DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
10

ELENA SÜGIS

Integration Methods for Heterogeneous Biological Data

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

10

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
10

ELENA SÜGIS

Integration Methods for Heterogeneous Biological Data

UNIVERSITY OF TARTU
Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in informatics on May 13th, 2019 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor

Prof.	Jaak Vilo University of Tartu, Estonia
Dr.	Hedi Peterson University of Tartu, Estonia

Opponents

Prof.	Laura Elo University of Turku, Finland
Prof.	Ewa Szczurek University of Warsaw, Poland

The public defense will take place on June 26th, 2019 at 14:15 in Liivi 2-405. The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906
ISBN 978-9949-03-057-6 (print)
ISBN 978-9949-03-058-3 (PDF)

University of Tartu Press
http://www.tyk.ee/

"The whole is greater than the sum of its parts."
Aristotle

ABSTRACT

The explosion of *omics* technologies has led to the large volumes of experimental biological data being produced in laboratories around the world. The experiments carried out in different biological domains, such as proteomics, genomics, transcriptomics, etc., allow to study biological process or disease only from a certain aspect without capturing the system as a whole. In order to get a systematic view on disease, it is necessary to combine these heterogeneous data. This leads to a growing need for a composition of the reliable integrated disease-specific data sets that researchers could effectively work with. The large amounts of diverse data require development of novel data science - based integration and analysis methods that can help, e.g. to identify which genes are related to disease or which drugs are toxic for early human development. These methods should provide flexible and effective ways for a description of biological systems at different levels, aiming to explain how individual heterogeneous data types relate to one another and to the studied phenotype or condition.

To this moment there has been a lack of the unified data integration methodology in the field of biological sciences. In this thesis we adapt two conceptual pipelines for data integration depending on the study objective and direct relation of individual data sets to the phenotype of interest. Both multi-staged data integration and transformation-based data integration use machine learning methods as core building blocks. However, a choice of one or the other approach highly depends on the study set up. While multi-staged data integration consists of a sequential analysis of individual data sets with a consequent combination of the results, a transformation-based integration involves data transformation into an intermediate form, such as a graph. Deep learning methods have gained popularity in various domains, being applied to learn useful low dimensional representations of images, text, videos, etc. However, due to the complexity of graph-structured heterogeneous data, e.g. lack of fixed node ordering or reference point, these methods cannot be directly applied on the graph data. Graph convolutional networks (GCN) are the state-of-the-art deep learning methods specifically developed for graphs. These methods leverage both information contained in the nodes and in the relationships between the nodes.

In this dissertation, we describe scenarios of the application of data integration methods to the practical tasks in computational analysis of biological data and illustrate how data integration improves understanding of biological processes. More precisely, we demonstrate how to combine and analyze different types of biological data in the example of three biological domains: immunology, toxicology, and Alzheimer's disease. Combining patient's data related to immune disease helps to uncover its pathological mechanisms and to find better treatments in the future. We analyze laboratory data from patients' skin and blood samples by combining them with clinical information. Subsequently, we bring together the results of individual analyses using available domain knowledge to form a more

systematic view on the disease pathogenesis. Toxicity testing is the process of defining the harmful effects of the substances for the living organisms. One of its applications is a safety assessment of drugs or other chemicals for early development of a human organism. In this work, we identify groups of toxicants that have a similar mechanism of actions. Additionally, we develop a classification model that allows to assess toxic actions of unknown compounds. In the frames of this work we describe an approach for the integration of the disparate but complementary heterogeneous data sets related to Alzheimer's disease into a novel Heterogeneous Network-based Alzheimer's disease-specific data set (HENA). This data set is based on FAIR standards and aims to provide bioinformatics researchers the possibility to exploit this unique resource in the context of research related to Alzheimer's disease. Additionally, this large heterogeneous graph-structured data set provides machine learning experts a possibility to benchmark novel methods for large graphs. We then apply a novel GCN-based method for heterogeneous graphs to node classification task in HENA to find genes that are potentially associated with the disease.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AD - Alzheimer's disease
ANOVA - Analysis of variance
AO - Adverse outcome
AOP - Adverse outcome pathways
BP - Biological pathways
cDNA - Complementary DNA
CNN - Convolutional neural network
ENSG ID - Ensembl database gene identifier
FVS - Fetal valproate syndrome
GE - Gene expression
GCN - Graph convolutional network
GNN - Graph neural network
GO - Gene ontology
GWAS - Genome-wide association study
HDACi - Histone deacetylase inhibitors
IGR - Intergenic region
IGRI - Intergenic region interaction
KEGG - Kyoto encyclopedia of genes and genomes
KEs - Key events
LDA - Linear discriminant analysis
MIE - Molecular initiating event
MINC - Migration of neural crest cell
MoA - Mode of action
mRNA - Messenger ribonucleic acid
NCC - Neural crest cell
PCA - Principal component analysis
PCBs - Polychlorinated biphenyls
PPI - Protein-protein interaction
RNA - Ribonucleic acid
TSA - Trichistatin A
qRT-PCR - Quantitative real-time polymerase chain reaction
SNP - Single nucleotide polymorfism
VPA - Valproic acid
Y2H - Yeast two hybrid
3D - Three dimensional

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

I. Elena Sügis*, Jerome Dauvillier*, Anna Leontjeva, Priit Adler, Valerie Hindie, Thomas Moncion, Vincent Collura, Rachel Daudin, Yann Loe-Mie, Yann Herault, Jean-Charles Lambert, Henning Hermjakob, Tal Pupko, Jean-Christophe Rain, Ioannis Xenarios, Jaak Vilo, Michel Simonneau, Hedi Peterson. HENA - Heterogeneous network-based data set for Alzheimer's disease. (The article is suggested to be accepted with minor modifications for publication in Nature Scientific Data, April 2019)

II. Liisi Šahmatova*, Elena Sügis*, Marina Šunina*, Helen Hermann, Ele Prans, Maire Pihlap, Kristi Abram, Ana Rebane, Hedi Peterson, Pärt Peterson, Külli Kingo, Kai Kisand. Signs of innate immune activation and premature immunosenescence in psoriasis patients. Scientific Reports, vol 7, no. 1, 2017, Springer Nature.

III. Nadine Dreser, Bastian Zimmer, Christian Dietz, Elena Sügis, Giorgia Pallocc, Johanna Nyffeler, Johannes Meisig, Nils Blüthgen, Michael R. Berthold, Tanja Waldmann, Marcel Leist. Grouping of histone deacetylase inhibitors and other toxicants disturbing neural crest migration by transcriptional profiling. Neurotoxicology, vol 50, 2015, pp. 56-70. Elsevier BV.

IV. Nina V. Balmer, Matthias K. Weng, Bastian Zimmer, Violeta N. Ivanova, Stuart M.Chambers, Elena Sügis (Nikolaeva), Smita Jagtap, Agapios Sachinidis, Jürgen Hescheler, Tanja Waldmann and Marcel Leist. Epigenetic changes and disturbed neural development in a human embryonic stem cell-based model relating to the fetal valproate syndrome. Human Molecular Genetics, vol 21 (18), 2012, pp. 4104-4114. Oxford University Press.

* - authors contributed equally

Other published work of the author

I. Valery Korzhik, Elena Sügis (Nikolaeva), Identification method based on protected biometric information. Intelligent Systems: Proceedings of the Eighth International symposium, 2008, ISBN 978-5-93347-332-9

1. INTRODUCTION

A rapid advance in technology and decreasing production costs led to the increasing amounts of experimental biological data being produced every day in the laboratories across the world. Scientists carrying out these experiments work in different *omics* domains, e.g. proteomics, genomics, transcriptomics, etc. Various types of experiments allow researchers to understand parts of the functional processes and disease mechanisms from different angles. The results of such studies are deposited in repositories, such as ArrayExpress [1, 2], IntAct [3, 4], ADNI [5], etc., designed for storing data sets of specific experimental data type, e.g. gene expression, protein-protein interactions, etc. The analysis of accumulated data sets of protein-protein interactions, gene expression and medical imaging have allowed to discover valuable knowledge about various biological processes, for example, identification of biomarkers that indicate a pathogenic processes, or pharmacologic responses to a therapeutic intervention [6, 7]. However, these discoveries provide only partial understanding when considered in isolation. In order to obtain a systematic view on a biological process in an organism, e.g. a disease, these individual data sets should be combined. The possibility of accessing the knowledge from various biological domains raises the question of how to combine these large heterogeneous data sets to obtain a meaningful understanding of the biological functions of an organism or a pathological mechanism of a disease described as a whole. One of the challenging tasks is to create a reliable disease-specific data sets that researchers could effectively work with. To ensure data quality for the subsequent analysis and reuse, individual data sets should be preprocessed, rigorously filtered, confirmed by experiments using different technologies and follow FAIR standards [8]. The potential application scenarios for such data sets include, for example, combination with newly-generated data as a complementary source, or use as an independent reference data set for the assessment of novel discoveries. Although modern technological advances allow quantitative studies of biological processes in different conditions, the way how these heterogeneous data types relate to one another and to the phenotype of interest still remains not completely understood. This issue drives the progress to develop and adapt data science-based methods that aim to integrate these heterogeneous data sets into a biologically meaningful multi-level analysis pipeline serving to describe biological system or a process [9, 10].

There is no unified classification of data integration methodologies in a field of biological sciences [11–14]. The amount and diversity of experimental, computational and domain knowledge data together with study setup dictate the choice of integration pipeline. In this thesis, we have used several integration approaches and have selected two generalized ones that are partially based on the classification proposed by Ritchie et al. [15]. These approaches are based on the relation of phenotype of interest and individual available data sets and can be applied in many different biological scenarios. We divide data integration approaches into

two groups: multi-staged data integration and transformation-based data integration. Both multi-staged data integration and transformation-based data integration use data science-based methods as core building blocks to discover relations in the data that would be difficult to detected otherwise. Multi-staged data integration implies a sequential analysis of individual data sets with a subsequent combination of the results and drawing a conclusion using domain knowledge. Depending on a nature of the relation between the studies data types, various techniques can be applied to model each type of relations and detect complex patterns in the data. There is an abundance of methods for regression and classification tasks, such as linear models [16, 17] or random forest [18], various testing strategies, e.g. t-test [19] or non-parametric Wilcoxon test [20–22], clustering methods, i.e k-means [23], that are well-established for the analysis of biomolecular and clinical data. A transformation-based integration involves data transformation into an intermediate form, such as a graph. While in multi-staged analysis data sets often come in a form of a numeric matrix accompanied by a categorical meta-data, where values represent experimental measurements, in transformation-based integration data is represented as a more complex structures, e.g. heterogeneous graphs [24], where nodes and edges are described by a set of attributes. Machine learning methods have became widely used in the biological sciences due to the possibility to build predictive models without making strong assumptions about the underlying mechanisms of biological processes [25–27]. Feature selection and feature engineering steps are often required for the most effective model performance. In the field of graph analytics traditional machine learning approaches usually rely on hand-crafted features and are limited by their inflexibility and high computational costs [28–31]. In recent years, deep learning methods have gained popularity in various domains, being applied to learn useful representations of images, text, videos, etc. [25, 32, 33]. The application of such models enables automatic feature extraction. However, due to the complexity of graph-structured heterogeneous data, e.g. lack of fixed node ordering or reference point, these methods cannot be directly applied on the graph data. Graph convolutional networks (GCNs) are the state-of-the-art deep learning methods specifically developed for graphs [34, 35]. GCNs are a type of neural network architectures designed to work directly on graphs and leverage their structural information, i.e. information contained in the relationships between the nodes.

In this thesis, we demonstrate data science-based approaches for integration and analysis of heterogeneous biological data in the application to emerging practical tasks. We show how understanding of the biological processes can be improved using data integration-based approaches. A variety of biological processes, serving as a base for defining physiological state of an organism, are described by different *omics* data. A schematic representation of a thesis structure is demonstrated on Figure 1.

In Chapter 2 we provide description of *omics* data origins and possible sources of experimental, computational, and domain knowledge-based biological data

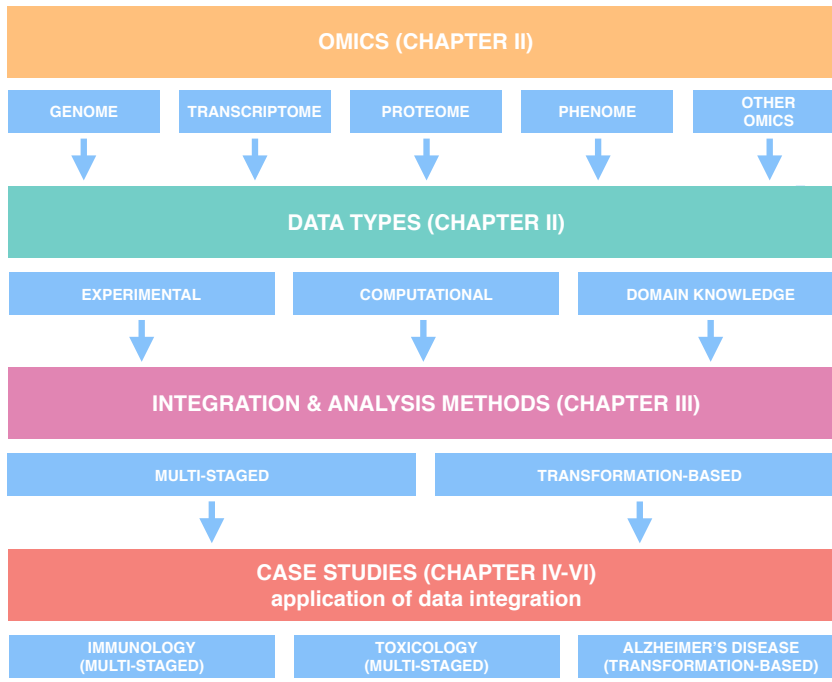


Figure 1. Schematic representation of the thesis structure.

(Figure 1). The availability of the data sets described in this section largely defines the methods for the integration and analysis. In this chapter we also give an intuition how each data type is related to one another. Given section describes data types that were used in the case studies described in Chapter 4-6. In Chapter 3 we introduce two major data integration approaches used in this thesis - transformation-based and multi-staged integration. We then provide the description of the data science methods used in integrative data analyses ranging from statistical hypothesis testing and supervised learning methods to state-of-the-art deep learning methods for graphs (Figure 1). These analysis methods are applied to the data described in Chapter 2. In the following Chapters 4 - 6 we illustrate how data integration-based analysis can improve the understanding of biological processes in application to three biological domains: Alzheimer's disease, toxicity testing and immunology (Figure 1). More specifically, in Chapter 4, based on Publication I, we introduce an approach for the integration of heterogeneous biological data sets related to Alzheimer's disease [36]. It is an age-related neurodegenerative disorder that progresses with age and eventually leads to death. Several approved drugs are used to reduce the symptoms of Alzheimer's disease, however, no current treatments can modify the underlying disease processes. We apply transformation-based integration, and describe a novel heterogeneous network-based data set for Alzheimer's disease (HENA). HENA is comprised of 64 data sets of six data types originating from nine data sources. These data types include

protein-protein interactions, gene co-expression, epistasis, genome-wide association studies (GWAS), gene expression in different brain regions, and positive selection data. HENA aims to provide researchers an opportunity to utilize this resource in studies related to Alzheimer's disease, and to allow machine learning experts to benchmark their methods using this feature-rich large graph data set. We also demonstrate an application of GCNs to node classification task in HENA to find genes that are potentially associated with the disease. In Chapter 5, based on Publication II, we demonstrate how application of multi-staged data integration concept can shed the light on the mechanism of a complex immune disease. Psoriasis is a skin condition causes cells to build up rapidly on the surface of the skin forming lesions. The mechanisms of this disease are not clearly understood. We study pathological mechanisms of the disease by combining laboratory data from patients' skin and blood samples and clinical information. Subsequently, we bring together the results of individual analyses using available domain knowledge. The Chapter 6, based on the Publications III and IV, is dedicated to the improvement of developmental toxicity testing strategies using multi-staged data integration. Toxicity testing is the process of defining harmful effects of the chemicals for the living organisms or cell lines. It serves for the examination, evaluation, and interpretation of the harmful effects of a substance. One of its applications is a safety assessment of drugs or other chemicals for early development of a human organism. Some of the substances can have similar end results while their mode of action differs. In this work, we identify groups of toxicants that have a similar mechanism of actions. Additionally, we develop a classification model that allows to assess toxic actions of unknown compounds.

The major author contributions of the current thesis based on the Publications I-IV could be summarized as follows:

1. Collection and combination of the heterogeneous data sets in the domain of Alzheimer's disease, toxicology and immunology (I-IV).
2. Integration-based data analysis applied to the domain of Alzheimer's disease, toxicology and immunology (I-IV).
3. Application of the state-of-the-art graph convolutional networks to the large heterogeneous graph-structured biological data sets (I).

2. BIOLOGICAL BACKGROUND

Homo sapiens as a species has around 10^{14} cells in its body allowing it to function as a living organism. Throughout the life of the organism cells are dividing and obtaining their specialization to become, for example, blood cells, muscle cells, neurons or even cancer cells. The fate of each cell depends on the set of proteins being produced inside it. It is the proteins that are responsible for the functional specialization of the cell [37].

The information about human organism is stored in each cell nucleus in the form of DNA. DNA is a long double stranded molecule that carry genetic information. The two DNA strands, known as polynucleotides, are composed of nucleotides. Each nucleotide is formed by one of four nucleobases: adenine (A), thymine (T), cytosine (C), guanine (G), a sugar molecule deoxyribose, and a phosphate molecule. DNA contains a set of instructions for producing proteins, the building blocks of our bodies. These instructions are inscribed in the structure of the DNA molecule through a set of functional regions called genes. The instructions are carried out by the cell's molecular machinery to produce new proteins and ultimately regulate functionality and fate of the cell. The flow of information from genes to proteins is described as follows: DNA is transcribed into RNA, RNA is translated into protein [38] (Figure 2).

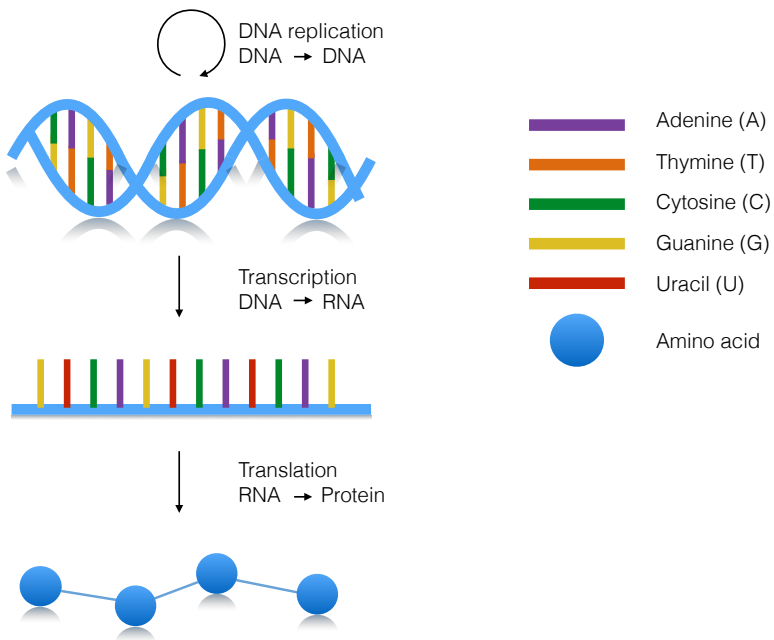


Figure 2. Flow of information from genes to proteins. Figure illustrates the process of DNA transcription into RNA and further RNA translation into protein. Individual coloured bars indicate nucleotide bases. The filled blue circles stand for protein amino acids.

Specific types of experiments are designed to catch different aspects of cellular functioning, e.g. on the DNA (genome), RNA (transcriptome) or protein (proteome) level, and its relation to the phenotype of interest (phenome). For example, the results of such experiments describe DNA sequence of an individual, detected mutations in the genome, measured levels of gene expression in different tissues or protein levels and interactions in various biological conditions, e.g. disease and healthy. The data sets produced from these experiments form so called *omics* data layers. The word *omics* in the context of this work refers to the set of studies and methods focused on the data sets of individual origin. In the Section 2.1 and 2.2 we will introduce individual *omics* layers and biological data types that originate from experiments and analysis of the corresponding *omics* data (Figure 3).

2.1. Omics data origins

As organism is a complex system, there is no single layer of information that is sufficient to completely explain the mechanisms by which genes at genome layer lead to complex phenotypes at phenome layer (Figure 3). It is the combination of the individual intermediate layers and their relationship that can bring the systematic view on how specific phenotype is formed and why [11, 13, 39]. Below we will describe the most frequently addressed *omics* data origins and describe information that individual *omics* levels can provide for biological studies.

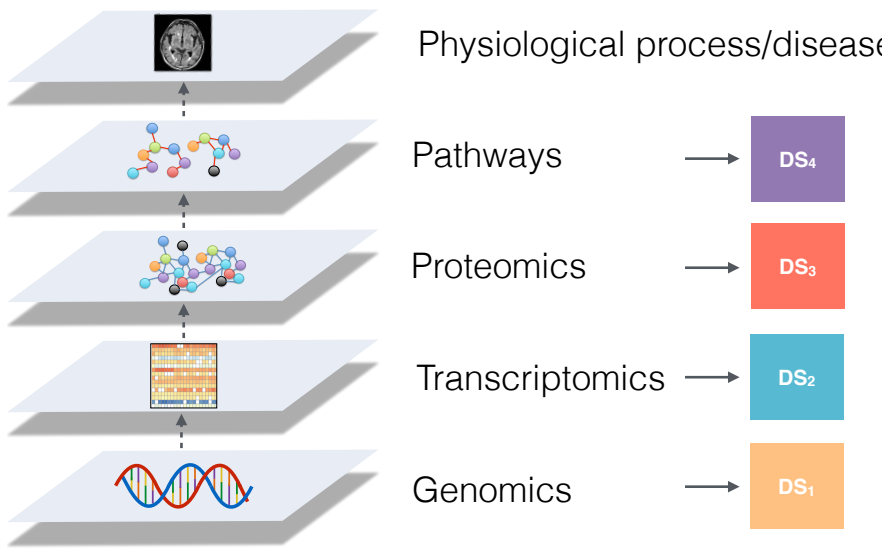


Figure 3. Combining individual *omics* layers.

2.1.1. Genome

A genome is a complete set of genetic information in an organism stored in its DNA. It contains coding regions for producing proteins and non-coding regions, e.g. regions producing functional RNA-s or RNA-s taking part in transcriptional and translational regulations [40]. It provides all information that is required for building and maintaining the organism. The genome is stored in a form of long DNA molecules tightly packed into chromosomes. The study of genomes, known as genomics, focuses on sequencing, assembling, and analysing the structure and function of genomes. Genomics and the corresponding types of data produced from such studies can be divided into several groups.

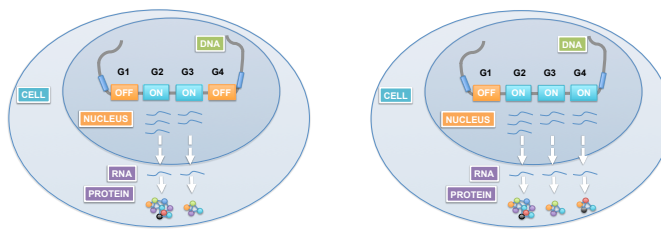
Genome sequencing and subsequent mapping of genes and genetic markers to their locations in the genome provides valuable information. These genetic markers represent short sequences and individual positions that can be used to identify individuals or species. An example of such marker is a single nucleotide polymorphism (SNP) that is a variation at a single position in a DNA sequence among individuals. This knowledge can be used to manipulate the genes and DNA segments of an organism. Another branch of genomic studies, functional genomics, answers the questions about the relationship between genotype and phenotype on genome-wide level. It investigates function of DNA at the levels of genes, RNA transcripts, and gene products and their interactions [41]. Genome-wide association study (GWAS) is a comparison of a genome-wide set of genetic variants, e.g. SNPs, in different individuals to find variants associated with a defined phenotype such as human disease, response to treatment, etc. [42]. Also, the joint effects of genes or genetic variants can produce a completely new trait in comparison with the traits in which individual effect of each gene or genetic variant takes place [43, 44]. Epistasis is an interaction between genes at two or more locations in the genome when the effect of one gene is dependent on the presence of the other gene [45]. The studies of epistasis allow to detect interacting genetic variants determining the trait or phenotype of interest such as disease associated phenotype. Another direction in genome studies is called comparative genomics. It is designed to compare the genomes of different organisms, e.g. human, mouse, chimpanzee, bacteria, etc. The goal of such studies is to find similarities and differences in sequences, and to identify evolutionary relations between species. Comparative genomics methods rely on the principle that common features of two organisms will often be encoded within the DNA that is evolutionary conserved between them [46, 47]. Data types related to genome level of information are described in Section 2.2.2.1, 2.2.2.4-2.2.2.5, 2.2.3.2 and 2.2.3.4.

2.1.2. Transcriptome

The total set of messenger RNA (mRNA) molecules expressed by genome is known as transcriptome. The mRNA carries genetic information copied from DNA template strand, that is later processed to create a protein. Transcriptome

describes the set of mRNA transcripts that are specific to some tissue or a particular cell type both qualitatively, describing what transcripts are present, and quantitatively, reflecting how much of each transcript is expressed [48]. In other words transcriptome shows which genes are expressed and to what extent in a particular tissue of interest.

Unlike genome, excluding mutations, transcriptome varies due to the developmental or environmental conditions. Transcriptome data includes all the transcripts in the cell or tissue type. It reflects the genes that are expressed, or in other words turned "ON", in cell at a given condition, e.g. healthy or disease (Figure 4) or time, e.g. early development, stage of disease, etc. Traditionally from the perspective of central dogma of molecular biology transcriptome was viewed as an intermediate step between genome and proteome. However, in the recent years other types of RNA molecules besides protein-coding mRNAs were discovered that can be dysregulated in the disease condition [49–51]. There are a few widely used laboratory techniques to measure transcriptomics data in the cell, such as hybridization microarrays, next generation sequencing methods, qRT-PCR, etc. The overview of data types obtained in transcriptomics studies are provided in the Section 2.2.1.3-2.2.1.4, 2.2.2.2-2.2.2.3, 2.2.3.1-2.2.3.2, 2.2.3.4.



HEALTHY CELL

Gene	Expression profile	Expression level
G1	OFF	0
G2	ON	30
G3	ON	20
G4	OFF	0

CANCER CELL

Gene	Expression profile	Expression level
G1	OFF	0
G2	ON	30
G3	ON	20
G4	ON	20

Figure 4. Gene expression in healthy and diseased cells. Figure illustrates the differences in the gene expression profiles and the variety of gene products (proteins) being produced in healthy and diseased cell. G1..G4 is short for gene 1..gene 4. ON/OFF state indicates if the gene is transcribed or not.

2.1.3. Proteome

Besides genetic information it is important to know the types and abundance of the proteins expressed by the organism. It is also critical to understand what proteins

are being produced in the cell in a given condition or tissue. The entire set of proteins produced by an organism, tissue or a cell form a proteome [37].

Proteomics, a study of proteomes, uncovers proteins' functions and their ability to physically interact with each other, e.g. forming complexes to execute some functioning role. Additionally, proteomics studies show that functionality of a protein is dependent of its 3D structure and modifications such as phosphorylation, lipidation, etc., that take place after translation [52, 53].

In essence, proteome is an expression of an organism's genome. In contrast with genome that consists of four nucleotides and is the same in every cell of the organism, proteome is comprised of twenty different amino acids and various post-translational modifications. The post-translational modification is an attachment or removal of chemical groups, e.g. sugars, phosphates, etc., to the protein molecules. The modifications in protein 3D structure allows it to physically interact or not interact with other proteins and protein complexes. The set of proteins produced by a cell varies depending on cell type, developmental stage, function, and location in the tissue. Proteome of an organism is constantly changing in response to various factors including organism's developmental stage, disease, and environmental factors. This property of proteome makes it very valuable in the studies of biological processes and in medical applications. Currently there are various technologies available in protein research such as mass-spectrometry, yeast two hybrid assays (Y2H), immunoassays, etc. These technologies allow to measure, for example, protein abundance and protein-protein physical interactions (Section 2.2.1.1-2.2.1.2, 2.2.1.6, 2.2.2.6, 2.2.3.1 and 2.2.3.4).

One of the sources for human proteome studies is blood plasma. It contains information about the proteins being present in the blood. This set of proteins can serve as an indicator of the physiological state of the body. For example, plasma contains such types of proteins as cytokines and chemokines involved in immune system response and inflammation. Thus plasma proteome is a great value for medical purposes to search for biomarkers in diagnostics and therapeutic use [54].

2.1.4. Phenome

A phenotype is a set of all observable physical and measurable biochemical characteristics describing a given individual, tissue, organism or species. A repertoire of all phenotypes of a given tissue, organism or species forms phenome [55, 56]. The observable physical characteristics can be represented by height, weight, eyes color, behaviour, disease state, etc., while biochemical characteristics are, for example, levels of hormones or metabolites. A notion of phenotype can be applied to characterize cell types in the organism. It can be defined as a set of morphological and additional biochemical characteristics such as expression of the certain genes or proteins, describing the cell (Figure 4). A phenotype is a physical characteristic, for example it could be height, weight, behaviour or a blood measurement. A phenome represents the entire repertoire of all phenotypes of given a given

tissue, organism or species. Although defining a phenotype for a study depends largely on a study setup, examples of data types that describe phenome level of information can be found in Section 2.2.1.5-2.2.1.7, 2.2.1.8.

The phenome is determined by the interplay between organism's genome and environmental factors. Knowledge about phenome or in more narrow sense phenotypes can be used in medical studies and practices, for example, to understand the etiology of the disease and to foresee how individuals react to therapies [11, 56]. Phenotypic level of information serves as a foundation for medical and biological studies. It defines the conditions in which the measurements are taken and analysed. Therefore an outcome of an experiment and a study is highly determined by how well phenotype of interest is defined. Poorly defined phenotypes might lead to inability to make right conclusions about studied biological question, e.g. identify genetic variants, difference in proteins or gene expression levels, etc.

2.1.5. Other omics

A fast progress in science and technology triggered the expansion and structuring of an existing knowledge. It led to the formation of the new, more specialized, *omics* fields. For example, epigenomics studies epigenetic modifications in the organisms; metabolomics investigates the metabolites present within an organism, cell, or tissue; metagenomics studies the genetic material containing in environmental samples. The specific data types related to those *omics* are not included into the current work.

