q9UK76	Hematological and neurological expressed 1 protein	HN1	nucleus		
Q9UK79	Herstatin	ERBB2	extracellular region, membrane	transmembrane receptor protein tyrosine kinase signaling pathway, protein amino acid phosphorylation	transmembrane receptor protein tyrosine kinase activity, ATP binding, non-membrane spanning protein tyrosine kinase activity, protein binding
Q9UNR6	Squalene epoxidase	SQLE	integral to membrane	oxidation reduction	squalene monooxygenase activity, FAD binding

A.3 PROTEIN-COMPLEXES

The following table lists the protein-complexes with their Functional Catalogue annotation. The data has been sourced from the Protein-Complexes FunCat annotation and the MIPS Functional Catalogue (http://mips.helmholtz-muenchen.de/proj/funcatDB/).

Complexname	Functional Catalogue Entries
SCF subcomplex (WEE1, SKP2, BTRC)	M phase (CELL CYCLE AND DNA PROCESSING → cell cycle → mitotic cell cycle and cell cycle control → mitotic cell cycle) modification by phosphorylation, dephosphorylation, autophosphorylation (PROTEIN FATE (folding, modification, destination) → protein modification) modification by ubiquitination, deubiquitination (PROTEIN FATE (folding, modification, destination) → protein modification) protein/peptide degradation (PROTEIN FATE (folding, modification, destination))
PGC-1-SRp40-SRp55-SRp75 complex	transcriptional control (TRANSCRIPTION → RNA synthesis → mRNA synthesis) mRNA processing (splicing, 5'-, 3'-end processing) (TRANSCRIPTION → RNA processing) nucleic acid binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)) nucleus (SUBCELLULAR LOCALIZATION)
MAML3-RBP-Jkappa-Notch4 complex	transcriptional control (TRANSCRIPTION → RNA synthesis → mRNA synthesis) DNA binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic) → nucleic acid binding) Notch-receptor signalling pathway (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → transmembrane signal transduction → non-enzymatic receptor mediated signalling) nucleus (SUBCELLULAR LOCALIZATION)
Polyadenylation complex (CSTF1, CSTF2, CSTF3, SYMPK CPSF1, CPSF2, CPSF3)	3'-end processing (TRANSCRIPTION → RNA processing → mRNA processing (splicing, 5'-, 3'-end processing)) RNA binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic) → nucleic acid binding) nucleus (SUBCELLULAR LOCALIZATION)
AR-AKT-APPL complex	ligand-dependent nuclear receptors (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → cellular signalling)
Rab5 GDP/GTP exchange factor complex	endocytosis (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transport routes → cellular import → vesicular cellular import)
PAR-3-PKCz-Tiam2 complex	directional cell growth (morphogenesis) (CELL FATE → cell growth / morphogenesis) neurogenesis (DEVELOPMENT (Systemic) → animal development) asymmetries and axis determination (DEVELOPMENT (Systemic) → animal development) neuron (CELL TYPE DIFFERENTIATION → animal cell type differentiation)

