UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Joanna Niklus

Current State-of-the-Art Bioinformatics Methods in Alzheimer's Disease Studies

Bachelor's Thesis (9 ECTS)

Supervisor: Hedi Peterson, PhD

Tartu 2017

Current State-of-the-Art Bioinformatics Methods in Alzheimer's Disease Studies

Abstract:

Alzheimer's disease is the most common form of dementia, mostly affecting people over the age of 65, world-wide. The studies have focused on finding the reasons for onset and possible cure.

The methods addressed in this thesis are mainly based on the microarray gene expression data. The data is analysed for differentially expressed genes and these are further analysed using visualisation or enrichment analyses.

This thesis hopes to provide an overview of the bioinformatical methods, used in the research of Alzheimer's disease. Resulting in a diverse list on bioinformatical methods, the analysis provides short descriptions and examples of the most used approaches among a chosen subset of articles related to Alzheimer's disease studies.

Keywords: alzheimer disease, data integration, bioinformatics methods

CERCS: B110, P170

Kaasaegsed bioinformaatika meetodid Alzheimeri tõve uuringutes

Lühikokkuvõte:

Alzheimeri tõbi on kõige levinum dementsuse vorm ning see esineb ülemaailmselt vanematel inimestel. Uuringud keskenduvad põhjuste ja ravi leidmisele.

Käsitletavad meetodid põhinevad geenieksпressiooni andmetel. Erinevalt avalduvad geenid eraldatakse ning kasutatakse edasistes analüüsides.

Käesolev bakalaureusetöö pakub ülevaadet Alzheimeri tõve uuringutes kasutatavatest bioinformaatilistest meetoditest. Tuleneval mitmekülgsete meetodite hulgal põhinev analüüs kirjeldab lähenemisi lühidalt ning toob välja näiteid valitud artiklite hulgast.

Võtmesõnad: alzheimeri tõbi, andmete integreerimine, bioinformaatika

CERCS: B110, P170

Contents

1 Introduction

Alzheimer's disease affects the brain, disrupting memory, thinking and judgement and causing problems in everyday life. This disease has been studied for over 100 years, however no cure or onset cause has been found. Usually, people over the age of 65 are affected, however there are cases, where people are affected, as early as 50. This earlier onset is predictable, when the person has certain gene traits. This type is also inheritable, affecting family members from different generations. Because of this type, the causes are thought to be genetic, however no certain knowledge of the disease mechanisms, is known. This complex disease is affecting people all over the world, and the number of cases is growing each year, as well as the costs related to caregiving and hospitalisation, for example. As the population is ageing, the patient numbers and costs will continue to grow, if no cure or method of prevention is found.

The methods involved in diagnosing Alzheimer's disease include mental fortitude and memory impairment tests, cerebrospinal fluid analysis and post-mortem brain tissue examination. However, the first signs of dementia may appear up to a decade later, when the disease has already progressed in the brain. The research related to better diagnosing techniques as well as better understanding of the onset and progression, are important. The methods used in this research involve differential gene expression, brain imaging scans and searching for possible drug targets. As the disease is thought to have genetic causes, the gene expression studies can shed light on why and how the disease progresses.

Bioinformatics provides many methods for analysing gene expression and the relevant pathways contributing to the progression of the disease, and discovering drug targets, for example. These methods are mainly based on tissue samples isolated from post mortem

brains, limiting the study of differential gene expression in earlier stages. Gene expression studies use microchips, enabling the analysis of multiple samples and providing more information in one reaction. These data are then analysed using computer algorithms and statistical methods to extract the genes that behave differently in diseased brains than in healthy brains. Numerous other studies conducted on these possible genes offer more insight and might provide targets for drugs.

There are problems linked to the availability of data from earlier stages, but also to how the studies are conducted. Not every method is compatible with certain hypotheses proposed during research. Furthermore, the submitted studies need to have certain quality of data presentation and descriptions of the conducted analyses and used methods. There are many new methods proposed and many new possibly related genes offered, however, the new methods need validation to be confidently used and the genes need further studies. Moreover, the novel genes offered, have little overlap between them.

2 Background

2.1 Biology

2.1.1 The central dogma of molecular biology

This short introduction into the basics of genetics is based on the writings of A. Lesk and A. Heinaru, except in cases otherwise cited [1, 2]. The Central Dogma of Molecular biology states that genetic information transmits from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) and from there to protein [Figure 1]. It does not transmit from protein to nucleic acid [3]. The DNA is the carrier of genetic information in cells. It

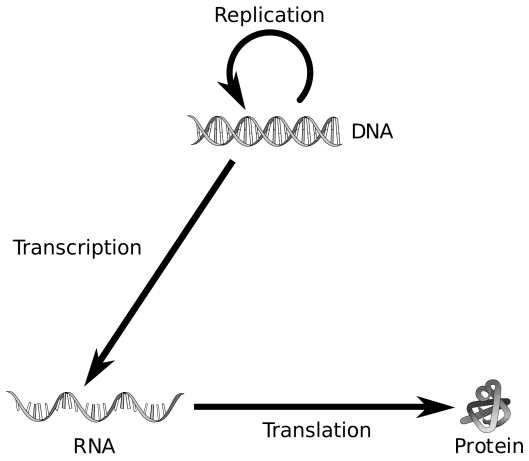

Figure 1. The central dogma of molecular biology. Genetic information transmits from nucleic acid to nucleic acid and from nucleic acid to protein. It does not transmit from protein to nucleic acid, nor from protein to protein. The solid lines show the usual transmission. By Philippe Hupé, via Wikimedia Commons; modified [4]

is a sequence based on the four-letter alphabet of adenine (A), guanine (G), cytosine (C) and thymine (T). These four nucleotides make up complementary pairs: A with T, and G with C. When pairing together they form hydrogen bonds: A forms two hydrogen bonds with T and G forms three hydrogen bonds with C. These four nucleotides connect

with one-another to make up a long, linear strand that has an anti-parallel (oriented in the opposite direction) and complementary sequence, forming the double helix. Each strand has a direction, named after the positions of the open ends, i.e. the 5' end and the 3' end. The sequence is read from 5' end to 3' end, and because of the differences of the ends and the specificity of the regulatory or replicatory proteins, the synthesising proteins move from 5' end toward 3' end.

Nucleotides form sequences, known as genes. Either strand of DNA can contain genes

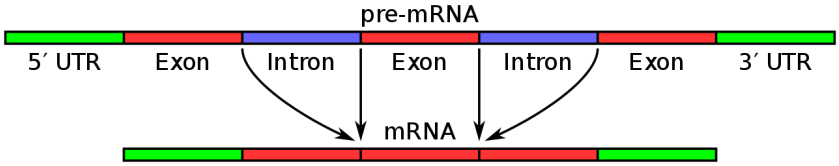

Figure 2. Splicing of the transcribed pre-mRNA. The pre-mRNA is modified by excising the introns and splicing together the exons. The mRNA is then transcribed into a protein. There are sequences in the ends of the strands that remain untranscribed (UTR). By Qef, via Wikimedia Commons [5]

and in eukaryotes one gene is often split into segments along one strand. Genes contain intervening regions called introns between regions that are expressed, i.e. exons [Figure 2]. As the strands are directed, the regions toward the 5' end from the gene are called upstream regions and similarly, regions towards the 3' end are called downstream regions.

The expression of genes is controlled by internal mechanisms that may turn the genes on or off. One of these mechanisms is regulatory genes, which can be found upstream from the gene they regulate. The regulation can be repressing, meaning the regulatory gene's product binds so that its target gene can no longer be expressed, therefore lessening the target gene's product. The regulation can also be activating, in which case the transcription is promoted.

RNA has a slightly different alphabet than that of DNA - instead of thymine, RNA has uracil (U), which also forms two hydrogen bonds with A. Moreover RNA is single-stranded, that is often folded, forming complementary structures with itself. After the transcription of the whole gene (introns and exons) the pre-messenger RNA (pre-mRNA) is synthesised and undergoes splicing, during which the introns are excised and the exons are spliced together, forming a strand of mRNA [Fig.2]. mRNA further undergoes translation, forming a protein. Proteins are strands of amino acids, determined by the sequence of the gene (mRNA). One amino acid corresponds to a three-nucleotide group, called a codon, making the number of possibilities for different sets of nucleotides into 64, which represent 20 standard amino acids. Among these 64 codons are 3 stop codons - a sequence such that, when the translator molecule encounters this, it stops the translation. However, there is only one sequence from which the translation starts.

Because of the intra-cellular influences (e.g. pH level) and molecular interactions, DNA, RNA and proteins have different structure. In the cell, the DNA has the form of linear double-stranded helix, densely and orderly packed into multiple chromosomes (the number is dependent on species). The single-stranded RNA has three main functions: mRNA, tRNA and rRNA. mRNA is what is used to translate specific proteins; transfer RNA (tRNA) transports amino acids to ribosomes during translation; and ribosomal RNA (rRNA) makes up the ribosome, in which the translation takes place. Each of these RNA types has a different structure which correlates with its purpose. The 3D native state of proteins is determined by the amino acid sequence, and the native state is what determines the function of the protein.

The biochemical functions of proteins are vast, they can be structural (e.g. membranes of organelles), catalytic (e.g. enzymes), regulatory (e.g. hormones) or control gene

transcription. The native state of proteins requires certain conditions to be met, e.g. the pH level and temperature (such as in the cell) in order for the proteins to fold spontaneously into the respective native state. This state can, however, unfold into a disordered and functionally passive structure. This process, called denaturation, happens when the conditions in which the protein is active, change. In very few cases the protein's structure recovers to its native state, when the normal conditions are restored. However, in irreversible cases of denaturation, the protein does not recover its natural state causing the aggregation of insoluble inactive proteins. These aggregates are linked to many diseases, one of such being AD, which is further discussed below.

2.1.2 Brain

Brain is the commanding organ responsible for tasks such as controlling the internal synchrony of functions. The brain is separated into two hemispheres and divided into several regions in regard to the respective functions. For example, the hippocampus and amygdala are associated with memory. The main cells native to the brain are neurons, which transmit the information among brain regions [6].

The structure of neurons is characterised by their central cell body called soma, input strands called dendrites and an output strand called axon [Fig.3]. The axon connects to other neurons' dendrite(s) and the connecting part is called a synapse [8]. An important feature of axons, is the myelin sheath that coats them [6]. It provides the rapidness of information transfer as well as functional insulation [6].

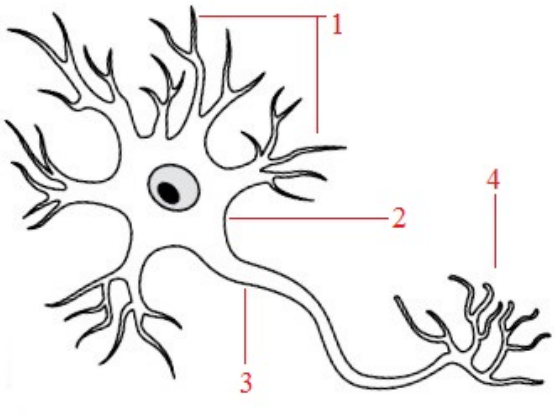

Figure 3. A schematic illustration of a neuron. 1) The dendrites through which the neuron receives the information; 2) The cell body, a.k.a. soma; 3) The axon, coated with the myelin sheath and used for transporting the signals out; 4) The axon terminals via which the neuron connects to other neurons' dendrite(s), for example. Via Wikimedia Commons; modified [7]

2.1.3 Alzheimer's disease

The first known description of AD originates from the year 1901, when Alois Alzheimer provided a detailed and extensive documentation of one of his 51 year old female patients, who showed signs of paranoid symptomatology, sleep disorders, memory disturbances, aggressiveness, crying and progressive confusion [9]. After the patient's death, Alzheimer further described the plaques and tangles in the brain [9]. In the later years similar cases were published and some of Alzheimer's earlier work was further specified [9].