2.2. Data types

In order to study biological process or an organism at individual *omics* levels described Section 2.1 diverse experiments are conducted by scientists working in different domains. The results of these experiments are data sets of various types (Figure 3). The development and application of the right data integration methods are highly dependent on the understanding on how the data were produced, what was measured and under which conditions. This section provides an overview of the data types used in the current thesis along with the technologies used for production and databases used to store data of similar type. Largely all the data types in bioinformatics studies, and particularly in data integration field, can be divided into three large groups - experimental, computational and domain knowledge-based data.

As experimental data we will refer data sets that were obtained in the biological laboratories using biotechnological means, e.g. microarray data, qRT-PCR, Y2H protein interaction data, sequencing data sets, etc. As computational data we will refer data sets that were obtained as a result of the analysis of the experimental data sets using well-established computational pipelines and methods, e.g. co-expression, GWAS data sets, positive Darwinian selection, epistasis data, etc.

Additionally, we will be using the notion of domain knowledge as a collection of domain expert knowledge and a set of well-established resources such as data collections and tools for characterization of the findings or for developing a new hypothesis. For example, we will be using aggregated resources gathering information about genes, proteins and biological pathways. A combination of experimental, computational and domain knowledge-based data described in Section 2.2.1, 2.2.2 and 2.2.3 in multi-staged and transformation-based data integration are later demonstrated in Chapter 4-6.

2.2.1. Experimental data

In this section we will describe experimental data types used in Publications I-IV. These data types are obtained from studies of individual *omics* layers described in Section 2.1. An application of these individual data types are later demonstrated in the integration setup in Chapter 4-6.

2.2.1.1. Protein-protein interactions

Studying the interacting partners of a protein can reveal its function and provide information about biological processes it participates in. For example, if most of the interacting partners of the protein are involved in a given biological process, there is a chance that the protein of interest is also involved. The process of function identification can be performed by screening a single protein with a known function against a set of proteins with unknown function or by scanning a protein with unknown function against a library of proteins with known function.

The yeast two-hybrid (Y2H) is a well established technique to detect protein-protein interactions (PPIs) [57]. Y2H relies on the detection of the physical interaction between two proteins. The interaction detection is based on the expression of so called reporter gene, which is activated when a specific transcription factor protein binds to its promoter. This transcription factor is comprised of a DNA-binding domain and an activation domain. In Y2H experiment the protein of interest, bait, is fused with the binding domain, and the protein library, prey is fused with the activation domain. The transcription of the reporter gene takes place only when both prey with activation domain and bait with DNA-binding domain are present in the promoter and therefore interact [57–61].

Alternative powerful technology for studying PPIs is protein mass spectrometry [62–64]. This method accurately measures mass of different molecules, i.e. proteins or protein complexes, in a sample. At the first step of process molecules in a sample are vaporized by heating and then ionized. The ions are sorted based on their mass-to-charge ratio by analyzer. The detector system catches ions and records their relative abundance. The advantage of this technology is lack of necessity to prepare library of proteins. However, mass spectrometry is limited to detect temporal or weak transient PPIs [65]. Both technologies can be used separately or in combination to complement individual results [66, 67]. Additionally,

gene expression microarrays and RNA sequencing technologies can provide indirect evidence for protein interaction [68]. Aggregated information about protein-protein interactions (Section 2.2.2.6) collected from various experiments is deposited in publicly accessible databases such as IntAct [4], STRING [69, 70], etc. In Chapter 4 we use PPI data in transformation-based study setup of Alzheimer's disease. We combine PPI data sets originating from different sources with gene co-expression (Section 2.2.2.2) and epistasis (Section 2.2.2.5) into heterogeneous network of interactions to identify genes potentially associated with disease.

2.2.1.2. Protein abundance

Blood plasma is one of the most accessible biofluids that can be used in the clinical diagnostics and for research purposes. Plasma contains variety of proteins carrying different functions, e.g. regulation of inflammation. The amount of a protein can be assessed by measuring its concentration, e.g. using Milliplex MAP multiplex assay [71,72]. This method allows detecting several proteins simultaneously [71]. It uses antibodies against several cytokines, i.e., small proteins that immune system uses for communication between cells, to quantify the level of those cytokines in blood plasma, i.e. blood without red blood cells [73]. Using a case-control study design, the concentration levels in disease group and healthy group can be measured and compared. The difference between the levels of concentrations in these conditions can be evaluated using statistical methods, e.g. Wilcoxon test (Section 3.2.7). The proteins with statistically significant differences in levels of concentration can potentially serve as biomarkers for the disease [74]. In Chapter 5 we analyze protein concentrations together with gene expression, fluorescent microscopy and other data sets to study psoriasis pathogenesis using multi-staged integration approach.

2.2.1.3. Gene expression microarrays

Protein-coding genes in our DNA define what proteins are produced in our body. An intermediate step of producing a protein is the creation of the DNA transcript or mRNA molecule. The relations between the mRNA levels and the amount and the variants of produced proteins are not always so straightforward [75] due to complex regulatory mechanisms. However, from the practical perspective it is more convenient to measure the levels of gene expression rather than the amount of proteins. Gene expression microarray is a high-throughput technology for measuring the expression of thousands of genes simultaneously [76]. Microarrays are the glass, plastic or silicon plates covered by different short oligonucleotide sequences called probes. Each of these sequences is a complementary DNA sequence to the specific gene [77]. Probes are designed to bind to the unique transcriptome sequences converted to the cDNAs by reverse transcriptase enzyme. Each microarray can contain thousands of copies of probes corresponding to the different genes, covering the vast majority of the genes in the organism. Obtained

cDNAs are labeled with the fluorescent molecules before being introduced to the slide to bind to the probes. The process of binding is called hybridization.

Quantification of the transcripts of interest is based on measuring the fluorescent intensity using a scanner. The intensity signal converted into numeric scale represents the abundance of the mRNA in the studied sample. There are two types of microarray technologies - two-channel and single-channel [78]. In two-channel microarrays two different fluorophores, e.g. Cy3 emitting green color and Cy5, emitting red color, are used to label the different samples, e.g. cancer and healthy. The labeled samples are then mixed together for hybridization at the same array. In the single-channel microarray, e.g Affymetrix Gene Chip, Illumina Bead Chip, only one fluorophore is used for measuring the signal. When using single-channel microarray each sample is measured using a separate array.

The limitation of the microarray technology is its dependency on the prior sequence knowledge. Only the mRNAs with the corresponding cDNA carrying probes introduced to the slide can be detected. It is not possible to detect the structural variations for discovering novel genes or transcripts. Detection of the very similar sequences such as gene isoforms is also limited due to the low sensitivity [79]. However, specifically designed microarrays can be used for those purposes, e.g. to detect single nucleotide polymorphisms and fusion genes [80]. Alternative high-throughput technology that overcomes substantially the limitations of the microarrays is RNA sequencing (RNA-seq). This method is not dependent on the defined probes and potentially could be used to detect all mRNA produced in the cell. However, there might be potential limitations, e.g. in detecting some low-expressed genes, in case the sequencing depth is not sufficient [81–84]. Despite that, gene expression microarray is a robust, relatively low-cost technology that has been used over two decades. Large collection of publicly reusable gene expression microarray data is currently deposited at ArrayExpress and Gene Expression Omnibus databases [1, 2, 85].

High-throughput gene expression data serve as a source for obtaining computational data types such as gene co-expression and gene differential expression (see Section 2.2.2.2 and 2.2.2.3 for details). In Chapter 4 and 6 we analyze and combine microarray data sets with other data in application of integration approaches to two biological domains: Alzheimer's disease and toxicology.

2.2.1.4. Quantitative real-time PCR

Quantitative real-time polymerase chain reaction (qRT-PCR) is a low-throughput method used in research and in clinical diagnostics to provide quantitative measurements of gene transcription, i.e RNA abundance in a cell [86, 87]. It is used for a broad range of applications, e.g. to determine how the expression of a particular gene changes in the response to alterations in environmental conditions or various stimuli.

This technology is often used as a gold standard to control or validate the find-

ings from the experiments that use high-throughput technologies such as gene expression microarrays. The robustness of the method relies on its ability to accurately amplify known DNA sequences of interest. The amount of an expressed gene (Section 2.1.2) in a cell can be measured by the number of an RNA transcript of the gene of interest. To apply qRT-PCR technique to RNA of interest, at the first step RNA should be converted to the complementary DNA (cDNA). This process is performed using reverse transcription, operated by the reverse transriptase enzyme.

The method relies on the basic principles of the polymerase chain reaction (PCR) [88]. PCR is a biochemical temperature-based technique to amplify a specific fragment of target DNA. Amplification process is cyclical, and the amount of DNA doubles in every cycle. The method consist of three major phases that are controlled by the change in temperature - denaturation, annealing and elongation (Figure 5).

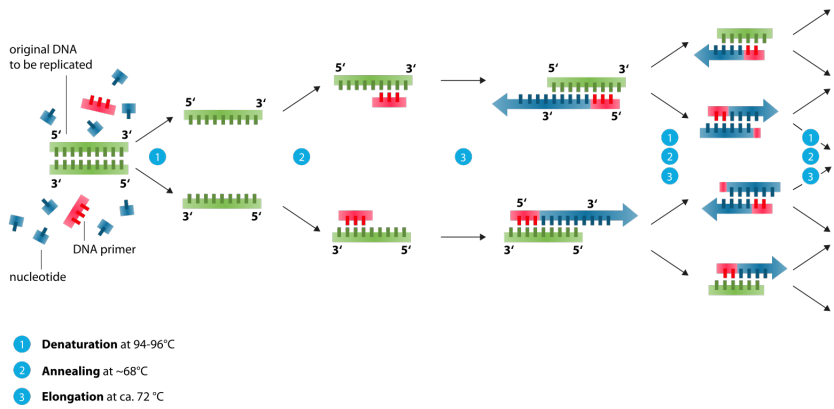


Figure 5. Principle of PCR (figure is adapted from Wikipedia [89]). The procedure consists of three parts: denaturation, annealing and elongation. At the denaturation stage the original double-stranded DNA is "melted" into two single strands at high temperature. At the annealing stage the temperature is lowered in order to allow DNA primers to bind to each of the separated strands of the original DNA. At the elongation stage the temperature is risen again to allow DNA polymerase to add nucleotides to the end of the primer sequence. The template DNA acts as a reference strand for the polymerase. There are two copies of the original DNA fragments by the end of the cycle.

The PCR reaction requires a single-strand template. At the denaturation stage the original double-stranded DNA that needs to be replicated is "melted" into two single strands at high temperature. Later at the annealing stage the temperature is lowered in order to allow DNA primers to bind to each of the separated strands of the original DNA. Primers are short nucleotide sequences complementary to the gene of interest that serve as a starting point for the complementary DNA synthesis. At the elongation stage the temperature is risen again to allow DNA polymerase, i.e a DNA building enzyme, to start adding nucleotides to the end of the primer sequence annealed to the template DNA. Primers serve as the indicators

of the direction in which polymerase molecule starts adding the nucleotides. The template DNA acts as a reference strand for the polymerase [86, 87]. As a result of such manipulations there are two copies of the original DNA fragments by the end of the cycle. To quantify the amount of the product in the end of the process the fluorescent intensity is used. In qRT-PCR the amplified product is measured at each step of the cycle. Quantification of the RNA can be done in two ways. A relative quantification approach is applied to calculate the expression levels of the gene of interest relatively to the expression of the stable reference gene. In an absolute quantification approach exact produced number of target DNA molecules is counted [86, 90].

Widely used relative quantification methods measure the difference in expression level of the gene of interest and the reference gene [91]. Relative quantification is easier to use since it does not require to know in advance the exact reference amount of the studied gene. However, the crucial aspect when applying the relative quantification method is the stability of the reference gene [92–94]. The output of the quantification is expressed in the number of cycles, denoted as CT values. The smaller is the number of cycles, the more gene transcripts were in the original sample, i.e. the higher is gene expression. Despite being a reliable method for measuring gene expression, it can be affected by the poor primers design or reference gene not being stable in a given condition [87, 95, 96]. We use qRT-PCR gene expression in the study of pathogenesis of psoriasis by identifying and later comparing expression levels in healthy and disease skin biopsy samples (Chapter 5). We also study expression of migration-related genes in toxicology studies to identify mechanisms of action of toxic compounds (Chapter 6).

2.2.1.5. Flow cytometry

Flow cytometry is the technique for the analysis of multiple parameters of individual cells, such as size and shape, within the heterogeneous population. It is used in a range of applications, e.g. for cell counting, cell sorting, biomarker detection and protein engineering. In Chapter 5 used flow-cytometric immunophenotyping to study changes in T cells subpopulations in psoriasis patients and healthy control individuals. During the procedure the flow of cells in a stream of a fluid is passed through a laser beam. The stream of a fluid is used to hydrodynamically focus the cell mixture through a small nozzle. The flow cytometer captures the light that is emerged from every cell as it passes the laser beam (Figure 6).

For the detection purpose cells are stained with specific fluorescent molecules such as fluorophore-labeled antibodies. The generated experimental data can be visualized as a two-dimensional dot plot. The areas on these plots can be sequentially separated, based on fluorescence intensity. These subsequent extraction of sub-populations of cells is called gating. Specific gating protocols exist for diagnostic and clinical purposes to discriminate between multiple populations of cells [97, 98].

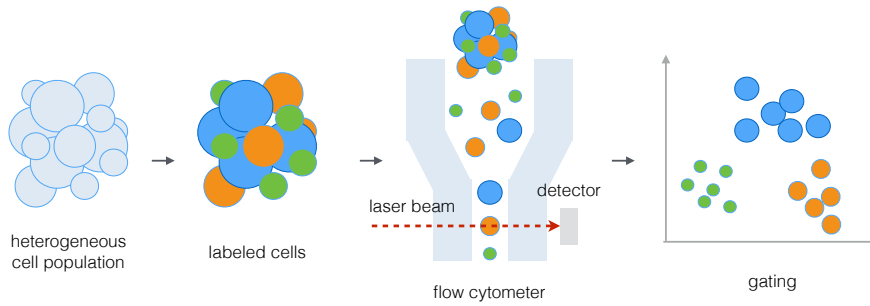


Figure 6. Schematic diagram of the working principle of the flow cytometry. Heterogeneous cell population is stained with fluorophore-labeled antibodies and passed through a laser beam. The flow cytometer's detector captures the light that is emerged from a cell when it passes the laser beam. The generated experimental data can be visualized as a two-dimensional dot plot. The sub-populations of cells on this plot can be separated based on fluorescence intensity.

2.2.1.6. Immunofluorescence microscopy

Immunofluorescence microscopy is a very robust and regularly used imaging technique in research and in medical practices to assess the localization and expression levels of proteins of interest [99].

It is used for a range of tasks such as immunophenotyping, cell sorting, cell cycle analysis, etc. The most common application of immunofluorescence microscopy is immunophenotyping. During this procedure individual populations of cells in the heterogeneous sample are identified and counted. The cell subsets are detected by labeling population-specific proteins with a fluorescent tag, known as fluorophore, on the cell surface. In clinical labs, immunophenotyping is used for diagnostics purposes [100, 101].

Modern advances in immunofluorescence microscopy allow to use this technology in broader range of applications, i.e. labeling of structures in living cells and measuring of the physiological state of a cell [102]. Immunofluorescence can be used to produce images of tissue sections, cultured cells or individual cells that are selected by a variety of methods.

2.2.1.7. MINC functional assay

The disruption of cells' ability to migrate can be used in toxicity testing strategies. To test if the compounds, e.g. drugs, disturb the migration capacity of neural crest cells (NCC), migration of neural crest cell (MINC) assay was designed [103]. This test system assesses, how many cells re-migrate into a cell-free area, i.e. scratch, within a neural crest cell monolayer after the treatment with the toxic compound. It is performed by analysing imaging data and counting the number of cells in the scratch area [103, 104]. While using these assays we can observe the endpoint of the compound's action, i.e. if it had an effect or not. This endpoint can indicate

toxic action of the studied compound, however in order to understand exact mechanism of action, MINC assay data should be combined with, for example, gene expression and known adverse outcome pathway data (Chapter 6).

2.2.1.8. Meta-data about the experiment

Every type of experiment is always accompanied by the corresponding meta-data, the descriptive information needed for data set to be understood and possibly integrated with other data sets. Meta-data records contain information about the purpose of experiment, experimental conditions, phenotypes, protocol, process of collecting the data, etc. Detailed, well-described and structured meta-data plays an important role in the tasks of data integration and increases the reproducibility of research.

2.2.2. Computational data

2.2.2.1. Genome-wide association studies

Genome-wide association studies (GWAS) are hypothesis free methods to identify genes associated with a phenotype, e.g. a disease.

GWAS examine the genomes to identify the common variants called single nucleotide polymorphisms (SNPs) in individuals both with and without a common phenotype, e.g. disease, using genome wide SNP arrays [105,106]. The main goal of this method is to identify SNPs that are more frequent in people with the disease than in other individuals. GWAS methods can scan genomes for thousands of SNPs at the same time to find the genes that can contribute to the person's risk for developing a disease, responding to the certain drugs and environmental factors [42]. GWAS is usually carried out in a case-control experimental setup comparing two large groups of individuals - case group with a particular phenotype, e.g. disease, and control group without a particular disease. It aims to identify SNPs that are more frequently observed in case group in comparison with control group. During the study both groups are being genotyped where the genetic constitution of each individual is determined and scanned for the majority of known SNPs. For each of these SNPs the frequency of each allele, i.e. variant form of a gene, is counted and compared in case and control groups (Table 2.1). The unit for reporting effect size in GWAS is the odds ratio between the odds of having a disease in individuals having a specific allele and the odds of having the disease for those who do not have the same allele (see equations 2.1) [107–110].

Table 2.1. Example of allele counts in case and control groups.

Allele counts		
	G	T
Cases	a	b
Controls	c	d

$$
\begin{aligned}
odds\ ratio\ (OR) &= \frac{odds\ of\ disease\ for\ individuals\ having\ allele\ G}{odds\ of\ disease\ for\ individuals\ having\ allele\ T} \\
&= \frac{a/c}{b/d} = \frac{a*d}{b*c}
\end{aligned}
\tag{2.1}
$$

Where the resulting odds ratios (OR) indicate the following:
OR = 1: no association between genotype and disease
OR > 1: G allele increases risk of disease
OR < 1: T allele increases risk of disease

When odds ratio is higher than 1 then the allele frequency in the case group is higher than in the control group. Additionally, chi-squared test is used to test whether the SNP association with the disease is significant [108]. GWAS approach has proven itself useful in finding genetic variations contributing to the diseases such as cancer, diabetes, heart disease and neurodegenerative disorders [111, 112].

GWAS detect SNPs and other variants in DNA associated with a phenotype of interest, e.g. disease, but cannot provide information which genes are causal [113, 114]. For further studies GWAS results should be combined with other biological data such as protein-protein interactions to identify the most reliable associations. These associations can be later link to the proteins and used for the development of novel drug targets. In Chapter 4 we describe how to incorporate the results of GWAS studies into the transformation-based data integration and machine learning model for classification of heterogeneous biological data.

2.2.2.2. Gene co-expression

Gene co-expression describes the correlation between gene expression levels across multiple samples and biological conditions. Genes that have similar expression profiles in the same conditions are considered to be co-expressed. This principle is often used to infer a function of a gene using "guilt-by-association rule" knowing the function of the group of correlated genes that demonstrate similar expression patterns [68, 115, 116]. This is based on the idea that genes that are co-expressed in some biological condition, e.g. cancer, might share a biological function [117, 118].

Co-expression can be measured using distance metrics such as Euclidean, Pearson correlation coefficient based, or Spearman rank correlation coefficient based (see detailed description of metrics and co-expression analysis in Chapter 3. Clustering methods such as k-means (Section 3.2.2.1) and hierarchical clustering (Section 3.2.2.2) are widely used to find groups of genes with similar expression patterns using one of the metrics. In Chapters 5 and 6 we demonstrate that clustering of co-expressed genes followed by functional enrichment analysis and the combination with domain knowledge about pathways help to understand the

biological processes these genes are involved in, i.e. psoriasis pathogenesis and adverse outcome pathways of various toxic compounds.

Co-expression analysis of multiple data sets can reveal more information about the behavior of the genes [119]. Methods such as Robust Rank Aggregation can be applied to identify co-expressed genes in a set of microarray experiments [120]. Additionally, in order to illustrate the connection between genomics and proteomics levels of biological evidence, it is important to mention, that gene co-expression has been used to predict or in some study set-up to validate protein-protein interactions [115, 121–124]. The main idea is that two genes which have correlated expression across various multiple conditions are more likely to encode interacting proteins [65]. These property could be used to identify potentially interacting proteins associated with the disease. However, co-expressed genes may be unrelated to protein interaction if the genes involved in two different biological processes were just activated by the same stimulus. Combination of gene co-expression with additional data types such as, for example, protein-protein interactions (Section 2.2.1.1), epistasis (Section 2.2.2.5), GWAS (Section 2.2.2.1), meta data about the tissue where genes are expressed, etc., allows to narrow down the search space for the potential proteins associated with the disease. In Chapter 4 we combine of co-expression data with other data sets in Alzheimer's disease study.

2.2.2.3. Differential expression

In humans and other organisms, nearly all cells contain the same sets of genes, however, different cells express different sets of genes, i.e have different transcriptomes. These differences in expression are responsible for the properties and behaviors of the cells and tissues in healthy and disease conditions.

Differentially expressed genes are genes that exhibit statistically significant change their expression levels between the conditions. These genes can describe the differences between phenotypes, e.g. healthy and disease, various tissues and cell types in the organism, indicate the influence of the toxic compounds on early human development in toxicity testing experiments, etc. In the current thesis we combine differential expression with other data types in multi-staged integration setup to study the pathogenesis of psoriasis (Chapter 5) and in toxicology studies (Chapter 6). Differential expression analysis is commonly performed using statistical methods such as linear models, ANOVA [16, 17], t-test [19] or non-parametric Wilcoxon test [20–22]. Statistical tests provide a p-value that serves as an estimation of a gene being significantly differentially expressed between the given conditions. The description of the statistical methods is provided in Chapter 3. Additionally, historically the approach known as the fold-change estimation was applied to identify differentially expressed genes. It evaluates an average log-ratio of expression values between the samples. In this approach genes are considered as differentially expressed if fold change is higher than an arbitrary

cut-off. However, in practice both measures - p-value and fold change are used in combination to detect the genes with significant changes between conditions.

2.2.2.4. Positive Darwinian selection

Darwinian selection, also known as positive selection, is based on the process of evolution by means of natural selection where phenotypes with increased chance of survival have a higher chance to be passed on to the next generations [125]. It is the process by which advantageous genetic variants propagate in a population. Understanding the process of adaptation is important for answering many biological questions, such as how species respond to environmental changes, e.g. climate or pathogens, and what mechanisms underlie genetic diseases [126]. Evolutionary adaptation occurs when an inheritable change in the phenotype makes it more preferable in the present environment.

Positive selection studies have contributed to understanding of the evolutionary basis of diseases such as Alzheimer's disease [127]. In case of Alzheimer's disease several genes were found to be associated with the regulation of transcription of the selected genes in immune cells, suggesting how disease related molecular mechanism may have evolved [127]. In Chapter 4 we demonstrate an application of positive selection data in transformation-based integration study setup of Alzheimer's disease.

The increasing number of completely sequenced genomes of various organisms together with an abundance of bioinformatics methods [125, 128, 129] provide a possibility to detect events of positive selection by means of comparative genomics. In particular, comparing the genomes from closely related species have proven to be effective in detecting genetic regions under the positive selection [130, 131]. Comparative genomics studies identified the genes that might have experienced positive selection during the evolution of human and other primates. These genes offer valuable insights for understanding the biological processes specific to humans [132]. Comparative genomics relies on the principles that common features of two organisms will often be encoded within the DNA that is conserved between the species, i.e. the DNA sequences encoding the proteins and RNAs responsible for functions that were conserved from the last common ancestor would be preserved in contemporary genome sequences [133].

Though positive selection cases are of great interest, they are difficult to detect and analyze [125, 134]. Signatures of positive selection can be detected by using statistical tests like maximum likelihood test, also called branch-site test, and machine learning algorithms implemented in such tools such as PAML [129], OmegaPlus [135], SweeD [136].

2.2.2.5. Epistasis

Epistasis can be defined as an effect of interaction between two or more variants of different genes on a phenotype deviating from their individual effects [45, 137].

In other words, epistasis describes how gene interactions can affect phenotypes [137]. An effect of the variant in one gene is masked by the presence of specific variant in another. These interactions are especially interesting in case of complex traits such as diabetes, multiple sclerosis and Alzheimer's disease. In Chapter 4 we have used epistatic interactions as one of data layers for the construction of heterogeneous network-based Alzheimer disease-specific data set.

Epistatic effects are difficult to detect due to the interplay of many factors such as an increased number of contributing genes and environmental effects. Genes with epistatic relationships tend to code for proteins that are involved in the same processes. Looking at epistatic relationships offers clues that help us to understand how genes and proteins function. The most common way of detecting the epistatic effects is by using linear regression [137, 138].

2.2.2.6. Aggregated protein-protein interactions

There are different experimental techniques to measure physical interactions between proteins [139]. The results of the majority of such experiments are reported in the scientific publications. Several public databases, e.g STRING [69, 70], IntAct [3, 4], accumulate data sets from these types of publications and directly deposited data sets. As a rule in such repositories information about repeatedly reported interactions is combined and the strength of each interaction is reported in a form of an aggregated score. Usually along with the aggregated score the reference to the original publication and the experimental method are mentioned. One widely used PPI database is IntAct [3, 4]. It is freely available and contains literature-curated data sets and data sets directly submitted by the users. Additionally, it provides expert-curated data sets, e.g data set of interactions in Alzheimer's disease, and computationally selected data sets, e.g. interactions of proteins with an established role in the presynapse [140] (Chapter 4). The confidence of each interaction in IntAct is reported in a form of a cumulative score normalized between zero and one across the entire database, where one is of the highest confidence. The cumulative score depends on the experimental score, detection method, number of publications and the interaction appeared.

2.2.3. Domain knowledge

Although individual experiments and data analysis provide valuable knowledge about the disease or biological process that was a subject of a study in a form of publications, it is important to accumulate and share such knowledge centrally. This section provides a description about the repositories and databases that serve for collecting, structuring and aggregating data of similar types.

2.2.3.1. Biological pathways

Genes provide the instructions which proteins should be produced in the cell to carry out nearly every task in our bodies (Figure 2). Proteins are the building

blocks of our bodies, they construct our muscles and organs, help our bodies move and defend us against diseases. Therefore it is important to study and reconstruct relations between genes and other substances in the cell. As a result of experiments scientists have accumulated knowledge about various types of interactions between proteins, genes, metabolites and other biological entities. The extracted knowledge of these interactions is aggregated in a form of biological pathways. It can describe both the physical and non-physical relations between genes, proteins, metabolites, etc.

Biological pathways are rather an abstract representation of the biological process or a disease in a form of a graph. The vertices in the graphs can represent various biological entities such as genes, proteins, metabolites, etc. The edges illustrate the relations between these entities such as chemical reactions, physical interactions, co-expression, etc.

Biological pathways can be broadly classified into three large groups by their functional specificity as metabolic pathways, gene regulation pathways and signal transduction pathways. Metabolic pathways are enzyme-mediated chemical reactions that are involved in either biosynthesis, a formation of new, more complex, biomolecules or breakdown of them into smaller units. Gene-regulation pathways describe the relations between the genes that lead to gene activation or inhibition. Signal transduction pathways show how a chemical or physical signal travels from a cell's exterior to its interior. Cells are able to receive various signals through receptors located on their surface. After interaction with the receptor the signal is transmitted inside the cell and triggers a particular reaction or process inside the cell. The knowledge about biological pathways has been gathered, classified and deposited into the databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [141, 142] and Reactome [143, 144].

Biological pathways are quite often used to represent domain knowledge to reduce the number of entities in the analysis in the data-driven studies. They are also used for the characterization of the obtained results according to the functional meaning of the pathway. For example identifying pathways involved in a particular disease or a part of the pathway that is affected may help to understand the disease mechanisms and improve diagnostics.

Additionally, to the three groups of biological pathways it is worth mentioning of a specific novel concept of *Adverse Outcome Pathway* (AOP) used in toxicity testing for human risk assessment [145]. An AOP is a schematic representation of a sequential chain of causally linked biological events that lead to an adverse, i.e. undesired harmful, effects related to health [10, 145, 146]. An AOP reflects an existing knowledge about the linked molecular initiating event (MIE) and the cascade of intermediate or key events (KEs) at the subcellular, cellular, tissue or organ level that lead to a specific adverse outcome (AO) [10]. AOPs are the central elements of toxicological studies that are used to support chemical risk assessment [10].

Biological pathways play an important role in the incorporation of the domain

knowledge into the analysis and in the interpretation of the results. It is often used as one of the data sources in multi-staged data integration analysis. We demonstrate a practical application of the biological pathways in combination with other computational and experimental data in Chapter 5-6.

2.2.3.2. Aggregated information about genes

Many genes in human and other organisms' genomes have been studied. As a result a variety of information was gathered about those genes such as their functions, location in the genome, possible variation, regulation, synonymous names, related pathways, etc. There are plenty of databases that provide the diverse gene-centered information such as Ensembl [147, 148], GeneCards [149, 150], Wikigenes [151], Entrez Gene [152, 153], etc. Historically various databases were creating their individual gene names that led to the problem of mapping those name spaces between each other during the analysis. One of the common practical problems in bioinformatics analysis is the one-to-many mapping of synonymous gene and protein names. Ensembl database currently provides unique universal identifiers for all genes and transcripts and the possibility of mapping to any other name space. These identifiers are used by the conversion web tools such as g:Profiler [154, 155] and David [156].

Gene annotations can help to interpret the analysis results in data driven research or give rise to the new hypothesis. For example, it can provide information about gene involvement in the disease pathway, its function or location in the genome.

2.2.3.3. Aggregated information about proteins

Additionally, to the gene-centered information it is important to know the aggregated information about the products that they produce. There are variety of the protein specific databases such as UniProt [157], HPRD [158], Human Protein Atlas [159]. These resources collect and provide information about the sequence of the proteins, their structure, isoforms, post-translational modifications, function, disease association, expression the different tissues and organs, etc. Protein annotations constitute another valuable source of information about the biological process of interest, e.g. disease or response to the certain medication.

