Epigenetic data analysis on Alzheimer’s disease

Pim van den Bogaerd
Epigenetic data analysis on Alzheimer’s disease

by

Pim van den Bogaerd

in partial fulfillment of the requirements for the degree of

Bachelor of Science
in Computer Science

at the Delft University of Technology,
to be defended publicly on Tuesday June 30, 2015 at 11:30 AM.

Supervisors: Prof. dr. ir. Marcel Reinders, TU Delft
Dr. Marc Hulsman, TU Delft
Client: Dr. Henne Holstege, VUmc
Bachelor Project Coordinator: Dr. ir. Felienne Hermans, TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.
# Contents

1 **Introduction** ........................................ 1

2 **Methods** ........................................ 2
   2.1 Data sets ........................................ 2
   2.2 Epigenetic states ................................ 3
   2.3 Mapping SNPs to genes and states .............. 3
   2.4 Epigenetic Hamming distances .................. 4
   2.5 State distribution distance ..................... 5
   2.6 Clustering ....................................... 6
   2.7 Classifier ....................................... 6

3 **Results** ........................................ 7
   3.1 Epigenetic states of SNPs differ between Alzheimer and control SNPs ...................... 7
   3.2 Heat map does not reveal an Alzheimer-specific state pattern .............................. 7
   3.3 Some Hamming distances show a slight Alzheimer-specific pattern ....................... 8
   3.4 State distribution distance shows a slight Alzheimer-related pattern ................. 9
   3.5 Clustering methods do not show Alzheimer-related patterns ............................ 9
   3.6 Classifier approaches do not yield Alzheimer-specific results ....................... 10
   3.7 Per-SNP transitions show Alzheimer-related pattern ................................ 10

4 **Discussion** ....................................... 17
   4.1 Closest gene mapping may not be appropriate .................................................. 17
   4.2 Statistical methods might show more results ................................................. 17

5 **Reflection** ....................................... 18

6 **Software used** .................................... 19
   A Maximum flow clustering .......................... 22
   B Distance histograms ............................... 23
   C Info sheet ........................................ 26
   D Original project description ..................... 27

**Bibliography** ....................................... 28
The bachelor project is the last course of the Computer Science bachelor program at TU Delft. For this course I investigated genetic variations that underlie Alzheimer’s disease, thereby contributing to the 100-plus research project of the VU University Medical Center (VUmc) in Amsterdam. This report concludes the bachelor project, which lasted ten weeks.

First of all, I would like to thank my supervisors prof. dr. ir. Marcel Reinders and dr. Marc Hulsman for the weekly appointments and the useful feedback that made this report possible. Also, I would like to thank the project leader, dr. Henne Holstege, for allowing me to participate in the 100-plus project. Further, I would like to express my thanks the bachelor project coordinator dr. Felienne Hermans.

Last but not least, my thanks go to the academic counsellors Gytha Rijnbeek and Jolien Kooijman for supporting me throughout the course of the bachelor project.

I hope that the contents of this report will help others who investigate genetic variations causing Alzheimer’s disease.

We thank the International Genomics of Alzheimer’s Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i-Select chips was funded by the French National Foundation on Alzheimer’s disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant n° 503480), Alzheimer’s Research UK (Grant n° 503176), the Wellcome Trust (Grant n° 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant n° 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer’s Association grant ADGC–10–196728.

Pim van den Bogaerd
Schiedam, June 2015
Mutations in one's DNA can influence the risk for developing Alzheimer's disease (Alzheimer). DNA consists of long strings of **nucleotides**, which together define the **genome**. However, next to these **nucleotides**, also the way in which DNA wraps in a cell affects the function of a cell. The information describing the DNA wrapping state is called the **epigenome**. A recent paper by Kundaje et al. [1] has provided, among other things, an epigenetic label (**state**) to each nucleotide.

One type of mutations on the DNA affects a single nucleotide; these are called **SNPs**. Some SNPs are related to Alzheimer whereas others are related to other diseases or traits such as hair color. In this report, we have investigated if Alzheimer SNPs are related to specific epigenetic states. Specifically, we searched for patterns to distinguish Alzheimer SNPs from other SNPs. For this, we explored numerous similarity approaches, such as differences in state counts, clustering methods, and a classifier.

Up till now, we did not find an obvious difference in epigenetic patterns between Alzheimer SNPs and control SNPs.
Introduction

Alzheimer’s disease is a brain disease causing about 70% of dementia cases [2]. Dementia is becoming an highly increasing problem due to population ageing. In the next decades, the number of people reaching the age of 99 is expected to increase significantly: it is expected that in 2060, this number has increased by a factor of almost 10 [3].

There is evidence that the genome 1 plays a large role in determining ones risk of getting Alzheimer. Variations in the genome have been identified that are likely related to Alzheimer, see e.g. [4].

The genome is represented using DNA molecules. These molecules consist of long strings of nucleotides. In the context of this paper, a nucleotide can be modelled as one of the four letters A, C, G and T. The nucleotides encode for proteins that control various processes in a human being. Such processes can develop traits such as hair color, as well as diseases. A variant or single nucleotide polymorphism (SNP) occurs when an individual possesses a different nucleotide at a specific position on his genome compared to other individuals.

Complimentary to the genome is the epigenome, which concerns, among other things, the spatial structure of DNA. The spatial structure of DNA refers to the way DNA wraps in three-dimensional space. In February 2015, a paper was published [1] in which the authors have attached one of fifteen states to each letter in a wide range of cell types. These states are related to the epigenome. The states are based on signals of certain proteins (histones) that cause the wrapping of DNA. A state can relate to e.g. the start of a gene. The different cell types are relevant for our study, since we can look separately at the epigenetic states in brain cell types.

There are also signs that the epigenome is related to Alzheimer [5].

An individual possesses many SNPs. Since SNPs often have little effects, large sample sizes are needed to derive significant results. There exist SNPs that are possibly related to Alzheimer, but for which this relation is not considered significant. The purpose of this report is to investigate whether the epigenome might give information that helps confirm whether such SNPs are related to Alzheimer.

The main research question is: Is there a difference in epigenetic states between SNPs related to Alzheimer and control SNPs?

To answer this question, it is necessary to define “difference” as well as to determine what genome locations to consider. We have defined several difference functions and mappings from SNPs to genome locations. To tackle the research question, we have approached it in multiple ways.

1We have added a glossary at the end of this report. The glossary contains the key terms and their definitions as used throughout this report.
2 Methods

2.1. Data sets

We have used GRCh37 as the reference genome. We will refer to the data sets by mnemonics.