After numerous research and investigation, the understanding of the disease has improved, however no cure to stop or reverse the progression has been reported and AD has become one of the most common form of dementia [9, 10]. Moreover, the costs and patient numbers of dementia are growing yearly [11].

The risk factors of AD include old age, environment (e.g. smoking, obesity, hyperten-

sion) and genetics (late and early onset) [12]. Late onset AD (also called sporadic AD) is the most common case, affecting people over the age of 65. Although some genetic risk factors have been identified (e.g. the inheritance of the $\varepsilon 4$ allele of APOE), they do not guarantee the development of the disease [12]. However in the case of familiar AD (early onset AD), in which symptoms develop before the age of 60, mutations in the APP and presenilin (PSEN1, PSEN2) genes attribute to the disease, so that many family members across multiple generations are affected [12]. This type of AD accounts for about 0.1% of the disease cases.

Currently, the presence of AD is confirmed by analysing the cerebrospinal fluid (CSF) of the patients for established biomarkers, including amyloid beta protein, tau protein and phospho-tau protein [13]. However, since CSF is obtained via lumbar punctures, both invasive and painful, there is a need for biomarkers that are more easily obtainable, more sensitive and more specific.

In addition to CSF analysis, mental fortitude and memory impairment tests are also conducted and combined with brain imaging to identify AD. However this does not provide the diagnosis with 100% certainty. The most definitive diagnosis remains the post-mortem brain tissue analysis [12].

AD progression has been divided into three stages: early stage, middle stage and late stage. As first symptoms of AD may manifest more than a decade later, when the disease has already spread in the brain, the early stage is often undetected [14]. Also, the time intervals of these stages differ for individuals and depend on multiple factors, for example hypertension, depressive symptomatology or chronic psychological stress [12, 14]. AD affects brain functions such as memory (forgetfulness, unable to recognize faces), thinking, orientation (lost in familiar places), comprehension (unaware of time and date), language (communication) and judgement (depression, aggression), pro-

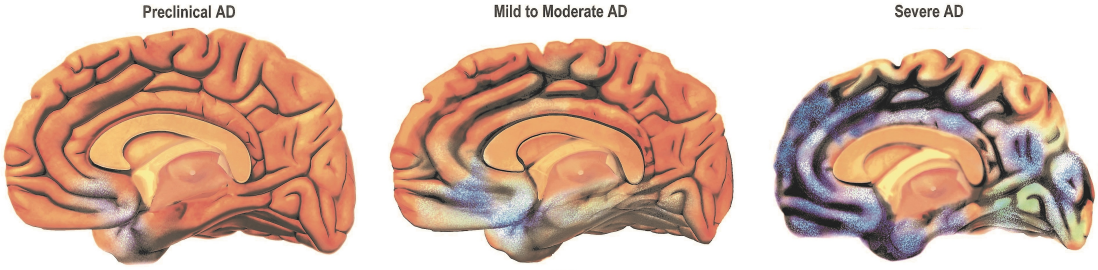

Figure 4. The progression of AD through the brain. The blue part shows the progression and affected regions over the years. As can be seen, the brain with severe AD is showing shrinkage of the cerebral cortex. Figure by National Institute on Aging, National Institutes of Health; modified [15]

gressing over the years [Figure 4]. The disease will eventually influence other physical abilities as far as the patient needing assistance in self-care, incapability of walking, difficulties swallowing, making one unable to eat without assistance, and finally ending with death [12].

AD affects brain regions such as entorhinal cortex, hippocampus, frontal cortex and amygdala and includes pathophysiological symptoms such as abnormal tau proteins forming neurofibrillary tangles [Fig.5], neuronal loss and atrophy, synaptic dysfunction, neurodegeneration and amyloid-β plaques [Fig.6] [14, 16].

In normal brains, tau proteins are involved in the assembly and stability of axonal microtubules [17]. In brains affected by Alzheimer's disease [Fig.5], the tau proteins have become hyperphosphorylated, which results in the progressive disruption of neuronal cytoskeleton and the formation of intracellular neurofibrillary tangles (NFT) [17].

The amyloid beta proteins in normal brains are located on the surface of the neurons and are related to neuronal growth, adhesion, cell mobility and regulation of transcription [17]. Mutations in the APP (Amyloid Precursor Protein) gene result in the different cleavage of the gene [Fig.6], producing neurotoxic amyloid-beta peptides [17]. These

amyloid-beta peptides adhere together and form extracellular amyloid plaques [17].

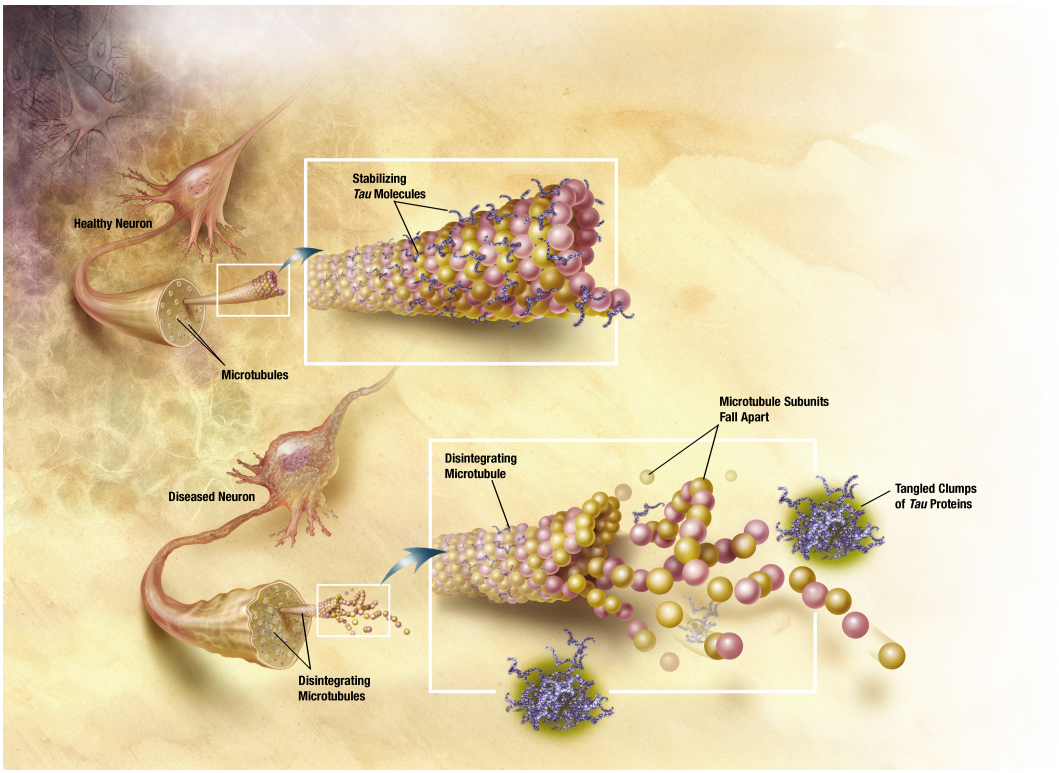

Figure 5. Illustration of normal functions of tau and causes of AD affected tau. The microtubules involved in information transmission between cells, can be seen disintegrating, which causes cells to lose connection and thereby disrupting the brain functions. By ADEAR, via Wikimedia Commons [18]

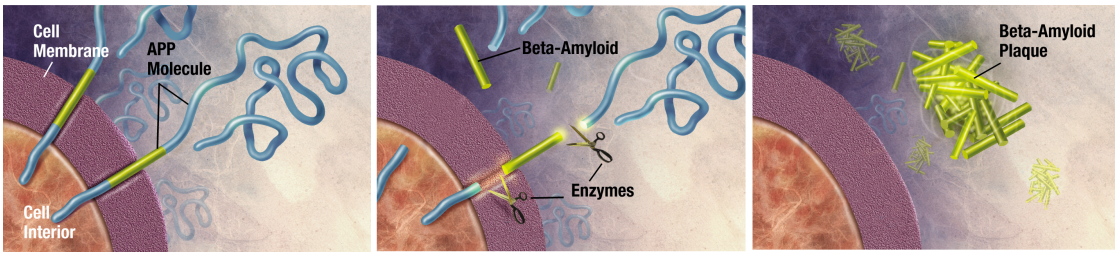

Figure 6. Figure showing the formation of Beta-Amyloid plaque. In healthy cells, APP molecule is differently cleaved and thus producing different products. In diseased brains, however, the differently cleaved products are not soluble and therefore form aggregates, known as Beta-Amyloid Plaques. By ADEAR, via Wikimedia Commons [19]

2.2 Bioinformatics

Bioinformatics is an interdisciplinary field combining mathematics, statistics, computer science and biology. The aim of bioinformatics is to solve biological problems with the help of computational resources and statistical methods.

2.2.1 Microarrays

Microarrays are designed for the simultaneous analysis of millions of sequences in one reaction. The surface has an orderly arrangement of immobilised probes with known sequences [Fig.7]]. The probes, also known as oligonucleotides, have a specific sequence and type (DNA or RNA) according to the specifics of the subsequent experiment [20]. There are three types of probes that are used by most of the designs: perfect match (PM), mismatch (MM) and control probes. The section that one probe occupies is called a feature, PM and MM probe together are called a probe pair. A number of probe pairs are selected per gene, hybridizing in different locations of the gene, thus forming a probe set. The number of probe pairs per gene as well as the length of the probes depends on the manufacturer of the array [20, 21].

Hybridization entails the binding of two complementary strands of different origins. The probe sequences on the array are synthetic, whereas the target sequences come from cells [21]. The more specific properties of microarrays are based on the manufacturer of the array, as well as the ensuing application and experiment type.

PM probes are entirely complementary to its target sequences, while MM probes have, in the centre of the sequence, one nucleotide that is complementary to the respective nucleotide of the PM probe [21]. The one mismatching probe is enough to disrupt the binding of the strands, and therefore provides a negative control for background hybrid-

ization [22]. Capturing the non-specific binding and the background was the theoretical design of the MM probes, determining the reliability of the PM probe [23].

Most commonly used microarrays are manufactured by Affymetrix and Illumina [Figure 7]. The photochemical synthesis directed manufacturing of Affymetrix GeneChip

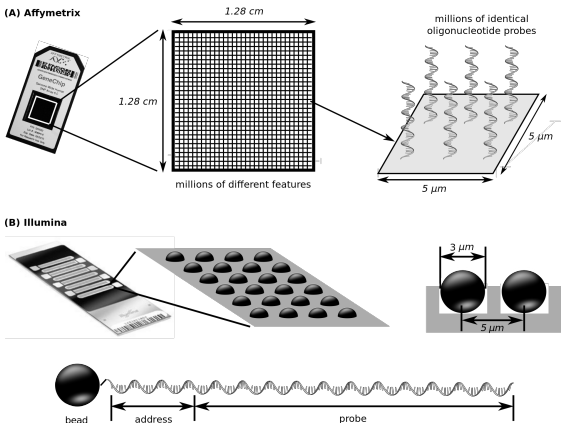

Figure 7. A comparison of Affymetrix and Illumina arrays. By Philippe Hupé, via Wikimedia Commons [24]

arrays is their main distinguishing feature [21]. The oligonucleotides synthesised onto the array are in accordance with the intended use, which may be gene expression profiling, whole-genome transcriptome mapping or custom genotyping [21]. The probes on the array are typically 25 nucleotides long and 11 probe pairs are usually selected per gene. The arrays, 1.28 cm^2 in size, are capable of containing more than 1.4 million different features [21].

Illumina uses silica beads containing around 10^5 copies of the same 75 base pair long oligonucleotide acting as probe. Their highly miniaturized arrays (1.4 mm across) can hold up to 50 000 beads and about 40 000 array elements per square millimetre. The localisation and identification are done post-assembly [25].