2.2.3.4. Gene Ontology

Gene Ontology [160] is a vocabulary-based unified hierarchical representation of genes and gene products attributes across species. Knowledge about the biological role of genes and proteins in one organism can often be projected to the other organisms. Gene Ontology representation for all the entities consists of three categories:

- Biological process
- Molecular function

- Cellular component

The biological process, e.g. *"regulation of neuron differentiation"* is a set of ordered molecular events [160]. It is characterized by the molecular functions. Molecular function, e.g.*"tau protein binding"*, is a type of biochemical activity of a gene product [160]. For example, it can include specific binding to ligands or structures. Cellular component, e.g. *"somatodendritic compartment"*, refers to the cellular compartment where a gene product carries out its molecular function. Gene Ontology terms refer to eukaryotic cell structure. Every term has a term name, e.g. *"regulation of neuron differentiation"* and a unique term accession number. Every accession number is represented as a seven digit identifier prefixed by word "GO:". For example, accession number GO:0045664 corresponds to the term *"regulation of neuron differentiation"*. Gene Ontology allows functional interpretation of experimental and computational data using enrichment analysis (Section 3.2.9).

2.3. Summary

In this chapter we have provided an overview of the biological data types that were used in the thesis and their origins. We have broadly divided the biological data into three categories: experimental, computational and domain expert knowledge. We have described the relations between individual data types represented as individual data layers.

The diversity of computational and experimental biological data requires an application of appropriate individual data analysis methods and approaches for the integration of these desperate data types. In Chapter 3 we will describe two approaches and the corresponding data science methods applied for biological data integration.

3. INTEGRATIVE ANALYSIS METHODS

When it comes to biological studies, the variety of experimental, computational and domain knowledge data dictates the application of the analysis methods suitable for the particular data type. Bearing in mind the nature of the relations between these data types, different machine learning techniques can be applied to model each type of relations and detect complex patterns in the data. The results of such individual analyses explain biological process or disease mechanism at individual levels. The task of integrative data analysis is to combine various heterogeneous data sets using machine learning methods in an effective way to discover relations in the data that could not be detected otherwise.

In this chapter we describe data integration approaches and the corresponding machine learning methods that are used as the building blocks in the analysis. The methods described in this chapter are applied for the analysis of the biological experimental data described in the Chapter 2.

3.1. Concepts of data integration in biological sciences

Depending on the study goals and design, data integration task can be approached in various ways. For example, each individual data set can be analysed separately, and joint conclusions are later drawn from the results of the individual analyses. Another way to approach this task is to transform individual data sets together into the unified data set, i.e. a graph, and subsequently apply analysis methods on the joint transformed data set. Although there is no universal classification of the data integration approaches [11–14], in order to ease the reference to these concepts for the reader we have partially adapted and modified the terminology proposed in [15]. Thus, we refer to the two approaches described above correspondingly as multi-staged data integration and transformation-based data integration (Figure 7). In the following Section 3.2 and 3.3 we provide an intuition about these data integration approaches and describe corresponding machine learning methods used in each of them. These methods include unsupervised and supervised classical machine learning methods applied in multi-staged data integration analysis, and a set of deep learning methods, graph neural networks, specifically designed for graph-structured data sets resulting from the transformation-based data integration.

3.2. Multi-staged data integration

Multi-staged integration is a sequential analysis of the individual data sets using various machine learning methods. This type of data integration is carried out in a linear or hierarchical manner, where the analysis is split into multiple steps. In the case of multi-staged data integration, the original experimental data sets are analysed using computational methods. The results of these analyses produces

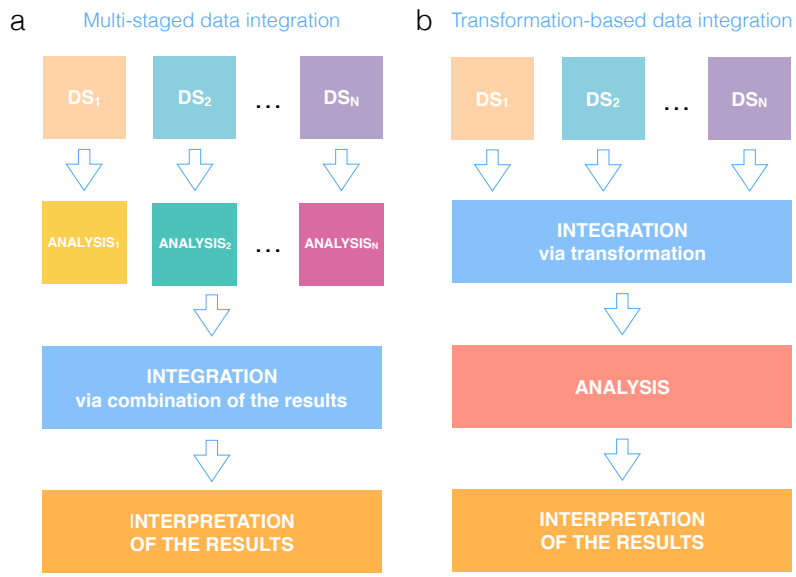


Figure 7. Schematic representation of data integration approaches. Figure illustrates: a) multi-staged data integration approach; b) transformation-based data integration approach. The colors of the individual data sets stand for the different data types.

computational data sets that are the derivatives from the original experimental biological data sets. The resulting computational data sets are then combined to find associations between different omic types, e.g. using domain-knowledge, and then interpretation of the result is performed. During the multi-staged data integration the relation of at most two data types is analyzed separately and then is associated with the phenotype of interest. This approach sequentially brings together the results of individual analyses, e.g. differential expression and image analysis, to capture a deeper knowledge about biological process or disease (Figure 7 a). The data sets under investigation could be of similar, e.g. a few gene expression microarray data sets, or different experimental type, e.g. gene expression, biological pathways, protein concentrations, etc., originating from various sources. This approach works well when the relation between two data types could be modelled in a linear manner [15]. In the following Sections 3.2.2 - 3.2.9 we describe methods applied in individual analyses of data types described in Section 2.2 to identify relations with a phenotype of interest. These methods are used as building blocks of the analysis pipeline (see Figure 7 a). Application of multi-staged data integration approach to the domain of immunology and toxicology is demonstrated in Chapter 5-6.

3.2.1. Principal component analysis

Measuring the levels of gene expression using various technologies, e.g. microarrays or qRT-PCR, has become a standard means for studying biological processes.

Principal component analysis (PCA) is mostly used as a tool in exploratory data analysis to visually identify the relationship between the samples in the study. In case of gene expression analysis PCA is applied to the transposed gene expression matrix in which columns represent genes and rows represent biological samples. The visualization of the results, e.g. in the form of scatter plot of a pair of principal components, will reflect the grouping of the samples. This type of analysis can be used to detect the outliers and observe the grouping patterns in the data. Although PCA can provide an understanding about the separation of biological samples, this analysis of gene expression data alone will not provide an a systematic understanding of a disease. In Chapter 5 and 6 we use PCA for developing initial understanding about the sample separation that is later accompanied by additional analyses. More precisely, in Chapter 5 we apply PCA on gene expression data obtained from skin biopsy of psoriasis patients and healthy control individuals. The results of PCA reveals separation of samples between psoriasis lesional skin and healthy control skin. However, from this analysis we are not able to infer the mechanism of psoriasis. In order to obtain more information it is necessary to combine additional analysis methods and use more data modalities, i.e. to identify genes that are differentially expressed in disease and control samples, to investigate whether there are changes on protein or cellular level and if these changes confirm mechanisms investigated on gene expression level.

PCA is a statistical procedure that represents an orthogonal linear transformation of the original data into a new coordinate system. Individual principal components (PCs) are uncorrelated linear combinations of original variables as shown in the equations (3.1) and (3.2) [161]. The first PC y_1 can be represented by the linear combination of the variables x_1, x_2, ..., x_p.

$$y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p \tag{3.1}$$

or in a form of a matrix

$$Y_1 = a_1^T X \tag{3.2}$$

The first PC accounts for the greatest possible variance in the data set. It is possible to make the variance of Y_1 as large as possible by increasing the weights (loadings) a_{11}, a_{12}, ... $a_{1\mathrm{p}}$. In order to prevent this scenario the weights are calculated with the constraint that their sum of squares is 1 (3.3).

$$a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1 \tag{3.3}$$

The other PCs are calculated in the similar manner, with the condition that they should be uncorrelated with the other PCs and that it accounts for the next highest variance.

As PCA creates a projection of the original data to the new reduced coordinate system while preserving the most variance in the data, it can be also applied as

a method for dimensionality reduction before further statistical modeling. Usually in practice first few components explain most of the variance in the data and therefore can be used to represent the data without loosing too much information.

Sometimes in gene expression analysis PCA can also be used to identify the genes that play the most important roles in separation of the samples by the principal component. It can be done by observing the loadings vector of the principal component. The genes with the largest corresponding positive and negative coefficients (loadings) give more impact on the separation of samples by the selected principal component.

3.2.2. Clustering methods

By analysing gene expression data it is also possible to identify genes that are co-regulated and might be involved in similar biological processes according to "guilt by association" rule. Usually these genes have similar expression patterns over different conditions or phenotypes. Various clustering methods can be applied to identify such groups of genes.

Clustering methods are unsupervised machine learning methods that allow to detect groups of similar objects, e.g. gene expression profiles, based on the applied similarity measure. After genes are grouped based on their expression patterns, these clusters can be characterized using domain knowledge (Chapter 6), i.e. Gene Ontology [160] and biological pathways from KEGG [141, 142] or Reactome [143, 144] databases. Characterization of the gene clusters allows to form new hypotheses and to understand the role of the particular gene or a group of genes in a certain biological process or disease (Chapter 5-6). The two most commonly used clustering methods are k-means and hierarchical clustering. Below we explain these methods.

3.2.2.1. K-means clustering

K-means [23] is a clustering algorithm that assumes a predefined number of clusters. On each step the algorithm assigns all data points to the closest cluster center using defined distance metric. Cluster centers are then recalculated, and the algorithm is repeated until it converges. The initialization can be performed by assigning cluster centers, e.g. randomly or using specific algorithms, or by initial partitioning of the data into given number of clusters and assigning the cluster number to each observation. To perform k-means clustering we need to estimate the number of clusters beforehand. In general it is not a trivial task. There are various techniques to select the number of clusters, e.g. based on within cluster sum of squares that serves as a measure of variability of the observations within each cluster or based on detection of so-called "anomalous patterns" as the candidates for the initial centroids [162–164]. The results of the k-means clustering is usually visualized using heatmaps. Heatmap represents a data matrix where the corresponding values are colour coded. Usually the lowest values correspond to

the "coldest" colours and the highest values to the "warmest" colours, however, arbitrary colour scheme can be chosen. Heatmaps allow to visually assess the trends in the data [165].

3.2.2.2. Hierarchical clustering

The hierarchical clustering method can be used as alternative method for the identification of patterns in the biological data. In comparison with k-means it does not require a predefined number of clusters. The result of the hierarchical clustering is depicted as a dendrogram, that represents a nested tree of partitions [23, 166].

The hierarchical clustering can be built using two main approaches: bottom-up (agglomerative) and top-down (divisive). In the agglomerative approach the closest objects are merged on each step of the algorithm to form clusters, and the distance to the other clusters is recalculated. The algorithm is repeated until one large cluster is formed. In the divisive approach all the data in a single cluster is divided into two clusters on each step of the algorithm.

The execution of the algorithm is heavily influenced by the way the distance between clustered objects is defined and measured. The choice of the distance metric, e.g. Euclidean, Manhattan, Canberra, correlation, etc., is defined for each particular task. Euclidean distance (equation 3.4) that measures the absolute difference between the objects, and correlation distance based on Pearson coefficient (equation 3.5) are the most commonly used metrics to compare the trends in the data.

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.4}$$

$$d_{cor}(x,y) = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.5}$$

where

$$x = (x_1, x_2, ..x_n), y = (y_1, y_2, ..., y_n),$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The way how clusters should be merged is defined by the linkage parameter [23], e.g. in single linkage the clusters with two closest objects are merged, in the complete linkage the furthest objects are merged, and in Ward's linkage the clusters are merged based on the sum of squared deviations to centroids, while minimizing the within-cluster sum of squares [167, 168]. However, while applying hierarchical clustering we have to bear in mind the possible uncertainty of the clustering results due to the statistical sampling error, i.e. when an analysed sample might not reflect the behaviour of the whole population [169]. In order to find the most reliable clusters bootstrap resampling techniques, i.e. pvclust, can be applied [170]. In multi-staged data integration study of psoriasis pathogenesis

described in Chapter 5 we use hierarchical clustering to identify genes with similar expression patterns based on qRT-PCR data. Later these groups of genes are used to develop a hypothesis about the functions of these groups of genes in the disease pathogenesis.

3.2.3. Robust rank aggregation

Often it is beneficial to understand the global behaviour of genes in a set of gene expression microarrays coming from different individual experiments. Gene co-expression analysis is one of the examples that greatly benefits from the aggregation of multiple data sets, as it can reveal functional relations between the genes. The results of individual microarray co-expression analyses are the lists of co-expressed genes that could be ranked based on their co-expression value, i.e Spearman or Pearson correlation coefficient. In order to understand the global relations between genes, the results should be aggregated. This integration can be done using Robust Rank Aggregation (RRA) method [120] as proposed in [119]. RRA is a rank aggregation method based on order statistics. It assumes a defined baseline model and re-ranks the genes based on the deviation from the baseline model. The baseline model considers that all input ranked lists would be a random permutation of the same set of genes. It means that for any gene from the input, its rank distribution across all list would be uniform. RRA tests the difference of the given gene rank distribution from the uniform distribution and uses test scores to re-rank the genes. Genes with observed smaller ranks more often than expected by the uniform distribution receive lower ranks, meaning that they are ranked in the top of the resulting list. The underlying probabilistic model makes the algorithm parameter free and robust to outliers, noise and errors. Test scores also provide a rigorous way to keep only the statistically relevant genes in the final list. Additionally, RRA can be applied in other types of meta-analysis of gene expression data. For example, it can be used for the aggregation of the result of differentially expressed genes coming from the experiments with the similar objectives.

3.2.4. Linear discriminant analysis

Linear discriminant analysis (LDA) [23,171] is a group of statistical methods used to find a linear combination of the original variables that separates two or more classes of objects or events. LDA can be applied to a variety of tasks in biology in order to classify and define groups of different biological objects. For example, it can be applied in the toxicology studies to classify the toxic compounds according to their mode of action. The obtained model can be applied for further mode of action prediction of unknown toxicants (Chapter 6).

Its purpose is to find the linear combination that gives the best possible separation between the groups. The resulting combination (LDA model) may be used as a linear classifier, or as a method for dimensionality reduction before classification, because it attempts to identify the transformation to the lower dimension

vector space while preserving the class structure of the original space [23, 172]. An optimal transformation, or in other words projection, (equation 3.6) in this case means the minimization of the within-class variance in the data set and maximizes the between-class variance simultaneously (equation 3.7) leading to maximum discrimination [173, 174]. LDA computes the directions, i.e. linear discriminants, that will represent the axes that maximize the separation between the classes.

$$y = W^T x \tag{3.6}$$

where W is the projection matrix that is calculated by maximizing the following objective:

$$J(W) = \frac{W^T S_B W}{W^T S_W W} \tag{3.7}$$

where S_B - is between class scatter matrix (a measure of between-class variance), and S_W is within class scatter matrix (a measure of within-class variance).

3.2.5. ANOVA

Often substantial information about the biological process or the disease can be obtained by studying the change in gene expression or the protein concentrations in blood plasma. Analysis of these changes helps to identify the difference between the biological conditions such as the difference between disease patients and healthy individuals or the difference between various patients' phenotypes. It is also used to discover potential disease biomarkers associated with the conditions [175–178]. Analysis of variance (ANOVA) is a method that is frequently applied by the researchers in clinical studies [175, 177–180]. ANOVA is a statistical method for comparing the means of measurements, such as gene expression or protein concentration in plasma, in two or more groups. It is based on comparison of the variability between groups to variability within groups. The groups can be defined by one or more factors and their categories [181]. For example, in Chapter 5 we apply ANOVA to investigate the difference in gene expression, protein levels and T cell subpopulations in psoriasis patients with different clinical features, i.e. duration of the disease that is divided into three groups "less than 10 years", "from 10 to 20 years" and "more than 20 years". While ANOVA shows whether there is an overall difference between the tested groups, it does not provide information what exact groups were significantly different. To identify such differences the post hoc tests for multiple comparisons are applied. Most commonly ANOVA is followed by Tukey's honestly significant difference [182].

3.2.6. Linear models for differential expression analysis

Differentially expressed genes (Section 2.2.2.3) exhibit significantly altered gene expression levels in the compared conditions, e.g. in disease vs. healthy tissue. They can serve as a potential biomarkers of disease and their characterization can

provide the hypothesis of the disease origins and mechanism. Differential expression analysis can be performed using a variety of methods. One of the methods that can be applied for such tasks is a linear regression model. It attempts to model the relationship between two or more variables by fitting a linear equation (3.8) to observed data. In case of the gene expression the main idea is to fit a linear model to the expression values of a gene. Linear models can be applied in case of multi-factor design when multiple condition grouping is present.

The method models the distribution of a response variable y_i in relation to one or more explanatory variables $x_{i1}, \ldots, x_{ip}$ as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i. \tag{3.8}$$

where y_i is a response variable, $x_{i1}, ..., x_{ip}$ are one or more explanatory variables, β_0 is a constant(intercept), $\beta_1, \ldots, \beta_p$ are regression coefficients and ε_i is an error term.

The linear regression model approach designed specifically for differential gene expression analysis, taking into account the nature of gene expression data, is implemented in a widely used software package limma [16]. It was initially designed for microarray data and later was extended for the application on RNA-seq data as well. It fits a linear model for each gene for a given series of arrays, where the coefficients of the model describe the differences between the RNA sources [183]. However, the method does not treat different genes independently, instead, information is borrowed between the genes. It is done by modifying gene-wise variances towards the global variance. This approach reduces the number of false positives for genes with very small variances and improves the possibility to detect differentially expressed genes with large variances [16]. In Chapter 6 we identified differentially expressed genes in normal early neural development and after treatment with toxic compounds, such as well-known anti-epileptic drug valproic acid. Further combination with PCA and functional enrichment analysis of differentially expressed genes allowed to characterize the valproic acid MoA on human cells.

3.2.7. Wilcoxon test

A Wilcoxon rank-sum test (also know as Mann-Whitney-Wilcoxon test) [20, 21, 184, 185] is a non-parametric test, where the test does not assume the samples (groups) to follow the normal distribution. It is applied to two independent data samples to test whether their distributions are identical.

Wilcoxon test can be applied in case-control studies to find out if the difference between measurement, such as the concentration of proteins in plasma or gene expression in two conditions, e.g. disease and healthy individuals, is significant (Chapter 5). The case-control studies of gene expression is a widespread case of a differential expression analysis.

Suppose we have two samples of size n and m from populations X and Y

with the cumulative distribution functions F_X and F_Y. During the testing procedure original data from two samples are pooled together and transformed into ranks where the smallest value gets the rank one and the largest rank $N = n + m$. Let Z be a vector that indicates the rank-order statistics of the combined samples and identifies the sample to which each observation belongs: $Z = (Z_1; Z_2; ...; Z_N)$, where $Z_i = 1$ if the ith random variable in the combined ordered sample originate from X and $Z_i = 0$ if it originates from Y, for $1, 2, ..., N$. After that the sum of the ranks is calculated for both samples as shown in the equation (3.9) [185]. These sums represent Wilcoxon test statistics.

$$W_N = \sum_{i=1}^{N} iZ_i \tag{3.9}$$

We can formulate the null hypothesis that we want to test as both samples come from the same distribution or formally $H_0 : F_Y(x) = F_X(x)$. If the null hypothesis is true, we would expect a similar mixture of ranks in both samples. The possible alternative hypotheses can be stated as following: $H_1 : F_Y(x) = F_X(x - \theta)$. If $\theta < 0$ F_Y is shifted to the left from F_X ($X > Y$) and in case $\theta > 0$ then F_Y is shifted to the right from F_X ($X < Y$). In case we consider two-sided case of Wilcoxon test and formulate our alternative hypothesis as following: $H_1 : F_X \neq F_Y$ ($X \neq Y$), i.e. $\theta \neq 0$.

The test can be concluded by calculating the p-values for the Wilcoxon rank-sum statistic W_{N} as illustrated in Table 3.1 (adapted from [185]).

Table 3.1. Defining p-value for Wilcoxon rank-sum statistic.

Alternative hypothesis	p-value
$\theta < 0$ ($X > Y$)	$P(W_{\mathrm{N}} \geq \mathrm{w_o})$
$\theta > 0$ ($X < Y$)	$P(W_{\mathrm{N}} \leq w_{\mathrm{o}})$
$\theta \neq 0$ ($X \neq Y$)	$2\times$ min$\{P(W_{\mathrm{N}} \geq \mathrm{w_o}); P(W_{\mathrm{N}} \leq \mathrm{w_o})\}$

where w_{o} - observed value of W_{N}

3.2.8. Multiple testing correction

When dealing with biological data, often multiple statistical tests are carried out repeatedly, e.g. to find the differentially expressed genes. The results of such test are concluded by identifying the cases where the p-value is less than some given critical value, usually corresponding to 0.05. A p-value of 0.05 means that there is a 5% chance of getting the observed or more extreme result, when the null hypothesis is true.

Multiple testing leads to the potential increase of false positive results, i.e. incorrect rejection of H_0 hypothesis. If multiple statistical tests are carried out

for the data for which null hypothesis holds, then about 5% of them will end up rejecting the null hypothesis generating false positive results.

Currently there is no generally accepted universal method to deal with the problem of multiple comparisons. However, a few approaches, e.g. a number of false discovery rate-based methods [186–190] and Bonferroni correction [191], are actively used to account for a false positive rate, also known as type I error [192]. These approaches are known as multiple testing correction procedures. False discovery rate (FDR) reduction based on Benjamini-Hochberg procedure [187, 188] is one of the popular methods that controls the expected proportion of incorrectly rejected null hypotheses in a list of rejected hypotheses. It helps to lower the number false positives among the significant results. The procedure converts tests' p-values into q-values (see algorithm 3.10), i.e. adjusted p-values. During this procedure original m p-values are ranked in an ascending order, i.e the smallest rank 1 is assigned to the smallest p-value and rank m to the largest. Then the q-values are computed using the equation (3.10). These corrected p-values are then compared with chosen critical value to identify the significant results [193].

$$q_i = min\{p_i * (m/i); q_{i+1}\} \quad for\ i = m-1; m-2; ...; 1 \tag{3.10}$$

In equation (3.10) m is the number of tests and i is the rank of the original p-value, $p_1 ... p_m$ are ordered p-values and $q_m = p_m$ is the largest adjusted p-value.

3.2.9. Functional enrichment analysis

Functional enrichment analysis is applied to characterize a group of genes using currently available domain knowledge. The aim of this analysis is to identify specific groups of genes that are over-represented in a large set of genes or proteins, and may have an association with a studied phenotype, e.g. disease. Often as a result of bioinformatics analysis, e.g. differential expression analysis, a group of genes is identified to be connected with the studied biological process or condition. In order to understand better the underlying biological process and the role that this set of genes or proteins play in these process, functional profiling of the gene set is performed. Functional profiling of the selected gene set would reflect mutual function(s) and the biological processes or pathways where this group of genes is involved.

Functional enrichment analysis uses Gene Ontology (GO) [160] as the main source of information about standardized annotation of gene products. Additionally, other databases as KEGG [141, 142], Reactome [143, 144] or The Human Protein Atlas [194] can be included for characterization.

The analysis is performed by identification of the enriched GO terms or pathways in a list of genes. The number of overlapping genes is found between the list of interest and the reference gene lists corresponding to GO category or pathway. The significance of the findings is assessed by computing the p-values from Fisher exact test. The approach is implemented in various tools including web-tools such

as g:Profiler [155]. In Chapter 6 we demonstrate how functional enrichment analysis can be applied in toxicology studies as a part of multi-staged data integration helping to provide more information about the results detected by differential expression analysis and PCA.

3.3. Transformation-based biological data integration

Currently in the field of *omics* integration there is an open question what integration methods should be used to provide flexible and effective way to describe biological system on the different levels. The answer to this question is highly dependent on the available *omics* data sets.

In the transformation-based integration approach multiple data sets are combined after transforming each data set into an intermediate form, such as a graph [195] or a matrix kernel [196, 197]. Application of kernel-based methods for *omics* integration is out of the scope of this work. In the current thesis we focus on the graph-based transformation methods.

The main idea of this integration technique is a representation of each individual data set as a graph and further combination of these graphs into one unified graph. Each individual graph $G = (V, E)$ is described by a set of vertices V, a set of edges E and an adjacency matrix A. In this graph $v_i \in V$ denotes a node and $e_{ij} = (v_i, v_j) \in E$ denotes an edge. The adjacency matrix A is $N \times N$ matrix with $A_{ij} = w_{ij} > 0$ if $e_{ij} \in E$ and $A_{ij} = 0$ if $e_{ij} \notin E$ [198]. These individual graphs are constructed from biological *omics* data sets originating from lab experiments or computational analysis.

The nodes in these graphs represent biological entities, e.g. genes or proteins, and the edges depict the type of biological interaction between a pair of nodes, e.g. co-expression or physical binding between proteins. The nodes and the edges can carry additional information about the nature of interaction introduced as the corresponding node and edge attributes, also known as features in the machine learning terminology (Figure 8). For example, node attributes can be described by a feature matrix $X \in R^{N \times D}$ where $X_i \in R^D$ representing the feature vector of node v_i, N is the number of nodes and D is the number of features.

An example of node attributes can be values describing a level of gene expression of the node in the disease, its alternative name, biological function, etc. The edge attributes can reflect information about its direction, type of interaction, interaction detection method, etc. An edge can also have a weight that describes some property of the connection between a pair of nodes i and j. For example, an edge can describe the strength of physical binding between two proteins. For the purpose of network analysis a set of node or edge attributes can be represented as a corresponding feature vectors.

The transformed data sets, i.e. individual graphs, can be simultaneously combined constituting a heterogeneous graph (Figure 9) that is described in Section 3.3.1. After that an appropriate analysis methods, e.g node classification, com-

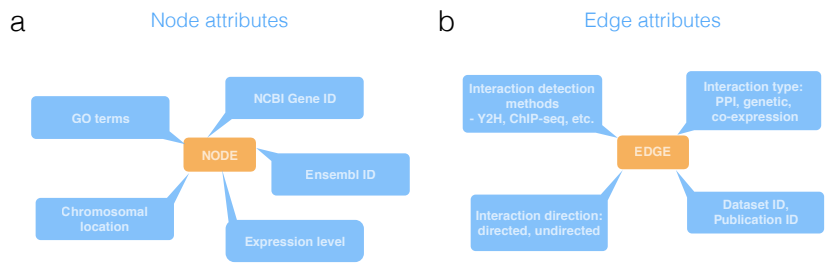


Figure 8. Example of node and edge attributes. Figure illustrates: a) A set of node attributes; b) A set of edge attributes.

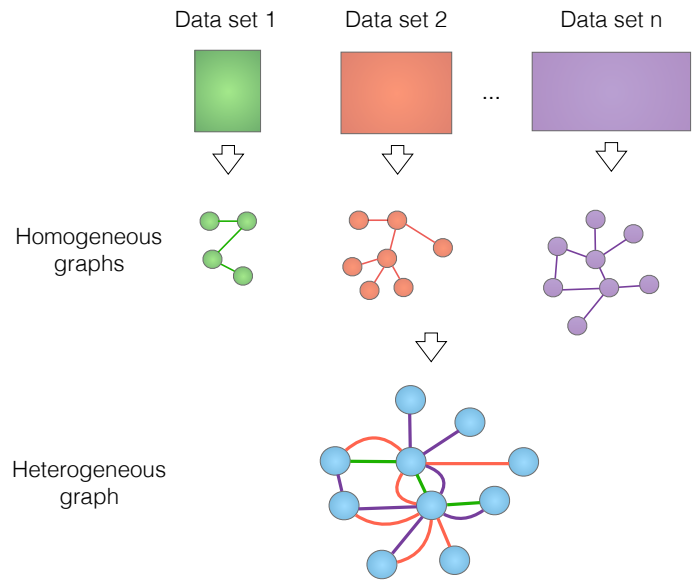


Figure 9. Combination of individual graphs into unified heterogeneous graph.

munity detection methods or link prediction methods can be applied on the transformed data (Figure 7 b). An advantage of this approach is that it preserves data type specific properties during the integration process [15]. It can be applied to integrate many types of data as long as data sets contain a common unifying feature such as a common gene name, patient identification number, etc.

3.3.1. Heterogeneous graphs

The result of transformation-based integration can be represented as a heterogeneous biological graph.

We have adapted the definition of a heterogeneous graph proposed by Lee et al. [24]. A *heterogeneous graph* is defined as $G = (V, E)$ consisting of a set of node objects V and a set of edges E connecting the nodes in G. A heterogeneous graph

also has a node type mapping function $\Theta : V \rightarrow T_V$ and an edge type mapping function defined as $\xi : E \rightarrow T_E$ where T_V and T_E denote the set of node types and edge types, respectively. The type of node i is denoted as $\Theta(i)$, e.g. a gene, a protein, or SNP in a heterogeneous biological network, whereas the type of edge $e = (i, j) \in E$ is denoted as $\xi(i, j) = \xi(e)$, e.g. a co-expression, physical binding, etc.

A homogeneous graph is simply a special case of a heterogeneous graph where $|T_V| = |T_E| = 1$.

As we briefly mentioned in the previous section additional information about nodes and edges is often available from the experiments, databases and in a form of domain expert knowledge (Figure 8). These attributes can be incorporated into the graph for further analysis. For the convenience let us introduce the notation of an attributed graph as following:

Let $G = (V, E, X, Y)$ be an *attributed graph* where V is a set of nodes, E is a set of edges, N and M are the numbers of nodes and edges in the graph. X is an $N \times D$ matrix of node input attributes where each x_i is a D-dimensional (row) vector of attribute values for node $v_i \in V$ and x_j is an N-dimensional vector corresponding to the jth attribute (column) of X, analogically, Y may represent a $M \times D$ matrix of edge input attributes.

3.3.2. Analysis of heterogeneous graphs

Graph-structured data emerge naturally in many different domains including biology. Many useful insights about the data can be obtained from using the graph structural properties as shown in the growing body of research focused on graph mining [35, 199]. However, in the real world examples graphs can be large, contain many different complex structural patterns, and be noisy, i.e. contain edges that do not represent real life relations, making effective graph mining to be a non-trivial task [24].