For the Alzheimer SNPs, we used two data sets. The first set of SNPs (D_alz21) consists of 21 identified SNPs detected in a large GWAS study [4].

The second set (D_igap682) is of size 682, obtained from the “combined” data file of the same article [6] with \( p \)-value greater than 0.001. The set was computed by picking the SNPs sorted on increasing \( p \)-value, but for each picked SNP removing the SNPs in a range \( \pm 50.000 \) around it. This mitigates linkage dependency among the picked SNPs.

International Genomics of Alzheimer’s Project (IGAP) is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyse four previously-published GWAS datasets consisting of 17,008 Alzheimer’s disease cases and 37,154 controls (The European Alzheimer’s disease Initiative – EADI the Alzheimer Disease Genetics Consortium – ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium – CHARGE The Genetic and Environmental Risk in AD consortium – GERAD). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer’s disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 & 2.

For the control SNPs, we used three data sets. The first two are of size 21 (D_gwas21) and 682 (D_gwas682) and are obtained by drawing random samples from the GWAS catalog [7]. The third data set is a random sample of size 682 (D_dbsnp682) drawn from dbSNP [8]. SNPs identified in GWAS studies are typically related to diseases, whereas dbSNP is a database of SNPs that are not necessarily related to diseases. For example, SNPs in dbSNP may cause a different hair color. We used both sources since disease SNPs may have a bias to epigenetically look like Alzheimer SNPs, because Alzheimer is also a disease. So, a comparison of Alzheimer SNPs with dbSNP control SNPs might give more significant results than a comparison with GWAS control SNPs. However, this does not imply the differences are Alzheimer-specific since they might be disease-specific in general. The comparison with GWAS control SNPs might indicate differences that are Alzheimer-specific.

Further, we used a sample of 682 stress response genes (D_stressgenes). These were obtained by using gene category data [9], collecting categories with “stress” in the definition. These categories were linked to genes using gene association data [10]. Genes on chromosome Y were discarded since in some cell types the epigenetic states on this chromosome were not measured. We used the Table Browser to map genes to their locations [11]. Some stress response genes were not on this list, and we have discarded these genes as well. The resulting set of genes was larger than 682, and we drew a random sample of size 682 from it. We also used a random sample of 682 genes drawn from the Table Browser (D_randomgenes) [11].

We used the lift over tool [12] to convert some SNPs into their GRCh37 locations. For the GWAS catalog, we have assumed chromosome 23 stands for chromosome X.
2.2. Epigenetic states

We used the model of fifteen states from Kundaje et al. [1]. The states are defined in Table 2.1. The states were determined for each of 127 cell types. Ten of these cell types are brain cell types¹. We have sometimes applied methods to brain cell types separately since Alzheimer is a brain-related disease.

2.3. Mapping SNPs to genes and states

SNPs are locations on the genome, i.e., a pair of chromosome and chromosome location. To associate the epigenetic state on each cell type to a SNP, we used a number of approaches. Each of these approaches assigns a vector of length 127 to each SNP, since there are 127 cell types and each approach assigns a single state to a SNP for each cell type.

- **Direct approach**: We mapped the SNPs into their epigenetic state for each cell type at the particular location of that SNP. Let \( s_{\text{direct}}(\text{snp}, \text{celltype}) \) denote the epigenetic state at the location of SNP \( \text{snp} \) in cell type \( \text{celltype} \).

- **Closest gene approach**: We used the state that occurs most on the closest gene. A gene is a range on the genome that can encode for proteins. This closest gene approach may be more interesting than using the SNP location itself, since a gene can encode for a protein that might influence the likelihood of developing Alzheimer’s disease. We mapped each location on the gene to its epigenetic state, and then picked the state that occurred most. To obtain a gene range, we used the RefSeq gene locations (obtained from [11]). Specifically, we looked at epigenetic states inside the coding range of a gene (denoted by the start, \( \text{cdsStart} \), and the end, \( \text{cdsEnd} \), location of the gene); see Figure 2.2. When multiple coding sequences are known for the same gene, we merged the entries by using the first start position and the last known end position of the gene. Moreover, we discarded genes with a coding region length of 0. Let this approach be denoted analogously by \( s_{\text{gene}}(\text{snp}, \text{celltype}) \).

- **Neighborhood approach**: The third mapping we used is the state which occurs most often on a range of -10.000 - 10.000 around the SNP. This method might give a different result than the first method. Let us use \( s_{\text{10k}}(\text{snp}, \text{celltype}) \) for this mapping.

- **Promotor approach**: We used the state which occurs most on a range of 0 - 1.000 nucleotides in front of the closest gene of the SNP, as an approximation of the promoter region. The promoter

¹The ten brain cell types are: brain hippocampus middle, brain substantia nigra, brain anterior caudate, brain cingulate gyrus, brain inferior temporal lobe, brain angular gyrus, brain dorsolateral prefrontal cortex, brain germinal matrix, fetal brain female, fetal brain male.
2.4. Epigenetic Hamming distances

Given a SNP, we can represent its epigenetic profile as a vector of states. Such a vector is of size 127, with each element containing the state in a cell type. Recall that a state can be represented as a number between 1 and 15. The state for a SNP in a cell type can be determined using a mapping from the previous section.

Given such state vectors for each SNP, we can determine the similarity between two SNPs using a distance function that operates on their state vectors. We investigated (variations of) the Hamming distance between these vectors. The regular Hamming distance between vectors $\vec{a}$ and $\vec{b}$ is defined as the number of elements $a_i$ with $a_i \neq b_i$. This implies that similar state vectors have a small Hamming distance, whereas dissimilar state vectors have a larger Hamming distance. If it turns out that Alzheimer SNPs have a smaller Hamming distance among each other compared to control SNPs, then this result may give rise to finding Alzheimer-specific epigenetic profiles.

We mapped SNPs into their state vectors using the methods described in the previous section. For each pair of state vectors, we calculated a Hamming distance. For the Alzheimer-Alzheimer and control-control comparisons, we only counted each pair once, and moreover did not count SNPs with themselves. These entries are not interesting since in these two cases, the Hamming distance is symmetric and the Hamming distance between two identical vectors is 0. We did normalize the resulting counts since the bars corresponding to the Alzheimer-control case would otherwise systematically obtain higher values, and thus render them incomparable to the other bars.

We used the following Hamming distances, given state vectors $\vec{a}$ and $\vec{b}$. For the examples, let us define $\vec{x} = [3, 15, 15, 4]$ and $\vec{y} = [15, 15, 14, 9]$.

1. **Hamming1.** The regular Hamming distance function as defined above. For example, the Hamming1 distance between $\vec{x}$ and $\vec{y}$ is 3. Imperatively, we can consider the states in the first vector one at a time, and for each state consider the state in the second vector at the same position. If the states are equal, increment the distance by one.