	neuron (CELL TYPE LOCALIZATION → animal cell type)
	brain (ORGAN LOCALIZATION → animal organ → nervous system → central nervous system)
Eps15-stonin2 complex	protein targeting, sorting and translocation (PROTEIN FATE (folding, modification, destination))
	vesicular transport (Golgi network, etc.) (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transport routes)
HOXA9-PBX2-MEIS1 complex	DNA binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic) → nucleic acid binding)
	blood cell (CELL TYPE DIFFERENTIATION → animal cell type differentiation)
	nucleus (SUBCELLULAR LOCALIZATION)
IL12B-IL12RB1-IL12RB2 complex	T-cell (CELL TYPE DIFFERENTIATION → animal cell type differentiation → blood cell → leucocyte → lymphocyte)
Phosphatidylinositol 3-kinase complex (PIK3CA, PIK3R1)	protein binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic))
	ATP binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic) → nucleotide/nucleoside/nucleobase binding)
	transmembrane receptor protein tyrosine kinase signalling pathways (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → transmembrane signal transduction → receptor enzyme mediated signalling)
	eukaryotic plasma membrane / membrane attached (SUBCELLULAR LOCALIZATION)
Transcription elongation factor complex (SUPT5H, CDK9, CCNT1)	phosphate metabolism (METABOLISM)
	transcription elongation (TRANSCRIPTION → RNA synthesis → mRNA synthesis → general transcription activities)
	transcription activation (TRANSCRIPTION → RNA synthesis → mRNA synthesis → transcriptional control)
	regulation by modification (REGULATION OF METABOLISM AND PROTEIN FUNCTION → regulation by)
	enzyme activator (REGULATION OF METABOLISM AND PROTEIN FUNCTION → regulation of protein activity → enzymatic activity regulation / enzyme regulator)
	nucleus (SUBCELLULAR LOCALIZATION)
Kinase-scaffold-phosphatase complex, PKA-AKAP79-CaN	protein targeting, sorting and translocation (PROTEIN FATE (folding, modification, destination))
	modification by phosphorylation, dephosphorylation, autophosphorylation (PROTEIN FATE (folding, modification, destination) → protein modification)
	assembly of protein complexes (PROTEIN FATE (folding, modification, destination))
	transmembrane signal transduction (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM)
	eukaryotic plasma membrane / membrane attached (SUBCELLULAR LOCALIZATION)

	cytoskeleton (SUBCELLULAR LOCALIZATION)
Mi-2/NuRD-MTA2 complex	DNA conformation modification (e.g. chromatin) (CELL CYCLE AND DNA PROCESSING → DNA processing → DNA restriction or modification)
	transcription repression (TRANSCRIPTION → RNA synthesis → mRNA synthesis → transcriptional control)
	modification by acetylation, deacetylation (PROTEIN FATE (folding, modification, destination) → protein modification)
	organization of chromosome structure (BIOGENESIS OF CELLULAR COMPONENTS → nucleus)
	B-cell (CELL TYPE DIFFERENTIATION → animal cell type differentiation → blood cell → leucocyte → lymphocyte)
	nucleus (SUBCELLULAR LOCALIZATION)
	lymphatic system (ORGAN LOCALIZATION → animal organ → immune system organs)
SMG-1-Upf1-eRF1-eRF3 complex (SURF)	control of mRNA stability (TRANSCRIPTION → RNA processing → mRNA processing (splicing, 5'-, 3'-end processing))
	modification by phosphorylation, dephosphorylation, autophosphorylation (PROTEIN FATE (folding, modification, destination) → protein modification)
	RNA transport (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transported compounds (substrates))
	nuclear transport (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transport routes)
	cytoplasm (SUBCELLULAR LOCALIZATION)
	nucleus (SUBCELLULAR LOCALIZATION)
GluR1-GluR2 heteromer complex	cation transport (H+, Na+, K+, Ca2+ , NH4+, etc.) (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transported compounds (substrates) → ion transport)
	ion channels (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transport facilities → channel / pore class transport)
	synaptic transmission (INTERACTION WITH THE ENVIRONMENT → membrane excitability)
	learning and memory (SYSTEMIC INTERACTION WITH THE ENVIRONMENT → animal specific systemic sensing and response → nervous system physiology → neuronal integrative functions)
	neuron (CELL TYPE LOCALIZATION → animal cell type)
	brain (ORGAN LOCALIZATION → animal organ → nervous system → central nervous system)
ERBB2-MEMO-SHC complex	cell motility (INTERACTION WITH THE ENVIRONMENT)
	microtubule cytoskeleton (BIOGENESIS OF CELLULAR COMPONENTS → cytoskeleton/structural proteins)
p27-cyclinD2-Cdk4 complex	cell cycle arrest (CELL CYCLE AND DNA PROCESSING → cell cycle → mitotic cell cycle and cell cycle control)
	regulation of protein activity (REGULATION OF METABOLISM AND PROTEIN FUNCTION)