As mentioned above, the hybridisation process involves the probes on the array, and sample material from the cells. Based on the RNA isolated from the cells, the cDNA are

synthesised, labelled fluorescently and washed over the array in controlled conditions [26]. Later, the array is scanned, and a composite image is created [26]. In two-colour microarrays, fluorescent labels for healthy and diseased samples differ in colour, enabling the calculation of differential expression [26]. One-colour microarrays feature one type of sample types and thus have one colour.

2.2.2 Data preprocessing and related algorithms

Since raw data includes inconsistencies, noise and missing values, the data has to be cleaned and levelled. Data inconsistencies can be caused, for example, by hairs, scratches or precipitation on the array, the scanner (less reliable below a certain intensity) or the image analysis software (peculiarities in calculations) [26]. As missing values and outliers (a value far away from other values) interfere with statistical tests and clustering, data cleaning and levelling is important for more meaningful results [26].

To choose a suitable method, amongst the pool of different and growing number of algorithms, one has to consider the type of the array used as well as the objective of the experiment.The following paragraphs introduce the general idea of data preprocessing and some of the most used methods are further described.

The conventional first step in preprocessing is correcting the data for background noise, and effects of individuals and batches. Individual effects include for example the age and sex of the person, the interval between death and collection of tissue, brain pH level, and cause of death. Batch effects include for example the time of hybridisation, used chemicals, humidity, and the scientist doing the experiment. It has been shown to be the main step of data preprocessing, as well as one with the biggest impact on the performance [27].

Next step is the normalization or the levelling of the data, i.e. median-centering the

value distributions across samples in the dataset, ensuring the better comparability of the data [28]. In the end of this step, the data from different datasets are comparable.
The last step is usually the summarization of the data, reducing the volume, but directly or implicitly preserving the outcomes of the analysis. Probe set values are respectively combined, forming a measure of expression for the genes [21].
Some of the most used preprocessing algorithms for Affymetrix arrays in relation to the AD studies include Robust multi-array average (RMA) [29], GC-corrected RMA (GCRMA) [30] and MicroArray Suite 5.0 (MAS5.0) developed by Affymetrix [31]. RMA uses global background adjustment and across-array normalization [29]. For gene expression measurement it uses log-transformed PM values [29]. For the combination of intensity values RMA uses median polish and fold change in the differential expression detection [29].
However, RMA does not use MM probes for calculations as MAS5.0 does. The Affymetrix algorithm subtracts MM probe values from the values of PM probes in the background correction step [32]. For the combination of intensity values, MAS5.0 uses the Tukey biweight estimator [33].
GCRMA is a variant of RMA with the same normalisation and summarisation steps, however it uses an estimate of non specific binding, taking advantage of sequence information for background correction [34].

2.2.3 Differential expression analysis

Expression change can be conveyed by three measures: intensity ratio, log ratio and fold change. The log ratio and fold change are derived from the intensity ratio, which represents the raw value for expression [26].
To calculate the intensity ratio for two-colour data, intensities of sample and control are

divided. This produces an asymmetrical distribution for the values, as down-regulated genes are represented in the range (0...1) and up-regulated genes belong in the range of (1...) [26].

The log ratio is calculated as a log-transformation of the intensity values. The base of the logarithm can be freely chosen (most commonly used is base 2), however the base has to be the same for all of the samples [26]. The log-transformation makes the value distribution more symmetric, as both up- and down-regulated genes fall in the range of 0 to infinity [26].

The calculation of fold change likewise changes the distribution to be more symmetrical and the of values range from 1 to infinity [26]. Fold change values stay the same as the intensity ratio in cases when the expression is higher than one, and in cases where the expression is lower than one, the fold change takes the value of the inverse intensity ratio [26].

2.2.4 Data visualization

An important step in data analysis, to better comprehend the information produced, is data visualisation [26]. For example, the construction of scatter plots can be very informative for initial evaluations of data, as well as comparing the data sets [26]. Further comprehension can be achieved by clustering the data, that is dividing the genes into a number of groups with similar expression patterns [26]. This reduces the dimensionality of the data, as well as makes the data more intuitive for the user [26]. Furthermore, when visualising the clustered data as a heat map, for example, a visual observation can determine, whether the control groups are clustered together or not, providing further information about the quality of methods used thus far in the analyses.

There are a number of ways to visualise different information, however some methods

work better with certain kind of data. For example, heatmaps are a better way to visualise the intensities of differential gene expression values than networks, for instance. From amongst the many methods, networks, heatmaps, scatterplots and boxplots, are introduced below.

Networks have the form of graphs and convey information about interactions. When thinking of it mathematically, it is a set of nodes or vertices V and a set of edges E, where vertices are connected via edges and an edge is specified by the vertices it connects [1]. When an edge has a value associated with it, the graph is labelled, e.g. the labels can stand for correlation coefficient [1]. In a directed graph, the vertices are ordered, e.g. the first vertex shows a regulator and the second vertex shows the gene it influences.

Biological networks include for example protein-protein interaction (PPI) networks, regulatory networks (containing gene expression control information); networks conveying the signal transmission inside, outside or within the cell; and metabolic networks, modelling the metabolism in organisms [35]. In the scope of this thesis the focus is mainly on PPI and regulatory networks. PPI networks are undirected and contain information about the interactions between different proteins. Regulatory networks are directed and contain the information of regulations, i.e. which genes are affected by which regulators.

Even though the biological networks contain various different information, there are certain patterns - network motifs - which are common to all the aforementioned, and these motifs are found to perform specific functions relative to the motif structure [36]. Uri Alon introduces in his review four network motif families, including 1) simple regulation; 2) feedforward loop (FFL); 3) single-input module (SIM); and 4) dense overlapping regulon (DOR) [36]. These network motifs, shown in Figure 8, have been found for

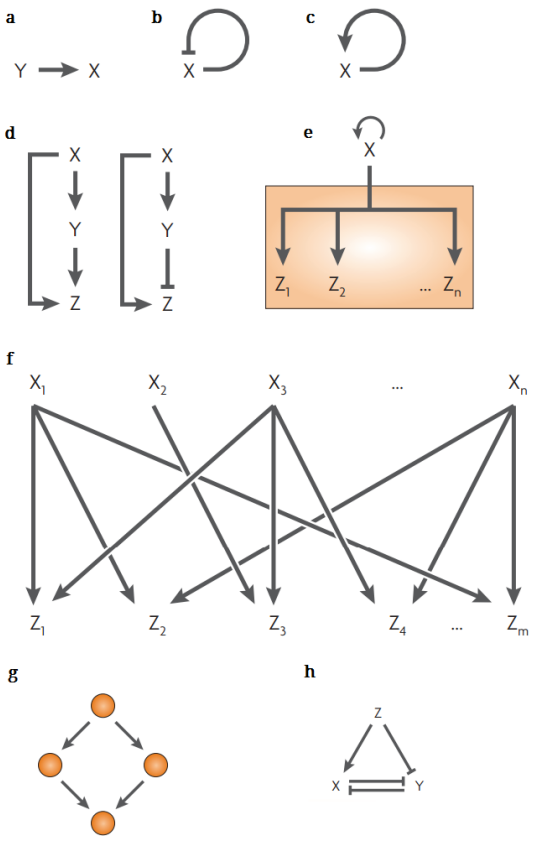

Figure 8. Four families of network motifs and their examples. **A**: Simple regulation. X regulates Y without any additional interactions. **B**: Negative autoregulation. X represses its own transcription. **C**: Positive autoregulation. X enhances its own production. **D**: Feedforward loop. X influences Y, which in turn influences Z, which is also influenced by X. These influences can be repressing or enhancing, thus there are eight possible feedforward loop structures, of which the most frequent two are shown. **E**: Single input module. X regulates a group of targets Y_n. **F**: Dense overlapping regulon. A set of regulators X_n regulate a group of targets Y_m. **G**: Diamond pattern. **H**: Regulated feedback with a double-negative-feedback loop. From the article by Uri Alon [36]

example in sensory networks, responding to stress and nutrient signals, developmental networks guiding the differentiations, protein modification networks and neuronal networks, suggesting the structural simplicity of complex biological networks Figure 8 [36].

When analysing biological networks, there are three important features in common with

non-random networks, including scale free and small-world properties and modularity [Fig.9] [37]. Scale free networks include a small number of highly connected nodes

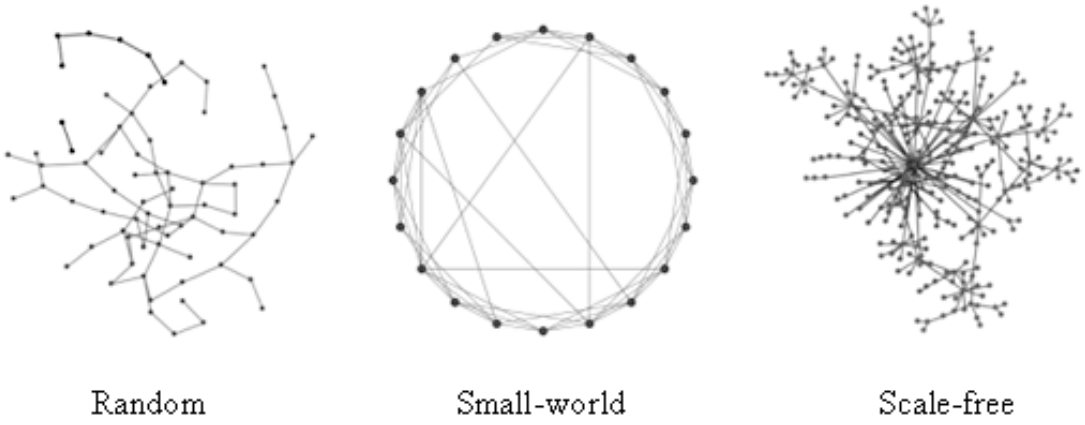

Figure 9. Examples of networks. It can be seen that the small-world network seems to have direct or indirect connections between all the nodes. The existence of network hubs is identifiable in the case of the scale-free network. By Utopiawiki (Own work), via Wikimedia Commons [38]

called 'hubs' and a lot of less connected nodes [37]. The scale free networks are remarkably resistant to accidental attacks (e.g. mutations), however when an attack is coordinated (e.g. a pathogen) on a hub-node, the network is very vulnerable [39]. Small-world networks have the tendency of having a shorter path between any two nodes than that expected in a random network of similar size and having a similar number of connections [37]. Finally the modularity refers conceptually to gene groups performing similar functions separable from the rest of the system, which means, they are in contrast to motifs, which can not be separated from the rest of the system [37].

Heatmaps can be thought of as a coloured matrix, with the rows representing genes, columns representing patients, for example, and the cells contain a colour corresponding to the value of expression [40]. However, without a certain ordering of the rows and columns, the interpretation of the matrix proves to be complicated, as the rows tend to

be in a random order, as can be seen from Figure 10 [40]. In order to retrieve mean-

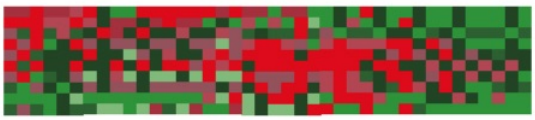

Figure 10. Random gene expression data. It can be difficult to perceive patterns from the randomly ordered rows and columns. Figure by MIT OpenCourseWare [41].

ingful data from gene expression heatmaps, the rows and columns are reordered with clustering algorithms [40]. One of the common algorithms is hierarchical clustering, which also provides a dendrogram to show the division levels of the values [40]. Clustering is based on the concept that gene expressions, arising from the similar functions and regulation, group the genes into clusters, an example of this is provided as Figure 11 [26].