2. **Hamming2.** A distance function based on the regular Hamming distance. A large part (68%) of the genome has been assigned state Quiescent [1, p.319] and this state means no significant epigenetic signals have been measured. Thus, although two vectors with many Quiescent states may be similar, such a result is not very informative.

We found that the regular Hamming distance considered control SNPs to have a smaller distance among each other, presumably because control SNPs are more often located within Quiescent (state 15) regions, leading to relatively many 15-15 cases. We tried to adjust for this using this modified distance function.
Concretely, the change from Hamming1 is that the distance is incremented also for each position where both vectors have state 15 (Quiescent). The distance function Hamming2 applied on \( \vec{x} \) and \( \vec{y} \) is 4.

3. Hamming3. Another distance function based on the regular Hamming distance to account for the high number of occurrences of the Quiescent state. This time, the number of occurrences \( n \) of both vectors attaining the Quiescent state is counted. The regular Hamming distance is then divided by \( 127 - n \). This value is then multiplied by 127 to get distances on the same scale. (A division by zero is resolved by setting the distance to 0 instead.) Division is required to get distances that are of the same order.

The idea behind this approach is that this function may also bump the distance between control SNPs since a large number of 15-15 cases cause division by a smaller number and thus result in a larger distance.

For example, the Hamming3 distance between \( \vec{x} \) and \( \vec{y} \) is 4. (We have multiplied by 4 rather than 127 since the example vectors are of length 4.)

4. Hamming4. A modified Hamming distance where we take into account the total occurrences of states in each cell type. This way, when comparing two states, we can add more to the distance when one or both states are rare, and add less when one or both states are common.

The distance is computed as follows. First, we count the states for each SNP in each cell type. This gives 127 vectors of length 15 where each vector contains the state counts of each state in a specific cell type.

Next, given two state vectors, determine for each cell type the two states and call the state counts \( x \) and \( y \) (as obtained from the vectors above). If the states are equal, add 0 to the distance. If the states are unequal, add \( 1/x + 1/y \) to the distance.

When comparing two states, this method adds a large number to the distance when one or both states are rare. This approach may also diminish the distance with a large number of 15-15 cases.

This distance function is not easily explained through an example, since it requires computing the total counts of many SNPs first.

2.5. State distribution distance

The Hamming distances we used require us to map SNPs to state vectors. We have considered multiple mappings, as explained in Section 2.3. However, all these mappings lose information by only taking the state occurring most in a region. For example, the closest gene mapping does not consider the relative occurrences of the other states: if, given a SNP and cell type, state A occurs most on the closest gene, then state A is used as the representative state of that SNP in that state. This happens both if state A clearly occurs most, and if state A happens to occur just a little more than state B.

To overcome this problem, we used another distance function Distribution between two SNPs (rather than state vectors). One possible advantage of this formula over the Hamming distances is that all states in a gene are taken into account. For example, if two states occur often in a gene in a specific cell type, this method does not only take into account the one that happens to attain the maximum count.

Concretely, when calculating the Distribution distance between two SNPs \( a \) and \( b \), we first fetched the state counts on the closest gene for each cell type. We put these counts into matrices of size 127 by 15, which for SNPs A and B we refer to as \( A \) and \( B \), respectively. Also define the length of the closest genes of \( a \) and \( b \) as \( s \) and \( t \), respectively. The Distribution distance between the two SNPs is then defined as follows:

\[
d(a,b) = \sum_{i=1}^{127} \sum_{j=1}^{15} (A_{ij}/s - B_{ij}/t)^2.
\]

This can be visualized as a distance function acting on histograms. Namely, for each cell type \( i \), consider the histogram over the 15 states representing the normalized state counts on both SNPs. (Such an histogram has two bars per state, and the bar heights per SNP sum to one.) The “distance”
of this histogram is the sum of squares of the differences of the bar heights per state. The distance between two SNPs is then the sum of these distances over all cell types.

2.6. Clustering

To find possible groups in the Alzheimer SNPs, we investigated three approaches: a maximum flow approach, a dendrogram approach, and MDS and TSNE plots. The former approach did not theoretically work out, and we have therefore moved its explanation to the appendix.

A dendrogram reorders a data list as defined by a distance function. A dendrogram also adds a binary tree to the reordered data, thereby revealing which data points have a small distance and which data points have a large distance. The dendrogram approach can show clustering since each node high in the tree divides the data points into two clusters that have high distance between each other.

Both MDS and TSNE are algorithms that accept a list of points and a distance function as input, and produce a list of points as output. The input points can live in a high dimension, whereas the output points are typically in a lower dimension (e.g. 2). Each output point is a representative of an input point.

The purpose of MDS and TSNE is to visually show clusterings among the input points, as defined by the distance function. If, among the input points, there are clusterings of points, then the set of output points should represent these clusterings by placing the representatives of the input points nearby each other. The interesting thing is that the output points can be produced as two-dimensional ones, which can easily be plotted. Therefore, clusterings in the original data can be spotted without trying to generate a plot in high dimensions.

2.7. Classifier

We used a classification algorithm (classifier) to predict whether the epigenetic profile of a SNP belongs to the class of Alzheimer SNPs or not. The algorithm first accepts a number of points with their pairwise distance, and the group of each point (this phase is called training the classifier). When the algorithm subsequently gets another point, it predicts the group that point belongs to, given the knowledge it has of the other points.

The \( k \)-nearest-neighbour classifier is a specific classifier that decides the class of a point by comparing the point to the \( k \) points nearest to it. The class occurring most among these \( k \) points is output by the algorithm as the class of the new point. Intuitively, if a new point is close to more points of one class than of another class, then the point most likely falls into the first class. In our case, we have two groups (Alzheimer and control), and to avoid ties, we used an odd value of \( k \): \( k = 3 \).

We used the classifier both with and without using prototypes. Without prototypes, SNPs are directly compared using the Hamming distance between each pair of SNPs. With prototypes, however, each SNP is represented by its distance to a fixed set of ‘prototype SNPs’, also according to the Hamming distance. Subsequently however, these distances for each SNPs are used to calculate an Euclidean distance between all SNPs. The advantage of using prototypes is that the Euclidean distance function, used for the classifier, adheres to basic distance properties such as the triangle inequality.
3 Results

3.1. Epigenetic states of SNPs differ between Alzheimer and control SNPs

We investigated whether there is a difference in state counts between locations of the genome related and unrelated to Alzheimer. To this end, we compared the frequencies of each state for both known Alzheimer SNPs (D_alz21) and GWAS control SNPs (D_gwas21); see Figure 3.1a. Each (SNP, cell type) pair adds 1 to the bar that corresponds to the state of that pair. Note that the states Tx and TxWk appear more in Alzheimer SNPs, while in contrast the Quiescent state occurs less often. For an interpretation of these states, see Table 2.1. The higher occurrence of Tx and TxWk states suggests that Alzheimer SNPs are more often located near genes that are actively transcribed.