	nucleus (SUBCELLULAR LOCALIZATION)
Chromosomal passenger complex CPC (INCENP, CDCA8, BIRC5, AURKB)	mitotic cell cycle (CELL CYCLE AND DNA PROCESSING → cell cycle → mitotic cell cycle and cell cycle control) chromosome segregation/division (CELL CYCLE AND DNA PROCESSING → cell cycle → nuclear and chromosomal cycle) DNA binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic) → nucleic acid binding) centromere / kinetochore (SUBCELLULAR LOCALIZATION → nucleus)
Psd3-Actn1 complex	regulation of protein activity (REGULATION OF METABOLISM AND PROTEIN FUNCTION) neuron (CELL TYPE LOCALIZATION → animal cell type)
Prolactin (PRL) - PRL receptor (PRLR) complex	non-enzymatic receptor mediated signalling (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → transmembrane signal transduction)
Glur4-cadherin-catenin complex	nervous tissue (TISSUE DIFFERENTIATION → animal tissue) neuron (CELL TYPE LOCALIZATION → animal cell type)
hMediator complex (MED23, CDK8, CCNC, MED7)	transcription activation (TRANSCRIPTION → RNA synthesis → mRNA synthesis → transcriptional control) regulation by binding / dissociation (REGULATION OF METABOLISM AND PROTEIN FUNCTION → regulation by) regulator of transcription factor (REGULATION OF METABOLISM AND PROTEIN FUNCTION → regulation of protein activity) nucleus (SUBCELLULAR LOCALIZATION)
Rab11-Fip2-Reps1 complex	protein targeting, sorting and translocation (PROTEIN FATE (folding, modification, destination)) receptor-mediated endocytosis (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transport routes → cellular import → vesicular cellular import → endocytosis)
ITGAV-ITGB3-NOV complex	protein binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)) cell adhesion (INTERACTION WITH THE ENVIRONMENT) vessels (ORGAN LOCALIZATION → animal organ → vascular organs)
REST-CoREST-mSIN3A complex	transcription repression (TRANSCRIPTION → RNA synthesis → mRNA synthesis → transcriptional control) neurogenesis (DEVELOPMENT (Systemic) → animal development) neuron (CELL TYPE DIFFERENTIATION → animal cell type differentiation) nervous tissue (TISSUE DIFFERENTIATION → animal tissue) nucleus (SUBCELLULAR LOCALIZATION) neuron (CELL TYPE LOCALIZATION → animal cell type)
RICH1/AMOT polarity complex, Flag-Rich1 precipitated	regulation by localization (REGULATION OF METABOLISM AND PROTEIN FUNCTION → regulation by)

	GTPase activator (GAP) (REGULATION OF METABOLISM AND PROTEIN FUNCTION → regulation of protein activity → enzymatic activity regulation / enzyme regulator → enzyme activator)
	G-protein mediated signal transduction (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → cellular signalling → enzyme mediated signal transduction)
	directional cell growth (morphogenesis) (CELL FATE → cell growth / morphogenesis)
	asymmetries and axis determination (DEVELOPMENT (Systemic) → animal development)
	cell junction (BIOGENESIS OF CELLULAR COMPONENTS)
	cell junction (SUBCELLULAR LOCALIZATION)
	epithelium (TISSUE LOCALIZATION → animal tissue)
Acinar cell-specific C complex	transcription activation (TRANSCRIPTION → RNA synthesis → mRNA synthesis → transcriptional control)
	embryogenesis (DEVELOPMENT (Systemic) → animal development)
	pancreas (ORGAN DIFFERENTIATION → animal organ → gastro-intestinal system)
	nucleus (SUBCELLULAR LOCALIZATION)
	pancreas (ORGAN LOCALIZATION → animal organ → gastro-intestinal system)
ITGA2-ITGB1-COL6A3 complex	protein binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic))
	cell adhesion (INTERACTION WITH THE ENVIRONMENT)
	eukaryotic plasma membrane / membrane attached (SUBCELLULAR LOCALIZATION)
	extracellular matrix component (SUBCELLULAR LOCALIZATION → extracellular / secretion proteins)
c-Src-Muc1 complex	regulation of signal transduction (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM)
	eukaryotic plasma membrane / membrane attached (SUBCELLULAR LOCALIZATION)
Ecsit complex (Ecsit2-Smad1)	TGF-beta-receptor signalling pathway (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → transmembrane signal transduction → receptor enzyme mediated signalling → transmembrane receptor protein serine/threonine kinase signalling pathways)
	embryogenesis (DEVELOPMENT (Systemic) → animal development)
nephrin-cadherin complex (Nphs1, Ctnnd1, Cdh3, Cd2ap)	cell-cell adhesion (INTERACTION WITH THE ENVIRONMENT → cell adhesion)
	intercellular junction (gap junction/adherens junction) (SUBCELLULAR LOCALIZATION → cell junction)
	epithelium (TISSUE LOCALIZATION → animal tissue)
NMDA receptor complex (NR2A, NR2B, NR1, PSD-95)	protein targeting, sorting and translocation (PROTEIN FATE (folding, modification, destination))
	ion channels (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transport facilities → channel / pore class transport)