3.3.2.1. Node Embeddings

Graph algorithms and approaches have become popular both in biology and in social studies due to the advantages of data representation in a network format [200]. Modern machine learning applications are using network-structured data to identify patterns and make predictions. To process network data effectively, the first critical challenge is to find appropriate network data representation and incorporate graph structure into a feature vector for further application of analysis methods [201]. These tasks are addressed by a field of *representation learning* that aims for automatic feature learning from the given data [32, 202]. Representation learning allows the replacement of manual feature engineering in machine learning tasks, e.g. when building classifiers or other predicting models [202].

Depending on the study domain and the machine learning task, such as link prediction or node classification, various structural graph properties can be incor-

porated into a machine learning model, e.g. information about local node neighbourhood. The major challenge is to find an effective way to encode these high-dimensional graph structural properties into a feature vector. Traditional machine learning approaches in biological and social network analyses have utilized hand-crafted graph features based on the global and local graph structural properties, for example, clustering coefficients, node degrees or other carefully engineered features. However, the downside of such approaches is that hand-crafted features are not able to adapt during the learning process and creating these features can be time consuming.

These issues required development of novel network representation methods that aim to learn low-dimensional vector representations for network nodes, i.e. embeddings, that encode information about graph structural properties [35, 202]. The objective is to map original node vectors to d-dimensional embeddings so that similar nodes in the graph are embedded close together (Figure 10). In this context node similarity is not defined by only topological close location in the graph, but rather by various node features, i.e. attributes. Once the embeddings are learned, they can be used in downstream analysis tasks.

Given a graph $G = (V,E)$ with V as the node set and E as the edge set, the objective of node embedding is to learn a function $f : V \rightarrow R^d$ such that each node $i \in V$ is mapped to a d-dimensional vector z_i where $d \ll |V|$ and $z_i \in R^d$, where R^d is a d-dimensional space. The node embedding matrix is denoted as Z. The learned node embeddings given as output can subsequently be used as input to mining and machine learning algorithms for a variety of tasks such as link prediction, node classification [35], community detection, etc.

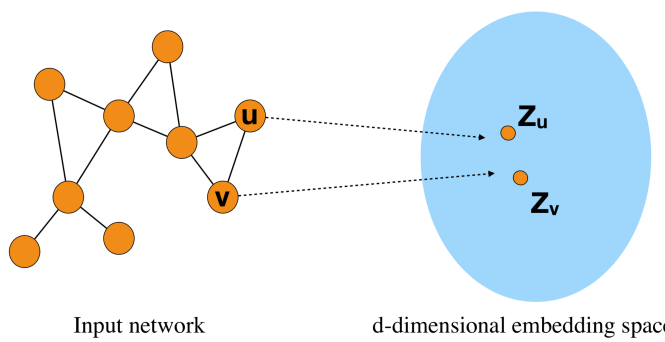


Figure 10. Low-dimensional vector representations for network nodes.

3.3.2.2. Graph convolutional networks

In the recent years deep learning methods have gained increasingly critical role in various domains, being used to learn useful low dimensional representations of images, text, videos, etc. [203–207]. These methods were inspired by the organization of the animal brain neural networks. They are built using artificial neurons (Figure 11 *A*), i.e. an approximation of a biological neuron, as computational units [208, 209]. Artificial neural networks consist of stacked layers of interconnected neurons (Figure 11 *B*), and the depth of the network corresponds to the number of hidden layers. The network receives a data at the input layer and then transforms it using non-linear function at each hidden layer. Each neuron at the hidden layer calculates the weights the input from all the neurons at the previous layer and sums them and then applies non-linear activation function, e.g. ReLU, to compute an output [25]. Training procedure starts by setting initial weights to small random numbers. Learning process involves minimization of a loss function $L(w)$ that computes the difference between the predicted value or label and a true label [25, 209]. In order to do that the internal weights of neural networks should be modified using some optimization technique. It is typically performed by using gradient descent technique [210]. The loss is backward propagated to the network resulting in small changes in the weights. The derivative (gradient) of the loss function dictated the weight update $w_{new} = w_{old} + \eta \Delta w$. The learning rate η is introduced as a small constant to ensure the smoothness of weight updates.

Graph neural networks (GNN) are deep learning methods that extend the application of existing neural networks to analyze graph-structured data [206,207,211]. Initially proposed by Scarselli et al. [212], a concept of GNN was later developed into a variety of methods [198, 207]. These methods can be divided into subgroups by the type of propagation rule, training method and a type of graphs that method can be applied to (Figure 12). In the current thesis we will focus on type of GNNs called graph convolutional networks (GCNs). GCNs is a set of new deep learning methods inspired by the application of well-established convolutional neural networks (CNNs) [205,213,214]. CNNs allow to reduce the number of model parameters compare to a fully connected network by applying convolutional operation to a small region of the input at a time and sharing the model parameters between regions [25, 205, 214]. However, unlike effective application to image, text and video types of data, due to complexity of graph-structured data, i.e. topological structure, lack of spatial locality like in grids, lack of fixed node ordering or reference point (Figure 13) and often multimodal features, these methods can not be directly applied on the graph data.

GCNs generalize an operation of convolution from traditional data such as images or grids to graph data [34, 35, 215, 216]. These methods leverage both information contained in the nodes and information contained in the relationships between the nodes. Firstly, we will provide a general description of GCN methods, and later introduce the particular cases, i.e. GraphSAGE and HinSAGE. The

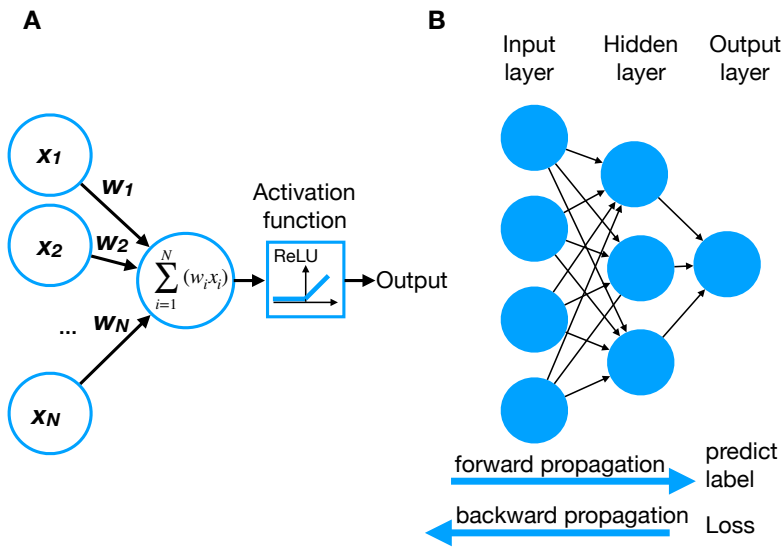


Figure 11. Artificial neural networks. *A*. Schematic representation of the artificial neuron. The unit receives input x_i with associated weight w_i from N neurons. The total input to a unit is the weighted sum over all inputs. *B*. Schematic drawing of artificial neural network with four inputs at the input layer and three neurons at the hidden layer. The loss is computed based on the true and predicted labels and backward propagated to the network.

central idea of GCN approaches is to learn a function f to generate a node v_i representation by aggregating its own features X_i and features of neighbouring nodes X_j , where $j \in N(v_i)$. Borrowing the concept of a convolution filter for image pixels, the main analogy between CNNs and GCN mechanisms can be explained as the aggregation of information from the neighbouring pixels or nodes into a low-dimension representation [198, 207] (Figure 14).

GCN aims to learn a state embedding $h_v \in R^d$, i.e node representation, that contains neighborhood information for each node. This state embedding h_v (equation 3.11) is a d-dimensional vector of node v that can be used to produce an output o_v (equation 3.12), e.g. the node label [207]. In the equation (3.11) f is a parametric function that updates node state based on the input neighborhood, x_v, $x_{ed[v]}$, $h_{ne[v]}$, $x_{ne[v]}$ are the features of v, features of its edges, states, and features of neighborhood nodes correspondingly. In equation (3.12) g is a local output function that describes in what form the output is produced [207].

$$h_v = f(x_v, x_{ed[v]}, h_{ne[v]}, x_{ne[v]}) \tag{3.11}$$

$$o_v = g(h_v, x_v) \tag{3.12}$$

In a compact form vectors H, O, X_E, X corresponding to all states, outputs,

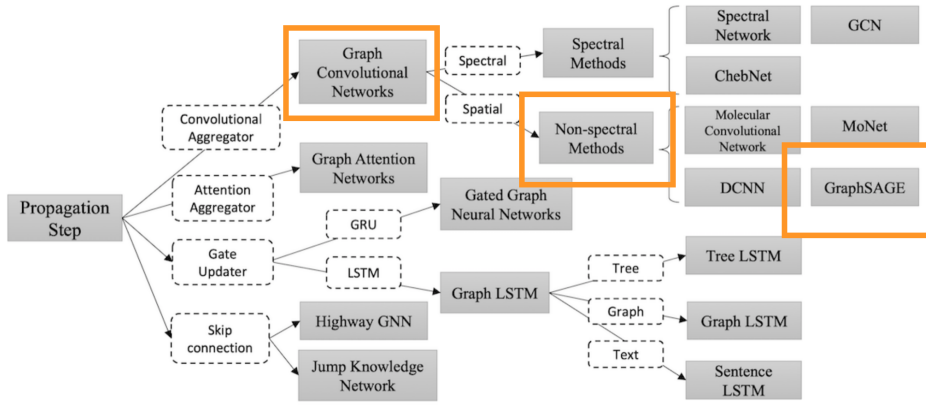


Figure 12. An overview of variants of graph neural networks. Image is adapted from Zhou et al. [207]. Figure demonstrates classification of GNN-based model types. A branch of methods indicated by the highlighted boxes is considered in the current thesis.

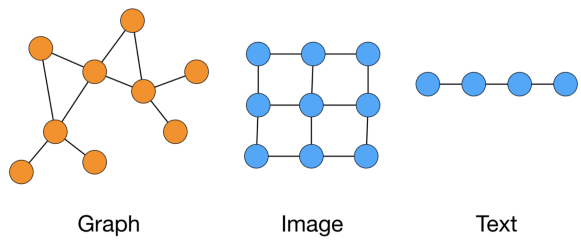


Figure 13. Differences in structure between graph, image and textual data.

edge and node features constructed by stacking can be compactly represented by equation (3.13) and equation (3.14), where F is a global transition function, and G is a global output function are stacked versions of f and g for all nodes in a graph [207].

$$H = F(H, X_E) \tag{3.13}$$

$$O = G(H, X) \tag{3.14}$$

The states are iteratively updated based on the states in the previous layer (see equation (3.15)).

$$H^{k+1} = F(H^k, X) \tag{3.15}$$

The computations described by f and g can be interpreted as feed forward neural networks with learnable parameters. In node classification task GCN (Figure 15) is trained on the labeled nodes of the graph, effectively propagating the node label information to unlabelled nodes by updating weight matrices W that are shared

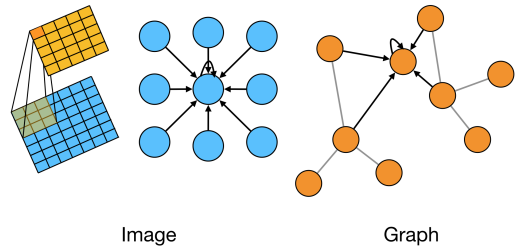


Figure 14. Low-dimensional representation analogy between images and graphs. In case of images the features of neighbouring pixels are used to create a representation, in case of networks - features of the neighbouring nodes.

across the nodes. Learning minimizes a loss function (3.16) [207] where p is the number of supervised nodes.

$$loss = \sum_{i=1}^{p} (t_i - o_i) \tag{3.16}$$

Similarly to CNNs methods applied for images, video and text, the parameter learning algorithm in GCNs is based on a gradient-descent [210] and can be generally described as following [34, 198, 207]:

- Perform forward propagation through the GCN. The states h_v^k are iteratively updated.
- Compute the cross entropy loss on known node labels. Backpropagate the loss. The gradient of weights W is computed from the loss.
- Update weight matrices W at each layer. The weights W are updated according to the gradient computed in the last step.

There are two sub-classes of GCNs (12), i.e. spatial and spectral [34,198,207]. Spectral-based approaches adapt graph signal processing methods [217] and define graph convolutions as application of filters, i.e. convolution operation is interpreted as removing noise from graph signals [198, 207, 215]. Spatial-based approaches denote graph convolutions as aggregating feature information from the neighbors. In this work we consider only spatial type of GCNs such as GraphSAGE and HinSAGE [35, 218]. GraphSAGE (SAmple and aggreGatE) proposed by Hamilton et al. [35] is a spatial-based GCN model that addresses several major issues that other GCN approaches suffer from. The most relevant to this work is scalability - most of the methods can not be applied to large graphs due to computational issues, while GraphSAGE is scalable to graphs with billions of nodes [202,211]. The second issue is the ability to work in an inductive setting as opposite to the transductive one [34, 207, 219]. Most of the approaches work with stationary graphs and are not generalizable in case new node is added to the network. In a nutshell the transductive learning does not generalize to unseen nodes, while inductive GraphSAGE can generate node embeddings, i.e. node vector representations, for nodes that were not seen during the training. The key idea that

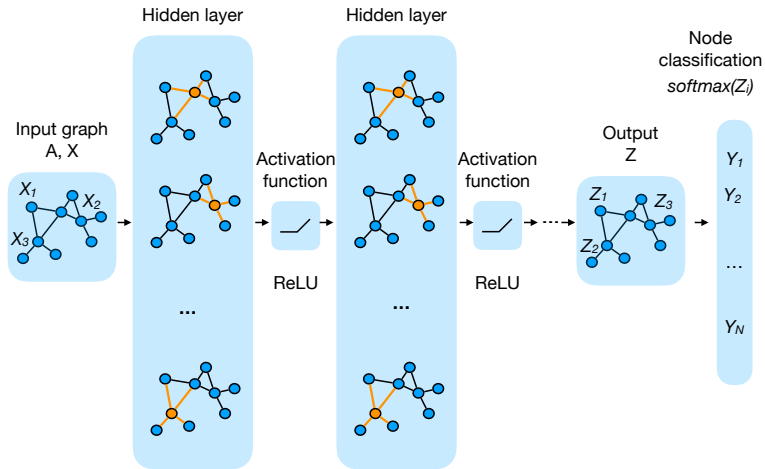


Figure 15. Schematic depiction of multi-layer GCN for node classification task. Input graph is characterized by adjacency matrix A and a matrix of node attributes X. The graph structure is shared over layers, labels are denoted by Y_i.

differs GrapgSAGE from other methods is that it generates node embeddings at each state by sampling and aggregating features from a node local neighborhood. However, it does not use the full set of neighbors, but a fixed-size set of neighbors uniformly sampled from the neighbouring nodes at each neighbourhood with the depth K, e.g. 1-hop, 2-hop, etc (Figure 16). These embeddings can be considered as a vector containing hash values describing the representation of a node v. This way in GraphSAGE node local neighbourhood defines a computational graph.

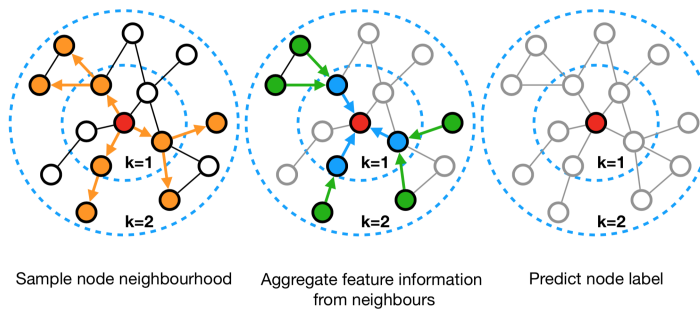


Figure 16. GraphSAGE sample and aggregate approach. Figure is adapted with modifications from Hamilton et al. [35].

Generally GraphSAGE learning process can be described as following:

- Sample head node local k-hop neighborhood $N(v)$ with fixed sample size.
- Derive node final embedding z_v at layer K by aggregating its neighbors features (equation 3.17).

- Use node embedding on the final layer to make predictions, compute loss and backpropagate errors to update weight matrices W_k.

$$\begin{aligned} h^k_{N(v)} &\leftarrow AGGREGATE_k(h^{k-1}_u, \forall u \in N(v)) \\ h^k_v &\leftarrow \sigma(W_k \cdot CONCAT(h^{k-1}_v, h^k_{N(v)}) \\ z_v &\leftarrow h^K_v \end{aligned} \tag{3.17}$$

GrapSAGE proposes three possible aggreator functions, i.e. *mean aggregator*, *LSTM aggregator* and *pooling aggregator*. Mean aggregator computes an element-wise mean of vectors $h^{k-1}_u, \forall u \in N(v)$, LSTM aggregator applied LSTM method-based [220] to a random permutation of the node's neighbours, finally when pooling aggregator is used, neighbors' vectors are fed through a fully-connected neural network independently followed by an element-wise max-pooling operation to aggregate information across the neighbors [35]. The embedding generation process or forward propagation algorithm using mean aggregator is formally described by the Algorithm 1. This algorithm assumes that a set of weight matrices $W_k, \forall k \in 1, ..., K$ is already learned [35]. At the fist step of the algorithm the node embeddings are initialized as the input node attributes. i.e. features, x_v (Figure 17). Then at each iteration of the algorithm embeddings of the node's neighbours are aggregated.

After feature vectors of the neighbouring nodes are aggregated, GraphSAGE concatenates the node current representation, $\mathbf{h}^{k-1}_v$, with the aggregated neighborhood vector, $\mathbf{h}^k_{N(v)}$ (Algorithm 1, Figure 17). This concatenated vector is fed through a fully connected layer with nonlinear activation function σ, which transforms the representations for the next step of the algorithm $\mathbf{h}^k_v$, $\forall v \in V$. On the next iterations the model adds information from the further neighbourhoods to the embeddings. However, the dimensionality of the embeddings remains constrained, i.e. mapping function compresses the neighbourhood information into the low dimensional vector. Through the incorporation of node features to the model, GraphSAGE also learns the topological structure of each node neighborhood as well as the distribution of node features in the neighborhood. The main advantage of the GraphSAGE algorithm is that the trainable parameters are shared across the nodes, i.e. the same aggregation functions and weight matrices W_k are used for the generation of the embeddings for all nodes. During this process only node neighbourhood structure and node input attributes are changing. This way GraphSAGE provides regularization and allows to generate embeddings for the nodes that have not been seen during the training [35]. The details of GraphSAGE loss computation and application of the model to node classification task are described in Section 3.3.2.4.

GraphSAGE uses a uniform sampling function and samples neighbourhood nodes with replacement in case the sample size is larger than a node degree. The size of a sample at each layer is defined by the user. Different uniform samples $N(v)^k$ are drawn at each iteration k in Algorithm 1. Without such sampling ex-

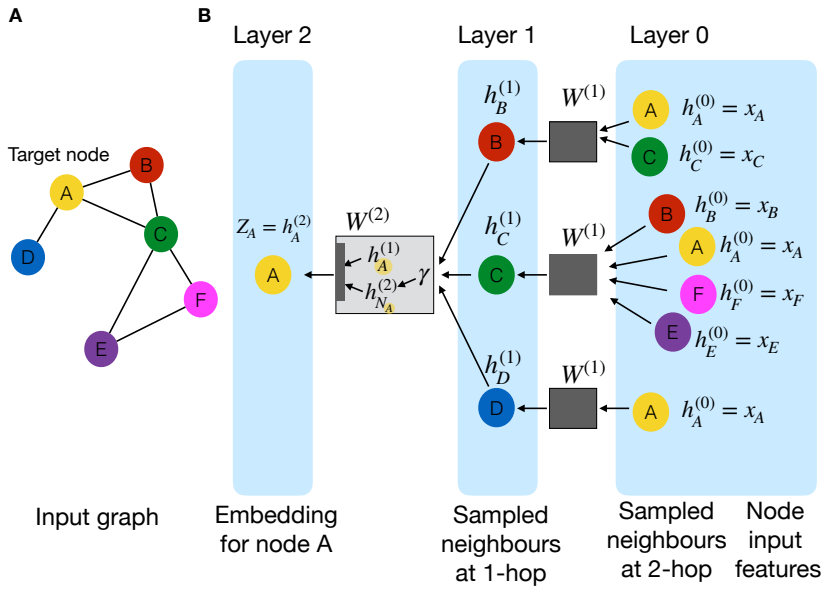


Figure 17. Information propagation in GraphSAGE algorithm from the local node neighbourhood (figure is adapted with modifications from Hamilton et al. [202]). **A** Illustration of the input graph. An example demonstrates computation of embedding for a node A. **B** Embedding for a node A is generated using neighbouring nodes' embeddings from 2-hop neighbourhood (K=2 in Algorithm 1). Grey boxes represent aggregation functions γ, weight matrices W_k and fully connected layer with nonlinear activation function σ.

pected runtime of a single batch is unpredictable and in the worst case $O(|V|)$. In contrast, the per-batch space and time complexity for GraphSAGE is fixed at $O(\prod_{i=1}^{K} S_i)$, where $S_i, i \in 1, ..., K$ and K are user-defined constants representing sample sizes at each layer [35, 198, 207].

GraphSAGE is developed for homogeneous feature-rich graphs, for example, biological graphs with a number of node attributes representing gene meta data (see Figure 8a). However, real life examples often imply working with heterogeneous graphs, that, for example, can contain nodes and edges of multiple types (Section 3.3.1). In the next Section 3.3.2.3 we will introduce extension of GraphSAGE model to heterogeneous graphs.

3.3.2.3. GCN extension for heterogeneous graphs

A HinSAGE [218] is an algorithm from the family of spatial GCNs - a generalisation of the GraphSAGE algorithm [35] for heterogeneous networks. HinSAGE is designed for the heterogeneous networks that can contain multiple types of edges and nodes, however, in this work we restrict the description only for the network with multiple edge types. For example, in the context of biological networks dif-

Algorithm 1: GraphSAGE embedding generation algorithm [35] (forward propagation) with Mean Aggregator.

Input: Graph $G(V,E)$; input features $\{\mathbf{x}_v, \forall v \in V\}$; depth K, weight matrices $\mathbf{W}^k$, $\forall k \in \{1,\ldots,K\}$; non-linearity σ; neighborhood function $N : v \rightarrow 2^V$

Output: Embeddings z_v for all $v \in V$

$\mathbf{h}_v^0 = \mathbf{x}_v, \forall v \in V$;
for $k = 1 \ldots K$ **do**
  for $v \in V$ **do**
    $\mathbf{h}_{N(v)}^k \leftarrow \frac{1}{|N(v)^k|} \sum_{u \in N(v)} (\mathbf{h_u}^{k-1})$;
    $\mathbf{h}_v^k \leftarrow \sigma(\text{CONCAT}(\mathbf{W}^k \mathbf{h}_v^{k-1}, \mathbf{W}^k \mathbf{h}_{N(v)}^k))$
  end
  $\mathbf{h}_v^k \leftarrow h_v^k / \|h_v^k\|_2, \forall v \in V$;
end
$\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in V$

ferent edge types can represent biological relations, i.e co-expression, physical interactions between the proteins. The main difference from the homogeneous GraphSAGE [35] is that HinSAGE takes into account individual edge types $t \in T$, where T is a set of all edge types in the given graph (Figure 18).

Recall that GraphSAGE forward propagates node information across the graph to compute node embeddings (see Algorithm 1). The intuition behind is that at each iteration the search depth for a node increases, aggregating information from further neighborhood of the node. HinSAGE operates in a similar manner, but models each edge type t separately (see Algorithm 2). It models the representation of the sampled neighbouring nodes at each layer for each edge type and combines it with the nodes representation from the previous layer. This concatenated vector is fed through a fully connected layer with nonlinear activation function σ similarly to GraphSAGE. In the Algorithm 2 $N_t(v)$ is a neighbourhood of node v connected by edges of type t, W^k and $W_{t,neigh}^k$ are trainable weight matrices, where the number of $W_{t,neigh}^k$ equals to the number of edge types per layer.

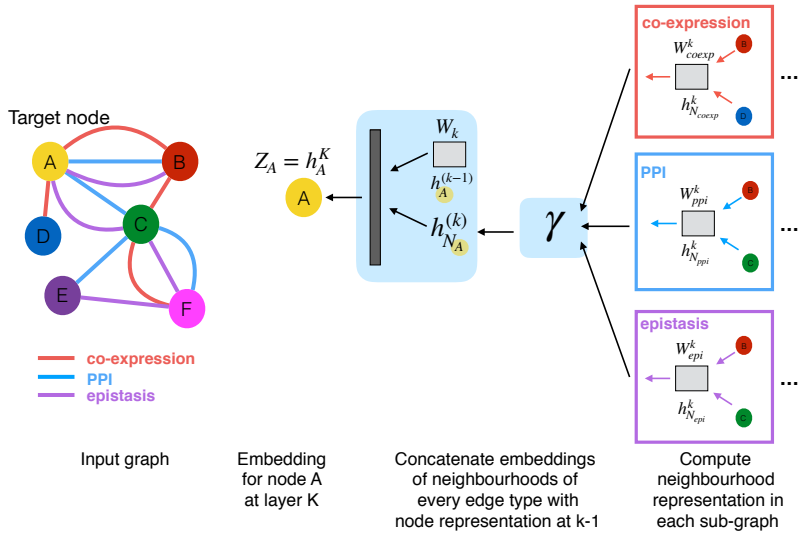


Figure 18. Generation of node embeddings in heterogeneous graph. HinSAGE algorithm generates node embeddings for each sub-graph, consisting of the same type of edges, and concatenates them similarly to GraphSAGE algorithm. Figure demonstrates an example of biological heterogeneous graph where nodes represent genes, and edges of different colors depict biological relationships, such as co-expression, PPI and epistatic interactions. Neighbourhood states of co-expression, ppi and epistasis sub-graphs are aggregated by function γ. The aggregated state of the neighbourhood $h_{N_A}^k$ is then concatenated with node A representation on the previous (k-1) layer and fed through fully connected layer nonlinear activation function σ.

Algorithm 2: HinSAGE embedding generation algorithm [218].

Input: Graph $G(V,E,T)$; edge type mapping $\xi : E \rightarrow T$, input features $\{\mathbf{x}_v, \forall v \in V\}$; depth K, weight matrices $\mathbf{W}^k$, $\mathbf{W}_{t,neigh}^k$ $\forall k \in \{1,\ldots,K\}$ and $\forall t \in T$; non-linearity σ; neighborhood function for the relationship t: $N_t : v \rightarrow 2^V$

Output: Embeddings $\mathbf{z}_v$ for all $v \in V$

$\mathbf{h}_v^0 = \mathbf{x}_v, \forall v \in V$;
for $k = 1 \ldots K$ **do**
  for $v \in V$ **do**
    for $t \in T_E$ **do**
      $\mathbf{h}_{N_t(v)}^k \leftarrow \frac{1}{|N_t(v)^k|} \sum_{u \in N_t(v)} (\mathbf{h_u}^{k-1})$;
    end
    $\mathbf{h}_v^k \leftarrow \sigma(\text{CONCAT}(\mathbf{W}^k \mathbf{h}_v^{k-1}, \frac{1}{|T(v)|} \sum_{t \in T} (\mathbf{W}_{t,neigh}^k \mathbf{h}_{N_t(v)}^k)))$
  end
  $\mathbf{h}_v \leftarrow h_v^k / \|h_v^k\|_2, \forall v \in V$;
end
$\mathbf{z}_v \leftarrow \mathbf{h}_v^K, \forall v \in V$

3.3.2.4. Node classification

Recent advances in network analysis demonstrate that methods based on GCNs outperform other traditional methods for many classification tasks on network data [34, 35]. For node classification task, such as node classification of genes into disease/non-disease classes, the major idea of the method is to train a model on a small subset of nodes, and then apply a classifier to the nodes with unknown labels [202]. The classification is performed by comparing node embeddings and computing a loss between the true-labeled and predicted output. GraphSAGE and HinSAGE use affinity scores to calculate loss. The affinity score between two given nodes can be defined as $A(z_u, z_v) = z_u^T z_v$. This is the cosine similarity between the two embeddings z_u and z_v [204]. The learning task is to maximize the score between nodes similar nodes, and minimize the score of the less similar pairs in the embedded space. Cross-entropy loss can be used to estimate performance of classification where the output is a probability value between 0 and 1. Binary cross-entropy loss for a node can be computed as following [35, 204]:

$$L(z_u) = -log(\sigma(z_u^T z_v)) - Qlog(\sigma(-z_u^T z_{v_n})) \tag{3.18}$$

In equation (3.18) where Q is an optional weight that defines the number of negative samples. This loss function implies that closer nodes would have similar embeddings, while nodes located further from each other in the projected space. Alternatively, the equation for loss function can be written in a more conventional way using label binary classification label, $y_i \in Z$, associated with each node. To learn to map nodes to their labels, embedding vectors, z_i, are fed through a logistic, or sigmoid, function $\hat{y}_i = \sigma(z_i^T \theta)$, where θ is a trainable parameter vector. The cross-entropy loss increases as the predicted probability diverges from the true label. Expression (3.18) can be rewritten as following :

$$L = \sum_{v_i \in V} y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)) \tag{3.19}$$

During the model training the gradient computed according to this equations (3.18) and (3.19) can be back-propagated to optimize model parameters W_k using stochastic gradient descent [35, 202, 210].

3.3.3. Summary

In this Chapter we have described a the transformation-based biological data integration approach and state-of-the-art deep learning algorithms for network-structured data. We have introduced a concept of representation learning in the context of graph-related machine learning tasks. We have also demonstrated how GCN approach can be extended to heterogeneous graphs and how it can be applied for node classification task.

4. INTEGRATING HETEROGENEOUS DATA SETS RELATED TO ALZHEIMER'S DISEASE (PUBLICATION I)

One of the emerging needs and associated challenges in the biological sciences at the present moment is methods for effective combination of the heterogeneous data sets, such as transcriptomics, proteomics or GWAS, describing the biological process or a disease at the individual levels [12, 15, 221, 222]. The integrated reliable data sources could be further explored by the scientific community and analysed together with other complimentary data sets. Well-designed and thoroughly described integrated data sets can be repurposed for a variety of applications, e.g. to generate or test novel hypotheses, to increase the reproducibility of research and provide general advancement in clinical and biological understanding of the disease through the analysis [223].

In this chapter we will describe the data integration approach developed for the combination of the heterogeneous data sets related to Alzheimer's disease. We will also demonstrate how transformation-based integration uncovers disease mechanisms that would not be detected otherwise. This approach can be applied in the integration tasks related to other diseases and biological processes.