The histogram using data sets (D_stressgenes), (D_randomgenes) and (D_igap682) is shown in Figure 3.1b. Remarkably, in these data sets, Alzheimer SNPs show a higher count of the Quiescent state whereas control and stress genes show higher Tx and TxWk counts. We are unsure why the relative occurrences of states Tx and Quiescent of Alzheimer and control SNPs are the opposite here. Possibly, the small size (21) of the data sets used in Figure 3.1a is a cause. The histogram also shows that stress response genes have a high number of Tx and TxWk states, which indicates that the behaviour observed in Figure 3.1a is not particular to Alzheimer.

We conclude that there are differences between the epigenetic states for different sets of SNPs and/or genes, but that these differences cannot directly be used to distinguish Alzheimer SNPs.

3.2. Heat map does not reveal an Alzheimer-specific state pattern

For the data sets of size 21, i.e., (D_alz21) and (D_gwas21), we created a heat map showing the states over the SNPs and cell types. The heat maps (Figures 3.2 and 3.3) are clustered in both dimensions by Hamming distance using hierarchical clustering. The numbers 1-21 (horizontal) represent the SNPs, whereas the 127 cell types (vertical) were obtained from the metadata file of the state data \[13\]. Brain cell types have been prepended with dots.

We noted that the Alzheimer heat map shows roughly a Tx-TxWk-Tx pattern for SNPs 7, 11, 12 and 19 \[1\], whereas this pattern does not appear in the control heat map. To test whether this pattern is particular to Alzheimer, we generated a random control sample of size 2000 from the GWAS catalog \[7\]. We calculated the state occurring most on the closest gene for each of the IGAP SNPs and these control SNPs \(s_{\text{gene}}\), and calculated the distances between the resulting vectors and the Tx-TxWk-Tx pattern vectors using two approaches.

Regarding the first approach, we first normalized the four patterns into a single one by picking the state occurring most for each cell type. We then defined the distance between a SNP and a pattern as the Hamming1 distance between the state vectors of the SNP and the normalized pattern.

\[1\]These SNPs are identified by rs10792832, rs28834970, rs11218343 and rs10838725 \[4, Table 2\].
3.3. Some Hamming distances show a slight Alzheimer-specific pattern

Regarding the second approach, we used the minimum Hamming1 distance between a SNP and each of the four patterns.

We sorted the SNPs on increasing distance with ties randomly shuffled. We generated a plot of the sorted list by means of a step function. More precisely, the x-axis represents the index of the SNPs in the sorted list. The y-axis represents the amount of Alzheimer SNPs that occur in the subset of the list that begins at the begin of the list, and ends at the index represented by the x value. In other words, for each Alzheimer SNP, the step function increases by one, but the function does not change for control SNPs.

If the pattern were particular to Alzheimer, we would expect the step function to increase rapidly for small x values, and less so for larger x values. For in that case, the Alzheimer SNPs have a smaller Hamming distance to the pattern than the control SNPs do; hence the Alzheimer SNPs appear more to the beginning of the list, and the function increases at these small indices.

However, the plots in Figure 3.4 show that this is not the case, since both functions increases roughly uniformly on its domain. In other words, the pattern does not seem particular to Alzheimer.

3.3. Some Hamming distances show a slight Alzheimer-specific pattern

Regarding the Hamming distances, we investigated whether Alzheimer SNPs are more “alike”, that is, whether they have a smaller Hamming distance to each other than control SNPs. We generated histograms to test this hypothesis.

The histograms in Figure 3.5 use the two data sets (D_alz21) and (D_gwas21). The histograms show the Hamming distance distribution on the closest gene of each SNP between the three combinations among these data sets, with histogram in Figure 3.5a considering all cell types and the one in 3.5b considering only brain cells. The distance counts are normalized because the Alzheimer-Control case considers more distances, see Section 2.4. The histograms in Figure 3.6 use the state at the location of the SNP itself.

Figure 3.5 shows that Control-Control have a more uniform distribution, whereas for brain cells only the Alzheimer-Alzheimer distances are more uniform. While we observe an enrichment at higher distances for Control-Control comparisons, we do not observe an enrichment for small distances for the Alzheimer-Alzheimer comparisons. Therefore, we conclude that the Hamming1 distance is not appropriate for separating Alzheimer SNPs from control SNPs. This is supported by the observation that the signal does not further improve when only looking at brain cells (which could be expected as Alzheimer is brain-related). We determined that this conclusion is also valid when observing the
epigenetic state directly at the SNP location (Figure 3.6).

However, a possible issue in these tests is the small number of SNPs that are tested. We therefore generated similar plots using data sets (D_{igap682}) and (D_{dbsnp682}) for the mapping _s_{10k} and _s_{promotor}; see Figures B.1 and B.2. With the _s_{10k} mapping, control SNPs tend to have a lower distance among each other. The _s_{promotor} mapping shows a slight shift of Alzheimer SNPs towards lower distances. Note that the _s_{promotor} mapping gives higher Hamming distances in general.

Further, we generated plots using distances **Hamming2**, **Hamming3** and **Hamming4** because they may give different results and possibly show a separation between Alzheimer and control SNPs. For these distance functions we used data sets (D_{igap682}) and (D_{dbsnp682}); these are shown in Figures B.3, B.4 and B.5. We counted the state occurring most on the closest genes in each case (_s_{gene}).

The **Hamming2** histograms show a shift of the Alzheimer-Alzheimer distances towards lower values when compared to Control-Control distances. Thus, using this modified Hamming distance, Alzheimer SNPs are more similar among each other than control SNPs are.

The **Hamming3** histograms show higher values than **Hamming1**. This is expected, because 15-15 cases are removed and the distance is normalized to the same range 0-127. Since 15-15 adds 0 to the distance in **Hamming1**, this means that the **Hamming3** distance is at least as high as the **Hamming1** distance. Note that, when considering all cell types, Alzheimer SNPs tend to have a smaller **Hamming3** distance than control SNPs, but the difference is small. Moreover, the pattern appears even less when only considering brain cells.

The **Hamming4** histograms do not show useful results. The distances tend to attain low values. The Control-Control distances show a slight shift towards higher values, but these distances also attain the highest count in the first bin. The brain cell types histogram is even less informative.