To perform clustering, one would create either a distance matrix or a similarity matrix,

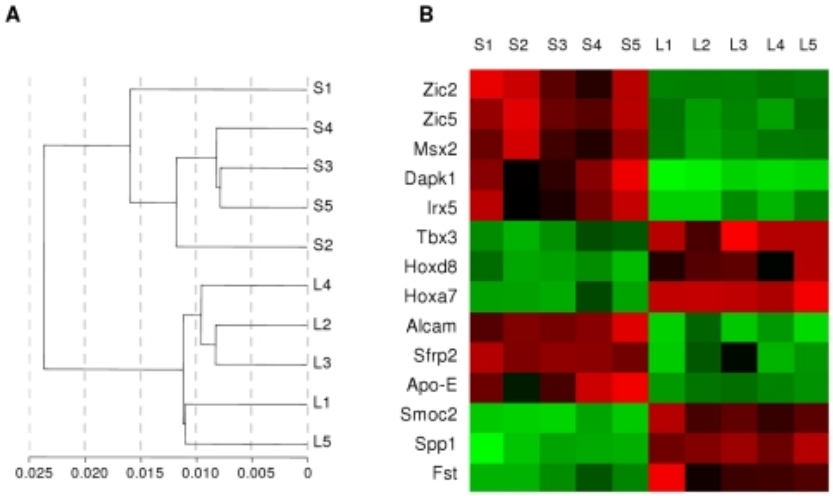

Figure 11. **A**: The dendrogram (hierarchical tree) showing the similarity divisions. **B**: The heatmap with clear clusters, depicting two-way clustering, i.e. the clustering of both rows and columns. The meaning of the genes is irrelevant in the scope of this illustration. By Rawlinson S, McKay I, Ghuman M, Wellmann C, Ryan P, Prajaneh S, Zaman G, Hughes F, Kingsmill V, via Wikimedia Commons [42]

using a distance metric [43]. These metrics can be, for example, Eucleidean distance or Pearson correlation coefficient [43]. The algorithm works by finding the closest clusters, merging them together, calculating a distance (or similarity) value for the new cluster and repeating these steps until all genes are clustered [26]. This can be done 'bottom up' or 'top down', that is the algorithm would start with all the nodes as single clusters and iteratively add the closest (most similar) nodes to the cluster or it would begin with one cluster containing all the nodes and iteratively remove the furthest (least similar) nodes.

Another commonly used clustering method is the k-means clustering, for which the number of groups (k) has usually been provided by the user [26]. The resulting clusters do not have a hierarchical structure, but they are geometrically very compact and close to the respective centroids [26].

Scatterplots are useful for the pairwise comparison of datasets in order to find disproportionately expressed genes [40]. A special case of scatterplot is the volcano plot, which has been used to visualise the fold changes of the data [Fig.12].

Boxplots are used for visualising the robust summary of a dataset's distribution. Figure

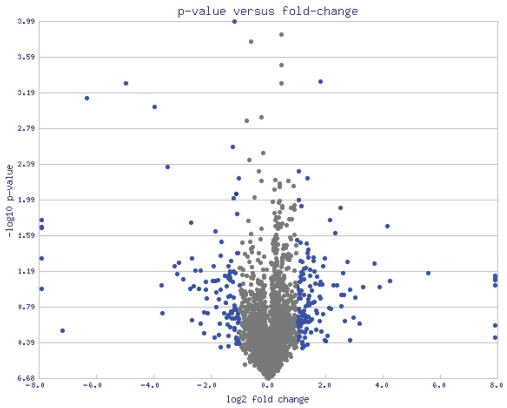

Figure 12. An illustration of a volcano plot. By Roadnottaken (Own work), via Wikimedia Commons; modified [44]

13 shows the elements of a boxplot. The whiskers are usually used with a multiplier of 1.5 [45].

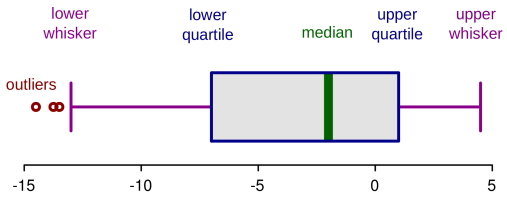

Figure 13. An illustration of the elements of a boxplot. By Ruediger85 (changed language). Original by RobSeb (Own work), via Wikimedia Commons [46]

2.2.5 Databases

The amount of biological data is growing fast and thus there is a need for good databases that not only store the data, but are also regularly updated and have a consistent standard for the presentation (as well as storage) of the data. Microarray data is only a small fraction of biological data, it also includes for example, sequences, interactions, functions and disease information. Thus, in order to regulate the structure of the data to be uploaded, minimum requirements ensuring the easy interpretation and verification of the data were proposed as the MIAME standards [47]. There are several databases following these guidelines, including, but not limited to, Gene Expression Omnibus (GEO) and ArrayExpress [48].

GEO at the NCBI and ArrayExpress share a similar purpose, they are both public repositories of high-throughput data and are both recommended by scientific journals [49, 50]. Researchers can submit for example, microarray and sequencing data, as well as query, review and download the data [50, 49]. ArrayExpress imports GEO datasets on a weekly basis [50].

The National Center for Biotechnology Information (NCBI) at the National Institutes of

Health (NIH) accommodates many databases relevant to biotechnology and bioinformatics, including GEO and PubMed for example [51]. NCBI has integrated the Entrez database retrieval system to provide access to many databases containing for example nucleotide sequence data and genomic mapping information [51].

The Gene Ontology (GO) provides annotation of genes, meaning the terms describing the gene product are assigned to corresponding genes. GO terms are divided into three main categories: molecular function, biological process and cellular component, that are in themselves three independent ontologies [52]. The terms are hierarchically organized, a node may have more than one parent and each node in GO ontologies is linked to additional other information, e.g. the SwissPROT or GenBank databases. The interlinking of databases tries to eliminate dated facts, because biological knowledge is improving daily [52].

Kyoto Encyclopedia of Genes and Genomes (KEGG) is similar to GO, in that it provides functional meaning to genes and genomes, assisting in interpreting biological datasets. KEGG is essentially a collection of databases, containing information about drugs, diseases, genes, cellular functions and much more, as well as offering tools for data analysis and query [53]. KEGG can be used for the identification of enriched gene pathways.

Similar to this is the Reactome peer-reviewed database [54]. Expert biologist, in collaboration with others, for example, GO (vocabularies), PubMed (research literature) and Entrez (genes), create the annotations for the pathways [54]. The website also provides tools for various activities, for example over-representation analyses, pathway analyses and visualisation of full pathway information [54].

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database includes information about protein-protein interactions, both known and predicted, and

covers more than 2 000 organisms [55]. The comprehensive interactions contained in STRING are derived from multiple databases, full-text articles and analyses, and scored with a confidence value [55].

2.2.6 Software

As a system for statistical graphics and computation, the open source programming language R can be obtained with a graphical run-time environment. The free, open source and open development Bioconductor project relying upon R, contains packages for analysing and understanding genomic and sequence data, amongst others. These additional packages can be obtained as R packages [56].

An open source software platform used for biological research is Cytoscape, which was originally designed only for this purpose. However, Cytoscape has since matured and expanded into being a general platform for complex network visualization and analysis in several domains [57]. For example, Cytoscape provides data file import (e.g. from GO), connections to public databases (e.g. NCBI Entrez Gene), analyses and visualisations in JavaScript environment via Cytoscape.js and several plugins for additional features [58, 59].

3 Analysis

I selected 15 articles found via PubMed and Google Scholar using the terms 'bioinformatics', 'alzheimer' and restricting the year to be since 2010. Among the articles I have used, 13 are from the years 2015 and 2016, one from 2010 and one from 2013. The following analysis is based on what I read and summarized from these articles.

3.1 The goals and objectives

The goals of the chosen articles were, in general, to identify novel genes; to further explore and expand the knowledge of less known previous findings; or to use novel approaches or previously unused methods in relation to AD studies. Ray and Zhang had the goal of identifying genes with differential topology in the pairwise regional coexpression networks of entorhinal cortex, hippocampus, middle temporal gyrus and posterior cingulate cortex [60]. The goals of Jamal and colleagues were to identify genes with potential relation to AD [61]. Zhang and others wanted to improve the knowledge of the molecular pathogenesis of AD by identifying AD related genes, sub-networks and pathways [62]. The purpose of the study by Puthiyedth and colleagues was to provide new knowledge about the regional specificity of AD by identifying significant genes in regions associated with AD [63].

The study by Yue and colleagues aimed to provide a new gene interaction analysis tool, which would provid higher credibility [64]. It was hypothesised that the understanding of the underlying mechanisms of AD could be improved by identifying DEGs from amongst data downloaded from GEO [65]. The study by L. Zhao and colleagues intended to determine the differences between the aging of male and female mouse brains in regions significantly affected by AD [66]. The objectives of Forabosco and col-

leagues were to analyse TREM2 in the networks of different brain regions and data sets to provide better understanding of TREM2 connections with known genes involved in AD [67]. Wang et al. studied nineteen brain regions for expanded descriptions of AD-associated molecular networks [68]. In the article by Y. Zhao and colleagues, the aim was to uncover the potential roles of miRNAs in AD pathogenesis and expand knowledge related to AD mechanisms [69].

Acquaah-Mensah and Taylor showed in their article the usefulness of in-situ hybridisation (ISH) data in differential gene expression studies [70]. The paper by Martinez-Ballesteros et al. aimed at providing gene expression patterns and deeper knowledge acquired by and integrated method consisting of descision trees, quantitative rules and hierarchical cluster analysis [71]. The goal of Song and colleagues was to evaluate the extent of AD-related gene identification of the NetWAS approach [72]. Nevado-Holgado and Lovestone proposed the hypothesis that the effect of Non-Steroidal Anti-Inflammatory Drugs (NSAIDs) on the risk of AD might be through previously unrecognised effects in addition to the known inflammation-suppressing effects [73]. The study by Hao and Friedman aimed at providing additional information for the effects of AD drugs, which had failed the clinical trials or were in clinical trials [74].

3.2 The data and methods used

3.2.1 Data

To better understand the disease and its mechanisms, it is important to know which genes are involved and how they are expressed compared to genes in normal, healthy brains. Thus the gene expressions of both healthy and diseased brains are needed. Out of all the 14 brain regions the datasets represented, hippocampus, entorhinal cortex and

posterior cingulate cortex were the most studied, with hippcampus studied 12 times, entorhinal cortex 5 times and posterior cingulate cortex 3 times.

When collecting cell samples from brain section slabs, the most accurate and efficient way is by laser capture microdissection (LCM). This method captures only specific cells from a cell population, while other methods may also gather the surrounding cells [75]. Capturing only one cell, makes the gene expression data more accurate, because in a tissue scrape, i.e. a population of cells as opposed to one cell, the majority of the cells might not be diseased, thus making the diseased cell's expression data suppressed [75]. One of the most mentioned across the selection of articles and also studied cell type is the pyramidal neuron. These cells occur in the forebrain structures (e.g. cerebral cortex and hippocampus) and their abundance in cortical structures, as well as their features suggest high involvement in cognitive processing [76].

Regarding the species mostly studied, humans were in the majority. The other species under investigation was the mouse, as lab-bred and as a data set from the Allen Brain Atlas (ABA) database. Breeding mice in controlled conditions, enabled Zhao and colleagues to use earlier stages of AD, a variety of ages and to control the distribution of gender [66]. ABA data was retrieved as mouse genome-wide in-situ hybridisation (ISH) image data with cellular level resolution for the article by G.K. Acquaah-Mensah and R.C. Taylor [70]. The high-throughput ISH method provides gene expression data across the brain that is tissue- or cell-type specific, the specifics of the technique are provided by Eichele and Diez-Roux [77]. Acquaah-Mensah and Taylor used, for their analyses, a set of genes from the mouse hippocampus that consisted of genes associated with inflammation, apoptosis and response to oxidative stress [70].