4.1. Bringing together disparate data sets related to Alzheimer's disease

Alzheimer's disease (AD) is a neurodegenerative disorder that progresses with age and is accompanied by a severe loss of memory and other cognitive abilities. Currently there are no effective treatments that are able to stop the disease completely. Furthermore, the development of the effective treatments is delayed because of an insufficient understanding of the disease at the systemic level in human.

Various experiments have been performed to capture the underlying disease mechanisms and identify disease biomarkers and potential new drug targets [111, 224–227]. Such experiments are carried out by the scientists working in different domains, e.g. proteomics, genomics, clinical diagnostics, etc. The results of such studies are deposited in the databases designed for collecting the data of similar types, e.g. IntAct [3, 4], ArrayExpress [1, 2], ADNI [5]. However, in order to obtain a comprehensive systematic view on the disease from these individual complementary data sets we need to integrate them in an effective way (Figure 19).

This task was addressed by AgedBrainSYSBIO[1] consortium. We formulated the aim of our study as following:

[1] AgedBrainSYSBIO - is a four year collaborative research project funded by the European Commission under the Health Cooperation Programme of the 7th Framework Programme (FP7); Grant Agreement No 305299; http://www.agedbrainsysbio.eu/

- To create a collection of reliable data sets of various experimental and computational origins, including both publicly available and newly generated disease-specific data sets.
- To develop an integrative approach for combining these data sets in a consistent manner.
- Investigate if application of the state-of-the-art machine learning methods on the integrated data will identify disease related mechanisms that would not be discovered otherwise.

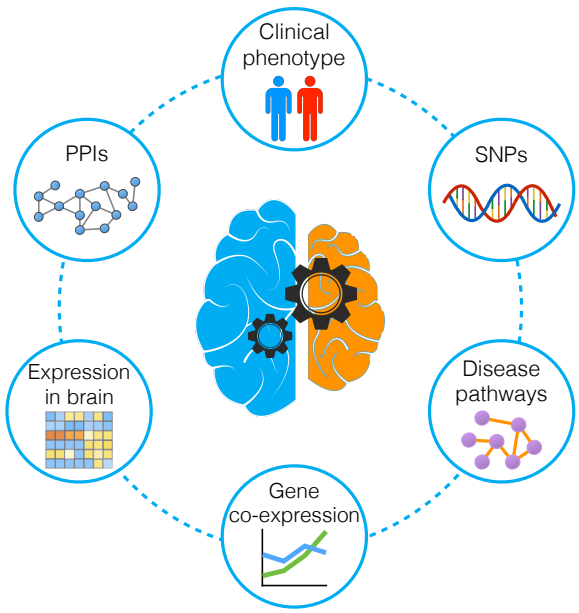


Figure 19. Data integration in Alzheimer's disease. Diagram represents the enrichment of knowledge about Alzheimer's disease by combing genomics, proteomics and clinical phenotype data.

4.2. Collecting and generating data sets related to Alzheimer's disease

The success of the data driven research relies highly on the quality and richness of data collections that researchers are working with. In the study of Alzheimer's disease we have applied systems biology approach to observe, combine and understand disease mechanism on various molecular levels. We have generated and collected 64 data sets of 6 data types originating from 9 data sources. Data sets included experimental and computational origins. Collection contains such data types as protein-protein interactions, gene co-expression in disease and healthy samples, gene co-expression in disease-related brain regions, epistasis, genome-wide association studies, gene expression in the whole brain regions, and positive

Darwinian selection data. For the reader's convenience the detailed description of each data type is provided in Section 2.2.

These data sets come from publicly available repositories as well as are generated by consortium members. For the combination of individual homogeneous co-expression data sets at the stage of data generation we performed multi-staged integration, when individual data sets were analysed and then the joint conclusion was drawn by the integration of the individual results. For example, we have computed co-expression in AD and healthy samples in microarray data sets coming from different individual experiments. In order to combine results of individual analyses, we have applied RRA method [120] (see Chapter 3 for details) to find genes with similar expression patterns over all individual experiments. Integration of such data can reveal functional relations between the genes.

To ensure the quality of the data collection we performed rigorous filtering and quality control of the combined data sets. Every interaction in the individual data sets contains a value (score) stating its importance. For example, the p-value in GWAS and epistatic data sets or MI score [228] for the PPI data sets from IntAct database [3,4]. We filtered the data sets using an appropriate threshold for each of the score type to ensure that only reliable interactions from each of the data sets could be integrated, i.e. we only kept the data record with p-value ≤ 0.05 and MI score ≥ 0.45.

The selected data are heterogeneous in their nature which makes it difficult to understand the relationships between the individual data types as well as relationship with clinical phenotype, i.e. AD. To obtain a comprehensive picture of the interactions within Alzheimer's disease we have applied transformation-based data integration approach [15].

4.3. Transformation-based data integration

Transformation-based data integration approach implies the combination of data sets of multiple data types after transforming each of them into an intermediate form, such as a graph [15]. This approach can be used to integrate many types of data as long as the data contain a common unifying feature.

Collected data sets can be divided into two groups - data sets describing biological relation, i.e. co-expression, PPI, epistasis, and data sets containing specific information about genes or SNPs association with the disease, i.e. GWAS or positive selection data sets.

In our work we represent each individual interactions data set as a graph (Figure 20), where nodes stand for proteins, genes, SNPs etc., and where edges illustrate the relationships between the entities e.g. protein-protein interaction, co-expression, epistatic interaction. Multiple individual graphs are then combined into one graph constituting HEterogeneous Network-based data set related to Alzheimer'd disease (HENA).

We use genes as a unifying feature to which all nodes, e.g SNPs, transcripts

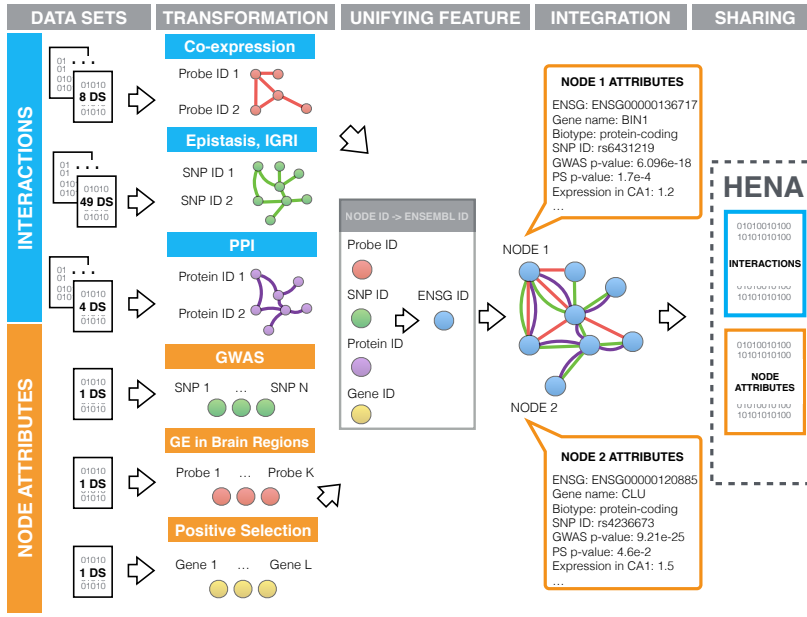


Figure 20. Transformation-based data integration pipeline. Preprocessed data sets contain information about the interactions between genes, proteins, SNPs, and information characterizing them, e.g. node attributes. Interaction data sets contain PPI, co-expression and epistatic interactions and inter-genic regions interactions (IGRI) as its sub-type. Node attributes originate from GWAS, positive selection and gene expression in brain regions from Allen Brain Atlas. Integration is performed using transformation-based approach. Data sets from the interaction group are converted into intermediate graphs, where nodes are genes, proteins, SNPs and the edges are the relations such as PPI (violet), epistatic interaction (green) or co-expression (red). All individual node identifiers are mapped to gene level and converted to ENSEMBL name space. Individual graphs are then combined into one heterogeneous graph with possible multiple edges between 2 nodes. Additionally, each node is characterized by a set of attributes such as ENSG ID, corresponding gene name, gene biotype, SNP ID, GWAS p-value, positive selection p-value and aggregated expression in 231 brain regions.

or proteins, can be mapped. We map identifiers from each individual data set to a common Ensembl database name space (please see Section 2.2 for details). The identifiers provided by Ensembl database (ENSG ID) are selected to ensure mapping of the multiple alternative gene and protein names to the unique identifier. We then combine individual graphs into a single heterogeneous graph with multiple edge types (please refer Section 3.3.1 for more information about the heterogeneous graphs). The combined data are represented as a single undirected graph where each interaction between a pair of nodes is described by a set of attributes outlined in Table 4.1.

Edge attribute	Description
ENSG.A	ENSG ID of the node A
ENSG.B	ENSG ID of the node B
Score	Value, associated with the interaction. It represents the importance of an interaction, i.e. strength, significance.
Interaction type	Type of biological relation between node A and B, i.e co-expression, PPI, epistasis.
Data source ID in HENA	Short name of the data source of the interaction. In case multiple data sets originate from one data source, data source name is appended by individual data set.

Table 4.1. List of edge attributes in HENA data set.

However, it is not always possible to map a SNP directly to the location of a gene, instead it can mapped to the intergenic region (IGR). In this case IGR is uniquely identified by two flanking genes. We refer epistatic interactions that contain IGR as intergenic region interactions (IGRI).

Combined disease related information about each node (gene), i.e. GWAS association, positive selection, expression in the brain regions, is described as node attributes shown in Table 4.2. Node attributes (or features) were collected for the genes present in the heterogeneous graph. Some sets of features are incomplete due to the absence of the required information in the various sources describing the genes. For each node a list of attributes is a vector of length 237.

Node attribute	Description
ENSG	ENSG ID of the node
Gene name	Alternative human readable gene name.
Biotype	Gene biotype according to Ensembl classification i.e. protein coding, long non-coding, etc.
SNP ID	SNP identifier from GWAS data set.
GWAS p-value	P-values corresponding to SNPs from GWAS data set.
PS p-value	P-value of gene association with positive Darwinian selection.
Brain region ID #1	Aggregated gene expression of the node in brain region ID #1
...	...
Brain region ID #231	Aggregated gene expression of the node in brain region ID #231

Table 4.2. List of node attributes in HENA data set.

4.4. Learning from the integrated data

To study if novel information about the disease can be inferred from HENA, we attempted to identify genes that were associated with Alzheimer's disease using information about genes and interactions of different types between pairs of genes. We also aimed to understand if network representation of heterogeneous data would be useful in this task. The study workflow is depicted on the Figure 21.

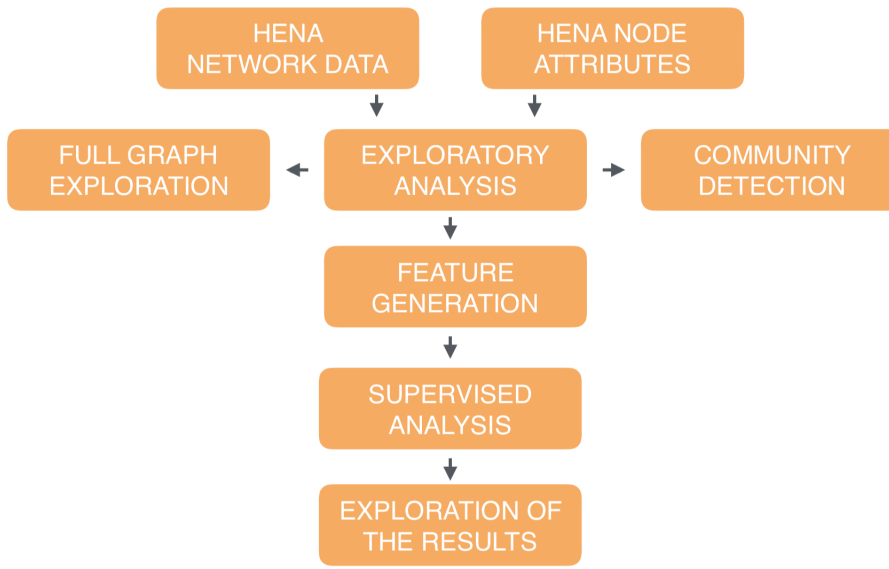


Figure 21. Identification of genes potentially associated with Alzheimer's disease in HENA data set. The diagram represents individual steps of the analysis for the identification of disease related genes.

The most straightforward way to identify genes that are associated with Alzheimer's disease would be to use a supervised machine learning method, where genes are labeled based on the association with Alzheimer's disease. Using the labeled set of genes we can train a model to find a decision boundary between two classes, and apply it to predict the association for the rest of the genes.

The problem is that such a set of confirmed positive and negative associations of genes and Alzheimer's disease is not yet well defined. Due to the limited technological advances the current research of Alzheimer's disease is not able to provide sufficient data to define the difference between patients and healthy individuals at the level of molecular interactions. One example to illustrate this issue is an inability to measure the presence or absence of the interaction between the proteins in a live brain regions.

4.4.1. Defining a node class

The genes, and proteins as their products, can be defined as associated with the disease based on genome-wide association studies and based on curated knowledge. Despite the substantial number of studies that has been carried out in the field of Alzheimer's disease, there is no clear agreement between the studies on the set of genes that are clearly associated with the Alzheimer's disease [111,229,230]. It is even more challenging to define the genes that are not related to the disease, which leads to inability to define the negative class, i.e. genes that are clearly not associated with the Alzheimer's disease. As the results of HENA study demonstrate, the noisiness of the gene labelling poses a serious challenge.

Therefore, we adopt the following strategy:

- We create three classes of genes: positive, negative and unknown. We use information about the nodes from HENA to assemble a set of genes associated with the disease based on GWAS data set and Alzheimer's related PPI data set collected in HENA. Negative gene label corresponds to the genes that are defined as essential non-disease genes in the evolutionary study by Spataro et al. [231]. The rest of the genes are labeled as unknown.
- Next, we explore one-class or anomaly detection problem in order to assess the quality of class separation.
- We use different feature sets, i.e. biological features and graph structural features, and apply two supervised approaches to classify the nodes with the unknown association with the disease.
- We apply supervised models, i.e. Random Forest and HinSAGE method, on set of genes labelled as unknown and suggest most likely candidates for Alzheimer's associated genes. Additionally, we perform qualitative analysis by exploring the existing body of research about the suggested genes.

4.4.2. Full graph exploration

Due to the fact that we are focusing on genes and their relation to the disease, we have excluded IGRs and corresponding IGRIs from this analysis. We also kept only strong co-expression interaction in disease-associated brain regions with Spearman correlation coefficient ≥ 0.5. The summary of the resulted graph is shown in Table 4.3.

	# nodes	# edges
full graph	24825	9740721
PPI subgraph	10445	52003
co-expression subgraph	14634	9671535
epistasis subgraph	13881	17183

Table 4.3. Number of nodes and edges for each sub-graph in HENA data set.

In Figure 22 we can see the whole picture. The graph consists of one largest component in the center, and many small components populating the "ring" closest to the central connected component. Alzheimer's genes, which are the minority, are shown with larger size red-colored nodes. They are seemingly uniformly distributed in the connected component as well as in the rings, which demonstrates that the problem of gene classification according to their association with the disease is not trivial.

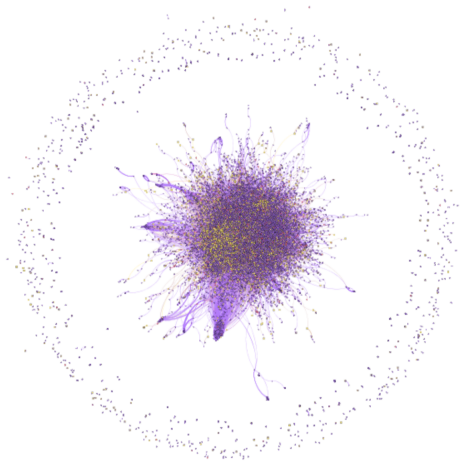


Figure 22. Visualization of HENA data set using Gephi platform [232]. Colors represent node types: red corresponds to the association with Alzheimer's disease, yellow node color indicates the genes that have no direct association with the disease, purple colored are genes with no information about their association with the disease.

4.4.3. Community detection analysis

In order to explore the connectivity of the nodes and identify potentially functionally related groups of nodes we performed clustering of the whole graph. We applied infomap clustering method [233, 234] to identify structurally related groups of nodes. By observing the resulted infomap communities we noted that Alzheimer's genes are often connected via one hop or two hops.

The analysis of communities also suggests that there are several big clusters that mostly have internal connections of one particular type, for example, a community of genes with PPI connections, or a separate cluster of co-expressions.

This observations does not contradict the previous knowledge about biological networks properties of being modular, small-world networks. In the biological terms it means that nodes sharing biological functionality are located within close proximity in the network [235, 236]. Taking into account that clustering method disregards the heterogeneity of the edges, we formed a hypothesis that the way the genes are connected reflects the type of the interaction they are involved in.

Based on the results of community detection analysis we collected additional node features that rely on graph structure.

4.5. Feature generation

To understand if network topological features can provide additional information for the model, we have created three sets of features that describe biological information, graph structural information and a combination of both.

4.5.1. Graph features

The results of community detection lead us to believe that we need to introduce gene features that exploit the graph-related information as it can improve the classification results. We collect the following features for each gene:

- an absolute number and a fraction of genes associated with Alzheimer's disease in a one-hop neighborhood via each edge type,
- same features for two-hop neighborhood,
- 256-dimensional node embeddings obtained by applying unsupervised GraphSAGE [35] (Section 3.3.2.2) on subgraphs consisting of one edge type, i.e. PPI, co-expression and epistasis. The resulting embeddings are concatenated, consisting in total of 768 features.

4.5.2. GraphSAGE embeddings

Recent advances in network analysis demonstrate that methods based on graph convolutional networks (GCN) outperform other traditional methods for many classification tasks on network data [34]. GraphSAGE model, proposed by Hamilton et al. [35], can generate node embeddings – node vector representations – for nodes that were not seen during the training (Section 3.3.2.1).

In this study we use unsupervised version of GraphSAGE, where for each node we learn embeddings based on both node features and the graph structure. Each node is then represented via a numeric vector of a specified fixed dimension (256 in our case) that captures the node properties in the graph in addition to the node features. Moreover, as we deal with a heterogeneous graph, we learn embeddings for each of the three subgraphs, where a subgraph consists of edges of a particular type. Therefore, the dimensionality of the resulting embedding vector for each node is 256×3 as we concatenate embeddings from subgraphs for each edge type.

4.5.3. Feature sets

To summarize, we represent each node via one of the following sets of node features: 1) Biological features here represent levels of aggregated gene expression in 231 brain regions and a value, representing genes, expressed higher in disease associated regions (CA1-CA4, DG, SptN, subiculum) compared with the rest of the brain regions. This value was obtained by the application of Wilcoxon test [20]. For further analysis we disregard genes that do not have the biological information about the expression in 231 brain regions for a fair comparison with

the graph-related set of features. The statistics of class distribution in the reduced data set is demonstrated in Table 4.4. 2) *graph-related features* that consist of a set of graph embeddings learned via unsupervised GraphSAGE and neihborhood related features discussed in the beginning of Section 4.5.1. Finally, 3) a union of both sets of features, which we refer to as *all features*.

	disease	non-disease	unknown	total
initial	1018	1430	22377	24825
reduced	854	1421	15286	17561

Table 4.4. Class distribution in the initial and reduced datasets.

In the following Section 4.5.4 we investigate these node features to select the most useful features for the task of gene classification according to their association with Alzheimer's disease.

4.5.4. Exploration of feature sets

As we mentioned earlier, positive and negative classes of genes are very noisy. Therefore, we adopt anomaly detection techniques. The common anomaly detection approach is to find the lower dimensional embeddings of the original data where anomalies are expected to be separated from the regular data. The lower dimensional embeddings are then reconstructed, which means that they are brought back to the original data space. Reconstruction of the data with the low dimension embeddings is expected to represent the underlying generative model of the data. The reconstruction error of a data point is defined as the error between the original data point and its low dimensional reconstruction. For the anomaly detection we adopt a particular type of neural networks called autoencoder [237]. An autoencoder learns a dense representation of the data by reconstructing it, where the goal of the learning is to minimise the difference between the data provided as an input and the reconstructed data. A typical architecture of an autoencoder resembles an hourglass, with the sizes of input and output layers (at the edges) equal to the dimensionality of the data feature vectors, and layers of lower dimensionality in between. We train three autoencoder models – using sets of biological, graph-related, and all the features (refer to Section 4.5.3 for the details). Each of the models is trained on a subset of negatively labeled genes and applied to the rest of the data points. We measure the mean reconstruction error of data points for each of the gene classes. We expect to see a larger difference in the reconstruction error of the positively labeled gene set.

The difference between the errors of different classes is quite small. The ideal scenario for the density plots is when the reconstruction error distribution for the negative genes does not overlap with the density distribution of the positive class. Figure 23 demonstrates the separation of three classes using three feature sets. From the density plots we can observe poor separation in general, however, we can

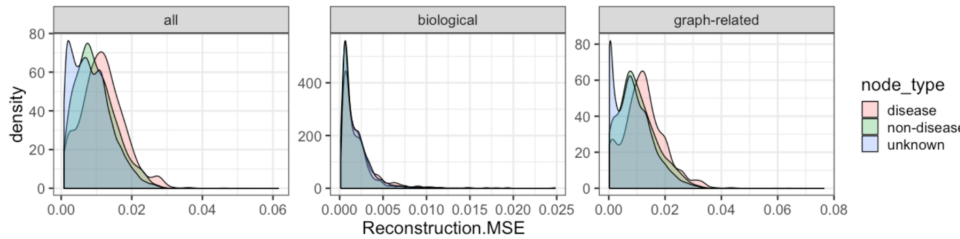


Figure 23. Reconstruction mean squared error (MSE) distribution density using three sets of features.

notice that graph-related and combined feature sets seems to be more informative for the classification task than the biological feature set.

4.6. Supervised and semi-supervised approaches

Next, we adopt a classical supervised strategy, where a Random Forest classifier is trained on the set of positive and negative genes for each of the feature sets (for the description of the feature sets see Section 4.5.3). The results are shown in Table 4.5. We then apply the model to a set of unknown cases and rank them according to the probability score. The higher the probability, the more likely the gene is associated with the Alzheimer's disease.

We also apply an algorithm from the family of Graph Convolutional Networks - a generalisation of the GraphSAGE [35] to heterogeneous networks- a HinSAGE [218] algorithm.

The HinSAGE model (see Algorithm 2) is a semi-supervised approach, where it is trained using labeled head nodes, i.e. genes of positive and negative classes, while the whole network structure is used for the feature vector aggregation (including connections to genes with unknown class as well as their features).

As in the case of random forest classifier, we rank the predictions of HinSAGE model for the unlabeled genes. The results of both models' performances are shown in Table 4.5.

	HinSAGE			
	Precision	Recall	F1 score	ROC AUC
all	0.40	0.98	0.57	0.60
graph	0.40	0.98	0.57	0.66
biological	0.37	0.94	0.54	0.56
	Random Forest			
all	0.50	0.50	0.50	0.82
graph	0.50	0.50	0.50	0.83
biological	0.43	0.27	0.33	0.56

Table 4.5. Comparison of HinSAGE and Random Forest classification model performance using different feature sets.

We have compared the results of random forest and HinSAGE model performance for biological, graph-related, and joined set of features. The results demonstrate that the performance of both models improves with the use of graph features. Both models showed the worse performance on a set of biological features. However, HinSAGE on biological features uses the initially provided non-graph features, and propagates this information exploiting the graph structure. It demonstrates better performance than Random Forest on a set of biological features. Random Forest on graph-related feature set slightly improves the performance, and HinSAGE on graph features is the best performing method taking to account F1 score, i.e. a harmonic average of the precision and recall. However, we can observe that the HinSAGE recall is substantially higher than Random Forest and the precision is slightly decreased. This can be influenced by a number of data properties and model parameters that require optimization. For example, current study implied building classification models for imbalanced data with noisy labels, meaning that there is no biological ground truth in case of defining a negative class based on the assumption that genes are not related to the disease. Model own parameters, i.e. a number of sampled nodes at each neighbourhood, selected prediction threshold, distribution of positive and negative class in the whole population of genes to estimate weights for positive and negative class in cross binary loss function, etc., require more experiments to explore model best setup. In the frames of the current work we aimed at exploring whether graph-related data features provide information useful for the classification and have observed improvements in metrics for both classification models. Undoubtedly, more investigation and careful data class definition is required which is a case for the future work.

Graph structure helps to capture complex relationships

To explore the classification results of the algorithms we performed qualitative analysis by checking the existing body of research. For this purpose we have explored additional independent resources to assess the result in a disease context. We have created a list of genes, shown to have a strong association with the disease, from the following resources:

- GWAS [238–240] and genome-wide association study by proxy (GWAX) study [241].
- list of Alzheimer's disease-specific autoantibodies in human sera [229]
- list of genes reported to be associated in Alzheimer's disease downloaded from MalaCards database [230]
- results of integrative transcriptome analysis [242]

The individual gene lists were extracted and combined. These external gene lists were assembled independently from HENA data set and were not used in the training and testing procedure to calculate the performance metrics, i.e. ROC and F1 measures, of random forest and HinSAGE models. Recall that we have classified 15286 genes labeled *unknown*. The genes present in these external gene lists were then searched among the predictions of the random forest and HinSAGE. Classification results, i.e. result for genes with unknown class, contained 169 genes that were present in the assembled list of 373 genes. Additionally, we have compared external gene lists with the genes labeled as *disease-related* and *non-disease* in classification task resulting correspondingly in overlaps of 88 genes and 34 genes. The latter again points to the difficulty of defining positive and negative class due to the lacking biological ground truth precisely determining the class separation.

For each model results, we have counted genes with the assigned probability ≥ 0.5 to be associated with Alzheimer's disease. Out of these 169 genes, random forest classified 14 genes to be associated with the disease while according to the probabilities assigned by HinSAGE, 154 genes were associated with the disease. Taking into account that GWAS studies are often used for thorough biological investigations and generation of novel hypotheses, we have focused on thee novel GWAS genes (ACE, ADAMTS4 and CLNK) recently reported in three independent GWAS meta analysis publications [238–240]. We have observed that HinSAGE classified these genes as related to Alzheimer's disease with the probabilities 0.81, 0.93 and 0.88, while random forest did not classify them as Alzheimer's disease related genes.

These qualitative findings of the subset of genes demonstrate that graph structure is a rich data source that helps to capture complex relationships and find the distinctive patterns that are not easily detectable otherwise.

4.7. Summary and impact

In our work we have combined the heterogeneous experimental data from various sources into one network-structured data set HENA. Combining HENA with the external complimentary data sets provides the possibility of discovering genes potentially involved in disease mechanisms.

Prediction of genes related to complex diseases, such as Alzheimer's disease, is not a trivial task. The ambiguity in the definition of the node class, i.e. relation to

the disease, and selection of the informative features strongly influences the model performance. However, we have demonstrated the advantage of state-of-the-art GCN-based methods over classical supervised approach for the classification of genes related to Alzheimer's disease. We have also demonstrated the advantage of using network structural information in node classification task compared to using biologically determined features alone.

4.8. Contribution

In this project my contributions were: collection, preprocessing and analyzing microarray data sets related to Alzheimer's disease (co-expression data sets in Alzheimer's disease and normal samples and co-expression in brain disease-associated regions were created), filtering of PPI data from IntAct, aggregation of gene expression from Allen brain atlas; developing a transformation-based integration for heterogeneous data sets; creating HENA data set; application of GCNs for gene classification; manuscript writing, preparation of figures.

5. STUDYING DISEASE PATHOGENESIS USING DATA INTEGRATION APPROACH (PUBLICATION II)

Disease pathogenesis is a complex biological process that can be investigated on different experimental levels including transcriptomics or proteomics studies, investigation of the specific clinical patients phenotypes, etc.

In this chapter we demonstrate how the application of data integration approaches for combining the results of these individual studies and experimental analysis allows to find hidden relationships between the data and the phenotype of interest and obtain the coherent view on the disease pathogenesis.

For the reader's convenience the individual data types and statistical methods mentioned in this chapter are described in the Chapters 2-3.

5.1. Studying the pathogenesis of psoriasis

Psoriasis is a chronic complex inflammatory disease that affects skin and is associated with systemic inflammation and many serious comorbidities such as metabolic syndrome or cancer [243, 244]. Important discoveries about psoriasis pathogenesis have enabled the development of effective biological treatments by targeting the members of T helper 17 pathway [245–248]. However, it has not been settled whether psoriasis is a T cell-mediated autoimmune disease, i.e. caused by the response of autoantigen-specific T cells, or rather autoinflammatory disease due to excessive stimulation of innate immune receptors [249–252].

Although autoimmunity and autoinflammation both lead to self-directed inflammation, their mechanisms of action differ. Autoimmunity involves adaptive immune cells targeting various autoantigens, while autoinflammation involves innate immune mechanisms without specific autoantigen involvement [253].

Given this problem, we applied multi-staged data integration framework to study this complex biological process on different levels ranging from transcriptomics to specific patient phenotypes (Figure 24). We analysed various types of experimental data, i.e. gene expression, protein concentration in plasma, and clinical meta-data about the patients at the individual levels exploring each data type association with the disease pathogenesis. Subsequently, we brought together the results of individual analysis using available domain knowledge to form a more systematic view on the disease.

We formulated the aims of our study as following:

- To find evidence confirming either autoimmune or autoinflammatory hypotheses of psoriasis pathogenesis.
- To understand the connection between specific phenotypes of psoriasis patients and signs of systemic inflammation.

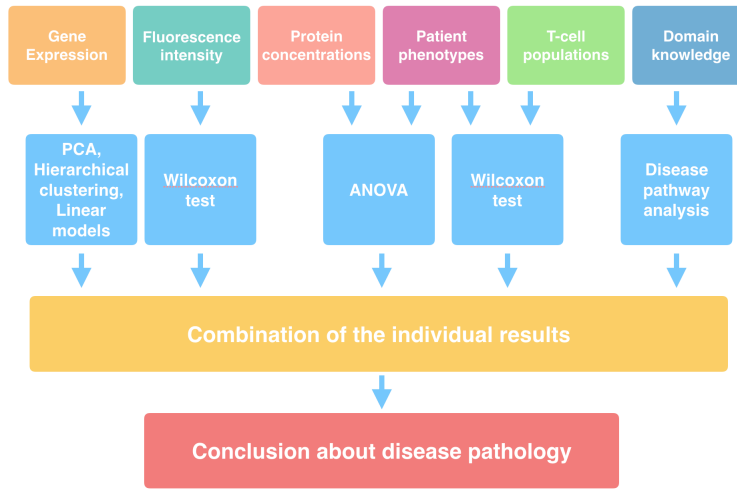


Figure 24. Integrative data analysis in psoriasis pathogenesis study. Individual data types were analysed separately using appropriate statistical methods implemented in R language [254]. The analysis results were combined and interpreted using domain knowledge.