So, it appears that **Hamming2** and **Hamming3** show that Alzheimer SNPs are epigenetically more similar, but the differences are small and cannot easily be used to distinguish Alzheimer SNPs. Note that the histograms use data set (D_{dbsnp682}) as the control SNPs. We also generated a plot for the **Hamming2** distances using data set (D_{gwas682}) instead, see Figure B.3. The histogram is similar to Figure B.3, but the Control-Control distances seem to attain slightly lower values. Possibly, the SNPs from dbSNP attain the Quiescent state more often, resulting in a higher **Hamming2** distance.

### 3.4. State distribution distance shows a slight Alzheimer-related pattern

The state distribution distance function **Distribution** is based on the distribution of states on a gene. In Figure 3.7 we show the comparison of (D_{igap682}) and (D_{gwas682}). We observe a small translation of the Alzheimer-Alzheimer distances towards lower values when compared to Control-Control distances. Thus, in the sense of this distance function, the Alzheimer SNPs are more similar among each other than control SNPs are. This distance function seems more promising than the Hamming distances.

### 3.5. Clustering methods do not show Alzheimer-related patterns

We generated MDS and TSNE plots using the two data sets (D_{igap682}) and (D_{gwas682}), using the distance function **Distribution** because it seemed more promising than the Hamming distances.

We tried to distinguish two groups in the Alzheimer SNPs by applying MDS and TSNE to data set (D_{igap682}). The two plots in Figure 3.8 are the resulting plots. They do not show that there is a clear clustering within these SNPs, since the points cannot be clearly separated.

We also created two such plots containing both the Alzheimer and the control SNPs, so as to try to find a possible clustering between the groups of SNPs. For the latter case, we used separate colors for the Alzheimer and control SNPs. Before plotting the output points, we shuffled the list of points so that there is no bias of either of the two colors in the plot. (Plotting the first set of points first and then the second set of points would possibly give a bias towards the second color, since these points may overlap the first set of points.)

The plots in Figures 3.9 show that the two groups of SNPs cannot be distinguished by their location. In other words, these algorithms do not show a clear difference between the two groups of SNPs.

It should be noted that TSNE does not always seem to display obvious clusterings in artificial data well when the set of input points is small. As a control experiment, we created a data set of size _N_,
split this data set into two equally-sized groups, with the distance within groups uniformly in the range \([0, 1]\), and between groups uniformly in the range \([10, 11]\). For \(N = 100\), the resulting plot did show some clustering, but the points were not separable by a straight line. For \(N = 300\), it was possible to separate the groups by a straight line. The points we used for both groups of SNPs are of size 682, and we have assumed that this peculiarity did not apply.

### 3.6. Classifier approaches do not yield Alzheimer-specific results

To further determine if we could distinguish Alzheimer from control SNPs, we applied a classification algorithm. We ran the \(k\)-nearest-neighbour classifier with \(k = 3\) on the data sets (D_{igap682}) and (D_{dbsnp682}). Specifically, we used most of the Alzheimer and control SNPs for training the classifier, and the rest as points to have their group predicted by the classifier. The number of correctly predicted SNPs is a measure of how well the two groups can be distinguished.

We used the \texttt{Hamming1}, \texttt{Hamming2} and \texttt{Hamming3} distances. Moreover, we used two approaches for each of the distance functions: with and without prototypes.

More concretely, we picked the last 20 SNPs of the Alzheimer SNPs and the last 20 SNPs of the control SNPs, and used them to test whether the classifier would be able to predict the group (Alzheimer or control) correctly: if the distance function were appropriate, then significantly more than 20 SNPs of the total of 40 SNPs would get their group predicted correctly.

In the non-prototype case, the points are vectors of length 127 with each element being a state value. One of the Hamming distance functions is then used to calculate the distance between points.

In the prototype case, we picked three points of each data set and calculated one of the Hamming distances for all remaining points (i.e., for each data set, the number of points used for training the classifier is \(682 - 20 - 3 = 659\)). Each SNP is then represented as a vector of length 6, containing the distance to each of the prototype points (as defined by one of the Hamming distances). We used the Euclidean distance function for the classifier.

We ran the classifier twice for the distance functions used, each time with and without using prototypes. The percentages of correctly predicted SNPs are displayed in Table 3.6. Note that the methods without prototypes yield the same result in both runs, since randomness does not play a role in these methods.

It appears that the number of correctly predicted is not significantly above 20. Indeed, the highest number achieved is 23 (57.5%), and a binomial random variable with \(p = 0.5, n = 40\) has a probability of approximately 0.134 of attaining a value \(\geq\ 23\). In other words, all results in the table are not significant.

### 3.7. Per-SNP transitions show Alzheimer-related pattern

We were wondering whether transitions between cell types for a single SNP would show a difference between Alzheimer and control SNPs. Using data sets (D_{igap682}) and (D_{gwas682}) and mapping \texttt{s\_gene}, we counted these transitions: for each SNP, we considered each combination of two cell types, and determined the state of the SNP in both cell types. This way, for both data sets we obtained a matrix of size 15 by 15, containing the transition counts. We calculated the difference of these matrices and generated a heat map; see Figure 3.11.

The heat map shows that Tx and TwK transitions occur more within Alzheimer SNPs, and that ReprPC and ReprPCWk transitions occur more within control SNPs. This might mean that Alzheimer SNPs are located to genes that are more related to transcription than control SNPs are.
3.7. Per-SNP transitions show Alzheimer-related pattern

Figure 3.2: Heat map showing the states of the 21 Alzheimer SNPs in each cell type. The brain cell labels have been prepended with dots.
3.7. Per-SNP transitions show Alzheimer-related pattern

Figure 3.3: Heat map showing the states of the 21 control SNPs in each cell type. The brain cell labels have been prepended with dots.
3.7. Per-SNP transitions show Alzheimer-related pattern

Figure 3.4: Step function that increases for each Alzheimer SNP in the list sorted on increasing Hamming distances to the pattern. The left figure uses the normalized pattern whereas the right pattern uses the minimum distance to each of the four SNPs. The green lines pass through the points (0, 0) and (2682, 682).

Figure 3.5: Histogram of \textit{Hamming1} distances among the three combinations of 21 Alzheimer and 21 control SNPs (mapping \textit{s\_gene}). The left histogram (a) considers all cell types whereas the right histogram (b) considers brain cells only.
3.7. Per-SNP transitions show Alzheimer-related pattern

Figure 3.6: Histogram of Hamming distance among the three combinations of 21 Alzheimer and 21 control SNPs (mapping s_direct). The left histogram considers all cell types whereas the right histogram considers brain cells only.