The articles by Acquaah-Mensah and Taylor, Yue et al., Ray and Lovestone, L. Zhang et al. and Puthiyedth all used the same dataset containing the gene expression data from

hippocampus, entorhinal cortex, middle temporal gyrus, posterior cingulate cortex, superior frontal gyrus and visual cortex [70, 64, 60, 62, 63]. Yue et al. and L. Zhang both used two datasets containing the hippocampal gene expression data [64, 62].
The microarrays used in the datasets mostly featured Affymetrix GeneChip arrays. Illumina arrays were in the minority. Other data production methods included quantitative Real-Time Polymerase Chain Reaction (qRT-PCR) and brain imaging methods, such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) scans.

3.2.2 Preprocessing

The most used preprocessing methods include the Robust Multi-array Average (RMA), RMA with GC background correction (GCRMA) and Microarray Analysis Suite 5.0 (MAS5.0). The uses of these methods varied, for example the RMA method was the most occurring and was most used for background corrections [64, 65, 62, 70]. Among other preprocessing phases, RMA was also used for the normalisation step [65, 68].
The GCRMA method is an alteration of the RMA method, its only difference is the background correction approach and was used once for this step [67]. Normalisation was the second step, in which this method was used [67, 69].
The MAS5.0 method was the only method used for the revision of PM-MM values, which is to be expected, because none of the other two methods use MM probes for their calculations [64, 62].

3.2.3 Analyses and methods

Analysing differential gene expression between healthy and diseased cells gives important information on how a disease has affected the cell, which genes are up-regulated, which are down-regulated and which ones have no change in expression. Up-regulated

gene produces more of its product than normal, which may lead to the accumulation of those gene products. Down-regulation on the other hand means that there is less of the gene's product is produced.

Differential expression

A variety of methods were used in the assortment of the articles, however, there were cases in which similar approaches were used. Machine learning was used in three articles, however their approaches differed; the differences are further discussed below [71, 61, 72]. Also used by more than one article was the open source statistical programming language and analytic tool, R and the related Bioconductor project. Bioconductor is popular for the processing and analyses of biological data and there are numerous packages available for different processing needs.

The use of Bioconductor framework spanned across multiple stages of the experiments, corroborating the versatility of the project. From amongst the numerous packages of Bioconductor, the most used in relation to the differential expression analyses were *limma* and *siggenes*. The *limma* package is used for data analysis from microarrays or RNA-Seq technologies, providing a variety of functions for reading, pre-processing, exploring and analysing gene expression data [78]. The *siggenes* package makes use of methods such as Significance Analysis of Microarrays (SAM) and Empirical Bayes Analysis of Microarrays (EBAM) to identify differentially expressed genes [79].

Other methods used include two-class significance analysis of microarrays (SAM), t - test, genome-wide relative significance (GWRS) together with genome-wide global significance (GWGS) and the Coloured (α, β)-k Feature Set.

The significance analysis of microarrays (SAM) was used by Ray and colleagues in order to process probe sets for differentially expressing genes [60]. The expression values

from the entorhinal cortex, hippocampus, middle temporal gyrus and posterior cingulate cortex were previously summarized by GC-RMA. This provided them with four sets of differentially expressed genes (DEGs), one for each region of the brain they used.

The article by Feng et al. used the t test to identify significant differential expression between diseased brain samples and normal samples [65]. The list of DEGs was reduced by limiting the fold change (FC) and p values of each gene, they obtained DEGs with fold changes larger that 2 or less than 0.5 and with less than 0.05 for p value [65].

Using a relatively novel method, Zang et al. used the genome-wide relative significance (GWRS) and genome-wide global significance (GWGS) to identify the robust gene signatures, which were then used to retain a number of top ranked genes for further analyses [62]. GWRS measures the degree of differential expression on a genome-wide scale using fold changes. Based on corresponding GWRS of a gene, GWGS was computed - a value marking the global significance of a gene across multiple studies. These genes were then ranked according to the degree of differential expression and 300 top scoring genes were selected.

Puthiyedth and colleagues used gene expression data from six brain regions (entorhinal cortex, hippocampus, middle temporal gyrus, posterior cingulate cortex, superior frontal gyrus and visual cortex) in two sets of analyses [63]. First, they identified the probes differentially expressed in each region separately using (α, β)-k Feature Set approach. Second, they combined the data from each region, in which the specific probe values from each of the regions were combined. The Coloured (α, β)-k Feature Set approach was then applied on the combined data in order to acquire the probes differentially expressed in every region [63]. They compared this combined region result with other commonly used methods, the RankProd and GeneMeta of Bioconductor.

The RankProd method uses the fold changes (FC) of genes for the ranking and com-

parison within regions. GeneMeta uses the false discovery rate (FDR) of genes for the production of a ranked gene list and is based on a meta-analysis method. Article written by Yue et al. uses RankProd to integrate their chosen datasets and detect from these data the DEGs [64]. Four different methods are then used on this integrated data to identify differential co-expression and obtain the gene association scores. The four methods used are the STRING database, Differentially Co-expressed Genes and Link (DCGL) package of R, Empirical Bayesian (EB) analysis and Weighted Gene Co-expression Network Analysis (WGCNA). These methods are further discussed below, as well as in the case study section.

The three articles using machine learning techniques, used different approaches, further introduced in this and the next paragraphs [71, 61, 72]. Martinez-Ballesteros et al. used Quantitative Association Rules (QAR), the C4.5 algorithm and the GarNet algorithm and presented the integration of these machine learning methods [71]. First, they used the C4.5 algorithm to obtain a classification of healthy and diseased genes from the dataset, which was then used to obtain the percentage of correctly classified instances. This percentage was used as minimum threshold for the selection of GarNet configurations. GarNet was then executed several times on a test set in order to obtain an accuracy value higher than that of the C4.5. Next, they ran the GarNet on the original dataset to obtain QAR providing information about the whole dataset and extracted the most frequent gene-AD state associations [71].

Jamal et al., however have used a selection of eleven machine learning methods in order to predict AD-connected genes from their gene pool [61]. They obtained 56 405 genes from the Entrez Gene database, which provided them with a positive dataset containing 458 genes reported as possible AD causes and a negative dataset containing 55947 genes not related to AD. Their machine learning methods include Naive Bayes (NB), NB

Tree, Bayes Net, Decision table, Decision table/NB (DTNB) hybrid classifier, Random Forest (RF), J48, Functional Tree, Locally Weighted Learning (J48 + k-nearest neighbour (KNN)), Logistic Regression and Support Vector Machine (SVM). They computed the features of networks, sequences and functions which were then used by the machine learning algorithms in order to generate classifiers [61]. For the training of the classifier models, ten-fold cross-validation was used, and the results were averaged across the generated models [61]. A gene was used in their following analyses, if it had been predicted by all of the methods [61].

Song et al. used the NetWAS approach to prioritize the previously found GWAS results [72]. The GWAS and NetWAS prioritised gene lists were compared with AD-associated genes from the Online Mendelian Inheritance in Man (OMIM) database, and it was found that NetWAS associations were in accordance with known annotations [72].

Integration of methods

Some of the articles used novel approaches by combining methods or algorithms and some introduced newly composed methods for data analysis, differential gene analysis or gene signature combining. Most of those are described below.

Based on a schematic network of AD, Hao and Friedman constructed a mathematical model comprising of partial differential equations [74]. This model can be used to simulate the effect of different drugs on AD, drugs that have been used in clinical trials or those that have failed them [74].

A novel approach was composed for the combination of gene signatures, using the multiplication of matrices [64]. Using four different approaches to find four sets of gene pair signatures, and processing these into uniformity, matrices of those signatures were formed. A new matrix, formed by the multiplication of those four matrices, comprised

a new and combined score for each gene pair [64].

In order to better analyse gene co-expression networks, a novel method based on differential topology was composed [60]. They based this method on the idea, that genes with differential topology between two region-specific co-expression networks would have region- or condition-specific functions [60]. Since AD is a progressive disease, meaning that two successive regions of the brain may not be in the same stage of the disease, this novel method could provide new knowledge of the disease and its progression, and of the earlier stage attributes [60].

Nevado-Holgado and Lovestone made use of fuzzy logic and the boolean operator "AND" for combining gene signatures [73]. They derived gene expression signatures from blood and brain, and of drug induced gene expression signatures. After the derivations they determined the overlap between these signatures. This approach tests if there are other ways besides suppressing inflammation with which Non-Steroidal Anti-Inflammatory Drugs (NSAID) could affect the disease risk. However, they also mention that this method does not take into account the direction of the dysregulations in gene expression [73]. Their findings suggest that NSAIDs could have additional influences than those of the known inflammation-suppression [73].

3.2.4 Additional analyses

Additional analyses such as differential co-expression, functional enrichment, pathway enrichment and network analyses that have been conducted by the subset of articles are further discussed below.

Differential co-expression analyses were conducted by Yue et al. on a set of acquired DEGs [64]. They made use of the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), Differentially Co-expressed Genes and Link (DCGL) package, Empirical

Bayesian (EB) analysis and Weighted Gene Co-expression Network Analysis (WGCNA) algorithm. The obtained scores were combined using the novel algorithm they proposed (previously discussed) producing a new score for each gene pair. They used the gene pair scores for the construction of five networks which then underwent a topological analysis, the clustering coefficients and short average paths were obtained and compared for the investigation of small-world network property [64].

The co-expression analysis used by Ray and W. Zhang was also used to introduce the topological overlap measure [60]. They compared region-specific co-expression networks and used the topological overlap measure to select genes of interest. Wang et al. used the WGCNA for the identification of gene modules with similar expression patterns [68]. They constructed a topological overlap matrix and employed the average linkage hierarchical clustering. The resulting tree was dynamically cut into highly connected modules [68].

Using the mouse ISH data, Acquaah-Mensah and Taylor used the reverse engineering of transcriptional regulatory networks, by the use of an ensemble of algorithms [70]. For the visualisation and analyses of the networks, they used Cytoscape and its packages. Yue et al. also used networks visualised and analysed with Cytoscape in their study [64]. They produced five differential co-expression networks and conducted topological analyses to find out which networks had the small-world property. Ingenuity Pathway Analysis (IPA) network analysis was used for the retrieval of biological connectivity information for significantly up- or down-regulated genes [66]. These connections were used for network generation, which were scored and ranked for comparison [66]. Jamal et al. constructed a human PPI network based on the interaction information collected from various databases with Cytoscape [61]. This network was used for the extraction of network properties of potential AD genes with their chosen 11 machine

learning algorithms [61]. The STRING database was used for the interaction analysis of found DEGs [69]. Interaction was considered significant if it had been experimentally validated and had a confidence score> 0.6. Based on these connections, a regulatory network was constructed with Cytoscape [69]. In order to construct a network, Y. Zhao et al. used the STRING database for the retrieval of significant genes, by selecting genes with a confidence score> 0.6 and used the Cytoscape for the following network construction [69]. For the detection of dense network regions, Zhang and colleagues utilized the Cytoscape plugin ClusterONE and clusters with a node count > 11 and a density score > 0.2 were selected [62].

Martinez-Ballesteros et al. conducted an enrichment analysis on the previously found gene set to obtain the respective GO terms [71]. They further restricted this set of terms and procured a set with significantly overrepresented terms. Y. Zhao and colleagues used GO for measuring the functional similarity of genes and used an R package called Ground-Operation Simulation (GOSim) for the calculations [69]. They chose $p < 0.05$ as the threshold in the identification of similar function location [69].

The popular Database for Annotation, Visualisation and Integrated Discovery (DAVID) was used for pathway and functional enrichment analyses. By providing the identified differentially expressed genes, or other genes of interest, DAVID returns the respective KEGG pathways or GO terms. For example, DAVID was used for the retrieval of enriched KEGG pathways based on a set of genes [64, 73]. DAVID was also used for the evaluation of biological and functional relevance of a set of co-expressed genes [67]. The pathways and functional terms were considered statistically significant when the respective p value was either less than 0.05 or 0.01 [65, 62].