5.2. Gene expression and cell fluorescent microscopy reveal the signs of innate immunity in psoriasis

We measured gene expression in psoriasis patients in disease-affected lesional and seemingly healthy non-lesional skin biopsy using qRT-PCR, and compared it to the expression in healthy individuals.

Principal component analysis (Section 3.2.1) of the gene expression data was able to separate the lesional skin samples from the other studied groups. At the same time psoriatic non-lesional skin samples mostly overlapped with control group. Although PCA did not distinguish healthy skin from psoriatic non-lesional skin, we were able to detect significant differences at the level of individual genes.

Our comparative gene expression and hierarchical cluster analysis revealed important gene circuits involving innate receptors. We used differential gene expression analysis to find the genes that could demonstrate the difference between these three conditions.

In our study we confirmed that gene expression of known psoriasis related Th17 pathway members and other pro-inflammatory cytokines associated with psoriasis pathogenesis (Figure 25) were significantly increased in psoriatic lesional skin in comparison to non-lesional and control skin. We also showed that several immunoregulatory genes and marker transcription factors were up-regulated in psoriatic skin. Additionally, hierarchical cluster analysis followed by multiscale bootstrap resampling [170] helped to find the groups of genes involved in biological processes and describe their roles in disease pathogenesis. For example, that antimicrobial peptides in psoriatic lesions may come from different cell types.

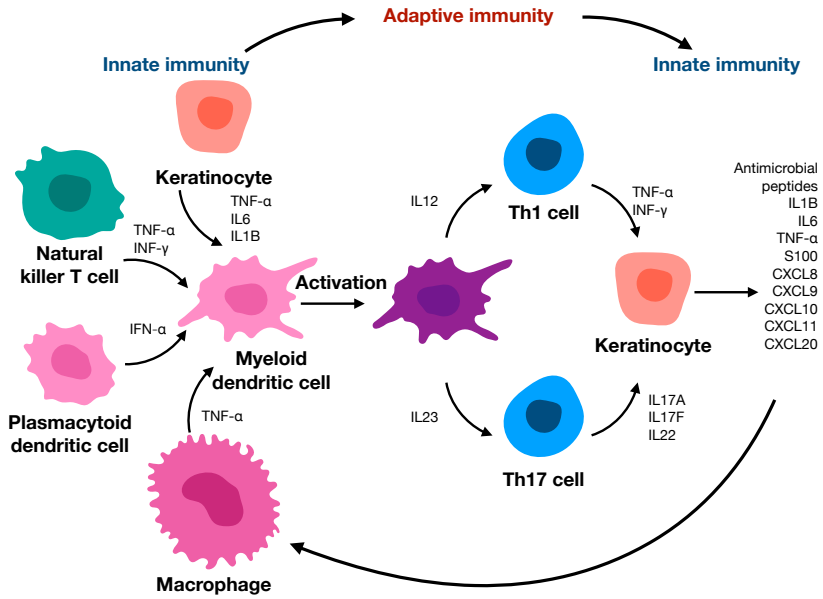


Figure 25. Key cells and mediators in the transition from innate to adaptive immunity in psoriasis. Pathway structure is adapted from Nestle et al. [253]

Our study is the first to describe increased expression of EOMES gene in psoriatic lesions and in non-lesional skin at mRNA as well as protein level. The product of this gene is an important transcription factor for the development of virtual memory CD8+ T cells. These cells have innate-like functions, such as the capability to rapidly produce inflammatory cytokines in the absence of antigenic recognition [256–258]. Immunofluorescent microscopy of psoriatic skin confirmed that a fraction of T cells indeed contained EOMES protein in their nuclei.

As autoinflammatory diseases are often associated with inflammation activated in response to stress signals, we studied the gene expression of the corresponding genes, e.g. innate receptors and inflammasome components. Our results confirmed recently discovered possibility of several innate receptors being involved in the pathogenesis of disease [259] (Figure 26a).

Another important regulator of innate immunity is caspase-1 enzyme encoded by CASP1 gene [260]. We used fluorescent microscopy together with staining to investigate its activation in the skin biopsies. We measured and compared fluorescent intensities in healthy control, lesional and non-lesional skin (Figure 26b-e). We applied Kruskall-Wallis test [261], followed by post-hoc Dunn's test [262] to determine which groups were significantly different. We didn't detect statistically significant differences between non-lesional and healthy skin. However, the fluorescence intensity was significantly higher in psoriatic lesion in comparison to healthy skin or non-lesional skin of psoriasis patients indicating CASP1 activation in lesional skin (Figure 26f). At the same time we were not able to iden-

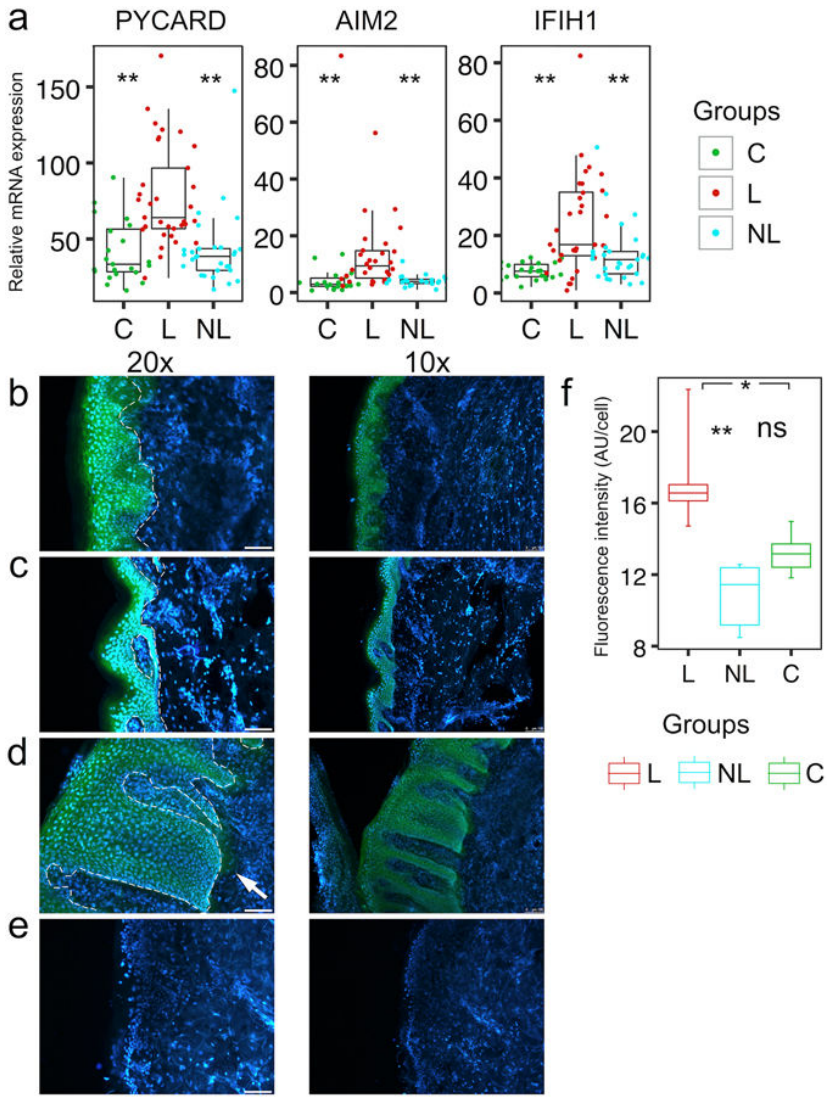


Figure 26. Innate receptors' up-regulation and inflammasome activation in psoriatic lesions (figure is adapted from Šahmatova & Sügis et al. [255]). (a) mRNA expression of genes encoding innate receptors relative to gene expression of ACTB in skin biopsy samples obtained from lesional (L) and non-lesional (NL) skin of psoriasis patients and from age-matched control individuals (C). Stars above the groups C and NL depict their significance level from L samples. **P < 0.001, *P < 0.05. (b–e) Fluorescence microscopic images illustrate caspase-1 activation in control skin (b) psoriatic non-lesional skin (c) or psoriatic lesional skin (d) biopsy frozen sections. (e) Represent control slides. White bars in 20x images represent 75 μm. White dotted lines indicate the basal membrane. (f) Green fluorescence intensity was measured from the slides in fixed skin areas (from slides of two different individuals and 2–3 different areas per slide) and divided by the number of cells per area.

tify statistically significant difference in mRNA levels of CASP1 gene in lesional, non-lesional and healthy control skin. Thus we were able to detect an activation of CASP1 only using extra level of information, i.e. results of analysis immunofluorescence microscopy data, that would not be possible from gene expression data alone.

The combined results from qRT-PCR and fluorescent microscopy analysis strongly suggest the increased activation of innate immunity in psoriatic lesions. The combination of these results in multi-staged analysis with the signs of innate immunity activation from other data sources provides the evidence for the autoinflammatory disease pathogenesis.

5.3. Protein concentration in plasma provide extra information about systemic inflammation

To study the signs of systemic inflammation further we measured the concentration of several cytokines and chemokines (IL-17A, IL-6, TNF-α, IL-1Ra and CXCL8) in plasma samples of psoriasis patients and healthy individuals. These proteins are involved in cell to cell communication in immune responses. Using Wilcoxon rank-sum test (Section 3.2.7) we detected that a concentration of a group of these proteins in psoriasis patients were elevated compared to control individuals. The results were consistent with the differential gene expression analysis.

5.4. Linking the experimental data to the patient phenotype

Additionally, in our study we aimed to find gene expression and plasma protein signatures for different phenotypic features of psoriasis. Comparisons of the selected phenotypes in skin and plasma samples were performed using multi-factor ANOVA (Section 3.2.5), followed by Tukey's honest significant difference test to find the groups that are significantly different from each other.

Gene expression in the psoriatic lesions, with an exception of TNFA gene, did not exhibit any statistically significant change in patients with different phenotypes. We identified a difference in expression levels of TNFA gene in patients with psoriatic arthritis. At the same time gene expression in non-lesional skin was more variable in psoriasis phenotypes. We managed to identify genes that were connected to the severity of the disease, nail involvement, duration of the disease and psoriatic arthritis. This way we showed that not only lesional skin but also the non-lesional skin reveals signs of inflammation that are associated with the variable clinical features of psoriasis. These clinical features constituted patient phenotypes in our study.

The analysis of protein concentrations in plasma in psoriasis patients with several phenotypes showed that the signs of systemic inflammation increased with

the severity of the disease as it was linked to increased levels of IL-17A. Additionally, analysis revealed that IL-6 is associated with sporadic form of psoriasis and is involved in the pathogenesis of the psoriatic arthritis.

5.5. Analysis of cell populations reveal the signs of premature senescence in psoriasis

Sustained inflammation is believed to have similar damaging effects on the immune cells, as observed during inflammageing, a chronic inflammation that involves the up-regulation of the inflammatory response with age [263, 264].

We studied the changes in the T cells subpopulations in the psoriasis patients and healthy control individuals using flow-cytometric immunophenotyping. We compared the proportions of subpopulations in healthy and disease samples using Wilcoxon rank-sum test (Section 3.2.7). We found that among innate-like virtual memory CD8+ T cells the terminally differentiated or senescent T cells had higher proportions in psoriasis patients similarly to the accumulation of these cells in aged individuals [265].

Also as the long-term consequences of systemic inflammation on circulating immune cells are not known in psoriasis patients, we studied T cell subpopulations in patients with different disease duration. We carried out the comparison of phenotypes using ANOVA followed by Tukey's honest significant difference test (Section 3.2.5). We identified a statistically significantly higher proportion of terminally differentiated or senescent CD8+ T cells in patients with longer disease duration.

These findings suggest premature immunosenescence, a decline of the immune system with age advancement, occurring during disease. It is associated with the implications to the psoriasis comordibidities such as several cancers [244]. As immunosenescence is generally associated with the increased risks of cancer development [266, 267], premature immunosenescence can play an important role in the inability to resist cancers in psoriasis patients.

This finding suggests that chronic inflammation together with the sustained innate receptor signalling drives premature immunosenescence in psoriasis patients, which may have implications on the health status of the patients.

5.6. Summary and impact

Integrated analysis of biological data types such as transcriptomics, proteomics, cell populations and clinical data about the patients allowed us to study the disease pathogenesis from different angles and obtain more systematic view on the inflammatory processes in the disease. Data integration allowed to identify an activity of the important regulators of innate immunity, signs of systemic inflammation increase in several phenotypes associated with the severity of the disease, and the

signs of premature senescence in psoriasis patients. The results of the study support the autoinflammatory hypothesis of the disease pathogenesis. These findings would not be possible to obtain by analysing individual experimental data types alone.

5.7. Contribution

My contribution to this project included statistical data analysis, data visualization and writing statistical methods part in the manuscript. I performed gene expression data analysis including preprocessing with filtering and imputation, hierarchical clustering, PCA and differential expression analysis. Additionally, I performed the comparison of the fluorescent intensities, comparison of the selected patients' phenotypes in skin, plasma samples and T cells subpopulations.

6. IMPROVING DEVELOPMENTAL TOXICITY TESTING STRATEGIES USING DATA INTEGRATION (PUBLICATION III, IV)

Toxicity testing is the process of defining harmful effects of the substances for the living organisms. One of its applications is safety assessment of drugs or other chemicals for early development of a human organism.

In this chapter we describe the multi-staged integrative data analysis approach and demonstrate how current testing strategies in toxicology can benefit from its application. The approach of combining multiple data sources and the corresponding analysis methods would possibly allow to identify key events in adverse outcome pathways (Section 2.2.3.1) related to the action of toxic compounds and define associated biomarkers [146, 268].

To address the problems stated above we investigated four groups of toxic compounds: heavy metals, pesticides, polychlorinated biphenyls (PCBs) and histone deacetylase inhibitors (HDACi).

The aims of this study were to investigate the following:

- whether the combination of transcritomics data with other data types would provide information related to the toxicants mode of action (MoA) and associated adverse outcome pathways;
- whether the same resulting action of structurally diverse toxicants, i.e. metals, pesticides, HDAC inhibitors, is characterized by unified transcriptional changes;
- whether patterns of transcriptional changes would reflect compound grouping;
- whether the analysis of a small group of migration-related transcripts would allow to classify the compounds according to their MoA.

For the reader's easier understanding of the data origins, their biological meaning and applied statistical methods used in this study, we provided their detailed description in Chapters 2-3.

In study III we have addressed the impact of the toxic chemicals on one of the essential biological processes during normal human early development - the ability of NCC to migrate correctly to different parts of the fetus where they differentiate to various tissues, e.g. bones, neurons, etc. [269]. The workflow of multi-staged integration for this study is outlined on the Figure 27.

The disturbance of neural crest functioning during early development can lead to severe birth defects [270]. Several chemical compounds are known to disturb the NCC migration process, i.e. NCC motility, causing developmental defects [271, 272]. The harmful action of the chemical can be assessed by measuring the cells ability to migrate using migration of neural crest cell assays (Section 2.2.1.7). The negative effect of a compound is defined as decreased or lack of

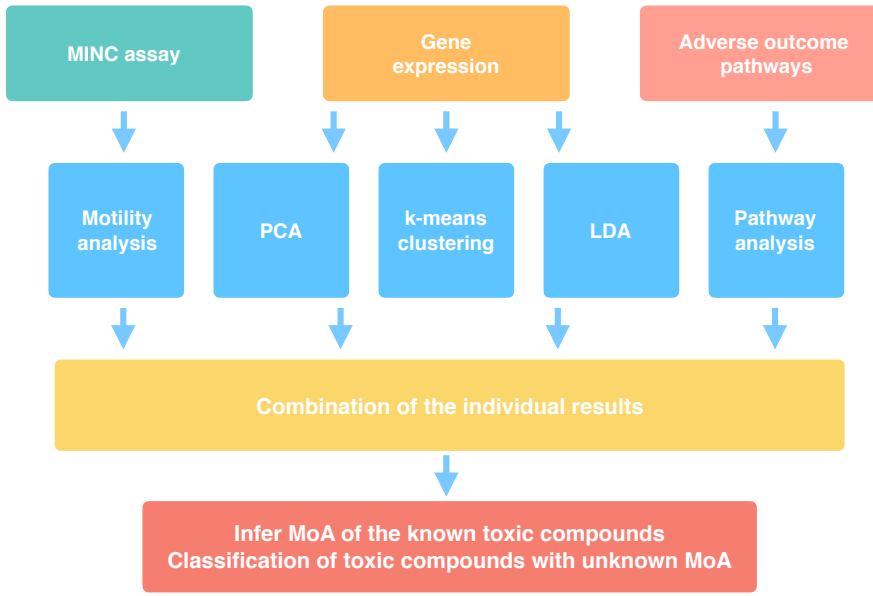


Figure 27. Data integration in the analysis of toxic compound disturbing the migration of NCC and their MoA. Individual data types were analysed separately using appropriate statistical methods implemented in R language [254].

cellular migration after treating the cells with the compound.

However, MINC functional assay is a "black box" approach that allows us to observe only the endpoint of the effect, i.e. if the cell motility was disturbed or not, without understanding how it was achieved. By using only the functional assay we are not able to judge about the mechanistic mode of action of the compound and to understand how it disrupts the migration pathway. Some pathways were identified to be important for the NCC migration. However, it is not clear which of them are affected by toxicants and whether all toxicants have effect on the specific group(s) of pathways.

One way to connect the final adverse outcome, e.g. disturbed NCC migration, to the intermediate changes inside the cell is to understand the effect of toxic compounds on the expression of the related genes. The possible outcome could be that toxic compounds leading to the same effect cause similar changes. Alternatively, each toxicant may trigger its own particular set of gene expression changes. In this scenario different changes eventually lead to the same end result. In practice it has been shown that several toxicity pathways may lead to the same end result [273]. Therefore the final outcome could be linked to different gene expression signatures. This way toxicants sharing a mode of action would trigger similar changes in the expression of relevant genes.

6.1. Grouping toxic compounds by their transcriptional signatures

In this study we first identified diverse toxicants that disturb the migration of NCC in the MINC assay (MINC assay is described in Section 2.2.1.7). Then we identified transcriptional changes triggered by these toxicants in 35 migration-related genes using qRT-PCR measuring technology (Section 2.2.1.4).

Grouping of toxicants according to their gene expression profiles using hierarchical clustering and PCA (Figure 28) identified the group of HDAC inhibitors as previously unrecognized drug class affecting NCC migration.

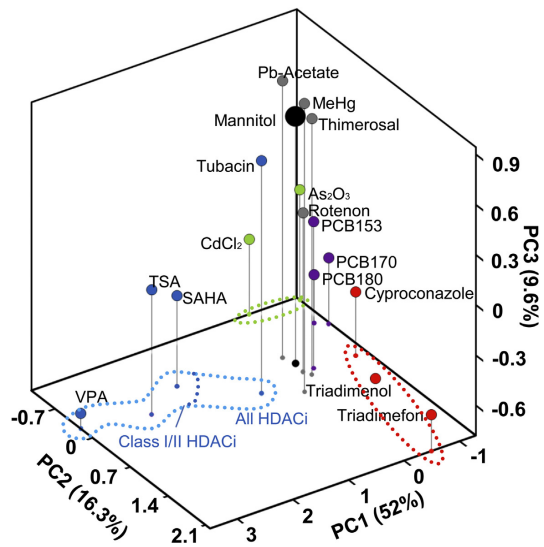
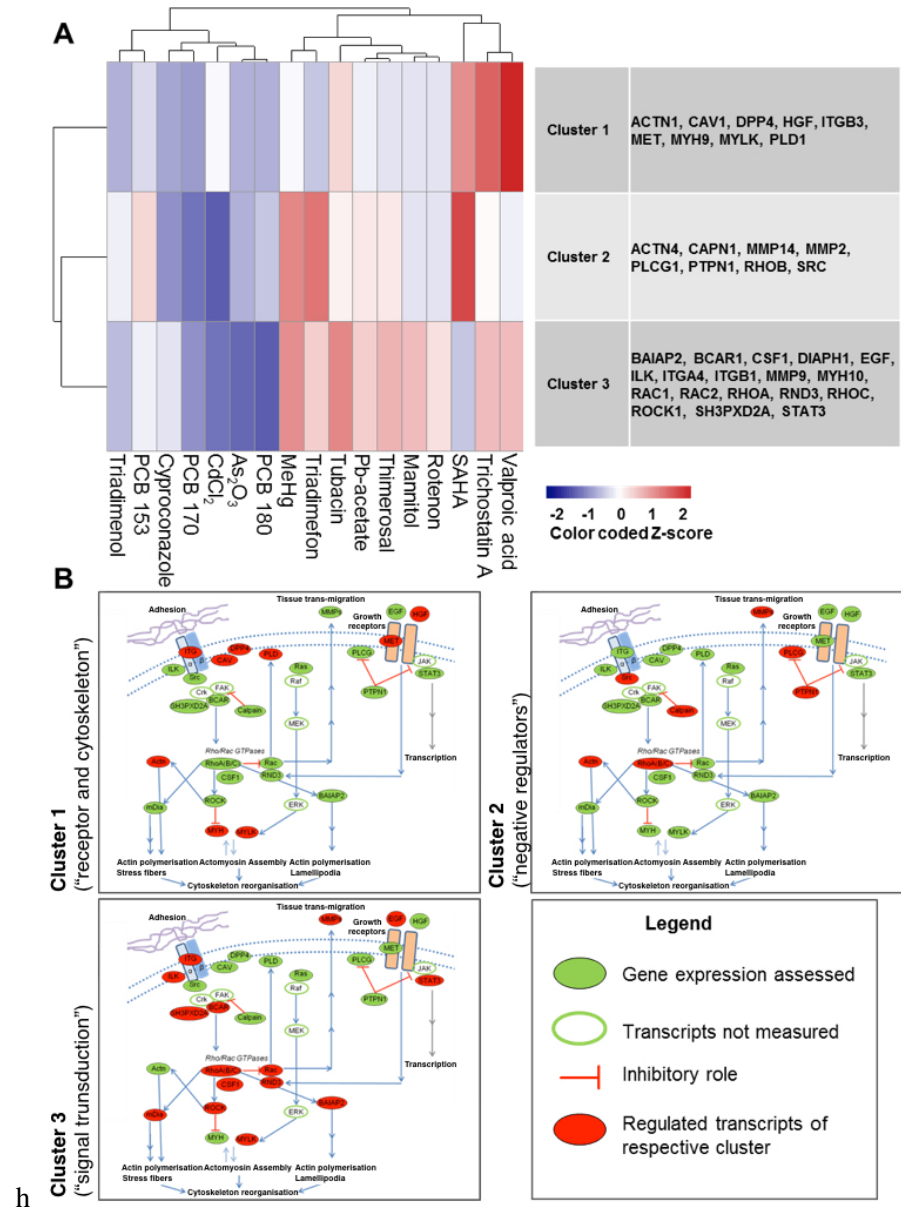


Figure 28. The grouping of toxic compounds based on the results of PCA (figure is adapted from Dreser et al. [274]). Projections of the individual experiments to the two-dimensional planes reflect separation by the corresponding PCs.

This analysis showed that similar gene expression patterns of HDACi compounds distinguished them from other migration inhibitors. We also demonstrated that a group of PCBs and triazoles could be identified by distinct gene expression signatures. Our study showed that a well-known drug for treating epilepsy and bipolar disease, valproic acid, disrupts the NCC migration through the mechanisms of histone deacetylase inhibition. To characterize the mechanisms of the intermediate changes in the cell we identified potentially involved intermediate pathways associated with the toxicants and their impact on the NCC migration. We used k-means clustering to find the co-regulated genes and performed pathway analysis of each cluster (Figure 29).

The results of the analysis demonstrated that HDACi group represents a new group of toxicants with the effects on neural crest during development.



h

Figure 29. Pathway analysis of migration-related clustered genes (figure is adapted from Dreser et al. [274]). (A) k-means clustering of gene expression profiles over individual compounds. Mean expression levels of the genes in each cluster are scaled (z-score) and color-coded. Dark red and blue colors indicate high and low levels of expression correspondingly. Expression profiles are hierarchically clustered using Euclidean distance and average linkage. The genes of each cluster are shown. (B) For each cluster its contribution in migration related processes is shown in separate window.

6.2. Classification of the compounds

Additionally, to our analysis we built LDA-based statistical model for compounds classification. This model allows the classification of the toxicants based on the small number of genes. The application of the model on the given data set revealed the compounds belonging to HDACi group being largely separated from the rest of the toxicants. The developed model can help to assign unknown individual compounds to the known toxicants and identify their mode of action.

6.3. Modeling neurodevelopmental defects caused by HDACi

Currently only limited amount of human data is available about how early development is affected by chemicals. The mode of action of such chemicals can be studied by mimicking the neurodifferentiation *in vitro*.

In the study IV we have investigated effects of HDACi group of toxic compounds on early neurodifferentiation (Figure 30). We have studied the mechanisms of action of these drugs including trichostatin A (TSA) valproic acid (VPA). Valproic acid is an anti-epileptic drug that can cause several pathologies related to early neurodifferentiation, such as fetal valproate syndrome (FVS) characterized by neural tube defects during development. Due to the fact that effects of HDACi are not well characterized in human cells, we have initiated this study to model the disturbed neural development triggered by HDACi using multi-staged data integration. Our goal was to identify a set of marker genes that would allow characterization of drug-induced effects such as present in VFS using various data. Additionally, we aimed to distinguish between short-term and long-term transcriptional changes, triggered as a result of disturbed neurodevelopment caused by HDACi. To answer these questions early neural differentiation was performed from human embryonic stem cells towards neural lineage over 10 days. The prolonged drug-induced effects were studies after drugs were washed out for 1-2 days.

We have integrated the results of differential expression analysis using linear models (Section 3.2.6) of gene expression microarray data (Section 2.2.1.3), PCA (Section 3.2.1) and gene functional enrichment analysis (Section 3.2.9) of the collected data during normal development and after the treatment with compounds. PCA of gene expression of untreated cells and cells treated with HDACi revealed that the strongest disturbance of differentiation was observed after toxicants exposure during early neural fate decision, i.e. during the period of day 0 to day 6. PCA plot revealed a group of samples treated with HDACi were separated from samples undergo normal neurodifferentiation. Differential expression analysis of microarray data showed that 2500 genes were altered after the treatments with HDACi. Further characterization of the differentially expressed genes was performed using functional enrichment analysis. We have identified statistically significantly over-represented GO terms related to biological processes associ-

ated with neuronal development, e.g. "nervous system development" and "central nervous system development", "neurogenesis", etc., in genes down-regulated after treatment with HDACi. In the enrichment analysis results of up-regulated genes we have observed terms such as "anatomical structure morphogenesis" and "anatomical structure development". The results of PCA and functional characterization of differentially expressed genes showed that new population of cells emerged after the treatment with HDACi.

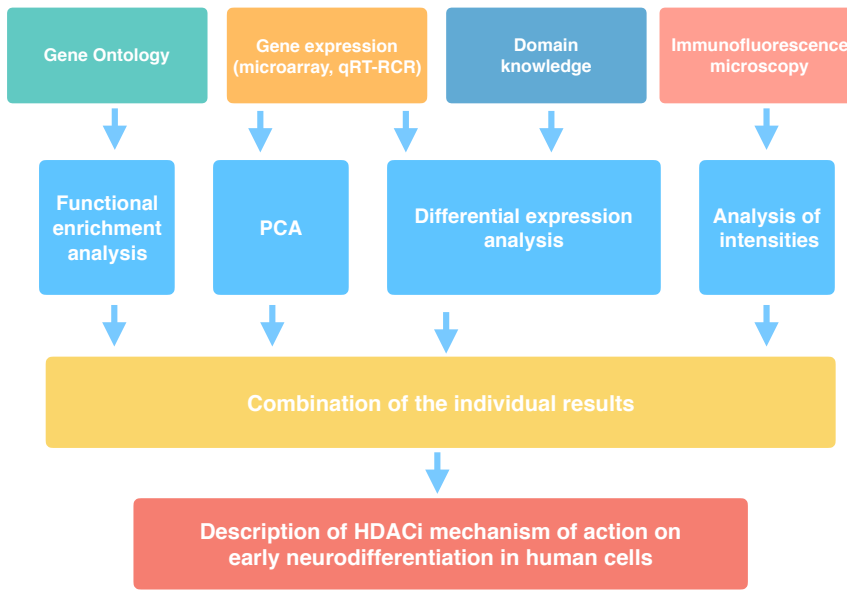


Figure 30. Data integration in the analysis of drug toxicity in early neurodifferentiation.

To investigate mechanisms of HDACi further, genes down-regulated by HDACi were compared with genes that were identified in mouse-knockout study resulting in neural tube defects. This combination of gene lists from our study and mouse resulted in the 14 potential candidates that might be associated with human neurodevelopmental defects.

As the next step of data integration we performed additional analysis of marker genes to compare the similarity of VPA and TSA MoA. Altered expression of these genes indicate disturbed neurodevelopment due to chemical-induced developmental neurotoxoicity. We performed differential expression analysis of qRT-PCR data and additionally analysed immunofluorescence imaging data of well-known neural marker PAX6. We have observed fully reproduced effects of VPA by TSA. Further investigation of HDACi MoA with respect of effects on histone modifications revealed that they can affect key differentiation factors indirectly by affecting histone code in their promoter region.

Taking to account individual results, we have characterized the HDACi short-

term and long-term effects on human cells. Study results suggested that long-term transcriptional effects of HDACi differed from those found in short-term studies, as they reflected altered neurodifferentiation due to secondary changes of the histone code for key transcription factors.

6.4. Summary and impact

The findings suggest that the combination of gene expression data with phenotypic functional cell-based assays, information about pathways related to neurodifferentiation and information from Gene Ontology characterizing gene functions, can improve the interpretation of the toxicity testing results. We showed that toxicants sharing a mode of action trigger similar changes of intermediate cellular markers. These findings allow detection of key events of adverse outcome pathways and definition of biomarkers. Additionally, proposed classification strategy can be applied to identify mode of action of the unknown toxic compounds. The usefulness of the proposed approach is demonstrated by the confirmation of HDACi group of compounds as neural crest toxicants.