Figure 3.7: Histogram of distance distribution among the three combinations of 682 Alzheimer and control SNPs. The left histogram considers all cell types whereas the right histogram considers only brain cells.

Figure 3.8: MDS (left) and TSNE (right) plots of 682 Alzheimer SNPs using distance distribution.
3.7. Per-SNP transitions show Alzheimer-related pattern

Figure 3.9: MDS (left) and TSNE (right) plot of 682 Alzheimer (red) and 682 control (blue) SNPs using distance function Distance

![MDS and TSNE plots](image)

Figure 3.10: Results of the classifier. The distance functions are defined in 2.4.

<table>
<thead>
<tr>
<th>Method</th>
<th>Percentage of correctly predicted points (run 1)</th>
<th>Percentage of correctly predicted points (run 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hamming1, with prototypes</td>
<td>42.5</td>
<td>40</td>
</tr>
<tr>
<td>Hamming1, without prototypes</td>
<td>57.5</td>
<td>57.5</td>
</tr>
<tr>
<td>Hamming2, with prototypes</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Hamming2, without prototypes</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Hamming3, with prototypes</td>
<td>47.5</td>
<td>45</td>
</tr>
<tr>
<td>Hamming3, without prototypes</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>
3.7. Per-SNP transitions show Alzheimer-related pattern

Figure 3.11: Heat map of differences of transition counts within SNPs. Yellow (red) colors refer to transitions occurring more in Alzheimer (control) SNPs.
Discussion

Our results indicate that there is not a clear difference between Alzheimer and control SNPs in terms of their epigenetic state. This is an unexpected result, as there do exist links between Alzheimer and epigenetic signals [5]. At the very least, we expected to be able to epigenetically distinguish Alzheimer SNPs in brain cell types.

4.1. Closest gene mapping may not be appropriate
We have doubts on whether the closest gene mapping is appropriate for measuring the similarity between the epigenetic profile of SNPs. DNA in a cell is wrapped in three dimensions, and it may happen that a SNP influences a gene that is nearby in three dimensional space but is not the closest gene when considering DNA as a one-dimensional string.

This is possibly the reason that the methods using the closest gene approach did not yield significant results. Perhaps elaborating more on one of the other mappings of this report, or on new mappings not mentioned in this report, may give other results and possibly allow one to distinguish Alzheimer SNPs from control SNPs.

4.2. Statistical methods might show more results
Our current approach mainly consists of inspecting the resulting plots to check for differences in Alzheimer SNPs and control SNPs. This means that this method is based on visual inspection by the human eye.

Also, the small sample size (21) of (D_alz21) and (D_gwas21) may be inappropriate to draw a conclusion. The number of points we used for prediction in the classifier may also be too small.

It might be that this approach does not yield results that may be uncovered by using formal statistical methods. Such an approach also gives an objective threshold for determining whether there is a significant difference.

In conclusion, we did not find the difference we hoped to find. A follow-up study might use different mappings and/or statistical methods to tackle the research question in another way, thereby possibly finding new results.
5

Reflection

During this bachelor project, I applied knowledge I gained during my study, and I also learned new concepts.

First of all, I used Python for the first time. Even though it was a new programming language for me, I did not experience much difficulty using it. The syntax is different from languages such as Java, but is nevertheless natural, and fundamental constructs such as conditions and loops are available in Python as well. Moreover, at times I used functional programming constructs, such as lambdas. Using lambdas meant I was able to reuse function definitions. For example, in a function containing loops, a lambda can be used as a generic version of a condition depending on the loop variables.

Secondly, to draw random samples from a large dataset, I was able to apply an efficient algorithm \[14\], p.245-246 that I learned during my minor. One method to draw a random sample of size \(n\) is to load the complete data set, shuffle it randomly, and pick the first \(n\) items. But for large data sets, such as the dbSNP one, I used the efficient algorithm. This algorithm loads the data set items one by one, and includes the item with a probability depending on the number of items picked so far and the number of items considered so far. This algorithm has linear runtime and does not require loading the complete data set into memory.

What I also learnt was that it is not obvious how to obtain information from data. We have tried several methods to distinguish Alzheimer SNPs, but we did not end up with a method that can be used to do so. The regular Hamming distance should be a measure of how different two vectors are, but in this context it did not yield clear results. We defined modified Hamming distances as well, but in fact they performed worse in the classifier than the regular Hamming distance did.

Also, I did have some difficulty with the biological context of the project. On high school, I did not do the final exam of the biology course, and I think this is one reason why I did not recognize many terms and concepts. Thanks to my supervisors, I was able to catch up with the relevant concepts of this project.

Last but not least, I learned to apply statistical concepts and computer science to a research project with concrete relevance to the present day and future society.
We used Python for data processing and generating most of the plots. In hindsight, we used a pipeline architecture where each script accepted an input file, and output either a file or a plot. We have described the scripts in Table 6.1. Initially, we also set up a testing framework for quality control, but we have not used it.

In the script that mapped SNPs to states (find_states.py), we used a modified binary search approach to obtain the closest gene. In this case, we were not simply looking for the index of an item in a list. Rather, the list contained gene ranges on the DNA, and given a SNP there might not even be a gene range that includes it. So, the script reduced the gene list to either the gene the SNP is on, or two genes of which it selected the closest one.

The Python packages we used include [pylab](wiki.scipy.org/PyLab) and [matplotlib.pyplot](www.matplotlib.org) for plotting. In some scripts we used [numpy](www.numpy.org) but we have not used this package extensively. For the classifier and clustering methods, we used the [sklearn](sklearn) package [15]. This package includes the MDS and TSNE algorithms, as well as the $k$-nearest-neighbour classifier algorithm.

We used R for generating the heat maps. Specifically, we used the [heatmap.2](gplots) function from the [gplots](gplots) package [16].
### Figure 6.1: Python scripts used.