	small GTPase mediated signal transduction (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → cellular signalling → enzyme mediated signal transduction → G-protein mediated signal transduction)
	glutamate signalling pathway (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → transmembrane signal transduction → non-enzymatic receptor mediated signalling → G-protein coupled receptor signalling pathway)
	regulation of signal transduction (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM)
	membrane excitability (INTERACTION WITH THE ENVIRONMENT)
	eukaryotic plasma membrane / membrane attached (SUBCELLULAR LOCALIZATION)
	neuron (CELL TYPE LOCALIZATION → animal cell type)
	brain (ORGAN LOCALIZATION → animal organ → nervous system → central nervous system)
GIPC1-NTRK1-RGS19 complex	retrograde transport (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transport routes → vesicular transport (Golgi network, etc.))
SHARP-CtBP1-CtIP complex	transcription repression (TRANSCRIPTION → RNA synthesis → mRNA synthesis → transcriptional control)
	DNA binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic) → nucleic acid binding)
	Notch-receptor signalling pathway (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → transmembrane signal transduction → non-enzymatic receptor mediated signalling)
	nucleus (SUBCELLULAR LOCALIZATION)
Ric-8A G alpha 13 complex	guanyl-nucleotide exchange factor (GEF) (REGULATION OF METABOLISM AND PROTEIN FUNCTION → regulation of protein activity)
Ubiquitin E3 ligase (CRY1, SKP1A, CUL1, FBXL3)	modification by ubiquitination, deubiquitination (PROTEIN FATE (folding, modification, destination) → protein modification)
	proteasomal degradation (ubiquitin/proteasomal pathway) (PROTEIN FATE (folding, modification, destination) → protein/peptide degradation → cytoplasmic and nuclear protein degradation)
	rhythm (e.g. circadian, ultradian) (INTERACTION WITH THE ENVIRONMENT → cellular sensing and response to external stimulus)
	nucleus (SUBCELLULAR LOCALIZATION)
ER-alpha-GRIP1-c-Jun complex	transcriptional control (TRANSCRIPTION → RNA synthesis → mRNA synthesis)
	ligand-dependent nuclear receptors (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → cellular signalling)
Pparalpha-Pric320 complex	transcription activation (TRANSCRIPTION → RNA synthesis → mRNA synthesis → transcriptional control)
	nucleus (SUBCELLULAR LOCALIZATION)
Smooth muscle dystroglycan complex	smooth muscle contraction (SYSTEMIC INTERACTION WITH THE ENVIRONMENT → animal specific systemic sensing and response → muscle contraction)