6.5. Contribution

In the study III I performed data analysis using PCA, k-means and hierarchical clustering to investigate if the individual toxic compounds can be grouped based on their mode of action. I also built an LDA-based statistical model for the classification of toxic compounds based on their transcriptional signatures. Additionally, I prepared part of the figures and participated in the manuscript writing and editing. In the study IV I performed differential expression analysis and functional enrichment analysis.

7. CONCLUSION

The aim of this work was to demonstrate the conceptual pipelines for data integration of heterogeneous biological data using data science methods, and to show how integrative data analysis provides more systematic view on the biological processes, helping to discover new knowledge.

In this work we have established two groups of data integration strategies - multi-staged and transformation-based that can be applied to various biological questions. We have described appropriate integration approaches for three practical study set-ups. In the Chapters 4-6 we have demonstrated how knowledge about biological processes can be enriched through combining experimental, computational and domain knowledge data using machine learning methods.

As a result of conducting this research, we have designed and implemented a multi-staged analysis for toxicology study and disease pathogenesis study.

Integrated analysis of the experimental and clinical data, i.e. presented in Chapter 5, provided an opportunity to understand psoriasis pathogenesis from different angles and characterize the inflammatory processes during the disease in a broader perspective. The results of the integrative analysis support the autoinflammatory hypothesis of the disease pathogenesis. The complex view on the pathogenesis also indicated potential implications for patient care that have not been attended previously.

The findings that we have presented in Chapter 6 suggest that the combination of gene expression data with other various data types and combination of analysis methods improves *in vitro* toxicity testing strategies. Additionally, the developed classification model can be applied to identify the mode of action of the unknown toxic compounds. The practical implications of the proposed approach is demonstrated by the experimental confirmation of HDACi group of chemical compounds as neural crest toxicants and investigation of its MoA in valproate syndrome, that would not be detected by analysing just a single data type.

We have designed a transformation-based data integration approach to combine diverse *omics* data sets related to Azlheimer's disease (Chapter 4). This approach allowed to integrate experimental and computational data from various sources into one heterogeneous network-based data set. Thus, we have generated a novel domain-relevant resource that researchers can incorporate into their analysis. This integrated resource allows researchers to get a systematic view on Alzheimer's disease by going from SNP to protein level. Additionally, we have applied state-of-the-art deep learning methods for heterogeneous graph structured data. The findings that we have presented suggest that by utilizing rich graph structure, GCN methods are able to uncover gene relation to the disease, that would not be detected by just using gene biological information alone.

For the effective *omics* data integration one has to understand data origins and the relations of one data type to another. In this work we summarized and described individual experimental and computational data types, their origins, the

way they relate to each other, and what we can learn from their combination.

BIBLIOGRAPHY

[1] Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, et al. Arrayexpress—a public repository for microarray gene expression data at the ebi. *Nucleic acids research*, 31(1):68–71, 2003.

[2] Nikolay Kolesnikov, Emma Hastings, Maria Keays, Olga Melnichuk, Y Amy Tang, Eleanor Williams, Miroslaw Dylag, Natalja Kurbatova, Marco Brandizi, Tony Burdett, et al. Arrayexpress update—simplifying data submissions. *Nucleic acids research*, 43(D1):D1113–D1116, 2014.

[3] Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, et al. Intact: an open source molecular interaction database. *Nucleic acids research*, 32(suppl_1):D452–D455, 2004.

[4] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(D1):D358–D363, 2013.

[5] Ronald Carl Petersen, PS Aisen, LA Beckett, MC Donohue, AC Gamst, DJ Harvey, CR Jack, WJ Jagust, LM Shaw, AW Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3):201–209, 2010.

[6] Kyle Strimbu and Jorge A Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

[7] Biomarkers Definitions Working Group, Arthur J Atkinson Jr, Wayne A Colburn, Victor G DeGruttola, David L DeMets, Gregory J Downing, Daniel F Hoth, John A Oates, Carl C Peck, Robert T Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, 69(3):89–95, 2001.

[8] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[9] Ali Ebrahim, Elizabeth Brunk, Justin Tan, Edward J O'brien, Donghyuk Kim, Richard Szubin, Joshua A Lerman, Anna Lechner, Anand Sastry, Aarash Bordbar, et al. Multi-omic data integration enables discovery of hidden biological regularities. *Nature communications*, 7, 2016.

[10] Mathieu Vinken. The adverse outcome pathway concept: a pragmatic tool in toxicology. *Toxicology*, 312:158–165, 2013.

[11] Prashanth Suravajhala, Lisette JA Kogelman, and Haja N Kadarmideen. Multi-omic data integration and analysis using systems genomics approaches: methods and applications in animal production, health and welfare. *Genetics Selection Evolution*, 48(1):38, 2016.

[12] Vasileios Lapatas, Michalis Stefanidakis, Rafael C Jimenez, Allegra Via, and Maria Victoria Schneider. Data integration in biological research: an overview. *Journal of Biological Research-Thessaloniki*, 22(1):9, 2015.

[13] Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015.

[14] Roger Higdon, Rachel K Earl, Larissa Stanberry, Caitlin M Hudac, Elizabeth Montague, Elizabeth Stewart, Imre Janko, John Choiniere, William Broomall, Natali Kolker, et al. The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. *Omics: a journal of integrative biology*, 19(4):197–208, 2015.

[15] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics*, 16(2):85, 2015.

[16] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[17] Marine Jeanmougin, Aurélien De Reynies, Laetitia Marisa, Caroline Paccard, Gregory Nuel, and Mickael Guedj. Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PloS one*, 5(9):e12336, 2010.

[18] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[19] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[20] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[21] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[22] Joshua S Yuan, Ann Reed, Feng Chen, and C Neal Stewart. Statistical analysis of real-time pcr data. *BMC bioinformatics*, 7(1):85, 2006.

[23] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[24] John Boaz Lee, Ryan A Rossi, Sungchul Kim, Nesreen K Ahmed, and Eunyee Koh. Attention models in graphs: A survey. *arXiv preprint arXiv:1807.07984*, 2018.

[25] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[26] Adi L Tarca, Vincent J Carey, Xue-wen Chen, Roberto Romero, and Sorin Drăghici. Machine learning and its applications to biology. *PLoS computational biology*, 3(6):e116, 2007.

[27] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Inaki Inza, José A Lozano, Rubén Armananzas, Guzmán Santafé, Aritz Pérez, et al. Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112, 2006.

[28] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[29] Lauri Eronen and Hannu Toivonen. Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC bioinformatics*, 13(1):119, 2012.

[30] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[31] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[32] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[33] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[34] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[35] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.

[36] C. Masters. Alzheimer's disease. *Nature Reviews Disease Primers*, 1, 2015.

[37] Jill Adams. The proteome: discovering the structure and function of proteins. *Nature Education*, 1(3):6, 2008.

[38] Suzanne Clancy and William Brown. Translation: Dna to mrna to protein. *Nature Education*, 1(1):101, 2008.

[39] Minseung Kim, Navneet Rai, Violeta Zorraquino, and Ilias Tagkopoulos. Multi-omics integration accurately predicts cellular state in unexplored conditions for escherichia coli. *Nature communications*, 7, 2016.

[40] ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *nature*, 447(7146):799, 2007.

[41] Manolis Kellis, Barbara Wold, Michael P Snyder, Bradley E Bernstein, Anshul Kundaje, Georgi K Marinov, Lucas D Ward, Ewan Birney, Gregory E Crawford, Job Dekker, et al. Defining functional dna elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–6138, 2014.

[42] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[43] Gibran Hemani, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K Henders, Allan F McRae, Jian Yang, Greg Gibson, Nicholas G Martin, Andres Metspalu, et al. Detection and replication of epistasis influencing transcription in humans. *Nature*, 508(7495):249, 2014.

[44] Wen-Hua Wei, Gibran Hemani, and Chris S Haley. Detecting epistasis in human complex traits. *Nature Reviews. Genetics*, 15(11):722, 2014.

[45] Patrick C Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics*, 9(11):855, 2008.

[46] Jessica Alföldi and Kerstin Lindblad-Toh. Comparative genomics as a tool to understand evolution and disease. *Genome research*, 23(7):1063–1068, 2013.

[47] Tarjei S Mikkelsen, LaDeana W Hillier, Evan E Eichler, Michael C Zody, et al. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69, 2005.

[48] J Adams. Transcriptome: connecting the genome to gene function. *Nat Educ*, 1(1):195, 2008.

[49] Ryan J Taft, Ken C Pang, Timothy R Mercer, Marcel Dinger, and John S Mattick. Non-coding rnas: regulators of disease. *The Journal of pathology*, 220(2):126–139, 2010.

[50] Manel Esteller. Non-coding rnas in human disease. *Nature reviews. Genetics*, 12(12):861, 2011.

[51] Xiaohui Yan, Zhongyi Hu, Yi Feng, Xiaowen Hu, Jiao Yuan, Sihai D Zhao, Youyou Zhang, Lu Yang, Weiwei Shan, Qun He, et al. Comprehensive genomic characterization of long non-coding rnas across human cancers. *Cancer cell*, 28(4):529–540, 2015.

[52] Vikas Pejaver, Wei-Lun Hsu, Fuxiao Xin, A Keith Dunker, Vladimir N Uversky, and Predrag Radivojac. The structural and functional signatures

of proteins that undergo multiple events of post-translational modification. *Protein Science*, 23(8):1077–1093, 2014.

[53] Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, 2016.

[54] Sudipto Saha, Scott H Harrison, and Jake Yue Chen. Dissecting the human plasma proteome and inflammatory response biomarkers. *Proteomics*, 9(2):470–484, 2009.

[55] Nelson Freimer and Chiara Sabatti. The human phenome project. *Nature genetics*, 34(1):15–21, 2003.

[56] David Houle, Diddahally R Govindaraju, and Stig Omholt. Phenomics: the next challenge. *Nature Reviews. Genetics*, 11(12):855, 2010.

[57] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein–protein interactions. *Nature*, 340(6230):245–246, 1989.

[58] Alexandra Naba, Celine Reverdy, Daniel Louvard, and Monique Arpin. Spatial recruitment and activation of the fes kinase by ezrin promotes hgf-induced cell scattering. *The EMBO journal*, 27(1):38–50, 2008.

[59] Igor Stagljar, Chantal Korostensky, Nils Johnsson, and Stephan te Heesen. A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. *Proceedings of the National Academy of Sciences*, 95(9):5187–5192, 1998.

[60] Sven Eyckerman, Annick Verhee, José Van der Heyden, Irma Lemmens, Xaveer Van Ostade, Joël Vandekerckhove, and Jan Tavernier. Design and application of a cytokine-receptor-based interaction trap. *Nature cell biology*, 3(12):1114–1119, 2001.

[61] Ulrich Putz, Paul Skehel, and Dietmar Kuhl. A tri-hybrid system for the analysis and detection of rna-protein interactions. *Nucleic acids research*, 24(23):4838–4840, 1996.

[62] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, et al. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582, 2014.

[63] Mass spectrometry: Introduction, principle of mass spectrometry, components of mass spectrometer, applications. `http://www.premierbiosoft.com/tech_notes/mass-spectrometry.html`. Online; accessed April 2019.

[64] Arne H Smits and Michiel Vermeulen. Characterizing protein–protein interactions using mass spectrometry: challenges and opportunities. *Trends in biotechnology*, 34(10):825–834, 2016.

[65] Jamie Snider, Max Kotlyar, Punit Saraon, Zhong Yao, Igor Jurisica, and Igor Stagljar. Fundamentals of protein interaction network mapping. *Molecular systems biology*, 11(12):848, 2015.

[66] Pierre Legrain and Jean-Christophe Rain. Twenty years of protein interaction studies for biological function deciphering. *Journal of proteomics*, 107:93–97, 2014.

[67] TM Cafarelli, A Desbuleux, Y Wang, SG Choi, D De Ridder, and M Vidal. Mapping, modeling, and characterization of protein–protein interactions on a proteomic scale. *Current Opinion in Structural Biology*, 44:201–210, 2017.

[68] Hui Ge, Zhihua Liu, George M Church, and Marc Vidal. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nature genetics*, 29(4):482, 2001.

[69] Berend Snel, Gerrit Lehmann, Peer Bork, and Martijn A Huynen. String: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research*, 28(18):3442–3444, 2000.

[70] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, 45(D1):D362–D368, 2017.

[71] Magpix® and luminex 200™ instruments provide equivalent multiplex detection of inflammation pathway analytes. `https://www.scribd.com/document/80702597/MAGPIX-and-Luminex-200-Instruments-Provide-Equivalent\-Multiplex-Detection-of-Inflammation-Pathway-Analytes`. Online; accessed April 2019.

[72] Kehui Wang, Pathik D Wadhwa, Jennifer F Culhane, Edward L Nelson, et al. Validation and comparison of luminex multiplex cytokine analysis kits with elisa: determinations of a panel of nine cytokines in clinical sample culture supernatants. *Journal of reproductive immunology*, 66(2):175–191, 2005.

[73] Jun-Ming Zhang and Jianxiong An. Cytokines, inflammation and pain. *International anesthesiology clinics*, 45(2):27, 2007.

[74] Abdul Hye, Joanna Riddoch-Contreras, Alison L Baird, Nicholas J Ashton, Chantal Bazenet, Rufina Leung, Eric Westman, Andrew Simmons, Richard Dobson, Martina Sattlecker, et al. Plasma proteins predict conversion to dementia from prodromal disease. *Alzheimer's & Dementia*, 10(6):799–807, 2014.

[75] Christine Vogel, Raquel de Sousa Abreu, Daijin Ko, Shu-Yun Le, Bruce A Shapiro, Suzanne C Burns, Devraj Sandhu, Daniel R Boutz, Edward M Marcotte, and Luiz O Penalva. Sequence signatures and mrna concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular systems biology*, 6(1):400, 2010.

[76] Michael J Heller. Dna microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4(1):129–153, 2002.

[77] David J Lockhart, Helin Dong, Michael C Byrne, Maximillian T Follettie, Michael V Gallo, Mark S Chee, Michael Mittmann, Chunwei Wang, Michiko Kobayashi, Heidi Norton, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675–1680, 1996.

[78] Donna K Slonim and Itai Yanai. Getting started in gene expression microarray analysis. *PLoS computational biology*, 5(10):e1000543, 2009.

[79] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1):e78644, 2014.

[80] Kirk J Mantione, Richard M Kream, Hana Kuzelova, Radek Ptacek, Jiri Raboch, Joshua M Samuel, and George B Stefano. Comparing bioinformatic gene expression profiling methods: microarray and rna-seq. *Medical science monitor basic research*, 20:138, 2014.

[81] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):13, 2016.

[82] David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121, 2014.

[83] Paweł P Łabaj, Germán G Leparc, Bryan E Linggi, Lye Meng Markillie, H Steven Wiley, and David P Kreil. Characterization and improvement of rna-seq precision in quantitative transcript expression profiling. *Bioinformatics*, 27(13):i383–i391, 2011.

[84] Hubert Rehrauer, Lennart Opitz, Ge Tan, Lina Sieverling, and Ralph Schlapbach. Blind spots of quantitative rna-seq: the limits for assessing abundance, differential expression, and isoform switching. *BMC bioinformatics*, 14(1):370, 2013.

[85] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[86] Christian A Heid, Junko Stevens, Kenneth J Livak, and P Mickey Williams. Real time quantitative pcr. *Genome research*, 6(10):986–994, 1996.

[87] Mikael Kubista, José Manuel Andrade, Martin Bengtsson, Amin Forootan, Jiri Jonák, Kristina Lind, Radek Sindelka, Robert Sjöback, Björn Sjögreen, Linda Strömbom, et al. The real-time polymerase chain reaction. *Molecular aspects of medicine*, 27(2):95–125, 2006.

[88] Kary B Mullis, Henry A Erlich, Norman Arnheim, Glenn T Horn, Randall K Saiki, and Stephen J Scharf. Process for amplifying, detecting, and/or-cloning nucleic acid sequences, July 28 1987. US Patent 4,683,195.

[89] Figure.polymerase chain reaction, accessed on Aug 2017.

[90] Willard M Freeman, Stephen J Walker, and Kent E Vrana. Quantitative rt-pcr: pitfalls and potential. *Biotechniques*, 26(1):112–125, 1999.

[91] SA Bustin. Quantification of mrna using real-time reverse transcription pcr (rt-pcr): trends and problems. *Journal of molecular endocrinology*, 29(1):23–39, 2002.

[92] Aleksandar Radonić, Stefanie Thulke, Ian M Mackay, Olfert Landt, Wolfgang Siegert, and Andreas Nitsche. Guideline to reference gene selection for quantitative real-time pcr. *Biochemical and biophysical research communications*, 313(4):856–862, 2004.

[93] J Huggett, K Dheda, S Bustin, and A Zumla. Real-time rt-pcr normalisation; strategies and considerations. *Genes and immunity*, 6(4):279, 2005.

[94] Milena Petriccione, Francesco Mastrobuoni, Luigi Zampella, and Marco Scortichini. Reference gene selection for normalization of rt-qpcr gene expression data from actinidia deliciosa leaves infected with pseudomonas syringae pv. actinidiae. *Scientific reports*, 5:srep16961, 2015.

[95] Bartłomiej Kozera and Marcin Rapacz. Reference genes in real-time pcr. *Journal of applied genetics*, 54(4):391–406, 2013.

[96] Paul D Siebert. *Quantitative rt-PCR*. Springer, 1998.

[97] Michael Brown and Carl Wittwer. Flow cytometry: principles and clinical applications in hematology. *Clinical chemistry*, 46(8):1221–1229, 2000.

[98] James V Watson. *Introduction to flow cytometry*. Cambridge University Press, 2004.

[99] Ian D Odell and Deborah Cook. Immunofluorescence techniques. *The Journal of investigative dermatology*, 133(1):e4, 2013.

[100] Anna M Aalbers, Marry M Van Den Heuvel-Eibrink, Irith Baumann, Michael Dworzak, Henrik Hasle, Franco Locatelli, Barbara De Moerloose, Markus Schmugge, Ester Mejstrikova, Michaela Nováková, et al. Bone marrow immunophenotyping by flow cytometry in refractory cytopenia of childhood. *Haematologica*, pages haematol–2014, 2014.

[101] Arthur A Nery, Isis C Nascimento, Talita Glaser, Vinicius Bassaneze, José E Krieger, and Henning Ulrich. Human mesenchymal stem cells: from immunophenotyping by flow cytometry to clinical applications. *Cytometry Part A*, 83(1):48–61, 2013.

[102] Zhe Liu, Luke D Lavis, and Eric Betzig. Imaging live-cell dynamics and structure at the single-molecule level. *Molecular cell*, 58(4):644–659, 2015.

[103] Bastian Zimmer, Gabsang Lee, Nina V Balmer, Kesavan Meganathan, Agapios Sachinidis, Lorenz Studer, and Marcel Leist. Evaluation of developmental toxicants and signaling pathways in a functional test based on the migration of human neural crest cells. *Environmental health perspectives*, 120(8):1116, 2012.

[104] Bastian Zimmer, Giorgia Pallocca, Nadine Dreser, Sunniva Foerster, Tanja Waldmann, Joost Westerhout, Sylvie Julien, Karl-Heinz Krause, Christoph van Thriel, Jan G Hengstler, et al. Profiling of drugs and environmental chemicals for functional impairment of neural crest migration in a novel stem cell-based test battery. *Archives of toxicology*, 88(5):1109–1126, 2014.

[105] Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, et al. Functional snps in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, 32(4):650, 2002.

[106] Daniel Nowak, Wolf-Karsten Hofmann, and H Phillip Koeffler. Genome-wide mapping of copy number variations using snp arrays. *Transfusion Medicine and Hemotherapy*, 36(4):246–251, 2009.

[107] Magdalena Szumilas. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3):227, 2010.

[108] Geraldine M Clarke, Carl A Anderson, Fredrik H Pettersson, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133, 2011.

[109] Alicia Martin. Genomics and personalized medicine. `https://web.stanford.edu/class/gene210/files/writeups/2012/gwas_notes_AM.pdf`, 2012. Online; accessed 18-August-2017.

[110] Jim Stankovich. Statistical analysis of genome-wide association (gwas) data. `http://bioinformatics.org.au/ws09/presentations/Day3_JStankovich.pdf`, 2009. Online; accessed 18-August-2017.

[111] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452–1458, 2013.

[112] Valgerdur Steinthorsdottir, Gudmar Thorleifsson, Inga Reynisdottir, Rafn Benediktsson, Thorbjorg Jonsdottir, G Bragi Walters, Unnur Styrkarsdottir, Solveig Gretarsdottir, Valur Emilsson, Shyamali Ghosh, et al. A variant in cdkal1 influences insulin response and risk of type 2 diabetes. *Nature genetics*, 39(6):770, 2007.

[113] Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *New England journal of medicine*, 363(2):166–176, 2010.

[114] Thomas A Pearson and Teri A Manolio. How to interpret a genome-wide association study. *Jama*, 299(11):1335–1344, 2008.

[115] Patrick Kemmeren, Nynke L van Berkum, Jaak Vilo, Theo Bijma, Rogier Donders, Alvis Brazma, and Frank CP Holstege. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular cell*, 9(5):1133–1143, 2002.

[116] Cecily J Wolfe, Isaac S Kohane, and Atul J Butte. Systematic survey reveals general applicability of" guilt-by-association" within gene coexpression networks. *BMC bioinformatics*, 6(1):227, 2005.

[117] Radek Szklarczyk, Wout Megchelenbrink, Pavel Cizek, Marie Ledent, Gonny Velemans, Damian Szklarczyk, and Martijn A Huynen. Weget: predicting new genes for molecular systems by weighted co-expression. *Nucleic acids research*, 44(D1):D567–D573, 2016.

[118] Yang Yang, Leng Han, Yuan Yuan, Jun Li, Nainan Hei, and Han Liang. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications*, 5:3231, 2014.

[119] Priit Adler, Raivo Kolde, Meelis Kull, Aleksandr Tkachenko, Hedi Peterson, Jüri Reimand, and Jaak Vilo. Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome biology*, 10(12):R139, 2009.

[120] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, 2012.

[121] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein–protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS computational biology*, 3(4):e43, 2007.

[122] Alpan Raval and Animesh Ray. *Introduction to biological networks*. CRC Press, 2013.

[123] Ronald Jansen, Dov Greenbaum, and Mark Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome research*, 12(1):37–46, 2002.

[124] Christian Von Mering, Roland Krause, Berend Snel, Michael Cornell, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399, 2002.

[125] Pavlos Pavlidis and Nikolaos Alachiotis. A survey of methods and tools to detect recent and strong positive selection. *Journal of Biological Research-Thessaloniki*, 24(1):7, 2017.

[126] Johan Nilsson, Mats Grahn, and Anthony PH Wright. Proteome-wide evidence for enhanced positive darwinian selection within intrinsically disordered regions in proteins. *Genome biology*, 12(7):R65, 2011.

[127] Towfique Raj, Joshua M Shulman, Brendan T Keenan, Lori B Chibnik, Denis A Evans, David A Bennett, Barbara E Stranger, and Philip L De Jager. Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. *The American Journal of Human Genetics*, 90(4):720–726, 2012.

[128] Kun Tang, Kevin R Thornton, and Mark Stoneking. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS biology*, 5(7):e171, 2007.

[129] Ziheng Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.

[130] Rasmus Nielsen, Carlos Bustamante, Andrew G Clark, Stephen Glanowski, Timothy B Sackton, Melissa J Hubisz, Adi Fledel-Alon, David M Tanenbaum, Daniel Civello, Thomas J White, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS biology*, 3(6):e170, 2005.

[131] Daven C Presgraves. The molecular evolutionary basis of species formation. *Nature reviews. Genetics*, 11(3):175, 2010.

[132] Eric J Vallender and Bruce T Lahn. Positive selection on the human genome. *Human Molecular Genetics*, 13(suppl_2):R245–R254, 2004.

[133] Ross C Hardison. Comparative genomics. *PLoS biology*, 1(2):e58, 2003.

[134] Ziheng Yang and Mario Dos Reis. Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution*, 28(3):1217–1228, 2010.

[135] Nikolaos Alachiotis, Alexandros Stamatakis, and Pavlos Pavlidis. Omegaplus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics*, 28(17):2274–2275, 2012.

[136] Pavlos Pavlidis, Daniel Živković, Alexandros Stamatakis, and Nikolaos Alachiotis. Sweed: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular biology and evolution*, 30(9):2224–2234, 2013.

[137] Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, 11(20):2463–2468, 2002.

[138] Thierry Schüpbach, Ioannis Xenarios, Sven Bergmann, and Karen Kapur. Fastepistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, 26(11):1468–1469, 2010.

[139] Benjamin Lehne and Thomas Schlitt. Protein-protein interaction databases: keeping up with growing interactomes. *Human genomics*, 3(3):291, 2009.

[140] IntAct. Synapse - interactions of proteins with an established role in the presynapse., feb 2019.

[141] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34, 1999.

[142] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2017.

[143] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2013.

[144] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, page gkx1132, 2017.

[145] Mathieu Vinken. Adverse outcome pathways as tools to assess drug-induced toxicity. In *In Silico Methods for Predicting Drug Toxicity*, pages 325–337. Springer, 2016.

[146] Anna Bal-Price, Kevin M Crofton, Magdalini Sachana, Timothy J Shafer, Mamta Behl, Anna Forsby, Alan Hargreaves, Brigitte Landesmann, Pamela J Lein, Jochem Louisse, et al. Putative adverse outcome pathways relevant to neurotoxicity. *Critical reviews in toxicology*, 45(1):83–91, 2015.

[147] Tim Hubbard, Daniel Barker, Ewan Birney, Graham Cameron, Yuan Chen, L Clark, Tony Cox, J Cuff, Val Curwen, Thomas Down, et al. The ensembl genome database project. *Nucleic acids research*, 30(1):38–41, 2002.

[148] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, et al. Ensembl 2012. *Nucleic acids research*, 40(D1):D84–D90, 2011.

[149] Michael Rebhan, Vered Chalifa-Caspi, Jaime Prilusky, and Doron Lancet. Genecards: integrating information about genes, proteins and diseases. *Trends in Genetics*, 13(4):163, 1997.

[150] Marilyn Safran, Irina Dalah, Justin Alexander, Naomi Rosen, Tsippi Iny Stein, Michael Shmoish, Noam Nativ, Iris Bahir, Tirza Doniger, Hagit Krug, et al. Genecards version 3: the human gene integrator. *Database*, 2010:baq020, 2010.

[151] Robert Hoffmann. A wiki for the life sciences where authorship matters. *Nature genetics*, 40(9):1047–1051, 2008.

[152] Gregory D Schuler, Jonathan A Epstein, Hitomi Ohkawa, and Jonathan A Kans. [10] entrez: Molecular biology database and retrieval system. *Methods in enzymology*, 266:141–162, 1996.

[153] Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, et al. Gene: a gene-centered information resource at ncbi. *Nucleic acids research*, 43(D1):D36–D42, 2014.

[154] Jüri Reimand, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(suppl_2):W193–W200, 2007.

[155] Jüri Reimand, Tambet Arak, Priit Adler, Liis Kolberg, Sulev Reisberg, Hedi Peterson, and Jaak Vilo. g: Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic acids research*, 44(W1):W83–W89, 2016.

[156] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44, 2009.

[157] UniProt Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014.

[158] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl_1):D767–D772, 2008.

[159] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, et al. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250, 2010.

[160] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[161] Andrew R Webb. *Statistical pattern recognition*. John Wiley & Sons, 2003.

[162] M Anthony Wong. Asymptotic properties of univariate sample k-means clusters. *Journal of Classification*, 1(1):255–270, 1984.

[163] Raffaele Giancarlo, Davide Scaturro, and Filippo Utro. Computational cluster validation for microarray data analysis: experimental assessment of clest, consensus clustering, figure of merit, gap statistics and model explorer. *BMC bioinformatics*, 9(1):462, 2008.

[164] Mark Ming-Tso Chiang and Boris Mirkin. Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. *Journal of classification*, 27(1):3–40, 2010.

[165] Raivo Kolde. Pheatmap: pretty heatmaps. *R package version*, 61, 2012.

[166] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[167] Trudie Strauss and Michael Johan von Maltitz. Generalising ward's method for use with manhattan distances. *PloS one*, 12(1):e0168288, 2017.

[168] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of Classification*, 31(3):274–295, 2014.

[169] Ryota Suzuki and Hidetoshi Shimodaira. An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters. In *The Fifteenth International Conference on Genome Informatics*, volume 34. Pacifico Convention Plaza Yokohama Japan, 2004.

[170] Ryota Suzuki and Hidetoshi Shimodaira. Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, 2006.

[171] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7(2):179–188, 1936.

[172] Aleix M Martínez and Avinash C Kak. Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):228–233, 2001.

[173] Corrado Priami and Melissa J Morine. *Analysis of biological systems*. World Scientific, 2015.

[174] Jieping Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6(Apr):483–502, 2005.

[175] Bin Zhang, Chris Gaiteri, Liviu-Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, et al. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707–720, 2013.

[176] Jason D Gray, Todd G Rubin, Richard G Hunter, and Bruce S McEwen. Hippocampal gene expression changes underlying stress sensitization and recovery. *Molecular psychiatry*, 19(11):1171, 2014.

[177] Xaquin Castro Dopico, Marina Evangelou, Ricardo C Ferreira, Hui Guo, Marcin L Pekalski, Deborah J Smyth, Nicholas Cooper, Oliver S Burren, Anthony J Fulford, Branwen J Hennig, et al. Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nature communications*, 6:7000, 2015.

[178] Cevat Yazici, Kader Köse, Serap Utaş, Esen Tanrikulu, and Nazan Taşlidere. A novel approach in psoriasis: first usage of known protein oxidation markers to prove oxidative stress. *Archives of dermatological research*, 308(3):207–212, 2016.

[179] Tite M Mikobi, Prosper Lukusa Tshilobo, Michel N Aloni, Pierre Z Akilimali, Georges Mvumbi-Lelo, and Jean Marie Mbuyi-Muamba. Clinical phenotypes and the biological parameters of congolese patients suffering from sickle cell anemia: A first report from central africa. *Journal of clinical laboratory analysis*, 2017.

[180] Konrad Janowski, Stanisława Steuden, and Jarosław Bogaczewicz. Clinical and psychological characteristics of patients with psoriasis reporting various frequencies of pruritus. *International journal of dermatology*, 53(7):820–829, 2014.

[181] Gerry P Quinn and Michael J Keough. *Experimental design and data analysis for biologists*. Cambridge University Press, 2002.

[182] Mary L McHugh. Multiple comparison analysis testing in anova. *Biochemia medica*, 21(3):203–209, 2011.