<table>
<thead>
<tr>
<th>Filename</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>classifier.py</td>
<td>Runs for the classifier, computes prototypes, counts the number of correctly predicted points. Draws a random sample from the GWAS catalog. Given the states of the four SNPs of Section 3.2, computes the normalized pattern (first approach). The same, but computes the minimum distance pattern (second approach). Given BED files of SNPs or stress response genes, generates an extended BED file containing states for each SNP or gene using a state mapping. Reads the gene category data and gene association data from Section 2.1, selects stress response genes, and outputs the gene names. Reads the GWAS catalog data, removes SNPs that have no chromosome or chromosome position attached or that have the &quot;CNV&quot; column set to &quot;Y&quot;, and outputs the filtered SNPs to a BED file. We used the lift over tool [12] to convert this BED file into version GRCh37, so that we did not have to do conversion when drawing random samples. Reads the 21 Alzheimer and control SNPs with states, and outputs the data in a format to use with heatmap.2 in R. The script also adds the cell type names. Script that other scripts use to generate histograms. We extracted this script for reuse purposes: e.g. it has logic for reading data files and it adds the legend. Scripts starting with hist_.py used histogram_opslaan.py to generate various histograms. Reads the file with IGAP SNPs, and selects SNPs as defined in Section 2.1. Generates MDS and TSNE plots for both only Alzheimer SNPs, and Alzheimer and control SNPs. Generates the step function plot related to the the normalized heat map pattern. The same, but for the minimum distance pattern. Reads the dbSNP data and outputs a BED file of the SNPs. Reads two groups of SNPs with states, and outputs one of the distance functions among and between groups, in brain cell types and in all cell types. Reads the gene names of stress response genes as output by find_stress_genes.py, and outputs a BED file of the gene ranges. Draws a random sample from the dbSNP data. Transforms state counts of IGAP SNPs, stress response genes and random genes into separate files. Transforms data of SNPs and states so that the histogram plotting function could use them. Draws a random sample from stress and random genes. Test script for TSNE. We noted from this script that TSNE shows clustering more clearly when many points are used, see Section 3.5. Computes the 15 by 15 matrix containing state transition counts within SNPs. Computes the difference of the matrices for the plot in Figure 3.11.</td>
</tr>
<tr>
<td>controle_snps_maken.py</td>
<td>Draw a random sample from the GWAS catalog.</td>
</tr>
<tr>
<td>find_patroon.py</td>
<td>Given the states of the four SNPs of Section 3.2, computes the normalized pattern (first approach).</td>
</tr>
<tr>
<td>find_patroon_4.py</td>
<td>The same, but computes the minimum distance pattern (second approach).</td>
</tr>
<tr>
<td>find_states.py</td>
<td>Generates an extended BED file containing states for each SNP or gene using a state mapping.</td>
</tr>
<tr>
<td>find_stress_genes.py</td>
<td>Reads the gene category data and gene association data from Section 2.1, selects stress response genes, and outputs the gene names.</td>
</tr>
<tr>
<td>gwascatalog_naar_bed.py</td>
<td>Reads the GWAS catalog data, removes SNPs that have no chromosome or chromosome position attached or that have the &quot;CNV&quot; column set to &quot;Y&quot;, and outputs the filtered SNPs to a BED file. We used the lift over tool [12] to convert this BED file into version GRCh37, so that we did not have to do conversion when drawing random samples.</td>
</tr>
<tr>
<td>heatmap.py</td>
<td>Reads the 21 Alzheimer and control SNPs with states, and outputs the data in a format to use with heatmap.2 in R. The script also adds the cell type names.</td>
</tr>
<tr>
<td>histogram_opslaan.py</td>
<td>Script that other scripts use to generate histograms. We extracted this script for reuse purposes: e.g. it has logic for reading data files and it adds the legend. Scripts starting with hist_.py used histogram_opslaan.py to generate various histograms.</td>
</tr>
<tr>
<td>hist_*.py</td>
<td>Reads the file with IGAP SNPs, and selects SNPs as defined in Section 2.1.</td>
</tr>
<tr>
<td>igap.py</td>
<td>Generates MDS and TSNE plots for both only Alzheimer SNPs, and Alzheimer and control SNPs.</td>
</tr>
<tr>
<td>mds_alzheimer_controle.py</td>
<td>Generates the step function plot related to the the normalized heat map pattern.</td>
</tr>
<tr>
<td>patroon_ranking.py</td>
<td>The same, but for the minimum distance pattern.</td>
</tr>
<tr>
<td>patroon_ranking_4.py</td>
<td>Reads the dbSNP data and outputs a BED file of the SNPs.</td>
</tr>
<tr>
<td>process_picked_vcf.py</td>
<td>Reads two groups of SNPs with states, and outputs one of the distance functions among and between groups, in brain cell types and in all cell types.</td>
</tr>
<tr>
<td>process_snps.py</td>
<td>Reads the gene names of stress response genes as output by find_stress_genes.py, and outputs a BED file of the gene ranges.</td>
</tr>
<tr>
<td>process_stress_genes.py</td>
<td>Draws a random sample from the dbSNP data.</td>
</tr>
<tr>
<td>read_vcf.py</td>
<td>Transforms state counts of IGAP SNPs, stress response genes and random genes into separate files.</td>
</tr>
<tr>
<td>state_count_histogram_stress_random_igap.py</td>
<td>Transforms data of SNPs and states so that the histogram plotting function could use them.</td>
</tr>
<tr>
<td>states_histogram.py</td>
<td>Draws a random sample from stress and random genes.</td>
</tr>
<tr>
<td>stress_random_682.py</td>
<td>Test script for TSNE. We noted from this script that TSNE shows clustering more clearly when many points are used, see Section 3.5.</td>
</tr>
<tr>
<td>tsne.py</td>
<td>Computes the 15 by 15 matrix containing state transition counts within SNPs.</td>
</tr>
<tr>
<td>write_matrix_K.py</td>
<td>Computes the difference of the matrices for the plot in Figure 3.11.</td>
</tr>
<tr>
<td>write_matrix_K_verschil.py</td>
<td></td>
</tr>
</tbody>
</table>
Glossary

**chromosome** A chromosome contains part of the DNA of an individual. A human has 25 chromosomes that together compose his/her complete DNA.

**DNA** DNA contains the genetic information of an individual and is contained in each cell. It is encoded by a string of nucleotides in each chromosome.

**epigenetic state** An epigenetic state is one of the fifteen predicates as defined in [1]. Each state can be represented as a number in the range 1 – 15, or as a mnemonic as defined in the article.

**epigenome** The epigenome defines e.g. the spatial structure of DNA, i.e., how DNA is wrapped in a cell.

**genome** The genome is the full DNA of an individual, in terms of the specific strings of nucleotides.

**GWAS** In a genome-wide association study (GWAS), one tries to seek for SNPs that are related to e.g. a particular disease, such as Alzheimer.

**linkage dependency** Linkage happens when two SNPs are located close to each other. Linkage makes it likely for an individual to possess both or neither of the SNPs. The SNPs thus have a dependency between them.

**nucleotide** A nucleotide is the unit using which DNA is encoded. It can be modelled as one of the letters A, C, G, T.

**reference genome** A reference genome is a genome that is published for comparison to other genomes. One can e.g. check a specific genome for mutations with respect to it.