In addition to DAVID, GeneGo MetaCore^{TM} database was also used for the identification of significant biological pathways [80, 60]. MetaCore^{TM} platform offers for ex-

ample information on interactions and pathways, gene-disease associations and several tools for data visualisation and analysis [80].

3.2.5 Statistics

Statistical methods are used for correction, calculation of different scores and measures (by which selection or discarding can be done), identification of differential expression through differences in expression levels, determining the statistical significance, etc. All of these methods, and more, are used throughout the articles. Subsequent descriptions cover most of the methods used.

An essential measure to determine the importance of the results in regard to some claim (null hypothesis) is the p-value. A cutoff value is chosen (e.g. 0.01) and is used for either rejecting ($p <= 0.01$) or failing to reject ($p > 0.01$) the corresponding claim. Another measure used is the fold change which is used to measure the change in values, e.g. gene expression between normal and diseased samples. The values of up-regulated genes fall between 1 and infinity and down-regulated gene value can be between 0 and 1. To remove this uneven distribution, the fold changes are log-transformed.

Bonferroni correction and false discovery rate (FDR) are the standard methods for the adjustment and correction of multiple testing. These methods entail the selection of probes based on a significance threshold, for example if the Bonferroni corrected p value is above the threshold of 0.0001, the probe is discarded [63, 67]. This is similar for FDR, e.g. if the FDR has a value smaller than that of the threshold, the probe is considered significant [68, 62].

Pearson's correlation coefficient was used for measuring the expression similarity between genes in order for the genes to be consider co-expressed (additional conditions applied) in the article by Ray and W. Zhang and to measure the coexpression significance of gene

pairs in the article by Feng and colleagues [60, 65].

Other less used statistical methods include Benjamini and Hochberg statistical tests, Fisher's exact test, Mann-Whitney U-test, Student t-test.

3.2.6 Visualisation

The visualisation of data is important to improve the understanding of it. Most common visualisations were networks, heatmaps and volcano plots, further introduced below. The constructed networks usually displayed the differentially expressed genes and their connections or the regulatory interactions between genes (regulatory network). Most commonly the networks were built using the Cytoscape platform, but there were cases in which Ingenuity Pathway Analysis (IPA) was used. Both provide extensive analyses of the networks, such as cluster analysis, topological analysis or the calculation of different network properties.

For example Acquaah-Mensah and Taylor visualized the obtained regulatory networks with Cytoscape and after analyses and examinations, identified a regulatory motif [Fig.14] consisting of three regulators all connected [70]. Yue et al. also used Cytoscape for the construction of their co-expression networks of which one was shown to exhibit small-world network properties and another exhibiting scale-free characteristics [64].

Jamal et al. used Cytoscape for the construction of human PPI network [61]. The PPI network of Zhang and colleagues was visualized by Cytoscape based on the STRING interaction data [62]. Y. Zhao and colleagues used Cytoscape for the visualization of miRNA regulatory networks [69]. The network visualization tool used by L.Zhao et al. was IPA – the Ingenuity Pathway Analysis, providing them with a molecular network [66]. IPA provides web-based analyses, integration and interpretation of data derived from the Ingenuity Knowledge Base [81]. Feng et al. used the online IPA database for

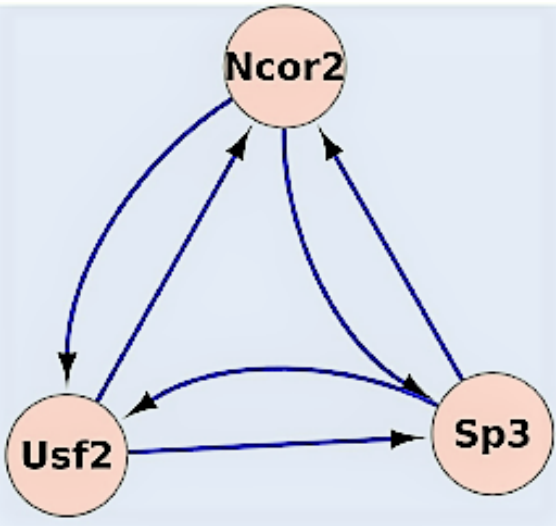

Figure 14. Part of Figure 2 from the paper by Acquaah-Mensah and Taylor [70]. The figure shows the regulatory motif featuring genes predicted to be each other's transcriptional targets.

obtaining PPI and protein-biomolecule interaction information and constructed a PPI network in which each edge was linked with at least 1 DEG and the chosen threshold for the co-expressed gene pair significance correlation was chosen as absolute Pearson Correlation Coefficient larger than 0.6 [65].

Heatmaps were the second most used visualisation techniques. They display the intensity of the differential expression, typically using one colour (e.g.red) to show over-expression and another colour(e.g. green) to show under-expression, with the interim stages being a mixture of those intensities. The construction of heatmaps was mostly conducted via the R environment. One important step in visualising the data as heatmap is the data clustering.

Hierarchical clustering used Spearman, Pearson correlation or Eucleidean distance as similarity/distance measures.

Martinez-Ballesteros et al. used hierarchical clustering as assessment of how well

the genes, found by the machine learning algorithm they used, classified as control or diseased based on their expression levels [71]. For the clustering of patients, they used Spearman correlation, for genes they used Pearson's correlation. In the article by Forabosco et al., hierarchical clustering tree was used for module detection [67]. To create this tree, they used the dissimilarity matrix, which was based on the topological overlap measure. Using a dynamic tree-cutting algorithm, they defined a set of modules based on branches of the clustering tree.

Wang et al. created a topological overlap matrix based on previously found adjacency matrix and employed hierarchical clustering algorithm to cluster probesets based on the topological overlap matrix [68]. Later, a tree-cutting algorithm was used and the hierarchical clustering dendrogram was cut into highly connected modules.

Hierarchical clustering based on Eucleidean distance was used by Y. Zhao et al. in order to determine the closest associations and cluster the found DEGs and miRNAs as well as distinguish the diseased and control tissues according to the expression values [69].

The article by L. Zhao et al. mentioned hierarchical clustering, in which Pearson's correlation was used for the distance calculations between assays [66]. One of their heatmaps can be seen on Figure 15.

The volcano plots were used for the visualisation of fold change values and p-values. L. Zhao et al. used the volcano plot to display statistically significant genes with large and small expression changes [66]. The x-axis displayed the values of fold changes and y-axis those of the p-values as can be seen from Figure 15 [66]. Martinez-Ballesteros et al. displayed the expression levels of the selected genes with a volcano plot [71]. They used $p < 0.05$ as a cutoff and identified up-regulated genes by having $FC > 1.5$ and $FC < 0.66$ for the representation of down-regulation [71].

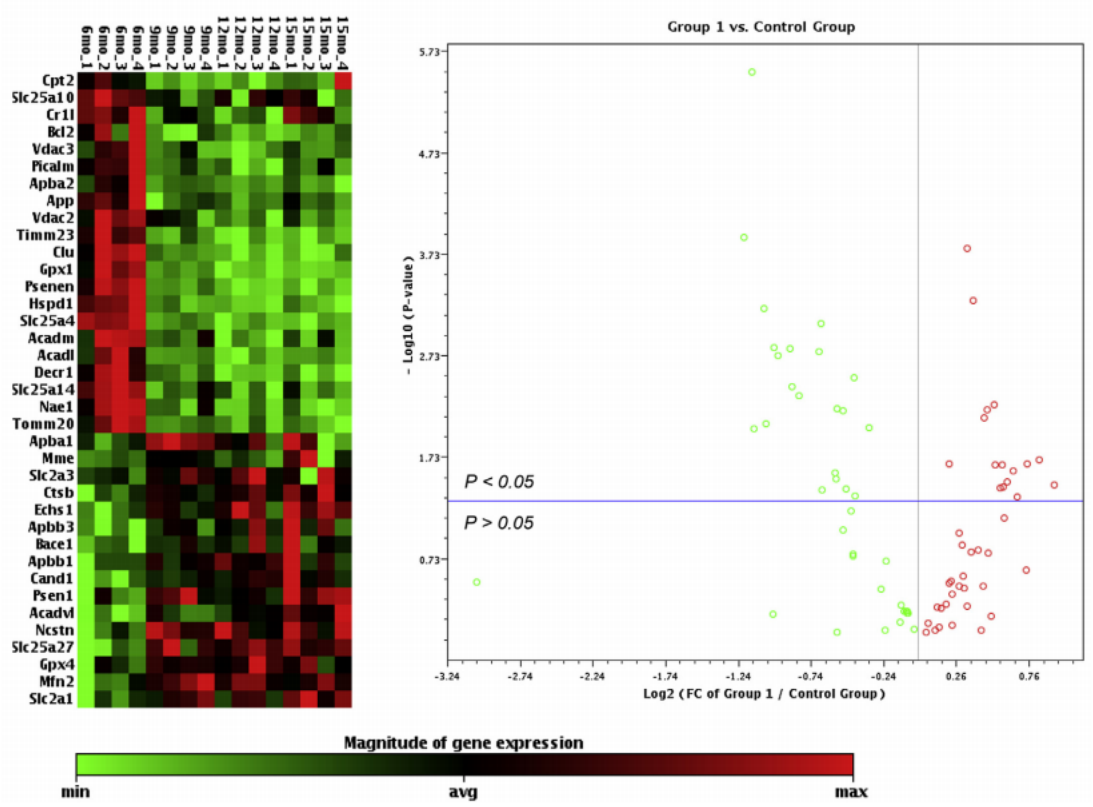

Figure 15. Part of Figure 3 from the study by L. Zhao and colleagues [66].The heat map displays hippocampal gene expression changes in female mice of different ages. As can be seen, four mice have been chosen per age group. The red signifies high expression and green signifies low expression. The volcano plot on the left displays the fold changes and p-values of the same age groups as the heatmap. The blue horisontal line in the volcano plot, separates the significantly changed genes ($p < 0.05$) and insignificantly changed genes ($p > 0.05$). The red dots represent up-regulated and green dots down-regulated genes.

3.3 Summary of the outcomes and their promises

The novel genes that were reported in the articles (not considering supplementary data) were summarised and have been presented as a Supplementary Table. With the use of machine learning approaches and molecular analyses, Jamal et al. identified thirteen genes with relation to AD and showed that an investigational AD-specific drug, AL-108 inhibited the novel drug targets found in the study [61]. The study by Zhang et al.

identified eight genes of which some have been proven to have relations to AD [62]. The study also reported four significant KEGG pathways. The analyses of Puthiyedth and colleagues provided a set of six genes and two miRNAs highly correlated to AD with the potential of improving the disease's knowledge [63]. Feng et al. reported a set of seven up- and four down-regulated significant genes, significant pathways and GO terms [65]. Y. Zhao and colleagues uncovered a set of target genes and their regulators, which could be used as potential therapeutic targets [69]. For the identification of novel AD-related genes, Ray and W. Zhang introduced a new network topology analysis method, which was used to examine gene co-expression networks of different brain regions [60]. The method identified a brain region to be less severely affected, which has also been confirmed in the literature [60].

The studies of L. Zhao et al. on the brains of mice from four age groups, provided knowledge of the differences between the ageing of female and male brains [66]. They also indicated that the brains of female mice, underwent changes related to ageing earlier than male mice brains [66]. The investigation into TREM2 by Forabosco and colleagues showed its mediating role of changing the microglial cytoskeleton, and connections with genetically implicated AD genes [67]. By analysing nineteen brain regions, Wang et al. identified novel networks and pathways shown to be in association with AD, as well as provided new knowledge on molecular mechanisms related to the regional vulnerability of AD [68].