[183] Colin S Gillespie, Guiyuan Lei, Richard J Boys, Amanda Greenall, and Darren J Wilkinson. Analysing time course microarray data using bioconductor: a case study using yeast2 affymetrix arrays. *BMC research notes*, 3(1):81, 2010.

[184] Chris Wild. The wilcoxon rank-sum test.

[185] Jean Dickinson Gibbons and Subhabrata Chakraborti. Nonparametric statistical inference. In *International encyclopedia of statistical science*, pages 977–979. Springer, 2011.

[186] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, 2003.

[187] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[188] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[189] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[190] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[191] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

[192] Yongchao Ge, Stuart C Sealfon, and Terence P Speed. Some step-down procedures controlling the false discovery rate under dependence. *Statistica Sinica*, 18(3):881, 2008.

[193] Warren J Ewens and Gregory R Grant. *Statistical methods in bioinformatics: an introduction*. Springer Science & Business Media, 2006.

[194] Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[195] Marc Jung, Hedi Peterson, Lukas Chavez, Pascal Kahlem, Hans Lehrach, Jaak Vilo, and James Adjaye. A data integration approach to mapping oct4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. *PLoS One*, 5(5):e10709, 2010.

[196] Teny Handhayani, Ito Wasito, Mujiono Sadikin, et al. Kernel based integration of gene expression and dna copy number. In *Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on*, pages 303–308. IEEE, 2013.

[197] Gert RG Lanckriet, Tijl De Bie, Nello Cristianini, Michael I Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[198] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019.

[199] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[200] Renzo Angles and Claudio Gutierrez. Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1):1, 2008.

[201] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. arxiv e-prints (nov. 2017). *arXiv preprint arXiv:1711.08752*, 2017.

[202] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[203] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[204] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[205] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[206] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

[207] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.

[208] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[209] Anders Krogh. What are artificial neural networks? *Nature biotechnology*, 26(2):195, 2008.

[210] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[211] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. *arXiv preprint arXiv:1806.01973*, 2018.

[212] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[213] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[214] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[215] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[216] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[217] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *arXiv preprint arXiv:1211.0053*, 2012.

[218] CSIRO data61 investigative analytics. stellar-ml v0.2.0: Machine learning on graphs. `https://github.com/stellargraph`, 2018.

[219] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.

[220] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[221] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):83, 2017.

[222] Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system: integrating'omics' data sets. *Nature reviews. Molecular cell biology*, 7(3):198, 2006.

[223] Barbara E Bierer, Mercè Crosas, and Heather H Pierce. Data authorship as an incentive to data sharing, 2017.

[224] Masaaki Hokama, Sugako Oka, Julio Leon, Toshiharu Ninomiya, Hiroyuki Honda, Kensuke Sasaki, Toru Iwaki, Tomoyuki Ohara, Tomio Sasaki, Frank M LaFerla, et al. Altered expression of diabetes-related genes in alzheimer's disease brains: the hisayama study. *Cerebral cortex*, 24(9):2476–2488, 2013.

[225] Rita Guerreiro, Aleksandra Wojtas, Jose Bras, Minerva Carrasquillo, Ekaterina Rogaeva, Elisa Majounie, Carlos Cruchaga, Celeste Sassi, John SK Kauwe, Steven Younkin, et al. Trem2 variants in alzheimer's disease. *New England Journal of Medicine*, 368(2):117–127, 2013.

[226] Kaj Blennow. Biomarkers in alzheimer's disease drug development. *Nature medicine*, 16(11):1218–1222, 2010.

[227] Randall J Bateman, Chengjie Xiong, Tammie LS Benzinger, Anne M Fagan, Alison Goate, Nick C Fox, Daniel S Marcus, Nigel J Cairns, Xianyun Xie, Tyler M Blazey, et al. Clinical and biomarker changes in dominantly inherited alzheimer's disease. *N Engl J Med*, 2012(367):795–804, 2012.

[228] IntAct. Intact.

[229] Eric Nagele, Min Han, Cassandra DeMarshall, Benjamin Belinka, and Robert Nagele. Diagnosis of alzheimer's disease based on disease-specific autoantibody profiles in human sera. *PloS one*, 6(8):e23112, 2011.

[230] Noa Rappaport, Noam Nativ, Gil Stelzer, Michal Twik, Yaron Guan-Golan, Tsippi Iny Stein, Iris Bahir, Frida Belinky, C Paul Morrey, Marilyn Safran,

et al. Malacards: an integrated compendium for diseases and their annotation. *Database*, 2013, 2013.

[231] Nino Spataro, Juan Antonio Rodríguez, Arcadi Navarro, and Elena Bosch. Properties of human disease genes and the role of genes linked to mendelian disorders in complex disease aetiology. *Human molecular genetics*, 26(3):489–500, 2017.

[232] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *Icwsm*, 8(2009):361–362, 2009.

[233] Martin Rosvall and Carl T Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331, 2007.

[234] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[235] Uri Alon. Biological networks: the tinkerer as an engineer. *Science*, 301(5641):1866–1867, 2003.

[236] Leland H Hartwell, John J Hopfield, Stanislas Leibler, and Andrew W Murray. From molecular to modular cell biology. *Nature*, 402(6761supp):C47, 1999.

[237] Dana H Ballard. Modular learning in neural networks. In *AAAI*, pages 279–284, 1987.

[238] Iris Jansen, Jeanne Savage, Kyoko Watanabe, Julien Bryois, Dylan Williams, Stacy Steinberg, Julia Sealock, Ida Karlsson, Sara Hagg, Lavinia Athanasiu, et al. Genetic meta-analysis identifies 10 novel loci and functional pathways for alzheimer's disease risk. *bioRxiv*, page 258533, 2018.

[239] Brian W Kunkle, Benjamin Grenier-Boley, Rebecca Sims, Joshua C Bis, Adam C Naj, Anne Boland, Maria Vronskaya, Sven J van der Lee, Alex Amlie-Wolf, Celine Bellenguez, et al. Meta-analysis of genetic association with diagnosed alzheimer's disease identifies novel risk loci and implicates abeta, tau, immunity and lipid processing. *bioRxiv*, page 294629, 2018.

[240] Riccardo E Marioni, Sarah E Harris, Qian Zhang, Allan F McRae, Saskia P Hagenaars, W David Hill, Gail Davies, Craig W Ritchie, Catharine R Gale, John M Starr, et al. Gwas on family history of alzheimer's disease. *Translational psychiatry*, 8, 2018.

[241] Jimmy Z Liu, Yaniv Erlich, and Joseph K Pickrell. Case–control association mapping by proxy using family history of disease. *Nature genetics*, 49(3):325, 2017.

[242] Towfique Raj, Yang I Li, Garrett Wong, Jack Humphrey, Minghui Wang, Satesh Ramdhani, Ying-Chih Wang, Bernard Ng, Ishaan Gupta, Vahram Haroutunian, et al. Integrative transcriptome analyses of the aging brain

implicate altered splicing in alzheimer's disease susceptibility. *Nature genetics*, 50(11):1584, 2018.

[243] Joel M Gelfand and Howa Yeung. Metabolic syndrome in patients with psoriatic disease. *The Journal of Rheumatology Supplement*, 89:24–28, 2012.

[244] C Pouplard, E Brenaut, C Horreau, T Barnetche, L Misery, M-A Richard, S Aractingi, F Aubin, B Cribier, P Joly, et al. Risk of cancer in psoriasis: a systematic review and meta-analysis of epidemiological studies. *Journal of the European Academy of Dermatology and Venereology*, 27(s3):36–46, 2013.

[245] Hsien-Yi Chiu, Yu-Pin Cheng, and Tsen-Fang Tsai. T helper type 17 in psoriasis: From basic immunology to clinical practice. *Dermatologica Sinica*, 30(4):136–141, 2012.

[246] Éric Toussirot. The il23/th17 pathway as a therapeutic target in chronic inflammatory diseases. *Inflammation & Allergy-Drug Targets (Formerly Current Drug Targets-Inflammation & Allergy)*, 11(2):159–168, 2012.

[247] Michelle A Lowes, Mayte Suárez-Fariñas, and James G Krueger. Immunology of psoriasis. *Annual review of immunology*, 32:227–255, 2014.

[248] Andrew Blauvelt, Kristian Reich, Tsen-Fang Tsai, Stephen Tyring, Francisco Vanaclocha, Külli Kingo, Michael Ziv, Andreas Pinter, Ronald Vender, Sophie Hugot, et al. Secukinumab is superior to ustekinumab in clearing skin of subjects with moderate-to-severe plaque psoriasis up to 1 year: Results from the clear study. *Journal of the American Academy of Dermatology*, 76(1):60–69, 2017.

[249] Yun Liang, Mrinal K Sarkar, Lam C Tsoi, and Johann E Gudjonsson. Psoriasis: a mixed autoimmune and autoinflammatory disease. *Current Opinion in Immunology*, 49:1–8, 2017.

[250] Joaquin J Rivas Bejarano and Wendell C Valdecantos. Psoriasis as autoinflammatory disease. *Dermatologic clinics*, 31(3):445–460, 2013.

[251] William Abramovits and Marcial Oquendo. *Autoinflammatory Disorders, an Issue of Dermatologic Clinics, E-Book*, volume 31. Elsevier Health Sciences, 2013.

[252] Anne Davidson and Betty Diamond. Autoimmune diseases. *New England Journal of Medicine*, 345(5):340–350, 2001.

[253] Frank O. Nestle, Daniel H. Kaplan, and Jonathan Barker. Psoriasis. *New England Journal of Medicine*, 361(5):496–509, 2009. PMID: 19641206.

[254] Robert Gentleman, Ross Ihaka, D Bates, et al. The r project for statistical computing. *URL: http://www. r-project. org/254*, 2009.

[255] Liisi Šahmatova, Elena Sügis, Marina Šunina, Helen Hermann, Ele Prans, Maire Pihlap, Kristi Abram, Ana Rebane, Hedi Peterson, Pärt Peterson,

et al. Signs of innate immune activation and premature immunosenescence in psoriasis patients. *Scientific Reports*, 7(1):7553, 2017.

[256] Florence Jacomet, Emilie Cayssials, Sara Basbous, Anaïs Levescot, Nathalie Piccirilli, Deborah Desmier, Aurélie Robin, Anne Barra, Christine Giraud, François Guilhot, et al. Evidence for eomesodermin-expressing innate-like cd8+ kir/nkg2a+ t cells in human adults and cord blood samples. *European journal of immunology*, 45(7):1926–1933, 2015.

[257] Jason T White, Eric W Cross, Matthew A Burchill, Thomas Danhorn, Martin D McCarter, Hugo R Rosen, Brian O'Connor, and Ross M Kedl. Virtual memory t cells develop and mediate bystander protective immunity in an il-15-dependent manner. *Nature communications*, 7:11291, 2016.

[258] Valérie Martinet, Sandrine Tonon, David Torres, Abdulkader Azouz, Muriel Nguyen, Arnaud Kohler, Véronique Flamand, Chai-An Mao, William H Klein, Oberdan Leo, et al. Type i interferons regulate eomesodermin expression and the development of unconventional memory cd8+ t cells. *Nature communications*, 6, 2015.

[259] Mari H Tervaniemi, Shintaro Katayama, Tiina Skoog, H Annika Siitonen, Jyrki Vuola, Kristo Nuutila, Raija Sormunen, Anna Johnsson, Sten Linnarsson, Sari Suomela, et al. Nod-like receptor signaling and inflammasome-related pathways are highlighted in psoriatic epidermis. *Scientific reports*, 6:22745, 2016.

[260] Stefan Winkler and Angela Rösen-Wolff. Caspase-1: an integral regulator of innate immunity. In *Seminars in immunopathology*, volume 37, pages 419–427. Springer, 2015.

[261] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[262] Olive Jean Dunn. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252, 1964.

[263] Claudio Franceschi, Paolo Garagnani, Giovanni Vitale, Miriam Capri, and Stefano Salvioli. Inflammaging and 'garb-aging'. *Trends in Endocrinology & Metabolism*, 28(3):199–212, 2017.

[264] T Fulop, G Dupuis, S Baehl, A Le Page, K Bourgade, E Frost, JM Witkowski, G Pawelec, A Larbi, and S Cunnane. From inflamm-aging to immune-paralysis: a slippery slope during aging for immune-adaptation. *Biogerontology*, 17(1):147–157, 2016.

[265] Branca I Pereira and Arne N Akbar. Convergence of innate and adaptive immunity during human aging. *Frontiers in immunology*, 7, 2016.

[266] Tamas Fulop, Anis Larbi, Rami Kotb, and Graham Pawelec. Immunology of aging and cancer development. In *Cancer and Aging*, volume 38, pages 38–48. Karger Publishers, 2013.

[267] Evelyna Derhovanessian, Rafael Solana, Anis Larbi, and Graham Pawelec. Immunity, ageing and cancer. *Immunity & Ageing*, 5(1):11, 2008.

[268] Philipp Kügler, Bastian Zimmer, Tanja Waldmann, Birte Baudis, Sten Ilmjärv, Jürgen Hescheler, Phil Gaughwin, Patrik Brundin, Wiliam Mundy, Anna K Bal-Price, et al. Markers of murine embryonic and neural stem cells, neurons and astrocytes: reference points for developmental neurotoxicity testing. *Alternatives to animal experimentation: ALTEX*, 27(1):16–42, 2010.

[269] Elisabeth Dupin and Lukas Sommer. Neural crest progenitors and stem cells: from early development to adulthood. *Developmental biology*, 366(1):83–95, 2012.

[270] Anna Keyte and Mary Redmond Hutson. The neural crest in cardiac congenital anomalies. *Differentiation*, 84(1):25–40, 2012.

[271] Francesca Di Renzo, Maria L Broccia, Erminio Giavini, and Elena Menegola. Antifungal triazole derivative triadimefon induces ectopic maxillary cartilage by altering the morphogenesis of the first branchial arch. *Birth Defects Research Part B: Developmental and Reproductive Toxicology*, 80(1):2–11, 2007.

[272] Leah C Fuller, Shannon K Cornelius, Charles W Murphy, and Darrell J Wiens. Neural crest cell motility in valproic acid. *Reproductive Toxicology*, 16(6):825–839, 2002.

[273] Grace Patlewicz, Nicholas Ball, Richard A Becker, Ewan D Booth, Mark TD Cronin, Dinant Kroese, David Steup, Ben Van Ravenzwaay, and Thomas Hartung. Food for thought: read-across approaches–misconceptions, promises and challenges ahead. *Alternatives to Animal Experimentation: ALTEX*, 31(4):387–396, 2014.

[274] Nadine Dreser, Bastian Zimmer, Christian Dietz, Elena Sügis, Giorgia Pallocca, Johanna Nyffeler, Johannes Meisig, Nils Blüthgen, Michael R Berthold, Tanja Waldmann, et al. Grouping of histone deacetylase inhibitors and other toxicants disturbing neural crest migration by transcriptional profiling. *Neurotoxicology*, 50:56–70, 2015.

ACKNOWLEDGEMENTS

I am very grateful to my supervisor Prof. Jaak Vilo for giving me an opportunity to enter the field of bioinformatics and to follow my dreams.

I want to thank my supervisor Dr. Hedi Peterson for introducing me to the field of biological network analysis, and over years becoming much more than a mentor and a colleague.

I want to thank my academic family, including but not limited to Tauno Metsalu, Konstantin Tretjakov, Liis Kolberg, Ahto Salumets, Dima Fishman for great discussions and brainstorms. I am also very thankful to a wizard of troubleshooting Ivan Kuzmin. Special thanks go to Priit Adler and Raivo Kolde for guiding me through the difficult understanding of the biological tasks and helping me to translate them into the computational language in the beginning of my journey in bioinformatics. I am very grateful to my friend Anna Leontjeva for the endless support accompanied by the deep expertise in various aspects of data science.

Additionally, I want to thank Prof. Ioannis Xenarios for the valuable on-time advises and guidance for the survival in academia and for the management of a large project. I was very lucky to meet this mentor. Also my scientific development progress would not be possible without my co-authors of the papers, especially Prof. Marcel Leist and Prof. Kai Kisand. It was a pleasure to work together.

My doctoral studies have been supported by AgedBrainSYSBIO consortium under funding from the European Union Seventh Framework Programme for research, technological development and demonstration (FP7, grant agreement No 305299), ESNATS consortium FP7 (grant agreement 201619), Tiger University program of the Information Technology Foundation for Education and European Social Fund Doctoral Studies and Internationalisation Programme DoRa.

I am very thankful to my friends Ekaterina Koryagina, Tatiana and Mikhail Shinkarev, Vladimir Berkhov. for being there for me and decreasing stress level many times. Academic friendships also have no boundaries, thank you Sanja Šćepanović, Pelle Jakovits and Yanina Timasheva for keeping these traditions.

I had a twisty and at times rocky path through my PhD, there were a lot ups and downs. I am very happy that I have my family standing by me on this journey. I am very grateful to my husband Andri for believing in me even when I stopped seeing a meaning in what I do; to my grandmother Lidia for the unconditional love, motivation and wise advice, and for my parents Ludmila and Sergey for being great role models.

SUMMARY IN ESTONIAN

Mitmekesiste bioloogiliste andmete ühendamine ja analüüs

Tehnoloogia kiire areng koos eksperimentaalandmete madalama tootmiskulukusega on viinud bioteadustes suure hulga andmete genereerimiseni. Peamisteks sellisteks valdkondadeks on erinevad „oomikad“ nagu näiteks proteoomika, genoomika ja trankstriptoomika. Need erinevad valdkonnad võimaldavad uurida elusolendeid, bioloogilisi süsteeme ning haigusi erinevatest aspektidest. Saadud andmeid talletatakse valdkonnaspetsiifilistes andmebaasides nagu näiteks ArrayExpress, IntAct ja ANDI. Sellistes andmebaasides hoiustatakse näiteks geeniekspressiooniandmeid, proteiin-proteiin interaktsioone kui ka paljusid teisi bioloogilisi andmeid. Taoliste andmete analüüs on võimaldanud saada väärtuslikku informatsiooni nii erinevate bioloogiliste protsesside kohta, kui ka leida biomarkereid, mis viitavad teatud patoloogilisele protsessile, iseloomustavad vastust ravimitele või aitavad hinnata terapeutilise sekkumise edukust. On aga selge, et eraldi neid andmeid analüüsides jääb tervikpilt nägemata. Selleks, et luua huvipakkuvast bioloogilisest protsessist (näiteks haigusest) süstemaatiline arusaam, on vaja neid andmestikke kombineerida. See aga tõstatab küsimuse, kuidas on ikkagi neid erinevaid andmestikke mõistlik kombineerida. Üheks esimeseks sammuks võiks olla usaldusväärsete ja mahukate haigusspetsiifiliste andmestike moodustamine. On ka selge, et tuleb kindlustada, et andmed oleksid kvaliteetsed ning seetõttu on väga oluline osa andmete eeltöötlemisel, mis kätkeb endas mitteusaldusväärsete andmete filtreerimist kui ka andmestike viimist vastavusse FAIR standarditele. Antud andmestikke oleks võimalik nii kasutada uute hüpoteeside testimiseks kui ka kombineerimiseks teiste andmetega. Üheks probleemiks on aga ka andmete heterogeensus, nimelt andmeid on toodetud väga erinevates eksperimentaalsetes tingimustes kui ka väga erinevate vahenditega ning pole selge, kuidas need suhestuvad uuritava fenotüübiga (nt haigusega). Sellest ajendatult on hakatud arendama andmeteaduspõhiseid meetodeid, et bioloogiliselt tähendusrikkalt kombineerida olemasolevaid andmeid. Tihtilugu käib see läbi mitmetasemelise analüüsi, mis siis aitab täpsemalt uuritavat bioloogilist protsessi kirjeldada.

Tänapäeval puudub ühtne ja standardne viis, kuidas bioloogilisi andmeid integreerida. Üldjuhul lähtutakse konkreetsetest saadavalolevates andmetest (eksperimentaalsed, arvutuslikud ja valdkonnaspetsiifilised), mis koos uuringudisainiga määravad ära, millist andmeanalüüsi lähenemist tuleks kasutada. Käesolevad väitekirjas oleme kasutanud kahte eri lähenemist, mis osaliselt põhinevat Ritchie jt poolt välja pakutud. Antud väitekiri kirjeldab, kuidas neid metoodikaid rakendada erinevate bioloogiliste küsimuste uurimiseks ning uute bioloogiliste teadmiste omandamiseks. Need lähenemised põhinevad uuritava fenotüübi ja individuaalsete saadaval olevate andmestike omavahelistel seostes ning meie hinnangul on need sobivad erinevate uuringudisainide puhul. Täpsemalt, need kaks gruppi on: mitmetasemeline ja transformatsioonipõhine integreerimine. Mõlemad koosne-

vad erinevatest andmeteaduse meetoditest, mis kombineerituna võimaldavad leida andmestikus ka muidu raskesti leitavaid seoseid. Mitmetasemeline andmete integreerimine tähendab, et andmestikud analüüsitakse eraldi ning seejärel kombineeritakse tulemused, mille põhjal tehakse juba üldistusi lähtuvalt valdkonnaspetsiifilistest teadmistest. Sõltuvalt uuritava tunnuse ja andmete vahelisest seose iseloomust võib rakendada seal erinevaid andmeteaduse metoodikaid. Nende hulka kuuluvad näiteks regressiooni ja klassifikatsiooni meetodid ,nagu näiteks lineaarsed mudelid, juhumetsad (ingl. k. Random Forest), mitmesugused testid nagu näiteks t-test ja mitteparameetriline Wilcoxon test, klasterdamismeetodid, nagu k-means. Transformatsioonipõhine integreerimine kätkeb endas aga andmete transformatsiooni vahepealseks vormiks, näiteks graafiks. Võrreldes mitmetasemelise analüüsiga, kus kasutatakse andmemaatrikseid, on transformatsioonipõhine lähenemine oluliselt kompleksem. Seal kasutatakse heterogeenseid graafe, milles tipud ja servad on kirjeldatud läbi erinevate bioloogiliste muutujate.

Masinõppe-põhised meetodid on juba bioteadustes laialt levinud, võimaldades ennustada bioloogilisi väärtusi tegemata tugevaid eeldusi täpsete bioloogiliste mehhanismide kohta. Sel juhul on muidugi erilisel kohal nii bioloogiliste tunnuste valik ning nende disain, mille edukal valikul on võimalik saada täpseid mudeleid. Klassikaliste masinõppe meetodite rakendamine graafianalüüsis on keeruline kuna need meetodid vajavad tihti käsitsi kombineeritud tunnuseid, on piiratud paindlikkusega ning arvutuslikult kulukad. Viimastel aastatel on ka populaarust kogunud süvaõppe (ingl. k. deep learning), mida rakendatakse edukalt piltide, teksti ja videote analüüsis. Sellised mudelite puhul pole enam tunnuste disain vajalik, kuna neid eraldatakse automaatselt. Kuid ka neid meetodeid ei saa otse graafidel rakendada näiteks kuna graafi tipud pole järjestatud. Kuid graafide jaoks on disainitud eraldi suvaõppemeetodid, mis on tuntud kui graafi konvolutsioonilised tehisnärvivõrgud (ingl. k. graph convolutional networks(GCN)). GCN-id suudavad efektiivselt ära kasutada graafide struktuuri ning tippude vahelist informatsiooni. Käesolevad väitekirjas näitame läbi praktiliste probleemide lahendamise, kuidas integreerida ja analüüsida heterogeenseid bioloogilisi andmeid kolmes valdkonnas: immunoloogias, toksikoloogias ja Alzheimeri tõve uuringutes. Teine peatükk keskendub erinevatele oomikaandmetele ning eksperimentidele, kuidas neid toodetakse. Liskas kirjeldame selles peatükis, kuidas on üks andmetüüp seotud teisega. Kolmandas peatükis me tutvustame eelmainitud kahte andmete integreerimise lähenemist – transformatsioonipõhist ja mitmetasemelist. Seejärel me anname ülevaate andmeteaduse metoodikatest, mida me kasutasime andmeanalüüsi juures, alates hüpoteeside testimisest kuni juhendatud õppe ja graafide spetsiifilise süvaõppeni. Järgnevates peatükkides (4-6), me näitame, kuidas integratsioonipõhine analüüs võimaldab paremini aru saada bioloogilistes protsessidest järgmistes kolmes valdkonnas: Alzheimeri tõbi, toksilisuse testimine ja immunoloogia.

Täpsemalt, neljandas peatükis, mis põhineb publikatsioonil I, kirjeldame andmete integreerimist rakendatuna Alzheimeri tõve uurimisele. Alzheimeri tõbi on vanusega seotud neurodegeneratiivne haigus, mis progresseerub vanusega ning

viib surmani. Alzheimeri tõve sümptomite vähendamiseks kasutatakse mitmeid ravimeid, kuid ükski olemasolev ravi meetod ei saa muuta haiguse aluseks olevaid protsesse. Me kasutame transformatsioonipõhist andmete integreerimist ja kirjeldame uudset heterogeenset graafil põhinevat andmestikku Alzheimeri tõve jaoks (HENA). HENA koosneb 64 andmestikust ning kuuest eri andmetüübist, mis pärinevad üheksast eksperimendist. Sinna kuuluvad andmed valk-valk interaktsioonide, geeni koekspressiooni, epistaasi, positiivse selektsiooni, aju eri regioonide geenikespressiooni kohta, kui ka ülegenoomsetest assotsiatsiooni uuringutest (GWAS) pärinevad andmed. HENA võimaldab teadlastel kasutada neid andmeid nii Alzheimeri tõve uurimiseks kui ka oma mudelite testimiseks. Me ka näitame, kuidas kasutada HENA-t, et leida haigusega seotud geene. Viiendas peatükis (põhineb publikatsioon II-l), me näitame, kuidas läbi mitmeetapilise andmeintegratsiooni saab uurida immuunhaiguste mehhanisme. Uuritavaks haiguseks on psoriaas, mis on krooniline põletikuline haigus ning mille tunnuseks on lööbeelemendid kehal. Haiguse mehhanismid pole veel selged. Et saada haigusest paremat selgust, me kombineerisime patsiendi verest ja nahast pärinevaid laboratoorseid andmeid kliinilise infoga. Seejärel viisime individuaalsete analüüside tulemused kokku kasutades lähtuvalt erialaspetsiifilistest teadmistest. Kuues paetükk (mis põhineb publikatsioonidel III ja IV), on pühendatud toksilisuse testimise strateegiate edasiarendusele. Seal me kasutame samuti mitmetasemelist andmete integratsiooni. Toksilisuse testimine tähistab protsessi, mille käigus hinnatakse, kas uuritavatel kemikaalidel esineb kahjulikke toimed organismile. See on vajalik näiteks ravimite ohutuse hindamisel. Selles töös me identifitseerisime sarnase toimemehhanismiga toksiliste ühendite rühmad. Lisaks me arendasime ka klassifikatsioonimeetodi, mis võimaldab hinnata uute ühendite toksilisust.

Autori panused artiklite I-IV põhjal on järgmised:

1. Alzheimeri tõve, toksikoloogia ja immunoloogia valdkonnas andmestike kogumine ja integreerimine (I-IV).
2. Integratsioonipõhise andmeanalüüsi läbiviimine nii Alzheimeri tõve, toksikoloogia kui ka immunoloogia valdkonnas (I-IV).
3. Graafil põhinevate konvolutsiooniliste tehisnärvivõrkude rakendamine suurte heterogeensete bioloogilisi andmeid sisaldavate graafide peal(I).

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name:	Elena Sügis
Birth:	December 29th 1984 Pärnu, Estonia
Languages:	Russian, Estonian, English, German, Spanish
Contact:	+372 56 96 25 36 elena.sugis@gmail.com

Education

2010 - 2019	University of Tartu, PhD candidate in Computer Science
2006 - 2008	Saint-Petersburg State University of Telecommunications, MSc in Telecommunications (*cum laude; practical value is noted*)
2002 - 2006	Saint-Petersburg State University of Telecommunications, BSc in Telecommunications (*cum laude*)

Employment

2014 - ...	University of Tartu, Institute of Computer Science, junior research fellow in bioinformatics
2018 - ...	Quretec ltd., business development manager and bioinformatics project manager
2016 - ...	Software&Data Carpentry, software and data carpentry instructor
2013 – 2018	Quretec ltd., researcher
2011 – 2012	University of Tartu, Institute of Computer Science, programmer
2009 – 2010	MegaFon OJSC, engineer
2008 – 2009	Russian Institute of Radionavigation and Time, design engineer
2006 – 2008	Svetlana JSC, design engineer
2004 – 2005	Leningrad Radio Research and Development Institute, design engineer

Scientific work

Main fields of interest:
Data integration, bioinformatics, data science, machine learning, graph convolutional networks, biological network analysis, toxicology, immunology, neurodegenerative disorders.

ELULOOKIRJELDUS

Isikuandmed

Nimi:	Elena Sügis
Sünd:	29. detsember 1984
	Pärnu, Eesti
Keeled:	vene, eesti, inglise, saksa, hispaania
Kontakt andmed:	+372 56 96 25 36
	elena.sugis@gmail.com

Haridus

2010–2019	Tartu Ülikool, informaatika doktorant
2006–2008	Peterburi Telekommunikatsiooni Riiklik Ülikool, MSc telekommunikatsioonis (*cum laude*, praktiline väärtus on tunnustatud)
2002–2006	Peterburi Telekommunikatsiooni Riiklik Ülikool, BSc telekommunikatsioonis (*cum laude*)

Teenistuskäik

2014 - ...	Tartu Ülikool, arvutiteaduse instituut, bioinformaatika nooremteadur
2018 - ...	Quretec OÜ, kliiniliste uuringute osakond, äriarendaja ja bioinformaatika projektijuht
2016 - ...	Software&Data Carpentry, tarkvara ja andmete analüüsi instruktor
2013 - 2018	Quretec OÜ, bioinformaatika, teadur
2011 - 2012	Tartu Ülikool, arvutiteaduse instituut, programmeerija
2009 - 2010	MegaFon OJSC, insener
2008 - 2009	Raadionavigatsiooni ja Aja Instituut (RIRT), insener-projekteerija
2006 - 2008	Svetlana JSC, insener-projekteerija
2004 - 2005	Leningradi Raadio Teadus- ja Arengu Instituut, insener-projekteerija

Teadustegevus

Peamised uurimisvaldkonnad:
Andmete integratsioon, bioinformaatika, andmeteadus, masinõpe, graafi konvolutsioonilised tehisnärvivõrgud, bioloogilise võrgustiku analüüs, toksikoloogia, immunoloogia, neurodegeneratiivsed haigused.

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.