**SNP** A single-nucleotide polymorphism (SNP, pronounced “snip”) is a mutation on the DNA that affects a single nucleotide.
Maximum flow clustering

Another idea we had for clustering was applying a maximum flow algorithm to determine a minimum cut that might split the groups in two parts. Here, SNPs are modelled by nodes, and there is an edge from a node to each of its $k$ nearest neighbours, measured using distance Distribution. Long distances between nodes are a candidate for a cut, so the edge capacities should be, e.g., inversely proportional to the distance.

However, we also need a source and a sink. We do not yet know what the cut is, so it is natural to connect the source and sink with every node. Let us say we give the edges to the source and the sink capacity $c$. Then any cut has capacity at least $nc$ if there are $n$ SNP nodes.

Indeed, each SNP node is either in the same partition as the source or as the sink. Hence, for each node an edge with capacity $c$ is in the cut. (Namely, to the sink if the node is in the same partition as the source, and to the source otherwise.) But now the source versus the other nodes is a minimum cut since its capacity is exactly $nc$. Therefore, this approach did not yield useful information with respect to clustering.
We have moved some histograms to this appendix so as not to pollute the main text.

Figure B.1: Histogram of Hamming1 distances among the three combinations of data sets (D-gap682) and (D_dbsnp682) (mapping s_10k). The left histogram considers all cell types whereas the right histogram considers brain cells only.
Figure B.2: Histogram of Hamming1 distances among the three combinations of data sets (D_igap682) and (D_dbsnp682) (mapping s_promotor). The left histogram considers all cell types whereas the right histogram considers brain cells only.

Figure B.3: Histogram of Hamming2 distances among the three combinations of data sets (D_igap682) and (D_dbsnp682) (mapping s_gene). The left histogram considers all cell types whereas the right histogram considers brain cells only.
Figure B.4: Histogram of Hamming3 distances among the three combinations of data sets (D_gap682) and (D_dbsnp682) (mapping s_gene). The left histogram considers all cell types whereas the right histogram considers brain cells only.

Figure B.5: Histogram of Hamming4 distances among the three combinations of data sets (D_gap682) and (D_dbsnp682) (mapping s_gene). The left histogram considers all cell types whereas the right histogram considers brain cells only.
Epigenetic data analysis on Alzheimer’s disease
VUmc
Presentation date: 30 June 2015
Description:

Mutations in one’s DNA can influence the risk for developing Alzheimer’s disease (Alzheimer). DNA consists of long strings of nucleotides, which together define the genome. However, next to these nucleotides, also the way in which DNA wraps in a cell affects the function of a cell. The information describing the DNA wrapping state is called the epigenome. A recent paper by Kundaje et al. [1] has provided, among other things, an epigenetic label (state) to each nucleotide.

One type of mutations on the DNA affects a single nucleotide; these are called SNPs. Some SNPs are related to Alzheimer whereas others are related to other diseases or traits such as hair color. In this report, we have investigated if Alzheimer SNPs are related to specific epigenetic states. Specifically, we searched for patterns to distinguish Alzheimer SNPs from other SNPs. For this, we explored numerous similarity approaches, such as differences in state counts, clustering methods, and a classifier.

Up till now, we did not find an obvious difference in epigenetic patterns between Alzheimer SNPs and control SNPs.

Member of the project team:
Name: Pim van den Bogaerdt
Role & Contribution: This project was entirely performed by Pim van den Bogaerdt - there were no other students who contributed to this project.

Client, Coach:
Name and affiliation of the client: Dr. Henne Holstege, VUmc
Name and affiliation of the project coaches: Prof. dr. ir. Marcel Reinders, Pattern Recognition & Bioinformatics Group, TU Delft; Dr. Marc Hulsman, Pattern Recognition & Bioinformatics Group, TU Delft

Contacts:
Pim van den Bogaerdt, E: pimvandenbogaerdt@hotmail.com

The final report for this project can be found at: http://repository.tudelft.nl
Original project description

The original project description was published on BEPsys [17], and is cited below. In the first meeting with my supervisors, we reformulated the problem to one involving epigenetic states. Our final research question was, as stated in the introduction: **Is there a difference in epigenetic states between SNPs related to Alzheimer and control SNPs?**

**WHY DO SOME PEOPLE DEVELOP DEMENTIA AT THE AGE OF 70 WHILE OTHERS REACH EXTREME AGES WITHOUT ANY SYMPTOMS?** Due to the increasing human longevity, the number of people confronted with the devastating effects of dementia increases each year. In spite of great efforts, there is currently no cure that prevents or delays dementia development. A successful cure should intervene in aberrant biochemical mechanisms that lead to cognitive decline. Detection of aberrant processes requires the knowledge of the biochemical mechanisms underlying long term retained cognitive health. Moreover, individuals who reach extreme ages without developing any signs of dementia may express molecular characteristics that protect against neuronal cell death. Heredity plays an important role in protection against dementia, since extreme old age without dementia often occurs within families. Therefore, the key to detect elements that protect against AD may be found in the genome. To detect alternative molecular processes that protect against neurodegeneration, we have initiated the ‘100-Plus Study’, (http://www.alzheimercentrum.nl/uploads/pdf/100plus-Informatiefolder.pdf) which aims to collect DNA samples and matching brain tissues from cognitively healthy centenarians. This unique cohort of DNA and brain tissues will allow us to detect genomic variants that are unique for cognitively healthy centenarians but that do not occur in genomes of patients suffering from Alzheimer Disease. We wish to compare the genetic information (exomes) from healthy centenarians with the genetic information of AD patients. We will search for genomic variants that uniquely recur in the genomes of centenarians and not in the genomes of AD patients and vice versa. Next we will predict the effect of the variant on the protein function and detect the molecular mechanisms that may be differentially active in cognitively healthy centenarians on the one hand and AD patients on the other. **BACHELOR PROJECT: VISUAL DATAMINING OF GENETIC DATA.**

There are currently many databases that list disease associated genomic variants (ClinVar, HGMD), that predict the effect of a detected variant on protein function (SIFT, Polyphen, Annovar), or on pathways (KEGG, Ingenuity). However, despite all these different databases it is difficult to determine the possible impact of a set of genetic variants on molecular pathway function. In this project, we ask the students to integrate the information in these databases such that we can do an integrated analysis of a set of detected genomic variants in one person. The integrated analysis should give an overview of the variant effects on different levels of cellular regulation. It further should enable the interactive exploration of the detected genetic variants. To facilitate the access to the databases you will make use of Ibidas, a powerful query language developed within the bioinformatics group. The way to visualize the different effects will be done in close collaboration with molecular biologists of the VUMC/Alzheimer Centre. Only a limited background in molecular biology (highschool) is expected.
Bibliography