Acquaah-Mensah and Taylor demonstrate the usefulness of in-situ hybridisation (ISH) expression data and its capability to offer unique insights [70]. Furthermore, the three-gene-hub they found [Fig.14] provides implication of diet-induced changes in gene expression [70]. By integrating three machine learning algorithms, Martinez-Ballesteros et al. showed the successful characterisation of information by the obtained rules of

the formed method [71]. They reported ninety genes of which some had previous connections to AD [71]. Song and colleagues reported the first use of NetWAS on any AD-related phenotype for the prioritization of the performed genome-wide association study (GWAS) results [72]. They also identifyed the need of multiple-mapping challenge solutions by new gene-based association tests [72]. Yue et al. introduced a novel approach based on the analysis of a combined co-expression network and reported the new method to have better credibility and strength compared with other methods used for the construction of networks [64]. The analyses of Nevado-Holgado and Lovestone provided additional information suggesting that NSAIDs might have an effect on gene expression pathways indirectly related to inflammation [73]. Their suggestion could introduce novel approaches for the therapeutic studies in dementia [73]. Hao and Friedman's simulations with different drugs used in (current and failed) clinical trials, indicated the efficacy of combined drug therapy [74].

3.4 Case study - Co-expression network-based analysis of hippocampal expression data associated with Alzheimer's disease using a novel algorithm

Authors: Hong Yue, Bo Yang, Fang Yang, Xiao-Li Hu, Fan-Bin Kong

DOI: 10.3892/etm.2016.3131

Journal: Experimental and Therapeutic Medicine

Volume, pages: 11, 1707-1715

The goal of the study. "to provide a novel tool for the analysis of gene interaction with a higher credibility and rapid transmission of information, concentrating on the scores of each gene pair across multiple approaches."

Datasets. Datasets were downloaded from ArrayExpress, of them together contained 54 patients and 30 normal controls. Details of the datasets are shown in Table 1.

Accession number	Sample size (cases/controls)	Platform	Featured brain regions
E-GEOD-1297	31(22+9)	Affymetrix HG-U133A	hippocampus
E-GEOD-28146	30(22+8)	Affymetrix HG-U133 Plus 2.0	hippocampus
E-GEOD-5281	23(10+13)	Affymetrix HG-U133 Plus 2.0	entorhinal cortex, hippocampus, medial temporal gyrus, posterior cingulate, superior frontal gyrus, primary visual cortex

Table 1. Modified from the paper. The details of used datasets.

Data preprocessing. The RMA method was used as background correction. MAS5.0 was used for the PM-MM value revision with median method. Gene expression level values were transformed for comparability. For data screening, they used the feature filter method from genefilter package of Bioconductor. Probes not matching any genes were discarded.

Differential gene expression detection. They applied the RankProd algorithm for the integration of array datasets and for the detection of differentially expressed genes. For the significantly differentially expressed genes, the percentage of false-positive was calculated and $pfp < 0.01$ was used as the cut-off value. This provided a list of 144 differentially expressed genes.

Four methods used for the constructions of differential co-expression networks. Their study consisted of applying four separate methods (described below) for co-expression analyses, provide a combined method and compare the results.

The first method, used was the use of STRING co-expression scoring. The scores for

each protein pair were obtained and based on these scores, the STRING network was constructed. This network consisted of 74 nodes and 166 edges.

Secondly, they used an R package called differentially co-expressed genes and links (DCGL). This method identifies co-expression interactions with the use of its sub-modules. These calculations used length-normalised Eucleidean distance for the measure of differential co-expression, Pearson correlation coefficient for the filtering of gene pairs and a binomial probability model for the estimation of differential expression significance. The network constructed had 16 nodes and 43 edges.

The third method was the Empirical Bayesian approach. The differentially co-expressed genes were identified by the control of FDR at the 0.05 value. The obtained pairwise correlations were visualised in the co-expression network. The network had 76 nodes and 88 edges.

The fourth method was the WGCNA. The method was used to perform an analysis on the correlation network, as well as the construction of the network. The network included 107 nodes and 2 271 edges.

Combining the scores of the various methods. This was done with the novel algorithm, which used the multiplication of the four matrices of respective methods to produce a new matrix with the new combined score. This was followed by the construction of the fifth - combined co-expression network. The network consisted of 37 nodes and 57 edges.

Network analyses. With the use of Cytoscape, a clustering coefficient, short average path length and fitting coefficient R^2 (a measure of degree distributions) were calculated for each of the five networks, which are shown in Table 2.

After the comparison of the topological parameters, it was found that the network to

Measure	STRING	DCGL	EB	WGCNA	Combined
R^2	0.786	0.037	0.477	0.071	0.810
Clustering coefficient	0.300	0.178	0.0	0.820	0.172
Mean shortest path length	2.925	1.783	2.038	1.578	3.618

Table 2. Table II. from the article. Topological parameters of co-expression networks constructed using four existing approaches and the new algorithm [64].

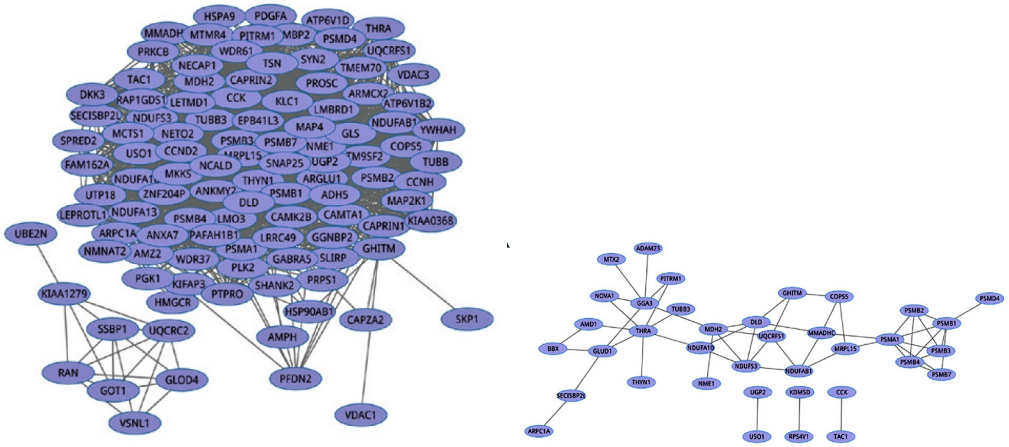

Figure 16. Sections taken from the article. These show the WGCNA network (left) and the combined network (right). The WGCNA shows small-world network properties and the combined network exhibits scale-free network properties.

show the greatest small-world characteristics was WGCNA, and the network to show the greatest characteristics of scale-free properties was the combined network.

Functional enrichment analysis. The differentially expressed genes identified with RankProd were used for the pathway enrichment analysis with DAVID online tool. From this list, the five top pathways (in decreasing order) in which the genes were enriched in, included proteasome, oxydative phosphorylation, Parkinson's disease, Huntington's disease and AD pathways. However the genes identified by the DCGL and EB methods, had no enriched pathways.

Closing remarks. It was mentioned by Yue et al. that the methods used depend on

the essence of the subsequent experiment. Therefore one has to choose carefully the method to be used, because different co-expression network analysis methods provide different results [64].

The novel combined method aimed for greater credibility and strength in gene interaction analyses [64]. Furthermore, the respective network exhibited scale-free properties, inherent to biological networks.

3.5 Trends and tendencies

The articles reported finding many novel genes related to AD. By my counting, out of the 170 genes proposed, only 10 were reported more than once. This small overlap between the genes provides further evidence of the complexity of AD. The novel methods described still need validation, as is stated by the authors. However, as the amount of people affected by AD grows each day, there is a pressing need for additional knowledge of the onset and progression of this disease.

More than 100 years of research into the causes and progression of this disease has provided little knowledge of prevention or reversion of the disease symptoms. This further corroborates the complexity of the disease

These 15 articles are but a small amount of those submitted yearly. The growing amount of information can cause overlooking of good methods because of unclear or lacking descriptions.

4 Conclusion

This thesis provides an overview of different methods used for the research of Alzheimer's disease. These methods are provided in groups of the corresponding tasks they are used for and the most popular methods are further discussed.

There are various approached which can be used for improving the knowledge of Alzheimer's disease, of which differential gene expression analyses are mostly introduced. These analyses combine the use of microarray data, data processing algorithms and knowledge integration from databases with validated information. Furthermore, there are numerous data visualisation methods which can be used for improving the comprehension of the findings. This thesis adresses the most used methods across a selection of articles related to Alzheimer's disease and bioinformatics methods.

During the analysis of the articles, some intricacies appeared. For example the complex descriptions of conducted experiments or of composed novel algorithms. The proposed novel methods, approaches or algorithms were mentioned to be in need of validation. The reconstruction of the methods from another point of view by other scientists, could bring out some discrepancies, not detected by the original authors. Therefore detailed and unambiguous descriptions of novel approaches are very important.

It was also noticed, that the proposed novel genes, had very little overlap amongst themselves. Different methods and approaches could be factors contributing to this observation. However it also indicates the necessity for additional analyses of these novel genes, which would benefit from unambiguous reports as well.

This thesis could provide an initial grasp on the studies, analyses and methods conducted with the purpose of understanding the mechanics of Alzheimer's disease onset and progression. As well as an overview of already conducted analyses and their results.

References

[1] Arthur M. Lesk. *Introduction to Bioinformatics*. 4th ed. Great Britain, Glasgow: Oxford University Press, 2014.

[2] Ain Heinaru. *Geneetika*. estonian. Tartu: Tartu Ülikooli kirjastus, 2012.

[3] Francis HC Crick. "On protein synthesis". In: *Symp Soc Exp Biol*. Vol. 12. 1958, p. 8.

[4] *The central dogma of molecular biology.* URL: `https://commons.wikimedia.org/wiki/File:Central_dogma_of_molecular_biology.svg` (visited on 11/05/2017).

[5] *The splicing of pre-mRNA*. URL: `location:https://commons.wikimedia.org/wiki/File:Pre-mRNA_to_mRNA.svg` (visited on 11/05/2017).

[6] Keith A. Johnson et al. "Brain Imaging in Alzheimer Disease". In: *Cold Spring Harbor Perspectives in Medicine* 2.4 (Apr. 2012).

[7] *Illustration of a neuron.* URL: `https://commons.wikimedia.org/wiki/File:1207_Neuron_Shape_Classification.jpg` (visited on 11/05/2017).

[8] Rodolfo Llinas. "Neuron". In: *Scholarpedia* 3.8 (Aug. 2008), p. 1490. URL: `http://www.scholarpedia.org/article/Neuron`.

[9] Hanns Hippius and Gabriele Neundörfer. "The discovery of Alzheimer's disease". In: *Dialogues in Clinical Neuroscience* 5.1 (Mar. 2003), pp. 101–108.

[10] Ahmet Turan Isik. "Late onset Alzheimer's disease in older people". In: *Clinical Interventions in Aging* 5 (2010), pp. 307–311.

[11] Martin Prince et al. "World Alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future". In: (2016).

[12] Béatrice Duthey. "Background paper 6.11: Alzheimer disease and other dementias". In: *A Public Health Approach to Innovation* (2013), pp. 1–74.

[13] Neeti Sharma and Anshika Nikita Singh. "Exploring Biomarkers for Alzheimer's Disease". In: *Journal of Clinical and Diagnostic Research : JCDR* 10.7 (July 2016), KE01–KE06.

[14] Reisa A. Sperling et al. "Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". en. In: *Alzheimer's & Dementia* 7.3 (May 2011), pp. 280–292.

[15] *AD progression in brain.* URL: https://flic.kr/p/DnazK6 (visited on 11/05/2017).

[16] Mark P. Mattson and Tim Magnus. "Ageing and neuronal vulnerability". In: *Nature Reviews Neuroscience* 7.4 (Apr. 2006), pp. 278–294.

[17] . "UniProt: the universal protein knowledgebase". In: *Nucleic Acids Research* 45.D1 (2017), p. D158. eprint: /oup/backfile/content_public/journal/nar/45/d1/10.1093_nar_gkw1099/3/gkw1099.pdf. URL: +http://dx.doi.org/10.1093/nar/gkw1099.