	smooth muscle (TISSUE DIFFERENTIATION → animal tissue → muscle)
	eukaryotic plasma membrane / membrane attached (SUBCELLULAR LOCALIZATION)
	smooth muscle (TISSUE LOCALIZATION → animal tissue → muscle)
	lung (ORGAN LOCALIZATION → animal organ → respiratory system)
Survivin homodimer complex	M phase (CELL CYCLE AND DNA PROCESSING → cell cycle → mitotic cell cycle and cell cycle control → mitotic cell cycle)
	spindle pole body/centrosome and microtubule cycle (CELL CYCLE AND DNA PROCESSING → cell cycle → cell cycle dependent cytoskeleton reorganization)
	anti-apoptosis (CELL FATE → cell death → apoptosis (type I programmed cell death))
	microtubule cytoskeleton (BIOGENESIS OF CELLULAR COMPONENTS → cytoskeleton/structural proteins)
	microtubule cytoskeleton (SUBCELLULAR LOCALIZATION → cytoskeleton)
ActRIIA-ActRIB-Smad3-Arip1 complex	transmembrane receptor protein serine/threonine kinase signalling pathways (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → transmembrane signal transduction → receptor enzyme mediated signalling)
	neuron (CELL TYPE LOCALIZATION → animal cell type)
	nervous tissue (TISSUE LOCALIZATION → animal tissue)
	nervous system (ORGAN LOCALIZATION → animal organ)
SMAD3-VDR complex	transcriptional control (TRANSCRIPTION → RNA synthesis → mRNA synthesis)
	ligand-dependent nuclear receptors (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → cellular signalling)
	TGF-beta-receptor signalling pathway (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → transmembrane signal transduction → receptor enzyme mediated signalling → transmembrane receptor protein serine/threonine kinase signalling pathways)
	nucleus (SUBCELLULAR LOCALIZATION)
TRF1-TIN2 complex	DNA binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic) → nucleic acid binding)
	organization of chromosome structure (BIOGENESIS OF CELLULAR COMPONENTS → nucleus)
	nucleus (SUBCELLULAR LOCALIZATION)
PYR complex	DNA conformation modification (e.g. chromatin) (CELL CYCLE AND DNA PROCESSING → DNA processing → DNA restriction or modification)
	transcriptional control (TRANSCRIPTION → RNA synthesis → mRNA synthesis)
	modification by acetylation, deacetylation (PROTEIN FATE (folding, modification, destination) → protein modification)
	DNA binding (PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic) → nucleic acid binding)

	embryogenesis (DEVELOPMENT (Systemic) → animal development)
	organization of chromosome structure (BIOGENESIS OF CELLULAR COMPONENTS → nucleus)
	blood cell (CELL TYPE DIFFERENTIATION → animal cell type differentiation)
	nucleus (SUBCELLULAR LOCALIZATION)
	erythrocyte (CELL TYPE LOCALIZATION → animal cell type → blood cell)
APLG1-Rababtin5 complex	vesicular transport (Golgi network, etc.) (CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES → transport routes)
EGFR-containing signaling complex	EGF-receptor signalling pathway (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM → transmembrane signal transduction → receptor enzyme mediated signalling → transmembrane receptor protein tyrosine kinase signalling pathways)
Ubiquitin-protein ligase (UBE2N, UBE2V2/MMS2)	modification by ubiquitination, deubiquitination (PROTEIN FATE (folding, modification, destination) → protein modification)
	regulation by modification (REGULATION OF METABOLISM AND PROTEIN FUNCTION → regulation by)
	kinase activator (REGULATION OF METABOLISM AND PROTEIN FUNCTION → regulation of protein activity → enzymatic activity regulation / enzyme regulator → enzyme activator)
	regulation of signal transduction (CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM)
Nephrin-cadherin complex (Nphs1, Ctnnd1, Cdh3, Cd2ap)	cell-cell adhesion (INTERACTION WITH THE ENVIRONMENT → cell adhesion)
	intercellular junction (gap junction/adherens junction) (BIOGENESIS OF CELLULAR COMPONENTS → cell junction)
	intercellular junction (gap junction/adherens junction) (SUBCELLULAR LOCALIZATION → cell junction)
	epithelium (TISSUE LOCALIZATION → animal tissue)
	kidney (ORGAN LOCALIZATION → animal organ → excretory apparatus)
BLM-TRF2 complex	DNA topology (CELL CYCLE AND DNA PROCESSING → DNA processing)
	cell aging (CELL FATE)
	organization of chromosome structure (BIOGENESIS OF CELLULAR COMPONENTS → nucleus)