[18] *The formation of tau tangles.* URL: https://commons.wikimedia.org/wiki/File:TANGLES_HIGH.jpg (visited on 11/05/2017).

[19] *The formation of Beta-Amyloid plaque.* URL: https://commons.wikimedia.org/wiki/File:Amyloid-plaque_formation-big.jpg (visited on 11/05/2017).

[20] *Microarray Technology: An introduction to DNA Microarray*. URL: http://www.premierbiosoft.com/tech_notes/microarray.html (visited on 11/04/2017).

[21] Dennise D. Dalma-Weiszhausz et al. "The Affymetrix GeneChip® Platform: An Overview". en. In: *Methods in Enzymology*. Vol. 410. DOI: 10.1016/S0076-6879(06)10001-4. Elsevier, 2006, pp. 3–28.

[22] Chris Seidel. "Introduction to DNA Microarrays". In: *Analysis of Microarray Data: A Network-Based Approach*. Ed. by Frank Emmert-Streib and Matthias Dehmer. 2008.

[23] Robert M. Flight, Abdallah M. Eteleeb and Eric C. Rouchka. "Affymetrix® Mismatch (MM) Probes: Useful After All". In: *BioMedical Computing (BioMedCom), 2012 ASE/IEEE International Conference on*. IEEE, 2012, pp. 6–13.

[24] *Comparison of Affymetrix and Illumina arrays.* URL: https://commons.wikimedia.org/wiki/File:Affymetrix_GeneChip_and_Illumina_BeadChip_designs.svg (visited on 11/05/2017).

[25] Arnold Oliphant et al. "BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping". In: *Biotechniques* 32.6 (2002), pp. 56–58.

[26] Jarno Tuimala and M. Minna Laine. *DNA microarray data analysis*. English. OCLC: 58384983. Espoo: CSC - Scientific Computing, 2003.

[27] Rafael A. Irizarry, Zhijin Wu and Harris A. Jaffee. "Comparison of Affymetrix GeneChip expression measures". In: *Bioinformatics* 22.7 (Apr. 2006), pp. 789–794.

[28] *About GEO DataSets - GEO - NCBI*. URL: https://www.ncbi.nlm.nih.gov/geo/info/datasets.html (visited on 18/04/2017).

[29] Rafael A. Irizarry et al. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data". In: *Biostatistics* 4.2 (Apr. 2003), pp. 249–264.

[30] Yunxia Sui et al. "Background Adjustment for DNA Microarrays Using a Database of Microarray Experiments". In: *Journal of Computational Biology* 16.11 (Nov. 2009), pp. 1501–1515.

[31] *Statistical Algorithms Description Document*. . 2002. URL: http://tools.thermofisher.com/content/sfs/brochures/sadd_whitepaper.pdf#/legacy=affymetrix.com (visited on 09/05/2017).

[32] Stuart D Pepper et al. "The utility of MAS5 expression summary and detection call algorithms". In: *BMC Bioinformatics* 8 (July 2007), p. 273.

[33] Sung E Choe et al. "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset". In: *Genome Biology* 6.2 (2005), R16.

[34] Zhijin Wu et al. "A Model-Based Background Adjustment for Oligonucleotide Expression Arrays". In: *Journal of the American Statistical Association* 99.468 (Dec. 2004), pp. 909–917.

[35] Georgios A Pavlopoulos et al. "Using graph theory to analyze biological networks". In: *BioData Mining* 4 (Apr. 2011), p. 10.

[36] Uri Alon. "Network motifs: theory and experimental approaches". In: *Nature Reviews Genetics* 8.6 (June 2007), pp. 450–461.

[37] E Alm. "Biological networks". en. In: *Current Opinion in Structural Biology* 13.2 (Apr. 2003), pp. 193–202.

[38] *Examples of complex networks.* URL: https://commons.wikimedia.org/wiki/File:Complex_networks.png (visited on 11/05/2017).

[39] Hui-Huang Hsu, ed. *Advanced data mining technologies in bioinformatics*. Hershey PA: Idea Group Pub, 2006.

[40] Dianne Cook et al. "Exploring gene expression data, using plots". In: *Journal of Data Science* 5.2 (2007), pp. 151–182.

[41] *An example of a heatmap.* URL: https://www.flickr.com/photos/mitopencourseware/4815736796 (visited on 11/05/2017).

[42] *An example of a clustered heatmap and the respective dendrogram.* URL: https://commons.wikimedia.org/wiki/File%3AAdult-Rat-Bones-Maintain-Distinct-Regionalized-Expression-of-Markers-Associated-with-Their-pone.0008358.g002.jpg (visited on 11/05/2017).

[43] Jonathan Pevsner. *Bioinformatics and functional genomics*. Third edition. Chichester, West Sussex, UK ; Hoboken, NJ, USA: John Wiley and Sons, Inc, 2015.

[44] *An example of a volcano plot.* URL: https://commons.wikimedia.org/wiki/File:Volcano_eg.jpg (visited on 11/05/2017).

[45] Hadley Wickham and Lisa Stryjewski. "40 years of boxplots". In: *Am. Statistician* (2011).

[46] *The elements of a boxplot.* URL: https://commons.wikimedia.org/wiki/File%3AElements_of_a_boxplot_en.svg (visited on 11/05/2017).

[47] A. Brazma et al. "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data". In: *Nature Genetics* 29.4 (Dec. 2001), pp. 365–371.

[48] *BioSharing: bsg-s000177: MIAME.* . URL: https://biosharing.org/bsg-s000177 (visited on 09/05/2017).

[49] *GEO Documentation - GEO - NCBI.* URL: https://www.ncbi.nlm.nih.gov/geo/info/ (visited on 26/04/2017).

[50] Nikolay Kolesnikov et al. "ArrayExpress update—simplifying data submissions". In: *Nucleic Acids Research* 43.D1 (Jan. 2015), pp. D1113–D1116.

[51] NCBI Resource Coordinators. "Database resources of the National Center for Biotechnology Information". In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D7–D19.

[52] Michael Ashburner et al. "Gene Ontology: tool for the unification of biology". en. In: *Nature Genetics* 25.1 (May 2000), pp. 25–29.

[53] Minoru Kanehisa and Susumu Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30.

[54] David Croft et al. "Reactome: a database of reactions, pathways and biological processes". In: *Nucleic Acids Research* 39.Database issue (Jan. 2011), pp. D691–D697.

[55] Damian Szklarczyk et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life". In: *Nucleic Acids Research* 43.Database issue (Jan. 2015), pp. D447–D452.

[56] Kurt Hornik. *R FAQ*. 2016. URL: https://CRAN.R-project.org/doc/FAQ/R-FAQ.html (visited on 08/05/2017).

[57] Paul Shannon et al. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks". In: *Genome Research* 13.11 (Nov. 2003), pp. 2498–2504.

[58] Max Franz et al. “Cytoscape.js: a graph theory library for visualisation and analysis”. en. In: *Bioinformatics* (Sept. 2015), btv557.

[59] Michael E. Smoot et al. “Cytoscape 2.8: new features for data integration and network visualization”. In: *Bioinformatics* 27.3 (Feb. 2011), pp. 431–432.

[60] Monika Ray and Weixiong Zhang. “Analysis of Alzheimer’s disease severity across brain regions by topological analysis of gene co-expression networks”. In: *BMC Systems Biology* 4 (Oct. 2010), p. 136.

[61] Salma Jamal et al. “Integrating network, sequence and functional features using machine learning approaches towards identification of novel Alzheimer genes”. In: *BMC Genomics* 17 (Oct. 2016).

[62] L. Zhang et al. “Potential hippocampal genes and pathways involved in Alzheimer’s disease: a bioinformatic analysis”. In: *Genetics and Molecular Research* 14.2 (2015), pp. 7218–7232.

[63] Nisha Puthiyedth et al. “Identification of Differentially Expressed Genes through Integrated Study of Alzheimer’s Disease Affected Brain Regions”. In: *PLoS ONE* 11.4 (Apr. 2016).

[64] HONG YUE et al. “Co-expression network-based analysis of hippocampal expression data associated with Alzheimer’s disease using a novel algorithm”. In: *Experimental and Therapeutic Medicine* 11.5 (May 2016), pp. 1707–1715.

[65] Bo Feng et al. “Analysis of Differentially Expressed Genes Associated With Alzheimer’s Disease Based on Bioinformatics Methods”. en. In: *American Journal of Alzheimer’s Disease & Other Dementias®* 30.8 (Dec. 2015), pp. 746–751.

[66] Liqin Zhao et al. “Sex differences in metabolic aging of the brain: insights into female susceptibility to Alzheimer’s disease”. eng. In: *Neurobiology of Aging* 42 (June 2016), pp. 69–79.

[67] Paola Forabosco et al. “Insights into TREM2 biology by network analysis of human brain gene expression data”. en. In: *Neurobiology of Aging* 34.12 (Dec. 2013), pp. 2699–2714.

[68] Minghui Wang et al. “Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer’s disease”. en. In: *Genome Medicine* 8.1 (Dec. 2016).

[69] Y. (1) Zhao et al. “Identification of Biomarkers Associated with Alzheimer’s Disease by Bioinformatics Analysis”. English. In: *American Journal of Alzheimer’s Disease and other Dementias* 31.2 (2016). 163, pp. 163–168.

[70] George K. Acquaah-Mensah and Ronald C. Taylor. “Brain in situ hybridization maps as a source for reverse-engineering transcriptional regulatory networks: Alzheimer’s disease insights”. en. In: *Gene* 586.1 (July 2016), pp. 77–86.

[71] María Martínez-Ballesteros et al. “Machine learning techniques to discover genes with potential prognosis role in Alzheimer’s disease using different biological sources”. In: *Information Fusion* 36 (2016), pp. 114–129.

[72] Ailin Song et al. “Network-based analysis of genetic variants associated with hippocampal volume in Alzheimer’s disease: a study of ADNI cohorts”. In: *BioData Mining* 9 (Jan. 2016).

[73] Alejo J. Nevado-Holgado and Simon Lovestone. “Determining the Molecular Pathways Underlying the Protective Effect of Non-Steroidal Anti-Inflammatory Drugs for Alzheimer’s Disease: A Bioinformatics Approach”. eng. In: *Computational and Structural Biotechnology Journal* 15 (2016), pp. 1–7.

[74] Wenrui Hao and Avner Friedman. “Mathematical model on Alzheimer’s disease”. In: *BMC Systems Biology* 10 (Nov. 2016), pp. 1–18.

[75] *Laser Capture Microdissection*. URL: `https://www.thermofisher.com/tr/en/home/life-science/gene-expression-analysis-genotyping/laser-capture-microdissection.html` (visited on 11/04/2017).

[76] Nelson Spruston. “Pyramidal neuron”. In: *Scholarpedia* 4.5 (May 2009), p. 6130.

[77] Gregor Eichele and Graciana Diez-Roux. “High-throughput analysis of gene expression on tissue sections by in situ hybridization”. en. In: *Methods* 53.4 (Apr. 2011), pp. 417–423.

[78] Gordon K. Smyth et al. “limma: Linear Models for Microarray and RNA-Seq Data User’s Guide”. In: ().

[79] Holger Schwender. “Identifying differentially expressed genes with siggenes”. In: *A Bioconductor Package* (2004).

[80] Sean Ekins et al. “Pathway mapping tools for analysis of high content data”. eng. In: *Methods in Molecular Biology (Clifton, N.J.)* 356 (2007), pp. 319–350.

[81] Andreas Krämer et al. “Causal analysis approaches in Ingenuity Pathway Analysis”. In: *Bioinformatics* 30.4 (Feb. 2014), pp. 523–530.

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Joanna Niklus**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   of my thesis

   Current State-of-the-Art Bioinformatics Methods in Alzheimer's Disease Studies

   supervised by Hedi Peterson

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

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

Tartu, 11.05.2017